OPEN ACCESS Freely available online

PLOS | ONE

# Assessing T Cell Clonal Size Distribution: A Non-Parametric Approach

**Olesya V. Bolkhovskaya[1], Daniil Yu. Zorin[2], Mikhail V. Ivanchenko[2]***

**1** Faculty for Radiophysics, Department of Statistics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, **2** Faculty for Computational Mathematics and Cybernetics, Department of Bioinformatics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

## Abstract

Clonal structure of the human peripheral T-cell repertoire is shaped by a number of homeostatic mechanisms, including antigen presentation, cytokine and cell regulation. Its accurate tuning leads to a remarkable ability to combat pathogens in all their variety, while systemic failures may lead to severe consequences like autoimmune diseases. Here we develop and make use of a non-parametric statistical approach to assess T cell clonal size distributions from recent next generation sequencing data. For 41 healthy individuals and a patient with ankylosing spondylitis, who undergone treatment, we invariably find power law scaling over several decades and for the first time calculate quantitatively meaningful values of decay exponent. It has proved to be much the same among healthy donors, significantly different for an autoimmune patient before the therapy, and converging towards a typical value afterwards. We discuss implications of the findings for theoretical understanding and mathematical modeling of adaptive immunity.

**Citation:** Bolkhovskaya OV, Zorin DY, Ivanchenko MV (2014) Assessing T Cell Clonal Size Distribution: A Non-Parametric Approach. PLoS ONE 9(10): e108658. doi:10.1371/journal.pone.0108658

**Editor:** Antonio A. Freitas, Institut Pasteur, France

**Received** April 29, 2014; **Accepted** August 25, 2014; **Published** October 2, 2014

**Copyright:** © 2014 Bolkhovskaya et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability:** The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. Experimentally obtained TCR libraries reported in References 6, 12, and 13 are available as supplementary files on the Publishers' web pages. Additionally, data from Ref. 12 are freely available from Dr. D. Chudakov's Lab webpage http://mitcr.milaboratory.com/datasets/. Libraries from References 14 and 15 were obtained by request from the corresponding authors.

**Funding:** The authors have no funding or support to report.

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

* Email: ivanchenko.mv@gmail.com

## Introduction

T lymphocytes are the key drivers of adaptive immune system, a powerful machinery able to detect, combat, and memorize pathogens in all their possible variety [1]. Specific recognition of potentially harmful foreign peptides is achieved by the highly selective binding of T cell receptors (TCRs) to peptide-MHC complexes (p-MHC), mounted on the surface of specialized antigen-presenting cells. Required diversity of recognition arises due to an astronomic number of distinct molecular variants of TCRs, potentially emerging through random-like V(D)J recombination during cell development in thymus. In result of allelic exclusion, a T cell typically expresses a single TCR variant, and all its daughter cells have identical antigen recognition properties, constituting a clonotype [2]. After massive positive and negative selection in thymus aimed to remove incapable or potentially autoimmune cells, about 5% of thymocytes finally enter periphery. Quantitative statistical assessment of TCR diversity is now becoming possible due to recent developments of next generation sequencing (NGS) techniques, capable of producing large TCR libraries. Current estimates claim about $10^{11}$–$10^{12}$ T cells compartmented in about $10^8$ clonotypes in healthy adult humans [3–6].

Efficient and adequate performance of this ensemble requires fine tuning of T cell populations, which is achieved by a plethora of mechanisms: cytokine, antigen-presentation, and cell regulation [7,8]. The nature of these mechanisms implies that competition for cytokines controls the total number of T-cells, while competition for p-MHC binding sites adjusts the frequencies of individual clones. The latter involves a cascade of processes: proteasomal degradation of proteins, peptide delivery and presentation on MHC complexes, p-MHC - TCR binding events, epitope recognition and signaling, cell fate decision by regulatory networks. Since their thorough experimental quantification faces tremendous difficulties, it becomes appealing to assess at least the outcome.

Attention is getting drawn to analyzing experimentally obtained clonal frequency (relative size) distributions, and there is a growing evidence that they are strongly non-Gaussian. Even the first sequencing results for human TCR repertoires (restricted to clonotypes responsive to a certain peptide, though) indicated that it has heavy tails that could follow a power law [9]. Parametric approaches quite successfully employed several Poisson abundance model distributions for mice with limited TCR diversity [10,11]. However, the studied TCR libraries contained only about a hundred of entries, which precluded a definite choice of the best model among the considered [10].

In contrast, the recently produced and publicly available NGS human TCR libraries contain $10^4$–$10^6$ distinct variants from 41 reportedly healthy [6,12] and one autoimmune [13–15] adult donors, the latter screened before and after chemotherapy and autologous hematopoietic stem cell transfer (HSCT). It gives an opportunity to bring the quantitative analysis of the peripheral T-cell pool statistics on a qualitatively new level.

