

Introduction to Coalescent Models

Biostatistics 666

Lecture 4

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

Previously ...

- DNA sequence variation
 - Types of DNA variants
- Allele frequencies
- Genotype frequencies
 - Hardy-Weinberg Equilibrium

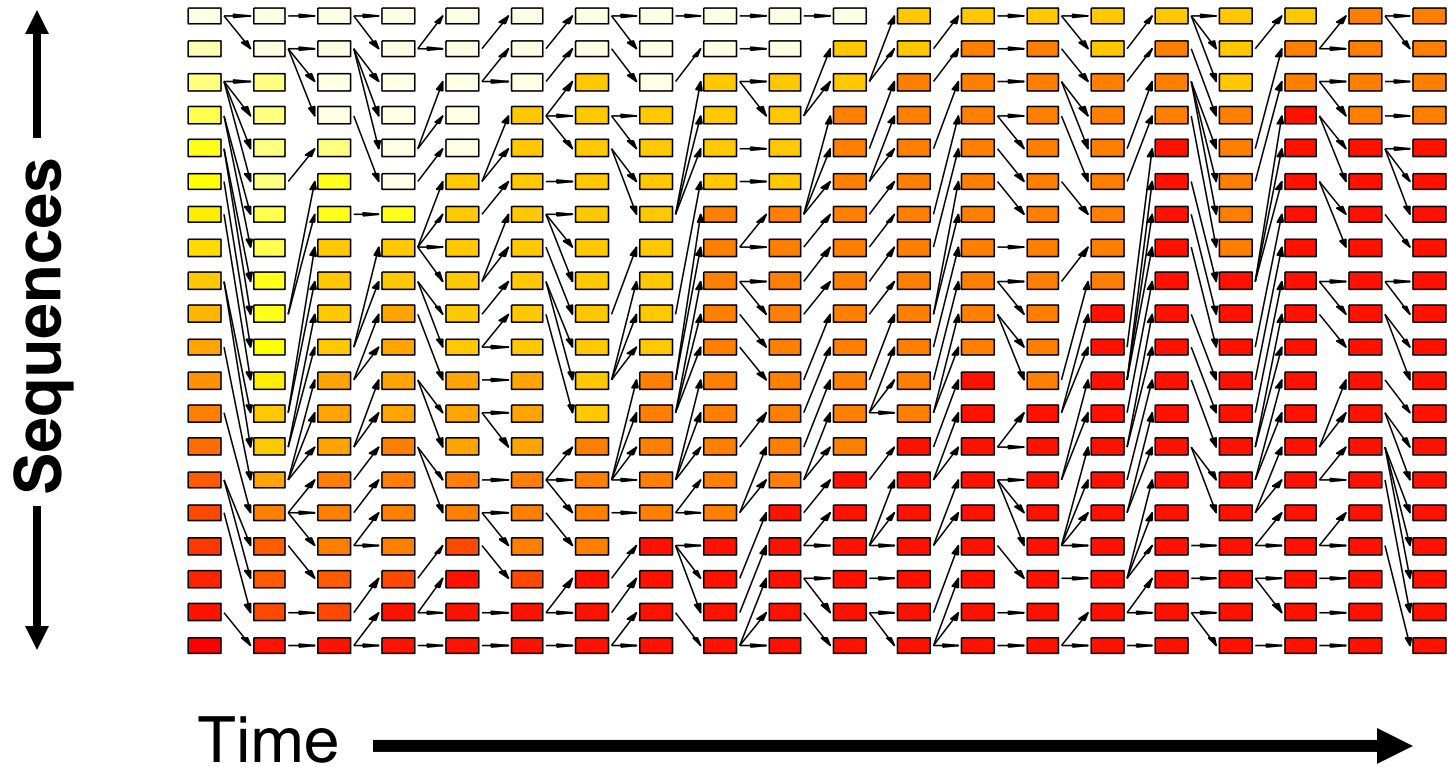
Making predictions...

