Immunological Diversity with Similarity

Rohit Arora, PhD[1], Harry M. Burke[2], and Ramy Arnaout, MD, DPhil[1,3]*

[1]Division of Clinical Pathology, Department of Pathology and [2]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215. [3]Harvard Medical School, 25 Shattuck St, Boston, MA 02115

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

Summary

A diverse immune repertoire is considered a hallmark of good health. However, other things being equal, current methods for measuring repertoire diversity do not distinguish between a repertoire that is composed of similar sequences, clonotypes, or clones and a repertoire that is composed of different ones, even though the latter is intuitively more diverse. Here we describe a framework for incorporating similarity into diversity measures, and illustrate using … define diversity with binding similarity as functional diversity, and measure functional diversity on 391 large-scale antibody and T-cell receptor (TCR) repertoires. We find that while repertoires often contain millions of unique sequences, functional diversity reveals a landscape defined by at most a few thousand unrelated CDR3 binding targets. Naïve/IgM repertoires have more unique sequences than memory/IgG, but memory/IgG repertoires are more functionally diverse. Functional diversity is sensitive to vaccination, infection, and aging, and unlike raw diversity is robust to sampling error. Finally, according to functional diversity, repertoires from different people overlap significantly, suggesting a definable ceiling for the functional diversity of humanity. Similarity redefines diversity in complex systems.

Introduction

Immune repertoires are famously diverse. Collectively, a person's ~10^{12} B and T cells express many millions of unique recombined antibody and TCR genes as part of millions of clonal lineages, more unique sequence than in the entire germline genome (Jiang et al., 2013). At the sequence level, the repertoires of any two people overlap by only a fraction of a percent, indicating still higher diversity in the population (Arnaout et al., 2011; Robins et al., 2010). Yet repertoires are formed from V, D, and J gene segments that almost all people share and that are expressed at similar frequencies across individuals targets (de Bourcy et al., 2017; DeWitt et al., 2016), and repertoires are shaped by similar antigenic exposures and a consequent need to recognize and bind similar. Squaring the diversity that is seen with the similarity that must exist is a major goal in immunology.

This goal has relevance for disease stratification and clinical management across a range of conditions. B- and T-cell diversity fall with age, as specific exposures expand a few select lineages at the expense of others (Messaoudi et al., 2004). Chronic infection appears to have a similar effect, impairing vaccination (Jiang et al., 2013). Low B-cell diversity is associated with physiological frailty, a syndrome seen alongside conditions that are traditionally considered to be unrelated to adaptive immunity (e.g., atherosclerotic cardiovascular disease), independent of chronological age (Gibson et al., 2009). In cancer, a rise in sequence-level T-cell diversity is thought to predict a successful response to immune-checkpoint inhibitors, drugs that make tumors more visible to the immune system (Hopkins et al., 2018).

Traditionally, diversity has been measured as a simple count of the number of different sequences, lineages, or clones in a sample, a measure known formally as species richness. However, species richness ignores a key feature of repertoire diversity, species frequency: the fact that some sequences are common and others rare. In an intuitive sense, a repertoire with a single dominant (e.g., leukemic) clone is less diverse than a repertoire that has the same number of clones but no dominant clone. To incorporate frequency into measurements of diversity, there exist a family of measures that includes Shannon entropy, the Simpson index (and related Gini coefficient), and the Berger-Parker index (Hill, 1973). These differ from each other in how much weight they place on frequency: i.e., how much more a large clone adds to the total diversity than a small one. Mathematically, weight can be represented as a parameter, q, in the so-called Hill framework, a master equation for diversity in which species richness, Shannon entropy, the Simpson index, and the Berger-Parker index, among others, have been

shown to correspond to different values of q (q=0, 1, 2, and ∞, respectively). It is understood that no single diversity measure is best: the different measures provide complementary information about a given complex system (Morris et al., 2014). Robust methods exist for correcting sampling error for species richness and the frequency-weighted measures, and these methods are becoming standard for measuring immunological diversity (Greiff et al., 2015; Kaplinsky and Arnaout, 2016).

However, there is a second key feature of repertoire diversity that the frequency-weighted measures fail to capture: species similarity. A repertoire made up of all-different sequences is intuitively more diverse than a repertoire that has the same number of sequences, present at the same frequencies as in the first repertoire, but all drawn from the same lineage or clone. In the literature, this fact is sometimes addressed indirectly by grouping sequences together before measuring diversity, for example by clustering reads, collapsing clones, or binning by V(D)J segment usage (DeWitt et al., 2016; Jiang et al., 2013; Ju et al., 2018; Kaplinsky et al., 2014; Vollmers et al., 2013). However, grouping usually imposes a binary threshold—in or out—on what is by nature a continuous and overlapping relationship among sequences and their encoded proteins. Grouping also usually zeros out or ignores any diversity that might exist within groups. It is unclear what is lost by ignoring similarity, or what might be gained from a more complete synthesis of diversity with similarity. This is true not only for the immunome, but for other complex systems such as microbiomes, images, and tissues. Here we sought to develop and explore a continuous framework for measuring diversity-with-similarity on B- and T-cell repertoires.

Results

Framework. We measured diversity-with-similarity on high-throughput B- and T-cell repertoires using a robust mathematical framework initially proposed for studying diversity in ecology and environmental settings (Leinster and Cobbold, 2012). This framework provides “with-similarity” counterparts for species richness and the frequency-weighted diversity measures: species richness with similarity (${}^{0}D_{s}$, which places a very small weight on frequency, and ${}^{\varnothing}D_{s}$, which like ${}^{0}D$ ignores frequency altogether), the exponential form of entropy with similarity (${}^{1}D_{s}$, henceforth simply "entropy with similarity," and likewise for other named indices), and so on. In ${}^{q}D_{s}$ notation, q is the frequency-weighting parameter, D_s denotes diversity-with-similarity, and D without the subscript means diversity without similarity, which we refer to as “raw diversity.”

Mathematically, the key innovation of diversity-with-similarity relative to raw diversity is inclusion of a similarity matrix whose entries quantify how similar each pair of species (sequences, clonotypes, etc.) is. Constructing the similarity matrix necessitates a choice of similarity measure. (Note the difference between similarity measures and diversity measures: a similarity measure is used to build the similarity matrix, which then is used to calculate diversity measures.) The choice of similarity measure depends on the biological feature(s) of interest. For example to study somatic hypermutation, one might use the Hamming distance. Just as different weightings provide complementary information about rare vs. frequent species—for example, the number of new thymic emigrants (species richness; ${}^{0}D$ or ${}^{\varnothing}D_s$) vs. large leukemic clones (Berger-Parker index; ${}^{\infty}D$)—different similarity measures are expected to reveal different systems-level features of repertoires' sequence-level configuration. Also as with raw diversity measures, expressing results for the new diversity-with-similarity measures as effective numbers, also known as number equivalents, (Macarthur, 1965; Hill, 1973; Jost, 2007; Marion et al., 2015), as opposed to as bits or nats (for ${}^{1}D_s$) or as various fractions (for ${}^{>1}D_s$), makes it possible to compare them to each other, regardless of weighting or similarity measure, on a single intuitive scale (Box 1).

<u>Similarity measure</u>. We were interested in the single most fundamental mechanistic feature of antibodies and TCRs: binding to specific targets (Fig. 1). Therefore for our similarity measure, we used a proxy for binding affinity that follows from the empirically observed changes in dissociation constant (K_d) associated with amino-acid substitution in antibody and TCR CDRs (Jankauskaite et al., 2018). We found that on average, a single amino-acid substitution at an antibody-antigen or TCR-peptide binding surface lowers affinity by 4-5 fold (geometric mean), with a long tail corresponding to rare orders-of-magnitude effects (Fig. 2a). We focused on CDR3, the third complementarity determining region, of IgH and TCRβ, since this is the single most important contributor to binding specificity (Xu and Davis, 2000); however, our approach can be applied to other regions. Because the relationship between sequence and specificity remains non-predictive and therefore complex, for any given sequence pair the similarity imputed from the observed distribution will be approximate; however, averaged over the many millions of pairs in each repertoire, it was expected to be a reasonably accurate first-pass repertoire-level view of immunological diversity with binding similarity.

Using this similarity measure, diversity-with-similarity is interpreted as the effective number of sequences in a repertoire if the sequences were equally common and had no binding overlap with each other (Box 1), or equivalently, the number of equally common non-overlapping binding

118 targets that a repertoire can recognize. We therefore refer to this version of diversity-with-
119 similarity as "functional diversity" (Fig. 1). Functional diversity can be interpreted in the context
120 of a "shape space" (Perelson and Oster, 1979) that contains all possible CDR3 binding targets,
121 with nearby targets having similar three-dimensional shapes and conformations (Fig. 1a). Each
122 CDR3 binds a (possibly overlapping) subset of targets; together, a repertoire's CDR3s cover
123 some part of shape space (Fig. 1b). Functional diversity measures the size of this region,
124 controlling for similarity and overlap in binding among different CDR3s (Fig. 1c).

125 **Validity.** We first established that our similarity measure behaved sensibly, with closely related
126 sequences scoring high and unrelated sequences scoring low (Fig. 2b). We next established
127 that it resulted in intuitive values for functional similarity by testing against expectations on
128 simple *in silico* repertoires. In a representative test, we constructed four repertoires with 34
129 unique sequences each and 752 sequences total (Fig. 2c-d). In each repertoire, a few
130 sequences were common (larger circles) while most were rare (smaller circles), representing
131 the long-tailed frequency distribution seen in real repertoires (Arnaout et al., 2011; Weinstein et
132 al., 2009). Importantly, the species-frequency distribution for all four repertoires was identical,
133 meaning that raw diversity was also identical across the repertoires, for all frequency
134 weightings. The only difference between the repertoires was in the pairwise similarity among
135 sequences.

136 For the first repertoire (Fig. 2d, top row), we chose closely related sequences from a single real-
137 world CDR3 clone. We expected that species richness with similarity—functional species
138 richness—would be close to 1. (We used ${}^{0}D_s$ here; ${}^{\varnothing}D_s$ performed similarly.) We observed a
139 value of 1.5; the extra 0.5 reflected sequence diversity within the clone. For the second
140 repertoire (Fig. 2d, second row), we swapped out half the unique CDR3s with CDR3s from a
141 different, unrelated real-world clone. As expected, we observed a rough doubling of functional
142 diversity, to 2.4. For the third repertoire (Fig. 2d, third row), we replaced all the sequences with
143 34 randomly chosen real-world CDR3s. We expected a functional diversity that was much
144 higher than in the first two repertoires but less than 34 because of the inherent sequence
145 similarities that make a CDR3 a CDR3, and, consistent with this expectation, observed a value
146 of 22. For the final repertoire (Fig 2d, bottom row), we replaced the CDR3s with random amino-
147 acid sequences (controlling for length), expecting a functional similarity of nearly 34, and this
148 again was observed (${}^{0}D_s$=32). In contrast to these differences in functional diversity, raw
149 diversity was indistinguishably 34 for all four repertoires. In every example, functional diversity fit
150 an intuitive sense of what diversity should mean (Fig. 2c), while raw diversity failed to detect a

difference. These results support the validity of our functional-diversity framework for immune repertoires.

Robustness. Sampling error—the "missing-species" problem (Bunge and Fitzpatrick, 1993(Bunge and Fitzpatrick, 1993))—is known to be a major potential confounder when measuring raw diversity, necessitating large (e.g.) blood volumes and/or post-hoc statistical correction for measurements on the sample to reflect repertoire diversity in the individual as a whole (Kaplinsky and Arnaout, 2016). A practical feature of functional similarity is that the values are smaller than those of raw diversity (reflecting clustering of similar sequences; Fig. 1-2). The effective coverage is therefore greater, meaning that less information about the functional diversity of an individual's overall repertoire is lost upon sampling than is the case for raw diversity. This observation suggested that functional diversity is more robust to sampling error, possibly even making it accurate enough to use without statistical correction, and thus useful for the sample sizes typically available for sequencing (10,000-1 million cells).

