Trees of Life (BIOL30004)
Lecture 1
An Introduction to Gene Genealogies

Mark Beaumont

Spring 2017

Gene genealogies: outline

- The aim of this part of the course is to provide an introduction to DNA sequence analysis at the **population** level.
- As with phylogenetics our understanding best comes from taking a genealogical perspective.
- Unlike phylogenetics we have a good model for what the shape of the genealogy should look like: coalescent theory.
- Unlike phylogenetics we are typically not interested in the genealogy itself, but with population-level phenomena: past changes in population size, migration rates, historical admixture of populations.
- The general teaching approach will be to focus on experimental or observational results from published papers, and use these as a vehicle for introducing theoretical ideas that can explain them.

Lecture outline

Lecture 1 Population-level gene-trees reconstructed from mitochondrial and Y-chromosome DNA.

Lecture 2 Coalescent theory I; basic nuts and bolts

Lecture 3 Coalescent theory II; application to Y-chromosome data; coalescence with varying population size.

Lecture 4 Coalescence in structured populations.

Lecture 5 Genealogy of recombining sequences and statistical inference with whole-genome data.

Lecture 6 Further methods of inference with whole-genome data.

Lecture 7 Reconciling gene-trees and population-trees: the bridge to phylogenetics.

Summary of Lecture 1

- Present the original paper that introduced ‘mitochondrial Eve’ (although they didn’t use that term).
- Give simple examples of the parsimony method for phylogeny construction, homoplasy, and rooted and unrooted trees.
- Describe recent high resolution gene trees for mitochondrial DNA and Y-chromosome DNA.
- Point out some apparent features of the gene trees, which leads into coalescent theory, the subject of the next lecture.

Mitochondrial Eve

Cann, Stoneking, and Wilson (1987) Mitochondrial DNA and human evolution, *Nature*, 325, 31–36.

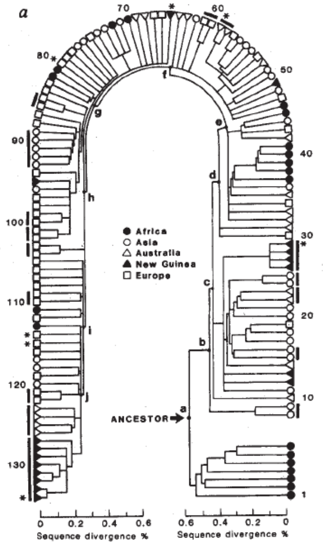
Extracted mtDNA from 147 people, representing 5 geographic regions.

Used restriction enzymes to genotype the mtDNA.

Restriction sites are groups of adjacent nucleotides that are recognised by a cutting enzyme.

If the site is cut you get a band on a gel (scored as '1'), otherwise it is not cut and you do not get a band (scored as '0'). Mutations can delete or create sites .

They used the pattern of 0s and 1s to construct a tree.



How did they construct this tree?

A short digression on tree-making . . .

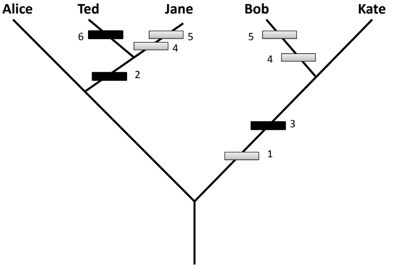
Method of Parsimony.

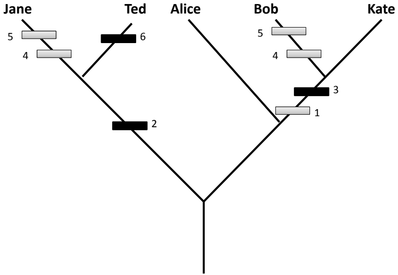
- The aim is to construct a branching tree that can represent the differences in the data: similar data occupy neighbouring parts of the tree.
- Note that such a tree need not imply any historical process that gives rise to the data. We could make a tree from mug shapes or tea-cosy designs. However, I am always going to think of a tree in an evolutionary context.

Method of Parsimony

Example Data
(adapted from Felsenstein, 2004.)

