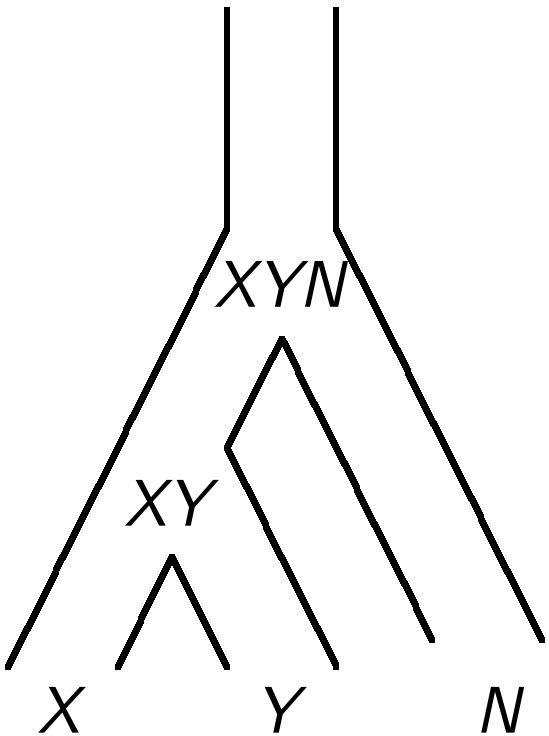
Site Patterns and Population History: the Intuition that Underlies Legofit

Alan R. Rogers

July 26, 2022

Notation for populations



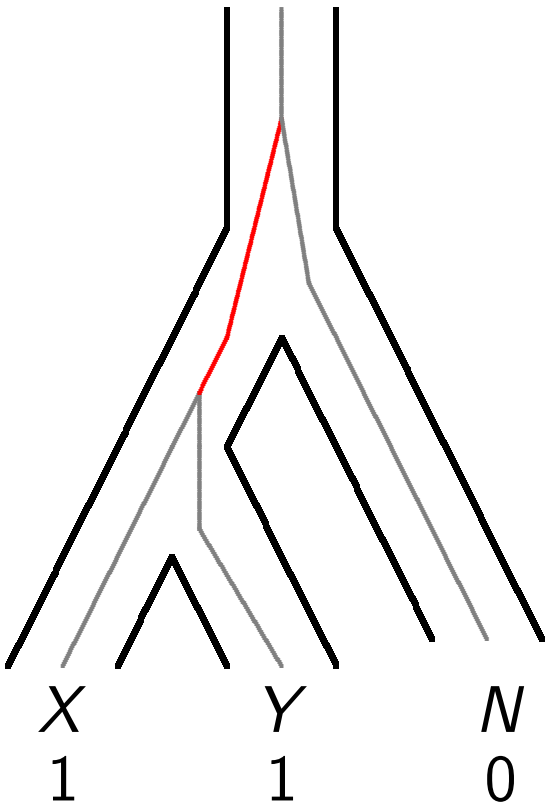
X, Y, and N are populations: African, European, and Neanderthal.

XY: population ancestral to X and Y.

XYN: ancestral to X, Y, and N.

Nucleotide site patterns

Pattern *xy*



Haploid sample: 1 nucleotide from each population.

Mutation on red segment would appear in samples from X, Y, not that from N.

Call this the *xy* site pattern.

I will write $xy \succ yn$ to mean that *xy* is more common than *yn*.

Why haploid samples?

