Statistical population genetics

Lecture 1: Genetic drift

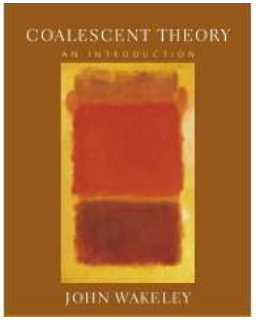
Xavier Didelot
Dept of Statistics, Univ of Oxford
didelot@stats.ox.ac.uk

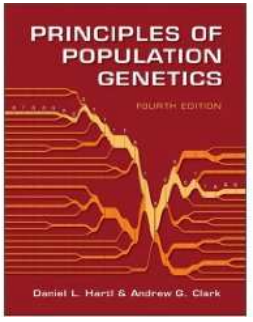
Structure

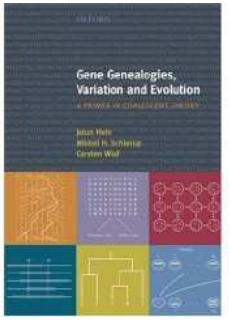
- Lecture 1: Genetic drift
- Lecture 2: Wright-Fisher model
- Lecture 3: Moran model
- Lecture 4: Derivation of the coalescent
- Lecture 5: Properties of the coalescent
- Lecture 6: Coalescent with mutations
- Lecture 7: Infinite alleles model
- Lecture 8: Infinite sites model

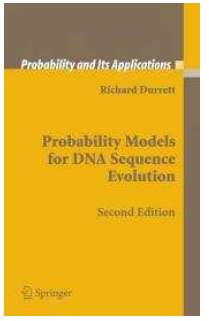
Supporting literature

- *Coalescent Theory: An Introduction*
  by Wakeley; **Chapters 3-4**
- *Principles of Population Genetics*
  by Hartl and Clark; **Chapter 7**
- *Gene Genealogies, Variation and Evolution*
  by Hein, Schierup and Wiuf; **Chapters 1-2**
- *Probabilistic Models for DNA Sequence Evolution*
  by Durrett; **Chapter 1**





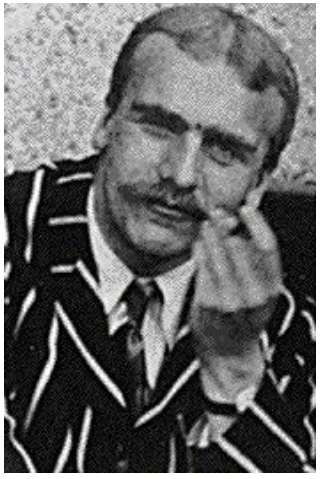




Historical overview

- Darwin (1859): evolution of **traits**
- Mendel (1866): traits are determined by **genes**
- Modern evolutionary synthesis (1900): birth of **population genetics**
- Fisher, Wright and Haldane (1930): **stochastic** formalization

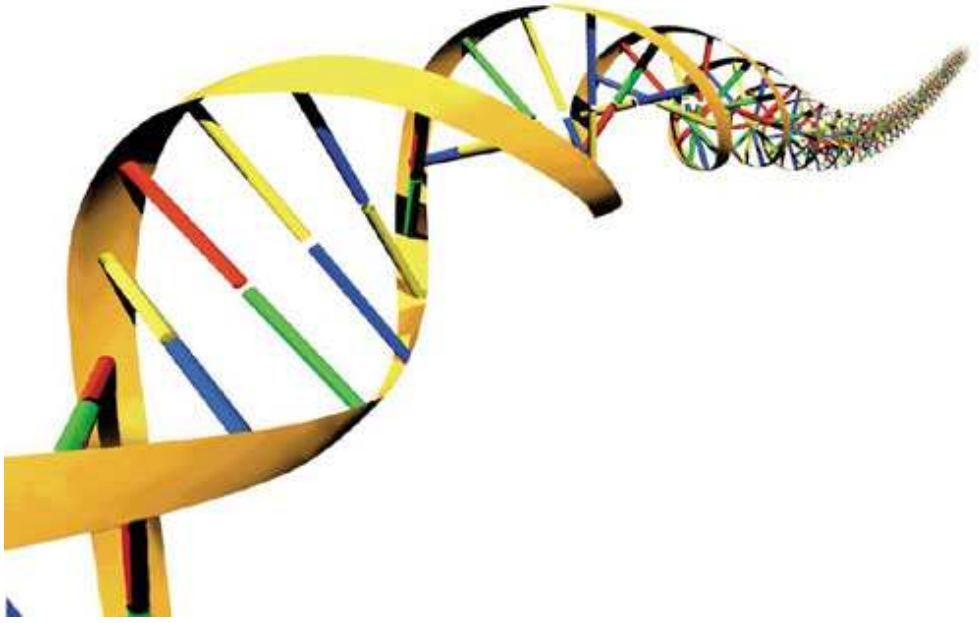




- Kimura (1950): **diffusion** approach
- Watson and Crick (1953): genes are encoded by **DNA**
- Sanger (1975): **DNA sequencing**
- Kingman (1980): **coalescent** approach