It is worth noting, however, that quantitatively meaningful estimates of power law distributions demand careful and critical

analysis [16]. In particular, assessing clonotype distributions requires accurate treatment of statistical sampling error, especially significant for low-frequency clonotypes abundant in the T cell repertoire. Otherwise, the blind cutoff of that part of repertoire would seriously restrict the volume of data and span coverage, undermining statistical significance of the result. This has proved to be characteristic of the recent report on the "fractal" organization of the T cell repertoire, with convincing evidence of heavy tailed distributions but quite questionable numerics of power law fits [17].

The purpose of this paper is to conduct a comparative non-parametric analysis to estimate the T cell clonal frequency distributions from a number of available human libraries without any prior assumption on the functional form. We aim to determine whether reconstructed clonal frequency distributions follow some functional dependence, common between different healthy individuals, whether it changes in the autoimmune patient before and after treatment, whether there are differences in healthy and autoimmune repertoire distributions, if any. We reveal the power law scaling over several decades of distributions and quantify the exponents.

## Materials & Methods

We analyze the sets of data coming from: (i) reportedly healthy donors: middle-aged two male and one female subjects [6], 38 subjects aged 9–90 years [12], and (ii) a middle-aged patient with ankylosing spondylitis who undergone chemotherapy and autologous HSCT, and was observed up to 25 months since the treatment [15] (see Table 1 for details). The respective TCR libraries were build by a common workflow that is extraction of peripheral blood mononuclear cells, isolation of RNA, cDNA synthesis, PCR amplification and sequencing of TCR$\beta$ CDR3 region [6,12–15].

It is essential that each step of profiling is prone to experimental and statistical errors that depend on implementation details and equipment (see [6,14,18] for a detailed discussion). We trust the quality assessment and error correction for the reads that the authors of the TCR libraries had performed and make use of their final post-processed data. Following the common approach we associate different TCR variants with different T cell clonotypes.

We focus our attention on the distribution of the multiply read TCR variants, since only they can be reliably distinguished from sequencing errors and allow a trustworthy frequency estimate. Instructively, even the most frequent clonotypes demonstrate significant variances of clonal frequencies in comparative deep sequencing of the same donor either within a week interval [6] or in parallel draws [14]. Therefore, the straightforward histogram analysis of the statistics of TCR reads is a questionable estimate of the clonotype frequency distribution, and it is essential to view the NGS repertoire profiling as a $K$-stage random sampling process, respective to sequencing steps. Notably, the novel cDNA barcoding procedure establishes a one-to-one correspondence between a T cell in a sample and a sequenced TCR variant [12], which allows for reducing the model to a single-step random sampling for their data.

To incorporate statistical uncertainty in our analysis we construct a non-parametric kernel distribution estimator for the complementary cumulative frequency distribution $\hat{F}(p^*) \equiv \mathrm{Prob}\ (p \geq p^*)$ [19,20]:

$$\hat{F}_c(p) = \frac{1}{N_c} \sum_{i=1}^{N_c} \hat{F}_{norm}(p - p_i), \qquad (1)$$

**Table 1.** Characteristics of the TCRB libraries.

| Subject, age | Timepoint (month) | No. of T cells ($\times 10^6$) | Sequences read ($\times 10^6$) | Distinct reads ($\times 10^3$) | Average frequency ($\times 10^{-4}$) | Unseen clonotype frequency CI95 ($\times 10^{-4}$) | Maximal frequency ($\times 10^{-2}$) |
|---|---|---|---|---|---|---|---|
| Male, 29 | – | 12.0 | 188 | 494 | 0.0002 | 0.8 | 0.1 |
| Male, 33 | – | 16.4 | 6.2 | 193 | 0.0005 | 1.0 | 0.2 |
| Female,33 | – | 18.2 | 1.1 | 94 | 0.001 | 3.0 | 0.6 |
| Group 1 (9–25 y) | – | 3.0 | 0.846 | 490 | 0.0058 | 0.037 | 0.25 |
| Group 2 (36–43 y) | – | 3.0 | 0.948 | 446 | 0.0047 | 0.0331 | 7.6 |
| Group 3 (61–66 y) | – | 3.0 | 0.999 | 308 | 0.0031 | 0.03 | 8.6 |
| Group 4 (71–90 y) | – | 3.0 | 0.995 | 354 | 0.0035 | 0.03 | 7.0 |
| Male A, 49 | 0 | 2 | 0.011 | 6.1 | 1.5 | 0.026 | 2.1 |
| Male A, 49 | 4 | 2 | 0.014 | 3.0 | 3.3 | 0.021 | 9.1 |
| Male A, 50 | 10 | 3.5 | 1.07 | 2.3 | 4.4 | 0.0015 | 19.5 |
| Male A, 51 | 25 | 3.5 | 1.06 | 168 | 0.59 | 0.0015 | 9.4 |

