Trees of Life (BIOL30004)
Lecture 5
Genealogies with Recombination

Mark Beaumont

Spring 2017

Summary of Lecture 5

- Main effects of recombination
- Haplotype structure
- Coalescent with recombination
- A useful approximation

The genealogy of the autosome — the effects of recombination

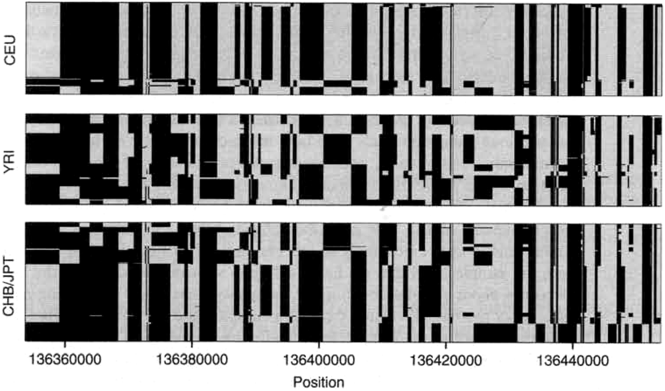
- The genealogical principles that we have described so far are also applicable to the nuclear genes, at least for any nucleotide position considered individually.
- The complicating factor is that recombination changes the genealogy for neighbouring regions of a chromosome.
- Generally, for regions of the chromosome that are more than about 100kb apart, we can think of the genealogies as independent — *i.e.* they can be modelled as independent realisations of the coalescent process.

The effects of a recombination event depend on whereabouts in the genealogy it happens

- A recent recombination event in the history of a sample may only affect a single lineage, and not be very noticeable when we look at sequence data.
- Recombination events that happen early in the genealogy may have large effects that are clearly noticeable in the data. For example if the genealogy becomes much longer there will be more derived mutations. Also there will be discontinuities in the sample chromosome copies that share a particular haplotype.
- This is particularly noticeable in regions of the genome that have been subject to selection, as shown in the next few examples. Here, the selected allele has caused a particular haplotype to increase in frequency, which then decays due to recombination.

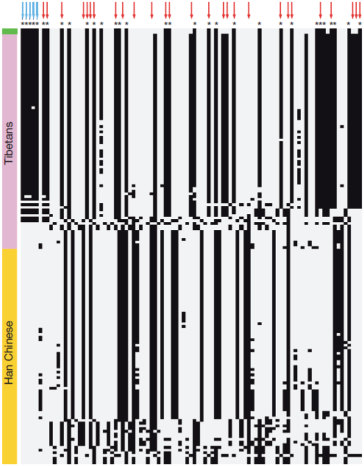
Haplotype Structure of the *Lactase* gene

(From McVean, 2007). Data shows a 100 Kb region around the *Lactase* gene, sampled in 3 different human populations. From the International Hapmap Consortium. The reference sequence allele is marked in grey.



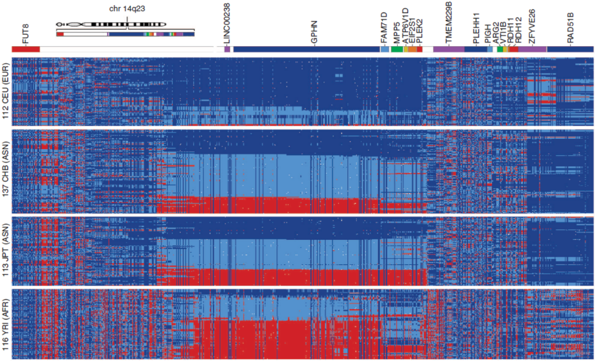
Haplotype Structure of EPAS1

This figure shows the haplotype structure of 95 SNPs in the EPAS1 gene in a sample of Tibetans and Han Chinese (Huerta-Sánchez *et al*, *Nature*, 2014). Variants of EPAS1 are implicated in adaptation to high altitude. The authors conclude from their study that this pattern is likely to be due to introgression from Denisovans. Ancestral alleles are white, derived are black.



Haplotype structure at the *Gephyrin* locus

Gephyrin is a multifunctional protein that binds other proteins together, and is believed to have a protective homeostatic role in maintaining physiological function. The local haplotype structure is described in a paper by Climer *et al* (2015, *Nature Communications*). Unlike the previous two pictures, for this example, we are looking at the diploid genotype of each individual. Homozygotes of the allele that is rarest in European populations are coloured red, heterozygotes are light blue, homozygotes of the alternate allele are dark blue.



