Bacterial pangenome notes

Benjamin H. Good

(Last updated: May 16, 2026)

I. METAPOPULATION MODEL WITH SELECTION AND HGT

We'll start by considering a simple metapopulation model describing the dynamics of a focal gene:

- **Metapopulation structure.** We'll assume that the global population of bacteria (N_g cells) is divided up into a total of D local populations, each of size $N_\ell = N_g/D$.
- **Focal gene frequency.** We'll track the presence or absence of a particular focal gene g in each cell. We'll let $f_{g,i}(t)$ denote the fraction of cells in deme i that g, and we'll let

$$\overline{f}_g(t) = \frac{1}{D}\sum_{i=1}^{D} f_{g,i}(t) \tag{1}$$

  denote the global frequency of gene g across all demes. In the limit of large D, this global gene frequency can approach a steady-state value even when the individual $f_i(t)$ are constantly turning over.
- **Pangenome.** If we consider a larger collection of genes of size $G_{\rm tot}$ (the *pangenome*), and we neglect correlations between genes, then the typical number of genes per cell is given by

$$G_1(t) \approx \sum_{g=1}^{G_{\rm tot}} \overline{f}_g(t) \tag{2}$$

  Similarly, the total number of genes in a sample of size n is given by

$$G_n(t) \approx \sum_{g=1}^{G_{\rm tot}} [1 - (1-\overline{f}_g(t))^n] \tag{3}$$

  while the typical number of gene differences between two cells is given by

$$\Delta G(t) = \sum_{g=1}^{G_{\rm tot}} 2\overline{f}_g(t)[1-\overline{f}_g(t)] \tag{4}$$

  In the limit of a large pangenome (large $G_{\rm tot}$), all three of these quantities can approach a steady-state value even if the individual $\overline{f}_g(t)$ are turning over.
- **Environmental fluctuations and selection.** We'll assume that for each gene g, there are two environmental states that are relevant (E_0^g and E_1^g) with characteristic residence times T_0^g and T_1^g, respectively. The fraction of demes that are in environment E_1^g at any given time is therefore given by

$$p_g \equiv \frac{T_1^g}{T_0^g + T_1^g} \tag{5}$$

We'll assume that the gene provides a fitness benefit s_g in environment E_1^g and a cost $-c_g$ in environment E_0^g. Similar to McInerney (biorxiv 2026.05.08.723712), we will think of c_g as an emergent quantity that depends on the characteristic genome size G_1 in Eq. (2), which will be determined self-consistently at the end. The nominal time- (or population-) averaged selection coefficient can therefore be denoted by

$$\overline{s}_g = s_g p_g - c_g(1-p_g) \tag{6}$$

- **Gene loss.** We'll assume that the focal gene can be lost from the genome at a per capita rate Λ per individual per gene per generation.

- **Gene gain via HGT.** We'll assume that individuals can gain the gene through HGT. For the present, we will parameterize[1] this rate as a function

$$h_g = h_{\rm int}\overline{f}_g + h_{\rm ext} f_g^{\rm ext} \tag{7}$$

which includes the contributions of internal vs gene HGT (i.e. from the same vs different species[2]) As above, it will be useful to view $h_{\rm int}$ as an emergent quantity that depend on the genome size G_1. All else being equal, a larger value of G_1 will lead to a lower effective rate of transfer (e.g. if accessory genes need to be shuttled on a finite number of plasmids with a fixed genome size, or sampled from environmental DNA at a fixed rate).

