Coalescent Theory: A brief Introduction

Ben Peter

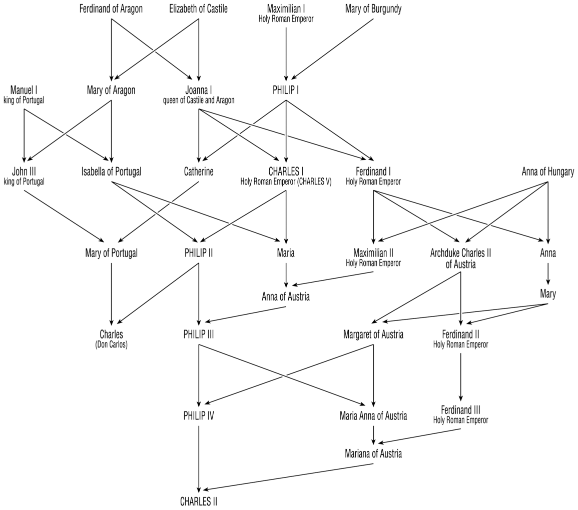
MPI for Evolutionary Anthropology

April 18, 2018

Recommended reading

- John Wakeley (2009): Coalescent Theory: An Introduction
- Joe Felsenstein (2016?): Theoretical Evolutionary Genetics `http://evolution.genetics.washington.edu`
- Rick Durett (2008): Probability Models for DNA Sequence Evolution `https://services.math.duke.edu/~rtd/Gbook/PM4DNA_0317.pdf`

Pedigrees



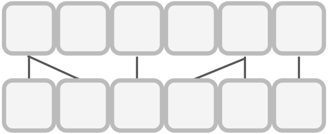
- backwards-in-time perspective
- identity-by-descent
- not available for most people

Wright (1922): The American Naturalist

Alvarez et al. (2009) PLoS One

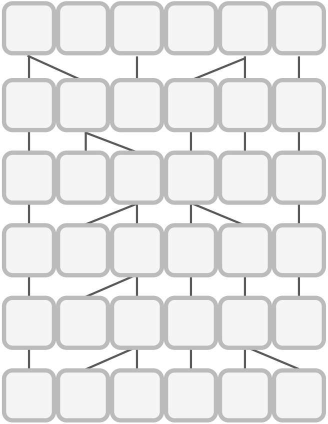
The population pedigree

The population pedigree



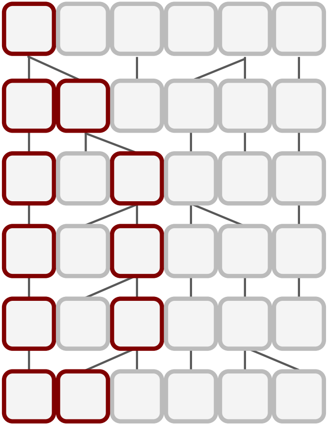
- Neutral WF-model with ancestry relationship

The population pedigree



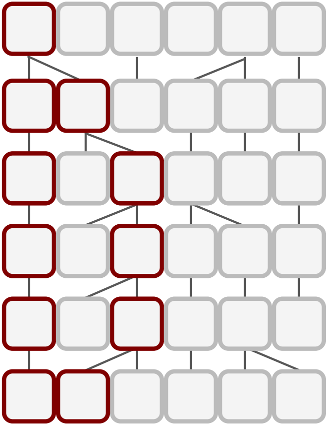
- Neutral WF-model with ancestry relationship
- Exists, but unobserved

The population pedigree



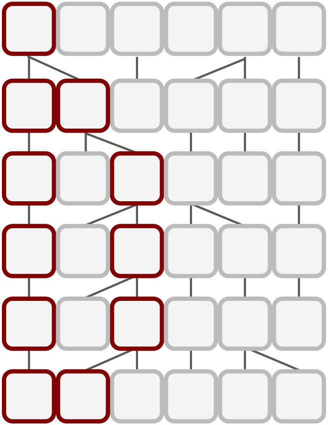
- Neutral WF-model with ancestry relationship
- Exists, but unobserved
- **Key:** most bits unimportant

