Natural Selection

02-715 Advanced Topics in Computational Genomics
Time Scales for the Signatures of Selection

- Proportion of functional changes
- Heterozygosity/rare alleles
  - High frequency derived alleles
  - Population differences
    - Length of haplotypes

6 mya
250 kya
75 kya
25 kya

Africa
Asia
Europe
Selective Sweep

Before sweep

No recombination

Recombination

Incomplete sweep

Complete sweep
Long Haplotypes

- *LCT* allele for lactase persistence (high frequency ~77% in European populations but long haplotypes)
Difficulties in Detecting Natural Selection

- **Confoundning effects of demography**
  - Population bottleneck and expansion can leave signatures that look like a positive selection

- **Ascertainment bias for SNPs**
  - Regions where many sequences were used for ascertainment may appear to have more segregating alleles at low frequencies with more haplotypes.

- **Recombination rate**
  - Strong signature for selection for regions with low recombination rates
Analysis of HapMap Data for Natural Selection (Sabeti et al., 2007)

• Look for evidence of recent selective sweep
  – Long haplotypes
  – Control for recombination rates by comparing the long haplotypes to other alleles at the same locus
  – EHH, iHS tests
EHH Test

• Extended haplotype homozygosity (EHH): EHH at distance $x$ from the core region is the probability that two randomly chosen chromosomes carry a tested core haplotype are homozygous at all SNPs for the entire interval from the core region to the distance $x$. 
Haplotype Bifurcation Diagram for Computing EHH
iHS Test

- iHS (integrated haplotype score):

\[
iHS = \frac{\ln \left( \frac{iHH_A}{iHH_D} \right) - E_p \left[ \ln \left( \frac{iHH_A}{iHH_D} \right) \right]}{SD_p \left[ \ln \left( \frac{iHH_A}{iHH_D} \right) \right]}
\]

- \(iHH\): integrated EHH
- \(iHH_A\): \(iHH\) for ancestral allele
- \(iHH_D\): \(iHH\) for derived allele
iHS Test
iHS: More Examples

(a) East Asians, rs6060371 (in SPAG4), $p_d = 0.742$, $2.3 \text{ cM/Mb}$

(b) CEPH, rs996521 (in SNTG1), $p_d = 0.808$, $0.28 \text{ cM/Mb}$

(c) Yoruba, rs995647 (in NCOA1), $p_d = 0.492$, $0.62 \text{ cM/Mb}$
Analysis of HapMap Data for Natural Selection

• Determining targets of selection among the candidate regions
  – Target alleles are likely to be derived alleles
  – Target alleles are likely to be highly differentiated between populations
  – Target alleles are likely to have biological effects, e.g., non-synonymous
### HapMap: Candidates for Natural Selection