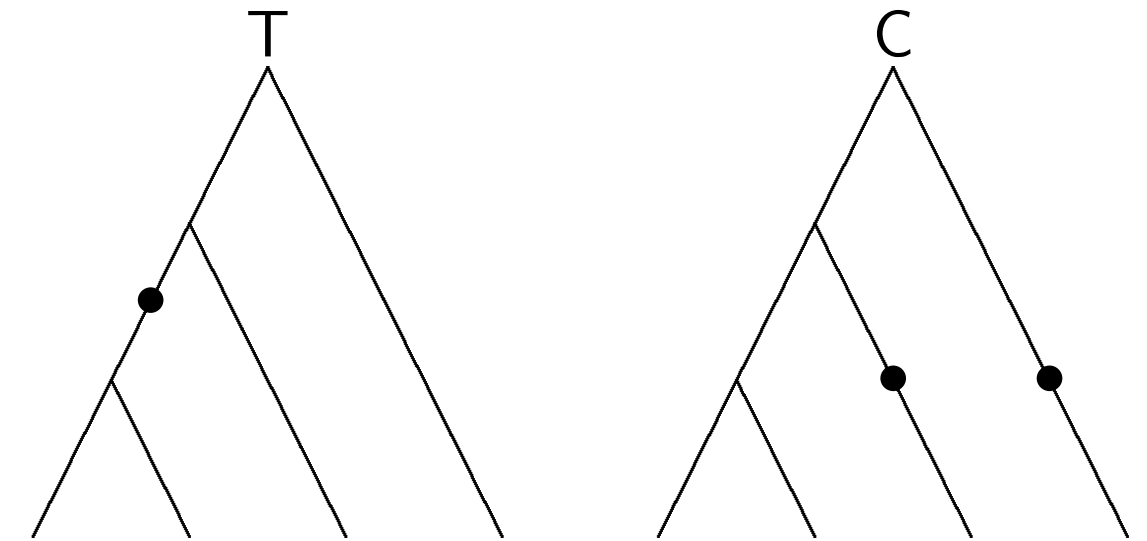
No variation within populations $\Rightarrow$ results not affected by recent history of population size.

The haploid samples are hypothetical; our real samples are larger. We use all the genomes in the real data to calculate the probability of observing site pattern xy in a hypothetical haploid sample:

$$\Pr[xy] = p_X p_Y (1 - p_N)$$

where p_i is the frequency of the derived allele in the sample from population i. These site pattern frequencies are our data.

