Population genetics: Inference using trees of individuals

Peter Beerli
Florida State University #MolEvol2018 MBL Woods Hole
The big overview

Phylogenetics
  Trees
  Mutation models
The big overview

Population genetics
Allele frequencies
Population models
The big overview

Phylogenetics

Trees
Mutation models

Population genetics

Allele frequencies
Population models
The big overview

- Phylogenetics
  - Trees
  - Mutation models

- Population genetics
  - Allele frequencies
  - Population models

Coalescence theory
Coalescence theory as a tool for population genetics

coa•alesce  |ˌkōəˈles|
verb [ intrans. ]
come together and form one mass or whole: *the puddles had coalesced into shallow streams* | *the separate details coalesce to form a single body of scientific thought.*
• [ trans. ] combine (elements) in a mass or whole: *to help coalesce the community, they established an office.*

DERIVATIVES

coa•ales•cence  |-ˈlesəns| noun

coa•ales•cent  |-ˈlesənt| adjective

ORIGIN mid 16th cent. (in the sense [bring together, unite]): from Latin *coalescere*, from *co-* (from *cum* ‘with’) + *alescere ‘grow up’* (from *alere ‘nourish’*).
Species trees

Pelophylax

- cypriensis
  - bergeri
  - cypriensis
  - cretensis
  - caralitanus
  - cerigensis
  - bedriagae2
    - epeiroticus
      - ridibundus
    - bedriagae
     - Cilicia
      - perezi
      - saharicus

0.0050
Tree of individuals of same species

- epiroticus
  - rid1
    - rid2
    - rid3
      - rid7
        - rid6
        - rid4
          - rid5
          - rid8
Little resolution

Tree building method should take into account that lineages are not independent of each other.
Interaction among individuals

Life cycle
Interaction among individuals

Wright-Fisher population model

- All individuals live one generation and get replaced by their offspring
- All have same chance to reproduce, all are equally fit
- The number of individuals in the population is constant
Sewall Wright evaluated the probability that two randomly chosen individuals in generation $t$ have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is
Sewall Wright evaluated the probability that two randomly chosen individuals in generation $t$ have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is $1.0$. 
Sewall Wright evaluated the probability that two randomly chosen individuals in generation $t$ have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

$$1.0 \times \frac{1}{2N}$$
Sewall Wright evaluated the probability that two randomly chosen individuals in generation $t$ have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

$$\frac{1}{2N}$$

The probability that two randomly picked chromosome do not have a common ancestor is

$$1 - \frac{1}{2N}$$
If we know the genealogy of the two individuals then we can calculate the probability as

$$P(\tau|N) = \left(1 - \frac{1}{2N}\right)^\tau \left(\frac{1}{2N}\right)$$

where $\tau$ is the number of generations with no coalescence. This formula is the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce:

$$\mathbb{E}(\tau) = 2N$$
Population model

Wright-Fisher

Present

Past
Population model

Past

Present
Population model

Wright-Fisher

Past

Present
10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of $2N=20.34$. 

Probability Distribution

$2N=20$

$P$

generations

$20$

$100$

$0.25$

$0.20$

$0.15$

$0.10$

$0.05$

$0.05$
For the time of coalescence in a sample of TWO, we will wait on average $2N$ generations assuming it is a Wright-Fisher population.

The model assumes that the generations are discrete and non-overlapping.

Real populations do not necessarily behave like a Wright-Fisher (the ‘ideal’ population).

We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.
Wright-Fisher

\[ \sigma^2_{\text{offspring}} \approx 1 \]
\[ \mathbb{E}(\tau) = 2N \]
\[ \text{generation time } g = 1 \]

Canning

\[ \sigma^2_{\text{offspring}} = x \]
\[ \mathbb{E}(\tau) = 2N/x \]
\[ g = 1 \]

Moran

\[ \sigma^2_{\text{offspring}} = \frac{2}{2N} \]
\[ \mathbb{E}(\tau) = \frac{1}{2}(2N)^2 \]
\[ g = 2N \]

You can generate graphs like this using the python program \textit{popsim}
(check out my faculty page for the link)
Sample larger than TWO

Past

Present

Wright-Fisher
Sample larger than TWO

Past

Present

Wright-Fisher
Sir J. F. C. Kingman described in 1982 the $n$-coalecent. He showed the behavior of a sample of size $n$, and its probability structure looking backwards in time.