Haplotype structure: summary

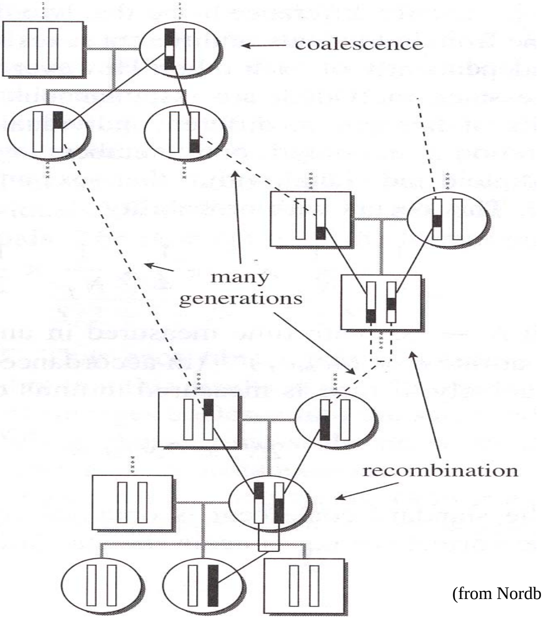
- We can see that there is a strong block structure.
- Part of this can be explained by presence of recombination hot-spots which itself leads to a block structure (Jeffreys *et al*, *Nature Genetics*, 2001).
- In addition recombination events occuring early in the genealogy of the genome can lead to large discontinuities in the clades that inherit a particular haplotype.

Genealogies with recombination I

Consider the ancestry of a single chromosome in an individual.

- This chromosome is a copy of what was transmitted by one parent.
- The copy that was transmitted (in sperm or egg) will have been produced by meiosis.
- In meiosis there will have been at least one crossing-over event in which *e.g.* the 'left' arm came from one parent (of this parent) and the 'right' one came from the other.
- So looking backward the genealogy of this chromosome has split into two.
- But if we consider the genealogy of each of these two ancestral chromosomes only a part of them has genetic material that is directly ancestral to the individual we are looking at.
- So when we model the genealogy we only keep a track of the ancestral bits (marked black in the next slide).

Genealogies with recombination II



(from Nordborg)

The coalescent with recombination 1

- Originally proposed by Hudson (1983).
- Assume that we represent a length of DNA sequence by a continuous line.
- Assume the probability of a recombination anywhere within the sequence in any generation is c.
- Assume recombination can occur uniformly within the sequence.

The coalescent with recombination 2

Time is scaled in units of $2N$ generations.

Define $\rho = 4Nc$, so recombinations in scaled time occur at a rate $\rho/2$ We start with a sample of k sequence copies. We can define 2 types of event:

Coalescence occuring at a rate $\frac{k(k-1)}{2}$

Recombinations occuring at a rate $\frac{k\rho}{2}$

The total rate is then $\frac{k(k-1)}{2} + \frac{k\rho}{2}$

The coalescent with recombination 3

It is useful to distinguish between genetic material that is ancestral to sample, and material that is not.

To simulate the coalescent with recombination:

1. Simulate time to the first event, using the total rate.
2. With probability $\frac{\rho}{k-1+\rho}$ choose a recombination; take a sequence, choose a position uniformly at random, and split into two sequences, dividing up ancestral material.
3. With probability $\frac{k-1}{k-1+\rho}$ choose a coalescence; merge two sequences, taking the union of ancestral material.
4. Repeat until there is one sequence (the grand most recent common ancestor, GMRCA) is reached.
5. The genealogy produced by this algorithm is known as the Ancestral Recombination Graph.

The coalescent with recombination 4

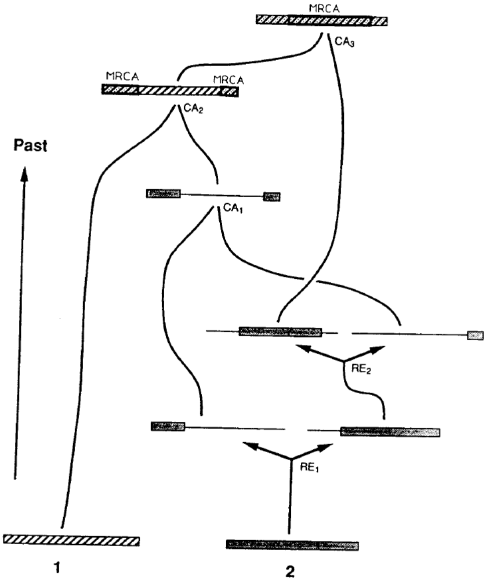
Some complications:

- The last MRCA of some part of the sequence may have been reached long before reaching the GMRCA
- Recombinations may produce lineages that contain no ancestral material at all.

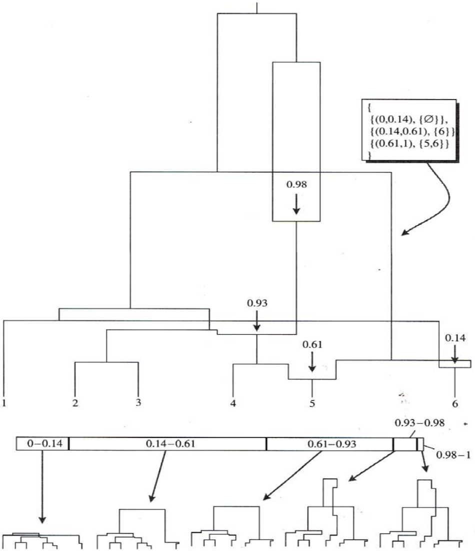
In most practical implementations the algorithm keeps a track of ancestral material and stops when the last MRCA is reached, and restricts recombinations to those that do not produce sequences with no ancestral material.

The ancestral recombination graph I

(Hudson, 1990)



The ancestral recombination graph II



(from Nordborg)

The ancestral recombination graph III

- This example shows the genealogy of a sample of 6 sequences.
- Recombination splits the sequence (scaled between 0 and 1) into 5 distinguishable genealogies.
- The lower part shows the 5 local genealogies; above it are the recombination break points; above these is the ancestral recombination graph, with recombination events shown by arrows.
- The box on the right lists as an example the descendants of that lineage: *i.e.* 0–0.14 contains no ancestral material, 0.14–0.61 is ancestral to 6, and 0.61–1 is ancestral to 5 and 6.

Implications for Genome Data

- The expected number of recombination events in the ancestry of a sample is

$$\rho \sum_{i-1}^{n-1} \frac{1}{i},$$

  so it is exactly the same as the expression for mutation, but using ρ rather than θ.

- Although on a fine scale the rate of recombination varies a lot across the genome (Jeffreys *et al*, 2001), averaged over *e.g.* a megabase it is about 1 centiMorgan/Mb/generation: 10^{-8} per base per generation. Which is very close to our current best estimates of the autosomal mutation rate of around $1{\cdot}6 \times 10^{-8}$ per base per generation.

- So recombination is quite a major force in determining genealogical structure.

- Unfortunately it is very difficult to work with the Ancestral Recombination Graph directly for statistical analysis.

- Instead we use a variety of approximate approaches.

How can we model autosomal genomic data?

In the remaining part of the course, I will be discussing how we can use coalescent theory to detect past changes in population size, look at past migration rates, and estimate times that populations have diverged from each other. So if recombination makes things difficult, how can we do it?

- Don't do it — stick with Y and mtDNA. The problem here is that we then just have two genealogies (for males and females), and we know that single genealogies are very variable, so may not tell us very much about the past history.
- Take bits of genome that are widely spaced apart, and short enough that we can ignore recombination.
- Look at each SNP independently and ignore linkage.
- Use tractable approximations of the ARG.
- Use simulation-based methods to simulate from the ARG and compare real data with the results of simulations.

Modelling recombination along the sequence

- Ideally we want to construct the genealogy sequentially along the sequence.
- If we could this then the problem becomes Markovian: the probability of the data at a particular site just depends on the local genealogy, which just depends on the genealogy of the immediately preceding site.
- In fact Wiuf and Hein (1999) developed a way of simulating the ARG exactly using an algorithm that was sequential.
- However, their algorithm is impractical because at every point along a sequence you potentially have to look back along the entire length of sequence that you have simulated (*i.e.* it is not Markovian).

