Using the site frequency spectrum to infer demography and natural selection

Emilia Huerta-Sanchez
Brown University
Population dispersal
Population dispersal
Population dispersal
Population dispersal
Heterozygosity

DeGiorgio et al. 2009
Number of variants per individual

Positive natural selection
Positive natural selection
Positive natural selection
Genetic Differentiation

Redrawn from Coop et al. 2009
Genetic Differentiation

Redrawn from Coop et al. 2009
Tons of modern human DNA sequence data

What are the relative contributions of genetic drift and natural selection?
What are the genes under positive selection?
Tons of modern human DNA sequence data

What are the relative contributions of genetic drift and natural selection? What are the genes under positive selection?

- Population size has not been constant in time
- The effect of demography can mimic the effect of natural selection
DNA sequence data

Sites sequenced

ATGACCCATAGTACCTTGTAAGTCGTCATA
ACGACATAGCTACCGACTCTCCGAGTGACAT
ACGACATATAGCAACTACTATGTAGCTGTAAGT
ACGATTTTCGCAAGCTATGTAGATCGTCAT
ACGACCTTTAGCAACTGTGCTAGACGTCTTA
ACGACCCATAGCAACCTGCTAGACAGTATT

AAGCAGCTGCTAGACGTCTTAAGACGATTT

Chimp reference
DNA sequence data

< Sites sequenced >

<table>
<thead>
<tr>
<th>Individuals</th>
<th>Sites sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A</td>
<td></td>
</tr>
<tr>
<td>A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A</td>
<td></td>
</tr>
<tr>
<td>A C G A C A T T A G C A A A C T A T G T A G C T G T A G T C A T G</td>
<td></td>
</tr>
<tr>
<td>A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G</td>
<td></td>
</tr>
<tr>
<td>A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A</td>
<td></td>
</tr>
<tr>
<td>A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A</td>
<td></td>
</tr>
<tr>
<td>A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A</td>
<td></td>
</tr>
</tbody>
</table>

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5
DNA sequence data

Sites sequenced

A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A
A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A
A C G A C A T T A G C A A A C T A T G T A G C T G T A G T C A T G
A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G
A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A
A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A
A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A

# sites

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5

Frequency in sample
DNA sequence data

Sites sequenced

The site frequency spectrum

The site frequency spectrum

<table>
<thead>
<tr>
<th>Frequency in sample</th>
<th># sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A
A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A
A C G A C A T T A G C A A A C T A T G T A G C T G T A G T C A T G
A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G
A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A
A C T A C C A T A G C A A A G T G T C A T G T A G A C G T C T T A
A C G A C C T T A G C A A A C T G T C T A G C A A G A A G T T T T A

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5
2D SFS

The site frequency spectrum (SFS) is shown in the plot above. The X-axis represents the site indices, and the Y-axis represents the HAN (Haplotype Allele Number). The color scale indicates the frequency of sites mutated. The graph shows a diagonal band, which is expected under neutral evolution, indicating a lack of selection pressure on the variants.
Neutrality

By chance …

[Diagram showing a sequence of illustrations with blue and red dots, indicating a process or change.]
Natural selection

Positive Selection
Natural selection

Positive Selection

Negative Selection
How is the SFS shaped by difference processes?

<table>
<thead>
<tr>
<th>Frequency in sample</th>
<th># sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
The genealogy of a sample
The Wright-Fisher model

\[X(t)\] = number of individuals carrying mutation at time \(t\)

\[X(t)\] is a **Markov chain** with transition probabilities:

\[P_{ij} = \text{Bin}(2N, i/2N)\]
If mutations are arriving at Poisson times
If mutations are arriving at Poisson times, the frequency of mutations follows the distribution $f(x) = \frac{1}{x}$. The graph illustrates the frequency of mutations over generations, showing the distribution of common and rare mutations.
The SFS in a sample of size $n$

The expected number of variant sites at frequency $i$

$$
\theta F(i, \gamma) = \theta \int_0^1 f(q) \Pr(i \mid q) dq
$$

$$
= \theta \int_0^1 \frac{1}{q \binom{n}{i}} q^i (1 - q)^{n-i} dq
$$

$$
= \frac{\theta}{i}
$$
The site frequency spectrum

Proportion of sites

Mutation frequency

rare

common

neutral
The site frequency spectrum from the coalescent
The site frequency spectrum from the coalescent
The site frequency spectrum from the coalescent
The site frequency spectrum from the coalescent
The effect of demography

The effect of demography

The effect of demography

Standard neutral model

Population growth model

Data from 1500 Danes and 1500 Chinese individuals

Synonymous sites
The effect of a bottleneck and population structure

bottleneck

Population structure

The effect of demography
If mutations are arriving at Poisson times

$$f(x) = \frac{1}{x}$$
The SFS in a sample of size $n$ with selection

The expected number of variant sites at frequency $i$

$$\theta F (i, \gamma) = \int_0^1 \frac{f(q, \gamma)}{2} \cdot \Pr (i \mid q) \, dq$$

$$= \int_0^1 \frac{1 - e^{-2\gamma(1-q)}}{1 - e^{-2\gamma}} \cdot \frac{1}{q(1-q)} \binom{n}{i} q^i (1 - q)^{n-i} \, dq$$
The SFS in a sample of size \( n \)

