Outline of lectures 22-24

Trees of Genes – The Coalescent.

1. If we look at a phylogeny from a distance, it looks just like a tree diagram made of thin lines. But (metaphorically) if we get closer, we can see that each line is a species, consisting of multiple populations, each with multiple individuals.

2. So what happens to molecular evolution as it gets down to looking at differences within populations? There is a tree of species, whose forks are speciation events. But we can also trace another kind of tree, the tree of ancestry for the copies (not the individuals, but the copies) within a single population or for individuals from multiple populations within a species.

3. Each lineage in that tree of ancestry of gene copies is a sequence of individual copies of genes that were descended from each other, going back into the past. Those 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. Those are the forks; they are not speciation events, but just cases where one copy of a gene in a parent got copied into two different offspring, ones which ended up being ancestral to copies that we sampled from the population.

4. As these lineages of ancestry of individual copies go back into the past, they accidentally encounter each other and (going backward in time) "join" by encountering such events and being descended in that parent from the same gene copy.

5. 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 mitochondrian 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 in the Bay Area, where they could get afterbirths (placentas) – as for the primitive DNA methods of the time 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 ancestral female, who has been named Mitochondrial "Eve".
   - She lived about 200,000 years ago (with a large uncertainty about that, maybe $\pm 100,000$ years).
   - She probably lived in Africa. (This is moderately supported by the data).

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

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

8. 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 (and Little-piece-of-junk-DNA Leticia) as well. Tracing back to one ancestor is not just a property of mitochondrial DNA or Y chromosomes.

9. 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).

10. When two lineages get back to a generation, we can assume that (for simple population genetics models) each is from a randomly chosen gene from that population. If there are N individuals in that generation there are (for autosomal loci) $2N$ copies. If a lineage happens to come from gene copy number 538 out of $2N$, the chance that the other lineage happens to come from that same copy is simply $1/(2N)$, as there are $2N$ it could come from.

11. So each generation the pair of lineages in effect tosses a coin with Heads pronbability $1/(2N)$, and when it finally comes up Heads, that is the number of generations back to when the two lineages merge ("coalesce").

12. Thus the time back to coalescence of two lineages is, on average $2N$ generations. So for $N = 10,000$ is averages $20,000$ generations but has the distribution (technically a Geometric Distribution) of the number of coin tosses until Heads when there is a Heads probability of $1/20,000$.

13. 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.

14. 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.

15. 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.

16. 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.

17. So far we have assumed a single population with constant population size, and no migration to or from other populations. We have also assumed we are looking at a single point on

the DNA, that recombination cannot separate different sites in the DNA, that that single point comes from only one ancestor. Here are how some of these complications affect the coalescenty genealogy of gene copies.

18. **Population growth** affects gene trees by making coalescence be faster when one gets back to periods in which the population size was 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.

19. **Migration** also affects coalescent 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.

20. **Recombination**. 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.

21. 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 pulled together a large kingdom during the pre-medieval Dark Ages, almost 300 years before William the Conqueror invaded England, and only abut 300 years after the fall of the Roman Empire.

22. 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 presumably does too (and hundreds of millions of other people share this distinction). In Asia the comparable figure is Genghis Khan.

23. If we computer simulate the ancestry of a chromosome as it is traced back up one lineage in my ancestry, say a lineage leading back to Charlemagne, the chromosome 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.*

24. Your genome starts out with 2×23 pieces (the chromosomes) and each generation about 33 recombinations occur per haploid genome. So going back 20 generations, your present genome has as its ancestors from $2 \times 23 + 2 \times 20 \times 33 = 1366$ pieces of genome, quite possibly all from different ancestors.

25. But you also have $2^{20} = 1,048,576$ ancestors then. Of course they might not be existing at exactly the same time. Are they all necessarily different people? (No!). But even if not, it is pretty much inevitable that, at that remove, *most of your ancestors did not contribute any genes to you*! So if you are proud to be a descendant of Count Rudolf the Gross, you may actually not have inherited the gene that perhaps predisposed him to want to cleave

people's heads in twain. Your copies of that gene may instead have come from Glenda the Cowherd and Rupert the Peddler.

26. Calculations of how much recombination will happen on a coalescent lineage that is going back to a root $4N_e$ generations ago suggest that two sites nearby in the genome will be inherited on very different coalescent trees if they are 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 indvidual 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!

27. So as you go along a chromosome, there are chunks of it that share the same coalescent tree and all go back to the same "Eve". But you don't have to go far along the chromosome to find yourself in a chunk that has a complertely different coalescent ancestor.

28. That means lots of ancestors, of both sexes and in many different generations, contributed to our gene pool. There is then no one tree of ancestry of our genes within the human species.

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

30. 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).

31. 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.

32. 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.

33. The alternative is called the Multiregional Hypothesis. According to it, *Homo sapiens* arose as all of the *H. erectus* populations kept exchanging genes, so that the innovations that arise in one geographiuc area diffused to all others. In that situation, no one region is then the place that *Homo sapiens* evolved, and the *H. erectus* populations do not go extinct – it was gradually transformed into modern humans.

34. This is in fact what we would have expected to happen, and probably happens often. 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.

35. 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.

36. 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.

37. 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 they lived at least tens of thousands of years apart.

38. 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.

39. Neanderthal genomes are now being finished. They show strongly suggestive evidence that about $4-5\%$ of the genes in the present human genome come from a Neanderthal ancestor, so may 1,000 loci.

40. ... but recent work in our department by Ben Vernot and Josh Akey suggest that the 4-5% is different in different people.

41. It is *not* the case that 4-5% of *people* "are" Neanderthals.

42. Sequences from a bone in a Siberian cave show a different archaic human, "Denisovans", who are as distinct from us and from Neanderthals as we are from Neanderthals. They seem to have contributed a small fraction of the ancestry of people from New Guinea and some other Asian groups.

43. Coalescent 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.

44. Note – making a tree for one genome versus another is not particularly valid if the species are close. Because ...

45. **Species differences and coalescents**. 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.

46. 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*.

47. 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.

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