Selection
Selection at work

- We have many examples of populations adapting to changing environmental conditions.
- The availability of sequence/genotype data for populations allows us to sample the genome for regions under active selection.
Selection at work: lactose intolerance

Worldwide prevalence of lactose intolerance in recent populations
(schematic)

- 0-15%
- 15-30%
- 30-60%
- 60-80%
- 80-100%
Selection at work: plasmodium resistance

Prevalence of the dhps 540E mutation in recent surveys of P. falciparum

Ninety surveys of 540E have been conducted since 2004. Red pie charts indicate those surveys where the prevalence of 540E exceeded 50% and countries where they occurred are shaded brown. The black pie charts indicate surveys recording prevalence of less than 50% and blue circles a prevalence of zero. Countries shaded pink indicate that 540E was detected at some stage but always <50% prevalence. Countries shaded blue indicate that it was not detected, while white indicates that no data are currently available.
Selection at work: insecticide resistance

Resistance is mobile
The majority of countries with ongoing malaria transmission now report mosquitoes resistant to one or more classes of insecticide.

Colombia
Resistance to two classes detected

Côte d’Ivoire
Resistance to four classes detected

India
Resistance to three classes detected

Countries with malaria transmission and insecticide resistance
Countries with malaria transmission and no reports of insecticide resistance
Selection at work: The high life

Cerro de Pasco (population 70,000), 14k ft above sea level
The challenge

- In each case, we would like to understand the genetic basis of the selection. What gene is under selection?
- It could be that a single mutation causes an increase/decrease in fitness. How can we identify the important mutation?
Evolution in a nutshell

• DISCLAIMER; There are many subtleties, and you should take each statement critically.

• Lamarck:
  - Evolutionary stress causes mutation (change in phenotypic trait) to happen. It is inherited and spreads through the population.
    • Ex: Giraffe necks become longer as they try to reach higher.
    • Weissman (allegedly) cut the tails of mice for 20+generations with no shortening, concluding that Lamarckism does not hold.
    • Given our knowledge of modern genetics, can this be valid a mechanism in principle, if not in practice?
Mutations happen by chance, and cause (phenotypic) variation in a population.

- Somatic variations are not inherited, even if they might be Lamrackian (occur in response to stress).
- Germline variations are inherited by the offspring.

Subsequently, selective pressure applies to the variants, making some individuals less fit than others.

Fit individuals survive.

Note: this argument was in the absence of any insight regarding the molecular basis for variation/inheritance.

Implicit in the argument is that all variations are subject to selection. Selection is the prime determinant of whether a variants survives or dies.
Neutral evolution (Kimura)

- Early assumption: most variation is mildly deleterious.
- Early genotyping surveys quickly revealed that the number of variable regions were far too many to be all deleterious.
- Kimura suggested that most alleles are selectively neutral.
- The presence of neutrally evolving alleles changes the landscape of genetics.
  - If a mutation is selectively neutral it plays no role for any phenotypic trait. Not useful for understanding the genetic basis of phenotypes (but useful as a molecular clock).
  - Mutations under selection have faster deterministic outcomes (fixation or elimination). By contrast, neutral mutations just drift for a while.
  - Therefore, we need to look for ‘interesting’ mutations against a null background of neutral ones.
How can you tell identify neutral mutations
We define a gene as a location on the genome that codes for proteins.

The genic information is used to manufacture proteins through transcription, and translation.

There is a unique mapping from triplets to amino-acids.
Translation

- The ribosomal machinery reads mRNA.
- Each triplet is translated into a unique amino-acid until the STOP codon is encountered.
- There is also a special signal where translation starts, usually at the ATG (M) codon.
The ribosomal machinery reads mRNA.

Each triplet is translated into a unique amino-acid until the STOP codon is encountered.

There is also a special signal where translation starts, usually at the ATG (M) codon.

Given a DNA sequence, how many ways can you translate it?
Gene Features

- The gene can lie on any strand (relative to the reference genome)
- The code can be in one of 3 frames.
Eukaryotic gene structure

RNA synthesis and processing
4-fold degenerate sites

- The redundancy of the genetic code implies that many base-pairs do not change the underlying amino-acid.
- Often, the redundancy is in the 3rd base-pair.
- It is a useful first guess that these are selectively neutral.
Neutral alleles

- Now that we know some molecular biology, ..
- How can we detect neutral alleles?

