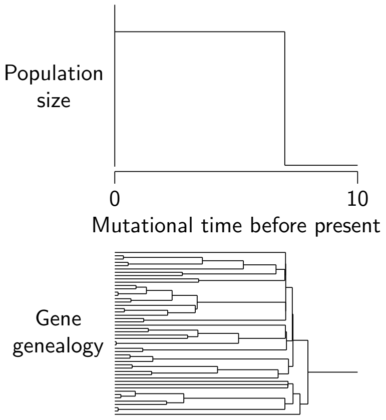
Gene Genealogies

Alan R. Rogers

February 6, 2023



Gene genealogy: ancestry of a sample of gene copies.

Coalescent event: when lines of descent coalesce at common ancestors.

Shape reflects history of population size.

Coalescent theory connects history to genetics via gene genealogy.

Hazards

The *hazard* of an event at time t is the conditional probability that it occurs then, given that it did not occur earlier.

We'll be interested in the hazards of coalescent events.

Preliminaries: 2 mathematical tricks

1. If the hazard of a coalescent event is h per generation, then we wait on average $1/h$ generations until the event happens.
2. There are $k(k-1)/2$ ways to choose 2 items out of k.

See *Lecture Notes on Gene Genealogies* (available online) for details.

Coalescence time in a sample of two gene copies

```
X -----------------------|
                         A------
Y -----------------------|

|--------- t ------------|
```

X and Y live now; ancestor A lived t generations ago.

How can we calculate $E[t]$—the expected "coalescence time?"

Coalescent hazard in a sample of 2 gene copies

```
               ________Population__________
Parents        0  Z  0  0  0  0  0  0  0  0
                  |
Offspring      0  X  0  Y  0  0  0  0  0  0
```

A population comprising 10 gene copies ($2N = 10$). Offspring generated by sampling with replacement from parents.

Z is the parent of X. What is the probability that Y has the same parent?

$$1/10 \qquad \text{or} \qquad 1/2N$$

$h_2 = 1/2N$ is the hazard of a coalescent event in a sample of 2 gene copies.

Expected coalescence time in a sample of 2 gene copies

We just saw that the hazard of a coalescent event is $h = 1/2N$ per generation.

Recall that if the hazard is h_2, the expected waiting time is $E[t_2] = 1/h_2$.

The expected coalescence time in a sample of 2 gene copies is

$$E[t_2] = 2N$$

generations.

Example: if $N = 10,000$, $E[t_2] = 20,000$ generations, or about 60,000 human years.

Coalescent hazard in a sample of K

```
W --------|
          |------------|
X --------|            |
                       |--------------|
Y ---------------------|              |---
                                      |
Z ------------------------------------|

|-- t_4 --|---- t_3 ---|------ t_2 ---|
```

With 4 gene copies, there are 3 coalescent intervals: t_4, t_3, and t_2.

The intervals are independent, so we can consider them one at a time.

Expected length of a coalescent interval containing 3 lines of descent

With 3 lines of descent (X, Y, and Z), there are 3 pairs of lines (XY, XZ, and YZ).

For each pair, we already know the coalescent hazard: $h_2 = 1/2N$.

We have 3 pairs, so the coalescent hazard is 3× as large: $h_3 = 3/2N$. (This argument is loose, but the answer is correct.)

Expected length of an interval with 3 lines of descent:

$$E[t_3] = 1/h_3 = 2N/3$$

Expected length of a coalescent interval containing i lines of descent

With i lines of descent, there are $i(i-1)/2$ pairs of lines. (See mathematical trick 2 above.)

Coalescent hazard of each pair: $h_2 = 1/2N$.

Hazard within an interval with i lines of descent:

$$h_i = \frac{i(i-1)}{4N}$$

Expected length of an interval with i lines of descent:

$$E[t_i] = 1/h_i = \frac{4N}{i(i-1)}$$

Coalescent intervals in a sample of size 4

Interval	Coalescent hazard	Expected length
4	$h_4 = \frac{4 \times 3}{4N} = 6/2N$	$2N/6$
3	$h_3 = \frac{3 \times 2}{4N} = 3/2N$	$2N/3$
2	$h_2 = \frac{2 \times 1}{4N} = 1/2N$	$2N$

Expected depth of a gene genealogy

"Depth" is the expected time (generations) since the last common ancestor (LCA).

Sample size	Mean depth of tree
2	$1/h_2 = 2N$
3	$1/h_3 + 1/h_2 = 8N/3$
4	$1/h_4 + 1/h_3 + 1/h_2 = 3N$
5	$1/h_5 + 1/h_4 + 1/h_3 + 1/h_2 = 16N/5$

In general, mean depth is

$$4N(1 - 1/K)$$

where K is the number of gene copies in the sample.

Deriving the formula

$E[t_i] = \frac{4N}{i(i-1)}$, so expected depth is

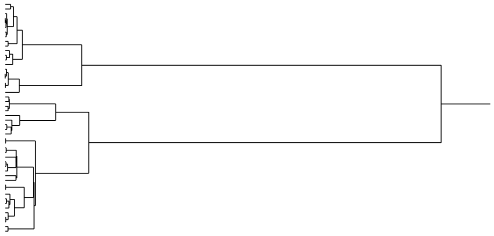
$$4N \sum_{i=2}^{K} 1/i(i-1)$$

To simplify this sum, convince yourself that

$$\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$$

Then substitute the right side into the sum above and write the sum out in expanded form. You should end up with $4N(1 - 1/K)$.