<table>
<thead>
<tr>
<th>Region</th>
<th>Chrposition (MB, HG17)</th>
<th>Selected population</th>
<th>Long Haplotype Test</th>
<th>Size (Mb)</th>
<th>Total SNPs with Long Haplotype Signal</th>
<th>Subset of SNPs that fulfil criteria 1</th>
<th>Subset of SNPs that fulfil criteria 2</th>
<th>Subset of SNPs that fulfil criteria 1, 2 and 3</th>
<th>Genes at or near SNPs that fulfil all three criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chr1:166</td>
<td>CHB + JPT</td>
<td>LRH, iHS</td>
<td>0.4</td>
<td>92</td>
<td>39</td>
<td>30</td>
<td>2</td>
<td>BLZF1, SLC19A2</td>
</tr>
<tr>
<td>2</td>
<td>chr2:72.6</td>
<td>CHB + JPT</td>
<td>XP-EHH</td>
<td>0.8</td>
<td>732</td>
<td>250</td>
<td>0</td>
<td>0</td>
<td>EDAR</td>
</tr>
<tr>
<td>3</td>
<td>chr2:108.7</td>
<td>CHB + JPT</td>
<td>LRH, iHS, XP-EHH</td>
<td>1.0</td>
<td>972</td>
<td>265</td>
<td>7</td>
<td>1</td>
<td>RAB3GAP1, R3HDM1, LCT</td>
</tr>
<tr>
<td>4</td>
<td>chr2:136.1</td>
<td>CEU</td>
<td>LRH, iHS, XP-EHH</td>
<td>2.4</td>
<td>1,213</td>
<td>282</td>
<td>24</td>
<td>3</td>
<td>PDE11A</td>
</tr>
<tr>
<td>5</td>
<td>chr2:177.9</td>
<td>CEU, CHB + JPT</td>
<td>LRH, iHS, XP-EHH</td>
<td>1.2</td>
<td>1,388</td>
<td>399</td>
<td>79</td>
<td>9</td>
<td>SLC30A9</td>
</tr>
<tr>
<td>6</td>
<td>chr4:33.9</td>
<td>CEU, YRI, CHB + JPT</td>
<td>LRH, iHS</td>
<td>1.7</td>
<td>413</td>
<td>161</td>
<td>33</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>chr4:42</td>
<td>CHB + JPT</td>
<td>LRH, iHS, XP-EHH</td>
<td>0.3</td>
<td>249</td>
<td>94</td>
<td>65</td>
<td>6</td>
<td>SLC30A9</td>
</tr>
<tr>
<td>8</td>
<td>chr4:159</td>
<td>CHB + JPT</td>
<td>LRH, iHS, XP-EHH</td>
<td>0.3</td>
<td>233</td>
<td>67</td>
<td>34</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>chr10:3</td>
<td>CEU</td>
<td>LRH, iHS, XP-EHH</td>
<td>0.3</td>
<td>179</td>
<td>63</td>
<td>16</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>chr10:22.7</td>
<td>CEU, CHB + JPT</td>
<td>XP-EHH</td>
<td>0.3</td>
<td>254</td>
<td>93</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>chr10:55.7</td>
<td>CHB + JPT</td>
<td>LRH, iHS, XP-EHH</td>
<td>0.4</td>
<td>735</td>
<td>221</td>
<td>5</td>
<td>2</td>
<td>PCDH15</td>
</tr>
<tr>
<td>12</td>
<td>chr12:78.3</td>
<td>YRI</td>
<td>LRH, iHS</td>
<td>0.8</td>
<td>151</td>
<td>91</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>chr15:46.4</td>
<td>CEU</td>
<td>XP-EHH</td>
<td>0.6</td>
<td>867</td>
<td>233</td>
<td>5</td>
<td>1</td>
<td>SLC24A5</td>
</tr>
<tr>
<td>14</td>
<td>chr15:61.8</td>
<td>CHB + JPT</td>
<td>XP-EHH</td>
<td>0.2</td>
<td>252</td>
<td>73</td>
<td>40</td>
<td>6</td>
<td>HERC1</td>
</tr>
<tr>
<td>15</td>
<td>chr16:64.3</td>
<td>CHB + JPT</td>
<td>XP-EHH</td>
<td>0.4</td>
<td>484</td>
<td>137</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>chr16:74.3</td>
<td>CHB + JPT, YRI</td>
<td>LRH, iHS</td>
<td>0.6</td>
<td>55</td>
<td>35</td>
<td>28</td>
<td>3</td>
<td>CHST5, ADAT1, KARS</td>
</tr>
<tr>
<td>17</td>
<td>chr17:53.3</td>
<td>CHB + JPT</td>
<td>XP-EHH</td>
<td>0.2</td>
<td>143</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>chr17:56.4</td>
<td>CEU</td>
<td>XP-EHH</td>
<td>0.4</td>
<td>290</td>
<td>98</td>
<td>26</td>
<td>3</td>
<td>BCAS3</td>
</tr>
<tr>
<td>19</td>
<td>chr19:43.5</td>
<td>YRI</td>
<td>LRH, iHS, XP-EHH</td>
<td>0.3</td>
<td>83</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>chr22:32.5</td>
<td>YRI</td>
<td>LRH</td>
<td>0.4</td>
<td>318</td>
<td>188</td>
<td>35</td>
<td>3</td>
<td>LARGE</td>
</tr>
<tr>
<td>21</td>
<td>chr23:35.1</td>
<td>YRI</td>
<td>LRH, iHS</td>
<td>0.6</td>
<td>50</td>
<td>35</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>chr23:63.5</td>
<td>YRI</td>
<td>LRH, iHS</td>
<td>3.5</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total SNPs</td>
<td></td>
<td></td>
<td></td>
<td>16.74</td>
<td>9,166</td>
<td>2,898</td>
<td>480</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>
Global Distribution of Positively Selected Allele SLC24A5 A111T
EHH, iHS, and Ascertainment Bias

