Extreme value statistics of T cell clone sizes

Pablo Mateu Hoyos,[1,*] Andrea Mazzolini,[2] Thierry Mora,[2] and Aleksandra Walczak[2]

[1] *Department of Physics, École Normale Supérieure, Paris, France*

[2] *Statistical Physics and Inference for Biology, Laboratoire de Physique de l'École Normale Supérieure, Paris, France*

Abstract. Very frequent T cell clonotypes in immune repertoires are a topic of great interest among immunologists. In this work, we study their statistics in the framework of an extreme value problem. We implement three models based on different biological and statistical hypothesis: the first model assumes a universal power law exponent and considers that frequencies inside the repertoire are independent; the second model imposes a normalisation over the total sum of the frequencies; and the third model introduces a mixture of the power law exponents across patients. We derive original analytical and numerical predictions which mostly go beyond the standard extreme value statistics. Upon comparison with real human data, we conclude that none of these models is consistent enough in describing the statistics of the top clones, and we propose an alternative model for the normalisation as further research on the topic.

I. INTRODUCTION

The immune system is the entity responsible for the protection against diseases, infection and any foreign or intern pathogen. It consists of a very complex and highly efficient structure of biological processes, at the top of whose complexity relies the adaptive immune system [1, 2].

The history of adaptive immunity dates back to the 1790s, when doctor Edward Jenner started vaccinating the English against smallpox virus and realised that milkmaids who had contracted cowpox virus before, exhibited some sort of natural protection against smallpox. Ever since Jenner's discovery, the study of adaptive immunity has been extended and integrated with several disciplines from biology, mathematics and physics. The more is known about the adaptive immune system, the more its almost unreachable heterogeneity is patent, and the more appealing it becomes. The main objects of interest in adaptive immunity are the cells from the adaptive immune system. One can study from their dynamics to their formation, evolution, statistical distribution, role in the immune response, etc.

In this work, we are concerned with the study of the most frequent ones. Concretely, we are concerned with studying the statistics of the most frequent T cell clonotype (top clone) in human repertoires. The question is: why is this an interesting thing to study?

The adaptive immune system is essentially integrated by B and T cells. Since they are able to provide personalised response, these cells represent the finest line of defence against pathogenic threats. But not only. Neither B or T cells alone are enough, but it is their combined action which makes adaptive immunity complete. B and T cells are organised inside the repertoire as clonotypes, and express themselves through a vast diversity receptors which cover for the plasticity of the adaptive immune response. From all this diversity, it is believed that clonotypes represented by a higher number of cells may be in charge of most of the pathogenic response [3]. Therefore, one expects these top clones to occupy a special position in the immune system.

Top clones are also interesting because of their strange nature. Studies of clonotype population dynamics show that heavy clonotypes are completely in the tail of the clonotype distributions, and their existence is to be considered an unlikely event inside immune repertoires [4–6]. In other words, top clones should be considered as extreme biological events and their statistics can be addressed in the context of an extreme value theory.

Throughout the work, we will realise the rather special nature of T cell top clones with respect to other extreme value problems. Nevertheless, identifying the extreme nature of top clones enables us to benefit from the theoretical framework of extreme value statistics. Not in vain, extreme value statistics equips the analysis of the top clone statistics with an standardise and deeply studied formalism with some known results in a wide range of typical cases [7]; which can be very useful to, first, design the problem in a systematic manner and, eventually, obtain results that help us in understanding the important biological features.

Besides, the connection of such a problem with extreme value statistics is not new, but additional work can be found both applied to different questions in a similar biological context [8], or to similar problems in disparate fields of physics [9–11].

With all this in mind, we start this report by introducing the reader to the necessary basics for its understanding. In Sec. II, we give an overview of the different biological concepts loosely mentioned here and the general dynamics of T cell repertoires, whereas in Sec. III we derive the more general results from extreme value statistics applied to T cell top clones. As a first nontrivial example, in Sec. IV (and Appendix A) we operate this general prescription in a simple model for independent frequencies and a universal power law exponent in the different clonotype distributions. We recognise its associated top clone density distribution as a Fréchet class in a classical extreme value theory and get reasonable

* pablo.mateu.hoyos@ens.psl.eu

results for the parameters. However, we identify that such a model is not biologically consistent, and propose a more biologically based model that considers normalisation of the frequencies as a global constraint in the joint distribution (Sec. V). We demonstrate how normalisation makes clonotype distributions no longer universal across patients, and discuss a refined version of the model replacing the universal exponent with a mixture over a normal distribution for the power law exponents. This is done in Sec. VI and represents the final result of this work. In the end, we realise that even the results obtained this way reveal themselves inaccurate and physically inconsistent; therefore, the model needs to be rejected. We conclude the report with a detailed analysis of the causes for incompatibility and with the proposition of an alternative model that might yield a solution (Sec. VII).

Advanced computational techniques for numerical integration and data fitting have been implemented in Secs. V and VI in order to compare the analytical predictions to the experimental data. Detail on these methods can be found in Appendix B, and the most relevant codes (to be indicated in the text) can be found in my GitHub repository `https://github.com/MateuPablo/EVS-T-cell-clones`.

II. THE BIOLOGY OF T CELL CLONES

The theory of extremes works independently of what T cells are, how they are produced or what is their importance in the immune response to a pathogenic threat, but assumes only some features of their statistical distribution. In this sense, the problem is general enough for one to forget all these questions, stick to the numbers and still get the correct result.

Nevertheless, interpreting this result requires building some intuition on how T cell repertoires behave and what their biological relevance is. Biological context is necessary to justify the assumptions made and, in turn, the theoretical framework of extreme value statistics may help us in understanding this biology.

In this section, we provide the reader with a brief but useful insight into the immune system [1, 2] and describe the general aspects of the population dynamics of T cell repertoires [4].

A. The immune system

Every human enjoys a highly sophisticated machinery to defend themselves against external invaders. Whenever a pathogen, that is, a disease-producing organism, enters our body, this machinery is automatically activated in order to prevent the pathogen from causing damage to our organism. This reaction is referred to as *immune response*, and the machinery in charge of it is called **immune system**.

Formally, the immune system can be defined as the network of organs, cells, proteins and physical and chemical barriers that protects organisms from diseases. Even if nearly all organisms have some kind of immune protection, its complexity varies widely between species. For vertebrates, this protection comes along in almost an infinity of different responses to different general and specific threats, which organise themselves in multiple layers. The first obstacle pathogens must overcome is the physical barriers formed by the skin, ciliated epithelia and mucous membranes. Then, the immune system is considered to present two more different lines of defences: the innate immune system and the **adaptive immune system**.

The innate immune system can be mostly seen as a line of pawns that attack pathogens able to penetrate the physical barriers. These pawns have numerous *receptors* in their cell surface which allow them to recognise characteristic molecules in a very wide range of common pathogens. However, pathogens displaying high mutation rates are able to escape the recognition of the innate immune system, making it unable to cover for the whole spectrum of possible invaders.

This deficiency is compensated by the adaptive immune system involving two more advanced types of cells, **B** and **T cells**, called *lymphocytes*. As well as innate cells, B and T cells exhibit receptors on their surface, namely BCRs and *TCRs*, responsible for pathogen recognition, but unlike innate immune cells, each type of B and T cell expresses multiple copies of only one type of receptor. To guarantee the needed receptor diversity, adaptive immune cells are generated via a stochastic recombination process called V(D)J recombination. Once generated, B and T lymphocytes undergo a "negative" selection that eliminates self-reactive receptors. Cells exhibiting non self-reactive receptors are spared by this mechanism, replicated by *clonal expansion* and allowed to migrate to the blood and peripheral organs in order to start operating. As a consequence, adaptive immunity not only provides a personalised response to each invader, but this response is also more refined than the one offered by innate immunity.

In this work, we are concerned with T cells. Hence, everything in the following will refer to T cells. The set of all the T cell population is called T cell **repertoire**. Cells exhibiting the same type of receptor are called **clones** or said to be of the same **clonotype**. In this work we will use both terms indistinctly. The total number of different clonotypes is called **diversity** of the repertoire. Every repertoire contains multiple clones of each clonotype. The amount of these clones is called **clone size**, and the "biggest" clonotype is said to be the **top clone** of the repertoire.

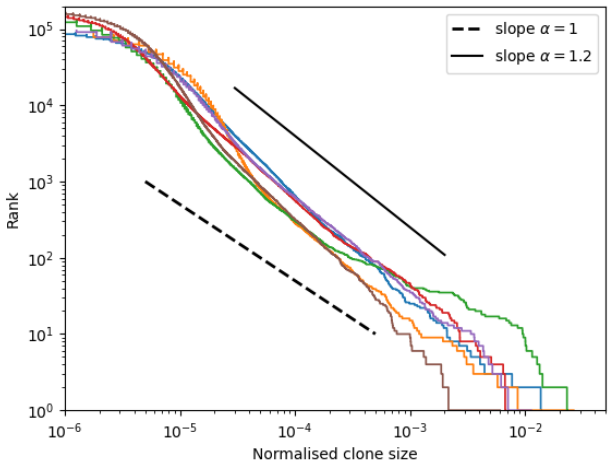

FIG. 1. Power law behaviour of the repertoires of 6 patients. The dashed black line displays a power law curve of slope $\alpha = 1$. The solid black line does the same for a slope of $\alpha = 1.2$. We observe that for this small sample the behaviour is more similar to the solid than to the dashed black line. Most of the patients' curves should be contained in between them. The repertoire shows some increase of the step between points at the end of the tail as a cause of the reduction of data for big frequencies. Seemingly, it can be appreciated the deviation from the power law suffered by all patients in the small frequency region, below a threshold of around $f = \cdot 10^{-5}$.

B. Clonotype population dynamics

Due to the variety of receptors, most of the clonotypes have very small size. Based on recent estimates, the human immune repertoire contains about 10^{11} total T cells, out of which a diversity of about $10^8 - 10^9$ is expressed [3]. This means that, on average, each clonotype has a size of about a hundred or thousand of copies. To compensate for the low size of most of its clones, adaptive immunity displays a mechanism called clonal selection, for which B and T cells are able to replicate and generate a temporary overpopulation of the specific type of cell needed when an immune response is triggered.

