

Introduction to Natural Selection

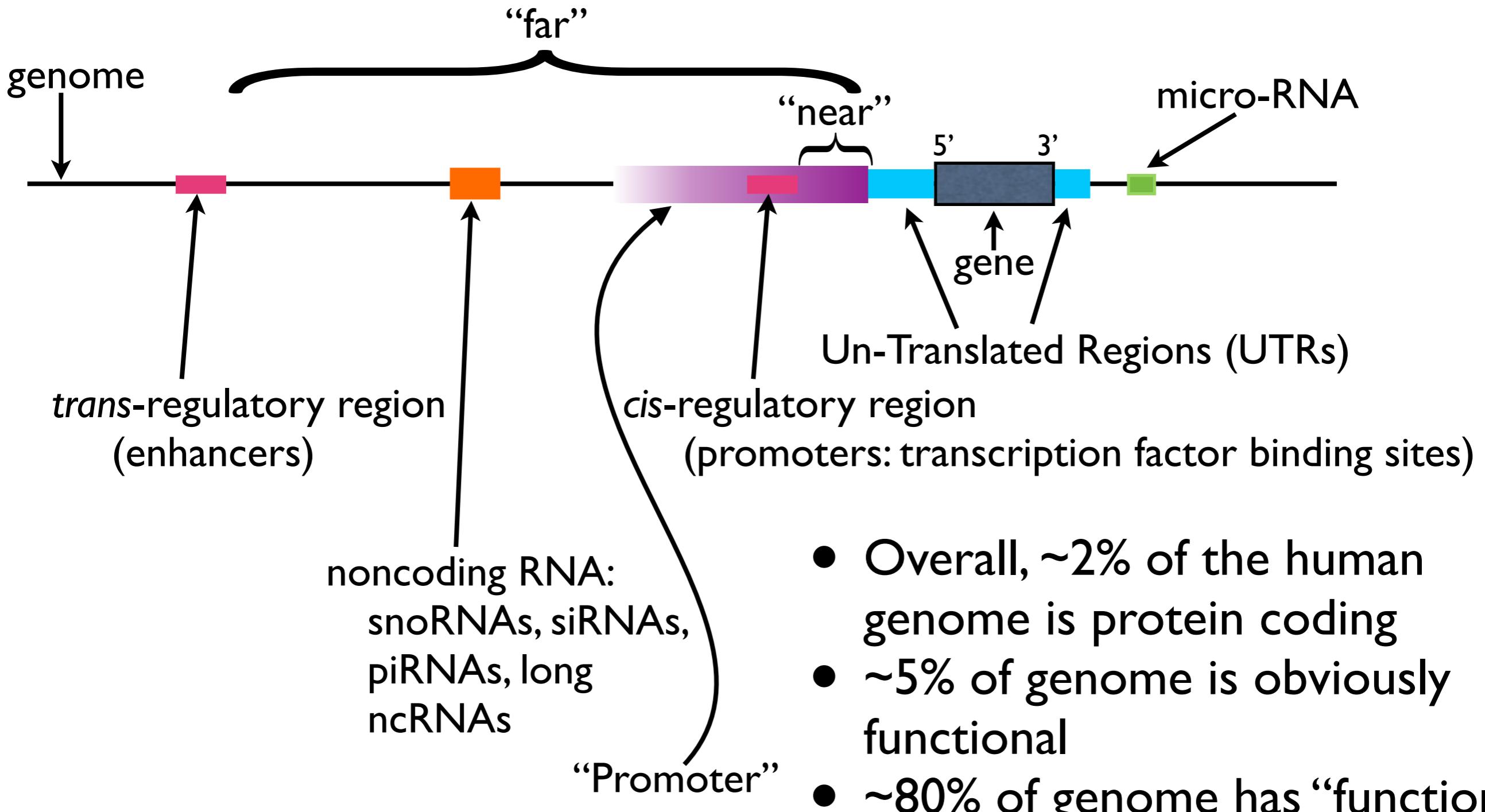
Ryan Hernandez

Goals

- Learn about the population genetics of natural selection
- How to write a simple simulation with natural selection

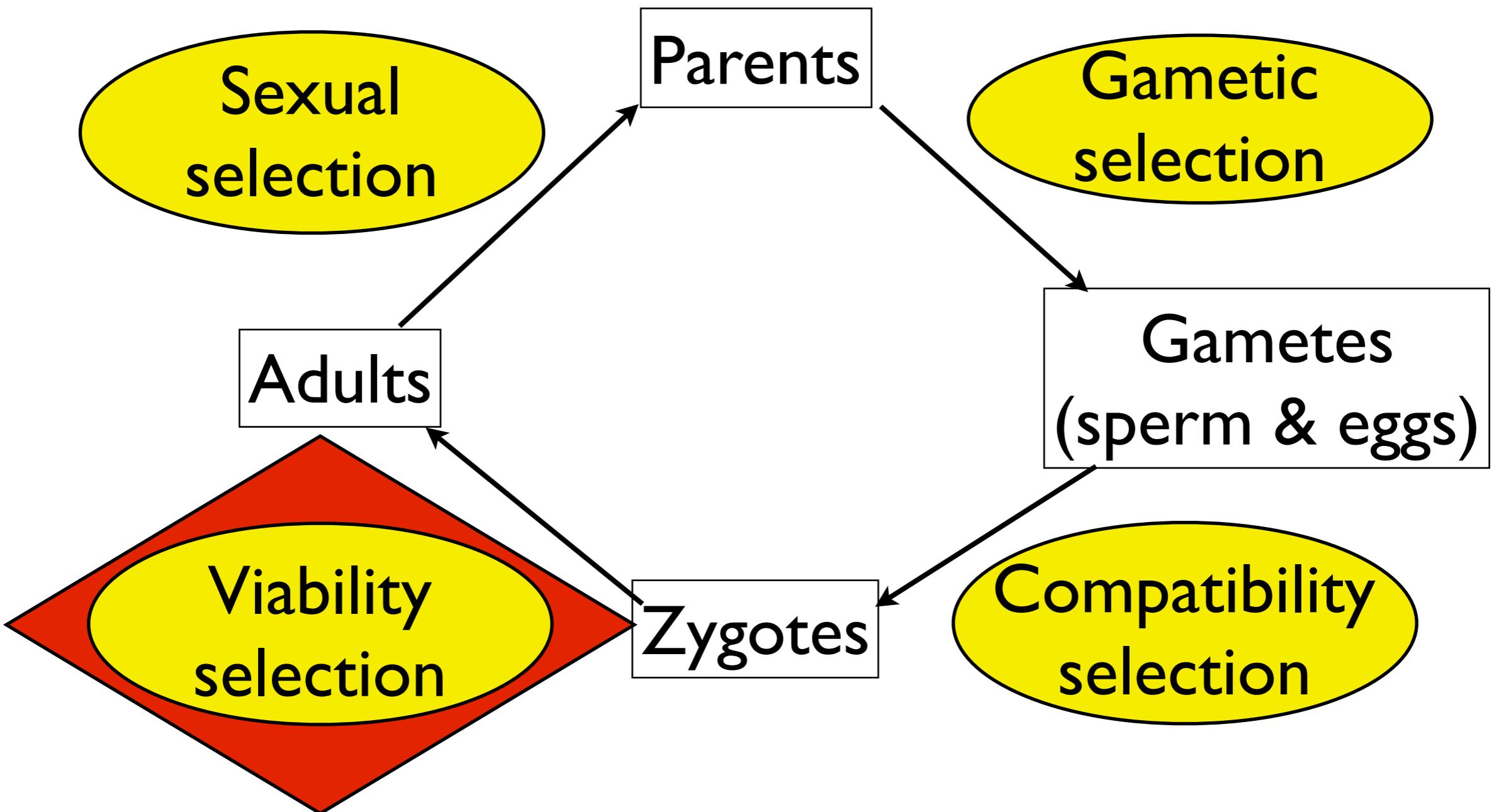
Basic Biology of Human Genome

Functional non-coding mutations



- Overall, ~2% of the human genome is protein coding
- ~5% of genome is obviously functional
- ~80% of genome has “functional activity”

Life Cycle



MODERN HUMAN GENOMICS: A CASE FOR RARE VARIANTS?

$$1.1 \times 10^{-8} \times 6 \times 10^9 = 66 \text{ [muts / person]}$$

$$\begin{array}{r} 66 \text{ [muts/p]} \\ \times 130M \text{ [p/y]} \\ \hline \end{array}$$

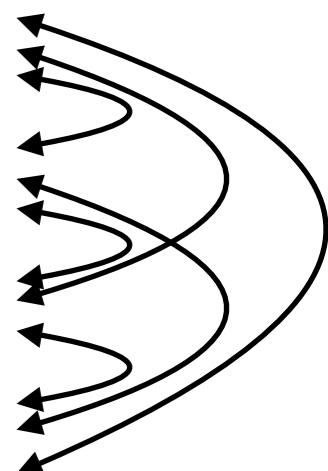
$$\begin{array}{r} \div 3B \text{ [bp]} \\ \hline \end{array}$$

$$\underline{\hspace{1cm}}$$

2.86 muts/bp/yr

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T

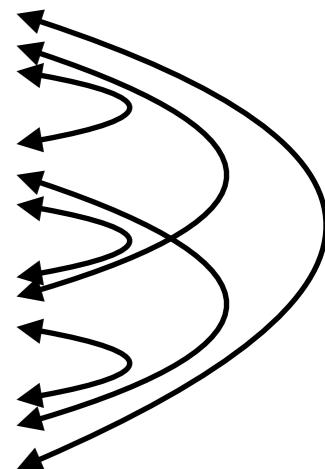


# Pairwise differences	3	4	3	3	3	3
------------------------	---	---	---	---	---	---

π = average pairwise diversity

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T

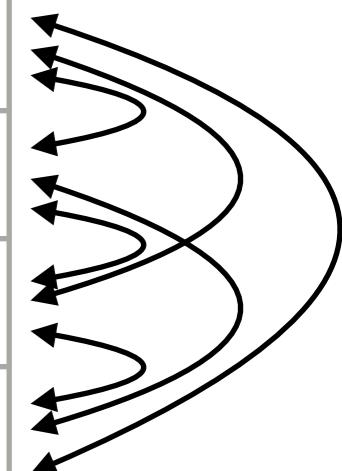


# Pairwise differences	3	4	3	3	3	3
# Compared	6	6	6	6	6	6

π = average pairwise diversity

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T



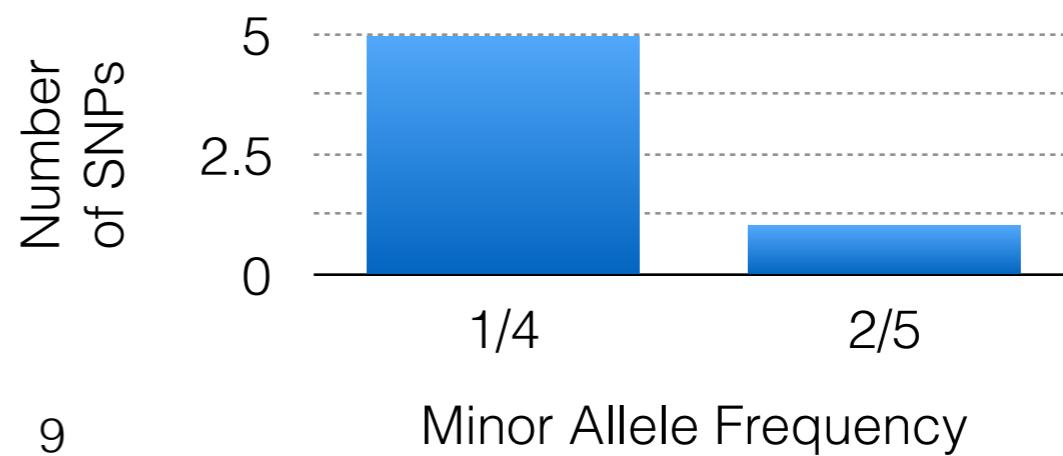
# Pairwise differences	3	4	3	3	3	3
# Compared	6	6	6	6	6	6
Avg. Pairwise Diff	0.5	0.67	0.5	0.5	0.5	0.5

Number of variants: 6 SNPs
 Diversity (π): 3.1667/L

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T
Minor Allele	A	C	A	G	C	T
MAF	0.25	0.5	0.25	0.25	0.25	0.25

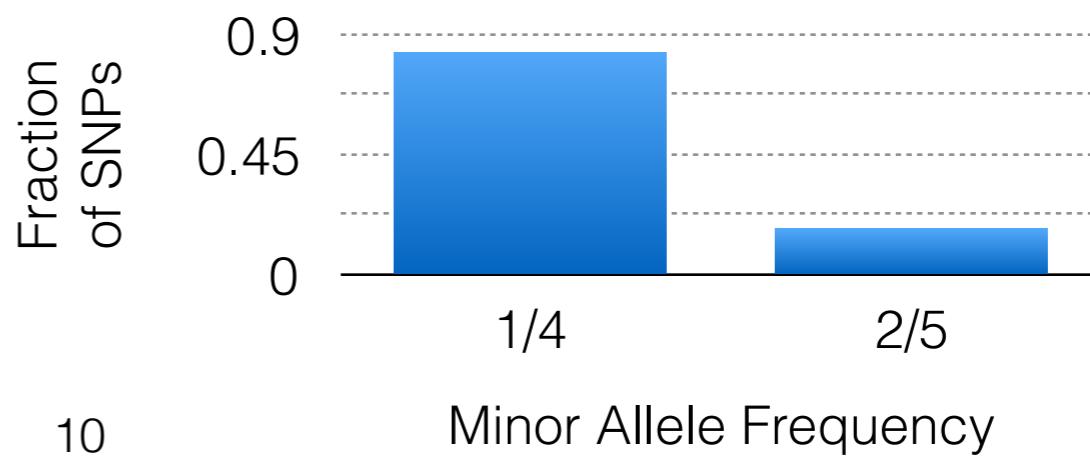
MAF	5	1
-----	---	---



SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T
Minor Allele	A	C	A	G	C	T
MAF	0.25	0.5	0.25	0.25	0.25	0.25

