Coalescent theory & Gene Genealogies

- The rapid accumulation of DNA sequence data since the 1980s has transformed the mainstream of population genetics research from prospective to retrospective, from demonstration of principles to inference of events that happened in the past.
- Coalescent theory arose from the necessity to infer the past from a sample taken from a present population.
- The essence of coalescent theory is to start with a sample, and trace it backward in time to identify evolutionary events that occurred in the past since the Most Recent Common Ancestor (MRCA) of the sample.
- Coalescent theory helps to understand the evolutionary causes that have influenced the DNA sequence variation in a sample of individuals, such as the demographic and mutational history of the ancestors of the sample.
- Coalescent theory represents the most significant progress in theoretical population genetics in the past two decades of this century.
- It is now widely recognized as a cornerstone for rigorous statistical analyses of molecular data from populations.

Coalescent theory & Gene Genealogies

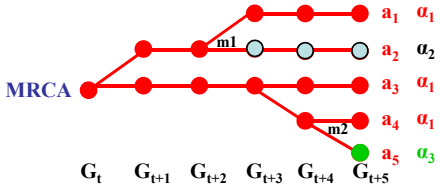
- Any two alleles from a sample of alleles may have a different history but descend from the same ancestral allele and have a common ancestor (CA) in the past
- The point at which this common ancestry for two alleles occurs is called **coalescence**
- If one goes back far enough in time in the population, then all alleles in the sample will coalesce into a single common ancestral allele
- The coalescent approach was suggested by John Kingman (1982a, b).
- The usefulness of the coalescent theory comes mainly from three features:
  1) it is a sample-based theory
  2) it developed highly efficient algorithms for simulating population samples under various population genetics models, allowing various aspects of a model to be examined numerically
  3) it is particularly suitable for molecular data, such as DNA sequence samples, which contain rich information about the ancestral relationships among the individuals sampled

six degrees of separation

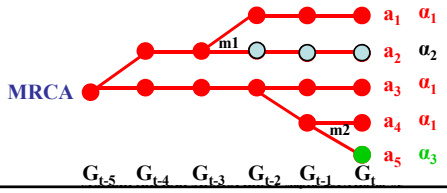
Coalescent theory & Gene Genealogies

- Consider a sample of ***n*** sequences of a DNA region from a population and assume that there is no recombination between sequences
- Coalescent theory suggests that these ***n*** sequences are connected by a single phylogenetic tree or genealogy, in which the root of the tree is the **most recent common ancestor (MRCA)** of these ***n*** sequences:

1) If one starts with the **MRCA** and looks **forward** in time, one sees that once in a while one of the existing sequences splits into two and along the way mutations accumulate. The impression from this prospective view is the **divergence** of sequences:

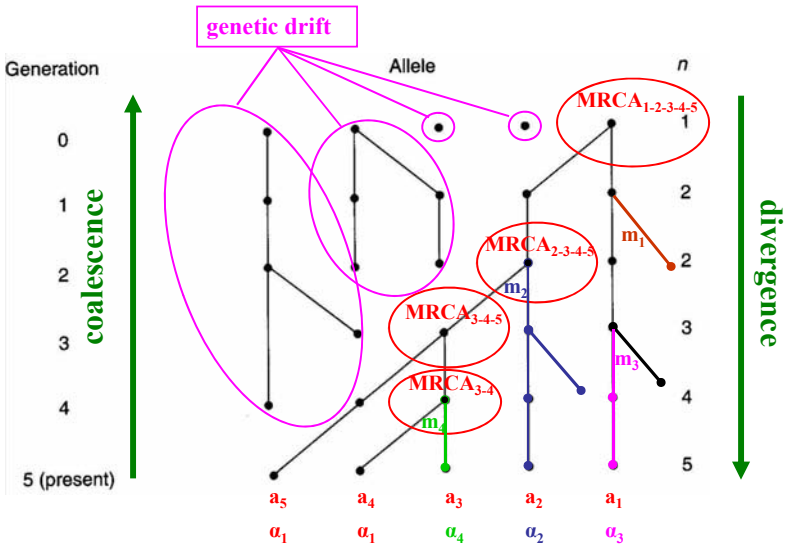


2) If one starts with the sample of sequences and looks backward in time, one sees that the number of ancestral sequences becomes fewer and fewer and the sequences become more and more similar. This retrospective view gives the impression of **coalescence:**



