Robust Estimates of Overall Immune-Repertoire Diversity from High-Throughput Measurements on Samples

Joseph Kaplinsky[1,3] and Ramy Arnaout[1-3*]

[1]Department of Pathology and [2]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center and [3]Department of Systems Biology, Harvard Medical School, Boston, MA 02215

*To whom correspondence should be addressed at rarnaout@gmail.com

Abstract

The diversity of a person's B- and T-cell repertoires is both clinically important and a key measure of immunological complexity. However, diversity is hard to estimate by current methods due to inherent uncertainty in the number of B- and T-cell clones that will be missing from a blood or tissue sample by chance (the missing-species problem), inevitable sampling bias, and experimental noise. To address these problems we developed Recon, a maximum-likelihood method that reconstructs the clone-size distribution of an overall repertoire from measurements on a sample. Recon improves over previous work, enabling highly accurate estimates of overall diversity by any measure, including species richness and entropy, even at 0.03x coverage, with error bars. It also enables power calculations, allowing robust comparisons of diversity between individuals and over time. We apply Recon to *in silico* and experimental immune-repertoire sequencing datasets as proof of principle for measuring diversity in large, complex systems.

Introduction

Recent technological advances are making it possible to study B- and T-cell repertoires in unprecedented detail[1]. Of special interest is repertoires' diversity, defined as the number of different B- or T-cell receptors on cells present in an individual, tumor (e.g., tumor-infiltrating lymphocytes), tissue (e.g., peripheral blood, bone marrow), or cell subset (e.g., IgG^+ B cells specific for influenza). This interest follows from observations that immune repertoire diversity correlates with successful responses to infection, immune reconstitution following stem-cell transplant, the presence or absence of leukemia, and healthy vs. unhealthy aging[2-5]. The reliability of such observations depends on the ability to measure diversity, and differences in diversity, in the overall population (tumor, tissue, etc.) accurately and with statistical rigor from clinical and experimental samples. Similar requirements arise in the study of cancer heterogeneity, microbial diversity, and high-throughput sequencing, as well as outside of biology.[6-9] However, measuring diversity is complicated, for two reasons.

First, diversity may mean many things. Conventionally, it refers to the number of different species in a population, a measure known as species richness. An example is the number of B-cell clones in an individual, where "clone" denotes cells with a common B- (or T-)cell progenitor. However, diversity can refer to several different measures that capture different features of the size-frequency distribution of species in the population. The Berger-Parker index (BPI), e.g., measures the dominance of the single largest clone (Fig. 1).[10] Several diversity measures have been used for immune repertoires. These include species richness, Shannon entropy (henceforth "entropy"), and the Simpson and Gini-Simpson indices[11-14]. Species richness is unique in that it takes no account of how common or rare each species is. In contrast, entropy and other measures systematically down-weight smaller clones to different extents. Species richness, entropy, the Gini-Simpson index, BPI, and other measures are related through a mathematical framework described by Hill[15,16]. Using simple mathematical transformations, this framework al-

lows each measure to be interpreted as the "effective number" of species of a given frequency, facilitating comparisons among different measures (Fig. 1b). For example, entropy, conventionally measured in bits, is converted into an effective number via exponentiation. Thus in the overall repertoire in Fig. 1, the effective number of clones is 7.4 by entropy and 2.9 by BPI (Fig. 1b). Different measures provide complementary information: e.g., two repertoires can have the same species richness but different entropies or BPIs (species richness and BPI bracket the Hill measures; Fig. 1d).[10] Thus, no single measure is likely to capture all the features of interest in a given repertoire. Consequently, methods for estimating diversity should provide complementary measures.

Second, the diversity of a sample (e.g. a blood sample) can differ markedly from the diversity of the overall repertoire from which it derives (e.g., the circulation). Although blood or tissue samples may contain many thousands of B or T cells, these are still only a fraction of the billions of cells in an overall repertoire. Consequently, some clones in the overall repertoire, especially small clones, usually go unsampled and thereby undetected in measurements of samples (Fig. 1a). Sample diversity therefore usually underestimates true diversity. This is called the missing-species problem (Fig. 1b)[17]. Weighted measures are less sensitive to missing species than species richness, since they down-weight the small clones that are most likely to be missing (Fig. 1b). However, use of weighted measures such as entropy as a substitute for species richness has potential drawbacks. First, it is unclear what information is lost by ignoring small clones. Second, for sample diversity measured by a weighted measure to be an accurate reflection of overall diversity, clone sizes—the number of cells per clone—in the sample must reflect clone sizes in the overall population; however, this is biased by sampling noise for many clone sizes. Note that sampling noise is intrinsic to sampling, and will affect measurements even for methods that can count every cell in a sample and perfectly assign sequences to clones. Consequently, depending on the distribution and measure, sampling can still misrepresent overall di-

versity, even with weighted measures (Fig. 1b and below).

Sampling noise is compounded by experimental error. Quantitation error due to inaccurate cell counts, amplification dropouts, and jackpot effects; sequence errors from amplification and sequencing; and annotation errors introduced during data processing add experimental noise to sample measurements. Required are methods that are robust not only to missing species and sampling noise, but to experimental noise as well.

Existing attempts at estimating the number of missing species have limitations. Fisher's gamma-Poisson mixture method, a parametric method that has been used on T-cell repertoires, involves a divergent sum that can result in large uncertainties in the number of missing clones and thereby overall species richness[18,19]. Moreover, because Fisher's method does not output overall clone sizes, it does not produce weighted measures. A nonparametric method by Chao, based on the Good-Turing estimator, avoids divergent sums and has been widely used in ecology; however, like Fisher's method, it provides only species richness[20,21]. So does extrapolating from curve fitting, which in addition is somewhat arbitrary[13,14,22,23]. Nonparametric approaches using maximum likelihood provide additional measures, but existing implementations either do not scale to complex populations like repertoires, risk overfitting or getting trapped in local maxima, or make restrictive assumptions about the clone-size distribution of the overall repertoire and therefore are not generalizable[24-26]. Moreover, because a higher-likelihood fit can often be had by adding small clones, previous approaches yield unbounded estimates for species richness, which are impractical[27].

We solve these problems with Recon—reconstruction of estimated clones from observed numbers—a generalized, high-performance, modified maximum-likelihood method that makes no assumptions about clone sizes in the overall repertoire, estimates any diversity measure, and leads naturally to sensible error bars that facilitate practical, statistically reliable comparisons

between samples, including between individuals and over time, for complex populations.

Results

Description. Recon is a modified maximum-likelihood method based on the expectation-maximization (EM) algorithm[6,28]. Briefly, an initial description of the overall distribution is refined iteratively based on agreement with the sample distribution, adding parameters as needed until no further improvement can be made without overfitting (Fig. 1c). The result is the overall clone-size distribution that, if sampled randomly, is statistically most likely to give rise to the sample distribution, given the above constraints (Fig. S1). The only assumptions Recon makes are that the overall repertoire is large relative to the sample and is well mixed.

The input is the observed clone-size distribution in a sample, provided as list of clone sizes and counts. This is easily generated from sequence data by counting clones that have the same number of sequences in the dataset for (at least semi-)quantitative sequencing. Recon outputs (*i*) the overall clone-size distribution; (*ii*) the diversity of the overall repertoire as measured by species richness, entropy, or other Hill measure, with error bars; (*iii*) the number of missing species, with error bars; (*iv*) the minimum detected clone size (below); (*v*) the diversity of the sample repertoire, for comparison; and (*vi*) a resampling of the overall distribution for comparison to the sample. Recon can be run on tumor clones, microbial species, sequence reads, or other (including non-biological) populations. Recon can also generate tables for power calculations and experimental design.

Recon marks five improvements over previous approaches. First, to avoid dependence on initial conditions or becoming trapped in local maxima, Recon "scans" a large number of initial conditions in each iteration of the algorithm. We verified that scanning produces substantially better estimates of overall clone sizes, missing species, and diversity measurements (Fig. S3). Se-

cond, it optimizes the average of the two best fits in each round (reminiscent of genetic algorithms). Third, it includes a check to prevent overfitting due to sampling noise. Fourth, it makes no assumptions about the overall clone-size distribution, making it widely applicable. And fifth, it improves over previous maximum-likelihood models in handling uncertainties, for example regarding bounds on overall diversity estimates.

Previous models overestimate species richness when coverage is low, as small clones added to the estimate result in overfitting of the sample distribution—in the limit leading to an estimate with infinite infinitesimal clones. Recon uses discrete clone sizes, which in the worst case ensures that estimates are bounded by the number of cells in the overall repertoire (clones cannot outnumber cells). Recon's use of both a noise threshold and the corrected Akaike information criterion provide tighter bounds, rejecting additional clones unless their expected contribution to the sample rises above sampling noise (by 3 standard deviations in our implementation) and outweighs the penalty of additional parameters. The trade-off is that for each sample, there is a minimum clone size that Recon can detect: if ≤1, Recon's species-richness estimate will include clones represented by just a single cell in the overall repertoire, if there are any; if >1, there may be clones in the overall repertoire that are too small to detect, although for a given sample, the smallest clones detected may be the smallest clones there are. In this case, U (Online Methods) gives an upper bound on species richness that includes clones that may be "hiding." See Supplementary Information for details.

Validation. We validated Recon on *in silico* repertoires that spanned a range of previously estimated overall diversities (10,000-10 million clones) and clone-size distributions: from steep, i.e., dominated by small clones, to flat exponentials; reciprocal-exponential distributions that derive from a generative model; and multiple bimodal distributions of small and large clones, with and without simulated experimental noise (Online Methods). These repertoires served as gold standards. We sampled a known number of cells from each, for a range of sample sizes

(10,000-1 million cells). We used Recon to reconstruct overall repertoires from each sample. We then compared the diversity of the reconstructed overall repertoire with the true overall diversity and sample diversity. We measured diversity by species richness, entropy, Simpson Index, and BPI (Fig. 1b).

First we compared sample diversity with overall diversity (Fig. 2a). For a given sample size, higher overall diversity means lower clonal coverage (the number of cells in the sample per clone in the overall repertoire). For each repertoire, the error, measured as the difference between sample and overall diversity, grew as coverage fell below 1x, because samples cannot contain more clones than cells. Consequently, for species richness, sample diversity underestimated true diversity by 50% at 1x coverage, 10 fold at 0.1x coverage and 30 fold at 0.03x coverage. The weighted measures performed little better, even for the flattest clone-size distributions that we tested, partly due to the absence of clones large enough to dominate these repertoires (e.g., leukemic clones; Figs. 2 and S2). We concluded that sample diversity is generally an unreliable proxy for true diversity below 1x coverage in the absence of dominant clones.

In contrast, Recon's estimates of overall diversity showed excellent agreement with true diversity even at <1x coverage, across the range of diversity measures (Fig. 2a, lower panels). For species richness, Recon's estimates were accurate to within 1% of the true diversity at 10x coverage, 10% at 3x coverage, and 50% at just 0.03x coverage—at which there is just one cell in the sample for every 30 clones in the overall repertoire. Error for entropy and other weighted measures was lower. Comparison of the top and bottom panels in Fig. 2a and Fig. S2a-c validates Recon's performance.

