Trees of Life (BIOL30004)
Lecture 6
Inferring demographic history from chromosomes

Mark Beaumont

Spring 2017

Summary of Lecture 6

1. How to infer past changes in population size from one individual.
2. Inferring past changes in population size using a single mitochondrial locus (for comparison).
3. Analysis of flycatcher demographic history using approximate Bayesian computation (ABC).
4. The site frequency spectrum.

A nice, easy-to-read review of all the different methods that can be used to analyse whole-genome data is in Schraiber and Akey (2015).

Population history from one individual I

- Using the SMC approximation for recombination, Li and Durbin (*Nature* 2010) developed a method for inferring past changes in population size using whole genome data from one individual.
- The maternal and paternal genomes can be thought of as randomly sampled chromosomes from the population (if the parents are unrelated). So in haploid terms this is a sample of size 2.
- So the only variable that will change along the sequence is the T_{MRCA}.
- The recombination break points can be modelled according to the SMC, and at these points the height of the tree changes.
- It is then possible to write down the likelihood (probability of getting the whole genome data, the recombination break points, and the T_{MRCA} along the sequence) in terms of the past population size.

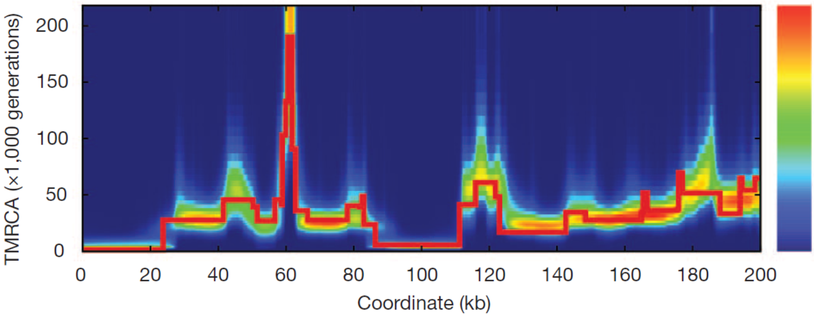
Population history from one individual II

- Li and Durbin then use a standard numerical (non-Monte Carlo) method to average over all possible break points and T_{MRCA}s to get the maximum likelihood estimate for the past population size.
- This is implemented in their PSMC package. (Pairwise sequentially Markovian coalescent.)

Population history from one individual III

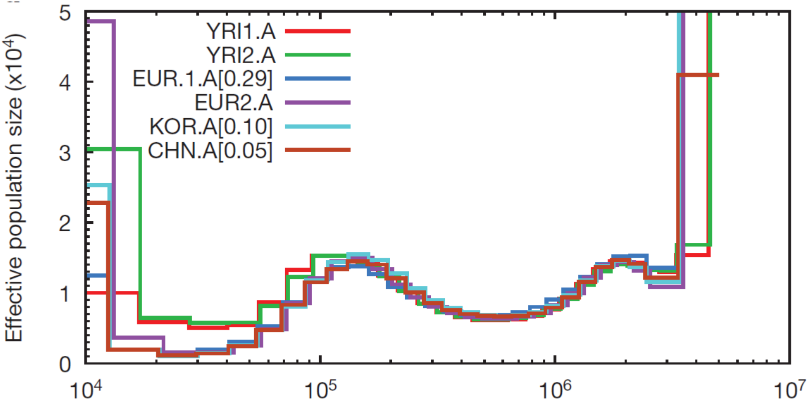
Example reconstruction of $T_{\mathbf{MRCA}}$

- They simulate a 200kb region using *ms* (*i.e.* data not using the SMC approximation).
- The figure shows a likelihood surface (represented as a heat map) for the estimated T_{MRCA}.
- The red line shows the true T_{MRCA}



Example inference of past population size change I

- This figure shows the results of applying the PSMC to data from two Yorubans (from Africa), 2 Europeans, a Korean and a Chinese.