Experimental evidence [3, 12–15] shows that the distribution of clonotypes is not uniform, but follows a power law behaviour

$$\rho(C) \propto C^{-1-\alpha} \tag{1}$$

of exponent $\alpha \approx 1 - 1.2$ for the clone size C distribution (see Fig. 1), which can be replicated through theoretical models of lymphocyte population dynamics [5, 16]. A power law behaviour like this reflects the fact that most clonotypes have very low abundance, but also allows for the existence of unlikely big clones. In general, very frequent clonotypes may be more important for immune protection than clonotypes expressed by very few cells [3]. Picture it: from all the 10^9 unique cells we have, the immune response might be mostly leaded by one (or ten) of them! Hence, being able to make predictions on the behaviour of these very frequent clonotypes becomes an immediately interesting topic. These top clones are our object of study.

III. EXTREME VALUE STATISTICS

It is clear from Sec. II that clonotypes in the tail of the power law distribution enter the category of *extreme events*. Extreme events are those which deviate significantly from the average behaviour of their representative distributions. The branch of statistics that deals with them is called **extreme value statistics (EVS)**.

Here, we give the first step towards the application of the EVS formulation to a T cell repertoire. First, we build some intuition around the theory. The uniqueness of each individual problem leads to very different theories for their extremes, but the formulation of an EVS theory expresses some shared traits among different problems. We discuss these common features and present the general framework for the EVS in the biological context of our problem. Ultimately, we give a short detail on how these features are inferred from experimental data.

A. The theory of extremes: preliminaries

Extreme value statistics is aimed at building theoretical predictions over the maximum values of a variable based on previous observations of this variable.

If we have a set N of variables $(x_1, \ldots, x_N)$, each of them drawn from a known parent distribution $\rho_i(x)$, then a properly constructed theory of extremes should be able to forecast the probability that the maximum of the set, $x_{\max} = \max(x_1, \ldots, x_N)$, exceeds or is under certain values. In other words, EVS lightens information on the probability distribution of the extreme value of the set.

To illustrate this idea, consider the following example. Imagine you want to determine the probability of an earthquake happening any day in your city, and you decide to face it as an extreme value problem. First, you define your variables as the seismic levels at different points of the day. Second, you need to build a model for the EVS of the seismic level. With this purpose, you make some theoretical assumptions on the statistical distribution of the daily seismic levels, introduce some parameters on which the seismic level should depend, and, eventually, calculate the probability distribution of the daily maximum value. In order to contrast your predictions, you need to collect, in parallel, daily measurements of the maximum seismic level over some big number of days. These measurements represent the experimental data to which your distribution should relate to. Upon comparison, you can not only validate or reject your hypothesis, but also infer the suitable values for the parameters. If the model is feasible, the picture becomes complete when you set the barrier for an earthquake event.

Extreme value statistics is a well understood theory for the case of independent and identically distributed (i.i.d) random variables [7]. In this situation, the theoretical functions respond to well-known analytic distributions for the maximums, which properly centred and scaled, can be shown, based solely on the asymptotic behaviour of the variable's parent distribution, to always fall into one out of three categories: a Gumbel distribution when x is unbounded and $\rho(x)$ decays faster than a power law for large x; a Fréchet distribution when x is upper unbounded and $\rho(x)$ decays as a power law for large x; and a Weibull distribution when x is bounded. These are usually called *universal classes*. The centring and scaling procedure is usually referred to as *renormalisation*, and essentially consists in performing a problem-specific linear change of variables such that the new variable absorbs the particular dependency and generalises the original distribution to one of the universal ones. Since it cannot be applied to this work, we won't give more detail on this. Nevertheless, it is important to understand that renormalisation is one of the major achievements of EVS and a very important consequence of the theory, with substantial benefits when it can be used.

A developed theory for the statistics of the extremes is still accessible for the first M maximums (order statistics) and for weakly correlated variables, but that is no longer the case when dealing with strongly correlated variables [7]. Recent studies suggest that results similar to the ones derived for i.i.d random variables can be replicated by permitting non-linear transformations in the renormalisation of the distributions [11]. Then, the EVS of i.i.d variables would be a particular case of a much more general theory of extremes. However, these non-linear changes of variables would need to ensure certain properties which were guaranteed for linear transformations and, in general, may not be anticipated from the variable's parent distribution. In the end, one mostly needs to face each problem separately.

B. T cell repertoires and extreme value statistics

In a T cell repertoire, the objects of study are T cell clones. As we know, each clonotype inside a repertoire is represented by a clone size, which is usually better defined in terms of its frequency rather than the number of individuals. The most natural interpretation for the number of variables, N, is the diversity of the repertoire. Altogether, we can model the patient's repertoire as a set of N different frequencies $(f_1, \dots, f_N)$, each representing a different type of T cell. As they are frequencies belonging to the same individual, they should satisfy the normalisation condition

$$f_1 + \cdots + f_N = 1 \tag{2}$$

to be be well-defined as frequencies. The top clone is represented by the maximum value among these frequencies,

$$f_{max} = \max(f_1, \dots, f_N). \tag{3}$$

Being lymphocyte frequencies, each of them must follow a different power law distribution of the type Eq. (1). Since all T cells experience the same selection process, it is reasonable to assume that

- frequencies of clonotypes follow identical distributions

This implies that all frequencies must come from the same power law distribution, which (properly normalised and adapted to frequency) reads

$$\rho(f) = \frac{\alpha}{f_{min}^{-\alpha}} f^{-(1+\alpha)}. \tag{4}$$

Seemingly, all frequencies are drawn from the same minimum value f_{min} to a maximum value of one.

To have an extreme value theory, we need to compute the probability curve of the top clone. For this, it is useful to define the cumulative distribution of f_{max} (CDF), which expresses the probability that f_{max} is under a certain value m:

$$\begin{aligned} G_N(m) &= P(f_{max} \leq m) = \\ &= P(f_1 \leq m, \dots, f_N \leq m) = \\ &= \int_{f_{min}}^{m} df_1 \cdots \int_{f_{min}}^{m} df_N \, P_{\text{joint}}(f_1, \dots, f_N). \end{aligned} \tag{5}$$

Then, the probability distribution (PDF) of the maximum is given by

$$g_N(m) = \frac{\partial G_N(m)}{\partial m}, \tag{6}$$

which expresses the probability that a certain measurement of the maximum, f_{max} has a given value m, for every possible value m.

From Eq. (5) it is clear that the result depends on the underlying statistical distribution of the variables, the possible correlations between them and their physical interpretation as clone frequencies, which sets boundaries to the the values m and f_{min} can assume; for all these factors shape the joint probability function. In general, one always takes the limit $N \to \infty$, which we stated in Sec. II B that is reasonable in this context.

Given the T cell repertoire of an arbitrary patient, Eq. (6) should return congruent predictions on the behaviour of the top clone. Nonetheless, the result of the calculation is naturally going to depend on the parameters α, f_{min} and N, which are in principle sensitive to every patient.

In order to make the problem analytically affordable, we make the two following assumptions:

- f_{min} is the same for every patient,
- α is the same for every patient.

These assumptions may seem very restrictive, but previous studies conveyed over the same data we have used point out the existence of patterns of universality across patients [15]. In any case, their validity will be put to the test in every specific case.

C. Experimental data for the extremes

The full data of our work is integrated by a set of 632 repertoires belonging to 632 different samples of blood from healthy donors, to which we will refer in the following as Emerson cohort [17].

In order to render the data useful, some preliminary analysis is needed. Most of the data analysis consists in cleaning the original repertoires and then finding each of the top clones. The reader can refer to Appendix B 1 for further detail and to my GitHub repository for the code. Some additional work over the full repertoires has been done, but it will be explained later on.

Sampling is a major challenge when analysing immune data [18]. Sampled abundances do not correspond exactly to clonotype abundances. This mismatch can be partially compensated working with frequencies, but even so it is important to distinguish between the true clonotype frequencies and the sampled ones, which are sensitive to biological and experimental noise and only provide a noisy reflection of the true ones. Whereas small clones are very subject to repertoire dynamics and sampling noise, top clones are less limited and it is precise enough to use the sampled values as approximate to their true values. This mostly ensures the validity of the comparison with top clone data from the cohort. We also assume that, as well as they do in the actual repertoire, clonotypes follow the power law behaviour Eq. (4) in the sampled repertoire. Typical sampled repertoires have sizes of around $N_r = 10^6$. Consequently, parameters such as f_{min} and N cannot be inferred directly from the data and need to be addressed through analytical methods.

The theory and assumptions exposed in this section are common to all the different models we will use to approach the statistics of T cell top clones. Seemingly, the collection of 632 top clones from the Emerson cohort serves as experimental data to compare with our theoretical predictions. For this, we will essentially fit these predictions to the data for the top clones in order to obtain the best estimates for the parameters of the model. In doing so, we will also give some insight into the analytical and computational tools used in the process.

IV. EXTREME VALUE STATISTICS OF T CELL CLONES WITH INDEPENDENT FREQUENCIES

The simplest EVS problem one can find is the one in which the variables are independent and identically distributed. The existence of deep theoretical knowledge in such a case and the analytical simplicity of the model motivate its application to the top clones of T cell repertoires as an interesting zeroth case. Based on the discussion from Sec. III, we expect to find a Fréchet-like distribution for the top clone statistics.

A. Standard model for top clone statistics

As stated in Sec. III B, clonotype frequencies from the same repertoire are always assumed to be identical. Additionally, we assume here that they are independent from them others. Two or more variables are said to be independent when their joint probability factorises to the product of the individual probabilities, that is, there is no correlation whatsoever in their drawing process. Therefore, the joint probability of such a model is defined by

$$P_{\text{joint}}(f_1, \dots, f_N) = \rho(f_1) \dots \rho(f_N), \tag{7}$$

where, naturally, $\rho(f)$ is given by Eq. (4) for every variable f_i.

The PDF for the top clones can be derived applying Eq. (5) to Eq. (7) and then using Eq. (6). The main particularity from this problem is the upper bound $m \leq 1$. However, as long as $m \gg f_{min}$ (which is true due to the extreme value condition), this bound does not have an influence. As a result, this represents an almost standard Fréchet type extreme value calculation, of which detailed resolutions can be found in usual EVS textbooks. For instance, the reader can refer to [7], and a detailed overview on this calculation can be found in Appendix A, where more emphasis is placed on the particularities of the problem and the reduction to a standard Fréchet distribution. In the end, the PDF we are looking for reads

$$g_s(m) = \frac{\alpha N}{f_{min}^{-\alpha}} m^{-(1+\alpha)} \exp\left(-N\left(\frac{m}{f_{min}}\right)^{-\alpha}\right), \tag{8}$$

where the index "s" stands for *standard* extreme value statistics (sEVS).

