# Inferring repertoire dynamics from repertoire sequencing

Maximilian Puelma Touzel and Aleksandra Walczak
*Laboratoire de Physique Théorique, ENS-PSL Research University, Paris, France*

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

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 sup-sampling, provides plausible parameter estimates for the unobserved repertoire. Using this replicate model as a baseline, we then formulate the differentially expressed condition using a prior distribution of 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 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.

- background: Next generation sequencing gives access to repertoire wide data containing information that can inform more comprehensive analysis repertoires and more robust vaccine design;
- Problem: quantification of repertoire response dynamics:
- To ongoing, natural stimuli, modelled as a point process of infections, giving rise to diffusion-like response dynamics. Or
- to a single, strong perturbation, such as a vaccine, giving rise to a stereotyped, transient response dynamics
- Conventional approaches
  - DESEQ2
  - Edge R
- Their limitations
- Our approach...

Final Paragraph:

Here, we first quantify the natural variation (either same-day replicate or different day non-perturbation containing) in a learned generative model of receptor RNA count-pair statistics. Next, we then construct a differential expression model by augmenting this null model with the hidden log 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 time-point comparisons provides the repertoire dynamics at ensemble-level of description. Finally, we show how our learned model can be used to infer a posterior probability distribution of fold change for any observed count pair, and thus any specific clone, that can be used to infer the temporal changes of that clone's frequency. 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 over-dispersed, roughly uniform over the range of expansion for singleton clones appearing in the differentially expressed condition. As an example, we use the posterior expansion probability to generate a list of clones significantly expanded by YF vaccination.

## RESULTS

### A repertoire model family learnable using RepSeq

#### *Model family definition*

We consider a family of generative models of pair count statistics of observed immune receptor RNA molecules obtained by sequencing blood samples taken in a reference and test pair of conditions, respectively. For our purpose, an immune repertoire is a finite set of $N$ clones with frequencies $\vec{f} = (f_1, \ldots, f_N)$, over the domain $f_i \in [f_{\min}, 1]$, where $f_{\min}$ is the minimum allowed frequency 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, $Z_f = \sum_{i=1}^{N} f_i$,

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

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 consisting of pair sampling and sequencing. We present a model, $P(n, n', f, f')$, based on the priors $\rho(f)$ and $\rho(f')$, with $f$ and $f'$ and $n$ and $n'$ denoting a clone's frequencies and receptor molecule counts in the reference and test condition, respectively. 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, $N_{\rm obs} \ll N$, of clones, which can still be large since $N$ is typically $10^6$ ($10^9$) for mouse (human). These *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 have no 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 pair counts from observed clones. The model estimate for the total number of clones is then $N = N_{\rm obs}/(1 - P(0, 0))$.

### *Handling normalization*

The $N - N_{\rm obs}$ *unobserved* clones influence the 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 the Methods, we show that $Z_f = 1$ is implicitly satisfied if

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

which also imposes self-consistency between $f_{\rm min}$ and $N$. We employ this constraint in our clone model. Equivalently, it is expressed using the frequency posteriors, which we separate into unobserved and observed contributions,

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

$Z_f$ and $Z_{f'}$ are insensitive to the precise values of a realization's 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. Indeed, we find that parameter learning in such cases is biased **??**. This sensitivity can nevertheless be incorporated into the model by using the 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 define the right-hand side of this realization-dependent constraint

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

and impose that $Z_f^{\mathcal{D}} = 1$. In contrast to learning with 4, learning with 5 leads to unbiased parameter estimates **??**. We impose an equivalent constraint for $\vec{f'}$, via the equivalent condition,

$$Z_{f'}^{\mathcal{D}} = Z_f^{\mathcal{D}} \,. \tag{6}$$

### *Model Assumptions*

Finally, we take the 'common dispersion' approach [1], 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*

Defined models were fit using a pair count dataset, $\mathcal{D} = \{(n_i, n'_i)\}_{i=1}^{N_{\rm obs}}$, by maximizing the log marginal likelihood of the data, $\sum_{i=1}^{N_{\rm obs}} \log P(n_i, n'_i|\theta)$, over the free parameters, $\theta$, subject to the above constraints.

Our method to determine differential expression proceeds in two steps, where in each we define, learn, and analyze an instance of this model family. 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 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. The parameters of $\rho(f)$ and the acquisition model here are set to those of the null model. As a result, $Z_f^{\mathcal{D}}$ is not unity, but in practice we find it is close, and thus so

is $Z^{\mathcal{D}}_{f'}$ on account of the constraint on $\vec{f}'$, eq.6. What is different here is $\rho(f')$: the test frequency, $f'$, is obtained from a transformation of the reference frequency, $f$. This transformation summarizes the effect of the dynamics assumed to act on clone sizes during the time period between the two samples. 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 transient immune response to yellow fever vaccination, we focus on the latter.

## A null model for replicate clone size variation

### *Model Description*

Using this model family, we defined a null model of count statistics and fit it to a pair of replicates, here defined as coming from the same blood sample. This model provides a baseline variability with which differential expression can after be assessed. The marginal pair count 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 $\theta_{\text{null}}$. The influence of $f$ on $P(n|f)$ is an explicit parametrization, e.g. where $P(n|f)$ is a Poisson distribution with mean proportional to $f$ (see Fig.2B). Current methods, e.g. [1], more accurately model $P(n|f)$ by accounting for its observed over-dispersion using a negative binomial distribution.

### *Testing the normalization constraints*

First, and in order to assess the effects of the constraints on the parameter learning, we took a model with $P(n|f)$ distributed as a negative binomial. We developed a sampling protocol for generic instances of this model (see Methods) and used it to validate our inference algorithm (Fig. 11). We used our model to generate a sample set of replicate pair repertoires for a given set of of parameter values. We performed three instances of our parameter learning procedure (See Methods), each subject to different constraints. The results show that the data agnostic constraint in fact bias the estimates, by allowing trials with large clones to create a second, high-bias mode. Learning with the data-specific constraint, by contrast, was insensitive to this effect.

