PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Cite this article: Laydon DJ, Bangham CRM, Asquith B. 2015 Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Phil. Trans. R. Soc. B* **370**: 20140291.
http://dx.doi.org/10.1098/rstb.2014.0291

Accepted: 3 May 2015

One contribution of 17 to a theme issue 'Within-host dynamics of infection: from ecological insights to evolutionary predictions'.

Subject Areas:
computational biology, immunology

Keywords:
T-cell receptor repertoire, diversity, species richness

Author for correspondence:
Becca Asquith
e-mail: b.asquith@imperial.ac.uk

THE ROYAL SOCIETY PUBLISHING

Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach

Daniel J. Laydon, Charles R. M. Bangham and Becca Asquith

Section of Immunology, Wright-Fleming Institute, Imperial College School of Medicine, London W2 1PG, UK

DJL, 0000-0003-4270-3321; BA, 0000-0002-5911-3160

A highly diverse T-cell receptor (TCR) repertoire is a fundamental property of an effective immune system, and is associated with efficient control of viral infections and other pathogens. However, direct measurement of total TCR diversity is impossible. The diversity is high and the frequency distribution of individual TCRs is heavily skewed; the diversity therefore cannot be captured in a blood sample. Consequently, estimators of the total number of TCR clonotypes that are present in the individual, in addition to those observed, are essential. This is analogous to the 'unseen species problem' in ecology. We review the diversity (species richness) estimators that have been applied to T-cell repertoires and the methods used to validate these estimators. We show that existing approaches have significant shortcomings, and frequently underestimate true TCR diversity. We highlight our recently developed estimator, DivE, which can accurately estimate diversity across a range of immunological and biological systems.

1. Introduction

The human T-cell receptor (TCR) repertoire—the range of different TCRs expressed—plays a vital role in host defence. By recombination, random insertion, deletion and substitution, the small set of genes that encode the T-cell receptor has the potential to create between 10^{15} and 10^{20} TCR clonotypes (a clonotype is a population of T cells that carry an identical TCR) [1,2]. However, the actual diversity of a person's TCR repertoire cannot possibly lie in this range. There are only an estimated 10^{13} cells in the human body [3], and many clonotypes are of high abundance due to strong selection forces (for example, thymic education or antigen specificity). The actual, or realized, diversity of the human TCR repertoire remains unknown. The term 'diversity' is commonly used to mean either the number of classes (also known as 'species richness'), or the degree of dispersion among those classes. In this study, we use the term 'species' to refer to a single TCR clonotype, and 'diversity' to refer to the number of TCR clonotypes.

TCRs are heterodimers and fall into two classes: TCR-αβ and TCR-γδ; γδ T cells constitute 1–10% of the T-cell repertoire [4]. A variable (*V*), joining (*J*) and constant region (*C*) constitute the TCR α- and γ-chains. The TCR β- and δ-chains are also made up of a *V*, *J* and *C* region, with an additional diversity (*D*) region [5]. One segment from each region is recombined, with additional nucleotide additions and/or deletions, to generate each rearranged TCR (figure 1). This recombination generates high T-cell diversity [1] and enables the recognition of millions of antigens [6]. While *V*(*D*)*J* gene rearrangement is believed to be random [7], some clonotypes are produced more commonly than others [2,8], leading to unequal frequencies of naive T-cell clonotypes and to 'public' clonotypes, i.e. clonotypes shared between people. This unequal frequency distribution is believed to be due to a process known as convergent recombination, whereby certain nucleotide sequences can be produced using a greater variety of recombination events; certain amino acid sequences can be made by a greater number of nucleotide triplets;

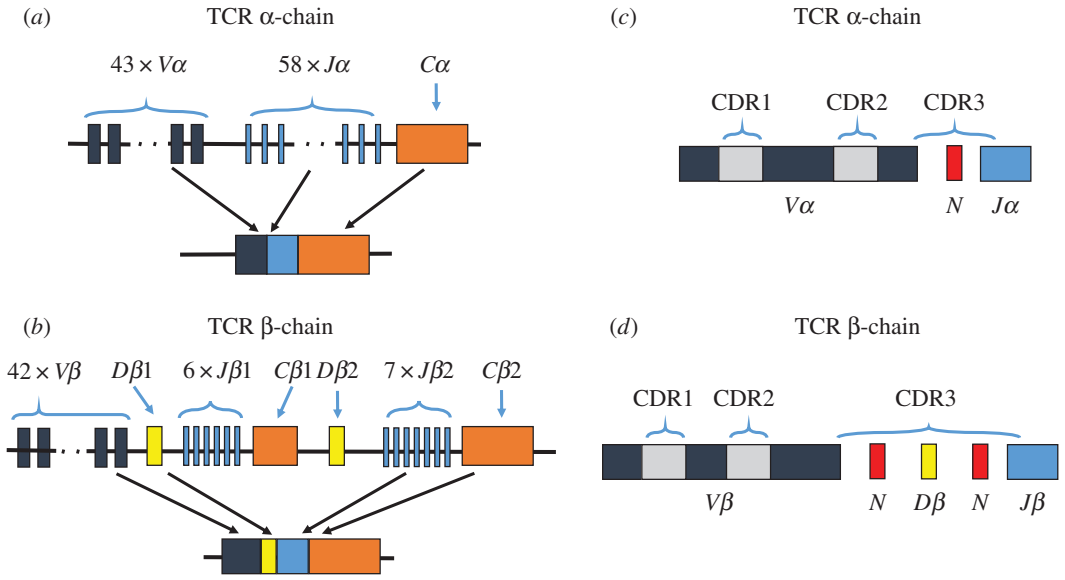

Figure 1. T-cell receptor gene rearrangement. (*a*) Variable (*V*), joining (*J*) and constant regions (*C*) constitute the TCR α-chain. (*b*) Variable (*V*), joining (*J*) and constant regions (*C*) constitute the TCR β-chain, with an additional diversity (*D*) region. Segments from each region are recombined, with additional nucleotide additions, to generate each rearranged TCR. These processes generate substantial T cell diversity. (*c*,*d*) Hypervariable complementarity-determining regions (CDR1-CDR3) of the α-chain (*c*) and β-chain (*d*). CDR1 and CDR2 regions are encoded on the *V* region, while the most variable CDR3 region straddles the *V*(*D*)*J* junction.

and certain TCRs require fewer nucleotide insertions, deletions or substitutions [9].

The third complementarity-determining region 3 (CDR3) of both the TCR α- and TCR β-chains straddles the *V*(*D*)*J* junction [10,11] (figure 1*b*), the primary site of antigen contact [5]. The CDR3 is the region most affected by recombination [12], and the CDR3 region of the β-chain accounts for most of the variation within a person's T-cell repertoire. Antigenic cross-reactivity of T cells results in a discrepancy between structural diversity (the number of different nucleotide or amino acid TCR combinations in the host) and functional diversity (the number of different antigens recognized by the T-cell repertoire) [1].

2. Why is T-cell receptor diversity important?

TCR diversity is associated with the effective control of viral infections and other pathogens [13–15]. The number of clonotypes observed in the blood in one person has been reported to decrease with age [16–19], viral challenge [15,20,21], immunization [22] and as a result of immune suppression, for example after haematopoietic stem cell transplantation (HSCT) [23]. TCR diversity has also been positively associated with autoimmunity in both mice [24,25] and humans [26]. Accurate quantification of diversity is important to assess the extent of immune convergence (sharing of clonotypes between people) [7,24,27–29].

