Inferring repertoire dynamics from repertoire sequencing

Maximilian Puelma Touzel
Laboratoire de Physique Théorique, ENS-PSL Research University, Paris, France and Mila, Université de Montréal, Montreal, Canada

Aleksandra Walczak
Laboratoire de Physique Théorique, ENS-PSL Research University, Paris, France

Thierry Mora
Laboratoire de Physique Statistique, ENS-PSL Research University, Paris, France,

Good, but need to add new, main takeaways. High-throughput sequencing provides access to expression-level detail of cell populations. Identifying signal in this data is nevertheless a challenge: intrinsic variability, vast sub-sampling, and indirect access together make difficult the reliable and accurate inference of the changes in population size due to environmental perturbations. Here, in the context of antigen-perturbed immune cell repertoires, we formulate access to the unobserved repertoire using a generative model of observed sequence count pairs from a reference and differentially-expressed condition. When applied to pairs of replicates, our model captures the natural variability in the system, giving reproducible behavior across donors and, in spite of sub-sampling, provides plausible parameter estimates for the underlying repertoire. Using this replicate model as a baseline, we then formulate the differentially expressed condition using a prior distribution on the ratio of a clone's frequency pair for pairs of repertoires sampled at different time points. The posterior distribution of this ratio for each observed clone is then obtained and used to characterize if and how strongly it participates in the response. Applying our approach to yellow fever vaccination as a model of acute infection in humans, we identify candidate clones participating in the response.

Next generation sequencing allows us to gain access to repertoire-wide data supporting more comprehensive repertoire analysis and more robust vaccine design [1]. Despite large-scale efforts[2], how repertoire statistics respond to such acute perturbations is unknown. Longitudinal repertoire sequencing (RepSeq) makes possible the characterization of repertoire dynamics. Despite the large number of samples (clones) in these datasets lending it to model-based inference, there are few existing model-based approaches to this analysis. Most current approaches (e.g. [3]) quantify repertoire response properties using measurement statistics that are limited to what is observed in the sample, rather than what transpires in the individual. Model-based approaches, in contrast, can in principle capture features of the actual repertoire response to, for instance ongoing, natural stimuli, modeled as a point process of infections, and giving rise to diffusion-like response dynamics. Another regime for model-based approaches is the response to a single, strong perturbation, such as a vaccine, giving rise to a stereotyped, transient response dynamics. In either case, a measurement model is needed since what is observed (molecule counts) is indirect. We also only observe a small fraction of the total number of clones, so some extrapolation is necessary. Finally, both the underlying clonal population dynamics and the transformation applied by the measurement is stochastic, each contributing its own variability, making inferences based on sample ratios of molecule counts inaccurate.

Inference of frequency variation from sequencing data has been intensely researched in other areas of systems biology, such as in RNAseq studies. There, approaches are becoming standardized (DESEQ2 [4],EdgeR [5], etc.) and technical problems have been formulated and partly addressed. The differences between RNAseq and RepSeq data, however, means that direct translation of these methods is questionable. Moreover, the known structure of clonal populations may be leveraged for model-based inference using RepSeq, potentially providing advantages over existing RNAseq-based approaches.

Here, we take a generative modeling approach to repertoire dynamics. Our model incorporates known features of clonal frequency statistics and the statistics of the sequencing process. The models we consider are designed to be learnable using RepSeq data, and then used to infer properties of the repertoires of the individuals providing the samples. To guide its development, we have analyzed a longitudinal dataset around yellow fever vaccination (some results of this analysis are published [6]). Yellow fever serves as model of acute infection in humans and here we present analyses of this data set that highlights the inferential power of our approach to uncover perturbed repertoire dynamics.

The paper is organized as follows. In Section A, we specify the generic form of the generative model of receptor RNA count pair statistics and its assumptions. Then in Section B, we apply it to replicate pairs to quantify the natural size variation. Next, in Section C we apply it to construct a differential expression model by augmenting this null model with the hidden log-frequency

fold-change of clone frequency between the two compared conditions, learning a prior on such change from the data. Seeing how the parameters of this prior vary across pair time-point comparisons provides the repertoire dynamics at ensemble-level of description. We also show how our learned model can be used to infer a posterior probability distribution of fold change for any observed clone. We show that the latter can be used to infer the temporal changes of that clone's frequency. Finally, in Section D we use this fact to infer from the posterior expansion probability a list of clones significantly expanded by YF vaccination across a given pair of time points.

RESULTS

A. A repertoire model family learnable from pair RepSeq datasets

Mathematically, dynamics in a wide variety of contexts is captured by a time-evolution operator, or propagator, $F_{t,t'}$, that evolves the state of the system from time t to time t'. Our model formulation of repertoire dynamics thus focuses on characterizing the transformation of clone frequencies between a pair of time points (called reference and test). All such frequencies are unobserved, however, and so we propose a family of generative models of the count pair statistics of what is actually measured: immune receptor RNA molecules obtained by sequencing blood samples. In this case, the model parameters are constrained by the corresponding pair of measured repertoires at the reference and test times.

Our method to determine differential expression proceeds in two steps, where in each we define, learn, and analyze an instance of this model family, first for same day replicates, then for a reference and test pair of conditions. With the latter model learned, we can arrive at estimates for quantities characterizing the repertoire as a whole, such as the total number of clones, N. A second application is to use the model to make statements about individual observed clones. Namely, what can be said about changes in a clone's frequency between the two time points based on its observed molecule count pair. All code used to produce the results in this work was custom written in Python 3 and is publically available online on GitHub.

Model family definition

We formulate an immune repertoire as a set of N clone frequencies $\vec{f} = (f_1, \ldots, f_N)$, each within the interval $[f_{\min}, 1]$, where $f_{\min}$ is the minimum allowed frequency of corresponding to a single lymphocyte. A prior density over clone frequencies is given by $\rho(f)$. N and $f_{\min}$ must be determined self-consistently when defining the corresponding joint density,

$$\rho_N(\vec{f}) \propto \prod_{i=1}^{N} \rho(f_i)\delta(Z_f - 1) \,, \tag{1}$$

where the Dirac delta-function, $\delta(x)$, is used to impose a normalization constraint on the sum of frequencies, $Z_f = \sum_{i=1}^{N} f_i$,

$$Z_f = 1 \,. \tag{2}$$

Accounting for normalization (see next section), the joint density over clones factorizes and so we focus on the statistics of single clones.

Each clone's frequency pair from the reference and test conditions impacts its chance of being picked up in a realization of the acquisition process. This process consists of pair blood sampling and standard bar-coded RepSeq RNA sequencing. Post-sequencing, consensus reduction produces a list of unique RNA receptor sequences along with their abundance (i.e. molecular count) in the sample. We present a model, $P(n, n', f, f')$, with f and f' and n and n' denoting a clone's frequencies and receptor molecule counts in the reference and test condition, respectively. The model consists of priors on reference size statistics, $\rho(f)$, and the statistics of the transformed sizes, $\rho(f'|f)$. The rest of the model is an observation model of the acquisition process.

In general, repertoires are dominated in number by small clones missed in the acquisition process. Thus, in any realization, $n+n' > 0$ for only a relatively small number, $N_{\text{obs}} \ll N$, of clones, which can still be large since N is typically 10^6 (10^9) for mouse (human). These N_{obs} *observed* clones are those captured in the blood sample and amplified above detection levels in the sequencer in at least one of the test and reference conditions. We assume we have no further experimental access to the *unobserved* clones that realize with $n + n' = 0$. Marginalizing over f and f' and conditioning on $n + n' > 0$, we obtain the model prediction for what we observe,

$$P(n, n'|n + n' > 0) = \frac{1 - \delta_{n0}\delta_{n'0}}{1 - P(0, 0)} P(n, n') \,, \tag{3}$$

i.e. the distribution of count pairs from observed clones, where $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. The model estimate for the total number of clones is then $N = N_{\text{obs}}/(1 - P(0, 0))$.

Normalization

The $N - N_{\text{obs}}$ *unobserved* clones influence the observed count statistics only via the presence of their frequencies in the two normalization constraints, $Z_f = 1$ and $Z_{f'} = 1$, so far unaccounted for in the model. In Appendix A, we show that $Z_f = 1$ is implicitly satisfied if