Eq. (8) shows that the statistics of the top clones effectively depend on three parameters: α, f_{min} and N; but these parameters are not independent. Indeed, the parameter dependency of the PDF can be fully expressed in terms of only α and $N f_{min}^{\alpha}$, which indicates that a relation of the type $N(\alpha, f_{min})$ could, in principle, be found.

Taking the average of Eq. (2) over the joint probability Eq. (7), it can be easily shown that N satisfies the relation

$$N\langle f \rangle = 1, \tag{9}$$

where $\langle\ \rangle$ expresses averaging over the power law distribution (4). Eq. (9) implies that N is now a function of α and f_{min} and, therefore, universal as well. Concretely, $N \propto f_{min}^{-1}$. Given the typical diversity of T cell repertoires, this establishes a value of about $10^{-9} - 10^{-10}$ for f_{min}.

B. Results

We are only left with fitting Eq. (8) to the Emerson data in order to find the best estimates for α and f_{min}.

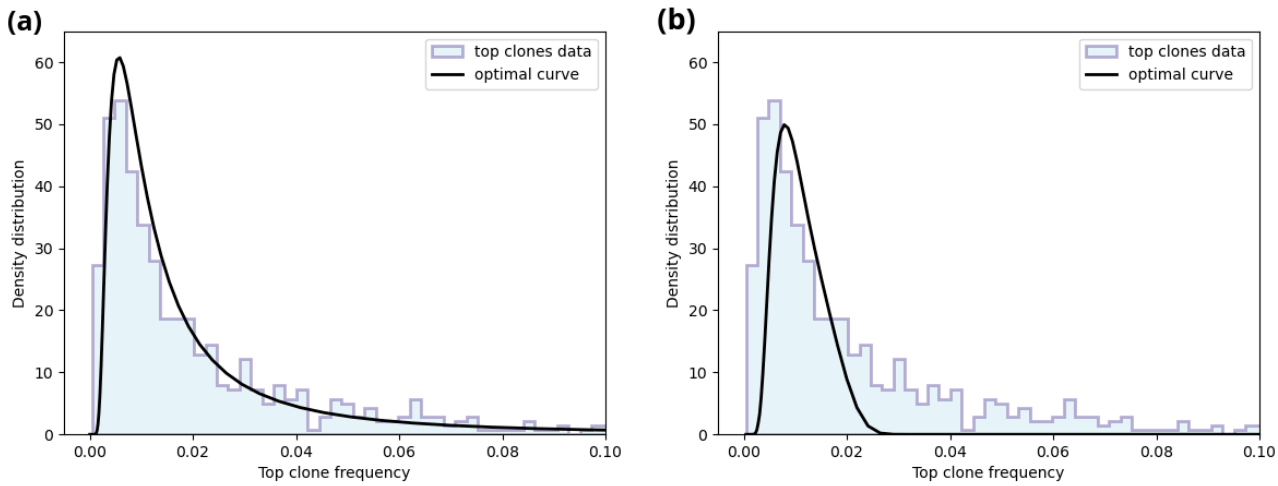

FIG. 2. Extreme value statistics of T cell top clones: universal power law exponent. (a) Fit of the standard extreme value prediction to Emerson data. The theoretical curve (black) (Eq. (8)) is plotted for $\alpha = 1.18$ and the optimal value of $f_{min} = 1.65 \cdot 10^{-9}$ obtained from the fit. An initial condition of $f_{min,0} = 10^{-10}$ is chosen. The fit returns a value of $R^2 = 0.87$. (b) Fit of the normalised extreme value prediction to Emerson data. The theoretical curve (black) (Eq. (19)) is plotted for the optimal parameters $\alpha = 1.12$ and $f_{min} = 3.25 \cdot 10^{-13}$ obtained from the fit. The fit is done for initial values of $f_{min,0} = 10^{-10}$ and $\alpha_0 = 1.2$, but it is very sensitive to small changes in them. The fit returns a value $R^2 = 0.11$. For both figures: the histogram of the Emerson top clones (skyblue) is displayed using a linear binning of $B = 200$ bins. Only the region of frequencies under 10^{-1} is shown. As it can be seen, only a few points describe the peak region, whereas most of the points belong to the tail (which extends until around $m = 0.5$).

While this is a simple fit that can be done using the standard Python functions and does not need any remarkable computational tool, one immediately realises that fitting both parameters together does not yield a unique pair of optimal values, but a continuum of pairs that seem to be optimal. This suggests that there may be a big region of indeterminacy between α and f_{min}, and therefore both parameters cannot be fitted simultaneously. To come across this difficulty, we have fixed α externally and fitted only to f_{min}, using the value $\alpha = 1.18$ obtained in [15]. The result is shown in Fig. 2a, and yields a value

$$f_{min} = 1.65 \cdot 10^{-9}. \tag{10}$$

Fig. 2a exhibits a clear agreement between the data and the behaviour described by Eq. (8). Moreover, the estimates for the two parameters f_{min} and α are compatible with the estimations from Sec. II B. In this sense, the model validates the assumptions made in Sec. III B on the universality of the parameters.

V. EXTREME VALUE STATISTICS OF T CELL CLONES WITH NORMALISED FREQUENCY SUM

Eq. (10) and Fig. 2a reveal that taking the clonotype frequencies as f_i as independent variables yields a good outcome. However, from a more physical point of view, it does not seem to be a consistent assumption.

In Sec. III B we argued that frequencies belonging to the same repertoire must obey the normalisation condition from Eq. (2). Concretely, what we physically require is that the clone sizes sum up to the total number of cells, and by taking the variables to be independent we are somehow deciding to neglect the effect such a constraint might have in the joint statistics. While this is be valid in the case in which the total number of cells, N_{cell}, does not have major restrictions and can be approximated as infinite, one expects that, if we require that N_{cell} is strictly finite, it may not be possible to ignore the constraint over the total number of cells anymore and the effect normalisation in Eq. (2) has might need to be included somehow in the model.

It has been shown that, in order to prevent the apoptosis caused by over-proliferation, there needs to be a finite upper limit to the total number of T cells that human repertoires can display [19]. Assuming $N_{cell} \sim \infty$ disregards an important biological restriction, thus normalisation is required.

Naturally, a condition such as Eq. (2) imposes a constraint over the frequencies and enforces to drop the joint probability given by Eq. (7), which would be no longer representative of the problem. By taking into account this necessary normalisation, we should be able to improve the results from Sec. IV and derive values for the parameters closer to the real quantities. But for this, a new model is required.

A. Normalised model for top clone statistics

There exist several ways of translating the normalisation in Eq. (2) to a model for the statistics of the extremes. Here, we introduce the correlation between variables as a delta function over the joint probability of the frequencies. The same choice can be found in works in other fields of physics in which similar normalisation arises in the context of EVS [9, 10]. The joint probability reads

$$P_{\text{joint}}(f_1,\dots,f_N) = \\ = \frac{1}{Z}\,\rho(f_1)\dots\rho(f_N)\,\delta\left(f_1+\dots f_N-1\right), \tag{11}$$

where $\rho(f_i)$ obeys again Eq. (4) and Z can be interpreted as a partition function that ensures global normalisation of the probability. The presence of the delta function enforces the sum of the frequencies to take the value of one in order to have a non-vanishing probability.

In order to operate with this expression, we need to adopt a suitable representation for the delta function. Given the linearity of the argument, the simplest choice is the Fourier representation, with which the delta function reads

$$\delta(f_1+\dots+f_N-1) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} d\mu\, e^{i\mu(f_1+\dots+f_N-1)}. \tag{12}$$

The joint probability can be thus completely factorised and takes the expression

$$P_{\text{joint}}(f_1,\dots,f_N) = \\ = \frac{1}{2\pi Z}\int_{-\infty}^{+\infty} d\mu\, e^{-i\mu}\,\rho(f_1)e^{i\mu f_1}\dots\rho(f_N)e^{i\mu f_N}. \tag{13}$$

Based on Eq. (13), the effect of correlations can be seen as the Fourier transform of a standard model with an effective clonotype parent distribution $\rho(f)e^{i\mu f}$.

B. Derivation of the statistics of the extremes

It is convenient to introduce the reduced notation

$$\langle L;m\rangle = \int_{f_{min}}^{m} df\rho(f)\,L(f), \tag{14}$$

for arbitrary $L(f)$, which encloses two interesting cases:

- $\langle L;1\rangle$ equals the usual average over $\rho(f)$
- $\langle 1;m\rangle^N$ equals the standard CDF (see Eq. (A2))

Applying Eq. (5) to Eq. (13) and using Eq. (14) leads to

$$G_N(m) = \frac{1}{2\pi Z}\int_{-\infty}^{+\infty} d\mu\, e^{-i\mu}\,\langle e^{i\mu f};m\rangle^N. \tag{15}$$

We can now place the exact form of the parent distribution $\rho(f)$. By taking the limit of large N, the final expression simplifies to the following compact form:

$$G_n(m) = G_s(m)\,H(m). \tag{16}$$

The index "n" reads for *normalised* EVS (nEVS). The first term in the RHS of Eq. (16) is given by the CDF of the independent frequency model (Eq. (A3) in Appendix A) and, naturally, represents the part independent from the model. On the other hand, $H(m)$ is a function that contains all the information about the imposed correlations. It reads

$$H(m) = \frac{1}{2\pi Z}\times \\ \int_{\mathbb{R}} d\mu\, e^{-i\mu}\exp\left\{N\log\left(\frac{\langle e^{i\mu f};m\rangle}{\langle 1;m\rangle}\right)\right\}. \tag{17}$$

The partition function Z is fixed by normalisation, using the fact that both $G_n(m=1)$, $G_s(m=1)=1$, and reads

$$Z = \frac{1}{2\pi}\int_{\mathbb{R}} d\mu\,\exp\left\{-i\mu + N\log\langle e^{i\mu f}\rangle\right\} \tag{18}$$

Then, applying Eq. (6), we get

$$g_n(m) = g_s(m)\,h(m), \tag{19}$$

where $h(m)$ is given by

$$h(m) = \frac{1}{2\pi\langle 1;m\rangle Z}\times \\ \int_{\mathbb{R}} d\mu\, e^{-i\mu(1-m)}\exp\left\{(N-1)\log\left(\frac{\langle e^{i\mu f};m\rangle}{\langle 1;m\rangle}\right)\right\}. \tag{20}$$

Similarly, $g_s(m)$ is the "free" term, given by Eq. (8), while $h(m)$ encompasses all the dependency on the normalisation.

The complexity of both Eqs. (17) and (20) hints at a similar complexity in the numerical implementation of the statistics. With this in mind, we have found different analytical simplifications of these expressions using saddlepoint approximations to first, second and even third order, but these have revealed physically inconsistent. As a consequence, Eq. (20) needs to be evaluated numerically.