Species diversity is also important in many systems outside T-cell immunology, for example, in estimating the repertoire of antibody classes [30,31], assessing the size of the metagenome in microbial communities [32,33] and measuring the rate of evolution of quasi-species of a pathogenic virus [34]. The original motivation for estimating diversity comes from population ecology, where the question of how many species there are in a given population gives rise to the 'unseen species problem': how many species are present, but unobserved, in the population of interest? Typically, there is a nonlinear relationship between the number of individuals (e.g. a T cell, a microbe) and the number of 'species' (e.g. a clonotype or viral variant), and so diversity cannot usually be estimated through linear scaling.

3. Why is estimating diversity difficult?

Estimating the diversity of the T-cell repertoire is difficult for many reasons. First, the repertoire is highly diverse. Given the number of T cells, (assumed to be of the order of 10^{12} [35,36]), a diversity of (say) 10^7 clonotypes [36] is unlikely to be directly observed owing to the limited volume of blood that can be taken from a person at any one time, and to the heavy-tailed frequency distributions with highly non-uniform clonotype abundances [19,37].

Second, the precise relationship between the diversity of different TCR-α and TCR-β sequences and the actual TCR functional diversity is unclear. Most recent studies focus on the CDR3 region [5,24,38–40], because it is the most variable region and because it is short enough to be captured in a single sequence read [30]. However, a T-cell receptor consists of pairings between either α and β chains or γ and δ chains; this adds a further level of diversity that is not routinely captured by many sequencing approaches. Furthermore, the relationship between TCR sequence and three-dimensional structural diversity and functional diversity are not fully understood [40,41].

Third, laboratory techniques that give absolute and unbiased estimates of clonotype frequency are technically challenging. Early studies measured TCR diversity qualitatively, where different clonotypes were identified visually as discrete bands on genomic southern blots [42–44]. Other approaches [45] used flow cytometry to measure the average observed frequency of each clonotype, reasoning that if this frequency was low then the population was more diverse.

Greater precision was achieved with spectratyping [22,46,47], where the number of different CDR3 lengths is used as a proxy for the number of clonotypes. The degree to which the frequency distribution of CDR3 lengths deviates

from normality is used as a metric of clonal expansion and thus of reduced diversity (because of limited lymphocyte capacity) [21,26]. Although this inference seems reasonable, expansion of some clonotypes does not imply the extinction of other clonotypes, merely their reduced relative frequency. Spectratyping produces incomplete sequence information [10] without further subcloning of the PCR product [2,48,49] which is low-throughput and labour-intensive [26,40].

High-throughput sequencing (HTS) allows greater sequencing depth and significantly more accurate quantification of TCR clonotype abundance [39], albeit at a greater expense than spectratyping [10]. However, HTS is still subject to PCR bias and sequencing error, with the consequences that clonotype abundances can be drastically distorted and that non-existent clonotypes can be recorded, thus falsely increasing the observed diversity [50].

4. Unbiased sequencing of T-cell receptor diversity is insufficient for diversity estimation

5′ rapid amplification of cDNA ends (RACE) is reported to suffer from markedly less bias than other HTS approaches [5,51]. Nevertheless, 5′ RACE (and unbiased sequencing more generally) is unlikely to be sufficient for diversity estimation. Diversity estimation usually makes use of two quantities: the relative abundances of observed species, and the extent to which each species is repeatedly observed in the sample. If PCR amplification is unbiased, then relative abundances will be preserved but the degree of repetition in the sample will not.

5. 'Exhaustive sequencing' cannot capture full repertoire diversity

Because all T cells within a sample of blood will not usually be detected in a single sequencing experiment, many researchers have used 'exhaustive sequencing' [37,38,52], i.e. the library is sequenced with the greatest possible depth, to maximize the number of reads per clonotype. It can then be justifiably concluded that further sequencing of the same library would not yield greater observed diversity. It is therefore tempting to conclude that the sample of blood contains a complete census of clonotypes in the periphery. However, such a conclusion would be false.

The principle that exhaustive sequencing does not capture full repertoire diversity was demonstrated by Warren *et al.* [52]. The authors exhaustively sequenced a library derived from a peripheral blood sample. However, upon sequencing a second library derived from the same blood sample, they found that 75% of the sequences returned were new, i.e. not contained in the first library. Furthermore, sequencing data were obtained from a second independent blood sample, and only 13% of the clonotypes observed in the second sample were observed in the first. This indicates that exhaustive sequencing of a single sample is incapable of capturing diversity, regardless of the apparent degree of repetition of species provided. That is, a saturating relationship between the number of reads and the number of clonotypes does not imply that there is a saturating relationship between the number of T cells and the number of clonotypes. The limiting factor is the number of TCRs present in the sample, not

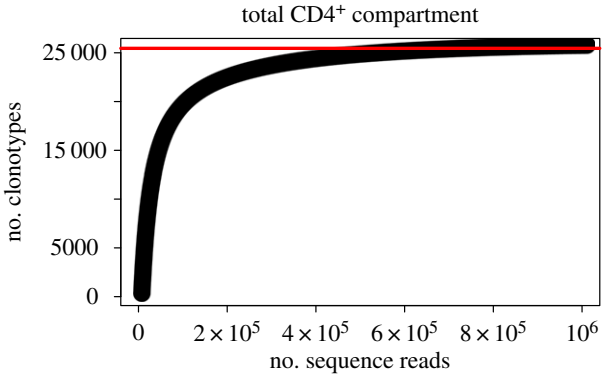

Figure 2. PCR amplification can lead to 'false saturation' of rarefaction curves. Example of 'exhaustive sequencing' of $CD4^+$ T cell compartment in a healthy donor. Unbiased sequence data was obtained through 5′ rapid amplification of cDNA ends (RACE) [53]. The rarefaction curve approaches saturation, falsely implying that further sequencing would not yield many additional clonotypes. However, the approximate saturation value of 2.5×10^4 is not a realistic estimate of total $CD4^+$ TCR diversity. For example, Robins *et al.* [47] frequently observed more than 10^5 clonotypes before estimating the number of unseen clonotypes. PCR amplification overestimates the repeated observation of TCR clonotype in the sample, leading to false saturation and substantial underestimates of TCR diversity. (Online version in colour.)

the extent of amplification or depth of sequencing. We have observed similar effects of 'false repetition' and 'false saturation' in our work [53]: figure 2 shows apparent saturation of the number of new clonotypes observed as the number of sequence reads increases. However, the number of clonotypes in the full data (2.5×10^4) is a drastic underestimate of TCR diversity, where between 10^5 and 10^6 distinct CDR3 sequences have been directly observed [47,52]. Finally, it has been noted [54] that exhaustive sequencing of either or both of the TCR α and β chains is insufficient to capture the full repertoire of a person.

6. Absolute T-cell counts are required for diversity estimation

Recent approaches have used DNA barcoding [29,50,51,55] or amplicon length discrimination [56,57] to resolve the problems of PCR amplification. Under DNA barcoding, a clonotype is identified by its nucleotide or amino acid sequence, but a second identifier is assigned to each individual short DNA sequence through the addition of a random DNA sequence label. Thus, the combination of a given clonotype nucleotide sequence and a given random label is unique. This allows identical T cells to be distinguished from identical sequence reads, and so preferential amplification is irrelevant. For example, if there are two amplicons that have identical CDR3 sequences and identical labels, it can be concluded that both amplicons have been derived from a single DNA sequence. The resulting data therefore consist of absolute—not relative—clonotype abundances, which are required for any abundance-based estimator. Furthermore, DNA barcoding can be extended to correct for sequencing error [50,55].

