Locating the targets of selective sweeps: theory and practice

Chip Aquadro
Cornell University

June 26, 2006
Cornell Probability Summer School
“Footprints” of natural selection can provide clues to function and novelty

1. Genes that evolve rapidly; candidates for positive selection
2. Genome scans for regions with “unusual” patterns of variability

adapted from Gil McVean
ancestry of current population

Genealogies along the sequence

...GAACCCGGATTTAAGGA...CCTTTCCACGGGATGGTTTCTCAGGGAAATCCCTATGGCTACCTG...
...GAACCAGGATTTTACGACTCTTTCCACGGGATGGACCTTCAGGGATTCCCTATGGCTACCTG...
...GAACCCAGGATTTAACGACTCTTTCCACGGGATGGACCTTCAGGGAAATCCCTATGGCTACCTG...
...GAACCCAGGATTTAAGGA...CCTTTCCACGGGATGGTTTCTCAGGGAAATCCCTATGGCTACCTG...

With no selection, all trees “roughly” the same “size”
So similar levels of variation
(“tree” topologies are different due to recombination)
Genealogies along the sequence

Localized reduction in the “size” of the gene tree which means a localized reduction in variation, and skew in allele frequency (excess of rare alleles)

Figure 2. Average nucleotide diversity along a recombining chromosome under the model of genetic hitchhiking, with and . Squares represent average heterozygosity at single nucleotide sites averaged over 50,000 replicates of the simulations. The expected value (continuous curve) was calculated using Equation 13 of Kim and Stephan 2000, with as the recombination rate between a nucleotide site and the site of selection. Directional selection occurs at position 20 kb with and = 0.005.

From Kim and Stephan 2002
40 kb region simulated with a selective sweep at position 20 kb

Theta pi (black), Theta S (blue), and Theta H (red)

\( N = 5 \times 10^5, s = 0.001, R=800, t=0.001 \) (from Kim & Stephan 2002)

40 kb region simulated under equilibrium neutral model

Theta pi (black), Theta S (blue), and Theta H (red)

\( N = 5 \times 10^5; R=800 \) (from Kim & Stephan 2002)
**Composite Likelihood Ratio Test (CLRT)**


1. Maximum composite likelihood of data under the model of selective sweep ($L_S$) is compared to that under neutral equilibrium ($L_N$).

2. The likelihood ratio is defined as $LR = \log(L_S/L_N)$.

3. The maximum likelihood estimates of the strength of selection ($\alpha = 2N_s$) and the location of the beneficial mutation ($X$) are obtained.

4. The null distribution of $LR$ is obtained by applying this test to data sets obtained from simulations under the standard neutral model. The neutral model is rejected when the observed $LR$ is greater than the 95 percentile of the null distribution.

**Departures from “neutrality” do not necessarily imply the action of natural selection.**

Effect of Non-equilibrium Demography

Tajima (1989) - points out that a negative value of $D$ may be obtained if the population recently experienced a bottleneck

Fu&Li (1992) - note that migration may introduce rare alleles, and expansion may produce an excess of rare alleles too

Fay&Wu (2000) - add that ancient subdivision with rare migration between subpopulations may result in an excess of high frequency derived alleles
False signals of selection due to Bottlenecks

Proportion of false positives in CLR test

Severity of bottleneck:
- 50%
- 80%
- 90%
- 99%
Can "real" sweeps be distinguished from demography?

**A possible solution: a Goodness-of-Fit Test**

1. Perform composite likelihood ratio test (CLRT) on empirical dataset

2. IF REJECTS, simulate 1000 replicates with selection using the number of segregating sites in the empirical dataset, as well as the estimated values of α and X given by CLRT in step (1)

3. Perform CLRT for each of the 1000 ‘selection’ simulation replicates

4. Each of these replicates is subsequently analyzed via a modification of a Goodness of Fit statistic ($\Lambda_{GOF}$), and the P-value of the fit of the observed data to the selection model estimated via Monte Carlo.

5. LR = ln($H_n/H_s$). Thus, our null distribution is simulated under a selection model, rather than a neutral equilibrium model as with the original CLRT.

**Jensen, Kim, Bauer DuMont, Aquadro & Bustamante. 2005. Genetics**
a) Rejection due to selection \( (P=0.64) \)
\[ \Lambda_{GOF}^{(0)} \]
Rejection due to demography \( (P=0.00) \)
\[ \Lambda_{GOF}^{(0)} \]
Replicates a better fit \( \Lambda_{GOF} \)
Replicates a worse fit

---

**GOF performance**

**SELECTION**

90% REDUCTION

- \( t_b = 0.0125 \)
- \( t_b = 0.025 \)
- \( t_b = 0.05 \)

99% REDUCTION

**MIGRATION**

- \( t_b = 0.0125 \)
- \( t_b = 0.025 \)
- \( t_b = 0.05 \)
Nonequilibrium demography and the CLRT test

- CLRT of Kim & Stephan is a popular approach as it incorporates multiple aspects of the data, and actually tests selection against neutrality.

- However, CLRT performs poorly under a several realistic demographic scenarios, returning a high rate of false-positives.

- A proposed GOF test allows for the discrimination between datasets that reject neutrality in favor of selection because of demography/population structure, from those that truly reject as the result of positive selection.

Jensen, Kim, Bauer DuMont, Aquadro & Bustamante. 2005. Genetics
**Drosophila melanogaster**'s ancestral range is in Africa. Within the last 10,000 years it colonized the rest of the world.