Another important change introduced by the constraint is that relation $N\langle f\rangle = 1$ (Eq. (9)) is no longer ensured. Indeed, averaging of Eq. (2) over the joint probability yields now a non intuitive expression, which does not bring any analytical or numerical advantage. Seemingly, one could come up with more complex relations involving constraints over the average frequency or the variance, but they may not have a logical interpretation and would surely complicate the problem. In any case, we do not expect the underlying biological process to respond to very exotic behaviours, but to obey an implementation similar to the one it would in an hypothetical unconstrained case. Even if the statistics of the extremes will be be affected, we believe it should be without requiring major changes on the statistics of the single average values. Based on this reasoning, we assume here that relation Eq. (9) still holds. Note that this is now an assumption, and not a consequence of the problem.

C. Result

It can be seen numerically that the constraint seems to remove the previous indeterminacy over parameters α and f_{min}. An intuitive explanation to this may be that, due to normalisation, a degree of freedom is taken from the joint probability and needs to be placed over the parameters for "conservation" of the degrees of freedom of the underlying problem. As a result, the free parameters of the model are again α and f_{min}, but in this case we need to fit them both at the same time to get the optimal result. An overview on the main challenges faced in the evaluation of Eq. (19) and the parameter-fitting to the data can be found in Appendix B 2. The code for the integral and a fit example are available in my GitHub repository. Here, we limit ourselves to present the results derived from this analysis.

Fig. 2b shows the fit of Eq. (19) to the Emerson top clone data. The fit returns the values

$$\begin{aligned} \alpha &= 1.12, \\ f_{min} &= 3.25 \cdot 10^{-13} \end{aligned} \tag{21}$$

for the parameters.

Obviously, the results are not what expected. In first place, the values for the parameters are not fully consistent, since one would imagine a bigger value for f_{min} based on Sec. II B estimates. In second place, there is a clear mismatch between the shapes of the data and the theoretical curve. Eq. (19) fails to describe the tail behaviour of the empirical distribution and compensates this by displacing the peak to higher frequencies, to the right of its experimental position.

In addition, the fit reveals to be very sensitive to the choice of the initial conditions and the upper and lower limits of the fitted parameters, for small changes result in very different outcomes, many of which are out of the physical barriers of the problem. Our intuition is that this may be another consequence of the model not working properly.

VI. EXTREME VALUE STATISTICS OF T CELL CLONES WITH NON-UNIVERSAL POWER LAW EXPONENT

The results displayed in Eq. (21) and Fig. 2b show that either the normalised model developed in Secs. V A and V B fails or the assumptions done in Sec. III B on the universality of parameters f_{min}, α or N are no longer valid.

The need for normalisation is based on biological evidence and it should be respected. Seemingly, its introduction via a delta function over the joint probability of i.i.d variables is the simplest approach that can be taken, and since other strong assumptions have been made, a full reformulation at this point might not be the best option.

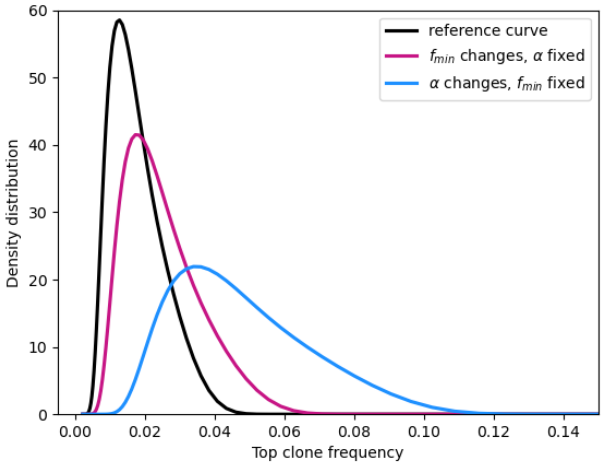

FIG. 3. Variation of the probability density curve with f_{min} and α. Reference values of $f_{min} = 10^{-9}$ and $\alpha = 1.2$ have been taken. The curve in black displays Eq. (19) for these values. To study how the parameters determine the shape, two additional curves are plotted: in magenta, for a value of $f_{min} = 10^{-8}$ while keeping α unchanged; in blue, it is f_{min} what remains fixed and α assumes a value of $\alpha = 1.1$. As seen, small relative changes of α can be much more influential than rather big ones in f_{min}.

Seemingly, we shall exclude not fixing N via $N\langle f \rangle = 1$. This way N is entirely determined by f_{min} and α, which allows for a big flexibility on its range of values, and ensures that $N \sim f_{min}^{-1}$. Dropping this assumption would imply the addition of a free parameter to the problem and, therefore, would lead to difficulties in the fit to the data. In particular, any other model would need, in principle, to fix N somehow, and this one we are using is definitely the simplest and more intuitive option of doing so.

Comparison between Figs. 2a and 2b shows that the tail behaviour of the probability distribution is remarkably model-dependent. In order to determine whether the parameters can be causing this mismatch, we have studied the variation of Eq. (19) with both α and f_{min}. The result is shown in Fig. 3, and indicates that small changes in α have significant effects in the shape of the distribution.

The hypothesis that α has a constant value across patients is based on the analysis from [15]. Note, however, that α was never the same for all patients, but the uncertainty in there obtained was small enough to be considered negligible both in the contexts of their experiment and Sec. IV of ours. Nevertheless, as Fig. 3 points out, that may not be the case anymore. Differences in the values of α should no longer be neglected and, as a consequence, we need to drop the assumption that α is the same for all patients from our theoretical formulation.

With α assuming different values for different patients, each individual patient may lead now to its own theory for the behaviour of the top clones. In other words,

Eqs. (4) and (19) are no longer universal.

A. Mixed model for the top clone statistics

To overcome this patient-dependency, we have performed a *mixture* over α of the particular patient distributions and studied the statistics of the resulting α-independent distribution.

We start by rewriting the PDF for the maximum in a way that points out the dependency on α of the distribution:

$$g_n(m;\alpha) \equiv g_n(m). \tag{22}$$

This way, we move from the unique density function represented by Eq. (19) to a set of distributions continuously parameterised by α. Given a probability density $\rho_\alpha(\alpha)$, the function

$$g_m(m) = \int_1^\infty d\alpha\, \rho_\alpha(\alpha)\, g_n(m;\alpha) \tag{23}$$

constitutes the corresponding *mixture distribution*. Naturally, the index "m" stands for *mixture* EVS (mEVS). Normalisation is guaranteed as long as both $\rho_\alpha(\alpha)$ and the $g_n(m;\alpha)$ have been properly normalised. As it can be seen, the mixture distribution is independent of α and, thence, universal.

In order to integrate Eq. (23), we need to know the form of the density function $\rho_\alpha(\alpha)$. It is our assumption here that $\rho_\alpha(\alpha)$ is such that the variable

$$\beta = \log_{10}(\alpha - 1) \tag{24}$$

is drawn from a gaussian distribution of mean $\langle\beta\rangle$ and variance σ^2,

$$\rho_\beta(\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{\beta - \langle\beta\rangle}{\sigma}\right)^2\right\}. \tag{25}$$

Thus, we can use the relation $\rho_\alpha(\alpha)d\alpha = \rho_\beta(\beta)d\beta$, which holds for changes of variable in density functions, to write Eq. (23) as

$$g_m(m) = \int_{-\infty}^{+\infty} d\beta\, \rho_\beta(\beta)\, g_n\left(m;\alpha\left[\beta\right]\right). \tag{26}$$

Note that, even if the density function changes, Eq. (22) remains the same, but we only need to place the exact form of α as a function of β, $\alpha\left[\beta\right]$. Note also that in the particular case in which $\sigma^2 = 0$, the gaussian distribution becomes a delta distribution centred at $\langle\beta\rangle$ and Eq. (26) reduces to Eq. (19) for a unique value of α.

B. Results

Dropping the universality of α and implementing a mixture implies that α is no longer a parameter of the model.

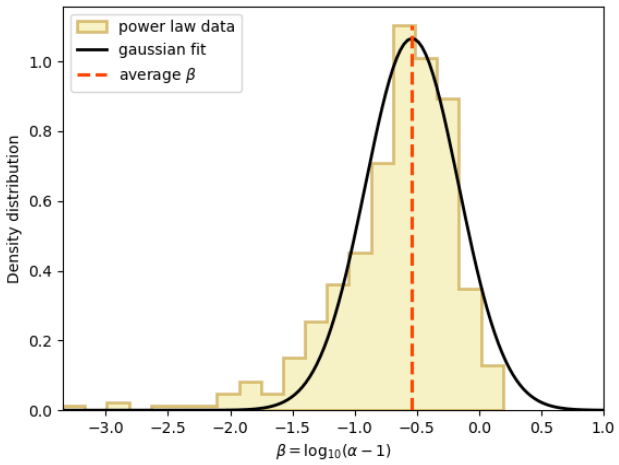

FIG. 4. Experimental probability density distribution of variable β and fit to a gaussian function. Yellow histogram: collected data for the β distribution from the fits to the power laws. The histogram is displayed with a binning of $B = 20$. Black line: gaussian fit to the histogram data. Eq. (25) is plotted for the optimal values $\sigma^2 = 0.14$ and $\langle\beta\rangle = -0.59$. The latter is shown as an orange dashed line. Two interesting behaviours can be appreciated. First, the empirical distribution seems more heavy tailed in the negative axis than the gaussian. Second, it falls abruptly for values of α over 2 ($\beta = 0$). The fit returns $R^2 = 0.94$.

Conversely, the assumption of a gaussian shape for the distribution of β implies the introduction of two new parameters: $\langle\beta\rangle$ and σ^2. Along with f_{min}, these are the parameters we need to fit now.

1. *Gaussian distribution from fit to power laws*

Looking at Eq. (26), one realises that both $\langle\beta\rangle$ and σ^2 are fully represented in the density function of β. Having data on this distribution should enable us to get the two parameters with a simple gaussian fit, leaving f_{min} as the only remaining parameter to fit directly from the mixture distribution and saving a decent amount of running time.

With this motivation, we have fitted each of the patients' repertoires to a power law distribution. To maximise the precision in these fits, we have closely followed the approach explained in [15]. A summary of the methods applied in there can be found in Appendix B 3. Seemingly, the code can be found in my GitHub repository.