Calling ancestral and derived alleles

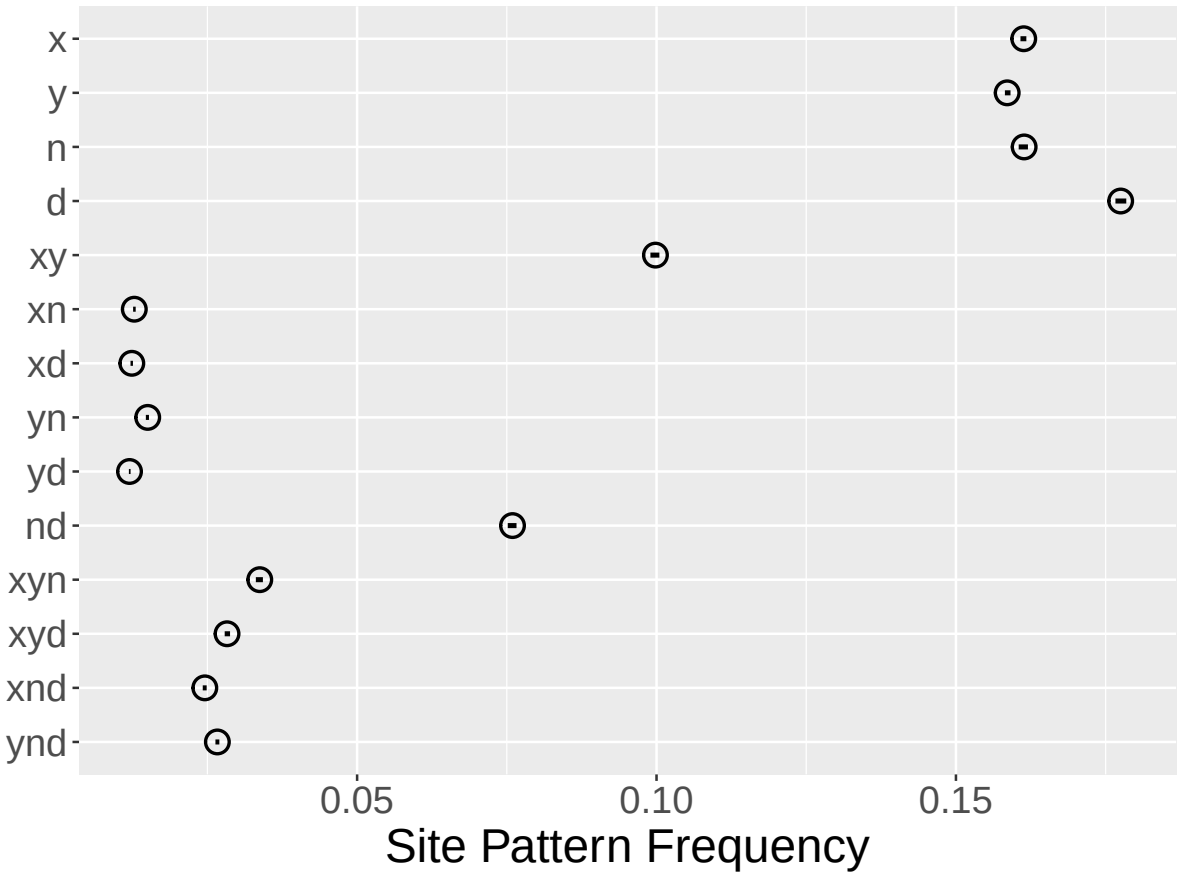


2 mutations needed if C is ancestral.

Only 1 needed if T is ancestral.

Prefer hypothesis requiring fewer mutations, because mutations are rare.

Observed site pattern frequencies

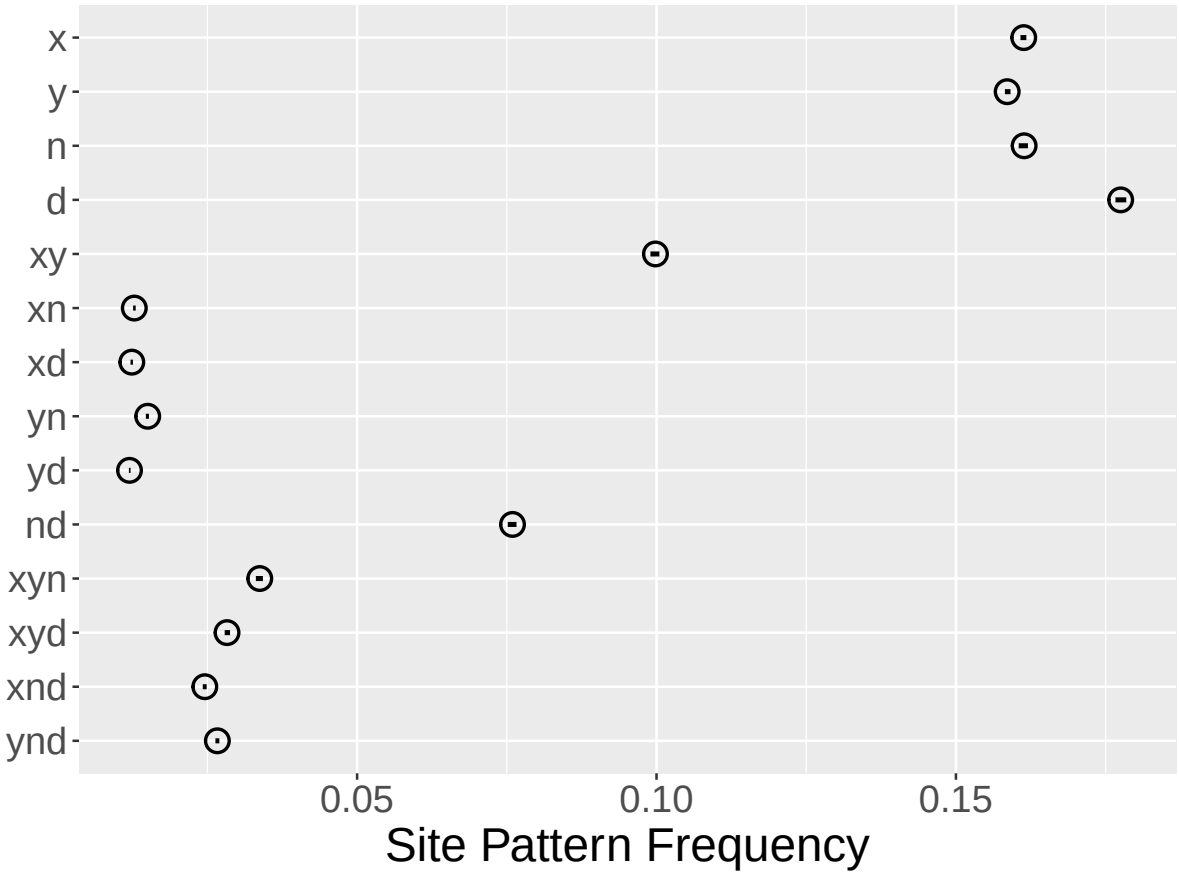


X, Africa; Y, Europe; N, Neanderthal; D, Denisovan.

