Inferring the immune response from repertoire sequencing

Maximilian Puelma Touzel
Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France and Mila, Université de Montréal, Montreal, Canada

Aleksandra M. Walczak* and Thierry Mora*
Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France

High-throughput sequencing of B- and T-cell receptors makes it possible to track immune repertoires across time, in different tissues, and in acute and chronic diseases or in healthy individuals. However, quantitative comparison between repertoires is confounded by variability in the read count of each receptor clonotype due to sampling, library preparation, and expression noise. Here, we present a general Bayesian approach to disentangle repertoire variations from these stochastic effects. Using replicate experiments, we first show how to learn the natural variability of read counts by inferring the distributions of clone sizes as well as an explicit noise model relating true frequencies of clones to their read count. We then use that null model as a baseline to infer a model of clonal expansion from two repertoire time points taken before and after an immune challenge. Applying our approach to yellow fever vaccination as a model of acute infection in humans, we identify candidate clones participating in the response.

High-throughput repertoire sequencing (RepSeq) of T-cell receptors (TCR) and immunoglobulins (Ig) promises to offer a detailed and comprehensive view of the properties of the adaptive immune system in health and disease [1–4]. In particular, RepSeq makes it possible to study the collective dynamics of immune repertoires. Longitudinal repertoire sequencing has been used to study repertoire dynamics following immunization [5–13], or to track minimal residual disease in a lymphoid malignancies [14]. This approach has demonstrated the potential to identify particular receptors involved in the immune response, offering an alternative to in vitro binding or functional assays [15, 16] or to RepSeq studies of cohorts with a common condition [17]. The natural temporal evolution of TCR repertoires has also been studied as a function of age across healthy individuals [18] or across time in single healthy individuals [19]. While these experiments yield a wealth of data, it is not always clear how to process and analyse them, or to link them to existing models of lymphocyte dynamics and proliferation [20–23]. As more londitudinal repertoire data are being produced, there is an increasing need to develop models and methods for extracting the relevant dynamical parameters from them.

A major difficulty with interpreting repertoire sequencing data is the inherent noise in the measurement of read counts associated to each TCR or Ig clonotype. While these read counts give a good representation the actual abundance of cells with the corresponding clonotype, at least for T-cells, library preparation and sequencing protocols create variability in these frequencies across experiments [24]. This experimental variability needs to be modeled and controled to be able to distinguish it from real variations in the repertoire. This problem is similar to that of differential expression analysis in RNAseq studies, to which a number of solutions have been proposed [25–28]. However, RepSeq differs from RNAseq in fundamental ways. In RepSeq, each class is a clonotype that is expressed by distinct cell clones, while in RNAseq each class is a gene expressed by all cells. In addition, in most cases RepSeq is based on relatively small samples of the entire repertoire, and only captures a small fraction of its full diversity, making extrapolation necessary.

Here, we take a Bayesian approach to uncover lymphocyte clonal dynamics. Our approach is based on a generative model that incorporates known features of the clonal frequency distribution and statistical properties of the sequencing process, with parameters inferred directly from RepSeq data. Once the model is learned, it can be used to study the expansion dynamics of particular clones, with confidence levels calculated from the model. We illustrate and apply our approach to longitudinal TCR repertoire datasets taken before and after yellow fever vaccination previously described in [12]. Here, we present our general inference method in detail, test its validity on synthetic data, and discuss its results for the immunization data. We also use the inferred models to estimate diversity and clone size statistics from data.

Results

Modeling repertoire variation

To describe the stochastic dynamics of an individual clone, we define a probabilistic rule relating its frequency f' at time t' to its frequency f at an earlier time t: $G(f', t'|f, t)$. In this paper, t and t' will be pre- and post-vaccination time points, but more general cases may be

*Corresponding authors. These authors contributed equally.

considered. It is also useful to define the probability distribution for the clone frequency at time t, $\rho(f)$ (Fig. 1A).

The true frequencies of clones are not directly accessible experimentally. Instead, sequencing experiments give us number of reads for each clonotypes, n, which is a noisy function of the true frequency f, described by the conditional probability $P(n|f)$ (Fig. 1B). Correcting for this noise to uncover the dynamics of clones is essential and is a central focus of this paper.

Our method proceeds in two inference steps, followed by a prediction step. First, using same-day replicates at time t, we jointly learn the characteristics of the frequency distribution $\rho(f)$ (Fig. 1A) and the noise model $P(n|f)$ (Fig. 1B). Second, by comparing repertoires between two time points t and t', we infer the parameters of the evolution operator $G(f', t'|f, t)$, using the noise model and frequency distribution learned in the first step (Fig. 1C). Once these two inferences have been performed, the dynamics of individual clones can be estimated by Bayesian posterior inference. These steps are described in the remaining Results sections. In the rest of this section, we define and motivate the classes of model that we chose to parametrize the three building blocks of the model, schematized in Fig. 1: the clone size distribution $\rho(f)$, the noise model $P(n|f)$, and the dynamical model $G(f', t'|f, t)$.

Distribution of lymphocyte clone sizes

The distribution of clone sizes in memory or unfractioned TCR repertoires has been observed to follow a power law in human [29–31] and mice [32, 33]. These observations justify parametrizing the clone size distribution as

$$\rho(f) = Cf^{-\nu}, \qquad f_{\min} \leq f < 1, \tag{1}$$

and C a normalizing constant. We will verify in the next section that this form of clone size distribution describes the data well. For $\nu > 1$, which is the case for actual data, the minimum $f_{\min}$ is required to avoid the divergence at $f = 0$. This bound also reflects the smallest possible clonal frequencies given by the inverse of the total number of lymphocytes, $1/N_{\text{cell}}$. The frequencies of different clones are not independent, as they must sum up to 1: $\sum_{i=1}^{N} f_i = 1$, where N is the total number of clones in the organism. The joint distribution of frequencies thus reads:

$$\rho_N(f_1, \ldots, f_N) \propto \prod_{i=1}^{N} \rho(f_i)\delta\left(\sum_{i=1}^{N} f_i - 1\right). \tag{2}$$

This condition, $\sum_{i=1}^{N} f_i = 1$, will be typically satisfied for large N as long as $\langle f \rangle = \int \mathrm{d}f\, f\rho(f) = 1/N$ (see Methods), but we will need to enforce it explicitly during the inference procedure.

Noise model for sampling and sequencing

The noise model captures the variability in the number of sequenced reads as a function of the true frequency of its clonotypes in the considered repertoire or subrepertoire. The simplest and lowest-dispersion noise model assumes random sampling of reads from the distribution of clonotypes. This results in $P(n|f)$ being given by a Poisson distribution of mean fN_{read}, where N_{read} is the total number of sequence reads. Note that for the data analyzed in this paper, reads are collapsed by unique barcodes corresponding to individual mRNA molecules.

Variability in mRNA expression as well as library preparation introduces uncertainty that is far larger than predicted by the Poisson distribution. This motivated us to model the variability in read counts by a negative binomial of mean $\bar{n} = fN_{\text{read}}$ and variance $\bar{n} + a\bar{n}^{\gamma}$, where a and γ control the over-dispersion of the noise. Negative binomial distributions were chosen because they allow us to control the mean and variance independently, and reduce to Poisson when $a = 0$. These distributions are also popular choices for modeling RNAseq variability in differential expression methods [26, 28].

A third noise model was considered to account explicitly for the number of cells representing the clone in the sample, m. In this two-step model, $P(m|f)$ is given by a negative binomial distribution of mean $\bar{m} = fM$ and variance $\bar{m} + a\bar{m}^{\gamma}$, where M is the total number of cells represented in the sample. $P(n|m)$ is a Poisson distribution of mean mN_{read}/M. The resulting noise model is then given by $P(n|f) = \sum_m P(n|m)P(m|f)$. The number of sampled cells, M, is unknown and is a parameter of the model. Note that this two-step process with the number of cells as an intermediate variable is specific to repertoire sequencing, and has no equivalent in RNAseq differential expression analysis.

