Summer Institute in Statistical Genetics
University of Queensland, Brisbane
Module 5: Population Genetic Data Analysis

Models for Selection

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

Feb 10, 2017

Contents:

1. Natural Selection
2. Selection and LD
3. A single-locus test of selection
4. Using F_{ST} to detect directional and balancing selection

1. Natural Selection

2. Selection and LD

3. A single-locus test of selection

4. Using F_{ST} to detect directional and balancing selection

In genetics, **selection** is the process whereby the survival and reproductive success of individuals depends on their genotypes.

The effects of selection can change over time and space as environment changes.

Some types of selection:

- *Positive* or *adaptive*: favours an allele;
- *Negative* or *purifying*: disfavours an allele;
- *Balancing*: favours multiple alleles in a location;
    - e.g. heterozygote advantage or frequency-dependent selection.
- *Diversifying*: favours different alleles in different locations.
- *Sexual selection*: is driven by reproductive success, involving e.g. between-male competition and/or female preferences.

Drift tends to be more important than selection in explaining patterns of gene frequency in small, isolated populations, or if the mutant is (almost) neutral. Selection becomes more important as the population size increases.

The effects of selection

The importance of selection in shaping the human genetic variation that we observe today has long been controversial:

- Since Darwin most geneticists believed that the effects of selection are ubiquitous
- Kimura (1970s) neutral theory: most variation is neutral and reflects migration and drift, not selection.

Today the picture is still not entirely clear.

- There has been progress in methods to identify the effects of selection on the genome, from dense SNP and sequence data.
- Evidence for strong selection has been identified and attributed e.g. to milk digestion or resistance to infectious disease.
- Still only a small fraction of the genome has been identified as showing evidence of selection, but this fraction is increasing and there are likely to be many more selection effects that have not yet been detected.

A mathematical model of selection

Consider a large, random mating population such that selection acts on a single diallelic locus with initial allele fractions p and q. If mutation occurs at rate μ in each direction then in the next generation we have

Genotype	AA	Aa	aa
Fraction in zygotes	p'^2	$2p'q'$	q'^2
Relative fitness	1	$1+s_1$	$1+s_2$

where $p' = p(1-\mu) + q\mu$ and $q' = p\mu + q(1-\mu)$. The relative fitnesses are proportional to the probabilities that individuals survive to adulthood and so the genotype fractions in the adult population have the form:

Genotype	AA	Aa	aa
Adult fraction	p'^2/c	$2(1+s_1)p'q'/c$	$(1+s_2)q'^2/c$

where c is the constant that makes these fractions sum to one.

Some special cases:

Deleterious recessive: $s_1 = 0, \quad s_2 = -s$
In the absence of mutation the deleterious a allele will be eliminated from the population, but if $\mu > 0$ then we obtain a mutation-selection *equilibrium*.

Deleterious dominant: $s_1 = s_2 = -s$
Mutation-selection equilibrium is again reached when $\mu > 0$, but with a much lower fraction of the deleterious allele than in the recessive case.

Heterozygote advantage: $s_1 > 0, \quad s_2 \leq 0$
Also known as heterosis or overdominance. The best-known example is sickle-cell anaemia in Africans; fitness coefficients have been estimated at $s_1 = 0.14$ and $s_2 = -0.84$.

Web app + R code for Wright-Fisher model with selection

You can run simple Wright-Fisher simulations with selection at:

http://genomicsresearch.org:3838/sample-apps/popgen/drift

- Set simulation parameters in panel on left. NB the two fitness parameters are, in our notation, $1+s_1$ and $1+s_2$.
- Left plot shows diploid genotype frequencies.
- Right plot shows heterozygosity, both expected assuming Hardy-Weinberg equilibrium, and observed.

The app is based on `R` code in file `rmsel.R`, which has an additional plotting function `plotall` that plots a histogram of the minor allele count over the simulation and reports the mean allele count after omitting the first 20% of iterations. The two plots generated by the web app are available as functions `plotgen` and `plothom`. Function `rmsel` generates a (large) matrix that must be stored and passed as a parameter to the plotting function. Non-default parameter settings should also be assigned in any call to a plotting function.

Mutation-selection equilibrium exercise

Examine how `mu` and `s2` affect the mean allele fraction at mutation-selection equilibrium in a deleterious recessive model. Keep `s1` at 0. Set `ngen` to be large to approach equilibrium – high precision is not needed, and you can assess the reliability of your answers by repetition of the simulation.

