

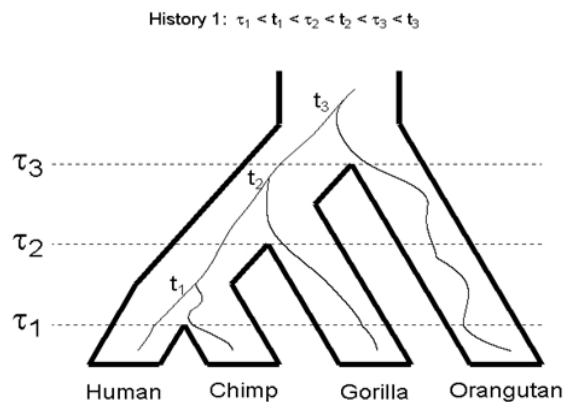
16

Coalescent Models

16.1: Learning Objectives

By the end of this chapter, the reader should be able to describe the coalescent process for a sample of gene copies, compute expected coalescence times under the standard neutral (Kingman) coalescent, incorporate a simple mutation model, and relate coalescent theory to allele-frequency distributions and effective population size in a Nigerian population-genetics context.

16.2: The Genealogical (Backward-in-Time) Perspective



A gene genealogy (gene tree) for four species — Human, Chimp, Gorilla, Orangutan — nested within, and potentially discordant with, the species tree.

Classical population genetics (Chapters 3–5) works **forward in time**, tracking how allele frequencies change from one generation to the next. **Coalescent theory**, developed principally by J. F. C. Kingman (1982), instead works **backward in time**: starting from a sample of n gene copies (alleles) drawn from a present-day population, it models the random process by which pairs of lineages successively find their **most recent common ancestor (MRCA)** — a **coalescence event** — as one traces genealogies back through the generations, until eventually all n sampled lineages have coalesced into a single ancestral lineage.

This backward perspective is often far more computationally and statistically efficient than simulating an entire population forward in time, because it need only track the (much smaller) sample's genealogy, not the whole population — and it directly produces the genealogical tree structure that underlies patterns of genetic variation, linkage disequilibrium (Chapter 15), and disease-mapping (Chapter 17) methods.

16.3: The Simple (Wright–Fisher) Coalescent Model

Consider a population of constant effective size N (for simplicity, assume N diploid individuals, so $2N$ gene copies per generation, reproducing under the idealised **Wright–Fisher model** of random mating, non-overlapping generations, and no selection). Sample two gene copies (lineages) from the present generation. Looking one generation back, the probability that these two lineages **coalesce** (i.e., that both copies were, in fact, copies of the same parental gene copy in the previous generation) is:

$$P(\text{coalescence in 1 generation}) = 1/(2N)$$

since the second lineage's parent gene copy is equally likely to be any one of the $2N$ gene copies in the previous generation, one of which is the first lineage's own parent.

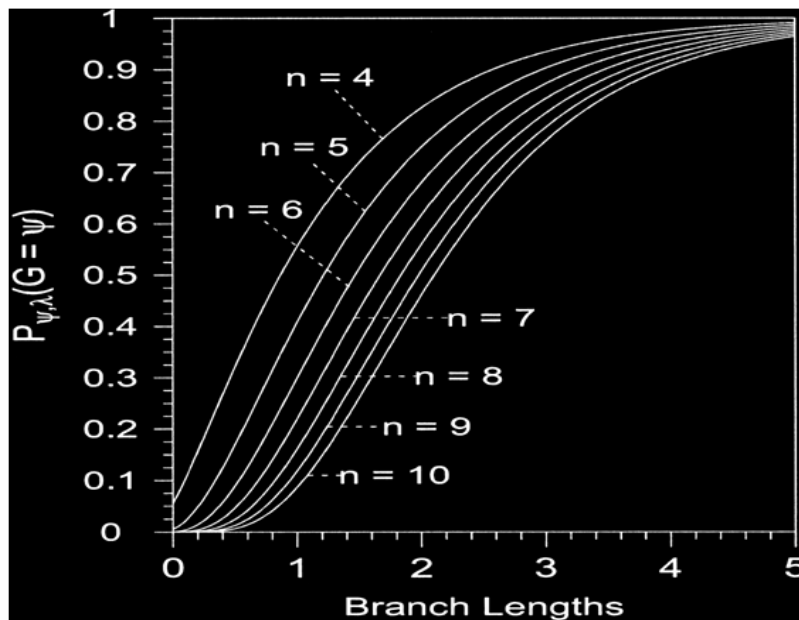
The probability that the two lineages have **not yet coalesced** after t generations, but do coalesce exactly at generation t , follows a geometric distribution:

$$P(T = t) = (1 - 1/2N)^{t-1} \times 1/(2N)$$

For large N , this geometric distribution is well approximated by an **exponential distribution** with rate $1/(2N)$, so that the expected time to coalescence of a pair of lineages is:

$$E(T) = 2N \text{ generations}$$

16.4: Coalescence Times for a Sample of n Lineages



The cumulative probability of coalescence as a function of branch length, shown for increasing sample sizes n .

For a sample of n gene copies, coalescence proceeds as a sequence of pairwise merging events, reducing the number of distinct lineages from n to $n-1$, then to $n-2$, and so on down to 1. When there are k lineages remaining (out of the original

n), any one of the $C(k,2) = k(k-1)/2$ possible pairs may coalesce in a given generation, so the coalescence rate at that stage is $C(k,2)/(2N)$, giving an expected waiting time (until the next coalescence event, reducing k lineages to $k-1$) of:

$$E(T_k) = 2N/C(k,2) = 4N/[k(k-1)]$$

The total expected time to the MRCA of the whole sample (the sum of the successive waiting times, from $k = n$ down to $k = 2$) is:

$$E(T_{MRCA}) = \sum_{k=2}^n 4N/[k(k-1)] = 4N(1 - 1/n)$$

Notably, as n grows large, $E(T_{MRCA})$ converges to $4N$ generations — the majority of the expected time to the MRCA of a large sample is actually accounted for by the final coalescence of just the last two remaining lineages (since $C(2,2) \dots$ rather $C(2,2)=1$ gives the largest, slowest expected waiting time, $2N$, dominating the sum), a key qualitative insight of coalescent theory: adding more samples yields diminishing genealogical “depth,” so modestly sized samples already capture most of a population’s deep genealogical history.

Theorem 16.1: (Expected Time to the Most Recent Common Ancestor)

For a sample of n lineages under the standard coalescent with constant effective population size N , $E(T_{MRCA}) = \sum_{k=2}^n \frac{4N}{k(k-1)} = 4N \left(1 - \frac{1}{n}\right)$.

Proof. Using the partial-fraction identity $\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$, the

sum telescopes:

$$\begin{aligned}\sum_{k=2}^n \frac{4N}{k(k-1)} &= 4N \sum_{k=2}^n \left(\frac{1}{k-1} - \frac{1}{k} \right) \\ &= 4N \left[\left(\frac{1}{1} - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \dots \right. \\ &\quad \left. + \left(\frac{1}{n-1} - \frac{1}{n} \right) \right].\end{aligned}$$

Every intermediate term cancels (each $-1/k$ is cancelled by the following term's $+1/k$), leaving only the first and last terms:

$$= 4N \left(\frac{1}{1} - \frac{1}{n} \right) = 4N \left(1 - \frac{1}{n} \right),$$

which is the stated result. As $n \rightarrow \infty$, $1/n \rightarrow 0$, so $E(T_{MRCA}) \rightarrow 4N$, confirming the “diminishing depth” property described above: the marginal contribution of each additional sampled lineage to the expected total depth shrinks rapidly, since $4N(1 - 1/n)$ is already within 10% of its limiting value $4N$ once $n = 10$. ■

Worked Example 16.2 (Estimating TMRCA for an HBB haplotype lineage in a Nigerian population sample)

A population geneticist samples $n = 15$ HBB-region haplotypes from a Nigerian cohort to study the genealogical history of a specific malaria-associated haplotype background, assuming an effective population size of $N = 12,000$ (a plausible historical effective size for a sub-Saharan African population, smaller than the Nigerian census population owing to long-term demographic history, Chapter 14). By Theorem 16.1, $E(T_{MRCA}) = 4(12,000) \left(1 - \frac{1}{15} \right) = 48,000 \times \frac{14}{15} = 44,800$ generations, or, at roughly 25 years per generation, approximately 1.12 million years —

illustrating that even a modest sample of 15 haplotypes is expected to trace back, on average, to a very ancient common ancestral haplotype, and that increasing the sample from 15 to, say, 30 would only push this estimate to $4(12,000)(1 - 1/30) = 46,400$ generations, a change of under 4%, concretely demonstrating the diminishing-returns property proved above.

