## Coalescent Theory

- the Wright-Fisher model considers changes in the ideal population as time moves forward
- coalescent theory (~1980+) looks backwards in time
- how long does it take for $k$ alleles to coalesce to $k - 1$ alleles, then $k - 2$, $k - 3$, …, and finally a single ancestral allele?

## Stochastic elements of the coalescent

- alleles randomly sample their parents in the previous generation
  - results in binomial variation in offspring number
- sample of loci from the genome
  - different loci have different genealogical histories
- sample of alleles from the population
  - different samples of the same locus may have different coalescent trees
- distribution of mutations on the genealogy
  - mutations allow estimates of coalescence times

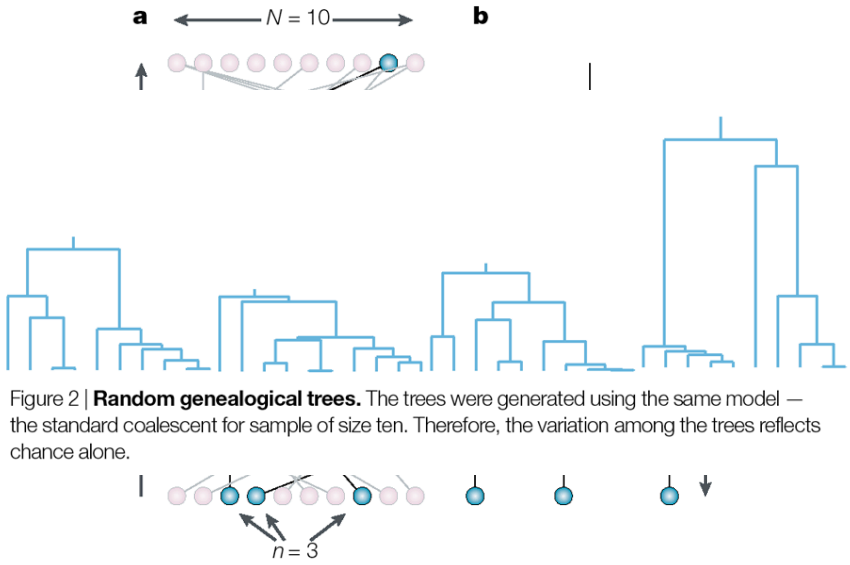



Figure 2 | **Random genealogical trees.** The trees were generated using the same model — the standard coalescent for sample of size ten. Therefore, the variation among the trees reflects chance alone.

Rosenberg & Nordborg 2002 Nat Rev Gen

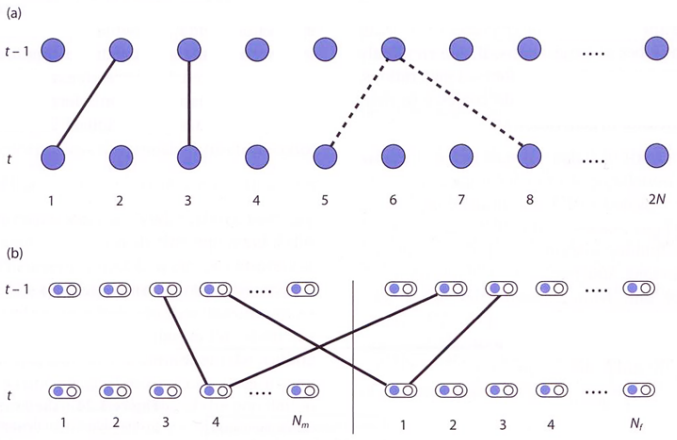



**Figure 3.23** Haploid (a) and diploid (b) reproduction in the context of coalescent events. In a haploid population, the probability of coalescence is $\frac{1}{2N}$ (dashed lines) whereas the probability that two lineages do not have a common ancestor in the previous generation is $1-\frac{1}{2N}$ (solid lines). In a diploid population, the two gene or allele copies in one individual in the present time have one ancestor in the female population ($N_f$) and one ancestor in the male population ($N_m$). Coalescent events in the diploid population arise when the gene copies in males and females are identical by descent. The haploid model with $2N$ lineages is routinely used to approximate the diploid model with $N = N_f + N_m$ diploid individuals.

