TnSeq Data Analysis

Sergey Kryazhimskiy

November 29, 2017

Pilot 2017-06

Identifying selected mutations (simple model, no reads)

We will model the selection coefficient X_r for replicate $r = 1, 2, \ldots, n$ of a particular transformant as

$$X_r = \mu + \delta + \varepsilon_r,$$

where μ is the true effect of the transposon-insertion mutation, δ is a defect potentially introduced by transformation, and ε_r is the measurement error. We make the following assumptions about these variables. μ is a parameter (not a random variable). ε_r are all i.i.d r.v.'s, normally distributed with mean 0 and variance σ^2_{err}. δ is an r.v. with a mixture distribution, such that

$$\delta = \begin{cases} 0, & \text{with probability } 1 - p_{\text{tr}}, \\ -\Delta, & \text{with probability } p_{\text{tr}}, \end{cases}$$

where Δ are normally distributed with parameters μ_{tr} and σ^2_{tr}. In other words, with probability p_{tr}, transformation introduces an artifact, which is a normally distributed (typically deleterious) effect; and with probability $1 - p_{\text{tr}}$, no such such artifact is introduced. The conditional probability of observing the selection coefficient X_r in the vicinity of value x_r in replicate experiment r, given that the current transformant's true selection coefficient Y in the given environment, is

$$\mathbb{P}_{X_r|Y}(dx_r) \equiv p\left(x_r|Y\right) dx_r = N_1\left(x_r; Y, \sigma^2_{\text{err}}\right) dx_r,$$

where $N_d\left(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{\Sigma}\right)$ is a d-dimensional multivariate gaussian probability density function with respect to variable $\boldsymbol{x}$, with mean $\boldsymbol{m}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The joint conditional probability density that the selection coefficients $\boldsymbol{X} = (X_1, \ldots, X_n)$ in replicate measurements of a particular transformant in a particular environment are in the vicinity of point $\boldsymbol{x} = (x_1, \ldots, x_r)$, given the transformant's true fitness Y in that environment, is

$$p\left(\boldsymbol{x}|Y; \sigma_{\text{err}}\right) = \prod_{r=1}^{n} N_1\left(x_r; Y, \sigma_{\text{err}}\right) = \frac{\exp\left\{-(n-1)\frac{\text{Var}\,\boldsymbol{x}}{2\sigma^2_{\text{err}}}\right\}}{\sqrt{n\left(2\pi\sigma^2_{\text{err}}\right)^{n-1}}} N_1\left(Y; \bar{x}, \frac{\sigma_{\text{err}}}{\sqrt{n}}\right), \tag{1}$$

where $\bar{x} = \frac{1}{n}\sum_{r=1}^{n} x_r$ and $\text{Var}\,\boldsymbol{x} = \frac{1}{n-1}\sum_{r=1}^{n}\left(x_r - \bar{x}\right)^2$.

Next, we will assume that a transformation artifact occurs with probability p_{tr}. If it occurs, its effect $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$ in environments $e_1, e_2, \ldots, e_K$ is drawn from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{\text{tr}} > 0$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\text{tr}}$. Suppose that the true selection coefficients of the current transformant in environments $e_1, e_2, \ldots, e_K$ are given by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. Then,

$$\mathbb{P}_{\boldsymbol{Y}}(d\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}) = (1 - p_{\text{tr}})\, \delta_{\boldsymbol{\mu}}(d\boldsymbol{y}) + p_{\text{tr}}\; N_K\left(\boldsymbol{y}; \boldsymbol{\mu} - \boldsymbol{\mu}_{tr}, \boldsymbol{\Sigma}_{\text{tr}}\right) d\boldsymbol{y},$$

where $\delta_{\boldsymbol{\mu}}(d\boldsymbol{y})$ is a point measure at $\boldsymbol{\mu}$.

Now consider all replicate measurements of the same transformant in all environments. We have $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1n_1})$ measurements of the selection coefficient of the transformant in environment 1, $\boldsymbol{X}_2 = (X_{21}, \ldots, X_{2n_2})$ measurements in environment 2, etc. up to environment K. The joint probability density of such observation is given by