16.5: Worked Example 16.1 (Coalescent time estimate, Nigerian population)

Suppose an autosomal locus's historical effective population size for a Nigerian ethnic group is estimated (from independent demographic/genetic data) at $N \approx 10,000$. For a sample of $n = 5$ gene copies (from 5 unrelated individuals, i.e., 5 of their 10 total gene copies at this locus, chosen for simplicity as unlinked/independent draws), find the expected time (in generations) to the MRCA of the sample, and convert to approximate years (assuming 25 years per generation).

Solution:

$$\begin{aligned} E(T_{MRCA}) &= 4N(1 - 1/5) = 40,000 \times 0.8 \\ &= 32,000 \text{ generations} \end{aligned}$$

In years: $32,000 \times 25 = 800,000$ years — illustrating the coalescent's characteristic result that, at neutral autosomal loci, gene genealogies typically trace back hundreds of thousands of years, far predating the emergence of modern population/ethnic group structure, and consistent with the deep, ancient genetic diversity long documented within African populations relative to non-African populations (whose out-of-Africa founding bottleneck, some 50,000–70,000 years ago, sharply reduced N and hence genealogical depth outside Africa).

16.6: Incorporating Mutation

Mutations are superimposed on the coalescent genealogy as a Poisson process along each branch, with rate μ per gene copy per generation. The total number of mutations expected on the genealogy of a sample of size n , summed over the whole tree, has expectation:

$$E(S) = \theta \times \sum_{k=1}^{n-1} 1/k, \theta = 4N\mu$$

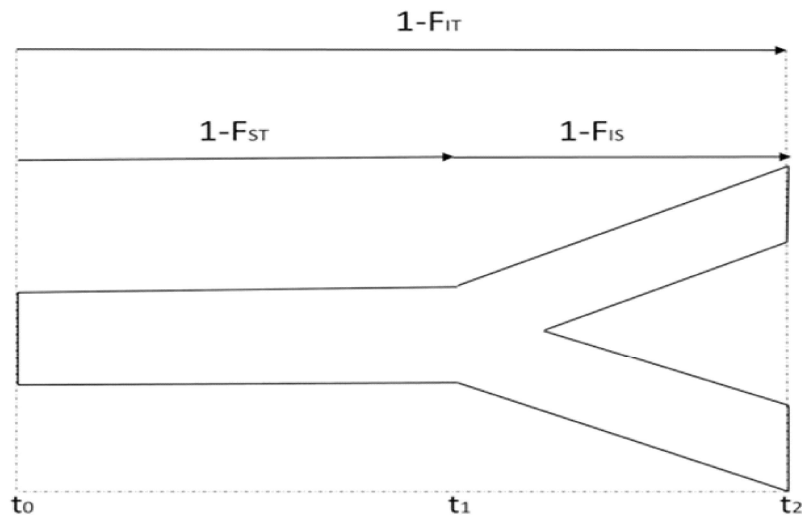
where θ (not to be confused with the recombination fraction θ of Chapter 7 — the same symbol is used by long convention for both quantities in population genetics, and context always disambiguates) is the **population-scaled mutation rate**, and S is the observed number of segregating (polymorphic) sites in a sample of DNA sequences. This relationship (due to Watterson, 1975) provides one of the standard estimators of θ (and hence, given an independently known mutation rate μ , of effective population size N) directly from observed DNA sequence variation — an approach widely used in estimating the effective population sizes and demographic histories of African, including Nigerian, populations from whole-genome sequence data.

