Coalescent theory
Statements such as the ones below can be made only if we have an underlying model that suggests what we should expect.
   - Recombination rates vary dramatically across the genome
   - There was a population bottleneck in Iceland

We would like models for populations.

Sometimes, even with a model, it is hard to compute expected values, etc. In this case, we resort to simulations.

We should be able to simulate populations.
Recall that a population sample can be thought of as a binary matrix.
- Rows \((n)\) are individuals. \(n<<N\) (population size)
- Columns are variant sites.

Suppose you are given some parameters about a population (mutation rates, size, time of evolution).

Can you quickly generate a population with those parameters?

What is the model, and how much time would it take?
Wright Fisher Model of Evolution

- Fixed population size from generation to generation
- Random mating
WF model assumptions

- Assumptions (implicit/explicit)
  - Discrete and non-overlapping generations
  - Constant population size (2N haplotypes) across generations
  - All individuals are equally fit.
  - No geographical or social structure. Random mating.
  - No recombination. Each haplotype is identical to its parent except at mutating positions.
  - We also make the infinite sites assumption.
Generating populations

- Forward simulation for generating a population of n<<2N haplotypes:
  - Start with a population of 2N haplotypes (random binary strings)
  - Simulate genealogy for T generations
  - Drop mutation according to fixed rate \( \mu \), each at a new site. (Let \( m \) be the total number of mutations)
  - Generate haplotypes
  - Sample n haplotypes

- How much time will it take to generate a random population? \( O(NTm) \)

- It turns out that this process can be accomplished in \( nm \) steps
Insight 1:
- Separate the genealogy from allelic states (mutations)
- First generate the genealogy (who begat whom)
Insight 2:
- Much of the genealogy is irrelevant, because it disappears.
- Better to go backwards
Coalescent approximation

- **Insight 3:**
  - Topology is independent of coalescent times
  - If you have $n$ individuals, generate a random binary topology
    - Iterate (until one individual)
      - Pick a pair at random, and coalesce

- **Insight 4:**
  - To generate coalescent times, there is no need to go back generation by generation
A brief digression on common distributions

- Exponential distribution
- Poisson distribution
Exponential: Consider the case of tossing coins until you first see HEADS.

- Let Probability [Heads]=p,
- Let q=1-p

Q: Number of steps to success?

\[
\begin{align*}
\Pr[\text{success in step 1}] &= p \\
\Pr[\text{success in step 2}] &= qp \\
\Pr[\text{success in step 3}] &= q^2p \\
\Pr[\text{success in step } k] &= q^{k-1}p
\end{align*}
\]
Expectation

\[ E(\# \text{ steps}) = p + 2qp + 3q^2p + \ldots + kq^{k-1}p + \ldots \]

\[ E(\# \text{ steps}) = \frac{1}{p} \]

Continuous case,

\[ E(\# \text{ steps}) = \int_{t=0}^{\infty} tq^t p \ dt = \int_{t=0}^{\infty} te^{-pt} p \ dt = \frac{1}{p} \]

HW: Show that variance \( \frac{1}{p^2} \)
Poisson distribution

- Ex: Throw darts at a line so that every unit interval has an average of $\lambda$ darts.
- $P[k] = \text{Pr}[\text{Interval has exactly } k \text{ darts}]$?

$$P_\lambda[k] = \frac{e^{-\lambda} \lambda^k}{k!}$$

- Mean: $\lambda$
- Var.: $\lambda$
Coalescent theory (Kingman)

- Input
  - (Fixed population (2N haploid individuals), random mating)

- Consider a sample of 2 individuals.
  - Probability that they coalesce in the previous generation (have the same parent) = \frac{1}{2N}

- Probability that they do not coalesce after \( t \) generations =

\[
\left(1 - \frac{1}{2N}\right)^t \approx e^{-t/2N}
\]
Coalescent theory

- Consider $k$ individuals.
  - Probability that no pair coalesces after 1 generation
    \[ \prod_{i=1}^{k-1} \left( \frac{2N - i}{2N} \right) = \prod_{i=1}^{k-1} \left( 1 - \frac{i}{2N} \right) \approx \left( 1 - \frac{k}{2N} \right) \]
  - Probability that no pair coalesces after $t$ generations
    \[ \left( 1 - \frac{k}{2N} \right)^t \approx e^{-\frac{k}{2N}t} = e^{-\left( \frac{k}{2} \right)t} \]
Coalescent approximation

- At any step, there are $1 \leq k \leq n$ individuals
- To generate time to coalesce ($k$ to $k-1$ individuals)
  - Pick a number from exponential distribution with rate $\frac{k(k-1)}{2}$
  - Mean time to coalescence

$$\text{Time to coalescence}$$

$$\text{for } k \text{ individuals } = \frac{2N}{\binom{k}{2}}$$
Typical coalescents

- 4 random examples with n=6 (Note that we do not need to specify N. Why?)