- EHH, iHS are haplotype based method
  - Less sensitive to ascertainment bias.
  - Good power for recent selective sweeps, but low power for older sweeps.
Composite Likelihood Test
(Nielsen et al., 2005)

• Likelihood models for null and alternative hypotheses

• Incorporates a scheme for correcting the ascertainment bias
Composite Likelihood Test 1

- $p = \{p_1, \ldots, p_{n-1}\}$: probabilities of derived allele frequencies for $n$ samples

- Likelihood model under neutral evolution

\[
CL_1(p) = \prod_{i=1}^{k} p_{x_i} = \prod_{j=1}^{n-1} p_{j}^{k_j}
\]

- Likelihood model under selective sweep

\[
CL_1(p; \nu \leftrightarrow b) = \prod_{i=\nu}^{b} p_{x_i}
\]

- Test statistic

\[
T_1 = 2\{\log CL_1(\hat{p}_{\nu \leftrightarrow b}; \nu \leftrightarrow b) - \log CL_1(\hat{p}; \nu \leftrightarrow b)\}
\]
Composite Likelihood Test 2

- Incorporate spatial distribution in allele frequencies due to recombinations

- Assumption: each ancestral lineage in the genealogy has an i.i.d. probability of escaping a selective sweep through recombination onto the selected background.
Ancestral Recombination Graph with Selective Sweep

neutral phase

selective phase

$T = \tau$

neutral phase

$T = 0$ (Present)
Composite Likelihood Test 2

• The probability of escaping through recombination

\[ P_e = 1 - e^{-\alpha d} \]

- \( d \): distance \( d \) between a given locus and the selected variant
- \( \alpha \): a parameter that is a function of recombination rate, effective population size, selection coefficient of the selected mutation (e.g., \( \alpha = r \ln(2N)/s \)
Composite Likelihood Test 2

- The probability that $k$ ($0 < k < n$) out of $n$ gene copies escaped the sweep:

$$P_e(k) = \binom{n}{k} P_e^k (1 - P_e)^{n-k}$$

- The probability of observing $B$ mutant alleles after a sweep:

$$p_B^* = P_e(n)p_B + \sum_{k=0}^{n-1} P_e(k) \left( p_{B+1-n+k, k+1} \frac{B + 1 - n + k}{k + 1} + p_{B, k+1} \frac{k + 1 - B}{k + 1} \right)$$
Simulation Study

- Distribution of test statistics under null hypothesis

Test 1

Test 2
Correcting for Ascertainment Bias

- Likelihood for allele frequencies after conditioning on ascertainment (i.e., unobserved true allele frequencies)

\[
L(\theta) \propto \Pr(X_i = \chi_i \mid \theta; \text{Asc}_i) = \frac{\Pr(\text{Asc}_i \mid X_i = \chi_i, \theta)\Pr(X_i = \chi_i \mid \theta)}{\Pr(\text{Asc}_i \mid \theta)}
\]

\[
\Pr(\text{Asc}_i \mid X_i = \chi_i, \theta) = 1 - \frac{C_{ii} + (n - C_{ii})}{\binom{n}{d}}
\]

\[
\Pr(\text{Asc}_i \mid \theta) = \sum_{j=1}^{n-1} \Pr(X_i = j \mid \theta) \Pr(\text{Asc}_i \mid X_i = j)
\]
Correcting for Ascertainment Bias
(Nielson et al., 2004)

- Illustration through simulation study (20 genes, 10,000 SNPs, 5 genes for ascertainment)
HapMap Data Analysis

• HapMap chromosome 2

• Test 1: requires a choice of window size

• Test 2: no need to fix the window size
Ascertainment Bias from HapMap Analysis

A: Chr 2 no ascertainment

B: Chr 2 ascertainment
Neandertals and Modern Humans
Selective Sweeps in Modern Human Genomes Compared to Neandertals