Another factor that prevents absolute quantification of TCR abundance is the sequencing of cDNA rather than genomic DNA, since a single T cell may express multiple mRNA copies. Therefore, cDNA is not suitable for diversity estimation.

Table 1. Comparison of diversity estimation approaches.

estimator	advantages	disadvantages
parametric (e.g. Poisson abundance models, Power laws)	can estimate clonotype frequency distribution	requires *a priori* assumptions on analytical form of clonotype frequency distribution lack of validation: goodness-of-fit to observed data does not confirm model accuracy
non-parametric abundance-based estimators (e.g. Chao1, ACE, capture–recapture)	no *a priori* assumptions required on analytical form of clonotype frequency distribution	cannot estimate clonotype frequency distribution biased by sample size inaccurate in highly diverse immunological populations
non-parametric incidence-based estimators (e.g. Chao2, ICE)	does not require absolute count data	lack of validation in immunological populations biased by sample size
DivE	accurate in multiple validations, across all immunological populations tested unbiased by sample size	time consuming: multiple models must be fitted

7. Unseen clonotypes: the problem

Even where data collection involves considerable sequencing depth, and where unbiased data have been obtained, estimators of the number of unseen clonotypes will need to be employed because of limits on blood volume that can be taken from donors. Several estimators of species richness (i.e. the number of species) developed in ecology have been applied to estimate TCR diversity, treating each clonotype as a 'species'. Such estimators fall into two broad categories: parametric estimators [58], where the shape of the species frequency distribution is assumed to follow some analytical form, and non-parametric estimators that make no such assumptions, and thus population frequencies cannot be inferred [59]. Since the true numbers of species or clonotypes are unknown, it is difficult to validate estimators of diversity, and so in common with ecological populations, it is often unclear which estimator should be used. We compare diversity estimators below and in table 1.

8. Differences between ecological and immunological data

There are important qualitative differences between T-cell repertoires and ecological populations in the uniformity of sampling. In ecological populations, data collection is frequently not random. For example, while placement of quadrats may be random, all of the individuals present in that quadrat are counted, leading to clustering of data [60]. Also, the probability of detection varies between species as it is influenced by colour, physical size, noise emission, geographical distribution, movement, variety of habitats and relationship to other species [61,62]. By contrast, in samples of T cells derived from blood, it is reasonable to assume that individual T cells have the same probability of detection; this assumption is less justifiable in solid tissue, as for example, lesions are non-randomly sampled.

In many ecological populations (e.g. plants, arthropods), the actual counting of individuals present in the sample is more straightforward than for populations of T cells, where sequencing introduces biases [19] and where it is difficult to distinguish sequencing errors from rare species [52]. The frequent implicit assumption that sequencing data comprised individuals that are equally detectable is often inappropriate. The probability that a given sequence read is recorded is conditional on two events: first, the probability that the T cell is sampled from blood, which is equal among T cells; and second, the probability that an amplicon from a T cell is amplified, which is not equal across all CDR3 sequences. This problem does not arise in ecological datasets.

The use of diversity indices developed in ecology that are used in T-cell repertoires is not restricted to species richness estimators. Similarity indices such as the Jaccard [63,64], Morisita-Horn [41,63], analysis of similarity (ANOSIM) [10] and dispersion metrics such as Simpson's diversity index [48,65], the Shannon entropy [20,66] and Renyi entropy [66] have been used to compare the TCR diversities between different people or between different T-cell phenotypes [65,67]. Many of the difficulties that arise in applying ecological species richness estimators to T-cell repertoires also confound the measurement of the extent of dispersion or similarity between repertoires, and ecological indices should be used with caution when analysing TCR repertoires.

9. Non-parametric abundance based species richness estimators

One of the most commonly used estimators is Chao1 [68] or its bias-corrected form (Chao1-bc) [69]. These estimators have been used to estimate TCR diversity in mice [70], and humans [71], making use of an amendment to the estimator [72] that takes account of the maximum upper bound of diversity.

The abundance-based coverage estimator (ACE) [73], which has been suggested as best practice [58] and is commonly used in ecology, has been used to estimate repertoire diversity in transgenic mice in the contexts of T-cell differentiation [64], and TCR specificity and self-recognition [63,74].

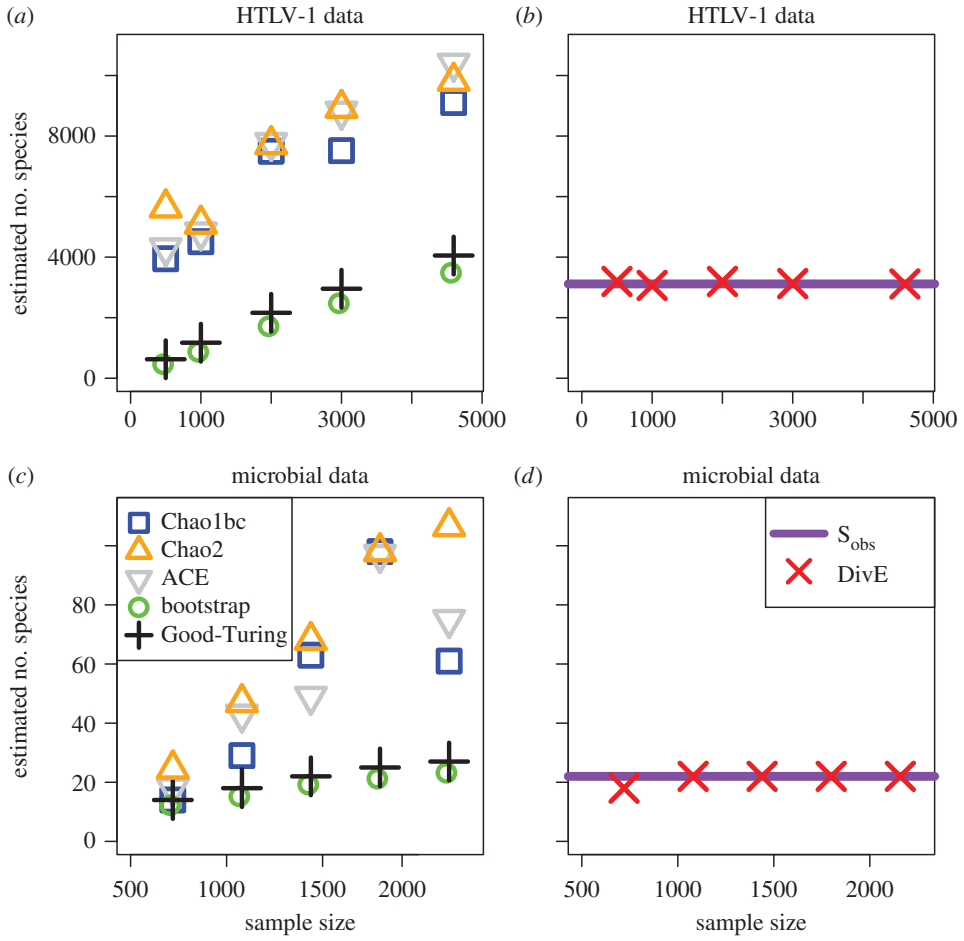

Figure 3. Performance of species richness estimators. (*a*,*c*) The Chao1bc (blue), Chao2 (orange), ACE (grey), Bootstrap (green) and Good-Turing (black) estimators are applied to *in silico* random subsamples of observed data. Examples for HTLV-1 and microbial data are shown. Estimates systematically increase with sample size. Chao2 estimates are calculated by randomly dividing each subsample into four *in silico* replicates. We observe the same bias with sample size where subsamples were divided into two and three *in silico* replicates (data not shown). (*b*,*d*) DivE (red) is applied to same subsamples as the other estimators. Performance of DivE was evaluated by comparing the error of estimates ($\hat{S}_{obs}$), to the (known) number of species S_{obs} in the full observed data (purple line) and by comparing estimates as a function of sample size. In all datasets, DivE accurately estimates the species richness of the full observed data from subsamples of that data and is unbiased by sample size.