MAF	5	1
-----	---	---



SEQUENCING DATA

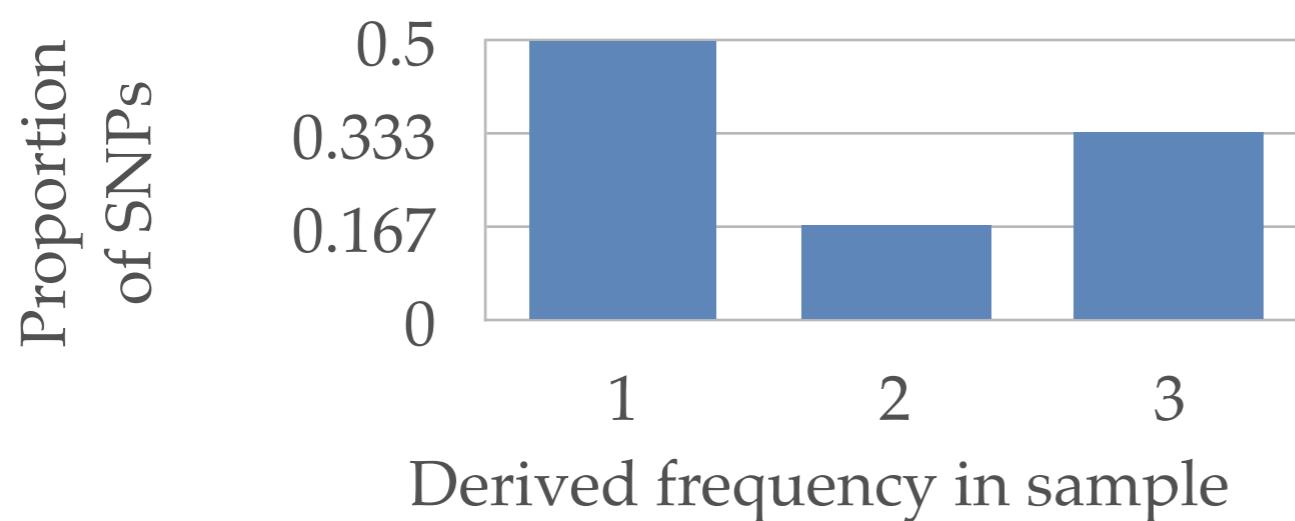
Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T
Chimp	A	C	A	G	C	T

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T
Chimp	A	C	A	G	C	T

SEQUENCING DATA

Chromosome	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
1	A	C	A	G	C	C
2	A	T	G	A	C	T
3	G	T	G	A	T	T
4	A	C	G	A	C	T
Chimp	A	C	A	G	C	T
Derived count	1	2	3	3	1	1

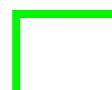


Site-Frequency
Spectrum (SFS)

Site-Frequency Spectrum

	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
1	C	A	T	T	C	G	A	A	G	C	G	A	T	C	A	G	G	C	T	A	T	A
2	C	A	T	T	T	G	A	G	C	G	A	T	C	A	G	G	C	T	A	T	A	
3	C	G	T	T	T	G	A	G	C	G	A	T	T	A	G	G	C	C	A	T	A	
4	C	A	T	T	C	G	A	G	C	G	A	T	C	A	G	G	C	T	A	T	A	
outgroup	T	A	C	C	C	A	G	G	A	G	A	T	A	C	G	C	A	T	T	T	A	

