Lecture 18 - Selection and Tests of Neutrality

- Gibson and Muse, chapter 5
- Nei and Kumar, chapter 12.6 p. 258-264
- Hartl, chapter 3, p. 122-127
The Usefulness of Theta

- Under evolution by genetic drift (i.e., neutral evolution), each estimator of theta is an unbiased estimator of the true value of theta:

\[ E(\hat{\theta}_W) = E(\hat{\theta}_\pi) = \theta \]

- Therefore, differences between estimates from each estimator can be used to infer non-neutral evolution (i.e., natural selection).
- Keep in mind that there are other estimators of theta as well. Each estimator, however, has the same expectation (i.e., the true value of theta)
Neutrality Tests Using Theta

Given that we observe a difference between estimates of theta using different estimators, can we perform a test for selection using this information?

Tajima’s $D$ (Tajima, 1989)

Calculate $D$ and perform a statistical test of the form:

$H_0$: $D = 0$ (No selection)

$H_A$: $D \neq 0$ (selection)
Tajima’s D - The Test Statistic

From the observed data, calculate the following:

\[ D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\text{var}(\hat{\theta}_\pi - \hat{\theta}_W)}} \]

\[
\text{var}(\hat{\theta}_\pi - \hat{\theta}_W) = \frac{n+1}{3(n-1)} - \frac{1}{\sum_{i=1}^{n-1}[1/i]} + \frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{n\sum_{i=1}^{n-1}[1/i]} + \frac{\sum_{i=1}^{n-1}[1/i^2]}{\left[\sum_{i=1}^{n-1}[1/i]\right]^2 + \sum_{i=1}^{n-1}[1/i^2]} S(S-1)
\]

If D is too negative or too positive, then H\(_0\) is rejected. Too positive or too negative is determined through simulation of DNA sequence samples under neutral evolution.
Nucleotide diversity in loblolly pine candidate genes for drought

<table>
<thead>
<tr>
<th>Gene</th>
<th>n</th>
<th>Total</th>
<th>Synonymous</th>
<th>Non-synonymous</th>
<th>Silent (non-coding + synonymous)</th>
<th>Tajima’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S</td>
<td>θ</td>
<td>π</td>
<td>S</td>
<td>θ</td>
</tr>
<tr>
<td>lp3-1</td>
<td>32</td>
<td>18</td>
<td>8.77</td>
<td>12.70</td>
<td>1</td>
<td>9.31</td>
</tr>
<tr>
<td>lp3-3</td>
<td>32</td>
<td>3</td>
<td>1.59</td>
<td>0.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dhn-1</td>
<td>32</td>
<td>13</td>
<td>5.08</td>
<td>4.15</td>
<td>4</td>
<td>8.88</td>
</tr>
<tr>
<td>dhn-2</td>
<td>31</td>
<td>14</td>
<td>6.64</td>
<td>7.76</td>
<td>6</td>
<td>15.31</td>
</tr>
<tr>
<td>mtl-like</td>
<td>32</td>
<td>9</td>
<td>5.55</td>
<td>5.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sod-chl</td>
<td>32</td>
<td>19</td>
<td>6.88</td>
<td>7.80</td>
<td>2</td>
<td>12.61</td>
</tr>
<tr>
<td>ferritin</td>
<td>32</td>
<td>7</td>
<td>2.92</td>
<td>1.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rd21A-2</td>
<td>31</td>
<td>26</td>
<td>7.02</td>
<td>7.68</td>
<td>7</td>
<td>13.28</td>
</tr>
<tr>
<td>sams-2</td>
<td>32</td>
<td>6</td>
<td>2.75</td>
<td>3.60</td>
<td>2</td>
<td>6.07</td>
</tr>
<tr>
<td>pal-1</td>
<td>31</td>
<td>6</td>
<td>3.81</td>
<td>2.62</td>
<td>1</td>
<td>4.13</td>
</tr>
<tr>
<td>ccoaomt-1</td>
<td>32</td>
<td>13</td>
<td>6.68</td>
<td>11.86</td>
<td>4</td>
<td>18.07</td>
</tr>
<tr>
<td>cpk3</td>
<td>32</td>
<td>8</td>
<td>3.16</td>
<td>3.55</td>
<td>3</td>
<td>9.49</td>
</tr>
<tr>
<td>pp2c</td>
<td>32</td>
<td>1</td>
<td>0.39</td>
<td>0.10</td>
<td>1</td>
<td>2.20</td>
</tr>
<tr>
<td>Aqua-MIP</td>
<td>32</td>
<td>5</td>
<td>2.06</td>
<td>1.74</td>
<td>2</td>
<td>7.03</td>
</tr>
<tr>
<td>erd3</td>
<td>32</td>
<td>6</td>
<td>1.70</td>
<td>0.43</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ug-2-498</td>
<td>32</td>
<td>10</td>
<td>8.65</td>
<td>5.26</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>10</td>
<td>4.6</td>
<td>4.79</td>
<td>2.2</td>
<td>7.09</td>
<td>7.06</td>
</tr>
</tbody>
</table>