Dynamical model of the immune response

Finally, we must specify the dynamical model for the clonal frequencies. In the context of vaccination or infection, it is reasonable to assume that only a fraction α of clones respond by expanding. We also assume that expansion or contraction does not depend on the size of the clone itself. Defining $s = \ln(f'/f)$ as the log-fold factor of expansion or contraction, we define:

$$G(f' = fe^s, t'|f, t)\mathrm{d}f' = \rho_s(s)\mathrm{d}s. \tag{3}$$

with

$$\rho_s(s) = (1-\alpha)\delta(s - s_0) + \alpha\rho_{\text{exp}}(s - s_0), \tag{4}$$

where ρ_{exp} describes the expansion of responding clones, and $s_0 < 0$ corresponds to an overall contraction factor ensuring that the normalization of frequencies to 1 is satisfied after expansion. In the following, we shall specialize to particular forms of ρ_{exp} depending on the case at hand.

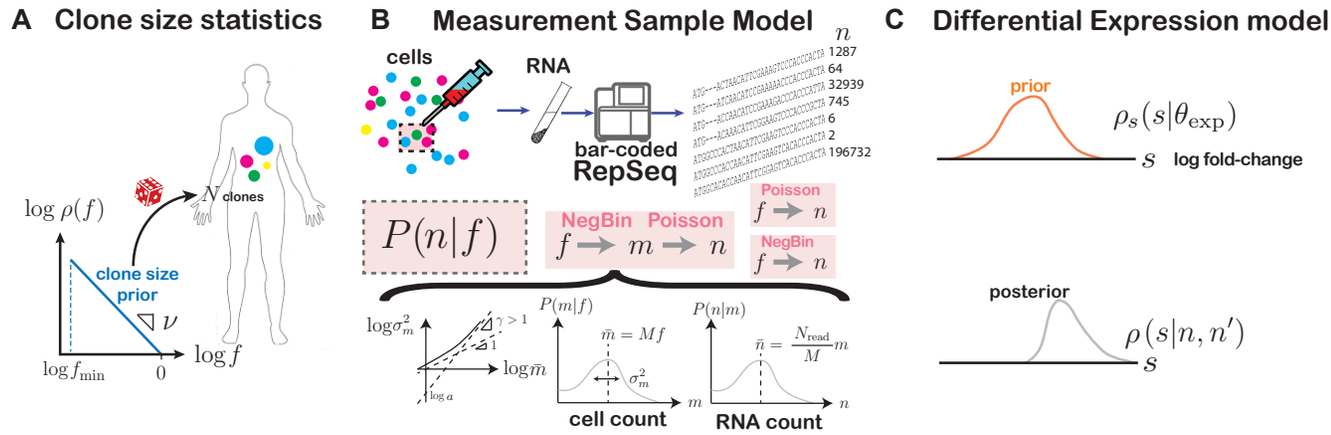

Figure 1: *Model components.* (A) Clone frequencies are sampled from a prior density of power law form with power ν and minimum frequency, $f_{\min}$. (B) Each clone's frequency f determines the count distribution, $P(n|f)$, that governs its mRNA count statistics in the observed sample. We consider 3 forms for $P(n|f)$: Poisson, negative binomial, and a two-step (negative binomial to Poisson) model. The negative binomial and two-step measurement models are parametrized through a mean-variance relationship specifying the power, γ, and coefficient, a, of the over-dispersion of cell count statistics. The mean cell count scales with the number of cells in the sample, M, while the mean read count scales with with the number of cells, m, and the sampling efficiency, M/N_{read}, with N_{read} the measured number of molecules in the sample. The parameters of the measurement model are learned on pairs of sequenced repertoire replicates. (C) Differential expression is implemented in the model via a random log fold change, s, distributed according to the prior $\rho(s|\theta_{\text{exp}})$. The prior's parameters, θ_{exp}, are learned from the dataset using maximum likelihood. Once learned, the model is used to compute posteriors over s given observed count pairs, which is used to make inferences about specific clones.

Inferring the noise profile from replicate experiments

To study variations arising from experimental noise, we analysed replicates of repertoire sequencing experiments. The tasks of learning the noise model and the distribution of clone sizes are impossible to dissociate. To infer $P(n|f)$, one needs to get a handle on f, which is unobserved, and for which the prior distribution $\rho(f)$ is essential. Conversely, to learn $\rho(f)$ from the read counts n, we need to deconvolve the experimental noise, for which $P(n|f)$ is needed. Both can be learned simultaneously from replicate experiments (i.e. $f' = f$), using maximum likelihood estimation. For each clone, the probability of observing n read counts in the first replicate and n' read counts in the second replicate reads:

$$P(n, n'|\theta_{\text{null}}) = \int_{f_{\min}}^{1} \mathrm{d}f\, \rho(f|\theta_{\text{null}}) P(n|f, \theta_{\text{null}}) P(n'|f, \theta_{\text{null}}), \tag{5}$$

where θ_{null} is a vector collecting all the parameters of both the noise model and the clone size distribution, namely $\theta_{\text{null}} = \{f_{\min}, \nu\}$ for the Poisson noise model, $\theta_{\text{null}} = \{f_{\min}, \nu, a, \gamma\}$ for the negative binomial noise model, and $\theta_{\text{null}} = \{f_{\min}, \nu, a, \gamma, M\}$ for the two-step noise model.

While Eq. 5 gives the likelihood of a given read count pair (n, n'), we need to correct for the fact that we only observe pairs for which $n + n' > 0$. In general, many clones in the repertoire are small and missed in the acquisition process. In any realization, we expect $n+n' > 0$ for only a relatively small number of clones, $N_{\text{obs}} \ll N$. Typically, N_{obs} is of order 10^5, while N is unknown but probably ranges from 10^7 for mouse to $10^8 - 10^{10}$ for humans [34, 35]. Since we have no experimental access to the unobserved clones ($n = n' = 0$), we maximize the likelihood of the read count pairs (n_i, n'_i), $i = 1, \ldots, N_{\text{obs}}$, conditioned on the clones appearing in the sample:

$$\hat{\theta}_{\text{null}} = \underset{\theta_{\text{null}}}{\operatorname{argmax}} \prod_{i=1}^{N_{\text{obs}}} \frac{P(n_i, n'_i|\theta_{\text{null}})}{1 - P(0, 0|\theta_{\text{null}})}. \tag{6}$$

While the condition $N\langle f\rangle = 1$ ensures normalization on average, we may instead require that normalization be satisfied for the particular realization of the data, by imposing:

$$Z = NP(0,0)\langle f\rangle_{\rho(f|n+n'=0)} + \sum_{i=1}^{N_{\text{obs}}} \langle f\rangle_{\rho(f|n_i, n'_i)} = 1, \tag{7}$$

where N is estimated as $N = N_{\text{obs}}/(1-P(0,0))$. The first term corresponds to the total frequency of the unseen clones, while the second term corresponds to a sum of the average posterior frequencies of the observed clones. Imposing either Eq. 7 or $N\langle f\rangle = 1$ yielded similar values of the parameter estimates, $\hat{\theta}_{\text{null}}$.

To test the validity of the maximum likelihood estimator, Eq. 6, we created synthetic data for two replicate sequencing experiments with known parameters θ_{null} under the two-step noise model, and approximately the same

number of reads as in the real data. To do so efficiently, we developed a sampling protocol that deals with the large number of unobserved clones implicitly (see Methods). Applying the maximum likelihood estimator to these synthetic data, we correctly inferred the ground truth over a wide range of parameter choices (Fig. S1).

Next, we applied the method to replicate sequencing experiments of unfractioned repertoires of 6 donors over 5 time points spanning a 1.5 month period (30 donor-day replicate pairs in total). For a typical pair of replicates, a visual comparison of the (n, n') pairs generated by the Poisson and two-step noise models with the data shows that the Poisson distribution fails to explain the large observed variability between the two replicates, while the two-step model can (Fig. 2A-C). The normalized log-likelihood of the two-step model was slightly but significantly higher than that of the negative binomial model, and much larger than that of the Poisson model (Fig. 2D). The two-step model was able to reproduce accurately the distribution of read counts $P(n)$ (Fig. 3A), as well as the conditional distribution $P(n'|n)$ (Fig. 3B), even though those observables were not explicitly constrained by the fitting procedure. In particular, $P(n)$ inherits the power law of the clone frequency distribution $\rho(f)$, but with deviations at low count numbers due to experimental noise, which agree with the data. Also, the two-step model outperformed the negative binomial noise model at describing the long tail of the read count distribution for clones that were not seen in one of the two replicates (see Fig. S2).