A simulated genealogy of 50 gene copies



Time Before Present →

Recent coalescent intervals are short; ancient ones are long.

The larger the sample, the shorter will be recent intervals.

Large samples don't help much, because few mutations appear on these short intervals.

Review

- What is a coalescent event?
- What is a hazard?
- What is the hazard of a coalescent event in an interval with i lines of descent?
- What is the expected length of such an interval?
- What is the expected depth of a gene genealogy with K tips?
- Why are coalescent intervals longer in large populations?
- Why are they shorter in large samples?

Relating Gene Genealogies to Genetics

Alan R. Rogers

February 6, 2023

Gene genealogies are not observable: we can never know the true genealogy of a sample. To *estimate*, we need a theory that relates gene genealogies to observable genetic data.
This lecture will:

1. add mutations to gene genealogies,
2. derive theory for the number, S, of segregating sites, and
3. the mean pairwise difference, π, between sequences.

A genealogy with 8 (not 9) mutations

```
--x----x---
           |
           |------x-
           |        |
------xx---         |
                    |-----x----
                    |
                    |
---x-x----------x---

|---t -----|---t ---|
     3          2
```

We are interested in mutations downstream of the root, because only these contribute to variation.

Expected # of mutations in a sample of size 3

```
--x----x---
           |
           |------x-
           |        |
------xx---         |
                    |---------
                    |
                    |
---x-x----------x---

|---t -----|---t ---|
     3          2
```

$$E[\#mutations|L] = uL$$

where u is mutation rate and L is total branch length.

$$L = 3t_3 + 2t_2$$

Unconditionally, $E[\#mutations] = E[uL] = uE[L]$.

For a tree with 3 tips

$$L = 3t_3 + 2t_2$$

Recall that $E[t_i] = 4N/i(i-1)$. Therefore,

$$\begin{aligned} E[L] &= 3E[t_3] + 2E[t_2] \\ &= (3 \times 2N/3) + (2 \times 2N) \\ &= 6N \end{aligned}$$

Expected number of mutations is $6Nu$ for a tree with 3 tips.

Expected length of tree with K tips

Expected length of the coalescent interval w/ i lines of descent

$$E[t_i] = \frac{4N}{i(i-1)}$$

Contribution to $E[L]$:

$$iE[t_i] = \frac{4N}{i-1}$$

Total expected length:

$$E[L] = \sum_{i=2}^{K} iE[t_i] = 4N \sum_{i=1}^{K-1} \frac{1}{i}$$

The expected number of mutations

$$\begin{aligned} E[\text{\# of mutations}] &= uE[L] \\ &= 4Nu \sum_{i=1}^{K-1} \frac{1}{i} \\ &= \theta \sum_{i=1}^{K-1} \frac{1}{i} \\ &= \theta \left\{ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{K-1} \right\} \end{aligned}$$

Mutation: the model of infinite sites

Mutation never strikes the same site twice, so the number of nucleotide differences between two sequences equals the number of mutations that separate them.

This is an approximation that works well with intraspecific data sets, because the mutation rate is so low that few sites mutate more than once.

The expected number, S, of segregating sites

S is the number of segregating (i.e. polymorphic) sites in a set of sequence data.

Under infinite sites, its expectation equals the expected number of mutations on the tree:

$$E[S] = \theta \left\{ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{K-1} \right\}$$

The effect of sample size is modest

K	$\sum_{i=1}^{K-1} 1/i$
2	1.00
3	1.50
5	2.08
10	2.82
100	5.17
1000	7.48

For practical purposes, $E[S]$ is θ times a number between 2 and 5.

The mean pairwise difference, π

The mean pairwise difference, π, is the mean number of nucleotide site differences between pairs of sequences in a sample.

In other words, it is the number of segregating sites in a sample of size 2.

Using our formula for $E[S]$,

$$E[\pi] = \theta \sum_{i=1}^{1} \frac{1}{i} = \theta$$

Two ways to estimate θ

$$\begin{aligned} \hat{\theta}_S &= \frac{S}{\sum_{i=1}^{K-1} \frac{1}{i}} \\ \hat{\theta}_\pi &= \pi \end{aligned}$$

Here $\hat{\theta}$ is read "theta hat." The "hat" indicates that these formulas are intended to estimate the parameter θ.

Since these formulas estimate the same parameter, we might expect them to be similar in real data. Are they?

Discrepancy between $\hat{\theta}_S$ and π

Mitochondrial sequence data published by Lynn Jorde's lab, describing 77 Asians and 72 Africans:

	Asian	African
S	82	63
$\sum_{i=1}^{K-1} 1/i$	4.915	4.847
$\hat{\theta}_S$ (per sequence)	16.685	12.998
π (per sequence)	6.231	9.208

Contrary to expectation, θ_S is much larger than π. Why?

Thinking about this discrepancy

S is equally sensitive to mutations anywhere in the gene genealogy. π, the MPD, is less sensitive to singletons than to mutations of intermediate frequency.

```
   00000 00001
   12345 67890
S1 AAACT GTCAT
S2 ..... A....
S3 ..... A...C
S4 ..G.. A....
S5 ..G.. A....
S6 ..G.. A....
     ^   ^
     |   |
     |    ------ Contributes 1 X 5 = 5 pairwise diffs
      ---------- Contributes 3 X 3 = 9 pairwise diffs
```

What does this imply about Jorde's data?

In the data, $\hat{\theta}_S \gg \pi$.

Suggests there are many young mutations (near the tips of the gene genealogy), where they affect $\hat{\theta}_S$ more than π.

As we'll learn later in the course, this implies a history of population growth.