$$\begin{aligned}
& p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K; \boldsymbol{\mu}, \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}, \boldsymbol{\sigma}_{\text{err}}\right) \\
= & \int_{\mathbb{R}^K} p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K | \boldsymbol{Y}; \boldsymbol{\sigma}_{\text{err}}\right)\; \mathbb{P}_{\boldsymbol{Y}}(d\boldsymbol{y}; \mu, \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}) \\
= & \int_{\mathbb{R}^K} \prod_{k=1}^{K} p\left(\boldsymbol{x}_k | Y_k; \sigma_{\text{err},k}\right)\; \left[(1 - p_{\text{tr}})\, \delta_{\boldsymbol{\mu}}(d\boldsymbol{y}) + p_{\text{tr}}\; N_K\left(\boldsymbol{y}; \boldsymbol{\mu} - \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}\right) d\boldsymbol{y}\right] \\
= & \int_{\mathbb{R}^K} \prod_{k=1}^{K} \frac{\exp\left\{-(n_{\text{rep},k} - 1)\frac{\text{Var}\; \boldsymbol{x}_k}{2\sigma^2_{\text{err},k}}\right\}}{\sqrt{n_{\text{rep},k}\left(2\pi\sigma^2_{\text{err},k}\right)^{n_{\text{rep},k}-1}}} N_1\left(y_k; \bar{x}_k, \frac{\sigma_{\text{err}}}{\sqrt{n_{\text{rep},k}}}\right) \\
\times & \left[(1 - p_{\text{tr}})\, \delta_{\boldsymbol{\mu}}(d\boldsymbol{y}) + p_{\text{tr}}\; N_K\left(\boldsymbol{y}; \boldsymbol{\mu} - \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}\right) d\boldsymbol{y}\right] \\
= & \frac{\exp\left\{-\sum_{k=1}^{K}(n_{\text{rep},k} - 1)\frac{\text{Var}\; \boldsymbol{x}_k}{2\sigma^2_{\text{err},k}}\right\}}{\sqrt{\prod_{k=1}^{K} n_{\text{rep},k}\left(2\pi\sigma^2_{\text{err},k}\right)^{n_{\text{rep},k}-1}}} \int_{\mathbb{R}^K} N_K\left(\boldsymbol{y}; \boldsymbol{M}, \boldsymbol{\Sigma}_{\text{err}}\right) \\
\times & \left[(1 - p_{\text{tr}})\, \delta_{\boldsymbol{\mu}}(d\boldsymbol{y}) + p_{\text{tr}}\; N_K\left(\boldsymbol{y}; \boldsymbol{\mu} - \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}\right) d\boldsymbol{y}\right] \\
= & \frac{\exp\left\{-\sum_{k=1}^{K}(n_{\text{rep},k} - 1)\frac{\text{Var}\; \boldsymbol{x}_k}{2\sigma^2_{\text{err},k}}\right\}}{\sqrt{\prod_{k=1}^{K} n_{\text{rep},k}\left(2\pi\sigma^2_{\text{err},k}\right)^{n_{\text{rep},k}-1}}} \Bigg[(1 - p_{\text{tr}})\, N_K\left(\boldsymbol{\mu}; \boldsymbol{M}, \boldsymbol{\Sigma}_{\text{err}}\right) \\
+ & p_{\text{tr}} \int_{\mathbb{R}^K} N_K\left(\boldsymbol{y}; \boldsymbol{M}, \boldsymbol{\Sigma}_{\text{err}}\right)\; N_K\left(\boldsymbol{y}; \boldsymbol{\mu} - \boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\Sigma}_{\text{tr}}\right) d\boldsymbol{y} \Bigg] \qquad (2)
\end{aligned}$$

where to obtain the last equality we used relationship (1). In equation (2), $\boldsymbol{M} = (\bar{x}_1, \ldots, \bar{x}_K)^T$ and $\boldsymbol{\Sigma}_{\text{err}}$ is a diagonal matrix with $\left(\sigma_{\text{err}}/\sqrt{n_{\text{rep},1}}, \ldots, \sigma_{\text{err}}/\sqrt{n_{\text{rep},K}}\right)$ on the diagonal, with $\bar{x}_k = \frac{1}{n_{\text{rep},k}} \sum_{r=1}^{n_{\text{rep},k}} x_{kr}$ and $\text{Var}\; \boldsymbol{x}_k = \frac{1}{n_{\text{rep},k}-1} \sum_{r=1}^{n_{\text{rep},k}} \left(x_{kr} - \bar{x}_k\right)^2$. Using properties of the multivariate

normal distribution, the integral in equation (2) can be taken analytically, which gives us

$$\begin{aligned} & p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K; \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\Sigma}_{\mathrm{tr}}, \boldsymbol{\sigma}_{\mathrm{err}}\right) \\ = & \frac{\exp\left\{-\sum_{k=1}^{K}(n_{\mathrm{rep},k}-1)\frac{\mathrm{Var}\ \boldsymbol{x}_k}{2\sigma^2_{\mathrm{err},k}}\right\}}{\sqrt{\prod_{k=1}^{K} n_{\mathrm{rep},k}\left(2\pi\sigma^2_{\mathrm{err},k}\right)^{n_{\mathrm{rep},k}-1}}} \\ \times & \left[(1-p_{\mathrm{tr}})\, N_K\left(\boldsymbol{M}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathrm{err}}\right) + p_{\mathrm{tr}} N_K\left(\boldsymbol{M}; \boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\Sigma}_{\mathrm{err}}+\boldsymbol{\Sigma}_{\mathrm{tr}}\right)\right] \end{aligned} \tag{3}$$