Sources of libraries: upper block [6], middle block [12], bottom block [13–15]. Values in the middle block are averages within each group.

doi:10.1371/journal.pone.0108658.t001

where $N_c$ is the number of identified distinct sequences ('distinct reads'), $p_i$ is the frequency of each TCR variant, $\hat{F}_{norm}(\cdot)$ is the kernel function that is the complementary cumulative normal distribution with the zero mean and standard deviation

$$\sigma_{p_i} = \left[\sum_{k=1}^{K} \sigma^2_{p_i:k}\right]^{1/2} \tag{2}$$

dependent on standard deviations $\sigma_{p_i:k}$ present at each of $K$ steps of repertoire profiling. Assuming that sampling a particular variant $i$ at step $k$ is a binomial process with probability $p_i$ we get $\sigma_{p_i:k} = \sqrt{\frac{p_i(1-p_i)}{N_k}}$, where $N_k$ is the total number of samples at this step. Note, that as the expected value of a binomial process $\propto p_i$, the expected frequency of a TCR derivative product will remain $p_i$ at each step of sequencing.

Obviously, the contribution of different steps to statistical errors is quite different (see Table 1 for indicative sample sizes). Three major bottlenecks accumulating statistical error in our case are sampling T cells, synthesized cDNA for PCR (typically, order of the number of sampled cells), and good sequence reads. A single PCR cycle efficiency was estimated to equal $\epsilon = 1.8$ [14], and there was a typical number of about 30 cycles performed. Viewing it as a branching process one obtains an expression for variance on exit $\sigma^2_{p_i,PCR} \approx (2/\epsilon - 1)\sigma^2_{p_i,cDNA}$ [21]. It yields a factor $2/\epsilon = 10/9$ correction to the variance of cDNA sampling.

If in question, the estimator of the probability density distribution can be also obtained then:

$$W_c(p) = -\frac{d}{dp}\hat{F}_c(p). \tag{3}$$

For each data set we estimate the upper 95% confidence bound (further, CI95) for the frequency of unseen clonotypes (viewing observation failure as an outcome of binomial samplings) as

$$p_0 = 1-(1-0.95)^s,\ s = \sum_{k=1}^{K} \frac{1}{N_k}, \tag{4}$$

where $s$ can be viewed as an effective inverse number of trials. Apparently, about and below this value the clonal frequency distribution cannot be reliably estimated from the available data.

To assess the potential impact of the additional sources and types of statistical sampling errors as well as laboratory errors that escaped post-processing we also estimated the clonotype frequency variances from three parallel samples from an autoimmune donor independently taken and sequenced 10 month after HSCT [14] and plugged them into the kernel distribution estimator (1). The result did not show significant deviations from the one obtained with the sampling bottlenecks approach.

Resulting frequency distributions can be analyzed by various approaches. Since they clearly exhibited intervals of power law decay $\hat{F}_c(p) \propto p^{\alpha}$, we performed linear polynomial fits in the double-log scales employing a standard least square method, determining respective exponents and their CI95.

To test applicability of the parametric methods, previously developed for assessing clonal size distributions [10], we developed a Poisson abundance model under an assumption of power law clonal frequency distribution within a certain range. In particular, one describes sampling distribution by a Multinomial law

$$P[\{m_i\}|D,\alpha] = \frac{D!}{(D-N_c)!\prod_{i=1}^{N_c} m_i}[f_\alpha(0)]^{D-N_c}\prod_{i=1}^{M}[f_\alpha(i)]^{m_i}, \tag{5}$$

where $D$ is the total diversity of the T cell repertoire, $m_i$ is the number of distinct TCR sequences found in $i$ copies, $M = \sum m_i$, $f_\alpha(i)$ is the parameter-dependent probability to obtain $i$ copies of a TCR variant in a sample. The latter can be calculated taking into account the typically small size of the sample comparing to the whole repertoire, which yields the Poissson distribution

$$f_\alpha(i) = \frac{e^{-p}p^i}{i!} \tag{6}$$

for the clonotypes with identical frequency $p$. If the frequencies are expected to follow some distribution, in particular, $W(p) \propto p^{\alpha-1}$, integrating out $p$ gives

$$f_\alpha(i) \propto \frac{\Gamma(i+\alpha)}{i!},\ i > -\alpha. \tag{7}$$

Diversity $D$ and distribution power law exponent $\alpha$ are then numerically estimated by maximizing the (log)likelihood function (5). The method is easily adapted to exclude low and high clonal frequency outliers that do now follow power law: it suffice to redefine $f_\alpha(0)$ in (5) as the probability to fall off the the specified range of frequencies.