Horizontal axis: rel. freq. of each site pattern

"Dots" w/i circles are 95% confidence intervals.

The pattern in the data



Lots of singletons (x, y, n, and d)

Two doubletons (xy and nd) especially common

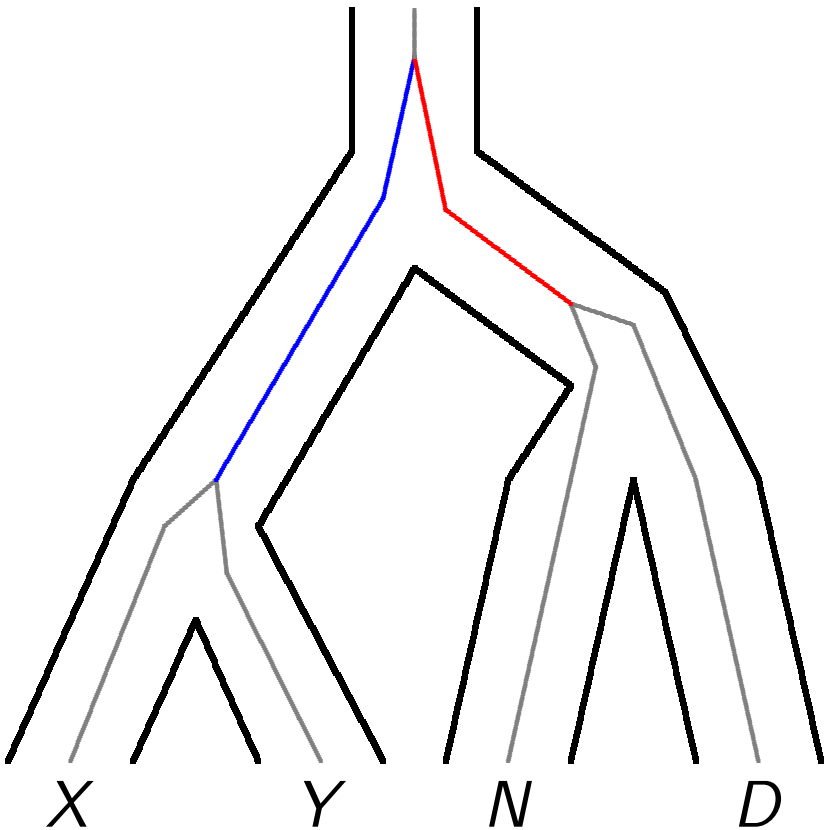
yn more common than other rare doubletons.

xyn more common than other tripletons.

How can we understand this pattern?