 = non-coding

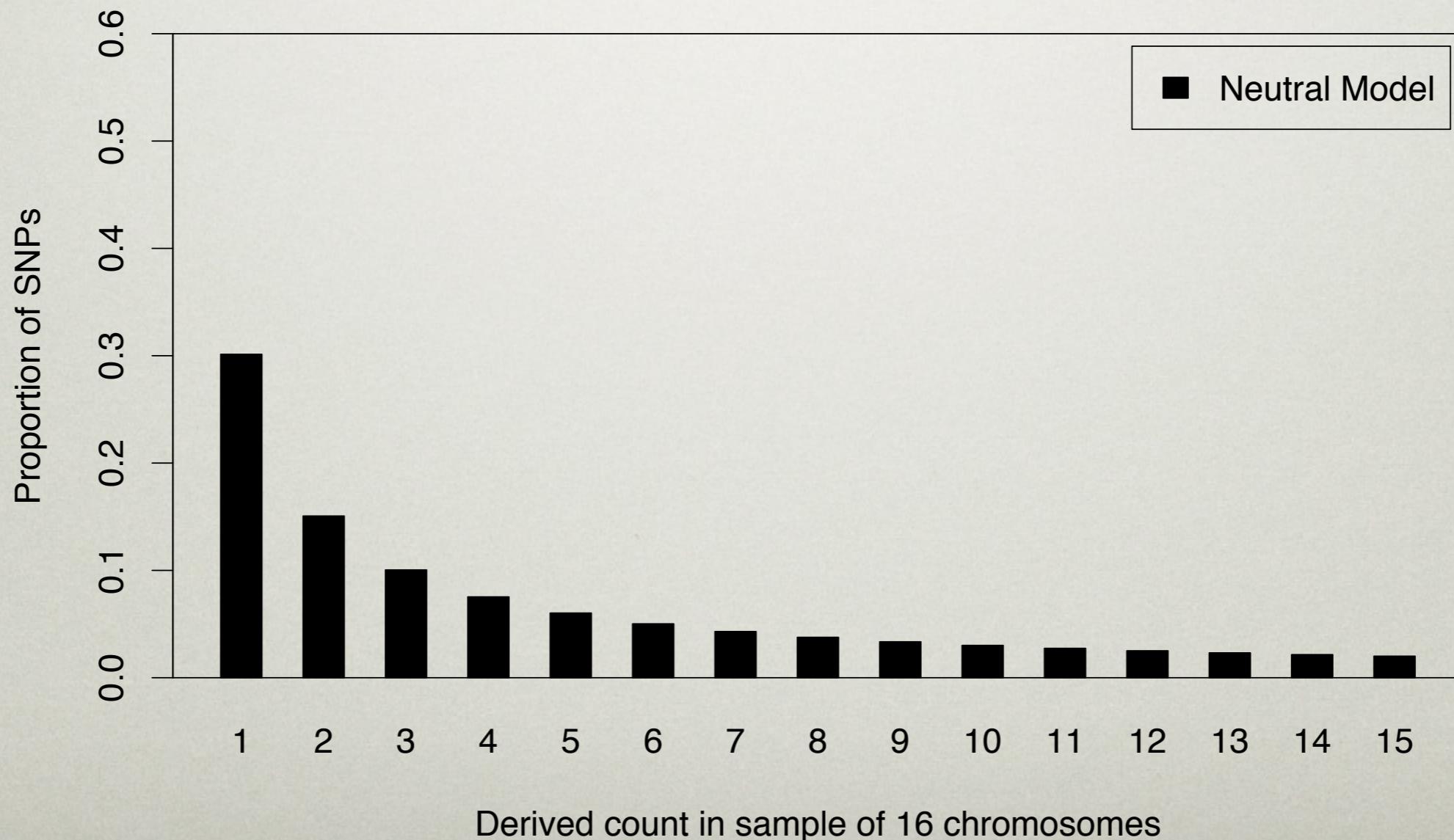
 = synonymous

 = nonsynonymous

* - Substitution between species

SITE-FREQUENCY SPECTRUM

The proportion of derived mutations at each frequency in a sample of chromosomes



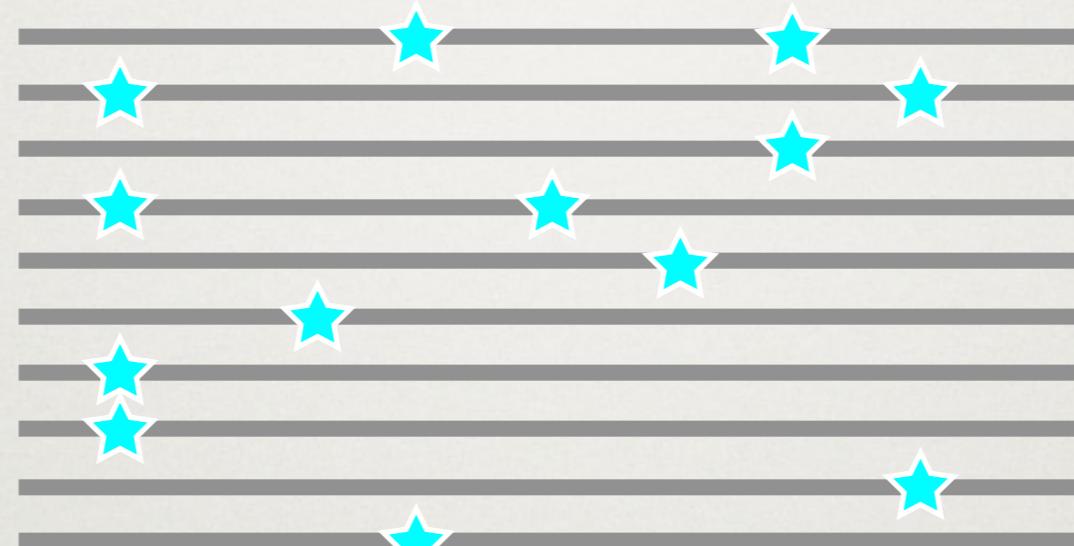
THE EFFECT OF NEGATIVE SELECTION

Chromosomes in
a population

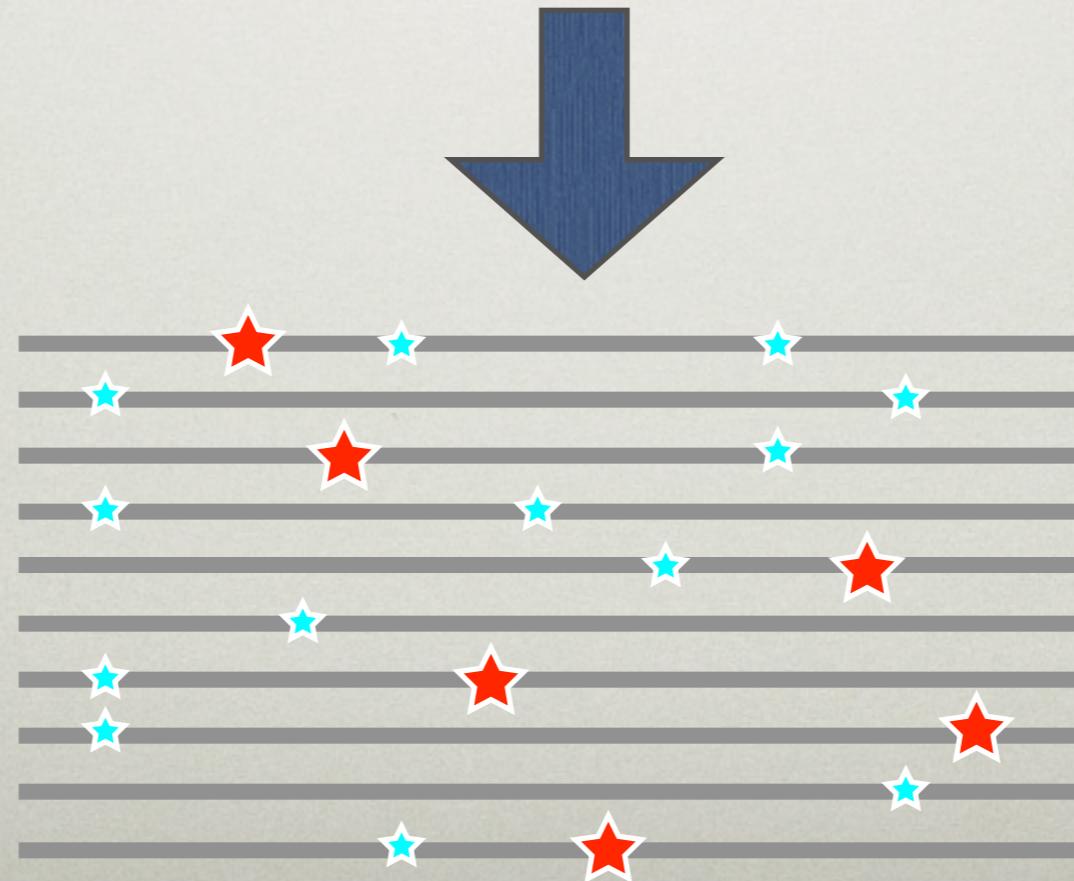


THE EFFECT OF NEGATIVE SELECTION

Chromosomes in a population with standing variation

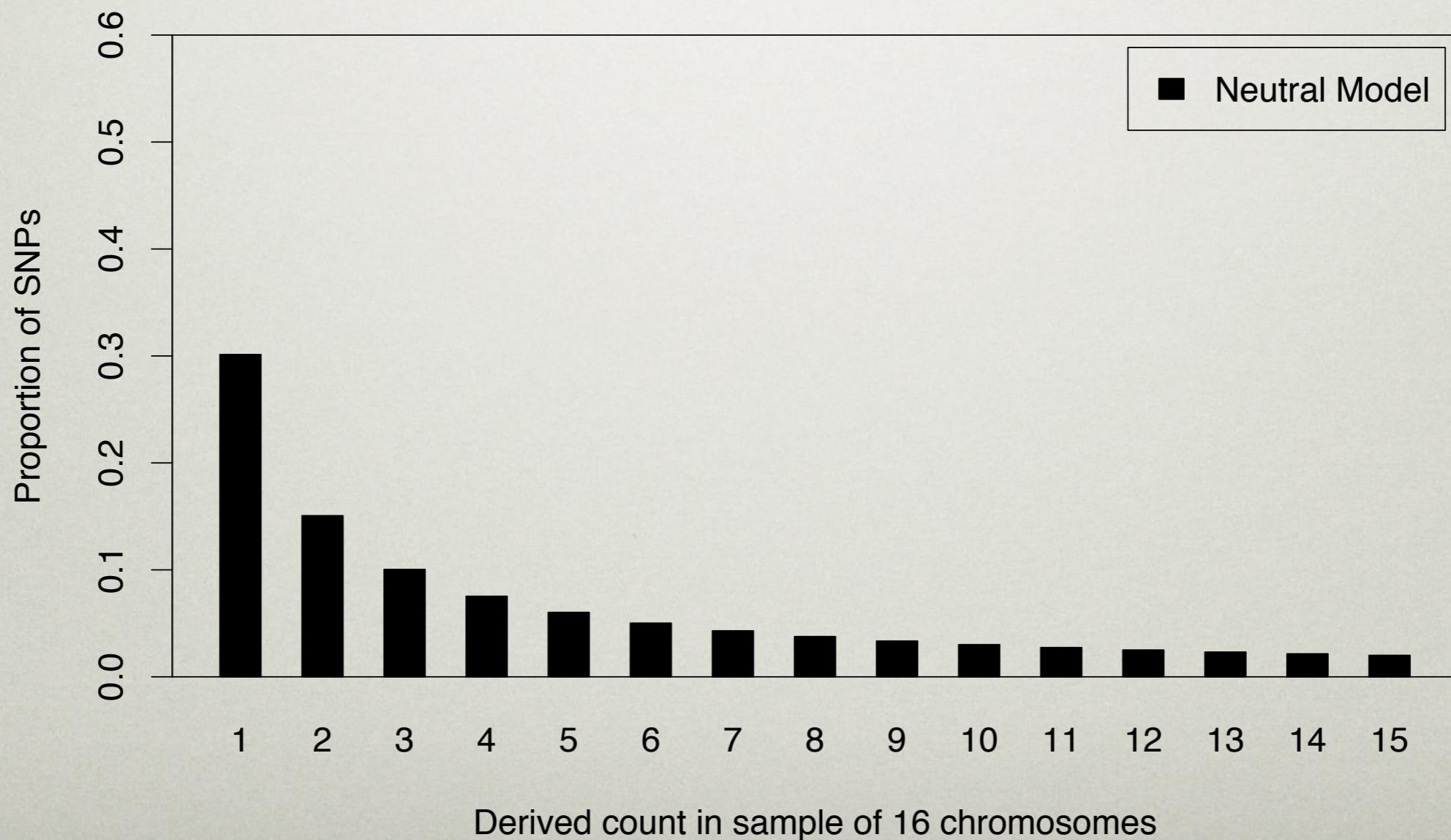


Deleterious mutations will arise in the next generation

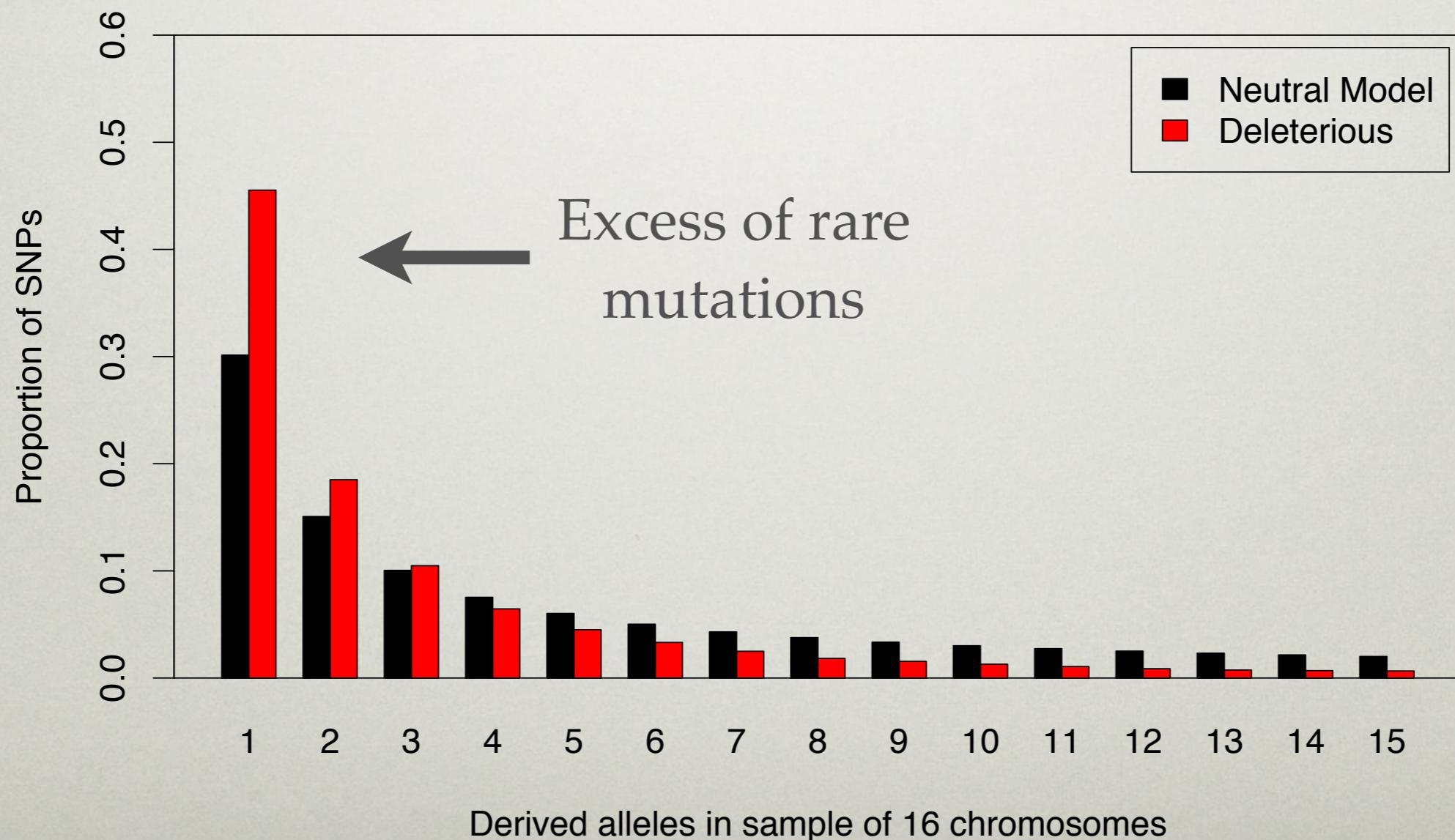


Negative selection:
the action of natural selection purging deleterious mutations.

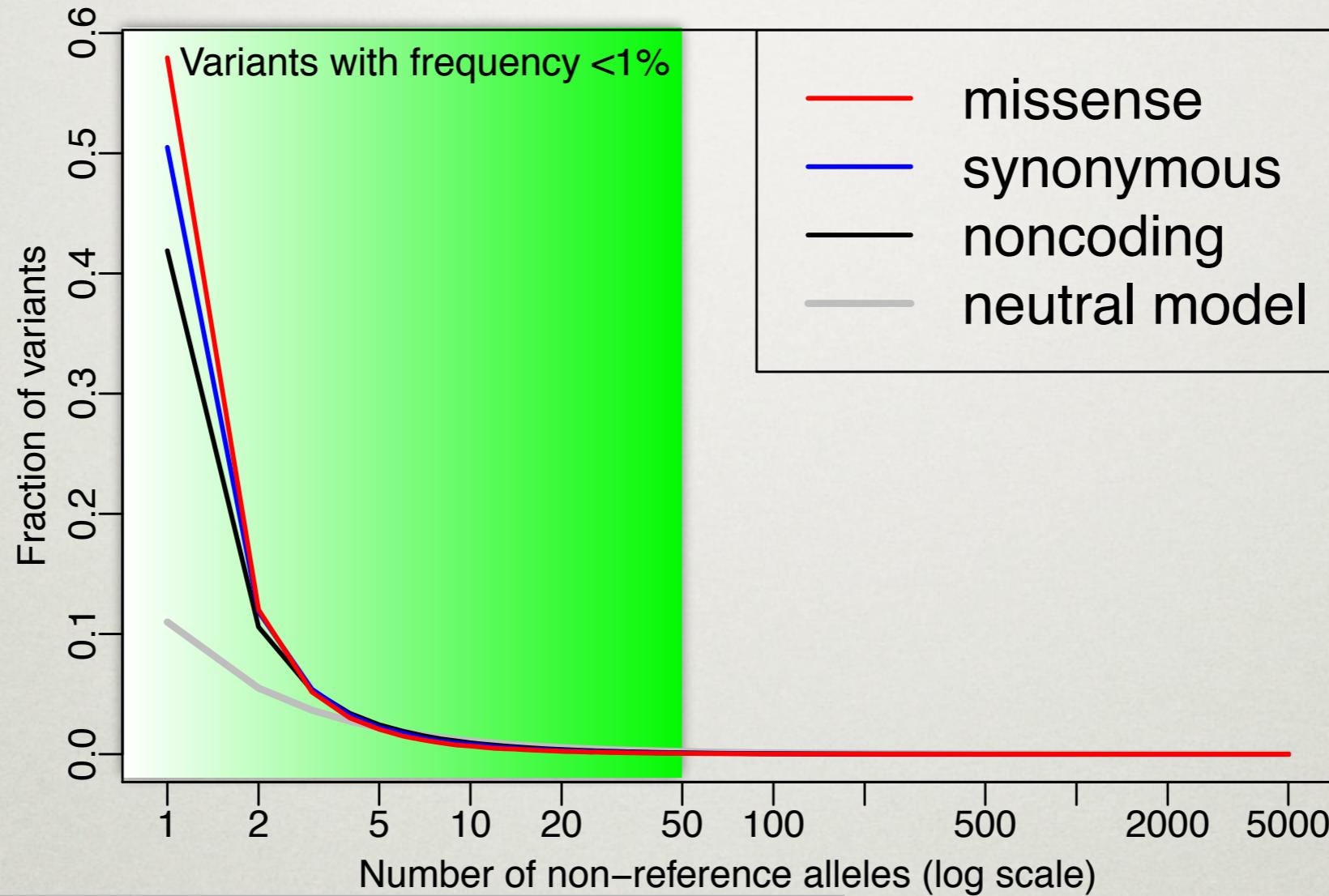
SITE-FREQUENCY SPECTRUM



SITE-FREQUENCY SPECTRUM



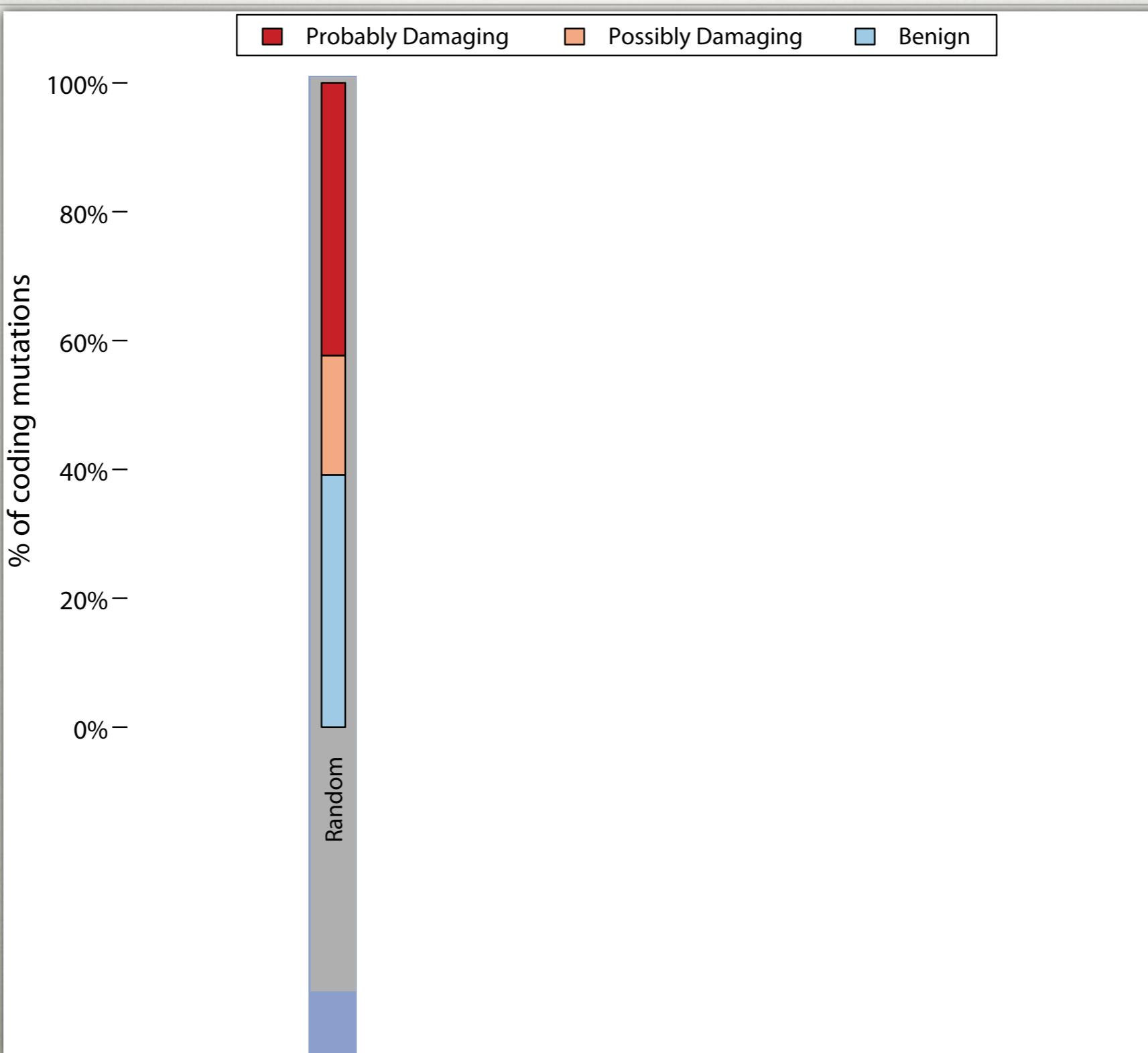
MAJORITY OF HUMAN GENETIC VARIATION IS RARE



