nature COMMUNICATIONS

ARTICLE

Received 12 Mar 2016 | Accepted 9 May 2016 | Published 15 Jun 2016

DOI: 10.1038/ncomms11881 OPEN

Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples

Joseph Kaplinsky[1,2,†] & Ramy Arnaout[1,2,3]

The diversity of an organism'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, because of 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 solve this problem, we developed Recon, a modified maximum-likelihood method that outputs the overall diversity of a repertoire from measurements on a sample. Recon outputs accurate, robust estimates by any of a vast set of complementary diversity measures, including species richness and entropy, at fractional repertoire coverage. It also outputs error bars and power tables, allowing robust comparisons of diversity between individuals and over time. We apply Recon to *in silico* and experimental immune-repertoire sequencing data sets as proof of principle for measuring diversity in large, complex systems.

[1] Department of Pathology, Beth Israel Deaconess Medical Center BIDMC East/Dana 615, 330 Brookline Avenue, Boston, Massachusetts 02215, USA. [2] Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02215, USA. [3] Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA. [†] Present address: Department of Micro- and Nanotechnology, Building 423, Room 220, Produktionstorvet, Technical University of Denmark, 2800 Kongens Lyngby, Denmark. Correspondence and requests for materials should be addressed to R.A. (email: rarnaout@gmail.com).

Recent technological advances are making it possible to study B- and T-cell repertoires in unprecedented detail[1]. Of special interest is repertoire diversity, defined as the number of different B- or T-cell receptors on cells present in an individual, tissue (for example, peripheral blood, bone marrow), tumour (for example, tumour-infiltrating lymphocytes) or cell subset (for example, influenza-specific IgG^+ B cells). This interest follows observations that immune-repertoire diversity correlates with successful responses to infection, immune reconstitution following stem-cell transplant, the presence or absence of leukaemia, and healthy versus unhealthy ageing[2–5]. The reliability of such observations depends on the ability to measure diversity—and differences in diversity—in overall B- or T-cell populations accurately and with statistical rigour from clinical and experimental samples. Similar requirements also arise in the study of cancer heterogeneity, microbial diversity and high-throughput sequencing, as well as beyond biology[6–9]. However, measuring diversity is more complicated than it may seem, for three reasons.

First, 'diversity' may refer to any of several different measures. The most familiar diversity measure is the number of different species in a population: the species richness. An example of species richness is the number of B-cell clones in an individual (where 'clone' denotes cells with a common B- or T-cell progenitor). Other diversity measures provide complementary information about the size-frequency distribution of species in the population. For example, the Berger–Parker index (BPI) measures clonality, that is, the dominance of the single largest clone (Fig. 1)[10]. Diversity measures that have been used on immune repertoires include species richness, Shannon entropy (henceforth 'entropy') and the Simpson and Gini-Simpson indices[11–14]. Of these, species richness is unique in that it takes no account of the frequency of each species. In contrast, entropy and other measures systematically down-weight or undercount rarer clones. The above measures (and many more) are related through a mathematical framework described by Hill[15,16]. Using simple mathematical transformations, this framework allows 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). The point here is that different diversity measures provide complementary information: two distinct repertoires can have the same species richness but different entropies or BPIs, and vice versa (Fig. 1d)[10]. Thus, no single measure is likely to capture all of the features of interest in a given repertoire. Consequently, methods for measuring immune-repertoire diversity should be capable of outputting any diversity measure.

Second, the diversity of a sample (for example, a 5-millilitre clinical blood sample) can differ markedly from the diversity of the overall repertoire from which it derives (for example, the 5 l of blood in the body). Although blood and tissue samples may contain thousands or millions of B or T cells, these are only a fraction of the billions of such cells that may comprise an overall repertoire. Consequently, some clones in the overall repertoire, especially small clones, almost always go unsampled and thereby undetected in measurements on samples (Fig. 1a). As a result, sample diversity usually underestimates true diversity (Fig. 1b). This phenomenon is known as the missing-species problem[17]. Weighted diversity measures (for example, entropy) are less sensitive to missing species than is species richness, as they down-weight the small clones that are most likely to be missing. However, using weighted measures as a substitute for species richness has drawbacks. First, it is unclear what information is