### *Refinements to the acquisition model*

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 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 the 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 size of a clone 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 at low counts, even for the additional dilution of counts due to PCR inefficiency during the sequencing of the sample [2]. We find that, indeed, this two-step model exhibits a better fit to data (see Fig.2B; Supp fig(suppmat figure from misha paper)), especially for clones captured with few counts (see Supp. Fig. 1). 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.

### *Acquisition model parameters*

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, $\theta_{\text{null}}$. For the two-step model, the parameters are $\theta_{\text{null}} = (\alpha, M, a, \gamma, f_{\text{min}})$, where $\alpha$ is the power law exponent, $M$ is the total number of cells, $a$ and $\gamma$ are the coefficient and power of the over-dispersion term in the mean variance relation of the negative binomial distribution of cells, and finally, $f_{\text{min}}$ is the minimum allowed clone frequency. We assessed the ability of the two-step model to capture the observed count pair statistics obtained from the same-day replicates, across days and donors (see supp mat for same results using the NB only model). Figure 3 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. The variability across donors and days is ... indicating a degree of regularity to the natural variability. 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

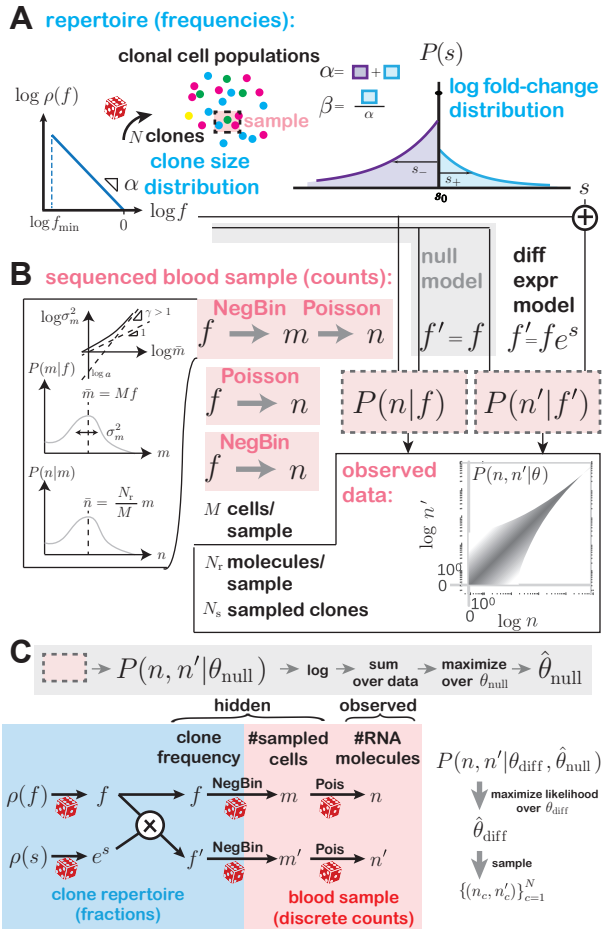


Figure 1. *Repertoire model.* (a) Clone frequency distribution, $\rho(f)$, is set as a power law, parameterized by the power, $\alpha$ and the minimum frequency, $f_{min}$. (b) (c) Differential expression model structure and learning procedure. Null model parameters are learned by marginalizing over clone frequency, $f$, and maximizing this marginal likelihood with respect to the parameters. Sampled repertoires can then be generated using the ML estimate. $P(n|f)$ is set as a negative binomial distribution in cell counts controlling the scale parameter of a Poisson distribution of molecule counts. In the differentially expressed condition (prime-decorated quantities), the parameters, $\theta_{\text{diff}}$, of the prior distribution, $\rho(c)$, of fold-change, $c$, are learned by maximizing the marginal likelihood (see section *Fold change prior*) with respect to $\theta_{\text{diff}}$, keeping the remaining parameters, $\theta_{\text{null}}$, fixed to their ML estimates, $\hat{\theta}_{\text{null}}$, previously obtained using same-day replicate data (see fig. 2).

learned values of $f_{\text{min}}$ give 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 around the optimum, *i.e.* the Fisher information, of the optimization problem without the normalization constraint. Must incorporate constraint in errorbar estimation.

We confirmed that the learned values of the parameters provide a good fit to the pair count statistics via the model's prediction of the conditional and marginal distributions of pair counts, whose correspondence with their empirical counterparts is not required by the fitting procedure (see Fig.4). For example, Fig. 4A 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 does well at capturing the count statistics in one condition when there are no observed counts in the other condition (need to add this plot).

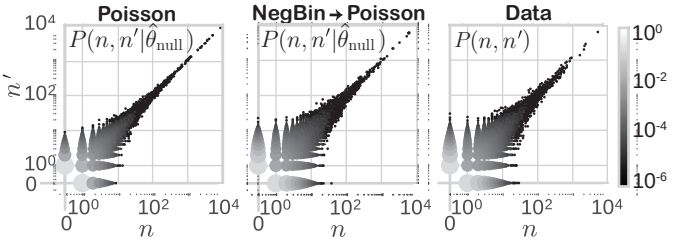


Figure 2. *Learning a null model.* Models and data comparison using molecule pair count statistics (example donor: S2; day 0-day 0 comparison). The learned model for Poisson distributed $P(n|f)$ (left) and for $P(n|f)$ set as a negative binomial distribution in cell counts controlling the scale parameter of a Poisson distribution of molecule counts (center; fig.??). Right: empirical pair count histogram.

*Diversity & observability Inference*

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 measures of diversity of the repertoire producing that sample (see Methods). 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 a uniform distribution of their frequencies; and the Simpson diversity, the expected number of shared clones between two realized repertoires. Discuss the variability here? We note that diversity estimates obtained from observed species abundance are affected by statistical bias [3]. We can also use the model to infer the observed fraction of clones picked up in our sample, relative to the full repertoire in the body, and the fraction of cells they take up. Generally, the estimates say that we catch only about 0.1% of clones, but that these take up to 1/3 of the number of available cells.

