

Trees of Life (BIOL30004)

Lecture 4

Genealogies and demographic history: population structure

Mark Beaumont

Spring 2017

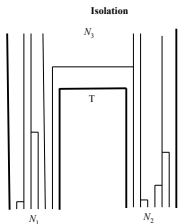
Summary of Lecture 4

1. Models of population structure
2. The coalescent with migration
3. F_{ST} and coalescent theory
4. Example: humpback whales

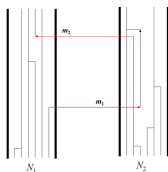
Population Structure

There are three basic patterns of population structure that are often considered.

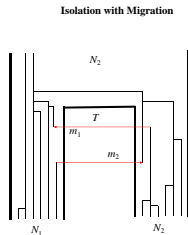
Isolation



Equilibrium Migration



Isolation with migration



The coalescent with migration

- ▶ I'll first describe how you could simulate a genealogy with migration to give an understanding of what the genealogy might look like.
- ▶ Discuss some example genealogies.
- ▶ Discuss the expected T_{MRCA} for a pair of samples under migration, which is rather counterintuitive.
- ▶ Introduce a quantity (discussed in the papers we will look at) called F_{ST} which is frequently used to describe population structure.

Migration I

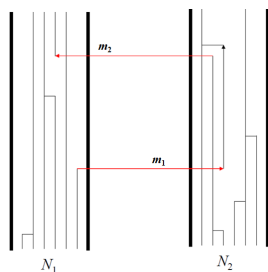
Take the simplest case of 2 demes exchanging genes (let's work with unscaled time). A 'deme' is a population geneticist's way of saying 'population'

Let's call the migration rate m . You can think of it as the probability that a randomly chosen gene copy in the deme originated from another deme in the previous generation. In each generation a lineage in deme 1 has a probability m_1 of being an immigrant from deme 2. For deme 2 it is m_2

Assume deme 1 has population size N_1 , and deme 2 has population size N_2

Assume that are n_1 gene copies sampled from deme 1; n_2 gene copies sampled from deme 2.

Migration II



There are then 4 possible events:

- ▶ Coalescence in population 1 with rate $\binom{n_1}{2}/(2N_1)$.
- ▶ Coalescence in population 2 with rate $\binom{n_2}{2}/(2N_2)$.
- ▶ Migration (backwards in time) from population 1 to population 2 with rate $n_1 m_1$
- ▶ Migration (backwards in time) from population 2 to population 1 with rate $n_2 m_2$

Migration III

The sum of these rates gives the total rate of an event (irrespective of what it is). We simulate events using the total rate, and then decide what the event is with probability equal to its rate as a proportion of the total rate. So:

The waiting time to an event is exponentially distributed with rate

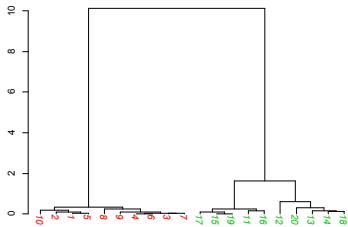
$$R = \binom{n_1}{2}/(2N_1) + \binom{n_2}{2}/(2N_2) + n_1 m_1 + n_2 m_2.$$

And (for example) the probability that it is a migration from population i is then $n_i m_i / R$.

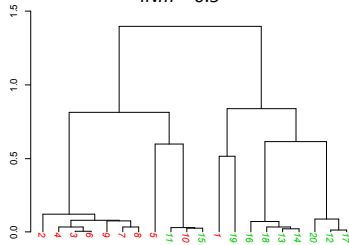
In this way we can build up a genealogy.