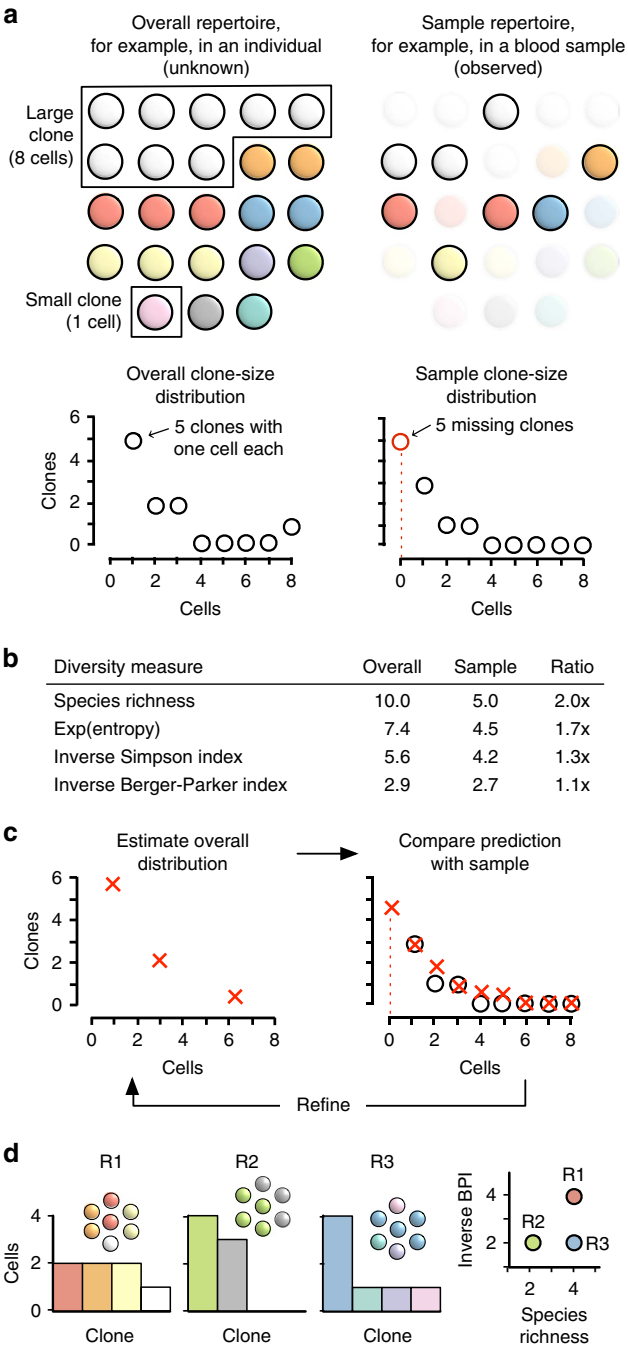

Figure 1 | Overall repertoires versus samples. (**a**) An overall repertoire (top left) and a random sample of this repertoire (top right), together with respective clone-size distributions from the overall repertoire and sample (bottom). Each circle denotes a cell; different colours denote different 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 (BPI); R2 and R3 have the same BPI but different species richness.

lost or biased by selectively ignoring small clones. Second, even using weighted measures, sample diversity will approximate overall diversity only when clone sizes (the number of cells per clone) in the sample approximate clone sizes in the overall population; however, clone sizes will inevitably be biased by the phenomenon of sampling noise. Note that unlike experimental error, which can be minimized, sampling noise is intrinsic to sampling, and will affect measurements even under perfect experimental conditions (for example, even if every cell in a sample is counted and perfectly annotated). Consequently, depending on the clone-size distribution and diversity measure, sampling can misrepresent overall diversity even when using weighted measures (Fig. 1b and below).

Third, real-world experiments will always exhibit some degree of experimental error, which manifests as noise in sample measurements. Sources include quantification error due to imprecise cell counts, amplification dropouts and jackpot effects; sequence error from amplification and sequencing; and annotation error introduced during data processing. Measuring diversity accurately requires methods that address not only the missing species problem and sampling noise, but experimental noise as well.

Existing methods for addressing the missing species problem either output only a single diversity measure (species richness) for the overall population, or else have known or suspected problems scaling to the complexity of immune repertoires. The first category includes Fisher's gamma-Poisson mixture method, a parametric method that has been used on T-cell repertoires, which involves a divergent sum that can result in large uncertainties[18–20]; the phenomenological approach of extrapolating from curve fitting[13,14,21,22]; and the Chao estimator (CE), a fast and simple calculation that avoids divergent sums and has been widely used in ecology[23,24]. The second category includes maximum-likelihood approaches such as the state-of-the-art methods of Norris and Pollock (NP)[25,26] and Wang and Lindsay (WL)[27]; however, to our knowledge, these have not been tested on, or are known not to scale to, highly complex populations like repertoires; or else make restrictive assumptions about the clone-size distribution of the overall repertoire and therefore are not generalizable[28]. Moreover, because a higher-likelihood fit can often be had by adding more small clones, existing maximum-likelihood approaches yield estimates that may overestimate diversity by orders of magnitude or be entirely unbounded—that is, they may find that the best estimate of diversity in the overall population is infinity[29].

We move beyond these shortcomings using a new algorithm, Recon—reconstruction of estimated clones from observed numbers—a generalized high-performance modified maximum-likelihood method that makes no assumptions about clone sizes or clone-size distributions 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 based on the expectation-maximization (EM) algorithm[6,30]. 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 subject to the no-overfitting constraint (Supplementary Fig. 1). The only assumptions Recon makes are that the overall repertoire is large relative to the sample and 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 data set 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 any 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 to overall diversity and (vi) a resampling of the overall distribution for comparison to the sample and plots thereof. Recon can be run on tumour clones, microbial species, sequence reads or other populations, including non-biological ones. Recon can also generate tables for power calculations and experimental design.

Recon embodies six improvements over the previous state of the art. First, to avoid dependence on initial conditions or becoming trapped in local maxima, Recon 'scans' a 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 (Supplementary Fig. 4). Second, Recon 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. Fifth, it improves over previous maximum-likelihood models in avoiding unbounded uncertainties, for example, regarding bounds on overall diversity estimates. And sixth, it is substantially faster (Fig. 2b,c).

Current methods tend to overestimate species richness when coverage is low, as small clones added to the estimate result in overfitting of the sample distribution—in the limit, as mentioned, 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). Beyond that, 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 adding more 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, in principle there may be clones in the overall repertoire that are too small to detect. In this case, Recon can be used to calculate a strict upper bound, U, on species richness that includes clones that may be 'hiding' (Methods and Supplementary Methods). However, we note that even in this case, in practice, for a given sample, the smallest clones detected may still be the smallest clones there are (the case for our *in silico* repertoires; below).

