# APPHYS 237 Problem Set 4

**DUE:** 3/9/21

## Problem 1: Measuring the DFE for de novo beneficial mutations, Part II

This problem is a continuation of the barcoded lineage tracking problem from last week's homework, now with some applications to real data.

The file `levy_blundell_etal_2015_barcode_trajectories.txt` contains the raw read count trajectories obtained from one such experiment.[14] In this experiment, half a million barcoded lineages were serially transferred in glucose limited media for 14 days, with bottleneck size of a 256-fold dilution rate ($\Delta t = 8$ generations/day) and a bottleneck size of $N_b \approx 7 \times 10^7$. We'll denote the read count trajectory for an arbitrary barcode $i$ by $R_{i,t}$, and we'll let $D_t = \sum_i R_i$ denote the total sequencing coverage in each timepoint. This defines a corresponding set of read count frequencies

$$\hat{f}_{i,\tau} \equiv \frac{R_{i,\tau}}{D_\tau} \,. \tag{19}$$

Noise in these read count trajectories reflects both the stochastic growth dynamics of the experiment, as well as noise in the data generation process (PCR amplification and sequencing). Levy, Blundell, *et al* argued that this compound process is well approximated by an effective branching process model that connects the read count frequencies at successive sequenced timepoints. In particular, given that we observe a lineage at frequency $\hat{f}_{i,\tau}$, the conditional probability at the next timepoint, $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$, can be approximated by a branching-process-like generating function:

$$H(z|\hat{f}_{i,\tau}) \equiv \int e^{-zf} p(f|\hat{f}_{i,\tau})\, df \approx \exp\left[-\frac{z\hat{f}_{i,\tau}[1 + (X_{i,\tau} - \overline{X}_\tau)\Delta t_\tau]}{1 + z\kappa_\tau/D_\tau}\right] , \tag{20}$$

where $\Delta t_\tau$ is the number of generations between the timepoints, $X_{i,\tau}$ is the fitness of lineage $i$ at timepoint $\tau$, $\overline{X}_\tau$ is the mean fitness of the population at that timepoint ($\overline{X}_\tau \approx \sum_i X_{i,\tau}\hat{f}_{i,\tau}$), and $\kappa_\tau$ is an effective parameter capturing the net effects of genetic drift and measurement noise. As we saw in class, this function is difficult to invert *exactly* to get the probability distribution $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$. But for large $R_{i,\tau+1}$, it can be approximated by the asymptotic expansion,

$$p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau}) \sim \frac{\left[(1 + (X_{i,\tau} - \overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}\right]^{1/4}}{(4\pi\kappa_\tau/D_\tau)^{1/2} f_{i,\tau+1}^{3/4}} \exp\left[-\frac{\left(\sqrt{\hat{f}_{i,\tau+1}} - \sqrt{(1 + (X_{i,\tau} - \overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}}\right)^2}{\kappa_\tau/D_\tau}\right] \tag{21}$$

Both representations of the probability distribution will be useful at different stages of the problem below.

(a) We'll first use the measured data to verify that Eq. 20 is a good approximation. Consider the first timepoint ($\tau = 0$), where few of the lineages will have any beneficial mutations. This means that we can assume that $X_{i,\tau} \approx \overline{X}_\tau \approx 0$. Then consider the set of all lineages

[14] Levy, Blundell, *et al*, (2015), "Quantitative evolutionary dynamics using high-resolution lineage tracking," *Nature* **519**:181–186.

with exactly 50 reads in the first timepoint. By construction, these should all have the same conditional distribution, $p(\hat{f}_{i,1}|\hat{f}_{i,0})$. Use the observed frequencies of these lineages at the next timepoint ($\hat{f}_{i,1}$) to show that the conditional distribution is consistent with the approximation in Eqs. 20 and 21.

**Hint:** consider the empirical generating function, $\hat{H}(z) = \frac{1}{n}\sum_i \exp\left(-z\hat{f}_{i,1}\right)$, evaluated for $z$ near "typical" values of $1/\hat{f}_{i,1}$. (Can you explain why this should be a robust moment to estimate for a positive random variable in a finite sample?) Rearrange Eq. 20 as a linear function of $1/z$, so that you can use linear regression[15] to estimate the slope and intercept.

