Miscellaneous rare mutation stuff

Benjamin H. Good[1]

[1] *Department of Applied Physics, Stanford University, Stanford, CA 94305*

(Dated: December 16, 2020)

This document contains some additional results about rare mutations that don't have a home yet. The basic results concern time-dependent population sizes. This yields some applications to demographic inference, single-locus selection (pN/pS), recombination, and some haplotype homozygosity results.

SITE FREQUENCY SPECTRA FROM A TIME-VARYING POPULATION SIZE

Consider a single-locus branching process model with a time-varying population size:

$$\frac{\partial f}{\partial t} = sf + \sqrt{\frac{f}{N_e(t_p - t)}} \cdot \eta(t) \tag{1}$$

where the function $N_e(\tau)$ gives the population size τ generations before the present day (t_p). We then consider a lineage founded by a single individual at time $t_0 < t_p$, such that

$$f(t_0) = \frac{1}{N_e(t_p - t_0)} \tag{2}$$

The generating function $H(z,t) = \langle e^{-zf(t)} \rangle$ at some later time t satisfies the partial differential equation,

$$\frac{\partial H}{\partial t} = \left[sz - \frac{z^2}{2N_e(t_p - t)} \right] \frac{\partial H}{\partial z}, \quad H(z, t_0) = e^{-z/N_e(t_p - t_0)} \tag{3}$$

This equation can be solved using the method of characteristics. The solution is relatively standard, but we'll just have to be careful about the definition of time. We define a characteristic curve, $z(\tau_R)$, which satisfies

$$\frac{\partial z}{\partial \tau_R} = sz - \frac{z^2}{2N_e(t_p - t + \tau_R)}, \quad z(0) = z, \tag{4}$$

Then the function

$$h(\tau_R) = H(z(\tau_R), t - \tau_R) \tag{5}$$

satisfies the PDE

$$\frac{\partial h}{\partial \tau_R} = 0 \tag{6}$$

and hence

$$H(z,t) \equiv h(0) = h(t - t_0) \equiv H(z(t - t_0), t_0) = e^{-z(t-t_0)/N_e(t_p - t_0)} \tag{7}$$

It remains to solve for $z(\tau_R)$. For future purposes, we will actually solve a more general equation,

$$\frac{\partial z}{\partial \tau_R} = s(t_p - t + \tau_R)z - \frac{z^2}{2N_e(t_p - t + \tau_R)}, \quad z(0) = z, \tag{8}$$

with a time-varying selection coefficient, $s(\tau)$, measured in the same time units as $N_e(t)$. It will then be useful to define a function,

$$\tilde{z}(\tau_R) = z(\tau_R) e^{-\int_0^{\tau_R} s(t_p - t + \tau') d\tau'} \tag{9}$$

Then $\tilde{z}(\tau_R)$ satisfies

$$\frac{\partial \tilde{z}}{\partial \tau_R} = -\tilde{z}^2 \cdot \frac{e^{\int_0^{\tau_R} s(t_p - t + \tau')dt'}}{2N_e(t_p - t + \tau_R)}, \quad \tilde{z}(\tau_R) = z \tag{10}$$

which has the solution,

$$\tilde{z}(\tau_R) = \frac{z}{1 + z\int_0^{\tau_R} \frac{d\tau'}{2N_e(t_p - t + \tau')} e^{\int_0^{\tau'} s(t_p - t + \tau'')d\tau''}} \tag{11}$$

and hence

$$H(z,t) = \exp\left[-\frac{1}{N_e(t_p - t_0)} \cdot \frac{z e^{\int_0^{t-t_0} s(t_p - t + \tau'')d\tau''}}{1 + z\int_0^{t-t_0} \frac{d\tau'}{2N_e(t_p - t + \tau')} e^{\int_0^{\tau'} s(t_p - t + \tau'')d\tau''}}\right] \tag{12}$$

Then evaluating at $t = t_p$ and defining $\tau = t_p - t_0$, we have

$$H(z,\tau) = \exp\left[-\frac{1}{N_e(\tau)} \cdot \frac{z e^{\int_0^{\tau} s(\tau'')d\tau''}}{1 + z\int_0^{\tau} \frac{d\tau'}{2N_e(\tau')} e^{\int_0^{\tau'} s(\tau'')d\tau''}}\right] \tag{13}$$