3

Coalescent theory & Gene Genealogies



4

Coalescent theory & Gene Genealogies

- The **most recent common ancestor (MRCA)** is determined via incorporation of DNA sequences into **gene trees** or **gene genealogies**:

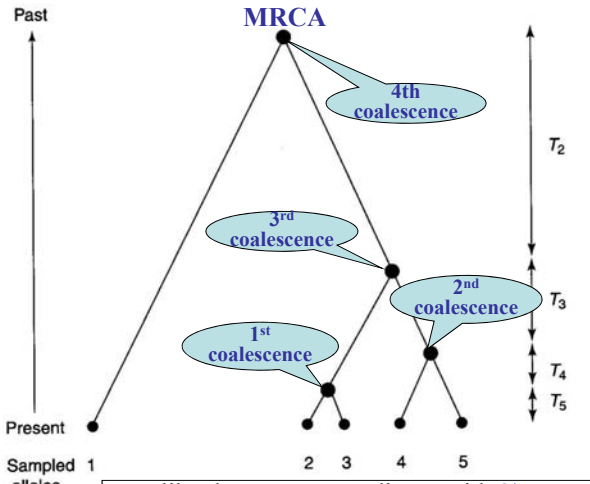


Figure 8.13. An example of a gene tree for five sampled alleles. The four large circles indicate coalescent events (after Hudson, 1990). T_i is the expected time in which there are i alleles, and these intervals are shown proportional to their expected time as given in expression 8.11*c*.

- Unlike the common pedigree with 2^t ancestors the number of common ancestors is decreasing during the coalescence because only one parent is transmitted a specific allele to a given offspring

Coalescent theory & Gene Genealogies

- The coalescent theory of two sequences from a population (of diploid organisms) is based on the neutral Wright-Fisher model, that assumes:
  - –nonoverlapping generations
  - –random mating
  - –constant population size
  - –random reproduction (resulting in a Poisson distribution of progeny)

The coalescent process can be described as follows:

- In the IAM the total probability that any 2 alleles descend from the **same** ancestral allele in the previous generation is $$(\frac{1}{2N})(\frac{1}{2N})2N = \frac{1}{2N}$$
- Therefore, the probability that **2 alleles** came from **different** ancestral alleles is $$1 - \frac{1}{2N}$$
- Furthermore, the probability that **3 alleles** came from **different** ancestral alleles is $$(1-\frac{1}{2N})(\frac{2N-2}{2N}) = (1-\frac{1}{2N})(1-\frac{2}{2N})$$
- Furthermore, the probability that **4 alleles** came from **different** ancestral alleles is $$(1-\frac{1}{2N})(1-\frac{2}{2N})(\frac{2N-3}{2N}) = (1-\frac{1}{2N})(1-\frac{2}{2N})(1-\frac{3}{2N})$$
- In general, the probability that ***n*** sampled alleles have ***n* different** ancestral alleles in the previous generation is $$\Pr(n) = \prod_{i=1}^{n-1}\left(1-\frac{i}{2N}\right)$$
- The probability that ***n*** sampled alleles have ***n* different** ancestral alleles in the previous ***t*** generations is $\mathbf{[Pr(}n\mathbf{)]}^t$

Coalescent theory & Gene Genealogies

- The probability that 2 alleles came from **different** ancestral alleles ***t*** generations before present is $$\Pr(2)^t = \left(1-\frac{1}{2N}\right)^t$$
- The probability that 2 alleles have the **same** ancestral allele (**that is, they coalesce**) ***t*+1** generations before present is $$\Pr(2)^t[1-\Pr(2)] = \left(1-\frac{1}{2N}\right)^t\frac{1}{2N}$$
- The probability that ***n*** sampled alleles have ***n*−1** ancestral allele (**that is, 2 out of n alleles coalesce**) ***t*+1** generations before present is $$\Pr(n)^t[1-\Pr(n)] = \left[\prod_{i=1}^{n-1}(1-\frac{i}{2N})\right]^t\frac{1}{2N}$$