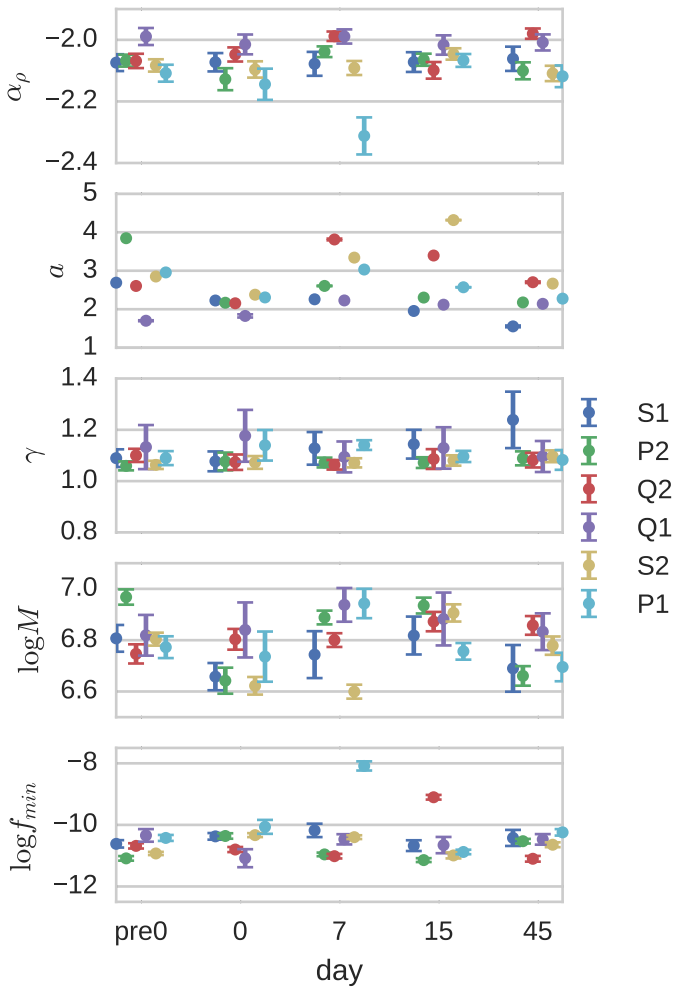


Figure 3. *Learned null model parameters* plotted separately for each donor and time point. Error bars are the inverse standard deviation of a Gaussian approximation around the maximum of the likelihood acting as a lower bound for the variance of the estimates.

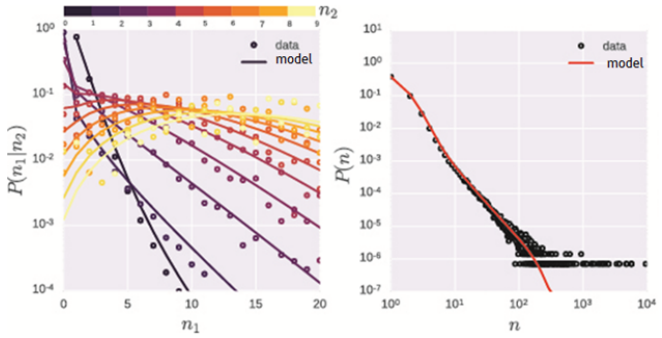


Figure 4. *Null model marginals and conditionals.* The marginal, $P(n_1|\theta_n) = \sum_{n_2} P(n_1, n_2|\theta_n)$ (a), and conditional $P(n_1|n_2\theta_n) = P(n_1, n_2|\theta_n)/P(n_1|\theta_n)$ (b), distributions.Add conditional $P(n|n' = 0)$ and $P(n'|n = 0)$. clean up figure (remove grey etc.)

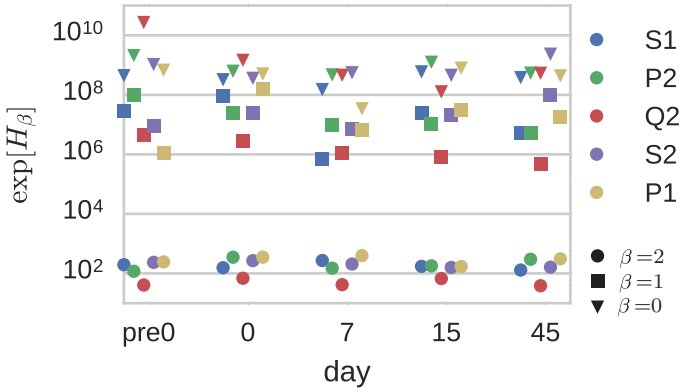


Figure 5. *Diversity estimates.* Shown are diversity estimates obtained from the Renyi entropies, $H_\beta$, of the inferred clone frequency distributions for $\beta = 0$ (estimated total number of clones, $N$), $\beta = 1$ (Shannon entropy) and $\beta = 2$ (Simpson index), across donors and days.

*Pushing the null model into the temporal domain*

Both the intrinsic and driven fluctuations of the population dynamics of immune receptor repertoires imply that clone frequencies will vary across days, even in the absence of any antigen-driven response. If the timescale of this diffusive effect falls within the transient response time, the variability of pair count statistics over this window will be structured by both types of fluctuations. in such a short time window, however, we cannot infer the diffusive component reliably, and we omit this component. Fitting the presented model, which lacks an explicit diffusive component, in the window might inaccurately attribute differential expression to what is actually natural clone size dynamics. However, in the case of our yellow fever dataset, we can learn a pair of null models, each for two time points separated by the same length of time just before and after vaccination, respectively, and compare the variability of each to that of the null model previously learned from same-day replicates (see fig. X). While we find that indeed the null model obtained from separated time points has higher variability, it is significantly lower than the variability of null models learned for pairs of time points with greater time separation. This suggests that our inference of differential expression discussed in the next section is unbiased by the diffusive contribution to the population dynamics.