## Results

We assess clonal frequency distributions ascribing each distinct TCR$\beta$ CDR3 sequence to a distinct T cell clonotype. We make use of publicly available TCR libraries obtained from human donors as described in Materials & Methods. Aiming to avoid prior assumptions on the clonal frequency distribution, we choose a non-parametric approach. As it is outlined in Materials & Methods we construct an estimator based on a complementary cumulative normal distribution kernel. For each clonotype the Gaussian is centered at the measured frequency and has the variance calculated from the binomial sampling model as described in the above. Altogether, it allows to incorporate statistical sampling error, different for clonotypes with different frequencies, avoiding the blind cutoff of the less frequent ones.

First, we analyze NGS TCR libraries obtained from reportedly healthy individuals: 38 subjects aged 9–90 years [12] and two male and one female, middle-aged ("Male 1", "Male 2", and "Female") [6]. Representative examples for donors from different age groups shown in Fig. 1, illustrate the general result: the major parts of the complementary cumulative frequency distributions $F(p) \equiv \mathrm{Prob}(p_i \geq p)$ exhibit a power law decay over at least two decades:

$$\hat{F}(p) \propto p^{\alpha}, \tag{8}$$

where the exponent values are $\alpha = \{-1.16,\ -1.16,\ -1.0\}$. Noteworthy, deviations from the power law are observed not only in the low frequency range (which can be accounted for significant statistical sampling uncertainties), but for abundant clonotypes as well. The latter hints that the clonal size distribution might consist of at least two different parts.

Parametric approach that makes use of Poisson abundance models [10] also produces good results, once a power law distribution is assumed (Fig. 1, inset). However, turning from

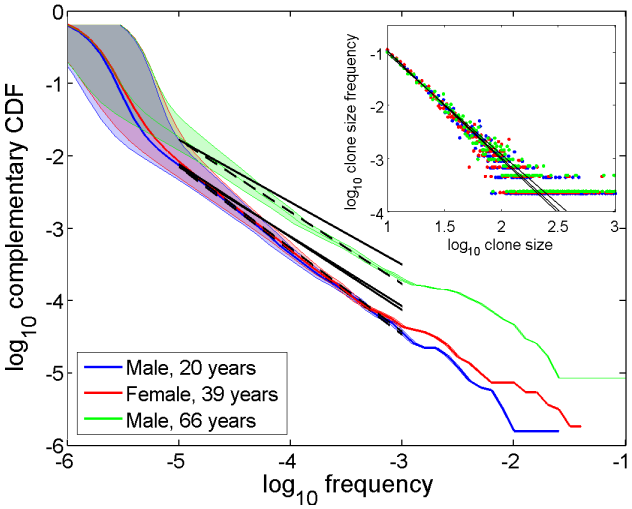


**Figure 1. Clonal statistics for healthy donors.** *Main figure.* Representative complementary cumulative clonal frequency distributions $\hat{F}(p)$ (CDF) for three healthy individuals across different age groups [12] in double log scale: male (20 years), female (39 years), and male (66 years). Shaded areas indicate CI95 intervals for clonal frequencies. Black dashed lines show power law dependencies $\hat{F}(p) \propto p^{\alpha}$, $\alpha \approx -1.16$, $\alpha = -1.16$, and $\alpha = -1.0$, respectively, indicating a good fit of experimental data over two decades (cf. Fig. 2 and Table 2 for more details). Black solid lines show power law dependencies with the exponents derived from the Poisson abundance model fit (5)–(7). *Inset.* Parametric approach [10]: frequency distribution of clonotypes binned by detected size in double log scale (colors code the same donors as in the main figure) and Poisson abundance method fits (black solid lines).
doi:10.1371/journal.pone.0108658.g001