1st pass: no frills

Early *N*-*D* split



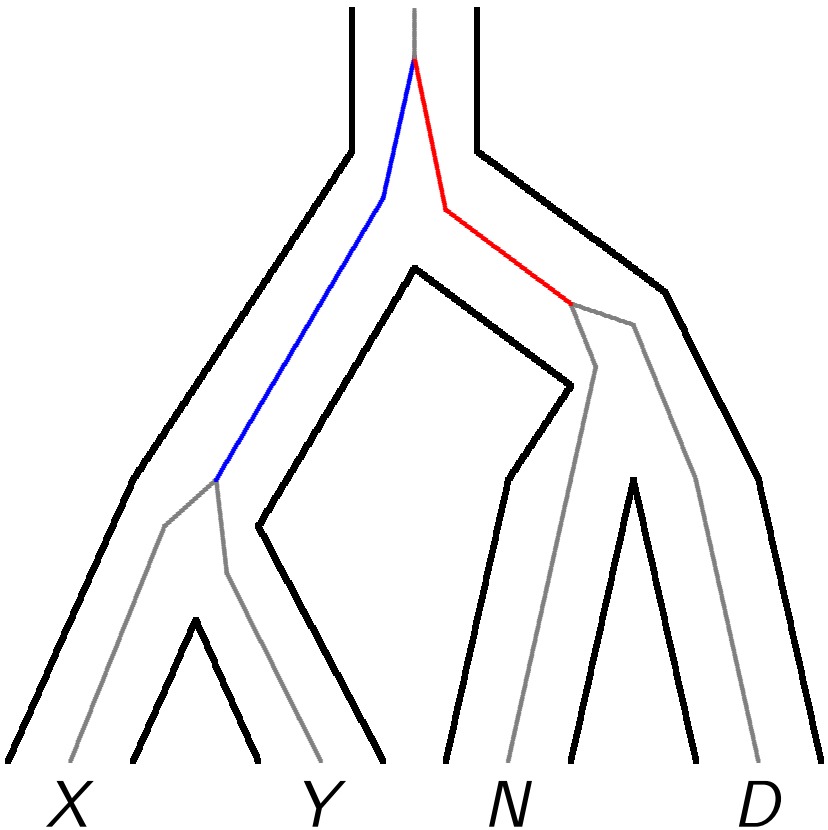
No gene flow; gene genealogy matches population tree.

Many other genealogies are possible, but this one will be common.

Captures large-scale pattern; misses subtleties.

Why are *xy* and *nd* so common?

Early *N*-*D* split

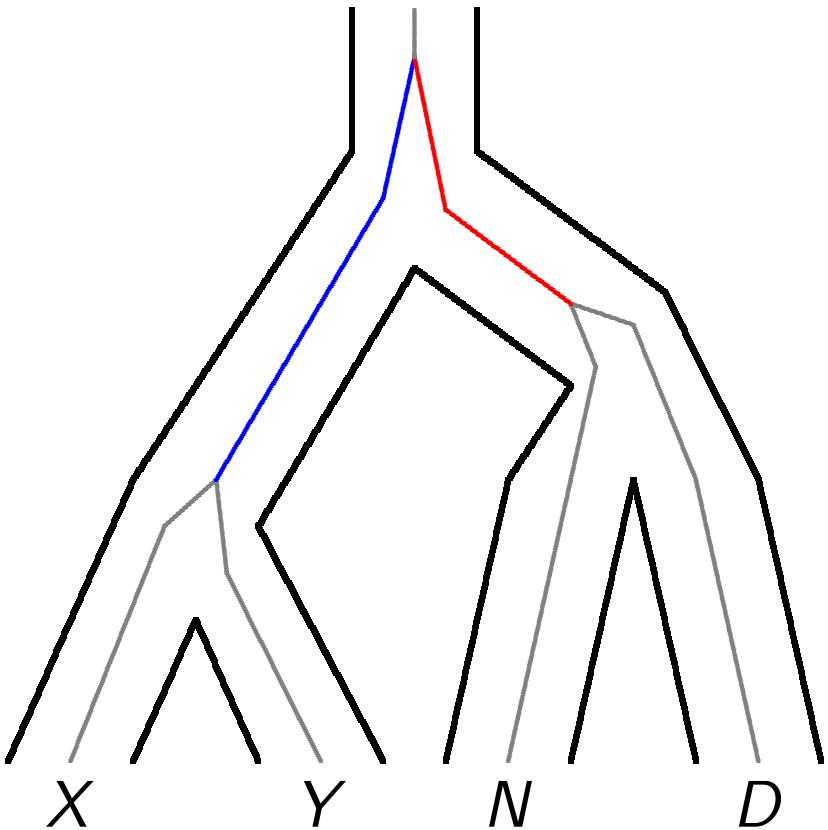


Mutation on blue $\rightarrow$ *xy*;
mutation on red $\rightarrow$ *nd*.

xy and *nd* are common because *X* and *Y* are closely related, as are *N* and *D*.

Why is $xy \succ nd$?