	Restriction sites					
People	1	2	3	4	5	6
Alice	1	0	0	1	1	0
Bob	0	0	1	0	0	0
Jane	1	1	0	0	0	0
Ted	1	1	0	1	1	1
Kate	0	0	1	1	1	0



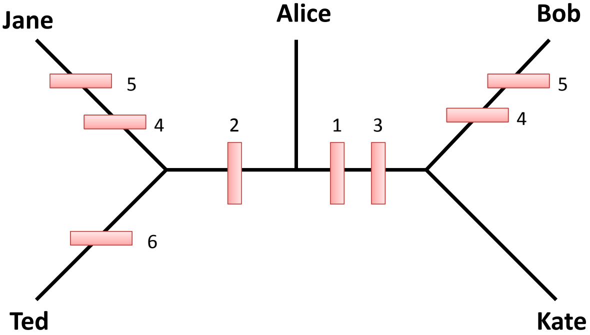


Method of Parsimony

- The aim is to construct a tree which connects each of the sequences together with the fewest possible changes.
- In these pictures black bars denote changes that give the ‘1’ state and grey bars are changes to the '0' state.
- These two examples are maximum parsimony trees for the same data.
- Note that sites 4 and 5 are required to change twice. This is an example of homoplasy, where some states arise more than once in the tree.
- Note also that there are two roots shown in this example: one that separates Alice, Ted and Jane from Bob and Kate; and one that separates Ted and Jane from from Alice, Bob and Kate.
- In fact we can put the root anywhere, so many algorithms give unrooted trees.
- In these two examples the root in both cases is 100110 for sites 1,. . . ,6 but more generally the type of the root may change depending on where we put it.

Unrooted Tree Example

- The unrooted tree below corresponds to the two trees given earlier.
- Imagine this tree is made of string; we can pick it up at any point (the root) and drag it up the slide to form a rooted tree.

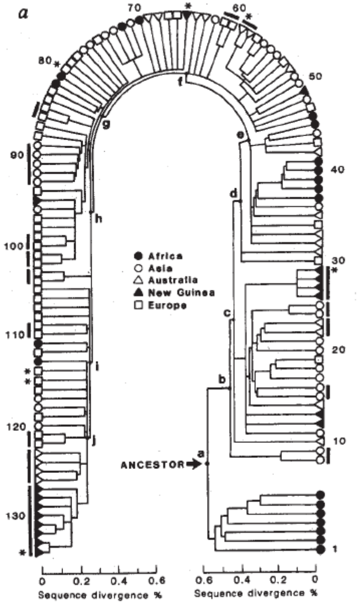


Rooting Trees

There are two types of method that are commonly used to find the root of a tree:

- The commonest approach in population genetics is to use an outgroup — a taxon that we can be confident will give the basal lineage in any reconstructed tree. For example, in human population genetics DNA from chimpanzees is often used to provide the outgroup.
- Another approach is to assume that there is a 'molecular clock' so that the rate of change from the ancestral sequence is expected to the same for all taxa. An example of this type of method is 'midpoint rooting', where we find the position in the unrooted tree that is half-way between the two most divergent taxa.
- Because there may be a number of equally parsimonious reconstructions leading to different unrooted trees, typically we would define branch lengths by the number of changes along a branch averaged over all equally parsimonious reconstructions, and then find the midpoint on this.

Back to Cann *et al.*. . .



Cann *et al* Summary I

- The restriction enzymes gave 467 independent sites, of which 195 were polymorphic.
- There were 133 distinct types (haplotypes) among the 147 people analysed.
- 7 haplotypes were found in more than 1 person.
- In constructing the tree they ignored all singleton sites (where the 1 or 0 at the site was present in just one individual), because this is not informative about topology (only branch lengths), so worked with a table of 93 sites.
- Many trees are equally parsimonious. (See Maddison, 1991, and Templeton, 1993 for critiques of the Cann *et al* analysis.)
- Branch lengths calculated using all 195 polymorphic sites.
- They used the midpoint method to root the tree.
- They argue that the most recent common ancestor (MRCA) is likely to have been in Africa.

Cann *et al* Summary II

- They used archaeological evidence of the timing of colonisation of different geographic regions to calibrate their tree.
- They estimated that the common ancestor of all surviving DNA types existed 140,000 to 290,000 years ago.

Some Questions

- To what extent can we read history from the reconstructed mitochondrial tree?
- Even if the MRCA can be validly placed in Africa, does this mean that modern humans originated in Africa?
- Do we expect the same tree for nuclear genetic sequences?
- How does the date of the MRCA relate to the origin of modern humans?

