

# Introduction to Wright-Fisher Simulations

Ryan Hernandez

# Goals

- Simulate the standard neutral model, demographic effects, and natural selection

# 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

- **Conclusion I:**

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

$$P=p^2$$
$$Q=2p(1-p)$$
$$R=(1-p)^2$$

# Hardy-Weinberg Principle

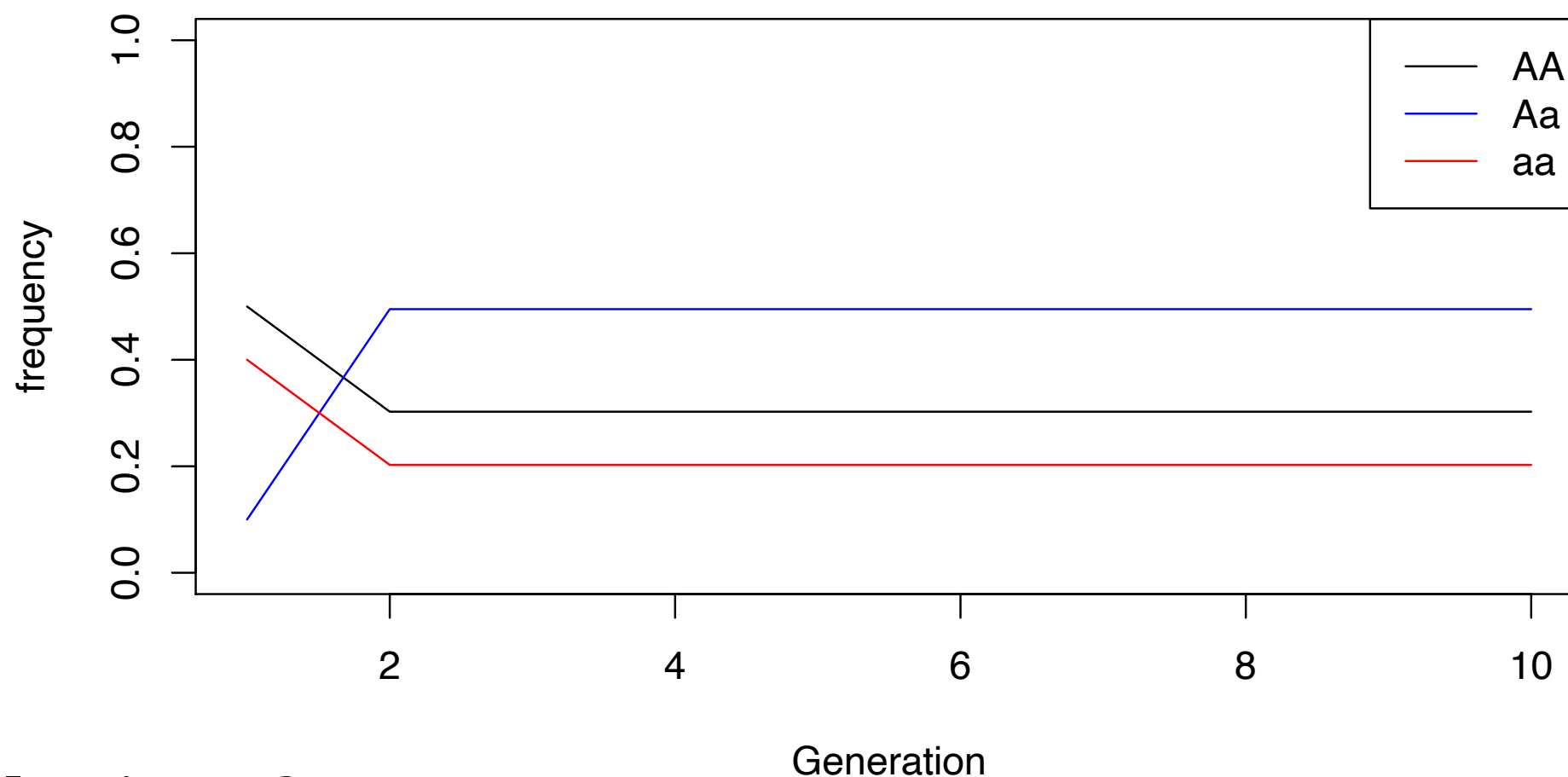
- Imagine a population of diploid individuals

$\textcircled{A}\textcircled{A}$      $\textcircled{A}\textcircled{a}$      $\textcircled{a}\textcircled{a}$   
 $P = 0.5$      $Q = 0.1$      $R = 0.4$   
 $p = P + Q/2 = 0.55$

$$p^2 = 0.3025$$

$$2p(1 - p) = 0.495$$

$$(1 - p)^2 = 0.2025$$



- 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

# 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*.

# 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  $t$ :

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

# Wright-Fisher Model

- A simple R function to simulation genetic drift:

Initial pop size  
Starting frequency  
number of simulated generations

```
WF=function(N, p, G){  
  t=array(NA,dim=G);  
  t[1] = p;  
  for(i in 2:G){  
    t[i] = rbinom(1,N,t[i-1])/N;  
  }  
  return(t);  
}
```

- Run it in R using:

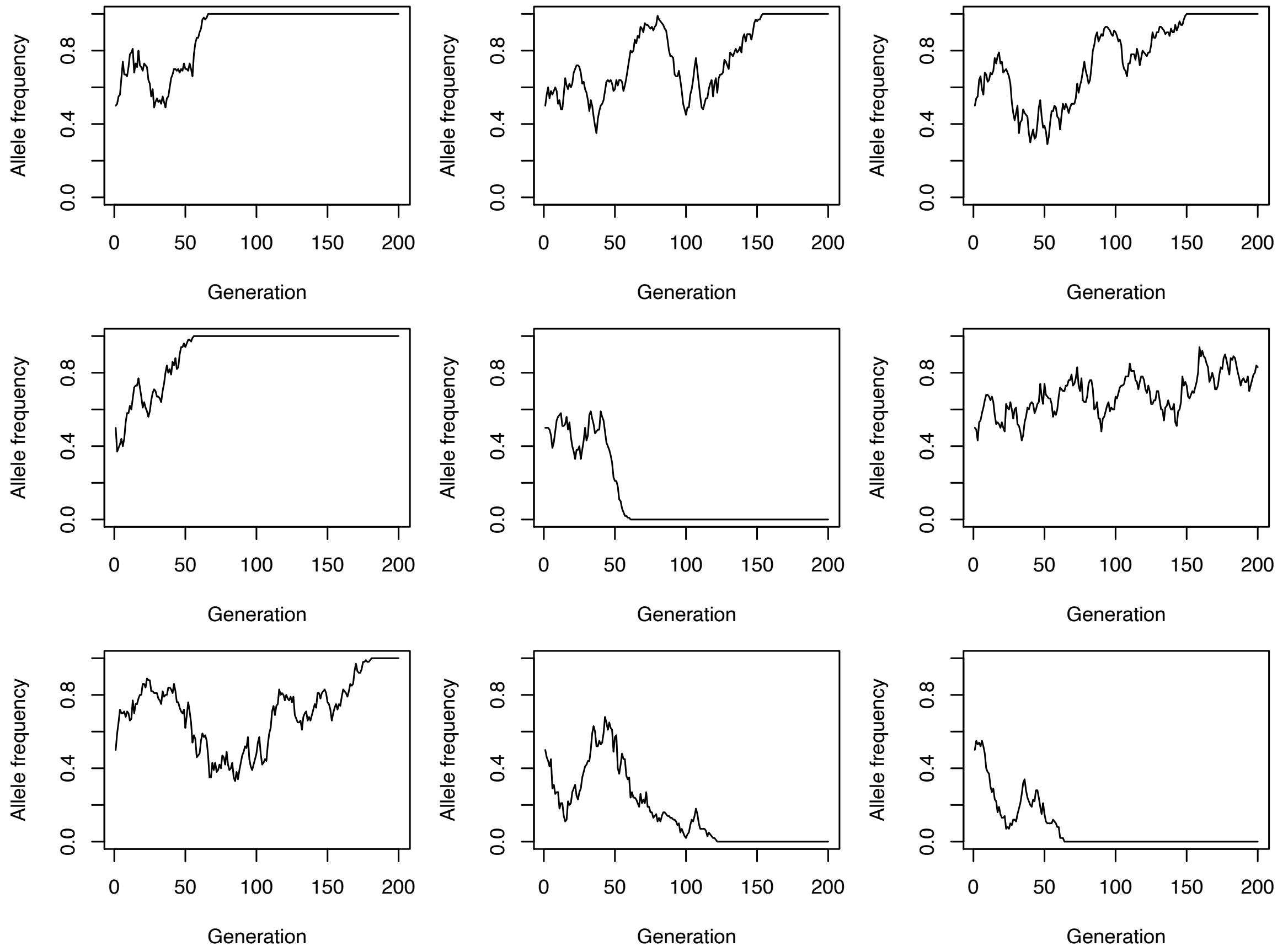
```
f=WF(100, 0.5, 200)  
plot(f)
```



# Questions

- What happens in your plot?
- Were you able to get any fixations or losses?

# Wright-Fisher Model



# Demographic Effects

- Population changes size at a given generation

# 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

- A simple R function to simulation demographic effects:

```
WFdemog = function(N, p, G, Gd, v){  
  t=array(NA,dim=G);  
  t[1] = p;  
  for(i in 2:G){  
    if(i == Gd){  
      N = N*v;  
    }  
    t[i] = rbinom(1,N,t[i-1])/N;  
  }  
  return(t);  
}
```