To test this possibility, we systematically downsampled from a representative TCRβ repertoire and two representative IgH (IgG) repertoires, one prepared from mRNA and one from genomic DNA, each with ~10^6 unique sequences, and compared raw vs. functional diversity on the subsamples to those of the full sample (Fig. 3). (We wished to consider possibility of lower diversity from mRNA than DNA, since transcriptionally less active cells may be less likely to be sampled.) For TCRβ, we found that functional species richness saturated at a sample size of ~30,000 sequences and functional entropy at ~10,000 sequences (Fig. 3b, first and third columns). Functional diversity for higher frequency weightings (*q*) saturated with even fewer sequences. For IgH from mRNA, functional species richness did not saturate but did plateau, with a final increase of ≤2 percent per order of magnitude. Assuming that each unique sequence corresponds to a cell and 10^{10} B cells in the body, this final measured rate of increase means that the individual's total functional species richness is no more than 50 percent higher than the value measured on the sample (Fig. 3, middle row). This is the maximum expected sampling error. For IgH from DNA, functional species richness had begun to plateau at the full sample size, resulting in the value for the individual being no more than three times as much as in the sample (maximum three-fold error). Meanwhile, functional entropy saturated at 30,000 cells for IgH from mRNA and 300,000 cells for IgH from DNA. This behavior was in marked contrast to that of raw diversity, which did not saturate or plateau for species richness (Fig. 3, white symbols), consistent with previous reports and illustrating the need for statistical correction (Kaplinsky and Arnaout, 2016).

We then asked whether the robustness of functional diversity can be expected to generalize for any IgH or TCR repertoire. We reasoned that a “meta-repertoire” comprising sequences drawn uniformly (i.e., without regard to frequency) from many individuals will be more diverse, by any measure, than any single repertoire (which will have fewer sequences, and in which the same sequence may appear multiple times). Downsampling from a meta-repertoire therefore provides an upper bound or worst-case scenario for the sampling requirements for any single repertoire. To build meta-repertoires, we pooled and then uniqued CDR3s from 114 different IgH repertoires from 79 individuals including Americans of African, European, and Hispanic descent (Bild et al., 2002; DeWitt et al., 2016; Vollmers et al., 2013) to build an IgH meta-repertoire of roughly 36 million unique sequences—as many as or more than ever observed or currently estimated to be in a typical individual—and similarly for CDR3s from TCRβ repertoires from 69 healthy individuals (of mostly European but some Asian descent) (Emerson et al., 2017) to build a TCRβ meta-repertoire of 10 million unique sequences, and downsampled from each of these meta-repertoires as above (Fig. 3, large circles). We found that functional diversity plateaued for all q, saturated for $q \geq 1$ and reflected overall diversity to within a few percent from sample sizes of 50,000 for TCRβ and IgH RNA and 100,000 for IgH DNA for $q=0$, and 30,000 for TCRβ and IgH RNA and 300,000 for IgH DNA for $q \geq 1$ (Fig. 3, large colored circles). Together, these results confirmed that functional diversity measured on samples is an accurate measure of overall functional diversity in the individual, at conventional sample sizes.

Raw and functional diversity. We measured raw and functional diversity on 141 healthy human subjects (Fig. 4). For IgH, we found a (geometric) mean functional species richness ($^{\square}D_s$) of 677 (range, 487-916) from mRNA and 2,205 (range, 2,042-2,485) from DNA, suggesting that on average, the human antibody repertoire is capable of recognizing the equivalent of no more than a few thousand unique non-overlapping heavy-chain CDR3 binding targets. (As above, lower diversity from mRNA was not unexpected, since inactive cells, which produce less IgH mRNA than active cells, may be underrepresented.) For TCRβ the mean functional diversity was 140 targets (range, 115-167). Functional diversity can be thought of as clustering similar sequences together, although functional clusters can overlap and sequences can belong to multiple clusters. An indication of the average size of these clusters can be obtained by taking the ratio of raw to functional diversity measures. For species richness, we found that IgH typically had hundreds of sequences per cluster, while TCRβ had thousands. Thus by both functional diversity and average functional-cluster size, IgH CDR3 repertoires are roughly 5-10 times as diverse as TCRβ (for small q). Repertoires with higher raw diversity might

be expected to be more functionally diverse, but we found no consistent trend across all repertoire types. Thus, functional diversity generally complements raw diversity, adding information not captured by raw diversity alone.

Naïve vs. memory B cells. We next sought to investigate what this information might add to our understanding of adaptive immunity. We began with two widely studied B-cell subsets, naïve (IgM) and memory cells (predominantly IgG). Previous studies have shown that naïve repertoires have higher raw diversity than memory repertoires (DeWitt et al., 2016). This is at least superficially consistent with the fact that only a subset of naïve cells are selected to enter into the memory compartment. However, in a functional sense there is a case to be made that memory/IgG repertoires should be more diverse, since somatic hypermutation differentiates memory cells from naïve cells, and indeed from each other. Using well-characterized publicly available repertoires from DNA from three healthy human subjects, we confirmed that by raw species richness, naïve ($CD27^-IgM^+$) B-cell repertoires are ~10 times as diverse as memory ($CD27^+IgM^-$) repertoires (DeWitt et al., 2016) (Fig. 5a-b). Yet by functional species richness, we found that memory repertoires were at least as diverse as naïve (Figs. 5a-b). Comparing raw and functional diversity for 34 IgM and 32 IgG repertoires from mRNA (repertoires with less than 100,000 total sequences were discarded) from 28 additional healthy individuals from a separate dataset showed a similar pattern as for the three DNA repertoires: in all but a few outliers, IgM had higher raw diversity but IgG had higher functional diversity (Fig. 5c-d). For raw diversity, the IgM:IgG ratio rose from ~3:1 at q=0 to peak at 10:1 around q=1, due to a large fraction of rare IgG sequences (Fig. 5d). This effect was more pronounced for naïve:memory (Fig. 5b). For functional diversity, the absence of a peak in the IgM:IgG ratio suggests that these many rare sequences must nonetheless be similar to others in the repertoire, possibly because they are members of clones (Fig. 5b,d).

Cytomegalovirus (CMV) exposure. CMV is a herpesvirus to which half of the adult population has been exposed and results in life-threatening opportunistic infections in newborns, transplant recipients, and immunocompromised individuals (Emery, 2001). In most healthy individuals, it causes a chronic infection marked by clonal expansion of both B cells and T cells and a consequent fall in raw diversity, an effect also seen during aging (see below) (de Bourcy et al., 2017; Qi et al., 2014). We measured raw and functional TCRβ CDR3 diversity for 120 individuals: 69 CMV-seronegative and 51 CMV-seropositive subjects aged 19-35 (Emerson et al., 2017), the narrow age range helping control for any age-related effects. There was a clear trend toward lower diversity in the CMV-seropositive group relative to the CMV-seronegative

group by both raw and functional diversity, for all weighting parameters (Fig. 6a). Combining raw with functional diversity facilitated identification of two subgroups among the subjects with known CMV status: subjects with a high raw Berger-Parker Index (${}^{\infty}D$) were almost always CMV-seronegative (Fig. 6b), whereas subjects with low functional Berger-Parker Index (${}^{\infty}D_s$) were almost always CMV-seropositive (Fig. 6c). The reverse—low ${}^{\infty}D$ or high ${}^{\infty}D_s$—did not distinguish between the groups. Using both measures gave a better indication of CMV status than did either one alone (Fig. 6d). The conclusion is that CMV is unlikely in the absence of large clones/expanded lineages, as has been reported, but is likely only if the large clones/expanded lineages that are present exhibit high similarity to other clones/lineages in the repertoire, or else are indeed very large (Fig. 6e). Again, the addition of functional diversity offers insight that raw diversity alone does not.

Flu vaccination. Vaccination with a seasonal trivalent influenza vaccine (TIV) triggers clonal expansion in B cells. Previous work on five vaccinees showed likely flu-specific memory IgG lineages emerging by day 7 post-vaccination (Vollmers et al., 2013). We found that combining raw and functional diversity reveals a signature of clonal expansion and selection without the need for lineage tracking (Fig. 7). We measured raw and functional diversity for IgM and IgG at day 0 (pre-administration) and day 7 from all 14 vaccinees in Vollmers' dataset. We found that for most subjects, for IgG, raw species richness rose from day 0 to day 7 while functional species richness fell (Fig. 7a-b). This means that even as the number of sequences increased, many of the new sequences were similar to each other (or to existing sequences), and they tended to replace different-looking sequences. Meanwhile, there was no obvious pattern in IgM (Fig. 7c). Together, these results are what we would expect from clonal expansion and selection in a memory response, and thus represent a repertoire-level signature of these phenomena. Interestingly, in most cases, raw and functional entropy both fell (Fig. 7a-b, right panels). This suggests that most of the new sequences at day 7 were rare, while at the same time a subset of sequences and functional clusters grew. Thus overall, the addition of functional diversity reveals a key feature of clonal dynamics, which is not evident from raw diversity alone.

Aging. To explore the effect of age, we measured raw and functional diversity for TCRβ CDR3 repertoires from 41 healthy individuals aged 6-90 years old (Britanova et al., 2014) (Fig. 8). We found that raw diversity falls with age regardless of weighting parameter; a fall in raw species richness had been reported previously (Britanova et al., 2014). Functional diversity also fell, regardless of weighting parameter. However, for species richness, four septuagenarians bucked the trend (Fig. 8, arrows): even as their raw species richness was unremarkable relative to that

of other individuals of similar age, their functional species richness was similar to that of children. Only one of these four had a high likelihood of being CMV-negative; the probability that all four were CMV-negative was low. We therefore consider CMV unlikely as an explanation for their high functional species richness. Unlike their peers, these four appear to have retained functional diversity among their rarest sequences. (The alternative hypothesis is that these four saw a rise in functional species richness from a lower level earlier in life, but we think this unlikely given the overall downward trend across individuals.) We considered but excluded PCR/sequencing artifacts as the cause, as we expected such artifacts would have led to larger raw species richness, which was not observed. Thus, functional diversity identified for further study individuals who were unremarkable by raw diversity alone.

Discussion

Diversity both affects function, and reflects it. In the adaptive immune system, the defining tradeoff is breadth vs. depth: a repertoire must be sufficiently diverse to contain sequences that can recognize a given target and lead to useful clones, but not so diverse that cells that express such sequences are too rare to find the target on biologically relevant timescales (Schober et al., 2018; Zarnitsyna et al., 2013). To monitor immunological diversity, either diagnostically or therapeutically, we must be able to measure it, and to measure it, we must define it. It is increasingly recognized that a reasonable definition of immunological diversity must account for differences in species frequency. Here we argue such a definition must also account for species' pairwise similarity, and show that binding similarity, which leads to what we call functional diversity, provides useful insight into repertoire function.