## Differential expression model

### *model description*

Here, we use 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-fold change, $s$. 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* $s$. As discussed in the Model Description section, 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. Similar to the learning of the null model, here the corresponding marginal pair count distribution is

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

where we fix the null model parameters to their learned values, $\hat{\theta}_{\text{null}}$, and collect the parameters of $\rho(s)$ into $\theta_{\text{diff}}$. We satisfy the normalization constraint on $\vec{f}'$ by introducing into the given $\rho(s)$ an additional shift parameter, $s_0$, set to ensure, $Z^{\mathcal{D}}_{f'} = Z^{\mathcal{D}}_{f}$. Subject to this constraint, we maximize this marginal likelihood over $\theta_{\text{diff}}$ to obtain the optimal $\rho(s)$.

Constraints on cell population dynamics (e.g. division rates) suggest that the transformed frequencies in the differentially expressed condition should be bounded relative to those in the reference condition. In the absence of normalization, they can differ drastically, depending on the $\rho(s)$.

Once learned, the differential expression model provides for any observed clone the posterior distribution of log fold-change conditioned on the clone's observed count pair. It is calculated from the model by marginalizing $f$, and using Bayes' rule,

$$P(s|n, n') = \frac{P(n, n'|s)\rho(s)}{P(n, n')} \,, \tag{9}$$

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

### *Example 1: base prior and mouse vs. human repertoires*

To illustrate the behavior of the differential expression model, we present an application where

$$\rho_{s_0}(s) = \alpha\frac{1}{\bar{s}}e^{\frac{s_0 - s}{\bar{s}}}\Theta(s - s_0) + (1 - \alpha)\delta(s - s_0) \,, \tag{10}$$

(see fig. 6a). This choice describes a differentially expressed condition arising from a stimulus to which some fraction, $\alpha$, of the repertoire expands. Some of these clones respond strongly, most respond weakly, and all together with a characteristic effect size of log 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. We inferred the parameters of the model from model-sampled synthetic $(n, n')$ data for both a small (mouse-like) and large (human-like (still to do!)) synthetic repertoire over a range of biologically plausible parameter values (see Methods for parameter values). In fig. 6b, we show the inference problem of a mouse ($N = 10^6$) repertoire for $(\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$, showing the errors are distributed optimally $\alpha$ and $\bar{s}$, i.e. they are given by the corresponding Fisher information, as expected from the fact that the MLE is efficient. To illustrate the structure of the posteriors computed from these learned models, in fig.6c, we show how the mass in the posteriors moves as we move in orthogonal directions in $(n, n')$ space. In particular, we see for example, that the width of the posterior narrows when counts are both large, and that the model ascribes no fold-change to clones with $n' < n$.

### *Learning $\rho(s)$ on real data*

We find that the regions of high likelihood in the parameter space of $\rho(s)$ do not highlight a particular $(\bar{s}^*, \alpha^*)$ pair, but an entire family of pairs with $\bar{s}^*$ trading off with $\alpha^*$. The bounds of this region are determined by the particular form chosen for $\rho(s)$. For the base prior introduced above, the region extends down to a minimum (maximum) possible $\alpha^*$ ($\bar{s}^*$). No maximum fraction less than 1 is given for $\alpha^*$. results for other choices of prior.

### *Inferring global properties of the differentially expressed repertoire*

Our model incorporates both observed and unobserved parts of the repertoire. As a result it provides estimates for global properties of the response. Namely, the fraction of repertoire that responds, $\alpha$. Obtaining such estimates from existing state-of-the-art approaches,

one might be tempted to use the number of significantly changed clones identified in the sample (say with EdgeR). Normalizing by knowledge of the total number of clones likely under estimates the fraction. Normalizing instead by the number of observed clones (even when adjusted for the presence of large clones (e.g. edgeR's TMM)) is likely to overestimate this fraction of clones.

We can also ask our model to estimate the fraction of the repertoire (e.g. the fraction of cells) that respond.

$$\alpha N \langle f \rangle_{P(f|s>0)} \,. \tag{11}$$

Again, we might be tempted to use the sum of molecule copies of the observed clones determined to have been significantly expanded. And again, we would only have insufficiently accurate estimates of the normalization factor at our disposal: the sample size, or some external estimate. To get a sense of these inaccuracies, we can apply these other approaches to samples from our repertoire model where we know the ground truth. As expected both estimates are off by orders of magnitude (our re-inference of course gets the parameters exactly, though this is expected since it is the model used to generate the date). What can we say about actual data? We show a comparison of the results of edgeR with the range of plausible estimates given by our model. The model gives a range of possible values, but very much more narrow that the range between the naive upper and lower bounds produced by the two estimates accessible using models without an underlying repertoire dynamics.

*Prior solvable via expectation maximization*

In the previous example, we performed a grid search followed by a quadratic approximation 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 perform the latter analytically and then evaluate the former numerically. We choose a symmetric exponential as a tractable prior for this purpose:

$$\rho_{\bar{s}}(s) = e^{-|s|/\bar{s}}/2\bar{s} \tag{12}$$

with $s \in \mathbb{R}$, $\bar{s} > 0$. The expected value of the log likelihood function, often called the Q-function in EM literature, is

$$Q(\bar{s}|\bar{s}') = \sum_{(n,n')\in\mathcal{D}}^{N_{\rm obs}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n,n',\bar{s}') \log\left[P(n,n',s|\bar{s})\right] \,, \tag{13}$$

where $\bar{s}'$ is the current estimate. Maximizing $Q$ with respect to $\bar{s}$ is relatively simple since $\bar{s}$ appears only in

