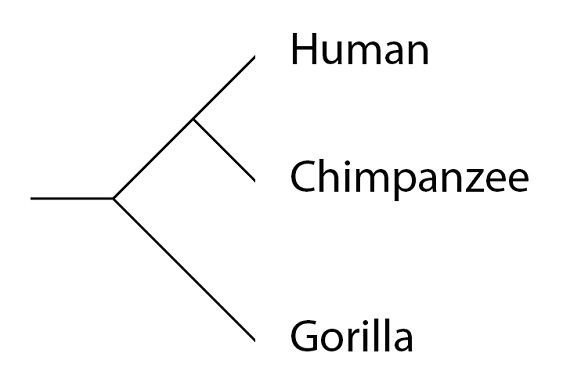
**Course: Advanced Bioinformatics**

**Module title: Genes and Species Tree**

**Module no. : 53**

**Species trees**

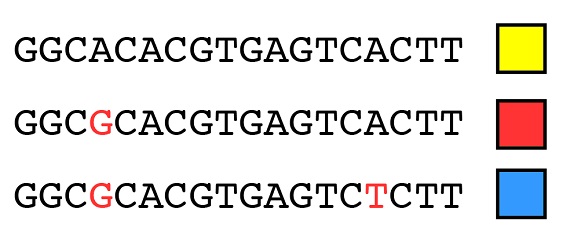
Let’s return to what is by now a familiar example of a phylogeny: that of humans, chimpanzees, and gorillas:



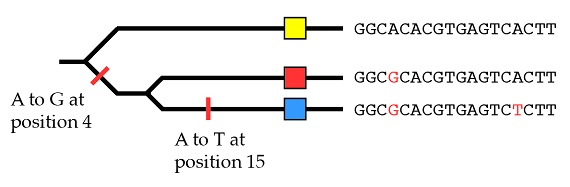
Phylogenies are also known as “species trees”, since “tree” is another name for phylogeny. A species tree shows us the overall pattern – which species share a common ancestral population more recently, and which share a common ancestral population more distantly in the past. In other words, as we noted in the last post in this series, a phylogeny is a measure of shared history and separate history for any two species. The longer two species have a common history, the more similar they are expected to be, on average. Humans and chimps, for example, continue to share a common history for several million years after the lineage leading to gorillas separates from the (human / chimpanzee) common ancestral population. This shared history is what *on average*, makes the chimpanzee and human genomes more similar to each other than either is to the gorilla genome.  Individual genes (and their alleles) may have a different history within species as they separate from one another. For this type of analysis, we need to examine phylogenies for individual genes – so called “gene trees.”

### ****Gene trees****

How variation (alleles) arise through mutation, it should be fairly intuitive that the same principles that can be used to group species into a phylogeny can also be used to group alleles of a single gene into a phylogeny. For example, consider the DNA sequence of three alleles of the same gene, which we can represent as the “yellow”, “red” and “blue” alleles (the colored boxes). Sequence differences that make these alleles distinct are highlighted in red text:



Using the same principles that we used for species as a whole, we can explain the origin of these three alleles by two mutation events (starting with a given that the yellow allele is the ancestral state):

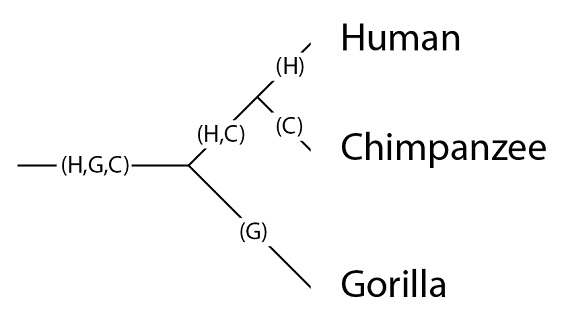


So, within a population, we can reconstruct the allele history of an individual gene using the same methods we have previously applied to species as a whole.

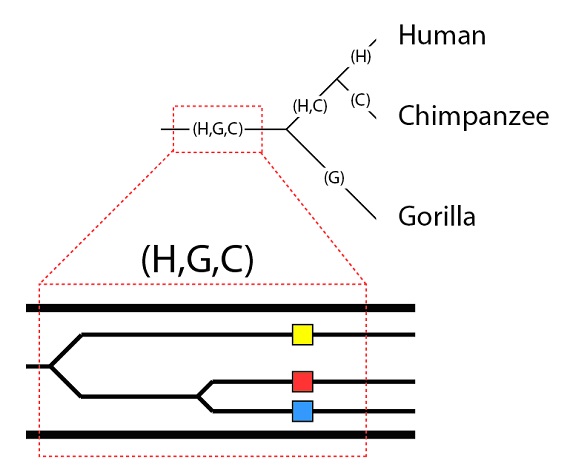
### ****Speciation with genetic variation along for the ride (or not)****

So, mutation is constantly producing new alleles (variation) within populations, and processes such as [natural selection](http://biologos.org/blog/evolution-basics-natural-selection-and-the-human-lineage-part-1) and [genetic drift](http://biologos.org/blog/evolution-basics-from-variation-to-speciation-part-2) work to either increase or decrease the frequency of alleles in populations over time. How speciation events occur, starting with populations that separate from one another, and accrue differences over time that may lead to the formation of distinct species. All that remains is to bring these ideas together: to consider what might happen to variation (alleles) within a population as it goes through a speciation event. To do that, let’s track our hypothetical alleles through the speciation events that led to humans, chimpanzees and gorillas.