(b) If we continue to focus on rare mutations (e.g,, $20 \leq R_{i,\tau} \leq 60$), then the vast majority should remain neutral even for $\tau > 0$. We can therefore use the statistics of these neutral lineages to estimate $\kappa_\tau$ and $\overline{X}_\tau$ using the same approach you outlined in (b). Specifically, estimate a separate value of $\kappa_\tau$ and $\overline{X}_\tau$ for lineages with $R_{i,\tau} = 20, \ldots, 60$, and average them together to obtain a single estimate of $\kappa_\tau$ and $\overline{X}_\tau$ for each timepoint. Plot your estimated values as a function of time. What is the estimated fold change in frequency of a neutral lineage over the course of the experiment?

(c) We can now use the fitted values of $\kappa_\tau$ and $\overline{X}_\tau$ (measured for the bulk population) to scan for outlier lineages that acquired a beneficial mutation. To do so, let's imagine that a beneficial mutation with effect $s$ occured in lineage $i$ some timepoint $t < t_\tau$. The lineage frequency at later timepoints can then be split into neutral and beneficial components,

$$\hat{f}_{i,\tau} = \hat{f}^0_{i,\tau} + \hat{f}^s_{i,\tau}\,, \tag{22}$$

where $\hat{f}^0_{i,\tau}$ and $\hat{f}^s_{i,\tau}$ are both described by Eq. 20 with $X_{i,\tau} = 0$ and $X_{i,\tau} = s$, respectively. Derive an exprssion for the generating function of $\hat{f}_{i,\tau+1}$, conditioned on the values of $\hat{f}_{i,\tau}$, $\hat{f}^0_{i,\tau}$, and $\hat{f}^s_{i,\tau}$. What is the effective lineage fitness $X_{i,\tau}$?

Unfortunately, we don't observe the sublineages $\hat{f}^0_{i,\tau}$ and $\hat{f}^s_{i,\tau}$ directly, so we'll have to estimate them from the observed values of $\hat{f}_{i,\tau}$. If the beneficial mutation establishes at time $t_0$, its frequency at later timepoints will be given by

$$f(t|s,t_0) \approx \frac{c}{N_b s} e^{\int_{t_0}^{t}(s-\overline{X}(t'))dt'}\,, \tag{23}$$

where $c$ is an $\mathcal{O}(1)$ constant that depends on the variance in offspring number in the experiment ($c \approx 1.8$ here, see SI p. 11 in Levy, Blundell, *et al* 2012). We can therefore approximate

$$\hat{f}^s_{i,\tau} \approx \begin{cases} 0 & \text{if } t_\tau < t_0 \\ f(t_\tau|s,t_0) & \text{if } f(t_\tau|s,t_0) < \hat{f}_{i,\tau}, \\ \hat{f}_{i,\tau} & \text{else.} \end{cases} \tag{24}$$

This completely specifies the model. The probability of observing a given lineage trajectory, conditioned on $s$ and $\tau$, is given by

$$p(\{\hat{f}_{i,\tau}\}|s,t_0) \approx p(\hat{f}_{i,0}) \prod_\tau p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau},s,t_0)\,. \tag{25}$$

[15] E.g., using the `linregress` function in the SciPy `stats` package.

(d) Parameter estimation can be done with a standard Bayesian approach. Write a formal expression for the posterior probability, $p(s, t_0|\{f_{i,\tau+1}\})$, relative to the posterior probability without a beneficial mutation ($t_0 = \infty$). You may leave your answer as a function of $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau}, s, t_0)$ and the prior probabilities $p_0(s, t_0)$. This ratio is known as the *posterior odds ratio*.

Numerically calculate the posterior odds ratio for trajectory 14 in the data file. For simplicity, we'll discretize $(s, t_0)$ values into a grid with spacing $\delta t_0 = 1$ and $\delta s = 0.005$, and we'll assume a flat prior

$$\frac{p_0(s, t_0)}{p_0(t_0 = \infty)} \approx \begin{cases} c f_0 N_b U_b^0 s \cdot \delta s \cdot \delta t_0 & \text{for } 0 \leq s \leq 0.4 \text{ and } -250 \leq t_0 < 100 \\ 0 & \text{else,} \end{cases} \tag{26}$$

where $f_0$ is the typical frequency of a lineage in the initial pool, and $U_b^0 \sim 10^{-5}$. For which values of $s$ and $t_0$ is the posterior odds ratio the highest? Does this make sense given the shape of the trajectory?

(e) Now use your approach in (d) to estimate $(s, t_0)$ values for the first 1000 trajectories in the experiment. Set $t_0 = \infty$ if the posterior odds ratio is less than one; otherwise take the values of $(s, t_0)$ that maximize the posterior odds ratio. How many beneficial mutations do you detect? Extrapolating the run time from this pilot data, estimate how long it would take your program to process all $\sim$500,000 trajectories in the experiment?