Fast-forward to the era of Next Generation Sequencing

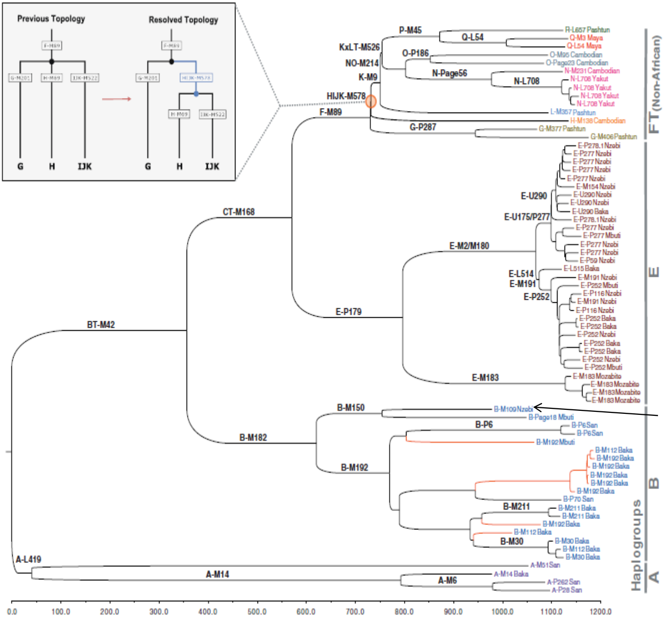
Poznik *et al* (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science*, 341, 562–565.

- This study uses next generation sequencing to sequence 9.9Mb of non-recombining Y chromosome at an average read depth of 250x.
- Similar methods are used to sequence whole mtDNA sequences (16,569bp at 250x) from the same males, plus an additional sample of 24.

Poznik *et al* summary

- The 9.9Mb region of Y chromosome was sequenced from 69 males.
- They identified 11,640 single nucleotide variants.
- The males were chosen from 7 globally diverse populations of the Human Genome Diversity Panel (HGDP) plus 2 additional African populations.
- They used a software package MEGA5 to derive a maximum likelihood tree. This will give an unrooted tree. Although the paper is not explicit on this, almost certainly they will have used the chimpanzee sequence as an outgroup for rooting the tree.
- They estimated a T_{MRCA} at around 120,000–156,000 years ago for the Y chromosome and around 99,000–148,000 years ago for the mtDNA.