The procedure described in Appendix B 3 has been applied to every repertoire and the whole collection of β estimates has been gathered. Then, we have fitted their distribution to the gaussian Eq. (25). The result is shown in Fig. 4 and yields the following parameters:

$$\begin{aligned} \langle\beta\rangle &= -5.39 \cdot 10^{-1} \\ \sigma^2 &= 1.40 \cdot 10^{-1}, \end{aligned} \tag{27}$$

and, consequently, $\alpha\left[\langle\beta\rangle\right] = 1.23$ and $\langle\alpha\rangle = 1.42$. Note

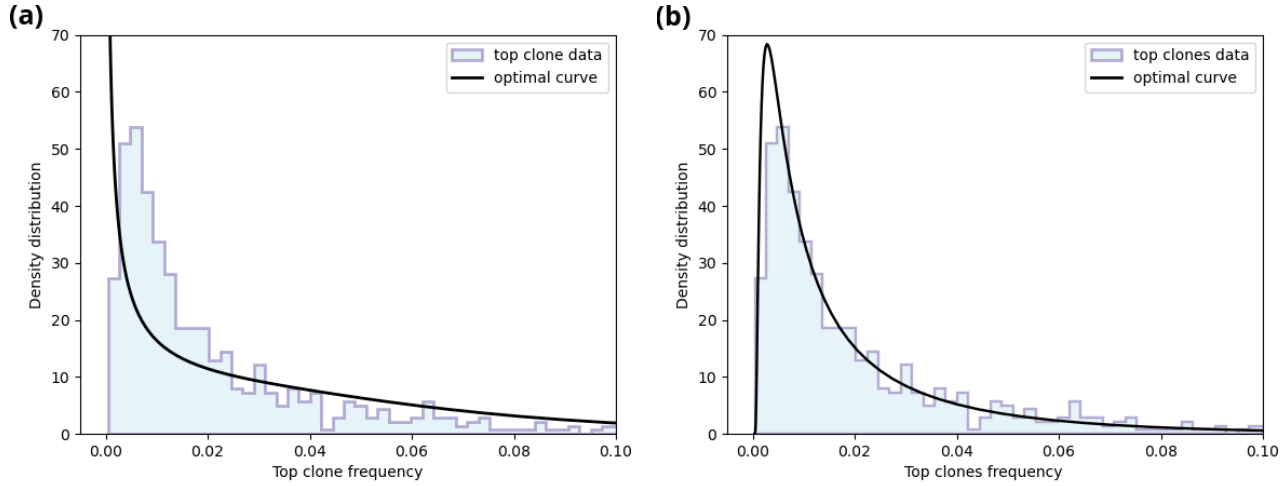

FIG. 5. Extreme value statistics of T cell top clones: non-universal power law exponents. (a) Fit of the mixture extreme value prediction to Emerson data, with gaussian parameters obtained from the fits to the power laws. The theoretical curve (black) (Eq. (26)) is plotted for the values $\langle\beta\rangle = -0.54$ and $\sigma^2 = 0.14$ obtained from the power law fits (see Figs. 4 and 9) and $f_{min} = 5.14 \cdot 10^{-9}$ obtained from the mixture distribution fit. An initial condition of $f_{min,0} = 10^{-10}$ is chosen. The fit returns a value of $R^2 = 0.54$, and is clearly misrepresenting the peak and part of the upper tail region. (b) Three-parameter fit of the mixture extreme value prediction to Emerson data. Again, Eq. (26) is plotted in black, now for the optimal parameters $\alpha = 1.49$, $f_{min} = 4.31 \cdot 10^{-4}$ and $\sigma^2 = 7.73 \cdot 10^{-2}$ obtained from the fit. The fit is done for initial values of $f_{min,0} = 10^{-10}$, $\alpha_0 = 1.2$ and $\sigma_0^2 = 10^{-3}$. The fit returns a value $R^2 = 0.97$, but is not physical in its predictions for the parameters. As in Fig. 2, the histogram of the Emerson top clones (blue) is displayed using a linear binning of $B = 200$ bins, and only the region of frequencies under 10^{-1} is shown.

that now these two only coincide if $\sigma^2 = 0$. In general, we will consider the best estimate of α not the average value, $\langle\alpha\rangle$, but the one evaluated at average β, $\alpha[\langle\beta\rangle]$.

The rather big value of σ^2 points out at a non-negligible diversity of exponents and serves as a check for the non-universality of α and the necessity of the mixture.

Similarly, Fig. 4 shows agreement between the experimental distribution and the optimal curve, which makes it easier to justify the choice of a gaussian distribution for the power law exponents. This is important to understand that, even if the particular analysis from Fig. 4 turns out to be incorrect, our hypothesis should still hold. In fact, assuming that α behaves randomly, one may expect it to respond to a gaussian behaviour. However, we know that α is physically limited to a lower bound of one and most of its values are concentrated in the region $1 - 2$. If, conversely, we assume a normal distribution for β, we get in return an skewed gaussian distribution for α. This choice is also naturally beneficial, since the use variable β not only extends α from $-\infty$ to $+\infty$, but also smoothens its behaviour significantly, which makes its study simpler. For example, now the region between $1 - 2$ for α becomes the whole negative real axis for β.

Afterwards, we have used these values inferred from the patient's repertoires to fit the mixture distribution to the top clones data. More detail on the numerical implementation of Eq. (26) and the fitting technique can be found in Appendix B 4. Additionally, the code for the integral and a fit example are accessible in my GitHub repository. The result is displayed in Fig. 5a. The optimal value of f_{min} returned by the fit is, in this case,

$$f_{min} = 5.14 \cdot 10^{-9}. \tag{28}$$

Even if the values obtained for f_{min}, $\alpha\,[\langle\beta\rangle]$ and σ^2 seem consistent with the foreseen ones, it is evident from Fig. 5a that Eq. (26) does not behave adequately at all, and our methods and premises need thus to be revised.

2. *Consistency of the gaussian distribution*

The disparity revealed in Fig. 5a points directly at the distribution of the power law exponents. In fact, a closer look at Fig. 4 shows that, while the normal distribution describes accurately the data for the exponents in the peak region, it underestimates the probability densities in the region of the smallest values. Given the high accuracy required for the fits of the repertoires to the power laws, we suspect that it is mostly this lack of precision what is causing the misbehaviour, but one could also think that a perfect gaussian is too simplistic and a more sophisticated distribution is necessary.

To convince ourselves that our guess is correct and that the normal distribution should be fine enough, we have applied the theoretical reasoning from Sec. VI A to the simple unconstrained distribution Eq. (8), for which non-mixed results (Fig. 2a) are already good, and fitted its

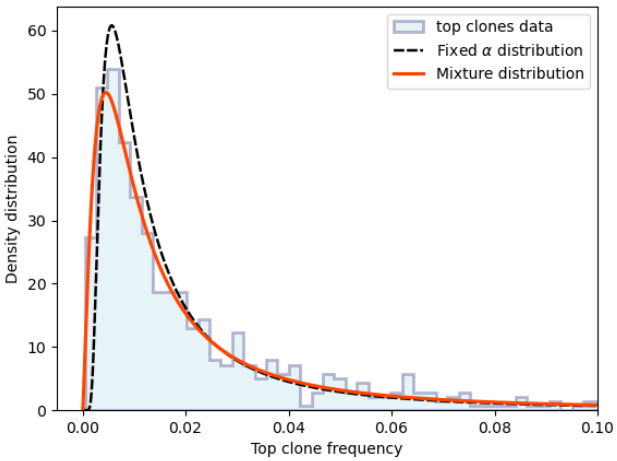

FIG. 6. Comparison of mixture distribution and α-universal distributions for the standard extreme value statistics of top clones. Red: mixture distribution of Eq. (8) evaluated at the optimal values $f_{min} = 6.64 \cdot 10^{-9}$, $\sigma^2 = 2.65 \cdot 10^{-2}$ and $\langle\beta\rangle = -0.71$ obtained from a three-parameter fit to Emerson data. The fit is accepted with value $R^2 = 0.96$. Black dashed line: same curve as in Fig. 2a. Comparison shows an improvement both in the returned parameters and the description of the peak behaviour.

mixture distribution to data, directly, without treating the exponents separately. The fit is shown in Fig. 6. The new theoretical curve (red curve) displays now a better behaviour with respect to the top clone distribution, especially in the peak region, and it also provides more accurate values for the parameters, namely

$$\begin{aligned} \langle\beta\rangle &= -7.10 \cdot 10^{-1} \\ \sigma^2 &= 2.65 \cdot 10^{-2} \\ f_{min} &= 6.64 \cdot 10^{-9}, \end{aligned} \tag{29}$$

and $\alpha\,[\langle\beta\rangle] = 1.03$, $\langle\alpha\rangle = 1.21$. The new f_{min} is closer to 10^{-8} than Eq. (10), as it should according to our estimates, and the new $\alpha[\langle\beta\rangle]$ is inside the interval $1 - 1.2$. Certainly, a gaussian distribution for the power law exponents betters the outcome of the universal exponent.

3. *Three-parameter fit*

Provided that the power law exponents cannot be shaped experimentally with enough precision and following the conclusions from Sec. VI B 2, we need to fit directly Eq. (26) to the top clone data, treating $\langle\beta\rangle$, f_{min} and σ^2 as free parameters at the same time. This is computationally a very expensive fit, whose result is shown in Fig. 5b. The parameters adopt values

$$\begin{aligned} \langle\beta\rangle &= -3.90 \cdot 10^{-1} \\ \sigma^2 &= 7.73 \cdot 10^{-2} \\ f_{min} &= 4.31 \cdot 10^{-4}, \end{aligned} \tag{30}$$

which imply $\alpha\,[\langle\beta\rangle] = 1.41$ and $\langle\alpha\rangle = 1.49$.

The black curve displayed in Fig. 5b seems now to follow the top clone data correctly, yet one immediately realises the results are, still, not acceptable. First, the value returned for $\langle\beta\rangle$ results in a big value for α, not outside its physical range, but sufficiently over the region of $1 - 1.2$. Second, even if the tail is now well replicated, the theoretical curve seems to overestimate the density value of the peak, causing it to be slightly displaced to the left with respect to the peak in the data.

Nonetheless, the most concerning fact comes from the value of f_{min}. This value is about five orders of magnitude out of the region it should be. More importantly, it is five orders above this region, which means that it is completely incompatible with any physical quantity. Such a big value for f_{min} implies that the lowest frequency of the repertoire would be greater than a significant number of frequencies actually observed in the subsampled repertoire and, hence, of a likely immense number in the whole organism. In fact, from the fit one realises that the value returned is essentially the lowest value for the top clone frequency, which we are using as the upper bound of the fitting procedure, and that the outcome is, again, very sensitive to the initial values chosen.