$$N\langle f \rangle = 1 \,. \tag{4}$$

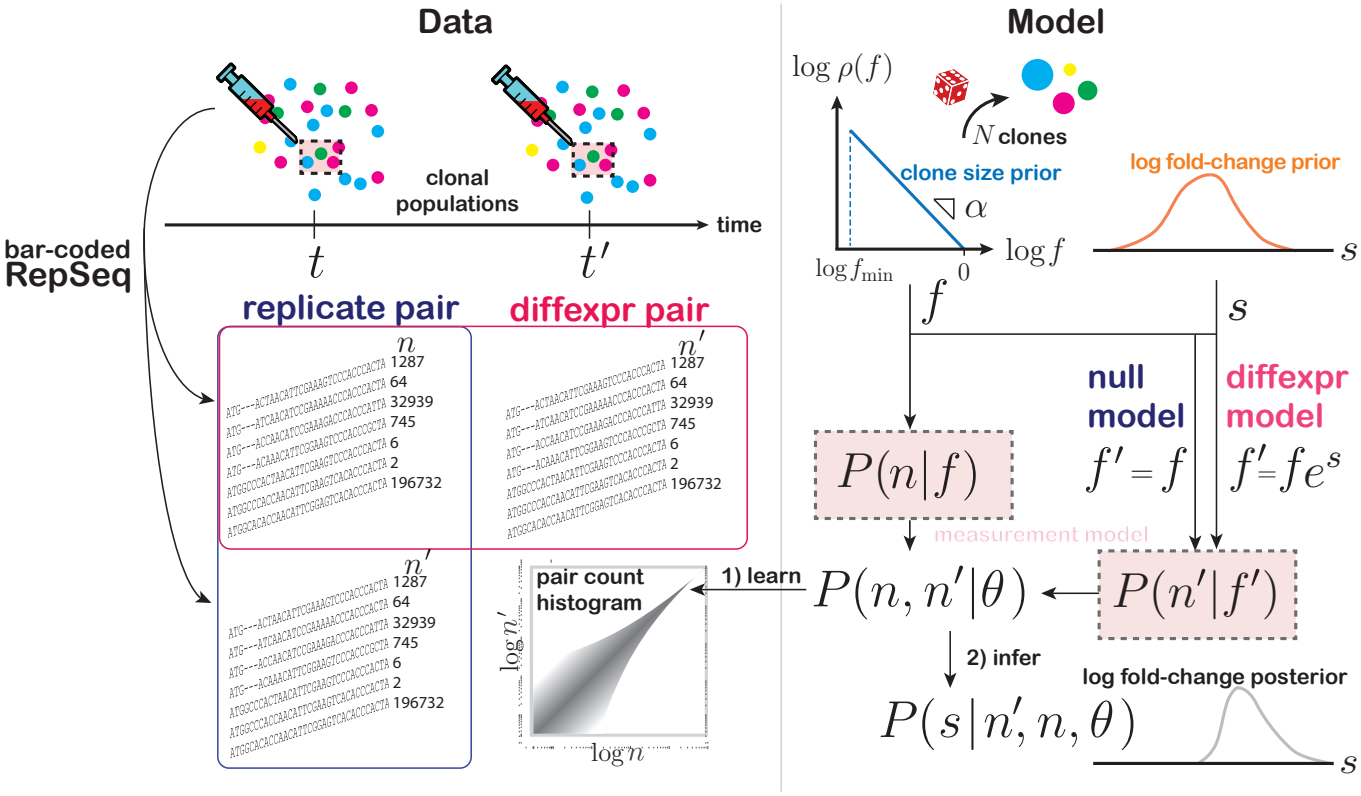

Figure 1. *Data and modeling method.* restructure Left: Data sets consist of pairs of bar-coded repertoire-sequenced (RepSeq) blood samples, either from same or different days. Each data set is summarized in its pair count histogram. Right: Clone frequency distribution, $\rho(f)$, is set as a power law, parameterized by the power, α_f and the minimum frequency, $f_{\min}$. With a choice of the measurement model, the null model is then specified and learned using a replicate pair data set. Using these parameters and choosing a form for the log fold-change prior, the differential expression model is specified and learned on a differentially expressed pair data set. Once learned, the model is used for posterior inference performed on all observed clones.

By employing this constraint, we impose the desired self-consistency between $f_{\min}$ and N. Equivalently, it is expressed using the frequency posteriors, which we separate into unobserved and observed contributions,

$$1 = NP(0,0)\langle f\rangle_{\rho(f|n+n'=0)} + N\sum_{n+n'>0} P(n,n')\langle f\rangle_{\rho(f|n,n')} \cdot$$

Z_f and $Z_{f'}$ are insensitive to the precise frequency values of a realized set of unobserved clones, and their average frequency is well approximated as the ensemble average in first term above. In contrast, the sum of frequencies of the observed clones might depend on the realization, especially in the case of large, outlying clones arising from power-law distributed clone sizes. This sensitivity can nevertheless be incorporated into the model using the observed dataset of pair of molecule counts, $\mathcal{D} = \{(n_i, n_i')\}_{i=1}^{N_{\rm obs}}$, by using an importance sampling approximation, $\sum_{n+n'>0} P(n,n') \approx \frac{1}{N}\sum_{i=1}^{N_{\rm obs}}$, so that the second term is $\sum_{i=1}^{N_{\rm obs}}\langle f\rangle_{\rho(f|n_i,n_i')}$. We thus define the right-hand side of this realization-dependent constraint

$$Z_f^{\mathcal{D}} = NP(0,0)\langle f\rangle_{\rho(f|n+n'=0)} + \sum_{i=1}^{N_{\rm obs}}\langle f\rangle_{\rho(f|n_i,n_i')} \,, \quad (5)$$

and impose that $Z_f^{\mathcal{D}} = 1$, in addition to $N\langle f\rangle = 1$. We note that while not equivalent, differences in values of parameters learned with each constraint separately were small, suggesting there is a high overlap in the respective regions of the parameter space satisfying the original Eq. 4 constraint and realization-dependent Eq. 5 constraint (Fig. 12). We impose the same constraint on $\vec{f}'$, via the equivalent condition,

$$Z_{f'}^{\mathcal{D}} = Z_f^{\mathcal{D}} \,. \quad (6)$$

Parameter sharing

We take the 'common dispersion' approach [5], in which we assume that n and n' are conditionally independent once the reference and test frequency are given,

and that their statistics depend only implicitly on clone identity (*i.e.* clonal sequence) via these frequencies.

Learning procedure

Models were fit using a count pair dataset, $\mathcal{D}$, by maximizing the log marginal likelihood of the data, $\sum_{i=1}^{N_{\text{obs}}} \log P(n_i, n_i'|\theta)$, over the free parameters, θ, subject to the above constraints. We proceed in two steps. In the first step, we consider a null model in which a replicate, e.g. same-day sample, is given for the test condition. In this case, the reference and test frequency are the same, $\rho(f'|f) = \delta(f' - f)$, for all clones and no additional constraint for $\vec{f}'$ is needed. The learned parameters of $\rho(f)$ and the acquisition model from this pair serves to define the baseline, e.g. pre-vaccination statistics. In the second step, we consider a model for differential expression in which a differentially expressed condition serves as the test, e.g. the reference and test condition being pre- and post-vaccination, respectively. Here, the parameters of $\rho(f)$ and the acquisition model here are set to those of the null model of reference day 0. As a result, $Z_f^{\mathcal{D}}$ is not unity, but in practice we find it is close, and thus so is $Z_{f'}^{\mathcal{D}}$ on account of the constraint on $\vec{f}'$, Eq. 6. What is different here from the replicate case is $\rho(f'|f)$: the test frequency, f', is obtained from a non-identity transformation of the reference frequency, f. This transformation summarizes the effect of the dynamics assumed to act on clone sizes in the period between the two sample times. In the absence of a strong perturbation, such as a vaccine or acute infection, this dynamics is dominated by the diffusive behavior of some stochastic population dynamics for which the transformation is given by the corresponding Green's function. For a strong, transient perturbation, in contrast, time-translation invariance is broken and a transformation tailored to the properties of the transient perturbation must be specified. In the context of immune response to yellow fever vaccination, we focus on the latter.

B. The replicate pair case: a null model capturing baseline clone size variation

Replicate variability

Using this model family, we defined a null model of count pair statistics and fit it to a pair of replicates. This model provides a baseline variability with which differential expression can after be assessed. The marginal count pair distribution of this null model is

$$P(n, n'|\theta_{\text{null}}) = \int P(n|f)P(n'|f' = f)\rho(f)\mathrm{d}f \ , \quad (7)$$

where we have collected the parameters into θ_{null}. The influence of f on $P(n|f)$ is an explicit parametrization (see Fig. 2). For example, $P(n|f)$ as a Poisson distribution with mean proportional to f. Current methods, e.g. [5], more accurately model $P(n|f)$ by accounting for its observed over-dispersion using a negative binomial distribution. The number of cells of a clone in the sample, m, is an additional random variable in the measurement process chain, which has so far been neglected. Thus, in a further refined choice for $P(n|f)$, we can explicitly account for this step by choosing $P(m|f)$ as a negative binomial distribution and then $P(n|m)$ as a Poisson distribution, giving $P(n|f) = \sum_m P(n|m)P(m|f)$. This two-step model, as a more explicit representation, more accurately captures the count statistics of the measurement process, especially at low counts. The latter fact arises from the power-law nature of frequency distribution, for which the frequency of most clones falls below the sampling depth so that the majority of clones are not captured in the sample. These low frequency clones are so numerous, however, that the sample is nevertheless dominated by them, each appearing at the minimum finite size. For a single-step model, the minimal clone count is a single molecule. For a two-step model, in contrast, the minimal clone size is a single cell, which gives a small, but variable number of molecules. The one and two-step models thus leave different signatures in the statistics at low counts, even for the additional dilution of counts due to PCR inefficiency during the sequencing of the sample [7]. We find that, indeed, this two-step model exhibits a better fit to data, especially for clones captured with few counts (see Fig. 13). We find that the fit for the over-dispersed, one-step model is not significantly worse however, while the one-step Poisson model is clearly a poor choice, as it fails to capture the over-dispersion (as visible in Fig. 2).