## Coalescent probabilities 1

- the present is time 0 (zero)
- probability that *two* alleles had a common ancestor in generation 1

$$= \frac{1}{2N}$$

- probability that *two* alleles **did not have** a common ancestor in generation 1

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

## Coalescent probabilities 2

- probability that *two* alleles have still not coalesced by generation $t$

$$= \left[1 - \left(\frac{1}{2N}\right)\right]^t$$

# Coalescent probabilities 3

- probability that *two* alleles had a common ancestor in generation $t+1$

$$= \frac{1}{2N}\left[1-\left(\frac{1}{2N}\right)\right]^{t}$$

probability of coalescence in generation $t$ + 1

probability of *NO* coalescence in generations 1 through $t$

Error on pg. 93

the exact probability of coalescence approximately equal to the exponential gives an expression

$$\left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N} \approx \frac{1}{2N}e^{-\frac{t}{2N}} \qquad (3.72)$$

that can be simplified by canceling the $\frac{1}{2N}$ constant term on both sides

$$\left(1-\frac{1}{2N}\right)^{t-1} \approx e^{-\frac{t}{2N}} \qquad (3.73)$$

Therefore, the exponential distribution approximates the probability of non-coalescence at each time *t*.

# Pg. 90 – errors in Figure 3.24

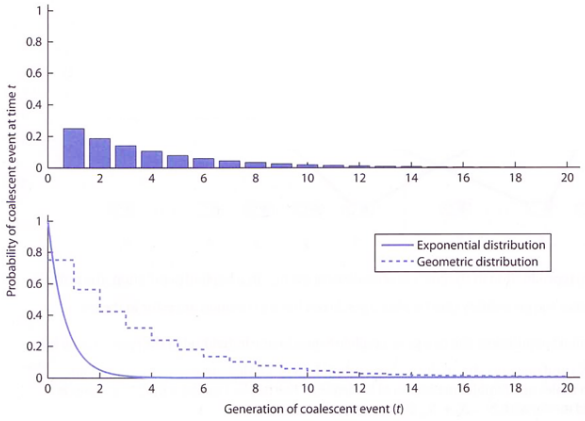
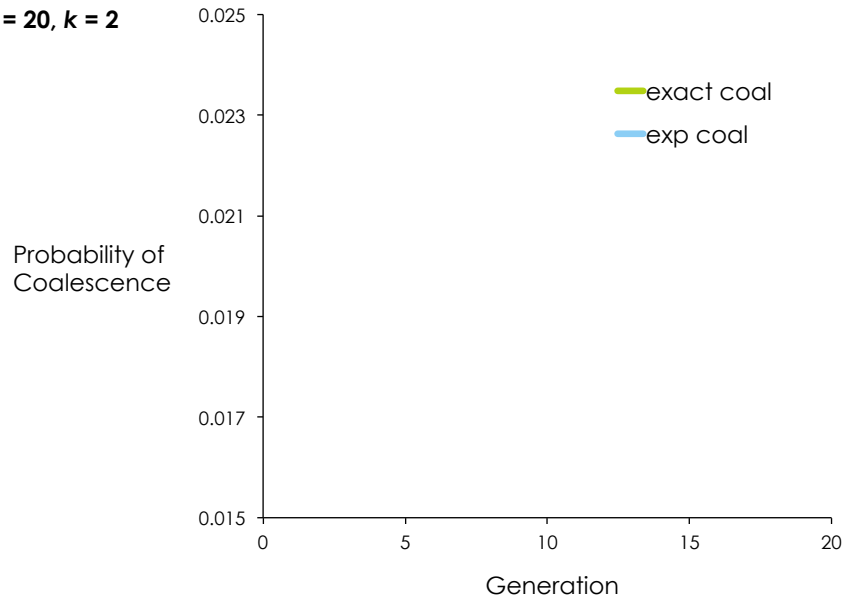



**Figure 3.24** The distribution of the probabilities of coalescence over time predicted by $\left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N}$ (top) for a population of $2N = 4$. The bottom panel shows the approximations of these coalescence probabilities based on geometric and exponential distributions with a probability of "success" of $1/4$.

***N* = 20, *k* = 2**



## Coalescent probabilities 4

- can we randomly choose a coalescence time from the exponential distribution?
  - need to solve for $t$ as a function of a random variable from 0 to 1

$$P_{NC} \approx e^{-t/(2N)}$$

$$\ln\left(P_{NC}\right) \approx -t/(2N)$$

$$\ln\left(P_{NC}\right) \times 2N \approx -t$$

$$t \approx -\ln\left(P_{NC}\right) \times 2N$$

## Coalescent probabilities 5

- what if we consider $k$ alleles and not just 2?
- probability that $k$ alleles had $k$ distinct parental alleles the previous generation

$$\Pr(k) = \prod_{i=0}^{k-1}\left(1 - \frac{i}{2N}\right) \approx \left(1 - \frac{\binom{k}{2}}{2N}\right)$$

