

II. Population Genetics

Introduction to Coalescent Models

Lecture 12

Instructor: Su-In Lee

Last Lecture

- Linkage Equilibrium
 - Expected state for distant markers
- Linkage Disequilibrium
 - Association between neighboring alleles
 - Expected to decrease with distance
- Measures of linkage disequilibrium
 - D , D' and Δ^2 or r^2

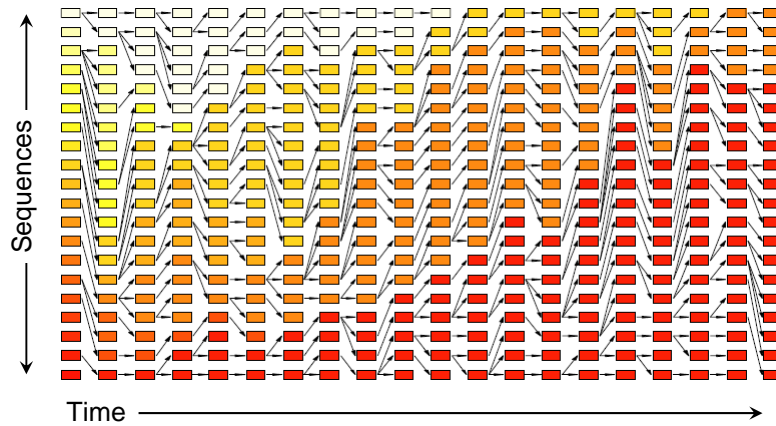
Today's Lecture

- Previously ...
 - Allele frequencies
 - Genotype frequencies
 - Hardy-Weinberg Equilibrium
- What next? – making predictions
 - What allele frequencies do we expect?
 - How much variation in a gene?
 - How are neighboring variants related?

Simple Approach: Simulation

1. N starting sequences
2. Sample N offspring sequences
 - Apply mutations according to μ
- Increment time
- If enough time has passed...
 - Generate final sample
 - Stop
1. Otherwise, return to step 1

Simulating a Population



Today

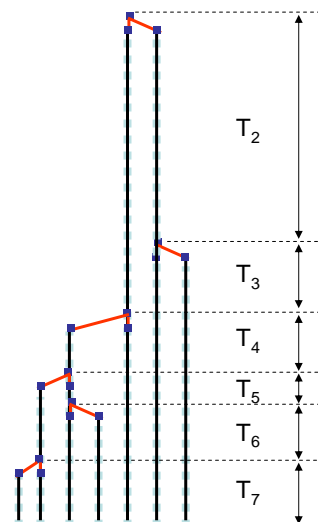
- Introduce coalescent approach
 - Framework for studying genetic variation
 - Provides intuition on patterns of variation
 - Provides analytical solutions

Goals

- Gene genealogies
 - Description of relatedness between sequences
 - Analogous to phylogenetic trees for species
- The shape of the genealogy depends on population history, selection, etc
- Together with mutation rate, genealogy predicts DNA variation

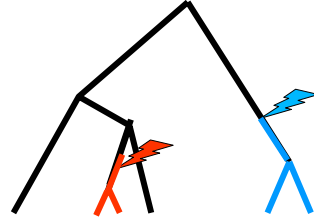
Genealogy

- History of a particular set of sequences
 - Describes their relatedness
 - Specifies divergence times
- Includes only a subset of the population
- Most Recent Common Ancestor (MRCA)
- **Coalescence**: the process in which, looking backward in time, the genealogies of two alleles merge at a common ancestor.



