PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire

Hillary Koch[1], **Dmytro Starenki**[2], **Sara J. Cooper**[2], **Richard M. Myers**[2], **Qunhua Li**[1]*

1 Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America

* qunhua.li@psu.edu

Abstract

Sequencing of the T cell receptor (TCR) repertoire is a powerful tool for deeper study of immune response, but the unique structure of this type of data makes its meaningful quantification challenging. We introduce a new method, the Gamma-GPD spliced threshold model, to address this difficulty. This biologically interpretable model captures the distribution of the TCR repertoire, demonstrates stability across varying sequencing depths, and permits comparative analysis across any number of sampled individuals. We apply our method to several datasets and obtain insights regarding the differentiating features in the T cell receptor repertoire among sampled individuals across conditions. We have implemented our method in the open-source R package powerTCR.

Author summary

A more detailed understanding of the immune response can unlock critical information concerning diagnosis and treatment of disease. Here, in particular, we study T cells through T cell receptor sequencing, as T cells play a vital role in immune response. One important feature of T cell receptor sequencing data is the frequencies of each receptor in a given sample. These frequencies harbor global information about the landscape of the immune response. We introduce a flexible method that extracts this information by modeling the distribution of these frequencies, and show that it can be used to quantify differences in samples from individuals of different biological conditions.

This is a *PLoS Computational Biology* Methods paper.

Introduction

Recent advances in high-throughput sequencing of the T cell receptor (TCR) repertoire provide a new, detailed characterization of the immune system. T cells, each displaying a unique

Citation: Koch H, Starenki D, Cooper SJ, Myers RM, Li Q (2018) powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire. PLoS Comput Biol 14(11): e1006571. https://doi.org/10.1371/journal.pcbi.1006571

Editor: Bjoern Peters, La Jolla Institute for Allergy and Immunology, UNITED STATES

Received: May 14, 2018

Accepted: October 19, 2018

Published: November 28, 2018

Data Availability Statement: The murine breast cancer data and Sarcoidosis patient data analyzed in this study are available at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession numbers GSE119670 and GSE100378, respectively. The glioblastoma and T1D data analyzed in this study are available in the immunoSEQ Analyzer database (https://clients.adaptivebiotech.com) at links https://clients.adaptivebiotech.com/pub/peakman-2017-naturecommunications and https://clients.adaptivebiotech.com/pub/Prins-2016-

Cancerimmunolres, respectively. Our method is implemented in the R package powerTCR, freely available on BioConductor (https://bioconductor.org/packages/devel/bioc/html/powerTCR.html) under an Artistic-2.0 license.

Funding: This work was supported by NIH training grant T32 GM102057 (CBIOS training program to The Pennsylvania State University) (to HK), by NIH grants R01GM109453 (to QL), and by funding from the HudsonAlpha Institute and the Breast Cancer Research Foundation of Alabama (to DS, SJC, and RMM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

TCR, are capable of responding to presented antigens and initiating an adaptive immune response. An immune response is described by rapid proliferation of T cell clonotypes whose TCRs are specific to the antigen. In humans, it is estimated that the body is capable of producing more than 10^{18} different TCRs [1, 2], where high diversity of the TCR repertoire implies a greater range of pathogens that can be fought off. A variety of studies have been published demonstrating the value in characterizing this immune response for purposes such as describing tumor cell origin [3] and predicting response to cancer therapy and infection [4]. The applications of TCR sequencing are many, but this type of data presents new needs for analysis techniques not met by existing tools for other kinds of genomic experiments.

Several groups have identified that the distribution of larger clone sizes in a sample can be approximated by a power law [5–8], which means that the number of clones of a given size decays approximately as a power of the clone size. This heavy-tailed distribution comes as a consequence of extensively proliferated clones actively participating in an ongoing immune response. More recent work has aimed to quantify statistically the diversity of the TCR repertoire, initially through the use of various estimators borrowed from ecology, such as species richness, Shannon entropy [9], and clonality. These estimators are known to be highly sensitive to sample size and missing observations. Given that the TCR repertoire is mostly populated by rare clonotypes, many of the clonotypes in the system are absent from any one sample. This presents a challenge to many of the ecological estimators. Model-based approaches to approximating the clone size distribution have also been proposed, with the goal of providing added stability and consequently more statistical power. Some examples are the Poisson-lognormal model [10], Poisson mixture models [11, 12], and a heuristic ensemble method [13]; however, these models lack a biologically meaningful interpretation, and further do not sufficiently account for the power law-like nature of the data. That is, power law distributions are heavier-tailed than the Poisson or even the lognormal distribution, leading to systematic bias in the model fit.

Previous research has also identified the imperfectness of the power law behavior for the clone size distribution below some clone size threshold [7, 8]. To handle the imperfectness, [7, 8] proposed to model large clones above the threshold using a type-I Pareto distribution, which is a member of the power-law distribution family, and omitted the clones with frequency below that threshold. The threshold is either user-specified or determined from the data based on a goodness-of-fit measure. Indeed, this model has certain biological basis. Through a stochastic differential equations setup that models the birth, death, selection, and antigen-recognition of cells active in the immune system, Desponds et al. [8] showed that the upper tail of the clone size distribution at equilibrium approximately follows a type-I Pareto distribution (Fig 1A). Unlike the Poisson and lognormal models, parameters in this model are related to relevant actors in the immune response, and can reveal certain biological insights into immune response, such as average T cell lifetime [8]. Yet, the resulting model excludes all clones below a certain frequency threshold.

However, even small clones may provide information; for example, Desponds et al. [8] indicated that the generation of new T cells affects the landscape of smaller clone sizes. Other studies have shown that low-frequency clones may support a diverse immune system and present a potential to mobilize against antigens, as in some cases having a clone size distribution highly dominated by a few clones has been correlated with unfavorable clinical outcome [14, 15]. With this in mind, we sought a means to exhaust all available data and consider modeling the complete clone size distribution.

To address this question, we propose a novel statistical tool, called powerTCR, to characterize the full distribution of the TCR repertoire. Our method models large clones that are above the threshold, where the power law begins, using the generalized Pareto distribution (GPD), which contains the type-I Pareto distribution as a special case, but provides a more flexible fit. It also models the small clones below the threshold using a truncated Gamma distribution. It determines the threshold in a data-driven manner simultaneously with the characterization of clone size distribution. Our final model contains parameters that are analogous to those found in the type-I Pareto model of Desponds et al. [8], relating our model to the biological interpretation of the dynamics of the immune system. Altogether, this allows our model to more accurately describe the shape of the clone size distribution for both large and small clones. Such a model is well suited for providing a global view of the state of the immune repertoire. It can also be employed to perform comparative analysis of healthy and compromised individuals to identify descriptors of strengths and deficiencies in the immune system.

Results

The discrete Gamma-GPD spliced threshold model

Our goal is to model the clone size distribution of a sample immune repertoire. Fig 1A shows a typical distribution plotted using the repertoire of a Sarcoidosis patient in [16]. If the data are truly Pareto distributed, this plot would appear linear [17, 18]. However, noting the linear behavior is only true for the far upper tail of the data, this suggests that these data are a departure from the Pareto distribution. This imperfect power law implicates the use of a heavy-tailed distribution above some threshold and a lighter-tailed distribution below that threshold. Here, we model the tail part with a GPD. The GPD, introduced by [19], is a classical distribution typically used to model the values in the upper tail of a dataset. This formulation results in a distribution with density

$$f(x) = \frac{1}{\sigma}\left(1 + \xi\frac{x-u}{\sigma}\right)^{-(1/\xi+1)}, \qquad (1)$$

where $u \in (-\infty, \infty)$ is a threshold that typically needs to be prespecified, $\sigma \in (0, \infty)$ is a scale parameter, and $\xi \in (-\infty, \infty)$ is a shape parameter. The GPD has support $x \geq u$ when $\xi \geq 0$

