The Coalescent: Expectations of the Past
Dem260 Math Demog
Spring 2020
Lecture 11

Joshua R. Goldstein

April 9, 2020

Agenda for today

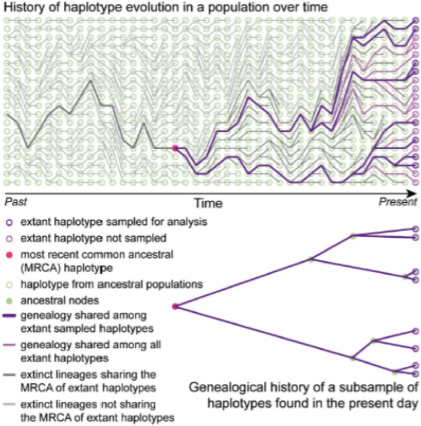
1. Big picture: What is “coalescent theory”?
2. Time to TMRCA
3. Inferring population size

Big picture

Nomenclature:

- Coalescent theory is not a theory.
- It's a model for the probability of different histories
- "the" coalescent is a bit confusing. We're not inferring the actual history of common ancestry, just the probabilities

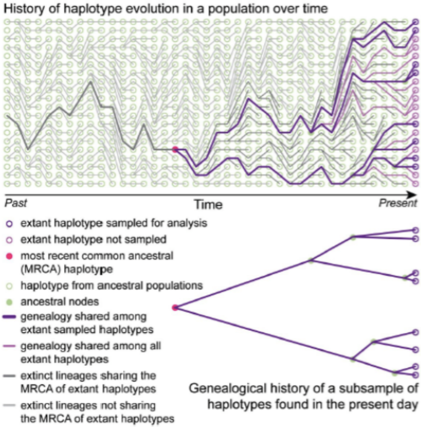
An actual "picture"


History of haplotype evolution in a population over time
Past
Time
Present
extant haplotype sampled for analysis
extant haplotype not sampled
most recent common ancestral (MRCA) haplotype
haplotype from ancestral populations
ancestral nodes
genealogy shared among extant sampled haplotypes
genealogy shared among all extant haplotypes
extinct lineages sharing the MRCA of extant haplotypes
extinct lineages not sharing the MRCA of extant haplotypes
Genealogical history of a subsample of haplotypes found in the present day

From book by A. Cutter (2019).

- Top panel is a Fisher-Wright instance, ordered so that lines don't cross.
- Haplotype is a sequence (we are diploids, each contributing 2 haplotypes). But let's just think of each line as an individual, for now.
- We can find The Most Recent Common Ancestor (TMRCA) of sample (dark purple). Who and when would the MRCA of the top two individuals be?

More on the “picture”


History of haplotype evolution in a population over time
Past
Time
Present
extant haplotype sampled for analysis
extant haplotype not sampled
most recent common ancestral (MRCA) haplotype
haplotype from ancestral populations
ancestral nodes
genealogy shared among extant sampled haplotypes
genealogy shared among all extant haplotypes
extinct lineages sharing the MRCA of extant haplotypes
extinct lineages not sharing the MRCA of extant haplotypes
Genealogical history of a subsample of haplotypes found in the present day

From book by A. Cutter (2019).

- Our sample $\neq$ even all *extant* descendants of the MRCA. What does this mean?
- Our sample $\neq$ all of the descendants of the MRCA. What does this mean?
- If we chose two descendants at random, would we always get same MRCA?

When we model coalescence we are thinking backwards in time.

Our first question: When was MRCA?

If we sample two individuals (today), how long ago was their MRCA?
(Note: question is not “who”)

- Our answer will in terms of the probability of MRCA being 1 generation ago, 2 generations ago, etc.
- We’ll assume Fisher-Wright (constant N, each gen randomly picks parents)
- The answer is surprisingly simple :)

Our answer (1)

Let's assume we have N lines in Fisher-Wright (Note: I'm not using $2N$.)

- The chance that two sampled people have same parent is $1/N$, right?
- Thus $P(T_{MRCA} = 1) = 1/N$.
- What is $P(T_{MRCA} = 2) =$?
- What is $P(T_{MRCA} = n) =$?

Our answer (2)

Let's go to continuous time (reasonable if pop is big and time scale is long).
Hazard of coalescence $= c = 1/N$. Probability of coalescence at time $t = \ell(t)h(t) = e^{-ct}c$
What is expected time of coalescence? Think life expectancy.

Our answer (2)

Let's go to continuous time (reasonable if pop is big and time scale is long).
Hazard of coalescence $= c = 1/N$. Probability of coalescence at time $t = \ell(t)h(t) = e^{-ct}c$
What is expected time of coalescence? Think life expectancy.
$E(T_{MRCA})$ if two samples: $1/c = 1/(1/N) = N$

TMRCA

Let's simulate

- 1 time, without random seed, letting N = 40, ngen = 200, mu = 0
- Average over 100 FW simulations
- What is variance of outcome? Is it what we would expect from exponential?

Break

Mutation and inference of TMRCA and N

- Say mutations occur at a constant rate μ (10^{-8}?)
- Each year we would expect μ mutations, and over T years we would expect $T\mu$ mutations.
- Say we observe that two people differ at k sites of the genome.
    - When was TMRCA?
    - How big is the population?

Mutation, continued

- When was TMRCA?
- How big is the population?

Picture (Λ) Tree length $= 2T$ Expected number of mutations:

$$E(k) = E(2T\mu) = \bar{T}2\mu$$

Since,

$$\bar{T} = E(TMRCA) = N$$

If we observe on average $\bar{k}$ mutations, then

$$E(k) = N2\mu \rightarrow \hat{N} = \frac{\bar{k}}{2\mu}$$

Inference of population size, simulation

- We do FW with mutations
- Average pairwise differences
- Divide by 2μ to get our estimate
- We can repeat a bunch of times and see average estimate converges to the truth

Break-out.

Break number 2

Coalescence of a sample of n individuals

- This is covered on pages 42 and 43 of Gillespie
- We'll just do one quick example, accepting the result

A sample of 3: Note we're using N (instead of $2N$)

```
        *            -------
       / \
      /   \
     /     \         T(2) : E(T(2)) = N
    /       \
   /         *       -------
  /         / \
 /         /   \     T(3) : E(T(3)) = N * 2/[3*2]
*         *     *    -------

                      ...
                     -------
                     T(n) : E(T(n)) = N * 2/[n * (n-1)]
```

Question: If we sample 4, how much of time to TMRCA is do we have 4 branches, 3 branches, and 2 branches?

For next time

- Varying population sizes
- A real life example inferring population history from real mitochondrial DNA sequences.