An approximation for recombination

- McVean and Cardin (2005), in an influential paper, noted that if you had a rule that two sequences were not allowed to coalesce if there was non ancestral trapped material then you could simplify the algorithm of Wiuf and Hein so that it could be applied truly sequentially.
- They called this the Sequentially Markov Coalescent (SMC) approximation.
- An even better approximation (called SMC$'$) was developed by Marjoram and Wall (2006), but I won't discuss that here.

Constructing the genealogy with the SMC I

This is taken from the paper by McVean and Cardin (2005). We are working in scaled time and we are also scaling the sequence to be of length 1.

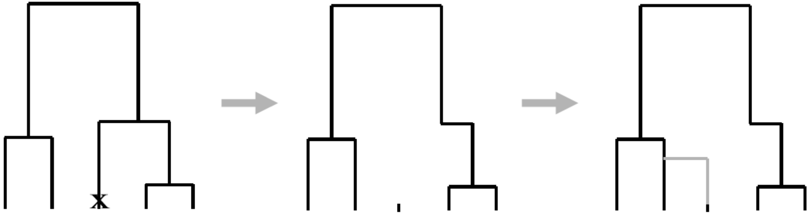
- Simulate a standard coalescent genealogy.
- This is the genealogy of the left side of the sequence. It has a total branch length of T_0.
- Recombination can occur anywhere in this genealogy at a rate $\rho T_0/2$ times the length of the sequence, which we have set at 1. Remember $\rho = 4Nr$ for recombination rate r and time is scaled in units of 2N generations. So the position in the sequence of the first recombination can be obtained by simulating an exponentially distributed random number with this rate.
- At the recombination point we choose a point in the genealogy where the recombination occurs.
- The lineage above this point is then erased.

Constructing the genealogy with the SMC II

- The lineage then is free to coalesce with any of the remaining lineages at a rate proportional the number of lineages in the genealogy at any time (this will be decreasing up the tree). It may coalescent with the MRCA at a point higher than the current T_{MRCA}.

SMC Illustration

This figure from Mcvean and Cardin (2005) illustrates what is happening. The cross marks the recombination point.

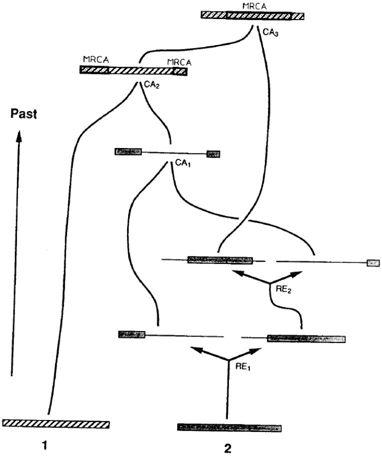


Consequences of the SMC approximation

- The SMC genealogy is remarkably similar to to full ARG.
- The main difference is that under the SMC algorithm you cannot have ‘trapped non-ancestral material’.
- I.e. once a recombination event has changed the coalescent time at some point along the sequence for a pair of sequences, the same coalescent time cannot then be recovered later on in the sequence.
- This means that LD in the SMC approximation tends to decay faster than in the ARG.
- However with the SMC approximation it is feasible to carry out likelihood-based analysis (at least for simple cases).

Trapped non-ancestral material

Hudson's example figure shows coalescence with trapped non-ancestral material. The consequence is that these two non-adjacent bits of chromosome have exactly the same T_{MRCA} in the descendants. (Hudson, 1990)



The next lecture

Examples of whole genome analysis in population genetics.

Further Reading I

McVean, G. (2008).
Linkage disequilibrium, recombination and selection.
Handbook of Statistical Genetics, Third Edition, 909–944.

Climer, S., Templeton, A. R., & Zhang, W. (2015).
Human gephyrin is encompassed within giant functional noncoding yinyang sequences.
Nature communications, 6, 6534

Hudson, R. R. (1990).
Gene genealogies and the coalescent process.
Oxford surveys in evolutionary biology, 7, 1–44.

Hudson, R.R. (1983)
Testing the constant-rate neutral allele model with protein sequence data.
Evolution, 37, 203-217

Further Reading II

Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., . . . & Wang, B. (2014).
Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.
Nature, 512(7513), 194–197.

Jeffreys, A. J., Kauppi, L. Neumann, R. (2001)
Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.
Nature Genetics, 29, 217–222.

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

Marjoram, P., & Wall, J. D. (2006).
Fast coalescent simulation.
BMC genetics, 7(1), 16.