- 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
3. Increment time
4. If enough time has passed...
 - Generate final sample
 - Stop.
5. 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

Aim ...

- Gene genealogies:
 - Descriptions 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

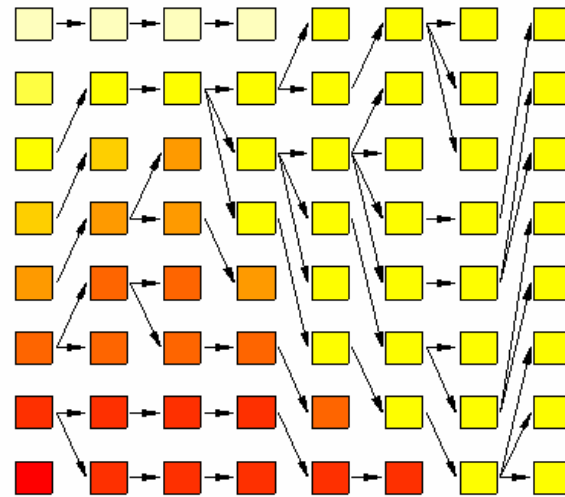
Genealogy

- 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)

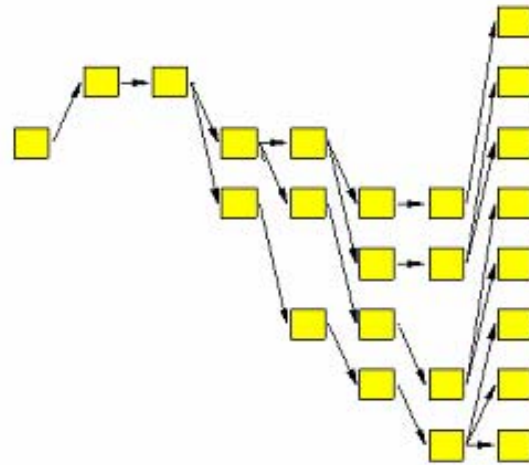
Coalescent approach

- Generate genealogy for a sample of sequences.
 - Introduces computational and analytical convenience.
- Instead of proceeding forward through time, go backwards!

History of the Population



Genealogy of Final Population



Levels of Complexity

- History of the population
 - Includes sequences that are “extinct”
- History of all modern sequences
 - Includes sequences that we haven’t sampled
- History of a subset of modern sequences
 - Minimalist approach!

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)

Other Parameters

- Selection
 - For gene of interest
 - For neighboring gene
- Demographic parameters
 - Migration
 - Population Structure
 - Population Growth

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

Neutral Variants

- Variants that have do not affect fitness
- Accumulate inexorably through time
 - Lost through genetic drift
- Do not affect genealogy

Example: Modeling Accumulation of Mutations

- Population of identical sequences
- Sample one descendant after t generations
- How many mutations have accumulated?
 - Hint: depends on mutation rate μ and time t
- Tougher questions
 - How many mutations have been fixed?
 - How much variation in the total population?

So far ...