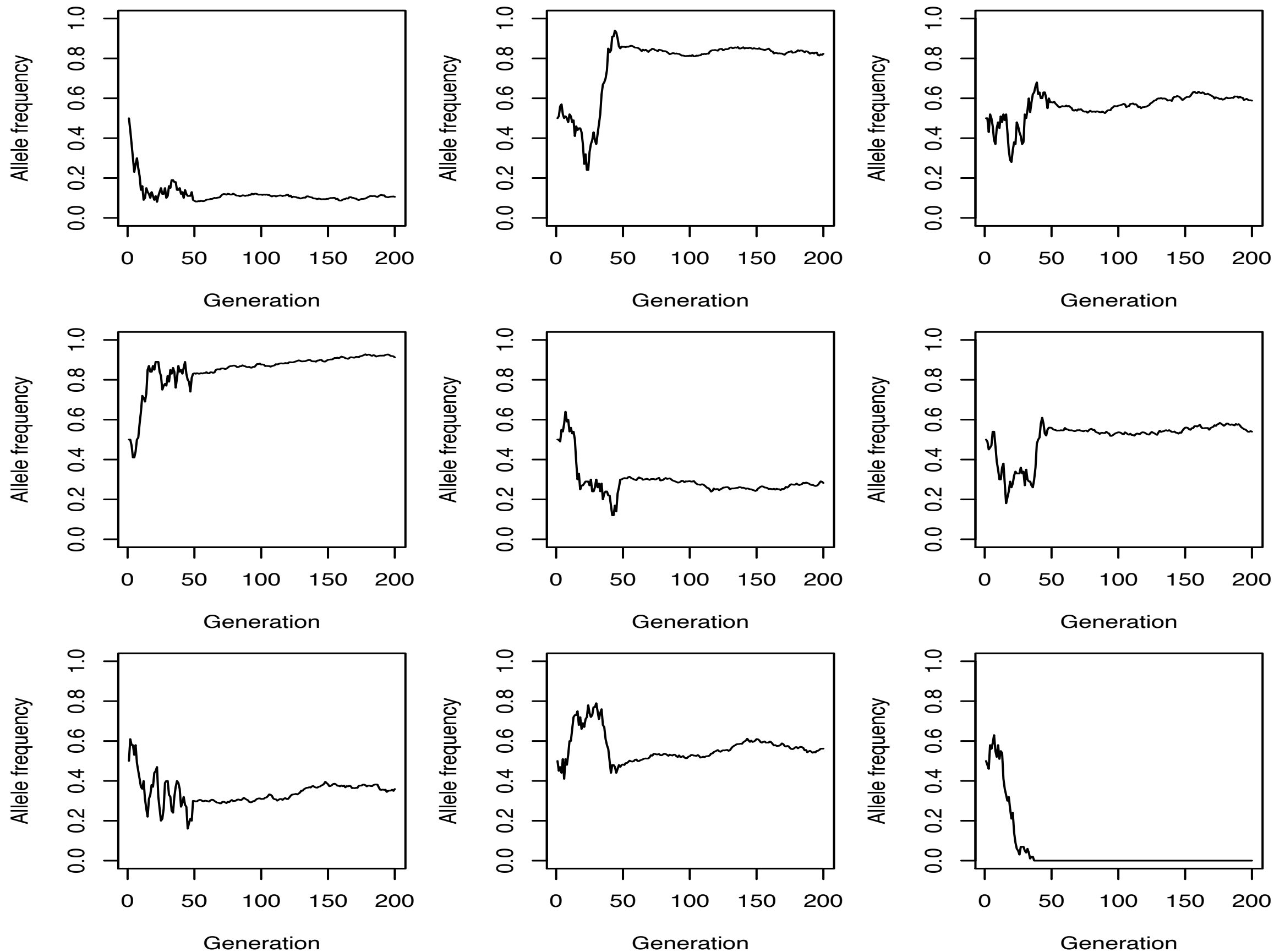
Initial pop size  
Starting frequency  
Generations to simulate  
Gen demographic event happens  
Magnitude of size change

- Run it using:

```
f=WFdemog(100, 0.5, 200, 50, 100)  
plot(f)
```

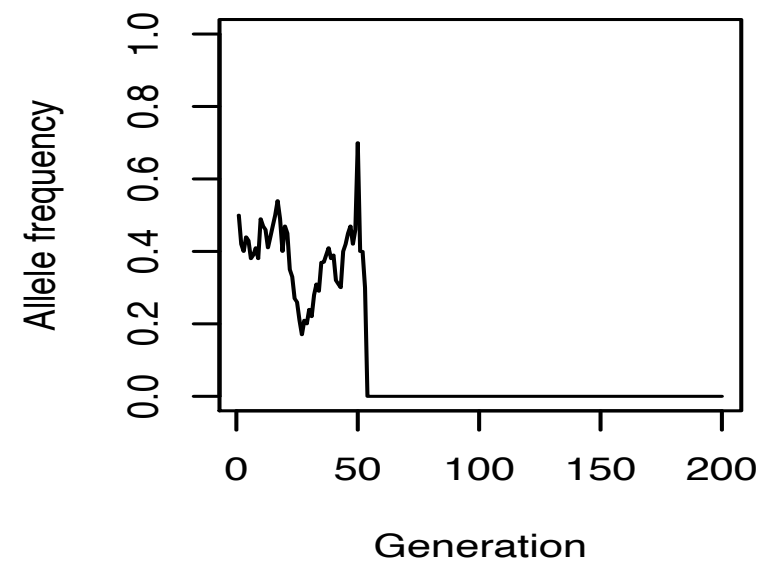
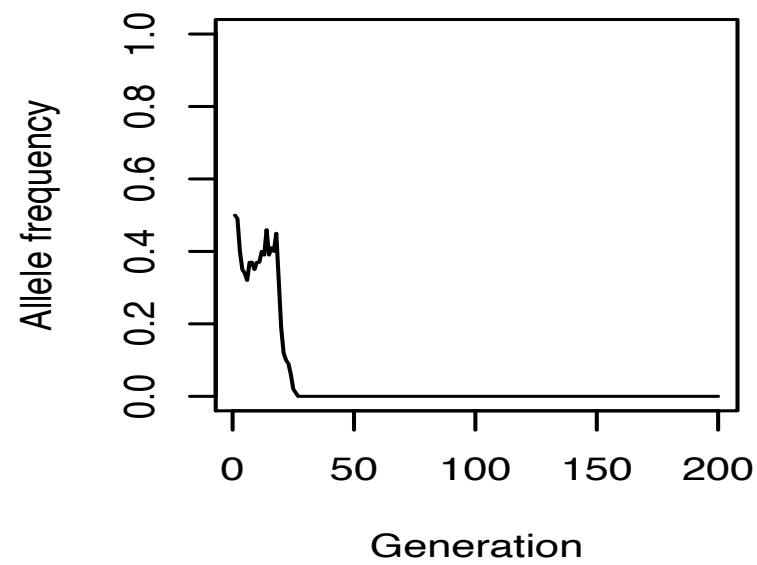
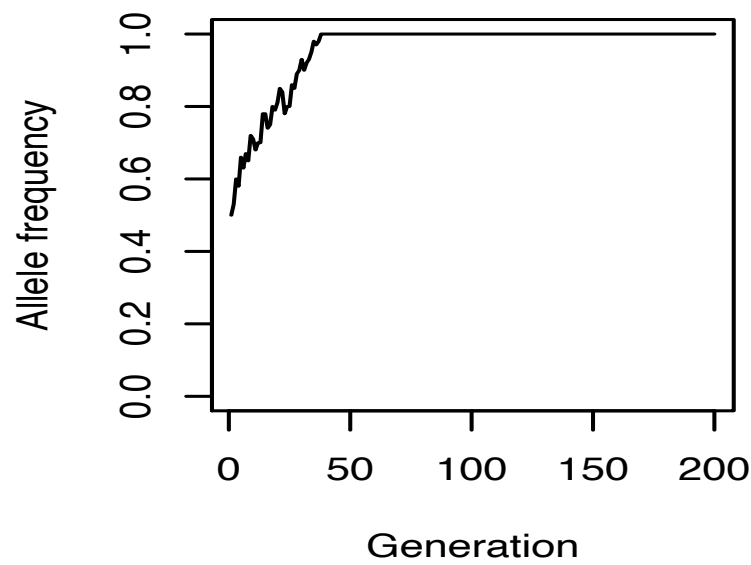
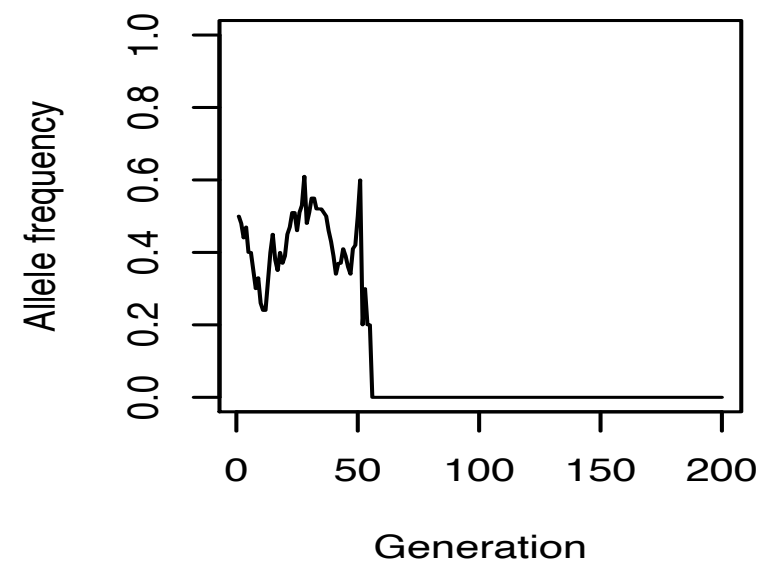
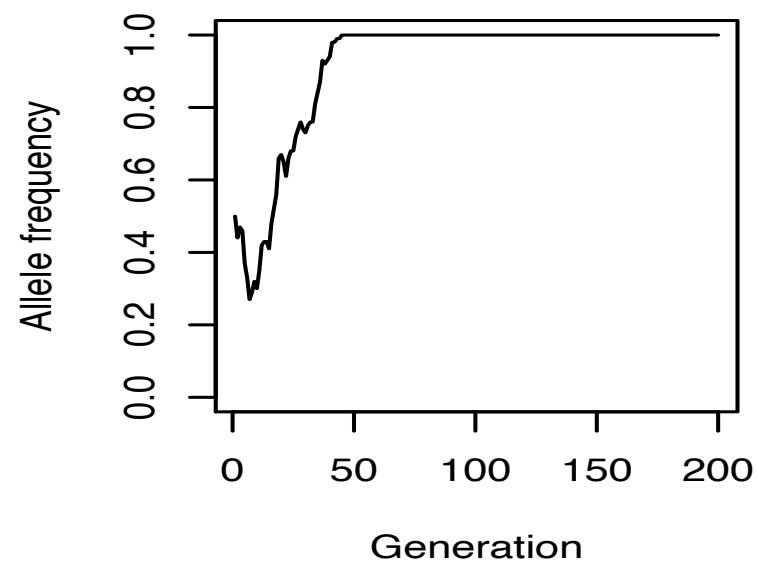
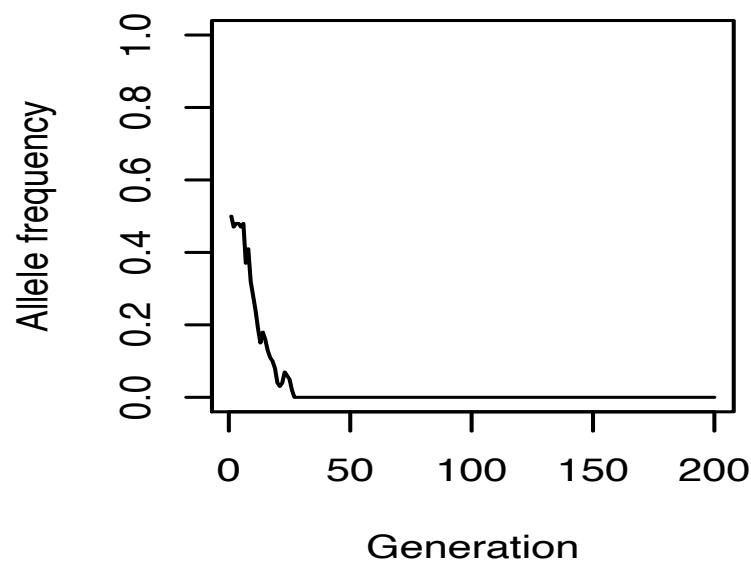
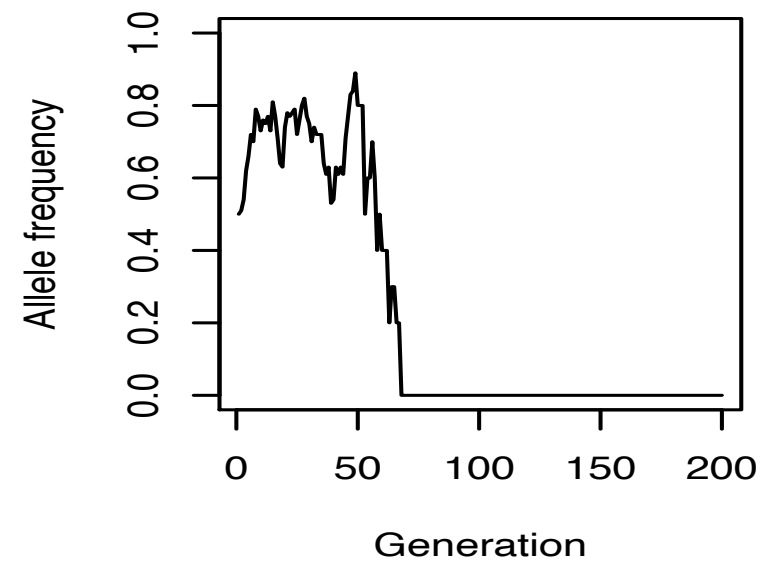
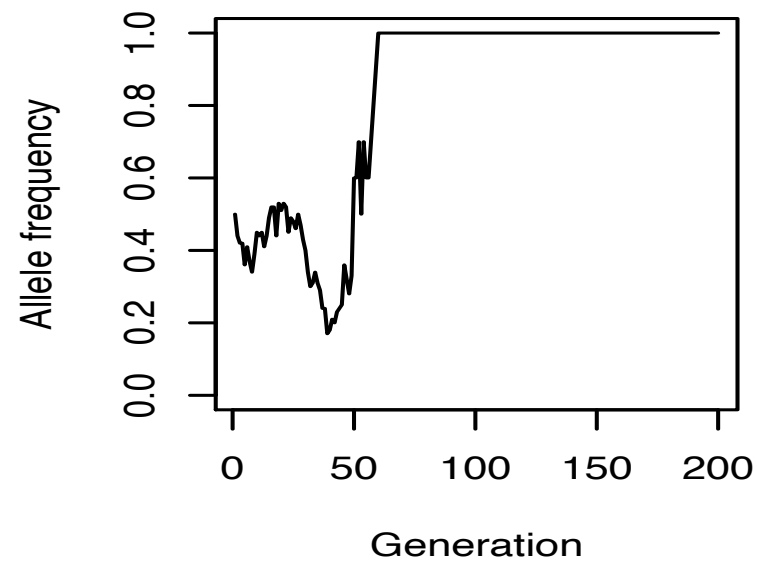
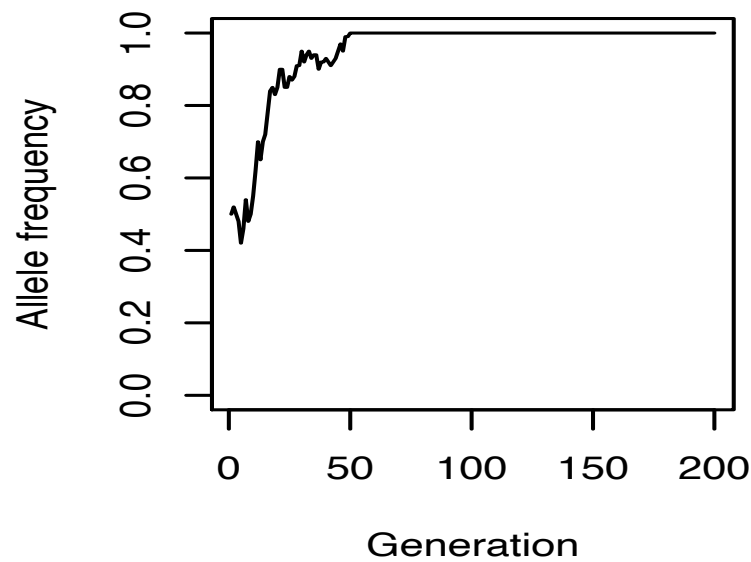
- What happens in your plot?
  - Were you able to get any fixations or losses?
- Can you simulate a 10-fold contraction?
  - How does it change the trajectory?