Some example genealogies

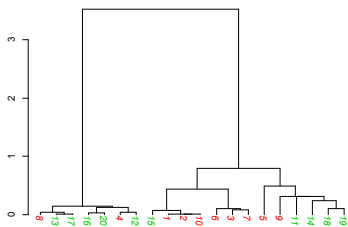
$4Nm = 0.1$



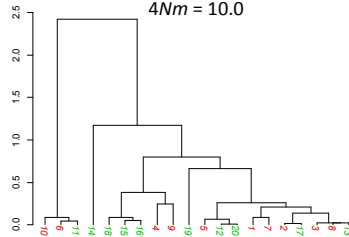
$4Nm = 0.5$



$4Nm = 1.0$



$4Nm = 10.0$



Some interesting results with migration

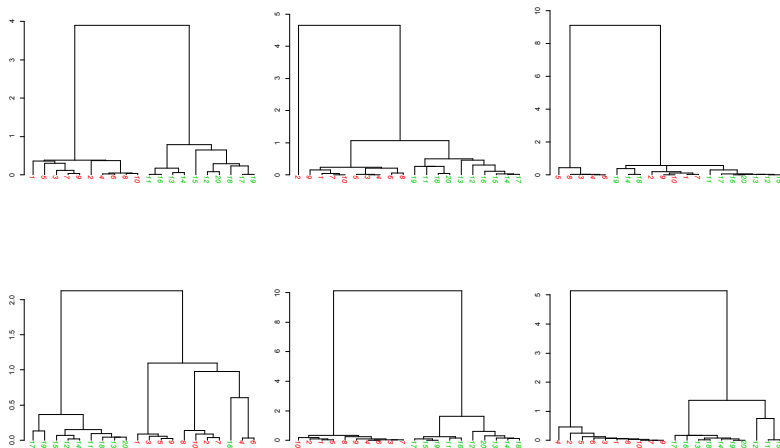
If we assume there are d demes each of size N , with equal migration rate, m , then:

- ▶ The expected time back to the most recent common ancestor for 2 lineages from the **same** deme, t_w , is $2Nd$
- ▶ The expected time for 2 lineages from **different** demes, t_b , is $2Nd(1 + \frac{d-1}{4Nmd})$

(for a derivation see the Hein *et al* book.)

I.e. irrespective of the migration rate the expected time back to a common ancestor remains the same for 2 genes sampled from the same deme. The intuition is that either you coalesce in the deme or you migrate out. If m is low then most of the time you coalesce, on a timescale of $2N$. However if you migrate out, then you can take a very long time to coalesce. When you do the maths these effects balance, so that the average stays the same at $2Nd$.

Example plots with $4Nm=0.1$



F_{ST} I

The quantity F_{ST} is used very commonly in population genetic analysis.

Unfortunately, although it is very widely used, there is little consensus about what it actually is, or how it can be defined (Rousset, 2012).

Many people think of it as a statistic for describing how genetically differentiated two samples are: it varies from 0 (identical gene frequencies) to 1 (maximally different gene frequencies). However you can also think of it as a parameter in a model, but there is not a lot agreement of what this parameter is.

The most widely used estimator of F_{ST} is that of Weir and Cockerham (I won't give it here, but the Rousset article gives useful pointers).

F_{ST} II

For low-mutation-rate markers (like SNPs...) it is an unbiased estimator of this quantity:

$$\frac{t_w - t_b}{t_b}$$

We saw t_w and t_b in the earlier slide: they are the expected T_{MRCA} for pairs of genes taken within and between demes.

It is very useful to then define the parameter F_{ST} as:

$$F_{ST} = \frac{t_w - t_b}{t_b} = \frac{1}{1 + \frac{d}{d-1} 4Nm}$$

When d gets very big (the 'infinite island model') this reduces to a classic formula:

$$F_{ST} = \frac{1}{1 + 4Nm}$$

F_{ST} III

To add yet another complication (or very interesting connection), if F_{ST} is defined by this equation, then a rather different way to think about it is that it is the probability that two gene copies in a deme coalesce backward in time before either of them migrates.

We can easily see this using the relative rates argument I used earlier for simulating a genealogy with migration.

The total rate of migration or coalescence within a particular deme is

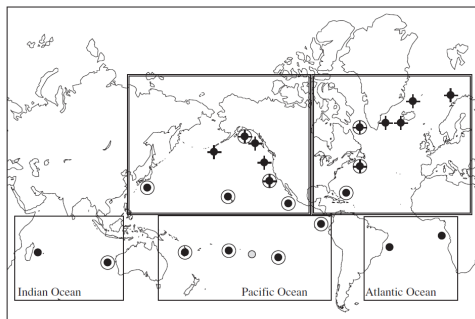
$$2m + \frac{1}{2N}.$$

So the chance that the first event is a coalescence is

$$\frac{\frac{1}{2N}}{2m + \frac{1}{2N}} = \frac{1}{1 + 4Nm} = F_{ST}$$

Example: Global diversity and oceanic divergence of humpback whales (*Megaptera novaeangliae*)

Jackson *et al* (2014) report a study in which they sequenced mtDNA control region (the most variable part) from almost 3000 whales taken from the North Pacific, North Atlantic, and Southern Hemisphere. They also sequenced 8 nuclear loci from 70 individuals worldwide.



Humpback whale: aims

Previous studies had indicated that there was evidence of long-term gene flow between different oceans. The aim of the study was to obtain a better resolution on estimates of gene flow by using nuclear genes as well as mitochondrial genes.

Although they carried out a number of analyses, I concentrate here on the F_{ST} -based analyses and on model-based estimates of gene flow between the different oceans.