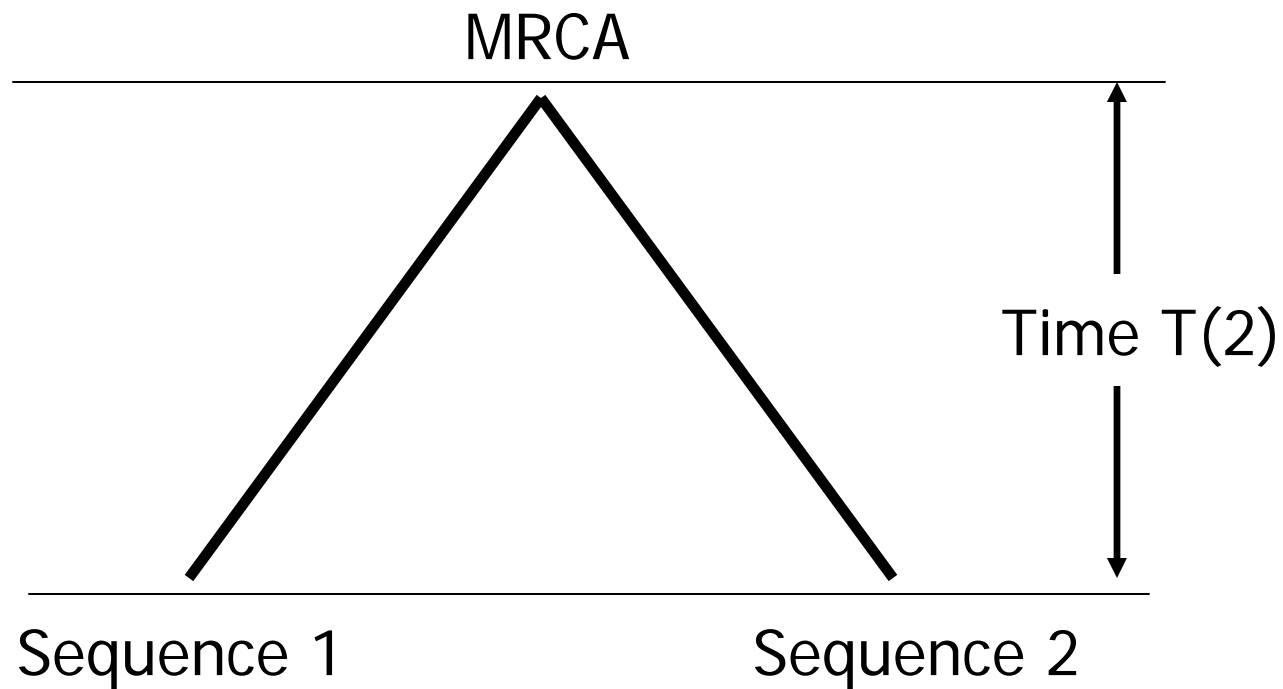
- Divergence of a single sequence
 - Accumulation of mutations
 - Depends on time t
 - Depends on mutation rate μ
 - Does not depend on population size N
 - Does not depend on population growth

- Next: A pair of sequences!