The population pedigree



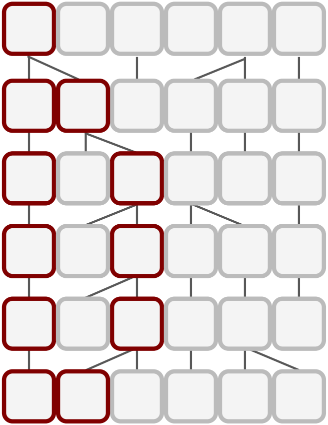
- Neutral WF-model with ancestry relationship
- Exists, but unobserved
- **Key:** most bits unimportant
- Backwards in time: each generation: coalesce (Y/N)

The population pedigree



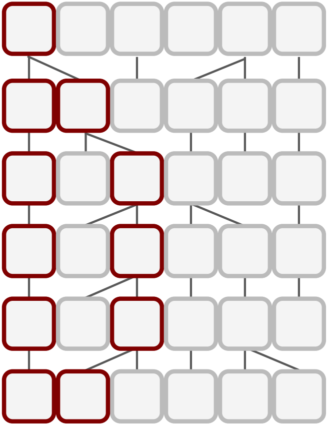
- Neutral WF-model with ancestry relationship
- Exists, but unobserved
- **Key:** most bits unimportant
- Backwards in time: each generation: coalesce (Y/N)
- $\mathbb{P}(\text{coalescence}) = \frac{1}{2N}$

The population pedigree



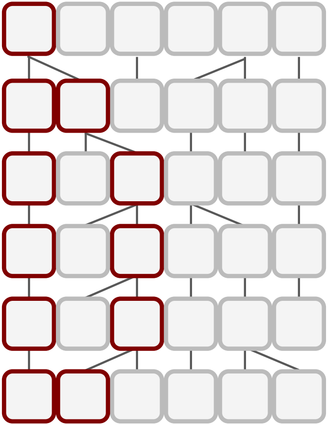
- Neutral WF-model with ancestry relationship
- Exists, but unobserved
- **Key:** most bits unimportant
- Backwards in time: each generation: coalesce (Y/N)
- $\mathbb{P}(\text{coalescence}) = \frac{1}{2N}$
- $T_2 \sim \text{Geometric}\left(\frac{1}{2N}\right)$

Implications



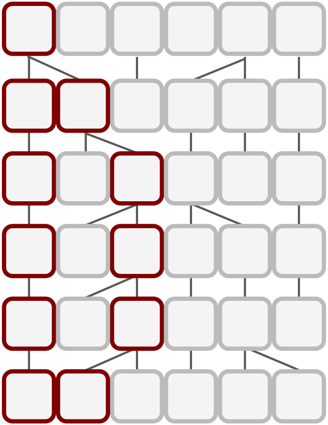
- If we go back far enough, there will be a common ancestor
- Takes about $2N$ generations

Implications



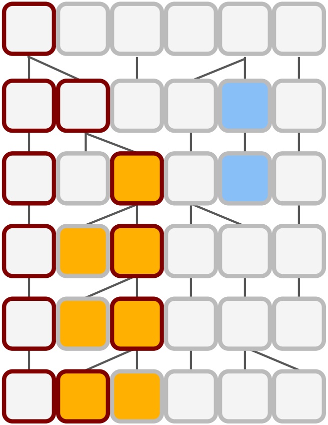
- If we go back far enough, there will be a common ancestor
- Takes about $2N$ generations
- $2N$ is typically large
    - $r_{\text{coalescence}} = \frac{1}{2N}$
    - $T_2 \sim \text{Exp}\left(\frac{1}{2N}\right)$

