

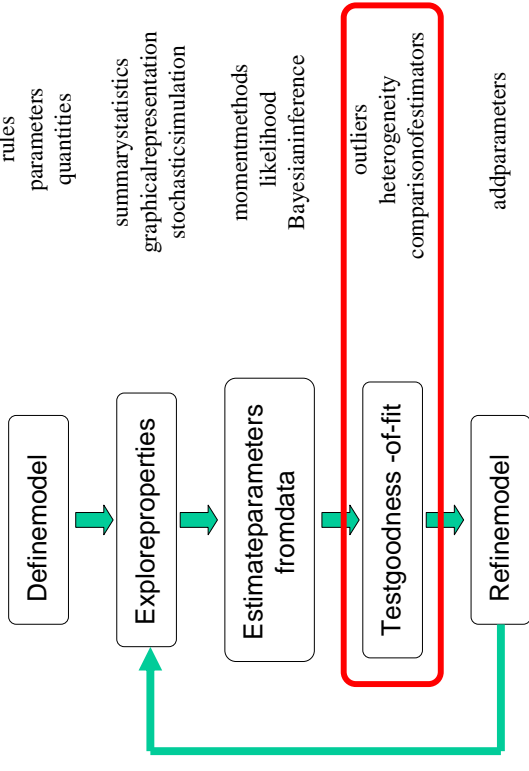
Population genetic inference

Date	Topic	
22 nd Jan	Good questions in population genetics	GM
29 th Jan	Principles of population genetic inference	GM
5 th Feb	Recombination in the coalescent	JH
12 th Feb	Natural selection	GM
19 th Feb	Demographic models	GM
26 th Feb	Combinatorics of the coalescent	JH
5 th March	Population genetics of disease mutations	GM
12 th March	Model organisms	GM

Reading

Nielsen, R. 2001. Statistical tests of neutrality in the genome. *Genomics*, **86**: 641-647
 Kreitman, M. 2000. Methods to detect selection in populations with applications to the human genome. *Hum. Genet.* **106**: 519-559

Statistical inference



Test statistics and hypothesis testing

- Let H_0 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 of the data
- Define a rejection region R such that the probability of observing a value of T that lies in R given that H_0 is true is equal to the desired rejection probability α
 - 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 $H_0 = \text{The assumed model is correct}$
 - May include statements about parameter values

Model-testing in population genetics

- Goodness-of-fit tests
 - Tests for differences between summary statistics at a single locus
- Watterson homozygosity test, Tajima D test, Fu and Li D^* test, Fay and Wu H test, Haplotype-based tests
 - Tests for heterogeneity between loci, or classes of mutations
- HKA test, McDonald-Kreitman test, variance tests, Lewontin D test, Krakauer test
 - Likelihood ratio tests
 - Calculate increase in likelihood due to addition of extra parameters
 - 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



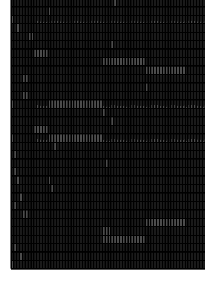
Watterson (1977) suggested using homozygosity as a summary of the distribution. The observed homozygosity can be compared to the distribution expected from Ewens' sampling formula

$$\Pr\{n_1, n_2, \dots, n_k \mid k, n\} = \frac{n!}{k! \cdot n_1! \cdot n_2! \cdot \dots \cdot n_k!} \quad \text{Ewens (1972)}$$

$$\text{Homozygosity} = \sum_i \frac{1}{n_i} \quad \Pr(H \leq 0.137) < 0.05$$

Haplotypes and Watterson's test

- If there is no recombination, Ewens' sampling formula can be applied to haplotype diversity
 - e.g. Kaessmann *et al.* (2000) 69 worldwide sequences of human Xq13 (10.2 kb). 33 segregating sites



$$K = 20$$

$$\hat{\theta} = 9.1$$

$$H = 0.133$$

$$P(H \leq 0.133) = 0.88$$

- But recombination violates the assumption of Ewens' formula
 - Can lead to Type I error (false rejection of null hypothesis)

Tajima's D test (1989)