# Wright-Fisher Model with Expansion



# Wright-Fisher Model with Contraction

- Run it using: `WFdemog(100, 0.5, 200, 50, 0.1)`





# 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?
  - Alleles change frequency!

# Natural Selection

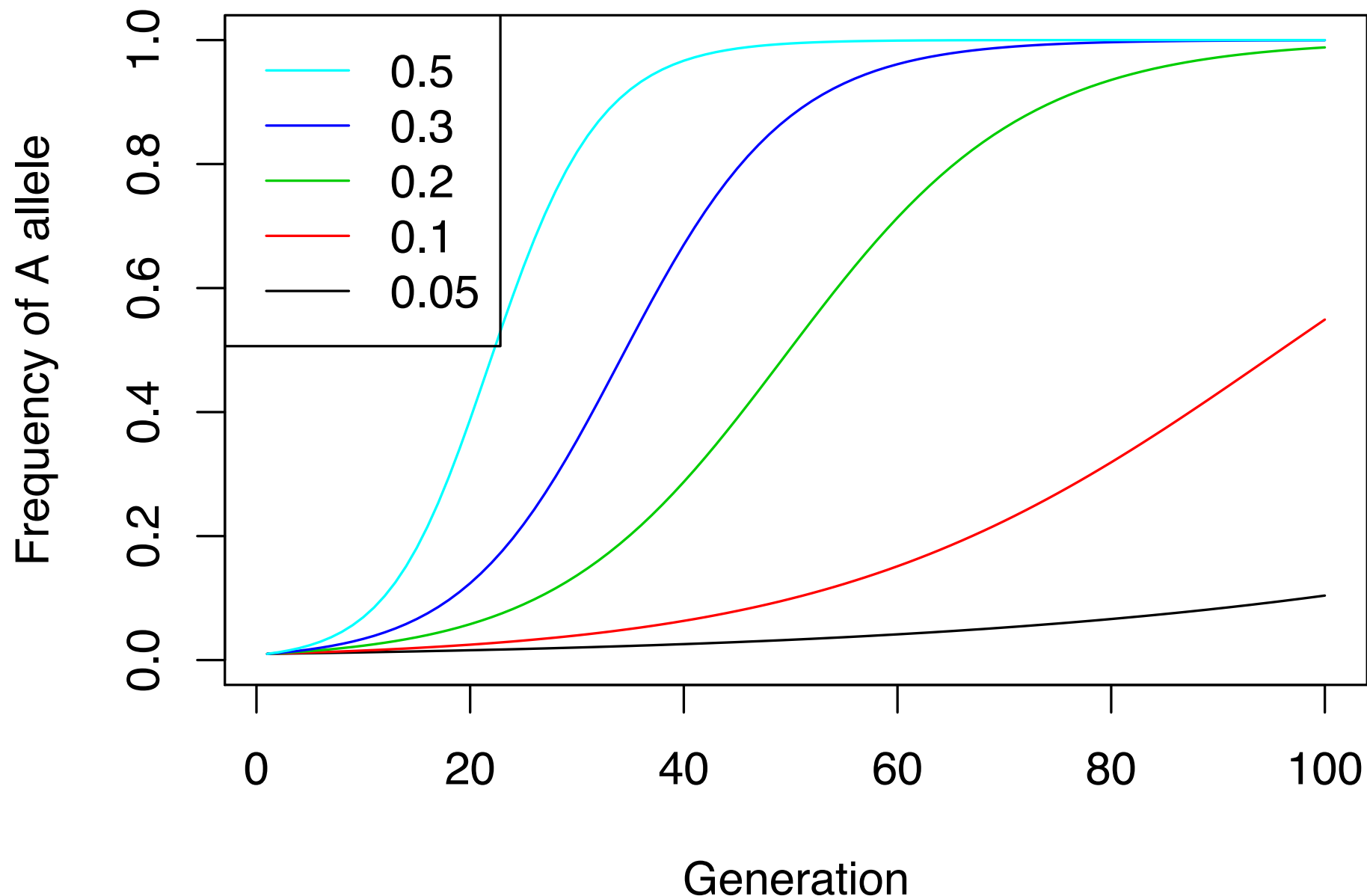
Genotype	AA	Aa	aa
Frequency	$p^2$	$2pq$	$q^2$
Fitness	1	$1+hs$	$1+s$

- 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:

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

# 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!

# Simulating Natural Selection

- First write an R function for the change in allele frequencies:

```
fitfreq = function(q, h, s){  
  p=1-q;  
  return((q^2*(1+s) + p*q*(1+h*s))/(1 + s*q*(2*h*p+q)));  
}
```

- Now use this in an updated WF simulator:

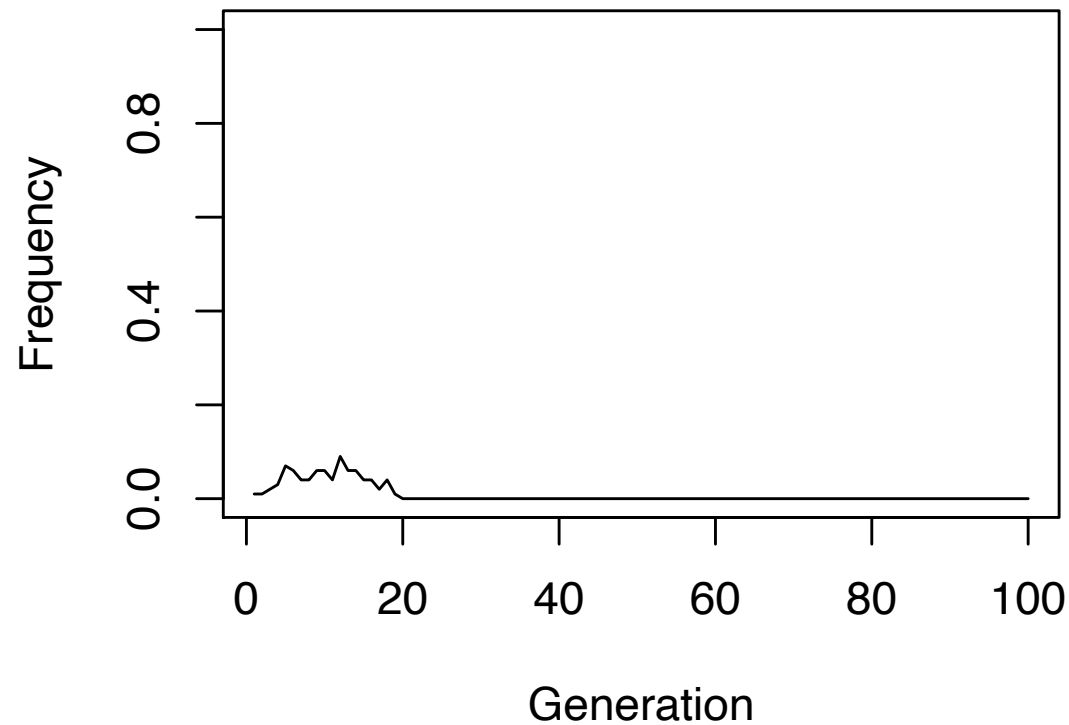
```
WF.sel=function(N, q, h, s, G){  
  t=array(NA, dim=G);  
  t[1] = q;  
  for(i in 2:G){  
    t[i] = rbinom(1, N, fitfreq(t[i-1], h, s))/N;  
  }  
  return(t);  
}
```

- Can you simulate a trajectory for 100 generations with these characteristics:
  - Population size = 100
  - Initial frequency is 1%
  - Allele has a 50% fitness advantage
- What happens in your plot?
  - Were you able to get any fixations or losses?

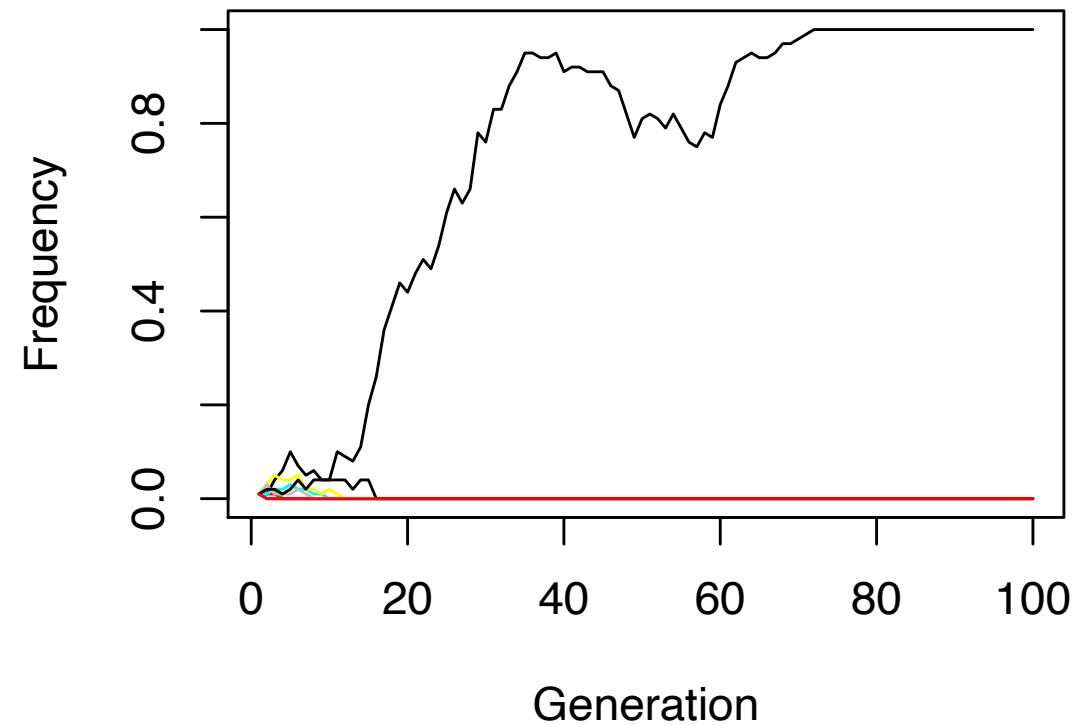
# Natural Selection

`WF.sel(100, 0.01, 0.5, 0.1, 100)`

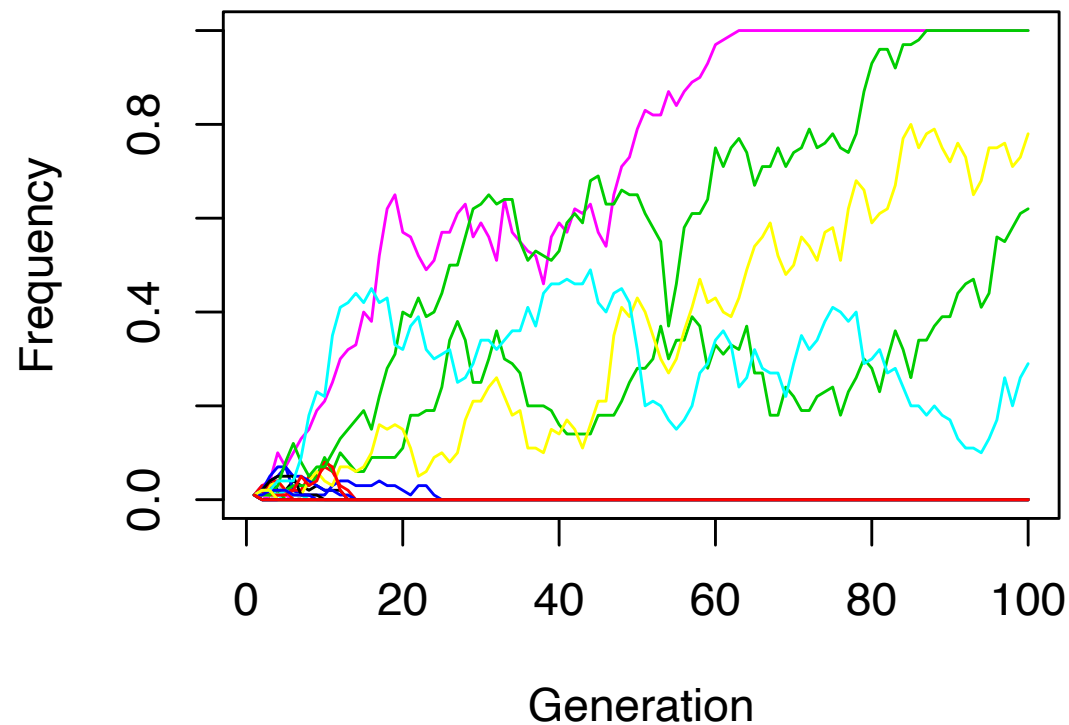
**1 simulations**



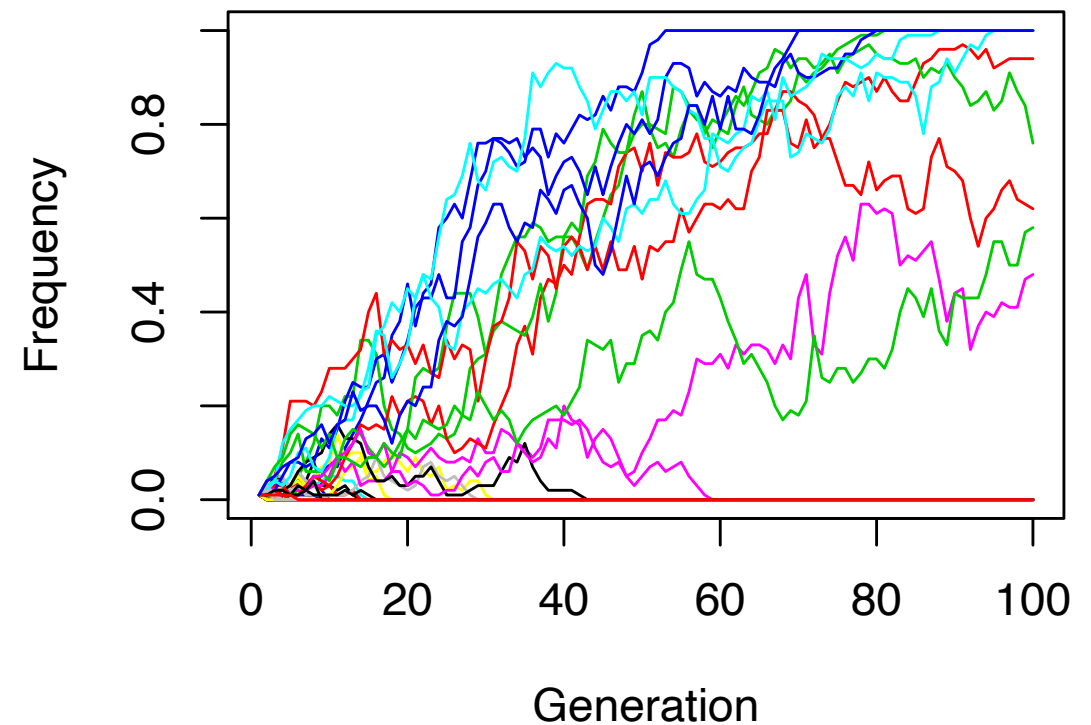
**10 simulations**



**50 simulations**



**100 simulations**



# Simulating Natural Selection

- How would you simulate both selection AND demographic effects?