Pairwise similarity can be seen as governed by a tunable parameter that helps define the similarity matrix, analogous to how q is a tunable parameter that governs the effect of differences in frequency (Chao et al.). In our study, the similarity matrix is defined by the average single amino-acid change in K_d, an average based on over 1,300 independent measurements, and the assumption of multiplicative independence. This source data is not systematic, but to our knowledge is the best available. While our study is to our knowledge the first of its kind, it follows a long tradition of attempts to estimate the number of binding targets that can be recognized by the adaptive immune system immunization (Bachmann et al., 1994; Obar et al., 2008). In these past studies, typically a sample of B or T cells was diluted until binding to/protection from a given target was abolished, using whatever thresholds the investigators deemed appropriate. If the limiting frequency for binding/protection was found to

315 be, for example, 1:3,000, the conclusion was that the repertoire could recognize 3,000 different
316 targets. This conclusion was based on the assumption that on average, all targets behave the
317 same as the one under study. Such studies gave functional diversities of 100 to 100,000 targets
318 for various T-cell populations and for B cells after antigen exposure, a wide range (Bachmann et
319 al., 1994; Obar et al., 2008). The width of this range may reflect real differences in the
320 frequencies of cells that are specific for different targets, or variability in stringency or
321 experimental setup. Interestingly, this "how-many-can-fit" logic seems not to have been used
322 when testing so-called natural antibodies, which bind many targets at low affinity (Frank, 2002;
323 Holodick et al., 2017; Notkins, 2004). For example, when one in five natural antibodies were
324 found to bind insulin (Chen et al., 1998), this was not taken to mean that the repertoire could
325 recognize only five targets, because of presumed overlapping specificity of these antibodies for
326 other targets. Meanwhile, theoretical studies have suggested a need for ≤10,000 binding
327 targets, and fewer for T cells than B cells, because of major histocompatibility complex (MHC)
328 restriction (Langman and Cohn, 1987; Zarnitsyna et al., 2013)

329 For raw diversity, we found that a typical repertoire contains on the order of 10 million unique
330 CDR3s, well above the upper end of previous estimate for the number of binding targets. These
331 findings are in line with other recent estimates that were likewise based on a combination of
332 deep sequencing and statistical correction (Britanova et al., 2014; DeWitt et al., 2016; Kaplinsky
333 and Arnaout, 2016). On average, these raw species richnesses mean that each of 100-100,000
334 putative binding targets can be bound by 100-100,000 unique CDR3s ($10^7/10^5$ to $10^7/10^2$). From
335 a medical perspective, such redundancy is good for treatment, because it supports the
336 prevailing view that there are many ways to design an antibody- or TCR-based drug that will
337 recognize a given target, but potentially a complicating factor for attempts to diagnose specific
338 diseases based on repertoire sequence, because it suggests that signatures of exposure to a
339 given target may be quite variable.

340 One of our key findings is that functional diversity is much lower than raw diversity: repertoires
341 contain only a few hundred functional clusters for TCRβ CDR3s and at most a few thousand for
342 IgH. The fact that functional diversity is based on K_ds suggests that functional diversity should
343 correlate with the number of structurally unique, non-overlapping target clusters that CDR3s can
344 recognize (Fig. 1). Yet our measurements of functional species richness lie at the low end of the
345 range of prior estimates. We propose two explanations. First, our measurements are limited to
346 CDR3s; variability in the rest of the antibody or TCR protein must add to the total number of
347 potential binding targets. This possibility is testable by extending our method to more or indeed

all of the antibody or TCR sequence. Second, functional diversity may be providing a less detailed description of shape space than limiting-dilution studies: i.e., functional diversity may be a coarse graining of the target-binding landscape (Fairlie-Clarke et al., 2009; Smith et al., 1997) (Fig. 9). A pair of sequences may be similar enough to lie near each other in shape space, but only one may bind a given target above a threshold level of specificity in a binding study. In short, binding studies may be counting peaks while functional diversity is counting mountains. If true, our results suggest that the landscape of TCRβ CDR3 binding is more clustered than that of IgH, such that there are on average several times as many functional IgH clusters as TCRβ. This prediction is testable, at least in principle, through large-scale systematic binding assays to measure K_d, or by measuring binding as a binary outcome at multiple stringency thresholds. Both explanations may contribute. To our knowledge ours is the first attempt a quantitative summary of this landscape using data from large-scale binding studies and high-throughput repertoire sequencing.

Why is CDR3 functional diversity higher for IgH than for TCRβ? We hypothesize that it is for the same three reasons that there is more sequence diversity for IgH than TCRβ. First, humans have 23 D_H gene segments vs. only 2 D_β segments, and D is the largest germline contributor to CDR3. V and J segments tend to directly contribute little more than the canonical starts and ends of CDR3s, and besides there are similar numbers of V-J combinations in IgH as TCRβ (49×6=294 and 48×13=624, respectively). Second is somatic hypermutation, which diversifies IgH but not TCRβ. And third, IgH CDR3s are longer than TCRβ CDR3s, allowing for a larger number of possible sequences. Further analysis will be needed to test these hypotheses.

We have shown how functional diversity complements raw diversity to offer insight into the difference between naïve and memory repertoires, to aid in identification of disease states, and to illustrate clonal selection and other repertoire dynamics. We hope these examples will encourage others to use and/or expand our framework to investigate repertoire dynamics in other conditions, in other subsets, in the other chains (TCRα and IgL), and in other model systems such as zebrafish (Weinstein et al., 2009) and mouse (Arnaout et al., 2011; Kaplinsky et al., 2014). We draw attention to the fascinating difference between the number of unique sequences, which ran into the millions in most of the repertoires we investigated, and the much smaller numbers of what we call functional clusters (the effective numbers of functional diversity). The result is a “functional degeneracy” among sequences that are organized into functional clusters. Characterizing these clusters is an interesting topic for future work.

Functional similarity also offers a new perspective on similarities and differences between people. To show that functional diversity is robust to sampling, we generated “meta-repertoires” by pooling sequences from scores and in some cases over a hundred people, including people from different ethnic backgrounds. Surprisingly, and somewhat unexpectedly given the low sequence overlap between pairs of individuals, the functional diversity of these meta-repertoires never exceeded the functional diversity of any given repertoire by more than a few fold; moreover, the functional diversity trended toward saturating in samples of just a million sequences (Fig. 3). Together, these findings predict that any two individuals share a majority of their functional clusters, in stark contrast to the vanishingly small fraction of sequences they share. Further, these findings suggest that the functional diversity of the entire population is only a few hundred clusters for TCRβ CDR3s and a few thousand for IgH, and imply that these clusters can be sampled exhaustively by sequencing fewer than 20 individuals. It will be fascinating to test this finding with additional ethnically and geographically diverse populations, to further examine our prediction that, contrary to conventional wisdom, the functional limits of the adaptive immune system are in a practical sense both finite and within reach.

The focus of this study was binding similarity, but we expect that the utility of the diversity-with-similarity framework will extend to other facets of immunology (e.g., somatic hypermutation) and to other fields, most readily metagenomics, sociology, oncology, and cellular cartography (Almendro et al., 2014; Heindl et al., 2016; Koopmans and Schaeffer, 2013; Li et al., 2012; Taraska, 2015). We hope this study will serve as a template for incorporating similarity into the study of other complex systems.

Acknowledgements

This research was supported by grants from the National Institutes of Health (NIAID K08AI11495801) and the American Heart Association (15GPSPG23830004), contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by grants UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 from NCATS, as well as support from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The authors thank the investigators, staff, and participants of the MESA study for their valuable contributions (a full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org), as

well as the Research Computing Group at the High Performance Computing Cluster at Harvard Medical School.

Online methods

High-throughput repertoires. We obtained 391 quantitative high-throughput IgH and TCRβ repertoires from 202 human subjects. These included IgH from naïve and memory B cells from DNA (*n*=3 individuals) (DeWitt et al., 2016); TCRβ chains from DNA from healthy subjects known to be serologically negative for cytomegalovirus (CMV) (*n*=69 individuals) (Emerson et al., 2017) and from healthy subjects whose CMV serostatus was unknown (*n*=41 individuals) (Britanova et al., 2014); pooled barcoded IgG and IgM heavy chains from mRNA from healthy subjects before and seven days after administration of one of two influenza vaccines (*n*=28 individuals) (Vollmers et al., 2013); quantitative pooled TCRβ chains from DNA for subjects who were otherwise healthy but serologically CMV positive (*n*=51 individuals) (Emerson et al., 2017); and IgH chains from DNA for subjects enrolled in the Multi-Ethnic study of Atherosclerosis (MESA; *n*=41 individuals) (Bild et al., 2002). CDR3 annotation was performed using our in-house pipeline as previously reported (Kaplinsky et al., 2014) and standard tools (e.g. IMGT). Details for obtaining these datasets are available from the primary publications referenced above.

Similarity measures. A functional measure of similarity between polypeptides is how well they bind the same target. We were interested in similarity as a function of the number of amino acid substitutions (i.e., as a function of edit distance). The effect of substitutions on binding is complex and depends on the position and identity of the specific amino acids involved; many substitutions may have little or no effect, while a few may abolish binding entirely (Lunzer et al., 2010). When comprehensive data are available, detailed statistical models can offer reasonable predictions of the effect of specific amino-acid substitutions (Hopf et al., 2017; Salinas and Ranganathan, 2018; Lee et al., 2008). However, this type of data does not yet exist across entire antibody and TCR repertoires, and so simpler models are required. These models are not expected to precisely predict the effects of specific substitutions, but should accurately reflect the effects of substitutions when averaged over many pairs of proteins, such as the millions of pairs in megacell-scale repertoires.

To develop a model for our similarity measure, *s*, we downloaded SKEMPI 2.0, which is to our knowledge the largest and best-curated database of experimentally measured effects of amino-

443 acid substitution on protein-protein binding (Jankauskaite et al., 2018). Each entry includes a
444 Protein Data Bank (PDB) identifier (Berman et al., 2000), the type of structural region (Levy,
445 2010) that contains the substitution(s), one or more PDB coordinates, and (in nearly all cases)
446 the dissociation constant (K_d) of each member of the pair (referred to in the database and Fig.
447 2a as "wild type" and "mutant"). We extracted entries for all single amino-acid substitutions for
448 which K_d for both wild type and mutant were recorded, and considered only entries that involved
449 binding between antibody and antigen (n=797) or TCR and peptide/MHC (n=531; total
450 n=1,328). Although amino-acid substitutions anywhere in a protein may affect binding,
451 substitutions at the core of the binding interface are more likely to affect binding than
452 substitutions elsewhere (Levy, 2010). Therefore we split the data into core (n=584) and non-
453 core (n=744) groups and analyzed the effect of substitution binding, measured as
454 $|\log_{10}(K_{d\mathrm{mut}}/K_{d\mathrm{wt}})|$, separately for each group.

455 As expected, the probability distributions for the two groups differed substantially from each
456 other (Mann-Whitney U p-value 2.0×10^{-33}). Substitution of a core residue had a 13-fold
457 (geometric) mean effect on binding, consistent with prior reports (Whittaker et al., 2001), while
458 substitution of a non-core residue had a 4-fold effect. Both probability distributions were long-
459 tailed, and were reasonably well described by exponential probability-density functions (i.e., of
460 the form ke^{-kx}). We confirmed that the distributions for antibody-antigen core residues (n=352)
461 and TCR-peptide/MHC core residues (n=232) were similar to each other, that the distributions
462 for antibody-antigen non-core residues (n=445) and TCR-peptide/MHC non-core residues
463 (n=299) were also similar to each other, that within each of the antibody-antigen and TCR-
464 peptide/MHC subgroups the distributions for core and non-core residues were different, and that
465 these results held separately for human and non-human (nearly all of which were mouse)
466 sequences (using the Structural Antibody Database (Dunbar et al., 2014) and the Structural
467 TCR Database (Leem et al., 2018) to assign species), all using Mann-Whitney U and visualized
468 as histograms. Using PyMol v2.2.0 (The PyMOL Molecular Graphics System, Version 2.0
469 Schrödinger, LLC), we next manually reviewed nine structures containing substitutions in
470 human IgH or TCRβ CDR3s (1BD2, 1OGA, 3BN9, 3QDJ, 3SE8, 3SE9, 4I77, 5C6T, 5E9D) and
471 found that to a good approximation, a constant fraction 0.15±0.05 of CDR3 amino acids consist
472 of core residues, with no obvious difference between chain types. To estimate the effect of a
473 single amino-acid substitution in a CDR3 in our datasets, we therefore combined core and non-
474 core distributions with a weighting of 0.15:0.85.