Validation. We validated Recon on *in silico* repertoires that spanned nearly five orders of magnitude of overall diversity (300 to 10 million clones) and a wide range of clone-size distributions: from steep, that is, 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, 1,711 in all, with and without simulated experimental noise (Methods). These repertoires served as gold

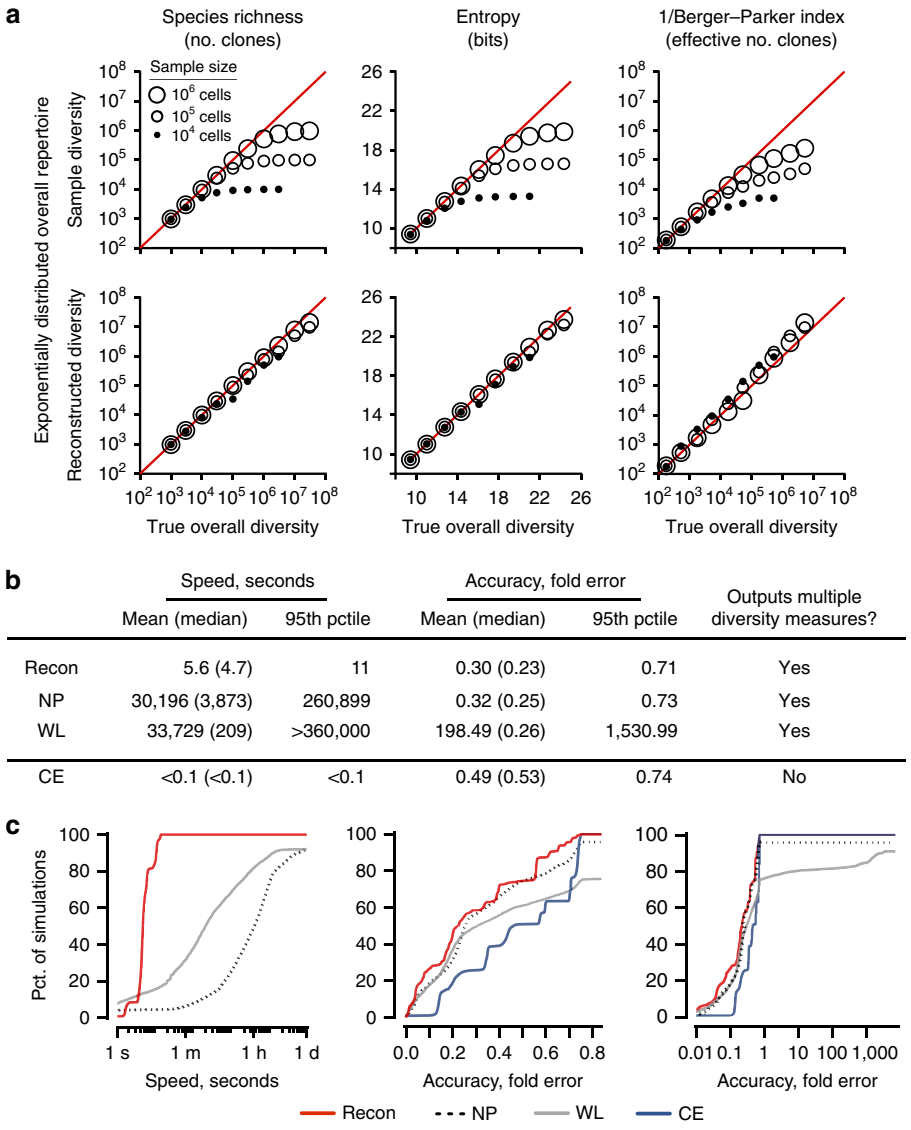

Figure 2 | Comparison of diversity estimates. (**a**) Sample diversity (top) and Recon's estimate (bottom) of overall diversity versus true overall diversity for three different sample sizes—10,000 cells (filled circles), 100,000 cells (small open circles) and 1 million cells (large open circles)—for a representative gold-standard distribution without noise (shown in Supplementary Fig. 2e, left panel; see Supplementary Fig. 2 for additional examples). Coverage is defined as the number of cells in the sample divided by the effective number of clones in the overall population. Red line, unity (zero error). Left-to-right: species richness, entropy and the inverse Berger–Parker index. (**b**) Performance summary of Recon versus two other state-of-the-art methods for estimating any overall diversity measure (NP and WL) as well as a method for estimating only species richness (CE) on 3,200 noisy distributions, 100 realizations of noise for each of 32 combinations of exponential and multimodal distributions (Methods), coverage (0.05–0.3x) and overall diversity (100,000 to 3 million clones in the overall population). (**c**) Cumulative distribution of performance for distributions in **b** showing Recon is much faster than NP and more accurate than WL, which could be off by orders of magnitude.

standards. We sampled a known number of cells from each, for coverage ranging from 0.01x to 10x, and used Recon to reconstruct overall repertoires from each sample. (Coverage is the number of cells in the sample divided by the number of clones in the overall repertoire.) 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, to illustrate the extent of the problem Recon solves, 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, defined as the difference between sample and overall diversity, grew as coverage fell below 1x, because samples cannot have 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 (for example., leukemic clones; Fig. 2 and Supplementary Fig. 2). 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 across the range of

diversity measures, even at <1x coverage (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. Recon was also robust to noise (Fig. 2b,c).