A tougher example ...

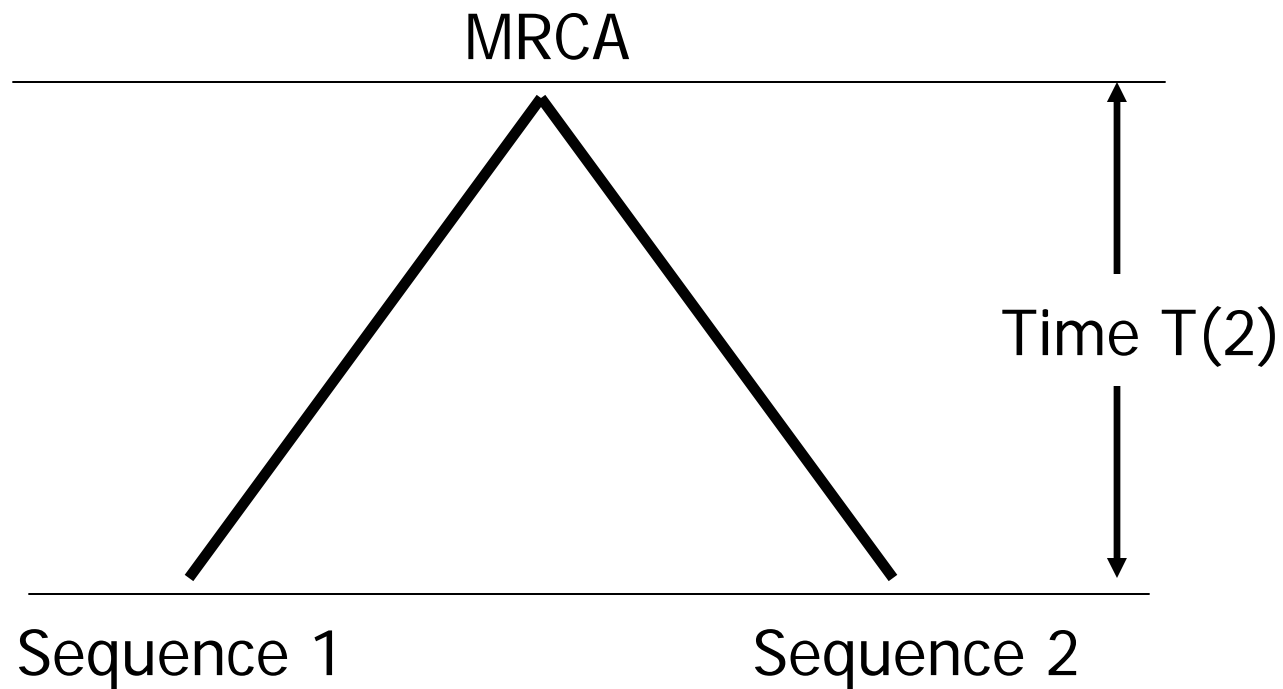
- Sample of two sequences
 - 100 bp each...
- How many differences are expected?
 - Population of size, $N = 1000$
 - Mutation rate
 - $\mu = 10^{-8}$ / bp / generation
 - $\mu \approx 10^{-6}$ / 100 bp / generation