General findings:

\[
\text{coalescence rate} = \binom{n}{2} = \frac{n(n - 1)}{2}
\]

Once a coalescence happened $n$ is reduced to $n - 1$ because two lineages merged into one. He then imposed a continuous approximation of the Canning’s exchangeable model to get results.
An analogy
Samples larger than two
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample $n$ and the total population size $N$. 
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample $n$ and the total population size $N$.

Using Kingman’s coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample size \( n \) and the total population size \( N \).

Using Kingman’s coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

\[
P(u_j | N) = e^{-u_j \lambda}
\]

with the scaled coalescence rate

\[
\lambda = \binom{k}{2} \frac{1}{2N} \times \text{Prob(others do not coalesce)}
\]
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample size $n$ and the total population size $N$.

Using Kingman’s coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

$$P(u_j|N) = e^{-u_j \lambda}$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2} \frac{1}{2N} = \frac{k(k-1)}{2(2N)} = \frac{k(k-1)}{4N}$$
We are now able to calculate the probability of a whole relationship tree (Genealogy $G$). We assume that each coalescence is independent from any other:

$$P(G|N)$$
We are now able to calculate the probability of a whole relationship tree (Genealogy $G$). We assume that each coalescence is independent from any other:

$$P(G|N) = P(u_0|N, i_1, i_2) \times$$
We are now able to calculate the probability of a whole relationship tree (Genealogy $G$). We assume that each coalescence is independent from any other:

$$P(G|N) = P(u_0|N, i_1, i_2) \times P(u_1|N, i_3, i_4)$$
We are now able to calculate the probability of a whole relationship tree (Genealogy $G$). We assume that each coalescence is independent from any other:

$$P(G|N) = P(u_0|N,i_1,i_2) \times P(u_1|N,i_3,i_4) \times P(u_3|N,i_3,4,i_5)$$
We are now able to calculate the probability of a whole relationship tree (Genealogy $G$). We assume that each coalescence is independent from any other:

\[
P(G|N) = P(u_0|N, i_1, i_2) \times P(u_1|N, i_3, i_4) \times P(u_3|N, i_{3,4}, i_5) \times P(u_4|N, i_{1,2}, i_{3,4,5})
\]
We are now able to calculate the probability of a whole relationship tree (Genealogy \( G \)). We assume that each coalescence is independent from any other:

\[
P(G|N) = P(u_0|N, i_1, i_2) \\
\times P(u_1|N, i_3, i_4) \\
\times P(u_3|N, i_{3,4}, i_5) \\
\times P(u_4|N, i_{1,2}, i_{3,4,5})
\]

\[
P(G|N) = \prod_{j=0}^{T} e^{-u_j} \frac{k_j (k_j - 1)}{4N} \cdot \frac{2}{4N}
\]
Each interval $u_j$ is independent of the others, the expected length of the interval is the inverse of the coalescent rate. Thus we can sum these expectations to get to expectation of the depth of the genealogy.

\[
E(\tau_{\text{MRCA}}) = \text{Sum of the expectation of each time interval} = \sum_{j=0}^{J} \frac{4N}{k_j(k_j - 1)}
\]

\[
\lim_{k \to \infty} E(\tau_{\text{MRCA}}) = 2N + \frac{2}{3}N + \frac{1}{3}N + \frac{1}{5}N + \frac{2}{15}N + ... = 4N
\]

\[
\lim_{k \to \infty} \sigma(\tau_{\text{MRCA}}) = 4N
\]
If we know the genealogy $G$ with certainty then we can calculate the population size $N$. Finding the maximum probability $P(G|N, k)$ is simple, we evaluate all possible values for $N$ and pick the value with the highest probability.
If we know the genealogy $G$ with certainty then we can calculate the population size $N$. Finding the maximum probability $P(G|N, k)$ is simple, we evaluate all possible values for $N$ and pick the value with the highest probability.
What is it good for?