Next we compared Recon to Chao's estimator[20,21], asking which method provided better estimates of true species richness for samples of exponentially and bimodally distributed repertoires, with and without noise. Although both methods performed well, Recon outperformed

Chao throughout, with estimates that were as close or closer than Chao's for 64% of exponential repertoires and 80% of bimodal distributions without noise, and 97% of repertoires with noise (Fig. 2b-c, S2d-g). Because Chao outputs only species richness, entropy and other diversity measures could not be compared. An alternative method that does output other measures could not be compared because computationally it does not scale above hundreds of clones (with millions of parameters, it would also risk overfitting)[25,26].

Re-sampling from a reconstructed repertoire makes it possible to test for self-consistency by comparing the clone-size distribution of the new sample to that of the original sample. As expected, we found good agreement between predicted and observed frequencies of clone sizes (Fig. 3). This included agreement on the number of missing clones. Because our gold-standard distributions were pre-defined, numbers of missing clones were known to us (though unknown to Recon). Recon's ability to estimate them accurately contributed to the accuracy of its overall diversity estimates. Of note, the number of missing clones depended strongly on the number of singlets and doublets in the sample: large singlet-to-doublet ratios, with enough of both for low sampling noise, gave more accurate estimates.

Error bars and power calculations. Detecting reliable differences in overall diversity requires bounds. Recon outputs two types of bounds: error bars for the effective number of clones greater than or equal to a minimum detected clone size, described below, and an species-richness upper bound for all clones, U (Supplementary Information).

For error bars, we sampled each gold standard systematically for a range of coverage and sample sizes ≤10 million cells. For each, we used Recon to estimate overall diversity as a function of coverage. Because higher coverage produces better estimates, the resulting error profile can be represented as a funnel plot that converges to the true overall diversity (Fig. 4a). The funnel's upper and lower bounds correspond to the largest and smallest values of estimated di-

versity that are consistent with the true diversity. For wider error bars, we used the proportional error of the worst fit at each level of coverage to define the bounds, making each funnel symmetric. In this way, error profiles were made for each gold standard, a separate profile for each diversity measure.

To make an error bar for an overall diversity estimate, Recon finds the true diversity for which the estimated diversity is at the lower bound, and the true diversity for which it is at the upper bound. These respectively define the upper and lower error bars (Figs. 4b, 4c). Combining error profiles across all samples suggests that 1x coverage generally produces error bars of ±20% for overall species richness (Fig. 4d), consistent with our validation (Fig. 2).

We used this error-bar framework to build tables for estimating the coverage required to detect differences between two samples. Specifically, given an order-of-magnitude estimate of the overall diversity for two samples, we determined the minimum sample size for which error bars for overall diversity estimates from these samples would not overlap at detection thresholds ranging from 10% to 5 fold (Table 1). This is the minimum sample size required to reject the null hypothesis that two estimates that differ by a given amount are actually from the same overall repertoire, at a confidence level of p~0.05 (Supplementary Information). While detecting larger differences requires fewer cells, for a given overall diversity there is a minimum sample size below which the number of non-singlets is expected to be too small for Recon to run (Table 1). So an experiment designed to detect a 20% (1.2x) difference in species richness between two samples, in which the samples are drawn from overall repertoires that have ~1 million clones, will require at least 485,204 cells from each sample for analysis. This is the number of cells in the sample that are in small (≤30 cells) clones that Recon requires to perform reconstruction; if 300,000 of the 485,204 cells in a sample belong to a single large clone, e.g. because of leukemia, the remaining 185,204 cells of the non-leukemic clones will be sufficient to detect a 40% difference in the species richness of the non-leukemic portion of the repertoire, but not ≤30%.

To further test Recon and our error-bar framework, we ran it on a sample distribution previously identified as causing difficulties for overall species-richness estimation by multiple existing methods, corresponding to an overall population of ~3,000 species sampled at ~0.8x coverage (Supplementary Information)[27]. Three- and four-point mixture models, a logit normal model, a log-gamma model, and a beta model gave estimates of 2,930-3,494, with non-overlapping error bars. Recon gave an estimate of 3,006, with error bars of 2,790-3,277, bracketing most of the other estimates, suggesting Recon can resolve other methods' inconsistencies in difficult cases.

Experimental data. We applied Recon to six experimental datasets: four of paired heavy-and-light chain and two of heavy chain (Online Methods). We used the authors' clone definitions—clusters of reads with ≥96% nucleotide identity in heavy-chain complementarity determining region 3 (CDR_H3)[29] or reads with identical CDR_H3s and V_H annotations [31]—with the caveats that clone assignment is difficult, some cells may not have been sequenced, artifacts are possible, and sequencing is only semi-quantitative. Because such datasets reflect the current state of the field and are used for diversity measurements, we treated them as (imperfect) samples and used Recon to estimate diversity for the corresponding overall repertoires (Table 2). Resampling showed good fits (Fig. 5). For four of the six repertoires, we found that most clones were missing from the sample (i.e., enough sampling would have approximately doubled the number of clusters). Entropy was almost identical between samples and overall repertoires, resulting from very large clones and/or PCR jackpot effects that contribute disproportionately to the entropy calculation. In these datasets, overall species richness captures information lost during sampling that entropy does not.

Availability. Recon is available subject to license agreement at http://arnaoutlab.github.io/Recon.

Discussion

High-throughput technologies enable highly detailed descriptions of B- and T-cell repertoires. That these descriptions are generally of samples, and not e.g. blood or tissue repertoires overall, may seem an inconsequential distinction when samples contain many cells. However, it is critical for estimating overall diversity. Unless the number of cells in a sample exceeds the number of clones in the overall repertoire by ~3-10-fold (Fig. 3), sample and overall diversity may bear little relation (Figs. 2a, S2a-c). This discrepancy is not a technological shortcoming but an inherent constraint of random sampling: smaller clones will be missed and larger clones overcounted (Fig. 1a). In humans, overall repertoires may contain many millions of clones. Because routine blood samples rarely contain more than a few million B and T cells of any sort combined, they are too small for sample diversity to serve as a reliable proxy for overall diversity. Thus conclusions drawn only from sample diversity measurements warrant caution.

This caveat applies for all diversity measures. Entropy, often used to measure sample diversity in immune-repertoire studies, is less prone to undercounting. However, in our validation repertoires even BPI, the Hill measure least prone to undercounting and most robust to missing species, underestimates overall diversity by an order of magnitude for levels of coverage encountered in experiments (Figs. 2, S2); it is unsurprising, then, that sample entropy can also underestimate overall entropy in these repertoires (Figs. 2, S2). Additional caveats apply to experimental datasets. Insufficient read clustering will overestimate species richness. For clone sizes defined proportional to the number of reads, PCR jackpot effects can produce artificially large "clones," overestimating entropy. These biases, not mutually exclusive, may explain some of the differences between species richness and entropy in the experimental datasets we studied (Table 2). Better quantitation (e.g., via barcoding and robust clonality modeling) would mitigate these biases but not the bias intrinsic to sampling, which Recon addresses.

Recon outperforms previous approaches at estimating species richness even for large, complex clone-size distributions and in the presence of experimental noise (Figs. 2, 3, S2). Moreover it performs as well for entropy and BPI (Figs. 2a, S2a-c)—measures that Chao, Fisher's method, and others do not provide. This is an important improvement, since species richness marks but one end of the spectrum of Hill measures (BPI marks the other, with entropy between). Until it is clearer how different measures correspond to specific biological and clinical processes of interest, single measures may mislead (Fig. 1d). Recon offers investigators the full suite.

Error bars and power tables are necessary steps toward being able to test for such correspondences and evaluating diversity as a biomarker. Recon's error bars and tables for entropy, BPI, and other measures mean differences can be assessed for any measure or noise level. Recon's error bars perform well by practical tests, bracketing the number of missing species (Fig. 3) and squaring previous models[27]. Its power tables offer guidance for sample requirements and suggest expected limitations for different studies. For example, measuring the species richness of naïve repertoires of ~10^7 clones[30,31] will likely require phlebotomy or apheresis samples; even then, detecting 5-fold differences is probably the limit (Table 1). Meanwhile, measuring diversity for effector/memory subsets should require only routine blood draws (2-6mL), which should detect sub-fold differences. For marrow, spleen, tumor, granuloma, or abscess samples, the investigator must decide whether the sample is well mixed, a Recon requirement.

High-throughput technologies hold much promise for measuring diversity in repertoires, cancer, and other complex populations, but current limitations warrant caution. Because most sequencing experiments are still only semi-quantitative, the number of reads does not always reflect the number of cells. Chimerism and sequencing/annotation errors mean not all clusters are clones. Incomplete cell lysis and sequencing inefficiencies can underestimate sample size. These limitations affect the calculation and interpretation of diversity estimates and upper bounds; the examples we have shown should be interpreted accordingly, even as they illustrate application of

our method. Our results suggest that overcoming these limitations will improve our understanding of diversity, a defining characteristic of complex systems.

Online Methods

Core algorithm. Mathematically, the problem is to find the B- or T-cell clone-size distribution in the individual (the "parent" or "overall" distribution) that is most likely to give rise to the clone size distribution that is observed in the sample (the sample distribution) (Fig. 1). From the parent distribution, we can then calculate overall diversity according to any diversity measure in the Hill framework. The core of our method is the expectation-maximization (EM) algorithm, in which a rough approximation of the parent distribution is refined iteratively until no further improvement can be made without overfitting[28].

The EM algorithm begins by assuming a parent distribution in which clones are all the same size, taken from the mean of the observations. To perform the fit, we need to know not just the observed clone frequencies but also the number of missing species, which is unknown and therefore must first be estimated. Following previous work[32], we estimate the number of missing species by calculating the expected clone size distribution for a (Poisson) sample of the parent distribution (see "Sampling" below) and applying the Horvitz-Thomson estimator [33]. We then fit the clone size of the parent distribution using maximum likelihood, recalculate the number of missing species, and repeat these steps until a self-consistent number of missing species is obtained. This completes the first iteration of the algorithm, yielding the uniform parent distribution that is most likely to give rise to the sample distribution.

In the second iteration, we refine this uniform parent distribution by adding a second clone size. We estimate the number of missing species for this new two-size distribution, fit the two clone sizes and their relative frequencies by maximum likelihood, and, as in the first iteration of the algorithm, repeat until there is no further improvement[32]. The result is the two-clone-size parent distribution that is most likely to give rise to the sample distribution.

In subsequent iterations, we continue to refine the parent distribution by adding clone sizes and

refitting as above, iterating until no more clone sizes can be added without overfitting (using the corrected Akaike information criterion as a stop condition). The result is the desired MLE. Note that whereas the sample distribution generally traces out a smooth curve, the MLE parent distribution is spiky, reflecting the resolution limits the information about the parent distribution contained in the sample distribution.

Sampling. We assume that each clone in the individual contributes cells to the sampled population according to a Poisson distribution. This will be true if (*i*) clones are well mixed in the blood or evenly distributed in the tissue being sampled, (*ii*) the parent population is sufficiently large that the Poisson estimate for the probability of e.g. a singleton contributing >1 cell is negligible, and (*iii*) no single clone is a large fraction (~30% or more) of the parent population. In practice, condition (*iii*) is satisfied by counting large clones directly (see "Fitting").

Fitting. The largest clones may be represented by hundreds or even thousands of cells in a sample. For such large clones, sampling error is small: the relative size of the clone in the sample and in the individual will be about the same. As a result, clones that are large enough to have sufficiently small sampling error do not have to be fit by EM, and instead can simply be added to the MLE. We found that using a threshold of 30 cells, and therefore applying EM only to clones that contribute ≤30 cells to the sample and then adding larger clones back to the resulting MLE gives results that are indistinguishable from applying EM on the entire sample distribution, but with vast gains in speed.