Class	Fraction of variants < 1%
Missense	92.6%
Synonymous	88.5%
Non-coding	82.3%

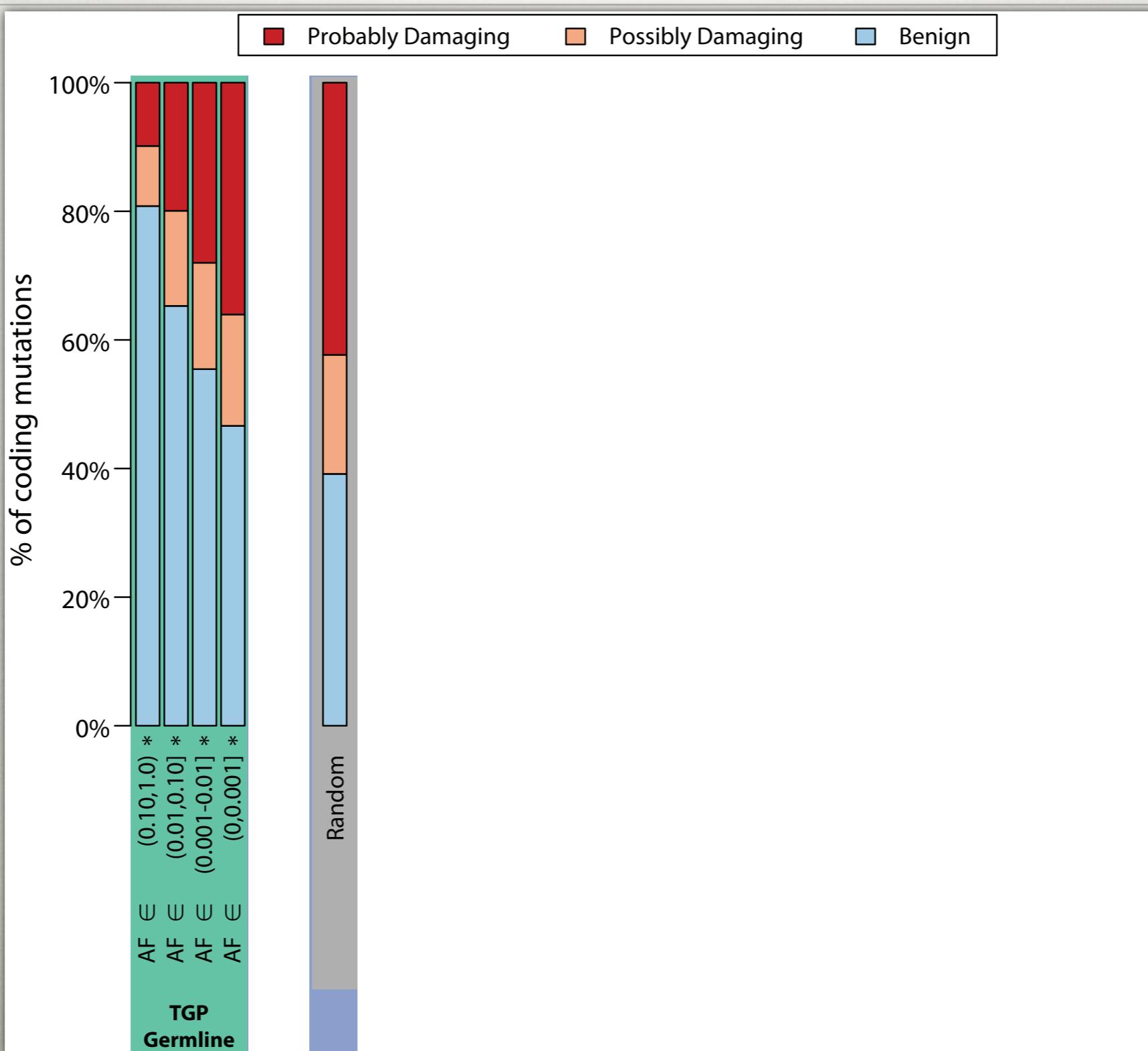


OBSERVED EFFECT OF SELECTION

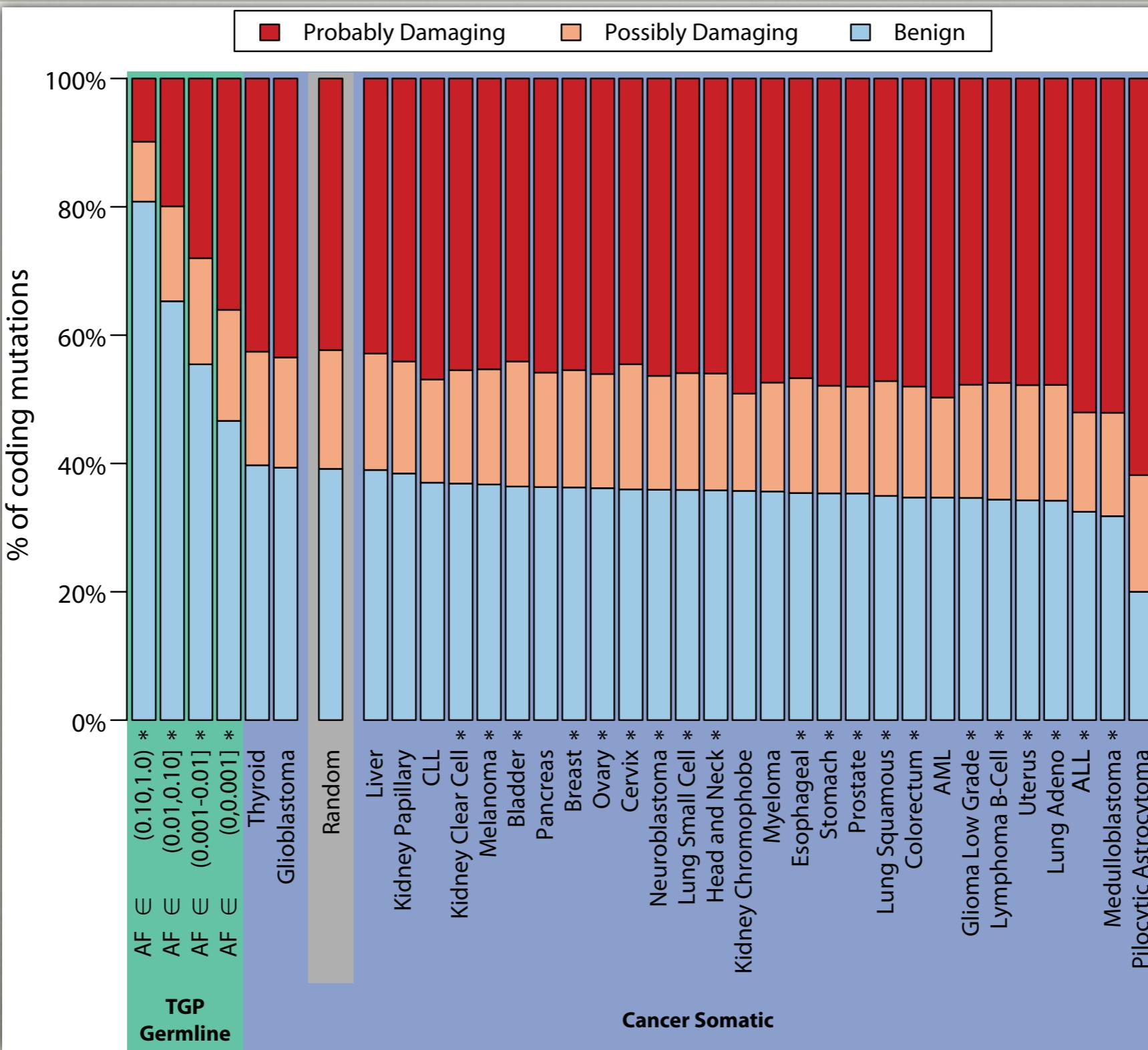


Zach
Szpiech

OBSERVED EFFECT OF SELECTION



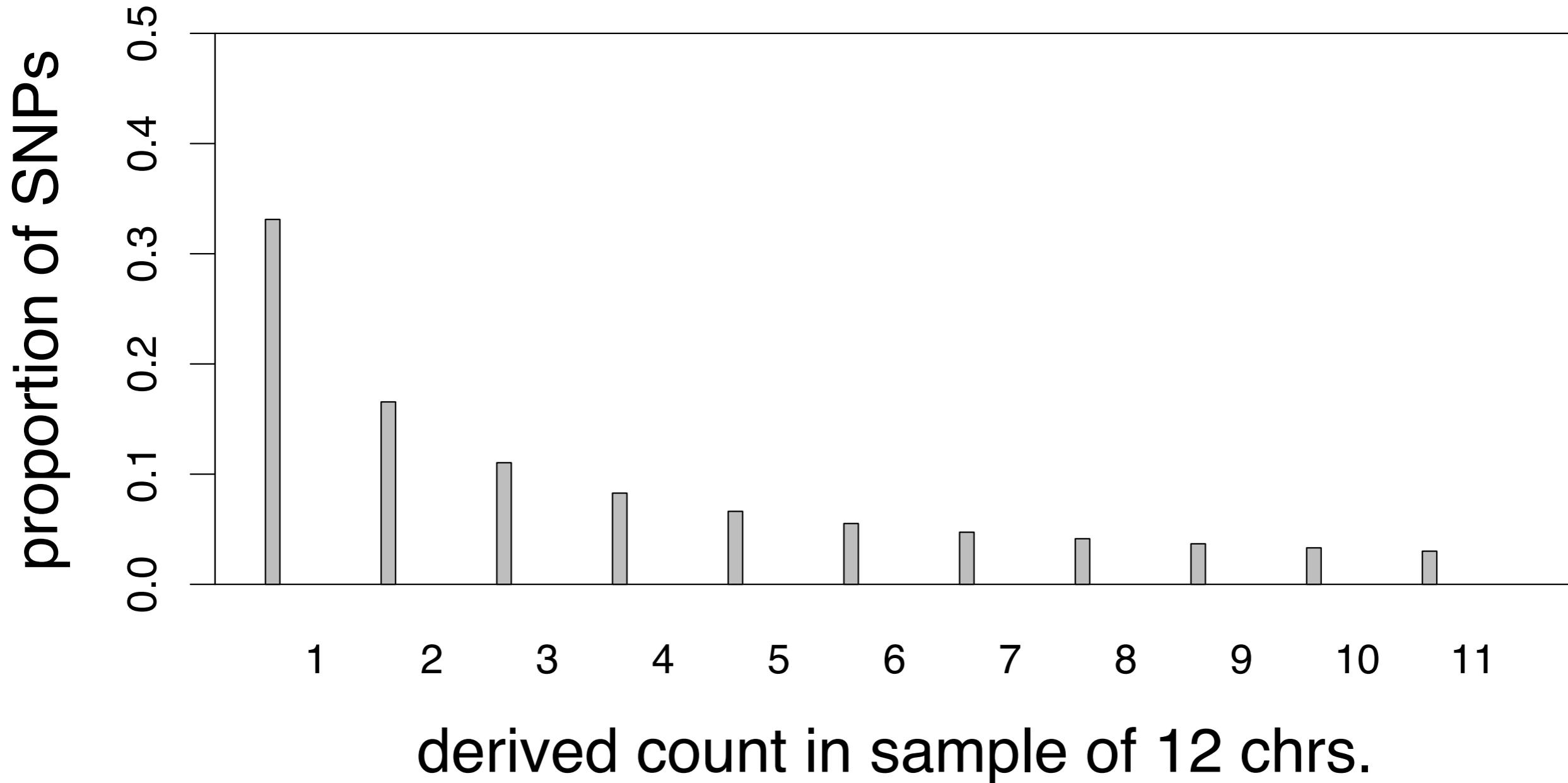
OBSERVED EFFECT OF SELECTION



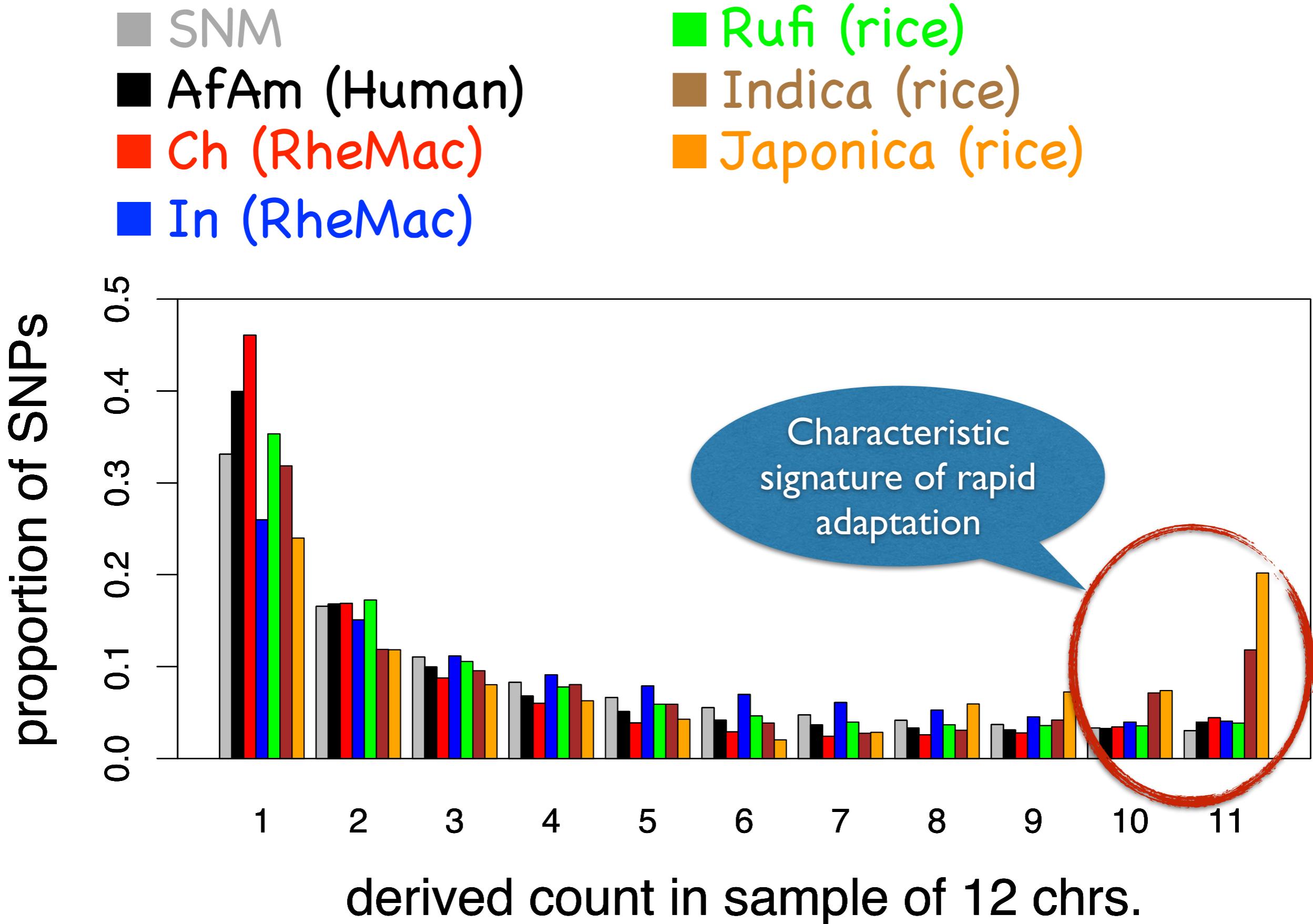
Zach
Szpiech

Site-Frequency Spectrum

The proportion of SNPs at each frequency in a sample of chromosomes.

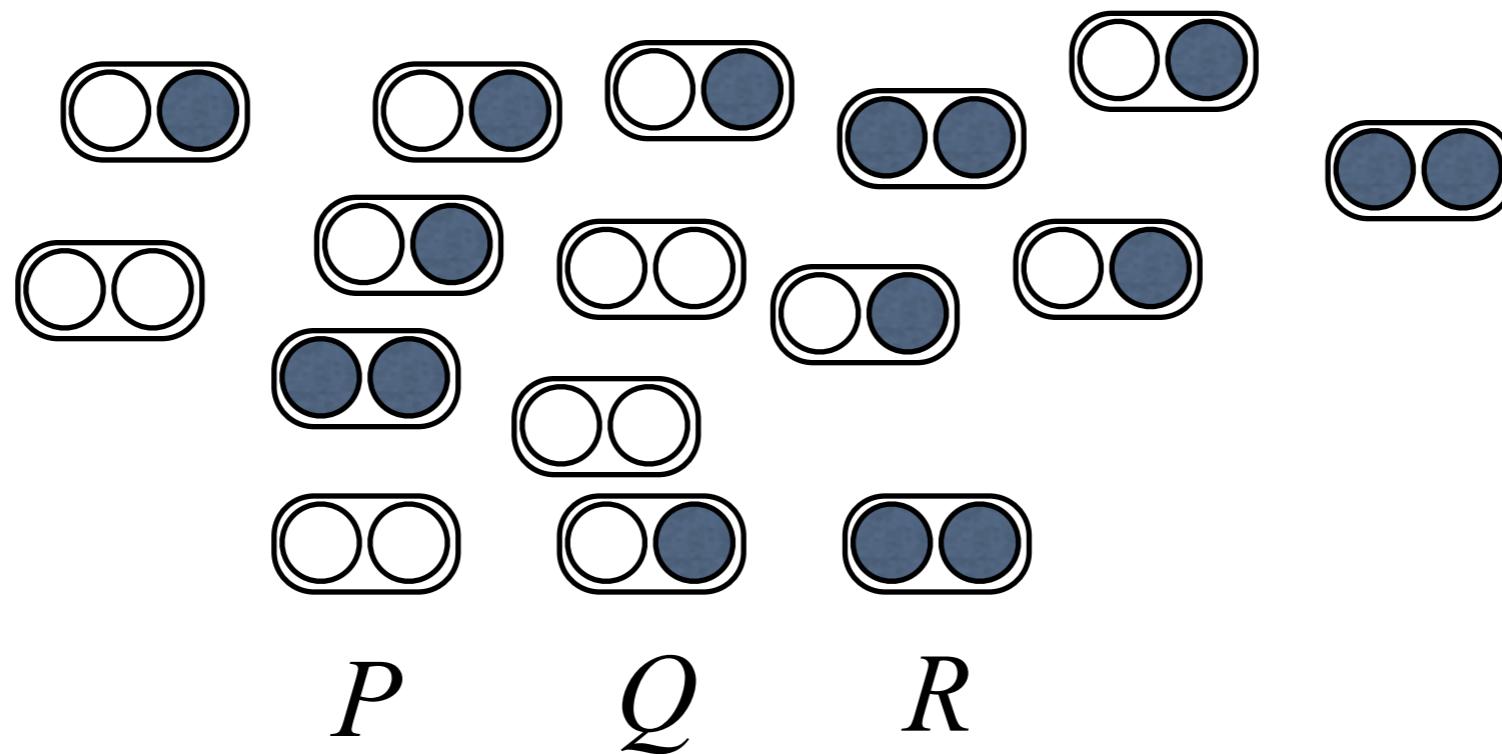


Site-Frequency Spectrum



Population Genetics

- Imagine a population of diploid individuals



- Principles of random mating:
 - Any two individuals are equally likely to mate and reproduce to populate the next generation.
 - Either chromosome is equally likely to be passed on.