To visualize self-consistency, we resampled from the overall repertoires we reconstructed from our gold-standard distributions in order to compare the resulting sample clone-size distributions to those of the original samples. We found excellent agreement between predicted and observed frequencies of clone sizes across the range of overall diversities and levels of coverage, including on numbers of missing clones (Fig. 3). Recon's ability to estimate the number of missing clones accurately was a key contributor to the accuracy of its overall diversity estimates. The number of missing clones depended strongly on the number of singlets (clones represented by a single cell) and doublets (two cells) in the sample: large singlet-to-doublet ratios, with enough of both for low sampling noise, gave more accurate estimates.

In head-to-head comparisons on 3,200 *in silico* samples with experimental noise (Methods), Recon was both faster and more accurate than the prior methods NP and WL, which like Recon, can be used to estimate overall diversity by multiple diversity measures (Fig. 2b,c and Supplementary Fig. 3). Specifically, Recon's median runtime of 4.7 s (95th percentile, 11 s) was >40x faster than WL and >800x faster than NP, both of which often took hours and sometimes days to complete (Fig. 2c). Recon's median error of 0.23x was smaller than that of NP (0.25x) and WL (0.26x), which was often off by orders of magnitude (mean, 198x; 95th percentile, >1,500x). Recon was also more than twice as accurate as CE (0.53x median error; Fig. 2b,e), which is fast but limited to outputting species richness.

Error bars and power calculations. Detecting reliable differences in overall diversity requires that diversity estimates have reliable bounds. Recon outputs two types of bounds: error bars on overall diversity (more precisely, on the effective number of clones greater than or equal to a minimum detected clone size) and a maximum-possible overall species richness, U (Supplementary Methods).

To build error bars, we first sampled gold-standard repertoires systematically across three orders of magnitude of coverage (0.01x–10x). For each sample, we used Recon to estimate overall diversity. Because higher coverages produce better estimates, the resulting error profile converges with increasing coverage to the true overall diversity (Fig. 4a). The upper and lower contours of this profile correspond to the largest and smallest values of estimated diversity that are consistent with a given true diversity. To make an error bar for a given estimated diversity, Recon uses the contours of the error profile to find the true diversities for which the estimated diversity is at the lower bound and the upper bound. These respectively define the upper and lower error bars (Fig. 4b,c). Following cross-validation, we adjusted our error profile slightly so that error bars reflect 95% confidence intervals (Methods). Combining error profiles across all samples suggests that ≥1x coverage generally produces error bars of ≤10% for overall species richness (Fig. 4d), consistent with our previous observations (Fig. 2).

Recon uses this error-bar framework to determine the coverage required to confidently detect differences in diversity between samples (for example, between individuals or over time). Given an order-of-magnitude estimate of the overall diversity for two samples, it outputs the minimum sample size for which error bars for overall diversity estimates from these samples would not overlap, at detection thresholds ranging from, for example, 1.1x to 5x (Table 1). This sample size is the minimum 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 $P = 0.05$ (t-test; Supplementary Methods). Not surprisingly, detecting larger differences requires smaller sample sizes; less obviously, 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. So an experiment designed to detect a 1.1x (10%) difference in species richness between two samples, in which the samples are drawn from overall repertoires that have ~100,000 clones, will require ≥313,792 cells from each sample for analysis. This is the number of cells in the sample that are in small (≤30 cell-) clones that Recon requires to perform reconstruction; if half of the cells in a sample of 314,000 cells belong to a single large clone, for example, because of leukaemia, the remaining half comprising the non-leukaemic clones will be sufficient to detect a 20% difference in the species richness of the non-leukaemic portion of the repertoire (which requires ≥153,543 cells), but not 10%.

To test Recon and our error-bar framework beyond exponentially and multimodally distributed samples, 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 Methods)[29]. Three- and four-point mixture models, a logit normal model, a log-gamma model and a beta model gave variable estimates that ranged from 2,930 to 3,494 overall species, with non-overlapping error bars that ranged from 2,867 to >10,000. In contrast, Recon returned an estimate of 3,014 overall species, with error bars (2,709–3,513) that bracketed the range of other models' estimates, suggesting Recon improves on multiple methods beyond WL and NP in arbitrary and/or difficult cases.

Experimental data. Having validated Recon, we next applied it to six experimental data sets: four of paired heavy-and-light chain sequence and two of heavy-chain sequence (Methods). We used the authors' clone definitions—clusters of reads with ≥96% nucleotide identity in heavy-chain complementarity determining region 3 (CDR_H3)[31] 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, artefacts are possible, and sequencing is only semi-quantitative. Because such data sets reflect the current state of the art in the field and are used for diversity measurements, we considered them as (imperfect) samples and used Recon to estimate diversity for the corresponding overall repertoires (Table 2). As with our gold-standard samples, resampling showed excellent agreement with the observed data (Fig. 5). For four of the six repertoires, we found that missing species accounted for the majority of clones: that is, half of all clones are unseen, and species richness in the sample underrepresents overall species richness by 2x. Entropy was generally very similar between samples and overall repertoires, resulting from very large clones and/or PCR jackpot effects that contribute disproportionately to the entropy calculation. Thus, in these data sets, overall species richness, estimated using Recon, captures information lost during sampling that entropy does not.