16.7: The Allele Frequency Spectrum

Under the coalescent with mutation, the expected distribution of allele frequencies at neutral, segregating sites in a sample — the **site frequency spectrum (SFS)** — follows a characteristic shape: rare variants (those carried by only one or a few copies in the sample) are expected to be far more numerous than common (intermediate-frequency) variants, because most branches in a typical coalescent genealogy are short, terminal (external) branches on which only a few descendant lineages carry any resulting mutation. Departures from the neutral coalescent-

predicted SFS — for example, an excess of rare variants relative to neutral expectation — are used as one signature of recent population growth or of purifying selection, both of considerable interest in interpreting the substantial rare-variant burden increasingly documented in Nigerian and other African genomic sequencing studies (e.g., under initiatives such as H3Africa).

Variable population size and the bottleneck effect. All of the coalescent results above (Sections 16.3–16.7) assumed a *constant* effective population size, N , through time. Real human populations, however, have experienced dramatic size changes — most notably the “out-of-Africa” bottleneck, a sharp reduction in effective population size experienced by the ancestors of non-African populations roughly 50,000–70,000 years ago as a small founding group migrated out of the ancestral, larger African population:



A population bottleneck: effective population size drops sharply from an ancestral value to a much smaller founding size before recovering, illustrated as a coalescent genealogy whose branches are compressed during the bottleneck interval (t_0 to t_1) and expand again afterward.

A bottleneck compresses coalescence times during the period of small population size (Section 16.4's $E(T) = 2N$ formula shows that smaller N mechanically produces faster coalescence), leaving a lasting signature in the genealogy: a burst of coalescent events concentrated in a narrow time window, followed by a period of much slower coalescence both before (in the ancestral population) and after (once the population re-expands) the bottleneck. This is precisely why genetic diversity is systematically lower in non-African populations than in African populations (Chapter 14, Section 14.2): the founding bottleneck removed much of the genetic variation carried by the larger ancestral African population, a loss that many thousands of subsequent generations have not fully restored. Modern coalescent-based demographic inference methods (Chapter 20) work by fitting exactly this kind of piecewise-constant (or continuously varying) population-size history to observed genome-wide genetic variation data.

Theorem 16.2: (Expected Time to the Most Recent Common Ancestor)

Under the standard (Wright-Fisher) coalescent for a sample of n lineages, the expected total time (measured in units of N generations, per Section 16.3) to the most recent common ancestor of the sample is $E[T_{MRCA}] = 2(1 - 1/n)$.

Proof. Under the coalescent, the waiting time T_k during which exactly k ancestral lineages are present (before the next coalescence event reduces k to $k-1$) is exponentially distributed with rate $C(k,2) = k(k-1)/2$, so $E[T_k] = 2/[k(k-1)] = 2[1/(k-1) - 1/k]$ (Section 16.4). The total time to the most recent common ancestor is the sum of these waiting times as the number of lineages decreases from n down to 1: $T_{MRCA} = T_n + T_{n-1} + \dots + T_2$. Taking expectations

and using linearity, $E[T_{\text{MRCA}}] = \sum_{k=2}^n E[T_k] = \sum_{k=2}^n 2[1/(k-1) - 1/k]$. This is a telescoping sum: writing out the terms, $2[(1/1 - 1/2) + (1/2 - 1/3) + \dots + (1/(n-1) - 1/n)]$, all intermediate terms cancel, leaving $E[T_{\text{MRCA}}] = 2[1 - 1/n]$. As n grows large, $E[T_{\text{MRCA}}]$ approaches 2, showing that most of the expected time to the common ancestor is contributed by the earliest coalescence events (when only 2 or 3 lineages remain), not by events near the tips of the genealogy where many lineages are present — a fact with direct implications for how sample size affects genealogical inference.

16.8 Chapter Summary

- The coalescent models gene genealogies backward in time; a pair of lineages coalesces with probability $\approx 1/(2N)$ per generation, giving expected pairwise coalescence time $2N$ generations.
- For a sample of n lineages, expected time to the MRCA is $4N(1-1/n)$ generations, converging to $4N$ as n grows — most genealogical depth is captured by even modest samples.
- Mutation is superimposed as a Poisson process on the genealogy; the population-scaled mutation parameter $\theta=4N\mu$ links observed sequence diversity to effective population size.
- The site frequency spectrum, predicted by the neutral coalescent, provides a basis for detecting population growth or selection, both relevant to interpreting Nigerian genomic diversity data.