Scanning. In the standard EM algorithm, the exact sizes and frequencies of clones in the final MLE can vary depending on the sizes and frequencies used at the start of each iteration, reflecting different relative maxima. To find global maxima, we developed a "scanning" approach in which we applied EM to many starting clone sizes and frequencies (110 in our implementation), ranking results by maximum likelihood (after first adjusting likelihoods according to the number

of ways to choose clones in each distribution; see Supplementary Information). In each round we perform an additional fit with starting clone sizes and frequencies at an average of the two top-ranked results. We then select the resulting best-ranked fit from the 110 starting points.

Diversity measures. Species richness, entropy, the Gini-Simpson Index, BPI, and indeed many other diversity measures are related to each other through the mathematical framework of the so-called Hill numbers[15,34]. These form a series in which the index reflects the extent to which counts are weighted toward large clones. Species richness, in which large and small clones are counted equally and so large clones are unweighted, has an index of zero and is denoted ${}^0D$ (pronounced "D-zero"). Other measures, or simple mathematical transformations thereof, correspond to larger indices; these include entropy ($\ln({}^1D)$), the Simpson Index ($1/{}^2D$), and BPI ($1/{}^\infty D$).

We calculated ${}^0D$, ${}^1D$, ${}^2D$, and ${}^\infty D$ for sample and overall distributions from *in silico*-sampled synthetic gold-standard distributions (see "Validation" below and in the main text) and from several published data sources (see "Experimental Data" in the main text). These ${}^qD$ are a function of frequencies of clone frequencies p_i, where *i* ranges over each clones and the frequencies are normalized to $\sum_i p_i = 1$, defined as ${}^qD(p) = \left(\sum_i^S p_i{}^q\right)^{1/(1-q)}$[34].

We calculated ${}^0D$ by simply counting the number of different clones, ${}^1D$ according to $\exp\left(-\sum_i p_i \ln p_i\right)$, ${}^2D$ according to the definition, and ${}^\infty D$ as the reciprocal of the frequency of the largest clone (the above definition reduces to these expressions for the value $q = 0$ and in the limits $q \to 1$ and $q \to \infty$).

Validation. We validated our method by generating a wide range of biologically plausible synthetic parent distributions of 10^9 cells *in silico*, sampling from these distributions to produce samples of different known sizes, using the samples to estimate overall diversities according to

the above measures, and comparing these estimates against the (known) calculated diversities of the original parent distributions. We studied three families of test distributions in detail: exponential distributions (of the form $f(x) \propto e^{-sx}$, where x denotes clone size, $f(x)$ is the frequency of clones of that size, and s is a parameter that controls the steepness of the distribution), which are simple distributions that describe the shape of observed sample distributions phenomenologically, "reciprocal-exponential" distributions ($f(x) \propto \frac{1}{x}e^{-sx}$), which are the analytical solution to a simple biologically plausible model of the dynamics of most B- and T-cell clones, and bimodal distributions with the largest clones an average multiple of the size of the smallest clones (e.g. 20x). We tested these distributions systematically by varying the steepness from very steep (*s*=1.2) to nearly flat (*s*=0.12) and different multiples for the bimodal distributions, encompassing the a range of biologically plausible clone-size distributions, with and without noise. For distributions with noise, noise added to each count *n* with mean of zero and standard deviation $1.22 \cdot \sqrt{n}$.

Error bars. Error bars define the range of overall diversity values that, given the inevitable error involved in reconstructing parent distributions from samples of a given size, are consistent with our algorithm's estimate. We determined error bars for each diversity measure (species richness, entropy, etc.) as follows (Fig. 4). First, we generated an *in silico* parent population with known diversity. Second, we took samples of this known distribution at systematically increasing sample sizes and, for each sample size, used our algorithm to estimate the overall diversity (Fig. 4a). These steps resulted in a reference table for how error falls with increasing sample size for a given level of diversity. Given a test sample, its coverage, and the overall diversity by a given measure (estimated from our algorithm), we can then look up (or interpolate) the largest and smallest diversity values that are consistent with the estimate (Fig. 4b, c). These upper and lower bounds define the desired error bar (and error bars). We note that estimates are more accurate for more peaked clone-size distributions, and that most real-world distributions are no-

ticeably peaked. Nevertheless we chose to study the flatter *in silico* parent distribution in detail, in order to provide wider error bars.

Experimental datasets. We found and downloaded six publically available datasets. Four were from paired heavy-and-light-chain sequencing experiments: two of IgG^+ B cells (from two subjects), one of memory B cells post-influenza vaccination, and one of tetanus-toxoid-specific plasmablasts[29]. Following that study's methods, we clustered reads with ≥95% heavy-chain complementarity-determining region 3 (CDR3) nucleotide identity (the study treated clusters as clones). The other two datasets were of pooled PCR of heavy-chain genomic DNA from bone-marrow plasma cells from a healthy subject and non-myeloma plasma cells from a subject with multiple myeloma, with clones defined as sequences with identical CDR3s at the amino acid level and identical V_H nucleotides[35]. We estimated the total number of IgG^+ B cells, post-vaccination memory B cells, tetanus-specific plasmablasts (and plasma cells), bone-marrow plasma cells in a healthy patient, and non-myelomatous plasma cells to be 75 million, 260 million, 3.5 million, 6 million, and 3 million, respectively, for N (See below)[36-41].

Minimum detected clone sizes and upper bounds (U). The smallest clone size in the reconstructed clone-size distribution is described by two parameters: the mean number of cells that each clone of this size contributes to the sample, $m_{\min}$, and the fraction of all clones that are of this size, w_m. The size of this smallest detectable clone in the overall repertoire is $m_{\min}$ scaled to the total number of cells: $m_{\min}N/S$. This is Recon's minimum detected clone size. It is possible that there are clones smaller than this size in the overall repertoire, but because they contribute a mean of zero cells to the sample they are not detected and therefore do not contribute to Recon's estimate of overall species richness. An upper bound on species richness that includes clones smaller than the minimum detected clone size, U, is obtained by assuming that all cells in clones that could be smaller than this are singlets: $U = R_{\max}w_m m_{\min}N/S$, where $R_{\max}$ is Recon's upper error bar estimate of overall species richness (Supplementary Information). We calculated

these quantities for our validation and experimental data.

Acknowledgements

The authors thank Rima Arnaout for critical reading of the manuscript and Ruth Moorman and Sheldon Simon for generous support. RA is supported by awards from the National Institutes of Health, American Heart Association, and BIDMC.

Table and Figure Legends

Table 1. Power calculations. Table entries give the minimum number of cells that must be analyzed in order to be able to detect a given fold-difference in species richness between two samples at p=0.05 (row headings), given an expected overall species richness (column headings). As noted in the main text, these numbers exclude cells that might belong to large clones (here, of clone size ≥30 in the sample). Minima required for reliable reconstructions are in gray. See Supplementary Information for details.

Table 2. Diversity estimates for experimental datasets from humans. Summarized are Recon's estimates of overall diversity for six datasets; its estimate of the number of missing species; comparisons to sample diversity, for species richness and entropy (given as effective numbers; 2^{bits}); the minimum detected clone size (see main text); and upper bound for species richness that includes potential "hiding" clones. Cell-surface phenotypes were as follows: IgG^+ B cells, $IgG^+CD2^-CD14^-CD16^-CD36^-CD43^-CD235a^-$; post-vaccination memory B cells, $CD19^+CD3^-CD27^+CD38^{int}$; tetanus-specific plasmablasts, $CD19^+CD3^-CD14^-CD38^{++}CD27^{++}CD20^-$; plasma cells, $CD138^+$. See references for details.

Figure 1. Overall repertoires vs. samples. (a) shows an overall repertoire (top left) and a repertoire from a random sample of this repertoire (top right), together with respective clone-size distributions from the overall repertoire and sample (bottom). Each circle denotes a cell; different

colors denote different clones. Note that five clones are missing from the sample entirely, represented by the open red circle at a clone size of zero in the sample clone-size distribution. (b) Sample diversity underrepresents overall diversity across a range of diversity measures. (c) Recon reconstructs the overall repertoire by estimating the number of missing clones and iteratively updating until the predicted clone size distribution in the sample (red crosses) matches the observed clone-size distribution in the sample (open circles), stopping short of overfitting. (d) Different diversity measures are complementary. Repertoires R1, R2 and R3 each have a total of 7 cells. R1 and R3 have the same species richness but different inverse Berger-Parker index (inv. BPI); R2 and R3 have the same Berger-Parker index but different species richness.

Figure 2. Recon vs. other methods. (a) the sample diversity (top) and Recon's estimate (bottom) of overall diversity as a function of the actual overall diversity for three different sample sizes—10,000 cells (filled circles), 100,000 cells (small open circles), and 1 million cells (large open circles)—for a representative gold-standard distribution without noise (given in Fig. S2e, left panel; see Fig. S2 for additional examples). Coverage is defined as sample size/overall diversity; for a fixed sample size, coverage falls as overall diversity increases. The red line represents zero error; the further the points fall off the line, the larger the error. Left-to-right: species richness, entropy, and the inverse Berger-Parker index. For a given sample size, Recon's estimates remain accurate for overall diversities approximately 30 times larger than the sample size (0.03x coverage). In contrast, sample diversity is accurate only until overall diversity is approximately equal to the sample size (1x coverage; see Fig. 4d). (b) Comparison of Recon and Chao's species-richness estimates for a variety of validation repertoires showing the percent in which Recon or Chao were closer to the true value. (c) Representative comparison of species richness estimates by Recon and Chao's estimator in the presence of noise for four overall repertoires of species richness from 300,000 to 10 million clones at constant sample coverage of 0.3x. Each violin plot shows a kernel density estimate from fits to 100 realizations of the noise

with mean 0 and standard deviation $1.22 \cdot \sqrt{n}$ on each count. Circles indicate estimates without noise. The true diversity of each overall repertoire is shown by a red bar.

Figure 3. Predictions vs. simulated observations, *in silico* gold standards. Shown are fits to observations from representative gold-standard distributions of the shape shown in Figure S2e, left panel. Left-to-right: overall distributions with increasing numbers of clones. Top-to-bottom: increasing sample size measured in coverage of the number of clones in the overall population. Open black circles denote observed clone-size distributions, which was the input data given to Recon. The open red circle denotes the number of missing clones, which was not known to Recon. Red crosses denote Recon's prediction of the clone-size distribution in the sample, based on its reconstruction of the clone-size distribution of the overall repertoire. This includes a prediction for the number of missing clones, plotted as the number of clones of size zero, with error bars as shown.

Figure 4. Error bars. (a) shows a schematic representation of Recon's diversity estimates (open circles) from a single gold-standard *in silico* repertoire with overall diversity d for many different sample sizes. These results are used to make error bars as follows. Given a test sample, Recon first estimates the overall diversity, d_R, and the coverage (=sample size/d_R). (b) Recon then looks up the maximum ($d_\oplus$) and minimum ($d_\ominus$) diversities that are consistent with its estimate (d_R); schematically, this is where the edges of the funnel plots for $d_\oplus$ and $d_\ominus$ intersect. (c) Higher coverage gives smaller error (arrows). (d) Combining the results from all gold-standard repertoires into a single plot suggests the rule of thumb that 1x coverage gives error bars of 20% for species richness. Shown are results for sample sizes ≤10 million cells (which corresponds to all the B or T cells in 10-50ml of blood), the range for which our implementation of Recon is optimized. We repeated this process separately for entropy and other measures.

Figure 5. Predictions vs. observations, experimental data. Shown are Recon's estimates of overall diversity for six experimental datasets. These included (a, b) immunoglobulin heavy (IgH)- and light-chain (IgL) paired-chain sequencing experiments from IgG^+ B cells from the blood of two different subjects, (c) pooled-DNA IgH sequencing experiments on the bone-marrow plasma cells from a healthy adult, (d) IgH+L of post-vaccination memory B cells, (e) IgH+L tetanus toxoid-specific plasmablasts, and (f) pooled-DNA IgH sequencing experiments on the bone-marrow plasma cells from a multiple myeloma patient (only the non-myeloma cells). Details, including references, are presented in Table 2.

Figure S1. The Recon algorithm. Steps in the flowchart are as described in the main text, Online Methods, and Supplementary Information.

Figure S2. Recon diversity vs. other estimates showing fits to additional gold standard repertoires plotted as for Figure 2. (a)-(c) Comparisons of sample diversity (top) to Recon diversity (bottom) plotted as in Figure 2a for (a) a steep exponential clone size distribution (b) a bimodal distribution in which the overall distribution contains a population of small clones and a population 31 times as large and (c) a bimodal distribution in which the overall distribution contains a population of small clones and a population 20 times as large. (d)-(g) Comparison of species richness estimates by Recon (middle) and Chao's estimator (right) shown as in Figure 2b for an example additional gold standard overall distributions (left) for (d) a steep exponential clone-size distribution, (e) a shallow exponential clone-size distribution, (f) a bimodal distribution in which the overall distribution contains a population of small clones and a population 31 times as large, and (g) a bimodal distribution in which the overall distribution contains a population of small clones and a population 20 times as large.

Figure S3. Scanning. Probability densities of the ratio of estimated missing species/true missing species demonstrating the benefit of using additional starting points. Fits using, in each round of

fitting, 2 starting weights (green), 2 starting means (black) and 110 combinations of starting weights and means (yellow) show that multiple starting point result in a sharper peak of the probability distribution function (pdf) near 1.0, and diminished trapping in local minima away from 1.0. Pdfs are plotted using Gaussian kernel density estimates over 800 samples from gold-standard distributions (see main text).

References

1 Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*, **32**, 158-168 (2014).

2 Gibson, K. L. *et al.* B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* **8**, 18-25 (2009).

3 Wang, C. *et al.* Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Repertoires. *J. Immunol.* **192**, 603-611 (2014).

4 Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra119 (2013).

5 Ademokun, A. *et al.* Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* **10**, 922-930 (2011).

6 Bunge, J., Willis, A. & Walsh, F. Estimating the Number of Species in Microbial Diversity Studies. *Annu. Rev. Stat. Appl., Vol 1* **1**, 427-445 (2014).

7 Daley, T. & Smith, A. D. Modeling genome coverage in single-cell sequencing. *Bioinformatics.* **30**, 3159-3165 (2014).

8 Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Meth.* **10**, 325-327 (2013).

9 Horswell, S., Matthews, N. & Swanton, C. Cancer heterogeneity and "the struggle for existence": diagnostic and analytical challenges. *Cancer Lett.* **340**, 220-226 (2013).

10 May, R. M. in *Ecology and Evolution of Communities* (ed J. M. M. L. D. Cody) (Harvard University Press, 1975).

11 Sherwood, A. M. *et al.* Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. *Cancer Immunol. Immun.* **62**, 1453-1461 (2013).

12 Robert, L. *et al.* CTLA4 blockade broadens the peripheral T-cell receptor repertoire. *Clin. Cancer Res.* **20**, 2424-2432 (2014).

13 Arnaout, R. *et al.* High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE* **6**, e22365 (2011).

14 Laydon, D. J. *et al.* Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput. Biol.* **10**, e1003646 (2014).

15 Hill, M. O. DIVERSITY AND EVENNESS - UNIFYING NOTATION AND ITS CONSEQUENCES. *Ecology* **54**, 427-432 (1973).

16 Jost, L. Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427-2439 (2007).

17 Bunge, J. & Fitzpatrick, M. Estimating the Number of Species: A Review. *J. Am. Stat. Assoc.* **88**, 364-373 (1993).

18 Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.* **12**, 42-58 (1943).

19 Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099-4107 (2009).

20 Chao, A. & Lee, S. M. ESTIMATING THE NUMBER OF CLASSES VIA SAMPLE COVERAGE. *J. Am. Stat. Assoc.* **87**, 210-217 (1992).

21 Chao, A. NONPARAMETRIC-ESTIMATION OF THE NUMBER OF CLASSES IN A POPULATION. *Scand. J. Stat.* **11**, 265-270 (1984).

22 Warren, R. L. *et al.* 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 (2011).

23 Klarenbeek, P. L. *et al.* Human T-cell memory consists mainly of unexpanded clones. *Immunol. Lett.* **133**, 42-48 (2010).

24 DeWitt, W. *et al.* Replicate immunosequencing as a robust probe of B cell repertoire diversity. arXiv:1410.0350v1 (2014).

25 Norris, J. L. & Pollock, K. H. Nonparametric MLE under two closed capture recapture models with heterogeneity. *Biometrics* **52**, 639-649 (1996).

26 Norris, J. L. & Pollock, K. H. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environ. Ecol. Stat.* **5**, 391-402 (1998).

27 Link, W. A. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123-1130 (2003).

28 McLachlan, G. J. & Krishnan, T. *The EM algorithm and extensions*. 2nd edn. (Wiley-Interscience, 2008).

29 DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166-169 (2013).

30 Wiegel, F. W. & Perelson, A. S. Some scaling principles for the immune system. *Immunol. Cell Biol.* **82**, 127-131 (2004).

31 Zarnitsyna, V. I., Evavold, B. D., Schoettle, L. N., Blattman, J. N. & Antia, R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front. Immunol.* **4**, 485 (2013).

32 Bohning, D. & Schon, D. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. R. Stat. Soc. Ser. C.-App.* **54**, 721-737 (2005).

33 Armitage, P. & Colton, T. *Encyclopedia of biostatistics*. 2nd edn (John Wiley, 2005).

34 Leinster, T. & Cobbold, C. A. Measuring diversity: the importance of species similarity. *Ecology* **93**, 477-489 (2012).

35 Tschumper, R. C. *et al.* Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. *Oncotarget* **3**, 502-513 (2012).

36 Perez-Andres, M. *et al.* Human peripheral blood B-cell compartments: a crossroad in B-cell traffic. *Cytom. Part B.-Clin. Cy.* **78 Suppl 1**, S47-60 (2010).

37 Lavinder, J. J. *et al.* Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc. Natl. Acad. Sci. U S A* **111**, 2259-2264 (2014).

38 Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **15**, e538-548 (2014).

39 Hindorf, C. *et al.* EANM Dosimetry Committee guidelines for bone marrow and whole-body dosimetry. *Eur. J. Nucl. Med. Mol. I.* **37**, 1238-1250 (2010).

40 Galotto, M. *et al.* Stromal damage as consequence of high-dose chemo/radiotherapy in bone marrow transplant recipients. *Exp. Hematol.* **27**, 1460-1466 (1999).

41 Terstappen, L. W., Johnsen, S., Segers-Nolten, I. M. & Loken, M. R. Identification and characterization of plasma cells in normal human bone marrow by high-resolution flow cytometry. *Blood* **76**, 1739-1747 (1990).

Table 1

	10,000	30,000	100,000	1 million	3 million
1.1	21,490	59,565	175,656	2,734,933	7,922,299
1.2	14,142	29,638	73,197	485,204	1,813,627
1.3	14,142	24,495	44,721	248,742	746,381
1.4	14,142	24,495	44,721	141,421	420,253
1.5	14,142	24,495	44,721	141,421	244,949
2	14,142	24,495	44,721	141,421	244,949
5	14,142	24,495	44,721	141,421	244,949

Table 2

Subset	Source	Method	Cells	Species richness		Missing species	Entropy (eff. no.)		Min clone size, cells	Upper bound, clones
				Sample	Overall		Sample	Overall		
IgG^+ B cells, individual 1[22]	healthy adult	IgH+L single-cell	61,000	2,759	6,357 (5,569-7,527)	3,598 (2,810-4,768)	696	700 (700-700)	300	2 million
IgG^+ B cells, individual 2[22]	healthy adult	IgH+L single-cell	47,000	2,211	6,770 (5,335-9,174)	4,559 (3,124-6,963)	345	348 (347-348)	400	4 million
memory B cells (IgG, IgM, and IgA)[22]	healthy adult vaccinee	IgH+L single-cell	8,000	336	516	180	21	21	20,000	10 million
tetanus toxoid-specific plasmablasts[22]	healthy immunized adult	IgH+L single-cell	2,000	159	706	547	3.5	3.5	200	100,000
bone-marrow plasma cells[24]	healthy adult	IgH pooled DNA	25,943	14,337	36,276 (29,347-46,143)	21,939 (15,010-31,806)	13.44	14.39 (14.33-14.46)	80	3 million
non-tumor plasma cells[24]	multiple myeloma patient	IgH pooled DNA	30,426	325	530	205	0.51	0.53	200	70,000