In the limit that $N_e(\tau)$ is large (i.e., the diffusion limit), the exponential can be safely expanded to first order

$$H(z,\tau) = 1 - \frac{1}{N_e(\tau)} \cdot \frac{z e^{\int_0^{\tau} s(\tau'')d\tau''}}{1 + z\int_0^{\tau} \frac{d\tau'}{2N_e(\tau')} e^{\int_0^{\tau'} s(\tau'')d\tau''}} \tag{14}$$

which can be rewritten in the form

$$H(z,\tau) = 1 \cdot [1 - p_s(\tau)] + p_s(\tau)\left[\frac{1}{1 + z f_s(\tau)}\right] \tag{15}$$

where we have defined two functions,

$$f_s(\tau) = \int_0^{\tau} \frac{d\tau'}{N_e(\tau')} e^{\int_0^{\tau} s(\tau'')d\tau''} \tag{16}$$

and

$$p_s(\tau) = \frac{2 \cdot \frac{1}{2N_e(\tau)} e^{\int_0^{\tau} s(\tau'')d\tau''}}{\int_0^{\tau} \frac{d\tau'}{N_e(\tau')} e^{\int_0^{\tau} s(\tau'')d\tau''}} = \frac{\frac{1}{N_e(\tau)} e^{\int_0^{\tau} s(\tau'')d\tau''}}{f_s(\tau)} \tag{17}$$

This generating function is then easily recognized as a mixture of a delta function at zero and an exponential distribution with mean $f_s(\tau)$:

$$p(f|\tau) = \delta(f)[1 - p_s(\tau)] + p_s(\tau)\frac{1}{f_s(\tau)} e^{-f/f_s(\tau)} \tag{18}$$

Given this solution, the ite frequency spectrum at low frequencies can be written as an integral over the possible ages of a new mutation

$$p(f) = \int_0^{\infty} N_e(\tau)\mu \cdot p(f|\tau)d\tau = \int_0^{\infty} \frac{\mu d\tau}{f_s(\tau)^2} \cdot e^{\int_0^{\tau} s(\tau'')d\tau''} \cdot e^{-f/f_s(\tau)} \tag{19}$$

Now we'll specialize to particular selection coefficients. Let $s(\tau) = -s$. Then we have

$$f_s(\tau) = \int_0^{\tau} \frac{e^{-s\tau'}}{2N_e(\tau')} d\tau' \tag{20}$$

and

$$p(f) = \int_0^\infty \frac{\mu d\tau}{f_s(\tau)^2} \cdot e^{-s\tau} \cdot e^{-f/f_s(\tau)} = \tag{21}$$

For a neutral mutation, this reduces to

$$p(f) = \int_0^\infty \frac{\mu d\tau}{\left[\int_0^\tau \frac{d\tau'}{2N_e(\tau')}\right]^2} e^{-f/\int_0^\tau \frac{d\tau'}{2N_e(\tau')}} \tag{22}$$

which can be compared to the pairwise coalescence time,

$$T_c = \int_0^\infty d\tau \, e^{-\int_0^\tau \frac{d\tau'}{N_e(\tau')}} \tag{23}$$

We an learn a few things from this expression right away. First, the distribution of pages of a mutation is given by

$$p(\tau|f) = \frac{1}{f_s(\tau)^2} e^{-f/f_s(\tau)} d\tau \tag{24}$$

which matches the results of Slatkin 2000 (w/ a slightly different derivation). The most likely age $\tau^*(f)$ is given by

$$\int_0^{\tau^*(f)} \frac{d\tau'}{N_e(\tau)} = f \tag{25}$$

i.e., the time it takes for drift to change the frequency by an amount f. Our formula allows us to generalize this to the case of selection. (discussed more below).

We can also ask whether it is possible to invert $p(f)$ to obtain $N_e(\tau)$. To see this, it is helpful to rewrite the $p(f)$ integral as

$$p(f) = \int_0^\infty 2N_e(\tau)\mu \cdot \frac{1}{2N_e(\tau) f_s(\tau)^2} e^{-f/f_s(\tau)} \, d\tau \tag{26}$$

and then change variables to the "drift timescale"

$$\xi = f_s(\tau) = \int_0^\tau \frac{d\tau'}{2N_e(\tau')} \tag{27}$$

This is a monotonically increasing function of τ and can therefore be inverted. We can think of ξ and measuring time in units of frequency change (the non-equlibrium version of $\tau \sim Nf$). Making this change of variables, we see that the SFS reduces to

$$p(f) = \int_0^\infty N_e(f_s^{-1}(\xi)) \frac{1}{\xi^2} e^{-f/\xi} \, d\xi \tag{28}$$