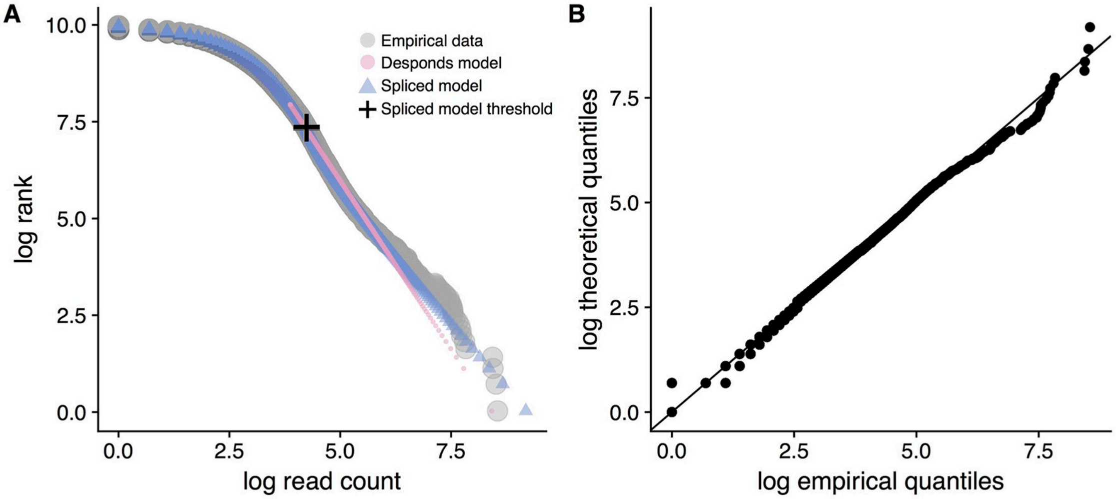

Fig 1. Examining the power law behavior of a sample clone size distribution. A: A sample repertoire (gray) of a Sarcoidosis patient and the fitted curves based on the Desponds method (pink) and our method (blue). The Desponds method only fits the data above the threshold it estimates (approximately 15% of the data). Our method fits all of the data. The cross marks the threshold estimated by our method. A complete collection of plots for every dataset analyzed are in S1 Text. B: The QQ-plot of the theoretical fit using our model against the empirical data for the same dataset in A shows that our model can achieve a good fit.

https://doi.org/10.1371/journal.pcbi.1006571.g001

and $u \leq x \leq u - \sigma/\xi$ when $\xi < 0$. We model the bulk part with a Gamma distribution with the upper tail truncated at the threshold. The Gamma distribution has a flexible shape and can fit many different clone size distributions. The threshold and the parameters in the two distributions are estimated from the data simultaneously. This setup, where data above and below an unknown threshold are drawn from the "bulk" and "tail" distributions respectively, falls into a class of models called spliced threshold models. The typical motivation for the model is the belief that the data above and below the threshold are driven by different underlying processes. We refer the interested reader to [20] for a thorough review of the general spliced threshold model, and its applications in fields such as insurance, hydrology, and finance.

Denote the proportion of data above the threshold u as ϕ. Let the bulk model distribution function be $H_c(x|\boldsymbol{\theta}_b)$ and the tail model distribution function be $G_c(x|\boldsymbol{\theta}_t)$, where subscripts b and t denote the bulk and tail model parameter vectors, respectively. Then the distribution function of the model is given by

$$F_c(x) = \begin{cases} (1-\phi)\dfrac{H_c(x|\boldsymbol{\theta}_b)}{H_c(u|\boldsymbol{\theta}_b)} & \text{for } x \leq u \\ 1-\phi+\phi G_c(x|\boldsymbol{\theta}_t, u) & \text{for } x > u \end{cases} \tag{2}$$

with corresponding density

$$f_c(x) = \begin{cases} (1-\phi)\dfrac{h_c(x|\boldsymbol{\theta}_b)}{H_c(u|\boldsymbol{\theta}_b)} & \text{for } x \leq u \\ \phi g_c(x|\boldsymbol{\theta}_t, u) & \text{for } x > u \end{cases}. \tag{3}$$

Because the clone size distribution is count data that typically exhibit numerous ties in the less frequently observed clonotypes, it is appropriate to treat this as a discrete problem. We modify the model in order to account for any quantized or censored data. Let ψ and Ψ be the density and distribution function of a continuous distribution, and let d be the interval length at which the data are censored. We obtain a quantized analog of ψ by letting

$$\Pr(X = x) = \Psi(x+d) - \Psi(x), \quad x \in k \cdot d, \quad k \in \mathbb{Z}.$$

This results in a discrete model with distribution function

$$F(x) = \begin{cases} (1-\phi)\dfrac{H(x|\boldsymbol{\theta}_b)}{H(u-d|\boldsymbol{\theta}_b)} & \text{for } x \leq u-d \\ 1-\phi+\phi G(x|\boldsymbol{\theta}_t, u) & \text{for } x \geq u \end{cases} \tag{4}$$

and corresponding probability mass function

$$f(x) = \begin{cases} (1-\phi)\dfrac{h(x|\boldsymbol{\theta}_b)}{H(u-d|\boldsymbol{\theta}_b)} & \text{for } x \leq u-d \\ \phi g(x|\boldsymbol{\theta}_t, u) & \text{for } x \geq u \end{cases}. \tag{5}$$

where $h(x|\boldsymbol{\theta}_b) \sim$ discrete Gamma(α, β), $g(x|\boldsymbol{\theta}_t) \sim$ discrete GPD(u, σ, ξ), and $d = 1$, which specifies that we model integer data (see Methods for the functional form of the discrete Gamma distribution and the discrete GPD). This discretization step turns out to be important for accurate estimation in our scenario. See S2 Text for a comparison between the performance of the discrete and continuous models in settings resembling true clone size distributions.

Biological interpretation of the model

The relationship between the discrete Gamma-GPD spliced threshold model and the type-I Pareto model in Desponds et al. [8], hereafter referred to as the Desponds et al. model, allows us to draw connections between some of our parameters and the dynamics of immune response underpinning their approach. First, results from [8] show that the threshold at which the power law begins is indicative of the point over which a clone's large size can be attributed to active immune response, as opposed to noise in the body that arises from processes such as self-recognition. The threshold fitted from the data provides an objective way to narrow down which clonotypes from a sample repertoire should be interrogated further. This notion is convenient for studying factors such as CDR3 (complementarity-determining region 3) amino acid motifs or specific V, D, and J genes important for combating certain antigens, which are typically determined based on a heuristic abundance cutoff. For example, [21] studies the 1,000 most abundant CDR3 amino acid motifs across all sampled peripheral blood mononuclear cell (PBMC) libraries, while [16] determines CDR3 amino acid motifs from clones that are present with 10 or more reads in a sampled repertoire. The threshold u estimated with our model, however, introduces a means to select motifs that does not rely on heuristics and automatically scales with sequencing depth.

Moreover, the shape parameter ξ of the GPD is inversely related to the shape parameter α_d used in the Desponds et al. model (see Methods). As explained by Desponds et al., a small α_d, i.e. a large ξ, implies increased average T cell lifetime and antigenic noise strength. They further show that antigenic noise strength grows as a consequence of a higher initial concentration of antigens and a higher rate at which new antigens are introduced. Interestingly, ξ also positively correlates with the familiar clonality estimator (1-Pielou's evenness [22]). Indeed, as ξ increases, the clone size distribution becomes heavier-tailed—that is, more skewed towards dominating clones. This trend is in line with that of the clonality estimator, which favors a more uniform clone size distribution as clonality approaches 0 and a distribution dominated by expanded clones as clonality approaches 1. To numerically validate this relationship, we simulated the data from our model and computed the clonality (see Methods). We observed a high correlation between clonality and ξ (Spearman's $\rho \approx 0.9$), confirming that ξ reflects the skewness towards dominating clones (see S3 Text).