Genealogy of two sequences



Mutations between MRCA and Sequence 1?

Genealogy of two sequences



Total mutations in genealogy?

Number of mutations S

- Distributed as Poisson, conditional on total tree length
 - $E(S) = \mu E(T_{\text{tot}})$
 - $\text{Var}(S) = E[\text{Var}(S|T)] + \text{Var}[E(S|T)]$
 $= \mu E(T_{\text{tot}}) + \mu^2 \text{Var}(T_{\text{tot}})$
- T_{tot} is the total length of all branches

Estimating $T(2)$

- Probability that two sequences have distinct ancestors in previous generation

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

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

Probability of MRCA at time $t+1$

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

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

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

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

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

Probability of Coalescence at Time $t+1$

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

Time to next coalescent event

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

$$\text{Decay Rate } \lambda = \frac{\binom{n}{2}}{N}$$

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

$T(j)$

- For convenience, measure time to next coalescent event in units:
 - N generations for haploids
 - $2N$ generations for diploids

$$E(T_j) = 1 / \binom{j}{2}$$

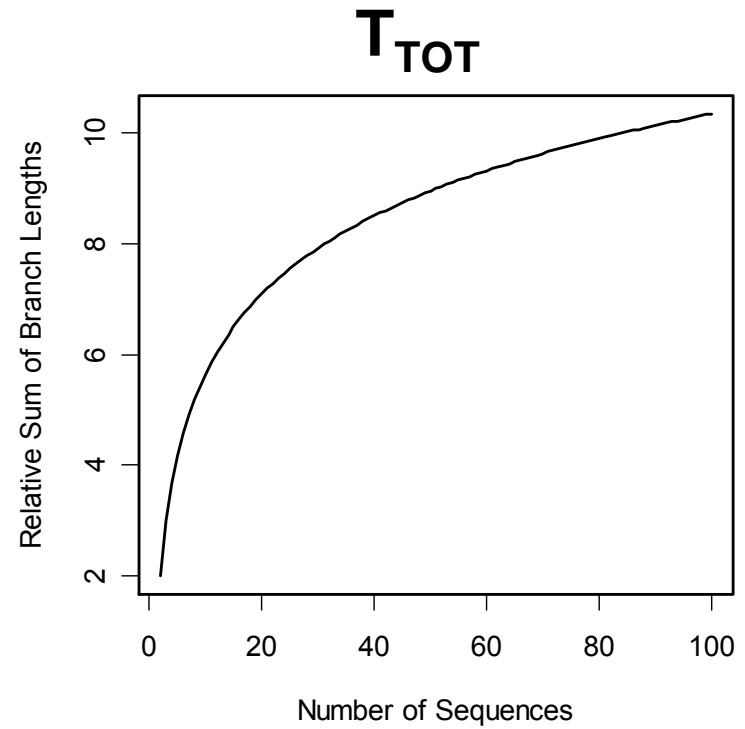
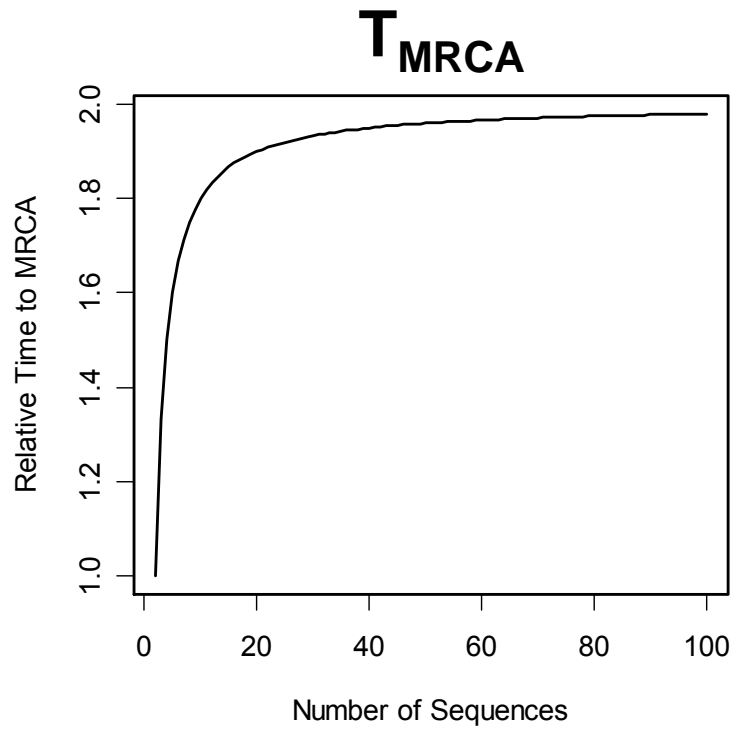
- How would you calculate time to MRCA of n sequences?

Total "Time in Tree"

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

$$\begin{aligned} E(T_{tot}) &= \sum_{i=2}^n iT(i) = \sum_{i=2}^n \frac{2i}{i(i-1)} \\ &= \sum_{i=2}^n \frac{2}{i-1} = \sum_{i=1}^{n-1} \frac{2}{i} \end{aligned}$$

T_{MRCA} VS. T_{TOT}



Number of Segregating Sites

- Commonly named S
- Total number of mutations in genealogy
 - Assuming no recurrent mutation
 - A function of the total length of the genealogy
 - T_{tot}

Expected number of mutations

- Factor N for haploids, $2N$ for diploids

$$\begin{aligned} E(S) &= 2N\mu \sum_{i=2}^n iE(T(i)) \\ &= 4N\mu \sum_{i=1}^{n-1} 1/i \\ &= \theta \sum_{i=1}^{n-1} 1/i \end{aligned}$$

- Population geneticists define $\theta=4N\mu$ (for diploids)
 - For gene mapping , θ is usually recombination rate
 - Population geneticists, use r for recombination rates

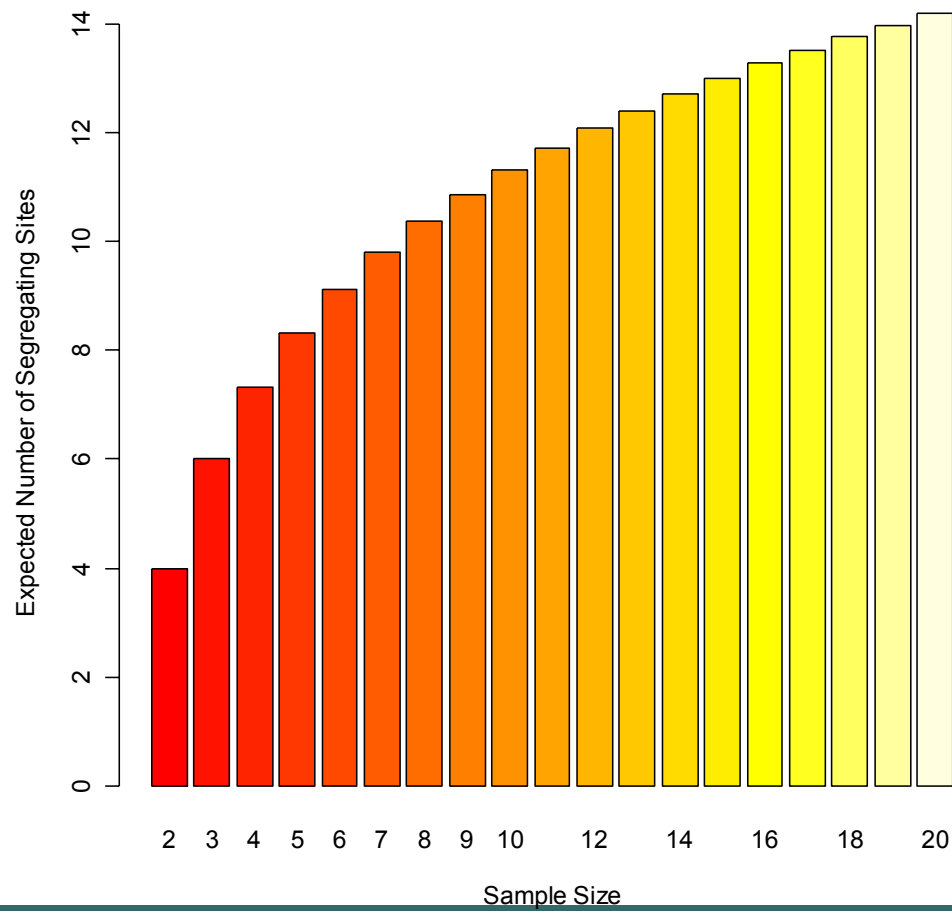
Expected number of mutations

- Factor N for haploids, $2N$ for diploids

$$\begin{aligned} E(S) &= 2N\mu \sum_{i=2}^n iE(T(i)) \\ &= 4N\mu \sum_{i=1}^{n-1} 1/i \\ &= \theta \sum_{i=1}^{n-1} 1/i \end{aligned}$$

- Population geneticists define $\theta=4N\mu$ (for diploids)
 - For gene mappers, θ is usually the recombination rate
 - Population geneticists, use r for recombination rates

$E(S)$ as a function of n



Parameters

$N = 10,000$ individuals

$\mu = 10^{-4}$

$\theta = 4$

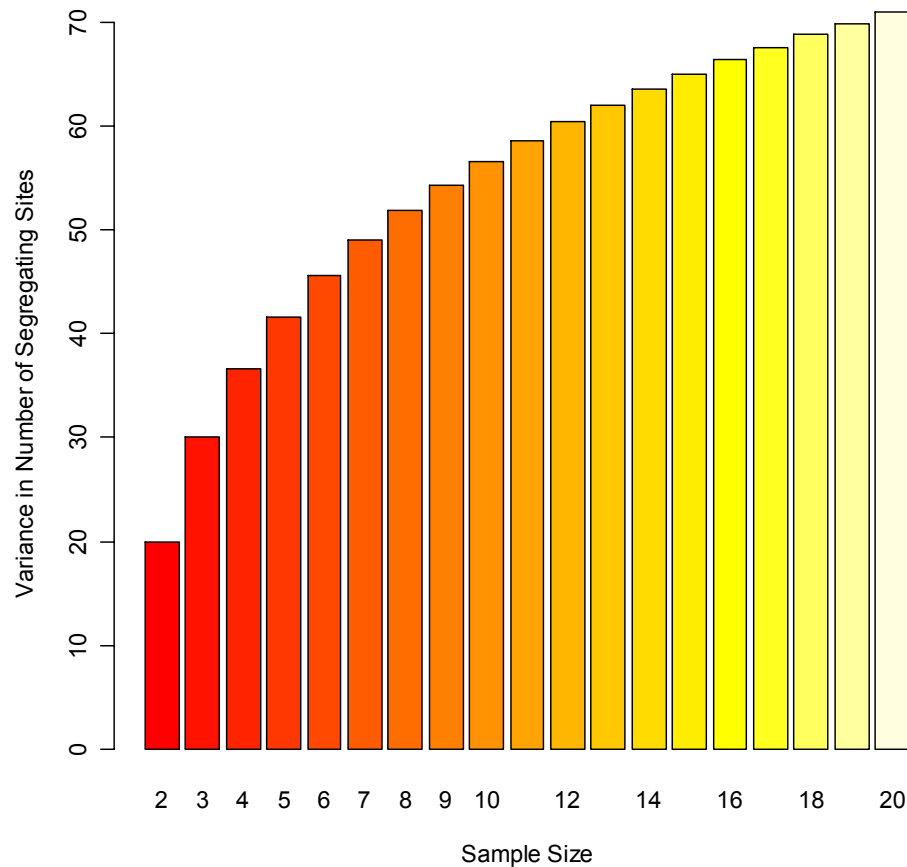
More about S ...