Fig. 1

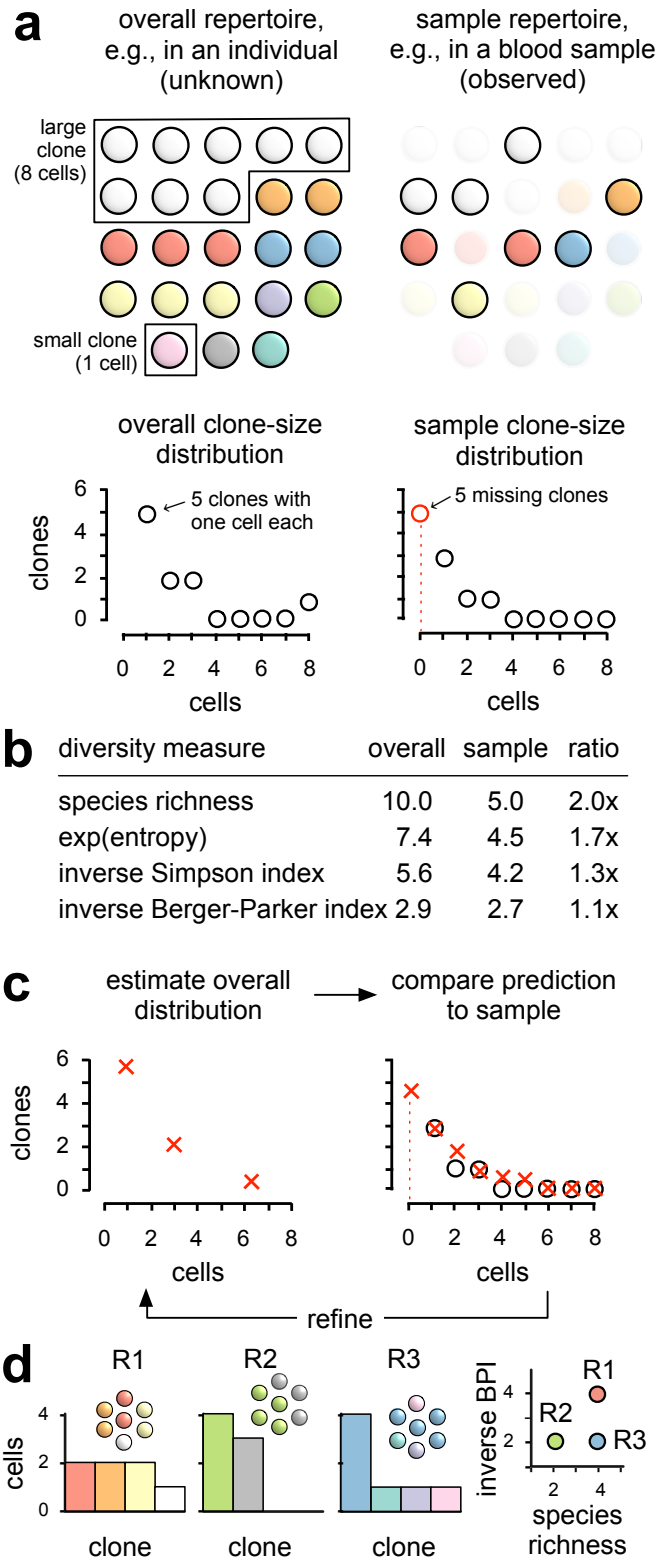
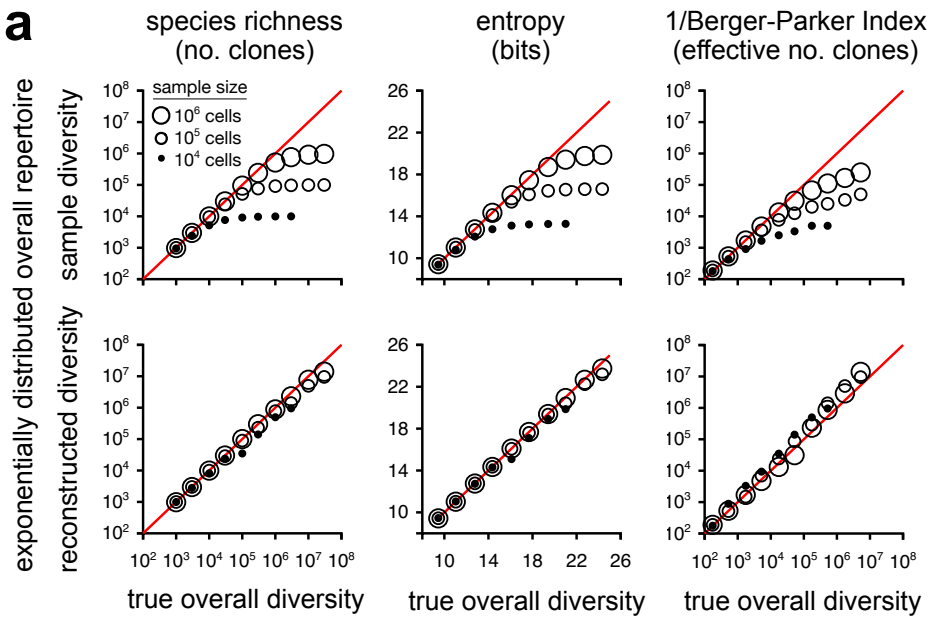


Fig. 2



b

repertoire type	Recon	Chao	tie	*n*
exponential, without noise	46%	36%	18%	56
bimodal, without noise	65%	20%	15%	60
with experimental noise	97%	3%	0%	400