It is worth noting that our model acquires a theoretical gain via the threshold stability property of the GPD [23]. That is, for any generalized Pareto distributed data, the shape parameter ξ remains constant regardless of changes in u. In our context, this means that at decreasing sequencing depths, though the threshold u would decrease due to fewer cells being sampled, the shape parameter ξ in principle would be stable against the variation in sequencing depth. We will demonstrate this gain in stability on a murine tumor dataset. See Methods for our extension of the threshold stability property to the case of the discrete GPD.

In the following sections, we inspect four different datasets using our model. We compare our results to results from the Desponds et al. model to demonstrate the practical and theoretical benefits of our approach. We also make comparisons to results from the widely used richness, Shannon entropy, and clonality estimators. See Methods for information on computation of competing methods.

Discrimination between tumors from MHC-II positive and control murine breast cancers

The expression of major histocompatability complex II (MHC-II) proteins in tumors correlates with boosted anti-tumor immunity. As part of a study of how MHC-II expression impacts tumor progression and functional plasticity of T cells [24], the CDR3 of TCRβ-chains of tumor infiltrating lymphocytes (TILs) were sequenced from breast cancer tumor tissue from six BALB/c mice [25]. Three of the mice were grafted with MHC-II expressing tumor cells and three control animals received parental MHC-II-negative cells. Samples were collected at 21 days after the date of treatment. Table A in S4 Text summarizes the number of unique clonotypes observed and the total number of reads in each sample.

Our method quantifies response to treatment and differentiates treatment groups. We asked whether the MHC-II expression would cause a global change in the TCR repertoire of the host mice. We analyzed the TCR repertoire in the individuals' tumor tissue using our proposed model and the Desponds et al. model. Tables B and C in S4 Text show parameter estimates from our model and the Desponds et al. model, respectively. Results from our model support a claim that increased expression of MHC-II induces an increased rate of clonal expansion at the tumor site, as indicated by the uniformly larger estimates of the tail shape parameter ξ in the treatment group.

Because the TCR repertoire is often used as an indicator of an individual's immune potency, we next evaluated how well our method discriminates between samples from case and control groups. We compared our method to the Desponds et al. model and the richness, Shannon entropy, and clonality estimators. For our model, we computed the pairwise Jensen-Shannon distance (JSD) between the fitted distributions of each pair of samples, and then used it as the distance measure to cluster samples with hierarchical clustering (see Methods) using Ward's method. Clustering of the Desponds et al. model was done similarly by using a generalization of JSD to continuous distributions, trading the summation for an integral. For all ecological measures, we computed pairwise Euclidean distances between estimates which we then used for hierarchical clustering of the samples. Fig 2 shows that our approach clusters the data by treatment and control groups. This is contrasted against all other methods, which display an incorrect construction of the expected true relationships among the experimental subjects. In the case of the ecological estimators, poor clustering likely occurs because a one-number summary of repertoire diversity cannot capture intricacies in the data, and varying sequencing depths across samples may further bias results. In the case of the Desponds et al. model, on the other hand, poor clustering is likely a result of the lack of a robust fitting procedure and a less flexible model. Moreover, the exclusion of smaller clones reduces power to discriminate between samples. See S5 Text for a demonstration of reduced clustering performance when only using the tail part of our spliced model.

Our results are robust to variation in sequencing depth. One challenge with TCR experiments is that the sequencing depth directly affects the number of TCR variants that are discovered. It is not simple to obtain the same sequencing depth across individuals, and it is also undesirable to discard data in favor of maintaining an equal number of reads across TCR repertoire libraries. An ideal scheme for comparative analysis achieves stable classification across different sequencing depths. Thus, we compared the stability of our model against the

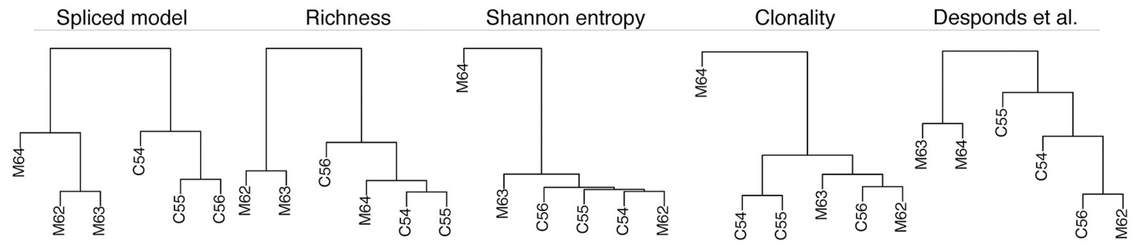

Fig 2. Hierarchical clustering of mouse tumor samples based on different quantifications of the TCR repertoire. The MHC-II-positive individuals are labeled with an 'M', while the control individuals are labeled with a 'C'.

https://doi.org/10.1371/journal.pcbi.1006571.g002

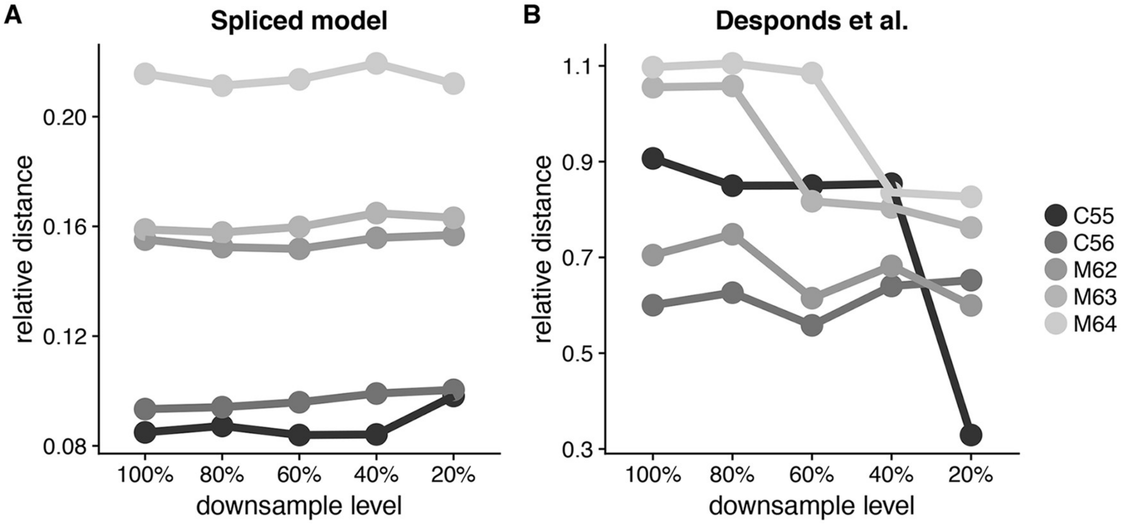

Fig 3. Robustness to variation in sequencing depth on the mouse tumor data. Relative JSD between sample C54 and the remaining samples across downsampling levels, using A: our model and B: the Desponds et al. model.

https://doi.org/10.1371/journal.pcbi.1006571.g003

Desponds et al. model. To do this, we randomly downsampled reads in the mouse data to 80, 60, 40, and 20% of total reads from the original samples. We fit our model and the Desponds et al. model in each case, then performed hierarchical clustering according to JSD. The clustering induced using our model was the same at every downsampling level, while clustering induced using the Desponds et al. model changed each time. S6 Text contains the dendrograms from both models at each downsample level, but Fig 3 summarizes this information by illustrating the relative JSD between sample C54 and the remaining samples across downsample levels. The spliced threshold model clearly maintains a similar trend across downsample levels, not to mention that the relationship inferred distinguishes between treatment and control groups. However, the downsampling study reveals that relative distances between sampled individuals observed using the Desponds et al. model are quite erratic.

The higher stability of our results is due in part to the threshold stability property of the GPD. It is also attributed to the fact that our method models the full clone size distribution. Doing so circumvents the threshold selection problem encountered when applying the Desponds et al. model, a recurrent issue when fitting models such as the type-I Pareto distribution [26]. For their approach, Desponds et al. select the threshold that minimizes the Kolmogorov-Smirnov statistic [27], a commonly used goodness-of-fit strategy for fitting a power law distribution [28]. However, the results on this dataset show that this strategy does not always yield stable results across different sequencing depths.