**Bonus:** estimate $(s, t_0)$ values for all $\sim$500,000 trajectories in the experiment.

(f) Finally, we can use your detected beneficial mutations to estimate the distribution of fitness effects, $U_b \rho(s)$. The number of beneficial mutations in an interval $s \pm \delta s$ that establish and rise to detectable frequencies is given by

$$n(s) \approx \left[ N_b \int_0^{t^*(s)} e^{-\overline{X}(t)}\, dt \right] \cdot U_b \rho(s) \delta s \cdot \frac{s}{c} \tag{27}$$

where $t^*$ is the latest the mutation could establish and still perturb the frequency of the lineage. Write an approximate expression for $t^*(s)$, and then rearrange Eq. 27 to write $U_b \rho(s) \delta s$ as a function of the observed values $n(s)$. Plot your estimated DFE using the beneficial mutations you detected in (f).

## Problem 2: Genealogies from sequences of neutral mutations

In class, we saw how we can use coalescent theory to go from genealogies to sequences of neutral mutations. In this problem, we will consider how to go in the opposite direction. Suppose we draw a sample of $n = 6$ individuals from a population and observe mutations at one or more sites. We'll consider a few different imaginary scenarios with $S = 1$, 2, and 3 variable sites.

```
(a) A     (b) AG     (c) AG     (d) AG     (e) AGTG
    A         TC         AG         AG         AGCG
    A         AG         AC         AC         ACCG
    T         TC         TC         TC         TCCA
    T         AG         TC         TC         TCCG
    T         TC         TC         TG         TCCA
```

(a) Draw two genealogies that are consistent with the mutation pattern in (a), assuming that each mutation happens only once ($\mu T_c \ll 1$).

(b) Repeat for pattern (b) above.

(c) Repeat for pattern (c) above.

(d) Try to repeat for pattern (d). Is it possible to draw a consistent genealogy where each mutation happens only once? How is (d) different from (c) and (b), in terms of the number of distinct haplotypes that are observed? (A version of this idea, known as the ***four gamete test*** is frequently used to diagnose recombination or recurrent mutation events in DNA sequence data.)

(e) Draw a genealogy that is consistent with the mutation pattern in (e).

### Problem 3: Sexual vs asexual selection on a highly polygenic trait

Suppose that we create a population by crossing two diverged strains of yeast, and we evolve the resulting hybrid offspring in an environment that selects for higher values of a particular trait. We'll assume that the fitness components of this phenotype are controlled by a large number $L$ of mutational differences between the two strains, each contributing a small fitness effect $\pm s/2$. For simplicity, we'll assume that the positive and negative mutations are evenly distributed between the two parents, and that the recombination rate is sufficinetly high that the different mutations are assigned to offspring independently. Under these assumptions, the variance in fitness of the offspring are normally distributed with mean 0 and variance $V = Ls^2/4$. The goal of this problem is to consider what happens in the so-called ***infinitesimal limit***, where we let $L \to \infty$ and $s \to 0$ while keeping the variance $V = Ls^2/4$ constant. (Formally, we can achieve this by setting $s = \sqrt{V/L}$ and thinking about an asymptotic expansion for large $L$.)

(a) Let's first consider the case where we evolve the hybrid offspring asexually. For simplicity, we'll neglect the possibility of additional mutations in the offspring, so that we essentially have a pooled fitness assay similar to Problem 5 of Problem Set 1. What is the initial rate of fitness increase of the population ($\partial_t \overline{X}$)?

(b) Suppose that the population was founded by a large but finite number of hybrid offspring, so that there is a maximum possible fitness within this initial pool. Some of these lineages will drift to extinction while rare, while others will establish and start to grow to higher frequencies. Calculate the maximum fitness you expect to observe among these established lineages, assuming that the population was founded from $N_0$ hybrid offspring. This gives an estimate of the maximum fitness that can be achieved in an asexual experiment before we have to wait for additional mutations.

*(Hint: see if you can use the same reasoning from Part E in Problem 2 of Problem Set 1.*

(c) Now let's imagine that the evolution step is performed with continual rounds of sexual reproduction, with a sufficiently high rate of recombination that the fitness-influencing sites are effectively unlinked ($r_{ij} \gg \sigma$). How does the mean fitness of the population grow in this scenario? How long do we have to wait before the population reaches the maximum asexual fitness from part (b)? How much do the frequencies of mutations change over this timescale?