Example inference of past population size change II

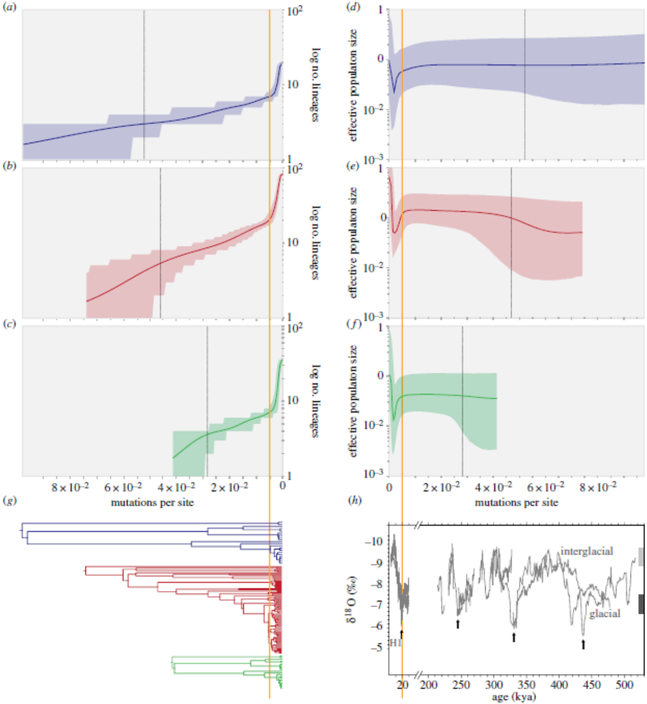
- Between the current time and 100k years ago, the African sample has a flatter trajectory than the European and Asian samples. Perhaps this reflects the out-of-Africa bottleneck (as with the *D. melanogaster* example.
- The history starts to be very similar going back beyond 100k years ago.
- Calibration will depend on assumed generation times and mutation rates (they use 25 years and $2{\cdot}5 \times 10^{-8}$ per generation, respectively).
- There is evidence of recent population growth, but this varies between the samples. (A lot of the variability is because of uncertainty in very recent times and older times, which is a consequence of having a pair of chromosomes.)
- Inference of population size change will be strongly confounded with population structure, which also affects coalescence times (Orozco-terWengel, 2016).

Genetic signatures of past response to climate change

- Groom *et al* (2014), in a study of *Homalictus* bee species in the Pacific, use sequence analysis of the mitochondrial cytochrome oxidase *c* subunit I (COI) to investigate signals of past changes in population size, and relate this to evidence of past climate change.
- Their sample sizes were large: 425 from Fiji, 177 from Vanuatu, and 143 from Samoa.
- They estimated past changes in population size using a likelihood-based method.

Bayesian Skyline Coalescent Analysis I

- The methodology is described in Drummond *et al* (2005), and implemented in a general-purpose phylogenetic and population-genetic package, BEAST.
- BEAST uses a MCMC method (as with Migrate-n, previous lecture) to obtain samples from the posterior distribution of past population trajectories, and has a number of tools for summarising the results.



(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
log no. lineages
10^2
10
1
effective populaton size
1
0
10^{-2}
10^{-3}
8×10^{-2}
6×10^{-2}
4×10^{-2}
2×10^{-2}
0
mutations per site
interglacial
glacial
H1
δ^{18}O (‰)
–10
–9
–8
–7
–6
–5
20
200
250
300
350
400
450
500
age (kya)

Summary of plots I

Distribution of the number of lineages

- The left set of figures show the posterior mean number of lineages at any time (measured in mutations per site) back in the past for bees sampled from the the three populations: Fiji, Vanuatu and Samoa.
- BEAST draws gene-trees from the posterior distribution. For each gene tree you can get the distribution of number of lineages back in time. The average is shown on the plot.
- The confidence intervals are coloured (again, obtained from the sampled trees).
- The number of lineages necessarily monotonically declines. In the standard coalescent, as we saw in Lecture 2, the decline is almost quadratic in the number of lineages at any time, so should appear as a straight line in a log-plot. Declining populations will lose lineages more rapidly, and expanding populations will lose lineages more slowly.
- Corresponding to these plots are 'maximum credibility trees' (trees with most support for the clades within them, based on comparison of the posterior sample of trees).