- 4 fold degenerate sites in DNA should be selectively neutral. (Not always, though. Why?)
Often, we can classify mutations as being
- coding, synonymous
- non-coding deleterious
- non-coding benign
- regulatory
A note on functional analysis

- Regulation is complex
Alleles under selection: a PopGen perspective
Basic principles of selection

- More offspring are produced than can survive
- Different offspring have different levels of ‘fitness’
- ‘fit’ individuals are more likely to survive and pass on their genotypes
- If a mutation is deleterious, it is quickly eliminated.
- If a mutation is advantageous, it is quickly driven to fixation
- Environmental change can change a neutral mutation into an adaptive one.
Coalescent under selection

Time

Out-group population

Control population

Case population

migration to high altitude

MRCA

Now
Coalescent should be different in selection

- Selection coefficient $s$:
- Probability of choosing favored allele $\alpha (1+s)$
- Number of descendant grows exponentially (actually logistically) with time.
**Digression: estimating $\theta$**

- Recall that $\theta = 4N\mu$, is the population-scaled mutation rate.
- Given a SNP matrix, can you identify the scaled mutation rate?

\[
\begin{array}{cccccccccc}
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

$m$

$n$
Watterson’s estimate

- Let $S$ be the number of mutations in the history of a population sample.
- If we make the infinite sites assumption, then $S$ can be estimated.
- Recall that
  - $E(S_n) = \mu E(T_{tot})$
  - $E(S_n) = \mu 2N \sum_k 2/(k-1) = 4N \mu (\gamma + \ln (n-1))$
  - Watterson’s estimate
    - $\theta_W = S_n / (\gamma + \ln (n-1))$
Tajima’s estimate of $\theta$

- Define $\pi_{ij} = \text{heterozygosity between two individuals}$
- Note: heterozygosity = # differing sites = hamming distance

\[
\begin{align*}
i: & \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \\
j: & \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \\
\pi_{ij} & = 2
\end{align*}
\]

- Average heterozygosity can be empirically estimated from a sample as

\[
\hat{k} = \frac{1}{\binom{n}{2}} \sum_{ij} \pi_{ij}
\]
Estimating Average heterozygosity

- Assuming an underlying coalescent model of evolution, what is the average heterozygosity?
- Q: Given 2 randomly picked individuals, what is the expected time to coalescence?
  - A: 2N
- Q: Given 2 individuals what is the expected number of mutations in the lineages connecting them?
  - A: \( \mu 2 2N = \theta \)
- Therefore, the average heterozygosity \( k \) is an estimate (Tajima’s estimate) of \( \theta \)
Two estimates of $\theta$

- Let $S$ be the number of mutations. Watterson’s estimate
  - $\theta_w = S_n / (\gamma + \ln (n-1))$
- Tajima’s estimate. Let $\pi_{ij}$ be the heterozygosity between individuals $i$ and $j$. The average heterozygosity is an estimate of $\theta$

$$\hat{k} = \frac{1}{\binom{n}{2}} \sum_{ij} \pi_{ij}$$
Tajima’s D statistic

• Tajima proposed the difference of the two as a test of selection (Tajima’s D statistic)
• Tajima’s D = (k - $\theta_w$)/(std. dev)
• The actual statistic involves a normalization
• Under neutral evolution, D=0
• What do we expect under positive selection?
Under positive/negative selection

- The branch that has the positive allele begins to dominate
- Over time the smaller branch will disappear
- What will happen to heterozygosity?
Under positive/negative selection

- What happens to nearby mutations.
- They hitchhike along with the selective allele
- The area round the selected region sees a loss of heterozygosity
- This phenomenon is also called selective sweep

![Diagram showing selection and neutral processes]
Tajima’s D under selection?

- Under positive selection, there is a loss in average heterozygosity?
  - $D = \sim k - \theta_w < 0$

- Under balancing selection, there should be a gain in average heterozygosity?
  - $D > 0$
When does Tajima’s D fail?

1. When the population is growing, what will happen to average heterozygosity?
2. What happens when the selection event is a recent one?
THE ALLELE FREQUENCY SPECTRUM
Allele Frequency Spectrum

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Frequency

Count

Scaled Count

Frequency
Scaled SFS of Neutral Evolution

\[ E(\xi_i) = \frac{\theta}{i} \quad \forall i = (1, \ldots, n - 1) \]

\[ E(\xi'_i) = iE(\xi_i) = \theta \quad \forall i = (1, \ldots, n - 1) \]

(Fu, 1995)

* average of 500 simulated population samples
Generalization via the allele frequency spectrum

- Consider a sample in which 0 is the ancestral allele, and 1 is the derived allele.
- Let $\xi_i$ be the number of sites with exactly $i$ derived alleles.
  - $\sum_{i=1}^{n-1} \xi_i = m$
- Fu had the remarkable result that
  $$\text{Exp}(\xi_i) = \frac{\theta_i}{i}$$
- Let $\hat{\theta}_i = i\xi_i$ denote the scaled frequency.
- Any linear combination of $\hat{\theta}_i$ is an estimate of theta.
Many estimates of theta can be explained by Fu’s observation

- Watterson’s estimate simply counts all such sites. Therefore, $\alpha_i = 1/i$
Tests of neutrality

- Many independent estimates of theta.
- Difference of any two should be 0 under neutral evolution.
- Difference ≠ 0
  - indicates departure from neutrality
  - Possibly indicates selection, but not always.
  - Sometimes the result is not easy to interpret.
Suppose you knew the true rate theta in a location under an environment of no selection. It would be easier to interpret results then.