However, Hsieh *et al.* [74] note that ACE is based on the probability that uniform sampling would produce the observed frequency distribution.

Weinstein *et al.* [30] used a capture–recapture approach to estimate the size of the antibody repertoire in zebrafish, and this approach was extended in Glanville *et al.* [31] to estimate antibody diversity in humans. The latter study also used a technique that allows sequencing of reads long enough to span all three CDR regions, which would allow more direct data on T-cell repertoires to be collected. No validation of the capture–recapture method was performed in either study.

Non-parametric abundance-based species richness estimators have been validated using ecological populations that have been extensively sampled and where approximate species richness is assumed to be known [75]. However, the accuracy of these ecological estimators has been questioned in immunological populations. We recently compared the performance of widely used non-parametric species richness estimators (the Chao1bc [69], ACE [73], Bootstrap [76] and Good-Turing [77] estimators) from population ecology when applied to immunological and microbiological systems [53]. We considered three distinct sets of data: the clonal distribution of cells naturally infected with human T-lymphotropic virus type-1 (HTLV-1), operational taxonomic units (OTUs) of Bifidobacteria in the gastrointestinal tract of infants, and T-cell receptor repertoires. In the case of HTLV-1, a 'species' is a clone, defined as a population of infected cells that share a genomic site of proviral integration.

For each set of data, we found that all estimators were biased by sample size (figures 3 and 4). This is problematic as estimates of species richness would increase if, for example, greater blood volumes were drawn or technique sensitivity was improved. Furthermore, there was strong evidence that the estimators underestimated diversity. Firstly, the estimators frequently produced estimates from subsamples that were lower than the diversity of the full observed sample. Secondly, in almost all cases, only a small number of unseen 'species' was predicted in addition to those observed. Such estimates are implausible in the HTLV-1 and T-cell repertoire datasets where there is such a vast potential diversity.

10. Non-parametric incidence-based species richness estimators

The incidence-based coverage estimator [78] was used to estimate the diversity of regulatory T cells in transgenic mice [64], although no validation of this estimator was performed.

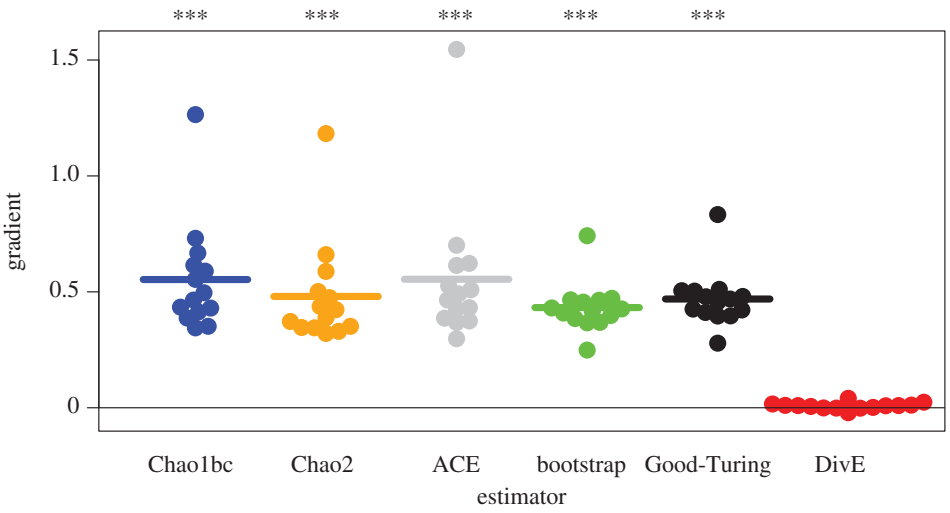

Figure 4. Comparison of estimators: effect of sample size on estimated HTLV-1 diversity. Gradients measuring increase in estimated HTLV-1 clonal diversity against increase in sample size. Gradients for each estimator were calculated by linear regression. All estimators except DivE show large gradients that are significantly positive, indicating a bias with sample size. ***$p < 0.0001$; two-tailed binomial test ($n = 14$).

Qi *et al.* [19] used multiple replicate libraries to compute the Chao2 estimator [79], resulting in an estimate of 10^8 clonotypes. The Chao2 estimator makes use of incidence (i.e. presence or absence) data, as opposed to abundance data, across different replicates. The estimator therefore provides a method of avoiding the distorted abundances due to PCR amplification mentioned above.

To validate their approach, the authors created an *in silico* Zipf distribution of clonotype frequencies from which to sample. They took random samples of varying sizes and found that the estimated diversity accurately estimated the number of clonotypes in their *in silico* distribution. Although indirect and using only one *in silico* distribution, this validation suggests that their method holds promise. However, we have applied Chao2 to HTLV-1 and microbial OTU data, and we again observed a bias with sample size, as seen with the other non-parametric estimators we tested (figures 3 and 4).

11. Parametric species richness estimators

Robins *et al.* [47] frequently observed as many as approximately 10^5 TCR clonotypes using single-molecule DNA sequencing, and employed a method originally devised by Efron & Thisted [80] and amended in Ionita-Laza *et al.* [81] to estimate a peripheral blood diversity of 3 to 4 $\times$ 10^6 clonotypes, including 1 $\times$ 10^6 antigen-experienced T-cell clonotypes, where the latter is approximately one order of magnitude higher than estimated previously [36]. Their method assumes that individual T-cell clonotypes enter the sample according to a Poisson process with clonotype-specific rates, which are inferred from the observed clonotype abundances. The method predicts the number of new sequences that would be observed in a subsequent sample. Hence their method does not merely provide an estimate of TCR diversity, but also the relationship between sample size and diversity. Therefore, the authors were able to validate their method. While this validation was direct, in that observation was compared with the predicted number of additional clonotypes, it was limited to only a single additional sample.

Several recent studies have made use of the class of Poisson abundance models (PAMs). Sepúlveda *et al.* [59] noted that species frequency data come from a multivariate hypergeometric distribution (i.e. a multinomial distribution where samples are taken without replacement). Because the size of a sample is dwarfed by the size of the total population (and therefore sampling does not drastically alter clonotype relative abundances), these authors approximated the multivariate hypergeometric distribution using a Poisson distribution. Incorporation of a varying sampling rate for clonotypes of varying frequencies leads to the class of PAMs [41,82]. They applied their method to previously published data on mice with different phenotypes, and evaluated the consistency of their method by excluding clonotypes above successively higher cut-off frequencies. Worryingly, there was wide variation in diversity estimates across all phenotypes depending on the specific PAM used. Rempala *et al.* [41] focused on one such model, the bivariate Poisson-lognormal distribution, and concluded that under-sampling in their repertoire datasets is more severe (and thus the population is more diverse) than would be estimated using the Good-Turing estimator [77]. Other extensions of the class of PAMs have been developed [41,82] that estimate the similarity between populations in the presence of unseen clonotypes.