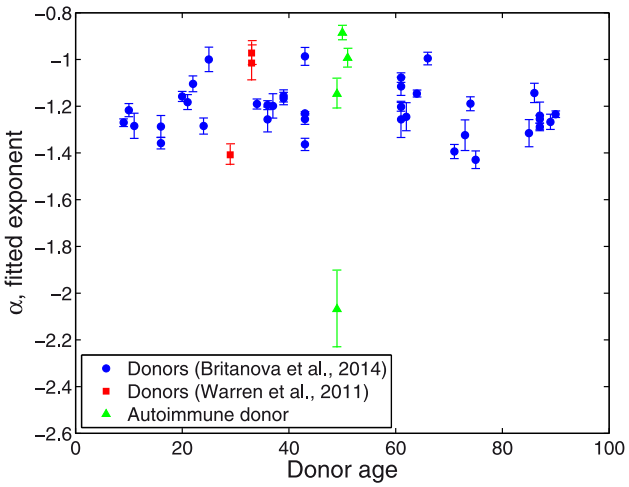


**Figure 2. Power law exponents.** Exponents $\alpha$ of power law fits with respective CI95 indicated vs. age of individuals. Blue circles: healthy donors from [12], least square fits performed over the interval $\log_{10} p \in [-5.5, -3.5]$ or $\log_{10} p \in [-5.0, -3.0]$, whichever produced better quality. Red squares: healthy donors from [6], least square fits performed over the interval $\log_{10} p \in [-4, -2]$. Green triangles: autoimmune patient before and after treatment [13–15], least square fits performed over the interval $\log_{10} p \in [-3.5, -2.5]$ for the time point before treatments and $\log_{10} p \in [-4, -2]$ for the three time points after.
doi:10.1371/journal.pone.0108658.g002

probability density to cumulative distributions, one notices moderate but systematic underestimation of the exponent value (Fig. 1, main). A possible reason for that could lie in employing probability estimators, more sensitive to statistical sampling errors than cumulative ones, used in the developed non-parametric method.

The fitted exponents along with respective CI95 are shown in Fig. 2 vs. donors age (see also Table 2 for details). In all cases they were obtained over two decades of clonal frequencies, specific ranges being $\log_{10} p \in [-5.5, -3.5]$ or $\log_{10} p \in [-5, -3]$ (whichever produces a better fit) for the data from [12], and $\log_{10} p \in [-4, -2]$ for the data from [6]. Overall, the power law exponents fall into quite a narrow range $\alpha \in [-1.43 - 0.97]$. Notably, no pronounced age dependence is observed. One can only point it out that all values among the young (9–20 years) and elderly (70–90 years) groups belong to the lower half of the interval: $\alpha < -1.18$, though it could simply be due to moderate pool of donors.

Secondly, we follow the dynamics of the clonal frequency distribution from an autoimmune patient("Male A") who undergone autologous HSCT and demonstrated a stable remission since [13–15]. Before the HSCT the patient's TCR repertoire had been analyzed for 2 years, and a number of stably hyperexpanded clones associated with inflammatory response were found. NGS was performed right before HSCT (the reference 0 time point), 4, 10, and 25 months after the procedure. Reportedly, the treatment led to a major resetting of the repertoire with only about 10% of earlier observed clonotypes detected afterwards. Previously hyperexpanded clones (including a pro-inflammatory one) drastically decreased their frequencies, while the other clones grew large and remained so over the study [15]. Interestingly, the share of hyperexpanded clonotypes in the repertoire even increased after HSCT, when remission reportedly occured, indicating that their existence is not a trait of an abnormal state.

Reconstruction of frequency distributions before and after the treatment reveals a drastic change in the power law exponent from $\alpha \approx -2.07$ before HSCT to $\alpha \approx -0.99$ after 25 months, though the power law fit of the major part of the distribution (spanning across two decades except for the horizontal axis for time point 0) remains plausible at all times (Fig. 3, Table 2). Remarkably, before HSCT the exponent was considerably below those obtained in reportedly healthy adult individuals (Fig. 2). Conversely, at all time points after HSCT the exponent consistently remained in the interval, specific for healthy donors.

Assessing the possible effect of technical differences in NGS procedures on the clonal frequency distribution, we note the following. The values obtained from healthy donors libraries in [6] and [12] agree well. Qualitatively different values of the power law exponent $\alpha$ are estimated from the TCR libraries assembled before, 4 and 10 months after HSCT from the same autoimmune donor by the same NGS protocol [13]. Moreover, the values of $\alpha$ 4–25 months after the therapy are typical of those for healthy donors, obtained by the protocols different in details. Therefore, differences in protocols do not seem to bias the value of the power law exponent significantly.