Hardy-Weinberg Principle



Godfrey H. Hardy:
1877-1947

Wilhelm Weinberg:
1862-1937

- **Assumptions:**

- Diploid organism
- Sexual reproduction
- Non-overlapping generations
- Only two alleles
- Random mating

- **Conclusion I:**

Both allele AND genotype frequencies will remain constant at HWE generation after generation... forever!

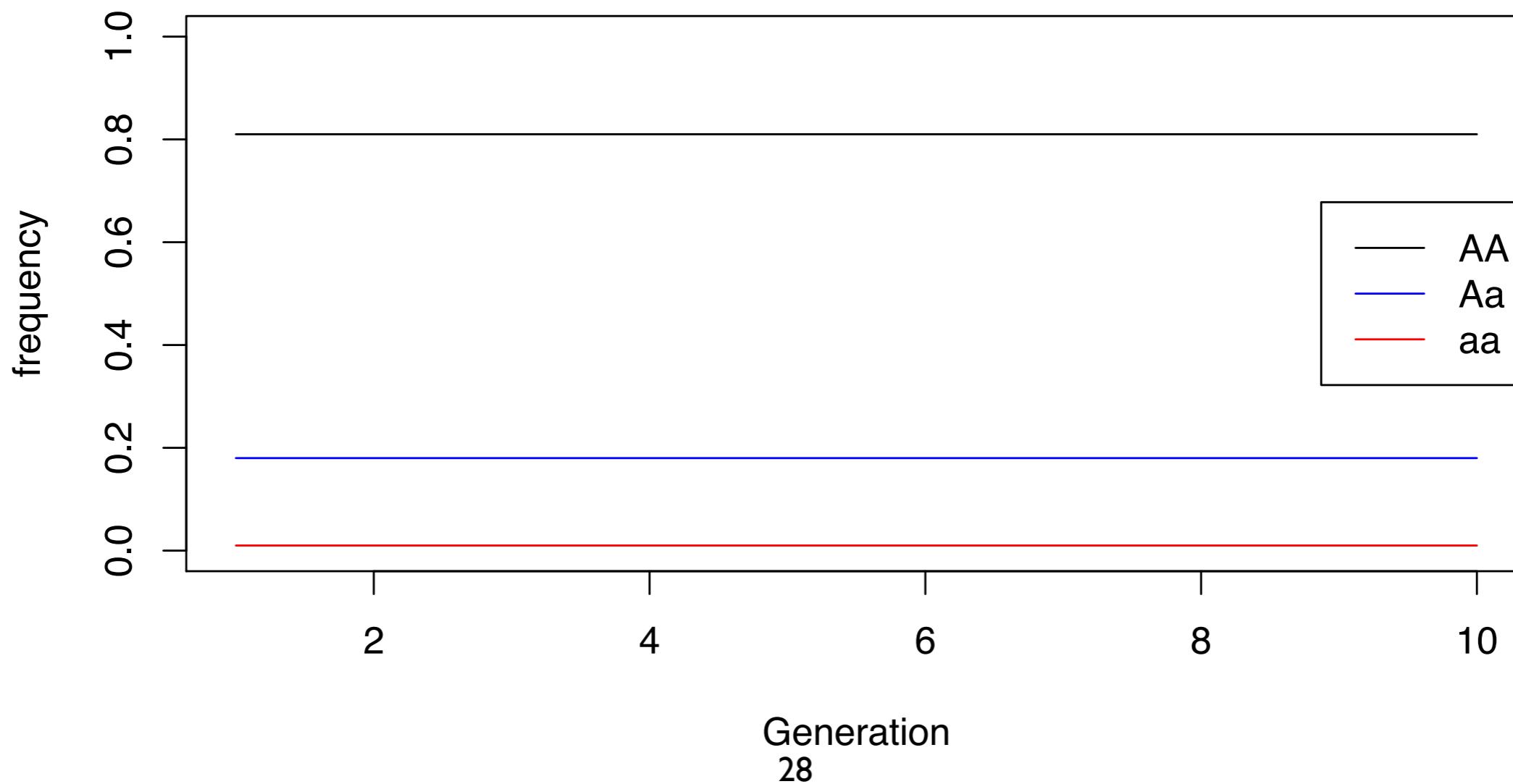
- Identical frequencies in males/females
- Infinite population size
- No migration
- No mutation
- No natural selection

$$\begin{aligned}P &= p^2 \\Q &= 2p(1-p) \\R &= (1-p)^2\end{aligned}$$

Hardy-Weinberg Principle

- Imagine a population of diploid individuals

$$\begin{array}{c} \textcircled{A} \textcircled{A} \quad \textcircled{A} \textcircled{a} \quad \textcircled{a} \textcircled{a} \\ P = 0.81 \quad Q = 0.18 \quad R = 0.01 \end{array}$$

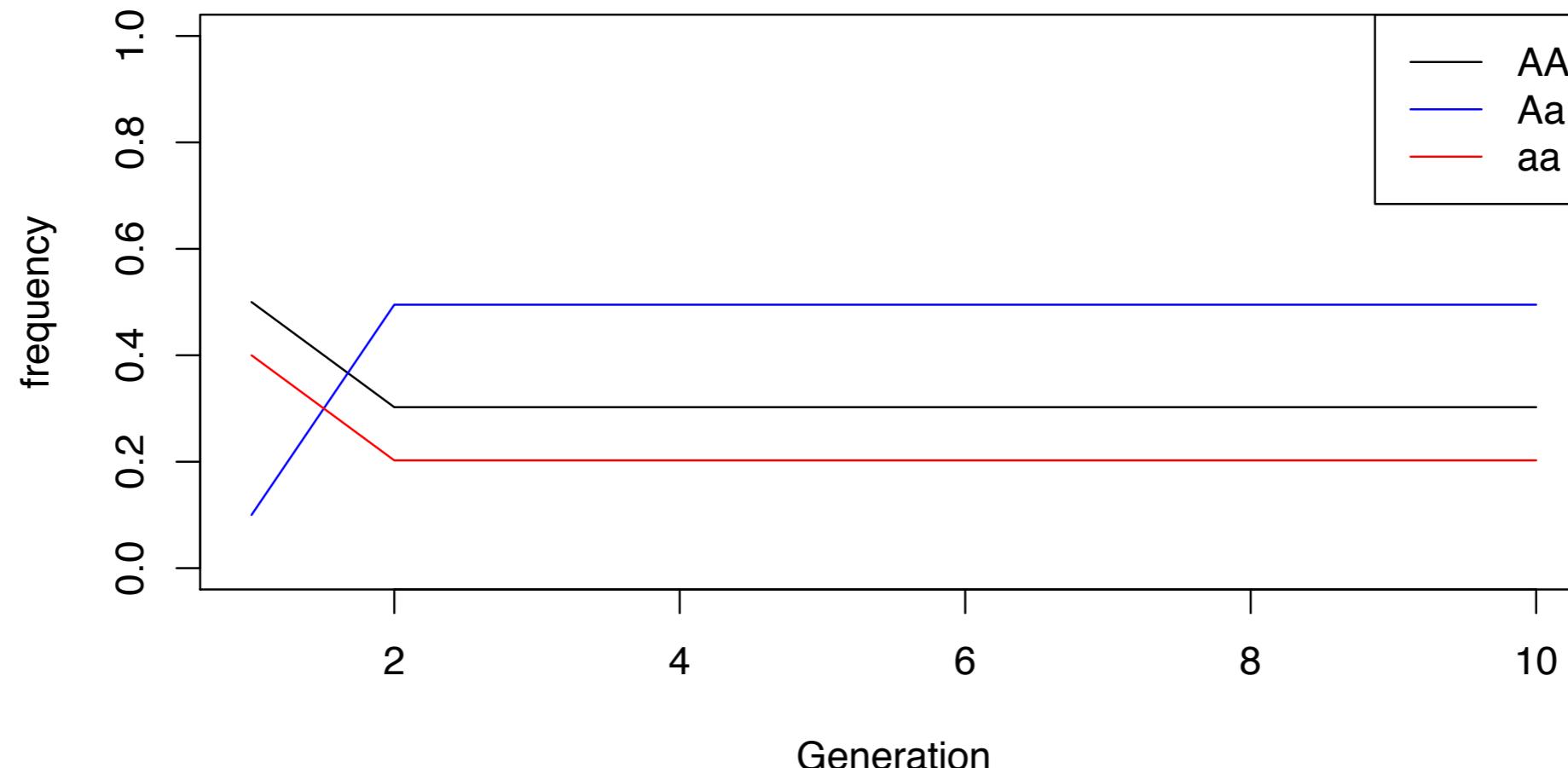


Hardy-Weinberg Principle

- Imagine a population of diploid individuals

$$\begin{array}{ccc} \textcircled{A}\textcircled{A} & \textcircled{A}\textcircled{a} & \textcircled{a}\textcircled{a} \\ P = 0.5 & Q = 0.1 & R = 0.4 \\ p = P + Q/2 = 0.55 \end{array}$$

$$\begin{aligned} p^2 &= 0.3025 \\ 2p(1-p) &= 0.495 \\ (1-p)^2 &= 0.2025 \end{aligned}$$



- Conclusion 2: A single round of random mating will return the population to HWE frequencies!

Hardy-Weinberg Principle



Godfrey H. Hardy:
1877-1947

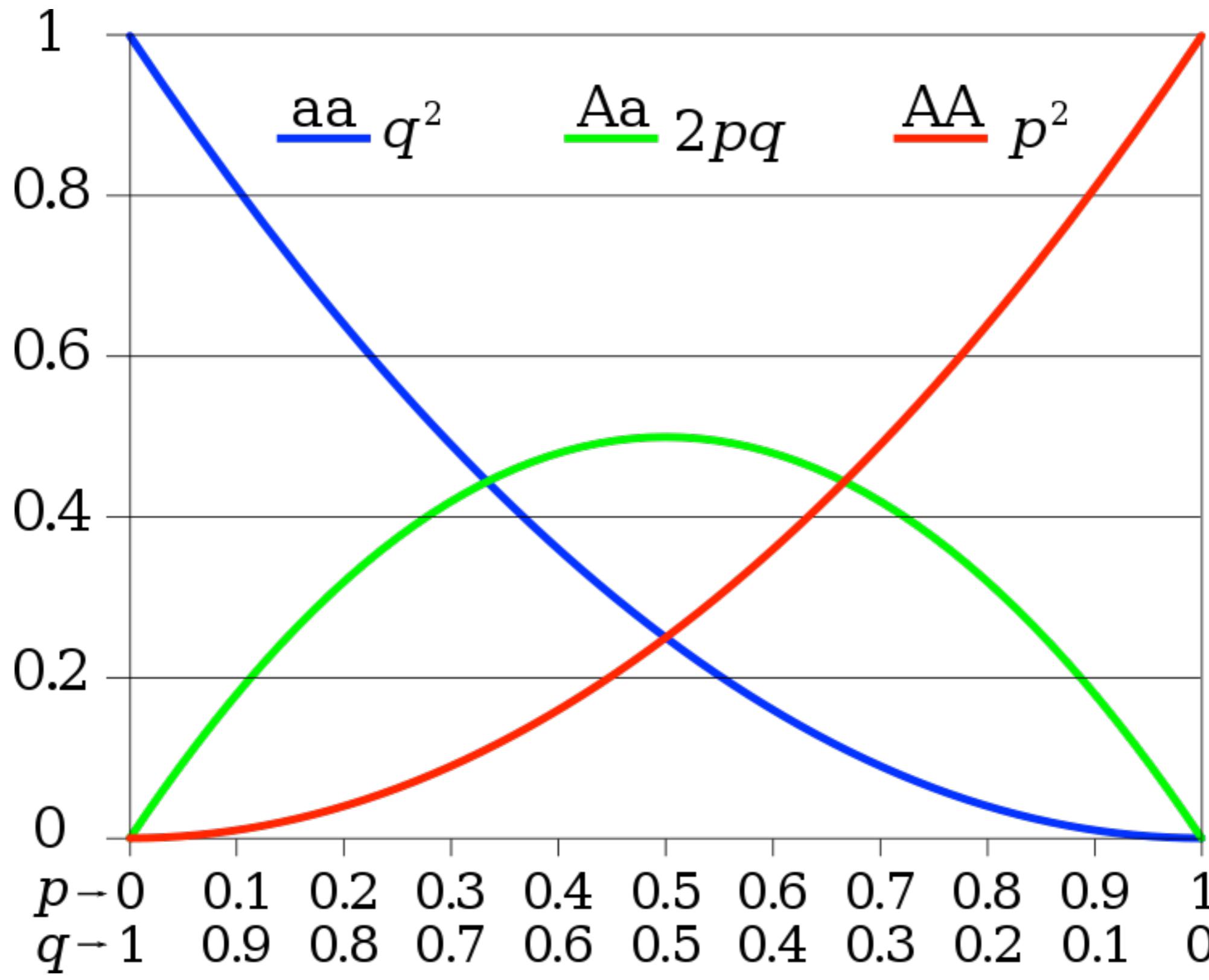
Wilhelm Weinberg:
1862-1937

- **Assumptions:**

- Diploid organism
- Sexual reproduction
- Non-overlapping generations
- Only two alleles
- Random mating