Using a compound Poisson process model used originally to estimate gene capture diversity [83], Wang *et al.* [39] estimated TCR diversity in the context of T-cell fate and differentiation. This method can also model the relationship between the number of clonotypes and the number of T cells. They estimated approximately 10^6 unique TCRα and TCRβ CDR3 nucleotide sequences, approximately one third of that predicted by Arstila *et al.* [36]. Their method was validated previously in the context of gene capture diversity [83] using *in silico* distributions, choosing lognormal, exponential and gamma distributions of varying diversities. It is unclear how this validation translates to T-cell immunology.

In addition to the capture–recapture approach used in Weinstein *et al.* [30], Klarenbeek *et al.* [37] fitted multiple Poisson mixture models to HTS data to estimate β-chain diversity in the $CD4^+$ and $CD8^+$ T-cell compartments. Extending the distribution fitted to the observed data to model the number of unseen clonotypes, the authors estimated that

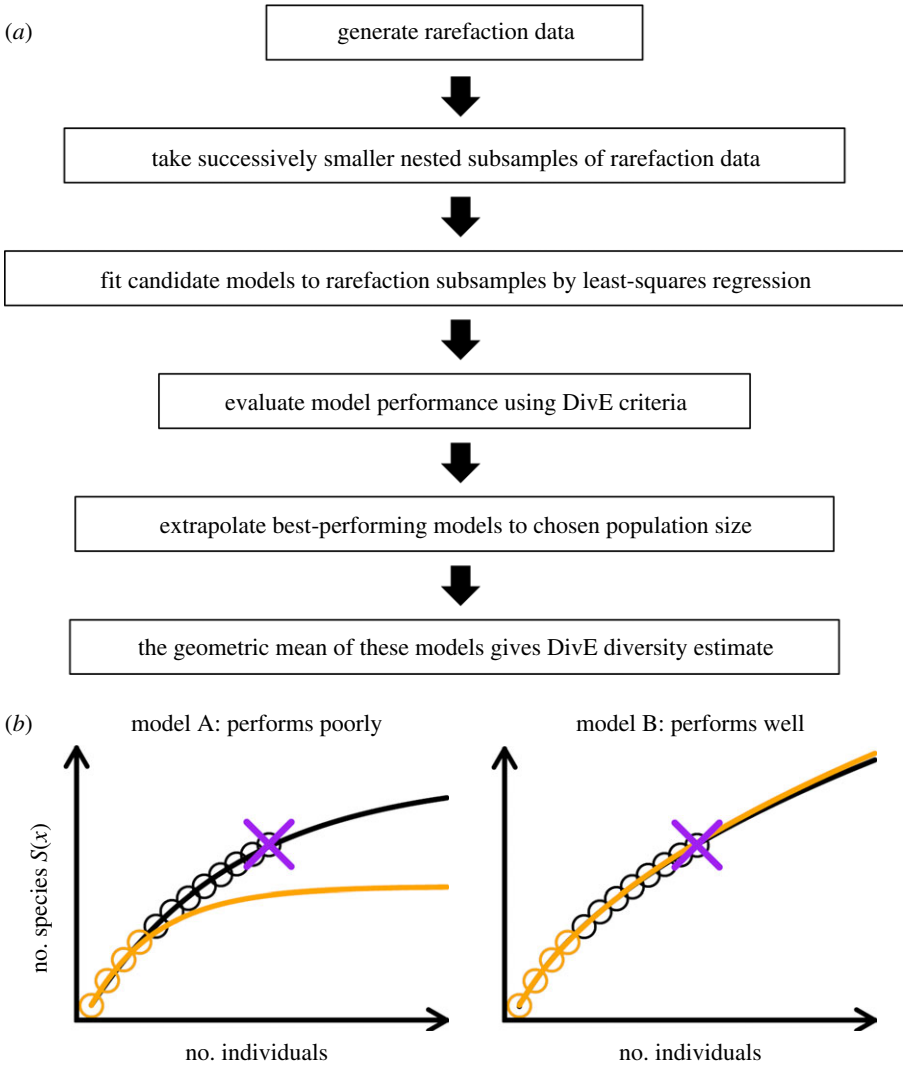

Figure 5. Outline of DivE species richness estimator. (*a*) Flow chart describing the process to calculate the DivE species richness estimate. (*b*) Full rarefaction curves shown in black and nested rarefaction subsample shown in orange. Data are denoted by circles, model fits by solid lines. Models are scored according to the following criteria: (i) discrepancy—mean percentage error between data points and model prediction; (ii) accuracy—error between full sample species richness (purple cross) and estimated species richness from subsample; (iii) similarity—area between subsample fit (orange) and full data fit (black) and (iv) plausibility—we require that $S'(x) \geq 0$ and $S''(x) \leq 0$. Model A performs poorly as criteria (ii) and (iii) are not satisfied. Model B performs well as all criteria are satisfied.

the memory compartment consists mainly of unexpanded clones and is far more diverse than thought previously [36] (only 2 and 3–10 times less diverse than the naive repertoire in $CD4^+$ and $CD8^+$ T-cell compartments, respectively). Their estimates are also remarkable in that they predict that more than 90% of memory clonotypes are relatively small.

Power laws have been used to model the form of the T-cell repertoire [84,85]. An advantage of this method is that the fitted parameters are relatively easy to interpret. It can be shown that one parameter quantifies the proportion of the repertoire occupied by clonotypes of a single T cell, and the other provides a measure of dispersion. Power law characterizations of the T-cell repertoire could be extended to estimate the number of unseen clonotypes in a similar manner to Klarenbeek *et al.* [37] by extending the modelled distribution.

Parametric approaches are often evaluated using goodness of fit to the observed data, for example using χ^2-tests or Akaike's information criterion (AIC_c) [30,59,86,87]. While these methods are useful for comparative purposes, they do not validate the resulting model's accuracy. A major limitation of all parametric approaches is that the estimated diversity is dependent on the assumed form of the clonotype distribution.

12. A new approach to T-cell receptor diversity estimation: DivE

We developed an estimator named DivE [53] which uses rarefaction curves (figure 5). Similar to a species accumulation curve, an individual-based rarefaction curve is created by cumulating the number of species as the number of observed individuals (e.g. a T cell) increases, in a single resample. Species counts are averaged over multiple resamples of the data to obtain the expected number of species as a function of the number of individuals. Sample-based rarefaction curves plot the expected number of species against the number of samples.

DivE involves fitting multiple simple mathematical models, many of which are well known in ecological studies [88,89], to rarefaction curves, and to nested subsamples of these curves. Novel criteria are then used to determine the most appropriate model; as well as assessing the quality of fit to seen data these criteria also assess the quality of fit to unseen data, i.e. how well a given model can predict the full dataset from random subsets thereof. The best-performing models are then aggregated and extrapolated to a user-specified population size to produce the diversity estimate (figure 5).