Definitions

- The **genome** of an organism is its whole hereditary information and is encoded in the DNA (or, for some viruses, RNA).
- A given fraction of a genome is called a **gene**.
- A gene can take several values, or **alleles**.
- **Population genetics** is the study of the forces that produce and maintain genetic variation within a population (eg. a species).



Definitions

- We will start with two such forces: **random drift** and **mutation**.
- **Random drift** is the process by which allele frequencies vary in the population following the birth and death of organisms.
- **Mutation** is the process by which a gene occasionally changes from one allele to another.
- Other forces include **natural selection**, **demographic stochasticity**, **recombination**, etc.

Wright-Fisher model

This model, introduced by Fisher (1930) and Wright (1931), is the most widely used model in population genetics.

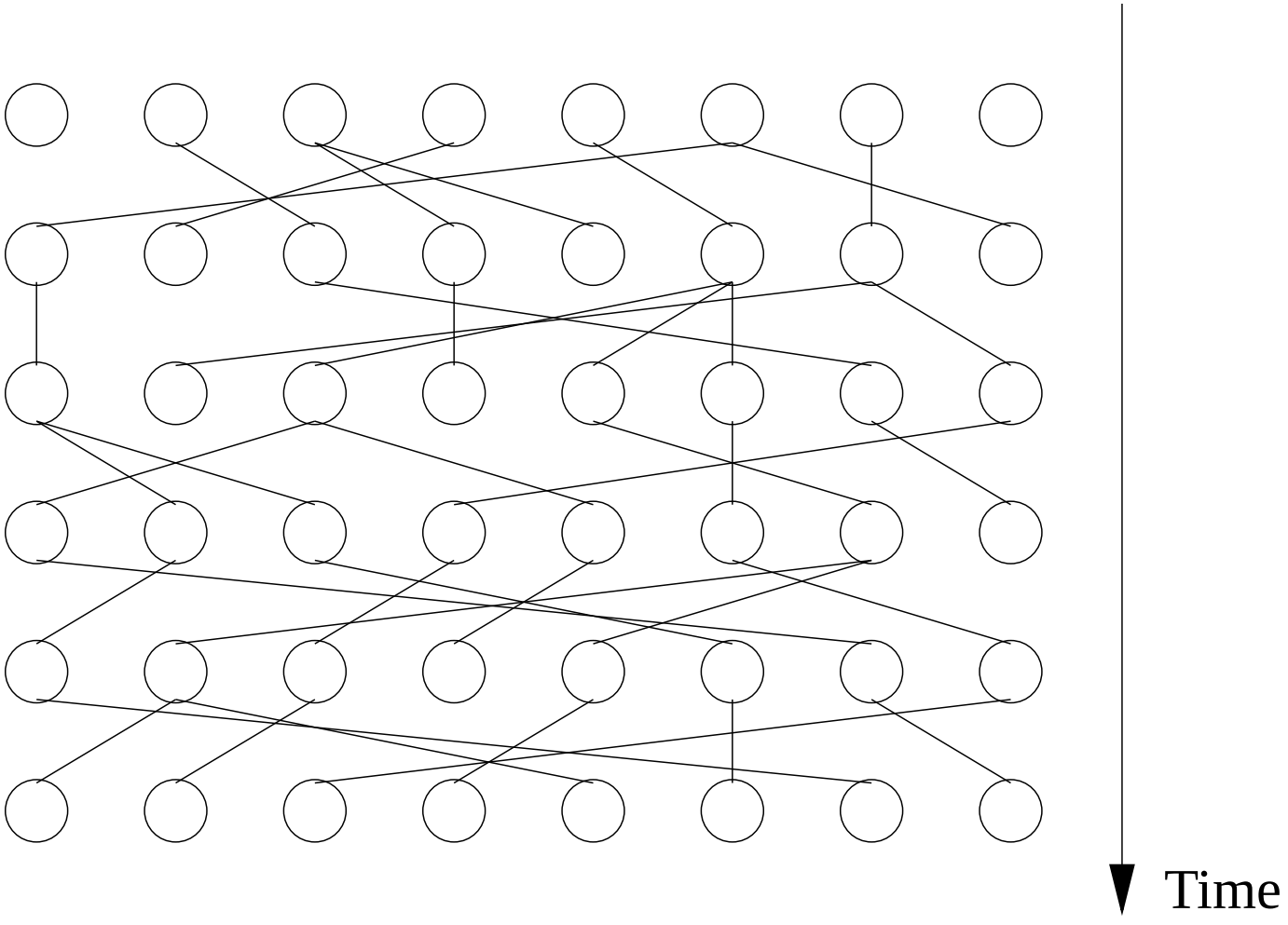
Definition (Wright-Fisher model)**.**
A population evolves according to the Wright-Fisher model if:

(i) *The population has a* ***constant size*** M*;*

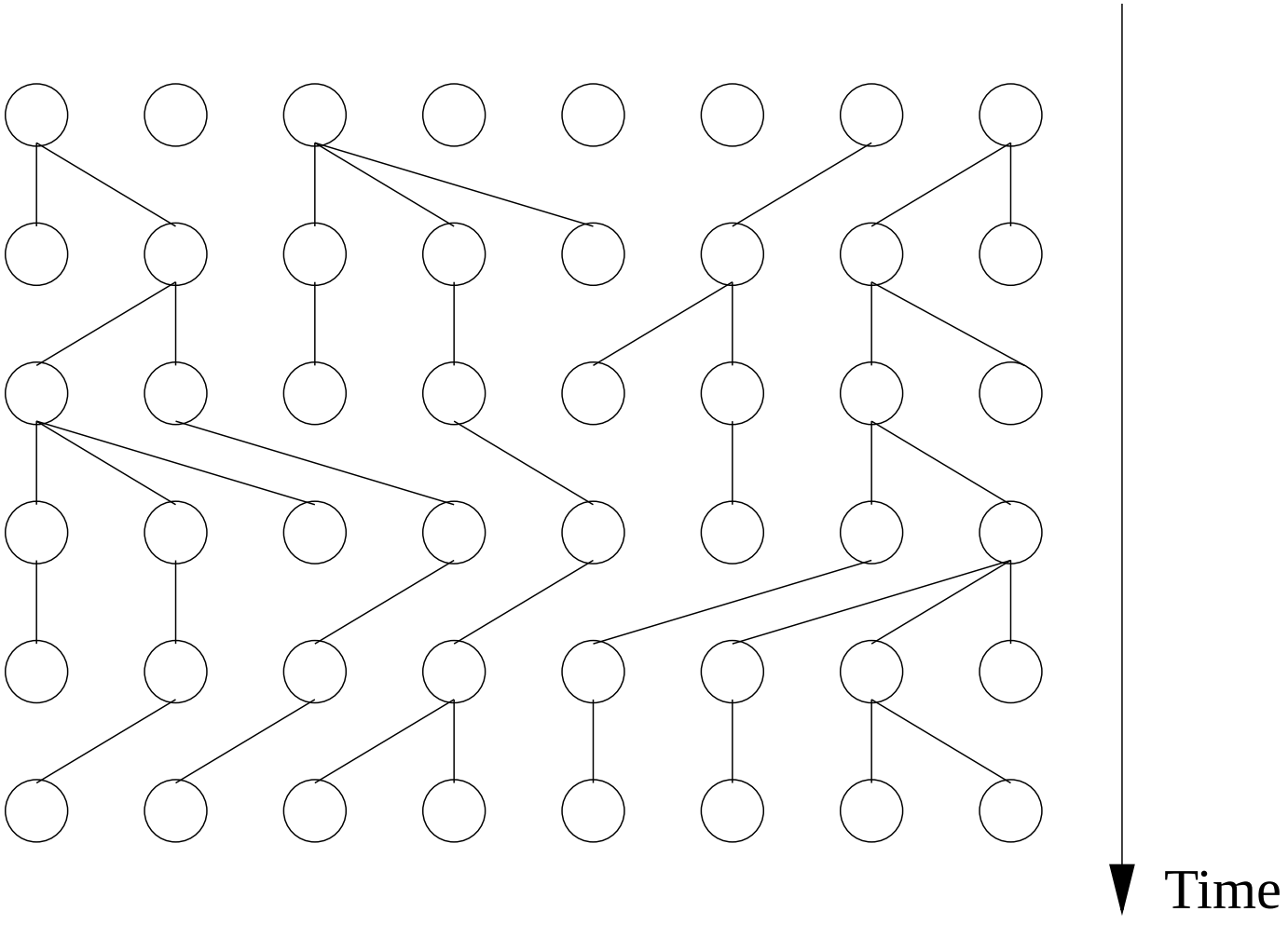
(ii) *Generations are* ***non-overlapping****;*