II. DYNAMICS IN THE MONOMORPHIC DEME LIMIT

One key question is whether environmental variation can maintain the gene in the population, even if it is locally unfavored ($\overline{s}_g < 0$). We can start by thinking about this in the limit where the local populations are sufficiently small that they are almost always fixed for one gene variant or the other (an analogue of the "weak mutation limit" in ordinary population genetics). Such a scenario might seem to reasonably describe the diversity within a single person's microbiome, for example.[3] In this case, it will suffice to keep track of four global variables:

$$\begin{aligned}
P_{11}(t) &\equiv \text{fraction of demes with } E_i(t) = E_1^g \text{ and } f_i(t) \approx 1 \\
P_{10}(t) &\equiv \text{fraction of demes with } E_i(t) = E_0^g \text{ and } f_i(t) \approx 1 \\
P_{01}(t) &\equiv \text{fraction of demes with } E_i(t) = E_1^g \text{ and } f_i(t) \approx 0 \\
P_{00}(t) &\equiv \text{fraction of demes with } E_i(t) = E_0^g \text{ and } f_i(t) \approx 0
\end{aligned} \tag{8}$$

with $\overline{f}_g(t)$ given by

$$\overline{f}_g(t) \approx P_{11}(t) + P_{10}(t) \tag{9}$$

---

[1] Ideally we might like to derive this from the microscopic dynamics of cells within a patch (e.g. migration + subsequent HGT). But we'll skip this for now.

[2] Species here refers to the global population being modeled. One could imagine a variant of the model where this global population refers to a particular subpopulation of a larger species (e.g. human vs mouse E. coli), that still comprises many individual demes.

[3] This might sound a bit weird, since even the nominal population sizes are still large in this case (e.g. $N_\ell \sim 10^{9-12}$), but it may still be self-consistent if the fixation probabilities $p_{\rm fix}(s_g)$ and $p_{\rm fix}(c_g)$ account for local hitchhiking and clonal interference.

In the monomorphic limit, these four fractions evolve as

$$\frac{dP_{11}}{dt} = +N_\ell \cdot h_g(\overline{f}_g) \cdot p_{\text{fix}}(s_g) \cdot P_{01} - N_\ell \cdot \Lambda \cdot p_{\text{fix}}(-s_g) \cdot P_{11} + \frac{1}{T_0^g} P_{10} - \frac{1}{T_1^g} P_{11} \tag{10a}$$

$$\frac{dP_{10}}{dt} = +N_\ell \cdot h_g(\overline{f}_g) \cdot p_{\text{fix}}(-c_g) \cdot P_{00} - N_\ell \cdot \Lambda \cdot p_{\text{fix}}(c_g) \cdot P_{10} + \frac{1}{T_1^g} P_{11} - \frac{1}{T_0^g} P_{10} \tag{10b}$$

$$\frac{dP_{01}}{dt} = +N_\ell \cdot \Lambda \cdot p_{\text{fix}}(-s_g) \cdot P_{11} - N_\ell \cdot h_g(\overline{f}_g) \cdot p_{\text{fix}}(s_g) \cdot P_{01} + \frac{1}{T_0^g} P_{00} - \frac{1}{T_1^g} P_{01} \tag{10c}$$

$$\frac{dP_{00}}{dt} = +N_\ell \cdot \Lambda \cdot p_{\text{fix}}(c_g) \cdot P_{10} - N_\ell \cdot h_g(\overline{f}_g) \cdot p_{\text{fix}}(-c_g) \cdot P_{00} + \frac{1}{T_1^g} P_{01} - \frac{1}{T_0^g} P_{00} \tag{10d}$$

where $p_{\text{fix}}(s)$ is the fixation probability of a mutation with fitness effect s.

A. Steady-state solution with no external HGT or hitchhiking

Another useful limit is one where we assume that there is no external HGT ($h_{\text{ext}} \approx 0$) and local selection is strong enough that the "wrong" gains and losses never fix [$p_{\text{fix}}(-s_g) \approx 0$ and $p_{\text{fix}}(-c_g) \approx 0$]. In this case, we can solve for the steady state solution to Eq. (10) where the time derivatives vanish.