Discussion

High-throughput technologies enable highly detailed descriptions of B- and T-cell repertoires. That these descriptions are generally of samples, and not for example, blood or tissue repertoires overall, may seem to be a distinction without a difference when samples contain many cells. However, and perhaps

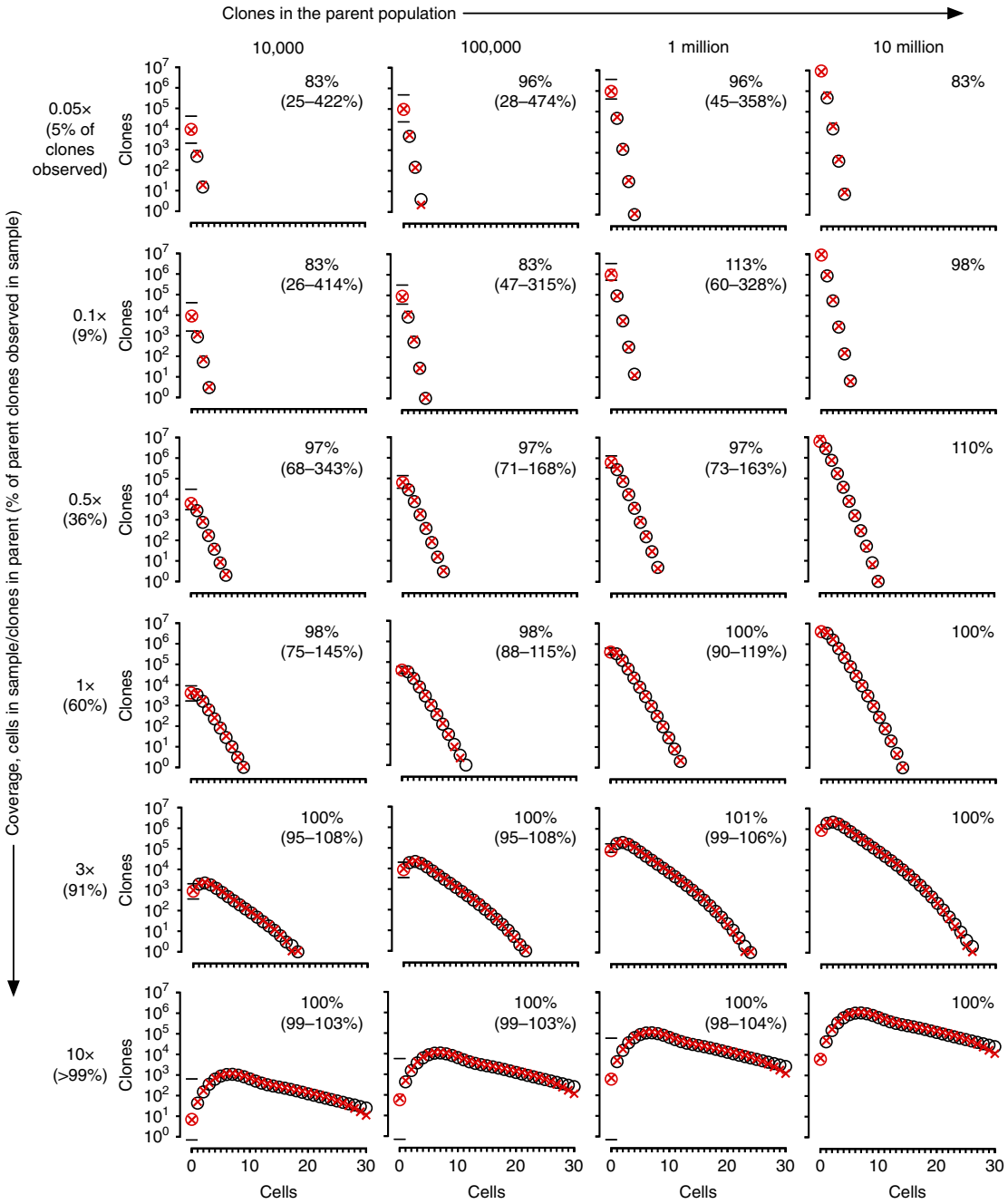

Figure 3 | Predictions versus simulated observations, *in silico* gold standards. Shown are fits to observations from representative gold-standard distributions of the shape shown in Fig. 2e, left panel. Left-to-right: overall distributions with increasing numbers of clones. Top-to-bottom: increasing 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.

counterintuitively, it turns out to be critical for estimating overall diversity. Unless the number of cells in a sample exceeds the number of clones in the overall repertoire by ~3- to 10-fold (Fig. 3), sample and overall diversity may bear little relation (Fig. 2a, Supplementary Fig. 2a–c). Importantly, this discrepancy is not a technological shortcoming but an inherent constraint of random sampling (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 gold-standard 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 (Fig. 2 and Supplementary Fig. 2). It is unsurprising, then, that sample entropy can also