Model validation

The null models were fit by maximizing the log marginal likelihood of the data, $\sum_{i=1}^{N_{\text{obs}}} \log P(n_i, n_i'|\theta_{\text{null}})$, over the parameters, θ_{null}. For the two-step model, the parameters are $\theta_{\text{null}} = (\alpha_f, M, a, \gamma, f_{\min})$, where α_f is the power law exponent, M is the total number of cells, a and γ are the coefficient and power of the over-dispersion term in the mean variance relation of the negative binomial distribution of cells, and finally, $f_{\min}$ is the minimum allowed clone frequency. We developed a sampling protocol for this model (see Appendix B) and used it to validate our learning procedure (Fig. 14), by verifying that it correctly learns ground truth.

Next, we assessed the ability of the two-step model to capture the measured count pair statistics of pairs of sequenced blood samples obtained from different volumes of the same sample, across days and donors. Figure 3

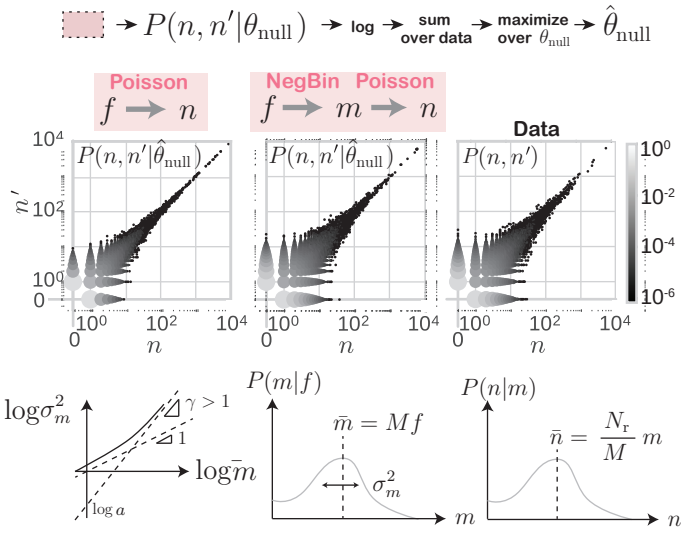

Figure 2. *Measurement model comparison.* (a) The null model learning procedure. With a measurement model specified, the count pair distribution is calculated and the likelihood of the data is maximized over the parameters, θ_{null}. (b) Data sampled from the learned models under different measurement models (left: one-step Poisson distribution, center: two-step Negative binomial distribution to Poisson distribution). Data is shown at right. (c) Parametrization of the two-step measurement model. Left: the mean-variance relationship specifying power, γ and coefficient, a, of the over-dispersion. Center: the cell count distribution with mean scaling with the number of cells in the sample, M. Right: the molecule count distribution with mean scaling with the number of cells and the sampling efficiency, M/N_r, with N_r the measured number of molecules in the sample.

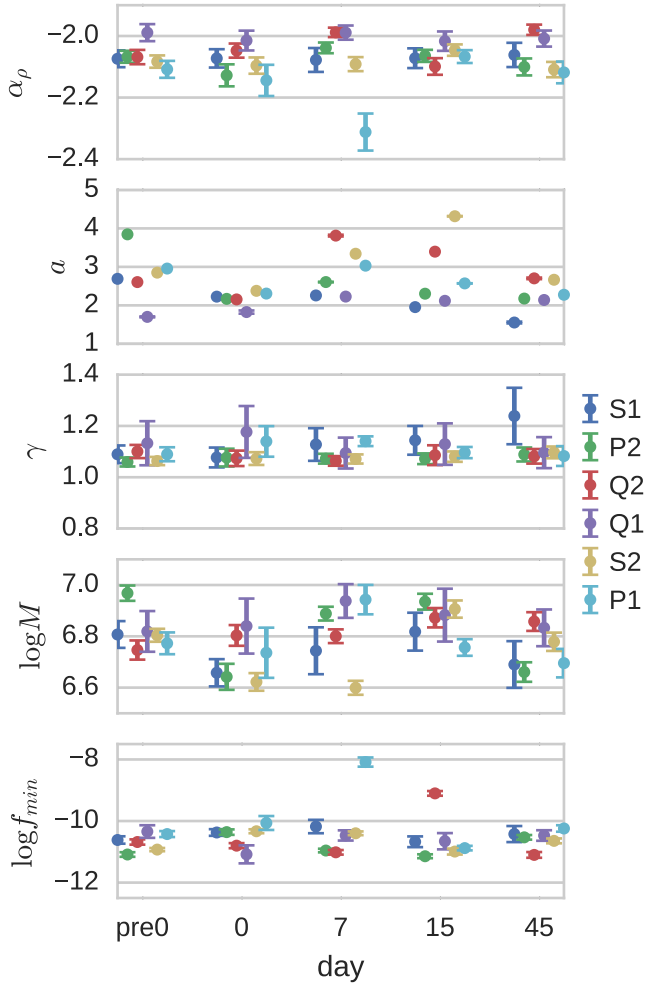

Figure 3. *Learned null model parameters.* Data is plotted separately for a pair of replicates across donors and time points. Error bars are obtained by projecting onto the respective parameter axis (not from projecting the Fisher Information in the manifold satisfying the constraint).

shows the learned values for 30 null models calculated from same-day replicates from 6 donors sampled over 5 time points spanning a 1.5 month period. While there is variability across donors and days, there is a surprising degree of regularity to the natural variability. In particular, despite estimates for M and f_{min} being very indirect, the learned values are within an order of magnitude of the expectation references okay?[8–10]. The learned values of M are consistent with rough estimates obtained from the known sample volume (personal communication, M. Pogorelyy), and the reciprocal of the learned values of f_{min} $(10^{10} - 10^{11})$ are plausible estimates for the total number of lymphocytes in the body. The uncertainty associated with these estimates was assessed using the lower-bound provided by the curvature of the likelihood function, *i.e.* the Fisher information, around the optimum and in the hyperplane locally satisfying the normalization constraint.

In addition to verifying the self-consistency of the learning procedure, we validated the model by confirming that the learned values of the parameters provide a good fit to the count pair statistics via the model's prediction of the conditional and marginal distributions of count pairs (see Fig. 4). While not explicitly required by the fitting procedure, the correspondence with data is good. For example, Fig. 4b shows that marginal, $P(n)$, inherits the power law of the clone frequency distribution, but exhibits deviations from this law at low count number consistent with the data and the prevalence of clones in the sample with putative frequencies less than $1/N_{\text{obs}}$. We also note that the model even does well at capturing the count statistics in one condition when there are no observed counts in the other condition (see Fig. 4a).

Inference of diversity and fraction observed

The learned models provide a clone frequency distribution, $\rho(f)$, that, together with the observed number of clones in one sample can be used to estimate mea-

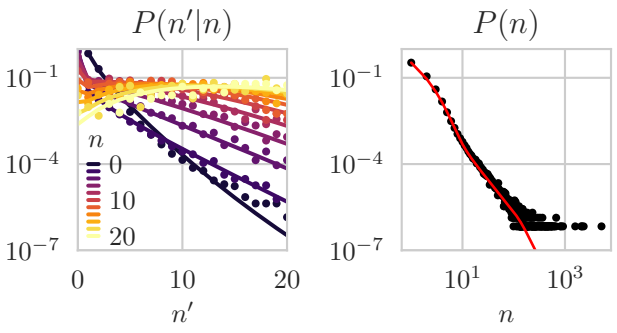

Figure 4. *Null model marginals and conditionals.* The marginal, $P(n|\theta_{\text{null}}) = \sum_{n'} P(n, n'|\theta_{\text{null}})$ (a), and conditional $P(n|n', \theta_{\text{null}}) = P(n, n'|\theta_{\text{null}})/P(n|\theta_{\text{null}})$ (b) distributions. Lines are analytic predictions of the learned model. Dots are estimated frequencies. (Donor S2, day 0/day 0 comparison).

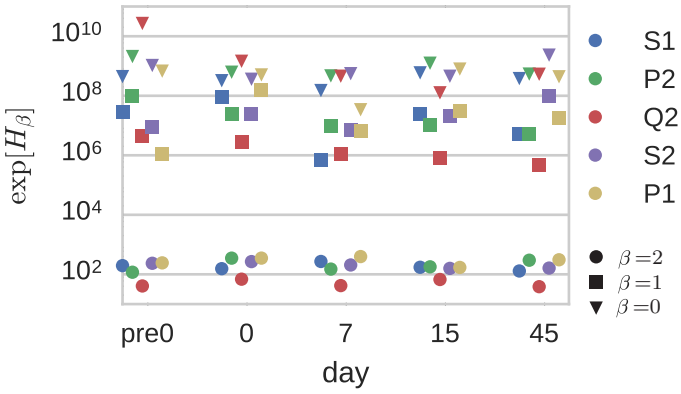

Figure 5. *Diversity estimates.* Shown are diversity estimates obtained from the Renyi entropies, H_β, of the inferred clone frequency distributions for $\beta = 0$ (estimated total number of clones, N), $\beta = 1$ (Shannon entropy) and $\beta = 2$ (Simpson index), across donors and days.

sures of diversity of the repertoire producing that sample (procedure outlined in Appendix D). In Fig. 5, we show the values, across donor and days, of the $\beta = 0, 1, 2$ Hill diversities as three distinct diversity metrics: the species richness, i.e. the total number of clones; the Shannon diversity as the effective number of clones obtained from the Shannon entropy of $\rho(f)$, if one assumes a uniform distribution instead; and the Simpson diversity, the expected number of shared clones between two realized repertoires. ...Thierry will discuss the variability here.... We note that diversity estimates obtained from observed species abundance are affected by statistical bias [8].