1. What (approximately) is the equilibrium fraction of the deleterious allele if $\texttt{mu} = 0.001$ and $\texttt{s2} = -0.8$?
2. Keeping $\texttt{s2} = -0.8$, how does the equilibrium fraction vary if you change the mutation rate? Can you guess a simple functional form for the (approximate) relationship between `mu` and the equilibrium allele fraction?
3. Now keep `mu` at 0.00075 and describe how the equilibrium allele fraction changes as `s2` changes between 0 and -1.

Balancing selection: in the absence of mutation, find values of `s1` and `s2` that lead to approximately stable non-zero frequencies for both alleles.

1. Natural Selection
2. Selection and LD
3. A single-locus test of selection
4. Using F_{ST} to detect directional and balancing selection

Factors that affect LD: selection

- **LD between multiple causal genes:** If two or more genes are involved in determining a phenotype that affects survival or reproductive success, then there can be LD between the genes even if they are unlinked ($\rho = 1/2$).
- **Extensive, high LD due to recent positive selection:** if an allele has been subject to recent positive selection, it will have increased rapidly in frequency. At first, an entire chromosome carrying the allele will be favoured. Over time, the part of the chromosome favoured by selection becomes narrower due to recombination, but for many generations there can remain an extensive genomic region in high LD with the selected allele.
  - Thus, regions of high LD (or highly-conserved haplotypes) are a useful indicator of effects of selection.

The effects of selection on LD can be investigated in simulations of the W-F model with selection.

LD under selection exercise

By running the R script `ldsel.R` you define `ldsel` which simulates a haploid, two-locus Wright-Fisher model with selection. As for `ldsim` (Lecture 1), function `ldsel` generates a (large) matrix that must be stored and passed as a parameter to the plotting function `plotld` to obtain plots of haplotype frequencies, D' and r^2 over the simultaion. Again, non-default parameter settings should be assigned in the call to `plotld`.

1. Previously you used `ldsim` to examine the effects on r^2 and D' of different values for the recombination fraction `rho` and mutation rate `mu`.
2. Let `rho = 0` (no recombination) and explore small, positive values for both `mu = 0.001` and `s`, which generates selective advantage for an allele at one of the two loci.
   - Which allele is selected for?
   - What happens to the two measures of LD before and after reaching mutation-selection equilibrium ?
   - How does the effect change if you know increase `rho`?

1. Natural Selection
2. Selection and LD
3. A single-locus test of selection
4. Using F_{ST} to detect directional and balancing selection

Tajima's D

- Tajima's D statistic is the difference between two different estimators of the scaled mutation rate $\theta = 2N\mu$:
  1. Watterson's estimator (Lecture 2) $\hat{\theta}_W = S_n / \sum_{j=1}^{n-1} \frac{4}{j}$,
  2. Δ_n = average number of pairwise differences among n sequences.

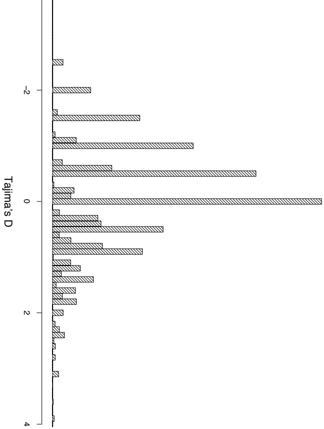
  $$D = (\Delta_n - \hat{\theta}_W)/SD,$$

  where SD is the value required to standardise D (variance $= 1$).
- The probability that a diallelic site differs in two random sequences is θ, so $\mathbb{E}[\Delta_n] = \mathbb{E}[\hat{\theta}_W] = \theta$ and thus $\mathbb{E}[D] = 0$.

Significantly positive or negative values of D indicate more or less genetic diversity given the number of segregating sites S:

- more diversity implies an excess of sites with both alleles common, which could be due to balancing selection.
- less diversity implies an excess of sites with a rare allele, which could be due to a selective sweep eliminating common variants.

Unfortunately, these signals could also reflect population decline or growth, respectively. However demographic effects should impact all loci, whereas selection effects may be locus-specific.