This species tree has the following populations: the population that is ancestral to all three species, designated “(H,G,C)” for “(Human, Gorilla, Chimpanzee)”; the population ancestral to both humans and chimps (H,C) and the lineages (populations) that lead to the present day species after their last speciation event with the species on the phylogeny (H), (G) and (C):



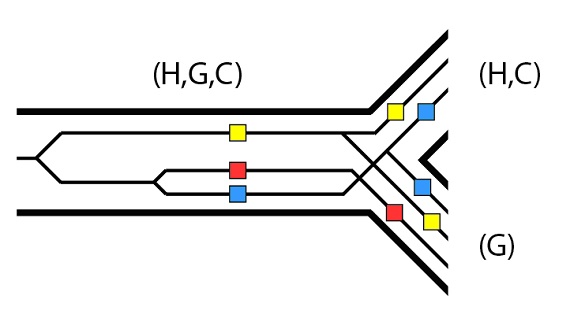
It’s important to keep in mind that a single line on the phylogeny is in fact a population, and populations can have genetic variation. Let’s place our three alleles into the (H,G,C) population:



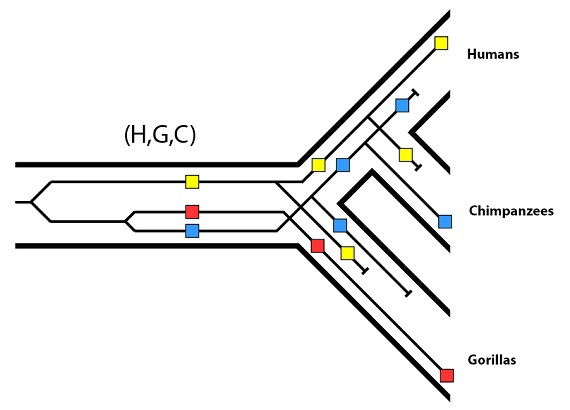
Now we are set to explore possibilities for how these alleles will be inherited (or not) through the speciation events that will occur. One possibility is straightforward – all three alleles will be inherited by all three species. This possibility is called “complete lineage sorting” since it represents a perfect segregation of all alleles into all lineages. This requires that all three alleles be present in the subpopulations that divide into separate lineages, and that no alleles be lost over time in any lineage. While this is certainly possible, it is by no means certain. When populations separate it is unlikely that all alleles in the original population will be represented in both subpopulations after the divide. Also, it is possible that selection or genetic drift may cause alleles to be lost over time in one lineage but not another. Anything other than perfect segregation of all alleles into all lineages is called “incomplete lineage sorting” – and for a large genome, it is a given that at least some genes will exhibit this effect.

### ****Incomplete lineage sorting – a worked example****

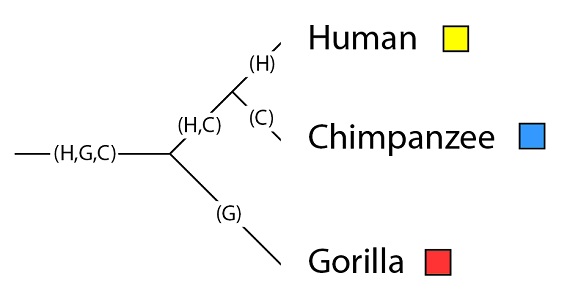
The first challenge to complete lineage sorting that these three alleles will face is the speciation event that separates the (H,C) and (G) lineages. For the purposes of this example, let’s suppose that the red allele is excluded from the population that forms the (H,C) lineage, but that all three alleles persist in the (G) lineage. You will recall that this is an example of the [**founder effect**](http://biologos.org/blog/evolution-basics-from-variation-to-speciation-part-2) – a nonrandom sampling that can exclude alleles from a new subpopulation by chance:



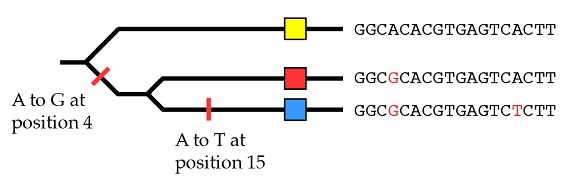
Now let’s examine one possible scenario following on from the (H,C) / (G) speciation event. In the (G) lineage, the yellow and blue alleles are lost over time. At the (H) / (C) speciation event, both the blue and yellow alleles segregate into both lineages, but in the (C) lineage, the yellow allele is later lost. Similarly, the blue allele is later lost in the (H) lineage:



For this particular gene, then, we have the following final pattern:



And at last we see the issue: the gene tree for these alleles is at odds with the species tree. Recall that in the gene tree, the red and blue alleles are more closely related to each other than they are to the yellow allele:



In the species tree, however, the two closest relatives (chimpanzees and humans) do not have the two most closely related alleles – they have more distantly related alleles.

Now that we have worked this example, hopefully the reason behind the discrepancy is clear – there is no guarantee that alleles will sort in a lineage to match up with the overall species pattern. If a gene has variation in a population undergoing speciation events, it is expected that some of the time it will assort with a pattern that does not match the species pattern – in some cases, it will have a gene tree that is “discordant” with the species tree. For a population with thousands of genes with multiple alleles present, it is a given that some alleles will assort into a discordant pattern. Far from being a problem for evolution, discordant trees are predicted by evolution. It would be a problem if we did not observe them – but in fact we do, and as we shall see next time, we observe them in precisely the pattern that matches what we would expect based on species trees.