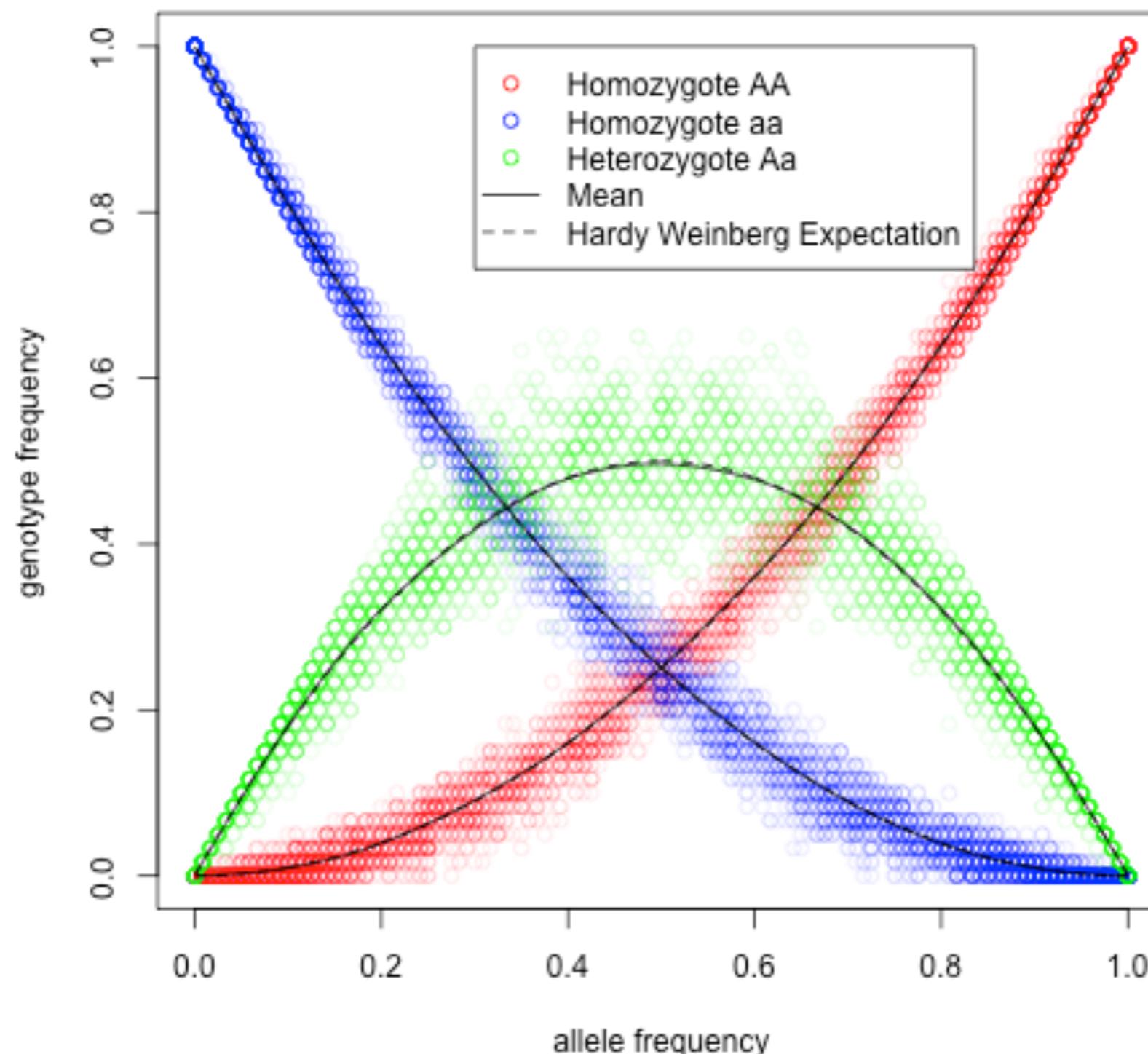
- Identical frequencies in males/females
- Infinite population size
- No migration
- No mutation
- No natural selection

Hardy-Weinberg Equilibrium



Hardy-Weinberg Equilibrium

HapMap YRI (Africans)



Genetic Drift

- In finite populations, allele frequencies can and do change over time.
- In fact, **EVERY** genetic variant will either be lost from the population ($p=0$) or fixed in the population ($p=1$) some time in the future.
- The most common model for finite populations is the **Wright-Fisher model**.
- This model makes explicit use of the *binomial distribution*.

Bernoulli Distribution



Jacob Bernoulli
1655-1705

- One of the simplest *probability distributions*
- A *discrete probability distribution*
- Classic example: tossing a coin
- If a coin toss comes up heads with probability p , it results in tails with probability $1-p$.
- If X is a Bernoulli Random Variable, x is an observation we write:
$$f(x|p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$
- The *Expected Value* is $E[X] = p$, and the *Variance* is $V[X] = p(1-p)$.

Binomial Distribution

- We introduced the Bernoulli Distribution, where we imagine a coin flip resulting in heads with probability p .
- But if we flipped the coin N times, how many heads would we expect?
- What is the probability that we get heads all N times?
- The number of “successes” in a fixed number of trials is described by the *Binomial Distribution*.
- Written out, if the probability of each success is p , then the probability we observe j successes in N trials is:

$$P(j|N, p) = \binom{N}{j} p^j (1-p)^{N-j}; \binom{N}{j} = \frac{N!}{j!(N-j)!}$$

Binomial Mean and Variance

- The mean of a Binomial Random Variable is:
 - $E[J] = Np$
- With variance:
 - $V[J] = p(1-p)/N$

Wright-Fisher Model



Sewall Wright:
1889-1988

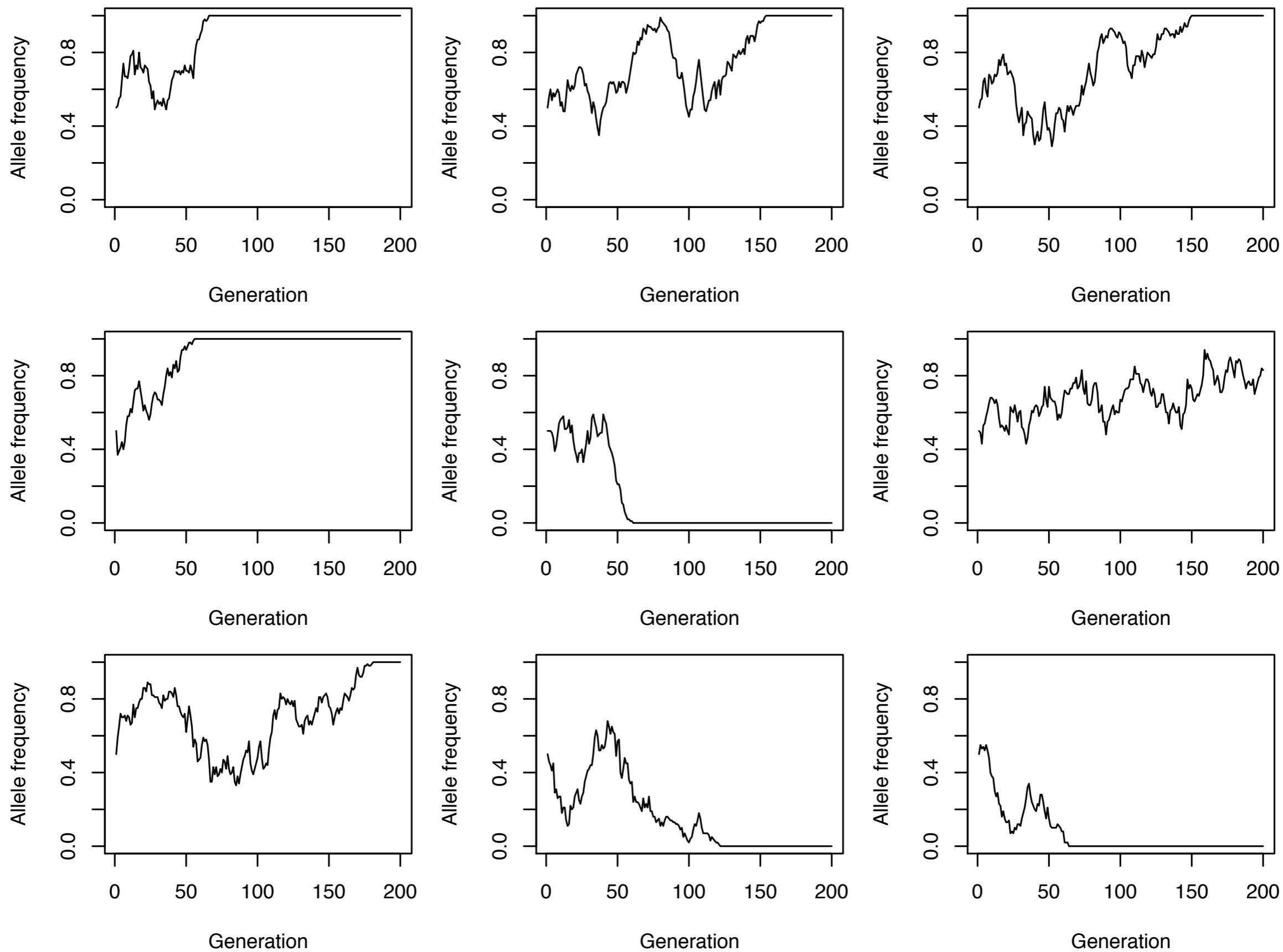


Sir Ronald Fisher
1890-1962

- Suppose a population of N individuals.
- Let $X(t)$ be the #chromosomes carrying an allele A in generation t:

$$\begin{aligned} P(X(t+1) = j | X(t) = i) &= \binom{N}{j} p^j (1-p)^{N-j} \\ &= \text{Bin}(j|N, i/N) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \end{aligned}$$

Wright-Fisher Model ($N=100$)



Demographic Effects

- What do you think will happen if a population grows? Or shrinks?

Wright-Fisher Model



Sewall Wright:
1889-1988

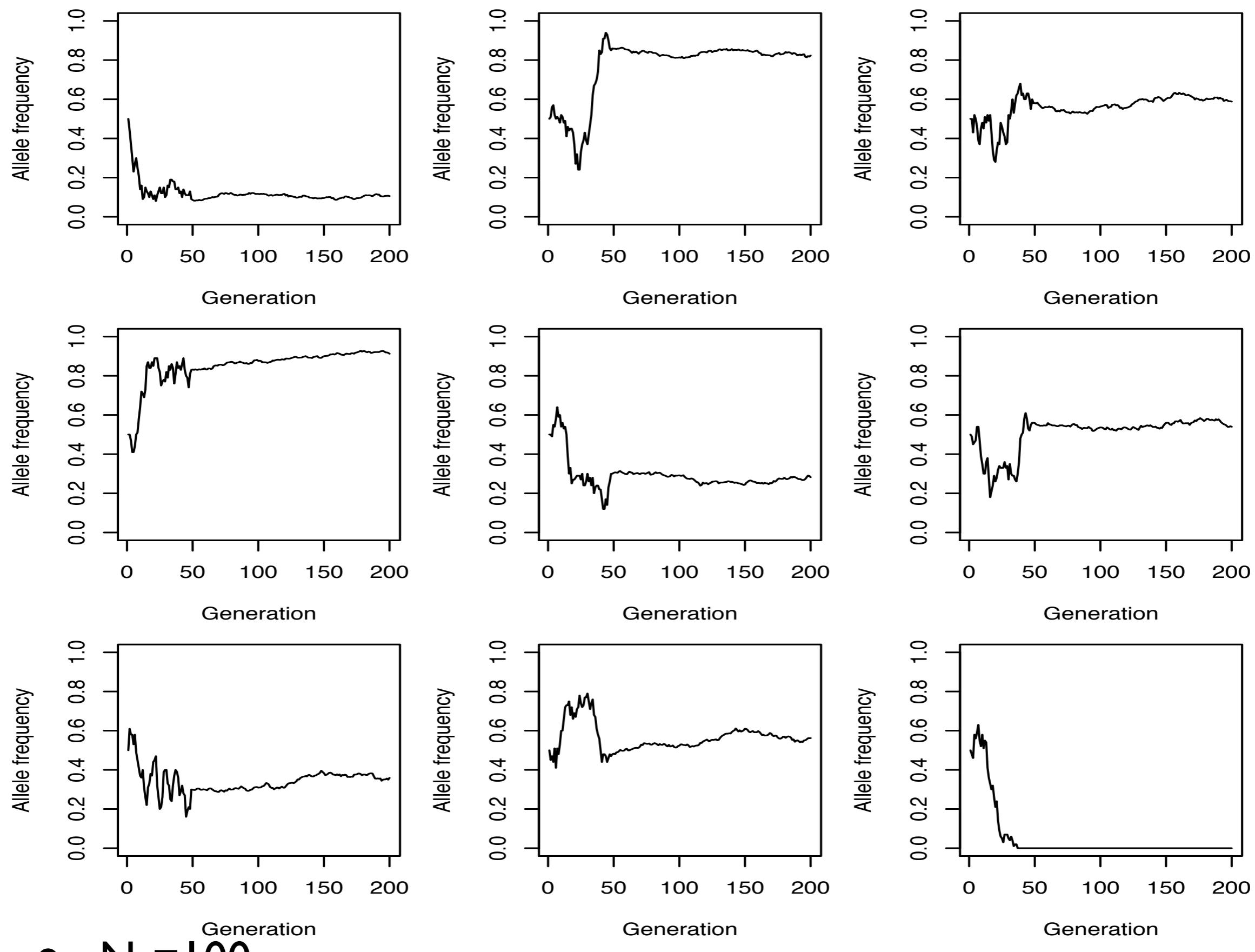


Sir Ronald Fisher
1890-1962

- Suppose a population of N individuals.
- Let $X(t)$ be the #chromosomes carrying an allele A in generation t:

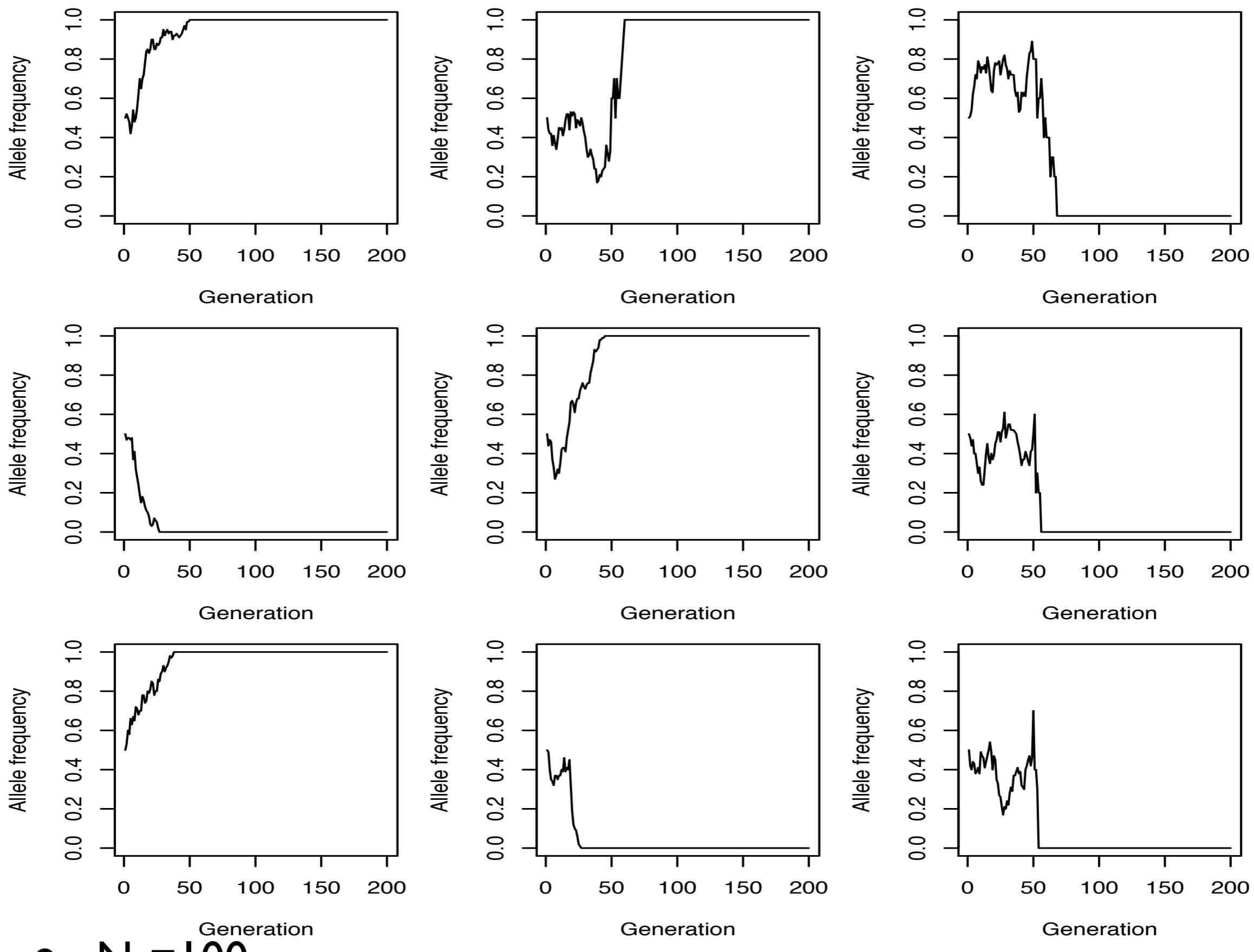
$$\begin{aligned} P(X(t+1) = j | X(t) = i) &= \binom{N}{j} p^j (1-p)^{N-j} \\ &= \text{Bin}(j|N, i/N) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \end{aligned}$$

Wright-Fisher Model with Expansion



• $N_0=100$

Wright-Fisher Model with Contraction



• $N_0 = 100$

Hardy-Weinberg Principle

- Assumptions:
 - Diploid organism
 - Sexual reproduction
 - Non-overlapping generations
 - Only two alleles
 - Random mating
 - Identical frequencies in males/females
 - Infinite population size
 - No migration
 - No mutation
 - No natural selection
- What happens when we allow natural selection to occur?
- Allele frequencies can change!

Natural Selection

- Usually parameterized in terms of a **dominance coefficient (h)**, and a **selection coefficient (s)**, with wildtype fitness set to 1:

Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Fitness	1	$1+hs$	$1+s$

- $h=1$ is completely dominant
- $h=0$ is completely recessive
- $h=0.5$ is “genic” selection, or “codominance”, or “additive” fitness

Natural Selection

Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Fitness	1	$1+hs$	$1+s$

- How do we model the change in allele frequencies?
- What is fitness?
 - Refers to the average number of offspring individuals with a particular genotype will have.
 - Wild-type individuals have on average 1 offspring, while homozygous derived individuals have on average $1+s$ offspring.

Natural Selection

Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Fitness	1	$1+hs$	$1+s$

- In this case, s is the *absolute fitness*.
- If the population size is fixed, then we need to consider *relative fitness*.
 - That is, how fit is an individual genotype relative to the population.
- For this, we need to know average *population fitness*!

$$\bar{w} = p^2(1) + 2pq(1 + hs) + q^2(1 + s) = 1 + sq(2hp + q)$$

Natural Selection

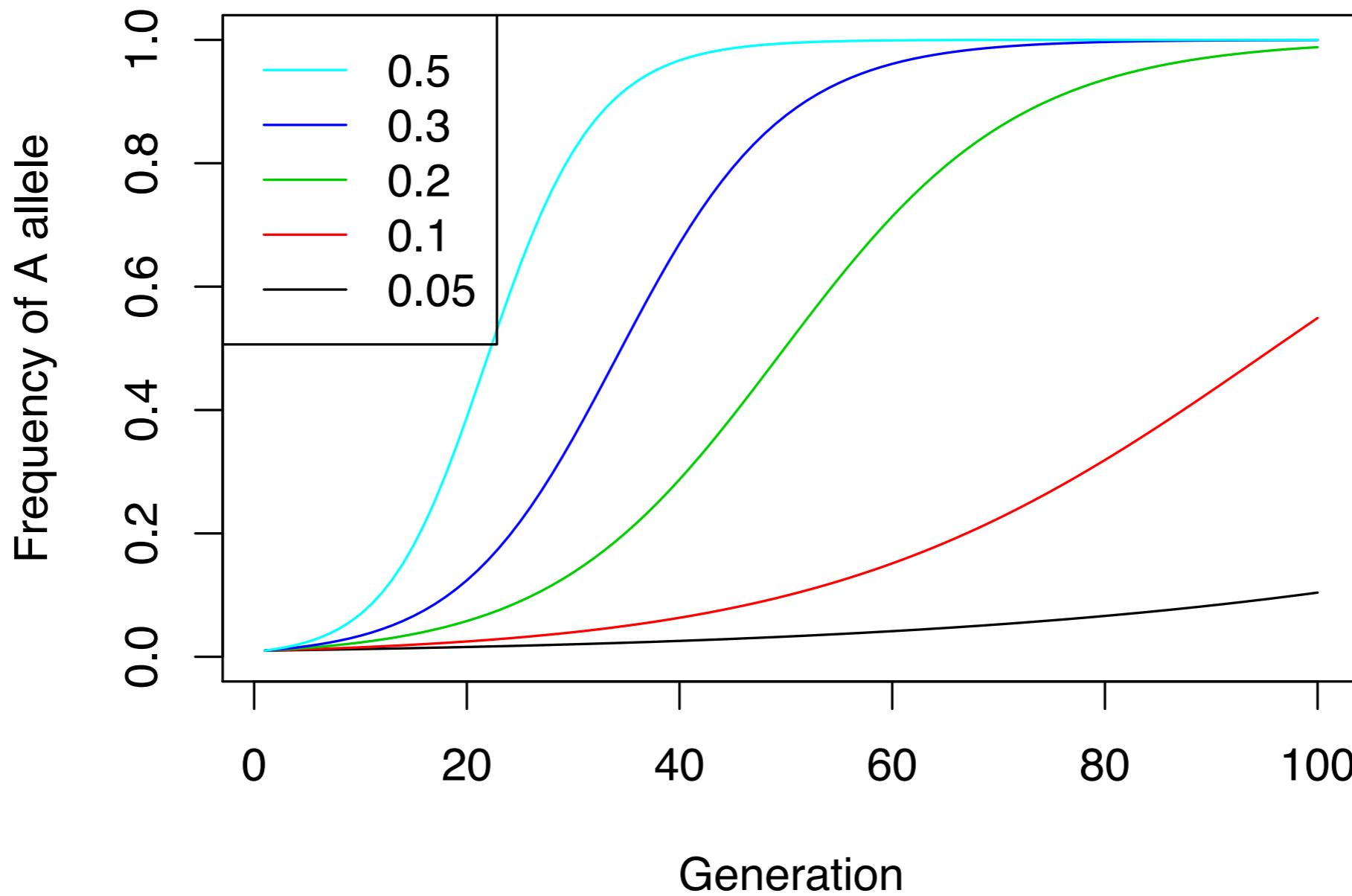
Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Fitness	1	$1+hs$	$1+s$

- The expected frequency in the next generation (q') is then the density of offspring produced by carriers of the derived allele divided by the population fitness:

$$q' = \frac{q^2(1 + s) + pq(1 + hs)}{1 + sq(2hp + q)}$$

Natural Selection

- Trajectory of selected allele with various selection coefficients under genic selection ($h=0.5$) in an “infinite” population

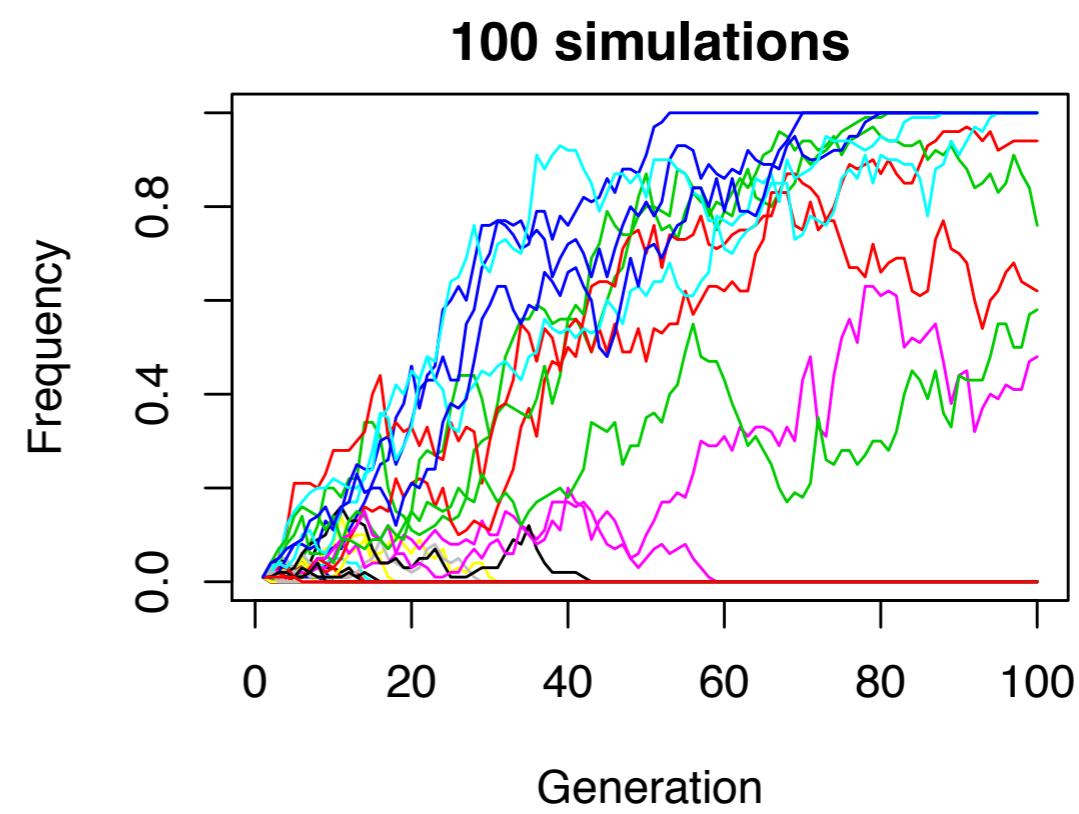
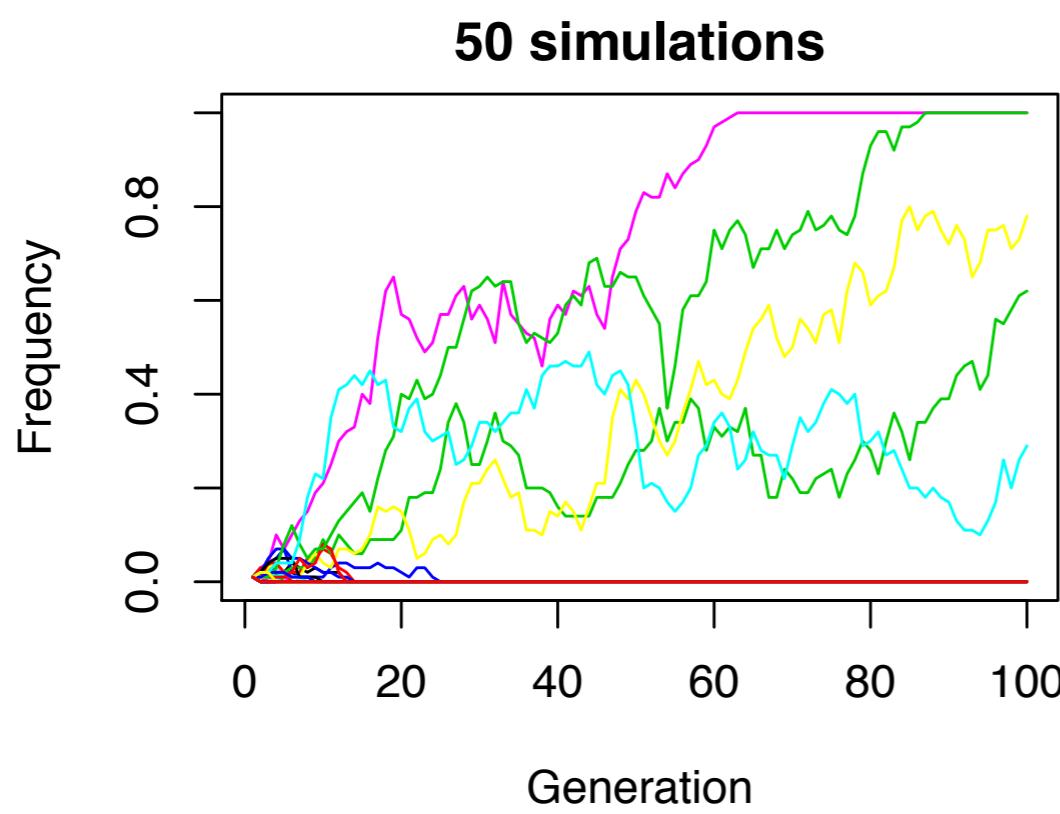
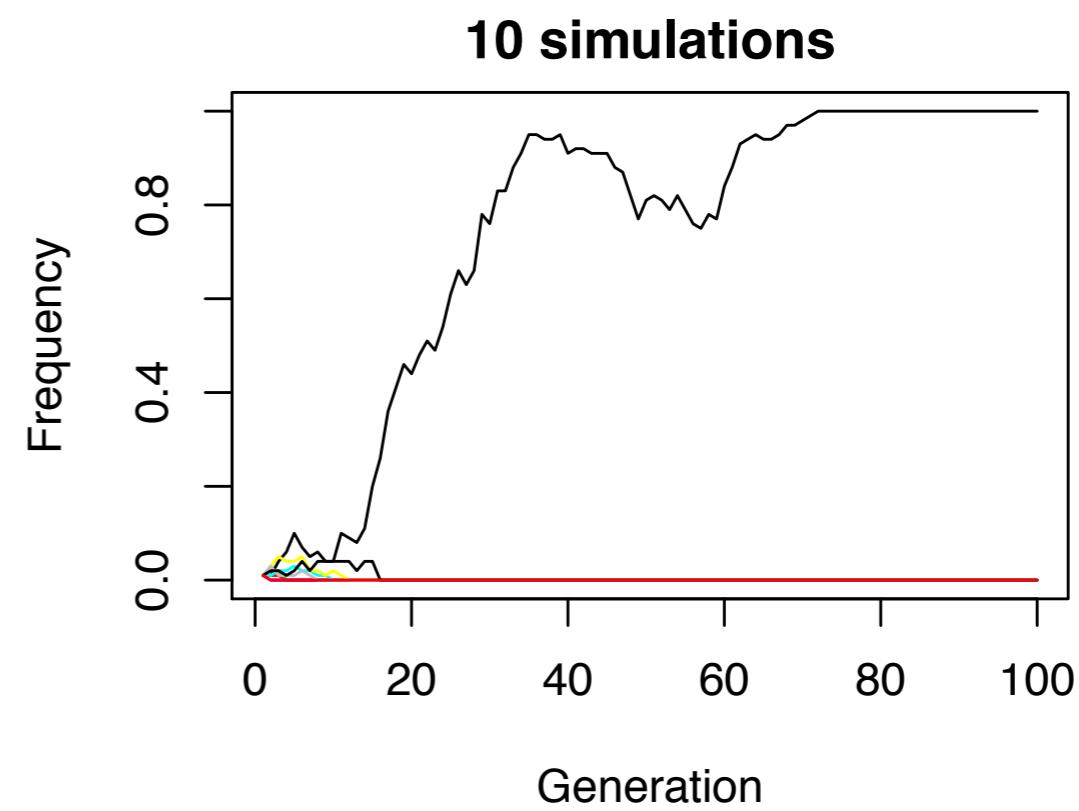
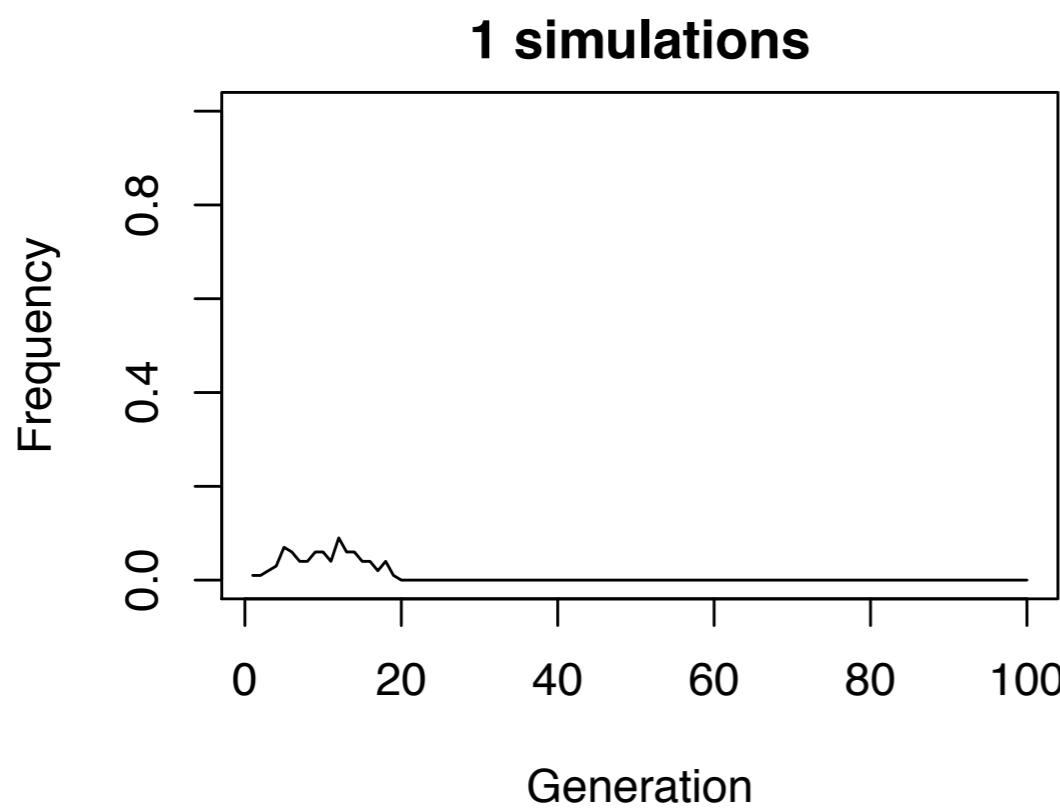


