Introduction to Wright-Fisher Simulations

Ryan Hernandez
Goals

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

• Start with single sites, and build in multiple sites
Hardy-Weinberg Principle

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

- Conclusion 1:
  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

\[
P = p^2 \\
Q = 2p(1-p) \\
R = (1-p)^2
\]
Hardy-Weinberg Principle

- Imagine a population of diploid individuals

\[ P = 0.81 \quad Q = 0.18 \quad R = 0.01 \]
Hardy-Weinberg Principle

- Imagine a population of diploid individuals

\[ P = \frac{0.5}{0.4} = 0.1 \]

\[ Q = 0.1 \]

\[ R = 0.4 \]

\[ p = P + Q/2 = 0.55 \]

\[ p^2 = 0.3025 \]

\[ 2p(1 - p) = 0.495 \]

\[ (1 - p)^2 = 0.2025 \]

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

- 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

Godfrey H. Hardy: 1877-1947
Wilhelm Weinberg: 1862-1937
Wright-Fisher Model

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

$$P(X(t + 1) = j | X(t) = i) = \binom{N}{j} p^j (1 - p)^{N-j}$$

$$= \text{Bin}(j | N, i/N) = \binom{N}{j} \left( \frac{i}{N} \right)^j \left( \frac{N - i}{N} \right)^{N-j}$$
Wright-Fisher Model

- A simple R function to simulate genetic drift:

```
WF=function(N, p, G){
  t=array(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)
```
Wright-Fisher Model

![Graphs showing allele frequency over generations.](image-url)
Demographic Effects

• Population changes size at a given generation
• Suppose a population of $N$ individuals.

• Let $X(t)$ be the #chromosomes carrying an allele A in generation $t$:

$$P(X(t+1) = j|X(t) = i) = \binom{N}{j} p^j (1 - p)^{N-j}$$

$$= \text{Bin}(j|N, i/N) = \binom{N}{j} \left( \frac{i}{N} \right)^j \left( \frac{N - i}{N} \right)^{N-j}$$
Wright-Fisher Model

- A simple R function to simulation demographic effects:

```r
WFdemog = function(N, p, G, Gd, v){
  t=array(,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);
}
```

- Run it using:

```r
f=WFdemog(100, 0.5, 200, 50, 100)
plot(f)
```
Wright-Fisher Model with Contraction

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

• What happens when we allow natural selection to occur?

• Alleles change frequency!

• Identical frequencies in males/females
• Infinite population size
• No migration
• No mutation
• No natural selection
### Natural Selection

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>$p^2$</td>
<td>$2pq$</td>
<td>$q^2$</td>
</tr>
<tr>
<td>Fitness</td>
<td>1</td>
<td>1+$hs$</td>
<td>1+$s$</td>
</tr>
</tbody>
</table>

- 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!
Simulating Natural Selection

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

  ```r
  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:

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

\( \text{WF}\_\text{sel}(100, 0.01, 0.5, 0.1, 100) \)
Natural Selection

WF.sel(100, 0.01, 0.5, 0.1, 100)
Simulating Natural Selection

• How would you simulate both selection AND demographic effects?

• Now use this in an updated WF simulator:

```r
WF.demsel=function(N, q, h, s, G, Gd, v){
  t=array(,dim=G);
  t[1] = N*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);
}
```
Wright-Fisher Model with Contraction

- Run it using: \texttt{WF.demsel(100,0.5,0.5,0.1,100,50,100)}
Wright-Fisher Model with Contraction

- Run it using: \texttt{WF.demsel(100,0.5,0.5,0.1,100,50,100)}
What parameters generated these?
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?
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 ending read depth for the experiment
  • and the amino acid identity of the allele
Figure S5. Related to Figures 3 and 4

(A) An illustration of the neutral simulations for a hypothetical allele with a starting frequency of 0.02, an ending read depth of 500 reads, and an amino acid identity of Arginine. The gray area depicts the range of trajectories that this allele could take if it were neutral. If an ending allele frequency were observed to be above or below this neutral expectation, it is deemed positively or negatively selected, respectively. The black dots indicate the upper and lower bounds for the ending allele frequency that would still be considered neutral. These upper and lower bounds correspond to relative fitness values of 0.380 and 0.699, respectively, which means neutrality cannot be rejected for any observed fitness value that resides between this interval.

(B) The observed distribution of fitness values for alleles found to be under negative (blue), neutral (gray), or positive (gold) selection. Neutral alleles were sometimes found to have relatively extreme fitness estimates (left and right tails of gray distribution). Likewise, alleles under significant positive or negative selection were sometimes found to have fitness estimates close to zero (right tail of blue distribution, and left tail of gold distribution).

(C) Correlation between the estimated fitness and the true fitness under our simulation framework. Each point corresponds to one simulation.

(D) Correlation plot of alleles between biological replicates. Alleles that have strong experimental evidence of selection show high repeatability. Correlation coefficients are shown for the whole data-set and those passing our QC criteria (note that these criteria have inherent biases, such as requiring the sign of the selection coefficient to be consistent, toward increasing the correlation coefficient). The experimental error rate (amino acids that appear to mutate outside the randomization site) is 2.24%. Note that this error rate is calculated on the amino acid level and does not consider synonymous mutations, or multiple mutations within the same codon.

Fernandez, et al., Cell (2016)
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: $p_1 = p' = p_0 \frac{w_1}{w}$.
- We can then write:

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

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

$$\frac{p_t}{q_t} = \frac{p_0 w_1}{q_0 w_2} = \left( \frac{p_0}{q_0} \right) \left( \frac{w_1}{w_2} \right)^t$$
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)
Figure S5. Related to Figures 3 and 4

(A) An illustration of the neutral simulations for a hypothetical allele with a starting frequency of 0.02, an ending read depth of 500 reads, and an amino acid identity of Arginine. The gray area depicts the range of trajectories that this allele could take if it were neutral. If an ending allele frequency were observed to be above or below this neutral expectation, it is deemed positively or negatively selected, respectively. The black dots indicate the upper and lower bounds for the ending allele frequency that would still be considered neutral. These upper and lower bounds correspond to relative fitness values of 0.380 and 0.699, respectively, which means neutrality cannot be rejected for any observed fitness value that resides between this interval.

(B) The observed distribution of fitness values for alleles found to be under negative (blue), neutral (gray), or positive (gold) selection. Neutral alleles were sometimes found to have relatively extreme fitness estimates (left and right tails of gray distribution). Likewise, alleles under significant positive or negative selection were sometimes found to have fitness estimates close to zero (right tail of blue distribution, and left tail of gold distribution).

(C) Correlation between the estimated fitness and the true fitness under our simulation framework. Each point corresponds to one simulation.

(D) Correlation plot of alleles between biological replicates. Alleles that have strong experimental evidence of selection show high repeatability. Correlation coefficients are shown for the whole data-set and those passing our QC criteria (note that these criteria have inherent biases, such as requiring the sign of the selection coefficient to be consistent, toward increasing the correlation coefficient). The experimental error rate (amino acids that appear to mutate outside the randomization site) is 2%. Note that this error rate is calculated on the amino acid level and does not consider synonymous mutations, or multiple mutations within the same codon.

Hidden: Adaptive sites with tiny allele frequency changes

Hidden: Neutral sites with large allele frequency changes

Fernandez, et al., Cell (2016)
Existing forward simulators

- **SFS_CODE**: Hernandez (2008)
  - Command-line flexibility… shameless plug!

- **SLIM 2**: Haller & Messer (2017)
  - R-like scripting environment that provides control over most aspects of the simulated evolutionary scenarios

- **FWDPP**: Thornton (2014)
  - C++ library of routines intended to facilitate the development of forward-time simulations under arbitrary mutation and fitness models