Figure 4 shows the learned values of the parameters for all 30 pairs of replicates across donors and timepoints. While there is variability across donors and days, there is a surprising degree of consistency. Despite being inferred indirectly from the characteristics of the noise model, estimates for the number of cells in the samples, M, are within one order of magnitude of their expected value based on the known concentration of lymphocytes in blood (about one million cells per sample). Likewise, $f_{\min}$ is very close to the smallest possible clonal frequency, $1/N_{\text{cell}}$, where $N_{\text{cell}} \approx 4 \cdot 10^{11}$ is the total number of T cells in the organism [36].

The inferred models can also be used to estimate the diversity of the entire repertoire (observed or unobserved). The clone frequency distribution, $\rho(f)$, together with the estimate of N can be used to estimate Hill diversities (see Methods):

$$D_\beta = \left(\sum_{i=1}^{N} f_i^\beta \right)^{\frac{1}{1-\beta}} = \left(N \langle f^\beta \rangle \right)^{\frac{1}{1-\beta}}. \tag{8}$$

In Fig. 5, we show the values, across donor and days, of three different diversities: species richness, i.e. the total number of clones N ($\beta = 0$); Shannon diversity, equal to the exponential of the Shannon entropy ($\beta = 1$); and Simpson diversity, equal to the inverse probability that two cells belong to the same clone ($\beta = 2$). In particular, estimates of $N \approx 10^9$ fall between the lower bound of 10^8 unique TCRs reported in humans using extrapolation techniques [34] and theoretical considerations giving upper-bound estimates of 10^{10} [35] or more [37].

Learning the repertoire dynamics from pairs of time points

Now that the baseline for repertoire variation has been learned from replicates, we can learn something about its dynamics following immunization. The parameters of the expansion model (Eq. 4) can be set based on prior knowledge about the typical fraction of responding clones and effect size. Alternatively, they can be inferred from the data using maximum likelihood estimation (Empirical Bayes approach). We define the likelihood of the read count pairs (n, n') between time points t and t' as:

$$P_{\text{exp}}(n, n'|\theta_{\text{null}}, \theta_{\text{exp}}) = \int_{f_{\min}}^{1} \mathrm{d}f \rho(f) \int \mathrm{d}s \rho_s(s|\theta_{\text{exp}}) P(n|f, \theta_{\text{null}}) P(n'|fe^s, \theta_{\text{null}}), \tag{9}$$

where $\theta_{\text{exp}} = \{\alpha, s_0, \bar{s}\}$ characterizes $\rho_s(s)$ (Eq. 4) with $\bar{s}$ parametrizing $\rho_{\text{exp}}(s)$, and where $\theta_{\text{null}} = \hat{\theta}_{\text{null}}$ is set to the value learned from replicates taken at the first time point t. The maximum likelihood estimator is given by

$$\hat{\theta}_{\text{exp}} = \underset{\theta_{\text{exp}}}{\operatorname{argmax}} \prod_{i=1}^{N_{\text{obs}}} \frac{P_{\text{exp}}(n_i, n_i'|\hat{\theta}_{\text{null}}, \theta_{\text{exp}})}{1 - P_{\text{exp}}(0, 0|\hat{\theta}_{\text{null}}, \theta_{\text{exp}})}. \tag{10}$$

This maximization was performed via gradient-based methods. In Methods we give an example of a semi-analytic approach to finding the optimum using the expectation maximization algorithm.

In addition to normalization at t, we also need to impose normalization at t':

$$Z' = NP(0,0)\langle f' \rangle_{\rho(f'|n+n'=0)} + \sum_{i=1}^{N_{\text{obs}}} \langle f' \rangle_{\rho'(f'|n_i, n_i')}, \tag{11}$$

with $\rho(f'|n, n') \propto \int \mathrm{d}f \rho(f) G(f'|f) P(n|f) P(n'|f')$ is the posterior distribution of the f' given the read count pair. In practice, we impose $Z = Z'$, where Z is the normalization of the first time point given by Eq. 7. Intuitively, this normalization constraint sets s_0 so that the expansion of a few clones is compensated by the slight contraction of all clones.

We first tested the method on synthetic data generated with the expansion model of Eq. 9, with an exponentially distributed effect size for the expansion with scale parameter, $\bar{s}$:

$$\rho_{\text{exp}}(s') = \frac{1}{\bar{s}} e^{-s'/\bar{s}} \Theta(s'), \tag{12}$$

where $\Theta(s') = 1$ if $s' > 0$ and 0 otherwise (Fig. 6A). We simulated small, mouse-like and large, human-like repertoires (number of clones, $N = 10^6$ and $N = 10^9$; number

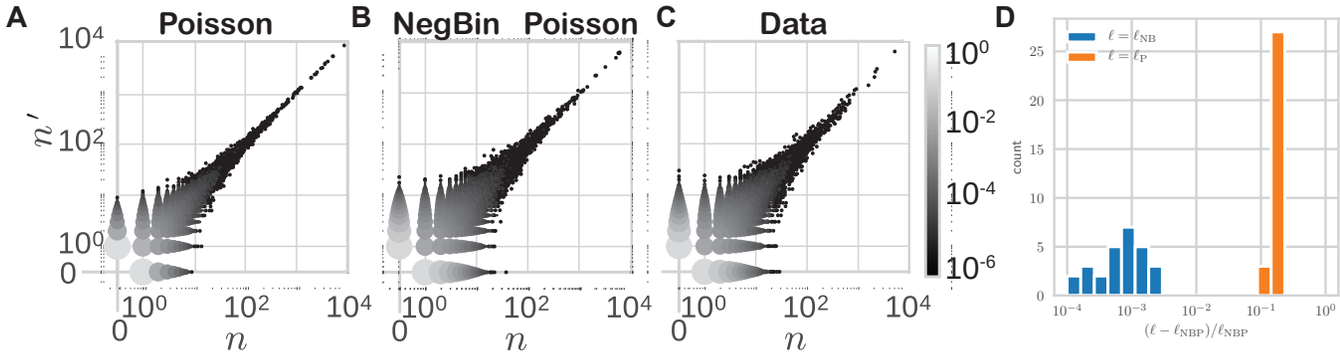

Figure 2: *Comparison of measurement models.* Pair count distributions sampled from learned (A) negative binomial and (B) Poisson models, compared to (C) data. (D) shows the log likelihoods, ℓ (logarithm of the argument of the argmax in Eq. 6) of the Poisson (P) and negative binomial (NB) models relative to that of the two-step model (NBP). (Example dataset: day-0 replicate pair from donor S2.)

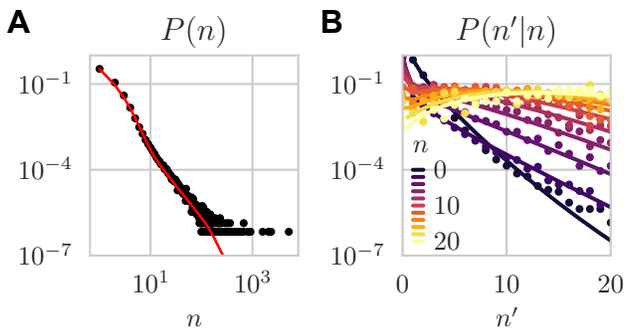

Figure 3: *Count distributions.* (A) Marginal count distribution, $P(n|\theta_{\rm null}) = \sum_{n'} P(n, n'|\theta_{\rm null})$, and (B) conditional count distribution, $P(n|n', \theta_{\rm null}) = P(n, n'|\theta_{\rm null})/P(n|\theta_{\rm null})$. Both marginal and conditional distributions are quantitatively predicted by the model. Lines are analytic predictions of the learned model. Dots are estimated frequencies. (Same data as Fig. 2; two-step noise model).

of reads/sample $N_{\rm reads} = 10^4$ and $N_{\rm reads} = 2 \cdot 10^6$, respectively), using $\nu = 2$ and $f_{\rm min}$ satisfying $N\langle f\rangle_{\rho(f)}=1$. For the parameter-free Poisson measurement model, we analyzed the differential expression model, Eq. 9, over a range of biologically plausible parameter values. In Fig. 6B, we show the parameter space of the inference of a mouse repertoire generated with $(\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$ and $s_0 = s_0(\alpha, \bar{s})$ fixed by the normalization constraint $Z' = Z$. The errors are distributed according to a diagonally elongated ellipse (or 'ridge'), with a covariance following the inverse of the Hessian of the log-likelihood. The imprecision of the parameter estimates is due to the small number of sampled responding clones. With $\alpha^* = 0.01$ and $N_{\rm obs} \approx 10^4$ sampled clones, only a few dozens responding clones are detected. For human-sized repertoires, millions of clones are sampled, which makes the inference much more precise (see Fig. S3).

