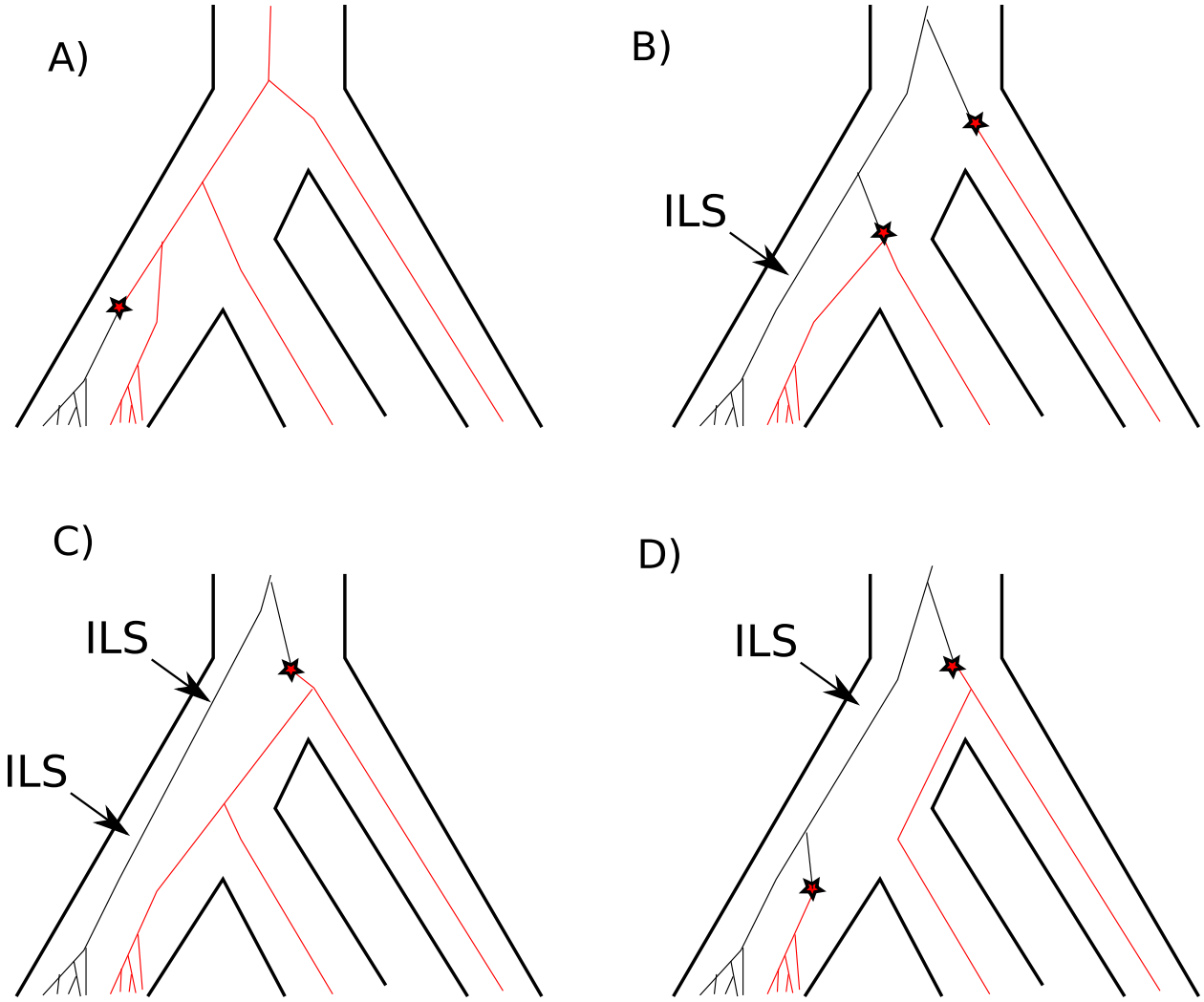

We observe alleles that are polymorphic in humans (left branch), and the same allele observed in one chimp (middle) and one gorilla (right). Depicted are the four most parsimonious scenarios: in any case it requires one mutation; and in other cases it requires at least two more rare(-ish) events: incomplete lineage sorting (ILS) and/or another mutation. The likelihoods are, I think:

$$p_A > p_B > p_C > p_D, \tag{1}$$

(see below); but D will create the bump on the right.

A The standard model, mutation sometime in the tree leading to modern humans. Frequency distribution skewed towards low frequency.

B Two mutations, one leading to chimps and humans; the other on the gorilla lineage. Requires the polymorphism in humans to be fairly old, so has a flat frequency distribution.

C One mutation, but long ago. Requires the polymorphism to be old, again.

D Two mutations, one in chimp-gorilla and one in humans. The red allele, being recently derived, has a skew towards low frequencies; which we miscall as high derived frequencies.

1 Asymmetric mutation

Suppose possibly asymmetric mutation rates: $\mu_\uparrow$ from black to red, and $\mu_\downarrow$ from red to black. Suppose that the probability of ILS in the whole human lineage is η_C, and in the human-chimp ancestor is η_G. Finally, suppose that the typical coalscence times in human, human-chip ancestor, and human-chimp-gorilla ancestor are, respectively, T_H, T_C, and T_G. The the relative rates of A–D are, roughly:

$$A : \mu_\downarrow T_H \qquad B : \mu_\uparrow^2 T_C T_G \eta_C \tag{2}$$

$$C : \mu_\uparrow T_G \eta_C \eta_G \qquad D : \mu_\uparrow^2 T_H T_G \eta_G. \tag{3}$$

The human-chimp divergence is like 7 mya, and the human-chimp-gorilla only a couple more million years back; and ancestral effective population sizes are thought to have been $\sim 10\times$ larger than the human effective population size; so η_G is not so small at all: like, 0.2? On the other hand, we have lots of samples, so the chance of more than one lineage making it back to the human-chimp ancestor is pretty good, maybe of similar order? If $\mu_\uparrow/\mu_\downarrow \approx 10$, this suggests that the ratio of A to C is about 10:1.

In C and D the relevant numbers of generations for mutation should be around 10^5; for T_H in D this is thanks to the large number of samples; and for T_G in both this is thanks to the large effective population size of the ancestral population. Even if $\mu_\uparrow \approx 10^{-7}$, this still leaves D about 1000 times less likely than A.

As for B, note that the mutation on gorilla has the full branch to occur on, i.e. $\sim 10^7$ years; so $\mu_\uparrow T_G$ is only a small factor. Because there are fewer lineages to mutate, the mutation in human-chimp is less likely than the one occurring in A by ad additional factor; and the incomplete lineage sorting is another factor. Incorporating $\mu_\uparrow/\mu_\downarrow \approx 10$, this suggest a $A : D$ ratio of only about 10:1.

2 Selection

Selection (or biased gene conversion) will do several things: effectively increase the mutation rate of the beneficial allele; make ILS *less* likely; and distort the frequency spectrum.

. . . effects on A–D?