Derivation. From normalization, we have

$$P_{00} = 1 - P_{11} - P_{10} - P_{01} \tag{11}$$

If we add Eqs. (10a) and (10c), the gain and loss events cancel out and we have

$$\frac{1}{T_0^g}(P_{00} + P_{10}) - \frac{1}{T_1^g}(P_{11} + P_{01}) = 0 \implies P_{01} = p_g - P_{11} \tag{12}$$

From Eq. (10b), we have

$$P_{10} = \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\text{fix}}(c_g) T_0^g} P_{11} \tag{13}$$

so that

$$\overline{f} = P_{11} + P_{10} = \left[1 + \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\text{fix}}(c_g) T_0^g}\right] P_{11} \tag{14}$$

and hence

$$P_{10} = \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\text{fix}}(c_g) T_0^g} \left[1 + \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\text{fix}}(c_g) T_0^g}\right]^{-1} \overline{f} \tag{15a}$$

$$P_{01} = p - \left[1 + \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\text{fix}}(c_g) T_0^g}\right]^{-1} \overline{f} \tag{15b}$$

Finally, if we add Eqs. (10a) and (10b), the environmental switching terms cancel out and we have

$$N_\ell h(\overline{f}) p_{\text{fix}}(s_g) P_{01} - N_\ell \Lambda p_{\text{fix}}(c_g) P_{10} = 0 \tag{16}$$

Using our expressions for P_{01} and P_{10} from Eq. (15) then yields

$$N_\ell h(\overline{f}) p_{\rm fix}(s_g) \left(p - \left[1 + \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \right]^{-1} \overline{f} \right) = N_\ell \Lambda p_{\rm fix}(c_g) \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \left[1 + \frac{\frac{T_0^g}{T_1^g}}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \right]^{-1} \overline{f} \tag{17}$$

or, after some algebra:

$$\overline{f}_g = p_g + \frac{1 - p_g}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} - \frac{N_\ell \Lambda p_{\rm fix}(c_g) T_0^g}{N_\ell h_{\rm int} p_{\rm fix}(s_g) T_1^g} \frac{1}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \tag{18}$$

$$\overline{f}_g = 1 - \left[1 - p_g \left(1 - \frac{1}{N_\ell h_g(p_g) p_{\rm fix}(s_g) T_1^g} \right) \right] \frac{N_\ell \Lambda p_{\rm fix}(c_g) T_0^g}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \tag{19}$$

$$\overline{f}_g = \frac{1}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} + \left(p_g - \frac{1}{N_\ell h_{\rm int} p_{\rm fix}(s_g) T_1^g} \right) \frac{N_\ell \Lambda p_{\rm fix}(c_g) T_0^g}{1 + N_\ell \Lambda p_{\rm fix}(c_g) T_0^g} \tag{20}$$

Interpretation. The steady-state frequency depends on two compound parameters:

$$\rho_b \equiv N_\ell \cdot h(p_g) \cdot p_{\rm fix}(s_g) \cdot T_1^g \tag{21}$$

$$\rho_d \equiv N_\ell \cdot \Lambda \cdot p_{\rm fix}(c_g) \cdot T_0^g \tag{22}$$

which describe the ability of natural selection to fix the correct gene state within a single environmental state. A particularly key role is played by ρ_b. If $\rho_b \gtrsim 1$ (i.e. selection tends to recover the gene in a good environment), then $\overline{f}_g$ reduces to a weighted average between a minimum frequency

$$\overline{f}_{g,\rm min} = p_g \left(1 - \frac{1}{\rho_b} \right) \geq 0 \tag{23}$$

and a maximum frequency of one:

$$\overline{f}_g \approx \overline{f}_{g,\rm min} \cdot \left(\frac{\rho_d}{1 + \rho_d} \right) + 1 \cdot \left(\frac{1}{1 + \rho_d} \right) \tag{24}$$

with a weighting factor controlled by ρ_d. In the limit of large ρ_d, the gene frequency approaches the nominal niche frequency

$$\overline{f}_g \approx p_g \tag{25}$$

which is valid when $\rho_d \gg 1/\overline{f}_{g,\rm min} \gtrsim 1$. This shows that (sufficiently slow) environmental fluctuations can maintain the gene in the global population even when $\overline{s}_g < 0$.