Coalescent theory & Gene Genealogies

Remember, the overall probability that ***n*** sampled alleles in any generation have ***n*** different ancestral alleles in the previous generation is

$$\Pr(n) = \prod_{i=1}^{n-1}\left(1-\frac{i}{2N}\right) \quad \text{or} \quad (1-\frac{1}{2N})(1-\frac{2}{2N})...(1-\frac{n-1}{2N})$$

assuming that $\mathbf{1/N^2}$ is too small and can be ignored, then it can be approximated as

$$1-\frac{1}{2N}(1+2+...+n-1)$$

since the sum of the first $\boldsymbol{n-1}$ integers equals $\boldsymbol{n(n-1)/2}$, then it can be written as

$$1-\frac{n(n-1)}{4N}$$

This is the probability of **absence** of a coalescence, therefore, the probability of **presence** of a coalescence is $\Pr(C) = 1-[1-\frac{n(n-1)}{4N}] = \frac{n(n-1)}{4N}$

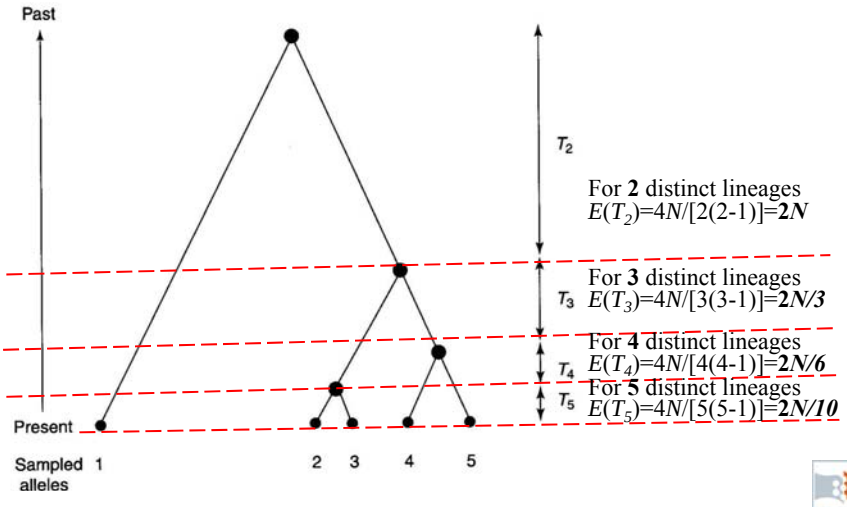
For ***n*** alleles the probability of no coalescence for the first $\boldsymbol{t-1}$ generations followed by coalescence in the ***t*th** generation is $\mathbf{(1-Pr(\mathit{C}))^{t-1}Pr(\mathit{C})}$

This is a geometric distribution of coalescence times, and the mean of this distribution is $\overline{T_n} = \sum_{t=1}^{\infty} t(1-\Pr(C))^{t-1}\Pr(C) = \frac{1}{\Pr(C)} = \frac{4N}{n(n-1)}$