Implications

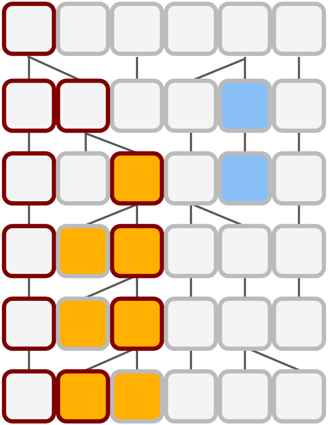


- If we go back far enough, there will be a common ancestor
- Takes about $2N$ generations
- $2N$ is typically large
  - $r_{\text{coalescence}} = \frac{1}{2N}$
  - $T_2 \sim \text{Exp}\left(\frac{1}{2N}\right)$
- Rescaling time:
  - $r_{\text{coalescence}} = 1$
  - $T_2 \sim \text{Exp}(1)$

Poisson process formulation



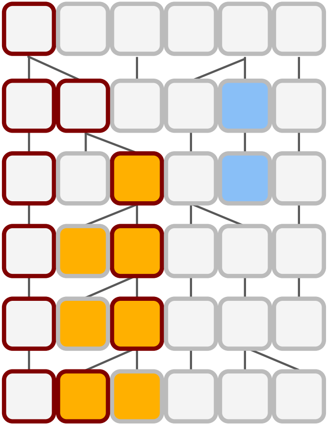
Poisson process formulation



Two types of events:

- Coalescence
- Mutation

Poisson process formulation



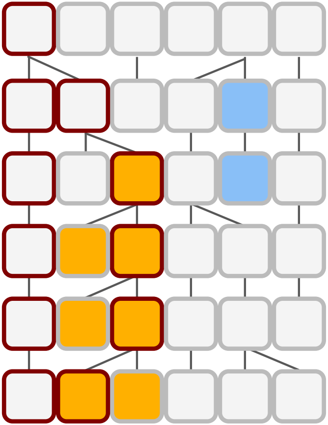
Two types of events:

- Coalescence
- Mutation

Rates:

- $r_{\text{coalescence}} = 1$
- $r_{\text{mutation}} = 2 \times 2N \times \mu = \theta$

Poisson process formulation



Two types of events:

- Coalescence
- Mutation

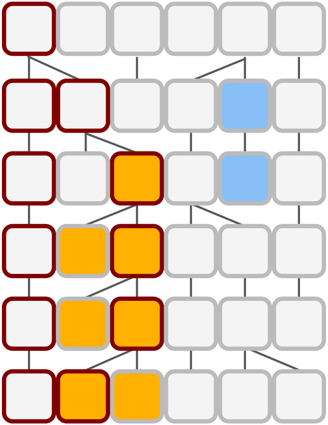
Rates:

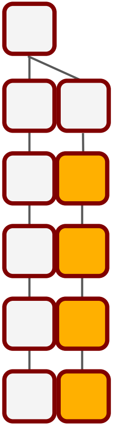
- $r_{\text{coalescence}} = 1$
- $r_{\text{mutation}} = 2 \times 2N \times \mu = \theta$

Define $\pi = \#\text{differences}$

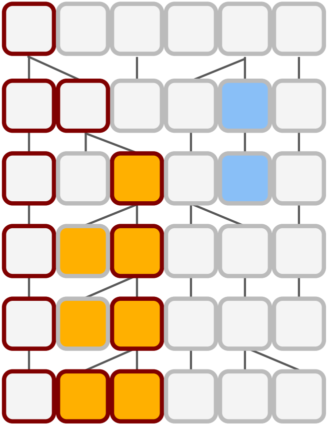
- $\pi \sim \text{Geometric}\left(\frac{1}{1+\theta}\right)$
- $\mathbb{E}[\pi] = \theta$

