

Shannon Quinn  
Feb 23, Scribe 12  
Linkage Disequilibrium and Coalescent Models

Calculate LD for a generation  $t$  over  $N$  iterations

$$D_{AB}^{(t+N)} = (1-r)^N * D_{AB}^t$$

- $(1-r)$  is the recombination rate [ $0 \leq r \leq 1$ ]
- number of recombinations must be an odd number
- recombination rate increases with distance between loci
- spots in the genome where recombination occurs more frequently are recombination hotspots

Simulation studies

- given  $n$  sequences from generation  $t$ , predict generation  $t + 1$
- we can randomly sample from generation  $t$  to produce offspring, without assuming any recombination
- can also have mutations between generations

Coalescent Approach

- perform the simulation backwards: infer genealogical sequence
- construct a phylogenetic tree in which each leaf represents the sequence of an individual
- child nodes have inherited sequences from parents and all ancestors
- height of the tree (and each node) represents the amount of time between generations
  - MRCA = most recent common ancestor
- Subtrees are either disjoint or contained in each other

We need five parameters

1. Mutation rate ( $\mu$ )
2. Population size ( $N$  for haploid population,  $2N$  for diploid)
3. Time ( $t$ )
4. Sample size ( $n$ )
5. Recombination rate ( $r$ )

From the given sequences, how many mutations occurred from the MRCA?

- Estimate length of time (height of the tree)

In generation  $t - 1$ , take a random sample of one of  $2N$

- If  $N$  is large, probability that  $seq_1$  and  $seq_2$  have the same ancestor is small =  $1/2N$
- Probability that they have different ancestors =  $1 - (1/2N)$
- Probability that  $seq_1$  and  $seq_2$  have distinct ancestors through  $t$  generations backwards =  $(1 - (1/2N))^t$
- Therefore, in generation  $t + 1$ , we have  $1/2N * (1 - (1/2N))^t$
- When  $N$  is large

$$\approx (1/2N) * \exp((-1/2N) * t)$$

Now, considering n sequences

- $P(n)$  = n sequences have n distinct ancestors in the previous generation

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

- And when N is large

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

Probability some sort of coalescence occurs, or the probability of some common ancestor in generation  $t + 1 = P(n)^t * (1 - P(n))$

$$= \frac{\binom{n}{2}}{2N} \left(1 - \frac{\binom{n}{2}}{2N}\right)$$

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

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

- Both these values are used to approximate time between coalescent events

$E(T_7)$  = expected number of generations to pass when 7 distinct sequences have a single coalescent event between any 2 of them

$$= \frac{2N}{\binom{7}{2}}$$

Total length of the three =  $T_1 + T_2 + \dots + T_n$

$T_{\text{tot}}$  = total length of all branches added together

$$= \sum_{i=2}^n iT_i$$

Assume mutation rate is  $\mu$

- S is number of mutations
- $\text{rate}(S) = \mu T_{\text{tot}}$

$$= 4N \sum_{i=1}^{n-1} \frac{1}{i}$$

$$E(T_{\text{tot}}) = 4N\mu \sum_{i=1}^{n-1} \frac{1}{i}$$

$(4N\mu)$  can be represented as  $\theta$

How can we infer  $\theta$ ?

- Assuming the number of mutations observed is the expected number (assuming  $E(S) = S$ ), and given a set of sequences, we have:

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

- Another way of estimating  $\theta$

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