(iii) *Generation* $t+1$ *is formed from generation* t *by* ***uniformly sampling with replacement****.*

Wright-Fisher model



Wright-Fisher model



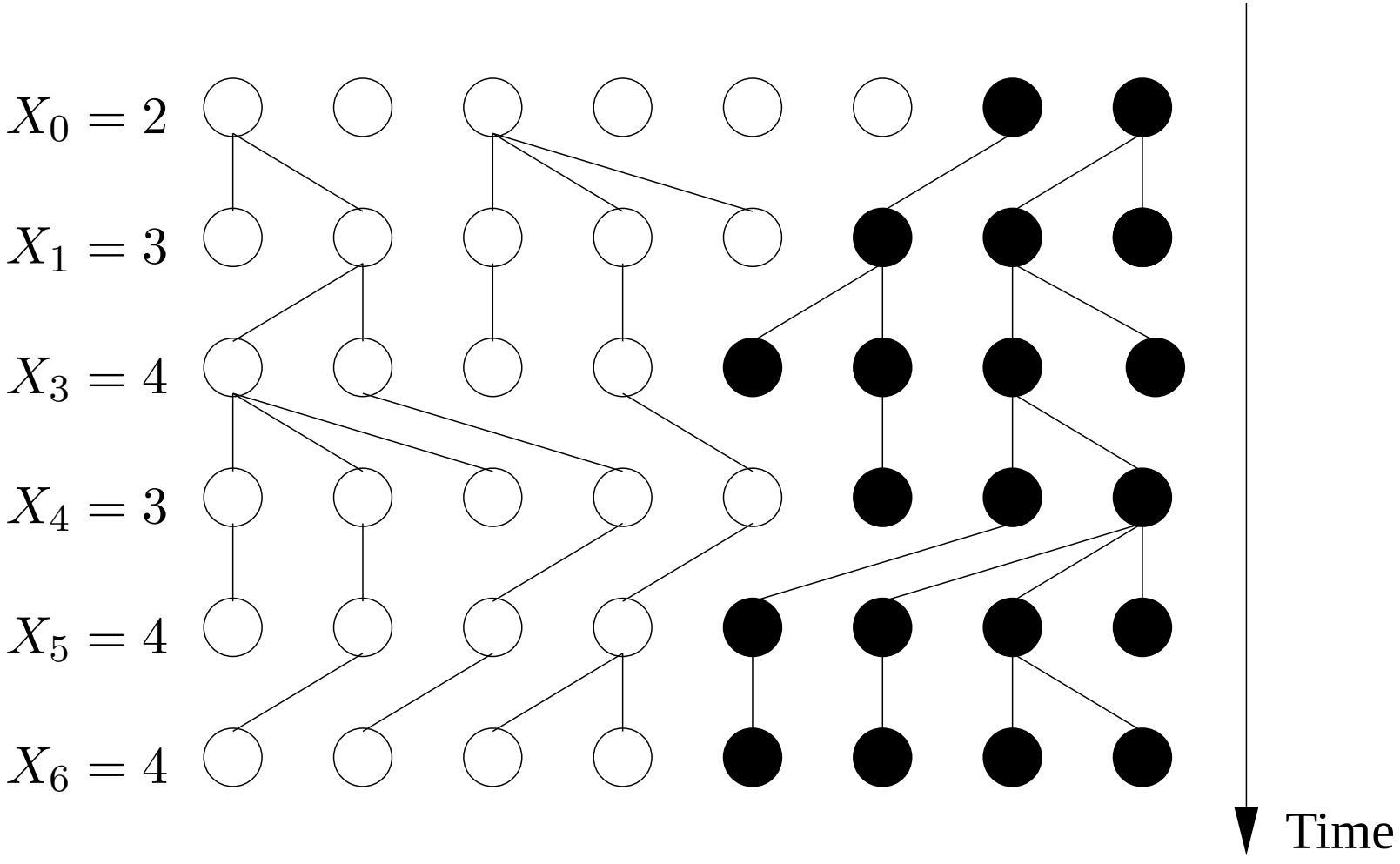
Biallelic WF model

- Assume that there are **two alleles** A and a.
- Let X_t denote the number of genes with allele A in generation t.
- It follows from the definition of the Wright-Fisher model that:

$$\mathbb{P}(X_{t+1} = j | X_t = i) = P_{i,j} = \left(\begin{array}{c} M \\ j \end{array} \right) (i/M)^j (1 - i/M)^{M-j}$$