From pairs to trees

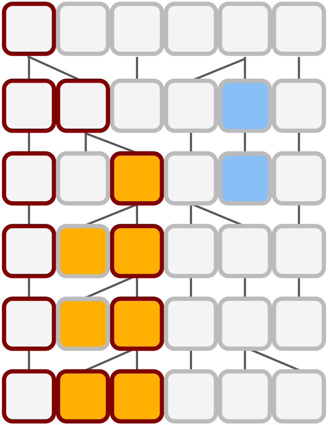


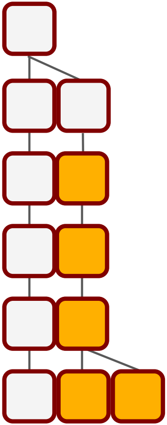


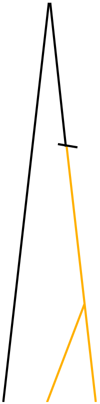
From pairs to trees



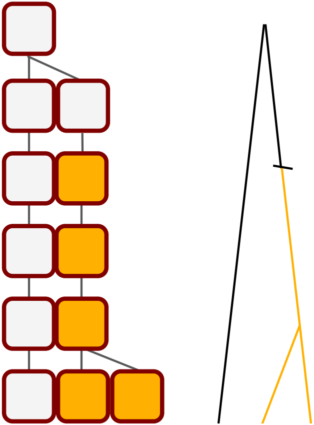
From pairs to trees



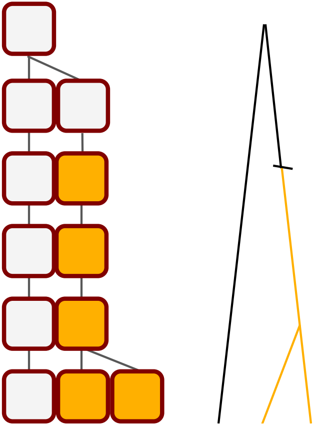




From pairs to trees



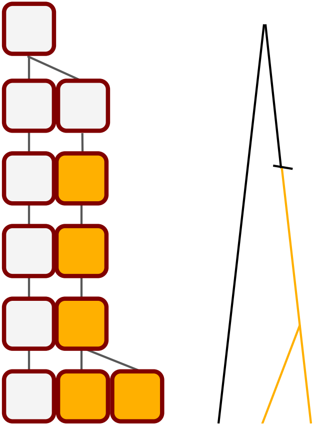
From pairs to trees



Two types of events:

- Coalescence
- Mutation

From pairs to trees



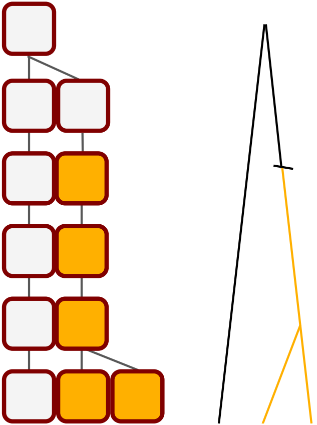
Two types of events:

- Coalescence
- Mutation

Rates:

- $r_{\text{coalescence}} = 1 + 1 + 1$
- $r_{\text{mutation}} = 3 \times 2N \times \mu = \frac{3}{2}\theta$

From pairs to trees



Two types of events:

- Coalescence
- Mutation

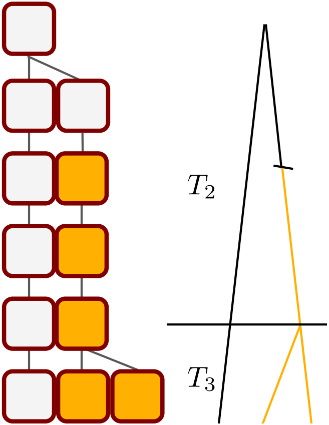
Rates:

- $r_{\text{coalescence}} = 1 + 1 + 1$
- $r_{\text{mutation}} = 3 \times 2N \times \mu = \frac{3}{2}\theta$

After coalescence:

- $r_{\text{coalescence}} = 1$
- $r_{\text{mutation}} = 2 \times 2N \times \mu = \theta$