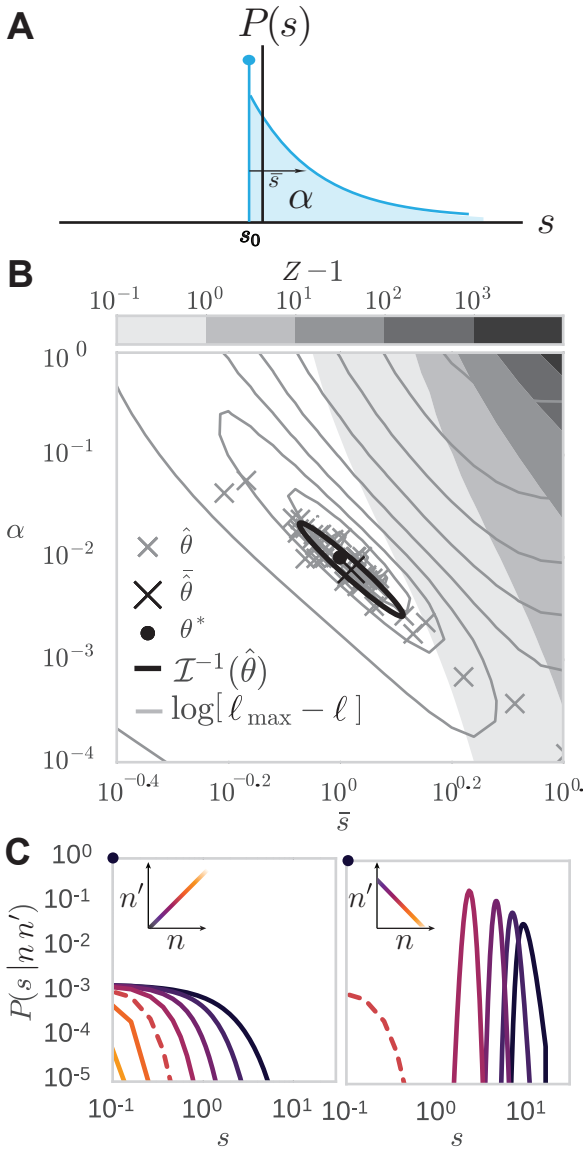


Figure 6. *Inference on synthetic data.*(a) $\rho(s)$, eq. 10, is parametrized by an effect size, $\bar{s}$, describing the expansion of the responding fraction, $\alpha$ of the repertoire. Expansion is relative to the functions center, $s_0$, which is fixed by the homeostatic constraint $\langle f \rangle = \langle f' \rangle$. (b) Inferring $\theta^* = (\bar{s}^*, \alpha^*) = (1.0, 10^{-2})$ (black dot). Maximums of the log-likelihood, $\ell_{\rm max} = \ell(\hat{\theta}_r)$, 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, $\ell_{\rm max}$. 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: $N = 10^6$,$\epsilon = 10^-2$.)

$\rho_{\bar{s}}(s)$ which is a factor in $P(n, n', s|\bar{s})$. For each $s$,

$$\frac{\partial \log\left[\rho_{\bar{s}}(s))\right]}{\partial \bar{s}} = \frac{1}{\rho_{\bar{s}}(s)} \frac{\partial \rho_{\bar{s}}(s)}{\partial \bar{s}} \tag{14}$$

$$= \frac{|s| - \bar{s}}{\bar{s}^2}\,, \tag{15}$$

so that $\frac{\partial Q(\bar{s}|\bar{s}')}{\partial \bar{s}} = \sum_{(n,n')\in\mathcal{D}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n, n', \bar{s}') \frac{\partial \log[\rho_{\bar{s}}(s))]}{\partial \bar{s}} = 0$ implies

$$\sum_{(n,n')\in\mathcal{D}} \int_{-\infty}^{\infty} \mathrm{d}s \rho(s|n, n', \bar{s}') \frac{|s| - \bar{s}^*}{\bar{s}^{*2}} = 0 \tag{16}$$

so that $\bar{s}^* = \frac{1}{N_{\mathrm{obs}}} \sum_{(n,n')\in\mathcal{D}} \bar{s}_{(n,n')}$, where

$$\bar{s}_{(n,n')} = \int_{-\infty}^{\infty} \mathrm{d}s |s| \rho(s|n, n', \bar{s}'). \tag{17}$$

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

$$\rho_{\bar{s}}(s) \leftarrow \rho_{\bar{s}^*}(s). \tag{18}$$

The number of updates typically required for convergence was small.

*Posteriors of log fold-change*

We can explain the shape of the posteriors by breaking up the model components into three groups: $P(n|f)\rho(f)$, 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 $f$ 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 fixed $s$, 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 $\rho(f')$. Which distribution, $\rho(f)$ or $\rho(s)$, dominates the shape of the posteriors then depends on the relative magnitude of their scale parameters. (Reader continues at their own peril).

## I. REST IS WORK IN PROGRESS

Figure 7. *Supp. Fig: Self-consistent reinference of diffexpr model.*

Figure 8. *Analysis of sloppiness of model: description of max-Likelihood manifold and possibly reduced description*

Figure 9. *Evolution of parameters.*

**Ensemble-level application: Time-tracking of ensemble parameters**

**Clone-level application: Identification of responding clones**

Here we infer the posteriors from the learned differential expression model and show their utility by using them to obtain a list of significantly expanded clones as a result of yellow fever vaccination.

Figure 10. *Posteriors.* Some example posteriors. Distributions of slow, smed, shigh, and Pval. Volcano plot.

- discuss posteriors and expansion probability over observed repertoire (Fig. 10):
  - nature of (0,n) posteriors a result of balance of rho(f) and rho(s) priors.
  - The distribution of posterior expansion probability shifts for different priors but maintains the rank of expansion across clones.
- discuss significance tables.
  - How much does the structure of the prior change the order and size of significance tables: e.g. shift/other parameters of prior correlates with size of list, i.e. location of cutoff.
  - ...
- Is there any bias in the method?
  - Sample from model and do ROC-like analysis showing quality of discrimination in tables. E.g. tends to underestimate large fold change.
  - ...

## DISCUSSION

- 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...
  - Nature of over-dispersion and order of NegBin/Pois.
- diffexpr results and discussion:
  - data strongly constrains prior expansion, not contraction. Implications...
  - Shift constraint and the relevance of homeostatis.
- application results and discussion:
  - posterior sensitivity to balance between $\gamma$ (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.
  - Shift constraint and the relevance of homeostatis.
- Clinical use (reference Misha paper)
- drawbacks of approach: need replicate data.