- X_t is a **Markov chain** with transition matrix $P = \{P_{i,j}\}_{i,j \in [1..M]^2}$.
- In principle the entire behavior of X_t can be derived from our knowledge of P and the initial state X_0.
- In practice the matrix P does not lend itself readily to simple answers.

Biallelic WF model



Biallelic WF model

- $X_{t+1}|X_t = i$ is binomial with parameters M and i/M. It follows that:

$$\begin{aligned}\mathbb{E}(X_{t+1}|X_t = i) &= M \cdot i/M = i \\ \text{var}(X_{t+1}|X_t = i) &= M \cdot i/M \cdot (1 - i/M) \\ &= i(1 - i/M)\end{aligned}$$

- Let $p_t = X_t/M$ be the allele frequency. We have:

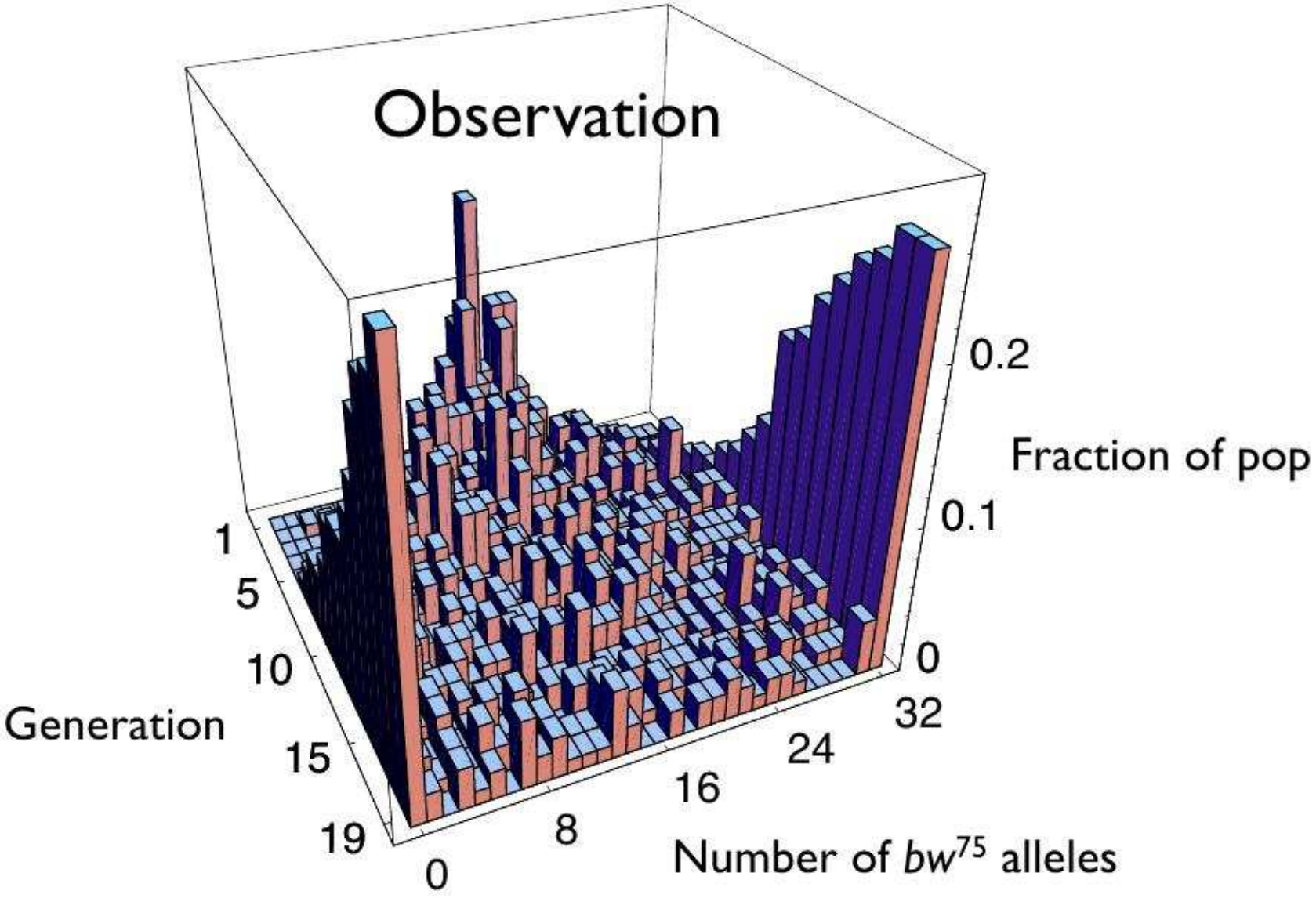
$$\begin{aligned}\mathbb{E}(p_{t+1}|p_t) &= p_t \\ \text{var}(p_{t+1}|p_t) &= p_t(1 - p_t)/M\end{aligned}$$