C. The differentially-expressed pair case: quantifying variation from baseline

An ensemble-level log-frequency fold-change model

Here, we introduce a frequency transformation and leverage the learned variation from the null model to define and learn a model for differential expression. We set the clone frequency of the test condition as $f' = fe^s$, where e^s is a multiplicative factor that we parametrize with a log-frequency fold-change, s (the natural base is used mathematical convenience). We incorporate s as an additional random variable in the variable chain of the model by providing a prior on s, $\rho(s)$. n and n' are now conditionally independent given f *and* s. The form of $\rho(s)$ depends on the application of the model. When used to describe a transient, selective perturbation relative to baseline, the form of $\rho(s)$ should contain a responding fraction with some effect size, alongside the non-responding component. The parameter values of these components can be chosen based on prior knowledge about typical sizes and fractions. Alternatively and best in the case of imprecise prior knowledge, $\rho(s)$ can be interpreted as part of the model and its parameter values learned directly from the data as was done with the prior on clone frequencies, $\rho(f)$ (the Empirical Bayes method). While the parameter were learned via gradient-based methods to maximize the likelihood, in Appendix F we give an example of an semi-analytic approach to finding the optimum using the expectation maximization algorithm.

The realistic range of values of the transformed frequencies in the differentially expressed condition are set by the properties of cell population dynamics. For example, finite division rates imply that the transformed frequencies should be bounded relative to those in the reference condition. In the absence of normalization of f', they can differ drastically, depending on the form of $\rho(s)$, so here we must impose normalization as we did with f. We satisfy the normalization constraint on $\vec{f}'$ by introducing into the given $\rho(s)$ an additional shift parameter, s_0, set to satisfy $Z^{\mathcal{D}}_{f'} = Z^{\mathcal{D}}_{f}$.

Model validation

Similar to the learning of the null model, here the corresponding marginal count pair distribution is

$$P(n, n'|\hat{\theta}_{\text{null}}, \theta_{\text{diff}}) = \iint P(n|f)P(n'|f' = fe^s)\rho(f)\rho(s)\text{d}s\text{d}f \ . \tag{8}$$

Since the null model results demonstrate a near universal measurement model, here we fix the measurement model parameters to the values from the fitted null model, $\hat{\theta}_{\text{null}}$,

for each donor. We also have collected the parameters of $\rho(s)$ (including the shift, s_0) into θ_{diff}. We then maximize this marginal likelihood over θ_{diff} subject to the $Z^{\mathcal{D}}_{f'} = Z^{\mathcal{D}}_{f}$ constraint to obtain the estimate, $\hat{\theta}_{\text{diff}}$.

Synthetic repertoire example To illustrate the behavior of the differential expression model, we present the case of a simple form for $\rho(s)$,

$$\rho_{s_0}(s) = \alpha \frac{1}{\bar{s}} e^{\frac{s_0 - s}{\bar{s}}} \Theta(s - s_0) + (1 - \alpha)\delta(s - s_0) \qquad (9)$$

(see Fig. 6a). This choice describes a differentially expressed condition arising from a stimulus to which some fraction, α, of the repertoire expands. Some of these clones respond strongly, most respond weakly, and all together with a characteristic effect size of log-frequency fold-change, $\bar{s}$, relative to non-responding clones in the remaining $1 - \alpha$ fraction of the repertoire. $s_0 < 0$ shifts the probability mass to lower values of s. We set s_0 using the equal average frequency normalization constraint, ensuring that the sum of frequencies in the differentially expression condition equals that in the reference condition (for details see Appendix E).

We inferred the parameters of the model from model-sampled synthetic (n, n') data for both small (mouse-like) and large (human-like) synthetic repertoires over a range of biologically plausible parameter values. In Fig. 6b, we show the parameter space of the inference of a mouse ($N = 10^6$) repertoire generated with $(\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$, showing the errors are distributed roughly optimally in $(\bar{s}, \alpha)$, i.e. with a covariance similar to the inverse of the corresponding Fisher information of one sample, as expected from the fact that the maximum likelihood estimator is efficient. The order of magnitude difference in axes ranges indicates that for this parametrization of $\rho(s)$, the Hessian of the likelihood is poorly conditioned. While second-order optimization methods are more efficient, this fact makes them ill-suited to our parametrization of the problem so we employ only first-order optimization methods for parameter learning.

For this particular mouse-like repertoire, the learned parameter estimates are imprecise due to the fact that the chosen value of $\alpha = 0.01$ and approximately 10^4 sampled clones means it is based on only tens to hundreds of responding clones. For human-sized repertoires, millions of clones are sampled making the inference much more precise (see Fig. 15). ...Need to add how many reads. Maybe exclude.

Posterior analysis

Once learned, the differential expression model provides for any observed clone the posterior distribution of log-frequency fold-change conditioned on the clone's observed count pair. It is calculated from the model by

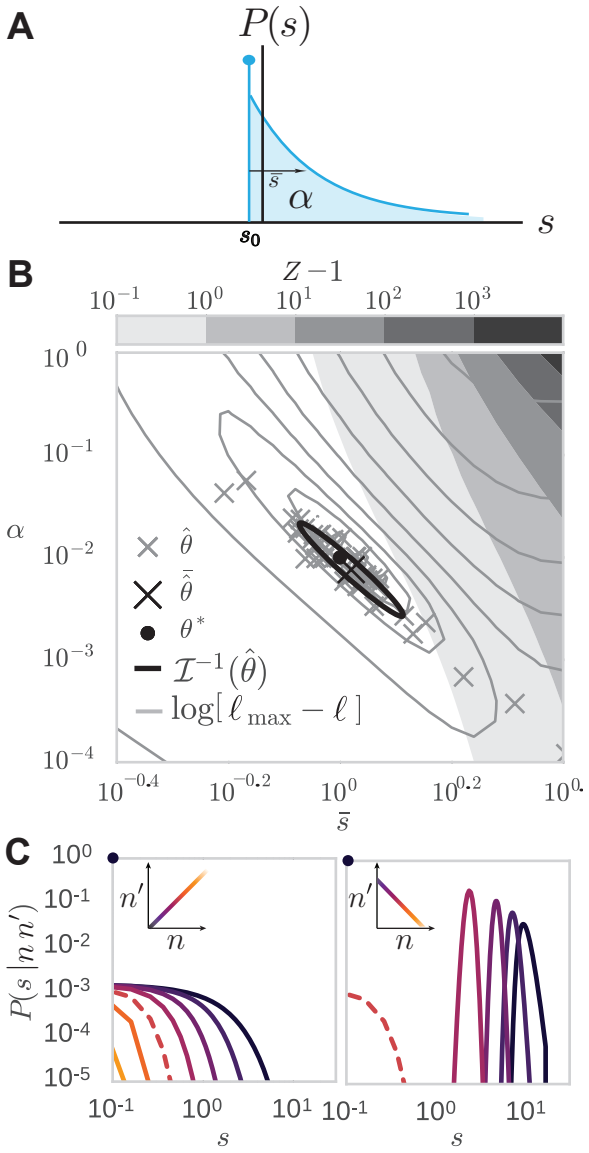

Figure 6. add $-s_0$ to x axis label*Inference on synthetic data.*(a) $\rho(s)$, Eq. 9, is parametrized by an effect size, $\bar{s}$, describing the expansion of the responding fraction, α of the repertoire. Expansion is relative to the functions center, s_0, which is fixed by the homeostatic constraint $\langle f \rangle = \langle f' \rangle$. (b) Inferring $\theta^* = (\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$ (black dot). Maximums of the log-likelihood, $\ell_{\max} = \ell(\hat{\theta}_r)$, for many realizations, $r = 1, \ldots, 50$, are given by gray crosses, with their average, $\bar{\hat{\theta}}$, shown as the black cross. The log-likelihood, $\ell(\theta)$, for one realization is shown over logarithmically-spaced gray contours decreasing from the maximum, $\ell_{\max}$. The inverse Fisher information, $\mathcal{I}^{-1}$, for a realization is shown as the black-lined ellipse centered around its maximum, $\bar{\hat{\theta}}$, provides a lower bound to the variance of our ML estimate. The gray scale contours increasing to the upper-right denote the excess in the used normalization, $Z = e^{s_0}$, above 1. (c) Posteriors of the learned model, $P(s|n, n')$ over pairs (n, n') for $n' = n$, with n varying over a logarithmically-spaced set of counts (left), and for n' given by the reverse order of this set (right). The black dot in both plots denotes the $1 - \alpha$ non-responding component, $\delta(s - s_0)$. (Parameters: $N = 10^6, \epsilon = 10^{-2}$.)

marginalizing f, and using Bayes' rule,

$$P(s|n,n') = \frac{P(n,n'|s)\rho(s)}{P(n,n')} \ , \qquad (10)$$

where $P(n,n'|s) = \int P(n|f)P(n'|f'=fe^s)\rho(f)\mathrm{d}f$.

To illustrate the wide range of possible posterior shapes as a function of the observed count pairs, in Fig. 6c, we show how the mass in the posteriors moves as we move in orthogonal directions in the space of observed count pair, (n,n'). In particular, we see for example that the width of the posterior narrows when counts are both large, and that the model ascribes a fold-change of s_0 to clones with $n' \lessapprox n$.