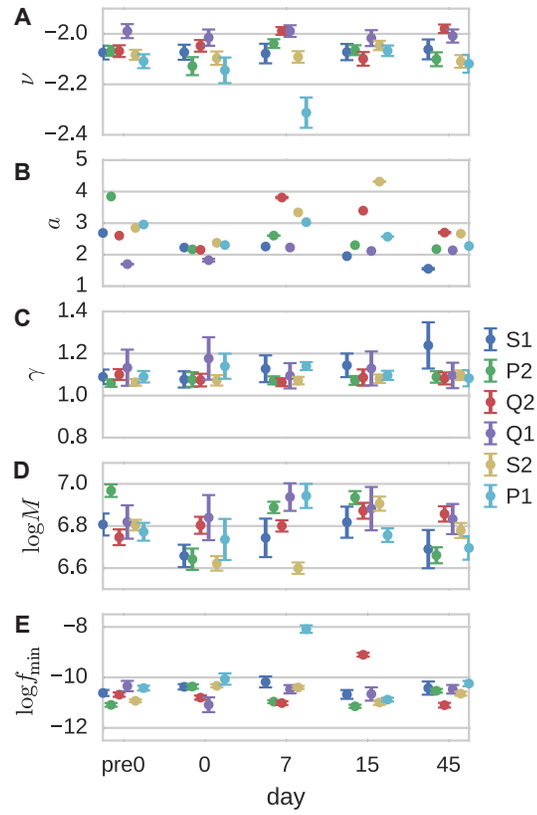

Figure 4: *Inferred null model parameters.* Inferred values: for (A) the power-law exponent ν of the clone size distribution; (B) and (C) linear coefficient and exponent of the mean-variance relationship of the noise; (D) effective number of cells; and (E) minimal clonal frequency. Each point is inferred from a pair of replicates for a given donor and time point. Error bars are obtained by inverting the Hessian of the log-likelihood projected onto the hyperplane locally satisfying the normalization constraint.

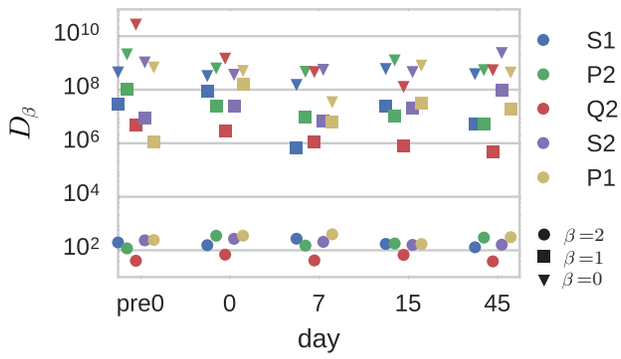

Figure 5: *Diversity estimates.* Shown are diversity estimates obtained from the Hill diversities, D_β, of the inferred clone frequency distributions for $\beta = 0$ (estimated total number of clones, N), $\beta = 1$ (Shannon entropy) and $\beta = 2$ (Simpson index), across donors and days.

Once learned, the model can be used to compute the posterior probability of a given expansion factor by marginalizing f, and using Bayes' rule,

$$\rho(s|n, n') \propto \rho_s(s) \int P(n|f)P(n'|fe^s)\rho(f)\mathrm{d}f. \qquad (13)$$

We illustrate different posterior shapes from synthetic data as a function of the observed count pairs in Fig. 6C. We see for instance 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 $n' \lessapprox n$.

Note that the value of the true responding fraction α is correctly learned from our procedure, regardless of our ability to tell with perfect certainty which particular clones responded. By contrast, a direct estimate of the responding fraction from the number of significantly responding clones, as determined by differential expression software such as EdgeR [26], is likely to misestimate that fraction. We applied EdgeR to a synthetic repertoire of $N = 10^9$ clones, a fraction $\alpha = 0.01$ of which responded with mean effect $\bar{s} = 1$, and sampled with $N_{\mathrm{read}} = 10^6$. EdgeR found $6,880$ significantly responding clones (corrected p-value 0.05) out of $N_{\mathrm{obs}} = 1,995,139$, i.e. a responding fraction $6,880/1,995,139 \approx 3 \cdot 10^{-3}$ of the observed repertoire, and a responding fraction $6,880/10^9 \approx 7 \cdot 10^{-6}$ of the total repertoire, underestimating the true fraction $\alpha = 10^{-2}$.

Inference of the immune response following immunization

Next, we ran the inference procedure on sequences obtained from human blood samples across time points following yellow fever vaccination. To guide the choice of prior for s, we plotted the histograms of the naive log fold-change $\ln n'/n$ (Fig. S4). These distributions show

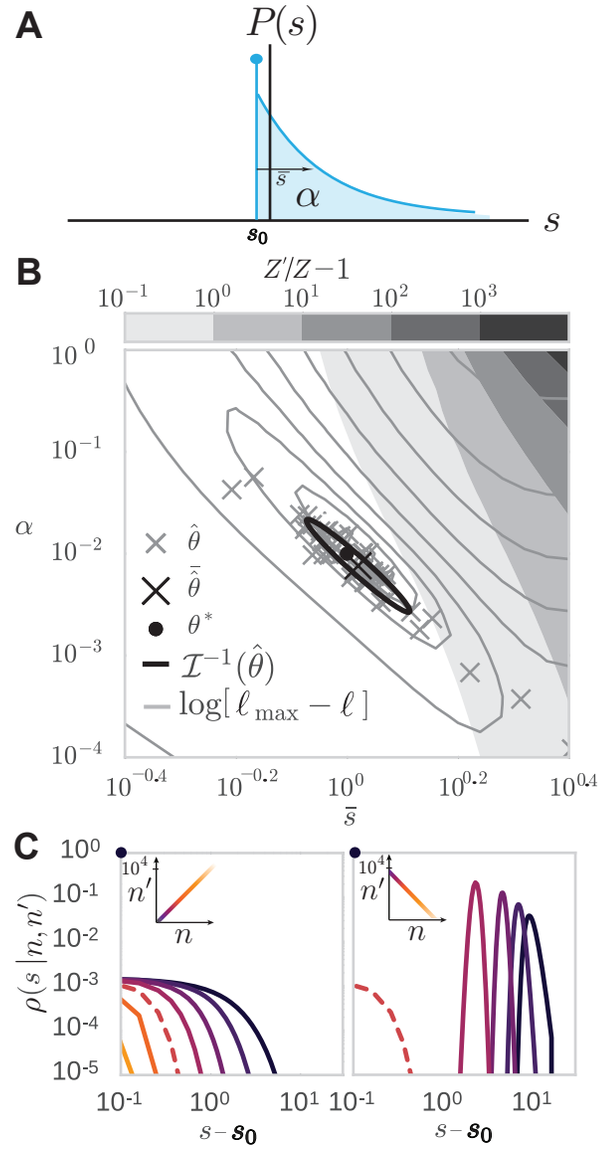

Figure 6: *Inference of clonal expansion on synthetic data.*(A) The prior distribution of expansion log fold-changes, $\rho_s(s)$, is parametrized by a mean effect size, $\bar{s}$, describing the expansion of the responding fraction, α of the repertoire (Eq. 12). Expansion is relative to a basal s_0 fixed by the normalization constraint. (B) Re-inference of the expansion parameters from synthetic data generated with value $\theta^*_{\mathrm{exp}} = (\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$ (black dot). Maximums of the log-likelihood, $\hat{\theta}_{\mathrm{exp}}$, for 50 realizations (gray crosses), along with their average $\bar{\hat{\theta}}$ (black cross). The log-likelihood for one realization is shown over logarithmically-spaced gray contours decreasing from the maximum. The inverse Fisher information, $\mathcal{I}^{-1}$, for the same realization is shown as the black-lined ellipse centered at $\bar{\hat{\theta}}$, which provides a lower bound to the variance of our ML estimate. The gray scale contours increasing to the upper-right denote $Z'/Z - 1$, the excess in the used normalization. (C) Posteriors of the learned model, $\rho(s|n, n')$ 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 contribution of the non-responding component, $\propto \delta(s - s_0)$, to the posterior. (Parameters: $N = 10^6$, $\epsilon = 10^{-2}$.)