If we know the genealogy $G$ with certainty then we can calculate the population size $N$. Finding the maximum probability $P(G|N,k)$ is simple, we evaluate all possible values for $N$ and pick the value with the highest probability.
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k - 1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

\[
p(G \mid N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}
\]
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$. 

$$p(G|N, n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k - 1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G \mid N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k - 1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$
If an oracle gives us the true relationship tree $G$ then we can calculate the population size $N$.

$$p(G|N, n) = \prod_{k=2}^{n} \exp \left( -u_k \frac{k(k - 1)}{4N} \right) \frac{2}{4N}$$
There are at least two problems with the oracle-approach:

- There is no oracle to give us clear information!
- We do not record genealogies, our data are sequences, microsatellite loci!
- What about the variability of the coalescence process?
All genealogies were simulated with the same population size $N_e = 10,000$
Variability of the coalescent process

MRCA = most recent common ancestor (last node in the genealogy)
All individuals have the same fitness (no selection).

All individuals have the same chance to be in the sample (random sampling).

The coalescent allows only merging two lineages per generation. This restricts us to have a much smaller sample size than the population size.

\[ n << N \]

Yun-Xin Fu (2005) described the exact coalescent for the Wright-Fisher model and derived a maximal sample size \( n < \sqrt{4N} \) for a diploid population. Although this may look like a severe restriction for the use of the coalescence in small populations, it turned out that the coalescence is rather robust and that even sample sizes close to the effective population size are not biasing immensely.
Kingman’s $n$-coalescent is an approximation.
Kingman’s $n$-coalescent is an approximation.
Kingman’s \(n\)-coalescent is an approximation.
Kingman’s \( n \)-coalescent is an approximation.
Observations

- Large samples coalesce on average in $4N$ generations.
- The time to the most recent common ancestor (TMRCA) has a large variance.
- Even a sample with few individuals can most often recover the same TMRCA as a large sample.
- The sample size should be much smaller than the population size, although severe problems appear only with sample sizes of the same magnitude as the population size, or with non-random samples because Kingman’s coalescence process assumes that maximally two sample lineages coalesce in any generation.
- With a known genealogy we can estimate the population size. Unfortunately, the true genealogy of a sample is rarely known.
Genealogy and data

Our data looks like this:
Finding the best genealogy from such data is difficult.
Genetic data and the coalescent

- Finite populations loose alleles due to genetic drift
- Mutation introduces new alleles into a population at rate $\mu$
- With $2N$ chromosomes we can expect to see every generation $2N\mu$ new mutations. The population size $N$ is positively correlated with the mutation rate $\mu$.
- With genetic data sampled from several individuals we can use the mutational variability to estimate the population size.
The observed genetic variability

\[ S = f(N, \mu, n). \]

Different \( N \) and appropriate \( \mu \) can give the same number of mutations. For example, for 100 loci sampled from 20 individuals with 1000bp each, we get:

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \mu )</th>
<th>( 4N\mu )</th>
<th>( \hat{S} )</th>
<th>( \sigma_S^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1250</td>
<td>10^{-5}</td>
<td>0.05</td>
<td>153.95</td>
<td>16.25</td>
</tr>
<tr>
<td>12500</td>
<td>10^{-6}</td>
<td>0.05</td>
<td>152.89</td>
<td>16.05</td>
</tr>
</tbody>
</table>

Using genetic variability alone therefore does not allow to disentangle \( N \) and \( \mu \).

With multiple dated samples and known generation time we can estimate \( N \) and \( \mu \) independently.
By convention we express most results as the compound $N\mu$ and an inheritance scalar $x$, for simplicity we call this the mutation-scaled population size

$$\Theta = xN\mu,$$

where $\mu$ is the mutation rate per generation and per site. With a mutation rate per locus we use $\theta$.

