Outline of lectures 20-22

Trees of Genes – The Coalescent.

1. What happens to molecular evolution as it gets down to looking at differences within populations? We can trace a tree of ancestry for the copies (not the individuals, but the copies) within a population.

2. Each lineage is a sequence of individual copies of genes that were descended from each other, going back into the past. These lineages join with each other as we go back, when two copies are replicates of the same individual copy of the gene in some generation.

3. The most highly publicized such tree came in 1987 when Becky Cann, Mark Stoneking and Allan Wilson made a tree of human mitochondria. Mitochondria are effectively haploid, and inherited only from the mother. From the point of view here, each mitochondrion is inherited as a unit, copied from the one in the mother. Cann et al. used restriction sites patterns in humans, sampled mostly from local hospitals where they could get placentas (as they needed a big chunk of human tissue). They saw three phenomena:

   - All of the 149 mitochondria they looked at were descended from a single female, who has been named Mitochondrial “Eve”.
   - She lived about 200,000 years ago (with a large uncertainty about that, maybe ±100,000 years).
   - She probably lived in Africa. (This is moderately supported by the data).

4. Subsequent studies by Wilson’s lab and others have found the same pattern when full DNA sequence data on mitochondria were used.

5. There was a big reaction to their study because in part it resonated with notions of Adam and Eve, and with earth-mother-godesses.

6. Actually it is inevitable that any small region of DNA will be descended from lineages that converge as we go backwards, and lead to one ancestral copy. So there will be Cytochrome Sams and Hemoglobin Harriets as well. Tracing back to one ancestor is not just a property of mitochondrial DNA or Y chromosomes.

7. If we trace each gene back to the gene in the previous generation that it was copied from, and continue to trace back generation by generation, the lineages randomly converge. Ultimately there is only one gene copy that is the ancestor of everyone. (We are assuming no recombination – for that see below).

8. The English probabilist (and science administrator and University head) (Sir) J.F.C. Kingman showed in 1982 what the random process of formation of a tree of lineages is expected to look like. He called such a random tree “the $n$-coalescent” and the name “coalescent” has stuck. The lineages combine at random, it taking longer and longer for them to combine as the go back. The whole process takes an average of about $4N_e$ generations, and the last two lineages (going backwards) take about half of that.
9. The process is like “bugs in a box”. We have a box full of hyperactive, indiscriminate, voracious, and insatiable bugs. They run around and collide at random. When two bugs collide, one eats the other and then resumes running. This process in fact has exactly the same mathematics as the coalescent: the number of bugs drops rapidly at first, the more slowly as there are fewer bugs to collide with.

10. A random sample of 149 lineages is overwhelmingly likely to have the whole population’s gene ancestor as its root. So mitochondrial Eve is likely to be the mitochondrial ancestor of everyone.

11. For mitochondria you take off the 4 because they are effectively haploid (that gets us down to 2), and because they trace back only to females (which loses another factor of 2): for them it takes about \( N_e \) generations, where \( N_e \) is the effective population size. That means that with a human generation time of 25 years or so, mitochondrial Eve was surprisingly recent unless human population sizes were about 12,000, which is rather small.

12. Population growth affects gene trees by making coalescence be faster when one gets back to periods in which the population size is small. So there are ways of using coalescents to make inferences about past population sizes. However studies in our lab and others suggest that it will take a lot of loci to get a clear picture of past population sizes.

13. Migration also affects gene trees. The more there is, the less consistency one will see between the present location of populations and their placement on the tree. The amount of migration needed to scramble the placement of individuals on the gene tree is about \( 4N_em = 1 \). That means that \( m = 1/(4N_e) \), so only 1/4 of a migrant individual arrives in each population each generation. (Or less gruesomely, one arrives about every 4 generations). Higher rates of migration than this mere trickle will homogenize the genes and lead to the species behaving much like one big random-mating population.

14. The above picture is true if the gene does not recombine. If it recombines, the lines do not only converge, they can also split as one goes back, every time there is a recombination within the locus. (Mitochondria and most of the Y chromosome both don’t recombine). More properly, the genes at one end of the sequence then have a slightly different tree from the ones at the other end. So as one “walks” along the genome, the tree gradually changes.

15. It happens that I have a famous ancestor. Do I have genes in direct descent from him? The ancestor is Charles the Great (Charlemagne), Emperor of the Frankish Kingdom in the year 800. He is one of the great figures of European history, who almost pulled together a large kingdom during the pre-medieval Dark Ages. How do I know I am descended from him? Has my genealogy been done? No. In fact genealogists working backwards from anyone in Western Europe ultimately come to Charlemagne up one line or another, back about 47 generations! As part of my ancestry is Western European, my ancestry presumable does too (and hundreds of millions of other people share this distinction). In Asia the comparable figure is Genghis Khan.