The resulting distribution was again long-tailed, with most substitutions having small effects and a few having effects of many orders of magnitude (Fig. 2a). There were small spikes in the tail for substitutions with ≳60-fold effects, i.e. $|\log_{10}(K_{d\mathrm{mut}}/K_{d\mathrm{wt}})| \gtrsim 1.8$. A review of sources cited by SKEMPI suggested that these spikes likely reflect ascertainment bias: selective experimentation on amino acids with unusually strong effects (e.g. Pons et al., 1999; Taylor et al., 1998). To counteract such bias, we therefore built a high-confidence dataset using 1.8 as the cutoff. Ascertainment bias in two- and three-amino-acid substitutions is expected to follow the square and cube of the bias in single-amino-acid substitutions, respectively, precluding rigorous conclusions from being drawn from independence testing. However, comparison with those groups was broadly consistent with either multiplicative ($s=c^m$, where s=similarity; c=the cost of binding, i.e., 1/(fold effect); m=edit distance) or additive ($s=c/m$ for $m\geq 1$) independence. Because additive effects result in higher pairwise similarities and therefore smaller repertoire diversities than multiplicative effects, the multiplicative-independence model is more conservative for studying the effects of similarity on diversity. We therefore chose the multiplicative model for further analysis.

To determine the similarity between two CDR3s with edit distance m, we sampled independently from the high-confidence dataset m times, and multiplied the costs together. We confirmed that on average, the results of this stochastic sampling were the same as deterministic calculation of $s=c^m$ with $c\approx 0.55$. We performed sensitivity analysis based on lower-confidence cutoffs (down to c=0.48) and alternative assumptions (up to c=0.60). This resulted in somewhat higher or lower diversity values, but qualitative patterns were robust to these perturbations.

Diversity measures. We calculated ${}^qD$ as previously described (Hill, 1973; Kaplinsky and Arnaout, 2016) and ${}^qD_s$ according to Leinster and Cobbold (Leinster and Cobbold, 2012). We corrected ${}^qD$ for sampling error using Recon (default settings) as previously described (Kaplinsky and Arnaout, 2016). We note Hill's framework (Hill, 1973) has inspired several methods for incorporating similarity into diversity measurements, each of which retains useful features of Hill's framework (Chao et al., 2018; Chiu and Chao, 2014; Leinster and Cobbold, 2012; Scheiner, 2012). Two of the new frameworks were introduced with explicit discussion of how to decompose population-level diversity into within- and between-group components (Leinster and Cobbold, 2012; Chiu and Chao, 2014). Each of these has advantages and disadvantages over the other (discussed in Botta-Dukát, 2018) . We chose Leinster and

Cobbold's framework here because we found it easier to apply and interpret. For readability, we made a minor change to the notation, from ${}^{q}D^{Z}$ to ${}^{q}D_{s}$.

Use of this framework raised two issues that we addressed. First, its q=0 measure, ${}^{0}D_{s}$, depends on frequency albeit to a very small extent, unlike the Hill framework's q=0 measure, ${}^{0}D$, which is species richness (and is independent of frequency). Therefore, as a more direct comparison to species richness, we calculated ${}^{0}D_{s}$ both with frequency information and without it (i.e., setting the frequencies of each of the n species to $1/n$). We refer to the latter as ${}^{\varnothing}D_{s}$ ("*D*-null"). Second, it has been shown that this framework can result in unreasonably low diversity values when most of the off-diagonal entries of the similarity matrix are far from zero, resulting in an insensitivity to q (Botta-Dukát, 2018; Chiu and Chao, 2014). We expected most of our off diagonals to be close to zero, since our similarity measures directly or indirectly involve exponential decay, which generates small values, but confirmed that most of the off diagonals in our similarity matrices were indeed close to zero by plotting histograms. Consequently, our measures were sensitive to q, as desired and expected.

Robustness analyses. For robustness analyses, IgH and TCRβ were analyzed separately. The upper-bound/worst-case scenario for IgH was evaluated by constructing a "meta-repertoire" by combining IgG sequences of subjects before vaccination (n=28 individuals; Vollmers et al. 2013), sequences from memory cells from healthy subjects from public database (n=3; DeWitt et al. 2016), and sequences from subjects enrolled in MESA study (n=41; Bild et al. 2002), and sampling from this meta-repertoire without regard to the frequency of sequences. We chose IgG/memory sequences where possible because those sets exhibited higher functional diversity than naïve sets, and we were interested in maximizing diversity. We ignored the frequency of sequences for the same reason: uniform frequency maximizes diversity, other things equal. For TCRβ, we constructed a meta-repertoire by combining sequences from CMV seronegative individuals (n=69; Emerson et al. 2017) and again sampling at uniform frequency. We chose CMV seronegative individuals for the same reason as we chose memory/IgG sequences above: seronegative individuals exhibited higher diversity. For both IgH and TCRβ, including all sequences lowered diversities slightly. The representative samples were from subject D3 for IgH (from DNA), subject SRR960344 for IgH (from mRNA), and subject Keck0070 for TCRβ (CMV seronegative). CDR3 sequences were sampled proportional to their frequency in the repertoire.

538 **Figure titles and legends**

539 **Figure 1: Functional diversity. (a)** Each dot represents a binding target (e.g. an epitope) with
540 a different shape. Nearby targets have similar shapes (inset). Targets form clusters of similarity.
541 **(b)** Each colored region represents the targets that can be bound by one of six unique CDR3
542 sequences in a representative repertoire; this repertoire has a raw species richness of 6.
543 Together, the colored regions cover the part of shape space that can be bound by the
544 repertoire. (The unbound region might include, e.g., self-antigens.) Note the substantial overlap
545 in binding targets for the orange, yellow, green, and blue CDR3s. This overlap reflects binding
546 similarity among these CDR3s. **(c)** Because of this similarity, the repertoire covers only the
547 region denoted by the four identical non-overlapping squares. The functional species richness of
548 this repertoire is therefore 4: this repertoire has the same species richness as a repertoire
549 comprising four CDR3s that have zero overlap in binding specificity.

550 **Figure 2: Validity. (a)** Single amino-acid mutations in antibody and TCR molecules have a
551 range of effects on affinity, as measured by change in dissociation constant, K_d (gray). This was
552 well fit by a simple exponential (black line), providing parameterization for the similarity metric.
553 **(b)** CDR3s with high sequence identity have high similarity, while different CDR3s have low
554 similarity. Shown are two clones, represented by red and white subnetworks, each composed of
555 17 unique CDR3 sequences drawn from clonotypes of two real IgH repertoires. Node size
556 corresponds to the frequency of each sequence; edges connect pairs of sequences that differ at
557 a single amino-acid position. **(c)-(d)** Functional similarity agrees with an intuitive sense of
558 repertoire diversity. Each of the four repertoires in (c) has the same number of unique
559 sequences, present at the same frequencies; as a result, they all have identical raw diversity
560 (for every value of q) despite their obvious quantitative and qualitative differences. In contrast,
561 functional diversity increases with the number of, and increasing difference between,
562 repertoires' constituent sequences. Node size denotes sequence frequency. Shades denote
563 different clonotypes. Comparing the third and fourth rows, note that even when two repertoires
564 have the same number and frequencies of unique sequences, the repertoire whose sequences
565 are more different from each other (random peptides) has the higher functional diversity.

566 **Figure 3: Robustness.** Results for raw and functional species richness (q=0; ${}^0D$ and ${}^0D_s$) and
567 raw and functional entropy (q=1; ${}^1D$ and ${}^1D_s$). Raw, white shapes; functional, colored shapes.
568 Large symbols give an upper bound/worst-case scenario based on sampling meta-repertoires;
569 small symbols give results for a representative sample from DNA (circles) and mRNA

570 (triangles). First row: sample diversity is plotted as the effective number of sequences. For *q*=0,
571 functional diversity plateaus for TCRβ and IgH RNA and trends toward a plateau for IgH DNA at
572 the tested sample sizes; all three plateau for *q*=1; raw diversity does not plateau for either *q*=0
573 or *q*=1. Second row: discovery rate is the probability that the next sampled sequence will add to
574 the diversity. For example, for the IgH DNA representative sample, for *q*=0 raw diversity, at a
575 sample size of 1 million sequences there is still a probability of ~0.5 (a 50 percent chance) that
576 the next sequence to be sampled will be one that has not yet been seen and will therefore add
577 to the diversity. Third row: maximum error is the maximum fraction by which the diversity in the
578 sample can underestimate the diversity in the individual from whom the sample was taken.
579 Horizontal dashed lines indicate the threshold for two-fold error. For example, for the worst-case
580 scenario for TCRβ, *q*=0 functional diversity measured on a sample of 10,000 sequences will be
581 no more than a two-fold underestimate of diversity in the individual as a whole; in other words,
582 the sample value will be at least 50 percent of the overall value.

583 **Figure 4: Diversity in individuals.** Raw (black lines; left vertical axis) and functional (colored
584 bars; right vertical axis) species richness (q=∅) for 179 CDR3 repertoires from healthy
585 individuals representing **(a)** IgH from mRNA (all isotypes: IgA, IgG, IgM, IgD, and IgE), **(b)** IgM
586 and **(c)** IgG from mRNA from the subjects in (a), **(d)** IgH from DNA (all isotypes), **(e)** naïve IgH
587 from DNA, **(f)** memory IgH from DNA, and **(g)** TCRβ from DNA. See Methods for references.
588 Matched pairs of symbols below the horizontal axis denote replicates. Note the difference in the
589 scale for functional diversity between IgH and TCRβ. Note also a general lack of correlation
590 between raw and functional species richness, except in (c).

591 **Figure 5: Naïve vs. memory. (a)** Diversity profiles for naïve (red) and memory (black)
592 compartments from three deeply sequenced individuals. A diversity profile is a way to show
593 diversity across a range frequency-weighting parameter values at once. By raw diversity (left),
594 the naïve compartment is more diverse across the range of weightings. By functional diversity
595 (right), this distinction disappears. In **(b)**, this disappearance is highlighted by plotting the ratio of
596 naïve:memory diversity for raw diversity (red) and functional diversity (black). According to
597 functional diversity, the naïve compartment is no more diverse, and indeed sometimes
598 somewhat less diverse, than the memory compartment. This reversal is even more prominent in
599 comparisons of repertoires from an additional 28 healthy subjects **(c,d)**.

600 **Figure 6: Infection. (a)** Diversity profiles showing effective number of species as a function of
601 weighting parameter *q* for diversity without similarity (left) and diversity with similarity (right)

showing a trend toward lower diversity in CMV-seropositive individuals (red) relative to CMV-seronegative individuals (black), especially for large *q*. **(b)** Raw Berger-Parker Index ($q=\infty$), which measures the largest clones, showing that high diversity—an absence of large clones—is rare in CMV-seropositive individuals. **(c)** Functional Berger-Parker Index, showing that low diversity—the presence of large clones with similarity to other clones in the repertoire—is rare in CMV-seronegative individuals. **(d)** Combining raw and functional Berger-Parker Indices (first principal component of PCA, which explains 72 percent of variance) illustrates both of the trends in (c): for the third of subjects beyond the cutoffs indicated by the horizontal dashed lines CMV serological status is assigned with an accuracy of 95 percent. **(e)** Schematic representation of the three classes revealed by combining diversity with and without similarity. Each circle is a clone; each collection of clones is a representative repertoire. Top: subjects without large clones are almost always CMV seronegative. Bottom: subjects with large clones that are similar to other clones in the sample (shown in red) are almost always CMV seropositive. Middle: repertoires with large clones that are not similar to other clones in the repertoire may be either CMV seropositive or CMV seronegative. Receiver-operator characteristic (ROC) analysis gave an area under the curve (AUC) of 0.79.

Figure 7: Vaccination. Raw and functional diversity together reveal clonal expansion and selection without needing lineage analysis. **(a)** In the IgG compartment, raw species richness rises while functional species richness falls in most vaccinees (left). Meanwhile raw and functional entropy both fall (right). The difference vs. species richness suggests most new sequences at day 7 are rare. (b) Meanwhile, the IgM compartment changes less by these measures.

