### Readings


### Test statistics and hypothesis testing

- Let $H$ be a hypothesis (or statement) about a population parameter
  - E.g., $\theta = 1$, or the human population started expanding 10,000 years ago
- Let $T$ be a statistic of the data
  - Can be any function, but ideally low dimension informative summary (e.g., number of segregating sites, difference between two estimators of $\theta$)
- Define a rejection region $R$ such that the probability of observing a value of $T$ that lies in $R$ given that $H$ is true is equal to the desired rejection probability $\alpha$
  - E.g., given the hypothesis that $\theta = 5$ and a sample size of 20 (with no recombination) 95% of observations would have between 6 and 48 segregating sites.
  - In population genetics, rejection regions are often estimated by simulation
- In goodness-of-fit tests $H = \text{The assumed model is correct}$
  - May include statements about parameter values

### Statistical inference

1. Define model
2. Explore properties
3. Estimate parameters from data
4. Test goodness-of-fit
5. Refine model

### Issues
- Summary statistics
- Graphical representation
- Stochastic simulation
- Moment methods
- Likelihood
- Bayesian inference
- Outliers
- Heterogeneity
- Comparison of estimators
- Add parameters

### Model-testing in population genetics

- **Goodness-of-fit tests**
  - Tests for differences between summary statistics at a single locus
    - Watterson homozygosity test, Tajima $D_{\text{test}}$, Fu and Li $D^*$ test, Fay and Wu $H$ test, Haplotype-based tests
  - Tests for heterogeneity between loci, or classes of mutations
    - HK$A$ test, McDonald-Kreitman test, variance tests, Lewontin-Krakauer test
- **Likelihood ratio tests**
  - Calculate increase in likelihood due to addition of extra parameters to model
  - Less well developed due to computational burden. Possible for population growth, selective sweeps
Watterson (1977)

“There is no single statistic which is best for testing the most general departures from neutrality”

Moral: “You have to know what you are looking for!”

Frequency distribution tests

Ewens’ sampling formula and the Watterson Homozygosity test

<table>
<thead>
<tr>
<th>Frequency</th>
<th>HLA-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td></td>
</tr>
</tbody>
</table>

Watterson (1977) suggested using homozygosity as summary of the distribution. The observed homozygosity can be compared to the distribution expected from Ewens’ sampling formula

\[
\Pr\{n_1, n_2 \ldots n_k \mid k, n\} = \frac{n!}{k! n_1 n_2 \ldots n_k}
\]

Ewens (1972)

Homozygosity = \(\sum x_i^2 = 0.137\) \(\Pr(H \leq 0.137) < 0.05\)

Haplotypes and Watterson’s test

- If there is no recombination, Ewen’s sampling formula can be applied to haplotype diversity
  - e.g. Kaessmann et al. (2000) 69 worldwide sequences of human Xq13 (10.2kb), 33 segregating sites
    - \(K = 20\)
    - \(\hat{\theta} = 9.1\)
    - \(H = 0.133\)
    - \(P(H \leq 0.133) = 0.88\)

- But recombination violates the assumptions of Ewens’ formula
  - Can lead to Type I error (false rejection of null hypothesis)
Tajima’s $D$ test (1989)

- Two estimators of theta
  - Watterson’s estimate: number of segregating sites
  - Average pairwise diversity: sensitive to intermediate allele frequencies

\[
E[\pi] = \theta \\
E[S] = \theta \sum_{i=1}^{t} \frac{1}{i} \\
D = \frac{\pi - S / a_n}{\sqrt{\text{Var}(\pi - S / a_n)}}
\]

- Difference normalised by s.d. like Z statistic

- Negative values of $D$ indicate an excess of rare mutations, positive values indicate an excess of intermediate frequency mutations
- Critical regions obtained by coalescent simulation

\[
\pi < S / a_n \quad \text{if} \quad D < 0 \\
\pi > S / a_n \quad \text{if} \quad D > 0
\]

Factors influencing test power

- Test power is more influenced by the number of segregating sites than the number of sequences
- Recombination breaks up correlations in genealogical history between linked sites, reducing the influence of evolutionary stochasticity

\[
T_{\text{tot}} = 4N_e a_n
\]

- Critical regions for tests can be estimated using a lower bound for the population recombination rate, $4N_e r$
  - The assumption of no recombination is generally conservative

An example: human mtDNA

- Ingman et al. (2000) 52 complete mtDNA molecules from a worldwide sample (linguistic groups)
- 521 segregating sites excluding D-loop

\[
\pi = 44.2 \\
a_{52} = 4.52 \\
S / a_{52} = 115.3 \\
\sqrt{\text{V}(d)} = 31.8 \\
D = \frac{44.2 - 115.3}{31.8} = -2.23
\]

Probability of observing such an extreme value under neutrality = 0.01

Human mtDNA have an excess of low-frequency variants

→ Population growth, selection, or sampling?