Differentiation between subtypes of Sarcoidosis patients

Sarcoidosis is an inflammatory disease that typically is accompanied by an accumulation of activated CD4+ T cells in the lungs. A particularly acute form of Sarcoidosis, called Löfgren's syndrom (LS), occurs with additional, more severe symptoms. A known signature of LS is the bombardment of the lungs with CD4+ T cells, which is expected to significantly alter the entire landscape of the TCR repertoire. We applied our method to TCR repertoire data of LS and non-LS Sarcoidosis patients [29], originally described in [16]. In this study, bronchoscopy with the bronchoalveolar lavage was performed on a cohort of 9 LS and 4 non-LS individuals and prepared for TCR α– and β–chain sequencing.

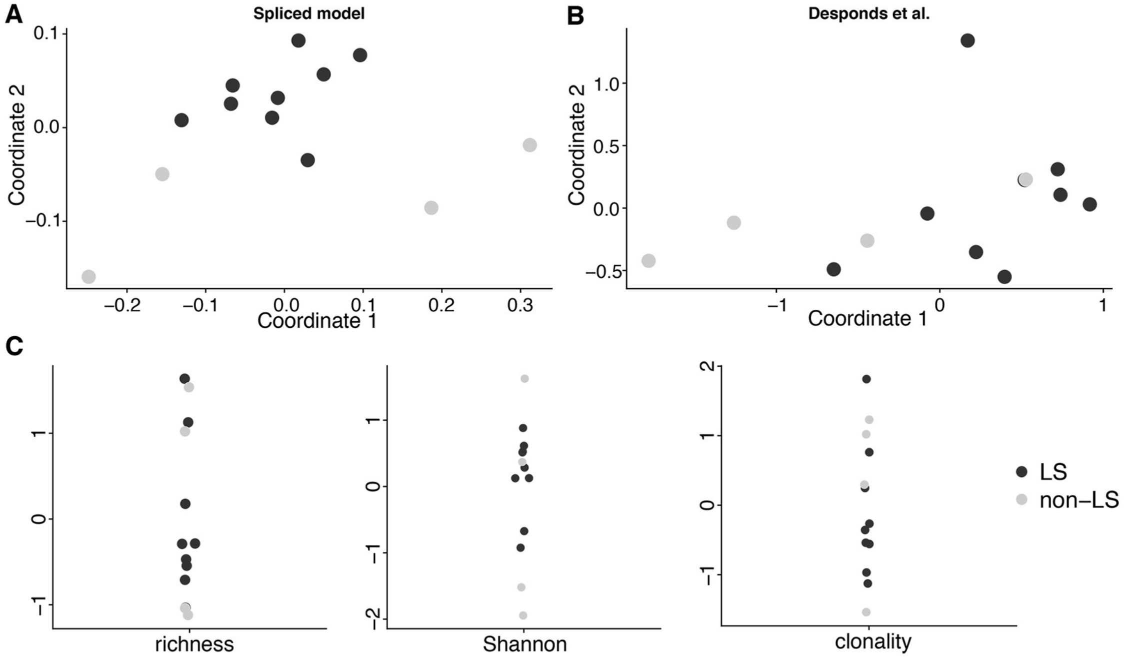

Fig 4. Differentiation of TCR repertoires between LS and non-LS Sarcoidosis patients. A–B: Multidimensional scaling representation of distances between TCR repertoires computed with JSD for the spliced threshold model and the Desponds et al. model. C: Centered and scaled estimates of richness, Shannon entropy, and clonality.

https://doi.org/10.1371/journal.pcbi.1006571.g004

We compared the TCR distribution between LS and non-LS Sarcoidosis patients using our method and the competing methods. In order to visualize closeness of samples, we generated a distance matrix using JSD between fitted distributions using our method and the Desponds et al. model. The estimated parameters are in Tables E and F in S4 Text respectively. We then applied non-metric multidimensional scaling (MDS) to the distance matrix and plotted the first two coordinates. For the ecological estimators, we simply plotted centered and scaled estimates. As shown in Fig 4A, results from our model cluster LS patients into a tight group distinct from non-LS patients, bolstering the claim that LS patients exhibit a signature immune response. On the other hand, competing methods fail to uncover any pattern (Fig 4B and 4C).

Relationship between the landscape of the clone size distribution and clinical outcome in glioblastoma patients

We applied our method to data collected during a clinical trial of 13 glioblastoma patients receiving autologous tumor lysate-pulsed dendritic cell (DC) vaccine therapy [30], first detailed in [31]. Three intradermal injections were administered to patients at biweekly intervals. TCRβ-chains from PBMC samples were sequenced for the patients prior to vaccinations and two weeks following the final injection. Patients were followed up with and their time to progression (TTP) and overall survival (OS) were recorded. TTP was defined as the time from the first DC vaccination until MRI-confirmed tumor progression. OS was calculated as the time from the first DC vaccination until the patient's death from any cause. We investigated whether current tools using TCRs sequenced only from blood samples indicate anything about patients' survival time and time to progression.

We first fit our model to the pre- and post-treatment samples. In both cases, we classified the patients into two groups using the hierarchical clustering based on our model, the

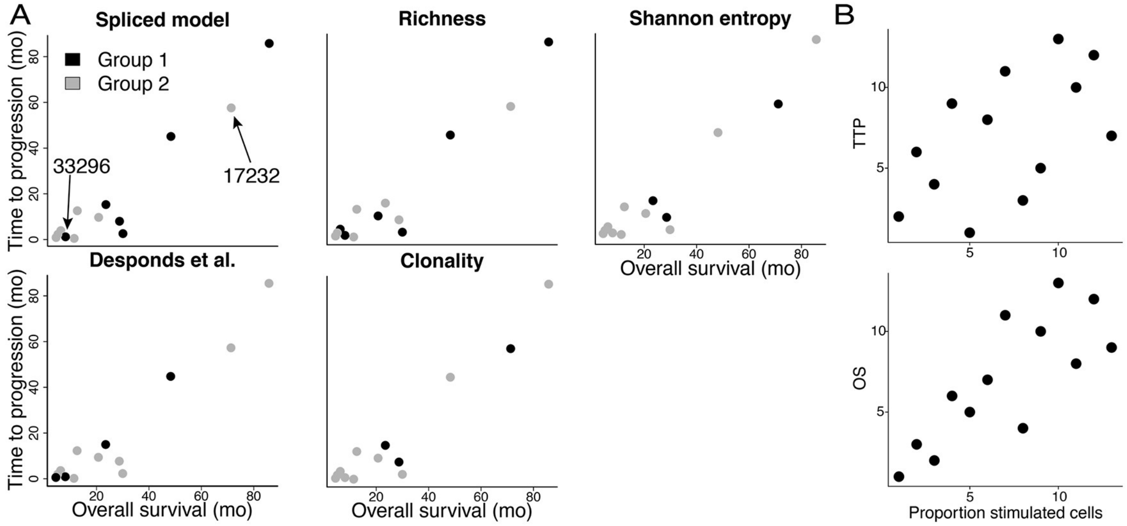

Fig 5. Association with time to progression (TTP) and overall survival (OS) time in glioblastoma patients. A: Patient classification based on TCR repertoire quantification and its relationship to TTP and OS. Patients are classified into two groups (black and gray) using hierarchical clustering according to the estimates of TCR repertoires from the post-treatment samples. Anomalous patients 33296 and 17232 are called out on the Spliced model plot. B: Relationship between the proportion of highly stimulated clones inferred by our method and TTP or OS.

https://doi.org/10.1371/journal.pcbi.1006571.g005

Desponds et al. model, and the richness, Shannon entropy, and clonality estimators. No clear grouping with respect to either TTP or OS could be observed from any clustering on the pre-treatment samples, whether by the model-based methods or the selected estimators (see S7 Text). However, among post-treatment samples, our method tends to cluster together patients with better clinical outcome (Fig 5A). This may indicate that the DC therapy alters the landscape of the TCR repertoire into a form that promotes favorable clinical outcome.