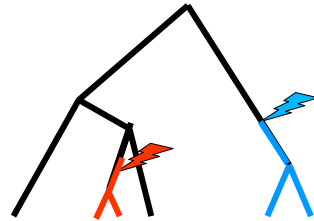
Two Mutations on a Tree

- Subtrees are either disjoint

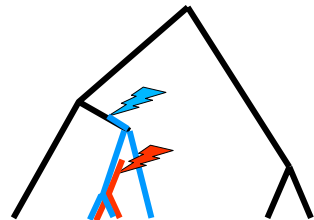


Two Mutations on a Tree

- Subtrees are either disjoint



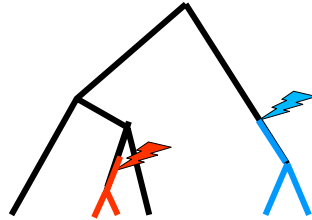
or contained in one another



Two Mutations on a Tree

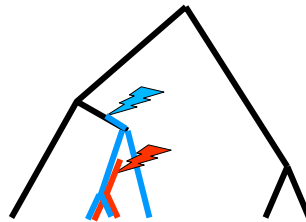
- Subtrees are either disjoint

Haplotypes: 00
01
10



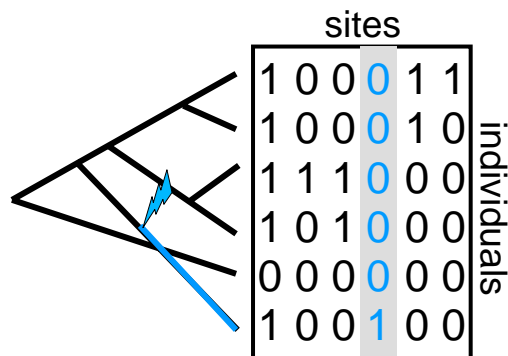
or contained in one another

Haplotypes: 00
01
11



An Unknown Tree

- Typical data: matrix M of *haplotypes*, w/o tree
- A *phylogeny*
 - A tree
 - mapping of sites \rightarrow branches, individuals \rightarrow leaves



Parameters we will focus on...

- Mutation rate (μ)
- Population Size
 - Haploid population (N chromosomes)
 - Diploid population (2N chromosomes)
- Time (t)
- Sample size (n)
- Recombination rate (r)

Mutation Model

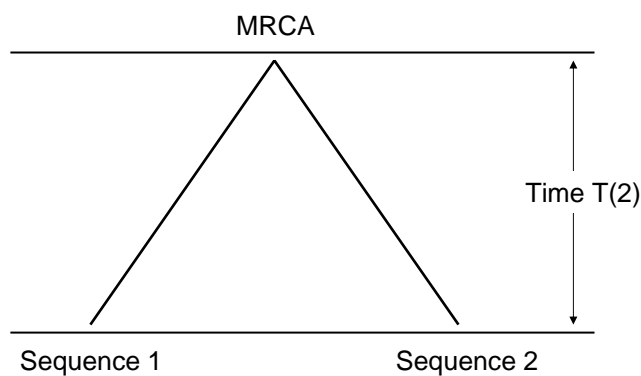
- The mutation process is complex
 - Rate depends on surrounding sequence
 - Reverse mutations are possible
- Two simple models are popular
 - Infinite alleles
 - Every mutation generates a different allele
 - Infinite sites
 - Every mutation occurs at a different site

Mutation Model

- Focus on infinite sites model
 - Mutation rate in genomic DNA is $\sim 10^{-8}$ / bp
 - Recurrent mutations should be very rare
- Scaled mutation rate parameter, e.g.:
 - 1000 bp sequence
 - 10^{-8} mutations per base pair per generation
 - $\mu = 10^{-5}$ per sequence per generation

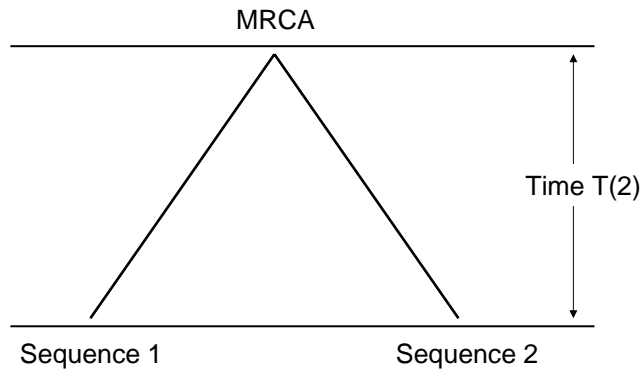
Genealogy of two sequences

- Mutations between MRCA and Sequence 1?