## Coalescent probabilities 6

- probability that $k$ alleles *do not* coalesce for $t$ generations

$$P_{NC} = \left(1 - \frac{\binom{k}{2}}{2N}\right)^t \approx e^{\left[-\frac{\binom{k}{2}}{2N}t\right]}$$

$$t \approx -\ln\left(P_{NC}\right) \times \frac{2N}{\binom{k}{2}} = -\ln\left(P_{NC}\right) \times \frac{4N}{k\left(k-1\right)}$$

## Coalescent probabilities 7

- probability that $k$ alleles *do not* coalesce for $t$ generations, and then *one pair* coalesces to give $k$ - 1 alleles at $t$ + 1 generations

$$= \Pr\left(k\right)^t \left[1 - \Pr\left(k\right)\right] \approx \frac{\binom{k}{2}}{2N} e^{\left[-\frac{\binom{k}{2}}{2N}t\right]}$$

- distribution has mean and variance:

$$\text{Mean} = \frac{4N}{k(k-1)} \text{generations} \quad \text{Var} = \frac{16N^2}{\left[k\left(k-1\right)\right]^2} \text{generations}^2$$

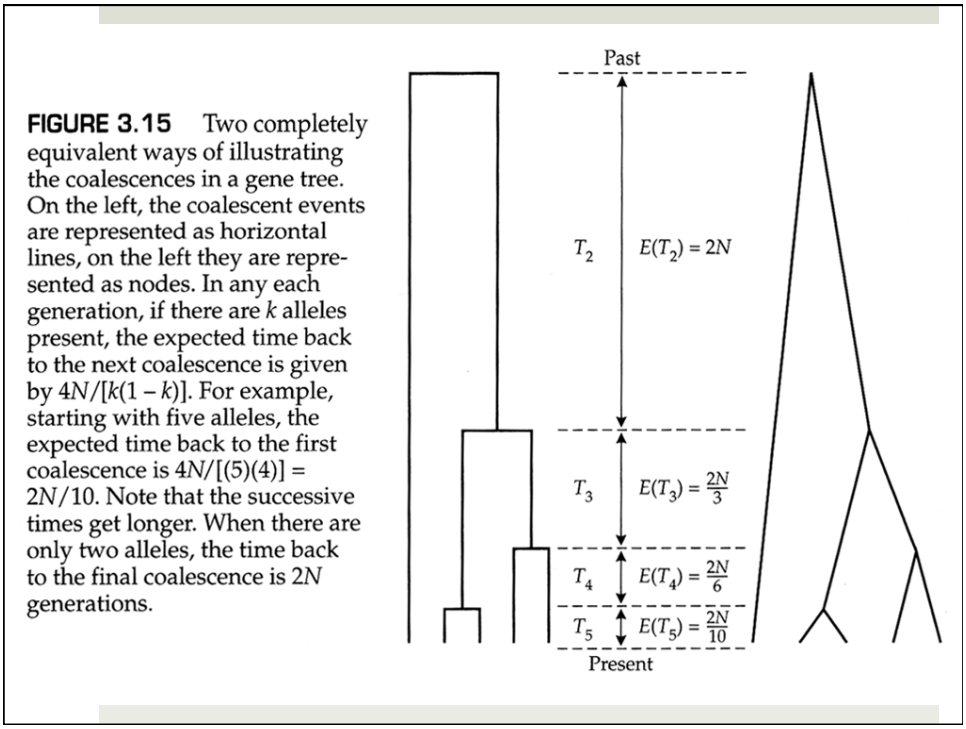
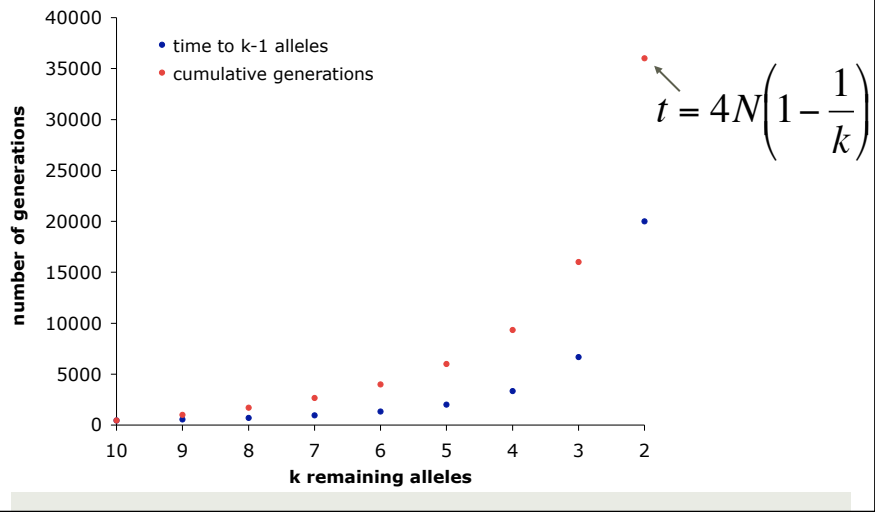



**FIGURE 3.15** Two completely equivalent ways of illustrating the coalescences in a gene tree. On the left, the coalescent events are represented as horizontal lines, on the left they are represented as nodes. In any each generation, if there are $k$ alleles present, the expected time back to the next coalescence is given by $4N/[k(1-k)]$. For example, starting with five alleles, the expected time back to the first coalescence is $4N/[(5)(4)] = 2N/10$. Note that the successive times get longer. When there are only two alleles, the time back to the final coalescence is $2N$ generations.

## expected coalescence times for $k$ = 10 and $N$ = 10,000

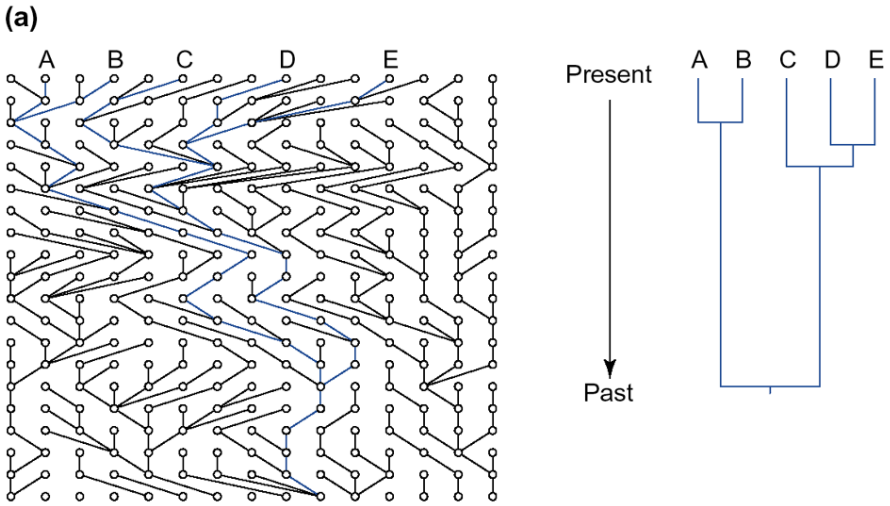



$$t = 4N\left(1 - \frac{1}{k}\right)$$

## The coalescent with population growth

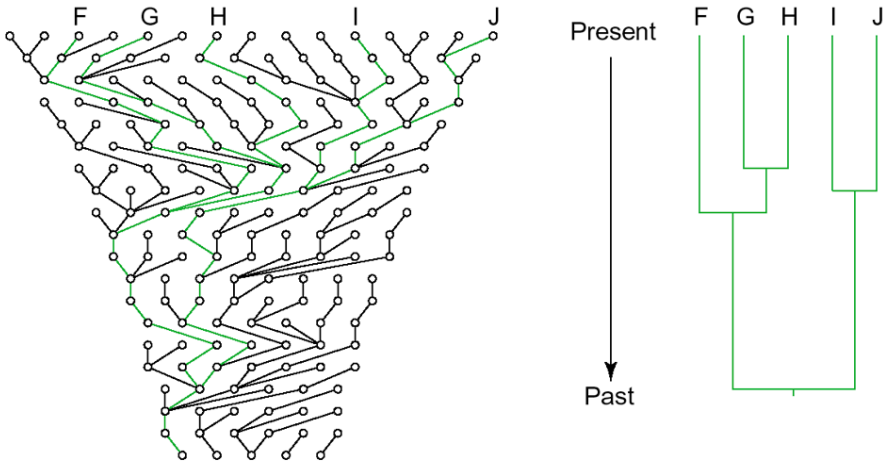
- coalescent trees are expected to be sparse (few lineages) near the root for populations of constant size
- in a growing or shrinking population, the distribution of coalescence times differs from expectations for the ideal population
- expanding populations have more nodes closer to the root of the tree
  - takes longer for alleles to "find each other" in a growing population

## constant



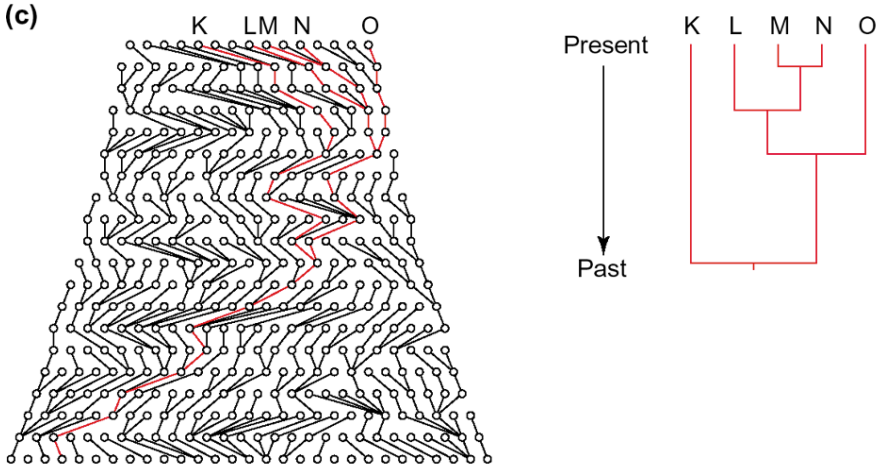
## growing

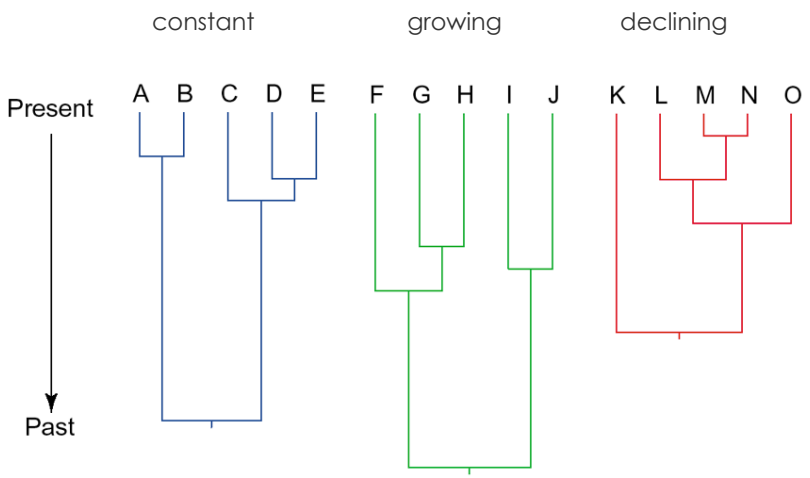
**(b)**




## declining

**(c)**






## So, what's the point?

- coalescent modeling
  - simulate genealogies under a given set of population parameters
  - "hang" random mutations on those trees in equal number (or at the same rate) as in the observed data
  - evaluate whether the observed data could have been produced by a random coalescent process (the null hypothesis)

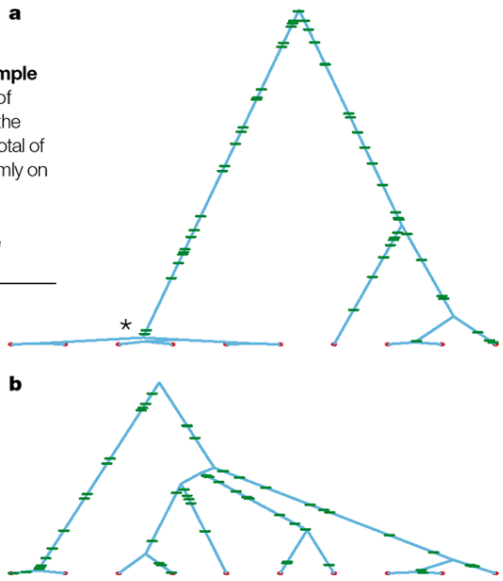



Figure 5 | **Simulated random genealogies with a sample size of ten.** The figure is based on data from the study of Hudson *et al.*[32]. Both genealogies were simulated from the standard coalescent model (without recombination). A total of 55 mutations (green horizontal lines) were placed randomly on each tree. **a** | One set of at least five sequences with no polymorphism; (marked by an asterisk) **b** | All sets of sequences contained polymorphisms. Red dots denote individual haplotypes.

- *Sod* locus in *Drosophila*
- 5 of 10 alleles sampled from Barcelona population were identical, yet the total sample had 55 segregating sites
- only 1.1% ($p = 0.011$) of simulated genealogies had five or more identical sequences
- so, neutral model rejected
- $p < 0.0001$ with recombination - why?

Hudson et al. 1994 Genetics

Rosenberg & Nordborg 2002 Nat Rev Gen