From pairs to trees



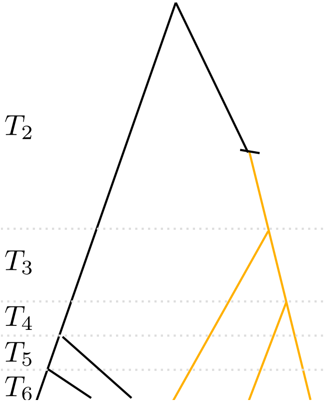
Two types of events:

- Coalescence
- Mutation

Rates:

- $r_{\text{coalescence}} = 1 + 1 + 1$
- $r_{\text{mutation}} = 3 \times 2N \times \mu = \frac{3}{2}\theta$

From pairs to trees



Two types of events:

- Coalescence
- Mutation

Rates:

- $r_{\text{coalescence}} = \binom{n}{2}$
- $r_{\text{mutation}} = n \times 2N \times \mu = \frac{n}{2}\theta$

Distributions:

- $T_n \sim \text{Exp}\left(\binom{n}{2}\right)$
- $S_n \sim \text{Geometric}\left(\frac{n-1}{\theta+n-1}\right)$
- $\mathbb{E}[S_n] = \frac{\theta}{n-1}$

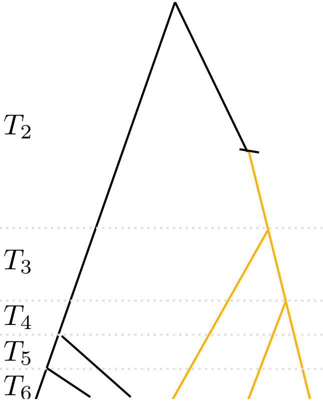
Infinite sites mutation model

- **Assumption:** Each mutation is at a new site

Infinite sites mutation model

- **Assumption:** Each mutation is at a new site
- **Implications:**
  1. $P_i \sim \text{Uniform}(0,1)$
  2. assuming mutations are rare ($\theta^2 \approx 0$)
  3. Good model for humans ($\theta \approx 10^{-3}$)
  4. More problematic for *Drosophila* ($\theta \approx 10^{-2}$)
  5. All mutations on genealogy are observable

Two measures of tree size



Tree height and total size

- $T_{MRCA} = \sum_{i=2}^{n} T_i$
- $T_{Tot} = \sum_{i=2}^{n} iT_i$

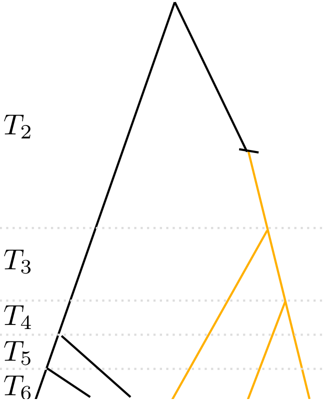
Expectation:

- $\mathbb{E}[T_{MRCA}] = \sum \binom{n}{2}^{-1} = 2\left(1 - \frac{1}{n}\right)$
- $\mathbb{E}[T_{Tot}] = \sum i \frac{2}{i(i-1)} = 2\sum \frac{1}{i}$

Implications:

- $\mathbb{E}[T_{MRCA}] \approx 2$
- $\mathbb{E}[T_{Tot}] \approx K + \log(n)$

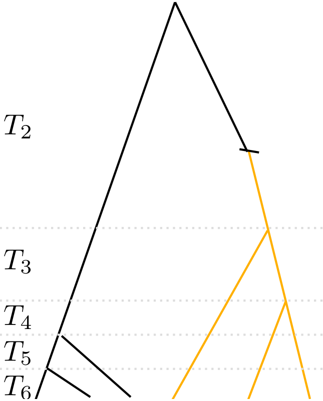
Total number of Mutations



Two formulations:

- $S \sim \sum_i S_i$
- $S \sim \text{Poisson}\left(\theta/2T_{Tot}\right)$

Total number of Mutations



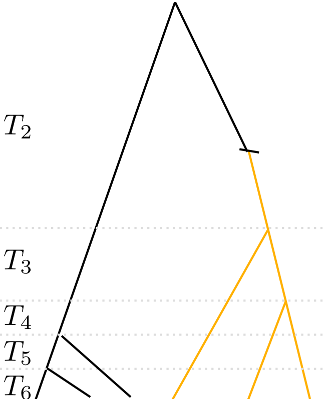
Two formulations:

- $S \sim \sum_i S_i$
- $S \sim \text{Poisson}\left(\theta/2\, T_{Tot}\right)$

Expectation:

- $\mathbb{E}[S] = \sum_i \mathbb{E} S_i = \theta \sum_{i=1}^{n-1} \frac{1}{i}$

Total number of Mutations



Two formulations:

- $S \sim \sum_i S_i$
- $S \sim \text{Poisson}\left(\theta/2 T_{Tot}\right)$

Expectation:

- $\mathbb{E}[S] = \sum_i \mathbb{E}S_i = \theta \sum_{i=1}^{n-1} \frac{1}{i}$
- $\mathbb{E}[S] = \theta/2 \mathbb{E} T_{Tot} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$

Trees as a random quantity

- Tree is a latent variable that will change across genome
- Number of possible trees is very large

Trees as a random quantity

- Tree is a latent variable that will change across genome
- Number of possible trees is very large
- $\mathcal{T}_{10} = 2.5 \times 10^9$

Trees as a random quantity

- Tree is a latent variable that will change across genome
- Number of possible trees is very large
- $\mathcal{T}_{10} = 2.5 \times 10^9$
- how is this addressed in practice?
  1. Focus on quantities independent of topology, e.g. branch lengths
  2. sample size of two, three or four
  3. Monte Carlo integration

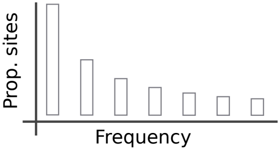
Some extensions

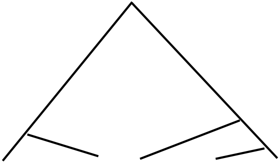
- Population size changes
- Population structure & Migration
- Recombination
- Linkage to selected site
- Analytical work often tricky, but simulations easy & very efficient

Part 2: Applications of the coalescent

1. The Site Frequency Spectrum (SFS)
2. Structured Populations
3. PCA
4. 2D-SFS

Neutral SFS

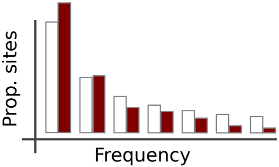


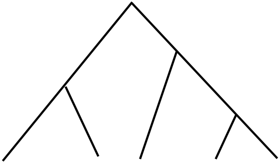


Definitions:
ξ_i : $\#$ mutations with i derived copies
η_i : $\#$ mutations with $i, n - i$ copies
Comments:

- For a particular tree: $\xi_i = 0$ is common
- Over many trees: $\mathbb{E}\xi_i = \frac{\theta}{i}$

Excess rare alleles





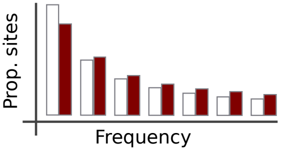
Definitions:

ξ_i : # mutations with i derived copies

η_i : # mutations with $i, n-i$ copies

- population growth
- negative selection

Excess common alleles



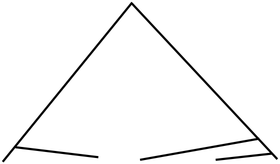
Definitions:

ξ_i : # mutations with i derived copies

η_i : # mutations with $i, n - i$ copies

Caused by:

- population decline
- positive selection



Inference Frameworks

There are a few frameworks that calculate expected SFS for (almost) arbitrary models:

- Gutenkunst et al. (2009): `dadi`, diffusion based

Inference Frameworks

There are a few frameworks that calculate expected SFS for (almost) arbitrary models:

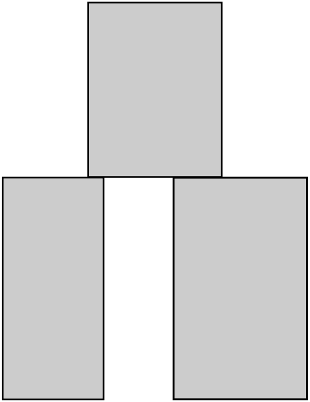
- Gutenkunst et al. (2009): `dadi`, diffusion based
- Kamm et al. (2014, 2018): `momi`, coalescent based

Inference Frameworks

There are a few frameworks that calculate expected SFS for (almost) arbitrary models:

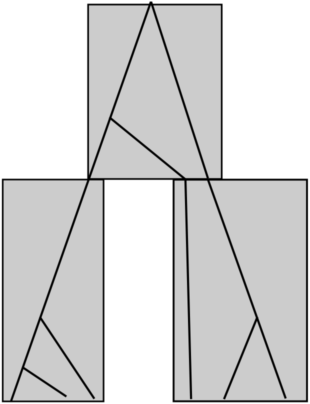
- Gutenkunst et al. (2009): `dadi`, diffusion based
- Kamm et al. (2014, 2018): `momi`, coalescent based
- Excoffier et al. (2012): `simcoal2`, simulation based

Population Structure: Isolation without migration



How do things change when we consider multiple populations?

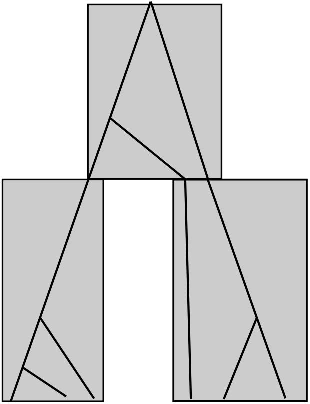
Population Structure: Isolation without migration



Rules:

1. Disconnected populations coalesce independently
2. State defined by number of lineages in each population (Markov process)
3. At a *merge* event, all surviving lineages are moved to new population
4. After *merge*, no labels are retained

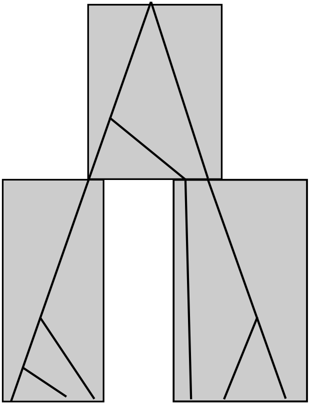
Coalescence times



Simplest case: $N_1 = N_2 = N_{ancestral}$

1. calculation for sample of size 2
2. $\mathbb{E} T_{11} = T_{22} = 1$
3. $\mathbb{E} T_{12} = 1 + t_{split}$

Coalescence times



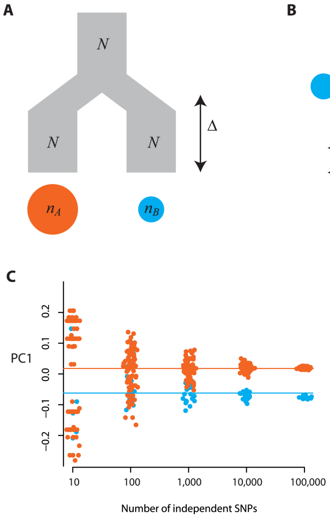
Simplest case: $N_1 = N_2 = N_{ancestral}$

1. calculation for sample of size 2
2. $\mathbb{E}\,T_{11} = T_{22} = 1$
3. $\mathbb{E}\,T_{12} = 1 + t_{split}$