Reconstructed Y-chromosome phylogeny



Non-African

W. African, Pygmy.

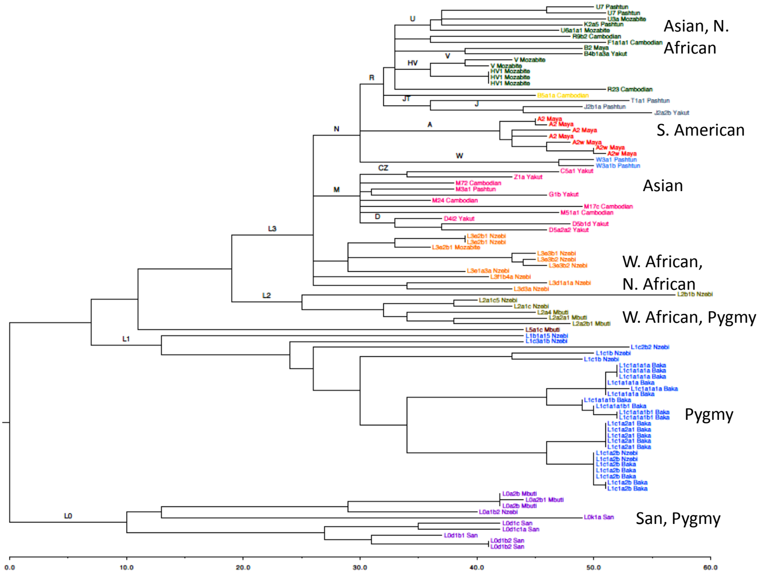
N. African

W. African

San, Pygmy

San, Pygmy

Reconstructed mtDNA phylogeny



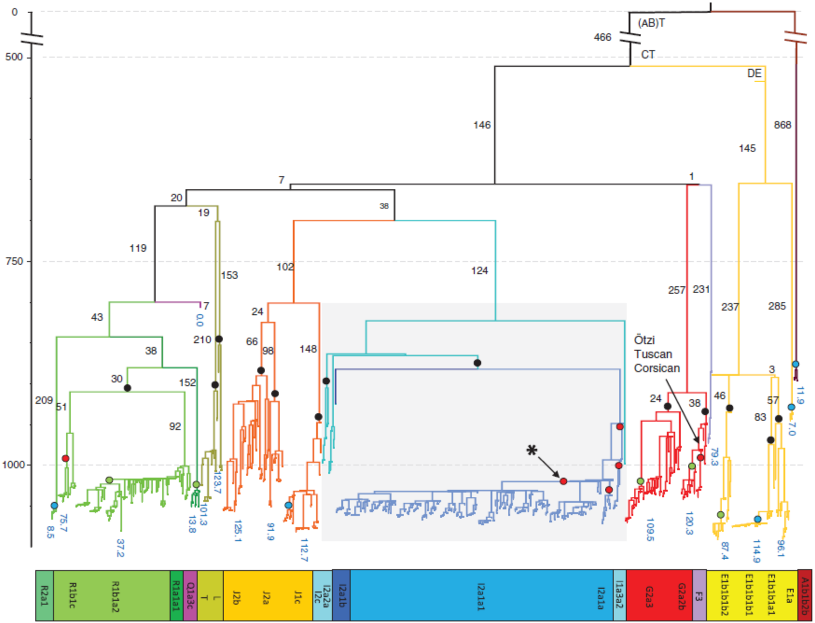
General features of these gene-trees

- There is strong clustering. Mostly this reflects different geographic origins of people. However some geographically similar groups also contain deeply branched clusters. *E.g.* Haplogroup B in the Y (mostly Baka pygmy), and the Baka group in the mtDNA phylogeny.
- Looking backwards in time from the present, and considering the number of lineages present at any time, we can see that typically there is initially a rapid rate of loss of lineages, as they join up together, and then we wait longer and longer for them to join up. (Logically this need not be the case — the rate of loss of lineages could be constant back in time, or it could accelerate back in time).

What does the gene-tree from a single population look like?

- The paper by Francalacci *et al*, (2013, *Science*, 341, 565–569) looks at the Y genealogy for 1204 Sardinians (plus 5 other Italians/S. Europeans).
- This is based on 4.5x sequencing of a 9MB region of the non-recombining Y.
- They used parsimony methods to construct a tree, and rooted it using the chimpanzee reference sequence.
- They estimated the time of the MRCA of their sample at around 180,000–200,000 years ago, and the MRCA of the mainly non-African group at 110,000 years ago.
- The apparent inconsistency with the conclusions of Poznik *et al* may be explained by the fact that Francalacci *et al* use a lower mutation rate (Cann, 2013).

The Sardinian Y-phylogeny



Overall summary so far

- The Francalacci *et al* results even more strikingly show the patterns noted for the Poznick *et al* paper.
- There is strong clustering in the data, even though 99.6% of the sample comes from the same small island.
- There is the same tendency for the rate of loss of lineages to be initially huge, then subsequently slowing down. *I.e* we see short terminal branches and much longer internal branches.
- The genealogy of the Sardinian sample seems to reflect aspects of the global genealogy (including similar T_{MRCA}).

Some further questions

- Why is the gene tree so strongly clustered even for individuals from the same population? Is this true of gene-trees in general?
- Should we expect the T_{MRCA} for Y and mtDNA to be the same?
- What about the T_{MRCA} for typical nuclear genes (autosomal genes)? Do we expect the same tree for autosomal DNA? How might it differ?
- What is the significance of ‘mitochondrial Eve’ and ‘Y-chromosome Adam’? Was there something special about these individuals? Do the dates of ‘Eve’ and ‘Adam’ tell us something about the origin of modern humans.

Coalescent Theory

To help explain these patterns, in the next lecture I will introduce you to what is known as coalescent theory — the theory of gene genealogies.

Further Reading I

Felsenstein, J. (2004)
Inferring Phylogenies.
Sinauer.

Cann, R. L., Stoneking, M., and Wilson A. C. (1987)
Mitochondrial DNA and human evolution
Nature, 325, 31–36.

Cann, R. L. (2013).
Y Weigh In Again on Modern Humans.
Science, 341(6145), 465–467.

Francalacci, P., Morelli, L., Angius, A., Berutti, R., Reinier, F., Atzeni, R., . . . & Sanna, D. (2013).
Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny.
Science, 341(6145), 565–569.

Further Reading II

Maddison, D. R., (1991).
African origin of human mitochondrial DNA reexamined.
Systematic Zoology, 40(3), 355–363.

Poznik, G. D., Henn, B. M., Yee, M. C., Sliwerska, E., Euskirchen, G. M., Lin, A. A., . . . & Bustamante, C. D. (2013).
Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females
Science, 341(6145), 562–565.

Templeton, A. R. (1993).
The “Eve” hypotheses: a genetic critique and reanalysis.
American Anthropologist, 95(1), 51–72.