Alternatively, if we have two populations: identical, but one is under selection, and the other is not, we can compare theta values
Case control scenarios in selection

Time

Out-group population
Control population
Case population

Now

migration to high altitude
ToS: HKA test (reduction in diversity)

- Compare locus b against another locus (a) evolving neutrally.
- Compute $\mu_a$, $\mu_b$ at loci a and b by computing the edit-distance against a second species.
- Compute $\theta_a$, $\theta_b$ by using Watterson’s estimate.
- Under neutral selection, $\theta_a/\theta_b = \mu_a/\mu_b$
- If locus (b) is under selection, $\mu$ should not change, but $\theta_b$ should decrease.
Comparing case-control populations

- If you do any estimate of theta, it should yield a higher value in controls versus cases.
Case control tests

\[ F_{ST} = \frac{\Pi_{Between} - \Pi_{Within}}{\Pi_{Between}} \]

- Note that the within-diversity should always be no more than the between-diversity
  - \( 0 \leq F_{ST} \leq 1 \)

- Why does it work?
  - \( F_{ST} \) is a measure of branch-length between the two trees.
Case control tests

\[ F_{ST} = \frac{\Pi_{Between} - \Pi_{Within}}{\Pi_{Between}} \]
Test of recent selection
Two genes have been implicated in resistance to the malarial parasite *Plasmodium falciparum*

- *Glucose-6-phosphate dehydrogenase (G6PD)*
  - A common variant G6PD-202A confers partial protection against malaria
- Likewise, TNFSF5-726C is a variant associated with protection against malaria.
- Sabeti et al. describe a test for identifying regions under selection, and test them on these loci
G6PD

- A core region of 15kb was identified, and 11 SNPs genotyped
- The core region was dense and had high LD (genealogy could be identified)
Extending core haplotypes

- As you add distant SNPs, the haplotypes begin to decay (reduce in frequency).
- For each core haplotype, do the EHH test
  - Define EHH (d): probability that two randomly chosen chromosomes with the core-haplotype are identical at distance d
  - Clearly EHH will decay due to mutations and recombinations
  - Claim: if the core haplotype is under selection, it will decay less than other haplotypes.
High values of EHH indicates selection

Note that EHH decays both due to mutation as well as recombination.

Mutation rates are different in different regions.

How do we choose cut-offs for EHH statistic?
Relative EHH

- Define: relative EHH:
  - EHH of core-haplotype/(average EHH of all other haplotypes)
  - Plot shows relative EHH for the 9 core haplotypes and simulated data

2/7/17 CSE280
EHH test at

- Decay of the 9 core haplotypes of G6PD region.
- Only one core haplotype (CH8) shows selection
- The other haplotypes serve as control
The EHH test helps in identifying recent positive selection.

Sabeti’s paper claims that for this data set, the other statistics don’t work as well.

Can this be tested?

Can you suggest where the test might fail to detect recent positive selection?
Tests for selection (ToS): Function altering mutations

- High proportion of function altering mutations is a sign of positive selection.
- Useful at very long distances, but needs many function altering mutations (low power)
- Let $K_a$=rate of function altering mutations, and $K_s$= rate of synonymous mutations.

$$\frac{K_a}{K_s} = \begin{cases} < 1 & \text{Purifying selection} \\ = 1 & \text{Neutral selection} \\ > 1 & \text{Positive selection} \end{cases}$$

**Fig. 2.** Excess of function-altering mutations in PRM1 exon 2. The PRM1 gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (7, 8).
The ‘functional change’ test works only if the number of function altering mutations is large enough to estimate Ka, and is usually done by species-species comparisons.
Typically, derived alleles should have lower frequency than ancestral alleles.

High frequency derived alleles might be indicative of positive selection.

By looking at related species (chimpanzee), one can infer the ancestral allele.

Duffy red-cell antigen region: Excess of derived (red) alleles in African populations.
• Intuitively, we see selection when an evolutionary lineage is favored
  - Most individuals come from that lineage
  - Results in a selective sweep with a loss of heterozygosity (captured by Tajima’s D)
  - Over time, these changes get fixed in a population, and this signal is lost
  - Also, if the selection is very recent, the selective sweep is not very strong.
  - For recent changes, LRH might work better, but their signal is lost earlier due to recombination events.