Can we find evidence for the genes and types of variation that played a role in adaptation to these new environments?

Genealogies along the sequence

"recent" adaptive mutation

...GAACCCGGATTTAAGGACTCTTTCACCGGACGGTCTTCCAGGGAATCCCTATGGCTATCTG...

...GAACCAGGATTTTACGACTCTTTACCGGACGGTCCTTCAGGGAATTTCCCCTATGGCTACCTG...

...GAACCGAGTCTTACGGACTCTTTACCGGACGGTCCTTCAGGGAATTTCCCCTATGGCTATCTG...

...GAACCGAGTCTTACCGGACGGACTCTTTACCGGACGGTCCTTCAGGGAATTTCCCCTATGGCTATCTG...

Localized reduction in the “size” of the gene tree which means a localized reduction in variation, and skew in allele frequency (excess of rare alleles)
Variation at microsatellite loci at equilibrium

Trajectory of expected heterozygosity and number of alleles (k) during and after a Selective Fixation of a new mutation
Microsatellite scan of X-chromosome of Drosophila melanogaster

Microsatellite results: frequency skew

Bottleneck two-phase model results

Heterozygosity

Allele Excess

Location of genes and putative ORF’s

Known genes
Annotated
Gene Finder

white roughest kirre Notch Fcp3c TOS CG18508

Microsatellite scan results: frequency skew

-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Zimbabwe
USA
China
Ecuador

Notch 5’ Notch 3’
Sequence polymorphism data collected for Zimbabwe and USA

Sliding window plots of variability and divergence

Zimbabwe

USA

ThetaW
Theta pi
divergence
Sliding window plots of tests of the frequency spectrum

- **Tajima's D**
  - **Zimbabwe**
    - $D = -0.412$
    - $P$-value = 0.01
  - **United States**
    - $D = 1.065$
    - $P$-value $\leq 0.0001^*$

- **Fu and Li's D**
  - **Zimbabwe**
    - $D = -0.664$
    - $P$-value = 0.004*
  - **United States**
    - $D = 0.627$
    - $P$-value $\leq 0.05^*$

- **Fay and Wu's H**
  - **Zimbabwe**
    - $H = -10.15$
    - $P$-value = 0.004*
  - **United States**
    - $H = -40.90$
    - $P$-value $\leq 0.0001^*$

**Likelihood surface of the location of a sweep in Zimbabwe**

- Log likelihood ratio = 6.0
- $P$-value = 0.447
Likelihood surface of the location of a sweep in the USA

Log likelihood ratio = 26.84
P-value < 0.001

Goodness of Fit Test result for USA CG18508 region data

CLRT = 26.9 (p<0.001 sel vs neutrality)

GOF test p = 1.0
(cannot reject selection)

Non-equilibrium demography we’ve considered cannot explain the “apparent” sweep
### Neutral Amino Acid sequence divergence at Fcp3C

**Type of variation (215 codons)**

<table>
<thead>
<tr>
<th></th>
<th>Synonymous</th>
<th>Non-synonymous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within species</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Between species</td>
<td>29</td>
<td>13</td>
</tr>
</tbody>
</table>

*P*-value = 0.309

*(D. melanogaster Zimbabwe + USA)*

### Excess Amino Acid sequence divergence at CG18508

**Type of variation (102 codons)**

<table>
<thead>
<tr>
<th></th>
<th>Synonymous</th>
<th>Non-synonymous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within species</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Between species</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

*P*-value = 0.007

*No excess of synonymous polymorphisms* *(D. melanogaster Zimbabwe + USA)*
What is known about CG18508?

- Four EST hits in *D. melanogaster*: two in testes and two in embryo

- Signal sequence predicted in both the *D. melanogaster* and *D. pseudoobscura* predicted proteins, but no other domains found

- Rapidly evolving! Alignments difficult for divergent species comparisons

Any functional significance of non-coding fixed difference between CG18508 and Fcp3C?

- Within the USA population fixed, derived difference is within the active domain of the transcriptional binding site of the *caudal* transcription regulator protein

- *caudal* binding site is not predicted in any of the Zimbabwe sequences or in *D. simulans*

- *caudal* is expressed in the embryo, testes, and ovaries which corresponds to expression of CG18508 and Fcp3C
Signatures of selection in and near Notch

-Synonymous site evolution at Notch in both species is an extreme example of species-wide trends with positive selection appearing to have played a key role in shaping this pattern both D. melanogaster and D. simulans

-There are two signatures of natural selection downstream of Notch.
  1) evidence of long term positive selection acting on amino acid fixations at CG18508
  2) evidence of a recent sweep in the USA population centered between the CG18508 and Fcp3C.

-Additional recent sweep within Notch 5th intron in USA


Conclusions

• The “targets” of strong “catastrophic” sweeps can be found by genome scans of microsatellite and DNA sequence variation

• Challenges:
  • Distinguishing demography from selection
    • Very difficult problem particularly for strong bottlenecks
    • GOF test is an improvement, adding LD may help further
    • How do we best sample when we find evidence of departures?
      • Complete sequencing leads to best MLE estimates

  • Adaptive trait locus mapping is like QTL mapping, but we have neither the trait nor the locus

  • Opportunity for further analysis posed by variation fixed between populations of D. melanogaster, but our task now turns to functional analyses of not only structural but regulatory variation.