16. If we computer simulate the ancestry of a chromosome as it is traced back up one lineage in my ancestry, it breaks into pieces and most pieces go off up other lineages. After only a modest number of generations all pieces are gone and no part of my genome comes from that ancestral lineage.
17. Calculations of how much recombination will happen on a lineage that is going back to a root $4N_e$ generations ago suggest that the tree will be very different if one goes a distance along the genome such that the recombination fraction $r > 1/(4N_e)$, or $4N_e r > 1$. For humans that is a surprisingly small distance. If $N_e = 100,000$ it is the distance one needs to go to get $0.0000025$ recombination. Since there is about one recombination per 100 million bases per individual gamete per generation, that is about 250 nucleotides (if one assumes recombination is evenly spread along the genome. Actually it is somewhat clumped, with “hot spots” and cold spots so that maybe 1,000-2,000 nucleotides is the distance at which one can expect very different new tree. But even that means that there are over 1,000,000 different gene trees for our genome!

18. That means lots of ancestors, of both sexes and in many different generations, contributed to our gene pool. Hemoglobin Harriet, Cytochrome Sam, little-piece-of-junk-DNA Lyle, and hundreds of thousands of others.

19. An example of “phylogeography” is shown using trees of the cytochrome oxidase I (COX I) in a rotifer *Brachionis plicata*.

20. In the mitochondrial tree the European and Asian sequences are all jumbled together with some of the African sequences. They form a group (a clade). The rest of the African sequences split off on both sides of the root. The time at which the European and Asian sequences start diverging is about 100,000 years ago. (These times are based on molecular clock calculations based on the numbers of differences between the sequences).

21. This suggests the Out Of Africa hypothesis: that a small random subset of African mitochondrial lineages left Africa about then, entering Europe and Asia (presumably through the Middle East), and becoming the ancestors of the European and Asian sequences.

22. What is very surprising about all this is that there were already *Homo erectus* populations throughout Europe and Asia at that time (as well as the African ones that are called *Homo ergaster*). Yet there is no sign that mitochondria from those *Homo erectus* got incorporated into the current human population.

23. The alternative is the Multiregional Hypothesis. All the *H. erectus* populations exchange genes, and the innovations that arise in one area diffuse to all others. No one region is then the place that *Homo sapiens* evolved, and the *H. erectus* populations do not go extinct – they are transformed into modern humans.

24. This is in fact what we would have expected to happen. When a lineage is transformed from one species to another, we mostly think of new mutations being incorporated that occur in various places throughout the species. We do not think of a small population of species A becoming species B, and then driving the parent species to extinction.

25. The Out Of Africa hypothesis is a bit strange compared to this view. It imagines a new species emerging within another, but not be transforming the parent species.

26. It is still early days yet, but a number of studies with other pieces of DNA (other loci in the nuclear genome) suggest that Out Of Africa is right. But the evidence is still quite
weak and one should not be too ready to jump to conclusions when we have such an important question.

27. The ancestors of different regions of the genome occurred at very different times and at different places. Mitochondrial Eve and Y-chromosome Adam did not know each other.

28. Recently work by Svante Pääbo and others has resulted in mitochondrial sequences from Neanderthals. These turn out to be substantially different from modern humans, which is surprising. They imply that much of Neanderthal ancestry diverged from that of modern humans, but they do not rule out that a minority fraction of human ancestry could come from Neanderthals. Coalescent arguments are used to figure out what fraction of our genome could come from Neanderthals and still be consistent with these sequences. As more loci are sequenced from Neanderthals, one would finally expect to see some that would be closer to modern humans, if there is a Neanderthal contribution to our ancestry. So far the followup sequencing has not shown a Neanderthal contribution.

29. The same phenomena occur in all species. The estimation of what all these population properties have been is going to require work on lots of loci, and this is barely under way.

30. When we look at a number of species on a phylogeny, and take a sample of individuals from each and look at its coalescent genealogy, if we could know that we would sometimes see discrepancies between the species tree and the coalescent tree. If a branch is a number of generations long that is a multiple of the population size $N$, coalescence will occur and only one lineage will get back into the immediate ancestor. But if the divergence time is smaller than $N$ generations, more than one lineage can reach the ancestor. There they coalesce randomly with lineages from the sister species, leading often to conflicts between the coalescent tree and the species tree.

31. An example shows inferred coalescent trees for two loci in *Drosophila* species in East Africa and the Indian Ocean islands. They differ but do agree that *Drosophila melanogaster* seems to have a bottleneck after it splits from the much bigger population of *D. simulans*.

32. Species trees can still be inferred because the discrepancies for different parts of the genome are different, and to some extent cancel each other out.

33. There is a lot more work on this under way. Stay tuned.