Summary of plots II

Bayesian skyline plots

- On the right side of the plot are shown estimates of the effective population size against time (again, in mutations per site).
- Underneath is a schematic of the change in the ratio of oxygen-18 to oxygen-16, δ^{18}O, which is a proxy for past temperatures. Note inverted scale: values associated with high temperatures are higher on the y-axis and vice versa.
- The yellow lines correspond to the last glacial maximum (LGM).
- There is evidence of population decline following the LGM, followed by a recent expansion.

Some caveats

- Correlation does not imply causation.
- The big uncertainty, if you read the paper, is in the determination of mutation rates. They give plausible reasons for the values they chose, but you need to bear in mind that the relationship between the scale on the x-axis for the genealogical plots and the time scale for the isotope plot is completely dependent on the calibration they use.
- You can see that the confidence intervals for population size estimates are very wide. This is a consequence of using 1 locus (mtDNA).

Example: Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome resequencing data

Nadachowska-Brzyska *et al*, *PLOS Genetics*, 2013

The pied flycatcher (*Ficedula hypoleuca* has an extensive range over northern Europe, whereas the collared flycatcher (*F. albicollis*) is more southerly, although they overlap in central Europe.

The group of Han Ellegren has performed whole genome resequencing from a number of individuals of both species in order to both look at evidence of natural selection and also to uncover the demographic history.

The resequencing work was first published in *Nature* (Ellegren *et al*, 2012), in which they concentrated on evidence for natural selection in the genome.

In this paper they aim to look at the speciation history.

The isolation with migration model

The basic approach taken in the flycatcher paper is to use the genomic data to fit the demographic parameters in variants of the isolation with migration (IM) model (Lecture 4).
Ideally it would be best to use MCMC as with BEAST or Migrate-n. A problem with MCMC-based genealogical methods, however, is that they do not scale well to many loci, and cannot be applied to the large amounts of sequence data that are now available.

One approach for handling larger amounts of data is to use a method known as approximate Bayesian computation (ABC). This is the approach taken in the flycatcher paper.

ABC in a nutshell I

A nice review of the approach is in Csillery *et al* (2010).

The basic method works as follows:

- Simulate parameter values from the prior distribution (this is the distribution of parameter values specified prior to the analysis).
- Simulate data sets, and compute summary statistics (such as Tajima's D *etc*).
- This gives a cloud of possible parameter values and summary statistics.
- Keep parameter values that gave rise to summary statistics that match the observed ones as closely as possible.

ABC in a nutshell II

- If the summary statistics have a certain statistical property known as sufficiency (basically, they contain all the information in the data), and if you can match the observed summary statistics exactly, then you will have a sample from the conditional distribution of the parameter values given the data (*i.e.* the Bayesian posterior distribution).
- In reality neither of these conditions typically holds, which is why the method is approximate.

The flycatcher data

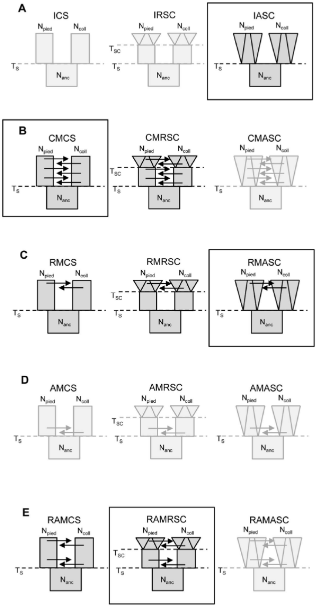
- Although the title of the paper contains the term "whole-genome", and their data comprise more than 10 million SNPs, they only use a tiny fraction (0.05% of the genome) in their analysis.
- In fact they looked at 267 loci (regions of the genome 2kb in length, separated from other loci by at least 500kb)
- They summarised the data by π, Tajima's D, F_{ST}, and the proportions of shared, fixed, and private polymorphisms. They calculated these for each locus, and then took the average and the variance. *I.e.* for these 267 loci in total they ended up with 12 summary statistics.