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

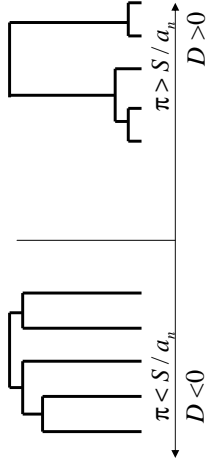
$$E[\pi] = \theta$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

$$D = \frac{\pi - S / a_n}{\sqrt{\text{Var}(\pi - S / a_n)}}$$

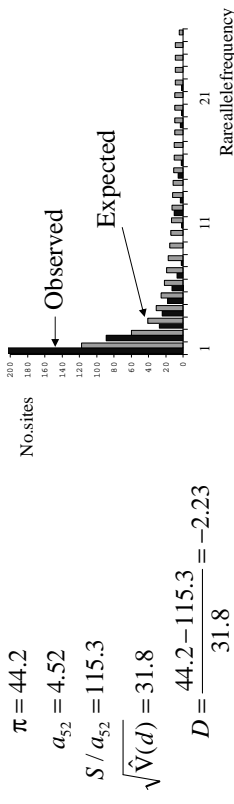
Difference normalised by s.d. like Z statistic

- 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



An example: human mtDNA

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



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?

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

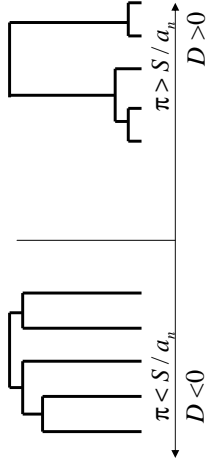
$$E[\pi] = \theta$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

$$D = \frac{\pi - S / a_n}{\sqrt{\text{Var}(\pi - S / a_n)}}$$

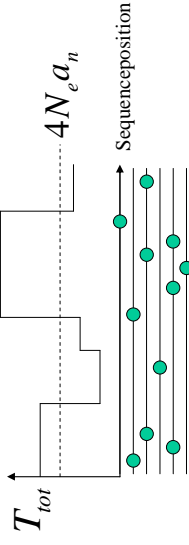
Difference normalised by s.d. like Z statistic

- 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



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



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

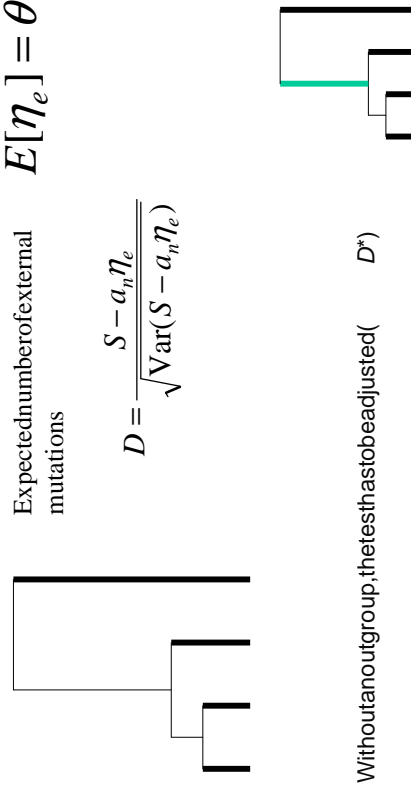
Other tests on the frequency spectrum

- Fu and Li (1993) D test

Expected number of external mutations

$$E[\eta_e] = \theta$$

$$D = \frac{S - a_n \eta_e}{\sqrt{\text{Var}(S - a_n \eta_e)}}$$



A general class of tests: Fu (1995)

- The expected number of mutations with a derived frequency of i in the samples is $E[\xi_i] = \frac{\theta}{i}$
- There is a potential to yield less 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

$$H = \frac{\pi - \theta_H}{\sqrt{\text{Var}(\pi - \theta_H)}} \quad \theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

Haplotype-based tests

Haplotype-based tests

- Depaulis & Veille (1998)
 - Number of haplotypes (K) given segregating sites $K_{\min} = 2, K_{\max} = S + 1$
 - Haplotype diversity (H) given segregating sites $H = 1 - \sum_i f_i^2$
- Both K and H are sensitive to recombination
 - Use lower or upper bounds of population recombination rate (4 $N_e \mu$) to derive critical regions by simulation
 - Power depends on balance between sample size and S
 - Test influenced by both haplotype structure and frequency spectrum