## METHODS

### Null model definition

The normalized clone size, $f$, is distributed according to the probability density function $\rho(f)$, bounded by $f_{\min} \leq f \leq 1$, where $f_{\min}^{-1}$ is an estimate of the total number of lymphocytes in an individual. Based on previous observations (Cite Weinstein 2009, Mora 2010, Mora 2016), $\rho(f)$ is set as a power-law, i.e. $\rho(f) \propto f^{-\gamma}$. We considered four different statistical models of cells and mRNA molecules contained in a sample. These used a negative binomial distribution, with parametrized over-dispersion: $\text{NegBin}(\mu, \sigma^2 = \mu + a\mu^{\beta})$ and a Poisson distribution, $\text{Poisson}(\bar{s})$.

The final and highest scoring model is of cells sampled from a negative binomial and mRNA molecules from a Poisson distribution. A clone of size $f$ appears in a sample containing $M$ lymphocytes on average as $fM$ cells. To account for over-dispersed count statistics, the number of cells is set to be Negative-Binomial distributed with mean $fM$ and variance $fM + a(fM)^{\gamma}$, with $a > 0$ the coefficient and $\gamma > 1$ the power controlling the over-dispersion. For each clone, the number $n$ of detected mRNA molecules (i.e. UMI) is distributed according to a Poisson distribution with mean $mN_{\text{reads}}/M$, where $N_{\text{reads}}/M$ is the average number of UMI per cell, obtained using the observed total number of molecules, $N_{\text{reads}}$.

We inferred the parameters of this null model, $\theta_n = (\alpha, M, a, \gamma)$, from day-0 replicates by maximizing the likelihood of the observed count pairs, $(n, n')$, where $n \sim \text{Poisson}(mN/M)$, $m \sim \text{NegBin}(fM, fM + a(fM)^{\gamma})$, for each replicate, and $f \sim \rho$ is common to both replicates. For a given pair, the likelihood, $P(n, n'|\theta_{\text{null}})$, is obtained by marginalizing over $m, m'$, and $f$.

### 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

$$\mathcal{Z} = \int_{f_{\min}}^{1} \cdots \int_{f_{\min}}^{1} \prod_{i=1}^{N} \rho(f_i)\delta(Z-1)\mathrm{d}^N\vec{f}\,, \tag{19}$$

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 $f_i$,

$$\delta(Z-1) = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{\mu(Z-1)} \tag{20}$$

$$= \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} e^{\mu f_i}\,. \tag{21}$$

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

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} \langle e^{\mu f} \rangle\,, \tag{22}$$

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

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-NI(\mu)}\,. \tag{23}$$

Note that $I(0) = 0$. With $N$ large, this integral is well-approximated by the integrand's value at its saddle point,

located at $\mu^*$ satisfying $I'(\mu^*) = 0$. Evaluating the latter gives

$$\frac{1}{N} = \frac{\langle f e^{\mu^* f} \rangle}{\langle e^{\mu^* f} \rangle} \,. \tag{24}$$

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

### Obtaining diversity estimates from the clone frequency density

For a set of clone frequencies, $\{f_c\}_{c=1}^N$, for a set of clones, the Hill family of diversities are obtained from the Renyi entropies, as $D_\beta = \exp H_\beta$, with $H_\beta = \frac{1}{1-\beta} \ln \left[\sum_c f_c^\beta\right]$. 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, $N_{1,samp} + P(n = 0)N = N$, where $N_{1,samp}$ 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 $1/\left(N\langle f^2 \rangle_{\rho(f)}\right)$.

### 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, and (3) starting with frequencies, sample and use to specify distribution of next random variable in the chain.

In detail, we first fix:

- the model parameters $(\alpha, M, a, \gamma)$, excluding $f_{\min}$. Separate $M$, $a$, and $\gamma$ 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), $\epsilon$. From this we get the effective total sample reads, $N_{\text{reads}} = \epsilon M$, 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:

$$N\langle f \rangle_{\rho(f)} = 1 \tag{25}$$

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 (see appendix for this procedure). 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 $N_{\text{obs}}$ is fixed in the former, while is stochastic in the latter.

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. 11).

### Null Model Inference

Given a data set, $\mathcal{D} = \{(n_i, n_i')\}_{i=1}^{N_{\text{obs}}}$, we infer the parameters of the null model, $\Theta_{\text{null}} = (\alpha, a, \gamma, M, f_{\min})$ by maximizing the marginal likelihood of the observable model, $P(n, n'|n + n' > 0, \Theta_{\text{null}})$, subject to the normalization constraint, $N\langle f \rangle_{\rho(f)} = 1$, where $N = N_{\text{obs}}/(1 - P(0, 0))$. Since in this case we have access to a realization, we could instead normalize conditioned on this realization. Since,

$$N \sum_{(n,n')>0} P(n, n') \approx N \sum_{(n,n')\in\mathcal{D}} \frac{\#(n, n')}{N} \equiv \sum_i^{N_{\text{obs}}}$$

we then have

$$1 = N\langle f \rangle_{\rho(f|\mathcal{D})} = P(0, 0)N\langle f \rangle_{\rho(f|n+n'=0)} + \sum_i^{N_{\text{obs}}} \langle f \rangle_{\rho(f|n_i,n_i')} \,. \tag{26}$$

### Differential expression model definition

We introduce a selection factor $s$ defined as the log-fold change between a clone's frequency on one day, $f$, and that on another, $fe^s$, and define $P(n_1, n_2|s, \theta_n)$ as before, but replacing $f$ by $fe^s$ in the definition of $m_2$. Given a

prior distribution $P(s|\theta_s)$ over $s$ parametrized by a set of parameters $\theta_s$ distinct from $\theta_n$, we used Bayes rule to obtain the posterior log fold-change probability function given an observed count pair, $P(s|n_1, n_2, \theta_n, \theta_s)$. We set the values of the parameters $\theta_s$ of this prior by again maximizing the likelihood of the count pair data given the model over $\theta_s$, $\int P(n_1, n_2|s, \theta_n)P(s|\theta_s)\mathrm{d}s$.