Simulations

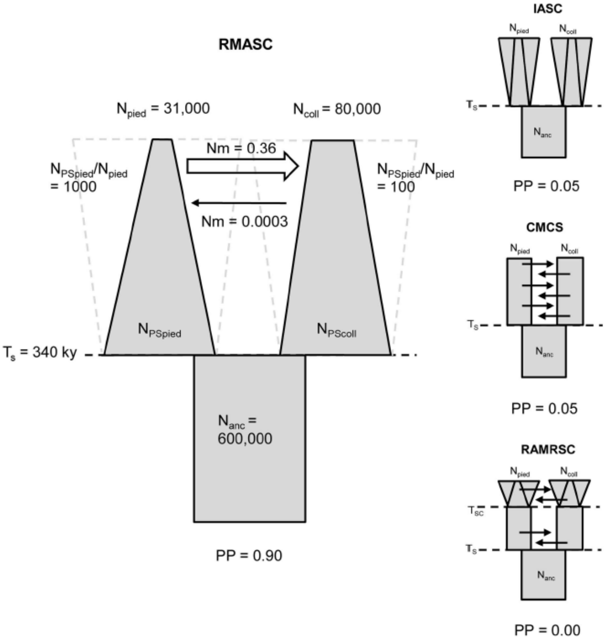
- They used *ms* (of course!) to perform the simulations.
- They considered 15 different demographic models.
- It is straightforward in ABC to estimate the posterior probability of models as well as parameter values (you just treat the model as a label and aim to find the distribution of model labels that are consistent with observed data).

A summary of the models studied

- The figure is just to give an overview of the types of model considered. They mostly differ in the amount of migration and when it happens.
- The arrows indicate migration.
- The triangles indicated changes in population size.
- The past is always at the bottom of the figures; present at the top.
- The boxed figures turned out to be the best fitting models and are compared in the next slide



The best fitting model



Main conclusions

- There has been a substantial decline in the effective size N_e for both species.
- The estimates of the current size are much smaller than census estimates (in the millions).
- Gene flow has occurred only since the last glacial maximum, and is unidirectional, from the pied flycatcher to the collared (about one individual per 3 to 6 generations).
- A relatively recent history of divergence leading to speciation (around 300k years).

Inferring Demographic History using the Site Frequency Spectrum I

- From a sample of SNPs you can count up the proportion in which the derived allele is seen once, twice, etc.
- This is called the site frequency spectrum (SFS).
- In relation to the coalescent, a mutation that occurs in a terminal lineage, results in one copy of the derived allele. A mutation that occurs in a lineage with two descendants, results in a lineage with two copies of the derived allele. . . *etc*.
- We can fit demographic models by comparing the observed SFS with the expected SFS under a demographic model (P_i — the expected proportion of SNPs with derived allele frequency count i).
- The SFS tells us about the expected shape of the coalescent genealogy.

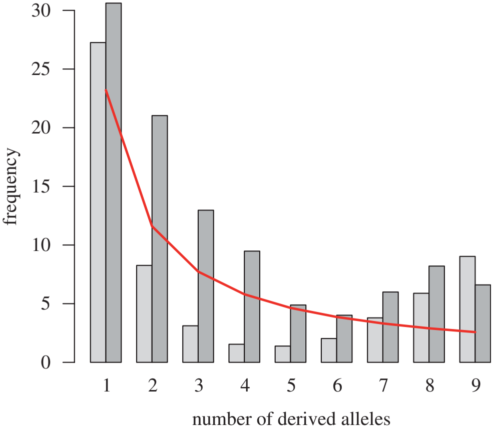
Inferring Demographic History using the Site Frequency Spectrum II

- There are a number of different methods for computing the expected SFS under a variety of demographic models:
  - For relatively simple models (*e.g.* up to two populations the most popular and fast method is implemented in the program $\partial a \partial i$, which is based on diffusion theory rather than the coalescent.
  - For more complex models we can simulate a large number of coalescent trees under each demographic scenario to estimate P_i (Excoffier *et al*, 2013).

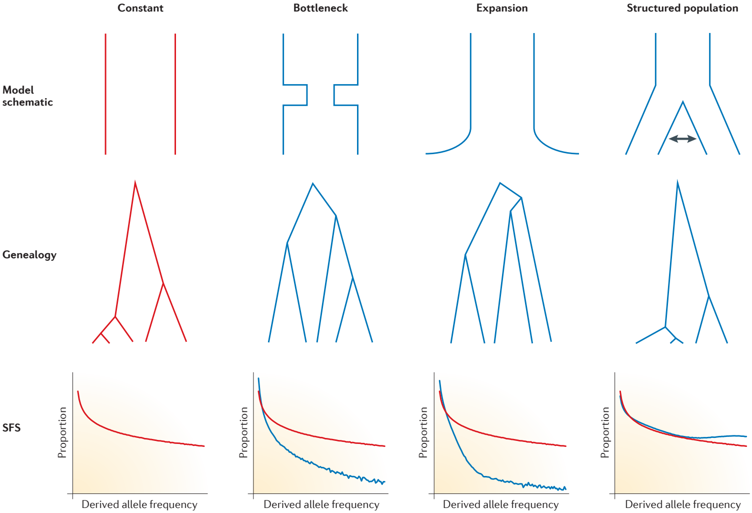
  (For more details see nice explanation in Schraiber and Akey, 2015)
- We can then compare the observed with the expected and find the parameters that maximise the likelihood of obtaining the observed SFS.

Example: SFS in Cattle

- Bos taurus (dark) and Bos indicus (light).
- 303 SNPS, 10 gene copies.
- Red line is the expected SFS under standard coalescent.
- From coalescent theory, the expected SFS is proportional to $1/i$ where i is the number of derived allele copies in the sample.



Example Schematic (from Scraiber and Akey, 2015)

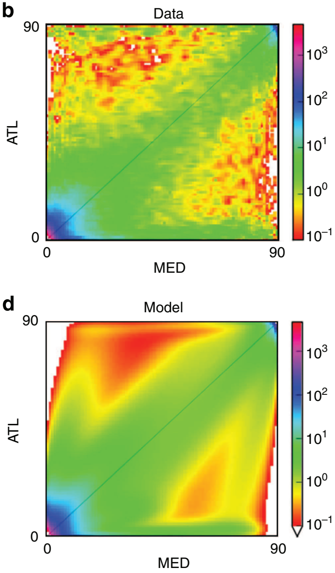


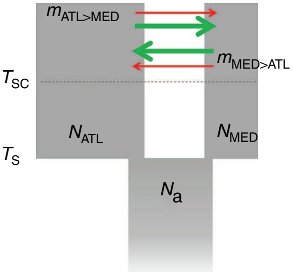
Example: Demographic history of Atlantic and Mediterranean populations of Sea Bass

- Study by Tine *et al* (*Nature Communications*, 2014).
- European sea bass forms two hybridising populations: Atlantic and Mediterranean.
- Used RAD sequencing to obtain 234,148 SNPs.
- Aim is to uncover the demographic history of the populations, but also identify regions that might be under selection.