- Very large variance

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

- 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

$\mu = 10^{-4}$

$\theta = 4$

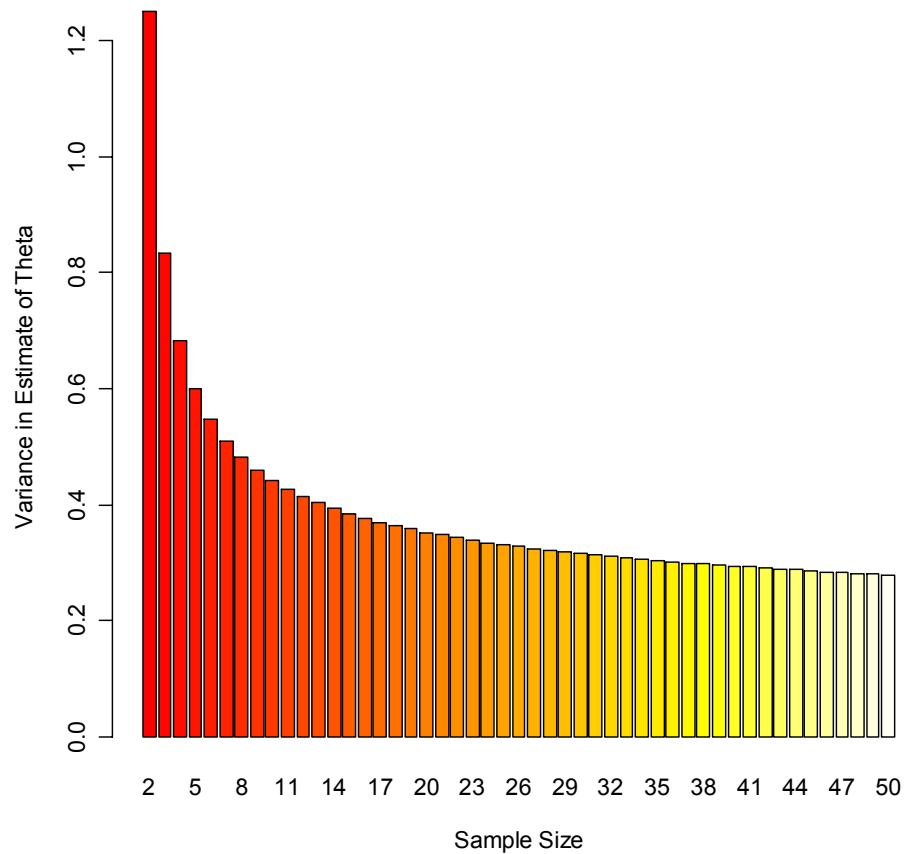
Inferences about θ

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

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

- Could 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

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

Today...

- 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.