Measured repertoire example

Here we run the above inference on actual sequences obtained from human blood samples across yellow fever vaccination. To guide the choice of prior for s, we plotted the histograms of the naive log-frequency fold-change $\ln n'/n$ (for $n', n > 0$) (Fig. 16). While these statistics are skewed relative to the unobserved statistics of $\ln f'/f$ by neglecting pairs with $n' = 0$ or $n = 0$, and by including the additional variability in the acquisition process, they nevertheless provide qualitative information about the underlying distribution of log-frequency fold-change. The histograms are qualitatively described as Gaussian distributions slightly offset from zero with tails between Gaussian and exponential form, depending on the day and donor (not shown). The Gaussian peaks are at least partially generated from the acquisition process. The tails are less affected by measurement noise, however, so the observed exponential form suggests right and left exponential tails for $\rho(s)$. Thus, here we add a contracting component to the $\rho(s)$ analyzed in the synthetic case above (Eq. 9). While the contraction achieved by $s_0 < 0$ affects all clones, the contraction due to this new component only affects a fraction of clones. This might arise from the fluctuations of clone frequency dynamics or ...more explanation on two kinds of contraction?... . In principle, inference using this form of $\rho(s)$ then reveals whether contraction is more homeostatic or clone-specific in nature.

Despite this added component, we can keep the number of parameters the same by setting its scale parameter equal to that of the expanding component. Since the contraction plays little role in normalization and other qualitative model features, this choice differs little from the case where the contraction scale parameter is specifically learned. Thus, we set

$$\rho_{s_0}(s) = \alpha \frac{1}{2\bar{s}} e^{\frac{|s_0 - s|}{\bar{s}}} + (1-\alpha)\delta(s-s_0) \ . \qquad (11)$$

We use this prior and day 0 as the reference and plot the resulting maximum likelihood parameter estimates for different test days in Fig. 7 and also across the 4 combinations of pair replicates from the reference and test days ($1 \to 1$, $1 \to 2$, $2 \to 1$, and $2 \to 2$).

There are a few remarks about how the learned values are dispersed. First, the day-0 replicates across donors gives an effect size less than the discretization (0.1), and so is effectively a Dirac delta function as expected. ...don't know why donor S1's values are in between.... More generally, except for the day-0 comparison for the reason above, the learned values across donors cluster along a ridge of high likelihoods, as did the synthetic estimates (Fig. 6). Consistent with the synthetic data, the uncertainty of these estimates, again quantified via the Fisher information (not shown), is relatively small on account of the large number ($\approx 10^6$) of observed clones, and orients along the ridge as expected. Replicate variability is much larger, but still confined to the ridge. The range of estimates for the effect size is relatively narrow compared with that for the fraction of responding clones that ranges over orders of magnitude across donors and even across replicates.

Imprecise estimation of ensemble-level differential expression

Our model incorporates both observed and unobserved parts of the repertoire. Consequently, it provides estimates for global properties of the response, namely the fraction of the total number of clones, α, that responds to the perturbation. Publicly available, state-of-the-art algorithms for differential expression also suggest an estimate for α from the outputted number of significantly differentially expressed clones in conjunction with some normalization. Normalizing by the number of observed clones (even when adjusted for the presence of large clones as is done in edgeR's TMM approach) is likely to overestimate α for the same reason: the sample preferentially excludes small, non-responding clones. To get a sense for the magnitude of these inaccuracies, we can apply one of these other algorithms (EdgeR) to samples from our repertoire model where we know the ground truth. As expected, such estimates are off by orders of magnitude ($\alpha^*_{EdgeR} = log - high$; actual $\alpha = 0.01$). Our re-inference of course precisely learns α ($\alpha^* = low - high$) since it employs the model used to generate the sampled data.

Regarding real data, even in our model-based approach, where we have learned the individual-specific variability, the method only constrains α to a range of values over a region in the parameter space highlighted as being consistent with the data. We conclude that, at least with the prior we have chosen, the data does not allow for a precise estimate of the responding fraction. Estimates obtained via more coarser methods should thus be interpreted with caution.

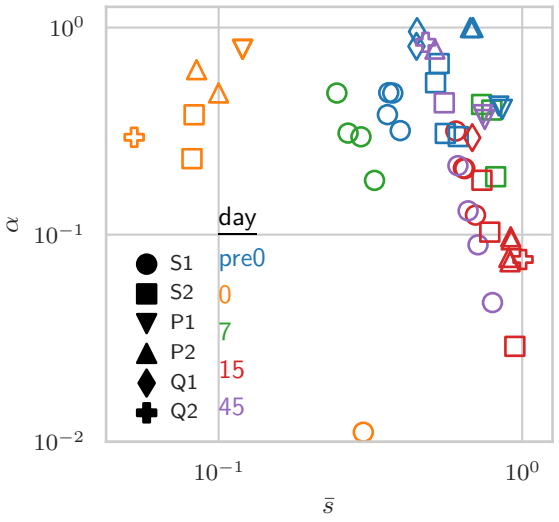

Figure 7. *Inference on actual data.* The optimal values of *alpha* and $\bar{s}$ across donors and days still missing values!. The background heat map is a spatial correlation of the resulting list of significantly expanded clones.

D. Identifying responding clones

A learned model of repertoire clone size statistics can be used to infer beliefs about differential expression given molecule count pairs of single observed clones. In this section we first describe the structure of these posteriors as a function of pair count, and then provide a threshold-based procedure for identifying clones with significant differential expression, i.e. significant log fold-change.

Measured posterior distributions of log-frequency fold-change

We can explain the shape of the posteriors by breaking up the model components into three interpretable factors: $P(n|f)\rho(f)$, which depends only on f, $\rho(s)$, which depends only on s, and the remainder depending on both f and s, $P(n'|f' = fe^s)$. $P(n|f)\rho(f)$ contributes an exponential cutoff in $\log f$ near n. $P(n'|f')$ contributes a similar cutoff in f' near n'. $\log f$ and s tradeoff in setting the value of f'. Thus for a given s, the cutoff in f shifts to lower values for larger values of s. For fixed f, the corresponding cutoff in s shifts to larger values for smaller values of f, until a maximum cutoff in s is reached for $f = f_{\min}$. This cutoff in s can more strongly bound the posterior as we consider (n, n') pairs with smaller n/n'. However, this large s cutoff can be gated by $\rho(s)$, depending on the form of its expansion ($s > 0$) tail. In fact, the form of the decay of the expansion component of $\rho(s)$ interacts with the decay of $\rho(f)$. Indeed, for power law $\rho(f)$, $\log f$ is distributed exponentially, so that $\log f$ and s tradeoff additively, not only in determining f' but also in determining $\rho(f'|f)$. Which distribution, $\rho(f)$ or $\rho(s)$, dominates the shape of the posteriors then depends on the relative magnitude of their scale parameters.

We computed the posteriors over all clones. For clones observed with $(0, n')$ with n' large, the posteriors exhibit a large s mode whose form has a maximum at a high or low positive s value depending on a competition between the two priors controlled by α_f that preferences high s and $\bar{s}$ that preferences low s (see Fig. 8). The average posterior reflects this, through its deviation from the prior in the large, positive s regime.

The structure of the contribution of singleton clones (the vast majority of all clones in the sample) reflect the balance between of the power of the power-law form of clone frequency density preferencing near maximal expansion and our conservative choice of prior preferencing expansion of a characteristic size. The uncertainty in the inferred expansion obtained from these posteriors is dispersed roughly uniform over the range of expansion for singleton clones appearing in the differentially expressed condition.

Figure 8. *Competition between α_f and $\bar{s}$ in shaping the posteriors, $\rho(s|n, n')$.*insert figure here

How do these posterior beliefs about log fold-change compare to the naive estimate, $s_{\text{naive}} = \ln n'/n$É We show this estimate in Fig. 17) as a function of a natural estimate from our model, the posterior median. s_{naive} reliably matches the median when both n and n' are large across a range of values of the median. This match demonstrates that our learning procedure identified parameters that respect the reliability of the fold change observed of large clones. The estimator increasingly over-estimates in magnitude where both fold change and counts are small. Indeed, where n and n' are small, a large fold change is more likely due to stochastic fluctuations in realization. Our posterior inference procedure averages out these fluctuations and so in principle gives a more accurate estimate. We confirmed this using synthetic data (not shown).

Identifying responding clones

A significance criterion on the inferred posteriors of log-frequency fold-change can be used to identify candidate clones that participate significantly in a response. In analogy with p-values, we use the posterior probability corresponding to the null hypothesis that they are not expanded, $P(s \leq 0|n, n')$ and set a threshold of significance $P_{\text{null}} \leq 0.025$. In Fig. 10, we show the results as a function of posterior effect size. The threshold in confidence sets a threshold in molecular count pair space (n, n'). The structure of the data in the plot recapitulates the structure of the posterior beliefs. Focusing

on expanding clones, the outer most line corresponds to clones with $(0, n')$ count pairs with $n' = 1, 2, \ldots$. The next furthest is for $n = 1$, where the effect size at which the response passes the threshold criteria is less. More generally, the larger n, the smaller is the lower bound on n' required for significance. The situation for significant contraction is analogous. Mapping the threshold back in (n, n') space gives a simple prescription for selection from the count pair histogram (Fig. 10 inset).