We explored a family of priors expressible as

$$\begin{aligned} P(s|\theta_s) = &\frac{\alpha\beta}{Z_+} e^{-\frac{|s-s_0|}{s_+}} \Theta(s - s_0) \\ &+ \frac{\alpha(1-\beta)}{Z_-} e^{-\frac{|s-s_0|}{s_-}} \Theta(s_0 - s) \\ &+ (1-\alpha)\delta(s - s_0) \ , \end{aligned} \tag{27}$$

with $Z_\pm \sim s_\pm$ (see Fig. **??**) so that in the most general prior, $\theta_s = (s_-, s_+, \alpha, \beta, s_0)$. See table **??** for reduced-parameter versions of this model that we considered.

### Identifying responding clones

In analogy with $p$-values, we used the posterior probability corresponding to the null hypothesis that they are not expanded, $p = P(s \leq s_{thr}|n_1, n_2, \theta_n, \theta_s)$ to rank the clones by the significance of their expansion, using a threshold of $p < 0.025$ and a threshold effect size of $s_{thr}$.

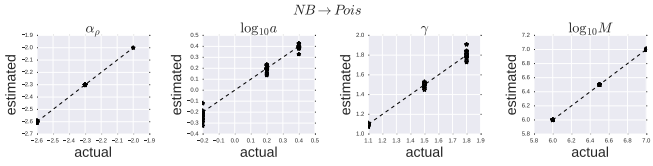


Figure 11. *Reinferring null model parameters.* Shown are the actual and estimated values of the null model parameters used to validate the null model inference procedure over the range exhibited by the data. A 3x3x3x3 grid of points were sampled and results collapsed over each parameter axis. $f_{min}$ was fixed to satisfy the normalization constraint.

### Alternative model sampling procedure

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.

#### *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-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. As a final step, unobserved clones, i.e. those with $(n, n') = (0, 0)$, are discarded.

#### *Effective Sampling*

For an efficient implementation, the procedure should avoid sampling the numerous clones that produce $(n, n') = (0, 0)$, since these are discarded. Such a procedure follows.

First, the $(f, s)$-plane is partitioned into two regions, $D = \{(f, s)|f < f_0, fe^s < f_0\}$ and its complement, $\bar{D}$, with $f_0$ chosen such that clones sampled from $\bar{D}$ are often *observed*, i.e. $n + n' > 0$ (a minority of clones sampled in $\bar{D}$ will nevertheless give $n + n' = 0$; these are discarded). In contrast, frequencies sampled from $D$ will be small, so that most will be unobserved, and we must condition on the clone being observed when sampling from this regime. Moreover, their average is unaffected by the long-tailed behaviour of the distribution in the large-frequency regime and thus is well-approximated by the corresponding ensemble average. We use this latter fact when computing the renormalization of the frequencies of the second condition.

We compute the mass in $D$ as $P_D = \int_f \mathrm{d}f \rho(f) \sum_s \rho(s) \mathbb{1}((f, s) \in D)$.

We sample $(f, s)$ in $\bar{D}$ until the sum of the first condition's frequencies, $\sum_i f_i$ added to the expected sum in $D$, $P_D N_{cl} \langle f \rangle_{P(f|D)}$, equals 1,

$$1 = \sum_{i=1}^{N_{\bar{D}}} f_i + P_D N_{cl} \langle f \rangle_{P(f|D)} \ , \tag{28}$$

where $N_{cl}$ is the total number of clones in the repertoire. The number sampled from $\bar{D}$, $N_{\bar{D}}$, is determined from that expression self-consistently by substituting $N_{cl} = N_{\bar{D}}/(1 - P_D)$ obtained from $N_{\bar{D}} + P_D N_{cl} \equiv N_{cl}$. The normalization for the second condition's frequencies is then

$$Z = \sum_{i=1}^{N_{\bar{D}}} f_i e^{s_i} + P_D N_{cl} \langle f e^s \rangle_{P(f,s|D)} \tag{29}$$

such that the second condition's frequencies are $f'_i = f_i e^{s_i}/Z$. Molecule counts are then sampled from $P(n|f)$ and $P(n'|f')$.

We then sample from $D$ conditioned on the clone being observed, i.e. having produced a finite number of molecules in either of the two conditions. We thus sample $N_D = P(n+n'>0|D)P_D N_{cl}$ clones from $P(f,s|D,n+n'>0)$. To avoid having to sample over the joint distribution of $n$ and $n'$, we condition on the 3 regions of finite counts in both conditions, $(n,0)$, $(0,n')$, and $(n,n')$, in which $n$ and $n'$ can be sampled independently. Note the presence of the normalization factor, $Z$, in

$$P(n+n'>0|D) = \int_f \mathrm{d}f \rho(f) \sum_s \rho(s)(1-P(n=0|f)P(n'=0|f'=fe^s/Z)) \tag{30}$$

and

$$P(f,s|D,n+n'>0) = \frac{\rho(f)\rho(s)(1-P(n=0|f)P(n'=0|f'=fe^s/Z))}{P(n+n'>0|D)P_D} \tag{31}$$

We then concatenate the $N_D$ sampled counts from $D$ and the $N_{\bar{D}}$ sampled counts (with $(n,n')=(0,0)$ realizations discarded) from $\bar{D}$ to obtain the full data set.

### equal frequency constraint

The constraint of equal frequencies in the two compared conditions, $\langle f_1\rangle = \langle f_2\rangle$ can be satisfied with a suitable choice of the shift parameter, $s_0$, in the prior for differential expression, $\rho(s)$. The ensemble average can be evaluated over $p(f,s|n_1+n_2>0)$, where $f_1 = f$ and $f_2 = fe^s$, where

$$P(f,s|n_1+n_2>0) = \frac{\sum_{n_1+n_2>0} P(n_1,n_2,f,s)}{\sum_{n_1+n_2>0}\int \mathrm{d}f \sum_s P(n_1,n_2,f,s)} \tag{32}$$

$$= \frac{\sum_{n_1+n_2>0} P(n_1,n_2,f,s)}{\sum_{n_1+n_2>0}\int \mathrm{d}f \sum_s P(n_1,n_2,f,s)} \tag{33}$$