We used three methods to validate the performance of DivE. We measured the extent to which DivE could: (i) estimate the diversity of the observed dataset from subsamples; (ii) estimate from a single dataset the diversity of additional independent HTLV-1 data, obtained using separate blood samples taken in immediate succession and (iii) provide consistent estimates given samples of unequal size. In each validation, the estimator performed better than the non-parametric abundance-based estimators we tested (figure 3). We believe the principal reason that DivE performs well is that candidate models are selected on their ability to consistently predict additional rarefaction data. The additional data (i.e. the full rarefaction curve) have no influence on fitted parameter values, and so DivE not only assesses goodness of fit but also evaluates the accuracy of the model. DivE has been provided as an R package [90], available at http://cran.r-project.org/web/packages/DivE/index.html [91].

Accurate extrapolation of rarefaction curves assumes that the sampled population is representative of the whole population to be extrapolated to [60,92,93]. This is a reasonable assumption in the case of T-cell sampling in the blood, i.e. T cells sampled in one blood draw are likely to be representative of all T cells in the peripheral blood. However, this is a poor assumption when trying to infer the TCR diversity in the whole body as T cells sampled in the blood may not be representative of T cells in lymphoid tissue, etc. The difficulty of inferring total population diversity from estimates in the blood is not unique to DivE and will adversely affect the accuracy of all estimators.

An alternative approach to rarefaction curve extrapolation that is based on a rigorous statistical footing has recently been developed [94–96]. However, to estimate the rarefaction curve, this method requires an input of species richness (usually provided by ACE or Chao1), which is the quantity we seek to estimate. Furthermore, the authors of these papers caution that this method is not suitable for extrapolation beyond two- or threefold.

We summarize the advantages and disadvantages of different diversity estimation methods in T-cell repertoire analysis in table 1.

13. Discussion

To estimate repertoire diversity it is essential to obtain unbiased data, with absolute counts of TCR clonotypes. If unbiased absolute count data are not available, neither relative abundances nor the degree of repetition of observations are credible, and so diversity estimators should not be applied. While Chao2 does not require abundance data, we have found that this estimator too is biased by sample size in immunological and microbiological data (figure 3).

We also caution against estimating diversity using severely under-sampled data, whether due to limited sequencing depth or low blood volume. To quantify 'under-sampling', we previously defined a parameter based on the curvature of the observed rarefaction curve (see [53] for further details). A linear rarefaction curve implies an implausible constant rate of species accumulation. As sampling depth increases, the rate of species accumulation should decrease as previously encountered species are repeatedly observed. Abundance-based estimators should not be applied when the rarefaction curve is close to linear.

Recent advances in HTS combined with DNA barcoding mean that unbiased absolute count data is now increasingly available. However, because of the enormous potential diversity of the TCR repertoire and the limited amount of blood that can be drawn from a donor at any given time, there will almost certainly be unseen TCR clonotypes regardless of the precision of data collection. Therefore, estimators of diversity must be employed. Existing parametric estimators suffer from the requirement of an *a priori* form of the species frequency distribution. Furthermore, each non-parametric estimator we have tested, either abundance- or incidence-based, was significantly biased by sample size.

Absolute count data allow important simplifying assumptions to be made about the relationship between the observed data and the underlying T-cell repertoire, namely that individual T cells have been sampled independently, randomly and with equal detection probabilities. These assumptions in turn allow the extrapolation of models fitted to individual-based rarefaction curves. The question of which model to fit, however, is non-trivial. DivE selects which models are most appropriate based on their ability to faithfully reproduce all observed rarefaction data from subsamples, providing a degree of robustness that we have not observed with classical non-parametric estimators. Crucially, the form of the model chosen depends on the data, and so DivE does not require *a priori* assumptions regarding the form of the clonotype frequency distribution, or regarding the relationship between the number of T cells and the number of TCR clonotypes.

We have validated DivE across three independent immunological and microbiological systems. In all systems, the estimator was accurate, and considerably more so than the non-parametric estimators we examined. We believe that this estimator will be an important tool to estimate T-cell repertoire diversity.

Ethics. HTLV-1 and microbial data were obtained as previously described [53]. Ethics approval was given by the UK National Research Ethics Service (NRES references 09/H0606/106 and 05/Q1702/119 for the HTLV-1 and microbial data respectively). All subjects provided full written informed consent.

Authors' contributions. D.J.L. performed the literature review and analysed the data. D.J.L., C.R.M.B. and B.A. conceived and designed the study, and wrote and revised the article.

Competing interests. We have no competing interests.

Funding. C.R.M.B. is a Wellcome Trust Senior Investigator (100291) and is supported by the Medical Research Council UK (K019090), the Imperial College National Institute for Health Research Biomedical Research Centre, and Leukaemia and Lymphoma Research (12038). B.A. is a Wellcome Trust Investigator (103865) and is funded by the Medical Research Council UK (J007439 and G1001052), the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement 317040 (QuanTI) and Leukemia and Lymphoma Research (15012).

References

1. Nikolich-Zugich J, Slifka MK, Messaoudi I. 2004 The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* **4**, 123–132. (doi:10.1038/nri1292)

2. Miles JJ, Douek DC, Price DA. 2011 Bias in the [α][β] T-cell repertoire: implications for disease

pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375–387. (doi:10.1038/icb.2010.139)