Strong haplotype structure
 $K \ll S$
High diversity H



Fragmented haplotype structure
 $K \approx S$
Low diversity H

Balancing selection, bottlenecks

Selective sweeps, population growth

Limitations of single-locus tests

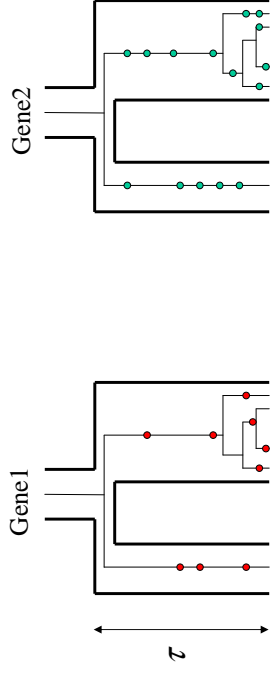
- Hard to reject neutral model
 - Single evolutionary history
 - Many parameters to estimate
- If do reject model, impossible to know whether locus reflects a wide pattern
- Need to look at variation *between* loci

genome

Heterogeneity tests

The HK A test

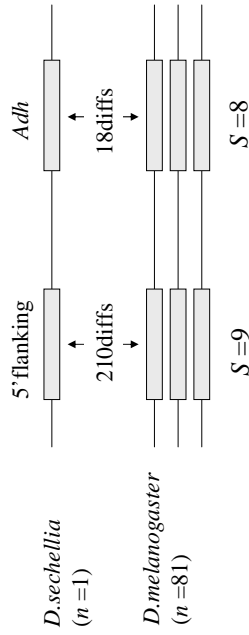
- Hudson, Kreitman and Aguadé (1987)
 - Compare polymorphism and divergence at two or more loci within a coalescent framework



$$E[S_1] = \theta_1 a_n \quad E[d_1] = \tau \theta_1 \quad E[S_2] = \theta_2 a_n \quad E[d_2] = \tau \theta_2$$

Estimate parameters and calculate goodness-of-fit test statistic

Adh in Drosophila



Solving the simultaneous equations

$$\hat{\tau} = 13.4 N_e \text{ generations}$$

$$\hat{\theta}_1 = 2.7$$

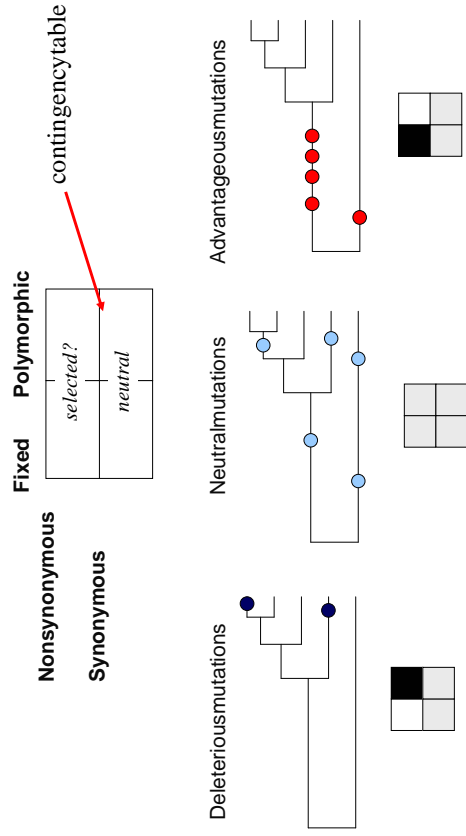
$$\hat{\theta}_2 = 0.7$$

$$\chi^2 = 6.09 \quad P = 0.016$$

- Fast/slow polymorphism on 4 leadstoatwo -fold difference in enzyme activity
- Cline in polymorphism: Fast more common in northern America and altitudes

The McDonald-Kreitman test (1991)

- Compare patterns of polymorphism and divergence at different classes of interspersed mutations



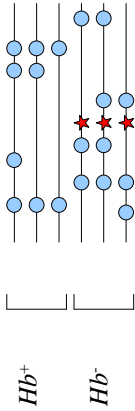
Other tests of heterogeneity

- 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?