and using this, $P(f|n_1+n_2>0) = \sum_s P(f,s|n_1+n_2>0)$. The shift enters in $P(n_1,n_2,f,s) = P(n_1|f)P(n_2|f,s)P(f)P_{s_0}(s)$ via $P_{s_0}(s)$. A convenient change of variables $s \leftarrow \Delta s + s_0$ maps $P_{s_0}(s)$ to $P_0(\Delta s)$, upon which

$$\langle fe^s\rangle = \int \mathrm{d}f \sum_{\Delta s} fe^{\Delta s+s_0} P(n_1|f)P(n_2|f,\Delta s+s_0)P(f)P_0(\Delta s) \tag{34}$$

$$= e^{s_0}\int \mathrm{d}f \sum_{\Delta s} fe^{\Delta s} P(n_1|f)P(n_2|f,\Delta s+s_0)P(f)P_0(\Delta s) \tag{35}$$

denoting the remaining integral, $\langle \tilde{fe^s}\rangle$, and performing the same change of variables on $\langle f\rangle$,

$$\langle f\rangle = \int \mathrm{d}f \sum_{\Delta s} fP(n_1|f)P(n_2|f,\Delta s+s_0)P(f)P_0(\Delta s)\ , \tag{36}$$

and so the condition can be written as $s_0 = \ln\langle \tilde{fe^s}\rangle - \ln\langle f\rangle$. To obtain $s_0$ from this implicit equation, we apply an iterative scheme beginning with $s_0 = 0$. We compute $P(n_2|f,\Delta s+s_0)$, and then the latter expression supplies $s_0$ in the next iteration. In practice, we take a bounded range of $\Delta s$ symmetric around 0. Thus, the only factor containing shift information is $P(n_2|f,\Delta s+s_0)$ appearing in both $\langle \tilde{fe^s}\rangle$ and $\langle f\rangle$. However, for correspondence with numerics, the $e^{\Delta s}$ factor must be defined over a shifted domain.

Alternatively, the average can be computed over the data directly, via $\langle fe^s\rangle = \frac{1}{N_{pairs}}\sum_{(n_1,n_2)}\langle fe^s\rangle_{P(f,s|n_1,n_2)}$. The should converge to the analytical result in the limit of many clones (and some other condition?).


## ACKNOWLEDGEMENTS

...would like to acknowledge discussions with ... This work was supported by ...


## AUTHOR CONTRIBUTIONS

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

## ADDITIONAL INFORMATION

The authors declare no competing financial interests.

### Appendix A: Appendixes

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 $P(n_i|f) = \sum_{m_i} P(n_i|m_i)P(m_i|f)$ with $m_i, m_i, n_i, n_i = 0,1,\ldots$ and $i=1,2$. Note that these two conditional distributions differ only in their average number of UMI per cell, $N_i/M$, due to their differing the observed total number of molecules, $N_i$. Together with $\rho(f)$, these distributions form the full joint distribution, which is conditioned on the clone appearing in the sample, i.e. $n_1+n_2>0$

which we denote $C$ for clarity,

$$P(n_1, n_2, f|C) = \frac{P(n_1|f)P(n_2|f)\rho(f)}{1 - \int \mathrm{d}f\rho(f)\mathrm{d}fP(n_1 = 0|f)P(n_2 = 0|f)}, \tag{A1}$$

with the renormalization accounting for the fact that $(n_1, n_2) = (0, 0)$ 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

$$P(q_{x0}|C) = \sum_{n_1>0} \int \mathrm{d}fP(n_1, n_2 = 0, f|C) , \tag{A2}$$

$$P(q_{0x}|C) = \sum_{n_2>0} \int \mathrm{d}fP(n_1 = 0, n_2, f|C) , \tag{A3}$$

$$P(q_{xx}|C) = \sum_{\substack{n_1>0,\\ n_2>0}} \int \mathrm{d}fP(n_1, n_2, f|C). \tag{A4}$$

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

$$P(f|q_{x0}) = \sum_{n_1>0} P(n_1, n_2 = 0, f|C)/P(q_{x0}|C) , \tag{A5}$$

$$P(f|q_{0x}) = \sum_{n_2>0} P(n_1 = 0, n_2, f|C)/P(q_{0x}|C) , \tag{A6}$$

$$P(f|q_{xx}) = \sum_{\substack{n_1>0,\\ n_2>0}} P(n_1, n_2, f|C)/P(q_{xx}|C). \tag{A7}$$

Using the sampled frequency, a pair of number of cells $(m_1, m_2)$ is obtained. For $q_{x0}$, $m_1$ is sampled from $P(m_1|f, n_1 > 0)$ and $m_2$ sampled from $P(m_2|f, n_2 = 0)$ with

$$P(m_1|f, n_1 > 0) = \frac{\sum_{n_1>0} P(m_1, n_1|f)}{\sum_{\substack{n_1>0,\\ m_1}} P(m_1, n_1|f)} , \tag{A8}$$

$$P(m_2|f, n_2 = 0) = \frac{P(m_2, n_2 = 0|f)}{\sum_{m_2} P(m_2, n_2 = 0|f)} , \tag{A9}$$

with $P(m_i, n_i|f) = P(n_i|m_i)P(m_i|f)$, for $i = 1, 2$. Note that by construction here, $m_1 > 0$, since $P(n_1 > 0|m_1 = 0) = 0$. The procedure is similar for frequencies sampled in $q_{0x}$. For frequencies sampled in $q_{xx}$, cell count pairs $(m_1, m_2)$ are sampled from $P(m_1|f, n_1 > 0)$ and $P(m_2|f, n_2 > 0)$, respectively.

Molecule counts for the three quadrants are then sampled as $(n_1, 0)$, $(0, n_2)$, and $(n_1, n_2)$, respectively, with $n_1$ and $n_2$ drawn from the renormalized, finite-count domain of the conditional distributions, $P(n_1|m_1)$ and $P(n_2|m_2)$, respectively, with $m_1 > 0$ and $m_2 > 0$.


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

[2] 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).

[3] Thierry Mora and Aleksandra Walczak, "Quantifying lymphocyte receptor diversity," , 1–10 (2016), arXiv:1604.00487.