Gene Genealogies
Coalescence Theory

Annabelle Haudry
Glasgow, July 2009
What could tell a gene genealogy?

- How much diversity in the population?
- Has the demographic size of the population changed? How? When?
- Is that gene under selection?
- Is/was there gene flow between sub-populations?

Could I describe the evolutionary model under which the species has been evolving?
Wright-Fisher: a simple model

Finite and constant population size
N diploid individuals, 2N genes
Random mating
Non overlapping generations
No selection, structure…
Only Drift

N = 5

2N = 10

N = 5

2N = 10
Genetic Drift

gene evolution

past

present

2N = 10
Genetic Drift

gene evolution

past

present

2N = 10
Genetic Drift

gene evolution

past

present

2N = 10
Genetic Drift

gene evolution

past

present

2N = 10
Genetic Drift

2N = 10

past

present
Genetic Drift

gene evolution

past

present

2N = 10
Genetic Drift

gene evolution

past

present

2N =10
2N = 10

past

gene evolution

present

2N = 10
Backward in Time

past

present

2N = 10

n = 3
Coalescent Events

2N = 10

n = 3

past

present
Coalescence probability of 2 genes at a given generation

\[ p(2, 2N) = \frac{1}{2N} \]
Wright-Fisher Model

Coalescence probability of 2 genes at a given generation

\[ p(2, 2N) = 1/2N \]

Coalescence probability of 2 genes \( j \) generations back, within a sample of \( n \) genes

\[ p(T_2 = j) = \frac{1}{2N} \left( 1 - \frac{1}{2N} \right)^{j-1} \frac{n(n-1)}{2} \]
Mathematical Theory: Kingman

Wright-Fisher Model

Coalescence probability of 2 genes at a given generation

\[ p(2, 2N) = 1/2N \]

Coalescence probability of 2 genes \( j \) generations back, within a sample of \( n \) genes

\[ p(T_2 = j) = \frac{1}{2N} \left( 1 - \frac{1}{2N} \right)^{j-1} \frac{n(n-1)}{2} \]

- probability 2 genes haven’t coalesced during the first \( j-1 \) generations
- number of possibles pairs of 2 genes within \( n \) genes
Coalescence probability of 2 genes at a given generation

\[ p(2, 2N) = \frac{1}{2N} \]

Coalescence probability of 2 genes \( j \) generations back, within a sample of \( n \) genes

\[ p(T_2 = j) = \frac{1}{2N} \left( \frac{1}{2N} \right)^{j-1} \frac{n(n-1)}{2} \]

\( T_2 \sim \text{geometric distribution} \)

For \( n \ll N \), approximation by an exponential distribution

\( T_n \) coalescence time between 2 genes: \( T_n \sim \text{exponential distribution} \):

\[ E(T_n) = \frac{4N}{n(n-1)} \]

so \( E(T_2) = 2N \)
Highly Variable Process

Same $T_n$, $T_{n-1}$, ..., $T_3$, but different $T_2$

distribution of $T_2$

Variable $T_n$, $T_{n-1}$, ..., $T_3$, $T_2$

All trees have same probability
Coalescent with Mutations

# mutations added on each branch of the tree following a Poisson process $m_i$:

$E(m_i) = \mu t$

$\mu$ : mutation rate/gene/generation
$t$ : length of branch
How to summarize a genealogy?

Infinite Site Model

5 mutations

5 segregating sites

15 genes/individuals
How to summarize a genealogy?

Infinite Site Model

5 mutations

5 segregating sites

15 genes/individuals

alignment
How to summarize a genealogy?

Estimations de $4N\mu$

- **nucleotide diversity** $\pi$ (Tajima 1983)
  $\pi = \text{average } \# \text{ of pairwise differences}$

  $E(\pi) = \mu t = \mu \cdot 2.2N = 4N\mu$

- **theta** $\theta_S$ (Watterson 1975)
  $S = \# \text{ segregating sites}$

  $\theta_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$
And a set of genealogies?

Expected distributions of summary statistics in a population under a standard model, 10,000 simulations with $\theta = 5$
And a set of genealogies?

Expected distributions of summary statistics in a population under a standard model, 10,000 simulations with $\theta = 5$
Effect of Demography

Gene genealogy affected by $N$ variations

- Formalization of the evolutionary history of populations
- Coalescence theory will describe expected diversity in a sample under different evolution scenarii
Effect of Demography

Affects polymorphism patterns ($\neq \theta_S$ and $\pi$)
$\theta_S$ and $\pi$ differentially affected

Molecular signatures of variations in population size
Tajima’s D

\[ D = \frac{\pi - \theta_S}{\sqrt{\text{Var}(\pi - \theta_S)}} \]

\[ D \sim \mathcal{N}(0,1) \text{ under standard coalescent} \]
Tajima’s D

\[ D = \frac{\pi - \theta_S}{\sqrt{\text{Var}(\pi - \theta_S)}} \]

\(D \sim \mathcal{N}(0,1)\) under standard coalescent

\[4N\mu\]

\[D \approx 0\]

\[D \geq 0\]

\[D \leq 0\]

equilibrium  
bottleneck  
demographic expansion

\(\theta_S\)

\(\pi\)

time
Hudson (2002): Program *ms* based on coalescence theory to generate simulated gene samples.

Simplest model = Wright-Fisher population

+ complex: - changes in population size
  - population structure
  - migration
  - recombination
  - partial selfing
  - selection...
Define the best model’s parameter

Real samples

α parameter tested in model

Simulated samples

Observed values ± 20%

Estimated Likelihood \( i (\alpha_i) \)
Example of Maize Domestication

Domestication bottleneck

$(k = 2.45)$

More severe bottleneck at some genes candidate selected genes?

$(k_1 = 0.15, k_2 = 2.45)$

Wright et al. (2005)
**Coalescence and phylogeny**

Coalescence: evolutionary history of **genes**

Phylogeny: evolutionary history of **species**

 Highly diverged species

Speciation history

Population history
Coalescence and phylogeny

Coalescence: evolutionary history of genes
Phylogeny: evolutionary history of species

Gene history ≠ Species history

Close related species

Incomplete lineage sorting
More details in: