Trees of Life (BIOL30004)
Lecture 3
Coalescent Theory II

Mark Beaumont

Spring 2017

Summary of Lecture 3

1. Summarise coalescent main features.
2. Apply theory to Sardinian Y-chromosome data.
3. Demographic history.
4. The coalescent with variable population size.
5. Bottlenecks and population growth in the fruit fly Drosophila melanogaster.

Main features of the coalescent I

- The expected time to the most recent common ancestor (T_{MRCA}) is $2N$ for a pair of genes.
- The expected T_{MRCA} for a large sample is approximately $4N$, so on average approximately half of the length of a gene genealogy is taken up with just two lineages.
- Genealogies are very variable (in fact the standard deviation in the T_{MRCA} is approximately $2N$ for a large sample). Most of this variability comes from the time taken for the last 2 lineages to coalesce.
- For a large sample size the coalescence rate declines approximately with the square of the number of lineages at any time, which means that there is an intense rate of coalescence to begin with and then this rapidly falls off.

Main features of the coalescent II

- For a given mutation rate, the number of mutations in a sample — *i.e.* the number of SNPs — just depends on the total length of the tree. So two different genetic regions (loci) may have very different numbers of SNPs not because of any variability in mutation rate, but just because they have different genealogies.

Application to the Y-chromosome data I

What is the effective population size for the Sardinian Y chromosome?

Francalacci *et al* do not provide an estimate for the effective population size, but we can do that from their data.

- There are 11763 SNPs in their data set, which we equate to the number of mutations in the genealogy.
- The expected sum of the branch lengths in the genealogy (in scaled time units) is $\sum_{i=1}^{n-1} 1/i$, and for $n = 1209$ this gives 7.67.
- Watterson's estimator

$$\hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

then gives 1532.8 as our estimate of θ.

Application to the Y-chromosome data II

- Since the Y chromosome is haploid $\theta = 2N\mu$ when N is the effective size of the male population.
- Using various arguments for calibrating the mutation rate, the authors estimate that it is $0{\cdot}53 \times 10^{-9}$ per nucleotide site per year.
- We can use the estimate of the male generation time given by the Poznik *et al* paper, of 31.5 years.
- The authors looked at $8{\cdot}97 \times 10^{6}$ nucleotides, so the total mutation rate per generation for this section of the Y is

$$8{\cdot}97 \times 10^{6} \times 0{\cdot}53 \times 10^{-9} \times 31{\cdot}5 = 0{\cdot}150$$

- So if we divide our estimate of θ by $2 \times 1{\cdot}5$ we get an estimate of the effective number of males, based on the Y chromosome.
- This is 5117.6

Application to the Y-chromosome data III

- For comparison, the Poznik *et al* paper estimates the effective size to be 4500 for their worldwide sample. But they used a mutation rate of $0{\cdot}82 \times 10^{-9}$, and if we use this for the Sardinian data, we get 3307.7.
- So, even if there are some uncertainties, given that these are completely independent data, with quite different methods for measuring genetic variation, the estimates of N are pretty close.

Is there any consistency between the effective population size that is estimated and the estimated T_{MRCA}?

- Both in the Poznik *et al* paper (at least for all the data) and the Francalacci paper the T_{MRCA} is estimated using a molecular clock argument (like Cann *et al* all those years ago ...) from the reconstructed tree. None of this requires coalescent theory.

Application to the Y-chromosome data IV

- By contrast the estimate of N from the Francalacci paper (and also the Poznik paper), uses Watterson's estimator, which depends on coalescent theory (just using the sequence data, not the branch length information).
- So it is worthwhile checking whether the empirical estimate of the T_{MRCA} matches what we expect given our estimate of N.
- The expected value is $2N$ generations (for haploid N). Assuming a generation time of 31.5 years, we get $2 \times 5117{\cdot}6 \times 31{\cdot}5 = 322,409$ for the Sardinian data (compare with their estimate of 200,000 years).
- For the Poznik *et al* data we get $2 \times 4500 \times 31{\cdot}5 = 283,500$ (compare with their estimate of 139,000 years).
- So there is a bit of a discrepancy here, in the same direction for both data sets. It looks as though there are more mutations in the genealogy than we expect, given the tree height. Why might that be?

Application to the Y-chromosome data V

- One explanation is that the demographic history of the population might not conform the assumptions of the coalescent model (constant population size; closed population).

Demographic history I

The phrase 'demographic history' is frequently used in the context of population genetics. A useful definition is provided by Hey and Machado (2003):

> *The reproductive history of a population or group of populations. This can include population sizes, sex ratios, migration rates, population splitting events, variation in reproductive rates and times among organisms, as well as variation over time in all of these quantities.*

Demographic history II

A key point to note:

- In general, although coalescent theory underpins the methods, we are not interested in the gene trees themselves. Why?
  - Because gene-trees are so variable, it is difficult to 'read-off' history from a reconstructed tree because it will vary from tree to tree.
  - Only in the case of the Y-chromosome and mtDNA can we get high resolution trees: in all other cases recombination breaks up the haplotypes, so there are many genealogies across the genome.
- So we treat trees as a 'nuisance parameter' and average over the possible trees that could have given rise to the data.
- So, ironically, given the title of this course, although gene-trees underpin the results, much of the remainder of my part of the course will not show any pictures of reconstructed gene trees.
- This is in contrast to Davide's part of the course, where the correct reconstruction of the phylogeny, and ancestral states, is the main focus.

The coalescent with variable population size I

- In the (unscaled) coalescent model we have seen that the distribution of waiting time until the first coalescence is exponentially distributed with a mean of

$$\frac{4N}{k(k-1)}.$$

  So the average waiting time is proportional to the population size.
- If the the population size is varying over time, then when the population size is large you have to wait a long time, and vice versa.
- This can be written down mathematically but requires a bit of calculus, and I won't give it here (look at the Hein *et al.* book, if you are interested in the details.

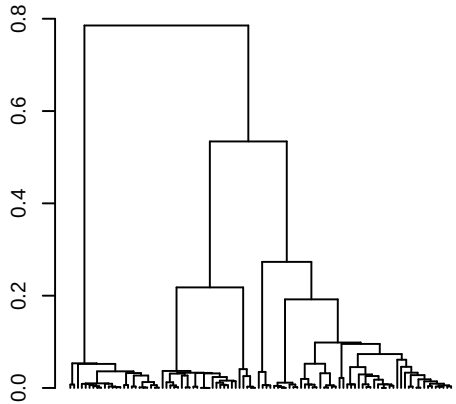
The coalescent with variable population size II

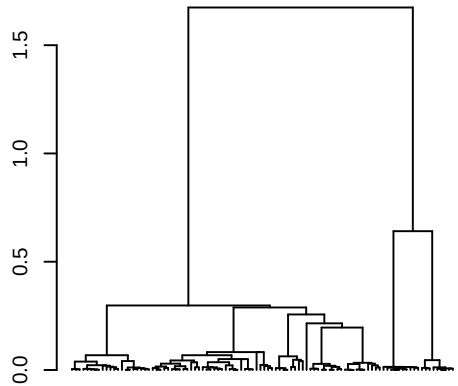
- A commonly-used model of population size change is the exponential model [note, don't confuse this with the exponential model behind the coalescent . . .].
- In this case, assuming time is scaled in units of twice the current population size N_C, we model the population size at any time in the past $N(t)$ relative to N_C as