## Discussion

Next generation sequencing tools have recently made a high-throughput analysis of T cell repertoire possible. Through the last several years most attention has been paid to estimating clonal diversity, and the reported numbers of distinct clonotypes have been renewing record each time the methods improved and analysis deepened. At the same time the statistics of clonal frequency distributions remained a secondary issue, addressed

**Table 2.** Power law exponent fits for complementary cumulative clonal frequency distributions and 95% confidence intervals.

| Subject, age | Timepoint (month) | $\alpha$ | CI95$_\alpha$ interval width |
|---|---|---|---|
| Male, 29 | – | −1.41 | 0.17 |
| Male, 33 | – | −0.97 | 0.10 |
| Female, 33 | – | −1.01 | 0.06 |
| Group 1 (9–25 y) | – | −1.21 | 0.07 |
| Group 2 (36–43 y) | – | −1.20 | 0.06 |
| Group 3 (61–66 y) | – | −1.15 | 0.07 |
| Group 4 (71–90 y) | – | −1.28 | 0.08 |
| Male A, 49 | 0 | −2.07 | 0.33 |
| Male A, 49 | 4 | −1.15 | 0.12 |
| Male A, 50 | 10 | −0.89 | 0.57 |
| Male A, 51 | 25 | −0.99 | 0.58 |

Sources of libraries: upper block [6], middle block [12], bottom block [13–15]. Values in the middle block are averages within each group.
doi:10.1371/journal.pone.0108658.t002

mostly to estimate the number of "unseen" small-frequency clonotypes by Fisher's techniques [4,22].

About ten years ago even very scarce data led to hypothesize that human T cell clonal frequency distributions have heavy non-Gaussian tails that could follow a power law [9]. However, the study was confined to only 141 clonotype and allowed to produce a power-law-like dependence over a single decade of magnitude at best, rendering the statistical validity of observations and fitted exponents questionable, according to the general practices [16]. A recent study by Meier et al. [17] generated much deeper libraries, reportedly containing about $10^5$–$10^6$ distinct TCR sequences. They also shared the much anticipated view of the self-similar properties of the repertoire, but admitted several shortcomings in their quantitative analysis. First, as we already mentioned in Introduction, the authors dismissed clonotypes with frequencies below 0.05%, restricting themselves to a single order of magnitude interval, insufficient for a reliable power law fit [16]. Second, they estimated clonal frequency distributions with unevenly sized bins, which distorted the result. Another recent report studied evolution of clonal distributions in the course of HSCT, also producing evidence of heavy non-Gaussian tails, but did not attempt to quantify it [23]. To achieve a progress one had to employ appropriate statistical inference techniques, taking into account sampling errors of sequencing but avoiding an overly strong data cutoff.

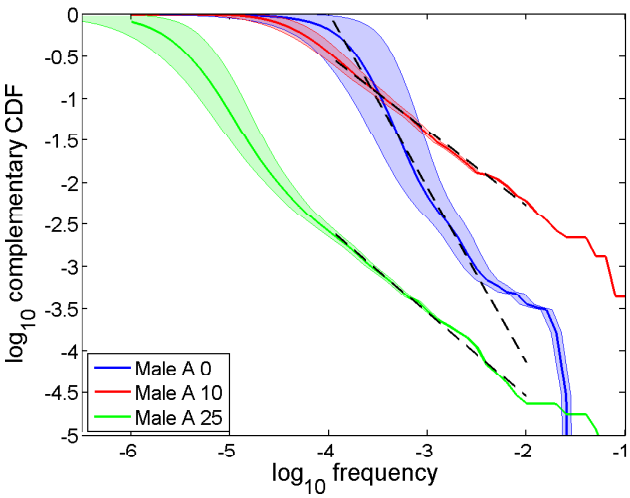


**Figure 3. Clonal statistics for an autoimmune patient.** Complementary cumulative clonal frequency distributions $\hat{F}(p)$ (CDF) for an autoimmune patient [13–15] in double log scale: right before the treatment (blue), 10 months after (red), and 25 months after (green). Shaded areas indicate CI95 intervals for clonal frequencies. Dashed lines show power law fits $\hat{F}(p) \propto p^{\alpha}$, $\alpha = -2.07$, $\alpha = -0.88$ and $\alpha = -0.99$, respectively. Least square fits performed over the interval $\log_{10} p \in [-3.5, -2.5]$ for the time point before treatments and $\log_{10} p \in [-4, -2]$ for the time points after.
doi:10.1371/journal.pone.0108658.g003

Parametric fitting approach offered several model distributions, which became popular choices in analyzing restricted TCR repertoires from mice, to name uni- and multi-variate Poisson abundance models with log-normal, exponential, and gamma sampling rates [10,11]. However, as it was demonstrated in [10], even a systematic analysis of the fit goodness cannot not yield a decisive answer in favor of one or another model, when experimental data are confined to about a hundred of clonotype frequencies. Besides, parametric estimators have two obvious limitations for analyzing clonal statistics *per se*. First, it requires to assume a specific functional distribution of TCR frequencies. Its biological meaning is unclear since there is no bottom-up theory predicting such a distribution from first immunological principles. Moreover, assuming a model distribution we limit our ability to infer these first principles in the top-bottom analysis. Second, it is not guaranteed that a single model distribution describes the whole range of frequencies, and optimal model distributions could, in principle, differ between species or even donors. On the other hand, when the kind of functional dependence(s) and their validity intervals are known *a priori*, by the virtue of another analysis, parametric fitting becomes indispensable.