F_{ST} analysis I

The table below shows F_{ST} (below diagonal) and ϕ_{ST} (above diagonal) estimates from the mitochondrial and nuclear data. The diagonal contains estimates of π per base position averaged over the sequences.

	Southern Hemisphere		North Pacific		North Atlantic	
	mtDNA	nuDNA	mtDNA	nuDNA	mtDNA	nuDNA
SH	0.0248	0.0016	0.3193	0.0451	0.1613	0.0990
NP	0.0858	0.0400	0.0113	0.0009	0.5164	0.1516
NA	0.0926	0.0610	0.1755	0.1030	0.0197	0.0016

There are a number of features of interest in these data:

- ▶ The ϕ_{ST} value tends to be much higher than F_{ST} for the mtDNA data.
- ▶ What is ϕ_{ST} ?
- ▶ Before answering this, a point to bear in mind is that F_{ST} is being referred to as a statistic here — specifically, the value of the Weir & Cockerham estimator.

F_{ST} analysis II

- ▶ This is only an unbiased estimator of F_{ST} as I defined it earlier when mutation rates are low.
- ▶ Why?
- ▶ Because (effectively. . .) it assumes that non-identical sequences within a population must have come from another population. But if the mutation rate is high this will not be true.
- ▶ When mutation rates are high the W&C estimator is too low.
- ▶ Back to ϕ_{ST} . . . This is a direct estimator of F_{ST} as defined by:

$$F_{ST} = \frac{t_w - t_b}{t_b}.$$

- ▶ It estimates t_w and t_b directly by substituting the average pairwise differences π_w and π_b (note the unknown mutation rates will cancel out).

F_{ST} analysis III

- ▶ You can see that there is not such a large discrepancy for the nuclear sequences because their mutation rate is lower.
- ▶ Concentrating on the ϕ_{ST} s, the other main pattern is that it is much bigger for the mitochondrial genes than the nuclear genes. Why is this?
- ▶ If we assume that the migration rate for males and females is the same (not necessarily a good assumption), and the number of males and females is the same, then the expected F_{ST} depends on $2N_f m$ for mitochondrial genes and $4(N_f + N_m)m$ for nuclear genes, a number 4 times larger. *I.e.* for large Nm we expect F_{ST} to be 4 times higher for mtDNA.

F_{ST} analysis IV

- ▶ What does this table tell us about the biology?
- ▶ There would appear to be more gene flow between the southern oceans and North Atlantic (nuclear ϕ_{ST} estimate of 0.05). The most restricted gene flow is between the Northern Pacific and the North Atlantic (nuclear $\phi_{ST} = 0.15$).
- ▶ (For comparison F_{ST} for the three most differentiated human groups, Africans, Europeans, Asians, is around 0.13.)

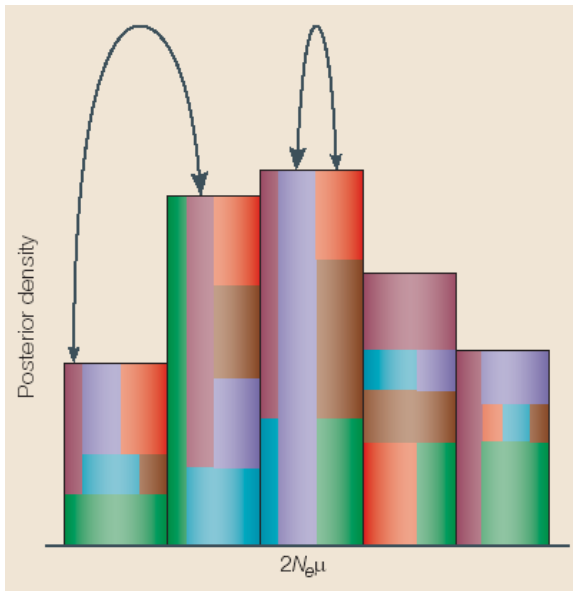
	Southern Hemisphere		North Pacific		North Atlantic	
	mtDNA	nuDNA	mtDNA	nuDNA	mtDNA	nuDNA
SH	0.0248	0.0016	0.3193	0.0451	0.1613	0.0990
NP	0.0858	0.0400	0.0113	0.0009	0.5164	0.1516
NA	0.0926	0.0610	0.1755	0.1030	0.0197	0.0016

Estimation of migration rates

- ▶ Although the F_{ST} estimates could then be used to estimate scaled migration, the authors take a likelihood-based approach, which should use the data more effectively.
- ▶ They use a program called `MIGRATE-N`, which uses Markov chain Monte Carlo (MCMC) to estimate the parameter values from the data.