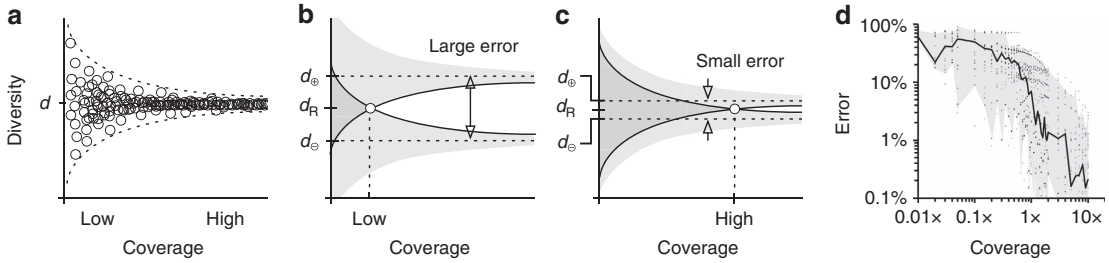

Figure 4 | Error bars. (**a**) A schematic representation of Recon's diversity estimates (open circles) from a single gold-standard *in silico* repertoire with overall diversity d for many different levels of coverage (= sample size/d). We used the absolute value of the proportional error of the worst fit at each level of coverage, making an error profile that is vertically symmetric around d. Given a test sample, Recon first estimates the overall diversity, d_R, and the coverage. (**b**) Using the error profile, it 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 errors from all 1,711 gold-standard repertoires into a single plot suggests that ≥1x coverage generally gives error bars of 5–10% for species richness (line, median; shaded area, 5th–95th percentiles).

Table 1 | Power calculations.

	10,000	30,000	100,000	1 Million	3 Million
1.1	34,634	118,418	313,792	2,211,303	16,230,339
1.2	19,103	56,989	153,543	1,277,637	10,598,339
1.3	14,142	28,206	85,156	649,124	1,947,385
1.4	14,142	27,711	70,415	639,665	1,919,012
1.5	14,142	27,238	64,982	630,590	1,891,799
2.0	14,142	24,495	64,977	510,381	1,524,687
5.0	14,142	24,495	44,721	141,421	244,949

Table entries give the minimum number of cells that must be analysed 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 grey. See Supplementary Methods for details.

Table 2 | Diversity estimates for experimental data sets from humans.

Subset	Source	Method	Cells	Species richness		Missing species	Entropy (clones)		*U*, clones (min. clone size, cells)
				Sample	Overall		Sample	Overall	
IgG^+ B cells, individual 1[31]	Healthy adult	IgH + L single-cell	61,000	2,759	5,870 (4,761–8,395)	3,111 (2,002–5,636)	696	700 (691–720)	1 Million (400)
IgG^+ B cells, individual 2[31]	Healthy adult	IgH + L single-cell	47,000	2,211	4,616 (3,374–7,000)	2,405 (1,163–4,789)	345	348 (327–373)	5 Million (700)
Memory B cells (IgG, IgM, and IgA)[31]	Healthy adult Vaccinee	IgH + L single-cell	8,000	336	473 (446–614)	137 (77–245)	21	21	14 Million (30,000)
Tetanus toxoid-specific plasmablasts[31]	Healthy immunized adult	IgH + L single-cell	2,000	159	239 (200–313)	80 (41–154)	3.5	3.5	300,000 (1,000)
Bone-marrow plasma cells[37]	Healthy adult	IgH pooled DNA	26,000	14,337	37,110 (27,350–58,916)	22,773 (13,013–44,579)	11,148	21,582 (20,891–22,572)	4 Million (80)
Non-tumour plasma cells[37]	Multiple myeloma patient	IgH pooled DNA	30,000	325	703 (563–1,081)	378 (238–756)	1.4	1.4	80,000 (80)

Summarized are Recon's estimates of overall diversity for six data sets; its estimate of the number of missing species; comparisons to sample diversity, for species richness and entropy (given as effective numbers; 2^{bits}); upper bound (U) for species richness that includes potential 'hiding' clones and the minimum detected clone size (see the main text). 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.

underestimate overall entropy in these repertoires (Fig. 2 and Supplementary Fig. 2). Additional caveats apply to experimental data sets. 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 affect species richness and entropy in the experimental data sets we studied (Table 2). Better quantification (for example, via barcoding and robust clonality modeling) would mitigate these biases but not the bias intrinsic to sampling, which Recon addresses.

Recon outperforms prior methods even for large, complex clone-size distributions, at fractional coverage, and in the presence of experimental noise (Figs 2 and 3 and Supplementary Fig. 2). Notably, Recon avoids WL's major failure modes: the 10–50% of cases in which WL unpredictably takes hours or days to run and/or overestimates diversity by orders of magnitude. Recon's characteristic runtime of seconds to a minute is especially faster than NP, and negligible relative to the hours-to-days of current sequence-processing pipelines. These advantages are not unexpected given that Recon was designed for handling samples from large, complex and arbitrary distributions.