Given the breadth of learned values for the parameters of $\rho(s)$, an natural question is how robust is the above significance procedure to that variation. In Fig. 11, we show the list overlap as a function of $(\bar{s}, \alpha)$. There is a ridge of high overlap values mirroring the high ridge of likelihoods on which we find most learned parameters. Showing the replicates for one donor shows that indeed these different values for $\bar{s}$ and especially α lead to virtually identical lists of candidates for response, confirming the robustness of the latter.

A final result will go here that uses the table to say something interesting. Candidates: correlation between precursor f and s_{median}. Compare selected clones with control from all clones. Also see how persistent clones (appearing on all days) participate in the response.

Figure 9. *Properties of significantly expanded clones.* The correlation between f and s_{median}.

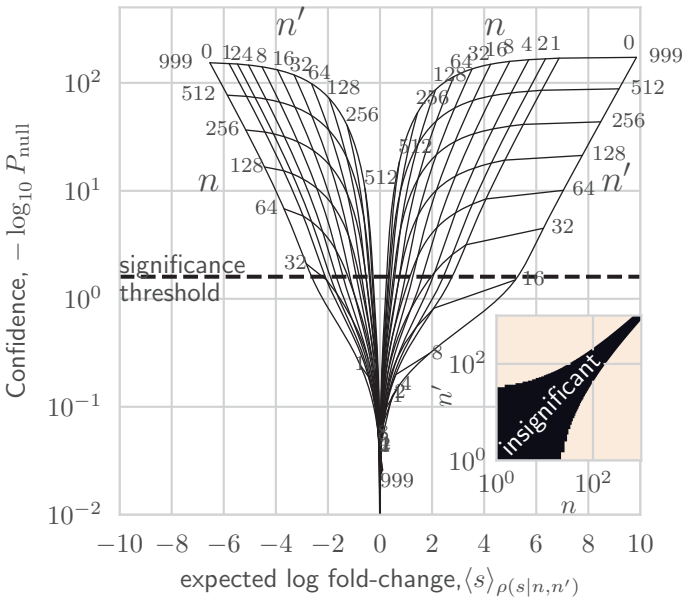

Figure 10. *Hummingbird plot of confidence of response versus average effect size.* A significance threshold is placed according to $P_{\text{null}} = 0.025$, where $P_{\text{null}} = P(s \leq 0)$ for expansion and $P_{\text{null}} = P(s \geq 0)$ for contraction. Inset shows the same threshold hold in (n, n')-space.

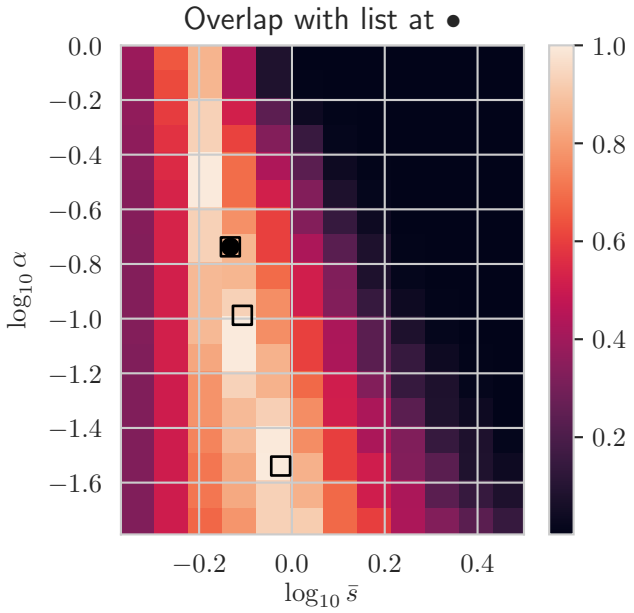

Figure 11. *Overlap in list of significantly expanded clones.* The optimal values of α and $\bar{s}$ for donor S2 and day-0 day-15 comparison for 3 replicates (square markers). The background heat map is the list overlap $\left|\ell_{(\bar{s},\alpha)} \cap \ell^{\text{ref}}_{(\bar{s},\alpha)}\right| / \left|\ell_{(\bar{s},\alpha)} \cup \ell^{\text{ref}}_{(\bar{s},\alpha)}\right|$ with the reference given by the list obtained using values of $\bar{s}$ and α at the black dot.

DISCUSSION

this outline is out-dated. Didn't get chance to work on this yet.

- Procedure Summary:
  - used replicates to determine experimental clone size variability
  - inferred repertoire change distributions
  - used to determine significantly expanded clones
  - validated using functional assay
- Natural variation results and discussion:
  - universal same-day variation. Implications...
  - Data tightly constrains power law frequency. Implications...
- diffexpr results and discussion:
  - data strongly constrains prior expansion, not contraction or responding fraction. Implications...

  - Shift constraint and the relevance of homeostatis.

- application results and discussion:
  - posterior sensitivity to balance between α_f (prior for maximum expansion) and $\bar{s}$ (prior for characteristic expansion) in $(0, n)$ and $(n, 0)$ pairs.
  - sensitivity of resulting tables. Note validation in Misha's paper.

- Clinical use (reference Misha paper)

- drawbacks of approach: need replicate data, ...

ACKNOWLEDGEMENTS

... would like to acknowledge discussions with ... MPT would like to thank M. Pogorelly for providing the R code used to obtain the EdgeR estimates in [6]. This work was supported by ...

AUTHOR CONTRIBUTIONS

M.P.T., A.W., T.M. ...

ADDITIONAL INFORMATION

The authors declare no competing financial interests.

APPENDICES

Appendix A: Normalization

Here we derive the condition for which the normalization in the joint density is implicitly satisfied. The normalization constant of the joint density is

$$\mathcal{Z} = \int_{f_{\min}}^{1} \cdots \int_{f_{\min}}^{1} \prod_{i=1}^{N} \rho(f_i)\delta(Z-1)\mathrm{d}^N\vec{f}\,, \tag{A1}$$

with $\delta(Z-1)$ being the only factor preventing factorization and explicit normalization. Writing the delta function in its Fourier representation factorizes the single constraint on $\vec{f}$ into N Lagrange multipliers, one for each f_i,

$$\delta(Z-1) = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{\mu(Z-1)} \tag{A2}$$

$$= \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} e^{\mu f_i}\,. \tag{A3}$$

Crucially, the multi-clone integral in Eq. A1 over $\vec{f}$ then factorizes. Exchanging the order of the integrations and omitting the clone subscript without loss of generality,

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} \langle e^{\mu f}\rangle\,, \tag{A4}$$

with $\langle e^{\mu f}\rangle = \int_{f_{\min}}^{1} \rho(f) e^{\mu f}\mathrm{d}f$. Now define the large deviation function, $I(\mu) := -\frac{\mu}{N} + \log\langle e^{\mu f}\rangle$, so that

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-NI(\mu)}\,. \tag{A5}$$

Note that $I(0) = 0$. With N large, this integral is well-approximated by the integrand's value at its saddle point, located at μ^* satisfying $I'(\mu^*) = 0$. Evaluating the latter gives

$$\frac{1}{N} = \frac{\langle f e^{\mu^* f}\rangle}{\langle e^{\mu^* f}\rangle}\,. \tag{A6}$$

If the left-hand side is equal to $\langle f\rangle$, the equality holds only for $\mu^* = 0$ since expectations of products of correlated random variables are not generally products of their expectations. In this case, we see from Eq. A5 that $\mathcal{Z} = 1$, and so the constraint $N\langle f\rangle = 1$ imposes normalization.

Appendix B: Null model sampling

The procedure for null model sampling is summarized as (1) fix main model parameters, (2) solve for remaining parameters using the normalization constraint, $N\langle f\rangle = 1$, and (3) starting with frequencies, sample and use to specify the distribution of the next random variable in the chain.

In detail, we first fix:

- the model parameters (α, M, a, γ), excluding $f_{\min}$. Separate M, a, and γ values could be defined for the reference and test condition, respectively. The empirical $P(n, n')$ for replicate data was found to be highly symmetric in n and n' across donors, however, supporting the assumption of a single acquisition model and so we neglect this complication.

- the desired size of the full repertoire, N.

- the sequencing efficiency (total sample reads/total sample cells), ϵ. From this we get the effective total sample reads, $N_{\text{reads}}^{\text{eff}} = \epsilon M$, that converts a clone's frequency to the average number of cells it appears with in the sample. (We could in fact define two sequencing efficiencies, one for each replicate, leading to different effective total number of reads in each replicate). Note that the actual sampled number of reads is stochastic and so will differ from this fixed value.

We then solve for remaining parameters. Specifically, $f_{\min}$ is fixed by the constraint that the average sum of all frequencies, under the assumption that their distribution factorizes, is unity:

$$N\langle f\rangle_{\rho(f)} = 1 \tag{B1}$$

This completes the parameter specification.