which looks like the normal exponential frequency distribution weighted by the time-varying probabilities of introducing new mutations backward in time. Making one more change of variables to $u = 1/xi$, this becomes

$$p(f) = \int_0^\infty 2N_e(f_s^{-1}(1/u))\mu \cdot e^{-uf} \, du \tag{29}$$

Normalization is always tricky, so let's rewrite this as

$$q(f) \equiv \frac{p(f)}{\pi} = \int_0^\infty \frac{N_e(f_s^{-1}(1/u))}{T_c} \cdot e^{-uf} \, du \tag{30}$$

where π is the pairwise heterozygosity (easily measurable) and T_c is the pairwise coalescent time scale above. This definition can then be recognized as the Laplace transform of the function,

$$G(u) \equiv \frac{N_e(f_s^{-1}(1/u))}{T_c} \tag{31}$$

Now suppose we use standard inverse Laplace transform techniques to estimate $G(u)$ from $q(f)$ – can we convert this back to an estimate of $N_e(\tau)$? Here's one potential way to do it. From the definition of the frequency time, we have

$$\frac{d\xi}{d\tau} = \frac{1}{2N_e(\tau)} \tag{32}$$

When viewed as a function of ξ, this becomes

$$\frac{d\tau}{d\xi} = 2N_e(f_s^{-1}(\xi)) = 2T_c \cdot G(1/\xi) \tag{33}$$

which can be integrated to obtain τ as a function of ξ:

$$\tau = 2T_c \int_0^{\xi} G(1/u)du \tag{34}$$

Then the implicit curve

$$\left(2\int_0^{\xi} G(1/\xi)d\xi\,, G(\xi)\right) = \left(\frac{\tau}{T_c}\,, \frac{N_e(\tau)}{T_c}\right) \tag{35}$$

traces out $N_e(\tau)/T_c$ as a function of τ/T_c. SO FORMALLY SPEAKING THIS CAN BE DONE. PRACTICALLY SPEAKING, WHAT FREQ RANGES ARE NEEDED TO PERFORM THIS INVERSION?

We can also look at moments like

$$P_n(k) = \left\langle \frac{(nf)^k}{k!} e^{-nf} \right\rangle \tag{36}$$

which gives the probability of observing exactly k alleles in a finite sample of n individuals. Using the results above, we find that

$$\left\langle \frac{(nf)^k}{k!} e^{-nf} \right\rangle = \int_0^{\infty} \frac{\mu e^{-s\tau}}{nf_s(\tau)^2} \left[1 + \frac{1}{nf_s(\tau)}\right]^{-(k+1)} d\tau \tag{37}$$

In particular, for $k = 1$, we have a simple formula for the number of singletons,

$$P_n(1) = \int_0^{\infty} \frac{\mu e^{-s\tau}}{nf_s(\tau)^2} \left[1 + \frac{1}{nf_s(\tau)}\right]^{-(k+1)} d\tau \tag{38}$$

which for neutral mutations yields an expression for Fu and Li's (1993) neutrality statistic,

$$\frac{P_n(1)}{\pi} = \int_0^{\infty} \frac{n}{2T_c} \left[1 + nf_s(\tau)\right]^{-2} d\tau \tag{39}$$

$$= \int_0^{\infty} \frac{n}{2T_c \left[1 + n\int_0^{\tau} \frac{d\tau'}{2N_e(\tau')}\right]^2} d\tau \tag{40}$$

$$= \int_0^{\infty} \left(\frac{n}{2}\right) \frac{d\zeta}{\left[1 + \frac{n}{2}\int_0^{\zeta} \frac{d\zeta'}{\eta_e(\zeta')}\right]^2} \tag{41}$$

where we have defined

$$\eta_e(\zeta) = \frac{N_e(\zeta T_c)}{T_c} \tag{42}$$

Thus, as expected, the fundamental timescale T_c is not identifiable from the shape of the SFS. IS THERE A SIMPLE METHOD TO INVERT THIS TO FIND $\xi(\zeta)$ AS A FUNCTION OF ζ?

For selection, it is clear that for $\tau \ll 1/s$, the selection part goes away. This corresponds to freqs of order

$$f_{\text{sel}} \sim \int_0^{1/s} \frac{d\tau}{2N_e(\tau)} \tag{43}$$

which is a generalization of the "drift barrier" for nonequilibrium demography. Above this point, selection will start to bias the SFS away from neutrality. CAN WE SHOW THIS A LITTLE MORE RIGOROUSLY?