- Expected time to coalesce?

\[
E \left[ \sum_{k=2}^{n} T(k) \right] = \sum_{k=2}^{n} E[T(k)] = 2N \sum_{k=2}^{n} \frac{2}{k(k-1)}
\]

\[
= 4N \sum_{k=2}^{n} \left( \frac{1}{k-1} - \frac{1}{k} \right) = 4N \left( 1 - \frac{1}{n} \right)
\]
Coalescent properties

- Expected time for the last step
  \[= 2N\]
- The last step is half of the total time to coalesce
- Studying larger number of individuals does not change numbers tremendously
- EX: Number of mutations in a population is proportional to the total branch length of the tree
  - \[E(T_{tot})\]

\[
E[T_{tot}(n)] = E\left[\sum_{k=2}^{n} kT(k)\right] = \sum_{k=1}^{n-1} \frac{2}{k} \rightarrow 2(\gamma + \log n),
\]

\[2N\]
Coalescent properties

- A significant fraction of the SNPs are ‘ancient’
- The time to MRCA is not sensitive to sample size
- Pick a sample of size $n$. Does it contain the MRCA of the entire population?
Sample MRCA versus true MRCA

- Proof sketch:
- Let $x$ be the fraction of individuals on the left side of the tree.
- By symmetry, $x$ is uniformly distributed in $[0..1]$ (formal proof required)

\[
\lim_{N \to \infty} \binom{N}{n} \frac{x^n}{N} = x^n
\]

\[
\Pr \left[ \text{Sample MRCA is not true MRCA} \right] = \int x^n + (1 - x)^n \, dx = \frac{2}{n + 1}
\]
EXPONENTIALLY GROWING POPULATION
Variants (exponentially growing populations)

- If the population is growing exponentially, the branch lengths become similar, or even star-like. Why?
- With appropriate scaling of time, the same process can be extended to various scenarios: male-female, hermaphrodite, segregation, migration, etc.
SIMULATING POPULATIONS USING THE COALESCENT
Simulating population data

- Generate a coalescent (Topology + Branch lengths)
Simulating population data

- Generate a coalescent (Topology + Branch lengths)
- For each branch length $t$, drop mutations with rate $\mu t$
- Based on infinite sites, each mutation is at a unique location
Simulating population data