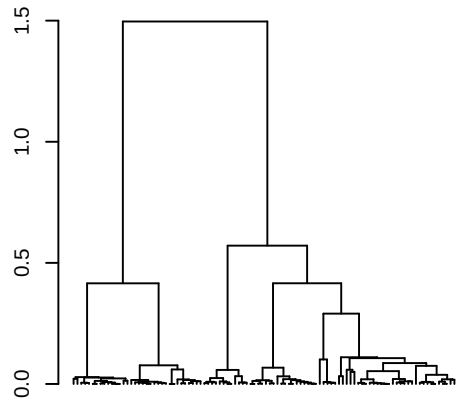
$$\frac{N(t)}{N_C} = e^{-\beta t}.$$

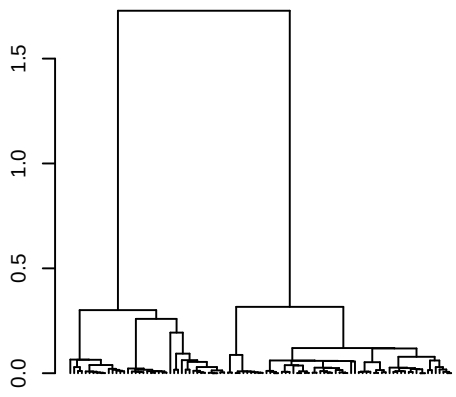
- In this case, because time is scaled, the growth rate $\beta = 2N_C b$, where b is the standard ecological growth rate. You can see that when $b = 0$ then $N(t)/N_C = 1$ and it behaves like the standard (scaled) coalescent model.
- Also note the minus sign: if the population size is growing (β is positive), then it is decreasing backward into the past, and vice versa.

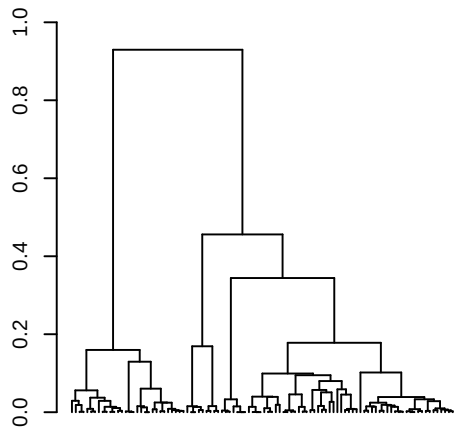
One of the earlier examples of trees with the standard coalescent