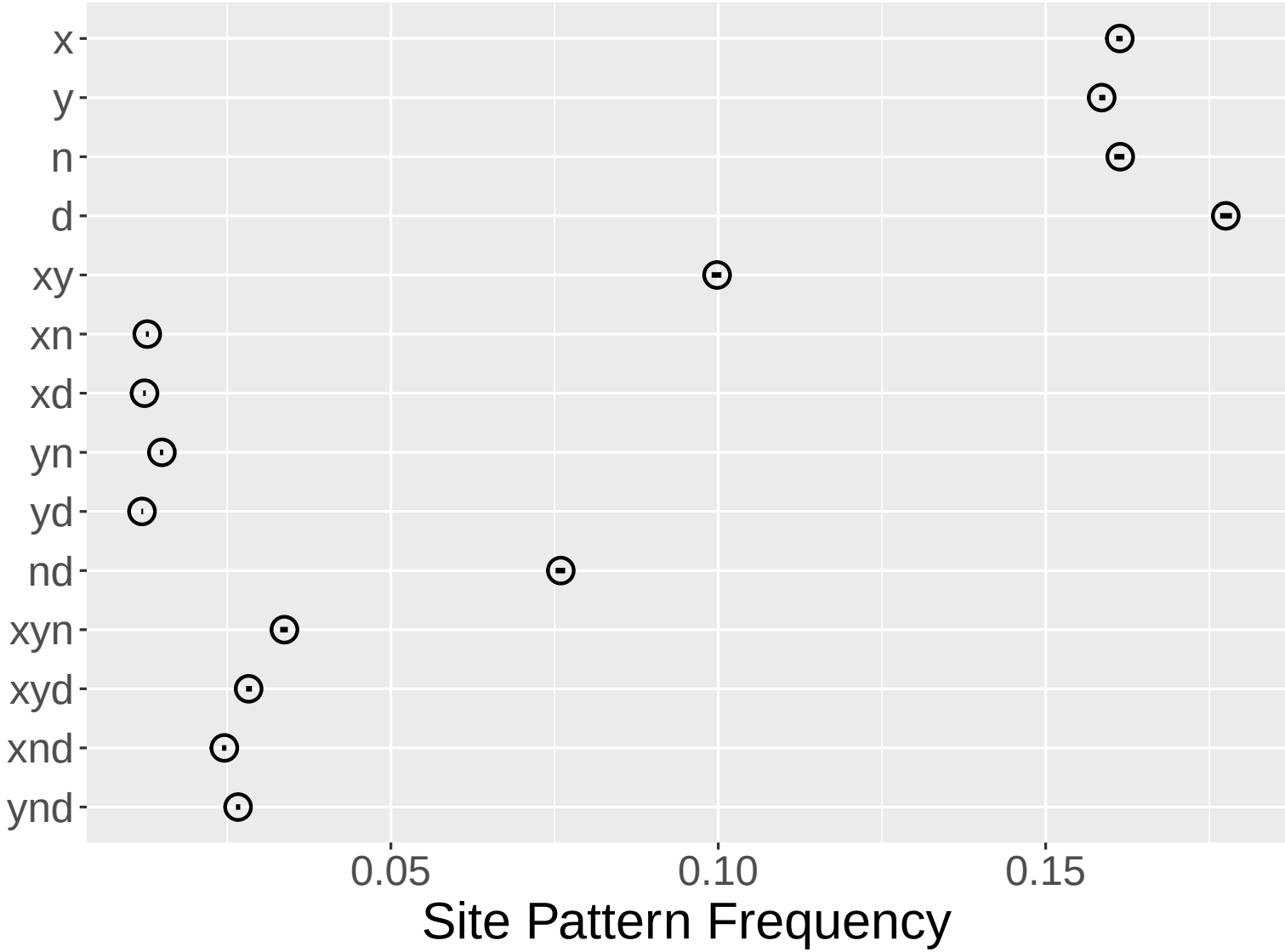
Early N-D split



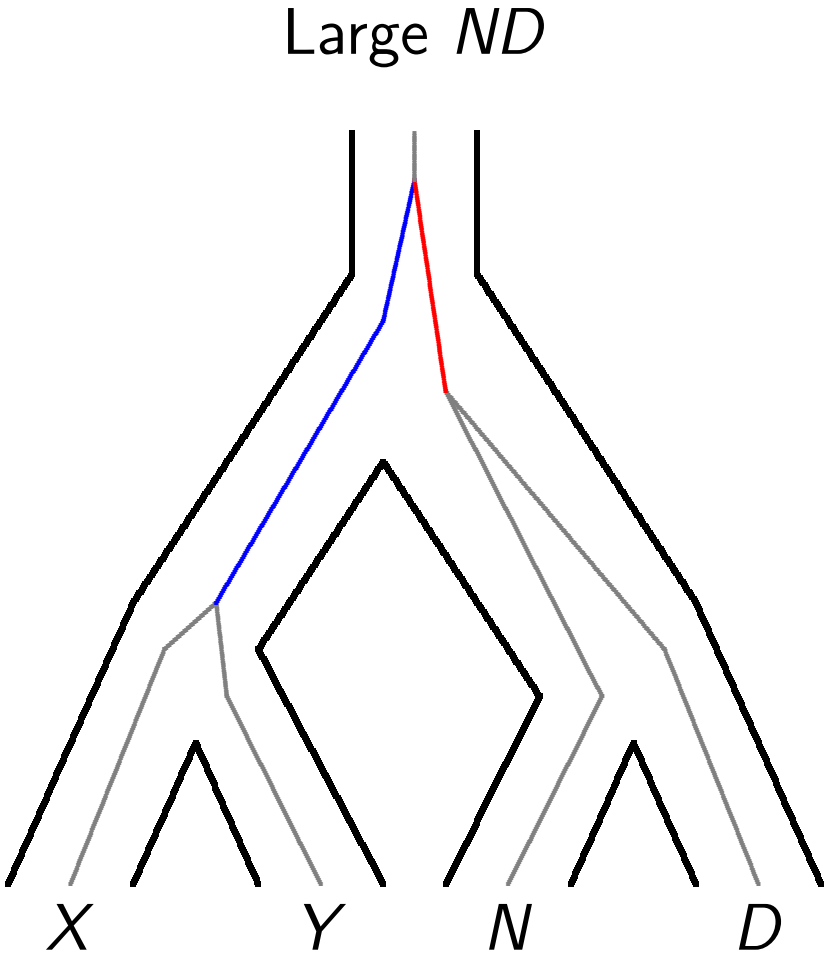
Blue branch is longer than red, because X and Y separated more recently than N and D.

Explains why $xy \succ nd$.