Genealogy of two sequences

- Total mutation in genealogy?



Estimating T(2)

- Probability that two sequences have distinct ancestors in previous generation:

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

- Probability of distinct ancestors for t generations is $P(2)^t$

Probability of MRCA at time t+1

$$P(2)^t (1 - P(2)) = \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t$$
$$\approx \frac{1}{2N} e^{-\frac{1}{2N}t}$$

For n>2

- Coalescence when two sequences have common ancestor
 - For simplicity, consider the possibility of multiple simultaneous coalescent events to be negligible
- Requirements for no coalescence:
 - Pick one ancestor for sequence 1
 - Pick distinct ancestor for sequence 2
 - Pick yet another ancestor for sequence 3
 - ...

Estimating P(n)

- Probability that n sequences have n distinct ancestors in previous generation:

$$P(n) = \prod_{i=1}^{n-1} \frac{2N-i}{2N}$$

$$\approx 1 - \frac{\binom{n}{2}}{2N}$$

- Assumptions

- N is large
- n is small
- Terms of order N^{-2} can be ignored

Probability of Coalescence at Time t+1

$$P(n)^t (1 - P(n)) \approx \frac{\binom{n}{2}}{2N} \left(1 - \frac{\binom{n}{2}}{2N} \right)^t$$

$$\approx \frac{\binom{n}{2}}{2N} e^{-\frac{\binom{n}{2}}{2N} t}$$

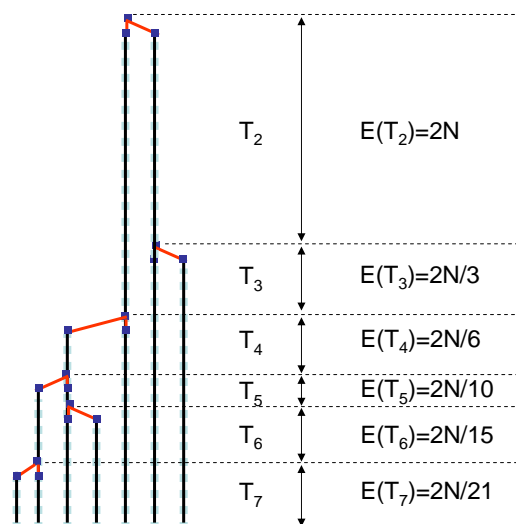
Time to next coalescent event

- Use an exponential distribution to approximate time to next coalescent event...

– Decay rate $\lambda = \frac{\binom{n}{2}}{2N}$

– Mean $\frac{1}{\lambda} = \frac{2N}{\binom{n}{2}}$

The coalescences in a gene tree when $N > 2$



Total "Time in Tree"

- Sum of all the branch lengths
- Total evolutionary time available
 - e.g. for mutations to occur

$$E(T_i) = \frac{2N}{\binom{i}{2}}$$

$$\begin{aligned} E(T_{tot}) &= E\left(\sum_{i=2}^n iT_i\right) = \sum_{i=2}^n iE(T_i) \\ &= \sum_{i=2}^n i \frac{4N}{i(i-1)} = \sum_{i=2}^n \frac{4N}{(i-1)} = 4N \sum_{i=1}^{n-1} \frac{1}{i} \end{aligned}$$

Number of segregating sites (mutations)