FINITE MUTATION RATES

Similarly, we can consider a finite mutation rate SFS,

$$\frac{\partial f}{\partial t} = \mu + sf + \sqrt{\frac{f}{N_e(t_p - t)}} \cdot \eta(t) \tag{44}$$

which will have a solution

$$H = \exp\left[-\mu \int_0^\infty \frac{z e^{\int_0^\tau s(\tau'')d\tau''}\, d\tau}{1 + z\int_0^\tau \frac{1}{2N_e(\tau')} e^{\int_0^{\tau'} s(\tau'')d\tau''}}\right] \tag{45}$$

Specializing to a $s(\tau) = -s$, we have

$$H(z) = \exp\left[-\mu \int_0^\infty \frac{z e^{-s\tau}\, d\tau}{1 + z\int_0^\tau \frac{e^{-s\tau'}}{2N_e(\tau')}}\right] = \exp\left[-\int_0^\infty 2N_e(\tau)\mu \cdot \frac{z \cdot \left(\frac{e^{-s\tau}}{2N_e(\tau)}\right) d\tau}{1 + z\int_0^\tau \frac{e^{-s\tau'}}{2N_e(\tau')}}\right] \tag{46}$$

whose mean and variance are given by

$$\langle f \rangle = \int_0^\infty e^{-s\tau}\, d\tau = \frac{\mu}{s}\,, \tag{47}$$

$$\mathrm{Var}(f) = \int_0^\infty 2\mu e^{-s\tau} \int_0^\tau \frac{e^{-s\tau'}}{2N_e(\tau')} d\tau' = \frac{\mu}{s}\int_0^\infty \frac{e^{-2s\tau}}{N_e(\tau)}\, d\tau \tag{48}$$

The coefficient of variation is therefore

$$c_V = \frac{\mathrm{Var}(f)}{\langle f \rangle^2} = \int_0^\infty \frac{s e^{-2s\tau}}{N_e(\tau)\mu} \tag{49}$$

which looks a little bit like a harmonic mean of $N_e(t)$ over a timescale of order $1/s$.

APPLICATION TO QLE REGIME

One place this comes up is in the QLE regime, where

$$\frac{\partial f_{AB}}{\partial t} = r f_{Ab} f_{aB} - r f_{AB} + \sqrt{\frac{f_{AB}}{N_e(t_p - t)}} \cdot \eta_{AB}(t) \tag{50}$$

with f_{Ab} and f_{aB} fixed. Plugging into the above formula with $\mu = r f_{Ab} f_{aB}$ and $s = r$, we have

$$\langle f_{AB} \rangle = f_{Ab} f_{aB}\,, \quad \mathrm{Var}(f) = f_{Ab} f_{aB} \int_0^\infty \frac{e^{-2r\tau}}{N_e(\tau)}\, d\tau \tag{51}$$

and

$$c_V = \int_0^\infty \frac{r e^{-2r\tau}}{N_e(\tau) r f_{Ab} f_{aB}} \tag{52}$$

If $c_V \ll 1$, then we are in the full QLE regime. If $c_V \ll 1$, we are in the clonal recombinant regime. In both cases, we have a new result for the decay of the LD curve,

$$\sigma_d^2 = \frac{\langle (f_{Ab} - f_{Ab} f_{aB})^2 \rangle}{\langle f_{Ab} f_{aB} \rangle} = \int_0^\infty \frac{e^{-2r\tau}}{N_e(\tau)}\, d\tau \tag{53}$$

that shows how LD curves decay for different demographies – it's a simple relationship to the laplace transform of the drift function, $1/N_e(\tau)$.

NOTE: THIS APPEARS TO BE STRONG PROOF THAT σ_d^2 CANNOT DECREASE AS A FUNCTION OF FREQUENCY IN THIS REGIME! (IN CONTRAST TO WHAT WE OBSERVE IN DATA!)

TODO: do this same thing for LE statistic? Show that it can't increase as function of data?

Is there an issue w/ denominator?

All this stuff is valid provided that single site frequencies do not change much on the timescale $1/r$ that matters. This occurs when

$$\int_0^{1/r} \frac{f_{Ab} d\tau}{N_e(\tau)} \sim \int_0^{\infty} \frac{f_{Ab} r e^{-r\tau}\, d\tau}{N_e(\tau)} \ll f_{Ab}^2 \tag{54}$$

or

$$\int_0^{\infty} \frac{r e^{-r\tau}}{N_e(\tau) r f_{Ab}} \ll 1 \tag{55}$$

similar to above.