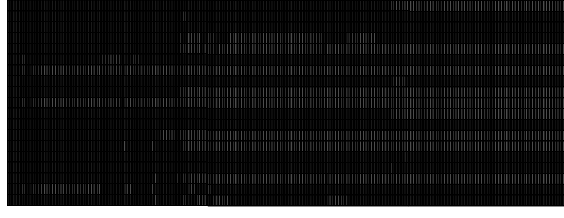
- What types of selection might we consider?
 - Directional selection (selective sweeps)
 - Balancing selection
 - Local adaptation (local selective sweeps)
- There are many ways of summarising data, how do we know which aspects will be most sensitive to the action of selection?
 - Models of selection
 - 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

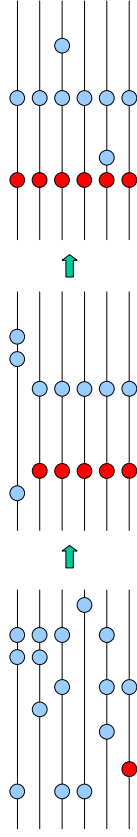
- Alleles that remain at a constant frequency by natural selection are called balanced mutations
 - e.g. Mutations causing malaria resistance in the hemoglobinopathies, Alcohol dehydrogenase in *Drosophila melanogaster*
- 
- Neutral mutations that occur at linked sites on one background cross onto the other by recombination
 - Balancing selection leads to an increase in local diversity and 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

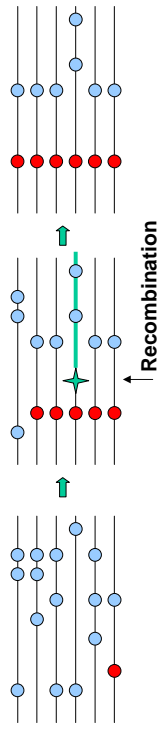
Standard neutral model applies

Strong linkage disequilibrium, high frequency derived mutations

Skewed allele frequency spectrum, low diversity

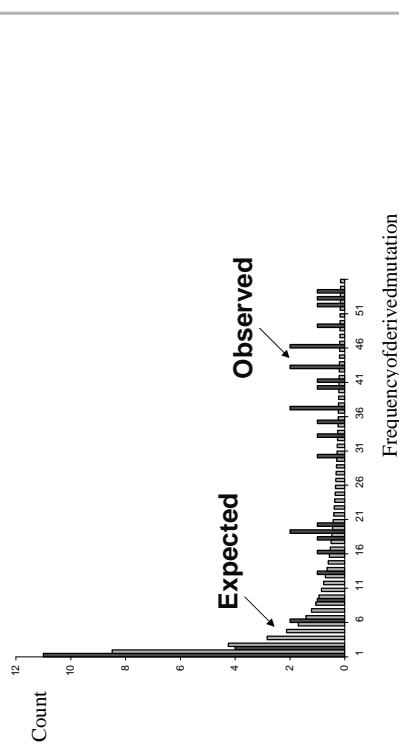
Fay and Wu's H test and hitch-hiking

- 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 the time to the most recent common ancestor, θ_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 selective sweep rapidly disappears after fixation on the site of advantageous mutation (and only has power some distance from the site of selection)



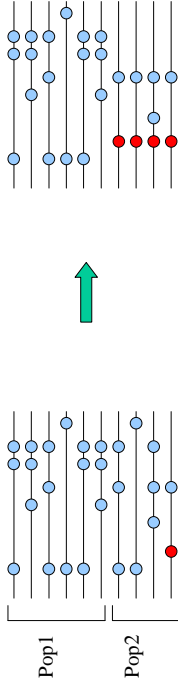
Monoamine-oxidase

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



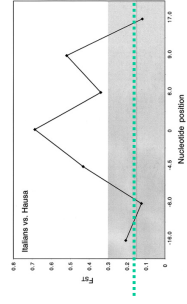
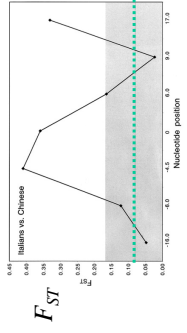
Local adaptation

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



- Tests 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
- Populations 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.....