Histogram of 10 000 values of (unstandardised) Tajima's D statistic simulated under the standard coalescent with $n = 6$, $m = 5$, and $\theta = 1.255$. Bin width $= 0.1$.

```
dsim <- function(niter=10000, nsamp=6, nloc=5, theta=1)
{
  ns1 <- nsamp-1
  acc <- 0
  for(i in 1:niter)
  {
    hap <- matrix(0,nsamp,nloc)
    w <- rexp(ns1,(2:nsamp)*(1:ns1)/2)
    L <- sum((2:nsamp)*w)
    param <- L*theta/(2*nloc)
    nmut <- rbinom(1,nloc,1-exp(-param))
    if(nmut==0) mut <- rep(0,ns1)
    else mut <- hist(runif(nmut),br=c(0,cumsum((2:nsamp)*w)/L),\
                                                  plot=F)$c
    loc <- 1
    for(j in 1:ns1)
    {
      hap[j+1,] <- hap[sample(j,1),]
      if(mut[j]>0) for(i in 1:mut[j])
      {
        hap[sample(j+1,1),loc] <- 1
        loc <- loc+1
      }
    }
    m <- apply(hap,2,sum)
    acc <- c(acc,sum(2*m*(nsamp-m)/nsamp/ns1)-nmut/sum(1/1:ns1))
  }
  acc[-1]
}
```

1. Natural Selection

2. Selection and LD

3. A single-locus test of selection

4. Using F_{ST} to detect directional and balancing selection

Loci subject to directional selection may respond differently in different environments, or an advantageous allele may arrive in different populations at different times.

- In either case the effect of selection at a locus may be detectable as an unusually high F_{ST} value.
- Conversely balancing selection can be detected as unusually low F_{ST}.

"High" and "low" values of F_{ST} can be assessed empirically with reference to the genome-wide distribution of F_{ST} estimates

- but it is then hard to assess significance or choose thresholds according to quantitative criteria;
- also hard to assess effects of sample and subpopulation sizes.

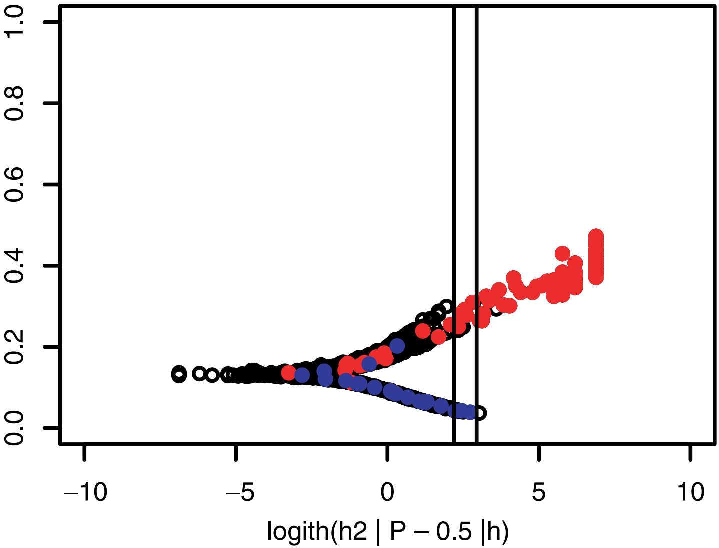
Beaumont and Balding (2004) addressed this problem using the likelihood model for F_{ST} introduced in my previous lecture.

- The b_j are nuisance parameters reflecting demography (subpopulation sizes and histories).
- A posterior distribution for an a_i that lies almost entirely above zero indicates directional selection.
- A posterior distribution for an a_i that mainly supports values below zero indicates balancing selection.

There has been much subsequent development of this model, for example Foll and Gaggiotti (2008), Guo et al. (2009), Coop et al. (2010), Galinsky et al. (2016) and Duforet-Frebourg et al. (2016). Some of the changes are:

- Different ways to decide if a value of a_i is "significant".
- Relating F_{ST} to principal components (which usually reflect geography) or other environmental covariates, rather than just a subpopulation label.

From Beaumont and Balding (2004):



Simulated data:
Red: directional
Blue: balancing
Open: neutral.

Vertical lines indicate 1% and 5% significance levels.
y-axis shows F_{ST} but inference is based on a_i.

References

Beaumont, M. and Balding, D. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, 13(4):969–980.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423.

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., and Blum, M. G. B. 2016. Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, 33(4):1082–1093.

Foll, M. and Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, 180(2):977–993.

Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., and Price, A. L. 2016. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, 98(3):456–472.

Guo, F., Dipak, K., and Holsinger, K. 2009. A bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*, 104.485:142–154.