Other tests on the frequency spectrum

- Fu and Li (1993) $D$ test

\[
E[\eta_e] = \theta \\
D = \frac{S - a_n \eta_e}{\sqrt{\text{Var}(S - a_n \eta_e)}}
\]

Without an outgroup, the test has to be adjusted ($D^*$)
A general class of tests: Fu (1995)

- The expected number of mutations with a derived frequency of $i$ in the sample is
  \[ E[\xi_i] = \frac{\theta}{i} \]

- There is a potentially endless supply of possible tests based around this result BUT
  - Much shared information
  - Large variance

- Fay and Wu (2000) suggest $H$ test is powerful for detecting selective sweeps
  \[ H = \frac{\pi - \theta_H}{\sqrt{\text{Var}(\pi - \theta_H)}} \]
  \[ \theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i \]

Haplotype-based tests

- Depaulis & Veuille (1998)
  - Number of haplotypes ($K$) given segregating sites
  - Haplotype diversity ($H$) given segregating sites
  \[ K_{\text{max}} = 2, \quad K_{\text{max}} = S + 1 \]
  \[ H = 1 - \sum_{i=1}^{n-1} i^2 \xi_i \]

- Both $K$ and $H$ are sensitive to recombination
  - Use lower or upper bounds of population recombination rate ($4N_e r$) to derive critical regions by simulation
  - Power depends on balance between sample size and $S$
  - Test influenced by both haplotype structure and frequency spectrum

Haplotype structure
- Strong haplotype structure: $K \ll S$; High diversity $H$
- Fragmented haplotype structure: $K = S$; Low diversity $H$

Balancing selection, bottlenecks
Selective sweeps, population growth

Limitations of single-locus tests

- Hard to reject neutral model
  - Single evolutionary history
  - Many parameters to estimate

- If do reject model, impossible to know whether locus reflects a genome wide pattern

- Need to look at variation between loci
Heterogeneity tests

The HKA test
- Hudson, Kreitman and Aguadé (1987)
  - Compare polymorphism and divergence at two or more loci within coalescent framework

\[
\theta = \frac{dE_{11}}{n} \quad \tau = \frac{dE_{22}}{n}
\]

Estimate parameters and calculate goodness-of-fit test statistic

Adh in Drosophila

\[
D. sechellia \quad (n = 1) \quad 210 \text{ diffs} \quad 18 \text{ diffs} \quad S = 9
\]

\[
D. melanogaster \quad (n = 81) \quad S = 8
\]

Solving the simultaneous equations
- \( \hat{\theta}_1 = 2.7 \)
- \( \hat{\theta}_2 = 0.7 \)
- \( \chi^2 = 6.09 \) \( P = 0.016 \)

* Fast / slow polymorphism in exon 4 leads to a two-fold difference in enzyme activity
* Cline in polymorphism: Fast more common in northern America and at higher altitudes

The McDonald-Kreitman test (1991)
- Compare patterns of polymorphism and divergence at different classes of interspersed mutations

Fixed Polymorphic

Nonsynonymous

Synonymous

Neutral mutations

Advantageous mutations

contingency table
Other tests of heterogeneity

- Variance in summary statistics
  - e.g. Is the variation in Tajima D statistic between loci greater than expected? (Frisse et al. 2001)

- Detection of outliers
  - e.g. Do certain loci show unusually large levels of geographic differentiation? (Lewontin and Krakauer 1973)

Selection: what are we looking for?

- What types of selection might we consider?
  - Directional selection (selective sweeps)
  - Balancing selection
  - Local adaptation (local selective sweeps)

- There are many ways of summarising data, how do we know which aspects will be most sensitive to the action of selection?
  - Models of selection
  - Formulation of test statistic

- How can we be sure that a deviation from the assumed model is due to selection?
  - Comparison to reference loci
  - Comparison across multiple populations
  - a priori knowledge

Balancing selection

- Alleles that are maintained at a constant frequency by natural selection are called balanced mutations
  - e.g. Mutations causing malaria-resistant haemoglobinopathies, Alcohol dehydrogenase in Drosophila melanogaster

- Neutral mutations that occur at linked sites on one background can only cross onto the other by recombination
- Balancing selection leads to an increase in local diversity and strong haplotype structure

β-globin

- Harding et al. (1997)
  - 349 chromosomes from African and European populations
- Very strong haplotype structure
- Intermediate frequency alleles
The hitch-hiking effect of beneficial mutations

A new advantageous mutation appears in the population

The mutation sweeps to high frequency, dragging linked mutations, and reducing variability in the region

Diversity recovers slowly; the first mutations to appear are at low frequency

Standard neutral model applies

Strong linkage disequilibrium, high frequency derived mutations

Skewed allele frequency spectrum, low diversity

Fay and Wu’s $H$ test and hitch-hiking

During selective sweeps, recombination can allow chromosomes to escape the sweep

Following fixation, sites linked to the selected mutation will have an excess of high frequency derived mutations

The estimator of theta we met earlier, $\theta_H$, is heavily influenced by high frequency derived mutations. A positive $H$ statistic is indicative of an excess of high frequency derived mutations.

This signal of a selective sweep rapidly disappears after fixation of the advantageous mutation (and only has power some distance from the site of selection)

Monoamine-oxidase

- Gilad et al. (2002)
  - Significant excess of high-frequency derived mutations
  - Locus associated with behavioural abnormalities +/- impulsive behaviour

Local adaptation

- Localised selective sweep due to geographically restricted selection pressure
  - e.g. malaria, drug or pesticide treatment

- Leads to local reduction of variability, strong linkage disequilibrium and strong geographical structuring of genetic variability around the locus of importance
**Duffy locus variation**

- Hamblin et al. (2002)
  - Null Duffy allele FY*O at fixation in Hausa
  - FY*A fixed in Chinese

- Test suggested by Lewontin and Krakauer (1973)
  - BUT need to account for sampling distribution of $F_{ST}$
  - Taylor et al. (1995) suggested an *a priori* approach

**Distinguishing selection from demographic effects**

- Demographic processes can mimic selection
  - Population growth can look like selective sweeps

- Population subdivision can look like balanced selection

- Differences between loci can distinguish between genome-wide effects and the local effect of natural selection
  - BUT need to know about variance of demographic processes……