The effects of population pedigrees on gene genealogies

John Wakeley*, Léandra King*,
Peter Wilton*, Sohini Ramachandran†

*Harvard University, †Brown University

University of Montreal, October 2013

Preview

Introduction

- Randomness in population genetics
- Classical versus coalescent approaches
- The population pedigree

Coalescence in fixed pedigrees

- Wright-Fisher pedigree simulations
- Per-generation probabilities of coalescence
- Large families and selective sweeps
- Pedigree effects in subdivided populations

Sources of Randomness in Population Genetics

1. Survival, Movement, Finding a Mate, Reproduction

   Complicated: fundamental parameters and rules largely unknown. Ultimately, these produce a population pedigree, or set of family relationships.

2. Genetic Transmission

   In diploid organisms, genetic transmission follows Mendel's Laws, which are well known.

3. Mutation (also Recombination)

   DNA replication error seems Markovian and has a very low rate, e.g. $\sim 10^{-8}$/site/generation in humans.

Questions of This Talk

1. To what extent can Mendelian inheritance alone account for the randomness in times to common ancestry observed among loci within genomes?

2. Are the effects of pedigrees on coalescence large enough in subdivided populations that we need to correct for them?

Population Genetic Models (*e.g.* Wright-Fisher Model)

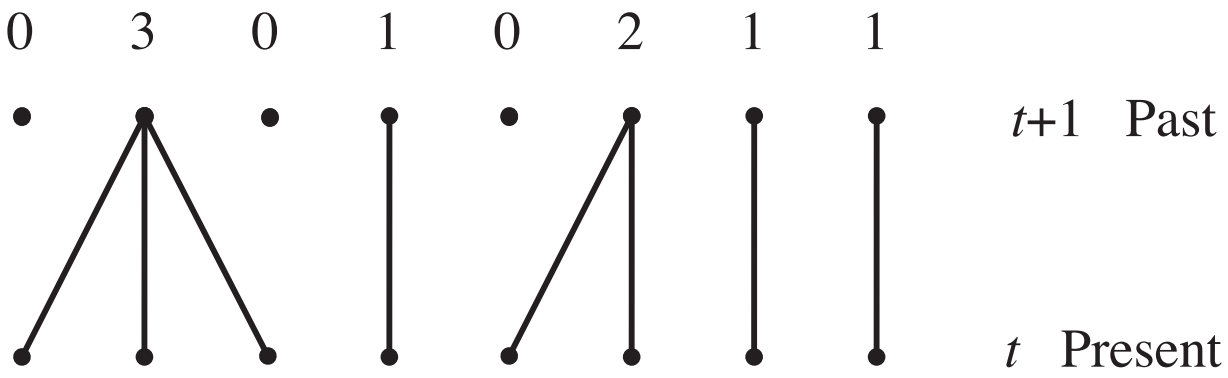
We combine Mendelian inheritance with assumptions about survival, reproduction, etc., to model the dynamics of genetic variation over time. Simple models typically assume:

1. The organism is haploid (or diploid and hermaphroditic).
2. The population is closed and well mixed (randomly mating).
3. The population size N is large and constant over time.
4. Genetic variation is selectively neutral.

Wright, Fisher, Kimura, and others developed a *diffusion theory* to predict changes in allele frequencies over time. Kingman, Hudson, Tajima, and others developed backward-time *coalescent theory*.

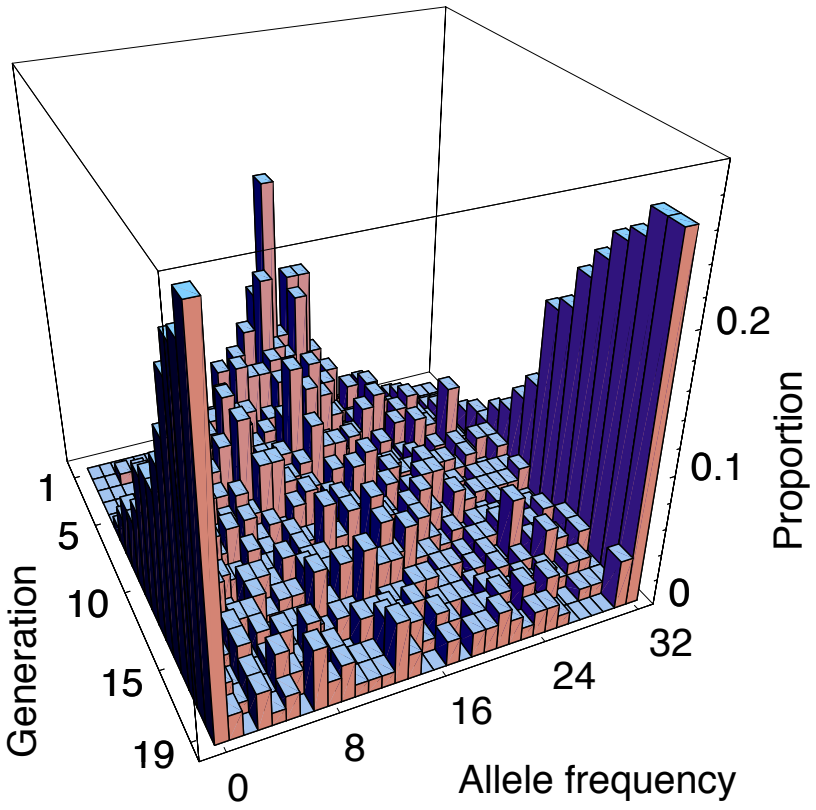
Exchangeable Population Models of Cannings (1974)

The number of "gene copies" $2N$ is constant over time. The "offspring numbers" $(\nu_1, \ldots, \nu_{2N})$ are exchangeable random variables, with the requirement that $\sum_{i=1}^{2N} \nu_i = 2N$.

