1 The evolutionary trajectory of selection
As we had discussed in class, the allele frequencies and therefore, the site frequency spectrum changes according to the ‘selection trajectory’. Specifically, time since the onset of selection, denoted by $\tau$, and the selection coefficient $s$. By understanding the number of generations since selection, we have a better understanding of the selective sweep.

Selection trajectory

**Input:** The input is a collection of SNP matrices sampled from a hard-sweep with selection coefficient $s$, and time $\tau$ since the onset of selection. You can choose a subset as training data where these parameters are known.

**Output:** Devise an algorithm to predict $\tau$.

**Project goals:** You should come up with a formulation that adequately models the problem, possibly a Markov Chain Monte Carlo approach that samples $\tau$. In the MCMC method, your goal would be to construct a likelihood function that depends upon the data. Other approaches are also possible.

**Primer design papers:** See Ronen [4]

**Data:** You can get SNP matrices at different selection coefficients and $\tau$ from Ali Akbari.

2 Distinguishing causes of non-neutrality
There are many statistics that identify ‘non-neutral’ evolution. However, the non-neutrality can be due to a population bottleneck, due to exponential growth in the population, or due to a selective sweep. Distinguishing between these is an important problem.

Selection trajectory

**Input:** Use a known simulator (Hudson’s ms), mPop, or others to simulate populations according to different demographic scenarios.

**Output:** Devise an algorithm to predict the ‘cause’ of non-neutrality.

**Project goals:** An alternative, but more achievable goal is to run different known tests of neutrality, and prepare a survey on how they are (or are not) confounded by these alternative causes. Your analysis should include an explanation for which tests are most robust, and which are most susceptible to other forms of non-neutrality.

**Primer design papers:** See Ronen [4],

**Data:** Use one of the popgen simulators to produce data.
3 Surveying literature on polygenic adaptation

Provide a critical discussion of different approaches to polygenic selection, including Berg and Coop [1], Turchin [6], and Corona [2].

4 Literature survey on population history estimates

Provide a critical discussion of approaches to estimating historical population sizes including Li and Durbin [3], as well as their more recent work, and Sheehan [5].

References


Appendix: Fixation time for alleles under selection
1 Wright Fisher (WF) or Coalescent model

We have a fixed population of \( N \) individuals going from generation to generation. The individuals have allele \( A \) or \( a \). In each generation, individual chooses a single parent from the previous generation, and adopts their allele.

**Discrete time modeling.** Let \( N_t \) denote the number of copies of allele \( A \) at time \( t \), so that \( N_0 = 1 \). Under selection with parameter \( s \),

\[
\Pr(A) = c(1 + s)\frac{N_{t-1}}{N} \\
\Pr(a) = c \left(1 - \frac{N_{t-1}}{N}\right)
\]

Where the normalizing constant

\[
c = \frac{1}{(1 + s)\frac{N_{t-1}}{N} + \left(1 - \frac{N_{t-1}}{N}\right)} = \frac{1}{1 + \frac{sN_{t-1}}{N}}
\]

Therefore,

\[
E(N_t) = N \Pr(A) = \frac{(1 + s)}{1 + \frac{sN_{t-1}}{N}} N_{t-1}
\]

**Continuous time.** The growth follows a logistic curve. Our goal is to get the rate of growth. In continuous time, suppose that we are sampling individuals at time \( t + \delta t \)

\[
\Pr(A) = c(1 + s)^{\delta t} \frac{N_t}{N} \\
\Pr(a) = c \left(1 - \frac{N_t}{N}\right)
\]

Where the normalizing constant

\[
c = \frac{1}{(1 + s)^{\delta t} \frac{N_t}{N} + \left(1 - \frac{N_t}{N}\right)} \approx \frac{1}{1 + \frac{s\delta t N_t}{N}}
\]

Therefore,

\[
E(N_{t+\delta t}) = N \Pr(A) = \frac{(1 + s\delta t)}{1 + \frac{s\delta t N_t}{N}} E(N_t)
\]

\[
E(N_{t+\delta t}) + \frac{s\delta t N_t}{N} E(N_{t+\delta t}) = E(N_t) + s\delta t E(N_t)
\]

\[
E(N_{t+\delta t}) - E(N_t) = s\delta t \left( E(N_t) - \frac{N_t}{N} E(N_{t+\delta t}) \right)
\]