Figure 8: Aging. Raw and functional species richness ($q=\varnothing$) for TCRβ CDR3 repertoires from 41 healthy individuals. Arrows denote four septuagenarians who bucked the trend of lower functional species richness with age. Note that for each individual, the raw species richness is ~10-fold higher than previously reported (Britanova 2014), likely because the method we used to correct for missing species (Recon) is more sensitive than the method used in the previous report (Fisher).

Figure 9: Binding landscape. (a) Schematic of the target-binding landscape. The gray distribution represents CDR3 sequences that bind a given target. Sequences are ordered by their similarity to each other. (In reality, similarity is a multidimensional property that makes it impossible to order sequences in a single dimension as shown here; this is done for illustrative

purposes only.) The height at each sequence denotes the affinity with which it binds the target (vertical axis; measured e.g. by K_d). Many more sequences bind the target at low affinity than at high affinity, resulting in a "mountain-and-peak" appearance. This schematic is useful for interpreting functional diversity as described in this study, and the raw diversity estimates based on previous binding studies, as described in the Discussion. (Note that in this schematic, the raw diversity as measured simply corresponds to the total number of sequences along, i.e. the width of, the horizontal axis.) At high affinity, very few sequences bind a given target. At medium affinity, more sequences bind, and can be binned into two small clusters, represented by the small circles. At low affinity, many sequences bind, and can be binned into a single large cluster, represented by the large circle. In **(b)**-**(d)**, many targets are shown. Each color corresponds to a different target; nearby targets are structurally similar. As in (a), each colored area denotes the sequences that bind a given target, as a function of binding affinity (vertical axis). Experiments usually detect the highest-affinity sequences: the peaks of the landscape (above the horizontal dotted line). The narrower the peak when it crosses the experimental threshold, the rarer specific sequences are, and the larger the number of targets that the repertoire will be estimated to bind. (For example, if 100,000 sequences are shown across the horizontal axis in each plot, and only one crosses the experimental threshold for a given target, the frequency of sequences specific for that target is 1:100,000, and the conclusion will be that there must be 100,000 such targets that the repertoire can bind. If 100 cross the experimental threshold, the conclusion will be that the repertoire can bind only 1,000 targets.) Functional diversity measures the overall contours of the landscape. Conceptually, this can be thought of as measuring the number and size of the "mountains" at a lower affinity threshold (horizontal solid lines). The differences in functional diversity between (b) memory IgH, (c) TCRβ, and (d) naïve IgH correspond to different landscapes. The raw species richness of memory IgH and TCRβ are comparable, represented here by the same width of all the plots. In addition, a similar number of sequences per target cross the experimental threshold, so estimates of the total number of targets that the repertoires can bind will also be comparable. However, less low-affinity overlap between the targets of the IgH sequences in (b) gives it higher functional species richness than the TCRβ repertoire in (c): here, six functional clusters (white circles) vs. three. (The sizes of the clusters are related to frequency-weighted functional diversity measures, i.e. larger q.) The sequences in the naïve IgH repertoire in (d) have only low affinity for the six colored targets, and many recognize more than one target (overlapping colored areas). Note the lower experimental threshold (horizontal dotted line), consistent with the ~10% or more of antibodies that recognize a target and the high degree of cross-reactivity in studies of natural

antibodies (Frank, 2002; Holodick et al., 2017). The functional diversity threshold controlled by the average effect on K_d of a single-amino-acid change in CDR3. If the effect were larger—or if it were amplified by e.g. raising it to a power when building the similarity matrix—the threshold would be higher, and vice versa.

672 **Tables and Figures**

Box 1: Interpreting effective numbers

Consider two repertoires of 100 clones each. In the first repertoire, one clone is large and accounts for 91 percent of all cells (e.g. a leukemic clone); the other 99 clones are small and account for the remaining 9 percent. In the second repertoire, all 100 clones are equally common, each accounting for 1 percent of cells. The Shannon entropies of the two repertoires are 1.0 bit and 6.6 bits. Entropy is converted to an effective number—${}^{1}D$—by exponentiation: the effective number of clones in the first repertoire is $2^{1.0}=2$, while in the second repertoire it is $2^{6.6}=100$. Thus per entropy, the first repertoire can be thought of as "effectively" consisting of just two clones: the 99 rare clones collectively count the same as the one large clone. In other words, the first repertoire has the same effective diversity as a repertoire that consists of just two clones that are equally common. The second repertoire already consists of clones that are equally common, so the effective number of clones in this repertoire, $2^{6.6}=100$, is the same as its species richness. Diversity with similarity is interpreted in a similar fashion: a repertoire with a ${}^{q}D_{s}$ of n species has the same effective diversity as a repertoire with n species that are equally common (as above), with the additional constraint that these species are now also completely unrelated to/ dissimilar from each other.

Figure 1

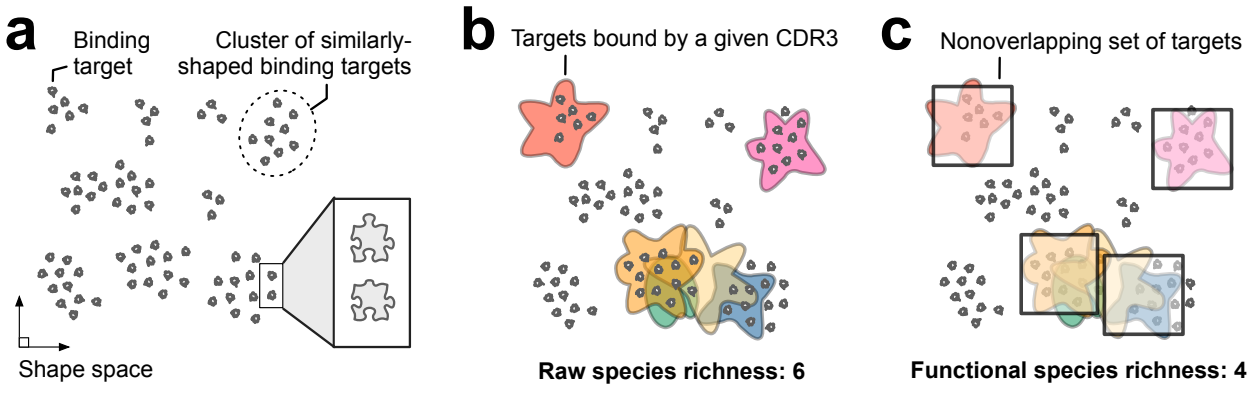
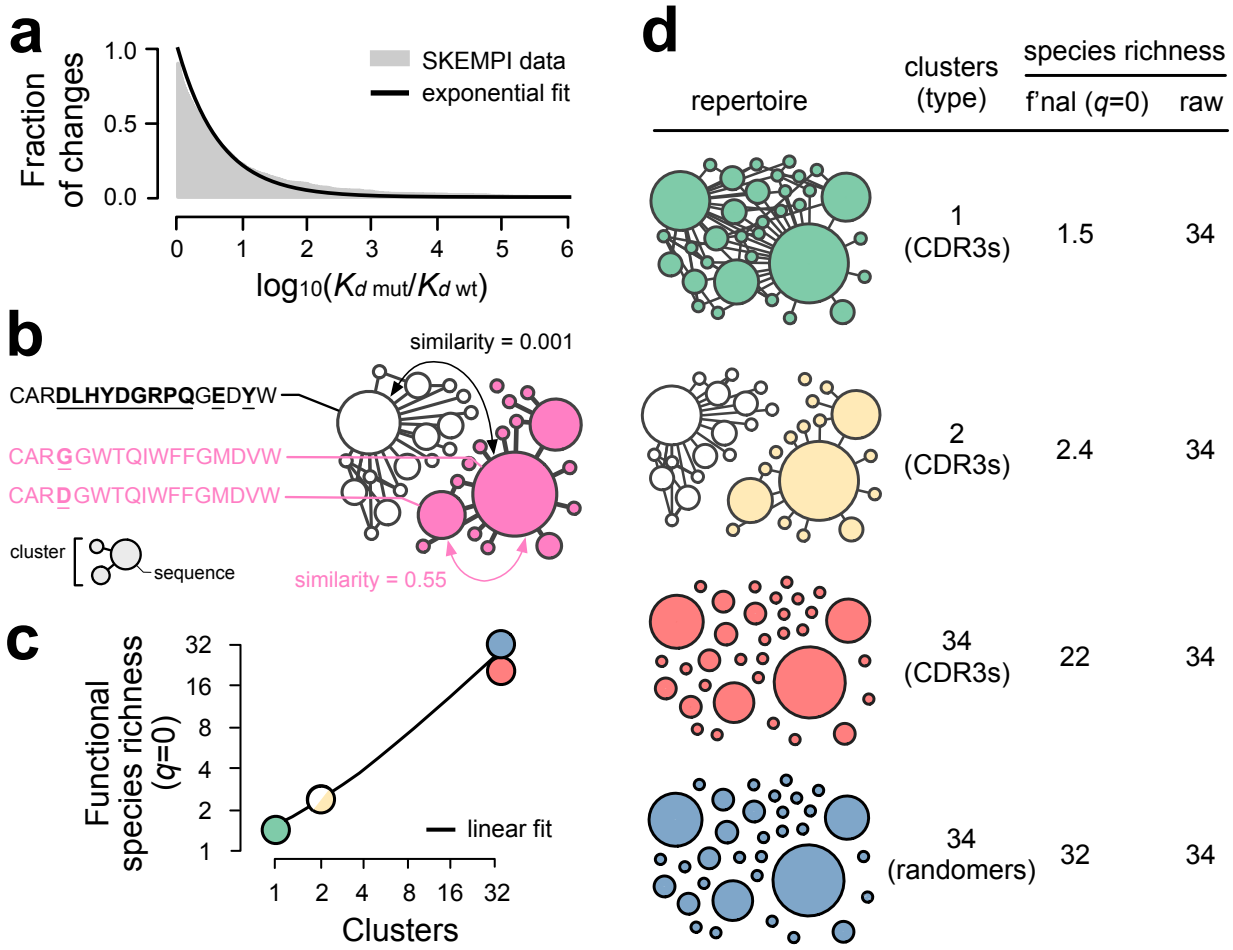
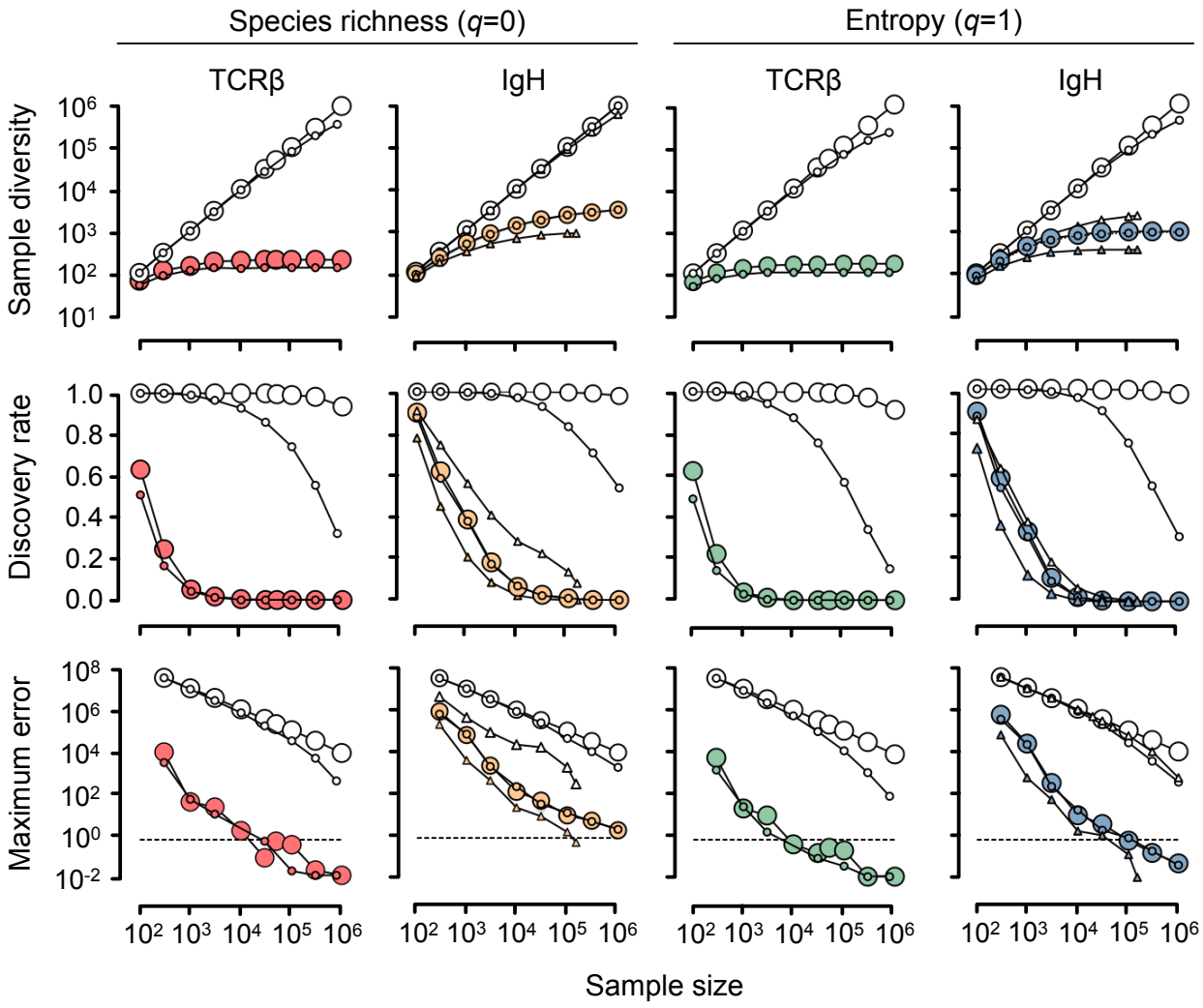