Identifying fraction and mean effect of selected mutations

To identify the fraction of beneficial mutations p^{b}, fraction of deleterious mutations p^{n}, and the fraction of deleterious mutations p^{d}, as well as the mean effect of deleterious and beneficial mutations, we will employ the following model. We will assume that, for each mutation m in strain s and environment e, the effect comes from a mixture distribution:

$$\mu_{sem} = \begin{cases} 0, & \text{with probability } p_{\mathrm{n},se}, \\ -X_{\mathrm{d},sem}, & \text{with probability } p_{\mathrm{d},se}, \\ X_{\mathrm{b},sem}, & \text{with probability } p_{\mathrm{b},se}, \end{cases}$$

where $X_{\mathrm{b},sem}$ and $X_{\mathrm{d},sem}$ are exponentially distributed random variables with parameters $\theta_{\mathrm{d},se}$ and $\theta_{\mathrm{b},se}$, respectively. Thus, the probability of observing the selection coefficient $X_{se\ell br}$ is given by

$$\begin{aligned} & \mathbb{P}\left(X_{se\ell br}; \boldsymbol{p}_{se}, \boldsymbol{\theta}_{se}, p_{\mathrm{tr},s}, \mu_{\mathrm{tr},se}, \sigma^2_{\mathrm{tr},se}, \sigma^2_{\mathrm{err},e}\right) = p_{\mathrm{n},se}\mathbb{P}\left(X_{se\ell br}; 0, p_{\mathrm{tr},s}, \mu_{\mathrm{tr},se}, \sigma^2_{\mathrm{tr},se}, \sigma^2_{\mathrm{err},e}\right) \\ + & \ p^{\mathrm{d}}_{se}\int_0^\infty \mathbb{P}\left(X_{se\ell br}; -x, p_{\mathrm{tr},s}, \mu_{\mathrm{tr},se}, \sigma^2_{\mathrm{tr},se}, \sigma^2_{\mathrm{err},e}\right)\ P_{\exp}\left(x; \theta_{\mathrm{d},se}\right)\ dx \\ + & \ p^{\mathrm{b}}_{se}\int_0^\infty \mathbb{P}\left(X_{se\ell br}; x, p_{\mathrm{tr},s}, \mu_{\mathrm{tr},se}, \sigma^2_{\mathrm{tr},se}, \sigma^2_{\mathrm{err},e}\right)\ P_{\exp}\left(x; \theta_{\mathrm{b},se}\right)\ dx \end{aligned} \tag{4}$$

Note on estimating the error distribution

Suppose we have measurements grouped into n groups $X_{11}, X_{12}, ..., X_{1m_1}, X_{21}, \ldots\ldots, X_{nm_n}$. Within group i, all measurements are i.i.d. with mean μ_i. Let us assume that the distributions for each group are the same, except with a shifted mean. Thus, the variance is the same in all groups, σ^2. Think of each group as replicate measurements of the same barcode, and different groups are measurements of the same mutation but different barcodes. The variance comes only from measurement noise, which is the same for all barcodes.

We would like to obtain a non-parameteric estimate of the underlying error distribution. Define $\bar{X}_i$ as the group mean, as usual. Let $Y_{ij} = X_{ij} - \bar{X}_i$. We want to pool all Y_{ij} together to estimate the distribution.

First, we note that $\mathbb{E}Y_{ij} = 0$. Next, we have

$$\begin{aligned}
\text{Var } Y_{ij} &= \mathbb{E}Y_{ij}^2 - \underbrace{\left(\mathbb{E}Y_{ij}\right)^2}_{=0} = \mathbb{E}X_{ij}^2 - 2\mathbb{E}\left(X_{ij}\bar{X}_i\right) + \mathbb{E}\bar{X}_i^2 \\
&= \left[\begin{array}{rcl} \mathbb{E}X_{ij}^2 & = & \sigma^2 + \mu_i^2 \\ \mathbb{E}\bar{X}_i^2 & = & \frac{\sigma^2}{m_i} + \mu_i^2 \\ \mathbb{E}\left(X_i\bar{X}\right) & = & \frac{\sigma^2}{m_i} + \mu_i^2 \end{array}\right] \\
&= \sigma^2\frac{m_i - 1}{m_i}.
\end{aligned}$$

Thus, variables $\tilde{Y}_{ij} = \sqrt{\frac{m_i}{m_i - 1}}Y_{ij}$ will have the same variance, and hence will be i.i.d.