We do, however, cluster one patient (ID: 33296) with low TTP and OS in the group with overall higher TTP and OS. Interestingly, this misplaced patient had the lowest estimated TIL count and tumor/PBMC overlap of the entire cohort (S4 Text, Table G). Tumor/PBMC overlap was defined as the total number of reads of shared CDR3s normalized by total reads in the tumor and PBMC samples. Similarly, patient 17232 displayed among the best clinical outcome but clustered with lower-performing patients. Patient 17232 had the highest TIL count and level of tumor/PBMC overlap in the whole cohort. This information taken as a whole suggests that, while the clone size distribution found in blood may indicate something about a patient's response to treatment, it still does not guarantee that T cells will infiltrate the tumor, an important factor for clinical benefit [32]. S8 Text highlights the clone size distributions of these two patients against all others.

Notably, inferred thresholds (minimum $u = 4$, maximum $u = 6$) on this dataset are much lower than on other datasets. This is likely because this dataset contains less deeply-sequenced samples than the others, which consequently reduces the threshold.

Noting that clones with size at or above the estimated threshold are considered active participators in the immune response, we sought to investigate whether any relationship existed between clinical outcome and the proportion of more highly stimulated cells. We defined the proportion of highly stimulated cells to be the total number of reads at or above the threshold, normalized by the total number of reads in the entire repertoire (S3 Text, Table I). We found correlations between this measure and both TTP (Spearman's $\rho = 0.54$) and OS (Spearman's $\rho = 0.80$). Rank scatterplots for these correlations are in Fig 5B. The positive correlation we uncovered suggests that this statistic could be a useful tool to quantify the antigen-specificity of the sample.

Relationships among sorted CD4$^+$ T cell subtypes in individuals with type 1 diabetes and healthy donors

Risk factors for type 1 diabetes (T1D) are known to be heritable, yet genes alone are not sufficient explanation for drivers of the disease. Studies of monozygotic twins have revealed that, given one twin has T1D, the other will only have it at most half of the time [33]. The CD4$^+$ T cell is viewed as the initiator of T1D as dysregulation of CD4$^+$ antigen-recognition drives the autoimmune disease. Seeking out apparently non-heritable determinants of T1D, [34] conducted a deeper investigation of the CD4$^+$ T cell. Briefly, the authors obtained PBMCs from 14 volunteer healthy donors (HDs) and 14 recently diagnosed patients with T1D. The cells were sorted using flow cytometry into distinct T cell subsets (true naïve; TN, central memory, CM; regulatory, Treg; and stem cell-like memory, Tscm) and TCRβ-chains were sequenced. The authors conducted a thorough analysis, finding shorter CDR3 sequence lengths and lower overall repertoire diversity among patients with T1D. However, on a per-individual basis, the authors were unable to uncover a relationship between repertoire diversity and disease status. Since the the spliced threshold model provides a new means to probe this complex data, we applied our approach to complement the original analyses.

Trajectory in clone size distribution across CD4$^+$ T cell subtypes. While [34] considered pairwise correlations in diversity indices of different CD4$^+$ T cell subsets from the same individual, their analysis provides no visualization of the data as a whole. Our method for comparative analysis, combined with MDS, allows this to be done naturally. Like the Sarcoidosis patient analysis, we applied our method and competing methods, and visualized the results as before (Fig 6A–6C).

It is known that TN cells propagate into both Tscm and CM cells, Tscm cells also propagate in CM cells, and Treg cells generally originate separately in the thymus [35–37]. Our approach appears to support a gradual change in repertoire shape throughout this differentiation process, as a clear trajectory is revealed in the MDS representation of the JSD between fitted

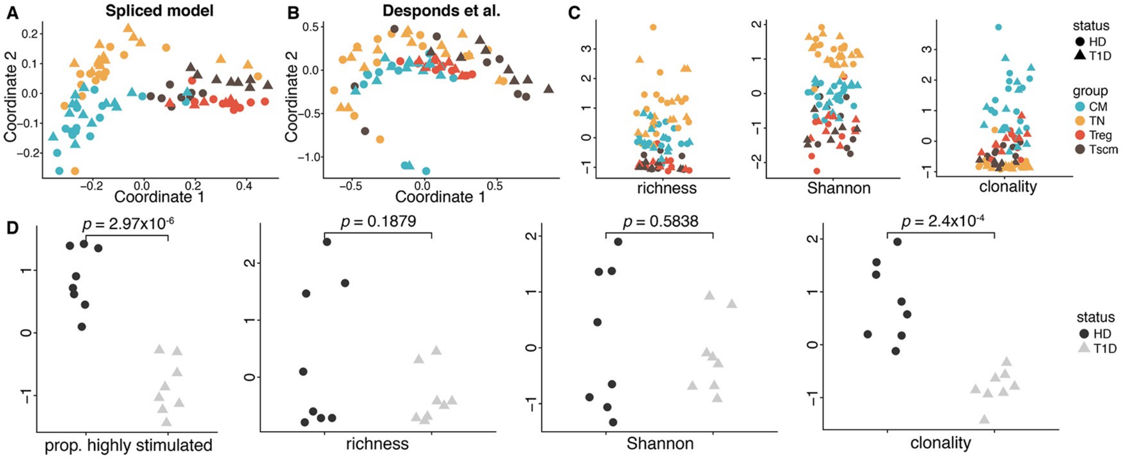

Fig 6. Trajectory of TCR repertoires among CD4$^+$ T cell subsets between T1D patients and HDs. A–B: MDS representations of JSD between TCR repertoires from the spliced threshold model and Desponds et al. model. C: Centered and scaled estimates of richness, Shannon entropy, and clonality for all CD4$^+$ T cell subsets. D: The proportion of highly stimulated clones derived from our method and centered and scaled diversity estimates in the Tscm subset.

https://doi.org/10.1371/journal.pcbi.1006571.g006

models in Fig 6A. On the other hand, the MDS representation obtained using the Desponds et al. model does not separate results by cell type or donor status. The ecological estimators recapitulate some results found in the original analysis [34], for example that TN are more diverse than other cells, but uncover nothing more.

Upon closer examination of the results from our model, we noticed that individuals from the T1D and HD groups are not well-separated in the cell types in Fig 6A, except in the Tscm subset. Tscm cells, identified *in vitro* in [35], are a stable, yet multipotent, subset of cells sustained via proliferation and turnover throughout the human lifetime and have been suggested to play a major role in establishing memory in immune response [35, 36]. More recently, Tscm cells have been implicated as a factor in development and treatment of autoimmune disorders such as acquired aplastic anemia [38] and systemic lupus erythematosis [39]. This motivated a deeper evaluation of the Tscm subset. Recalling that the proportion of highly stimulated clones, an estimator derived from our model, proved informative in our analysis of the TCR repertoires sequenced from PBMCs of individuals with glioblastoma, we decided to again apply it here to the Tscm subset alongside the ecological estimators.

As shown in Fig 6D, richness and Shannon entropy have difficulty differentiating the T1D and HD groups ($p = 0.1879$ and $p = 0.5838$ for a two-sided t-test, respectively), yet clonality and our measure uncover a clear split between the two groups ($p = 2.4 \times 10^{-4}$ and 2.97×10^{-6}, respectively). This indicates the potential for the Tscm TCR repertoire to be used as a biomarker for detecting the status of T1D, suggesting the relevance of Tscm cells in T1D pathogenesis.

Discussion

We have developed a model, the discrete Gamma-GPD spliced threshold model, and demonstrated its utility on several datasets. As shown in our analyses, several biologically relevant descriptive features can be obtained from our model. One is the tail shape parameter ξ, a measure of the weight of the upper tail of the clone size distribution, where a heavier tail of the fitted model implies a more dominated distribution of expanded clones. Another is the proportion of total reads at or above the estimated threshold, a possible measure of intensity of the immune response. The third is the estimated threshold, which is a useful guide to objectively identify CDR3 motifs for downstream analysis. This could involve denoting motifs as only those CDR3s found in TCRs with frequencies at or above the estimated threshold for a given sample, or it could mean studying TCR gene usage among that same group of clonotypes. Though the dynamics driving our model form a compelling argument for this interpretation of the threshold, we acknowledge that further biological validation on more datasets is still needed to confirm this.