Data again: *xy* and *nd* common, but $xy \succ nd$



An alternate hypothesis



Separation times are equal.

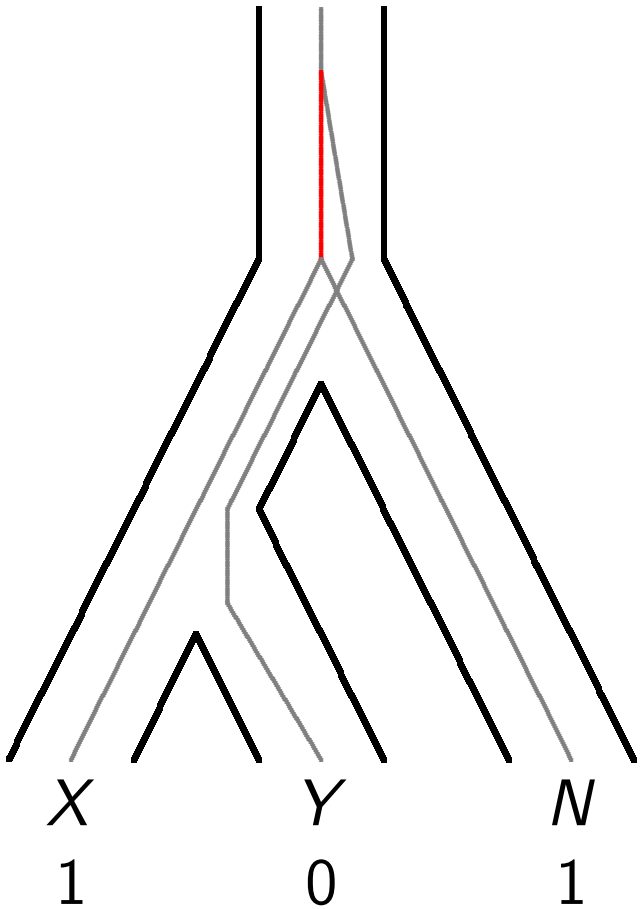
But ND is large, so coalescence is slow, and red branch is short.