3. Bianconi E *et al*. 2013 An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471. (doi:10.3109/03014460.2013.807878)
4. Girardi M. 2006 Immunosurveillance and immunoregulation by [γ][δ] T cells. *J. Invest. Dermatol.* **126**, 25–31. (doi:10.1038/sj.jid.5700003)
5. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. 2009 Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824. (doi:10.1101/gr.092924.109)
6. Wing JB, Sakaguchi S. 2011 TCR diversity and Treg cells, sometimes more is more. *Eur. J. Immunol.* **41**, 3097–3100. (doi:10.1002/eji.201142115)
7. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, Davenport MP. 2006 Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl Acad. Sci. USA* **103**, 18 691–18 696. (doi:10.1073/pnas.0608907103)
8. Venturi V, Price DA, Douek DC, Davenport MP. 2008 The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238. (doi:10.1038/nri2260)
9. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, Douek DC, Davenport MP, Price DA. 2010 Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proc. Natl Acad. Sci. USA* **107**, 19 414–19 419. (doi:10.1073/pnas.1010586107)
10. Epstein M, Barenco M, Klein N, Hubank M, Callard RE. 2014 Revealing individual signatures of human T cell CDR3 sequence repertoires with kidera factors. *PLoS ONE* **9**, e86986. (doi:10.1371/journal.pone.0086986)
11. Turner SJ, Doherty PC, McCluskey J, Rossjohn J. 2006 Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6**, 883–894. (doi:10.1038/nri1977)
12. Bercovici N, Duffour M-T, Agrawal S, Salcedo M, Abastado J-P. 2000 New methods for assessing T-cell responses. *Clin. Diagn. Lab. Immunol.* **7**, 859–864.
13. Messaoudi I *et al*. 2002 Direct link between MHC polymorphism, T cell avidity, and diversity in immune defense. *Science* **298**, 1797–1800. (doi:10.1126/science.1076064)
14. Davenport MP, Price DA, McMichael AJ. 2007 The T cell repertoire in infection and vaccination: implications for control of persistent viruses. *Curr. Opin. Immunol.* **19**, 294–300. (doi:10.1016/j.coi.2007.04.001)
15. Chen H *et al*. 2012 TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. *Nat. Immunol.* **13**, 691–700. (doi:10.1038/ni.2342)
16. Naylor K *et al*. 2005 The influence of age on T cell generation and TCR diversity. *J. Immunol.* **174**, 7446–7452. (doi:10.4049/jimmunol.174.11.7446)
17. Yager EJ, Ahmed M, Lanzer K, Randall TD, Woodland DL, Blackman MA. 2008 Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J. Exp. Med.* **205**, 711–723. (doi:10.1084/jem.20071140)
18. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK. 2013 Human lymphocyte repertoires in ageing. *Curr. Opin. Immunol.* **25**, 511–515. (doi:10.1016/j.coi.2013.07.007)
19. Qi Q *et al*. 2014 Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl Acad. Sci. USA* **111**, 13 139–13 144. (doi:10.1073/pnas.1409155111)
20. Meyer-Olson D *et al*. 2004 Limited T cell receptor diversity of HCV-specific T cell responses is associated with CTL escape. *J. Exp. Med.* **200**, 307–319. (doi:10.1084/jem.20040638)
21. Long SA, Khalili J, Ashe J, Berenson R, Ferrand C, Bonyhadi M. 2006 Standardized analysis for the quantification of Vβ CDR3 T-cell receptor diversity. *J. Immunol. Methods* **317**, 100–113. (doi:10.1016/j.jim.2006.09.015)
22. Höhn H, Neukirch C, Freitag K, Necker A, Hitzler W, Seliger B, Maeurer MJ. 2002 Longitudinal analysis of the T-cell receptor (TCR)-VA and -VB repertoire in CD8+ T cells from individuals immunized with recombinant hepatitis B surface antigen. *Clin. Exp. Immunol.* **129**, 309–317. (doi:10.1046/j.1365-2249.2002.01841.x)
23. Muraro PA *et al*. 2014 T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *J. Clin. Invest.* **124**, 1168–1172. (doi:10.1172/JCI71691)
24. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, Chain B, Cohen IR, Friedman N. 2014 T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* **24**, 1603–1612. (doi:10.1101/gr.170753.113)
25. Ferreira C *et al*. 2009 Non-obese diabetic mice select a low-diversity repertoire of natural regulatory T cells. *Proc. Natl Acad. Sci. USA* **106**, 8320–8325. (doi:10.1073/pnas.0808493106)
26. Matsumoto Y, Matsuo H, Sakuma H, Park I-K, Tsukada Y, Kohyama K, Kondo T, Kotorii S, Shibuya N. 2006 CDR3 spectratyping analysis of the TCR repertoire in *Myasthenia gravis*. *J. Immunol.* **176**, 5100–5107. (doi:10.4049/jimmunol.176.8.5100)
27. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH. 2010 Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Translational Med.* **2**, 47ra64. (doi:10.1126/scitranslmed.3001442)
28. Venturi V *et al*. 2011 A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* **186**, 4285–4294. (doi:10.4049/jimmunol.1003898)
29. Mamedov IZ *et al*. 2011 Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.* **3**, 201–207. (doi:10.1002/emmm.201100129)
30. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. 2009 High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810. (doi:10.1126/science.1170020)
31. Glanville J *et al*. 2009 Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl Acad. Sci. USA* **106**, 20 216–20 221. (doi:10.1073/pnas.0909775106)
32. Temperton B, Giovannoni SJ. 2012 Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.* **15**, 605–612. (doi:10.1016/j.mib.2012.07.001)
33. Shakya M, Quince C, Campbell JH, Yang ZK, Schadt CW, Podar M. 2013 Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* **15**, 1882–1899. (doi:10.1111/1462-2920.12086)
34. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. 2006 Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348. (doi:10.1038/nature04388)
35. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002 Lymphocytes and the cellular basis of adaptive immunity. In *Molecular biology of the cell* (eds B Alberts *et al*.), 4th edn. New York, NY: Garland Science. Available at http://www.ncbi.nlm.nih.gov/books/NBK26921/.
36. Arstila TP *et al*. 1999 A direct estimate of the human αβ T cell receptor diversity. *Science* **286**, 958–961. (doi:10.1126/science.286.5441.958)
37. Klarenbeek PL *et al*. 2010 Human T-cell memory consists mainly of unexpanded clones. *Immunol. Lett.* **133**, 42–48. (doi:10.1016/j.imlet.2010.06.011)
38. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. 2000 Size estimate of the αβ TCR repertoire of naive mouse splenocytes. *J. Immunol.* **164**, 5782–5787. (doi:10.4049/jimmunol.164.11.5782)
39. Wang C *et al*. 2010 High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl Acad. Sci. USA* **107**, 1518–1523. (doi:10.1073/pnas.0913939107)
40. Thomas N *et al*. 2014 Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30**, 3181–3188. (doi:10.1093/bioinformatics/btu523)
41. Rempala GA, Seweryn M, Ignatowicz L. 2011 Model for comparative analysis of antigen receptor repertoires. *J. Theor. Biol.* **269**, 1–15. (doi:10.1016/j.jtbi.2010.10.001)
42. Breit T, Wolvers-Tettero I, Beishuizen A, Verhoeven M, van Wering E, van Dongen JJ. 1993 Southern blot patterns, frequencies, and junctional diversity of T-cell receptor-delta gene rearrangements in acute lymphoblastic leukemia. *Blood* **82**, 3063–3074.
43. Mizoguchi A, Mizoguchi E, Saubermann LJ, Higaki K, Blumberg RS, Bhan AK. 2000 Limited CD4 T-cell diversity associated with colitis in T-cell receptor α mutant mice requires a T helper 2 environment. *Gastroenterology* **119**, 983–995. (doi:10.1053/gast.2000.18153)