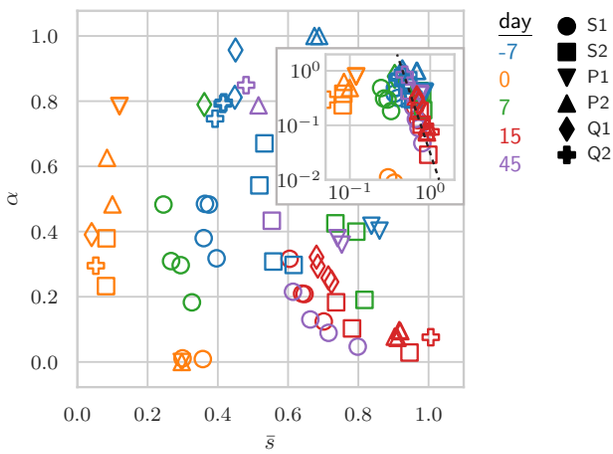

Figure 7: *Application to yellow fever vaccination data.* Optimal values of α and $\bar{s}$ across all 6 donors and days relative to the day of vaccination (day 0). Inset shows the same data on a logarithmic scale for comparison with Fig. 6b. Comparisons with days other than 0 fall on straight line (dashed line).

symmetric exponential tails, motivating us to model the statistics of expansion factors as:

$$\rho_{\text{exp}}(s) = \frac{1}{2\bar{s}} e^{-|s|/\bar{s}}, \tag{14}$$

with typical effect size $\bar{s}$.

We applied the inference procedure (Eq. 10) between the repertoires taken the day of vaccination (day 0), and at one of the other time points (day -7, day 7, day 15, and day 45) after vaccination. Since there are two replicates at each time point, we can make 4 comparisons between any pair of time points.

Same-day comparisons (day 0 vs day 0) gave effectively zero mean effect sizes ($\bar{s} < 0.1$, below the discretization step of the integration procedure), as expected (Fig. 7). Comparisons with other days yielded inferred values of α and $\bar{s}$ distributed along the same 'ridge' (Fig. 7), as observed on synthetic data (Fig. 6). The mean effect size $\bar{s}$ is highest at day 15, where the peak of the response occurs, but is also substantially different from 0 at all time points except day 0 (including before vaccination at day -7), with often high values of α. We speculate that these fluctuations reflect natural variations of the repertoire across time, as well as experimental batch effects. As a consequence of the natural diversity, values of the responding fraction α are not learned with great precision, as can be seen from the variability across the 4 choices of replicate pair, and are probably gross overestimations of the true probability that a naive T cell responds to an infection, which is believed to be of order $10^{-5} - 10^{-3}$ [38].

Identifying responding clones

The posterior probability on expansion factors $\rho(s|n, n')$ (Eq. 13) can be used to study the fate and dynamics of particular clones. For instance, we can identify responding clones as having a low posterior probability of being not expanding $P_{\text{null}} = \rho(s \leq 0|n, n') < 0.025$. P_{null} is the Bayesian counterpart of a p-value but differs from it in a fundamental way: it gives the probability that expansion happened given the observations, when a p-value would give the probability of the observations in absence of expansion. We can define a similar criterion for contracting clones.

To get the expansion or contraction factor of each clone, we can compute the posterior average and median, $\langle s \rangle_{n,n'} = \int \mathrm{d}s\, s\rho(s|n, n')$ and s_{median} ($F(s_{\text{median}}|n, n') = 0.5$, for the cumulative density function, $F(s|n, n') = \int_{-\infty}^{s} \rho(\tilde{s}|n, n')\mathrm{d}\tilde{s}$), corresponding to our best estimate for the log fold-change. In Fig. S5, we show how the median Bayesian estimator differs from the naive estimator $s_{\text{naive}} = \ln n'/n$. While the two agree for large clones for which relative noise is smaller, the naive estimator overestimates the magnitude of log fold-changes for small clones because of the noise. The Bayesian estimator accounts for that noise and gives a more conservative and more realistic estimate.

In Fig. 8A, we show a 'volcano' plot showing how both P_{null} and $\langle s \rangle_{n,n'}$ vary as one scans values of the count pairs (n, n'). Figure 8B shows the all count pairs (n, n') between day 0 and day 15 following yellow fever vaccination, with red clones above the significance threshold line $P_{\text{null}} = 0.025$ being identified as responding.

Given the uncertainty in the expansion model parameters $\theta_{\text{exp}} = (\bar{s}, \alpha)$, we wondered how robust our list of responding clonotypes was to those variations. In Fig. 8C, we show the overlap of lists of strictly expanding clones ($P(s \leq 0|n, n') < 0.025$) as a function of θ_{exp}, relative to the optimal value $\hat{\theta}_{\text{exp}}$ (black circle). The ridge of high overlap values exactly mirrors the ridge of high likelihood values onto which the learned parameters fall (Fig. 7). Values of $\hat{\theta}_{\text{exp}}$ obtained for other replicate pairs (square symbols) fall onto the same ridge, meaning that these parameters lead to virtually identical lists of candidates for response.

The list of identified responding clones can be used to test hypotheses about the structure of the response. For example, recent work has highlighted a power law relationship between the initial clone size and clones subsequent fold change response in a particular experimental setting [39]. We can plot the relationship in our data as the posterior mean log fold change versus the posterior initial frequency, f (Fig. 8D). While the relationship is very noisy, emphasizing the diversity of the response, it is consistent with a decreasing dependency of the fold change with the clone size prior to the immune response.

The robustness of our candidate lists rests on their insensitivity to the details of how the model explains typical expansion. In Fig. S6, we show how the posterior belief