In the end, even if the approach followed here enables to obtain more agreeing representations, the quality of the parameters is significantly lowered. Definitely, the three-parameter fit is not improving the results of Sec. VI B 1 as expected.

VII. DISCUSSION AND FURTHER RESEARCH

In Secs. V and VI we have developed, step by step, a full extreme value theory for the statistics of top clones of T cell repertoires that takes into account the strictly finite size of the immune repertoires and the normalisation of the frequencies. Such a theory should better the results from the simple case in Sec. IV, in which we decide to ignore this constraint, and therefore provide more accurate estimates on the parameters that shape immune repertoires, namely, the minimum frequency of the clonotypes, f_{min}, and the exponent of the power laws that describe their statistics, α. Nothing further from the reality, the results obtained with this prescription not only do not improve the previous ones, but in most cases either they do not properly describe the statistics or do not return physical values for the parameters.

In this section, we assess the possible causes for such inconsistencies. Assuming that there are no calculation mistakes, we have come up with a list of possible incongruities arising either from wrong biological hypothesis, wrong modelling or wrong numerical implementations. We also try to suggest some possible corrections.

A. Frequencies instead of clone sizes

Using clonotype frequencies as variables makes the comparison with experimental data more reliable (see Sec. III C), but at the same time implies some theoretical complications that may be easily avoided rewriting the problem with clone sizes instead.

Typical extreme value theories deal with the statistics of variables whose extremes are large in absolute numbers [7]. This, that is valid for clone sizes, does not happen for frequencies, for the top clones we are describing are huge in comparison to the rest of the repertoire, but constrained to values smaller than one.

Working with clone sizes provides an equivalent rewriting of the problem which brings us closer to typical extreme value problems and, more importantly, to formally similar studies which go beyond the standard extreme value statistics and have a stronger theoretical basis, and to which our results could be compared. Examples are [10, 11] and, especially, [9].

Besides, we hope that clone sizes would help to achieve further analytical progress. To gain a quick intuition on this: while normalisation of the power law in frequency space requires two parameters, α and f_{min}, in clone size space it would only require α. With clone sizes instead of frequencies large behaviour limits could be applied to the top clone size which do not hold for the top clone frequency. Comparing with [9], we see that this may lead to significant simplification of the equations.

B. Parent distribution of the clonotypes

Power law distributions characterise the bulk of the repertoires, but this description becomes imperfect below some clone size threshold (see Fig. 1). However, our theoretical reasoning does not really account for this deviation from the power law behaviour. On the other hand, this problem has been taken into account in the numerical analysis done in Appendix B 4 by removing the top ranked frequencies, but different studies point out that small clones may reveal useful information of the most frequent ones [5]. Thus, being able to study the whole sequence of clonotypes might be important, something which could be assessed with a more complex model for the clonotype dynamics.

Examples of this can be found in [5, 6]. In the first case, the power law behaviour is derived from simplified models of clonotype population dynamics, ignoring the small clone dynamics under a decided threshold, and improved behaviours are attempted through refinement of such models. A power law density with exponential cutoff, $\rho(C) \sim C^{-(1+\alpha)}e^{-C/C_{min}}$, is derived and compared to data. Although it seems to be more precise, it ends up revealing itself nonphysical in the theoretical interpretation of α. In the second case, an advanced model is proposed by splitting the whole repertoire dynamics in two regimes: above the threshold, clonotypes follow a generalised power law behaviour, whereas under the threshold they are modelled using a truncated Gamma distribution.

Note, however, that use of such precise clonotype dynamics would turn the simple calculation in Sec. IV to an analytically non-solvable one, for which the whole problem becomes significantly more complicated. In addition, nothing ensures that the modifications or new parameters introduced in such an approach would have an impact on the results or a real biological meaning, likely leading to further congruence problems.

C. Definition of the mixture

In principle, the diversity of power law exponents in the cohort could be assessed in many reasonable ways, and different distributions may result in adequate descriptions. The reason we have used a gaussian distribution is because it is the simplest generalisation of the Dirac delta (universal exponent).

Nevertheless, we have realised in Sec. VI B 1 that the gaussian distribution tends to underestimate the influence of values of α close to one, where in turn one expects to find the average and best estimates for α. Along with the negative results from the same Sec. VI B 1 and Sec. VI B 3, these suggest that, even if the normal distribution seems the most natural choice, it may be a bit forced and a different solution would need to be found. Such solution could range from adding one or more new parameters in the gaussian exponent that slightly displace the curve towards negative values of β, to the use of known skewed or modified gaussian distributions. It is true that the analysis of Sec. VI B 2 seems to claim the opposite, but one could think that the better results in there are mostly due to the already good results for the unconstrained model (see Sec. IV), while further improvement is needed otherwise, considering that the non-mixed results for the normalised model (see Sec. V) are poor.

In any case, the introduction of such distributions might not have a physical interpretation and does not ensure beforehand any better results. In conclusion, although finer mixtures could be devised, we believe that, if the model for the normalisation were correct, the mixture as defined in Sec. VI A should automatically generate good results.

D. Wrong or unnecessary imposition of the normalisation

The results displayed in Fig. 6 both for the α-universal and mixture distributions for an independent frequency model are quite good. This seems to suggest that the normalisation of the frequencies, if not fully neglected, should not be considered to have a strong influence. However, with the delta function introduced as in Eq. (11),

the normalisation condition is simple, but also very rigid.

This rigidity comes mainly from the fact that the normalisation is introduced in such a way that enforces conservation of the total number of cells and fixes N_{cell} to one specific value. However, even if the number of cells is very conserved across patients, one expects to find some variability. Our intuition is that other models in which the normalisation is introduced "softly", allowing for some flexibility in the total number of cells during the generation process, may have a stronger biological basis.

There exists also the possibility that attempting to include the normalisation in the theory is actually unnecessary. As a sort of Occam's razor, if the normalisation over-complicates the problem, maybe the correct solution is to neglect its effects. In any case, further research in the different corrections suggested here needs to be done before being able to conclude this.

E. Alternative model for the normalisation

We want to end the discussion by proposing an alternative model for the normalisation. Even though we haven't had time to work on it, we believe that much of what has been done throughout the work could be profited in there.

As briefly hinted in Sec. VII D, we can understand the normalisation introduced in Eq. (11) as a statement for the conservation of the total number of cells. The immune system generates clonotypes drawing them arbitrarily from their parent distributions, but then enforces the produced sequences to contain exactly N_{cell} total cells. Hereby, N_{cell} acts as a rigid constraint during the generation process, such that clonotypes are somehow intentionally selected to have the exact size needed for the total size to sum up to N_{cell}. Indeed, one would expect that such a mechanism has an strong effect on the statistics of the top clones.

The model we introduce here proposes an alternative normalisation procedure. According to this model, the immune system would generate repertoires by drawing N different clonotypes of sizes $(C_1, \dots, C_N)$, each from its correspondent parent distribution, the same way it was done before, but normalisation is imposed now by simply defining the (normalised) frequencies, f_i, as the sampled clone sizes divided by the total number of sampled cells. In other words, clonotypes are now selected without any restriction on their size or the total number of cells, and the obtained outcome is then renormalised with respect to N_{cell}. The resulting joint probability reads:

$$P(f_1, \dots, f_N) = \int \prod_{i=1}^{N} \left[dC_i \, \rho(C_i) \, \delta \left(f_i - \frac{C_i}{\sum_{j=1}^{N} C_j} \right) \right]. \tag{31}$$

This model introduces an analytical complication in the sense that now sum (integration) over all the possible values of C_i is needed in order to cover for the whole set of realisations. The advantage comes from the fact that, stated like this, the model is much less restrictive in the formation of immune repertoires than the previous one. Loosely speaking, normalisation is now imposed after the creation of the clonotypes and not in the meantime. This way, N_{cell} is respected, but in a manner such that some variability in the number of cells is naturally introduced during the generation process.

In the end, normalisation is ensured, and N_{cell} is provided with the same physical interpretation as in the previous model; but all of this is done following a softer and, therefore, more biologically consistent approach. In principle, this should translate into less influence over the top clone statistics and more realistic results.

In this section we have suggested some minor modifications that should help in the final result; but much development and refinement of the formalism, assumptions and methods has been done without any significant improvement. In addition, these modifications may not lead to any worthy correction or, even if they do, they may lack a satifying biological interpretation. At this point, we should accept that the model was wrong from the beginning, and hope that a new model for the normalisation (this one or a different one) brings solutions to the issues here discussed.

VIII. CONCLUSIONS

In this work, we have studied the statistics of the more frequent clonotypes of T cell immune repertoires (top clones). To that end, we have identified their nature as extreme biological events and we have analysed their behaviour in the framework of an extreme value theory.

We have started by presenting the typical formulation and the main results of a theory of extremes, and interpreting them in the context of T cell repertoires, stating some important assumptions over the relevant biological parameters (Sec. III).

Provided the necessary theoretical basis, we have proposed different models to describe the statistics of the top clones. Our first model has treated frequencies inside the repertoire as independent. Under such conditions, we have derived the theoretical probability distribution for the top clones and fitted it to the Emerson data (Sec. IV).

Independent frequencies ensure good results, yet such a model may not be consistent. In fact, evidence suggests normalisation of the repertoire frequencies is required, imposing a constraint over their total sum. With this in mind, we have implemented a second model enforcing normalisation via a delta function over the whole set of frequencies. However, the subsequent probability distribution has revealed inaccurate and unable to furnish consistent estimates for the parameters (Sec. V).

In order to correct this model, we have removed the universality of the power law exponents and considered instead that they are normally distributed. First, we have graphed the gaussian distribution directly from the

power laws, but the subsequent mixture distribution has returned very poor results. Consequently, we have compared the mixture distribution directly to the Emerson data. Unlike the previous case, the theoretical distribution seemed agreeing with the experimental distribution, but, ultimately, the parameters found have revealed again non physical (Sec. VI).

In summary, we can only but conclude that the models introduced here are not appropriate to describe the extreme value statistics of T cell clonotypes. Even if normalisation is required, the results seem to suggest that its effect in the distribution of the top clones may not be very strong. A model ignoring normalisation is inconsistent on a biological basis. Conversely, normalisation, stated in the simple way we have proposed in our model, applies a too rigid constraint over the full repertoire; thence, alternative models in which such constraint is managed in a less restrictive way should work out better predictions. Accordingly, further research has been proposed for one specific of such models.