To overcome these difficulties we proposed to employ non-parametric analysis, appropriate for the currently available extensive human TCR libraries. We have successfully implemented kernel density estimators, which conveniently incorporate the variances of the clonal frequency estimators, essentially dependent on the clonotype size. Variances can either be calculated by comparing data from several parallel samples taken and processed independently (which is preferable, as it counts both statistical and

experimental errors), or by estimating statistical sampling errors at the bottlenecks of an experiment. One has to keep in mind that non-parametric fitting cannot overcome limitations of NGS as the data source sequencing errors, PCR amplification bias, etc. Meanwhile, the developed error detection and correction algorithms along with the most recent cDNA barcoding technique [12] already seem to ensure substantial reliability (for a detailed discussion of the issue we refer a reader to [14,18]). It is worth stressing that non-parametric analysis does not dismiss parametric one, as the latter is required to evaluate and validate the functional dependence in the output data from the former.

Implementing this approach, we demonstrated for the first time that cumulative clonal frequencies distributions from 41 adult human donors can be fitted well with the power law $F(p) \propto p^{\alpha}$ over several decades of magnitude. For reportedly healthy individuals the exponents grouped about $\alpha \approx -1 \ldots -1.4$, while for an autoimmune donor it had been considerably less, about $\alpha \approx -2.1$ before HSCT and increased to the values typical of healthy donors $\alpha \approx -1$ afterwards. We were unable to identify any certain functional behavior of the frequency distribution for small size and hyper-expanded clonotypes, in the first place, due to insufficient statistics in current experimental data: the former yield reads of very low copy number or simply escape sequencing, the latter are too few. We, therefore, cannot exclude deviations from the power law scaling in these parts of distributions.

These findings put a challenge for theoretical immunology to identify the mechanism(s) behind such clonal distributions. Several candidates appear plausible. A major role in shaping individual clonal sizes belongs to T cell competition for access to cognate p-MHCs on antigen presenting cells, for survival and proliferation stimuli. Indeed, it is known that the same antigen can be recognized by several, sometimes hundreds of clonotypes [24,25], and, vice versa, a given TCR can recognize theoretically up to $10^6$ different p-MHCs due to inherent cross-reactivity [26,27]. A good question to address is whether existing mathematical models of T cell competition and clonal selection at the periphery [28–33] can reproduce power-law distributions, or their further development is needed. Another crucial mechanism is the positive and negative selection in thymus, and we await theoretical and experimental advances here. Some insight could also be expected from the statistics of V(D)J recombination, though it still remains a far stretch to infer the functional properties from TCR sequence.

Finally, our results should inspire studying a greater number of human TCR libraries to confirm or disprove that statistical T-cell clonal size distributions for healthy and autoimmune donors (at least for certain diseases) may exhibit drastically different power laws. Another issue to be addressed in a consistent manner is that clonal frequency distribution may possibly evolve significantly under treatment like HSCT, similarly to drastic changes in the other numerical measures [17,23]. We believe that the proposed approach will become a useful tool in the studies of immunity and autoimmunity development, complement the developing deep sequencing methods of individual diagnostics of infectious and autoimmune diseases, characterizing and understanding immunosenescence.

## Acknowledgments

The Authors thank D.M. Chudakov, D.A. Boilotin, Ya. Safonova, and J. Carneiro for illuminating discussions, and also the Max Planck Institute for the Physics of Complex Systems for providing stimulating scientific environment and hospitality.

## Author Contributions

Conceived and designed the experiments: OB MI. Performed the experiments: DZ MI. Analyzed the data: OB DZ MI. Wrote the paper: OB MI.