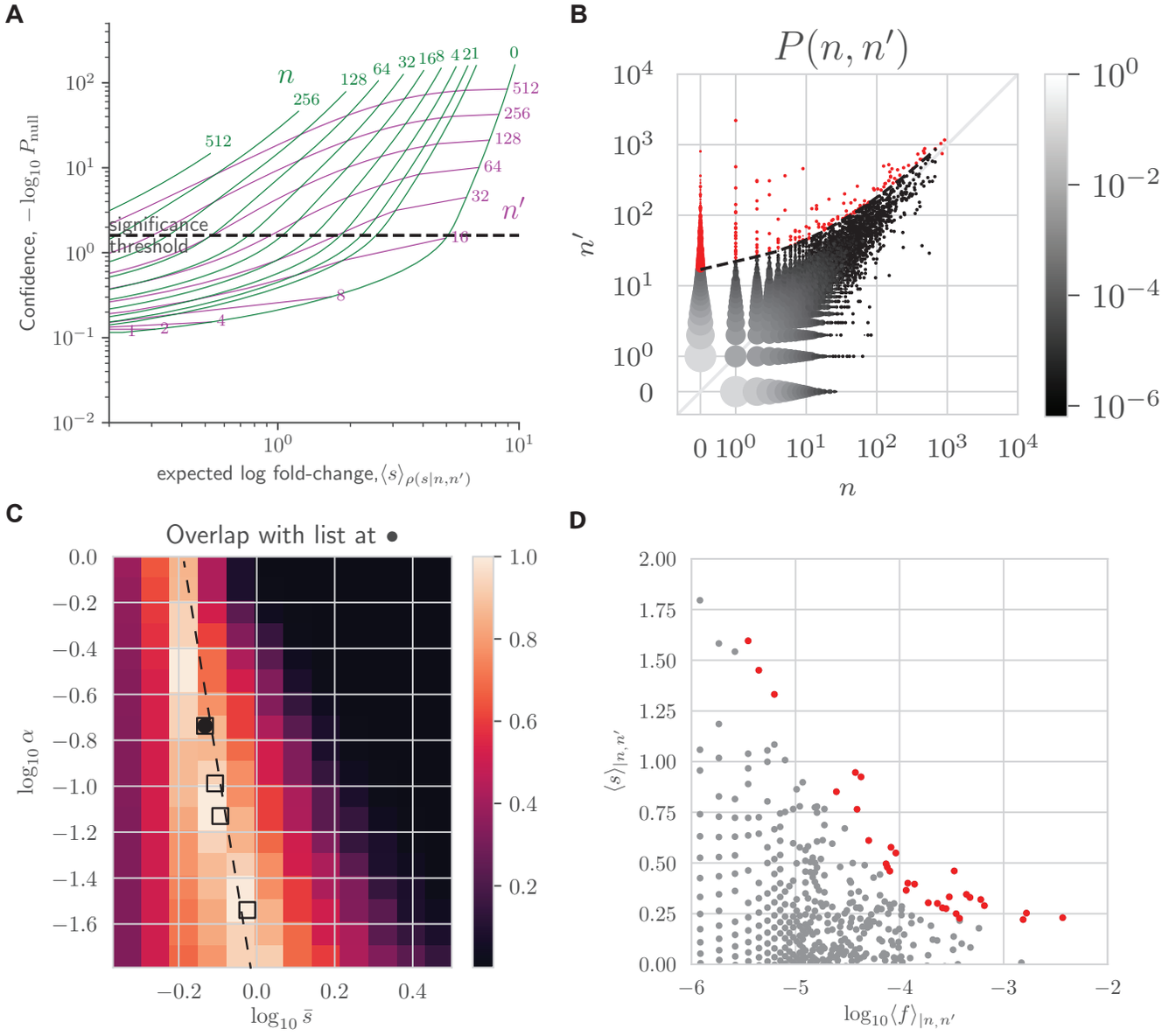

Figure 8: *Identifying responding clones.* (A) Plot of confidence of expanded response versus average effect size. A significance threshold is placed according to $P_{\text{null}} = 0.025$, where $P_{\text{null}} = P(s \leq 0)$. (B) The same threshold for significant expansion in (n, n')-space with identified clones highlighted in red. (C) 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 (the size of the intersection of the two lists divided by the size of their union) between a reference list obtained at the optimal $\hat{\theta}_{\text{exp}}$ (black dot) and lists obtained at non-optimal θ_{exp}. (D) Mean posterior log fold-change $\langle s \rangle_{\rho(s|n,n')}$ as a function of precursor frequency.

varies significantly for count pairs $(0, n')$, $n' > 0$, across a range of values of $\bar{s}$ and α passing along the ridge of plausible models (Fig. 8C). A transition from a low to high value of the most probable estimate for s characterizes their shapes and arises as $\bar{s}$ becomes large enough that expansion from frequencies near f_{min} is plausible, and the dominant mass of clones there makes this the dominate posterior belief. Thus, these posteriors are shaped by $\rho_s(s)$ at low $\bar{s}$, and $\rho(f)$ at high $\bar{s}$. Our lists vary negligibly over this transition, and thus are robust to it.

Discussion

Our probabilistic framework describes two sources of variability in clonotype abundance from repertoire sequencing experiments: biological repertoire variations and spurious variations due to noise. We found that in

a typical experiment, noise is over-dispersed relative to Poisson sampling noise. This makes the use of classical difference tests such as Fisher's exact test or a chi-squared test inappropriate in this context, and justifies the development of specific methods.

As a byproduct, our method learned the properties of the clone size distribution, which is consistent with a power law of exponent ≈ -2.1 robust across individuals and timepoints, consistent with previous reports [29–31]. Using these parameters, various diversity measures could be computed, such as the species richness (10^8–10^9), which agrees with previous bounds [34, 35], or the "true diversity" (the exponential of the Shannon entropy), found to range between 10^6 and 10^8.

The proposed probabilistic model of clonal expansion is described by two parameters: the fraction of clones that respond to the immune challenge, and the typical effect size (log fold-change). While these two parameters were difficult to infer precisely individually, a combination of them could be robustly learned. Despite this ambiguity in the model inference, the list of candidate responding clonotypes is largely insensitive to the parameter details. For clonotypes that rose from very small read counts to large ones, the inferred fold-change expansion factor depended strongly on the priors, and resulted from a delicate balance between the tail of small clones in the clone size distribution and the tail of large expansion events in the distribution of fold-changes.

While similar approaches have been proposed for differential expression analysis of RNA sequencing data [25–28], the presented framework was specifically built to address the specific challenges of repertoire sequencing data. Here, the aim is to count proliferating cells, as opposed to evaluating average expression of genes in a population of cells. We specifically describe two steps that translate cell numbers into the observed TCR read counts: random sampling of cells that themselves carry a random number of mRNA molecules, which are also amplified and sampled stochastically. Another difference with previous methods is the explicit Bayesian treatment, which allows us to calculate a posterior probability of expansion, rather than a less interpretable p-value.

Here we applied the presented methodology to an acute infection. We have previously shown that it can successfully identify both expanding (from day 0 to 15 after vaccination) and contracting (from day 15 to day 45) clonotypes after administering a yellow fever vaccine. However the procedure is more general and can also be extended to be used in other contexts. For instance, this type of approach could be used to identify response in B-cells during acute infections, by tracking variations in the size of immunoglobulin sequence lineages (instead of clonotypes), using lineage reconstruction methods such as Partis [40]. The framework could also be adapted to describe not just expansion, but also switching between different cellular phenoypes during the immune response, *e.g.* between the naive, memory, effector memory, *etc.* phenotypes, which can obtained by flow-sorting cells before sequencing [41]. Another possible application would be to track the clones across different tissues and organs, and detect migrations and local expansions [42].

The proposed framework is not limited to identifying a response during an acute infection, but can also be used as method for learning the dynamics from time dependent data even in the absence of an external stimulus [19]. Here we specifically assumed expansion dynamics with strong selection. However, the propagator function can be replaced by a non-biased random walk term, such as genetic drift. In this context the goal is not to identify responding clonotypes but it can be used to discriminate different dynamical models in a way that accounts for different sources of noise inherently present in the experiment. Alternatively, the framework can also be adapted to describe chronic infections such as HIV [43], where expansion events may be less dramatic and more continuous or sparse, as the immune system tries to control the infection over long periods of time.

Methods

Code

All code used to produce the results in this work was custom written in Python 3 and is publicly available online at `https://github.com/mptouzel/bayes_diffexpr`.

Normalization of the clonal frequencies

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

$$\mathcal{Z} = \int_{f_{\min}}^{1} \cdots \int_{f_{\min}}^{1} \prod_{i=1}^{N} \rho(f_i)\delta(Z-1)\mathrm{d}^N \vec{f}\,, \tag{15}$$

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,

$$\delta(Z-1) = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{\mu(Z-1)} \tag{16}$$

$$= \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} e^{\mu f_i}\,. \tag{17}$$

Crucially, the multi-clone integral in Eq. 15 over $\vec{f}$ then factorizes. Exchanging the order of the integrations we obtain

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \langle e^{\mu f}\rangle^N\,, \tag{18}$$

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

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-NI(\mu)} \ . \tag{19}$$

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

$$\frac{1}{N} = \frac{\langle f e^{\mu^* f}\rangle}{\langle e^{\mu^* f}\rangle} \ . \tag{20}$$

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. 19 that $\mathcal{Z} = 1$, and so the constraint $N\langle f\rangle = 1$ imposes normalization.

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: (a) the model parameters $\{\alpha, M, a, \gamma\}$, excluding $f_{\min}$; (b) the desired size of the full repertoire, N; (c) the sequencing efficiency (total sample reads/total sample cells), ϵ. From this we get the effective total sample reads, $N_{\text{reads}}^{\text{eff}} = \epsilon M$, which converts a clone's frequency to the average number of cells it appears with in the sample. 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:

$$N\langle f\rangle_{\rho(f)} = 1 \tag{21}$$

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_{read}/M, due to their differing number of observed total number of molecules, N_{read}. 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 $\mathcal{O}$),

$$P(n, n', f|\mathcal{O}) = \frac{P(n|f)P(n'|f)\rho(f)}{1 - \int \mathrm{d}f\rho(f)\mathrm{d}f P(n=0|f)P(n'=0|f)} \ , \tag{22}$$

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

$$P(q_{x0}|\mathcal{O}) = \sum_{n>0} \int \mathrm{d}f P(n, n'=0, f|\mathcal{O}) \ , \tag{23}$$

$$P(q_{0x}|\mathcal{O}) = \sum_{n'>0} \int \mathrm{d}f P(n=0, n', f|\mathcal{O}) \ , \tag{24}$$

$$P(q_{xx}|\mathcal{O}) = \sum_{\substack{n>0,\\ n'>0}} \int \mathrm{d}f P(n, n', f|\mathcal{O}). \tag{25}$$

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,

$$P(f|q_{x0}) = \sum_{n>0} P(n, n'=0, f|\mathcal{O})/P(q_{x0}|\mathcal{O}) \ , \tag{26}$$

$$P(f|q_{0x}) = \sum_{n'>0} P(n=0, n', f|\mathcal{O})/P(q_{0x}|\mathcal{O}) \ , \tag{27}$$

$$P(f|q_{xx}) = \sum_{n>0, n'>0} P(n, n', f|\mathcal{O})/P(q_{xx}|\mathcal{O}). \tag{28}$$

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. S1).