Similar to other estimators, our model requires that a repertoire be adequately sampled. Without adequate sampling, the differentiating features between TCR repertoires will be masked [8], and the estimated model parameters will not be reliable. Given the immense diversity of the TCR repertoire, one should in general be cautious about using any method to make inference about a sample TCR repertoire when few cells are sequenced. With sufficient samples, though, the spliced threshold model provides the user a meaningful high-level view of the TCR repertoire.

The diversity of the TCR repertoire and its responsiveness to stimuli provide a high-dimensional biomarker for monitoring the immune system and its adaptivity. Robust assessment of the clone size distribution through TCR sequencing is important for understanding this diversity. The discrete Gamma-GPD spliced threshold model is a flexible model that effectively captures the shape of the clone size distribution. It is especially appropriate since the heavy-tailed

GPD is a good fit to model the highly expanded clones that dominate many TCR repertoire samples. The method also provides a means to comparatively analyze a collection of TCR repertoire samples while maintaining convenient theoretical properties and interpretations.

Compared with existing approaches, our method is more flexible, utilizes the full clone size distribution, is less sensitive to sequencing depth, and identifies the threshold in a data-driven manner. The parameters estimated from our method are biologically relevant and instructive to the dynamics of immune response. Our results on multiple datasets also show that the spliced threshold model is powerful in a range of scenarios for comparing TCR repertoires across samples, revealing potential trends in the landscapes of clone size distributions of affected immune systems.

Methods

Estimation

We use maximum likelihood to estimate the parameters of our model. First, we more explicitly specify the form of our distribution. Letting $x \sim \text{Gamma}(\alpha, \beta)$, we write the probability mass function of a discrete Gamma distribution as

$$h(x) = \frac{1}{\Gamma(\alpha)}[\gamma(\alpha, \beta(x+1)) - \gamma(\alpha, \beta x)]$$

for $\alpha > 0, \beta > 0, x \in \mathbb{Z}$, and where $\gamma(\alpha, \beta x)$ is the lower incomplete gamma function

$$\gamma(\alpha, \beta x) = \int_0^{\beta x} t^{\alpha-1} e^{-t} dt.$$

If $x \sim \text{GPD}(u, \sigma, \xi)$, we write the probability mass function of a discrete GPD as

$$g(x) = \left(1 + \xi\frac{x-u}{\sigma}\right)^{-1/\xi} - \left(1 + \xi\frac{x+1-u}{\sigma}\right)^{-1/\xi}$$

for $u \in (-\infty, \infty)$, $\sigma \in (0, \infty)$, and $\xi \in (-\infty, \infty)$. The discrete GPD has support $x \geq u$ when $\xi \geq 0$ and $u \leq x \leq u - \sigma/\xi$ when $\xi < 0$, where $x \in \mathbb{Z}$. In all analyses presented here, we make no assumptions on the sign of ξ, although empirically we tend to observe $\xi > 0$.

To proceed, we employ a profile likelihood approach. Let u be the threshold, $\boldsymbol{\theta}_b$ be the bulk parameter vector $\{\alpha, \beta\}$, $\boldsymbol{\theta}_t$ be the tail parameter vector $\{\sigma, \xi\}$, and $\boldsymbol{\theta}$ be the parameter vector $\{\boldsymbol{\theta}_b, \boldsymbol{\theta}_t\}$. Let also h and H be the density and distribution function of a discrete Gamma distribution, respectively, and let g be the density of a discrete GPD. Then the complete data likelihood is given by

$$L(\{\boldsymbol{\theta}, u\} \mid \mathbf{x}) = \prod_{i=1}^{n}\left[(1-\phi)\frac{h(x_i \mid \boldsymbol{\theta}_b)}{H(u-1 \mid \boldsymbol{\theta}_b)}\mathbb{1}(x_i \leq u-1) + \phi g(x_i \mid \boldsymbol{\theta}_t, u)\mathbb{1}(x_i \geq u)\right]$$

and the profile likelihood of the model at u is denoted as

$$L_p(u) = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{x}, u).$$

A grid search over a suitable range of thresholds $\mathbf{u}^\star = (u_1, \ldots, u_k)$ may be implemented to maximize the profile likelihood. In this study, we adopted an approach similar to those of [19] and [40], searching for thresholds at or above the 75% quantile of the sample. The estimated parameters are then

$$\begin{aligned}\hat{u} &= \arg\max_{u^\star \in \mathbf{u}^\star} L_p(u^\star),\\ \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid u = \hat{u}), \text{ and}\\ \hat{\phi} &= \frac{n_u}{n},\end{aligned}$$

where n is the total number of clones and n_u denotes the number of clones with size greater than or equal to the threshold.

Computation for competing methods

The Desponds et al. model was fit as previously described [8]. Briefly, the model has density

$$f(x) = \frac{\alpha_d u^{\alpha_d}}{x^{\alpha_d + 1}} \tag{6}$$

and distribution function

$$F(x) = 1 - \left(\frac{u}{x}\right)^{\alpha_d} \tag{7}$$

where $u > 0$ is the threshold and $\alpha_d > 0$ is a shape parameter. For each sample TCR repertoire, a grid of potential thresholds $\mathbf{u}^\star = (u_1, \ldots, u_k)$ was constructed by considering every unique clone size in the repertoire. Then, for each u_i, the shape parameter is estimated as

$$\hat{\alpha}_d = n_i \left[\sum_{j=1}^{n_i} \ln \frac{x_j}{u_i}\right]^{-1} \tag{8}$$

where n_i is the number of clones with size larger than the threshold u_i. Once this value is computed for every threshold in $\mathbf{u}^\star$, the threshold and corresponding $\hat{\alpha}$ were chosen to minimize the Kolmogorov-Smirnov statistic.

The ecological estimators [9, 22] were computed as follows. For a sample X, let $S(X)$ be the sample richness, defined as the number of unique clonotypes in X, and let p_i be the number of cells of clonotype i normalized by the total number of cells in the sample. Then, the Shannon entropy of X is

$$H(X) = -\sum_{i=1}^{S(X)} p_i \ln p_i \tag{9}$$

and the clonality of X is

$$C(X) = 1 - \frac{H(X)}{\ln S(X)}. \tag{10}$$

ξ is inversely related to the the shape parameter of the type-I Pareto distribution

The Desponds et al. model, which is a type-I Pareto distribution, and the "tail" part of our model, which is a GPD, are closely related. In fact, the GPD contains the type-I Pareto distribution as a special case. We can write the distribution function of y, where $y \sim \text{GPD}(u, \sigma, \xi)$, as

$$F(y) = 1 - \left(1 + \xi \frac{y-u}{\sigma}\right)^{-1/\xi}. \tag{11}$$

Now, let $x \sim \mathrm{GPD}\left(u, \frac{u}{\alpha_d}, \frac{1}{\alpha_d}\right)$. Then

$$\begin{aligned} F\left(x;\ u, \frac{u}{\alpha_d}, \frac{1}{\alpha_d}\right) &= 1 - \left[1 + \frac{1}{\alpha_d}\left(\frac{x-u}{u/\alpha_d}\right)\right]^{-\alpha_d} \\ &= 1 - \left[1 + \left(\frac{x-u}{u}\right)\right]^{-\alpha_d} \\ &= 1 - \left(\frac{u}{x}\right)^{\alpha_d} \end{aligned}$$

which is exactly the distribution function of a type-I Pareto distribution with threshold u and shape α_d (Eq 7). Of course, this exact relationship only holds when $\sigma = \frac{u}{\alpha_d}$. Nevertheless, α_d and ξ perform the same function in their respective distributions, adjusting the weight of the tail. This relationship always holds—a larger ξ (smaller α_d) implies a heavier-tailed distribution, while a smaller ξ (larger α_d) implies a lighter-tailed distribution.