Hardy-Weinberg Principle

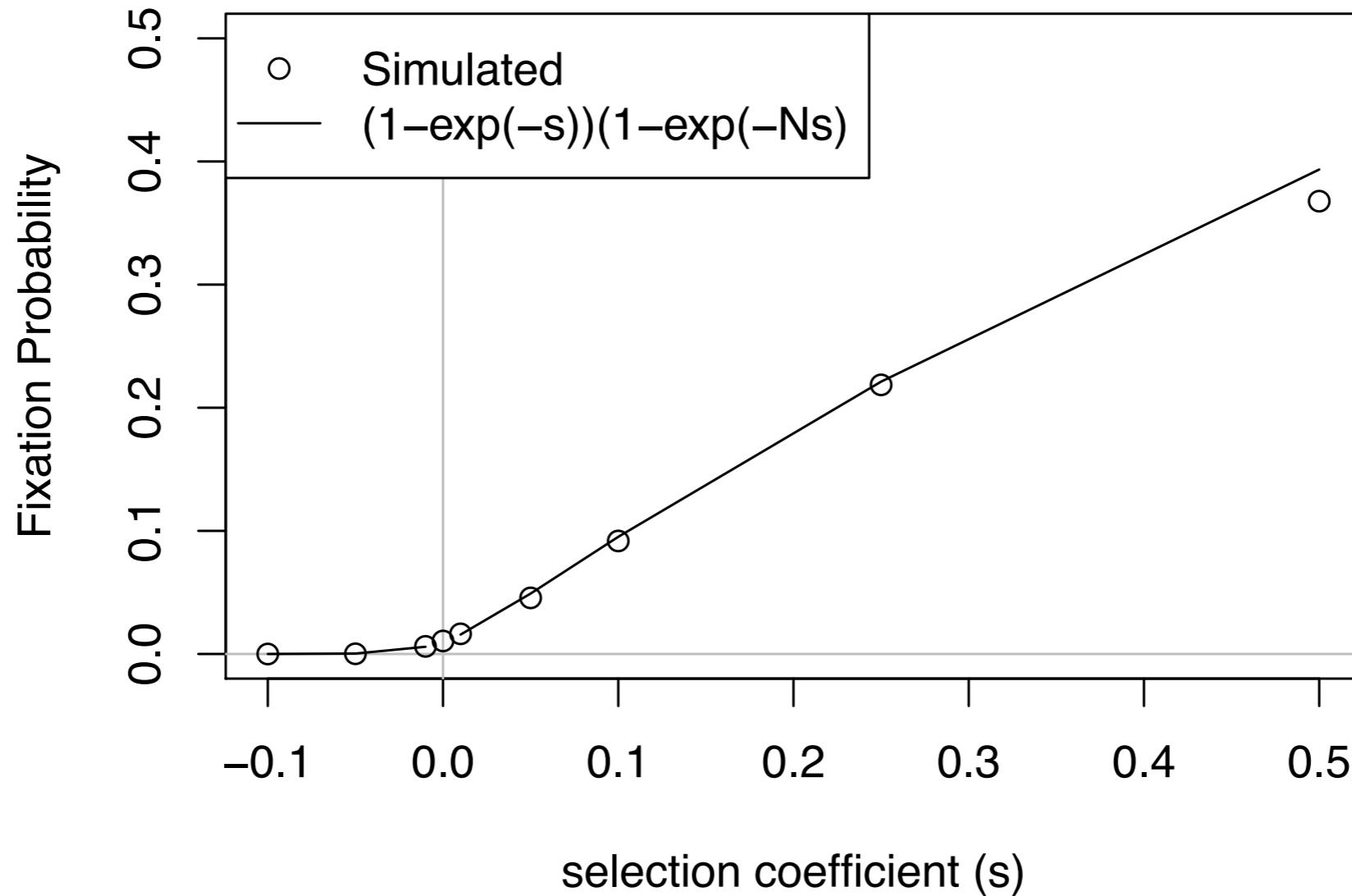
- Assumptions:
 - Diploid organism
 - Sexual reproduction
 - Non-overlapping generations
 - Only two alleles
 - Random mating
 - Identical frequencies in males/females
 - Infinite population size
 - No migration
 - No mutation
 - No natural selection
- What happens with natural selection in a finite population?
 - Directional selection AND drift!

Natural Selection

$N=100; s=0.1; h=0.5$



Natural Selection



- Estimating the probability of fixation of a new mutation ($p_0=1/N$)
- 5000 simulations: $N=100; h=0.5$
- $\Pr(\text{Fixation} \mid s=0, p_0) = p_0!!$

Natural Selection

Time-course data from artificial selection/ancient DNA

- Let's estimate some selection coefficients!
- Given 2 alleles at a locus with frequencies p_0 and q_0 , and fitnesses w_1 and w_2 (with w the population-wide fitness).
- Expected freq. in next generation is: $p_1=p'=p_0w_1/w$.
- We can then write:

$$\frac{p_1}{q_1} = \frac{p_0w_1/w}{q_0w_2/w} = \left(\frac{p_0}{q_0}\right) \left(\frac{w_1}{w_2}\right)$$

- Using induction, you could prove for any generation t :

$$\frac{p_t}{q_t} = \frac{p_0w_1/w}{q_0w_2/w} = \left(\frac{p_0}{q_0}\right) \left(\frac{w_1}{w_2}\right)^t$$

Natural Selection

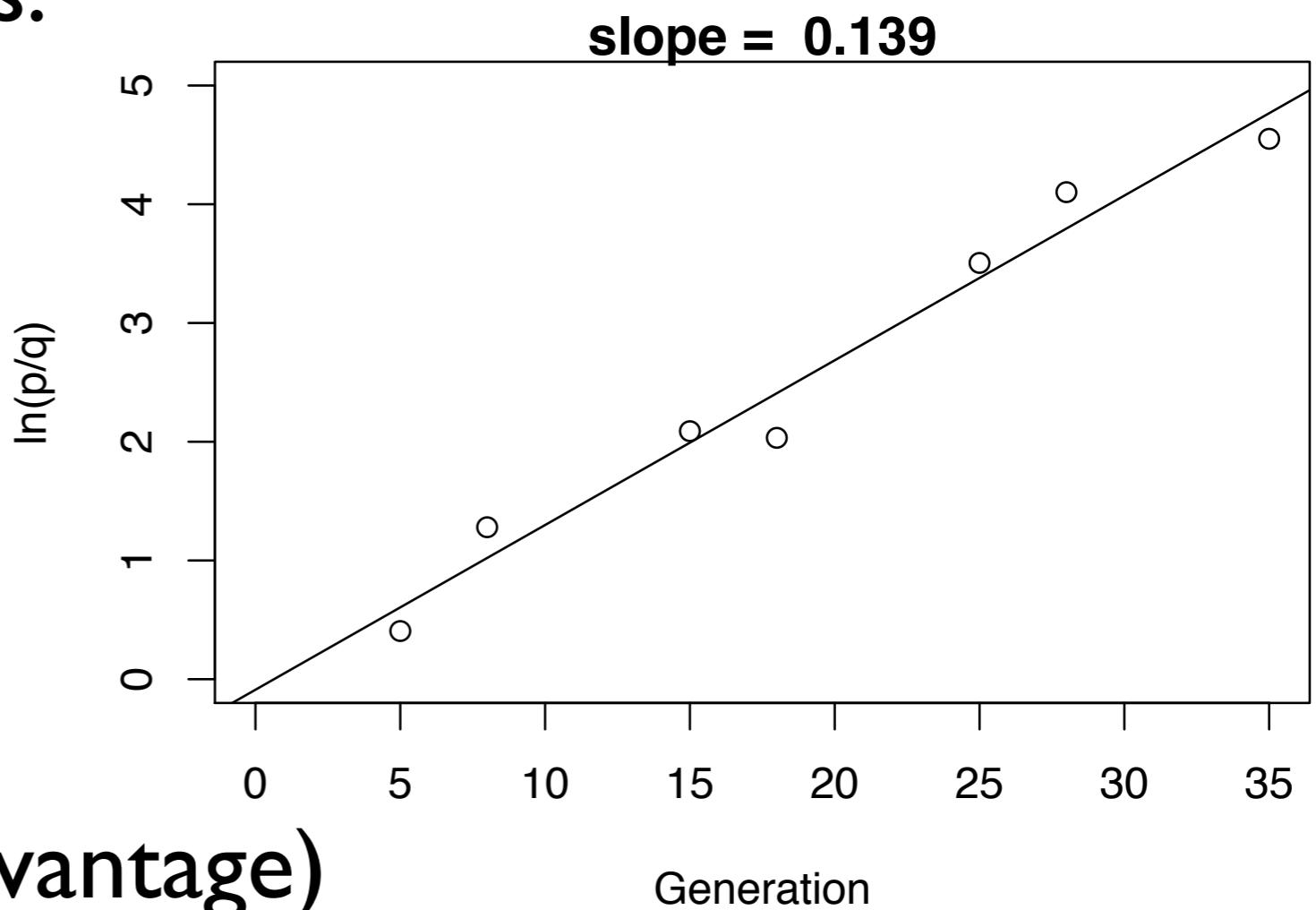
- Taking the natural log of this equation:

$$\log\left(\frac{p_t}{q_t}\right) = \log\left(\frac{w_1}{w_2}\right)t + \log\left(\frac{p_0}{q_0}\right)$$

- Which is now a linear function of t , the number of generations.
- Therefore, the ratio of the fitnesses $w_1/w_2 = e^{\text{slope}}$

Natural Selection

- Experiment: Set up a population of bacteria in a chemostat, and let them reproduce.
- Sample roughly every 5 generations.
- A slope of 0.139 implies:
 $w_1 = e^{0.139} = 1.15$
- Assume $w_2 = 1$.
- Thus, allele p has a 15% fitness advantage over allele q!
- (simulated with 20% advantage)



Summary

- Hardy-Weinberg Equilibrium requires many assumptions, all of which are routinely violated in natural populations.
- Nevertheless, the vast majority of variants are in HWE.
 - Deviations almost always due to technical artifacts!
- Simulating Wright-Fisher models is easy!
- Natural selection changes the expected allele frequency in the next generation.
 - But drift still acts in finite populations!