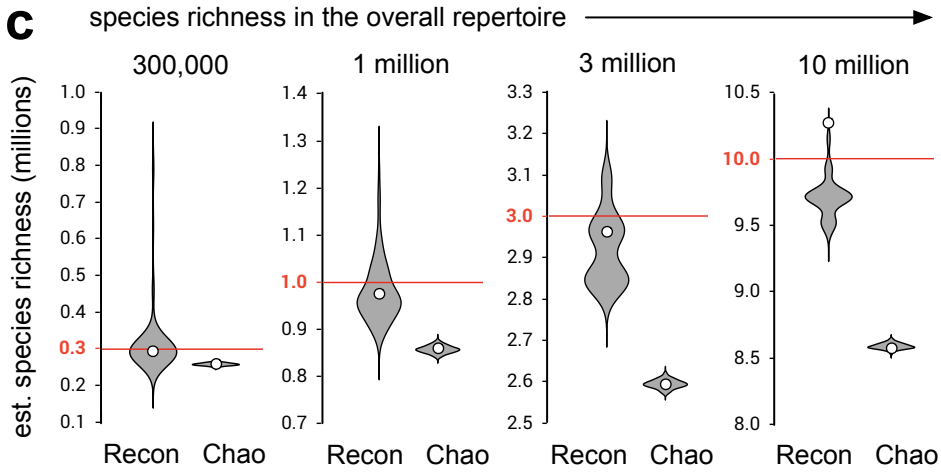


Fig. 3

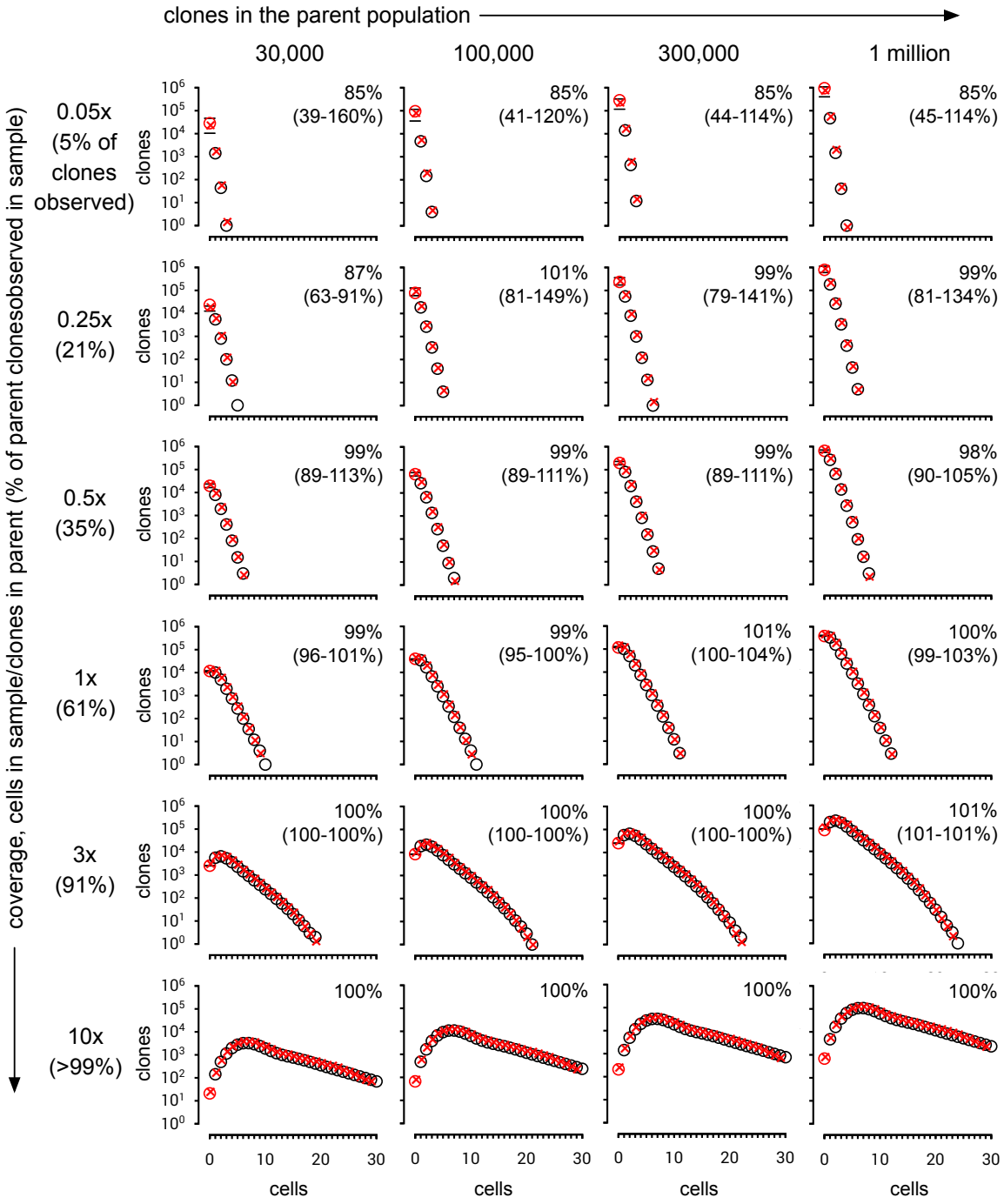


Fig. 4

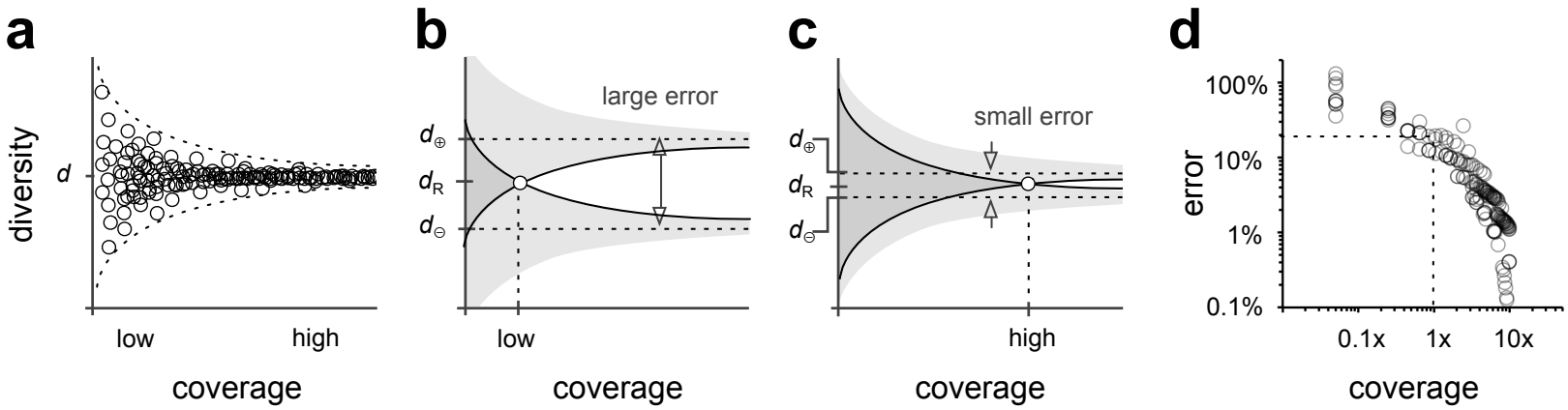
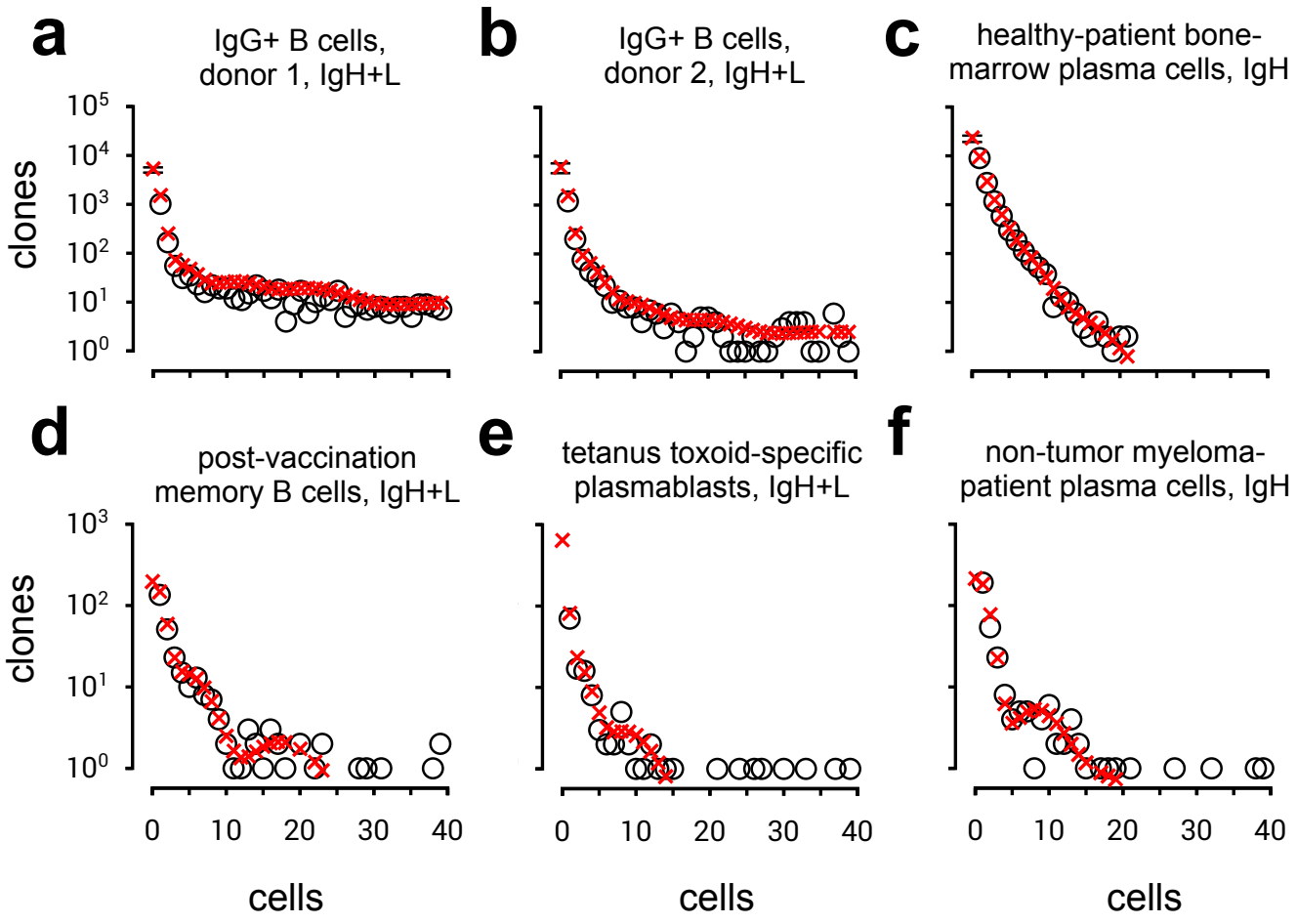

a
diversity
d
low
high
coverage
b
large error
$d_\oplus$
d_R
$d_\ominus$
low
coverage
c
small error
$d_\oplus$
d_R
$d_\ominus$
high
coverage
d
100%
10%
1%
0.1%
error
0.1x
1x
10x
coverage

Fig. 5



Kaplinsky et al., Supplementary Information

Contents

1. Detailed description of the Recon algorithm

Overview. The problem is, given the observation of the clone size distribution in a sample, to reconstruct the number of clones of each size in the parent or overall population from which a sample was taken (e.g. memory B cells in the peripheral blood).

By *clone size* we mean the number of cells that make up a clone. A clone made up of a single cell has clone size 1, while a clone made up of a million cells has clone size one million.

By the *clone size distribution* we mean the number of clones of each size (Fig. 1a). For the sample we use the notation n_i, where i indexes the clone size and n_i is the number of clones of that size. Thus n_1 is the number of clones represented in the sample by a single cell, n_2 the number of clones represented by 2 cells, and so forth. The number of clones that are present in the parent distribution but missing from the sample is represented by n_0, as they are represented by 0 cells in the sample. These are the *missing species*.

Experimentally observed clone size distributions are described by a sampling distribution from the parent population. The overall strategy of the Recon algorithm is to find a maximum-likelihood estimate (MLE) for parameters of a model describing the sampling distribution. The form of the model has an immediate interpretation in terms of the clone size distribution of the parent population.

Recon is based on a mixed-Poisson model for the contribution of each clone in the parent population to the sample:

$$p_i = \sum_j w_j \, Poisson(i;\, m_j)$$

$$= \sum_j w_j \frac{m^i \exp(-m_j)}{i!}$$

where w_j are *weights* and m_j are Poisson parameters. The weights w_j give the proportion of clones in the parent population with clone size *j*.

The parameters m_j give the mean number of cells a clone of size *j* contributes to a sample and are referred to as *means* below. They correspond to clone sizes in the parent population:

$$\text{Clone size } j \text{ in parent} = \text{number of cells in parent population} \times m_j / \text{ sample size}$$

The parameters therefore give a complete description of the clone-size distribution in the parent population.

If there are *k* different sizes in the parent population, so that the index *j* ranges from 1 to *k*, then there are a total of 2*k*-1 independent parameters, consisting of *k* independent sizes m_j and *k*-1 independent weights w_j, which sum to 1.

Assuming that the sample comes from a well mixed parent population, such as blood, this gives rise to a sampling distribution:

$$P(n_0, n_1, n_2 \ldots) = (n_{\text{total}}!/n_0!\, n_1!\, n_2! \ldots) \times (p_0^{n_0} p_1^{n_1} p_2^{n_2} \ldots) \quad (1)$$

Where n_{total} is the sum of all n_i.

The Recon algorithm addresses three fundamental problems in the search for parameters which maximize the likelihood (1) given the data n_i.

First is the need to determine the number of different clone sizes, k. This is addressed by starting with a homogeneous population in which all clones are the same size and refining the description of the population by adding clone sizes (incrementing k by 1) until no better fit can be obtained. A better fit must both (*i*) improve the fit by an amount that is larger than expected variation from sampling noise and (*ii*) improve the corrected Akaike Information Criterion (AICc). This loop is described in Steps 2, 7 and 8 below.

Second is that the likelihood is a non-linear function of the parameters and has local minima, whereas a global minimum is desired. In practice, for some fits, the AICc allows 20 or more parameters; searching such a high dimensional space requires a careful strategy to find a global minimum. To handle this problem, in Recon each step of the fit is run many times (110 in our implementation) from different starting points. These multiple different starting points often result in finding multiple local minima, from which the global minimum is selected. This loop is described in Steps 3 and 6 below.

Third, the likelihood of the data can be calculated directly from the $2k$-1 parameters of the mixed-Poisson distribution model given only the number of unseen species, n_0. Thus, the number of missing species must be jointly modeled as an additional parameter.

In order to handle this problem an expectation maximization (EM) approach (see references in main text) is used in which an expected value of n_0 is obtained from the remaining parameters and parameters are then refitted until self consistent values of parameters and of n_0 are obtained. This loop is described in Steps 4 and 5 below.

These three nested loops are shown in the flowchart in Fig. S1.

Step 1: Separate large from small clones. To simplify our calculations of n_0, the first step splits the observed clone size distribution into large clones and small clones. Our implementation uses a threshold of 30 cells.

Consider repeated sampling of the parent distribution. Any clone in the parent population that is large enough to contribute 30 cells to a sample will essentially always be represented in the sample; i.e., it will never contribute to the number of missing clones, n_0. Furthermore, the sampling error on such large clones will be relatively small, and the size of the clone in the parent population will scale linearly with the number of clones in the sample.

The main work of reconstruction must then be applied to the remaining small clones, whose contribution to the observed sample is less than 30 cells, and which correspond to clones in the parent population that are small enough to include clones that will contribute no cells to the sample and thus affect n_0. The remaining reconstruction steps are applied only to these small clones.

Step 2: Determine mean observed clone size. The mean size of all observed clones contributing to the fit (i.e. clones contributing less than 30 cells) is calculated. This is used to set the scale for initial guesses of clone sizes in step 3.

The initial parameters for the fit are set to empty lists of weights and means. This is recorded as the *current best fit*.

Step 3: Add a clone size to the parent distribution. Next, the algorithm adds a new distinct clone size to the parent population, in such a way that the new distribution maximizes the log likelihood. Because there are multiple maxima in the likelihood, this fitting (Steps 3-5) will be repeated for each of many starting points for the new clone size added to the same current best

fit in an attempt to find the best possible improvement. We used 110 starting points in our implementation.

Except on the first iteration of the fit, the weight of the new clone size is selected from the list of starting weights (Table S1). On the first iteration of the fit the newly added clone size is the only clone size, so the weight is 1.0. The mean for the new population is calculated by selecting a starting scale factor (Table S2) and multiplying by the mean size of the small clones.

Table S1: Starting weights

0.05
0.1
0.2
0.25
0.3
0.4
0.5
0.6
0.7
0.9
0.95

Table S2: Starting scale factors

0.001
0.02
0.07
0.3
0.6
0.9
1.1
1.2
1.3
1.4

The number of missing species is updated as

$$n_0 = n_{obs}/(1 - p_0)$$

where n_{obs} is the number of small clones observed in the sample (i.e. the sum of n_i for $0<i<30$).

Step 4: First EM step. Given the estimate of n_0, Recon maximizes log P, where P is given by Eq. (1) above:

$$P(n_0, n_1, n_2 \ldots) = (n_{\text{total}}!/n_0!\, n_1!\, n_2! \ldots) \times (p_0^{n_0} p_1^{n_1} p_2^{n_2} \ldots)$$

The n_i for $i>0$ are the observed number of clones represented by i cells in the sample, n_{total} is the sum of all n_i including n_0, and the p_i are the probabilities of a randomly selected clone giving rise to exactly n_i cells in the sample, as calculated from the mixed-Poisson model. In our implementation this is carried out using the L-BFGS-B minimization method from the scipy.optimize library.

Step 5: Second EM step. A new value for n_0 is estimated according to:

$$n_0 = n_{obs}/(1 - p_0)$$

.

This new value of n_0 is used to find maximum likelihood values for the parameters.

If the newly estimated value of n_0 is equal to the old value of n_0 then there has been no improvement, and so EM for the corresponding starting point is completed.

If instead the newly estimated value of n_0 differs from the old value of n_0 then Step 4 is repeated using the new estimate and starting from the parameter values given by the fit for the old n_0 estimate. This ensures that the end result of EM is a set of parameters that maximize likelihood and produce a self-consistent estimate for n_0.

The result is added to a list of *possible best fits.* As shown in Fig. S1, the algorithm returns to Step 4 until all starting points have been tried and the list of possible best fits contains 110 entries. Note that at this point the current best fit is not yet updated.

6: Compare the multiple minima that arise from the different starting points. After all 110 fits, each starting from different initial parameters, are complete, the MLE from among these 110 fits is selected.

The likelihood minimized in Step 4 treats n_0 as data, and maximizes likelihood given that data. However, solutions from different starting points will arrive at differing self-consistent values of n_0. In order to compare these solutions n_0 must be treated as a parameter rather than as data.

