MODEL FAMILY

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 of corresponding to a single lymphocyte. A prior density over clone frequencies is given by $\rho(f)$. N and $f_{\min}$ must be determined self-consistently when defining the corresponding joint density,

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

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

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

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, $N_{\text{obs}} \ll N$, 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 $N = N_{\text{obs}}/(1 - P(0,0))$.

The $N - N_{\text{obs}}$ *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 the desired self-consistency between $f_{\min}$ 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 realized set of unobserved clones, and their average frequency is well approximated as the ensemble average in first term above. In contrast, the sum of frequencies of the observed clones might depend on the realization, especially in the case of large, outlying clones arising from power-law distributed clone sizes. This sensitivity can nevertheless be incorporated into the model by using the approximation $\sum_{n+n'>0} P(n, n') \approx \frac{1}{N} \sum_{i=1}^{N_{\text{obs}}}$ so that the second term is $\sum_{i=1}^{N_{\text{obs}}} \langle f \rangle_{\rho(f|n_i, n'_i)}$. We define the right-hand side of this realization-dependent constraint

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

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

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

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

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. 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 immune response to yellow fever vaccination, we focus on the latter.

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

$$\mathcal{Z} = \int_{f_{\min}}^{1} \cdots \int_{f_{\min}}^{1} \prod_{i=1}^{N} \rho(f_i) \delta(Z-1) \mathrm{d}^N \vec{f} \,, \qquad (7)$$

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,

$$\delta(Z-1) = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{\mu(Z-1)} \qquad (8)$$

$$= \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} e^{\mu f_i} \,. \qquad (9)$$

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

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-\mu} \prod_{i=1}^{N} \langle e^{\mu f} \rangle \,, \qquad (10)$$

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

$$\mathcal{Z} = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\mu}{2\pi} e^{-N I(\mu)} \,. \qquad (11)$$

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

$$\frac{1}{N} = \frac{\langle f e^{\mu f} \rangle}{\langle e^{\mu f} \rangle} \,. \qquad (12)$$

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. Thus we see from eq.11 that $\mathcal{Z} = 1$, and so the constraint $N\langle f \rangle = 1$ imposes normalization.