

# Coalescent Models With Recombination

Biostatistics 666

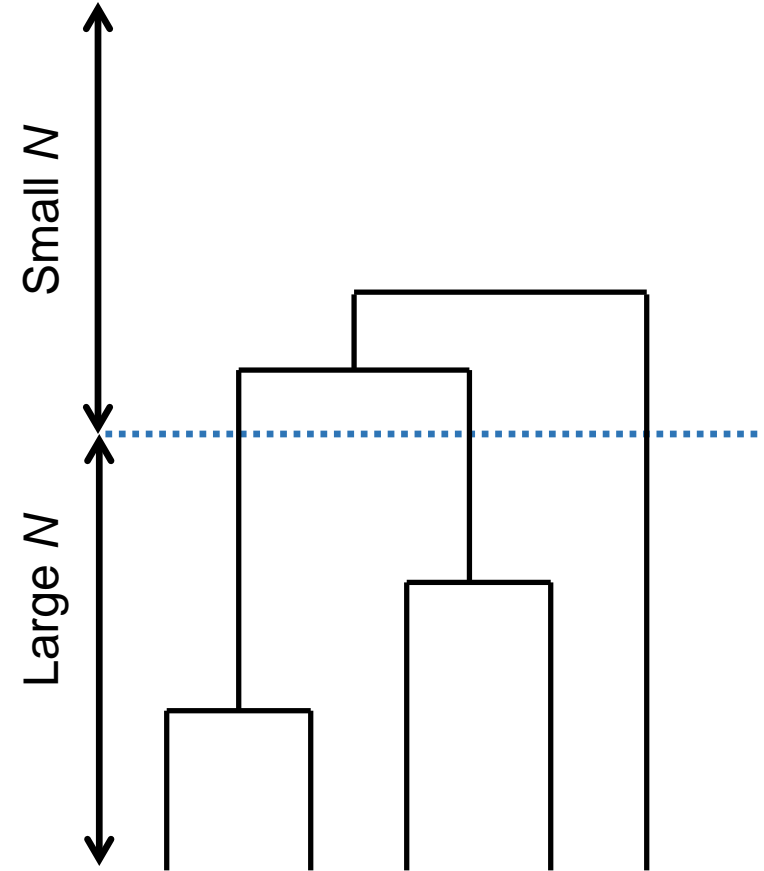
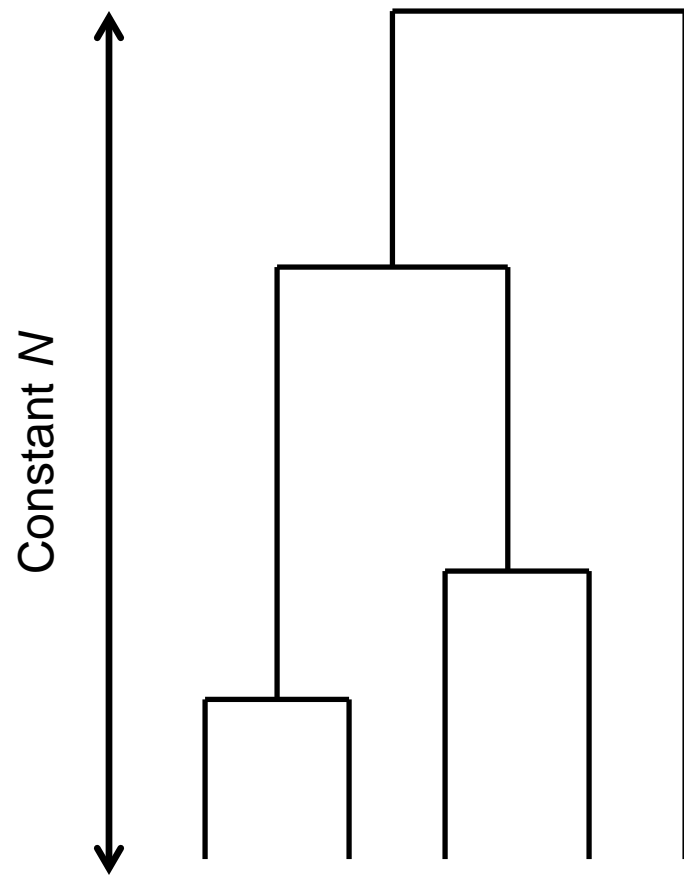
# So far ...

- Basic Properties of the Coalescent
  - MRCA
  - Coalescence times
  - Number of mutations
- Frequency spectrum of polymorphisms
- Predicting number of variants in a sample

# The Parameter $\Theta = 4N\mu$

- Occurs frequently in population genetic models
- Two estimators for  $\theta$ :
  - One, based on number of segregating sites  $S$
  - Another, using average of pairwise sequence differences
- The two estimators have the same expectation
  - For constant size, equilibrium populations
- Can give consistently different answers when...
  - There are deviations from neutral model, such as ...
  - ... population growth, bottlenecks, natural selection
- Comparison of the two estimates defines Tajima's  $D$

How do changes in  $N$   
change the coalescent tree?



Large  $N$

# Today ...

- Using the coalescent to learn about a population
- Further refining the coalescent
  - Recombination
  - Migration
- Discussion of potential applications

# Making Inferences About Population Parameters ...

- The coalescent is a very useful tool for simulating data
- Given a set of parameters (population size, mutation rates, etc.) it can generate a plausible sample of sequences
- But how do we learn about parameters, given a sample?

# Making Inferences About Population Parameters ...

- Select a set of summary statistics
  - For example, the number of observed variants  $S$
- For each of several parameters settings, carry out coalescent simulations
- Record how often each parameter setting results in a simulated dataset that matches the original.

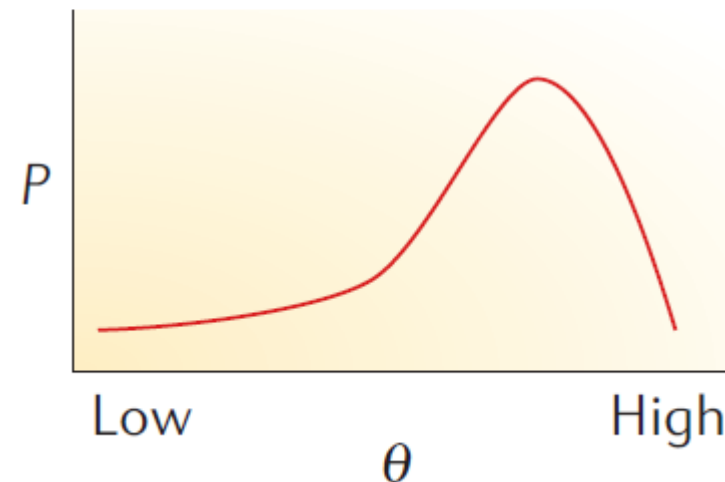
# Making Inferences About Population Parameters ...

Sequence Data

```
ACGATCGAT . . . . . ATAT
ACGATCGAA . . . . . ATAA
. . . . .
ACGATCGAT . . . . . ATAT
```

+  
Model

Probability of data given the mutation rate (likelihood function)

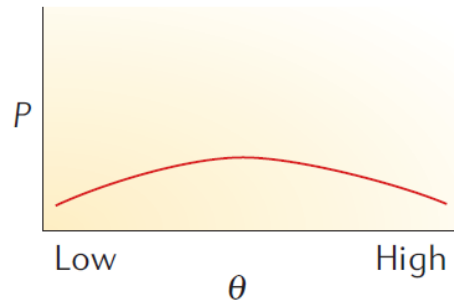


Marjoram and Tavaré (2006)

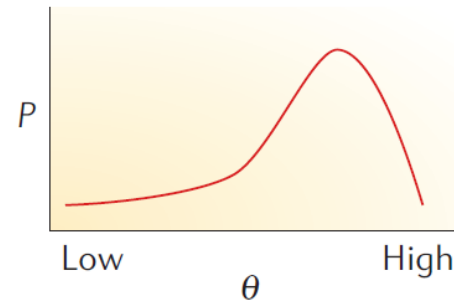


# Making Inferences About Population Parameters ...

Prior distribution of the mutation rate ( $\pi$ )



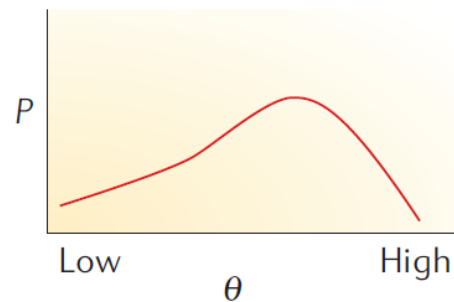
Probability of data given the mutation rate (likelihood function)



+

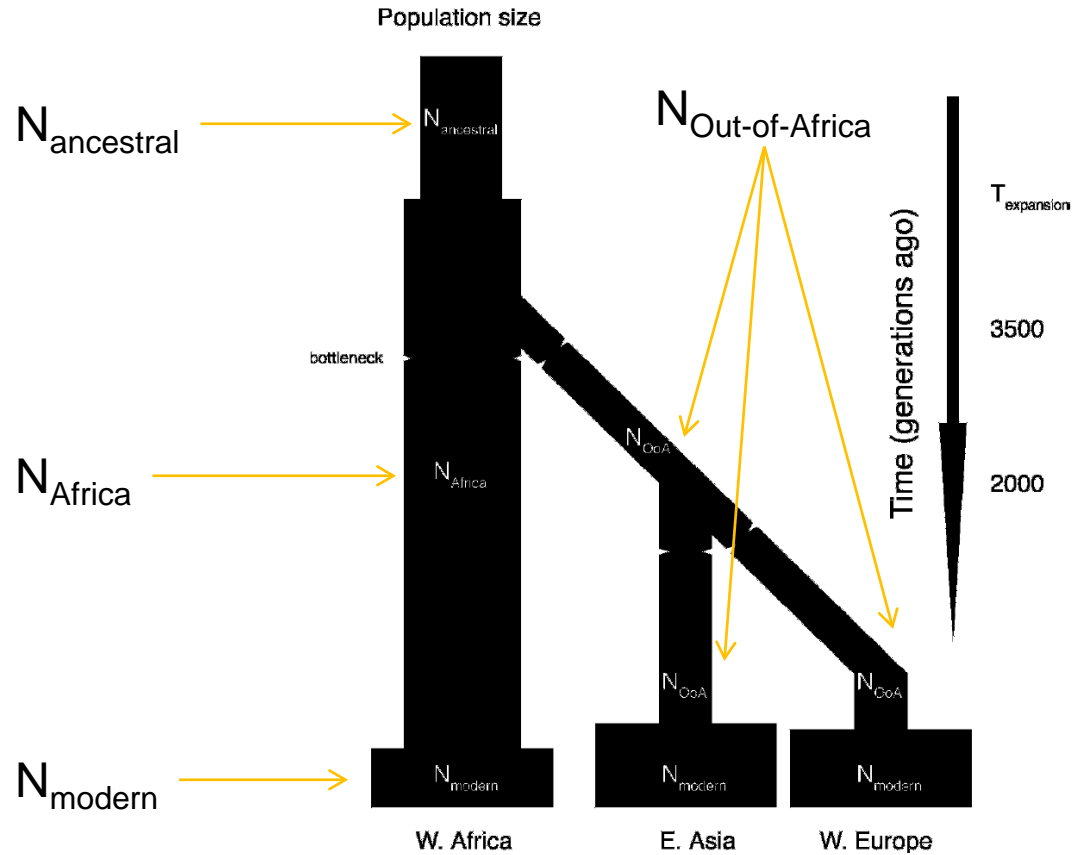


Posterior distribution of the mutation rate



Marjoram and Tavaré (2006)

# Inferences Not Restricted to $\theta$



Schaffner et al (2005)

# Inferences Not Restricted to $\theta$

Variable parameters	Best-fit model
$N_e$ (ancestral)	12,500
$N_e$ (African)	24,000
$N_e$ (non-African)	7700
$T$ (African expansion) (gens)	17,000
OoA bottleneck ( $F$ )	0.085
Asian bottleneck ( $F$ )	0.067
European bottleneck ( $F$ )	0.020
African bottleneck ( $F$ )	0.008
Africa ↔ Europe migration rate (per chromosome)	$3.2 \times 10^{-5}$
Africa ↔ Asia migration rate (per chromosome)	$0.8 \times 10^{-5}$
Recombination hotspot spacing (bp)	8500
Hotspot spacing shape parameter	0.35
Fraction of recombination in hotspots	88%
Gene conversion (initiation prob/bp)	$4.5 \times 10^{-9}$

Schaffner et al (2005)

# Recombination ...

- No recombination
  - Single genealogy
- Free recombination
  - Two independent genealogies
  - Same population history
- Intermediate case
  - Correlated genealogies

# The History of Two Sequences

Let's consider the potential history of two sequences, but this time... with a twist!

**Sequence A**

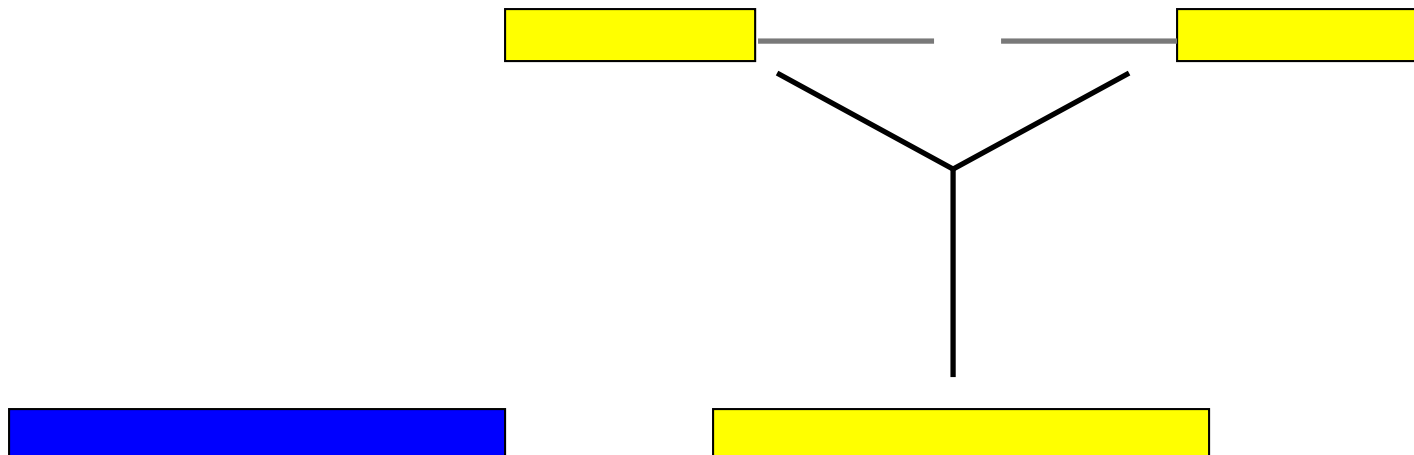


**Sequence B**



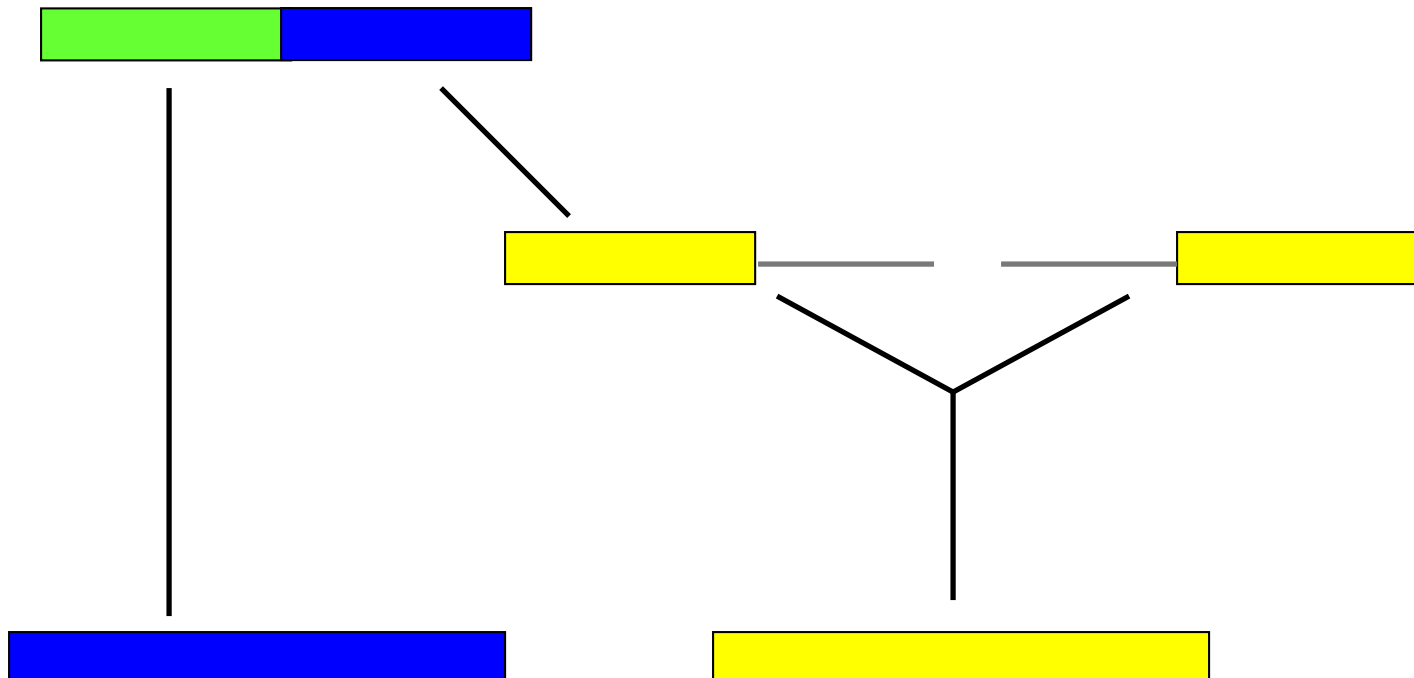
# The History of Two Sequences

Before we reach a common ancestor ... we find that sequence B is actually the result of recombination between two ancestral sequences

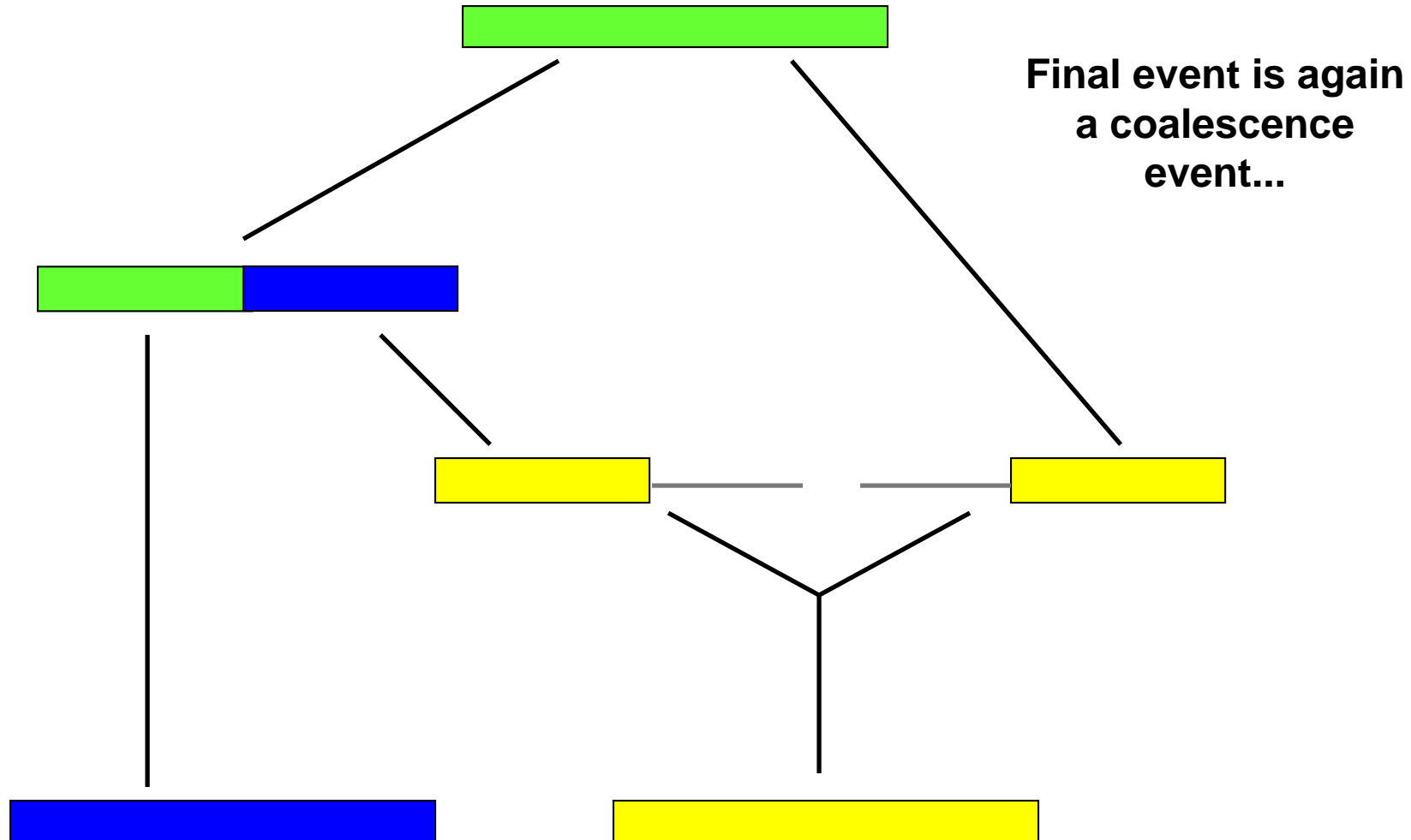


# The History of Two Sequences

The next event we encounter is a coalescence event, as expected ...



# The History of Two Sequences





# Potential Consequences ...

- Different portions of the sequence have different coalescence times
- Different portions of the sequence will show more or less variation

# Another Consequence ...

- Recombination and recurrent mutation can produce similar outcomes ...

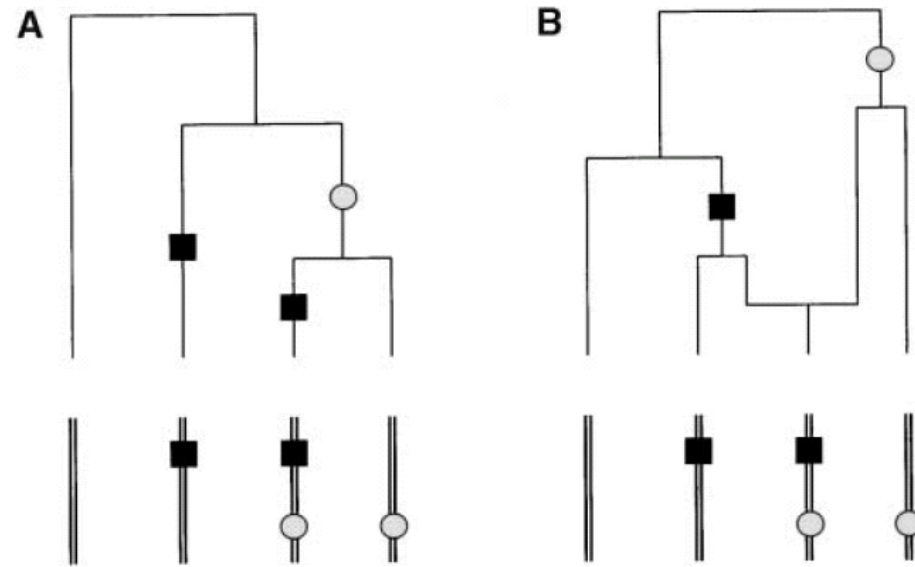


Figure from McVean et al (*Genetics*, 2001)

# Simulating the Coalescent with Recombination

- Assume the various alternative events are rare
- Time until the next event is approximately exponentially distributed
- Conditional on something happening, figure out whether it was:
  - Recombination
  - Coalescence

# Generating Genealogies

- Proceed backwards in time, until...
  - Coalescent event
    - Reduces number of ancestors by 1
  - Recombination
    - May increase number of ancestors by 1

$$P_{CA} \approx \binom{n}{2} / 2N$$

$$P_{rec} \approx nr$$

P(First Event is CA)

$$\begin{aligned} P(\text{no rec}) &= \frac{P_{CA}}{P_{CA} + P_{rec}} = \frac{\binom{n}{2} / 2N}{\binom{n}{2} / 2N + nr} \\ &= \frac{n-1}{4Nr + n - 1} \\ &= \frac{n-1}{R + n - 1} \end{aligned}$$

# Coalescent W/ Recombination

- Analytical results are difficult
- Typical approach is to ...
- First, simulate ancestral recombination graphs (ARG)
  - Coalescent tree with recombination events
- Study sample properties implied by simulated ARGs
  - For example, similarity in frequencies of neighboring SNPs

# Correlated Genealogies

- Produce correlation in
  - Allele frequencies
  - Number of mutations
  - Distribution of alleles among chromosomes
    - Linkage disequilibrium
- Use simulations to evaluate distributions as a function of recombination rate

# Example 1

- Consider a sample of  $n = 90$  chromosomes
- 2 locus coalescent, focus on samples where
  - $n_A = 30$
  - $n_B = 20$
- Over the next few slides, we will explore distribution of  $n_{AB}$ 
  - And consequently of  $D'$ ,  $r^2$



# Low Recombination

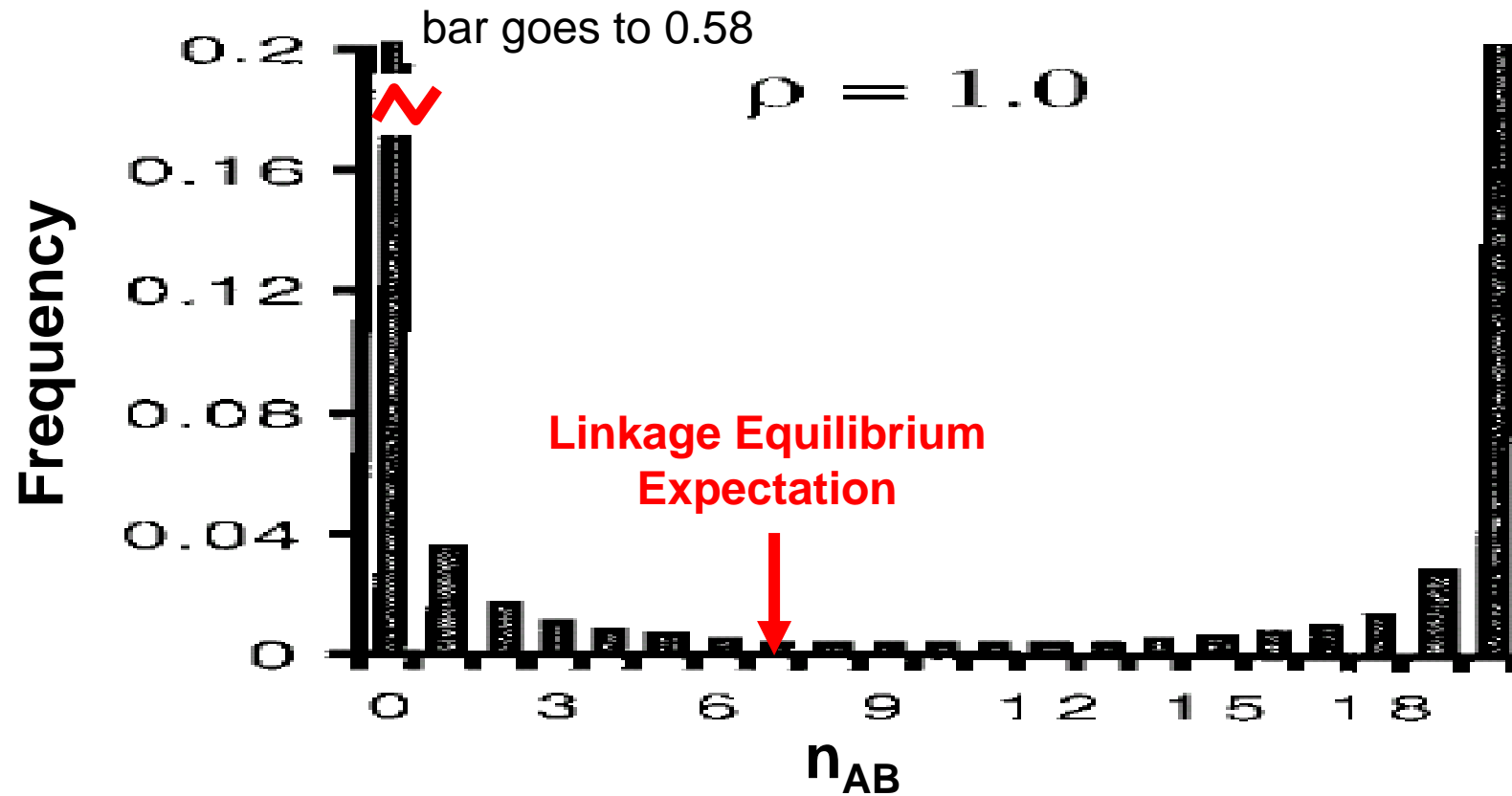


Figure from Hudson et al (Genetics, 2001)

# Higher Recombination

$$\rho = 10.0$$

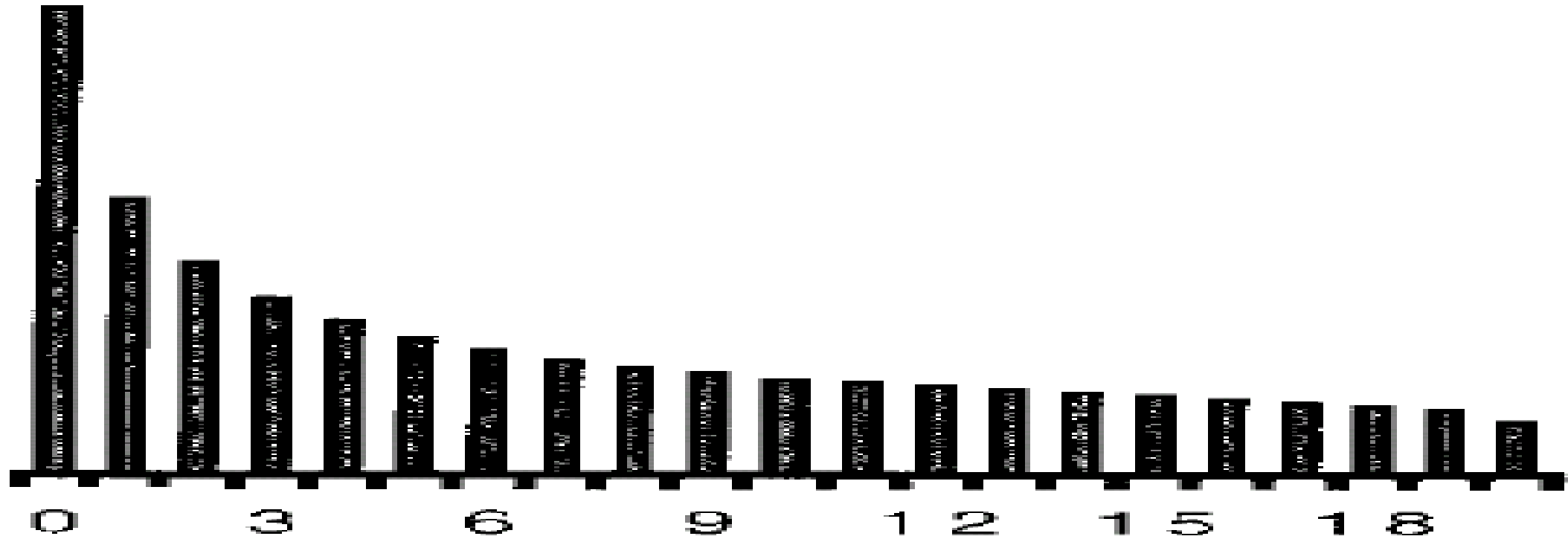


Figure from Hudson et al (Genetics, 2001)

# High Recombination Rate

$\rho = 100.0$

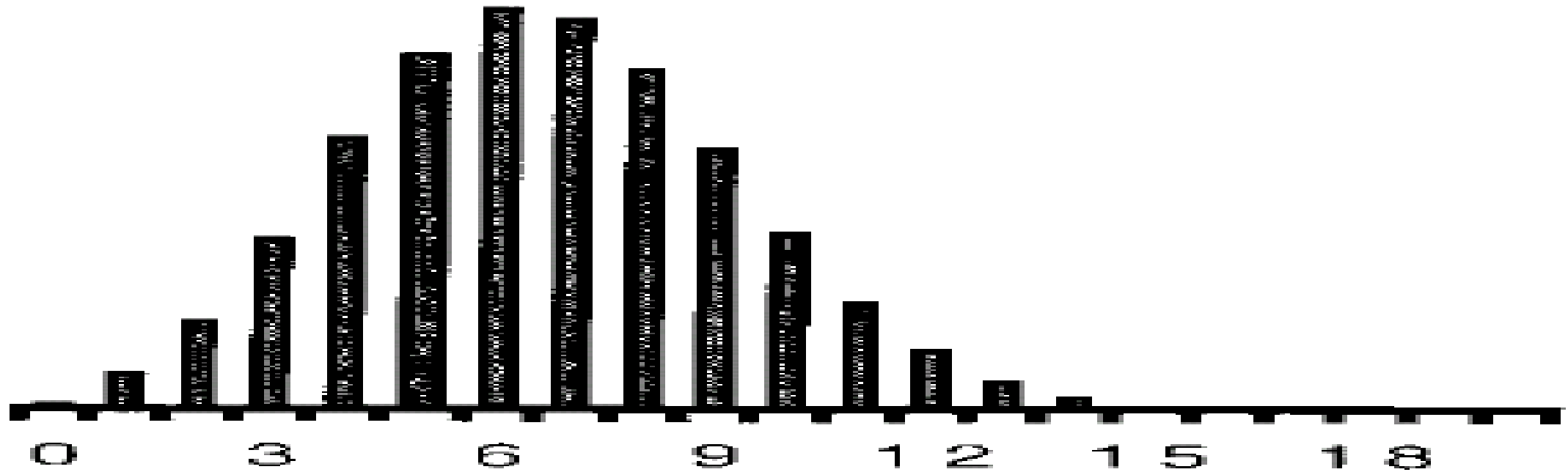
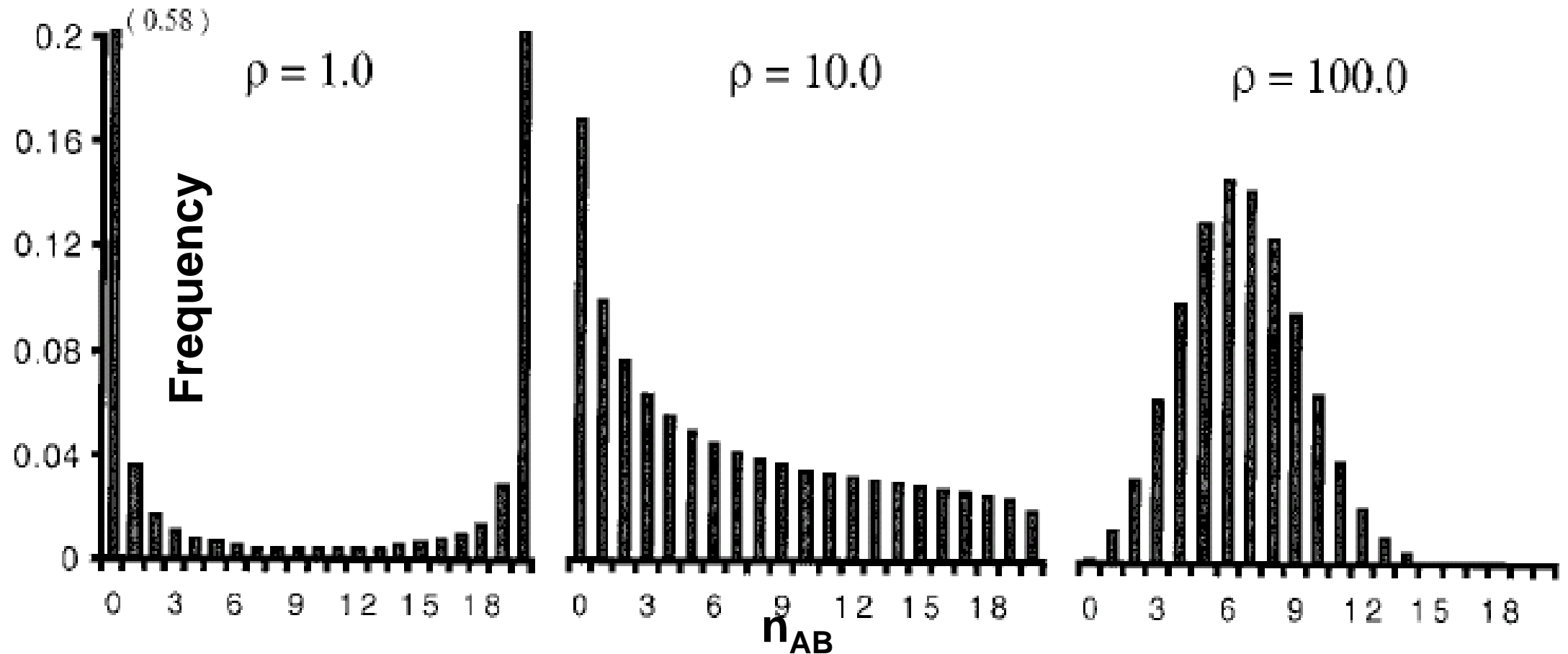


Figure from Hudson et al (Genetics, 2001)

# Impact of Recombination on Haplotype Distribution



# Some Notes ...

- If we are interested in studying the local recombination rate, neither  $r^2$  or  $D'$  retain all the information contained in  $n_A, n_B, n_{AB}$
- We can estimate  $R$  or  $\rho$  by finding the value that maximizes the probability of the observed sample configuration

# Estimating Recombination Rates

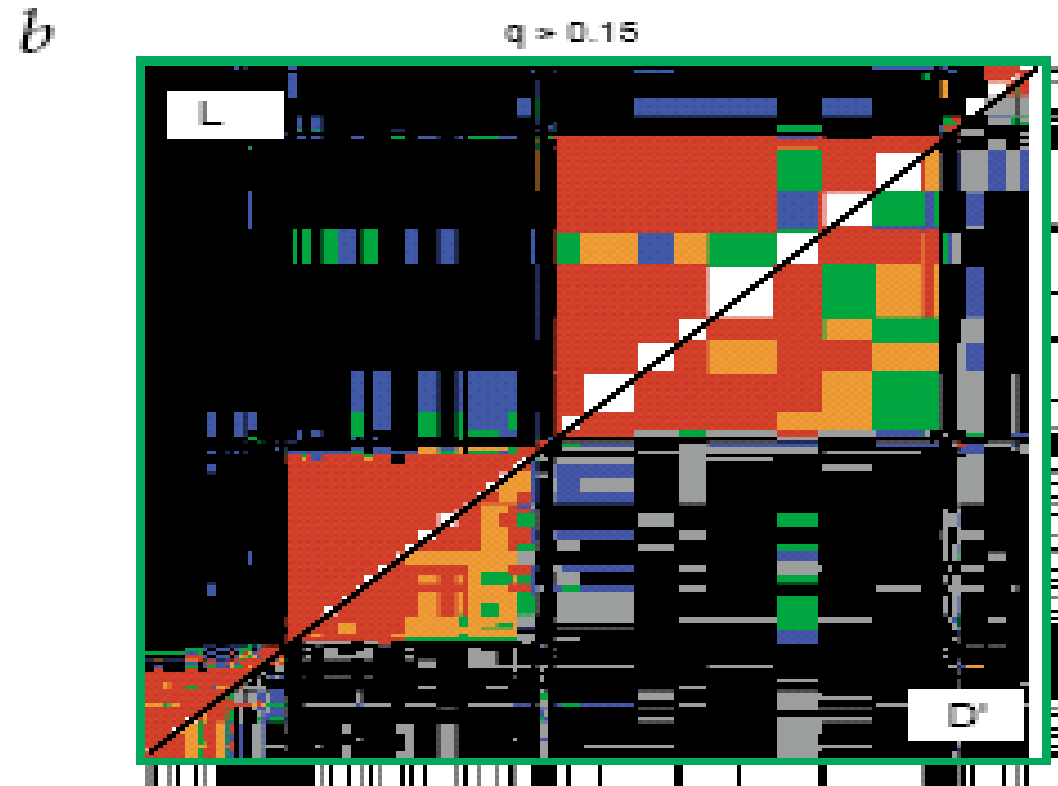
- McVean et al. (*Science*, 2004) estimated the following “pseudo-likelihood” for a sample of haplotypes:

$$\ell(4Nr) = \sum_{ij} \ell(n_i, n_j, n_{ij} | 4Nr_{ij})$$

(summation is over all pairs of markers)

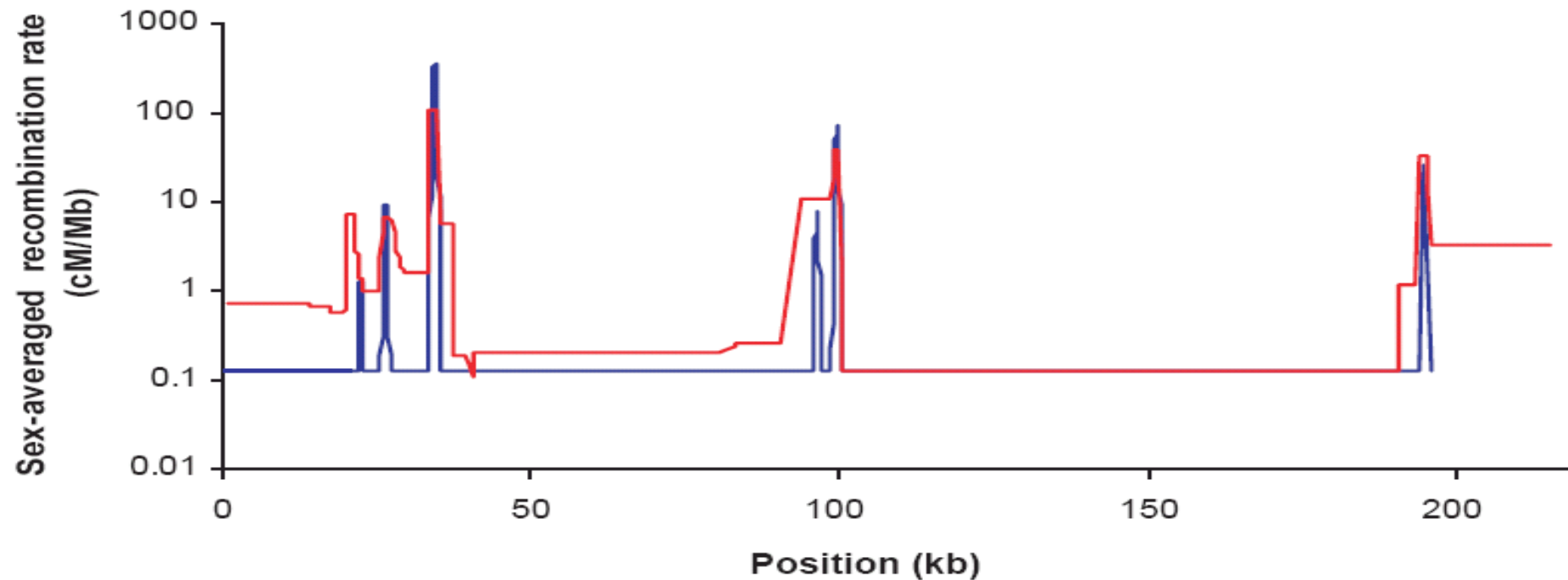
- Estimated recombination rates allow us to predict what other chromosomes or samples from the population might look like.

# Pairwise LD in HLA



Pairwise LD data from Jeffrey's et al (2001)

# Recombination Rate Within HLA



**Fig. 2.** Comparison between estimates of local recombination rates from population genetic data (red) and sperm analysis (blue) in the HLA region; data from (3). To convert the male crossing over rates to sex-averaged rates, we used the previous observation that the female crossing-over rate in this region is about four times that of males (42).



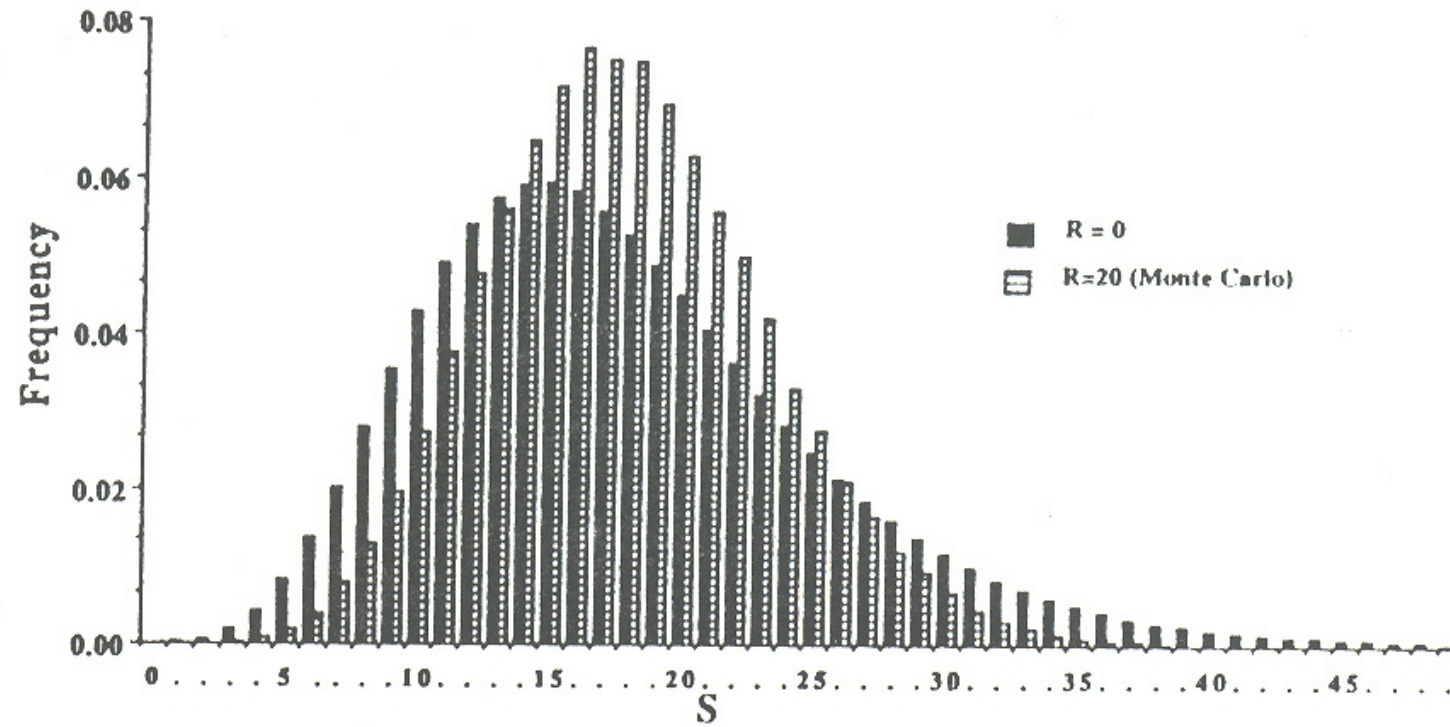
# Total number of mutations

- Recombination does not change expectation for  $S$ ...

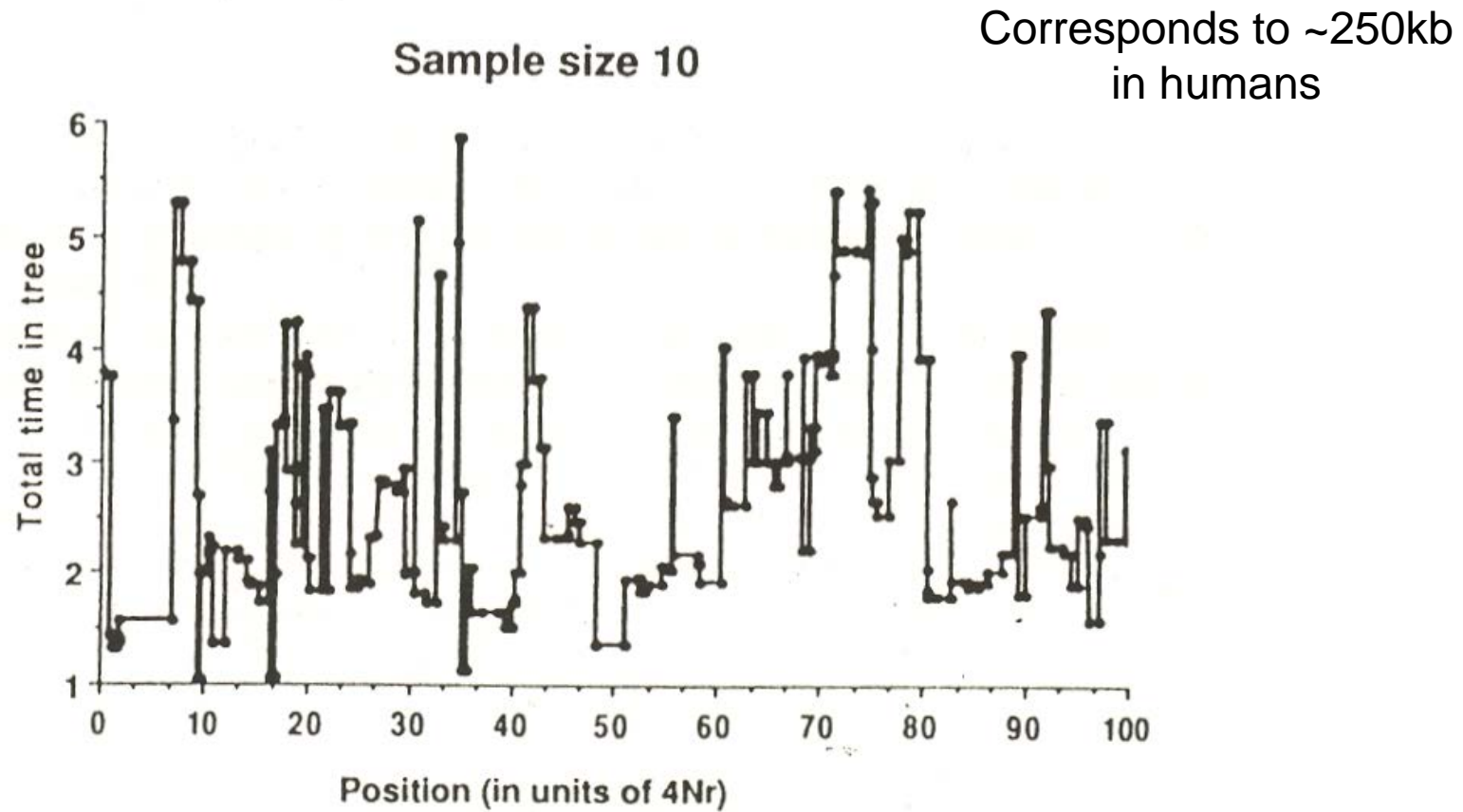
$$E(S) = 4N\mu \sum_{i=1}^{n-1} 1/i = \theta \sum_{i=1}^{n-1} 1/i$$

- ... but it reduces its variance.
  - With large  $r$ ,  $S$  is effectively averaged over multiple genealogies

# Number of Mutations



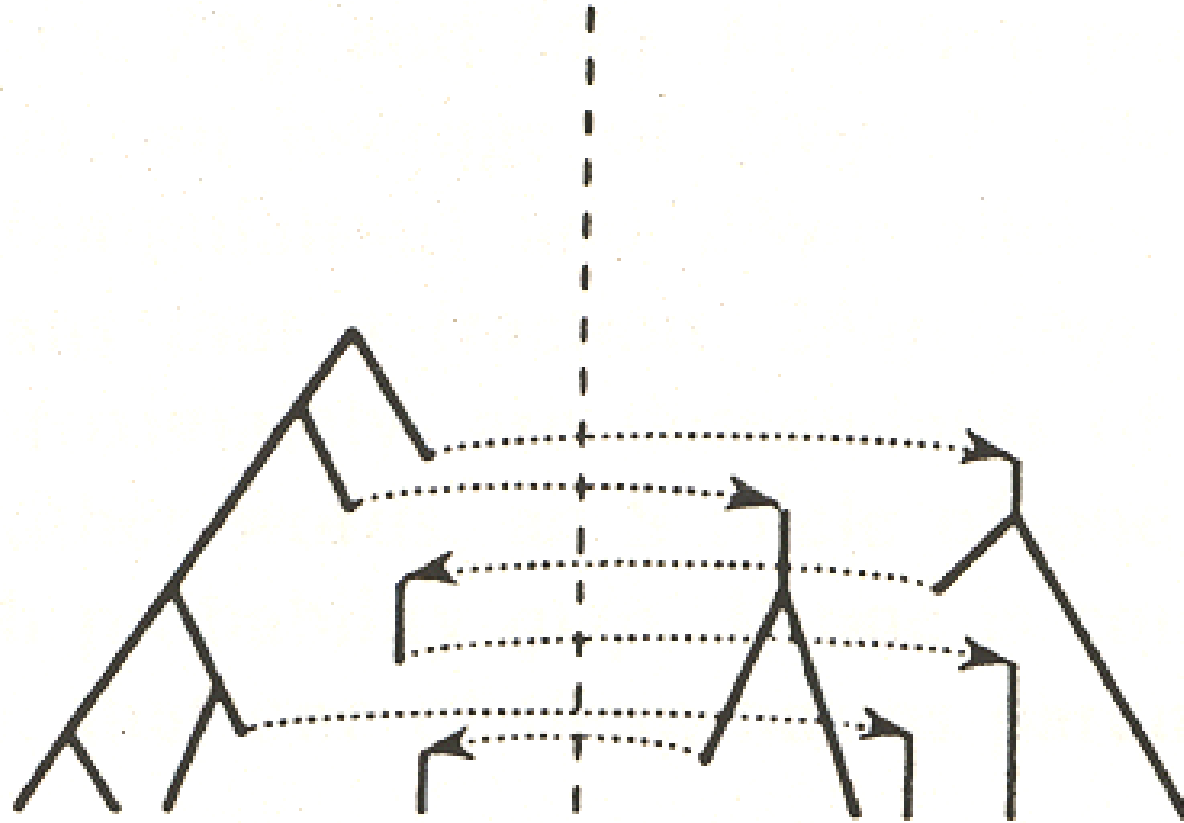
# Total Time in Tree



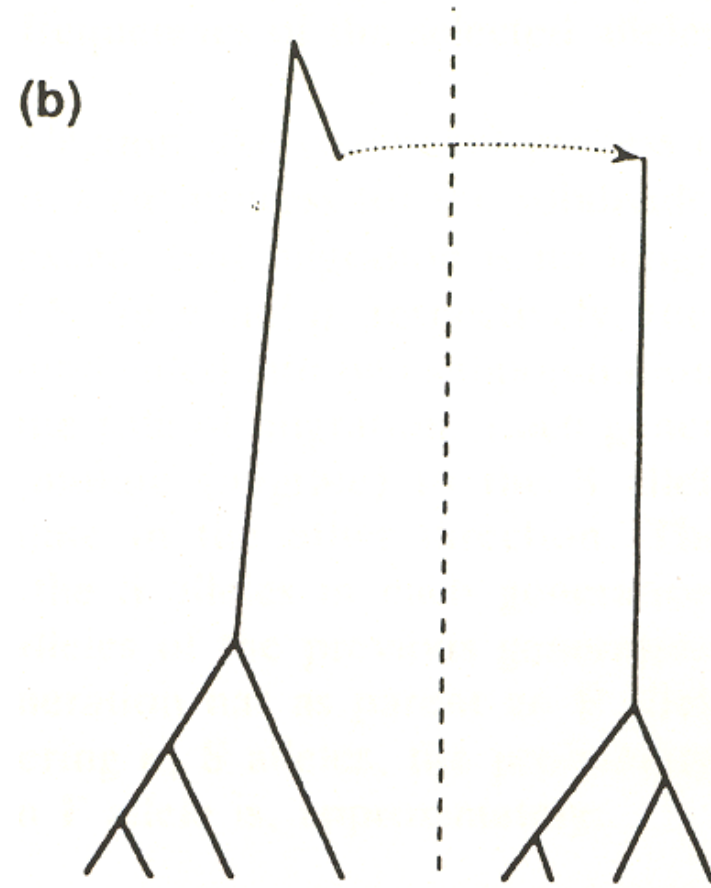
# Population Subdivision

- What if the population is not mating at random, but is made up of multiple small groups?
- Track migration among ancestors...

High Migration rate ...



Low Migration rate ...



# Formulae:

If the two subpopulations each have  $N$  diploids

Coalescent among  $n_1$  lineages in population 1

$$\binom{n_1}{2} / 2N$$

Coalescent among  $n_2$  lineages in population 2

$$\binom{n_2}{2} / 2N$$

Migration

$$(n_1 + n_2)m$$

# Conditional Probabilities

Conditional Probability of Coalescence

$$\frac{\binom{n_1}{2}}{\binom{n_1}{2} + \binom{n_2}{2} + (n_1 + n_2) \frac{M}{2}}$$

is the conditional probability of coalescence in population 1.

Here,  $M = 4Nm$ , is the scaled migration rate.



# Conditional Probabilities

Conditional Probability of Migration

$$\frac{n_1 \frac{M}{2}}{\binom{n_1}{2} + \binom{n_2}{2} + (n_1 + n_2) \frac{M}{2}}$$

is the probability of migration from population 1 to 2.

# Questions that Coalescent Can Tackle...

- Frequency spectrum of observed mutations
  - Impact of population growth
  - How many mutations are unique?
- Disequilibrium coefficient
  - Joint distribution of  $(p_A, p_B, D_{AB})$
  - Impact of population growth

# MS Computer Program

- Coalescent Simulator
  - by Richard Hudson at U. of Chicago
- Generates samples of sequences
  - Population and subpopulation sizes
  - Mutation rate ( $\theta = 4N\mu$ )
  - Recombination rate ( $R = 4Nr$ )
- <http://home.uchicago.edu/~rudson1/>

# Recommended Reading

- Richard R. Hudson (1990) “Gene Genealogies and the coalescent process”
  - from Oxford Surveys in Evolutionary Biology, Vol. 7. D. Futuyma and J. Antonovics (Eds). Oxford University Press, New York.

# Further Reading

- Marjoram P and Tavaré S (2006).  
Modern computational approaches for analysing molecular genetic variation data.  
*Nature Reviews Genetics* **7**:759-770
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ and Altshuler D (2005).  
Calibrating a coalescent simulation of human genome sequence.  
*Genome Research* **15**:1576-83