Or,

\[
\frac{E(N_{t+\delta t}) - E(N_t)}{\delta t} = s \left( N_t - \frac{N_t N_{t+\delta t}}{N} \right)
\]
Note that
\[ \lim_{\delta t \to 0} N_{t+\delta t} = N_t \]
Therefore,
\[ \frac{dE(N_t)}{dt} = s \frac{N_t(N - N_t)}{N} \]
Suppose the population goes from \( N_i \) to \( N_j \) in time \( \tau \). Then,
\[
\int_{t=0}^{\tau} s \, dt = \int_{N_t=N_i}^{N_j} dE(N_t) \left( \frac{1}{N_t} + \frac{1}{N - N_t} \right)
\]
\[ s \tau = \ln \frac{N_i}{N - N_i} \bigg|_{N_i}^{N_j} \]
This gives us the basic logistic regression equation for time as:
\[ \tau = \frac{1}{s} \ln \left( \frac{N_j(N - N_i)}{N_i(N - N_j)} \right) \] (1)
and for population growth as
\[ \tau s = \ln \left( \frac{N_j(N - N_i)}{N_i(N - N_j)} \right) \] (2)
\[ e^{\tau s} = \left( \frac{N}{N_j} - 1 \right) \left( \frac{N}{N_i} - 1 \right) \] (3)
\[ N_t \simeq \frac{N}{1 + Ne^{-ts}} \] (4)
where Eqn. 4 is obtained by choosing \( N_i = 1 \) (population at time 0), and \( N_j = N_t \). We can use Eqn. 1 to compute times to fixation, as well as the time to reach different population milestones.
\[ N_0 = 1 \]
\[ N_1 = \ln n \]
\[ N_2 = \frac{N}{\ln N} \]
\[ N_3 = \frac{N}{2} \]
\[ N_4 = N - \frac{N}{\ln N} \]
\[ N_5 = N - \ln N \]
\[ N_f = N - 1 \]

**Fixation, and other times.** To go from \( N_0 \) to \( N_f \), we have
\[ \tau = \frac{1}{s} \ln(N(N - 1)) = \frac{2}{s} \ln N \]
This time can be split into 3 different epochs: $N_0$ to $N_2$, $N_2$ to $N_4$, and $N_4$ to $N_f$. Let the respective times be $\tau_1, \tau_2, \tau_3$. Then,

$$\tau_1 = \frac{1}{s} \ln \left( \frac{N}{\ln N} \frac{(N - 1)}{(N - \frac{N}{\ln N})} \right) = \frac{1}{s} \ln \left( \frac{N - 1}{\ln N - 1} \right) \simeq \frac{1}{s} \ln N \ln \ln N$$

$$\tau_2 = \frac{1}{s} \ln \left( \frac{(N - \frac{N}{\ln N})\left(N - \frac{N}{\ln N}\right)}{\ln N \ln N} \right) = \frac{2}{s} \ln \ln N$$

$$\tau_3 = \frac{1}{s} \ln \left( \frac{(N - 1)\frac{N}{\ln N})}{(N - \frac{N}{\ln N})} \right) = \frac{1}{s} \ln \left( \frac{N - 1}{\ln N - 1} \right) \simeq \frac{1}{s} \ln N \ln \ln N$$

2 Sampling from a population where the beneficial allele has reached fixation

Consider the case when the beneficial allele is just driven to fixation, so that $N_0 = N$. Sample a pair of individuals. **Q: What is the expected time to coalescence for the pair?** Recall from standard coalescent theory that if the population remains constant at $N$, the time to coalescent is distributed exponentially, with success probability $\frac{1}{N}$. Thus, probability that we coalesce exactly at $t$ generations (assuming continuous time) is

$$P_t = \frac{1}{N} \left( 1 - \frac{1}{N} \right)^t \simeq \frac{1}{N} e^{-\frac{t}{N}}$$

Also, the expected time to coalesce is

$$\int_{t=1}^{t=\infty} \frac{t}{N} e^{-\frac{t}{N}} \, dt = N$$

Similarly, if there are $k$ individuals, the expected time for a coalescence to occur is $\frac{N}{(\frac{N}{k})}$, and the number of mutations created in the time that $k$ lineages become $k-1$ is thus

$$\mu k \frac{N}{(\frac{N}{k})} = \frac{2N\mu}{k-1}$$

This allow us to calculate the total number of polymorphic sites we expect to see as

$$\sum_{k=n}^{2} \frac{2N\mu}{k-1} \simeq 2N\mu \ln n$$

We want to do a similar calculation for the case when the individuals are sampled from the process under selection at a time when the beneficial allele has just fixed in the population. In this case, the population is decreasing as we go back in time. Specifically, using an approach similar to Eqn. 4, if a population of $N_i$ individuals becomes $N_j$ $t$ generations later, then

$$e^{st} = \frac{\frac{N}{N_j} - 1}{\frac{N}{N_i} - 1}$$

$$N_t \simeq \frac{N}{1 + e^{st} / N}$$
The last equation is derived choosing $i = 0$, $N_i = N$, and $N_j = N_t$. Now, the probability that two individuals coalesce at time $\tau$ is

$$P_\tau = \frac{1}{N_\tau} e^{-\int_{t=1}^{\tau} \frac{1}{N_t} dt}$$

(11)

Q: Can we solve this expression using Eqn. 10, and then find the expected time to coalesce for (a) 2 individuals, and (b) $k$ individuals?