Markov Chain Monte Carlo (MCMC) and Genealogies

- ▶ From coalescent theory we can calculate the probability, $p(G, D|\phi)$ of obtaining any particular tree G and data set D (see Beaumont and Rannala, 2003 and Marjoram and Tavaré, 2006 for accessible reviews of these methods), conditional (the '|' in the expression above) on a set of demographic and mutational parameters in ϕ .
- ▶ In a Bayesian calculation we want to invert the relation above, and compute $p(\phi, G|D)$, which is called the posterior distribution. If we could compute this (essentially impossible), it would give the probability of obtaining any set of parameter values and genealogy, given the data.
- ▶ Markov chain Monte Carlo (MCMC) allows you to sample values from $p(\phi, G|D)$ knowing only $p(G, D|\phi)$.
- ▶ These are autocorrelated values, but if you have enough of them you can build up an accurate picture of the posterior distribution.



(From Beaumont and Rannala, 2004)

MCMC and Genealogies

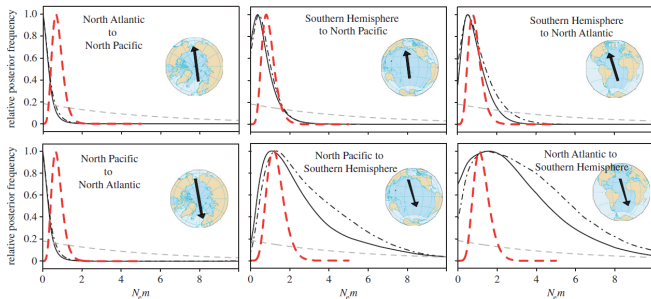
- ▶ In this picture the coloured blocks represent the probability of obtaining the data and a particular genealogy.
- ▶ The histogram bars represent the sum of this probability over all the genealogies, for a particular value of θ .
- ▶ With the MCMC algorithm you **jump** either from genealogy to genealogy keeping θ fixed, or from value of θ to value of θ keeping the genealogy fixed (or both, not shown here).
- ▶ The length of time you remain without jumping is **in proportion** to the size of the coloured blocks $p(G, D|\theta)$.
- ▶ If you ignore the the genealogies that are generated, you end up sampling from $P(\theta|D)$, marginal (averaged over) all the genealogies (*i.e.* giving the overall histogram in this figure).

MCMC with Migration

- ▶ In this case the model follows exactly the same coalescent model that we have described in this lecture.
- ▶ Rather than simulating coalescent genealogies, MIGRATE-N uses coalescent theory to compute the probability of jointly getting a genealogy and the data and uses MCMC to generate samples from the posterior distribution of migration rates and genealogies conditional on the data.
- ▶ We treat the genealogies as a nuisance parameter and throw them away, and look at the distribution of migration rates.

Estimated migration rates between oceans I

- ▶ This shows posterior distributions for the migration rates
- ▶ N_e refers to the population size of males and females together.
- ▶ The grey dashed line is the prior (the distribution of the parameter irrespective of the data).
- ▶ The two black curves are for two sets of mtDNA sequences.
- ▶ The red curve is for nuclear DNA.






Estimated migration rates between oceans II

- ▶ Generally the nuclear DNA posterior is much tighter.
- ▶ We can see that typically the value of Nm can be between 0.5 and 2 for the nuclear markers, and it is pretty similar in all directions.
- ▶ These are quite high values, and consistent with the ϕ_{ST} values in the previous table.
- ▶ You can see that an advantage of this approach is that it also provides some measure of accuracy of the estimates, which would not be straightforward from the ϕ_{ST} table.
- ▶ Look back at the coalescent genealogies for different values of Nm . Do you think these are particularly differentiated populations?

Next Lecture

We will look at the genealogy of recombining sequences, with a view to modelling whole-genome data.

Further Reading I

-  Beaumont, M. A., & Rannala, B. (2004).
The Bayesian revolution in genetics.
Nature Reviews Genetics, 5(4), 251–261.
-  Jackson, J. A., Steel, D. J., Beerli, P., Congdon, B. C.,
Olavarra, C., Leslie, M. S., ... & Baker, C. S. (2014).
Global diversity and oceanic divergence of humpback whales
(*Megaptera novaeangliae*).
*Proceedings of the Royal Society of London B: Biological
Sciences*, 281(1786), 20133222.
-  Marjoram, P., & Tavaré, S. (2006).
Modern computational approaches for analysing molecular
genetic variation data.
Nature Reviews Genetics, 7(10), 759-770.

Further Reading II



Rousset, F. (2013).

Exegeses on maximum genetic differentiation.

Genetics, 194(3), 557–559.