- Now use this in an updated WF simulator:

```
WF.demsel=function(N, q, h, s, G, Gd, v){  
  t=array(NA,dim=G);  
  t[1] = q;  
  for(i in 2:G){  
    if(i == Gd){  
      N = N*v;  
    }  
    t[i] = rbinom(1, N, fitfreq(t[i-1], h, s))/N;  
  }  
  return(t);  
}
```

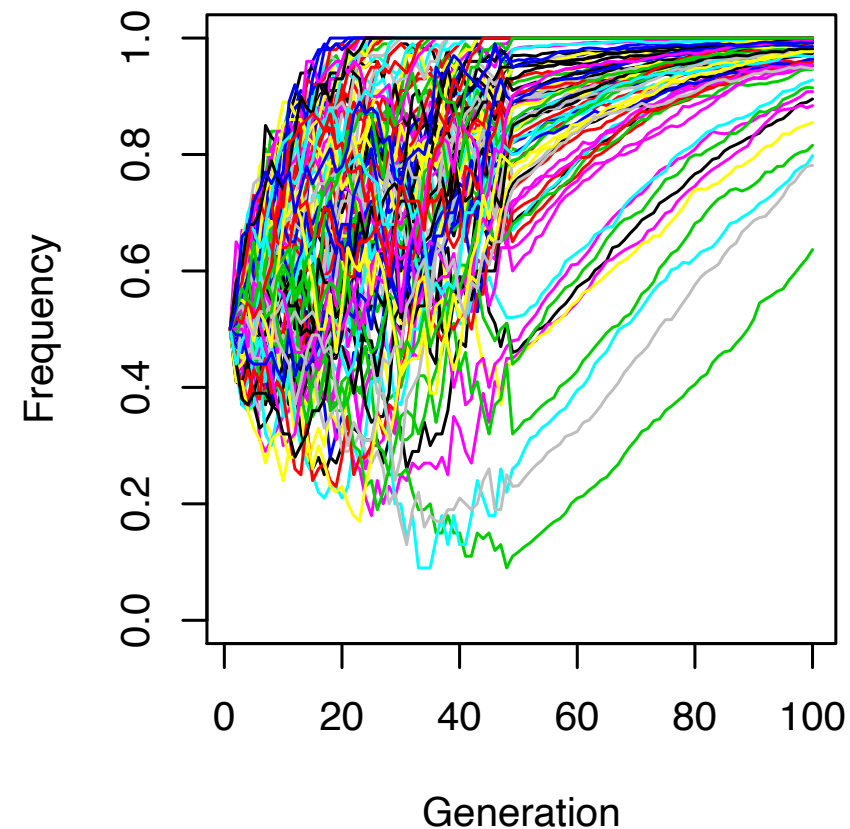
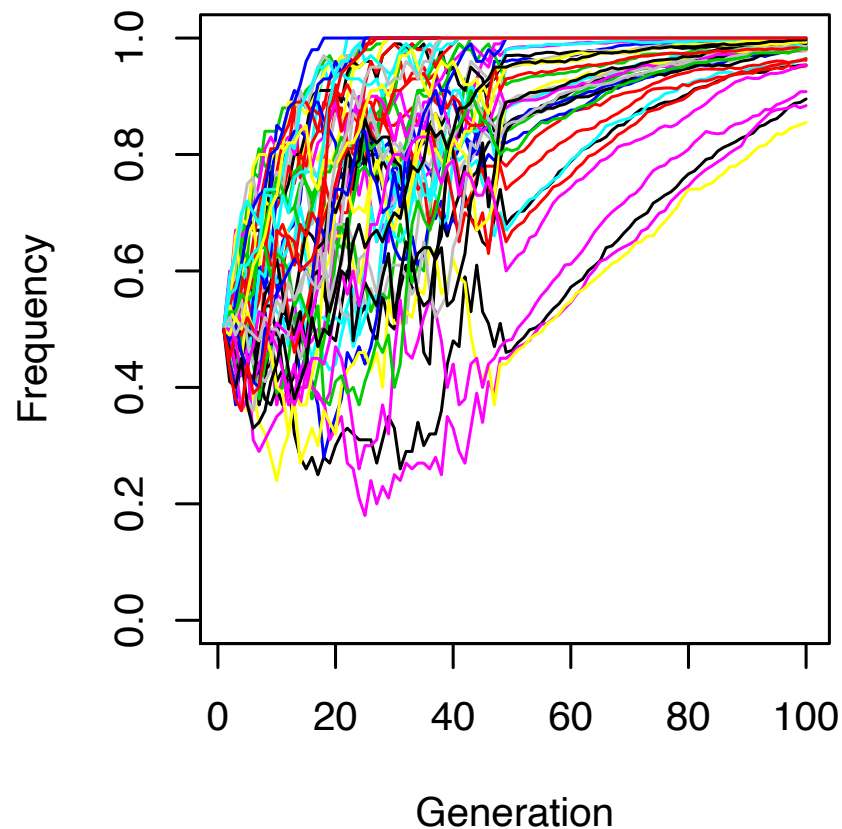
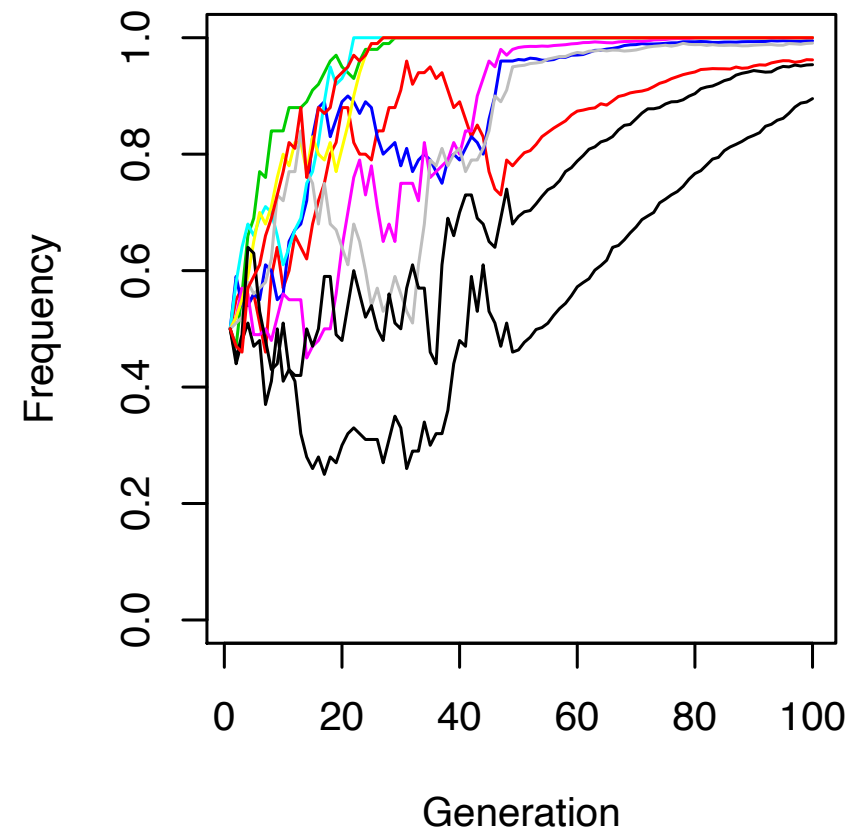
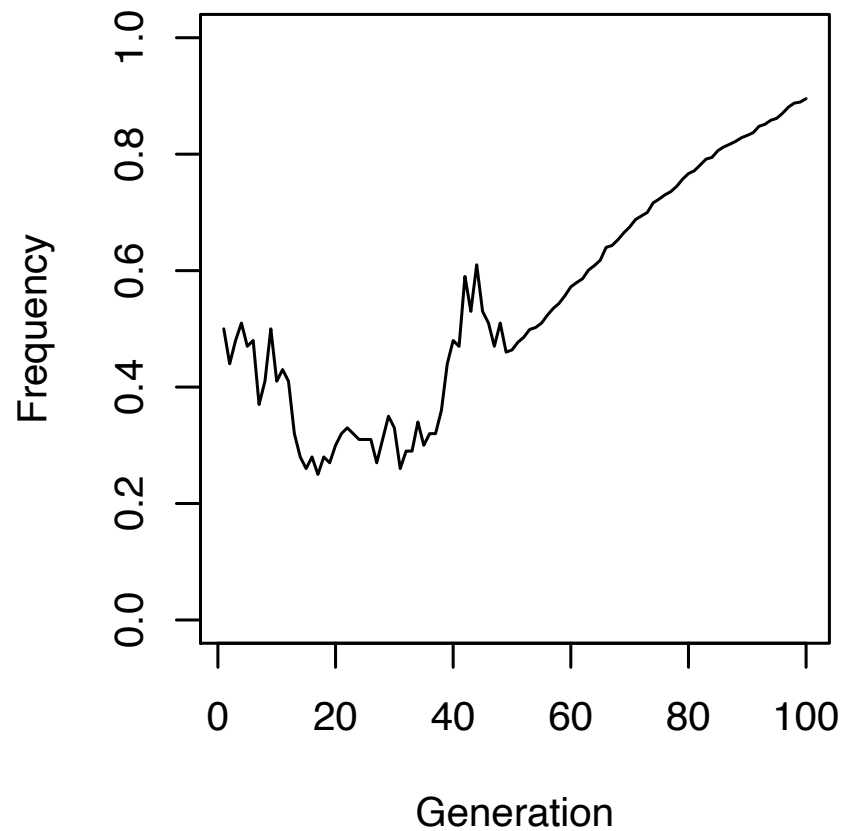
pop size  
initial freq  
dominance  
fitness  
gens to simulate  
Gen demographic event happens  
Magnitude of size change