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 $N_{\rm obs}$ is fixed in the former, while is stochastic in the latter.

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.

Obtaining diversity estimates from the clone frequency density

For a set of clone frequencies, $\{f_i\}_{i=1}^N$, the Hill family of diversities are obtained from the Rényi 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, $N_{\rm obs} + P(n = 0)N = N$, where $N_{\rm obs}$ 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 $1/\left(N\langle f^2\rangle_{\rho(f)}\right)$.

Inferring the differential expression prior

To learn 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 practice, we first perform the latter analytically and then evaluate the former numerically. We choose a symmetric exponential as a tractable prior for this purpose:

$$\rho_{\rm exp}(s|\bar{s}) = e^{-|s|/\bar{s}}/2\bar{s} \tag{29}$$

with $\bar{s} > 0$, and no shift, $s_0 = 0$. The expected value of the log likelihood function, often called the Q-function in EM literature, is

$$Q(\bar{s}|\bar{s}') = \sum_{i=1}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n_i, n'_i, \bar{s}') \log\left[P(n_i, n'_i, s|\bar{s})\right] \ , \tag{30}$$

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

$$\frac{\partial \log\left[\rho_{\rm exp}(s|\bar{s}))\right]}{\partial \bar{s}} = \frac{1}{\rho_{\rm exp}(s|\bar{s})} \frac{\partial \rho_{\rm exp}(s|\bar{s})}{\partial \bar{s}} \tag{31}$$

$$= \frac{|s| - \bar{s}}{\bar{s}^2} \ , \tag{32}$$

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

$$\sum_{i=1}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n_i, n'_i, \bar{s}') \frac{|s| - \bar{s}^*}{\bar{s}^{*2}} = 0 \tag{33}$$

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

$$\bar{s}_{(n,n')} = \int_{-\infty}^{\infty} \mathrm{d}s |s| \rho(s|n, n', \bar{s}'). \tag{34}$$

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 $\left.\frac{\partial^2 \log[\rho_{\rm exp}(s|\bar{s}))]}{\partial \bar{s}^2}\right|_{\bar{s}=\bar{s}^*} = -\bar{s}^{*-2} < 0$. Thus, we update $\rho_{\rm exp}(s|\bar{s})$ with $\bar{s} \leftarrow \bar{s}^*$. The number of updates typically required for convergence was small.

The constraint of equal repertoire size, $Z' = Z$ can be satisfied with a suitable choice of the shift parameter, s_0, in the prior for differential expression, $\rho_s(s)$, namely $s_0 = -\ln Z'/Z$. 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'.

Acknowledgements

MPT would like to thank M. Pogorelly for providing the R code used to obtain the EdgeR estimates in [12]. This work was supported by the European Research Council Consolidator Grant n. 724208.

[1] Benichou J, Louzoun Y (2011) Rep-Seq : uncovering the immunological repertoire through next-generation sequencing. *Immunology* pp 183–191.

[2] Georgiou G, et al. (2014) The promise and challenge of

high-throughput sequencing of the antibody repertoire. *Nature biotechnology* 32:158–68.

[3] Robins H (2013) Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology* 25:646–652.

[4] Calis JJ, Rosenberg BR (2014) Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in Immunology* 35:581–590.

[5] Vollmers C, Sit RV, Weinstein Ja, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences* 110:13463–8.

[6] Jiang N, et al. (2013) Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Science Translational Medicine* 5:171ra19.

[7] Parameswaran P, et al. (2013) Convergent antibody signatures in human dengue. *Cell Host and Microbe* 13:691–700.

[8] Thomas N, et al. (2014) Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* 30:3181–3188.

[9] Laserson U, et al. (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences* 111:4928–4933.

[10] Wang C, et al. (2015) B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proceedings of the National Academy of Sciences* 112:500–505.

[11] Qi Q, et al. (2016) Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Science Translational Medicine* 8:332ra46–332ra46.

[12] Pogorelyy MV, et al. (2018) Precise tracking of vaccine-responding T-cell clones reveals convergent and personalized response in identical twins. *Proc Natl Acad Sci* 115:12704–12709.

[13] Horns F, Vollmers C, Dekker CL, Quake SR (2019) Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *Proceedings of the National Academy of Sciences of the United States of America* 116:1261–1266.

[14] Logan AC, et al. (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proceedings of the National Academy of Sciences of the United States of America* 108:21194–21199.

[15] Dash P, et al. (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547:89–93.

[16] Glanville J, et al. (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature* 547:94–98.

[17] Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics* 49:659–665.

[18] Britanova OV, et al. (2016) Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *The Journal of Immunology*.

[19] Chu ND, et al. (2019) Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunology* 20:1–12.

[20] Chao DL, Davenport MP, Forrest S, Perelson AS (2004) A stochastic model of cytotoxic T cell responses. *Journal of Theoretical Biology* 228:227–240.

[21] Zilman A, Ganusov VV, Perelson AS (2010) Stochastic models of lymphocyte proliferation and death. *PLoS ONE* 5:1–14.

[22] Boer RJD, Perelson AS (2013) Quantifying T lymphocyte turnover. *Journal of Theoretical Biology* 327:45–87.

[23] Desponds J, Mora T, Walczak AM (2016) Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences* 113:274–9.

[24] Robins H, et al. (2012) Ultra-sensitive detection of rare T cell clones. *Journal of Immunological Methods* 375:14–19.

[25] Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion , with applications to SAGE data. pp 321–332.

[26] Robinson MD, Mccarthy DJ, Smyth GK (2010) edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

[27] Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11:R106.

[28] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15:550.

[29] Mora T, Walczak A (2018) in *Systems Immunology*, eds Das JD, Jayaprakash C (CRC Press), pp 185–199.

[30] Gerritsen B (2018) Ph.D. thesis (Utrecht University).

[31] Greef PCD, et al. (2019) The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *bioRxiv:691501*.

[32] Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R (2013) Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in Immunology* 4:485.

[33] Heather JM, Ismail M, Oakes T, Chain B (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics* 19:554–565.

[34] Qi Q, et al. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America* 111:13139–44.

[35] Lythe G, Callard RE, Hoare RL, Molina-París C (2016) How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology* 389:214–224.

[36] Jenkins MK, Chu HH, McLachlan JB, Moon JJ (2009) On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. *Annual review of immunology* 28:275–294.

[37] Mora T, Walczak AM (2019) How many different clonotypes do immune repertoires contain? ´.

[38] de Boer RJ, Perelson ASA (1993) How diverse should the immune system be. *Proceedings of the Royal Society B: Biological Sciences* 252:171.

[39] Mayer A, Zhang Y, Perelson AS, Wingreen NS (2019) Regulation of T cell expansion by antigen presentation dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 116:5914–5919.

[40] Ralph DK, Matsen FA (2016) Likelihood-Based Inference of B Cell Clonal Families. *PLoS Computational Biology* 12:1–28.

[41] Minervina A, Pogorelyy M, Mamedov I (2019) TCR and BCR repertoire profiling in adaptive immunity. *Transplant International* pp 0–2.

[42] Kadoki M, et al. (2017) Organism-Level Analysis of Vaccination Reveals Networks of Protection across Tissues. *Cell* 171:398–413.e21.

[43] Nourmohammad A, Otwinowski J, Łuksza M, Mora T, Walczak AM (2019) Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1. *Molecular Biology and Evolution* 36:2184–2194.