Correlation between ξ and clonality

We conjecture that ξ, the shape parameter of the GPD, positively correlates with clonality. We numerically validated this claim using a simulated cohort of 48 clone size distributions. That is, we generated samples of $n = 20,000$ clonotypes, where our 48 parameter settings were derived from every combination of $\alpha \in \{3, 5, 10\}$, $\xi \in \{.25, .5, .75, 1.1\}$, and $\phi \in \{0.1, 0.15, 0.2, 0.25\}$. We chose $\beta = 0.15$, $\sigma = \frac{\sqrt{\alpha}}{\beta}$, and $u = \lfloor Q_{\alpha,\beta}(1 - \phi)\rfloor$ in each simulation, where $Q_{\alpha,\beta}$ is the quantile function of the Gamma distribution with mean $\frac{\alpha}{\beta}$. To adjust for the effect of sample size on clonality, we downsampled the simulated data so that each sample contained the same number of reads (415,989 total reads per sample). We computed the clonality of each simulated TCR repertoire on these adjusted datasets.

Comparative analysis of multiple samples using the spliced threshold model

The relationship between a pair of TCR repertoires can be elucidated by evaluating the distance between their fitted spliced threshold models. Several methods to compare densities are available. We propose measuring the distance between each pair of distributions using Jensen-Shannon distance (JSD) [41]. This metric is a symmetric and smoothed adaptation of the well-known Kullback-Leibler divergence that does not require the distributions under comparison to share the same support.

Given discrete distributions P and Q, the JSD between P and Q is

$$JSD(P, Q) = \sqrt{\frac{1}{2}\left[\sum_i \left(P_i \ \ln \frac{P_i}{M_i}\right) + \sum_i \left(Q_i \ \ln \frac{Q_i}{M_i}\right)\right]}, \tag{12}$$

where $M_i = \frac{1}{2}(P_i + Q_i)$. The resulting distances allow analysis and visualization via MDS or hierarchical clustering of the samples. Throughout our study, we use Ward's method for hierarchical clustering.

Threshold stability of the discrete GPD

The threshold stability property of the GPD is well-established [23]. Here, we show that the property also holds for the discrete GPD. Let $X \sim$ discrete GPD(u, σ, ξ) and denote its distribution function as F with F_c as its continuous analog. Then we can write

$$\begin{aligned}
P(X - u \leq x + 1 | X \geq u) &= \frac{P(u \leq X \leq x + u + 1)}{P(X \geq u)} \\
&= \frac{F(x + u + 1;\ u,\ \sigma,\ \xi) - F(u;\ u,\ \sigma,\ \xi)}{1 - F(u;\ u,\ \sigma,\ \xi)} \\
&= \frac{F_c(x + u + 2;\ u,\ \sigma,\ \xi) - F_c(u + 1;\ u,\ \sigma,\ \xi)}{1 - F_c(u + 1;\ u,\ \sigma,\ \xi)} \\
&= \frac{\left(1 + \frac{\xi}{\sigma}\right)^{-1/\xi} - \left(1 + \xi\frac{x + 2}{\sigma}\right)^{-1/\xi}}{\left(1 + \frac{\xi}{\sigma}\right)^{-1/\xi}} \\
&= 1 - \left(1 + \xi\frac{x + 1}{\sigma + \xi}\right)^{-1/\xi} \\
&= F_c(x + 1;\ 0,\ \sigma + \xi,\ \xi) \\
&= F(x;\ 0,\ \sigma + \xi,\ \xi).
\end{aligned}$$

This states that if $X \sim$ discrete GPD(u, σ, ξ), then $X - u \sim$ discrete GPD(0, $\sigma + \xi$, ξ). Or, for our application, consider a clone size distribution, where clones larger than some threshold u are distributed according to the discrete GPD. At decreasing sequencing depths, this estimated u decreases, implying naturally that the size a clone in the sample must achieve to be considered “expanded” decreases. Still, while u shrinks, the threshold stability property states that ξ remains constant.

Supporting information

S1 Text. Visualization of all model fits. For each real data sample, a plot analagous to Fig 1A is provided.
(PDF)

S2 Text. Simulation for comparing the accuracy of parameter estimation for the discrete spliced threshold model, the Desponds et al. model, and the continuous spliced threshold model. Details on simulation study comparing the discrete model and Desponds et al. model, and simulation study comparing the discrete and continuous models.
(PDF)

S3 Text. Correlation between ξ and clonality. Information on simulation study that finds strong positive relationship between ξ and the clonality estimator.
(PDF)

S4 Text. Summaries for the four TCR repertoire datasets. Tables summarizing number of unique clonotypes per sample, total reads sequenced per sample, and other patient-specific information for the Sarcoidosis and glioblastoma datasets. Additionally, tables containing

fitted model parameter estimates for both our model and the competing model, as well as ecological estimator values, computed for every sample.
(PDF)

S5 Text. Comparative analysis using the GPD. By comparing results from our full model to those from only our tail model, we observe empirically the gains from including the full clone size distribution.
(PDF)

S6 Text. Dendrograms for downsampling study. We downsampled mouse tumor data to 80, 60, 40, and 20% of total reads. We used JSD to compute pairwise distances between the samples for our model fits and the Desponds et al. model fits at each downsample level and did hierarchical clustering using Ward's method. The dendrograms for each model at each downsample level are presented here.
(PDF)

S7 Text. Clustering of pre- and post-treatment glioblastoma patient samples. Clustering dendrograms generated on pre- and post-treatment glioblastoma samples. Groupings presented for the post-treatment samples here correspond to the colored groupings in Fig 5.
(PDF)

S8 Text. Clone size distributions of glioblastoma patients. This plot calls out patients 33296 and 17232 from the glioblastoma patients. Patient 33296 incorrectly clustered with the individuals with favorable clinical outcome, while patient 17232 incorrectly clustered with individuals with unfavorable clinical outcome.
(PDF)

Author Contributions

Conceptualization: Hillary Koch, Dmytro Starenki, Sara J. Cooper, Qunhua Li.

Data curation: Hillary Koch, Dmytro Starenki.

Formal analysis: Hillary Koch.

Funding acquisition: Hillary Koch, Sara J. Cooper, Richard M. Myers, Qunhua Li.

Investigation: Hillary Koch, Qunhua Li.

Methodology: Hillary Koch, Qunhua Li.

Project administration: Qunhua Li.

Resources: Dmytro Starenki, Sara J. Cooper, Richard M. Myers.

Software: Hillary Koch.

Supervision: Qunhua Li.

Visualization: Hillary Koch, Qunhua Li.

Writing – original draft: Hillary Koch, Dmytro Starenki, Sara J. Cooper, Qunhua Li.

Writing – review & editing: Hillary Koch, Dmytro Starenki, Sara J. Cooper, Richard M. Myers, Qunhua Li.

References

1. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human *αβ* T cell receptor diversity. Science. 1999; 286(5441):958–961. https://doi.org/10.1126/science.286.5441.958 PMID: 10542151

2. Naylor K, Li G, Vallejo AN, Lee WW, Koetz K, Bryl E, et al. The influence of age on T cell generation and TCR diversity. J Immunol. 2005; 174(11):7446–7452. https://doi.org/10.4049/jimmunol.174.11.7446 PMID: 15905594

3. Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott LL, Elco CP, et al. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. Sci Transl Med. 2015; 7(308): 308ra158–308ra158. https://doi.org/10.1126/scitranslmed.aaa9122 PMID: 26446955

4. Neller M, Burrows J, Rist M, Miles J, Burrows S. High frequency herpesvirus-specific clonotypes in the human T cell repertoire can remain stable over decades with minimal turnover. J Virol. 2012; p. JVI–02180.