- The expected allele frequency is **constant**.
- The allele frequency fluctuates **faster in small populations**.

Experiment of Buri

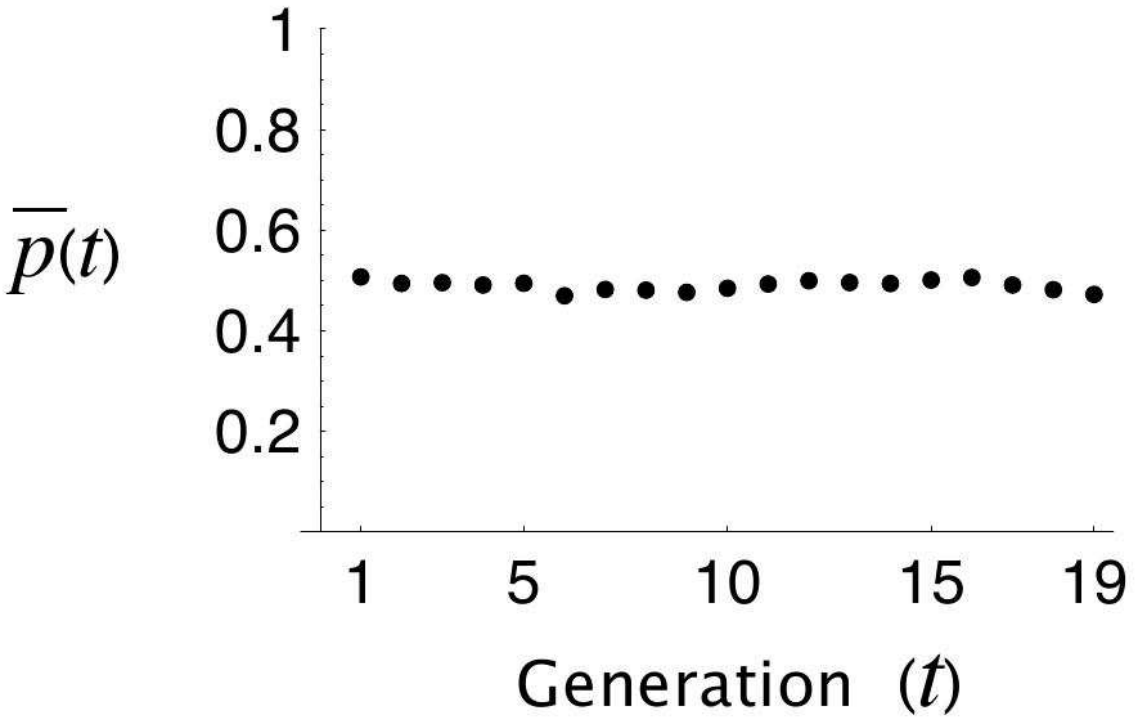
- In 1956, Buri (a student of Wright) followed the evolution of the number of bw^{75} alleles in 107 distinct populations of Drosophila melanogaster (easy to follow using eye color).
- Each population was started with $M = 32$ genes, half of which were of the bw^{75} allele. Population size $M = 32$ was forced to remain constant.
- The experiment was carried over 19 generations.