16.9: Computational Corner: Simulating Coalescent Genealogies

R.

```
# Simulate coalescence times for a sample of n lineages
(Section 16.4)
set.seed(2024)
simulate_coalescent <- function(n, Ne) {
  times <- numeric(n - 1)
  k <- n
  total_time <- 0
  for (i in 1:(n - 1)) {
    rate <- choose(k, 2) / Ne      # merge rate ~ Exp(C(k,2)/
    Ne) in coalescent units
    t_k <- rexp(1, rate = rate)
    total_time <- total_time + t_k
    times[i] <- total_time
    k <- k - 1
  }
  times
}
Ne <- 10000
coal_times <- replicate(500, simulate_coalescent(n = 10, Ne =
Ne))
mean_tmrca <- mean(coal_times[9, ])    # last (9th) event =
TMRCA for n=10
cat("Simulated mean TMRCA (generations):", mean_tmrca,
"\n")
cat("Theoretical E[TMRCA] = 4Ne(1-1/n):", 4*Ne*(1-1/10),
"\n")
```

Python.

```
import numpy as np

rng = np.random.default_rng(2024)
```

```

def simulate_coalescent(n, Ne):
    k = n
    total_time = 0.0
    for _ in range(n - 1):
        rate = (k * (k - 1) / 2) / Ne
        total_time += rng.exponential(1 / rate)
        k -= 1
    return total_time

Ne = 10000
tmrcas = [simulate_coalescent(10, Ne) for _ in range(500)]
print("Simulated mean TMRCA:", np.mean(tmrcas))
print("Theoretical E[TMRCA] = 4Ne(1-1/n):", 4*Ne*(1-1/10))

```

The Monte Carlo average TMRCA from 500 simulated genealogies should lie close to the analytic expectation $4N_e(1 - 1/n)$ derived in Section 16.4, providing a direct simulation-based check of the coalescent theory and previewing the more elaborate sequence-level coalescent simulations of Chapter 20.

Mini-Project

Mini-Project 16: Simulating Coalescent Genealogies and the Site Frequency Spectrum

Simulate gene genealogies under the standard coalescent and examine how sample size affects the time to the most recent common ancestor.

Step 1: Simulate coalescent genealogies for samples of size $n = 5, 10, 20$, and 50 lineages (Section 16.9).

Step 2: For each sample size, record the total time to the most recent common ancestor (TMRCA) across at least 500 simulated genealogies.

- Step 3:** Plot the distribution of TMRCA for each sample size and compare the means to the theoretical expectation.
- Step 4:** Superimpose mutations on your simulated genealogies (Section 16.6) and tabulate the resulting site frequency spectrum.
- Step 5:** Write a short paragraph on how the site frequency spectrum's shape might change under a recent population expansion.

Deliverable: *Simulation code, a TMRCA comparison plot across sample sizes, and a site frequency spectrum plot with written interpretation.*

16.10: Practice Questions

- Q1.** For an effective population size $N = 5,000$, find the expected pairwise coalescence time (in generations and in years, using 25 years/generation).
- Q2.** For a sample of $n = 10$ gene copies from a population with $N = 8,000$, find the expected time to the MRCA of the sample (in generations).
- Q3.** Explain, in words, why the expected time to the MRCA converges to $4N$ generations as sample size n grows large, rather than continuing to grow indefinitely with n .
- Q4.** If $\theta = 4N\mu = 0.008$ for a Nigerian population sample at a given locus, and the sample size is $n = 20$, use Watterson's formula (Section 16.6) to find the expected number of segregating sites, $E(S)$. (You may leave the harmonic sum as a decimal approximation; $\sum_{k=1}^{19} 1/k \approx 3.548$.)
- Q6.** Explain, in your own words, why the “out-of-Africa” population bottleneck illustrated in Section 16.7 left non-African populations with systematically lower

genetic diversity than African populations, using the relationship between effective population size and expected coalescence time (Section 16.4).