5. Burgos JD, Moreno-Tovar P. Zipf-scaling behavior in the immune system. Biosystems. 1996; 39(3): 227–232. https://doi.org/10.1016/0303-2647(96)01618-8 PMID: 8894123

6. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science. 2009; 324(5928):807–810. https://doi.org/10.1126/science.1170020 PMID: 19423829

7. Naumov YN, Naumova EN, Hogan KT, Selin LK, Gorski J. A fractal clonotype distribution in the CD8+ memory T cell repertoire could optimize potential for immune responses. J Immunol. 2003; 170(8): 3994–4001. https://doi.org/10.4049/jimmunol.170.8.3994 PMID: 12682227

8. Desponds J, Mora T, Walczak AM. Fluctuating fitness shapes the clone-size distribution of immune repertoires. P Natl Acad Sci U S A. 2016; 113(2):274–279. https://doi.org/10.1073/pnas.1512977112

9. Hill MO. Diversity and evenness: a unifying notation and its consequences. Ecology. 1973; 54(2): 427–432. https://doi.org/10.2307/1934352

10. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. J Theor Biol. 2011; 269(1):1–15. https://doi.org/10.1016/j.jtbi.2010.10.001 PMID: 20955715

11. Guindani M, Sepúlveda N, Paulino CD, Müller P. A Bayesian semiparametric approach for the differential analysis of sequence counts data. J R Stat Soc C-Appl. 2014; 63(3):385–404. https://doi.org/10.1111/rssc.12041

12. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. Nat Commun. 2016; 7. https://doi.org/10.1038/ncomms11881 PMID: 27302887

13. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, Darko S, et al. Quantification of HTLV-1 clonality and TCR diversity. PLoS Comput Biol. 2014; 10(6):e1003646. https://doi.org/10.1371/journal.pcbi.1003646 PMID: 24945836

14. Keane C, Gould C, Jones K, Hamm D, Talaulikar D, Ellis J, et al. The T-cell receptor repertoire influences the tumor microenvironment and is associated with survival in aggressive B-cell lymphoma. Clin Cancer Res. 2017; 23(7):1820–1828. https://doi.org/10.1158/1078-0432.CCR-16-1576 PMID: 27649554

15. Snyder A, Nathanson T, Funt SA, Ahuja A, Novik JB, Hellmann MD, et al. Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. PLoS Med. 2017; 14(5):e1002309. https://doi.org/10.1371/journal.pmed.1002309 PMID: 28552987

16. Mitchell AM, Kaiser Y, Falta MT, Munson DJ, Landry LG, Eklund A, et al. Shared *αβ* TCR Usage in Lungs of Sarcoidosis Patients with Löfgren's Syndrome. J Immunol. 2017; 199(7):2279–2290. https://doi.org/10.4049/jimmunol.1700570 PMID: 28827283

17. Newman ME. Power laws, Pareto distributions and Zipf's law. Contemp Phys. 2005; 46(5):323–351. https://doi.org/10.1080/00107510500052444

18. Cirillo P. Are your data really Pareto distributed? Physica A. 2013; 392(23):5947–5962.

19. Pickands J. Statistical inference using extreme order statistics. Ann Stat. 1975; p. 119–131.

20. Scarrott C. Univariate Extreme Value Mixture Modeling. In: Dey DK, Yan J, editors. Extreme Value Modeling and Risk Analysis: Methods and Applications. Boca Raton: CRC Press;. p. 41–67.

21. Sims JS, Grinshpun B, Feng Y, Ung TH, Neira JA, Samanamud JL, et al. Diversity and divergence of the glioma-infiltrating T-cell receptor repertoire. P Natl Acad Sci U S A. 2016; 113(25):E3529–E3537. https://doi.org/10.1073/pnas.1601012113

22. Pielou EC. An introduction to mathematical ecology. New York: Wiley-Inter-science; 1969.

23. Castillo E, Hadi AS. Fitting the generalized Pareto distribution to data. J Am Stat Assoc. 1997; 92(440): 1609–1620. https://doi.org/10.1080/01621459.1997.10473683

24. McCaw T, Li M, Starenki D, Cooper SJ, Meza-Perez S, Arend R, et al. The expression of class II major histocompatibility molecules on breast tumors delays T cell exhaustion, expands the T cell repertoire and slows tumor growth. Cancer Immunol Immunother. 2018. https://doi.org/10.1007/s00262-018-2262-5 PMID: 30334128

25. Gene Expression Omnibus;. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119670.

26. Scarrott C, MacDonald A. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT–Stat J. 2012; 10(1):33–60.

27. Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951; 46(253): 68–78. https://doi.org/10.1080/01621459.1951.10500769

28. Van Zyl JM. Application of the Kolmogorov–Smirnov Test to Estimate the Threshold When Estimating the Extreme Value Index. Commun Stat Simulat. 2011; 40(2):199–207. https://doi.org/10.1080/03610918.2010.533227

29. Gene Expression Omnibus;. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100378. Accessed 10 February 2018.

30. immunoSEQ Analyzer database;. Available from: https://clients.adaptivebiotech.com/. Accessed 17 November 2017.

31. Hsu MS, Sedighim S, Wang T, Antonios JP, Everson RG, Tucker AM, et al. TCR sequencing can identify and track glioma-infiltrating T cells after DC vaccination. Cancer Immunol Res. 2016; 4(5):412–418. https://doi.org/10.1158/2326-6066.CIR-15-0240 PMID: 26968205

32. Rosenberg SA, Spiess P, Lafreniere R. A new approach to the adoptive immunotherapy of cancer with tumor-infiltrating lymphocytes. Science. 1986; 233(4770):1318–1321. https://doi.org/10.1126/science.3489291 PMID: 3489291

33. Barnett AH and Eff C and Leslie R DG and Pyke DA Diabetes in identical twins. Diabetologia. 1981; 20(2):87–93. https://doi.org/10.1007/BF00262007 PMID: 7193616

34. Gomez-Tourino I and Kamra Y and Baptista R and Lorenc A and Peakman M T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. Nature Commun. 2017; 8(1): 1792. https://doi.org/10.1038/s41467-017-01925-2

35. Gattinoni L and Lugli E and Ji Y and Pos Z and Paulos CM and Quigley MF and Almeida JR and Gostick E and Yu Z and Carpenito C and Wang E and Douek DC and Price DA and June CH and Marincola FM and Roederer M and Restifo NP A human memory T cell subset with stem cell–like properties. Nature. 2011; 17(10):1290.

36. Ahmed R and Roger L and del Amo PC and Miners KL and Jones RE and Boelen L and Fali T and Elemans M and Zhang Y and Appay V and Baird DM and Asquith B and Price DA and Macallan DC and Ladell K Human stem cell-like memory T cells are maintained in a state of dynamic flux. Cell Rep. 2016; 17(11)2811–2818.

37. Kitagawa Y and Ohkura N and Sakaguchi S Molecular determinants of regulatory T cell development: the essential roles of epigenetic changes Front Immunol. 2013;(4)106. https://doi.org/10.3389/fimmu.2013.00106 PMID: 23675373

38. Hosokawa K and Muranski P and Feng X and Townsley DM and Liu B and Knickelbein J and Keyvanfar K and Dumitriu B and Ito S and Kajigaya S and Taylor JG VI and Kaplan MJ and Nussenblatt RB and Barrett AJ and O'Shea J and Young NS Memory stem T cells in autoimmune disease: high frequency of circulating CD8+ memory stem cells in acquired aplastic anemia. J Immunol. 2016;1501739.

39. Lee YJ and Park JA and Kwon H and Choi YS and Jung KC and Park SH and Lee EB Role of Stem Cell-Like Memory T Cells in Systemic Lupus Erythematosus. Arthritis Rheumatol. 2018. https://doi.org/10.1002/art.40524

40. Wong TST, Li WK. A threshold approach for peaks-over-threshold modeling using maximum product of spacings. Stat Sinica. 2010; p. 1257–1272.

41. Lin J. Divergence measures based on the Shannon entropy. IEEE T Inform Theory. 1991; 37(1): 145–151. https://doi.org/10.1109/18.61115