Coalescent theory & Gene Genealogies

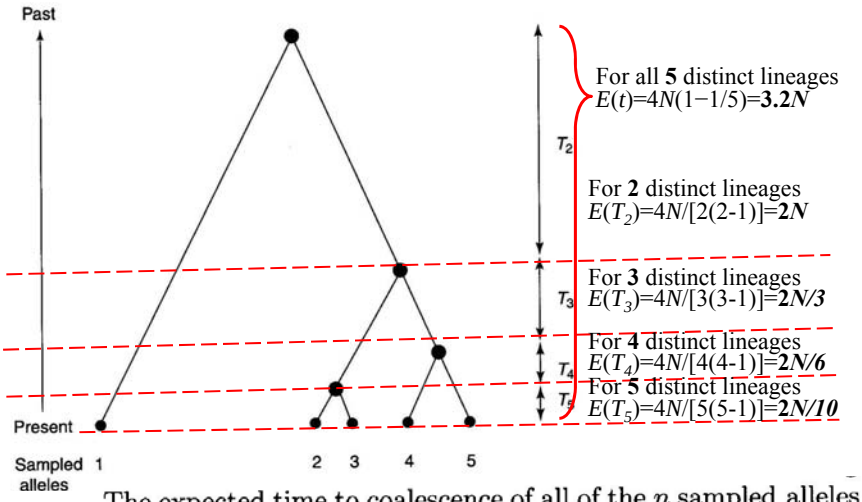
- Now you know that the expected time during which there are ***n*** distinct lineages is

$$E(T_n) = \frac{4N}{n(n-1)}$$



Coalescent theory & Gene Genealogies

the expected time increases as the number of lineages decreases.

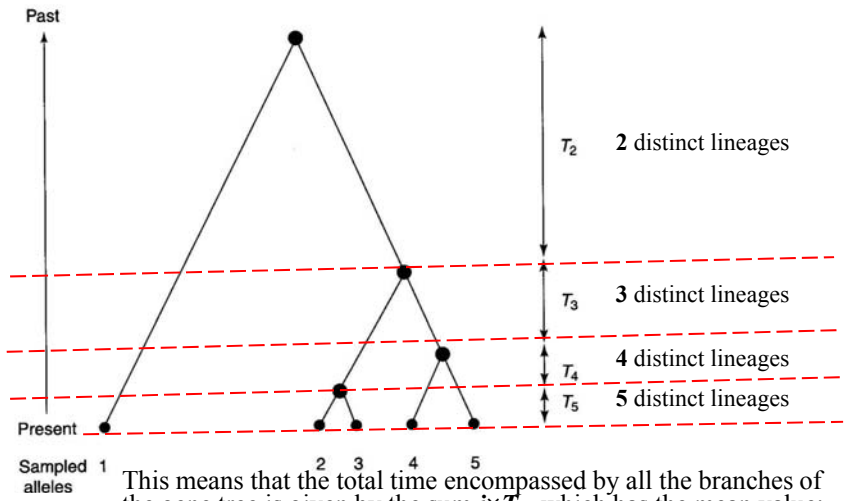


The expected time to coalescence of all of the n sampled alleles is

$$E(t) = \sum_{i=2}^{n} E(T_i) = 4N\left(1 - \frac{1}{n}\right)$$

Coalescent theory & Gene Genealogies

Note in the figure that the interval marked by each $\boldsymbol{T_i}$ has $\boldsymbol{i}$ coexisting lineages:



This means that the total time encompassed by all the branches of the gene tree is given by the sum $\boldsymbol{i}\times\boldsymbol{T_i}$, which has the mean value:

$$\sum_{i=2}^{n} i\overline{T_i} = \sum_{i=2}^{n} i\frac{4N}{n(n-1)} = 4N\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}\right)$$

Coalescent theory & Gene Genealogies

$$\sum_{i=2}^{n} i\overline{T}_i = \sum_{i=2}^{n} i\frac{4N}{n(n-1)} = 4N\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}\right)$$

In the IAM each new mutation in the branches of gene tree results in a distinctive allele in the sample. This is a reasonable assumption for DNA sequences.