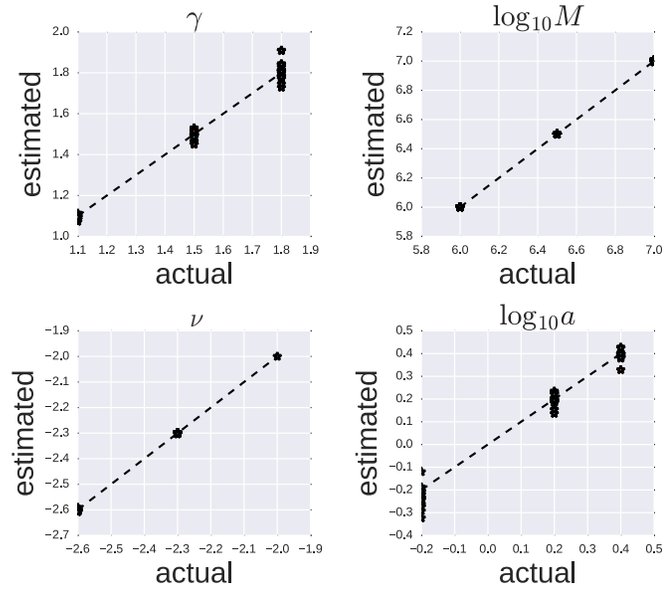

Figure S1: *Reinferring null model parameters.* Shown are the actual and estimated values of the null model parameters used to validate the null model inference procedure over the range exhibited by the data. A 3x3x3x3 grid of points were sampled and results collapsed over each parameter axis. $f_{\min}$ was fixed to satisfy the normalization constraint.

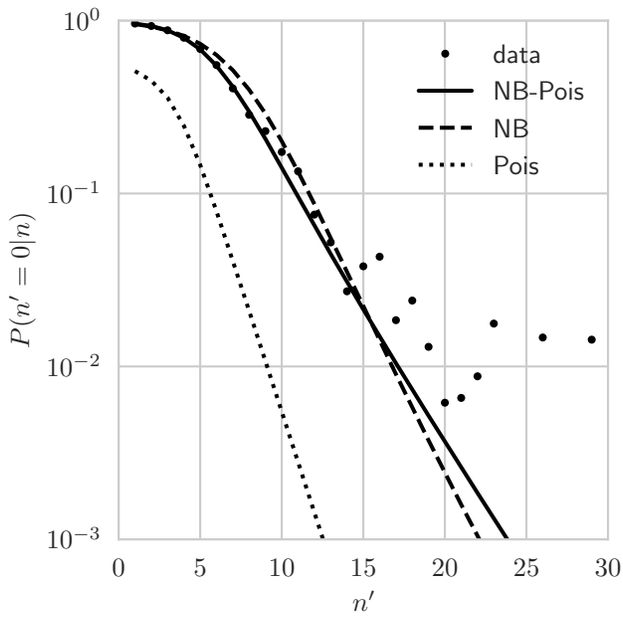

Figure S2: *Dependence of conditional distribution* $P(n' = 0|n)$ *on* n. Two-step negative binomial to Poisson model captures tail better than one-step negative binomial model. Poisson model fits poorly. (Example donor S2-day 0 replicate pair.)

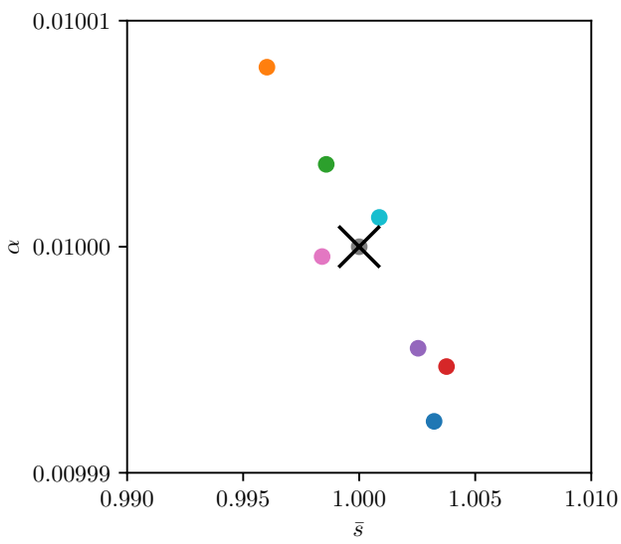

Figure S3: *Reinference of differential expression model for human-sized repertoire.* 10^9 clones were sampled using $N_{\text{read}} = 10^6$ and $(\bar{s}, \alpha) = (1.0, 0.01)$ (cross), and clones with $n = n' = 0$ were removed. Note the orders of magnitude higher precision compared to the synthetic mouse repertoire Fig. 6.

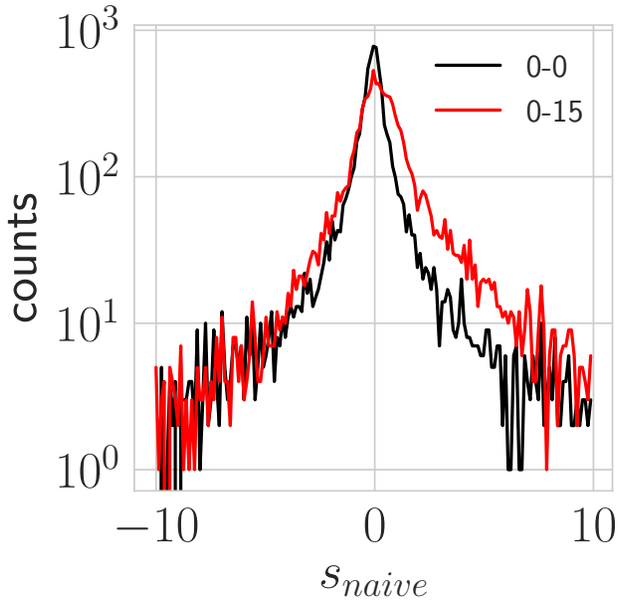

Figure S4: *Empirical histograms of naive log-frequency fold-change.* For example data: day-0/day-0 and day-0/day-15 pair comparisons averaged over donors.

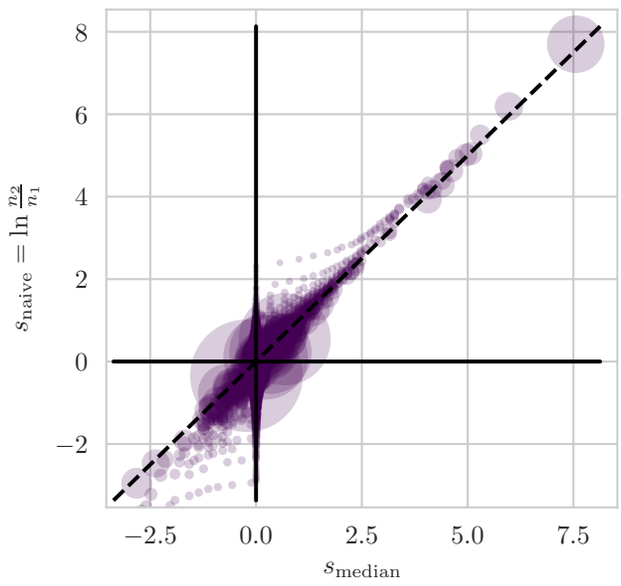

Figure S5: *Summary statistics of log-frequency fold-change posterior distributions.* Comparison of the posterior median log-frequency fold-change and the naive estimate, $\log n'/n$ (across clones with $n, n' > 0$). Each circle is a (n, n') pair with size proportional to pair count average $(n + n')/2$ and intensity proportional to the number of observed clones with that pair.

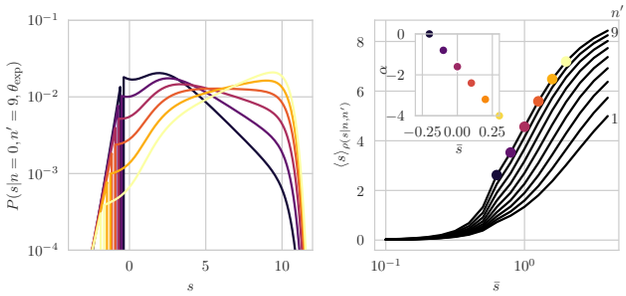

Figure S6: *Competition between ν and $\bar{s}$ in shaping the posteriors, $\rho(s|0,n')$.* A) Posteriors for $n' = 9$ over a range of $(\bar{s}, \alpha)$ pairs spanning the ridge shown in the inset in (B) and Fig. 8 along which the growth of $\bar{s}$ leads to $\rho(f)$ overwhelming $\rho_s(s)$ as the dominant explanation for observed expansion. (B) The posterior mean versus $\bar{s}$ for values of $n' = 1, \ldots, 9$, with the 5 values of $\bar{s}$ used in (A) shown for $n' = 9$.