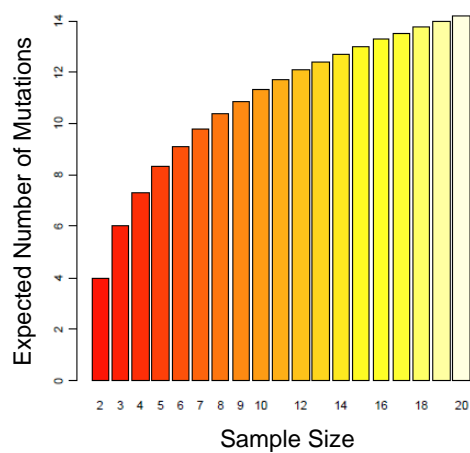
- Commonly named S
- Total number of mutations in genealogy
 - Assuming no recurrent mutation
 - A function of the total length of the genealogy T_{tot}
- Distributed as Poisson, conditional on total tree length
 - $E(S) = \mu E(T_{tot})$
 - $\text{Var}(S) = E[\text{Var}(S|T_{tot})] + \text{Var}[E(S|T_{tot})]$
 $= \mu E(T_{tot}) + \mu^2 \text{Var}(T_{tot})$
- T_{tot} is the total length of all branches

Expected number of mutations

$$\begin{aligned} E(S) &= E(\mu T_{tot}) \\ &= 4N\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i} \end{aligned}$$

- Population geneticists define $\Theta=4N\mu$

E(S) as a function of n



- Parameters
 - $N=10,000$ individuals
 - $\mu=10^{-4}$
 - $\Theta=4$

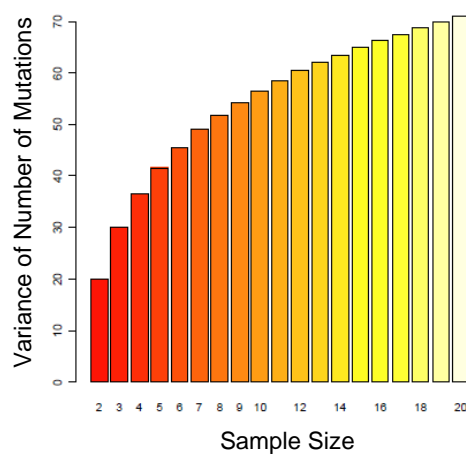
More about S...

- Very large variance

$$\text{Var}(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

- Most of the variance contributed by early coalescent events (i.e. with small n)

Var(S) as a function of n



- Parameters
 - $N=10,000$ individuals
 - $\mu=10^{-4}$
 - $\Theta=4$

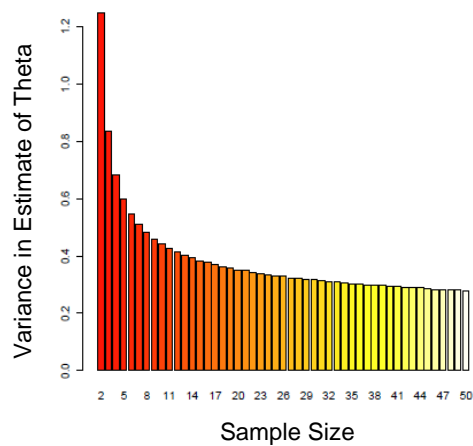
Inferences about Θ

- Could be estimated from S
 - Divide by expected length of genealogy

$$\hat{\theta} = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

- Could be then be used to:
 - Estimate N, if mutation rate μ is known
 - Estimate μ , if population size N is known

$\text{Var}(\hat{\Theta})$ as a function of N



- Parameters
 - N=10,000 individuals
 - $\mu=10^{-4}$
 - $\Theta=4$

Alternative Estimator for Θ

- Count pairwise differences between sequences
- Compute average number of differences

$$\tilde{\theta} = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=i+1}^n S_{ij}$$

Summary

- Probability of coalescence events
- Length of genealogy and its branches
- Expected number of mutations
- Simple estimates of Θ

Recommended Reading

- **Richard R. Hudson (1990)**
- *Gene genealogies and the coalescent process*
- Oxford Surveys in Evolutionary Biology, Vol. 7. D. Futuyma and J. Antonovics (Eds). Oxford University Press, New York.

Acknowledgement

- This lecture note is based on material by Profs Goncalo Abecasis (Univ of Michigan) and Itsik Pe'er (Columbia Univ).
- Chapter 3 of *Principles of Population Genetics* by Daniel L. Hartl & Andrew Clark