On the other hand, if $\rho_b < 1$ (i.e. selection can't always recover the gene in a good environment), then there is a possiblity that selection might not be able to maintain the gene in the population at all. We can search for this point by setting $\overline{f}_g \approx 0$ and solving for the corresponding value of ρ_b. This yields a critical value

$$\rho_b^* = \frac{\rho_d \cdot p_g}{1 + \rho_d \cdot p_g} \tag{26}$$

below which selection can't maintain the gene at all. We can also write this as a characteristic switching timescale,

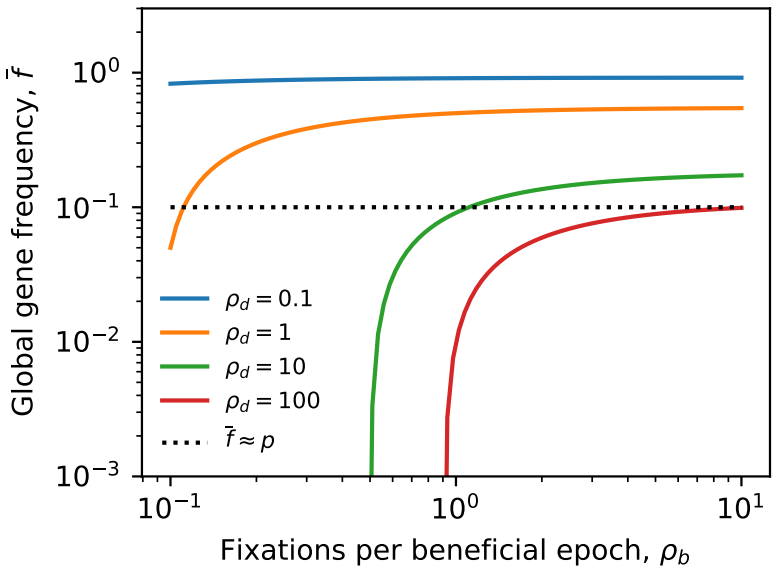
$$(T_0^g + T_1^g)^* = \frac{1}{p_g} \left(\frac{1}{N_\ell h_{\rm int} p_{\rm fix}(s_g) p_g} - \frac{1}{N_\ell \Lambda p_{\rm fix}(c_g)(1 - p_g)} \right) \tag{27}$$

or

$$(T_1^g)^* = \frac{1}{N_\ell h_{\text{int}} p_{\text{fix}}(s_g) p_g} - \frac{1}{N_\ell \Lambda p_{\text{fix}}(c_g)(1-p_g)} \tag{28}$$

below which the gene can't be maintained. We note that in this monomorphic regime, this critical timescale crucially depends on the absolute HGT rate, in addition to the relative values of s_i and T_i. Sufficiently low rates of gene flow can prevent a population from maintaining an occasionally useful gene even when local selection can efficiently act on it (similar to Mazel et al Am Nat 2007). This illustrates the big coordination problem that pangenomes face – how to get the right genes to the right people at the right time.

An example is plotted below for $p_g = 0.1$ and a range of different ρ_b and ρ_d values:



We can think about this in a few different ways. For a fixed value of h_{int} and $T_0^g + T_1^g$, there are certain environmental niches p_g that you can't optimize for. Alternatively, for a fixed value of T_0^g and T_1^g, there are a certain number of genes you can pack into the genome before h_{int} gets too low that you can't maintain those genes in the global population.

We can compare the critical timescale to the time it takes an individual without the gene to acquire one. This is given by

$$T_{\text{reacquire}} = \frac{T_0^g + T_1^g}{N_\ell h(\overline{f}_g) p_{\text{fix}}(s_g) T_1^g} = \frac{1}{N_\ell h(\overline{f}_g) p_{\text{fix}}(s_g) p_g} \tag{29}$$

so in units of $T_{\text{reacquire}}$

$$\frac{(T_0^g + T_1^g)^*}{T_{\text{reacquire}}} = \frac{\overline{f}}{p}\left(1 - \frac{h_{\text{int}} p_{\text{fix}}(s_g) p_g}{\Lambda p_{\text{fix}}(c_g)(1-p_g)}\right) \tag{30}$$

For most reasonable parameters, the rhs is of order one, so this is telling us that the switching time needs to be at least as long as the reacquisition timescale if the gene is to be robustly maintained.