

Scribed Notes Lecture 12

Vamsee Pillalamarri | 03-711

1. Introduction to Coalescent Models

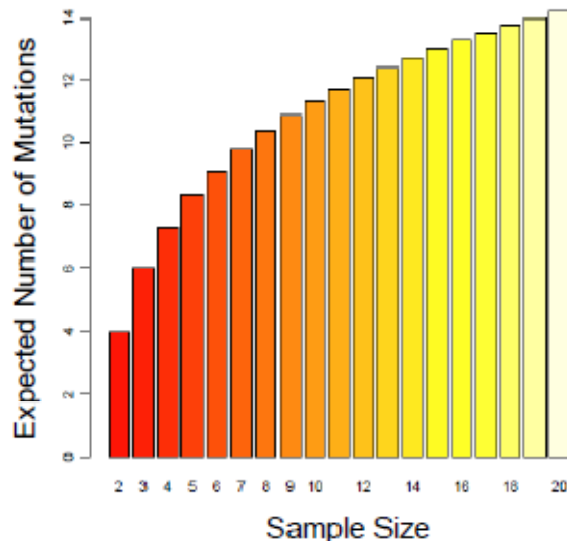
- a. Review of Last lecture
 - i. Previously we looked at calculating frequencies of alleles, genotypes (using HWE), Linkage Equilibrium and Linkage Disequilibrium, and measures of LD: D , D' , Δ^2 or r^2
- b. Now, we look at how to make Predictions about these various values:
 - i. –What allele frequencies do we expect?
 - ii. – How much variation in a gene?
 - iii. – How are neighboring variants related?
- c. One simple approach is to do a Simulation of a population growth:
 - i. Start with N sequences, then sample N offspring sequences (with mutation rate being μ , increment time step t , if $t = t_{final}$, then stop; otherwise, return to step1.
 - ii. You can then cluster sequences and populations using any clustering method.
- d. Coalescent Method – framework for studying genetic variation
 - i. Main Goals: **Gene Genealogies** (descriptors of relations between sequences – like phylogeny trees for interspecies relations)–they predict DNA variation using mutation rate and other parameters.
 - ii. Definitions:
 1. Genealogy: a tree that describes a subset of population sequences, and how they're related to each other.
 2. MRCA: Most Recent Common Ancestor
 3. Coalescence: the process in which, looking backward in time, the genealogies of two alleles merge at a common ancestor.
 - iii. We basically generate a genealogy for a sample of sequences, not the whole population, and proceed backwards in time from the leaves to the root.
 - iv. Parameters:
 1. Mutation Rate μ , Population size N or $2N$, Time t , sample size n , recomb rate r
 - v. *Mutation Model*: based on infinite alleles and infinite sites assumptions.
 1. Rate is equal to genomic DNA rate: $\sim 10^8$ muts / bp

2. Then, the scaled parameter $\mu = 10^{-5}$ mut / sequence, assuming 1000bp seq.
- vi. Genealogy time: $T(2)$ time difference between MRCA and sequences at leaves.
1. Probability of Coalescence Events, based on coalescent time T_N
 - a. Estimating $T(2)$: $P(2) = \frac{2N-1}{2N} = 1 - \left(\frac{1}{2N}\right)$
 - b. This is the prob of two sequences having distinct ancestors in previous generation. For t generations, it's $P(2)^t$
 - c. In general this equation is as follows:

$$P(n) = 1 * \left(\frac{2N-1}{2n}\right) \dots \left(\frac{2N-(n-1)}{2N}\right) \approx 1 - \left(\frac{nC2}{2N}\right)$$
 where $nC2 = n$ choose 2.
 - d. $P(n)^t \{1 - P(n)\} = \left(\frac{nC2}{2N}\right) e^{-\left(\frac{nC2}{2N}\right)t} \Rightarrow P(Tn)$ Where $T(n)$ is coalescence time from present generation. $P(Tn)$ is prob distribution of coalescent time. Thus, Coalescent time is exponentially distributed. Let $-\left(\frac{nC2}{2N}\right) = \lambda$. $E[Tn] = \frac{1}{\lambda}$.
 - e. So, the expected value of coalescent time from present generation is $E[Tn]$.
- vii. When coalescence happens, the total length of sequences reduces by 1. $n \rightarrow n-1$.
1. Total length = $\sum_{i=2}^n iTi$
- viii. For T_{n-1} generations, there will be $n-1$ branches. T_{tot} will be $E[T_{tot}] = \sum_{i=2}^n i * E[Ti] = 4N * \sum 1/i$
- ix. When we take μ into consideration, we can see that $E[Ti] = \mu 4N * \sum 1/i \dots$
- x. Thus, Parameter $\theta = \mu 4N$ that describes the **mutation rate** between generations. *Compare this to recombination rate $R = r 4N$ where r is the mutation rate of per generation mutation.*
- xi. Thus, the Expected number of mutations, where S is the term for mutations, is

$$E[S] = E[\mu T_{tot}] = 4N\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$
 where $\theta = 4N\mu$
- xii. When we estimate $E[S]$ as a function of n (sample size), see slide 30.

xiii. Graph: where $\theta = 4N\mu = 4 * (10,000) * (10^{-4}) = 4$



xiv. The expected number of mutations increases as sample size increases.

xv. The variance increases too, and there's *most* variance when n is small, pertaining to the most variance contributed by early coalescence events.

xvi. How do we infer θ ? We can use:

1. Divide S(number of mutations) by expected length of genealogy:

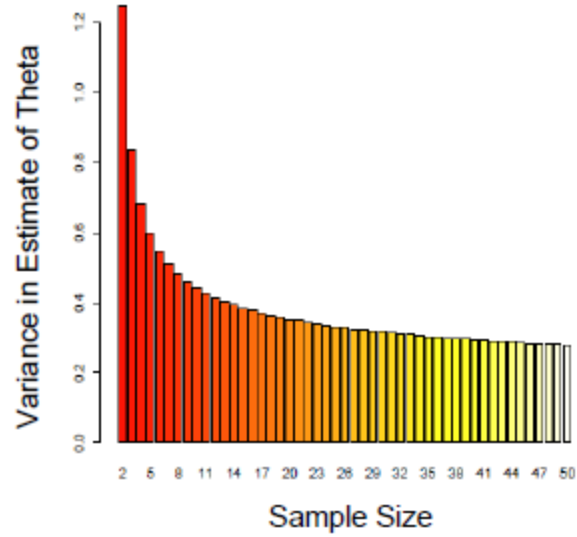
$$a. \hat{\theta} = \frac{S}{\sum_{i=1}^{n-1} \left(\frac{1}{i}\right)} \text{ where denominator is } T_{tot} \text{ (see above)}$$

xvii. We can then use θ to estimate N (given μ) or μ (given N) from the relation

$$\hat{\theta} = 4N\mu.$$

xviii. The variance of $\hat{\theta}$ as a function of N decreases as sample size n increases.

1. $var(\hat{\theta})$ is inversely proportional to n .



2. Thus, as the sample size increases, mutations (due to early coalescence events) goes down, thus decreasing the variance, and the values getting closer to the mean.

Thanks.