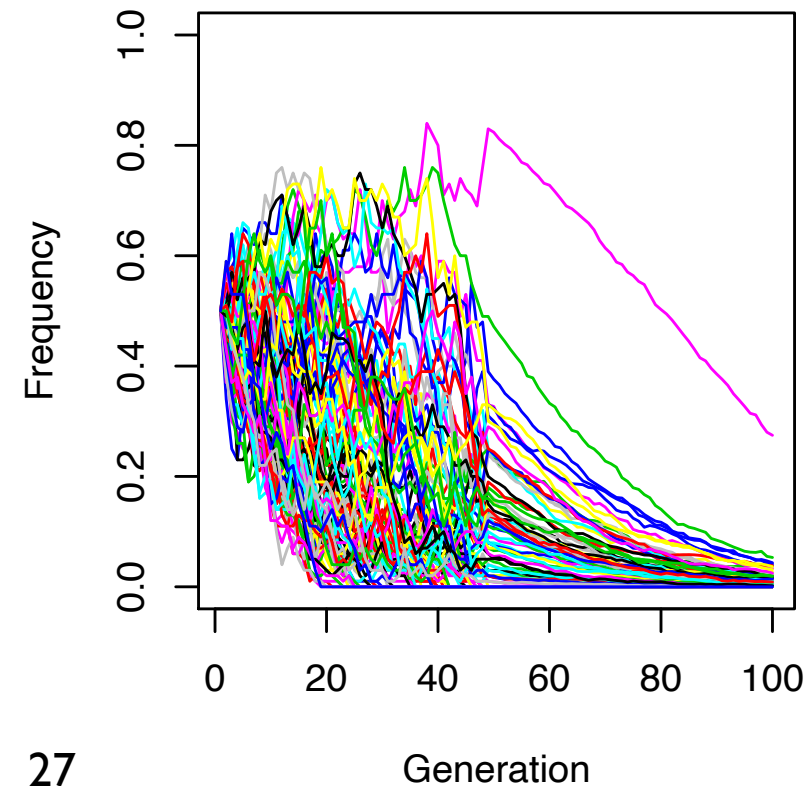
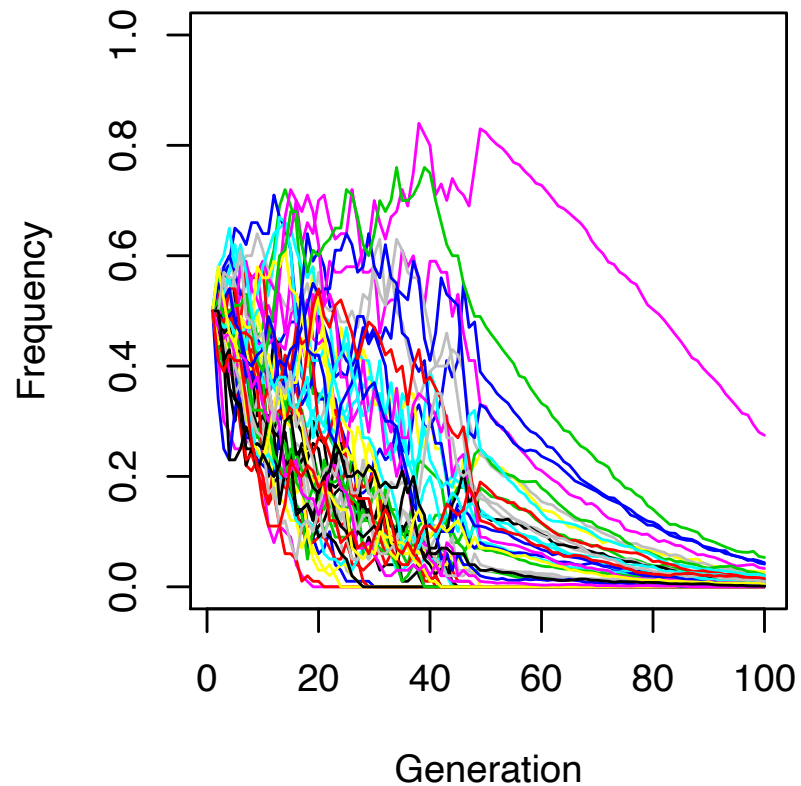
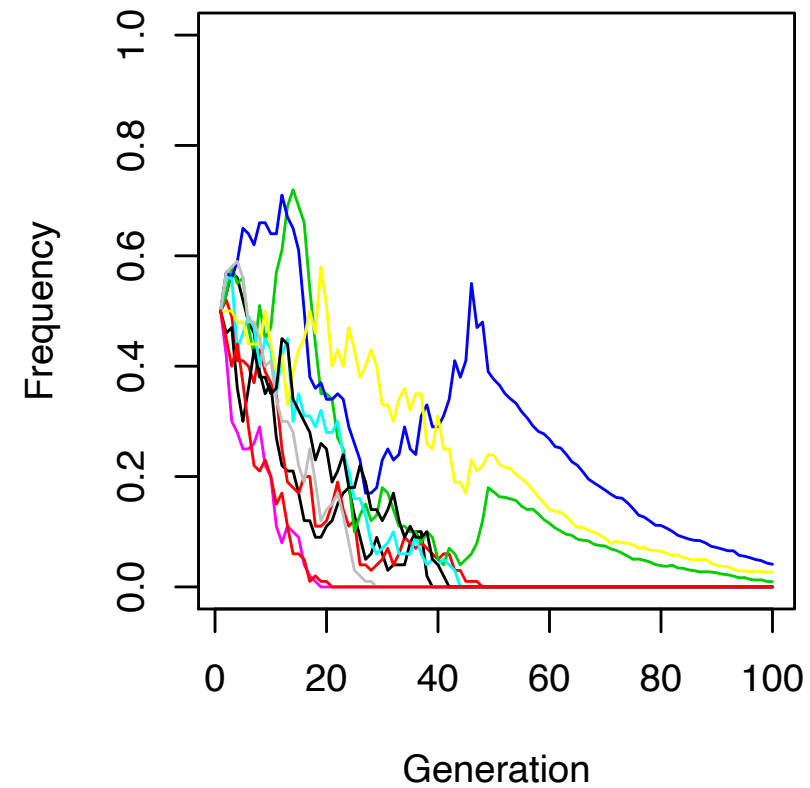
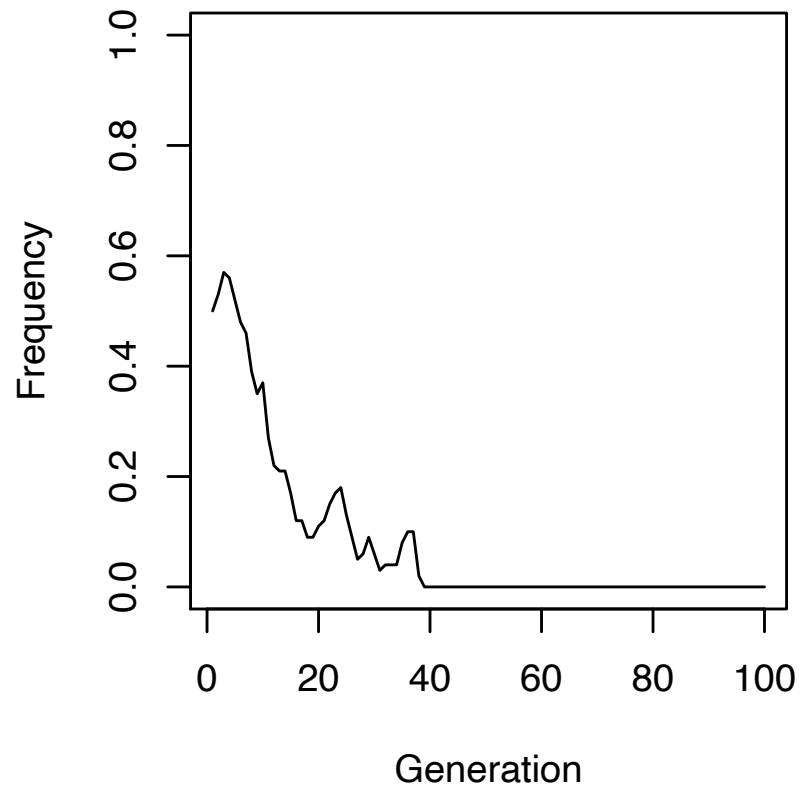
- Can you add 100-fold population growth at generation 50 to your previous simulation?
- What happens in your plot?
- What if the initial frequency is 50%?

# Wright-Fisher Model with Contraction

- Run it using: `WF.demsel(100,0.5,0.5,0.1,100,50,100)`

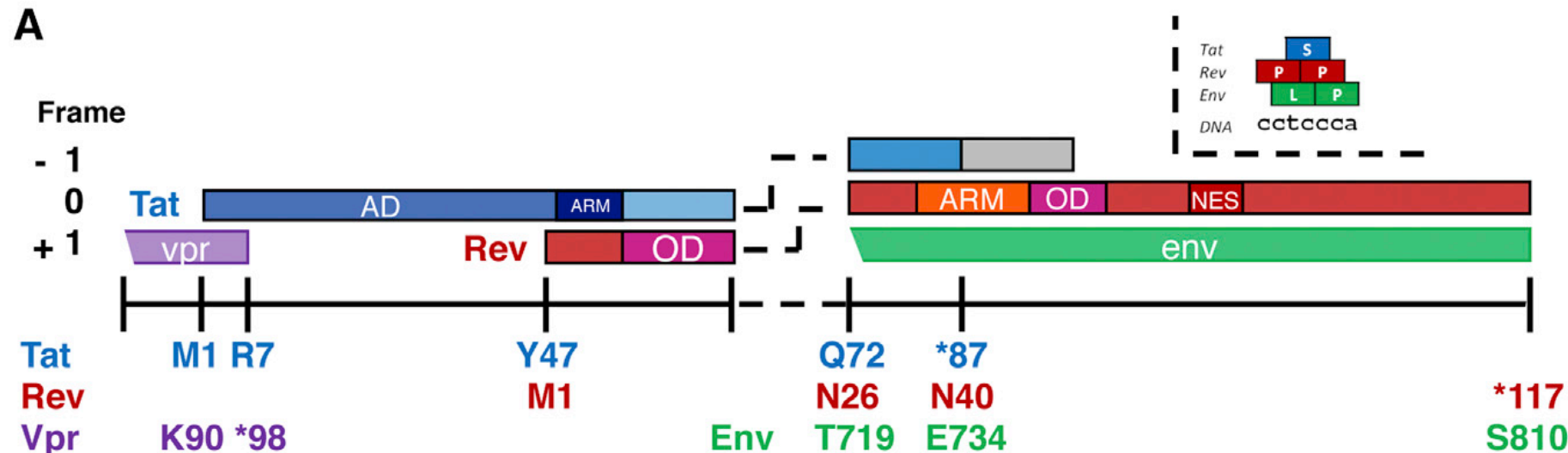


# What parameters generated these?

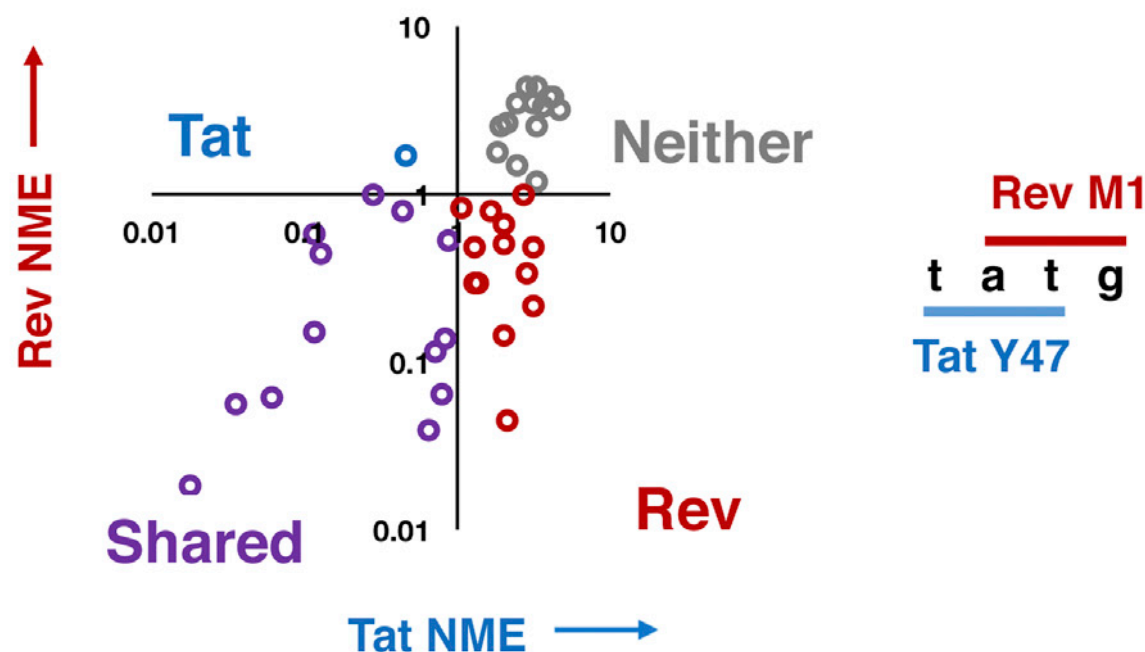


# Functional Segregation of Overlapping Genes in HIV

Jason D. Fernandes,<sup>1,2</sup> Tyler B. Faust,<sup>1,3</sup> Nicolas B. Strauli,<sup>4,5</sup> Cynthia Smith,<sup>1</sup> David C. Crosby,<sup>1</sup> Robert L. Nakamura,<sup>1</sup> Ryan D. Hernandez,<sup>4</sup> and Alan D. Frankel<sup>1,6,\*</sup>

