Inferring repertoire dynamics and using them to identify responsive clones

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,

Abridged: Here, we present a method to infer repertoire dynamics from repertoire-sequenced receptor RNA. We analyze the structure of the model used, justifying the ingredients and apply it to answer questions regarding the repertoire dynamics of yellow fever.

Null model ($\mathcal{M}_f$) sampling procedure:
$\mathcal{M}_f$: $P(n, n', f) \propto P(n|f)P(n'|f' = f)\rho(f)\mathrm{d}f\delta(Z_{f(\mathcal{M}_f)} - 1)$

- fix N, total number of clones in repertoire
- fix M, total number of cells in sample
- fix ϵ, sampling efficiency (mapping cells to reads, equivalent to specifying an effective number of reads via $N^{\text{eff}}_{\text{reads}} = \epsilon M$; but since $N^{\text{eff}}_{\text{reads}}$ will differ from the actual number of reads, I prefer to define ϵ)
- fix remaining parameters: α, a and γ.
- set $f_{\min}$ via normalization, $Z_{f(\mathcal{M}_f)} = 1$ with

$$Z_{f(\mathcal{M}_f)} = \langle \sum_{i=1}^{N} f_i \rangle = N\langle f \rangle_{\mathcal{M}_f} = N\langle f \rangle_{\rho(f)} \tag{1}$$

- compute $P(0, 0)$ from model with which $N_{\text{samp}} = N(1 - P(0, 0))$
- sample N_{samp} frequencies. In order to be able to sample n and n' independently, sample in the 3 regions of finite counts proportional to their probabilities, e.g. sample $P_{0x}/Z_x N_{\text{samp}}$ frequencies according to $\rho(f|0x) = \rho(f)P(n = 0|f)(1-P(n' = 0|f))/P_{0x}$ with $P_{0x} = \int \mathrm{d}f\rho(f)P(n = 0|f)(1-P(n' = 0|f))$ and $Z_x = P_{0x}+P_{x0}+P_{xx}$. Similarly for the two other regions $0x$ and xx, where x denotes $n > 0$.

Null model inference procedure:

- get N_{samp}, N_{reads}, N'_{reads} from dataset
- maximize marginal likelihood over $(\alpha, M, a, \gamma, f_{\min})$ subject to constraint that the normalization conditioned on the dataset, $\mathcal{D}$, $Z_{f(\mathcal{M}_f|\mathcal{D})} = 1$ with

$$Z_{f(\mathcal{M}_f|\mathcal{D})} = \frac{N_{\text{samp}}}{1 - P(0, 0)} P(0, 0)\langle f \rangle_{\rho(f|0,0)} + \sum_{i=1}^{N_{\text{samp}}} \langle f \rangle_{\rho(f|n_i,n'_i)} \tag{2}$$

Differential expression model ($\mathcal{M}_{f'}$) sampling procedure:
$\mathcal{M}_{f'}$: $P(n, n', f, s) \propto P(n|f)\ P(n'|f' = fe^s/Z_{f'(\mathcal{M}_{f'})})\ \rho(f)\mathrm{d}f\ P(s)\ \delta(Z_{f'(\mathcal{M}_{f'})} - 1)\ \delta(Z_{f(\mathcal{M}_{f'})} - 1)$

- repeat steps of null model sampling up to the last step (n.b. $Z_{f(\mathcal{M}_{f'})} = Z_{f(\mathcal{M}_f)}$ so step 5 gives us $Z_{f(\mathcal{M}_{f'})} = 1$, however, $P(0, 0)$ depends on $P(s)$ here and can therefore give a different N_{samp}).
- Note that, unlike $\langle f \rangle_{\mathcal{M}_f}$, $\langle fe^s \rangle_{\mathcal{M}_{f'}}$ diverges. We thus normalize instead on a realized repertoire. Sample f and s from $\rho(f)$ and $P(s)$ respectively N times.
- renormalize by $Z = \left(\sum_{i=1}^{N} f_i e^{s_i}\right) \Big/ \left(\sum_{i=1}^{N} f_i\right)$ so $f'_i = f_i e^{s_i}/Z$
- sample $P(n_i|f_i)$ and $P(n'_i|f'_i)$ and discard clones with $n + n' = 0$.

Differential expression inference procedure:

- get N_{samp}, N_{reads}, N'_{reads} from dataset
- use parameters from null inference (n.b. doesn't satisfy the $Z_{f(\mathcal{M}_{f'}|\mathcal{D})} = 1$ constraint).
- maximize marginal likelihood over parameters of $P(s)$, e.g. $\bar{s}$, subject to constraint, $Z_{f'(\mathcal{M}_{f'}|\mathcal{D})} = Z_{f(\mathcal{M}_{f'}|\mathcal{D})}$, with

$$Z_{f'(\mathcal{M}_{f'}|\mathcal{D})} = \frac{N_{\text{samp}}}{1 - P(0, 0)} P(0, 0)\langle fe^s \rangle_{\rho(f,s|0,0)} + \sum_{i=1}^{N_{\text{samp}}} \langle fe^s \rangle_{\rho(f,s|n_i,n'_i)} \tag{3}$$

which we satisfy by adding a shift, s_0, to $P(s)$. This is achieved with a recursive shift-finding procedure.