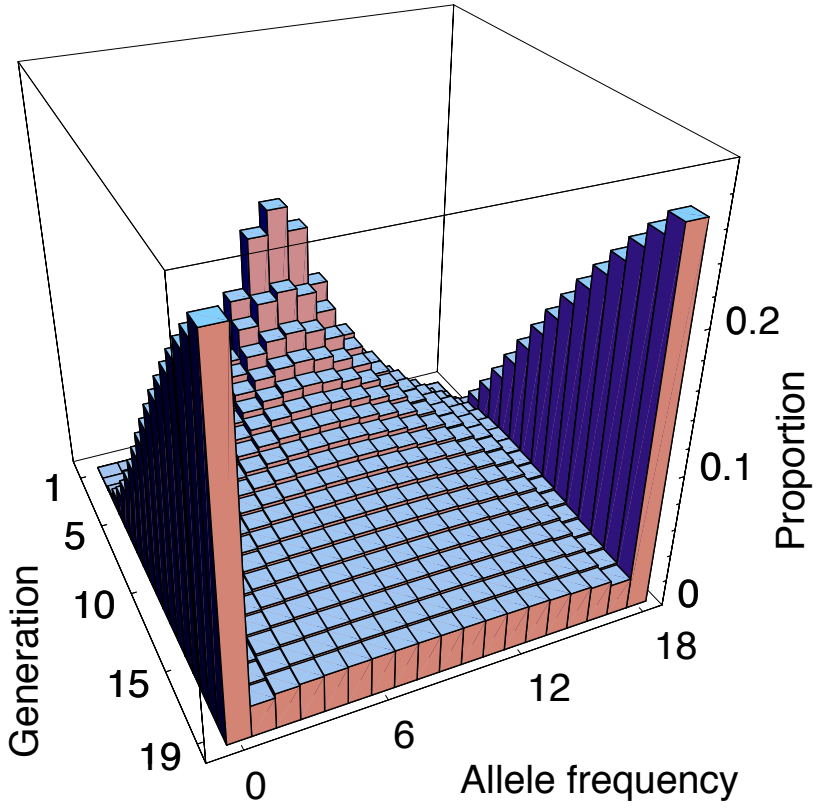


The Wright-Fisher model (of diploid hermaphrodites) can be reduced to a (haploid) Cannings model.

Buri (1956) Experiment: Application of Diffusion Theory



Experimental results, following allele frequencies over 19 generations in 108 independent populations, each with $N = 16$.



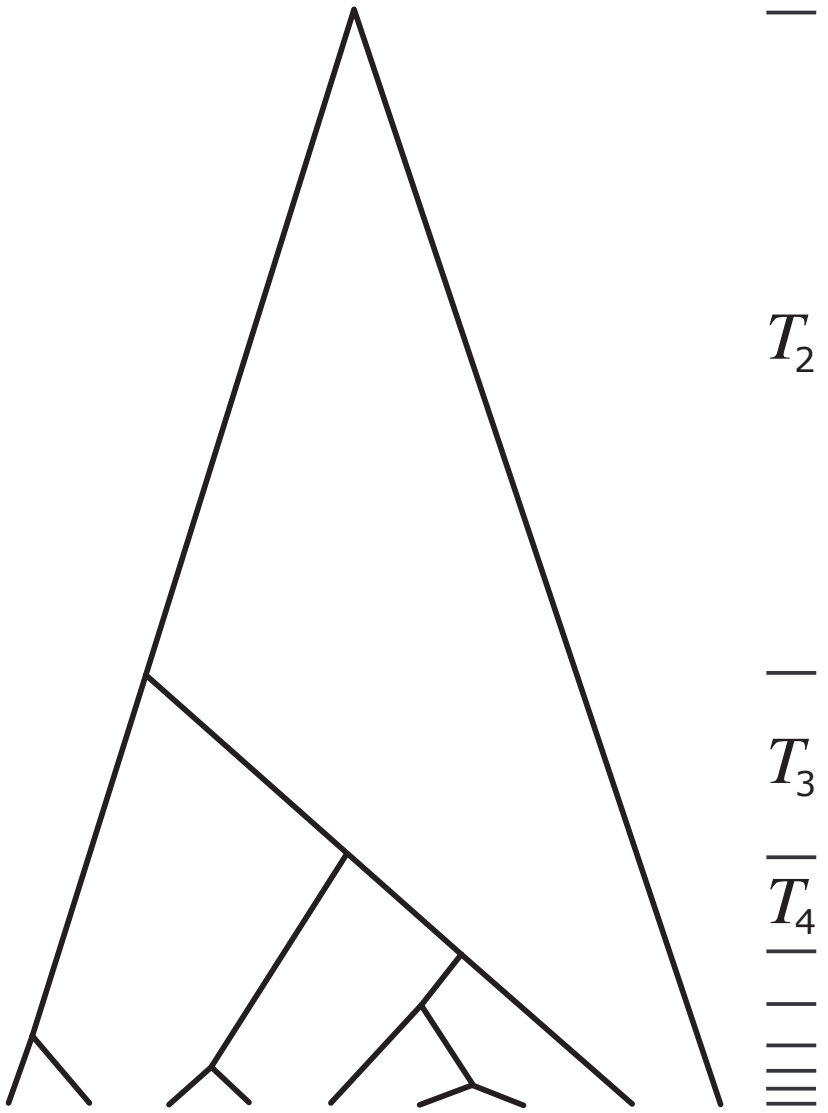
Prediction of single-locus diffusion theory about the future of one population, with "effective" size $N = 9$.

The (Standard Neutral) Kingman Coalescent

1. Finite sample size n, and population size $N \to \infty$.
2. Time measured in units of $2N_e$ generations.
3. T_i is distributed as

$$f_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}.$$

4. Random-joining tree.
5. Mutations occur with rate $\theta/2$ per unit time.



Application of Coalescent Theory to Genetic Data

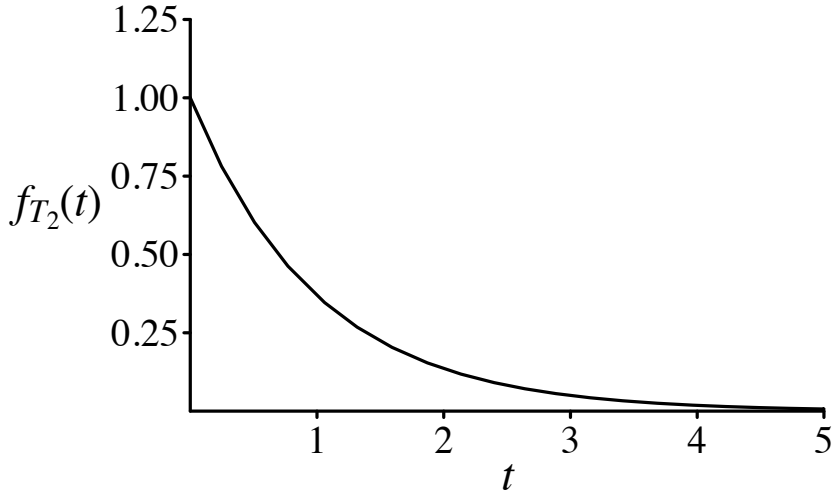
Kingman's coalescent has been used to explain variation in levels of polymorphism among loci within genomes.

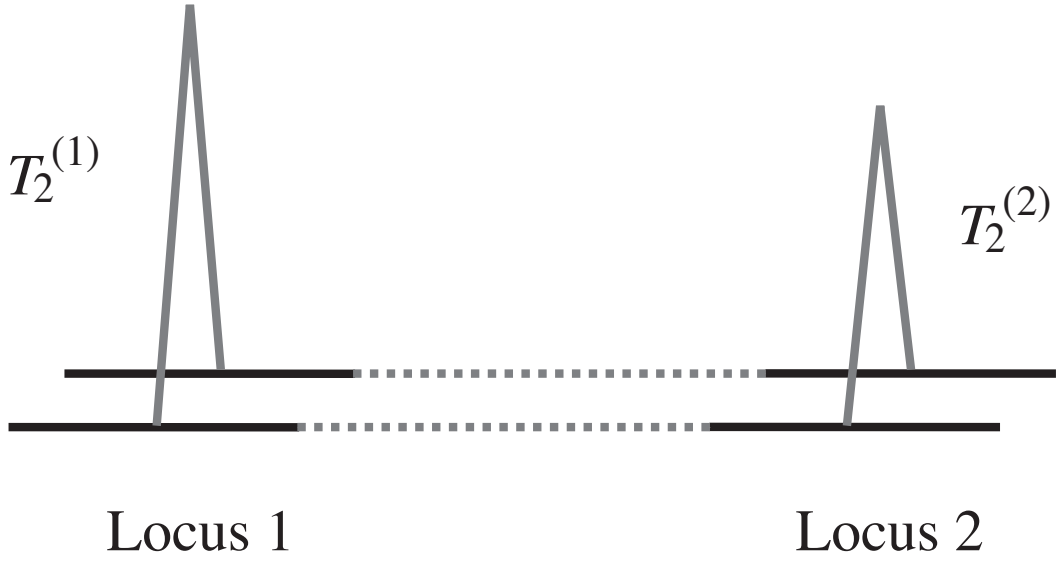
# SNPs	Poisson	Coalescent	Observed
0	8256	8767	8796
1	3040	2332	2247
2	617	663	668
3	99	200	214
4	16	66	102

Table 3 of The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**:928-933.

Interpretation: Independent Loci, Independent Genealogies

Exponential distribution of pairwise coalescence times (among loci).





(Also variation due to the random process of mutation.)

Two Conceptually Different Random Experiments

1. Repetition of the entire population process in independent replicate populations (as in Buri's laboratory experiment).

2. Repetition of the process of Mendelian inheritance among different genetic loci within a single population pedigree.

For natural populations (1) is entirely hypothetical, whereas (2) is realized when we take samples from different genetic loci.

Averaging in Kingman's Coalescent

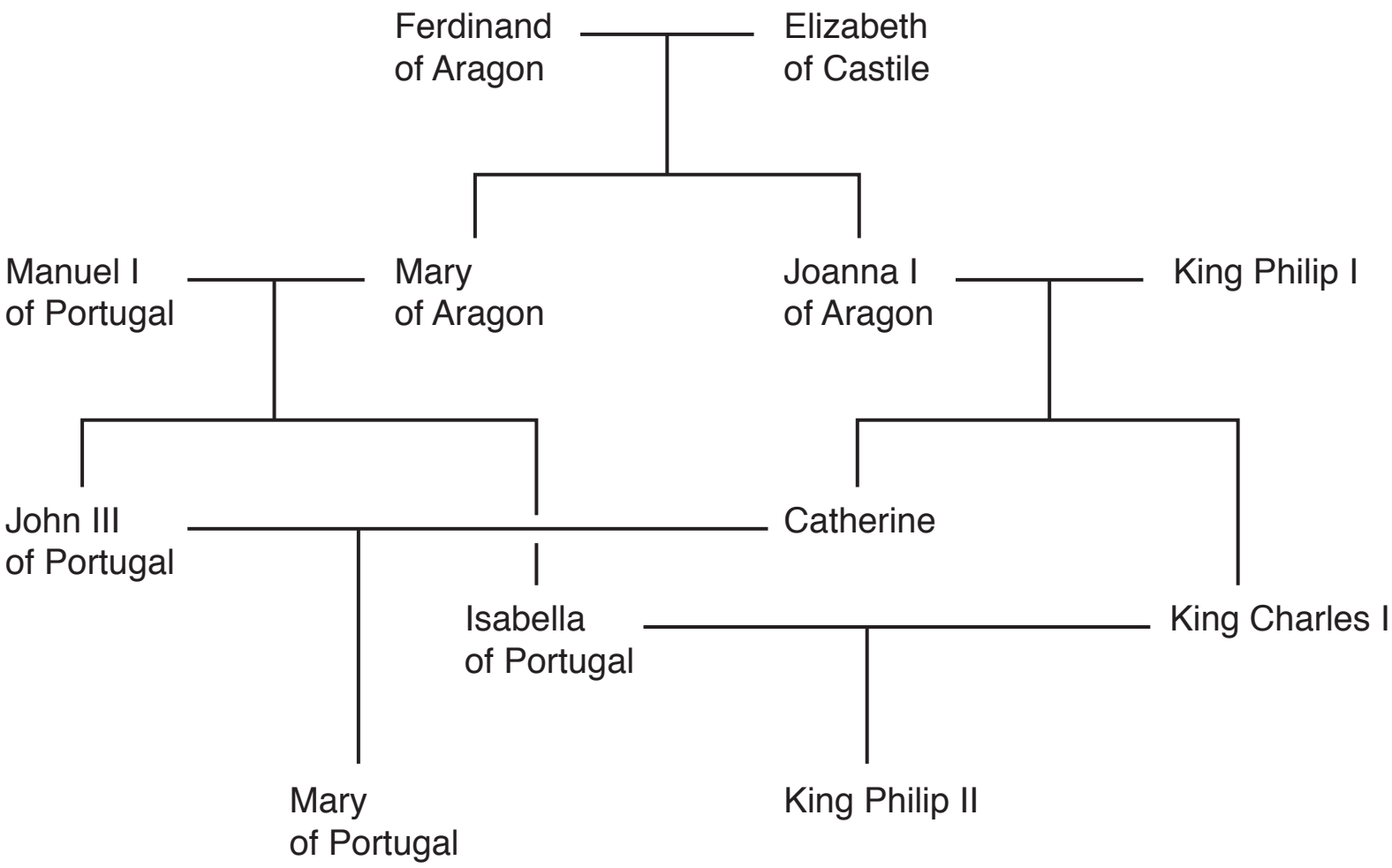
Kingman obtained the coalescent in the limit $N \to \infty$, from a time-homogeneous Markov model (subset of Cannings' model).

$$P(\text{coal}) = E\left[\sum_{i=1}^{2N} \frac{\nu_i(\nu_i - 1)}{2N(2N-1)}\right] = \frac{E[\nu_1(\nu_1 - 1)]}{2N - 1}$$

Homogeneity is achieved, as above, by averaging over the process of reproduction (i.e. over the pedigree in diploid biparental models).

Note that for this model $N_e := N/\sigma^2$, with $\text{Var}[\nu_1] \to \sigma^2$.

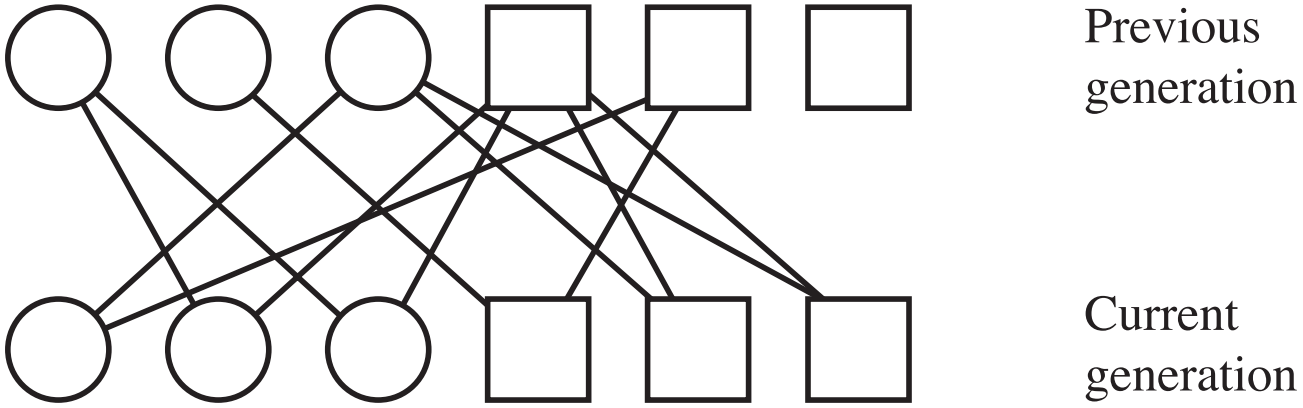
A Small Piece of the Human Population Pedigree



Alvarez et al (*PLoS One*, 2009 4:e5174) paper about Spanish Habsburg King Charles II ($F = 0.254$).

Simulations of Diploid, Two-Sex Reproduction

For this talk: random-mating (WF) and constant, equal sex ratio.



1. Generate the pedigree of the population.
2. Sample two individuals and follow gene copies back through the pedigree (1/2 to mother, 1/2 to father each generation).

Simulations of Coalescence-Probability Distributions

1. Generate 10000 pedigrees under two-sex Wright-Fisher model.

2. Sample one pair of individuals.

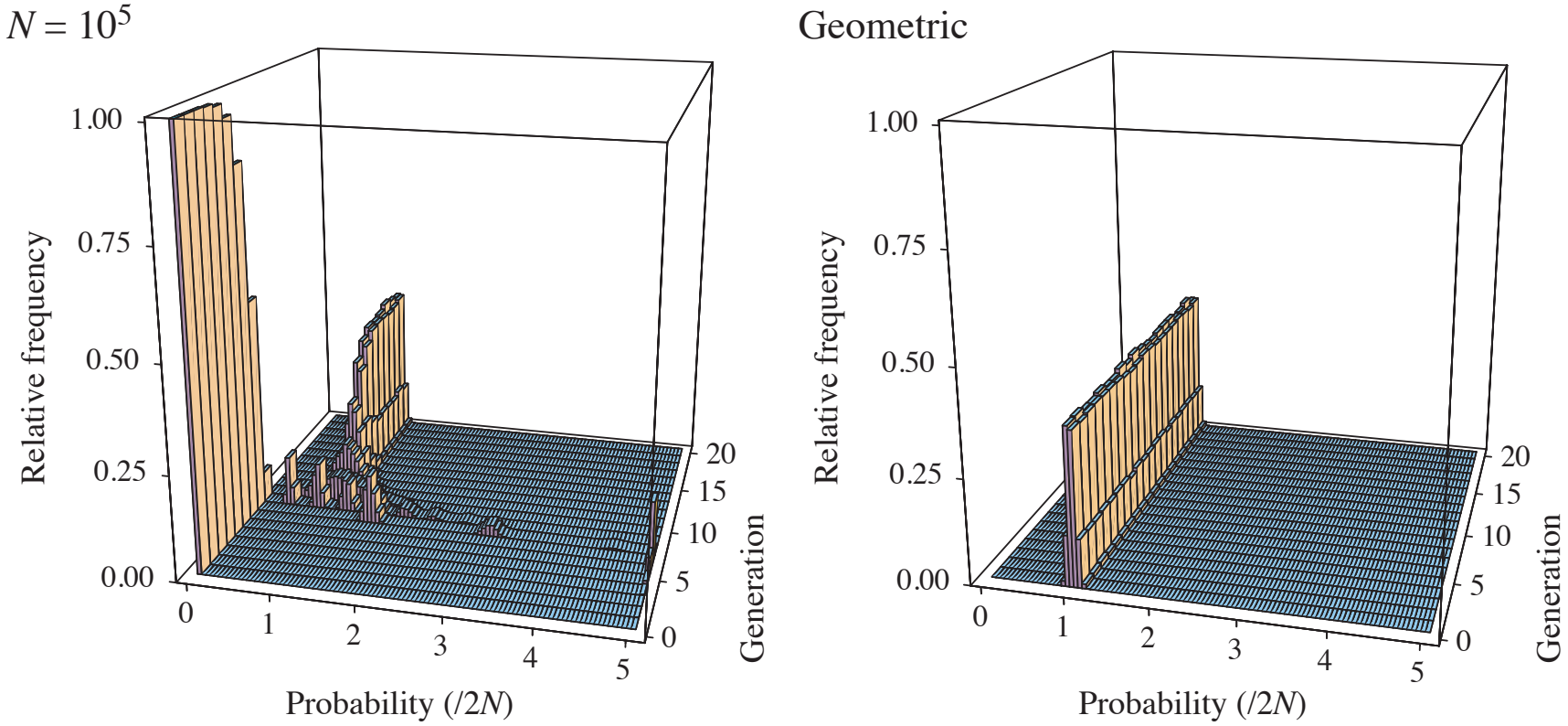
3. For each pedigree and sample-pair (1 & 2):

   Simulate 10^8 coalescence times.

   Record the fraction of times that coalescence occurred in each of the past 20 generations.

4. Make a histogram of these probabilites for each generation.

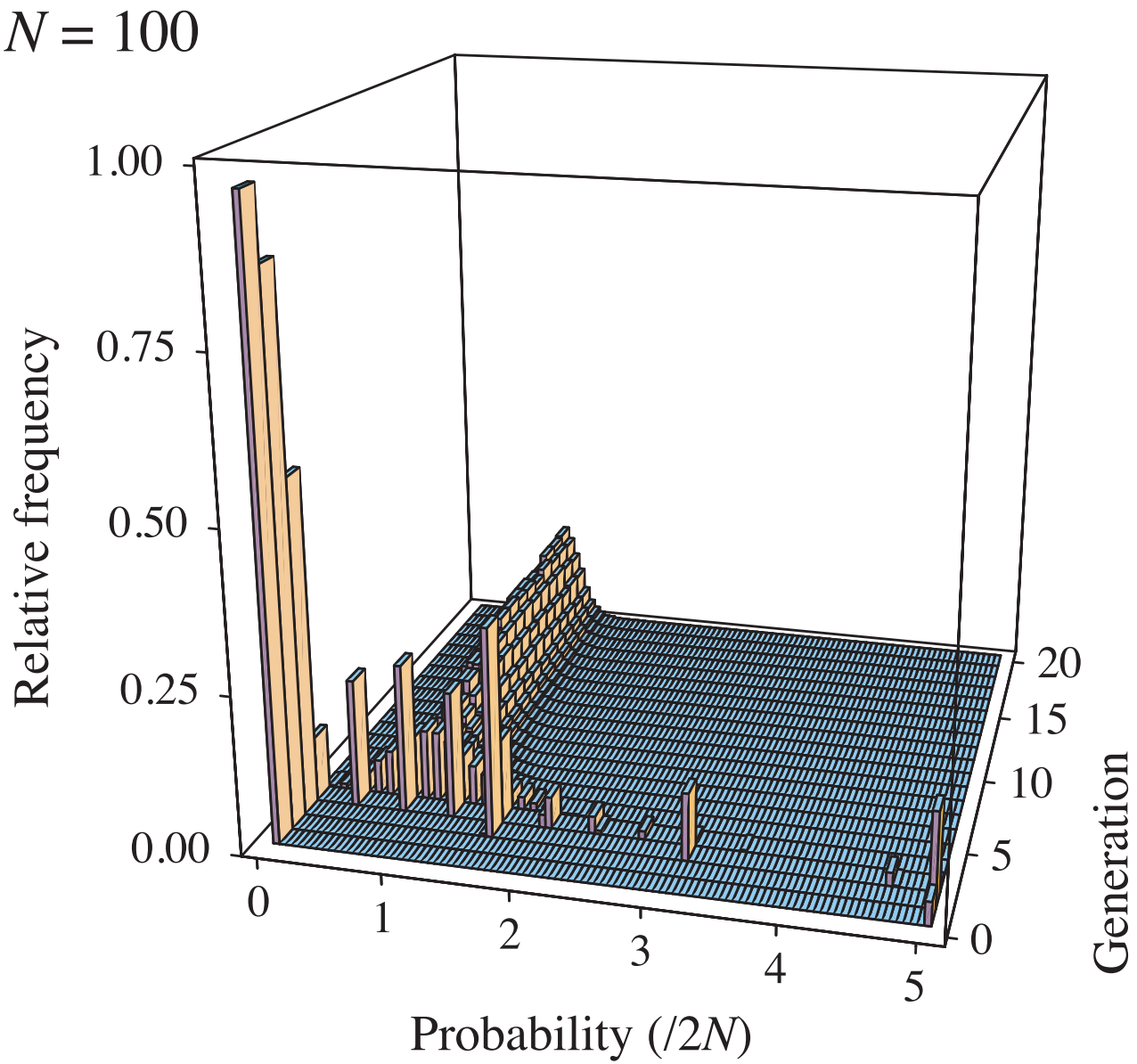
Pedigree-Coalescence Times v. Geometric Times



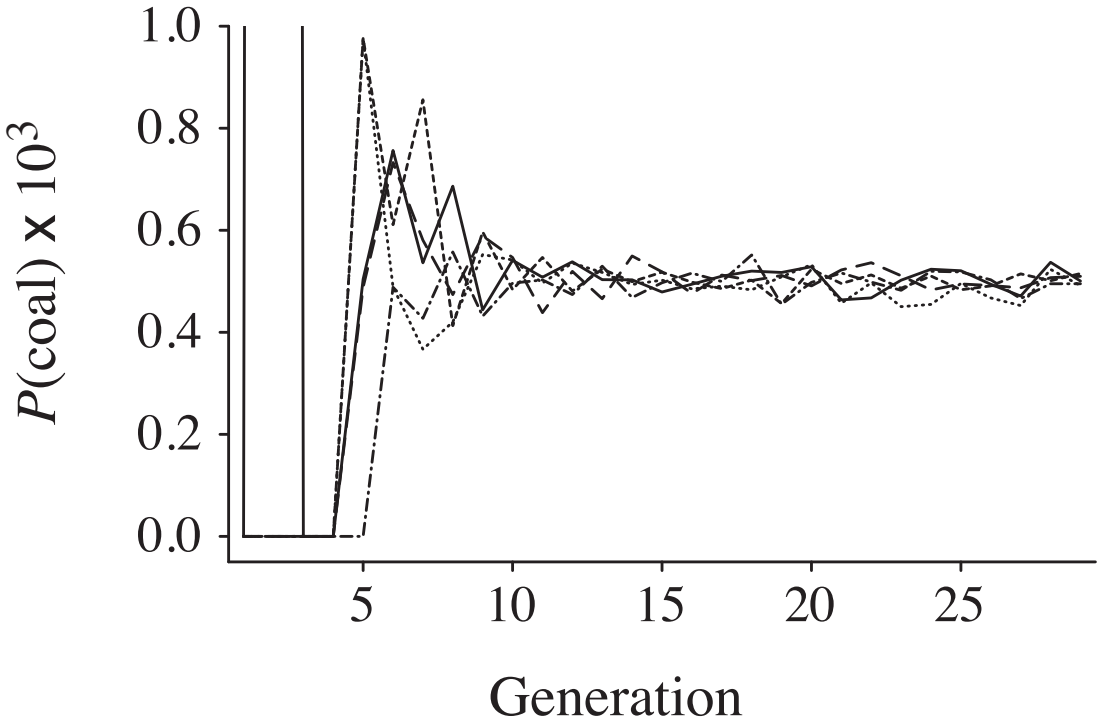
Left: Pairwise times on Wright-Fisher pedigrees with $N = 10^5$, nearly homogeneous after about $\log_2(2N)$ generations.

Right: Times from a geometric distribution with $p = 1/(2 \times 10^5)$.

Pedigree Effects are Stronger in Smaller Populations



Coalescence-Probability Trajectories for 5 Single Pedigrees



Within a single pedigree, the coalescence probability jumps wildly during the first $\log_2(2N)$ generations—here $\sim$11—then settles near $1/(2N)$ in the more distant past. For this example, $N = 1000$.

Summary and Next Steps

1. The distribution of coalescence times among independent loci on (a variety of types of) fixed pedigrees differs from the Kingman coalescent, but only

   in the most recent ($< \log_2 2N_e$) generations,

   and only detectable when N_e is small.

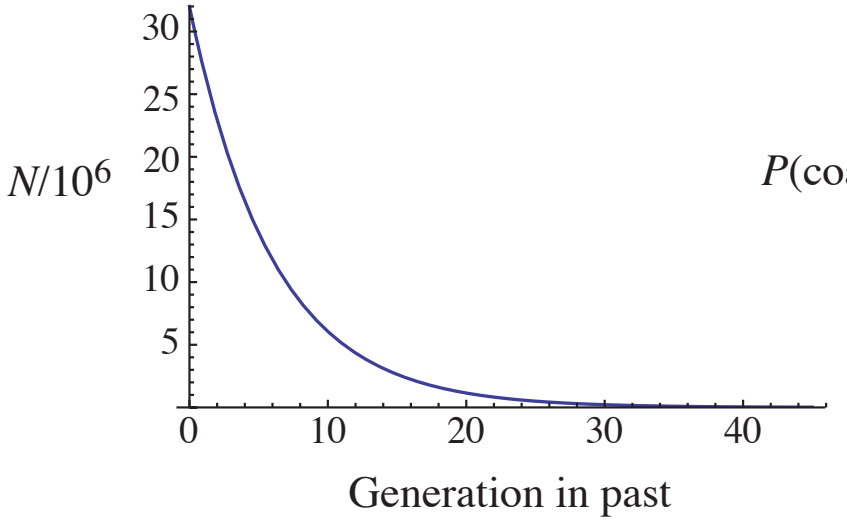
2. Thus the population pedigree may constrain coalescence times in ways not predicted by standard models.

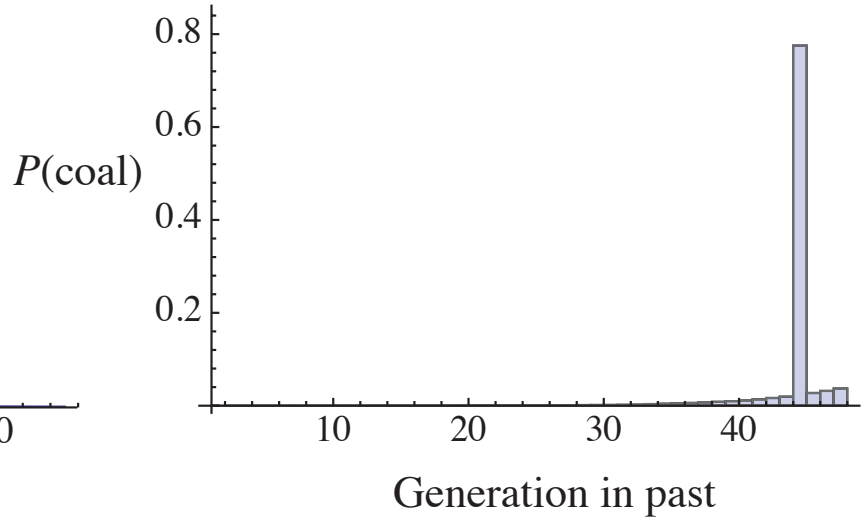
   Large families, selective sweeps, recombination.

   Migration, bottlenecks, range expansions.

Effect of a Very Large Family (∼Genghis Khan)

A single male has 2000 children at generation 43 in the past, in a population that is growing exponentially.

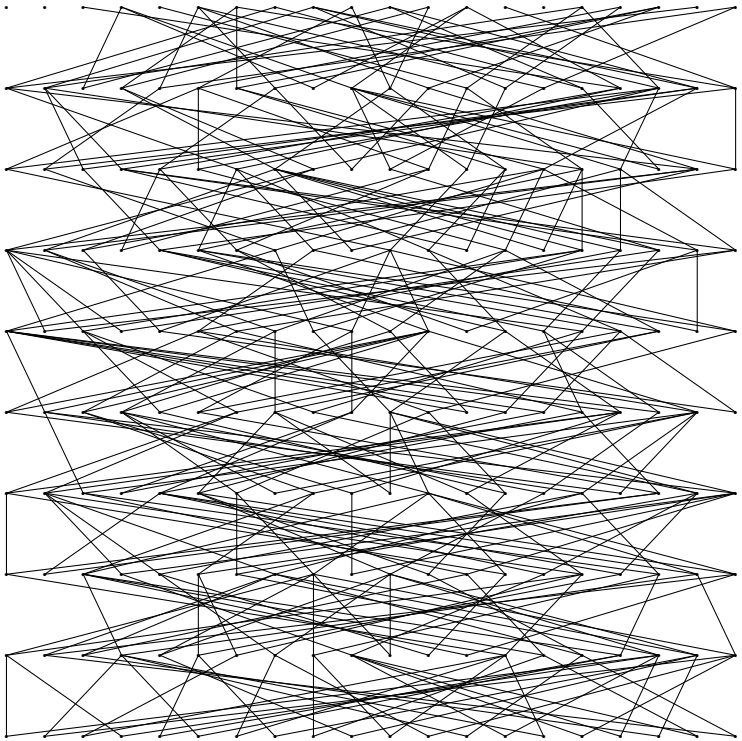




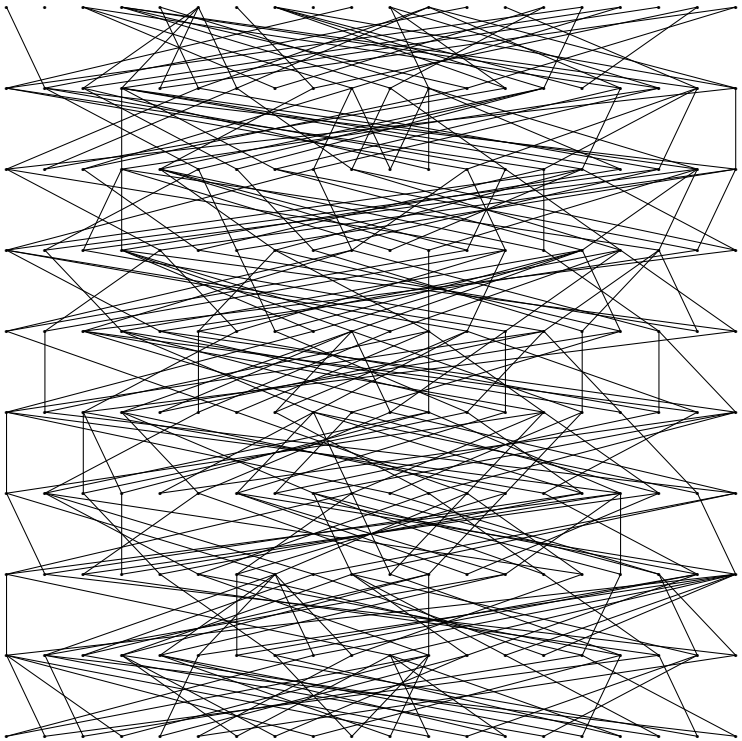
Note that P(coal) on the right is restricted to the cases in which coalescence occurs by generation 47 (roughly 1/1000 of the time).

Effect of a Selective Sweep on the Population Pedigree

Sweep of an advantageous dominant mutation, $N = 20$ and $s = 5$.



Neutral control pedigree.



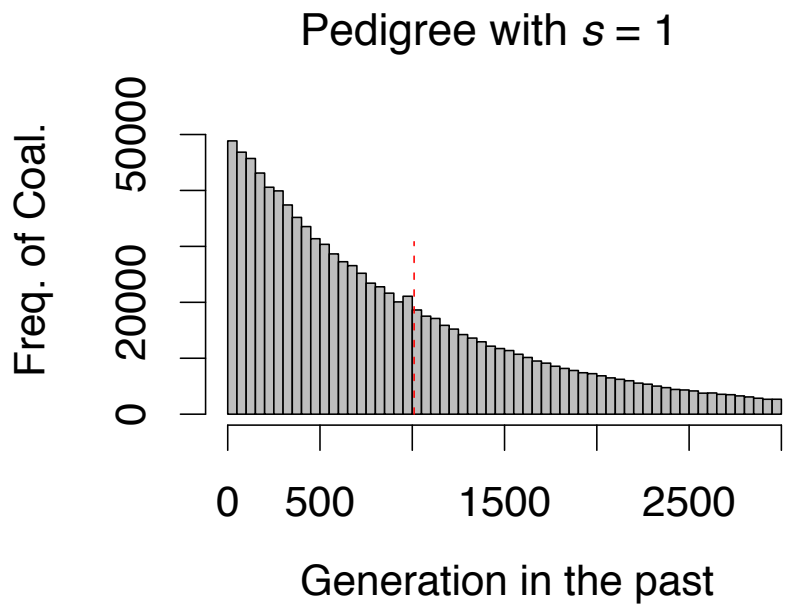
Pedigree with a sweep.

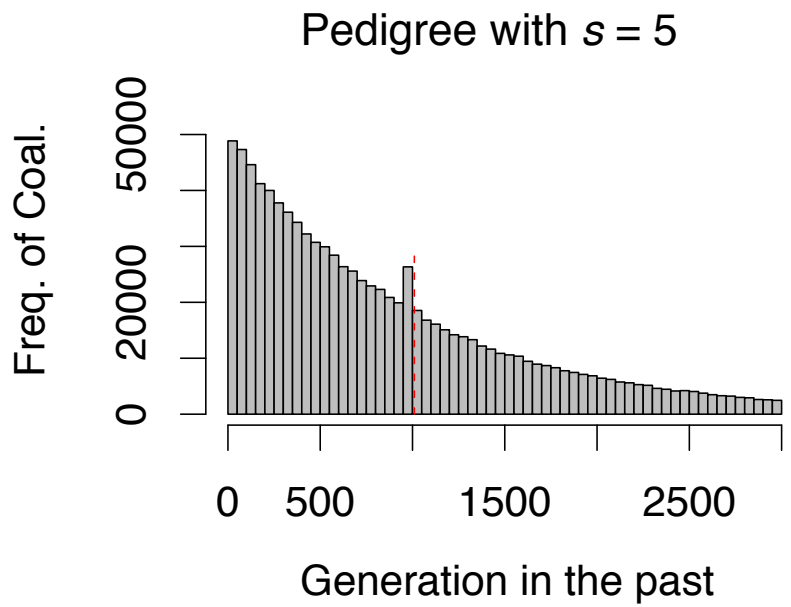
Effect of a Selective Sweep on Unlinked Loci

Advantageous dominant mutation introduced in generation 1000.

$N = 500$, and conditional on fixation before present time (0).

Distribution of pairwise coalescence times for two pedigrees under the monogamous WF model:





Thoughts about Population Subdivision and Migration

1. When local population sizes are small, the pedigree could have a strong effect of local patterns of genetic variation.

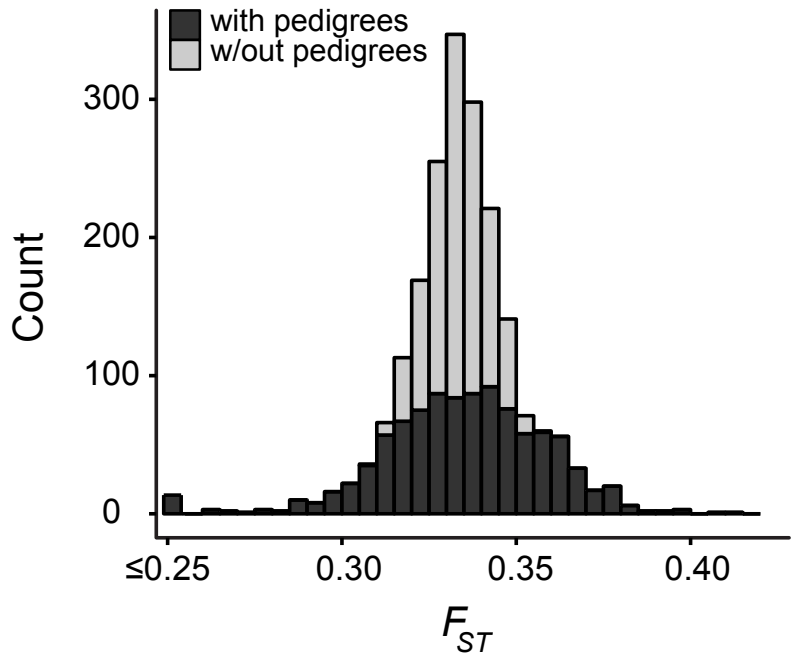
2. Since genetic lineages "migrate" through a fixed set of migrant individuals:

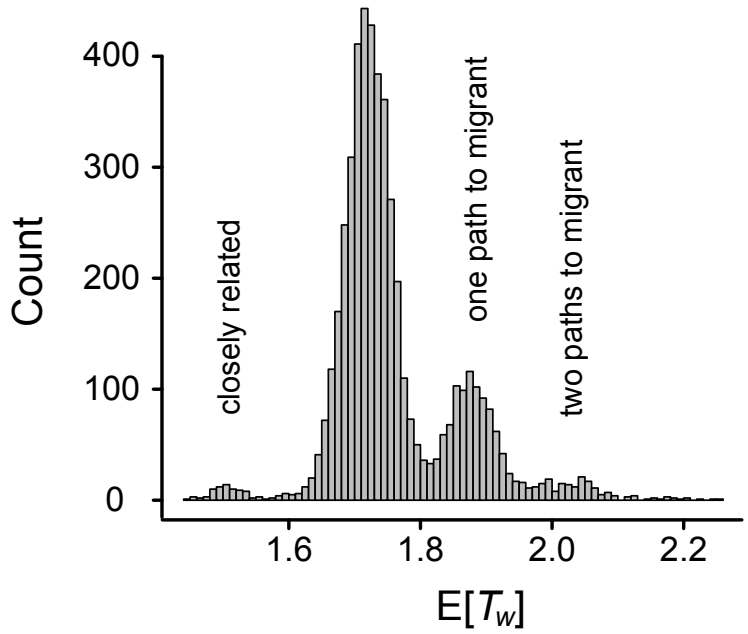
   Multi-locus genetic data might show evidence of the actual series of migration events.

   Single-locus measures of variation might have higher variances than standard migration models predict.

Distributions of F_{ST} and $E[T_w]$ under the Island Model

Two demes of size $N = 500$ with symmetric migration, $Nm = 1/4$





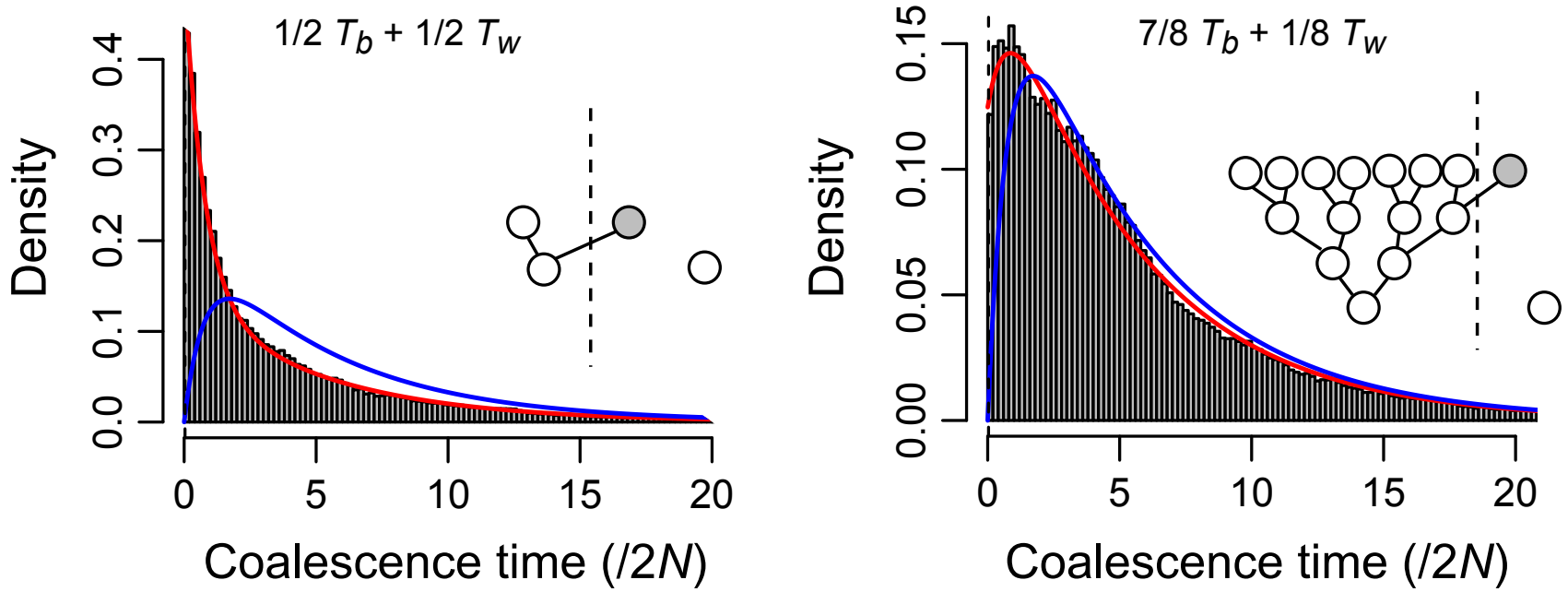
Left: F_{ST} from coalescence times for 1000 pedigrees/replicates.

Right: $E[T_w]$ among pairs of individuals on a single pedigree.

Mixture Distributions of Inter-Deme Coalescence Times

Two demes of size $N = 500$ with symmetric migration, $Nm = 1/4$.

The sampled individual from one deme has recent migrant ancestry.



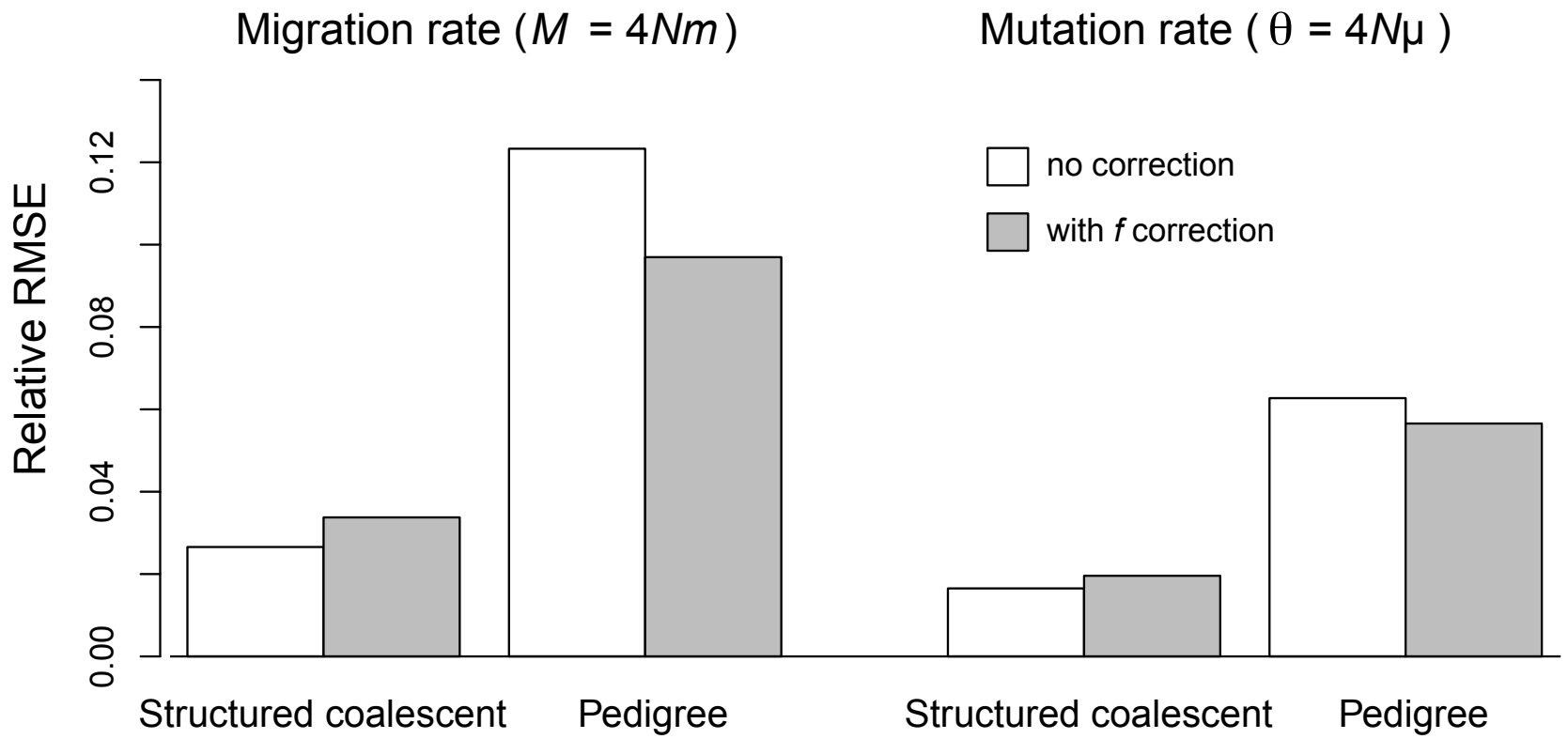
Blue: Expected distributions of between-deme coalescence times.

Red: Mixed distributions, with proportions given by the equations.

Inference in Structured Populations

Two demes of size $N = 250$, with $M = 1/4$ and $\theta = 1/2$.

ML estimates from pairwise differences at $10^4 + 10^4$ loci.



Relative root mean square error over 1000 estimates.

(thanks!)