We then sample from the corresponding chain of random variables. Sampling the chain of random variables of the null model can be performed efficiently by only sampling the $N_{\text{obs}} = N(1 - P(0,0))$ observed clones. This is done separately for each replicate, once conditioned on whether or not the other count is zero. Samples with 0 molecule counts can in principle be produced with any number of cells, so cell counts must be marginalized when implementing this constraint. We thus used the conditional probability distributions $P(n|f) = \sum_m P(n|m)P(m|f)$ with $m, n = 0, 1, \ldots$. $P(n'|f)$ is defined similarly. Note that these two conditional distributions differ only in their average number of UMI per cell, N_{reads}/M, due to their differing number of observed total number of molecules, N_{reads}. Together with $\rho(f)$, these distributions form the full joint distribution, which is conditioned on the clone appearing in the sample, i.e. $n + n' > 0$ (denoted $\mathcal{O}$),

$$P(n, n', f|\mathcal{O}) = \frac{P(n|f)P(n'|f)\rho(f)}{1 - \int \mathrm{d}f\rho(f)\mathrm{d}fP(n=0|f)P(n'=0|f)}\ , \tag{B2}$$

with the renormalization accounting for the fact that $(n, n') = (0, 0)$ is excluded. The 3 quadrants having a finite count for at least one replicate are denoted q_{x0}, q_{0x}, and q_{xx}, respectively. Their respective weights are

$$P(q_{x0}|\mathcal{O}) = \sum_{n>0} \int \mathrm{d}f P(n, n' = 0, f|\mathcal{O})\ , \tag{B3}$$

$$P(q_{0x}|\mathcal{O}) = \sum_{n'>0} \int \mathrm{d}f P(n = 0, n', f|\mathcal{O})\ , \tag{B4}$$

$$P(q_{xx}|\mathcal{O}) = \sum_{\substack{n>0,\\ n'>0}} \int \mathrm{d}f P(n, n', f|\mathcal{O}). \tag{B5}$$

Conditioning on $\mathcal{O}$ ensures normalization, $P(q_{x0}|\mathcal{O}) + P(q_{0x}|\mathcal{O}) + P(q_{xx}|\mathcal{O}) = 1$. Each sampled clone falls in one the three regions according to these probabilities. Their clone frequencies are then drawn conditioned on the respective region,

$$P(f|q_{x0}) = \sum_{n>0} P(n, n' = 0, f|\mathcal{O})/P(q_{x0}|\mathcal{O})\ , \tag{B6}$$

$$P(f|q_{0x}) = \sum_{n'>0} P(n = 0, n', f|\mathcal{O})/P(q_{0x}|\mathcal{O})\ , \tag{B7}$$

$$P(f|q_{xx}) = \sum_{\substack{n>0,\\ n'>0}} P(n, n', f|\mathcal{O})/P(q_{xx}|\mathcal{O}). \tag{B8}$$

Using the sampled frequency, a pair of molecule counts for the three quadrants are then sampled as $(n, 0)$, $(0, n')$, and (n, n'), respectively, with n and n' drawn from the renormalized, finite-count domain of the conditional distributions, $P(n|f, n > 0)$.

Using this sampling procedure we demonstrate the validity of the null model and its inference by sampling across the observed range of parameters and reinferring their values (See Fig. 14).

Appendix C: Differential model sampling

Since the differential expression model involves expansion and contraction in the test condition, some normalization in this condition is needed such that it produces roughly the same total number of cells as those in the reference condition, consistent with the observed data. One approach (the one taken below) is to normalize at the level of clone frequencies. Here, we instead perform the inefficient but more straightforward procedure of sampling all N clones and discarding those clones for which $(n, n') = (0, 0)$. A slight difference in the two procedures is that N_{obs} is fixed in the former, while is stochastic in the latter.

Direct sampling

The frequencies of the first condition, f_i, are sampled from $\rho(f)$ until they sum to 1 (i.e. until before they surpass 1, with a final frequency added that takes the sum exactly to 1). An equal number of log-frequency fold-changes, s_i, are sampled from $\rho(s)$. The normalized frequencies of the second condition are then $f'_i = f_i e^{s_i} / \sum_j f_j e^{s_j}$. Counts from the two conditions are then sampled from $P(n|f)$ and $P(n'|f')$, respectively. Unobserved clones, i.e. those with $(n, n') = (0, 0)$, are then discarded.

Appendix D: Obtaining diversity estimates from the clone frequency density

For a set of clone frequencies, $\{f_i\}_{i=1}^N$, for a set of clones, the Hill family of diversities are obtained from the Renyi entropies, as $D_\beta = \exp H_\beta$, with $H_\beta = \frac{1}{1-\beta} \ln \left[\sum_{i=1}^N f_i^\beta\right]$. We use $\rho(f)$ to compute their ensemble averages over f, again under the assumption that the joint distribution of frequencies factorizes. We obtain an estimate for $D_0 = N$ using the model-derived expression, $N_{\text{samp}} + P(n = 0)N = N$, where N_{samp} is the number of clones observed in one sample, and $P(n = 0) = \int_{f_{\min}}^1 P(n = 0|f)\rho(f)\mathrm{d}f$. For $\beta = 1$, we

compute $\exp(N\langle -f\log f\rangle_{\rho(f)})$ and for $\beta = 2$, we use $1/\left(N\langle f^2\rangle_{\rho(f)}\right)$.

Appendix E: Deriving update for shift, s_0

The constraint of equal repertoire size, $Z^{\mathcal{D}}_{f'} = Z^{\mathcal{D}}_{f}$ (Eq. 6), can be satisfied with a suitable choice of the shift parameter, s_0, in the prior for differential expression, $\rho_{s_0}(s)$, namely $s_0 = -\ln Z^{\mathcal{D}}_{f'}/Z^{\mathcal{D}}_{f}$. The latter arises from the coordinate transformation $s \leftarrow \Delta s + s_0$ that maps $\rho_{s_0}(s)$ to $\rho_0(\Delta s)$, and adds a factor of e^{s_0} to all terms of $Z^{\mathcal{D}}_{f'}$.

Appendix F: Prior solvable via expectation maximization

For learning the parameters of $\rho(s)$, we performed a grid search, refined by an iterative, gradient-based search to obtain the maximum likelihood. In a more formal approach, here we employ expectation maximization (EM) to obtain the optimal parameter estimates from the data by calculating the expected log likelihood over the posterior and then maximizing with respect to the parameters. In practise, we first perform the latter analytically and then evaluate the former numerically. We choose a symmetric exponential as a tractable prior for this purpose:

$$\rho_{\bar{s}}(s) = e^{-|s|/\bar{s}}/2\bar{s} \tag{F1}$$

with $s \in \mathbb{R}$, $\bar{s} > 0$, and no shift. The expected value of the log likelihood function, often called the Q-function in EM literature, is

$$Q(\bar{s}|\bar{s}') = \sum_{i=1}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n_i, n_i', \bar{s}') \log\left[P(n_i, n_i', s|\bar{s})\right] \ , \tag{F2}$$

where $\bar{s}'$ is the current estimate. Maximizing Q with respect to $\bar{s}$ is relatively simple since $\bar{s}$ appears only in $\rho_{\bar{s}}(s)$ which is a factor in $P(n, n', s|\bar{s})$. For each s,

$$\frac{\partial \log\left[\rho_{\bar{s}}(s))\right]}{\partial \bar{s}} = \frac{1}{\rho_{\bar{s}}(s)} \frac{\partial \rho_{\bar{s}}(s)}{\partial \bar{s}} \tag{F3}$$

$$= \frac{|s| - \bar{s}}{\bar{s}^2} \ , \tag{F4}$$

so that $\frac{\partial Q(\bar{s}|\bar{s}')}{\partial \bar{s}} = \sum_{i=1}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n_i, n_i', \bar{s}') \frac{\partial \log[\rho_{\bar{s}}(s))]}{\partial \bar{s}} = 0$ implies

$$\sum_{i=1}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n_i, n_i', \bar{s}') \frac{|s| - \bar{s}^*}{\bar{s}^{*2}} = 0 \tag{F5}$$

so that $\bar{s}^* = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} \bar{s}_{(n_i, n_i')}$, where

$$\bar{s}_{(n,n')} = \int_{-\infty}^{\infty} \mathrm{d}s |s| \rho(s|n, n', \bar{s}'). \tag{F6}$$

The latter integral is computed numerically from the model using $\rho(s|n, n', \bar{s}') = P(n, n', s|\bar{s}')/\int_{-\infty}^{\infty} P(n, n', s|\bar{s}')\mathrm{d}s$. Q is maximized at $\bar{s} = \bar{s}^*$ since $\frac{\partial^2 \log[\rho_{\bar{s}}(s))]}{\partial \bar{s}^2}\Big|_{\bar{s}=\bar{s}^*} = -\bar{s}^{*-2} < 0$. Thus, we update $\rho_{\bar{s}}(s)$ with

$$\rho_{\bar{s}}(s) \leftarrow \rho_{\bar{s}^*}(s). \tag{F7}$$

The number of updates typically required for convergence was small.

Appendix G: Identifying responding clones

In analogy with p-values, we used the posterior probability corresponding to the null hypothesis that they are not expanded, $p = \rho(s \le s_{thr}|n, n', \theta_n, \theta_s)$ to rank the clones by the significance of their expansion, using a threshold of $p < 0.025$ and a threshold effect size of s_{thr}.

Figure 12. *Supp. Fig: Not Done!Approximate equivalence of* $N\langle f\rangle = 1$ *and* $\mathcal{Z}^{\mathcal{D}}_{f}$.