## References

1. Janeway C, Travers P, Walport M, Shlomchik M (2001) Immunobiology. New York and London: Garland Science, 5th edition.
2. Khor B, Sleckman B (2002) Allelic exclusion at the tcr-beta locus. Curr Opin Immunol 14: 230234.
3. Arstila T, Casrouge A, Baron V, Even J, Kanellopoulos J, et al. (1999) A direct estimate of the human $\alpha\beta$ t cell receptor diversity. Science 286: 958–961.
4. Robins H, Campregher P, Srivastava S, Wacher A, Turtle C, et al. (2009) Comprehensive assessment of t-cell receptor beta-chain diversity in $\alpha\beta$ t cells. Blood 114: 4099–4107.
5. Freeman J, Warren R, Webb J, Nelson B, Holt R (2009) Profiling the t-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res 19: 1817–1824.
6. Warren R, Freeman JD, Zeng T, Choe G, Munro S, et al. (2011) Exhaustive t-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res 21: 790–797.
7. Surth C, Sprent J (2008) Homeostasis of naive and memory t cells. Immunity 29: 848–862.
8. Takada K, Jameson S (2009) Naive t cell homeostasis: from awareness of space to a sense of place. Nat Rev Immunol 9: 823–832.
9. Naumov YN, Naumova E, Hogan K, Selin L, Gorski J (2003) A fractal clonotype distribution in the cd8+ memory t cell repertoire could optimize potential for immune responses. J Immunol 170: 3994–4001.
10. Sepulveda N, Paulino C, Carneiro J (2010) Estimation of t-cell repertoire diversity and clonal size distribution by poisson abundance models. J Immunol Meth 353: 124–137.
11. Rempala G, Seweryn M, Ignatowicz L (2011) Model for comparative analysis of antigen receptor repertoires. J Theor Biol 269: 1–15.
12. Britanova O, Putintseva E, Shugay M, Merzlyak E, Turchaninova M, et al. (2014) Age-related decrease in tcr repertoire diversity measured with deep and normalized sequence profiling. The Journal of Immunology 192: 2689–2698.
13. Mamedov I, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, et al. (2011) Quantitative tracking of t-cell clones after haematopoietic stem cell transplantation. EMBO Mol Med 3: 201–207.
14. Bolotin D, Mamedov I, Britanova O, Zvyagin I, Shagin D, et al. (2012) Next generation sequencing for tcr repertoire profiling: platform-specific features and correction algorithms. Eur J Immunol 42: 3073–3083.
15. Britanova O, Bochkova A, Staroverov D, Fedorenko D, Bolotin D, et al. (2012) First autologous hematopoietic sct for ankylosing spondylitis: a case report and clues to understanding the therapy. Bone Marrow Transplant 47: 1479–1481.
16. Stumpf M, Porter M (2012) Critical truths about power laws. Science 335: 665.
17. Meier J, Roberts C, Avent K, Hazlett A, Berrie J, et al. (2013) Fractal organization of the human t cell repertoire in health and after stem cell transplantation. Biology of Blood and Marrow Transplantation 19: 366–377.
18. Baum P, Venturi V, Price D (2012) Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. Eur J Immun 42: 2834–2839.
19. Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27: 832–837.
20. Parzen E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33: 1065–1076.
21. Grimmett G, Stirzaker D (2001) Probability and Random Processes. Oxford University Press, 3rd edition edition.
22. Fisher R, Corbet A, Williams C (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12: 42–58.
23. van Heijst J, Ceberio I, Lipuma L, Samilo D, Wasilewski G, et al. (2013) Quantitative assessment of t cell repertoire recovery after hematopoietic stem cell transplantation. NATURE MEDICINE 19: 372–377.
24. Pacholczyk R, Ignatowicz H, Kraj P, Ignatowicz L (2006) Origin and t-cell receptor diversity of foxp3+ cd4+ cd25 + t-cells. Immunity 25: 249–259.
25. Kedzierska K, Day E, Pi J, Heard S, Doherty P, et al. (2006) Quantification of repertoire diversity of influenza-specific epitopes with predominant public or private tcr usage. JImmunol 177: 6705–6712.
26. Mason D (1998) A very high level of crossreactivity is an essential feature of the t cell receptor. Immunol Today 19: 395–404.
27. Baker B, Scott D, Blevins S, Hawse W (2012) Structural and dynamic control of t-cell receptor specificity, cross-reactivity, and binding mechanism. Immunol Rev 250: 10–31.

28. De Boer R, Perelson A (1997) Competitive control of the self-renewing t cell repertoire. Int Immunol 9: 779–790.
29. Callard R, Stark J, Yates A (2003) Fratricide: a mechanism for t memory-cell homeostasis. Trends Immunol 24: 370–375.
30. Stirk E, Molina-Paris C, van den Berg H (2008) Stochastic niche structure and diversity maintenance in the t cell repertoire. J Theor Biol 255: 237–249.
31. Stirk E, Lythe G, van den Berg H, Hurst G, Molina-Paris C (2010) The limiting conditional probability distribution in a stochastic model of t cell repertoire maintenance. Math Biosci 224: 74–86.
32. Stirk E, Lythe G, van den Berg H, Molina-Paris C (2010) Stochastic competitive exclusion in the maintenance of the nave t cell repertoire. J Theor Biol 265: 396–410.
33. Ivanchenko M (2011) Transient selection in multi-cellular immune networks. JETP Letters 93: 35–40.