Treating n_0 as data we use Eq. (1) above. For practical purposes, since the n do not depend on the parameters, we maximize

$$\log(p_0^{n_0} p_1^{n_1} p_2^{n_2} \dots) = \sum_{i=0}^{\infty} n_i \log p_i$$

Because the p_i are known functions of the mixed-Poisson model parameters this is a straight-forward procedure.

In contrast, treating n_0 as a parameter we have the likelihood to be maximized:

$$P'(n_1, n_2, n_3, \dots | n_0) = \left(\frac{n_{\text{obs}}!}{n_1!\, n_2!\, n_3! \dots}\right) \times (p'^{n_1}_1 p'^{n_2}_2 p'^{n_3}_3 \dots)$$

Here the p'_i are not equal to p_i, (as can be seen e.g. by considering normalization) and depend on n_0. It is not straightforward to calculate the p'_i from the mixed-Poisson model parameters.

In order to calculate P' in terms of the mixed-Poisson model parameters we write log P' in terms of log P:

$$\begin{aligned} P(n_0, n_1, n_2 \dots) &= \left(\frac{n_{\text{total}}!}{n_0!\, n_{\text{obs}}!}\right) p_0^{n_0} (1 - p_0)^{n_{\text{obs}}} \left(\frac{n_{\text{obs}}!}{n_1!\, n_2!\, n_3! \dots}\right) \times (p'^{n_1}_1 p'^{n_2}_2 p'^{n_3}_3 \dots) \\ &= \left(\frac{n_{\text{total}}!}{n_0!\, n_{\text{obs}}!}\right) p_0^{n_0} (1 - p_0)^{n_{\text{obs}}} P' \end{aligned}$$

Then

$$\log P = \log\left(\frac{n_{\text{total}}!}{n_0!\, n_{\text{obs}}!}\right) + n_0 \log p_0 + n_{\text{obs}} \log(1 - p_0) + \log P'$$

so

$$\log P' = \log P - \log\left(\frac{n_{\text{total}}!}{n_0! n_{\text{obs}}!}\right) - n_0 \log p_0 - n_{\text{obs}} \log(1 - p_0). \quad (2)$$

Taking the log of Eq. (1) we can write

$$\log P = \log\left(\frac{n_{\text{total}}!}{n_0!\, n_1!\, n_2!\, \ldots}\right) + n_0 \log p_0 + \log\left(p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots\right).$$

Substituting this expression for log *P* into Eq. (2) we find:

$$\log P' = \log\left(\frac{n_{\text{obs}}!}{n_1!\, n_2!\, n_3!\, \ldots}\right) - n_{\text{obs}} \log(1 - p_0) + \log\left(p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots\right)$$

The first term on the right does not depend on parameters, so in order to maximize *P'* we select the fit giving the maximum value of:

$$\log\left(p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots\right) - n_{\text{obs}} \log(1 - p_0).$$

Because this is written in terms of the p_i it can be evaluated in terms of the mixed-Poisson model parameters, so it is straightforward to maximize. Note that directly comparing the log likelihoods treating n_0 as data between fits that have different values of n_0 is in practice misleading, and leads to a severe bias against large values of n_0.

The 2 fits with the highest log likelihoods are passed to Step 7

Step 7: Fit a best average starting point

The starting weights that led to the two best fits are averaged together to produce the best average starting weights. The starting means that led to the two best fits are averaged together to produce the best average starting means. The best average starting weights and means are then fit using Steps 4 and 5 of the algorithm.

The log likelihood of the resulting fit is computed as in Step 6.

The resulting 111 fits ordered from highest to lowest log likelihood is passed to Step 8 as the list of *candidate best fits*.

Step 8: Check sampling noise and minimum clone size. If an estimate of the number of cells in the parent population is available then it is possible to set a minimum size for clones in the parent population—namely 1 cell. However, in general such estimates may not be available, and the Recon algorithm does not rely on such information.

If there is no restriction on the minimum clone size then the algorithm can produce a perfect fit to n_1 in the observed clones by fitting a large number of clones, each of which contributes an unrealistically small fraction of a clone to the observed distribution. It is therefore necessary to introduce a minimum mean clone size.

The expected number of cells contributed to n_1 by the clones with the smallest m parameter in the candidate best fit is compared against the expected noise in n_1 arising from the remaining clones. In our implementation, the noise threshold on the remaining clones is calculated as three times the standard deviation from Poisson sampling.

If the contribution from the smallest clones in the candidate best fit with the highest log likelihood is larger than this noise threshold then it is passed to Step 8.

Otherwise, it is removed from the list of candidate best fits and the next candidate best fit is tested until a fit is found for which the contribution from the smallest clones is larger than the noise threshold. This fit is then passed to Step 9.

Step 9: Test for improvement of the AICc. The AICc is defined as

$$\mathrm{AICc} = 2q - 2\ln P' + 2q(q+1)/(N-q-1)$$

Where $q = 2k-1$ is the number of parameters and N is the number of observations. N is taken as the number of distinct clone sizes that being fitted, which in the case of Recon is the number of small clone sizes, which is 29 in our implementation.

The new AICc of the new candidate best fit is compared against the AICc of the current best fit. Note that in this step the candidate best fit has two more parameters (one weight and one population size) than the current best fit. This is what necessitates the use of the AICc. (In previous steps, comparisons were made only between fits with the sample number of parameters, so a simple log likelihood comparison sufficed.)

If the candidate best fit is not an improvement then the algorithm exits with the current best fit as its final result.

Otherwise the algorithm records the candidate best fit as the new current best fit and returns to Step 3 to search for a further improvement with additional parameters.

2. Upper bound on species-richness estimates

Any reconstruction of missing species using a small sample from a large population suffers from a fundamental limitation. Species that are too rare to have an appreciable chance of appearing in the sample cannot be estimated based upon the sample. As shown by Mao and Lindsay, this results in upper confidence intervals for the missing species that are formally infinite.

As discussed, Recon addresses this problem by only estimating those species that are large enough to have an appreciable chance of influencing the sample distribution in a meaningful way. (Note that while mixing distributions are often approximated as continuous, in reality they are discrete, so smallest fitted population will often be practically meaningful.) In many cases this estimate will be of interest.

But this still leaves the estimate for *all* species unbounded. The number of individuals in a population is of course an upper bound for the number of species. In many cases of interest, such as analysis of immune repertoires, it is relatively easy to obtain reasonable estimates of the total number of individuals. For example, an estimate of total cells can be obtained by scaling a cell count against total tissue or blood volume, e.g., 10^{10} B cells in the body.

Below we show how the Recon fit can be combined with an estimate of the number of all individuals in a population to get a sharper upper bound on the number of all species.

Recon produces an overall clone-size distribution. The smallest clone size in this distribution is described by two parameters: the fraction of all clones that are of this size, w_{min}, and a mean number of cells that it contributes to the sample, m_{min}. Clone sizes smaller than this contribute a mean of zero cells to the sample; however, it is possible that there are smaller clones in the parent population, clones so small that they both do not contribute to the sample and are invisible to our algorithm. Recon's estimate of the number of missing clones would not count such clones because it is not necessary to assume that they exist in order to obtain the observed sample clone-size distribution. However, if they were to exist, they would result in an undercount of the species richness in the parent. The goal in this section is to bound this potential undercount. One can then test its plausibility, as described in the main text.

The maximum undercount U_{max}, and therefore the desired upper bound, is obtained for the case that all the cells in clones smaller than m_{min} are actually singlets. How many would that be? The answer is given by

$$U_{max} = Rw_{min}m_{min}N/S$$

where R is (Recon's upper bound of) the overall species richness estimate, N is the total number of cells in the overall repertoire, and S is the sample size. Note the ratio S/N is the fraction

of cells in the overall population that are sampled; scaling m_{min} by S/N (yielding $m_{min}*N/S$) thus gives the smallest clone size in the overall repertoire that Recon can distinguish from singlets. Error bars on R and uncertainty in N contribute to uncertainty in the upper bound. Because generally $N > R$, upper bounds are larger than Recon's estimates. We note, however, that in our experimental datasets (see main text) comparison of upper-bound estimates to the error expected given the coverage (S/U_{max}) excludes U_{max} as a plausible estimate, given the observed R (Fig. 4d).

An example of a limiting case will illustrate how the formula for U_{max} works. Suppose an organism contains $N = 10^{10}$ B cells, and further suppose that every one of these is a distinct clone, so that each clone in the parent is made up of 1 cell. If a sample of $S = 10^6$ cells is taken, then the observed clone size distribution will consist of 10^6 singletons, i.e. n_1 = 1,000,000 and remaining $n_i = 0$. The best that Recon could do here would be to take a single population (that is w =1.0) and note that the mean contribution, m, of each clone in the overall repertoire must be less than 10^{-3}.

The value of m comes from the fact that no clone is observed twice, so that $(10^{-3})^2 * S < 1$. Note that in fact the true mean contribution of each clone to the sample is 10^{-4}. Taking $m = 10^{-3}$ will result in a severe undercount, but is all that can be said with confidence given the sample size.

The unseen species estimated by Recon will be given by

$$n_0 = \frac{n_{obs}}{1 - p_0}.$$

In this example $n_{obs} = S = 10^6$. Recon's estimate of p_0 will be given by 1-$p_{>0}$, where $p_{>0}$ is the chance that a clone contributes to the sample. Therefore the estimate of 1- p_0 will be $p_{>0} = 10^{-3}$.

Again, the true value of p_0 is much greater, but this is the best estimate possible given the sample. This results in an estimate

$$n_0 = \frac{10^6}{10^{-3}} = 10^9$$

The estimate of the species richness R is then $S + n_0$, which is approximately 10^9. In this extreme case, Recon therefore underestimates the true species richness by a factor of 10.

However, the formula for U_{max} is able to recover the true population. Since Recon fits only a single weight and mean, $w_{min} = w = 1.0$ and $m_{min} = m = 10^{-3}$. Then

$$U_{max} = \frac{10^9 \times 1.0 \times 10^{-3} \times 10^{10}}{10^6} = 10^{10}.$$

As expected, in this case Recon adds no further constraint. If every individual in the sample is from a different species then the only sensible upper bound for the number of species is N.

Now consider a case in which $S = 10^6$ cells are again sampled, but now the observed distribution has n_1 = 900,000, n_2 = 35,640, n_3 = 6,667, n_4 = 1,500, n_5 = 400, n_6 = 100, n_7 = 10, n_8 = 5, n_9 = 1 and remaining n_i = 0..

In this case the sample contains 944,323 clones. Recon fits 19,919,406 missing species for a total richness of 20,863,729 detectable clones. The w_{min} of the fit is 0.252 and the m_{min} is 0.041. The upper bound is now:

$$U_{max} = \frac{2.09 \times 10^7 \times 0.252 \times 0.041 \times 10^{10}}{10^6} = 2.16 \times 10^9.$$

The upper bound of N can therefore be usefully reduced by a factor of almost 5.

3. Power calculations for species richness

To obtain the minimum number of cells suggested to power an experiment detecting a specified difference, we required a number of cells sufficient to separate the expected sample means by at least one error bar, where the error bar is calculated as described in the main text.

If experimentally reconstructed missing species from multiple identical samples with identical true overall diversity are taken to be normally distributed, then our calculation corresponds to a t-test at p=0.05 using our error bar as an estimate of the 3 times the standard deviation of this distribution.