a
Binding target
Cluster of similarly-shaped binding targets
Shape space
b
Targets bound by a given CDR3
Raw species richness: 6
c
Nonoverlapping set of targets
Functional species richness: 4

Figure 2


a
Fraction of changes
1.0
0.5
0.0
SKEMPI data
exponential fit
0 1 2 3 4 5 6
log10(Kd mut/Kd wt)
b
similarity = 0.001
CARDLHYDGRPQGEDYW
CARGGWTQIWFFGMDVW
CARDGWTQIWFFGMDVW
cluster
sequence
similarity = 0.55
c
Functional species richness (q=0)
32
16
8
4
2
1
linear fit
1 2 4 8 16 32
Clusters
d
repertoire
clusters (type)
species richness
f'nal (q=0)
raw
1 (CDR3s)
1.5
34
2 (CDR3s)
2.4
34
34 (CDR3s)
22
34
34 (randomers)
32
34

677 Figure 3



678

Figure 4

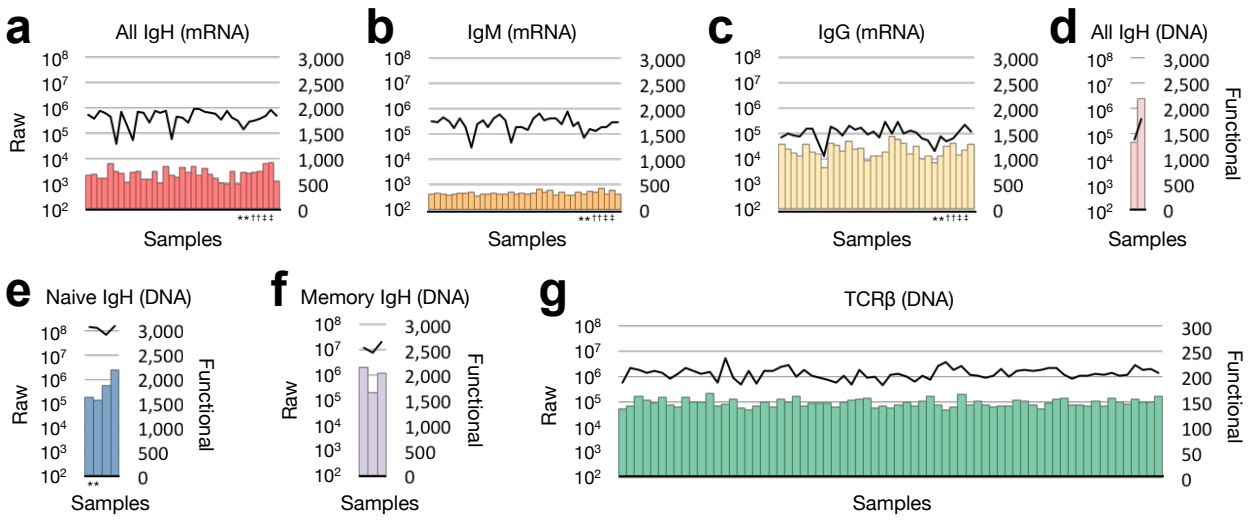
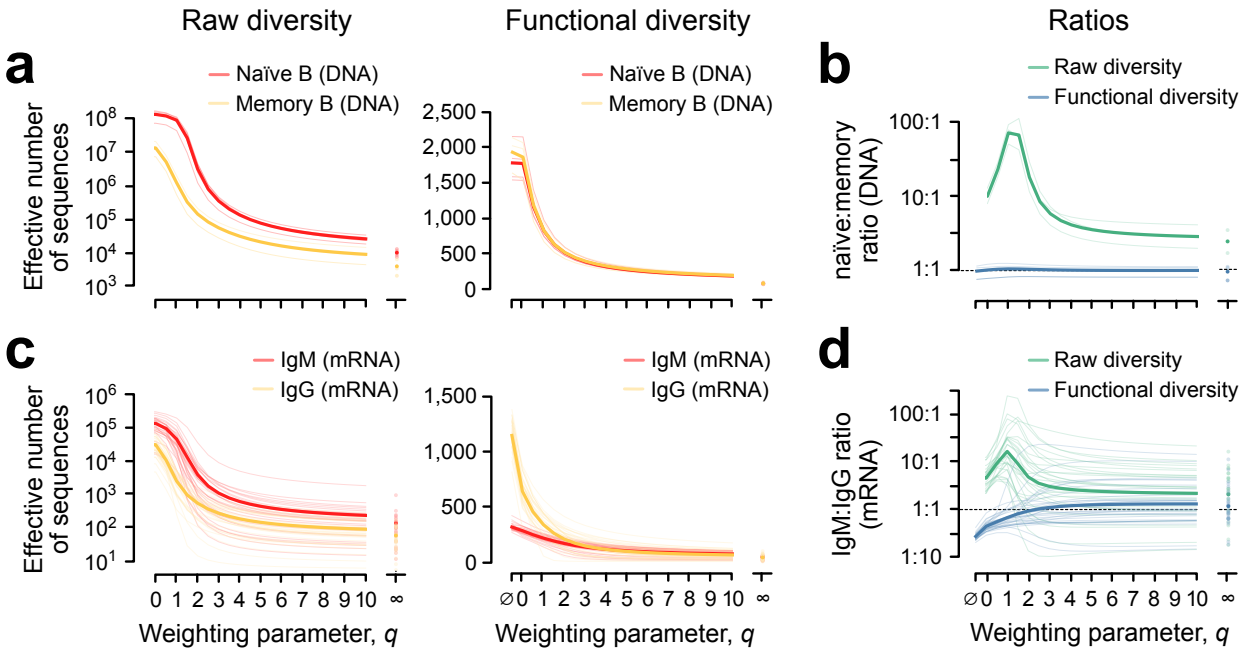
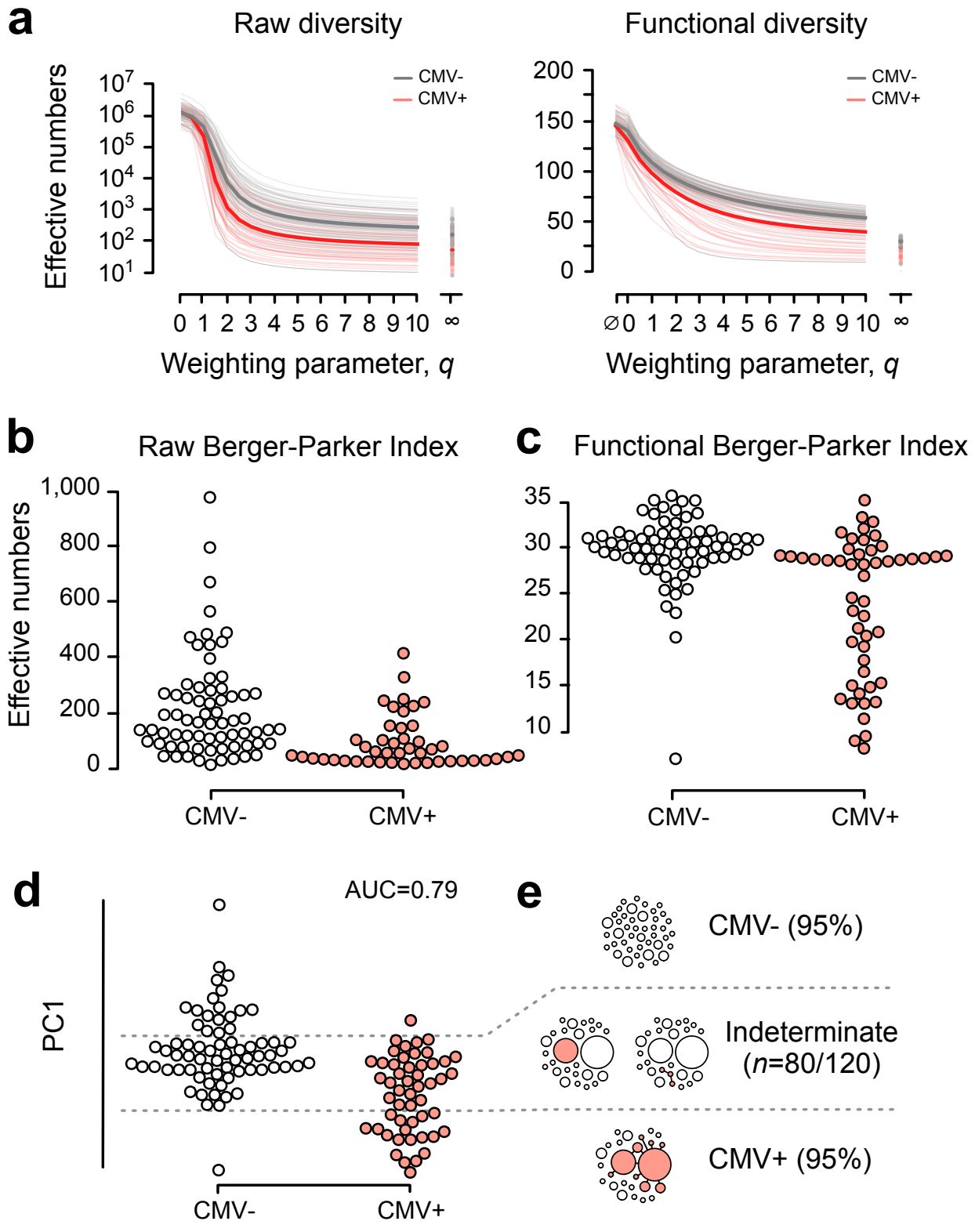

a
All IgH (mRNA)
b
IgM (mRNA)
c
IgG (mRNA)
d
All IgH (DNA)
e
Naive IgH (DNA)
f
Memory IgH (DNA)
g
TCRβ (DNA)
Raw
Functional
Samples
10^8
10^7
10^6
10^5
10^4
10^3
10^2
3,000
2,500
2,000
1,500
1,000
500
0
300
250
200
150
100
50
**††‡‡