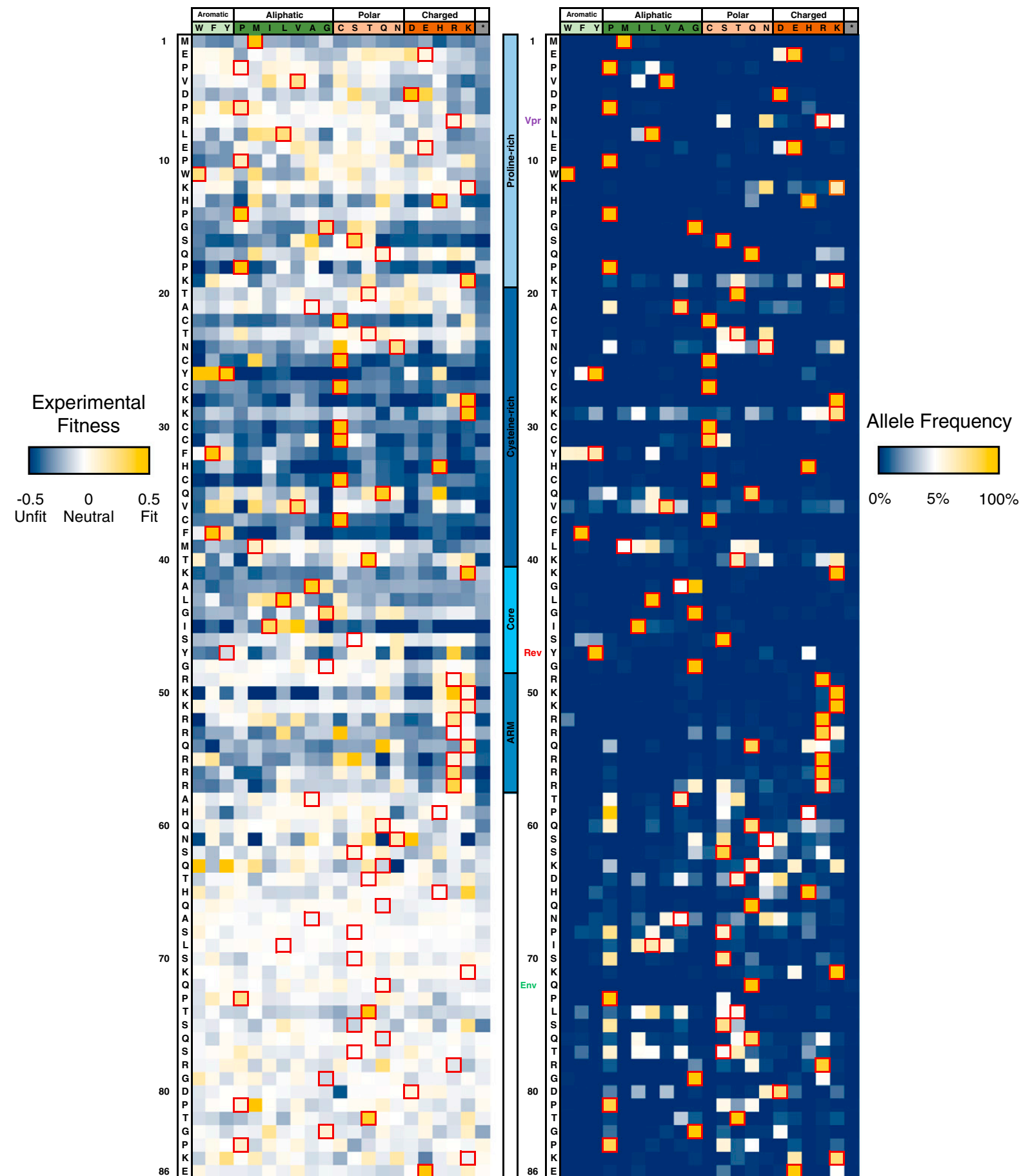


- HIV genes Tat and Rev overlap.
- At protein level, many overlapping sites are conserved in both, but some sites only conserved in Rev.
- Is joint conservation due to dual function or genetic code?



## Non-Overlap Deep Mutational Scanning

## Overlap Patient Conservation



## Functional Segregation of Overlapping Genes in HIV

Jason D. Fernandes,<sup>1,2</sup> Tyler B. Faust,<sup>1,3</sup> Nicolas B. Strauli,<sup>4,5</sup> Cynthia Smith,<sup>1</sup> David C. Crosby,<sup>1</sup> Robert L. Nakamura,<sup>1</sup> Ryan D. Hernandez,<sup>4</sup> and Alan D. Frankel<sup>1,6,\*</sup>

- In patient data, Tat sites that overlap with Rev are highly conserved.
- HIV can be engineered so that Tat and Rev do not overlap
- Deep mutational scanning in non-overlap context (all possible codons at each position) shows that many sites lack conservation in cell lines.
- Is this due to drift (neutral) or selection?

# Functional Segregation of Overlapping Genes in HIV

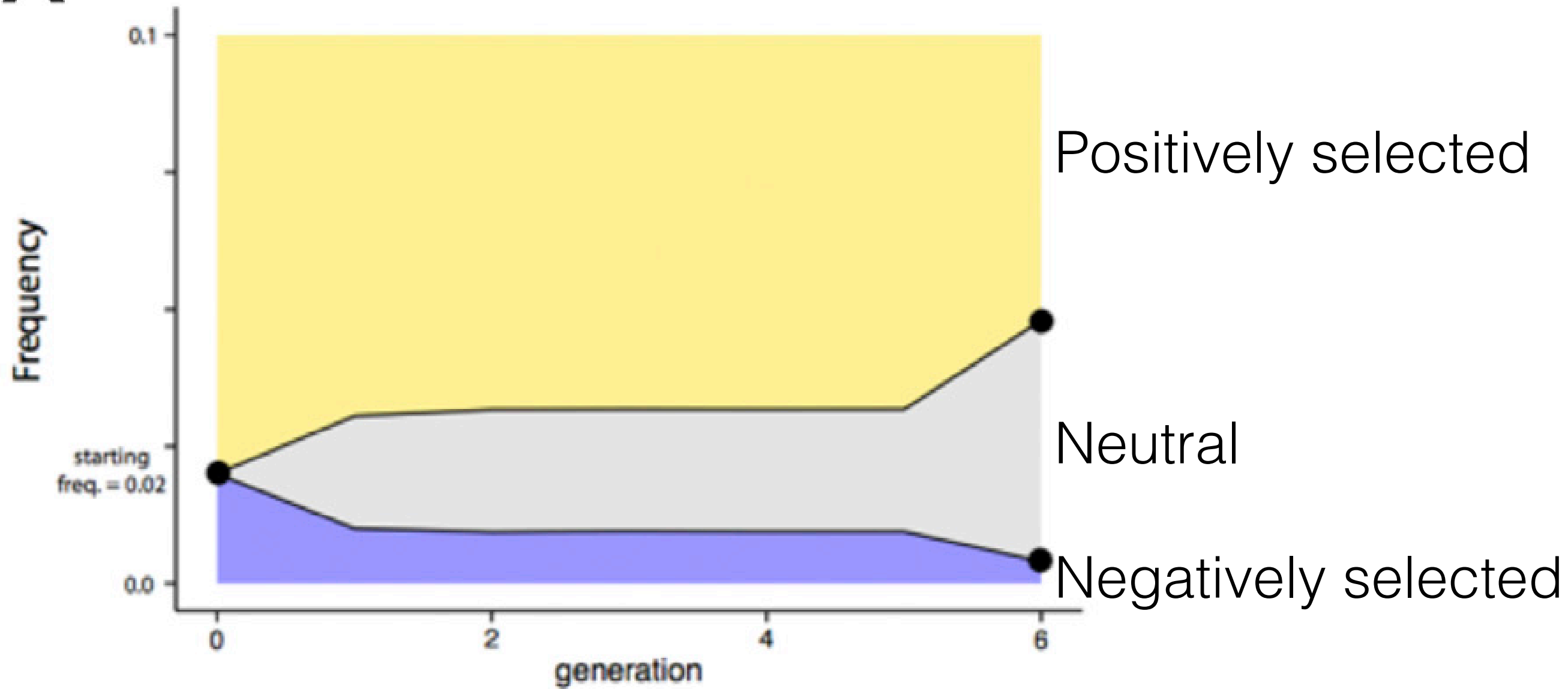
- **Deep mutational scanning:**

- Create exhaustive libraries with all possible codons at all overlapping positions
- Allow population mixture to evolve for **G** generations, then sequence to measure final frequencies of all amino acids
- Simulate to evaluate significance of allele frequency change

- **Factors you might want to include in your simulation:**

- the overall population growth function
- the number of generations
- the starting allele frequency
- the read depth for the experiment