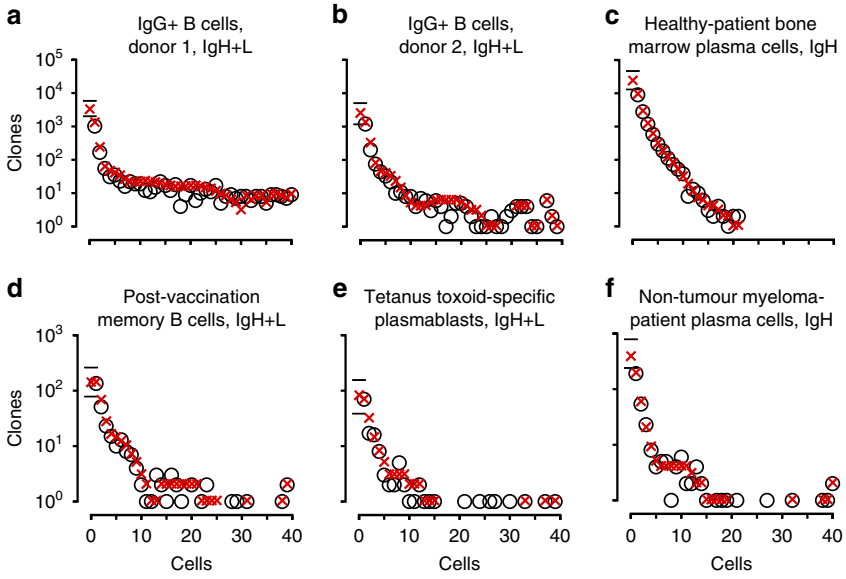

Figure 5 | Predictions versus observations, experimental data. Shown are Recon's estimates of overall diversity for six experimental data sets. These included (**a**,**b**) immunoglobulin heavy (IgH)- and light-chain (IgL) paired-chain sequencing experiments from IgG$^+$ B cells from the blood of two 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.

Error bars and power tables are necessary steps towards being able to compare diversity between samples and over time and thus for evaluating diversity as a potential 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 in validation studies and squaring previous models[29]. Its power tables offer guidance for sample requirements during experimental design and suggest expected limitations for different studies. For example, measuring the species richness of naïve repertoires of $\sim 10^7$ clones[32,33] will likely require phlebotomy or apheresis samples; even then, detecting 50% differences is probably the limit (Table 1). Meanwhile, measuring diversity for effector/memory subsets should require only routine blood draws (2–6 ml), which should detect sub-fold differences. For marrow, spleen, tumour, granuloma or abscess samples, the investigator must decide whether the sample is well mixed, which Recon requires.

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. Overcoming these limitations will improve our understanding of overall diversity, a defining characteristic of complex systems that we can now better measure.

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 EM algorithm, in which a rough approximation of the parent distribution is refined iteratively until no further improvement can be made without overfitting[30].

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[34], we estimate the number of missing species by calculating the expected clone-size distribution for a (Poisson) sample of the parent distribution (see the 'Sampling' below) and applying the Horvitz–Thomson estimator[35]. 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[34]. The result (pending a check for overfitting, below) 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 maximum likelihood estimate (MLE). Note that whereas the sample distribution generally traces out a smooth curve, the MLE parent distribution is spiky, reflecting the limited resolution that information in the sample distribution provides about the parent 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, for example, a singleton contributing >1 cell is negligible and (iii) no single clone is a large fraction ($\sim$30% or more) of the parent population. In practice, condition (iii) is satisfied by counting large clones directly (see the '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 $\leq$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. (Note that observing an absence of clones at a given size counts towards the number of observations used for calculating the Akaike information criterion.)

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 local maxima. To find global maxima, we developed a 'scanning' approach in which we applied EM to many starting clone sizes and frequencies (56 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 Methods). 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 starting points. Runtime and (to some extent) accuracy correlate with the number of 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[15,36]. These form a series in which the index reflects the extent to which counts are weighted towards 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 ${}^0D$ ('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 the 'Validation' below and in the main text) and from several published data sources (see the 'Experimental data' in the main text). These ${}^qD$ are a function of clone frequencies p_i, where i indexes each clone and the frequencies are normalized to $\sum_i p_i = 1$, defined as ${}^qD(p) = (\sum_i p_i^q)^{1/(1-q)}$ (ref. 36).

We calculated ${}^0D$ by simply counting the number of different clones, ${}^1D$ according to $\exp(-\sum_i p_i \ln p_i)$, ${}^2D$ 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 Recon against CE, NP and WL 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 species richness by the listed methods and the other above measures for all but CE (which outputs only species richness), and comparing these estimates against the (known) calculated diversities of the original parent distributions. We studied three families of test distributions in detail: (i) 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; (ii) '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; (iii) bimodal distributions with the largest clones an average multiple of the size of the smallest clones (for example, 20–30x) in the overall population. We tested these distributions systematically by varying the steepness from very steep ($s=0.2$) to nearly flat ($s=0.02$) exponential distributions and different multiples for the bimodal distributions, encompassing a range of biologically plausible clone-size distributions, with and without noise. We investigated three different modes of noise: (i) noise added to each count n with mean of zero and standard deviation $1.22\cdot\sqrt{n}$, (ii) a small baseline amount of noise added to all clone sizes and (iii) sporadic noise at random clone sizes (reminiscent of PCR jackpot effects). For completeness, we tested on both Macintosh (2.7 GHz Intel Core i5 running OS X 10.11.1) and Linux (2.3–2.8 GHz Intel Xeons running RHEL CentOS 6.6) platforms. NP and WL fits that were still incomplete after 100 h were terminated.

Error bars. Error bars define the range of overall diversity values that, given inevitable sampling error and any error in reconstructing parent distributions from samples of a given size, are consistent with Recon's estimate. We determined error bars for each diversity measure (species richness, entropy and so on) as follows (Fig. 4). First, we generated a wide range of exponentially and multimodally distributed *in silico* parent populations with known diversities of 3×10^2–1×10^7 species. Next, we took samples of these known distributions at systematically increasing coverage/sample sizes from 0.01x to 10x and, for each sample size, ran Recon to estimate the overall diversity, running on 1,716 samples in all (Fig. 4a). Five outliers (0.3%) were removed, leaving 1,711. For each overall diversity and coverage, the error was defined as the difference between the (true) overall diversity and Recon's overall diversity estimate. Given a test sample, the coverage, and Recon's estimate, one can then look up or interpolate from these errors the largest and smallest diversity values that are consistent with the estimate (Fig. 4b,c). These upper and lower bounds define the desired error bars on Recon's estimate.