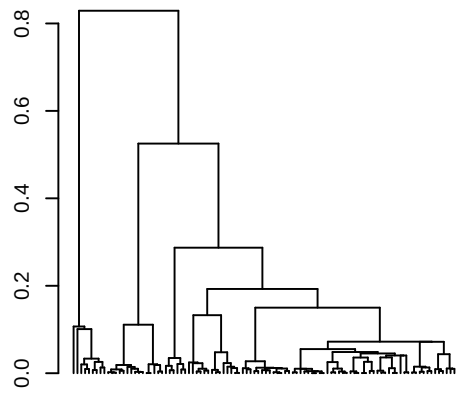






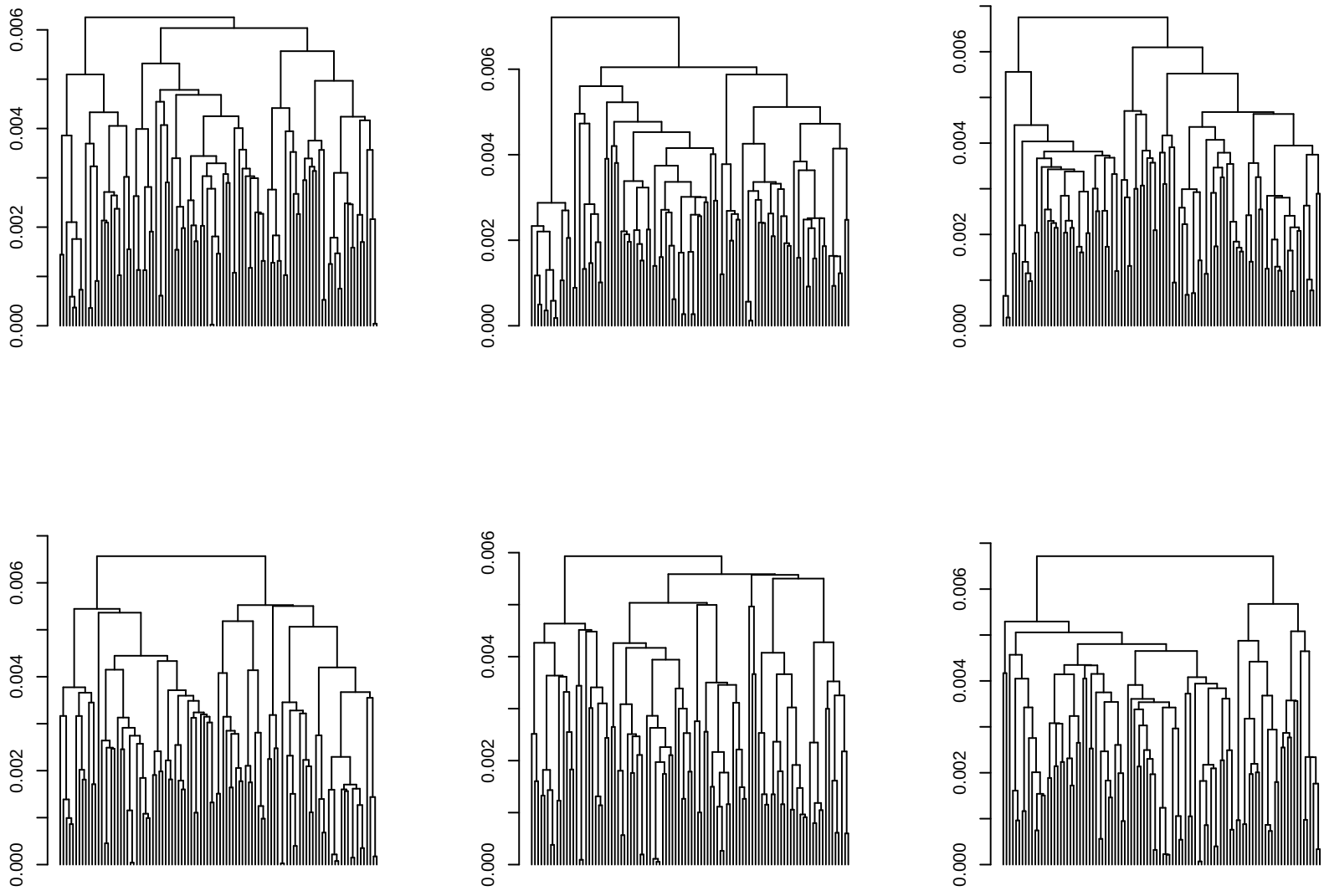






Example with growing population

$\beta = 1000$

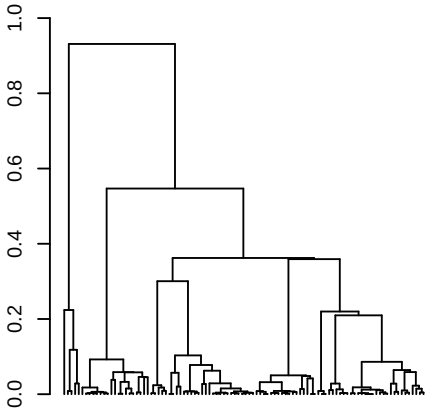


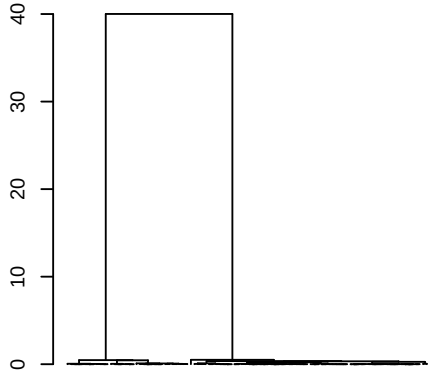
Some notes on the previous slide

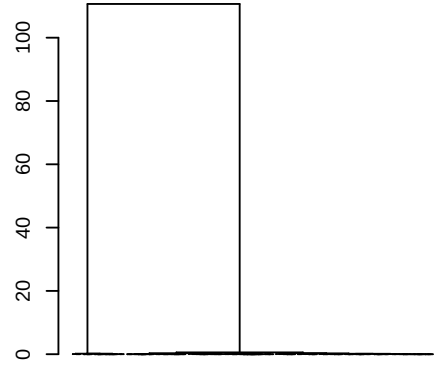
- Note the smaller variance in waiting times — e.g. the T_{MRCA} is generally around 0.006.
- Notice that the time of the MRCA in these scaled units is much shorter than with the standard coalescent. This is because it is scaled in units of the current population size, which is rapidly getting smaller into the past.
- You can hack the code I gave for lecture 2 by just changing the *ms* command from `"-T"` to `"-T -G 1000"`. [You might find it easiest to edit this directly with R using `File -> Open script...` Otherwise use a text editor such as notepad.
- As noted before, Hudson's *ms* scales time in $4N$ generations rather than $2N$. If we used the $2N$ scaling the time on the y-axis would be double that shown in the figure.

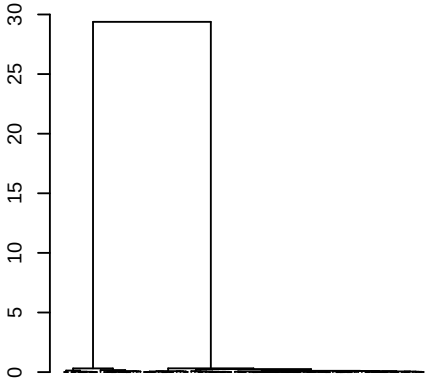
Example with declining population

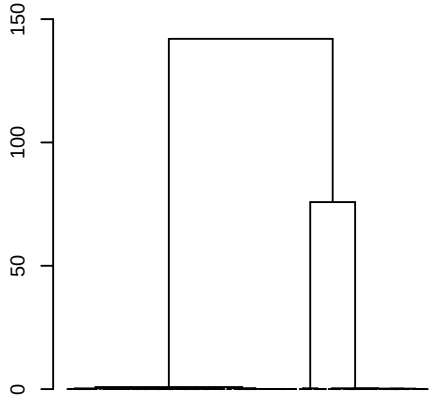
Corresponds to *e.g.* a population of initially constant size 20,000 declining to a current size of 100, over a period of 1000 generations.

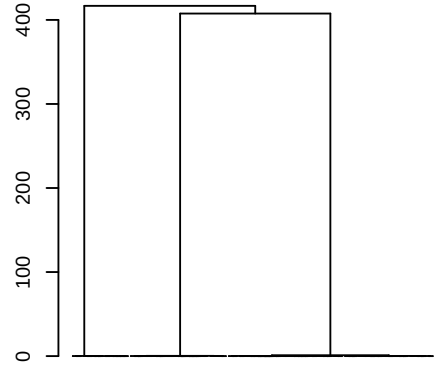












Notes on the genealogy of a declining population

- The *ms* command for this is `"-T -G -2.119 -eG 2.5 0.0"`.
- Note that we had to start with an initially constant size population. If we had used exactly the same model as before, but with a negative β (expanding into the past) there is a high chance the genealogy will get infinitely long.
- We get the opposite of the features seen in expanding populations: there is much more variability in the distribution of coalescence times.
- Note that time is scaled by $4N$.
- There is a bimodal distribution of coalescence times: those that occur on very short time scales corresponding to the current population size (of the order of 1 in *ms*-scaled time), or those on much longer time scales corresponding to the ancestral population size (of the order of 200).

The effect of demographic history on mutations and haplotype frequencies

- As before, mutations are place randomly on these genealogies.
- So for a growing population the terminal branches take up a much bigger proportion of the total branch length in the tree than in the standard coalescent.
- Mutations on these branches will only have one descendent: *i.e.* they will be unique; so-called 'singleton mutations'.
- By contrast with declining populations, either: 1) the T_{MRCA} is very recent, in which case there is very little genetic variability; or 2) there are deeply diverged haplotypes, differing by many mutations, corresponding to mutations along the branches closer to the root of the tree.
- In growing populations different loci (parts of the genome) have similar patterns of genetic variability, whereas in declining populations there is a lot of variability between loci.

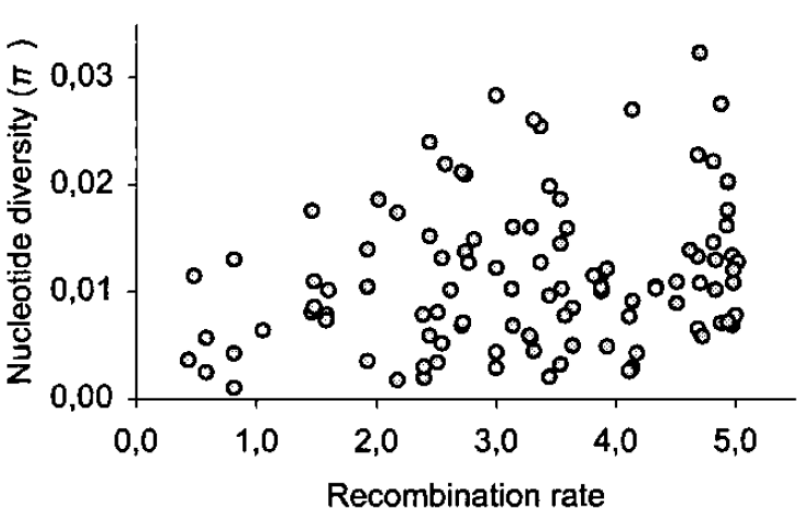
Example: Demographic history of African populations of *Drosophila melanogaster*

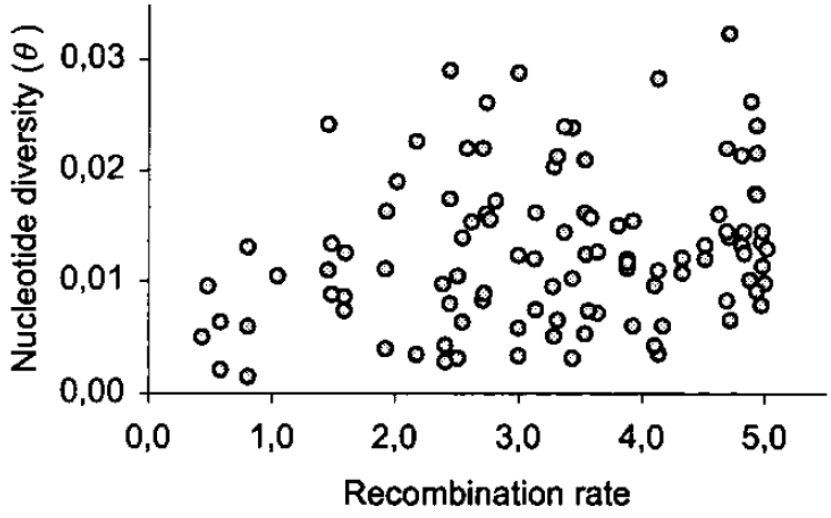
This example is based on the following papers in the reading list: Glinka *et al* (*Genetics*, 2003); Haddrill *et al* (*Genome Research*, 2005); Ometto *et al* (*Molecular Biology and Evolution*, 2005).

The fruitfly *Drosophila melanogaster* originates in sub-Saharan Africa. It is commensal with humans and appears to have expanded its range, following humans over the last 10,000–15,000 years. Much of the focus of these 3 papers is trying to identify the effects of selection, and also in examining European populations. However I will only focus on the demographic aspects, and on the African populations.

X-chromosome variation

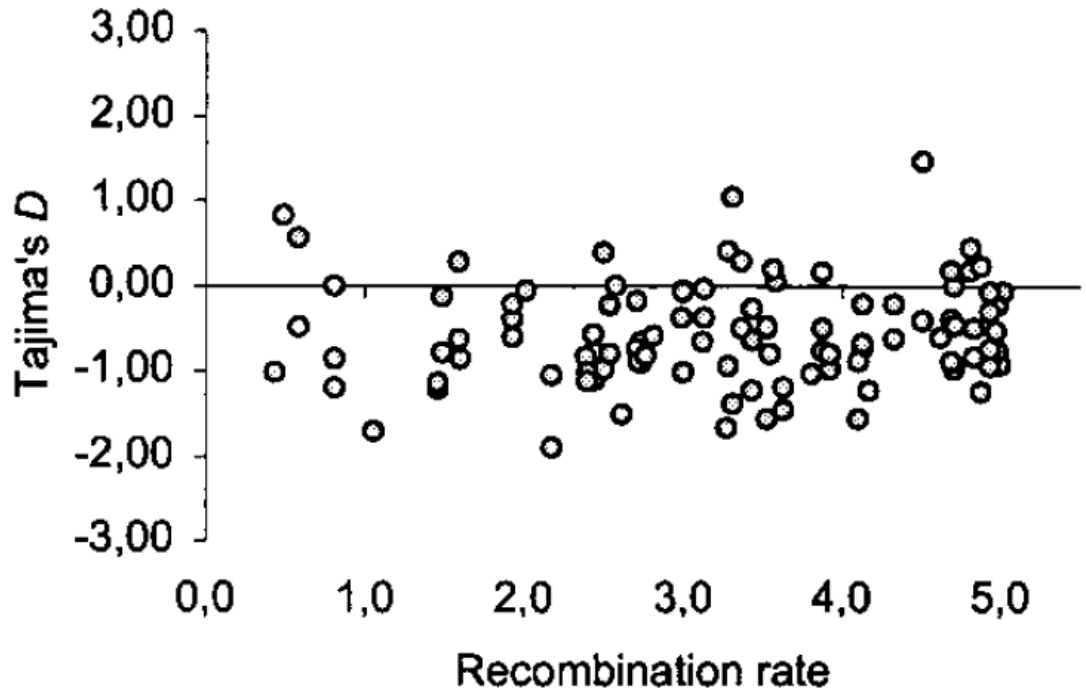
- Glinka *et al* (2003) genotyped 105 sections of the X chromosome in Drosophila from 12 inbred lines sampled in Zimbabwe. (We can think of the lines as 12 haploid individuals).
- The sections varied in size from 240–781bp over a 14Mb region.
- of the 54,944 sites sequenced, 2057 were polymorphic.
- Estimates of θ_W and θ_π across fragments are shown in the figures below (note their 'θ' is $\hat{\theta}_W$ and 'π' is $\hat{\theta}_\pi$.





Tajima's D

- They also estimated Tajima's D for each fragment, shown in the figure below
- It can be seen that many of the fragments tend to have negative Tajima's D.
- What is Tajima's D, and what is the significance of a negative value?



Tajima's D explained I

- The two estimators of θ, $\hat{\theta}_W$ and $\hat{\theta}_\pi$, use different information: the total number of mutations in the genealogy, and the average pairwise difference, respectively.
- In growing populations, more of the mutations will be in the terminal branches and so for a given $\hat{\theta}_W$ we expect $\hat{\theta}_\pi$ to be lower, because, for a mutation in a terminal branch only $(n-1)$ out of $\binom{n}{2}$ comparisons involve that mutation, whereas shared mutations are involved in many more comparison.
- In contracting populations, more of the mutations will be in the branches near the root. If the tree is balanced (for example), then $n^2/4$ comparisons involve that mutation — almost $n/4$ times as many in comparison with mutations in terminal branches.

Tajima's D explained II

- Tajima's D is equal to

$$\frac{\hat{\theta}_\pi - \hat{\theta}_W}{\text{estimated s.d. of this difference}}$$

- So a negative Tajima's D can indicate population growth, and a positive value can indicate contraction.
- Note there are many other factors such as selection, population structure, and population bottlenecks (contraction followed by expansion) that can also lead to discrepant Tajima's D in directions that are difficult to predict.

- Glinka *et al* use simulations (using *ms*) to establish that this distribution of Tajima's D is unlikely to have arisen by chance.
- Based on this and other observations (read the paper) they conclude that the African population shows a signature of recent population expansion.

A possibly more complex demography

- Haddrill *et al* studied 10 X-linked loci from 3 African populations. They concluded that the main features of the data can be explained by a bottleneck model alone.
- In a later paper, by the same group as that of Glinka *et al*, Ometto *et al* (2005) accept the Haddrill *et al* analysis, but note the present of large numbers of rare variants, and suggest this must be compatible with recent population growth.
- Read the additional notes on Blackboard for this story (and read the papers).

Next Lecture

We will look at models of population structure, using example data sets.

Further Reading I

Glinka, S., Ometto, L., Mousset, S., Stephan, W., & De Lorenzo, D. (2003).
Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach.
Genetics, 165(3), 1269–1278.

Haddrill, P. R., Thornton, K. R., Charlesworth, B., & Andolfatto, P. (2005).
Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations.
Genome research, 15(6), 790–799.

Hey, J., & Machado, C. A. (2003).
The study of structured populationsnew hope for a difficult and divided science.
Nature Reviews Genetics, 4(7), 535–543.

Further Reading II


Ometto, L., Glinka, S., De Lorenzo, D., & Stephan, W. (2005).
Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation.
Molecular Biology and Evolution, 22(10), 2119–2130.