* P < 0.05; ** P < 0.01
Diversity values are multiplied by 10³

Gonzalez-Martinez et al.
Genetics 2006
Tajima’s **D** - An Example

**Tajima’s D = 2.52**

**Tajima’s D = -2.10**

**ccoamt-1**, excess of high frequency variants. Balancing selection?

**erd3**, excess of rare variants. Genetic hitchhiking?
Other Neutrality Tests Using Theta

Fu and Li’s $D^*$

$$D^* = \frac{\hat{\theta}_w - \hat{\theta}_\eta}{\sqrt{\text{var}(\hat{\theta}_\pi - \hat{\theta}_\eta)}}$$

Zeng et al.’s $H$

$$H_{norm} = \frac{\hat{\theta}_\pi - \hat{\theta}_L}{\sqrt{\text{var}(\hat{\theta}_\pi - \hat{\theta}_L)}}$$

Both of these tests require an outgroup for their estimation.
Pitfalls of $D$, $D^*$, and $H_{\text{norm}}$ as Tests for Selection

- Moderate to low statistical power (i.e., the null hypothesis is not rejected when in reality it should be).
- Sensitive to the demographic history of the populations from which the samples were taken.
  - Population expansions tend to produce significantly negative values of $D$, while population bottlenecks tend to produce significantly positive values of $D$.
  - Moreover, any process that affects the shape and depth of the gene tree will affect these statistics.
Neutrality Tests - A Different Approach

• The previous approach was susceptible to any process that affected the genealogy for a sample of gene sequences.

• If a test uses information that is standardized by the genealogy, then this may not be a problem.
Expectations from Genealogies

Comparisons of within to between species polymorphisms, so-called polymorphism-divergence tests, can remove the dependency upon the genealogy for tests of selection.

Figure 3.8 Basis for tests of neutrality. Four possible scenarios for comparing the amount of intraspecific polymorphism (blue; right-hand arm of each panel) with the extent of divergence to a sibling species (orange; left arm). Under neutrality, there is an expectation that highly constrained genes are less polymorphic and diverge less between species. After adaptive divergence, diversity is reduced within a species, whereas balancing selection is one possible explanation for the maintenance of polymorphism, despite low divergence.
The Hudson, Kreitman, and Agaude (HKA) Test

• The HKA test uses data from two different species. Under neutral evolution, the expected number of differences ($D$) between two homologous sequences (one from each species) is given by:

$$E(D) = k\theta(T + 1)$$
$$\text{var}(D) = k\theta(T + 1) + k^2\theta^2$$

• Similarly, the expected number of polymorphisms within each species are given by:

$$E(S) = \hat{\theta}_w a_n$$
$$\text{var}(S) = \hat{\theta}_w a_n + b_n (\hat{\theta}_w)^2$$

• The rationale for the HKA test is that under neutral evolution, the diversity within each species depends upon theta, but that the divergence between species depends on theta AND time ($T$).
The HKA Test Statistic

• The equations from the last slide can be used to construct a $X^2$ statistic with $2L-2$ degrees of freedom ($L = \text{number of loci}, L \text{ must be } \geq 2$):

\[ X^2 = \sum_i \sum_j \frac{(S - E(S))^2}{\text{var}(S)} + \sum_i \frac{(D - E(D))^2}{\text{var}(D)} \]

• If $X^2$ is large enough, the null hypothesis of neutral evolution can be rejected. (If $E(S)$ and $E(D)$ are estimated from the same data as that being used in the test, $X^2$ must be bigger than 3.841 for $L = 2$ using a critical p-value of 0.05).
Interpretations of HKA Results

• A significant HKA result can be due to a number of causes:
  – Too much polymorphism within species at synonymous sites.
    • This is consistent with purifying selection.
  – Too much polymorphism between species at nonsynonymous sites
    • This is consistent with positive directional selection.

• The HKA is also conservative because it assumes complete linkage within each gene and free recombination between genes.
The MacDonald-Kreitman (MK) Test

• The MK test removes effects of the genealogy by dividing the types of polymorphisms at a given locus into four types:
  – Synonymous/Polymorphic
  – Synonymous/Fixed
  – Nonsynonymous/Polymorphic
  – Nonsynonymous/Fixed
The MK Table and Tests for Independence

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-synonymous</td>
<td>$N_F$</td>
<td>$N_P$</td>
</tr>
<tr>
<td>Synonymous</td>
<td>$S_F$</td>
<td>$S_P$</td>
</tr>
<tr>
<td>Ratio</td>
<td>$\frac{N_F}{S_F} = \frac{N_P}{S_P}$</td>
<td></td>
</tr>
</tbody>
</table>

Under neutrality, $\frac{N_F}{S_F} = \frac{N_P}{S_P}$

Under positive directional selection, $\frac{N_F}{S_F} > \frac{N_P}{S_P}$

A likelihood ratio test for independence in a contingency table is then used to test the null hypothesis of neutrality.

If we assume that $S_F = S_P$ under neutrality, the test statistic is given as:

$$G = 2 \sum_{\text{all loci}} \left[ \frac{N_F}{N_P} \right] \ln \left( \frac{\left[ \frac{N_F}{N_P} \right]}{\text{avg}[\frac{N_F}{N_P}]} \right)$$
dN/dS

• A related concept to the MK test are tests based on the dN/dS ratio.
• These tests assume that under neutrality, $dN/dS = 1.0$
• Statistical tests are then constructed to test whether or not an observed dN/dS deviates from 1