Experiment of Buri

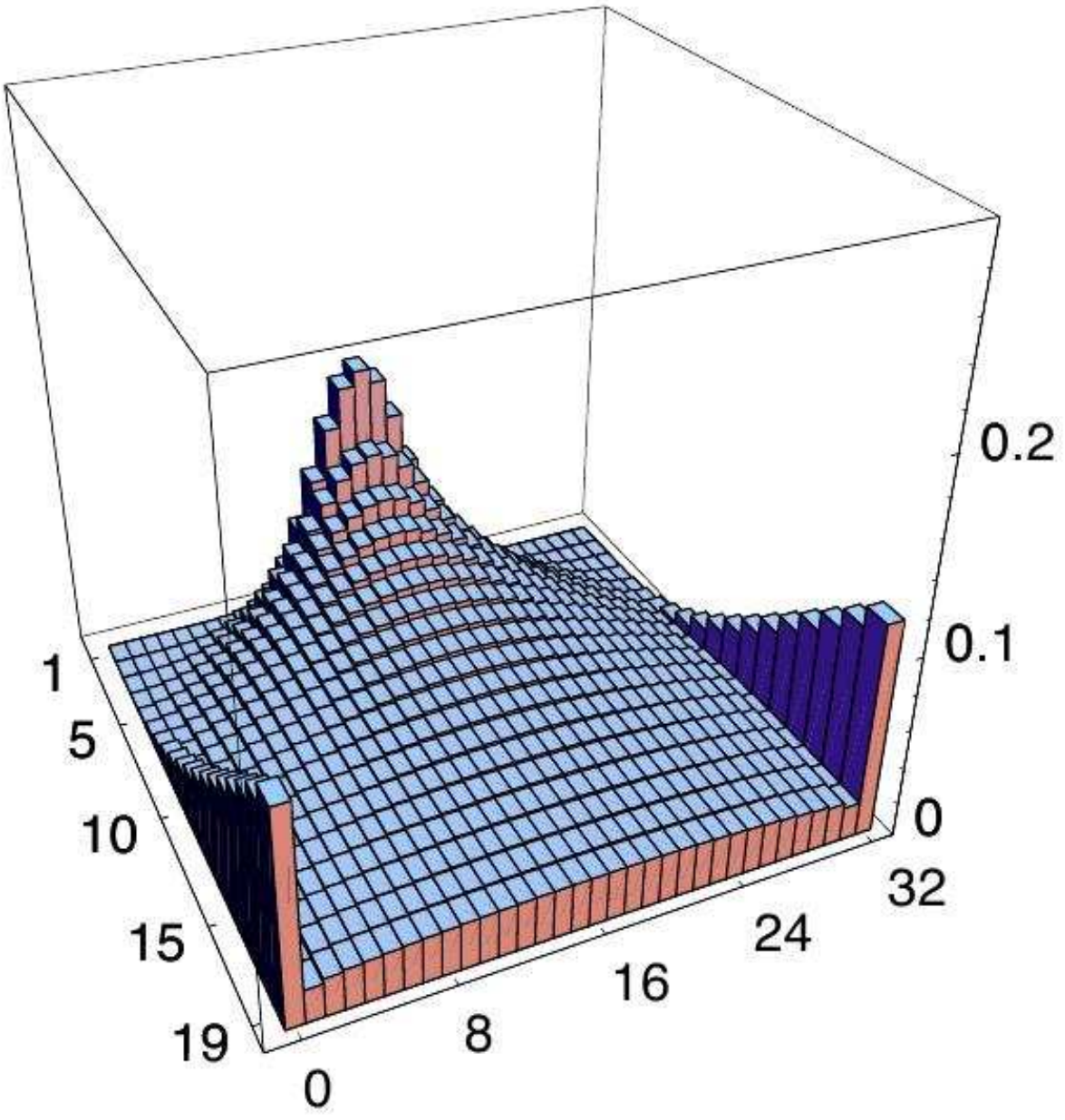


Experiment of Buri

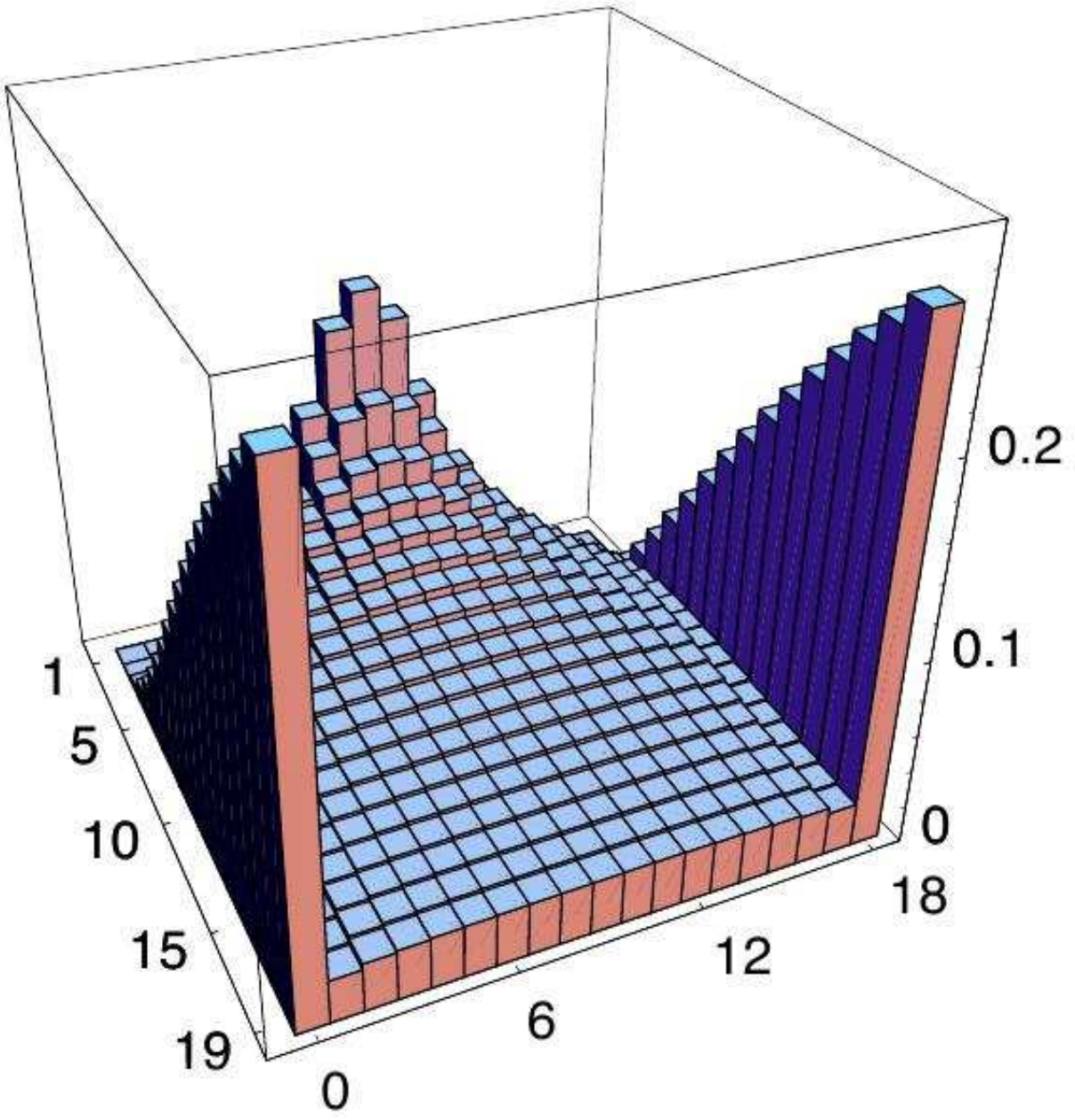
There is apparently no selection on this gene.



Simulation ($M = 32$)



Simulation ($M = 18$)



Effective population size

- Loss of allele happens **faster** in the data than expected with $M = 32$.
- In fact, the data fits closely our expectations for a population of 18 genes rather than 32.
- The number of genes required in the Wright-Fisher model for it to behave like a real population is called the **effective population size** (denoted M_e) of that population.
- The difference between actual and effective population sizes comes from the fact that **real populations do not follow the assumptions** of the Wright-Fisher model.

In bacteria...

- Huge census population sizes. Example: Numbers of Vibrio cells per cubic meter of seawater in temperate coastal regions range from 10^8 to 10^9, suggesting $M > 10^{20}$
- Estimates of M_e for bacteria range from 10^5 to 10^9
- Mismatch of many orders of magnitude between effective population size and census population size
- Many possible explanations
- Fraser et al. (2009). Science. 323:741-746

Summary

- **Population genetics** is the study of the forces that create and maintain genetic variation.
- One such force is **genetic drift**: allele frequencies vary due to birth and death.
- The **Wright-Fisher model** is a simple model of genetic drift.
- It can easily be **extended** to incorporate mutation, natural selection, etc.
- Real populations can often be modelled as a WF population with a given **effective population size**.