The expected number of variant sites at frequency \( i \)

\[
\theta F (i, \gamma) = \int_0^1 \frac{f(q, \gamma)}{2} \cdot \Pr(i \mid q) \, dq
\]

\[
= \int_0^1 \frac{1-e^{-2\gamma(1-q)}}{1-e^{-2\gamma}} \cdot \frac{1}{q(1-q)} \binom{n}{i} q^i (1 - q)^{n-i} \, dq
\]

The probability of seeing \( x_i \) sites at frequency \( i \)

\[
p (X_i = x_i \mid \theta, \gamma) = e^{-\theta F(i, \gamma)} \frac{(\theta F (i, \gamma))^{x_i}}{x_i!}
\]
The SFS in a sample of size $n$

The expected number of variant sites at frequency $i$

$$
\theta F (i, \gamma) = \int_0^1 \frac{f(q, \gamma)}{2} \cdot \Pr (i \mid q) \, dq
$$

$$
= \int_0^1 \frac{1-e^{-2\gamma(1-q)}}{1-e^{-2\gamma}} \cdot \frac{1}{q(1-q)} \binom{n}{i} q^i (1 - q)^{n-i} \, dq
$$

The probability of seeing $x_i$ sites at frequency $i$

$$
p (X_i = x_i \mid \theta, \gamma) = e^{-\theta F(i, \gamma)} \frac{(\theta F (i, \gamma))^{x_i}}{x_i!}
$$

The likelihood function

$$
L_u (\theta, \gamma \mid x) = \prod_{i=1}^{n-1} e^{-\theta F(i, \gamma)} \frac{(\theta F (i, \gamma))^{x_i}}{x_i!}
$$
If mutations are arriving at Poisson times

\[ f(x) = \frac{1}{x} \]
The site frequency spectrum

The site frequency spectrum shows the distribution of mutation frequencies. The x-axis represents the mutation frequency, with rare mutations on the left and common mutations on the right. The y-axis represents the proportion of sites. The bars indicate the number of sites for each mutation frequency category.
The site frequency spectrum

- **Mutation frequency**
  - Neutural
  - Positive selection

Proportion of sites vs. Mutations frequency.
The site frequency spectrum

Proportion of sites

Mutation frequency

neutral
positive selection
negative selection

rare
common
The site frequency spectrum

The site frequency spectrum shows the distribution of mutation frequencies across different sites.

- **Proportion of sites**
- **Mutation frequency**
- **rare**
- **common**

The spectrum categorizes mutations into three groups:
- **Neutral**
- **Positive selection**
- **Negative selection**

The chart displays the proportion of sites for each mutation frequency category.
f(x, β) = \frac{2}{K(β)x(1-x)} \log \left( \frac{1 - r_1(β)}{x - r_1(β)} \cdot \frac{r_2(β) - x}{r_2(β) - 1} \right)

The site frequency spectrum

Huerta-Sanchez et al. (2008)
Population sample

200 individuals of Danish nationality
An excess of rare mutations

Li* Y. et al. (2010) Nature Genetics
An excess of rare mutations

Li* Y. et al. (2010) Nature Genetics
An excess of rare mutations

Li* Y. et al. (2010) Nature Genetics
An excess of rare deleterious mutations

Li* Y. et al. (2010) Nature Genetics
X chromosome

Li* Y. et al. (2010) Nature Genetics
Distribution of selective effects of non-synonymous mutations

\[ f(x) = k \frac{1}{x} + (1-k) \int f_{sel}(x,-\gamma) \Gamma (\gamma; \alpha, \beta) d\gamma \]

A proportion \( k \) of mutations are neutral

A proportion \((1-k)\) are deleterious

Li* Y. et al. (2010) Nature Genetics
A larger proportion of weakly deleterious mutations

Estimated gamma densities

Boyko et al. (2008)
Plos Genetics

Li* Y. et al. (2010) Nature Genetics
Demographic inference, Gutenkunst et al. 2009

\[ \tau = t / 2N_A \]

\[ N_1 = \nu_1 N_A \]

\[ N_2 = \nu_2 N_A \]

\[ M = 2N_A m \]

\[ \theta = 4N_A \mu \]
Demographic inference

\[ \theta = 4N_{A \mu} \]

\[ M = 2N_{A \mu}m \]

\[ N_1 = v_1N_A \]

\[ N_2 = v_2N_A \]

\[ \tau = \frac{t}{2N_A} \]
\[
\frac{d}{dt} f(q,t; \Theta) = \frac{1}{2} \frac{d^2}{dq^2} \left\{ V(q; \Theta) f(q,t; \Theta) \right\} - \frac{d}{dq} \left\{ M(q; \Theta) f(q,t; \Theta) \right\}
\]
dadi
Fastsimcoalsim
Excoffier et al. 2013
Infer $\Theta$ by maximizing likelihood

$E(EFS|\Theta)$

History that best fits data

Demographic inference with dadi of fastsimcoal

Can be done using simulations, optimization techniques like gradient based or EM (fastsimcoal)

Convergence

Get from genotypes or genotype likelihoods

Can also be obtained using simulations or numerical solutions of pde

Observed SFS
Benefits of model-based approach over Structure

A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots

Daniel J. Lawson, Lucy van Dorp & Daniel Falush
We need to consider demography when detecting selection

- When the population size grows, genetic drift slows down and selection becomes more efficient
- A population bottleneck speeds up genetic drift and makes natural selection get less efficient
- Population size changes shift the site frequency spectrum away from the shape $\text{SFS}(n) \sim 1/n$ that is expected under neutrality
- Natural selection also shifts the SFS away from the equilibrium $1/n$ shape
- The SFS shape can be used to test for selection, but only after correcting for nonequilibrium demography