Appendix A: Derivation of the statistics of the extremes of T cell clonotypes with independent frequencies

Factorisation of the joint probability allows to write Eq. (5) in the compact form

$$G_N(m) = \left(\int_{f_{min}}^{m} df\, \rho(f) \right)^N . \tag{A1}$$

Placing the parent distribution Eq. (4), this integral can be easily solved explicitly, yielding

$$G_N(m) = \left(1 - \left(\frac{m}{f_{min}} \right)^{-\alpha} \right)^N . \tag{A2}$$

Note that $m \gg f_{min}$ and $\alpha \geq 1$, which implies that $(m/f_{min})^{-\alpha} \ll 1$. Hence, even if m is upper bounded to one, the important quantity is its relative value with respect to f_{min}, which behaves as a typical extreme value. We can then use the fact that $e^{-x} \approx 1 - x$ for $x \ll 1$ to write the CDF as

$$G_s(m) \simeq \exp\left(-N \left(\frac{m}{f_{min}} \right)^{-\alpha} \right). \tag{A3}$$

The exponential approximation in Eq. (A3) could be also seen as an exponential limit in the limit of large N. Eq. (A2) converges rapidly to Eq. (A3) as N increases. Given the typical sizes of the repertoires discussed in Sec. II B, this condition is implicitly assumed and Eq. (A3) is a good approximation. The interest of such an approximation relies in the fact that, under the linear transformation

$$z = N^{-\frac{1}{\alpha}} \left(\frac{m}{f_{min}} \right), \tag{A4}$$

the CDF distribution as expressed in Eq. (A3) reduces to a universal Fréchet distribution,

$$\widetilde{G_s}(z) = \exp\left(-z^{-\alpha} \right), \tag{A5}$$

something which is not possible otherwise. In the end, in spite of some small particularities, assuming independent frequencies makes of the T cell top clone statistics a simple Fréchet-like extreme value problem. Note, nonetheless, that this renormalisation requires knowledge on f_{min} and α that we do not have before the fits, so we cannot really benefit from the renormalisation.

Appendix B: Numerical methods

1. Repertoire data analysis

In first place, the whole dataset has been filtered to exclude all unnecessary information for the problem, keeping only the data on the T cells names (amino acid sequences of their TCRs), their frequencies and their number of counts.

The sampled repertoires contain essentially two defects that need to be corrected. One is that in most of the repertoires the frequencies are not normalised. This is probably due to the exclusion of some non-productive sequences, but not all of them. Hence, one needs to exclude all the non-productive sequences by tracking the presence of special characters (mostly $*$ or _) in their amino acid sequences. The other one is that some amino acid sequences appear multiplied in the data chain. This is probably due to the fact that they have been measured in separate times, and processed, without any prior knowledge, as separate TCRs, or to the fact that they come from to different nucleotide sequences. To fix this, we simply look for coincidences in the amino acid sequences and sum the frequencies of the identical clones. Then, we divide all the frequencies by their total sum in order to normalise them.

Last, we run over every patient's repertoire to find the most frequent clone and mark it as the top clone of the patient. This yields a collection of 632 top clones to compare with our calculations.

2. Extreme value statistics of T cell clones with normalised frequency sum

Here we summarise the more remarkable aspects of the numerical implementation of both the integral in Eq. (20) and the posterior fit of Eq. (19) to the Emerson data.

a. Integration over the frequencies

The integral $\langle e^{i\mu f}; m \rangle$ cannot be performed directly due to the typically small relative values of f_{min} with respect

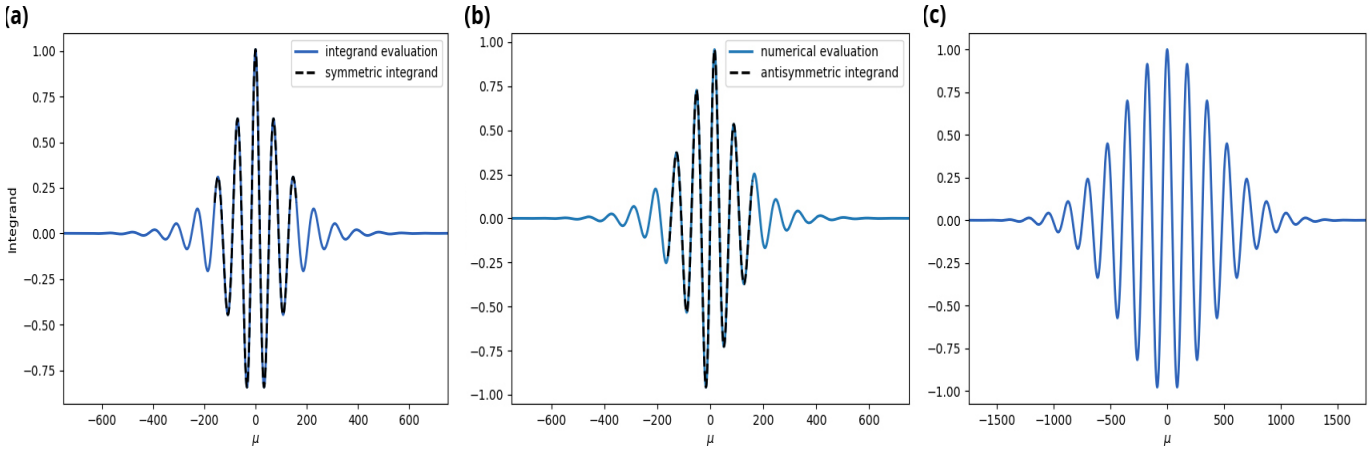

FIG. 7. Representation of the integrand in Eq. (20) as a function of μ, for different values of m. (a) Real part of the integrand for big a value of m. Blue line: evaluation of the integrand at $m = 10^{-1}$. Black line: same integrand as a function of $-\mu$. The result shows that the real part of the integrand is a symmetric function. As it can be seen, the exponential term of the integrand converges rapidly to zero for big values of m. (b) Imaginary part of the integrand for the same value of m. Black line: a similar approach proves that the integrand is now antisymmetric. (c) Integrand for a small value of m. Concretely, it is evaluated at $m = 5 \cdot 10^{-4}$. The range for the oscillations of the integrand increases rapidly with the decrease of m. Parameters used: $\alpha = 1.2$, $f_{min} = 10^{-10}$.

to m. To fix this, we have split the integration interval in n points logarithmically spaced between f_{min} and m. Logarithmic spacing reduces significantly the relative difference between lower and upper (sub)integration limits in the region close to f_{min} by creating a bigger concentration of points than in the big frequency regions. The resulting integral is given by

$$\langle e^{i\mu f}; m\rangle = \sum_{k=1}^{n-1} \int_{f_k}^{f_{k+1}} df\, \rho(f) e^{i\mu f}, \tag{B1}$$

where $f_1 = f_{min}$ and $f_{n-1} = m$. The lowest valid division is $n = 10$.

b. *Fourier transform: integration over μ*

The main difficulty in the integration over μ in this context comes from the complex-valued integrand. The integrand in Eq. (20) can be split into its real and imaginary parts. In Fig. 7 we have numerically studied its behaviour separately. We observe that the real part is even with respect to the integration variable, whereas the imaginary part is odd. This may be ensured by the fact that $g_n(m)$ needs to be real. Thus, the integral over the imaginary part vanishes, allowing us to write Eq. (19) only by its real part:

$$h(m) = \frac{1}{2\pi\langle 1; m\rangle Z} \times \int_{\mathbb{R}} d\mu \, \cos\left\{(N-1)\arg\frac{\langle e^{i\mu f}; m\rangle}{\langle 1; m\rangle} - \mu(1-m)\right\} \times \exp\left\{(N-1)\log\left|\frac{\langle e^{i\mu f}; m\rangle}{\langle 1; m\rangle}\right|\right\}. \tag{B2}$$

This eases significantly the task of Python in performing the integral and reduces in more than one half the compilation time. Then, compilation time can be further reduced with a suitable (mostly heuristic) choice among Python's default integrators.

Additionally, in Fig. 7, we have compared the behaviour of the integrand in Eq. (B2) for two representative small and big values of m, in order to set finite integration limits. As it can be seen, the exponential behaviour decays rapidly for values of μ over one thousand. With this, we have constrained the integration to the interval $\mu \in [-2000, 2000]$.

c. *Correct implementation of the integral*

Given the difficulties in the implementation and optimisation of the integral Eq. (19), it would not be surprising that our code is not working properly. This mistake would transfer afterwards to the evaluation of Eq. (26) and create an undesired domino effect.

In order to check this hypothesis, we have confronted the numerical evaluation of Eq. (19), done as explained in here, with numerically-simulated data for the top clones

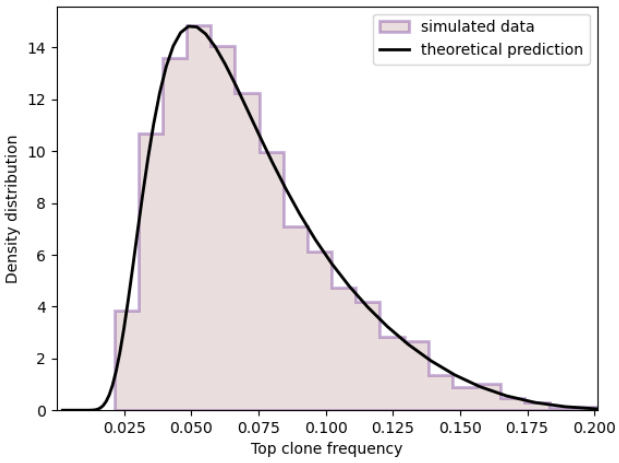

FIG. 8. Comparison of numerical integration to numerical data simulations. Violet: histogram of top clones data numerically generated by drawing from a power law distribution and imposing normalisation to one. This data corresponds to a joint distribution such as Eq. (11). Chosen binning: $B = 200$. Black: numerical integration of Eq. (19). Great agreement between both theoretical prediction and numerical data is observed. As an interesting feature, even if the distribution is wide in the peak region, the data falls is not fat tailed for high frequencies. Parameters used: $f_{min} = 10^{-5}$, $\alpha = 1.2$.

based on a delta-normalised model as the one stated via Eq. (11). These simulations generate very big sequences of clonotype frequencies drawn from a power law distribution, and discard those sequences that do not sum up to (approximately) one to impose normalisation. The computational cost is mostly delimited by the desired diversity, for which the simulation should be done for big values of f_{min} and typical values of α. The reader can refer to my GitHub repository for the code with a simulation example. In principle, if the integral is properly evaluated, both theoretical and generated distributions should match. One result is shown in Fig. 8. As it can be seen, both distributions match perfectly.

d. Fitting technique

Even with optimised integration, Eq. (19) displays very complex dependency on f_{min} and α. As a consequence, fitting takes very long time and Python default functions are unable to return any results.