Figure 5


Raw diversity
Functional diversity
Ratios
a
Naïve B (DNA)
Memory B (DNA)
Effective number of sequences
b
Raw diversity
Functional diversity
naïve:memory ratio (DNA)
100:1
10:1
1:1
c
IgM (mRNA)
IgG (mRNA)
d
IgM:IgG ratio (mRNA)
1:10
Weighting parameter, q

683 Figure 6



684

Figure 7

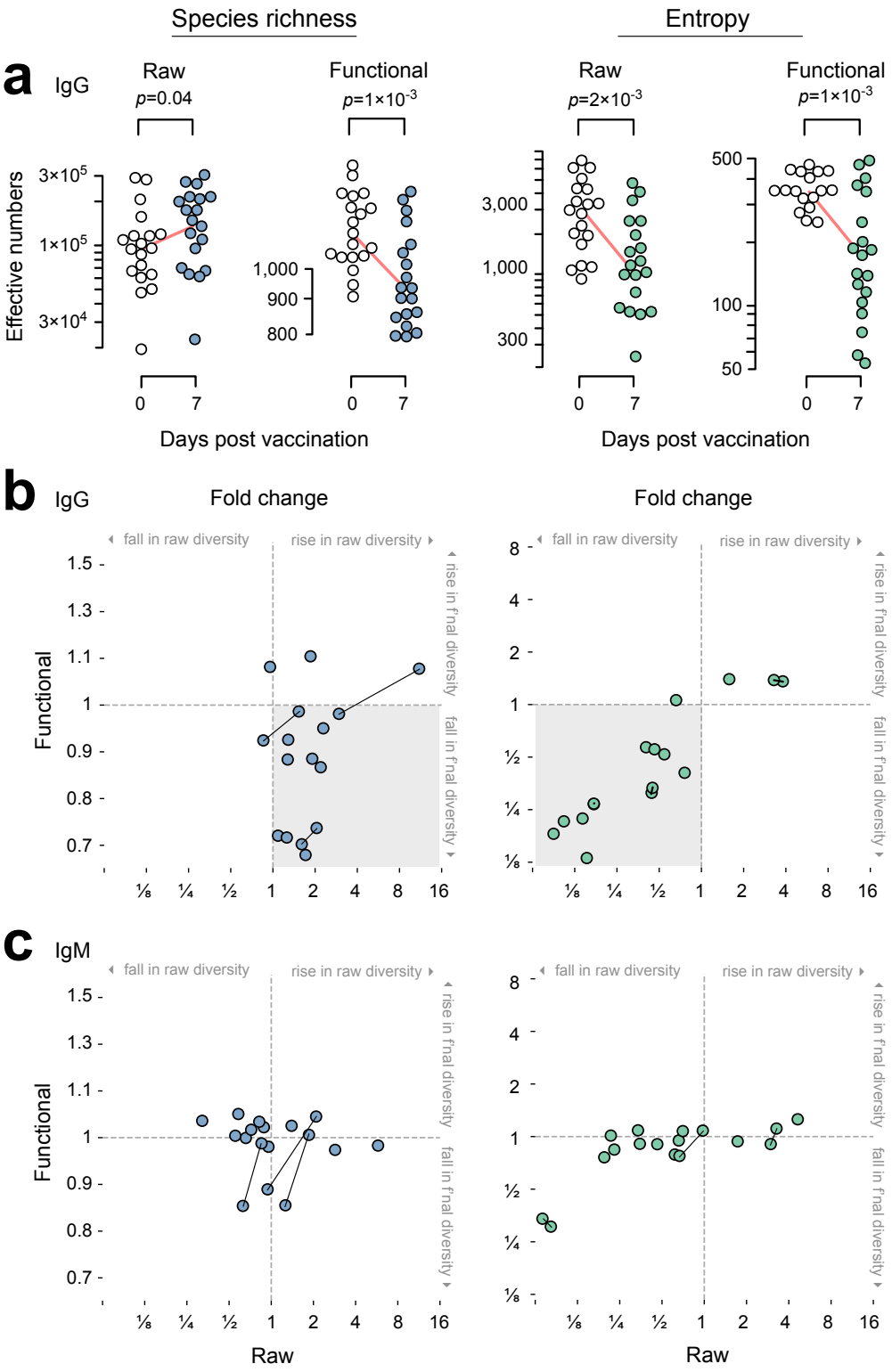

Species richness
Entropy
a
IgG
Raw
p=0.04
Functional
p=1×10-3
Raw
p=2×10-3
Functional
p=1×10-3
Effective numbers
3×10^5
1×10^5
3×10^4
1,000
900
800
3,000
1,000
300
500
100
50
0
7
Days post vaccination
b
IgG
Fold change
fall in raw diversity
rise in raw diversity
rise in final diversity
fall in final diversity
Functional
1.5
1.3
1.1
1
0.9
0.8
0.7
8
4
2
½
¼
⅛
16
c
IgM
Raw

Figure 8

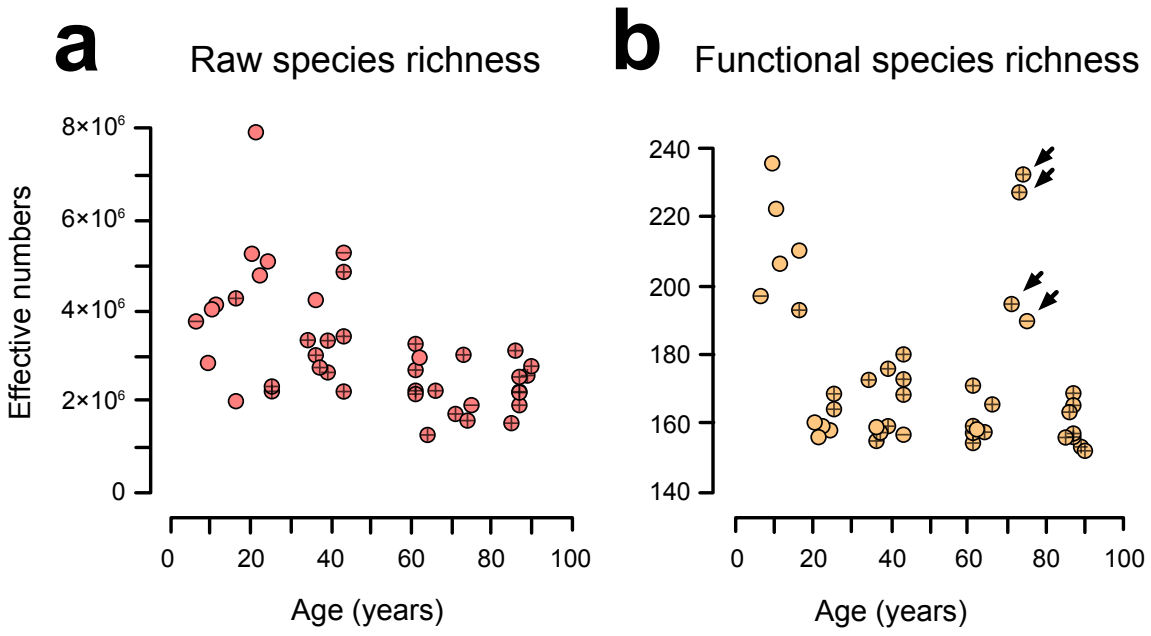
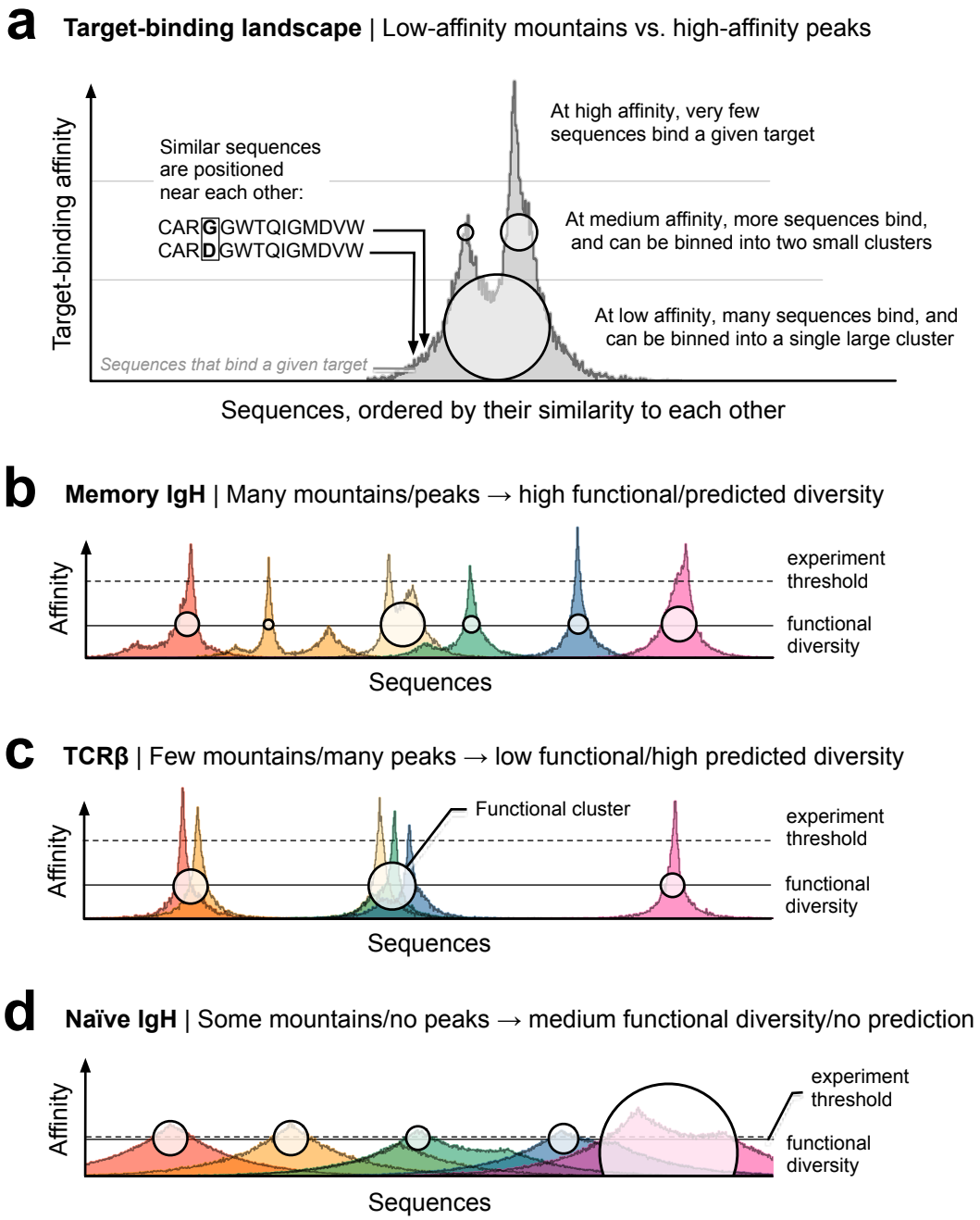

a
Raw species richness
Effective numbers
8×10⁶
6×10⁶
4×10⁶
2×10⁶
0
0
20
40
60
80
100
Age (years)
b
Functional species richness
240
220
200
180
160
140
0
20
40
60
80
100
Age (years)

Figure 9


a
Target-binding landscape | Low-affinity mountains vs. high-affinity peaks
Target-binding affinity
Similar sequences are positioned near each other:
CARGGWTQIGMDVW
CARDGWTQIGMDVW
At high affinity, very few sequences bind a given target
At medium affinity, more sequences bind, and can be binned into two small clusters
At low affinity, many sequences bind, and can be binned into a single large cluster
Sequences that bind a given target
Sequences, ordered by their similarity to each other
b
Memory IgH | Many mountains/peaks → high functional/predicted diversity
Affinity
experiment threshold
functional diversity
Sequences
c
TCRβ | Few mountains/many peaks → low functional/high predicted diversity
Affinity
Functional cluster
experiment threshold
functional diversity
Sequences
d
Naïve IgH | Some mountains/no peaks → medium functional diversity/no prediction
Affinity
experiment threshold
functional diversity
Sequences

References

Almendro, V., Kim, H.J., Cheng, Y.-K., Gönen, M., Itzkovitz, S., Argani, P., van Oudenaarden, A., Sukumar, S., Michor, F., and Polyak, K. (2014). Genetic and phenotypic diversity in breast tumor metastases. Cancer Res *74*, 1338–1348.

Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., and Koralov, S.B. (2011). High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. PLOS ONE *6*, e22365.

Bachmann, M.F., Kündig, T.M., Kalberer, C.P., Hengartner, H., and Zinkernagel, R.M. (1994). How many specific B cells are needed to protect against a virus? J. Immunol. *152*, 4235–4241.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.

Botta-Dukát, Z. (2018). The generalized replication principle and the partitioning of functional diversity into independent alpha and beta components. Ecography *41*, 40–50.

de Bourcy, C.F.A., Angel, C.J.L., Vollmers, C., Dekker, C.L., Davis, M.M., and Quake, S.R. (2017). Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. Proc Natl Acad Sci U S A *114*, 1105–1110.

Britanova, O.V., Putintseva, E.V., Shugay, M., Merzlyak, E.M., Turchaninova, M.A., Staroverov, D.B., Bolotin, D.A., Lukyanov, S., Bogdanova, E.A., Mamedov, I.Z., et al. (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J. Immunol. *192*, 2689–2698.

Bunge, J., and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. Journal of the American Statistical Association *88*, 364–373.

Chao, A., Chiu, C.-H., Villéger, S., Sun, I.F., Thorn, S., Lin, Y., Chiang, J.-M., and B. Sherwin, W. (2018). An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. Ecological Monographs.

718 Chao, A., Chiu, C.-H., Villéger, S., Sun, I.-F., Thorn, S., Lin, Y.-C., Chiang, J.-M., and Sherwin,
719 W.B. An attribute-diversity approach to functional diversity, functional beta diversity, and related
720 (dis)similarity measures. Ecological Monographs *0*.

721 Chen, Z.J., Wheeler, C.J., Shi, W., Wu, A.J., Yarboro, C.H., Gallagher, M., and Notkins, A.L.
722 (1998). Polyreactive antigen-binding B cells are the predominant cell type in the newborn B cell
723 repertoire. Eur. J. Immunol. *28*, 989–994.

724 Chiu, C.-H., and Chao, A. (2014). Distance-Based Functional Diversity Measures and Their
725 Decomposition: A Framework Based on Hill Numbers. PLOS ONE *9*, e100014.

726 DeWitt, W.S., Lindau, P., Snyder, T.M., Sherwood, A.M., Vignali, M., Carlson, C.S., Greenberg,
727 P.D., Duerkopp, N., Emerson, R.O., and Robins, H.S. (2016). A Public Database of Memory
728 and Naive B-Cell Receptor Sequences. PLoS ONE *11*, e0160853.

729 Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C.M.
730 (2014). SAbDab: the structural antibody database. Nucleic Acids Res. *42*, D1140-1146.

731 Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais, C.,
732 Klinger, M., Carlson, C.S., Hansen, J.A., et al. (2017). Immunosequencing identifies signatures
733 of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat
734 Genet *49*, 659–665.

735 Emery, V. (2001). Investigation of CMV disease in immunocompromised patients. J Clin Pathol
736 *54*, 84–88.

737 Fairlie-Clarke, K.J., Shuker, D.M., and Graham, A.L. (2009). Why do adaptive immune
738 responses cross-react? Evol Appl *2*, 122–131.

739 Frank, S.A. (2002). Immunology and Evolution of Infectious Disease (Princeton (NJ): Princeton
740 University Press).

741 Gibson, K.L., Wu, Y.-C., Barnett, Y., Duggan, O., Vaughan, R., Kondeatis, E., Nilsson, B.-O.,
742 Wikby, A., Kipling, D., and Dunn-Walters, D.K. (2009). B-cell diversity decreases in old age and
743 is correlated with poor health status. Aging Cell *8*, 18–25.

744 Greiff, V., Bhat, P., Cook, S.C., Menzel, U., Kang, W., and Reddy, S.T. (2015). A bioinformatic
745 framework for immune repertoire diversity profiling enables detection of immunological status.
746 Genome Med *7*.

747 Heindl, A., Lan, C., Rodrigues, D.N., Koelble, K., and Yuan, Y. (2016). Similarity and diversity of
748 the tumor microenvironment in multiple metastases: critical implications for overall and
749 progression-free survival of high-grade serous ovarian cancer. Oncotarget *7*, 71123–71135.

750 Hill, M.O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology
751 *54*, 427–432.

752 Holodick, N.E., Rodríguez-Zhurbenko, N., and Hernández, A.M. (2017). Defining Natural
753 Antibodies. Front Immunol *8*.

754 Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks,
755 D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–
756 135.

757 Hopkins, A.C., Yarchoan, M., Durham, J.N., Yusko, E.C., Rytlewski, J.A., Robins, H.S., Laheru,
758 D.A., Le, D.T., Lutz, E.R., and Jaffee, E.M. (2018). T cell receptor repertoire features associated
759 with survival in immunotherapy-treated pancreatic ductal adenocarcinoma. JCI Insight *3*.

760 Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J., and Moal, I.H. (2018).
761 SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and
762 thermodynamics upon mutation. Bioinformatics.

763 Jiang, N., He, J., Weinstein, J.A., Penland, L., Sasaki, S., He, X.-S., Dekker, C.L., Zheng, N.-Y.,
764 Huang, M., Sullivan, M., et al. (2013). Lineage structure of the human antibody repertoire in
765 response to influenza vaccination. Sci Transl Med *5*, 171ra19.

766 Jost, L. (2007). Partitioning diversity into independent alpha and beta components. Ecology *88*,
767 2427–2439.

768 Ju, C.-H., Blum, L.K., Kongpachith, S., Lingampalli, N., Mao, R., Brodin, P., Dekker, C.L., Davis,
769 M.M., and Robinson, W.H. (2018). Plasmablast antibody repertoires in elderly influenza vaccine
770 responders exhibit restricted diversity but increased breadth of binding across influenza strains.
771 Clin. Immunol. *193*, 70–79.

772 Kaplinsky, J., and Arnaout, R. (2016). Robust estimates of overall immune-repertoire diversity
773 from high-throughput measurements on samples. Nat Commun *7*, 11881.

774 Kaplinsky, J., Li, A., Sun, A., Coffre, M., Koralov, S.B., and Arnaout, R. (2014). Antibody
775 repertoire deep sequencing reveals antigen-independent selection in maturing B cells. Proc.
776 Natl. Acad. Sci. U.S.A. *111*, E2622-2629.

777 Koopmans, R., and Schaeffer, M. (2013). De-composing diversity: In-group size and out-group
778 entropy and their relationship to neighbourhood cohesion (WZB Berlin Social Science Center).

779 Langman, R.E., and Cohn, M. (1987). The E-T (elephant-tadpole) paradox necessitates the
780 concept of a unit of B-cell function: the protection. Mol. Immunol. *24*, 675–697.

781 Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and
782 Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in Proteins. Science
783 *322*, 438–442.

784 Leem, J., de Oliveira, S.H.P., Krawczyk, K., and Deane, C.M. (2018). STCRDab: the structural
785 T-cell receptor database. Nucleic Acids Res. *46*, D406–D412.

786 Leinster, T., and Cobbold, C.A. (2012). Measuring diversity: the importance of species similarity.
787 Ecology *93*, 477–489.

788 Levy, E.D. (2010). A simple definition of structural regions in proteins and its use in analyzing
789 interface evolution. J. Mol. Biol. *403*, 660–670.

790 Li, K., Bihan, M., Yooseph, S., and Methé, B.A. (2012). Analyses of the Microbial Diversity
791 across the Human Microbiome. PLOS ONE *7*, e32118.

792 Lunzer, M., Golding, G.B., and Dean, A.M. (2010). Pervasive cryptic epistasis in molecular
793 evolution. PLoS Genet. *6*, e1001162.

794 Macarthur, R.H. (1965). Patterns of Species Diversity. Biological Reviews *40*, 510–533.

795 Marion, Z.H., Fordyce, J.A., and Fitzpatrick, B.M. (2015). Extending the Concept of Diversity
796 Partitioning to Characterize Phenotypic Complexity. Am. Nat. *186*, 348–361.

797 Messaoudi, I., Lemaoult, J., Guevara-Patino, J.A., Metzner, B.M., and Nikolich-Zugich, J.
798 (2004). Age-related CD8 T cell clonal expansions constrict CD8 T cell repertoire and have the
799 potential to impair immune defense. J. Exp. Med. *200*, 1347–1358.

800 Morris, E.K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T.S., Meiners, T., Müller,
801 C., Obermaier, E., Prati, D., et al. (2014). Choosing and using diversity indices: insights for
802 ecological applications from the German Biodiversity Exploratories. Ecol Evol *4*, 3514–3524.

803 Notkins, A.L. (2004). Polyreactivity of antibody molecules. Trends Immunol. *25*, 174–179.

804 Obar, J.J., Khanna, K.M., and Lefrançois, L. (2008). Endogenous naive CD8+ T cell precursor
805 frequency regulates primary and memory responses to infection. Immunity *28*, 859–869.

806 Perelson, A.S., and Oster, G.F. (1979). Theoretical studies of clonal selection: minimal antibody
807 repertoire size and reliability of self-non-self discrimination. J. Theor. Biol. *81*, 645–670.

808 Pons, J., Rajpal, A., and Kirsch, J.F. (1999). Energetic analysis of an antigen/antibody interface:
809 alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme
810 interaction. Protein Sci. *8*, 958–968.

811 Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R.A., Weyand, C.M.,
812 Boyd, S.D., and Goronzy, J.J. (2014). Diversity and clonal selection in the human T-cell
813 repertoire. Proc Natl Acad Sci U S A *111*, 13139–13144.

814 Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R.,
815 Carlson, C.S., and Warren, E.H. (2010). Overlap and effective size of the human CD8+ T cell
816 receptor repertoire. Sci Transl Med *2*, 47ra64.

817 Salinas, V.H., and Ranganathan, R. (2018). Coevolution-based inference of amino acid
818 interactions underlying protein function. Elife *7*.

819 Scheiner, S.M. (2012). A metric of biodiversity that integrates abundance, phylogeny, and
820 function. Oikos *121*, 1191–1202.

821 Schober, K., Buchholz, V.R., and Busch, D.H. (2018). TCR repertoire evolution during
822 maintenance of CMV-specific T-cell populations. Immunol. Rev. *283*, 113–128.

Smith, D.J., Forrest, S., Hightower, R.R., and Perelson, A.S. (1997). Deriving Shape Space Parameters from Immunological Data. Journal of Theoretical Biology *189*, 141–150.

Taraska, J.W. (2015). Cell biology of the future: Nanometer-scale cellular cartography. J Cell Biol *211*, 211–214.

Taylor, M.G., Rajpal, A., and Kirsch, J.F. (1998). Kinetic epitope mapping of the chicken lysozyme.HyHEL-10 Fab complex: delineation of docking trajectories. Protein Sci. *7*, 1857–1867.

Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L., and Quake, S.R. (2013). Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. U.S.A. *110*, 13463–13468.

Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S., and Quake, S.R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. Science *324*, 807–810.

Whittaker, J., Groth, A.V., Mynarcik, D.C., Pluzek, L., Gadsbøll, V.L., and Whittaker, L.J. (2001). Alanine scanning mutagenesis of a type 1 insulin-like growth factor receptor ligand binding site. J. Biol. Chem. *276*, 43980–43986.

Xu, J.L., and Davis, M.M. (2000). Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity *13*, 37–45.

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

PyMOLThe PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.