If the mutations occur uniformly in time at rate ***u*** per nucleotide site per generation, then the expected proportion of segregating sites in the sample, ***E(S)***, must equal the mutation rate per nucleotide times the total length of all the branches of the gene tree:

$$E(S) = u\sum_{i=2}^{n} i\overline{T}_i = 4Nu\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}\right) = \theta\alpha_1$$

where $\theta = 4Nu$ and $\alpha_1 = \left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}\right) = \sum_{i=1}^{n}\frac{1}{i}$

This implies that θ can be estimated as $\theta = S/\alpha_1$, which is often called the **expected nucleotide polymorphism**

Coalescent theory & Gene Genealogies

Let's illustrate the effects of both genetic drift and mutation:
Consider 10-generations & $2N$=10

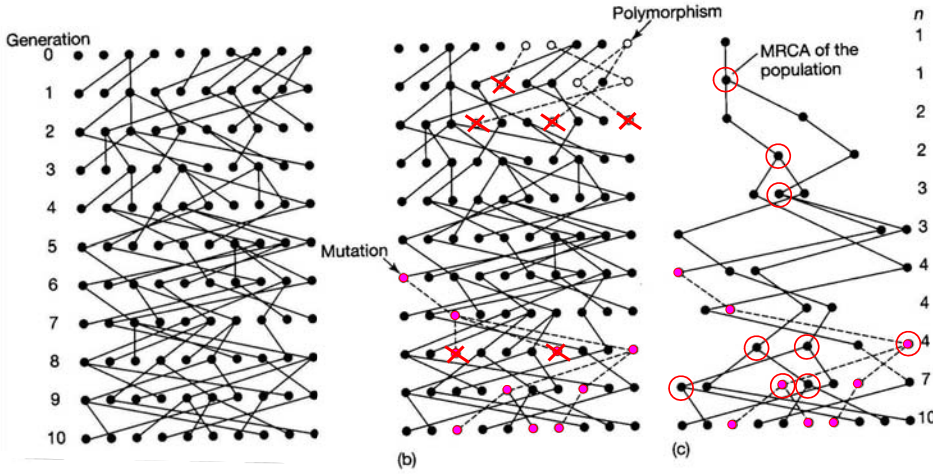
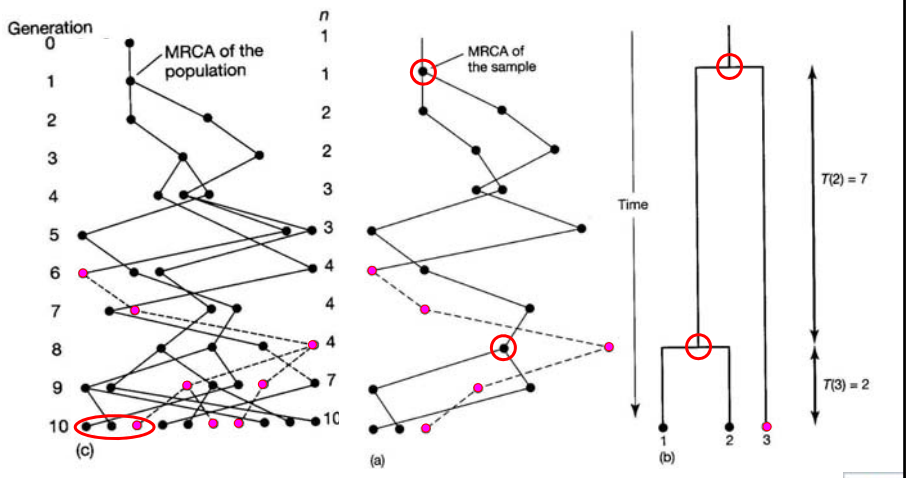


Figure 8.16. An illustration of the effects of both genetic drift and mutation in 10-generation example where $2N = 10$ (Nordborg, 2001) where it (a) gives the complete genealogy illustrating the effect of genetic drift, (b) includes a mutation in generation 6 and initial variation in generation 0, and (c) gives the genealogy showing only the ancestors of the 10 alleles in generation 10; n is the number of ancestral alleles from which the 10 alleles in the present generation are descended.

Coalescent theory & Gene Genealogies

Notice from the previous example that 7 of the 10 alleles in the present generation are unchanged from the MRCA (generation 1) , whereas 3 descend from the mutation. Let's look at the genealogy of the 3 leftmost alleles from this population:



Coalescent theory & Gene Genealogies

- The expected distribution of coalescent times is greatly affected by changes in population size:

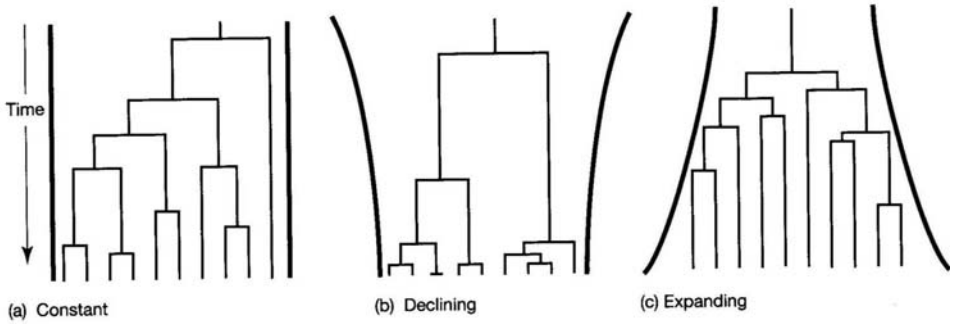


Figure 8.18. The theoretical distributions of coalescent times in genealogies with 10 contemporary samples under three scenarios of historical population change: (a) constant population size, (b) declining population size, and (c) increasing population size (Garrigan *et al.*, 2002). The distance between the thick lines indicates the relative population sizes at different times, and the genealogies reflect coalescent events in periods of relative small or large population sizes, respectively.

Coalescent theory & Gene Genealogies

b. Estimating Effective Population Size

Neutral theory and the coalescent approach can potentially be used to estimate the value of evolutionary parameters, such as effective population size and mutation rate. In the simplest form, at neutrality equilibrium, the level of diversity is a balance between mutation and genetic drift and is

$$\theta = 4N_e u$$

This expression can be solved for an estimate of the effective population size as

$$N_e = \frac{\theta}{4u}$$

For diversity data for mitochondrial sequences, then

$$\theta = 2N_{ef} u$$

and

$$N_{ef} = \frac{\theta}{2u}$$

where N_{ef} is the female effective population size (see p. 327).

17

Coalescent theory & Gene Genealogies

In both of these estimates, u is the mutation rate per generation, but the mutation rate is estimated from the rate of substitution per year. To put this estimate on a per generation scale, the denominator in both of the above expressions need to be multiplied by T, which is defined as the generation length in years, so that

$$N_e = \frac{\theta}{4uT} \qquad (8.12a)$$

$$N_{ef} = \frac{\theta}{2uT} \qquad (8.12b)$$

For example, in 50 random, noncoding, nuclear DNA segments in humans, $\theta = 0.000882$ (Yu *et al.*, 2002). The divergence between humans and chimpanzees for these 50 segments is 0.01221 (Yu *et al.*, 2003) and, assuming that humans and chimpanzees diverged 6×10^6 years ago, the estimate of mutation rate per year is $u = 0.01221/[(2)(\times 10^6)] = 1.02 \times 10^{-9}$ (the 2 in the denominator is included because divergence is occurring in both lineages). Assuming that the generation length in humans is 20 years, then using expression 8.12*a*, $N_e = 0.000882/[(4)(1.02 \times 10^{-9})(20)] = 10,800$ (Yu

(d)

$u = d/2T$

18

Coalescent theory & Gene Genealogies

- There are many coalescence methods have been developed recently to include all of the different evolutionary factors:
  - variation in population size
  - gene flow
  - inbreeding
  - recombination (that spreads the ancestry of a mutation over different chromosome)
  - balancing selection
  - selective sweeps
  - purifying selection
- The advanced methods are theoretically difficult and very computationally intense

Coalescent theory & Gene Genealogies

Conclusions

- The coalescence methods allow to estimate:
  - time to the MRCA (important for analysis of speciation, divergence)
  - evolutionary or long-term effective population size N_e (has very important conservation genetics applications!)
  - age of alleles (important for epidemiology!)
  - whether the population is constant, growing, or declining in size