• There are other tests of selection, each with its own peculiarity, and time range.
Reviewing selection tests

Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated.)
THE DYNAMICS OF SELECTIVE SWEEPS
Scaled SFS of Neutral Evolution

\[ E(\xi_i) = \theta / i \quad \forall i = (1, \ldots, n - 1) \]

\[ E(\xi'_i) = iE(\xi_i) = \theta \quad \forall i = (1, \ldots, n - 1) \]

(Fu, 1995)

* average of 500 simulated population samples
Scaled SFS of \textbf{Neutral Evolution}
Scaled SFS of a Selective Sweep

Time

\[ \tau = 150 \]

\[ \tau = 1000 \]

\[ \tau = 250 \]

\[ \tau = 2000 \]

\[ S = 0.08 \]

Fixation
But the scaled SFS varies with selection pressure and time! And, so does power to detect selection…

Achaz G., *Frequency Spectrum Neutrality Tests: One for All and All for One?* GENETICS 2009
Power to Detect Selection Varies

![Graph showing power to detect selection varying over time. The x-axis represents time in generations, ranging from 0 to 4000, and the y-axis represents power ranging from 0 to 1. Dotted line indicates s = 0.01. The graph shows two lines representing Tajima’s D (red) and Fay & Wu’s H (pink), with peaks at certain time points. Red line for Tajima’s D stays relatively flat compared to the pink line for Fay & Wu’s H.]

* Power at 5% FPR
Power to Detect Selection Varies

* Power at 5% FPR
Supervised Learning

We train >200 models!

Support Vector Machine (LR, Boosting, LDA, etc.)

Scaled SFS

x1000 'sweep'

x1000 'neutral'

(s=0.01, t=100), (s=0.01, t=200), (s=0.01, t=300), ...

Scaled SFS
Specialized tests, Increased Power

* Power at 5% FPR
Can we simplify?
How similar are the trained models?
How similar are the trained models?

Near-fixation similarity block

Post-fixation similarity block

Mean fixation time

Low similarity

High similarity
Two SVMs/models:

- $W_{\text{near}}$ trained on data from **near-fixation** similarity blocks.
- $W_{\text{post}}$ trained on data from **post-fixation** similarity blocks.

Define:

$$S = \max \{ \Pr(\xi|W_{\text{near}}), \Pr(\xi|W_{\text{post}}) \}$$

**Important – no prior knowledge necessary!**
General SVM Test

Time (ln(2N/s) generations)

- Power

- Time (generations)

- Tajima's D
- Fay & Wu's H
- OmegaPlus
- SweeD
- SFselect-s
- SFselect
Conclusion

- A number of statistics exist for detecting selection (negative as well as balancing)
- The tests work for different scenarios such as selection coefficient, and time since onset of selection. Can a single test dominate all of these tests?
- One could consider a discriminative approach as well, by collecting appropriate negative examples.
  - Neutrally evolving regions that do well on some of the tests.
- In practice, a key idea is to establish appropriate controls to understand the significance of a statistic.
- In spite of all this work, this area needs algorithms and possibly some machine learning ideas.
HARD SWEEPS, SOFT-SWEEPS, POLYGENIC AND MULTI-LOCUS SELECTION
Hard and soft-sweeps

**Hard sweep**
- ‘hard-sweep’: A single favorable mutation is rapidly driven to fixation.
- ‘soft-sweep’: multiple favorable mutations start to simultaneously increase in frequency. These are typically “standing variation”, already existent in the population.
- The waiting time for new mutation to occur is very large.

**Soft sweep (standing variation)**

![Graph showing probability of a sweep from standing variation]

Pritchard, Current Biology, 2010
Polygenic adaptation

- Most discussion has been on the sweeping of a single locus.
- The quantitative perspective is that of polygenic adaptation: multiple loci contribute to fitness.
- E.g., height:
  - highly heritable
  - recently, three GWAS identified a total of around 50 loci that contribute to adult height in Europeans.
  - Each associated allele affects total height by about 3–6 mm.
  - Together these loci explain about 5% of the population variation in height, after controlling for sex.
Hard sweep

Before selection

After selection

Polygenic adaptation

Current Biology
Conclusion

- It is possible, even likely, that the bulk of the adaptation is based on soft-sweeps, and polygenic adaptation. This would make it very hard to detect with current methods.
Re-iterating the themes of the class

- With the availability of genome sequencing, we have a remarkable opportunity to see evolution in action, and look at the ‘fossil records’ of historical events in our DNA
  - Selective pressures
  - Migration
  - Bottlenecks in population