- Generate Sequences

```
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Coalescent theory: example

- Ex: ~1400bp at Sod locus in Dros.
  - 10 taxa
  - 5 were identical. The other 5 had 55 mutations.
  - Q: Is this a chance event, or is there selection for this haplotype.
Coalescent application

- 10000 coalescent simulations were performed on 10 taxa.
- 55 mutations on the coalescent branches
- Count the number of times 5 lineages are identical
- The event happened in 1.1% of the cases.
- Conclusion: selection, or some other mechanism explains this data.
Looking at lineage specific mutations might help discard the candelabra model. How?

How do we decide between the multi-regional and Out-of-Africa model? How do we decide if the ancestor was African?
Human Samples

- We look at data from human samples
  - 3 populations were sampled at multiple regions spanning the genome
    - 54 regions (Average size 250Kb)
    - SNP density 1 over 2Kb
    - 90 Individuals from Nigeria (Yoruban)
    - 93 Europeans
    - 42 Asian
    - 50 African American
Population specific recombination

- D’ was used as the measure between SNP pairs.
- SNP pairs were classified in one of the following
  - Strong LD
  - Strong evidence for recombination
  - Others (13% of cases)
- Plot shows fraction of pairs with strong recombination (low LD)
- This roughly favors out-of-africa. A Coalescent simulation can help give confidence values on this.

*Gabriel et al., Science 2002*
Coalescent theory applications

Box 2 | Likelihood for trees

Basic statistics makes the distinction between phylogenetic and coalescent approaches apparent. The fundamental equation for likelihood inference in phylogenetics is

\[ L = P(D | G, \mu), \quad (1) \]

where \( L \) is the likelihood (the probability of the data, given the parameters), \( D \) is the data (typically DNA sequences), \( G \) is the tree and \( \mu \) is the collection of parameters in the mutation model. The objective of the analysis is to estimate the parameter \( G \).

The analogous equation in the coalescent setting is

\[ L = \sum_G \left[ P(D | G, \mu) \cdot P(G, \alpha) \right], \quad (2) \]

where \( \alpha \) is the collection of parameters (such as population sizes and migration rates) for the population process. The objective of the analysis is typically to estimate these parameters. The tree or genealogy, \( G \), is a so-called nuisance parameter, which we remove by averaging the likelihood over all possible values.

- Coalescent simulations allow us to test various hypothesis. The coalescent/ARG is usually not inferred, unlike in phylogenies.
Coalescent with Recombination

- An individual may have one parent, or 2 parents.
- The evolutionary history is not a tree, but an ancestral recombination graph (ARG)
ARG: Coalescent with recombination

- Given: mutation rate $\mu$, recombination rate $r$, population size $2N$ (diploid), sample size $n$.
- How can you generate the ARG (topology + branch lengths) efficiently?
- How will you generate sequences for $n$ individuals?
- Given sequence data, can you reconstruct the ARG (topology)
Recombination

- Define $r$ as the probability of recombining.
  - Note that the parameter is a scaled value which will be defined later
- Assume $k$ individuals in a generation. The following might happen:
  1. An individual arises because of a recombination event between two individuals (It will have 2 parents).
  2. Two individuals coalesce
  3. Neither (Each individual has a distinct parent)
  4. Multiple events (low probability)
We ignore the case of multiple (> 1) events in one generation.

\[ \text{Pr (No recombination)} = 1 - kr \]

\[ \text{Pr (No coalescence)} = \left(1 - \frac{k}{2N}\right) \]

Consider scaled time in units of 2N generations. Thus the number of individuals increase with rate kr2N, and decrease with rate \( \binom{k}{2} \).

The value 2rN is usually small, and therefore, the process will ultimately coalesce to a single individual (MRCA).
ARG

- Let $k = n$,
- Define $\rho = 4rN$
- Iterate until $k = 1$
  - Choose time from an exponential distribution with rate
    \[
    \frac{k\rho}{2} + \binom{k}{2}
    \]
  - Pick event as recombination with probability
    \[
    \frac{\rho}{\rho + (k - 1)}
    \]
  - If event is recombination, choose an individual to recombine, and a position, else choose a pair to coalesce.
  - Update $k$, and continue
Simulating sequences on an ARG

- Simulate the ARG
- Generate each of the constituent coalescents and revise mutation rates
- Generate sequences for each of the coalescents
- Concatenate
Generating samples from coalescent with recombination

\[ x_2 = 0.3 \]

\[ x_1 = 0.8 \]

\[ G(0) \quad G(0.3) \quad G(0.8) \]
Coalescent theory Review

- Under a specific model of evolution, coalescent theory allows us to simulate population data efficiently (linear in the size of the data).
- This allows us to compute many summary statistics, and test hypotheses.
Estimating (scaled) mutation rate

- Given a population sample evolving according to a coalescent without recombination, can you estimate $\mu$ (number of mutations per individual per generation)?
Watterson’s estimate

- Let $S$ be the number of mutations in the history of a population sample (diploid, $2N$ haplotypes).
- If we make the infinite sites assumption, then $S$ can be estimated.
- Recall that
  - $E(S_n) = \mu E(T_{\text{tot}})$
  - $E(S_n) = \mu 2N \sum_k \frac{2}{(k-1)} = 4N \mu (\gamma + \ln (n-1))$
  - Watterson’s estimate
    - $\theta_w = S_n / (\gamma + \ln (n-1))$
Tajima’s estimate of $\theta$

- Define $\pi_{ij} = \text{heterozygosity between two individuals}$
- Note: heterozygosity = # differing sites = hamming distance

\[
i: 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \\
j: 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \\
\]

$\pi_{ij} = 2$

- Average heterozygosity can be empirically estimated from a sample as

\[
\hat{k} = \frac{1}{\binom{n}{2}} \sum_{ij} \pi_{ij}
\]
Estimating Average heterozygosity

- Assuming an underlying coalescent model of evolution, what is the average heterozygosity?

- Q: Given 2 randomly picked individuals, what is the expected time to coalescence?
  - A: 2N

- Q: Given 2 individuals what is the expected number of mutations in the lineages connecting them?
  - A: $\mu 2 2N = \theta$

- Therefore, the average heterozygosity $k$ is an estimate (Tajima’s estimate) of $\theta$
Difference tests

- Under neutral evolution, there are many different estimates of $\theta$, all using coalescent theory.
  - You’ll explore these in homework 2.
- If you take any two and take the difference, the expected value is 0.
- Departure from neutrality is indicative of non-neutral evolution.
Coalescent theory: summary of results

- CT can be used to efficiently generate populations
- Test out possible departures from neutrality.
- The theory also helps estimate various parameters of a population sample
  - Scaled mutation rate, \( \theta \)
  - Effective population size, \( N \)
  - Time to MRCA (4N)
  - Likely genealogical history of the population sample
    (Perfect phylogeny, ancestral recombination graph)