44. Swanson SJ, Rosenzweig A, Seidman JG, Libby P. 1994 Diversity of T-cell antigen receptor V beta gene utilization in advanced human atheroma. *Arterioscler. Thromb. Vasc. Biol.* **14**, 1210–1214. (doi:10.1161/01.ATV.14.7.1210)
45. Wagner UG, Koetz K, Weyand CM, Goronzy JJ. 1998 Perturbation of the T cell repertoire in rheumatoid arthritis. *Proc. Natl Acad. Sci. USA* **95**, 14 447–14 452. (doi:10.1073/pnas.95.24.14447)
46. Koga M, Yuki N, Tsukada Y, Hirata K, Matsumoto Y. 2003 CDR3 spectratyping analysis of the T cell receptor repertoire in Guillain–Barré and Fisher syndromes. *J. Neuroimmunol.* **141**, 112–117. (doi:10.1016/S0165-5728(03)00212-1)
47. Robins HS *et al*. 2009 Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107. (doi:10.1182/blood-2009-04-217604)
48. Gil A, Yassai MB, Naumov YN, Selin LK. 2015 Narrowing of human influenza A virus specific T cell receptor α and β repertoire with increasing age. *J. Virol.* **89**, 4102–4116. (doi:10.1128/JVI.03020-14)
49. Bousso P *et al*. 2000 Diversity, functionality, and stability of the T cell repertoire derived *in vivo* from a single human T cell precursor. *Proc. Natl Acad. Sci. USA* **97**, 274–278. (doi:10.1073/pnas.97.1.274)
50. Bolotin DA *et al*. 2012 Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083. (doi:10.1002/eji.201242517)
51. He L *et al*. 2014 Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* **4**, 6778. (doi:10.1038/srep06778)
52. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. 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. (doi:10.1101/gr.115428.110)
53. Laydon DJ *et al*. 2014 Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput. Biol.* **10**, e1003646. (doi:10.1371/journal.pcbi.1003646)
54. Zarnitsyna V, Evavold B, Schoettle L, Blattman J, Antia R. 2013 Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front. Immunol.* **4**. (doi:10.3389/fimmu.2013.00485)
55. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2012 Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74. (doi:10.1038/nmeth.1778)
56. Gillet NA *et al*. 2011 The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* **117**, 3113–3122. (doi:10.1182/blood-2010-10-312926)
57. Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CRM, Bushman FD. 2012 Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28**, 755–762. (doi:10.1093/bioinformatics/bts004)
58. Bunge J, Fitzpatrick M. 1993 Estimating the number of species: a review. *J. Am. Stat. Assoc.* **88**, 364–373.
59. Sepúlveda N, Paulino CD, Carneiro J. 2010 Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J. Immunol. Methods* **353**, 124–137. (doi:10.1016/j.jim.2009.11.009)
60. Gotelli NJ, Colwell RK. 2001 Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391. (doi:10.1046/j.1461-0248.2001.00230.x)
61. May RM. 1988 How many species are there on earth? *Science* **241**, 1441–1449. (doi:10.1126/science.241.4872.1441)
62. Iknayan KJ, Tingley MW, Furnas BJ, Beissinger SR. 2014 Detecting diversity: emerging methods to estimate species diversity. *Trends Ecol. Evol.* **29**, 97–106. (doi:10.1016/j.tree.2013.10.012)
63. Hsieh C-S, Zheng Y, Liang Y, Fontenot JD, Rudensky AY. 2006 An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. *Nat. Immunol.* **7**, 401–410. (doi:10.1038/ni1318)
64. 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. (doi:10.1016/j.immuni.2006.05.016)
65. Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. 2007 Methods for comparing the diversity of samples of the T cell receptor repertoire. *J. Immunol. Methods* **321**, 182–195. (doi:10.1016/j.jim.2007.01.019)
66. Cebula A *et al*. 2013 Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* **497**, 258–262. (doi:10.1038/nature12079)
67. Venturi V, Kedzierska K, Tanaka MM, Turner SJ, Doherty PC, Davenport MP. 2008 Method for assessing the similarity between subsets of the T cell receptor repertoire. *J. Immunol. Methods* **329**, 67–80. (doi:10.1016/j.jim.2007.09.016)
68. Chao A. 1984 Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **11**, 265–270.
69. Chao A. 2005 Species estimation and applications. In *Encyclopedia of statistical sciences* (eds N Balakrishman, CB Read, B Vidakovic), 2nd edn, pp. 7907–7916. New York, NY: Wiley Press.
70. La Gruta NL, Rothwell WT, Cukalac T, Swan NG, Valkenburg SA, Kedzierska K, Thomas PG, Doherty PC, Turner SJ. 2010 Primary CTL response magnitude in mice is determined by the extent of naive T cell recruitment and subsequent clonal expansion. *J. Clin. Invest.* **120**, 1885–1894. (doi:10.1172/JCI41538)
71. Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ, Chudakov DM. 2013 Huge overlap of individual TCR beta repertoires. *Front. Immunol.* **4**. (doi:10.3389/fimmu.2013.00466)
72. Eren MI, Chao A, Hwang W-H, Colwell RK. 2012 Estimating the richness of a population when the maximum number of classes is fixed: a nonparametric solution to an archaeological problem. *PLoS ONE* **7**, e34179. (doi:10.1371/journal.pone.0034179)
73. Chao A, Lee S-M. 1992 Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **87**, 210–217. (doi:10.1080/01621459.1992.10475194)
74. Hsieh C-S, Liang Y, Tyznik AJ, Self SG, Liggitt D, Rudensky AY. 2004 Recognition of the peripheral self by naturally arising $CD25^+$ $CD4^+$ T cell receptors. *Immunity* **21**, 267–277. (doi:10.1016/j.immuni.2004.07.009)
75. Colwell RK, Coddington JA. 1994 Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345**, 101–118. (doi:10.1098/rstb.1994.0091)
76. Smith EP, Belle GV. 1984 Nonparametric estimation of species richness. *Biometrics* **40**, 119–129. (doi:10.2307/2530750)
77. Good IJ. 1953 The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264. (doi:10.1093/biomet/40.3-4.237)
78. Lee S-M, Chao A. 1994 Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50**, 88–97. (doi:10.2307/2533199)
79. Chao A. 1987 Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791. (doi:10.2307/2531532)
80. Efron B, Thisted R. 1976 Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435–447. (doi:10.1093/biomet/63.3.435)
81. Ionita-Laza I, Lange C, Laird MN. 2009 Estimating the number of unseen variants in the human genome. *Proc. Natl Acad. Sci. USA* **106**, 5008–5013. (doi:10.1073/pnas.0807815106)
82. Greene J, Birtwistle MR, Ignatowicz L, Rempala GA. 2013 Bayesian multivariate Poisson abundance models for T-cell receptor data. *J. Theor. Biol.* **326**, 1–10. (doi:10.1016/j.jtbi.2013.02.009)
83. Wang J-PZ, Lindsay BG, Cui L, Wall PK, Marion J, Zhang J, dePamphilis CW. 2005 Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. *BMC Bioinform.* **6**, 300. (doi:10.1186/1471-2105-6-300)
84. Naumov YN, Naumova EN, Hogan KT, Selin LK, 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. (doi:10.4049/jimmunol.170.8.3994)
85. Meier J *et al*. 2013 Fractal organization of the human T cell repertoire in health and after stem cell transplantation. *Biol. Blood Marrow Transplant.* **19**, 366–377. (doi:10.1016/j.bbmt.2012.12.004)
86. Hong S-H, Bunge J, Jeon S-O, Epstein SS. 2006 Predicting microbial species richness. *Proc. Natl Acad. Sci. USA* **103**, 117–122. (doi:10.1073/pnas.0507245102)

87. Zuendorf A, Bunge J, Behnke A, Barger KJA, Stoeck T. 2006 Diversity estimates of microeukaryotes below the chemocline of the anoxic Mariager Fjord, Denmark. *FEMS Microbiol. Ecol.* **58**, 476–491. (doi:10.1111/j.1574-6941.2006.00171.x)
88. Flather C. 1996 Fitting species–accumulation functions and assessing regional land use impacts on avian diversity. *J. Biogeogr.* **23**, 155–168. (doi:10.1046/j.1365-2699.1996.00980.x)
89. Scheiner SM. 2003 Six types of species-area curves. *Glob. Ecol. Biogeogr.* **12**, 441–447. (doi:10.1046/j.1466-822X.2003.00061.x)
90. R Developer Core Team. 2012 *R: a language and environment for statistical computing. 2.14.2 ed*. Vienna, Austria: R Foundation for Statistical Computing.
91. Laydon DJ, Sim A, Bangham CRM, Asquith B. 2014 *DivE: diversity estimator*. *R package version 1.0*. Available at http://CRAN.R-project.org/package=DivE.
92. Tipper JC. 1979 Rarefaction and rarefiction—the use and abuse of a method in paleoecology. *Paleobiology* **5**, 423–434.
93. Fager EW. 1972 Diversity: a sampling study. *Am. Nat.* **106**, 293–310. (doi:10.1086/282772)
94. Colwell RK, Mao CX, Chang J. 2004 Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* **85**, 2717–2727. (doi:10.1890/03-0557)
95. Xuan Mao C, Colwell RK, Chang J. 2005 Estimating the species accumulation curve using mixtures. *Biometrics* **61**, 433–441. (doi:10.1111/j.1541-0420.2005.00316.x)
96. Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, Longino JT. 2012 Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **5**, 3–21. (doi:10.1093/jpe/rtr044)