This can be handled using `lmfit`, an advanced nonlinear least squares minimisation function fitter [20]. `Lmfit` is able to yield consistent values for the parameters on top of detailed statistical analysis on the quality of the fit.

Next, fitting time can be drastically reduced by implementing multiprocessing techniques and fitting in logarithmic space. By this, we precisely mean fitting to the distribution of the variable $\log_{10}(m)$, rather than m itself. In linear space, only a few points correspond to the peak region, whereas most of them belong to the tail and do not contain any relevant information. By fitting in logarithmic space, the narrow peak region broadens and the number of data points decreases by a factor ten (concretely from 200 to only 20 points), reducing the noise from the tail and allowing for a higher accuracy in the fit. With all these implementations, fitting time reduces from two to four hours to a couple of minutes.

On top of this, the precision of the results can be improved by fitting the logarithmic reciprocals of the parameters. Thus, the parameter f_{min} has been obtained by fitting $\log_{10} f_{min}$ and the parameter α by fitting $\log_{10}(\alpha - 1)$. The use of logarithmic variables enlarges the small differences between possible outcomes and increases the number of values accessible by the fitter. Based on the small interval into which α and f_{min} should fall, this is rather necessary.

3. Fit to power law distributions

Fitting power laws is usually a very hard task. Depending on the exponent, most of the points may be contained in the low frequency region and mostly determine the behaviour of the distribution. Subsequently, the accuracy of the result is very sensitive to subsampling, whose effects are difficult to predict but, at the same time, need to be analysed carefully. We can resume the different steps taken as three (again, obtained from [15]):

1. exclude the small frequency data, likely to be very influenced by the sampling process. We set a threshold of $C_t = 16$ to the minimum acceptable clone size and remove all the frequencies from the repertoire corresponding to smaller counts

2. display the remaining data in a logarithmic rank-frequency plot. The rank is essentially the integral of the density Eq. (4), thence,

$$\log \text{Rank} \sim K - \alpha \log f. \tag{B3}$$

   Not normalised cumulative distributions (rank) of normalised clone sizes (frequencies) are the best choice for collapsing the patients distributions.

3. fit the data to the rank-frequency curve (Eq. (B3)) with α and K as free parameters. The fit is only accepted for a quality of over $R^2 = 0.95$. Seemingly, the fit is rejected if $\alpha < 1$. We consider those cases not physical.

An example of a valid fit for an arbitrary patient is shown in Fig. 9.

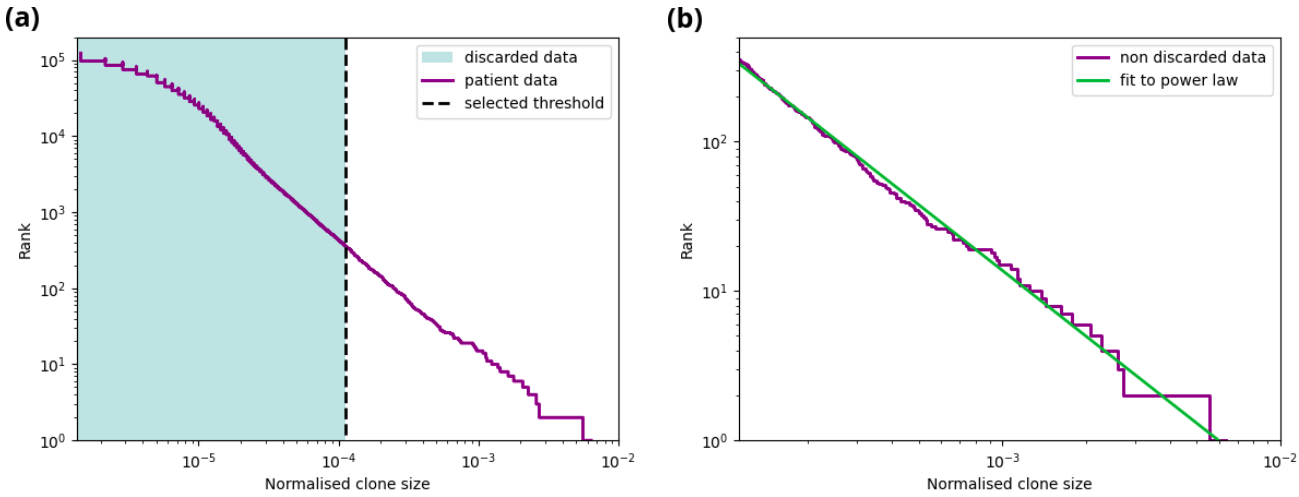

FIG. 9. Numerical analysis of the power law distribution of an arbitrary patient. (a) Data reading of the patient's repertoire. In purple: log-log plot of the patient's rank-frequency distribution. A universal threshold of $C_t = 16$ (black dashed line) is chosen. Note that its equivalent frequency changes from patient to patient due to differences in sampling size. Only the data above the threshold is considered to be a good representative of the power law behaviour. Below the threshold (blue shadowed region), the data is be very sensitive to sampling and exhibits strong deviation from the power law behaviour. As it can be seen, part of the power law region is discarded in order to improve the results. This way, only about 15% of the data is fitted. (b) Fit to the data above the threshold to a power law distribution. The remaining data is fitted to a power law rank-frequency distribution, as in Eq. (B3). Only the fits that provide $\alpha \geq 1$ with a significance of $R^2 \geq 0.95$ are accepted. This reduces the number of valid repertoires from 632 to 487. The resulting curve for the optimal values of α and K is plotted in green. In this case, the fit returns $\alpha = 1.46$ and $K = -3.24$, with $R^2 = 0.99$. Patient shown: HIP13722.

4. Extreme value statistics of T cell clones with variable power law exponents

The numerical methods described in Appendix B 2 still hold for the evaluation of the integrand in Eq. (26), but, given the complex dependency both of Eq. (22) and Eq. (25) on β, implementing the mixture is computationally very expensive and Python default integrators are either unable or quite slow in returning a result. More advanced integration tools need to be used.

The `differint` software package developed for Python provides a single repository for multiple numerical algorithms for the computation of fractional derivatives and integrals [21, 22]. Concretely, `differint` contains a Riemann-Liouville (`RL`) algorithm able to perform efficient and accurate numerical integral calculation via high-dimensional matrix operations.

Some minor changes are required in the original `RL` integrator in order to adapt it to the context of this problem. First, the integrand in Eq. (26) depends on multiple implicit and explicit parameters. We have modified the source code in order for RL to allow this. Second, we have slightly sped up the original algorithm by introducing just-in-time compilation via `Numba` in the matrix operations. We compare the final algorithm to standard `Scipy` integrators in Table I.

Integration via `RL` accelerates `Scipy`'s result by a factor $2-3$ without significant loss in precision. Thence, it seems to be the suitable function to integrate Eq. (26). Further optimisation is achieved by identifying in Fig. 4 the low density of β for values of $\beta > 1$, and setting a low upper integration limit to the mixture accordingly. On a different note, also the fitting technique of Appendix B 2 holds: we use again `Lmfit` as the fitting environment and we apply the same idea of logarithmic analogue variable to get the variance. With all these implementations, fitting time is drastically reduced from nearly three days with linear fitting and standard Python integrator to only four to six hours.

TABLE I. Comparison between `RL` and `Scipy` integration results for different m. Parameters: $\alpha[\langle\beta\rangle] = 1.2$, $f_{min} = 10^{-9}$ and $\sigma^2 = 10^{-2}$.

m	density function 10^{-3}	10^{-2}	10^{-1}	time[m]
`quad`	−−	−−	−−	∞
`fixed quad`	0.003291	1.026089	0.001786	5.7
`RL`	0.003260	1.026125	0.001783	2.0

[1] L. M. Sompayrac, *How the immune system works* (John Wiley & Sons, 2022).

[2] M. Moser and O. Leo, Vaccine **28**, C2 (2010).

[3] T. Mora and A. M. Walczak, Current Opinion in Systems Biology **18**, 104 (2019).
[4] G. Altan-Bonnet, T. Mora, and A. M. Walczak, Physics Reports **849**, 1 (2020).
[5] J. Desponds, T. Mora, and A. M. Walczak, Proceedings of the National Academy of Sciences **113**, 274 (2016).
[6] H. Koch, D. Starenki, S. J. Cooper, R. M. Myers, and Q. Li, PLoS computational biology **14**, e1006571 (2018).
[7] S. N. Majumdar, A. Pal, and G. Schehr, Physics Reports **840**, 1 (2020).
[8] A. Košmrlj, A. K. Chakraborty, M. Kardar, and E. I. Shakhnovich, Physical review letters **103**, 068103 (2009).
[9] M. R. Evans and S. N. Majumdar, Journal of Statistical Mechanics: Theory and Experiment **2008**, P05004 (2008).
[10] A. Bar, S. N. Majumdar, G. Schehr, and D. Mukamel, Physical Review E **93**, 052130 (2016).
[11] A. Shekhawat, Physical Review E **90**, 052148 (2014).
[12] T. Mora and A. M. Walczak, Syst Immunol , 183 (2018).
[13] V. I. Zarnitsyna, B. D. Evavold, L. N. Schoettle, J. N. Blattman, and R. Antia, Frontiers in immunology **4**, 485 (2013).
[14] J. A. Weinstein, N. Jiang, R. A. White III, D. S. Fisher, and S. R. Quake, Science **324**, 807 (2009).
[15] M. U. Gaimann, M. Nguyen, J. Desponds, and A. Mayer, Elife **9**, e61639 (2020).
[16] J. Desponds, A. Mayer, T. Mora, and A. M. Walczak, Mathematical, computational and experimental T cell immunology , 203 (2021).
[17] R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, *et al.*, Nature genetics **49**, 659 (2017).
[18] M. Bensouda Koraichi, S. Ferri, A. M. Walczak, and T. Mora, Proceedings of the National Academy of Sciences **120**, e2207516120 (2023).
[19] N.-p. Weng, Immunity **24**, 495 (2006).
[20] M. Newville, T. Otten, Renee Stensitzki, *et al.*, "lmfit/lmfit-py: 1.2.2," (2023).
[21] M. Adams, arXiv preprint arXiv:1912.05303 (2019).
[22] M. Adams, "differint/differint: 1.0.0," (2023).