Sea Bass Model and Joint Site Frequency Spectrum

- Right hand figures show the observed joint SFS (top) and predicted joint SFS from best-fitting model (bottom).
- The joint SFS is a 2-d version of the SFS.
- Below is the demographic model (more detail next slide).





Sea Bass: modelling and conclusions I

- The model assumes that there is an ancestral population (Atlantic) of size N_a, which splits T_s generations ago into the Atlantic and Mediterranean, of size N_{ATL} and N_{MED}. There is then secondary contact (*i.e.* onset of migration) T_{SC} generations ago, allowing for different migration rates in each direction.
- An additional feature of the model is they allow 2 classes of migration rate (a higher 'neutral' migration rate and lower 'selected' migration rate), and try to estimate the proportion in each category.
- Their best fitting model (using $\partial a \partial i$) gives an estimate of divergence time of 270,000 years ago, and date of secondary contact as 11,500 years.
- They estimate the Atlantic population to be 5 times larger than the Mediterranean population.

Sea Bass: modelling and conclusions II

- Migration from Atlantic to Med. is 5 times larger than the other way around.
- Around 35% of the genome did not freely introgress (with a migration rate around 5 times lower), giving ‘islands of differentiation’.

Comparison of methods

- **PSMC**
  - **Strengths** Makes full use of data; correctly accounts for linkage and recombination.
  - **Weaknesses** Currently can only be successfully applied to a sample of size two (*i.e.* a single diploid individual).
- **ABC**
  - **Strengths** Very flexible modelling framework; can incorporate site information and linkage/recombination information.
  - **Weaknesses** Dependent on choice of summary statistics; degree of approximation difficult to quantify; computationally challenging to simulate whole genomes.
- **SFS**
  - **Strengths** Can be very fast (if using *e.g.* $\partial a \partial i$; ideally suited to whole genome data; potentially very flexible.
  - **Weaknesses** Does not use linkage/recombination information; cannot distinguish between some demographic histories; difficult to quantify uncertainty.

Next Lecture

We will look at the gene-tree versus species tree problem.

Further Reading I

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & Franois, O. (2010).
Approximate Bayesian computation (ABC) in practice.
Trends in ecology & evolution, 25(7), 410–418.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrm, N., Kawakami, T., . . . & Uebbing, S. (2012).
The genomic landscape of species divergence in Ficedula flycatchers.
Nature, 491(7426), 756–760.

Excoffier, L., Dupanloup, I., Huerta-Snchez, E., Sousa, V. C., & Foll, M. (2013).
Robust demographic inference from genomic and SNP data.
PLoS Genet, 9(10), e1003905.

Further Reading II

Groom, S. V., Stevens, M. I., & Schwarz, M. P. (2014).
Parallel responses of bees to Pleistocene climate change in three isolated archipelagos of the southwestern Pacific.
Proceedings of the Royal Society of London B: Biological Sciences, 281(1785), 20133293.

Li, H., & Durbin, R. (2011).
Inference of human population history from individual whole-genome sequences.
Nature, 475(7357), 493–496.

McVean, G. A., & Cardin, N. J. (2005).
Approximating the coalescent with recombination.
Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360(1459), 1387–1393.

Further Reading III

Murray, C., Huerta-Sanchez, E., Casey, F., & Bradley, D. G. (2010).
Cattle demographic history modelled from autosomal sequence variation.
Phil. Trans. Roy. Soc. London. B., 365(1552), 2531–2539.

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013).
Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data.
PLoS Genet, 9(11), e1003942.

Orozco-terWengel, P. (2016)
The devil is in the details: the effect of population structure on demographic inference.
Heredity 116, 349–350

Further Reading IV

Schraiber, J. G. & Akey, J. M. (2015).
Methods and models for unravelling human evolutionary history.
Nat Rev Genet 16, 727–740.

Tine, M., Kuhl, H., Gagnaire, P. A., Louro, B., Desmarais, E., Martins, R. S., . . . & Dieterich, R. (2014).
European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation.
Nature communications, 5.