We established these error bars as 95% confidence intervals using Monte Carlo cross-validation. Briefly, we randomly partitioned the above 1,711 samples 70–30 into reference and validation sets 100 times, each time using the reference set to calculate error bars for the samples in the validation set and counting how often error bars bracketed true diversity. These raw error bars bracketed true diversity in $93.6\pm1.3\%$ of cases; adjusting them by raising the upper bar by 1.6% brought this figure to the desired conventional level for confidence intervals, 95% ($96.2\pm1.0\%$). Note that error bars bracketed true diversities despite the formal possibility of there being clones in the parent population too small to observe in the sample (see the 'Minimum detected clone sizes and upper bounds (U)' below), meaning in practice this was not an issue. The above procedure can be generalized to incorporate arbitrary models of experimental noise.

Experimental data sets. We found and downloaded six publically available data sets. 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[31]. Following that study's methods, we clustered reads with $\geq$95% heavy-chain complementarity-determining region 3 (CDR3) nucleotide identity (the study treated clusters as clones). The other two data sets 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[37]. 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)[38–43].

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, $m_{\min}N/S$, where N is the total number of cells in the overall repertoire and S is the number of cells in the sample (the sample size). 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 Methods). We calculated these quantities for our validation and experimental data.

Data availability. Data referenced in this study are available at http://www.nature.com/nbt/journal/v31/n2/full/nbt.2492.html#supplementary-information and http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=downloadSuppFile&path%5B%5D=469&path%5B%5D=852. Recon is available subject to license agreement at http://arnaoutlab.github.io/Recon. Other data supporting the findings of this study are available from the corresponding author upon request.

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. Its Appl* **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. Methods* **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. Cody, M. L. & Diamond, J. M.) Ch. 4 (Harvard Univ., 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. Immunother.* **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. Good, I. J. & Toulmin, G. H. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43,** 45–63 (1956).
19. 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).
20. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114,** 4099–4107 (2009).
21. 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).
22. Klarenbeek, P. L. *et al.* Human T-cell memory consists mainly of unexpanded clones. *Immunol. Lett.* **133,** 42–48 (2010).
23. Chao, A. & Lee, S. M. Estimating the number of classes via sample coverage. *J. Am. Statist. Assoc.* **87,** 210–217 (1992).
24. Chao, A. Nonparametric-estimation of the number of classes in a population. *Scand. J. Stat.* **11,** 265–270 (1984).
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. Wang, J. P. Z. & Lindsay, B. G. A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Stat. Assoc.* **100,** 942–959 (2005).
28. DeWitt, W. *et al.* Replicate immunosequencing as a robust probe of B cell repertoire diversity. *arXiv* 1410.0350v1 (2014).
29. Link, W. A. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59,** 1123–1130 (2003).
30. McLachlan, G. J. & Krishnan, T. *The EM Algorithm and Extensions*. 2nd edn (Wiley-Interscience, 2008).
31. DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnol.* **31,** 166–169 (2013).
32. Wiegel, F. W. & Perelson, A. S. Some scaling principles for the immune system. *Immunol. Cell Biol.* **82,** 127–131 (2004).
33. 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).
34. Bohning, D. & Schon, D. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. R. Stat. Soc. Ser. C Appl. Stat.* **54,** 721–737 (2005).
35. Armitage, P. & Colton, T. *Encyclopedia of Biostatistics*. 2nd edn (John Wiley, 2005).
36. Leinster, T. & Cobbold, C. A. Measuring diversity: the importance of species similarity. *Ecology* **93,** 477–489 (2012).
37. 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).
38. Perez-Andres, M. *et al.* Human peripheral blood B-cell compartments: a crossroad in B-cell traffic. *Cytometry B Clin. Cytom.* **78**(Suppl 1): S47–S60 (2010).
39. Lavinder, J. J. *et al.* Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc. Natl Acad. Sci. USA* **111,** 2259–2264 (2014).
40. Rajkumar, S. V. *et al.* International myeloma working group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **15,** e538–e548 (2014).
41. Hindorf, C. *et al.* EANM dosimetry committee guidelines for bone marrow and whole-body dosimetry. *Eur. J. Nucl. Med. Mol. Imaging* **37,** 1238–1250 (2010).
42. Galotto, M. *et al.* Stromal damage as consequence of high-dose chemo/radiotherapy in bone marrow transplant recipients. *Exp. Hematol.* **27,** 1460–1466 (1999).
43. 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).

Acknowledgements

We thank the reviewers, William A. Link for correspondence, and Rima Arnaout for critical reading of the manuscript. This work was supported by the National Institutes of Health (NIH) National Institute of Allergy and Infectious Disease (NIAID) grant K08AI114958-01 (RA), Ruth Moorman and Sheldon Simon (R.A.) and American Heart Association (AHA) grant 15GPSPG23830004 (R.A.).

Author contributions

R.A. and J.K. conceived of and carried out the work and wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Kaplinsky, J. & Arnaout, R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat. Commun.* 7:11881 doi: 10.1038/ncomms11881 (2016).