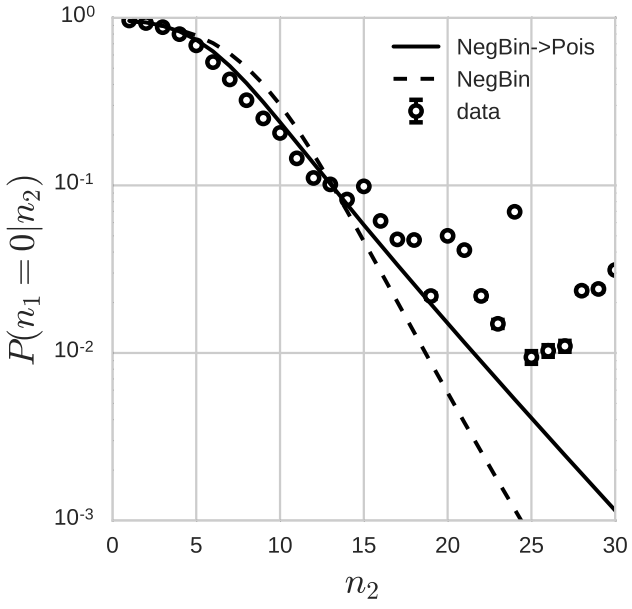

Figure 13. *Supp. Fig: Two-step model captures tail better than one-step model.*

[1] Jennifer Benichou and Yoram Louzoun, "Rep-Seq : uncovering the immunological repertoire through next-generation sequencing," Immunology , 183–191 (2011).

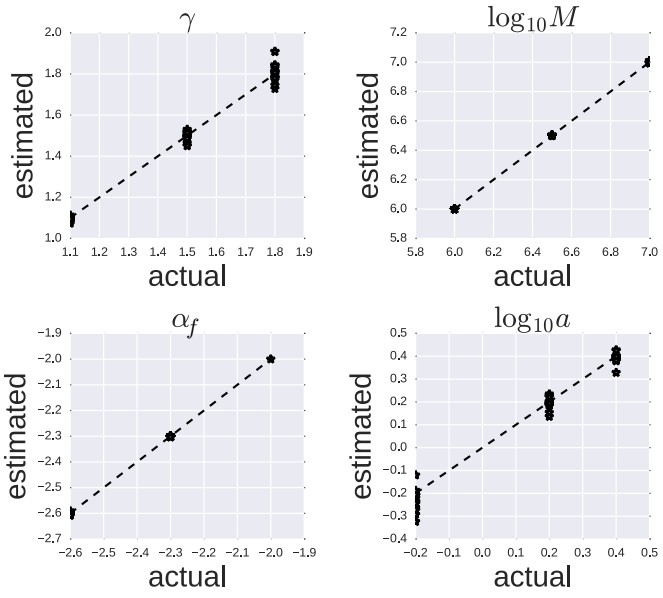

Figure 14. *Reinferring null model parameters.* Shown are the actual and estimated values of the null model parameters used to validate the null model inference procedure over the range exhibited by the data. A 3x3x3x3 grid of points were sampled and results collapsed over each parameter axis. f_{min} was fixed to satisfy the normalization constraint.

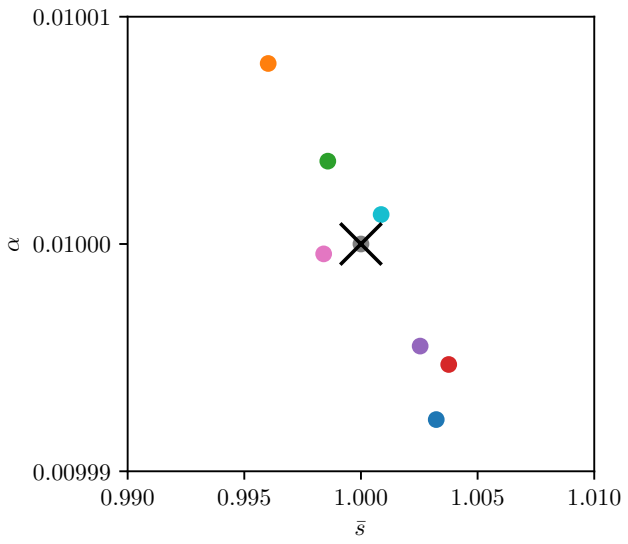

Figure 15. *Supp. Fig: Precise, self-consistent reinference of differential expression model for human-sized repertoire.*

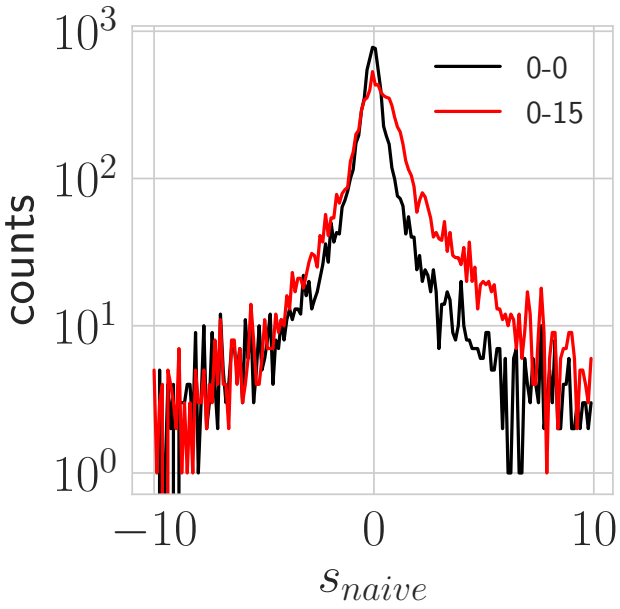

Figure 16. *Supp. Fig: Empirical histograms of naive log-frequency fold-change for day-0/day-0 and day-0/day-15 pair comparisons.*

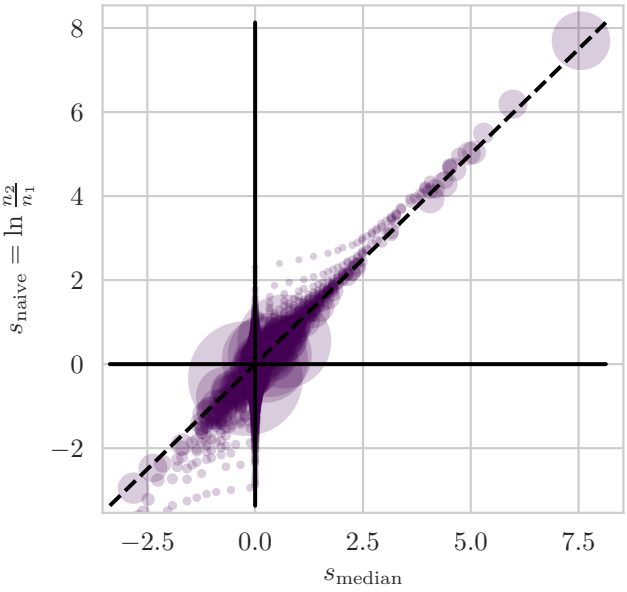

Figure 17. *Supp. Fig: Summary statistics of log-frequency fold-change posterior distributions.* (a) summary statistics. (b) comparing the posterior median log-frequency fold-change and the naive estimate, $\log n'/n$ (across clones with $n, n' > 0$).

[2] Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M Krams, Christina Pettus, Nikhil Haas, Cecilia S Lindestam Arlehamn, Alessandro Sette, Scott D Boyd, Thomas J Scriba, Olivia M Martinez, and Mark M Davis, "Identifying specificity groups in the T cell receptor repertoire," Nature Publishing Group **547**, 94–98 (2017).

[3] Nathaniel D Chu, Haixin Sarah Bi, Ryan O Emerson, Anna M Sherwood, Michael E Birnbaum, Harlan S

Robins, and Eric J Alm, "Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors," BMC Immunology **20**, 1–12 (2019).

[4] Michael I Love, Wolfgang Huber, and Simon Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology **15**, 550 (2014).

[5] Mark D Robinson and Gordon K Smyth, "Small-sample estimation of negative binomial dispersion , with applications to SAGE data," , 321–332 (2008).

[6] Mikhail V Pogorelyy, Anastasia A Minervina, Maximilian Puelma Touzel, Anastasiia L Sycheva, Ekaterina A Komech, Elena I Kovalenko, Galina G Karganova, Evgeniy S Egorov, Alexander Yu. Komkov, Dmitriy M Chudakov, Ilgar Z Mamedov, Thierry Mora, Aleksandra M Walczak, and Yuri B Lebedev, "Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins," Proceedings of the National Academy of Sciences **115**, 12704–12709 (2018).

[7] Katharine Best, Theres Oakes, James M Heather, John Shawe-Taylor, and Benny Chain, "Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding." Scientific reports **5**, 14629 (2015).

[8] Thierry Mora and Aleksandra M Walczak, "Quantifying lymphocyte receptor diversity," in *Systems Immunology: An Introduction to Modeling Methods for Scientists*, edited by J. Das and C. Jayaprakash (CRC Press, 2018).

[9] Thierry Mora and Aleksandra M Walczak, "How many different clonotypes do immune repertoires contain? ´," (2019), arXiv:arXiv:1907.08230v1.

[10] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-yeun Lee, and Richard A Olshen, "Diversity and clonal selection in the human T-cell repertoire," Proceedings of the National Academy of Sciences **111** (2014), 10.1073/pnas.1409155111.