Structure often described using

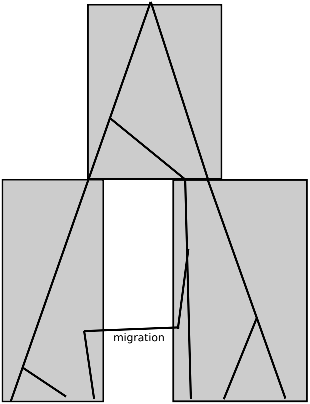
$$
\begin{aligned}
F_{ST} &= 1 - \frac{T_{within}}{T_{overall}} \\
&= 1 - \frac{T_{11} + T_{22}}{T_{12} + T_{11}/2 + T_{22}/2} \\
&= \frac{t_{split}}{2 + t_{split}}
\end{aligned}
$$

Principal componenet analysis



McVean (2009): PLoS Genetics

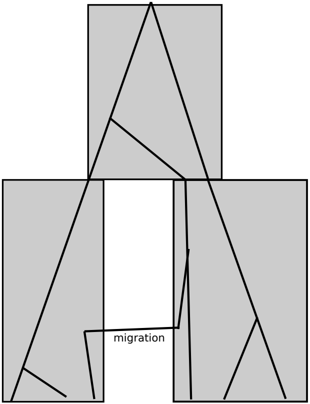
Isolation with migration



Rules:

1. Disconnected populations coalesce independently
2. State defined by number of lineages in each population (Markov process)
3. At a *merge* event, all surviving lineages are moved to new population
4. After *merge*, no labels are retained

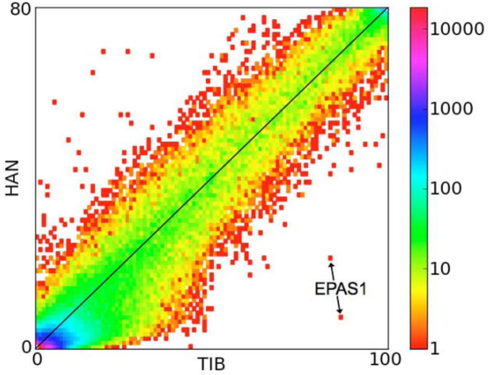
Isolation with migration



Rules:

1. Disconnected populations coalesce independently
2. State defined by number of lineages in each population (Markov process)
3. At a *merge* event, all surviving lineages are moved to new population
4. After *merge*, no labels are retained
5. Lineages *migrate* at some rate m $(2, 3) \rightarrow (1, 4)$

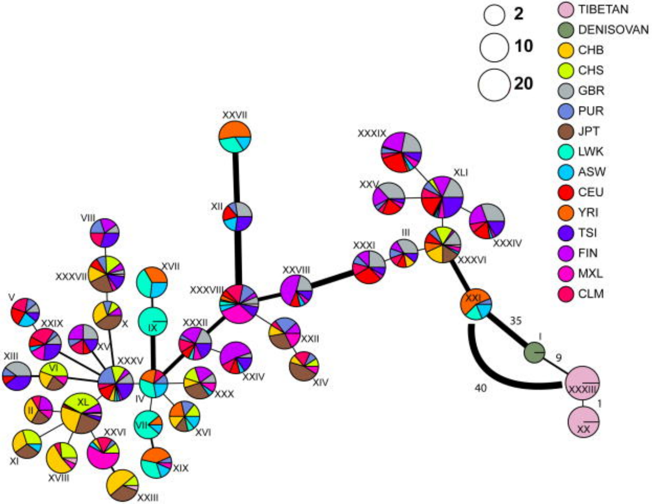
The 2D-SFS



Yi et al (2010): Science

- Matrix where M_{ij} is number of SNP at frequency i, j in pops A, B.
- For closely related populations, most mass is near diagonal.
- Outliers often biologically interesting

EPAS1



Huerta-Sanchez et al. (2014): Nature

Summary

1. Coalescent is a backwards-in-time model of evolution
2. Quantities directly related to sampling
3. Can easily simulate
4. Incorporate recombination, population size changes, migration
5. SFS is a key summary statistic (both for coalescent/diffusion)