- for diploids: $\Theta = 4N\mu$.

- for haploids: $\Theta = 2N\mu$.

- For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is 1 : 1 then $\Theta = N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript $N$ and call it the effective population size $N_e$, and consider $\Theta$ the mutation-scaled EFFECTIVE population size.
By convention we express most results as the compound $N\mu$ and an inheritance scalar $x$, for simplicity we call this the mutation-scaled population size

$$\Theta = xN\mu,$$

where $\mu$ is the mutation rate per generation and per site. With a mutation rate per locus we use $\theta$.

- for diploids: $\Theta = 4N\mu$.

- for haploids: $\Theta = 2N\mu$.

- For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is 1 : 1 then $\Theta = N\mu$.

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript $N$ and call it the effective population size $N_e$, and consider $\Theta$ the mutation-scaled EFFECTIVE population size.
Historical humpback whale population size

Humpback whales in the North Atlantic: Census population size around 12,000.
Historical humpback whale population size

using the data by Joe Roman and Stephen R. Palumbi (Science 2003 301: 508-510)

\[ \Theta = 2N_q\mu \]

Population size of the North Atlantic population, estimated using migrate

\[ N_q = \frac{\Theta}{2\mu} \]

with \( \mu = 2.0 \times 10^{-8} \text{bp}^{-1} \text{year}^{-1} \) and a generation time of 12 years

\[ N_e = N_q + N_\sigma \]

Sex ratio is 1:1

\[ N_B = 2N_e \]

ratio \( N_B/N_e \) assumed, using other data

\[ N_T = N_B \frac{N_{\text{juveniles}} + N_{\text{adults}}}{N_{\text{adults}}} \]

from catch and survey data (used a ratio of 1.6)

More modern estimates for mtDNA: 150,000 [improved estimation of mutation rate]; for nucDNA: 112,000 (45,000 – 235,000) [Conservation Genetics (2013) 14:103114]
Using the infinite sites model we use the number of variable sites $S$ per locus to calculate the mutation-scaled population size:

$$\theta_W = \frac{S}{n-1} \sum_{k=1}^{n-1} \frac{1}{k}$$

from a sample of $n$ individuals. For a single population the Watterson’s estimator works marvelously well, but it is vulnerable to population structure.

Watterson’s $\theta_W$ uses a mutation rate per locus! To compare with other work use mutation rate per site.
For Bayesian inference we want to calculate the probability of the model parameters given the data $p(\text{model}|D)$.

- **Coalescent** to describe the population genetic processes.
- **Mutation model** to describe the change of genetic material over time.
We calculate the Posterior distribution $p(\Theta|D)$ using Bayes’ rule

$$p(\Theta|D) = \frac{p(\Theta)p(D|\Theta)}{p(D)}$$

where $p(D|\Theta)$ is the likelihood of the parameters.
The probability density of a genealogy given parameters.

The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.
The Felsenstein equation, aka Likelihood calculation:

\[ p(D|\Theta) = \int_G p(G|\Theta)p(D|G)dG \]

- \( p(G|\Theta) \): The probability density of a genealogy given parameters.
- \( p(D|G) \): The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.
The Felsenstein equation, also known as the Likelihood calculation, is given by:

\[ p(D|\Theta) = \sum_G p(G|\Theta)p(D|G) \]

- \( p(G|\Theta) \): The probability of a genealogy given parameters.
- \( p(D|G) \): The probability of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.
The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.
Naive integration approach
Another naive integration approach
Metropolis-Hastings algorithm
Metropolis-Hastings algorithm
Metropolis-Hastings algorithm
Metropolis-Hastings algorithm
Metropolis-Hastings algorithm
Metropolis-Hastings algorithm
Inference of population size

Bayesian inference: $\Theta = 0.00903$
Watterson Estimator $\Theta_W = 0.01003$
Coalescent:


Sample size: