Gene genealogies within organismal pedigrees, and the robustness of Kingman's coalescent

John Wakeley*, Léandra King*, Peter Wilton*,
Bobbi S. Low†, Sohini Ramachandran‡

*Harvard University, †University of Michigan, ‡Brown University

CIRM June 11, 2012

Preview

Introduction

- Randomness in population genetics
- Dual models: diffusion and coalescence
- The population pedigree

Pedigree-coalescent results

- Wright-Fisher and Swedish-family simulations
- Power to reject the Kingman coalescent
- Per-generation probabilities of coalescence

Conclusions and future work

Sources of Randomness in Population Genetics

1. Survival, Movement, Finding a Mate, Reproduction

   Complicated: the fundamental parameters and rules are not well known. 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, $\sim 10^{-8}$/site/generation in humans.

Questions of This Talk

- General: To what extent can Mendelian inheritance alone account for the randomness in times to common ancestry that is observed among loci within genomes?

- New and preliminary:

  What are the effects of very large families or selective sweeps on coalescence times, as mediated by pedigrees?

  Are standard coalescent predictions for two loci with recombination accurate when the pedigree is fixed?

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

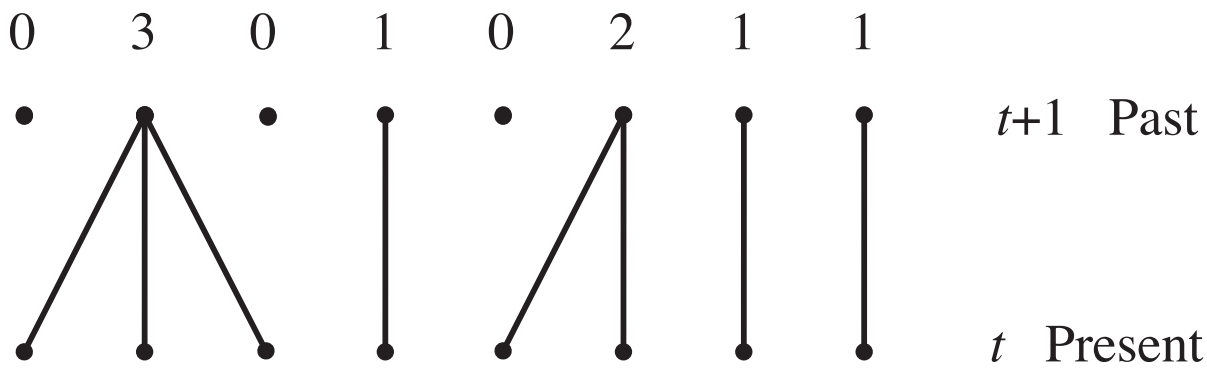
We combine Mendelian inheritance with assumptions about survival, reproduction, etc., to model the dynamics of genetic variation over time. Simple models typically assume:

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.

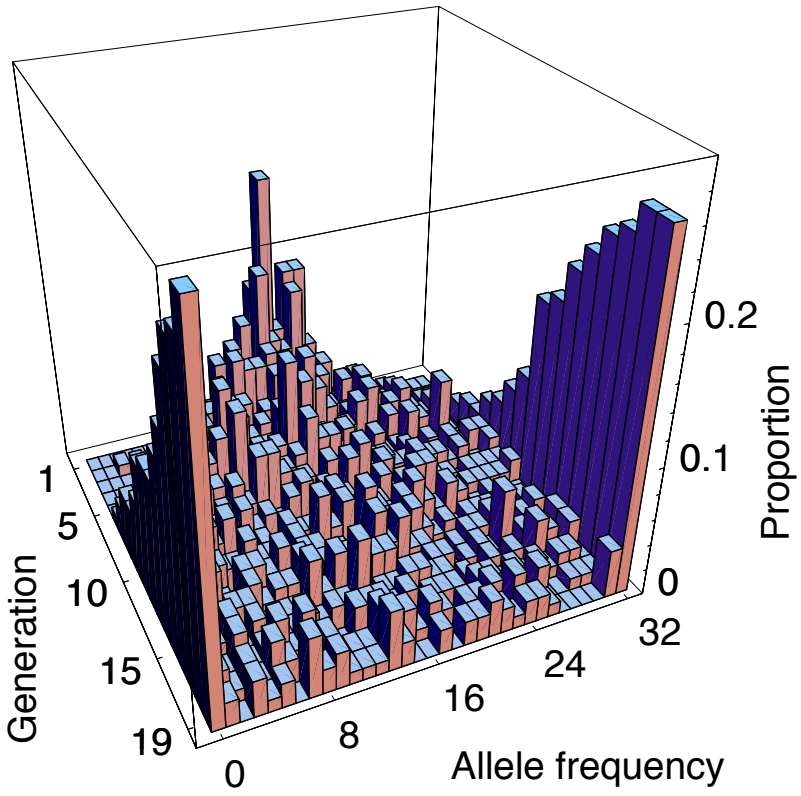
Exchangeable Population Models of Cannings (1974)

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 $\sum_{i=1}^{2N} \nu_i = 2N$.

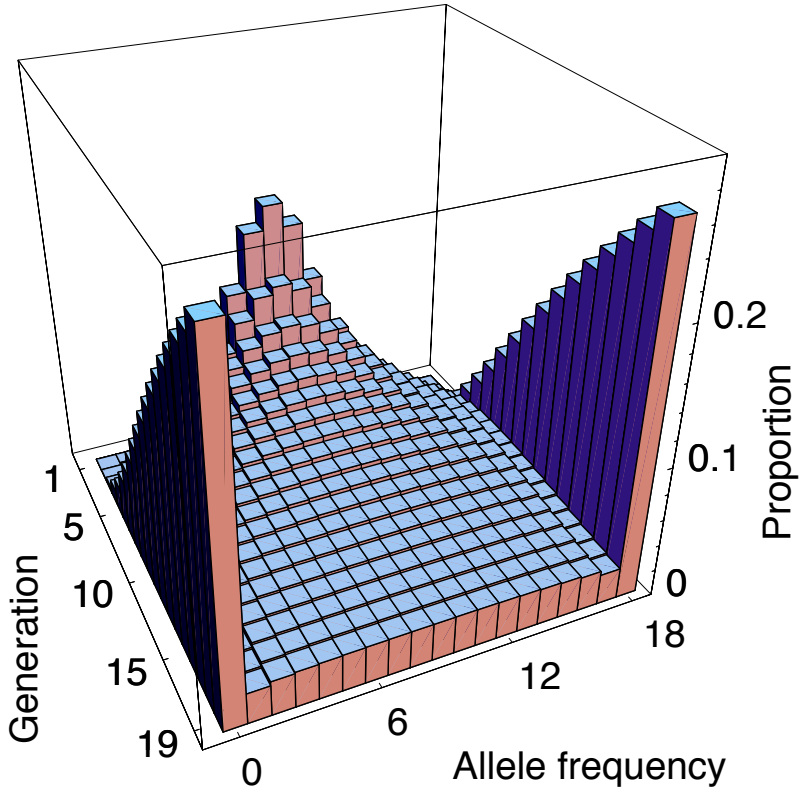


The Wright-Fisher model (of diploid hermaphrodites) can be reduced to a (haploid) Cannings model.

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 $N = 9$.

Application of Coalescent Theory to Genetic Data

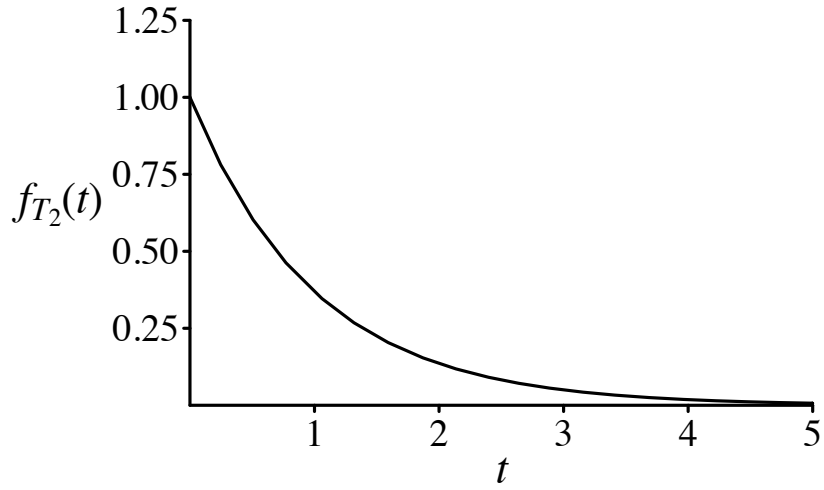
Kingman's coalescent has been used to explain variation in levels of polymorphism among loci within genomes.

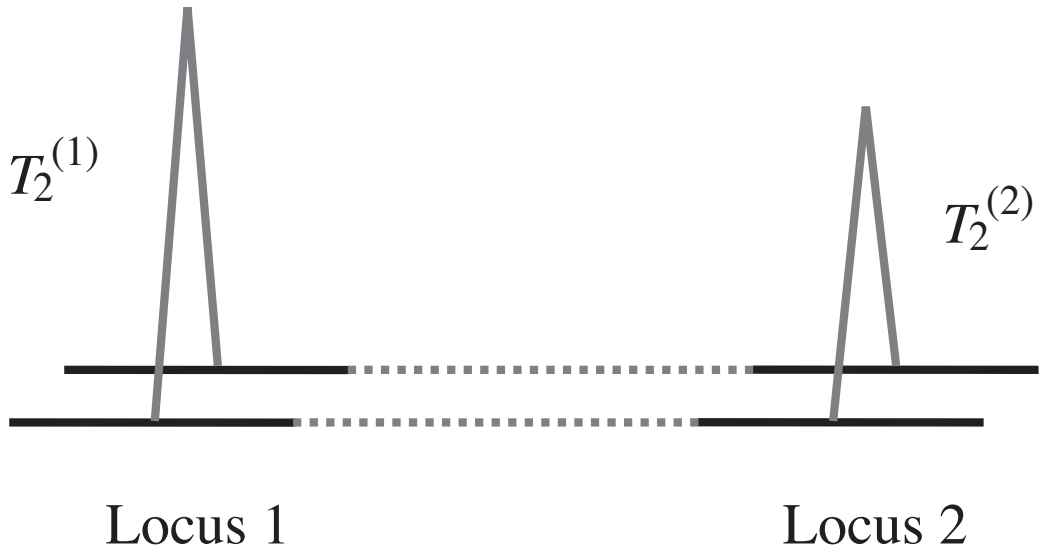
# SNPs	Poisson	Coalescent	Observed
0	8256	8767	8796
1	3040	2332	2247
2	617	663	668
3	99	200	214
4	16	66	102

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, there is 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

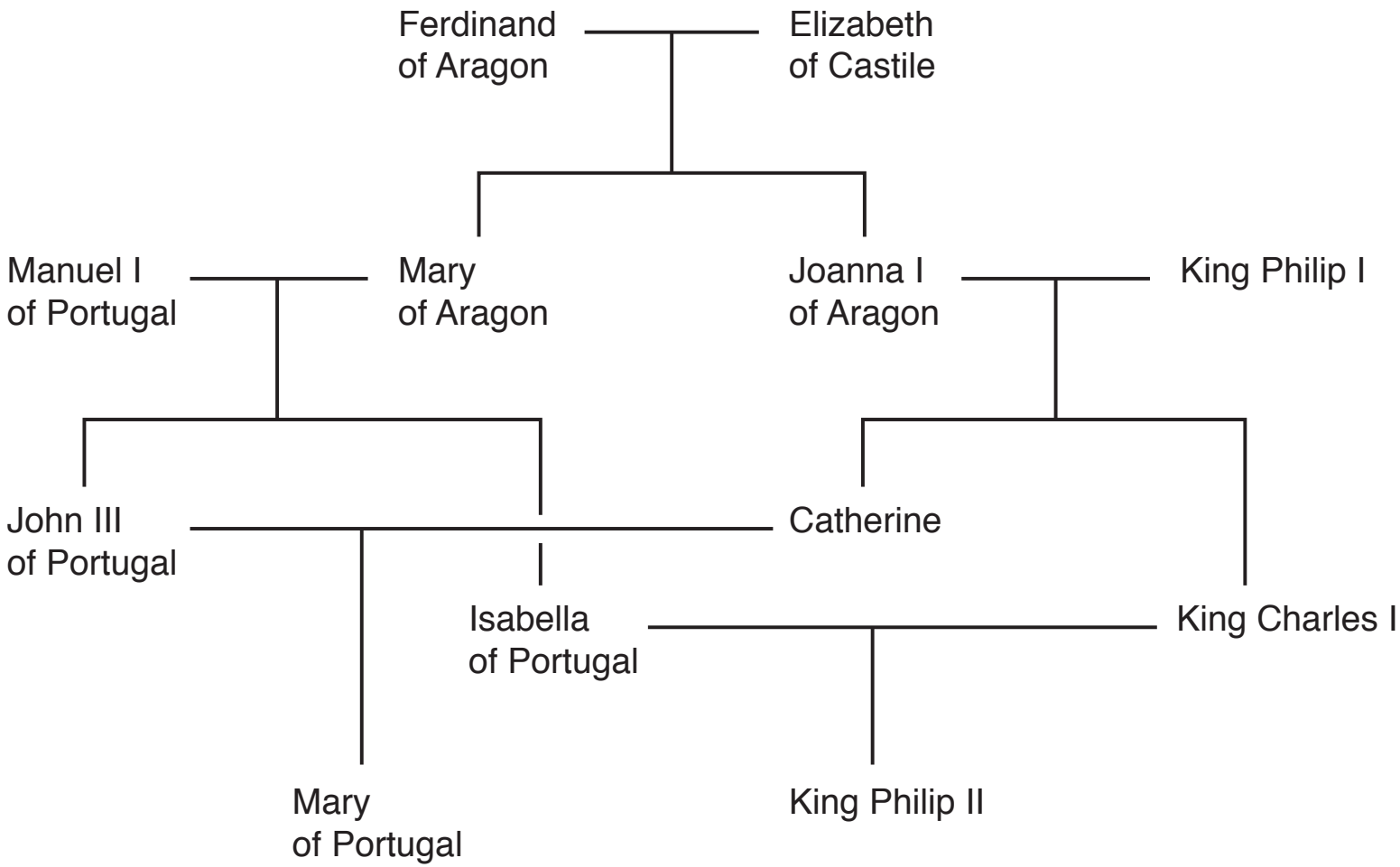
Kingman obtained the coalescent in the limit $N \to \infty$, from a time-homogeneous Markov model (subset of Cannings' model).

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

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

As an aside, note that here $N_e = N/\sigma^2$, where σ^2 is $\text{Var}[\nu_1]$.

A Small Piece of the Human Population Pedigree



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

Pedigree-Coalescent Simulations

1. Three different pedigree simulations based on the two-sex Wright-Fisher model.

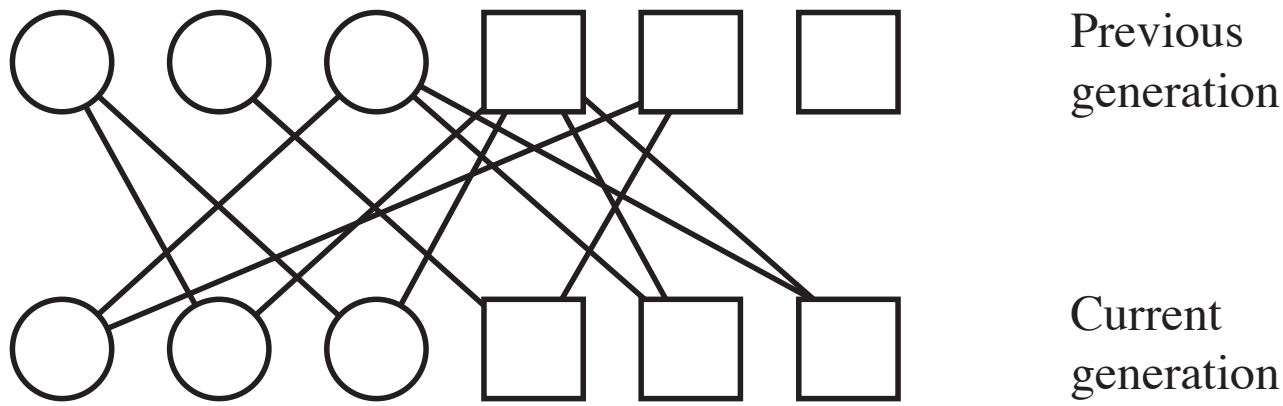
2. Pedigrees generated using 19th-century data from seven parishes in Sweden:

   All men married between 1824 and 1840 and all of their descendants, and all spouses, up to 1896.

   512 extended families were chopped into 1884 two-generation families containing a total of 4451 parents and 7889 offspring.

Simulations of Diploid, Two-Sex Reproduction

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



1. Generate the pedigree of the population ($\times 30N$ generations).
2. Sample two individuals and follow gene copies back through the pedigree (they go 50:50 to mother or father).

Simulations of Coalescence Times ($n = 2$) for 1000 Loci

"WF ind pairs"

1000 coalescence times for 1000 pairs of individuals.

"WF same pair"

1000 coalescence times for a single pair of individuals.

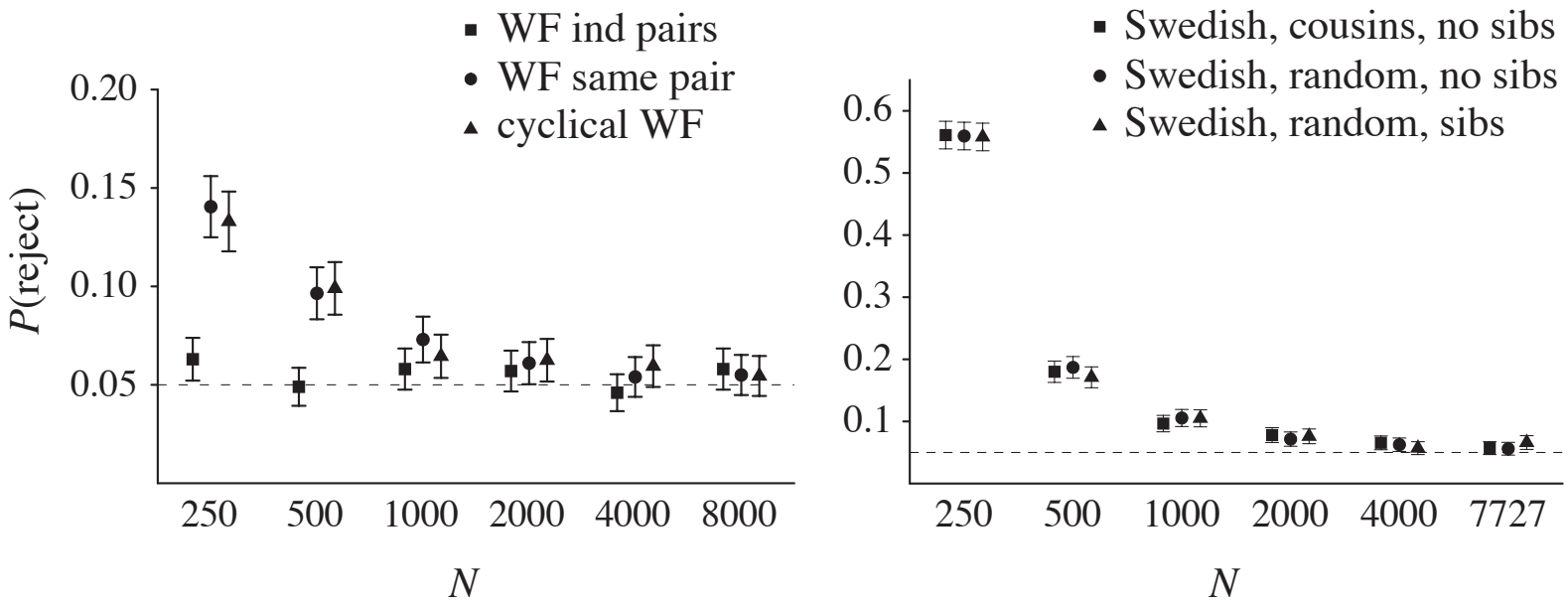
"cyclical WF"

Same as "WF same pair" but using a single realization of a one-generation pedigree $30N$ times.

"SF same pair"

Like "WF same pair" but using Swedish families, and attempting to account for (no) sib and cousin mating.

The Power to Reject the Kingman Coalescent



H_0: The pairwise coalescence times for the 1000 simulated loci are i.i.d. exponential random variables. (χ^2 test with $\alpha = 0.05$)

P(reject) is the proportion of pedigree-samples (out of 2000 total) for which the null model is rejected.

Simulations of Coalescence-Probability Distributions

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

2. Sample one pair of individuals.

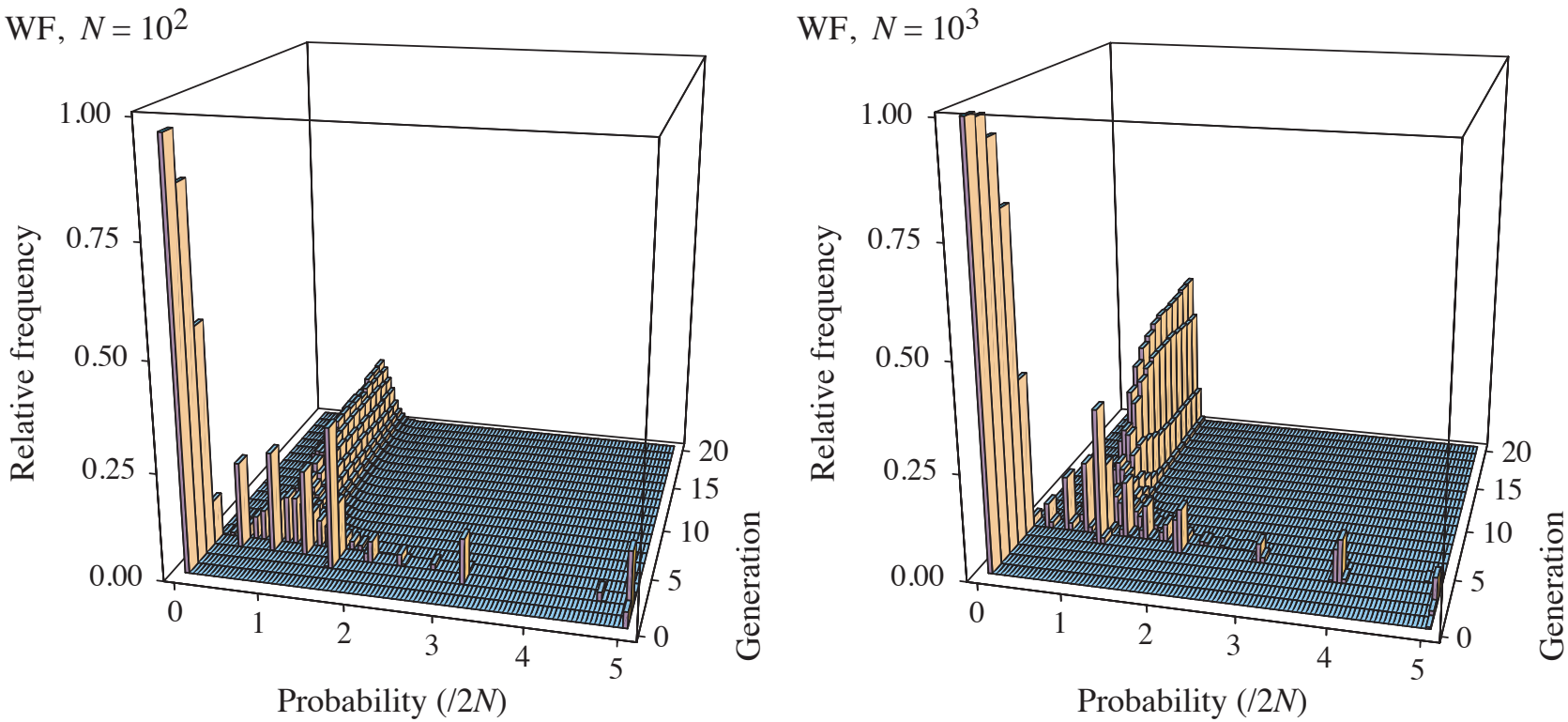
3. For each pedigree and sample-pair (1 & 2):

   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.

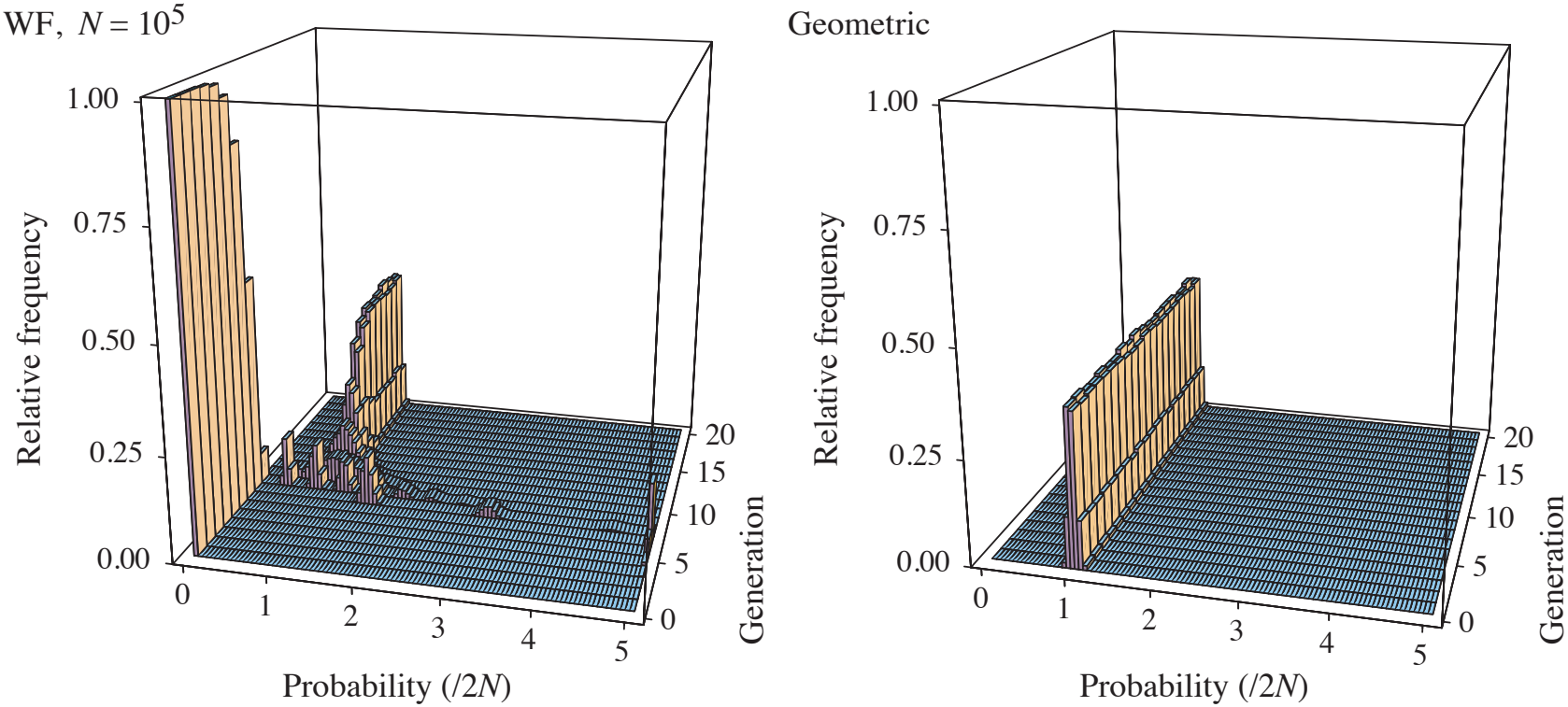
Population Size and Coalescence-Probability Distributions



The probability of coalescence is highly variable in the recent past.

After about $\log_2(2N)$ generations it becomes nearly homogeneous.

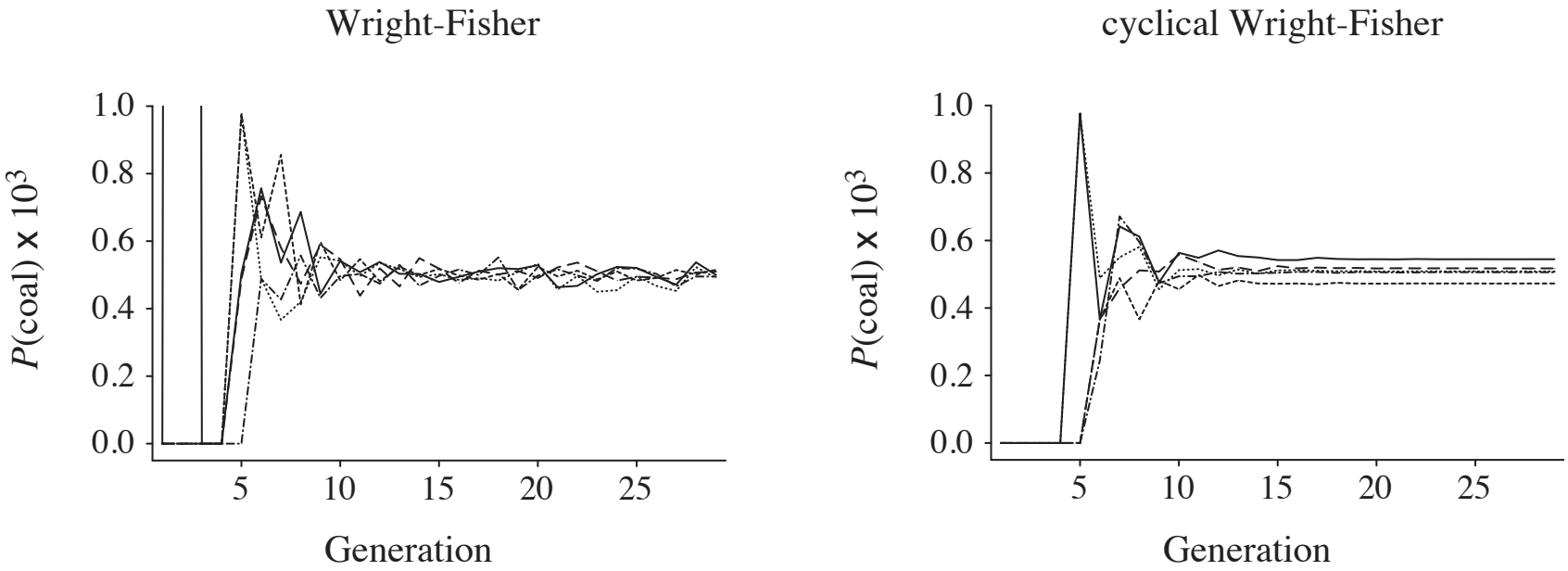
Pedigree-Coalescence Times v. Geometric Times



Left: Pairwise times on Wright-Fisher pedigrees with $N = 10^5$.

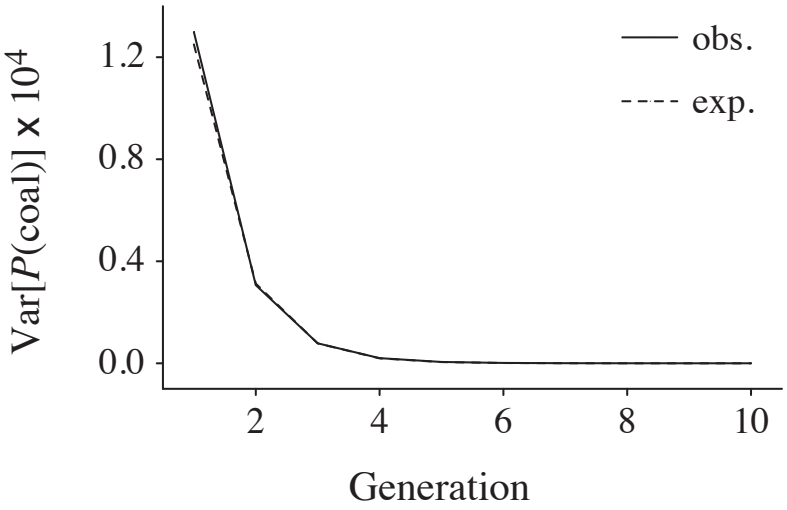
Right: Times from a geometric distribution with $p = 1/(2 \times 10^5)$.

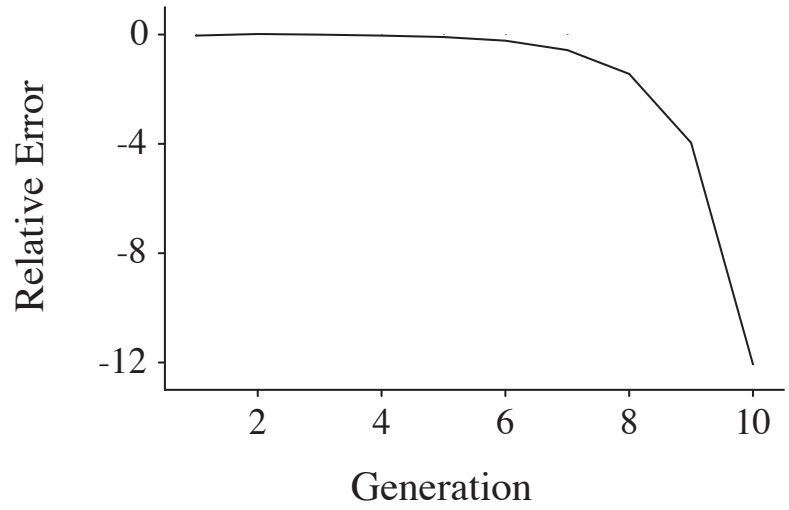
Coalescence-Probability Trajectories for 5 Single Pedigrees



Within a single pedigree, the coalescence probability jumps wildly in the initial $\sim \log_2(2N)$ generations, then settles near $1/(2N_e)$ in the more distant past.

Modeling Variation in Coalescence-Probabilities





$$P(\text{one shared ancestor at } g) \approx \frac{2^{2g}}{N}$$

$$P(\text{coal} \mid \text{one shared ancestor at } g) = \frac{1}{2^{2g+1}}$$

$$E[P(\text{coal})^2 \text{ at } g] \approx \frac{2^{2g}}{N}\left(\frac{1}{2^{2g+1}}\right)^2 = \frac{1}{2N}\frac{1}{2^{2g+1}}$$

Conclusions

1. Kingman's coalescent does a surprisingly good job at predicting the distribution of coalescence times among independent loci on a variety of kinds of fixed pedigrees.

2. Even so, a closer look at distributions of coalescence times uncovers structure, in the most recent $\sim \log_2 2N$ generations, that is inconsistent with the coalescent.

3. The population pedigree may constrain coalescence times in ways not predicted by standard models.

4. Current and future work:

   More on large families, selection, recombination.

   Migration, bottlenecks, range expansions.

   Understand why the coalescent works so well.