I also started looking at Poisson sampling moments like $\left\langle f_{AB} f_{Ab} f_{aB} e^{-n f_{Ab} - n f_{aB} - f_{AB}} \right\rangle$. Used the finite mutation version. Looked pretty hard because term in exponent sticks around. Is that always small when $2NRf_0 \gg 1$?

HAPLOTYPE HOMOZYGOSITY

We can use this as a model of haplotype homozygosity in the limit that H_1 is low. To be more specific, suppose that U is the rate of producing a new haplotype per individual per generation. We can imagine that this is implemented by $B \gg 1$ possible haplotypes, each of which has a probability U/B of being produced. Then we can model the joint distribution of haplotypes as a collection of independent branching processes of the form

$$\frac{\partial f_i}{\partial t} = \frac{U}{B} - U f_i + \sqrt{\frac{f_i}{N_e(t_p - t)}} \cdot \eta_i(t) \tag{56}$$

which has the same form as above. In the limit that $Bt_0\infty$, we therefore have

$$p(f_i) = \frac{1}{B} \int_0^{\infty} \frac{U e^{-U\tau} d\tau}{f_s(\tau)^2} \cdot e^{-f/f_s(\tau)}, \quad f_s(\tau) = \int_0^{\tau} \frac{e^{-U\tau'}}{2N_e(\tau')}\, d\tau' \tag{57}$$

The haplotype homozygosity is defined by

$$H_1 = \sum_{i=1}^{B} f_i^2 \tag{58}$$

so the mean is

$$\langle H_1 \rangle = B \langle f_i^2 \rangle \tag{59}$$

$$= \int d\tau \frac{U e^{-U\tau}}{f_s(\tau)^2} \int_0^{\infty} f^2 e^{-f/f_s(\tau)} df \tag{60}$$

$$= \int d\tau 2U e^{-U\tau} f_s(\tau) \tag{61}$$

$$= \int_0^{\infty} \frac{e^{-2U\tau}}{N_e(\tau)} \tag{62}$$

Provided that $H_1 \ll 1$, then the size of any single lineage is small and the above will be a good approximation. From our expression, we see that this will be true provided that the harmonic mean of $N_e(\tau)U$ is much greater than one over the time interval $(0, 1/U)$.

For the moment, let's assume that this is true. We can then think about two things: the coefficient of variation of H_1:

$$\mathrm{Var}(H_1) = B \langle f_i^4 \rangle = \int d\tau 24 U e^{-U\tau} f_s(\tau)^3 \tag{63}$$

and hence

$$c_V(H_1) = \left[\frac{\int d\tau 24 U e^{-U\tau} f_s(\tau)^3}{\left[\int_0^\infty 2U e^{-U\tau} f(\tau)\right]^3} \right] H_1 \tag{64}$$

TODO: SHOULD BE ABLE TO SHOW THAT THE PREFACTOR IS GENRALLY $\mathcal{O}(1)$. Thus, when $H_1 \ll 1$, we also have $c_v(H_1) \ll 1$.

CAN WE BREAK THIS DOWN INTO A POISSON PROCESS OF RVS THING LIKE FOR THE NORMAL BP?

We can also think about the probability of a very large fluctuation. In the normal equilibrium, we have $\sim 1/H_1$ lineages each with size of order $\sim H_1$. The probability that a single lineage reaches a size $\gtrsim cH_1$ for $c \gg 1$ is given by

$$\Pr[f \gtrsim cH_1] = \int d\tau \, U e^{-U\tau} \frac{1}{f_s(\tau)} e^{-cH_1/f_s(\tau)} \tag{65}$$

To evaluate this expression, we note that $f_s(\tau)$ monotonically increases with τ and reaches at most of order H_1 for $\tau \gtrsim 1/U$. Thus, we have approximately,

$$\Pr[f \gtrsim cH_1] \sim \frac{1}{f_s(\infty)} e^{-cH_1/f_s(\infty)} \sim \frac{1}{H_1} e^{-c\cdot\mathcal{O}(1)} \tag{66}$$

(WE CAN THINK ABOUT THE H1 FACTOR AS TELLING THE NUMBER OF ATTEMPTS / MULTIPLE HYPOTHESIS TESTING)

This shows that regardless of demography, the probability of having a large fluctuation is bounded by the average homozygosity.