Fig. S1

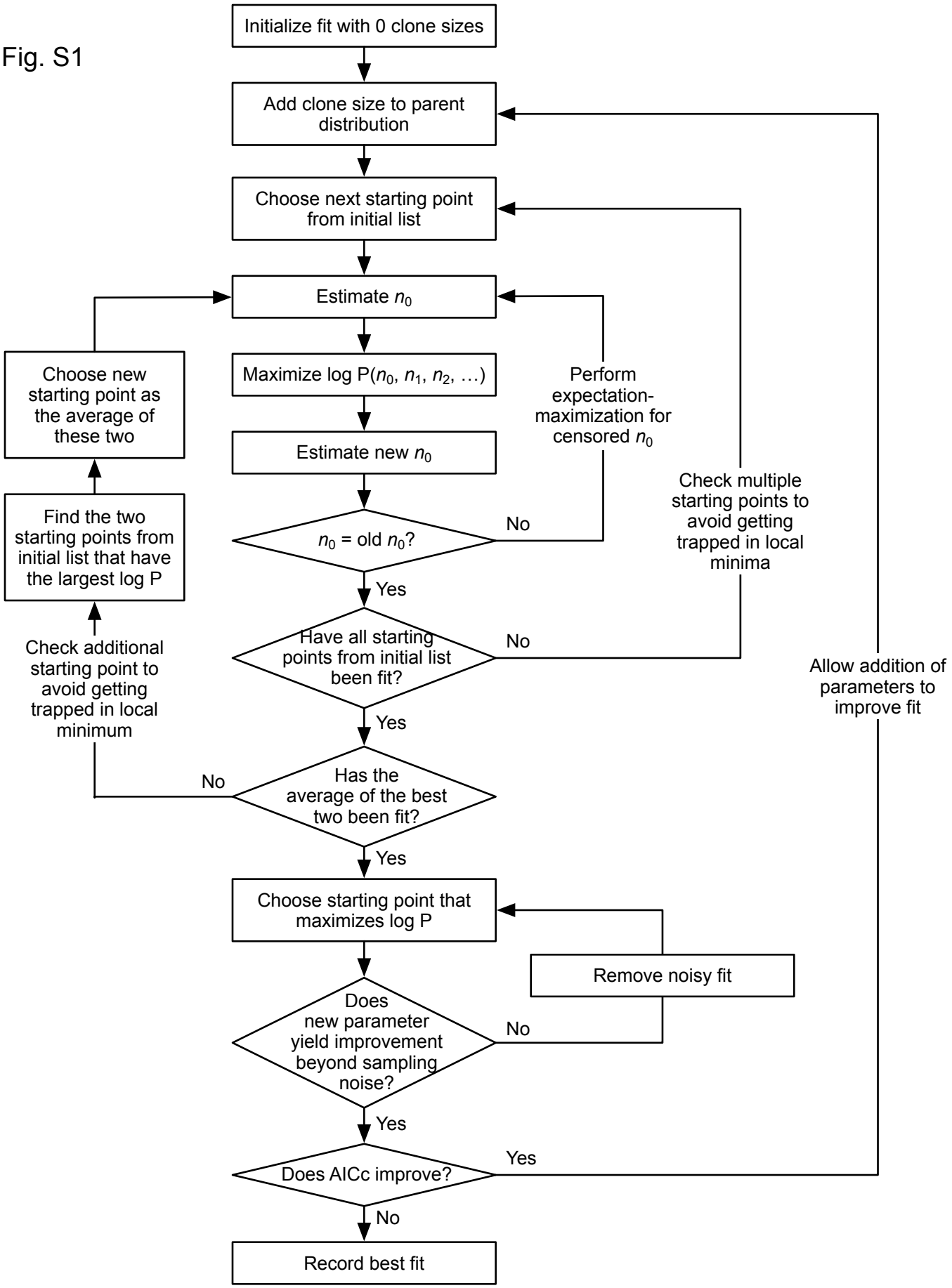


Fig. S2a-c

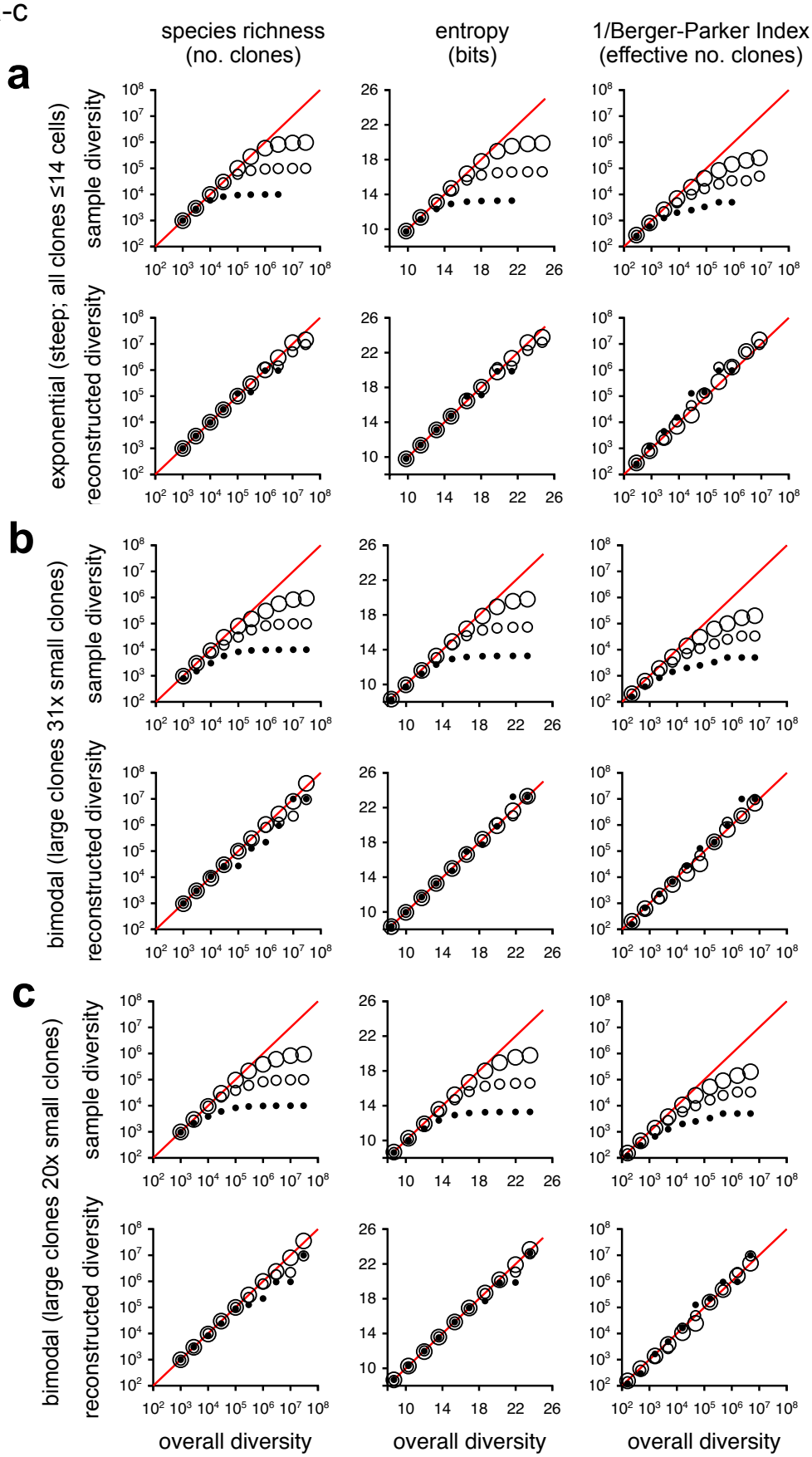


Fig. S2d-f

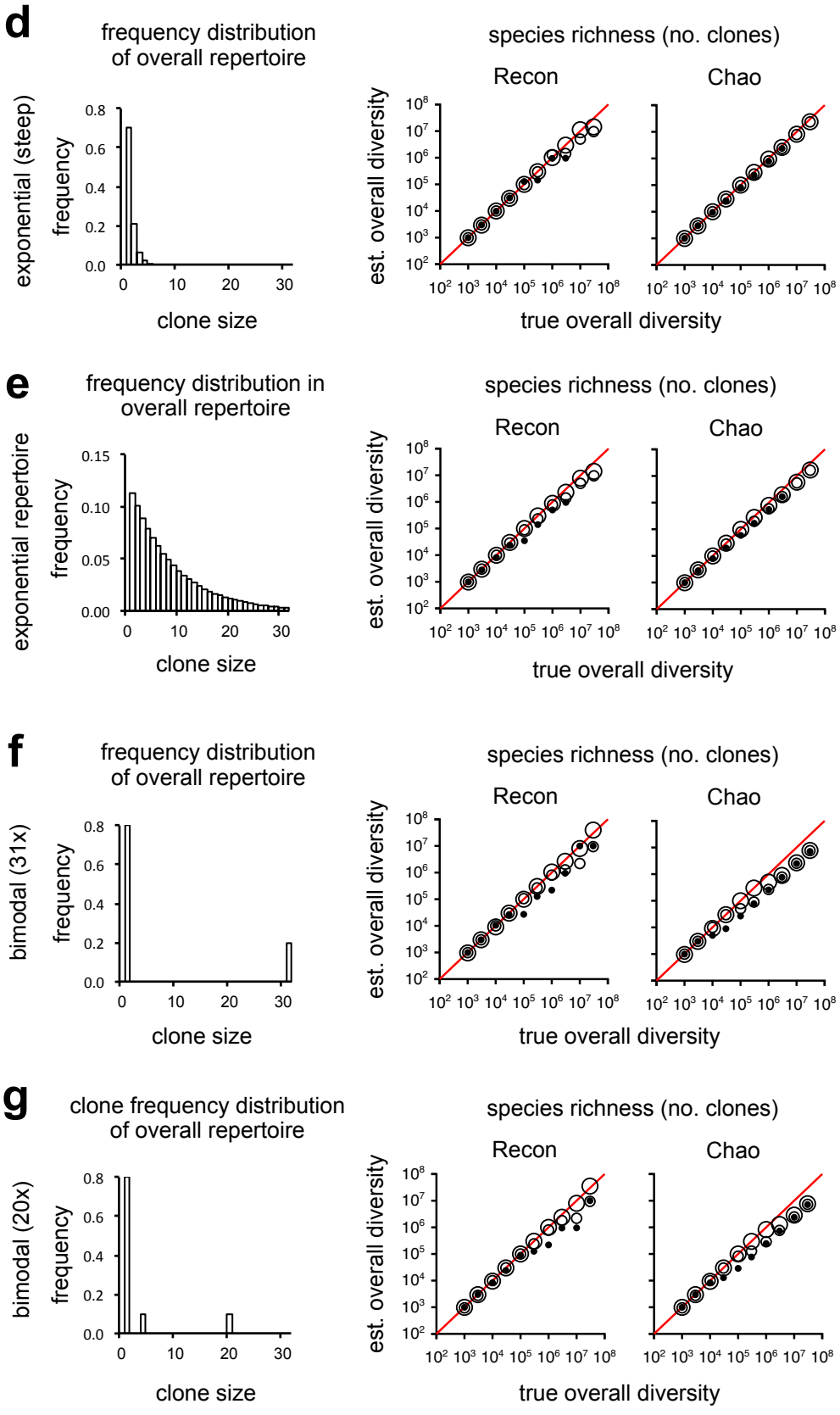


Fig. S3