**A**



# 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:  $p_1 = p' = p_0 w_1 / w$ .
- We can then write:

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

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

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



# Natural Selection

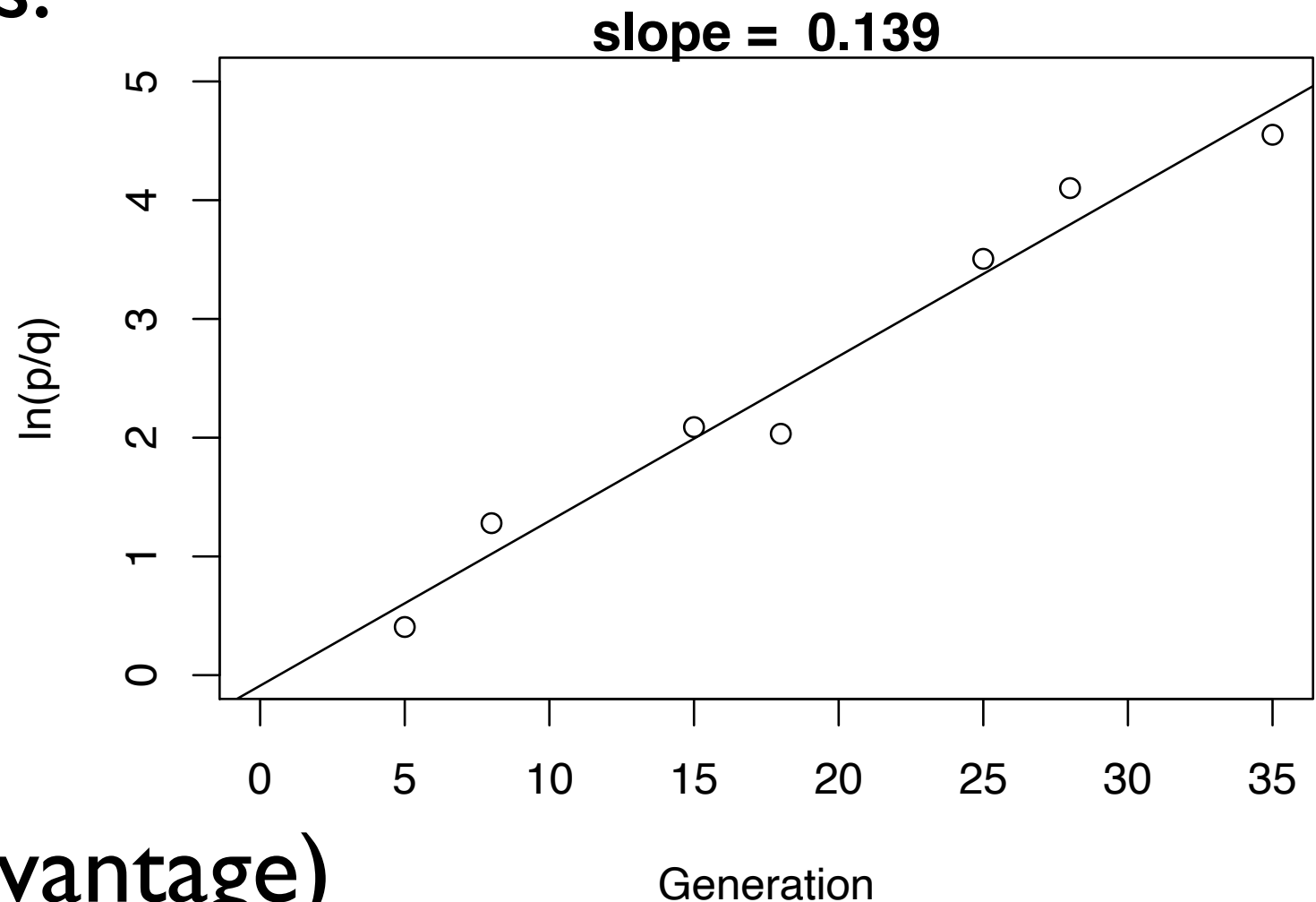
- Taking the natural log of this equation:

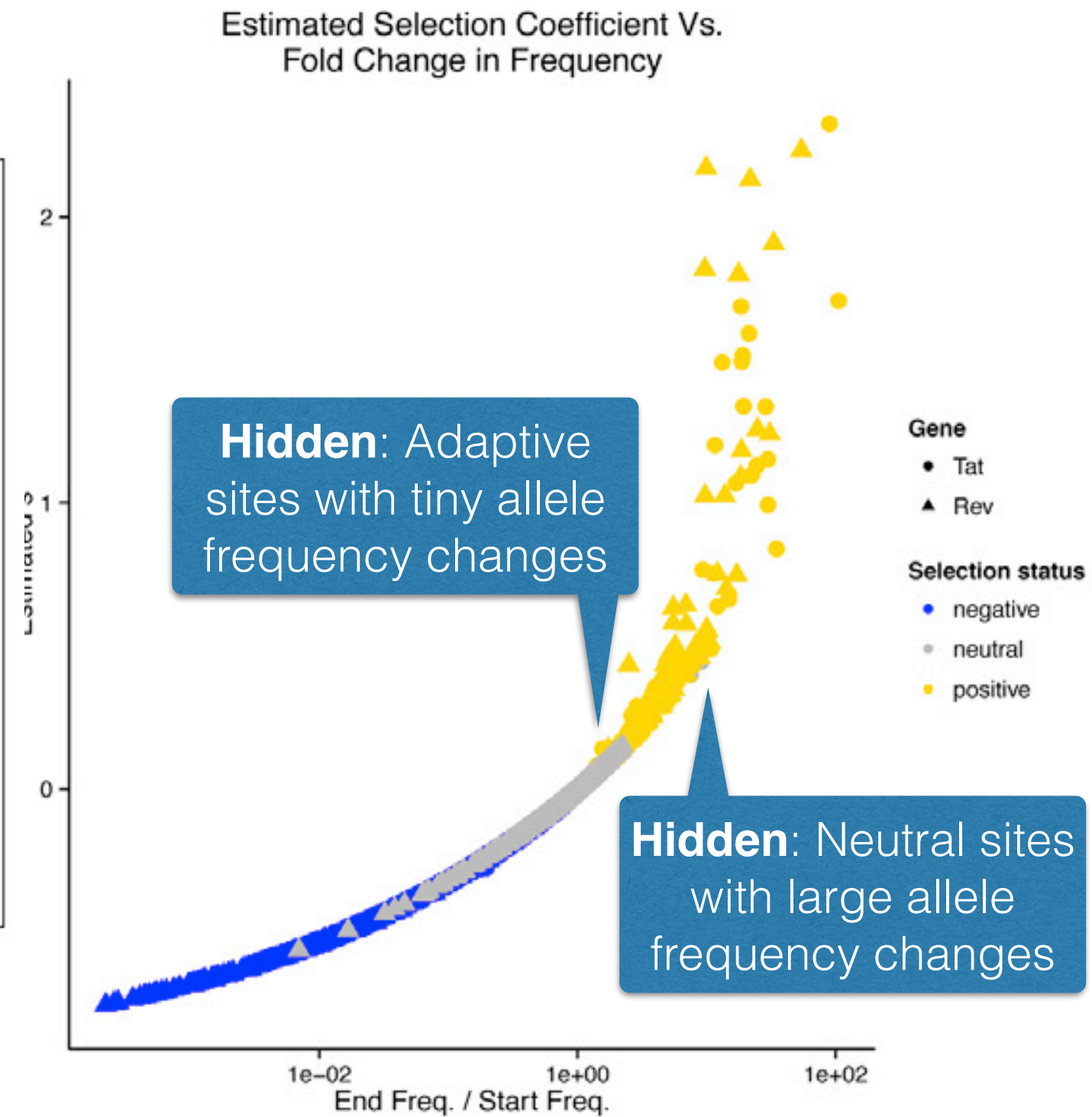
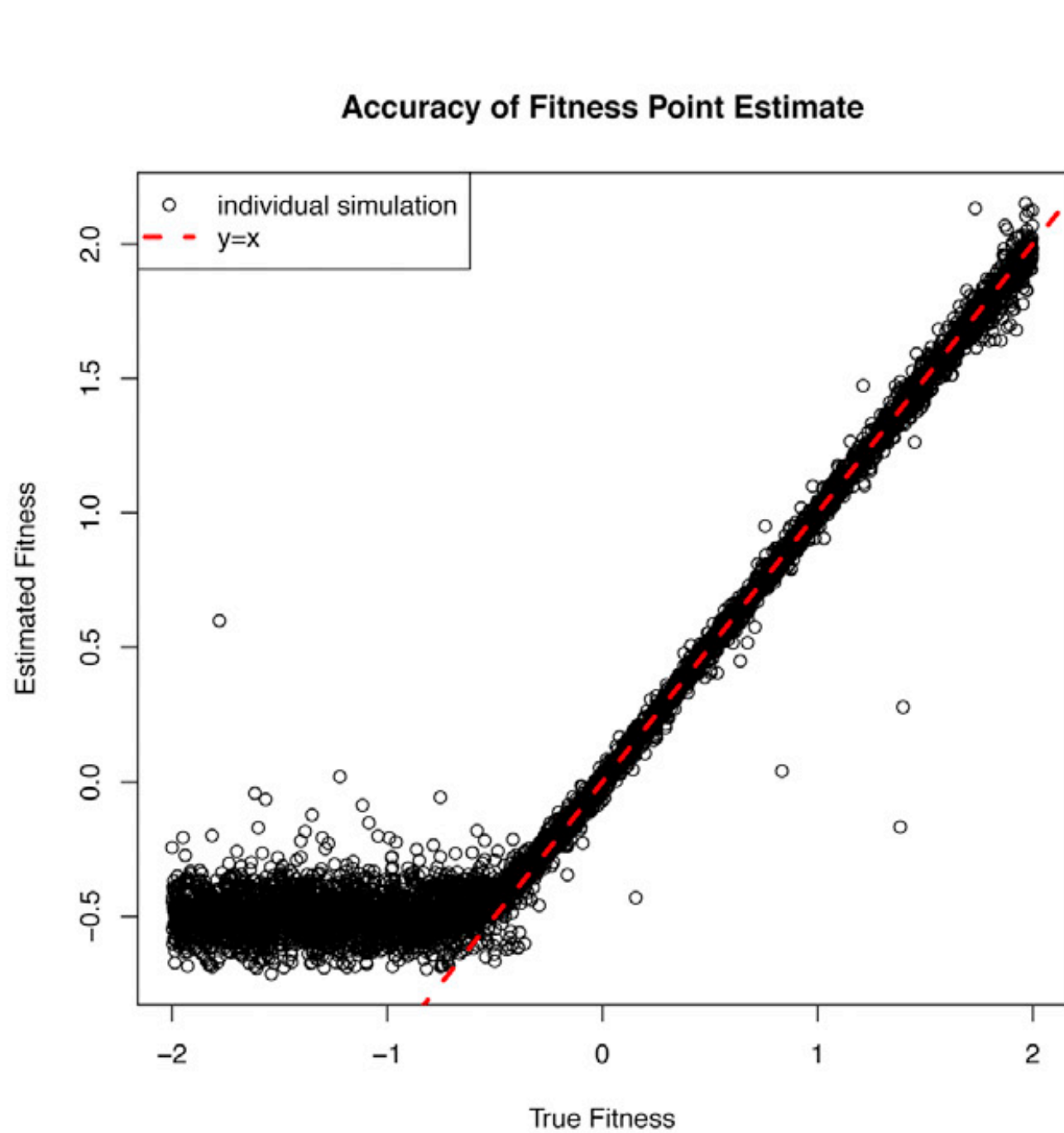
- $$\log \left( \frac{p_t}{q_t} \right) = \log \left( \frac{w_1}{w_2} \right) t + \log \left( \frac{p_0}{q_0} \right)$$

- Which is now a **linear function** of  $t$ , the number of generations.
- We can now estimate the ratio of fitnesses by regression!

# 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:  
 $w_1/w_2 = e^{0.139} = 1.15$
- Assume  $w_2=1$ .
- Thus, allele p has a 15% fitness advantage over allele q!
- (simulated with 20% advantage)





# Existing forward simulators

- **SFS\_CODE: Hernandez (2008)**
  - Command-line flexibility... shameless plug!
- **FWDPP: Thornton (2014)**
  - C++ library of routines intended to facilitate the development of forward-time simulations under arbitrary mutation and fitness models
- **SLiM: Now you know about this!**
  - Command-line, GUI, and R-like scripting environment that provides control over most aspects of the simulated evolutionary scenarios