$xy \succ nd$ because ND is larger than XY.

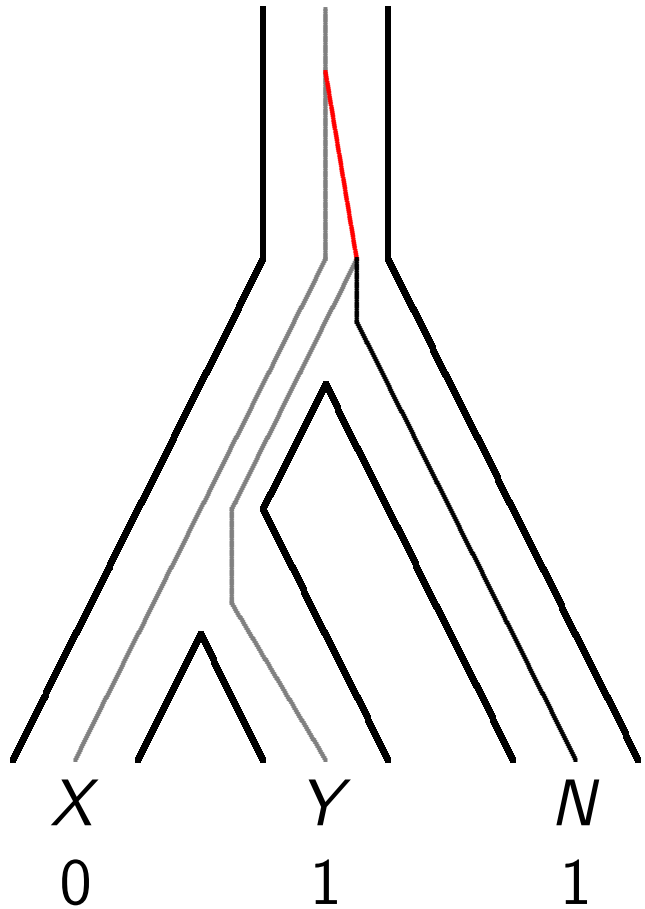
The two hypotheses are hard to tell apart.

Counterintuitive site patterns

Pattern *xn*



Pattern *yn*



Incomplete lineage sorting

Suppose that, as we trace the ancestry of our sample backwards in time, the lineages from X and Y don't coalesce until we reach XYN.

Then there are three lineages, X, Y, and N, in the same population.

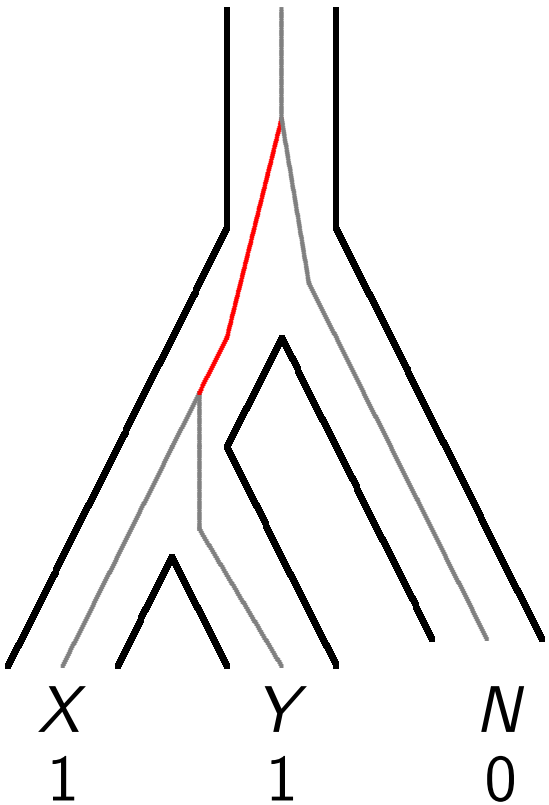
They can coalesce in any order.

Site patterns xy, xn, and yn are equally likely.

This process is called "incomplete lineage sorting."

Pattern *xy* can also arise another way

Pattern *xy*



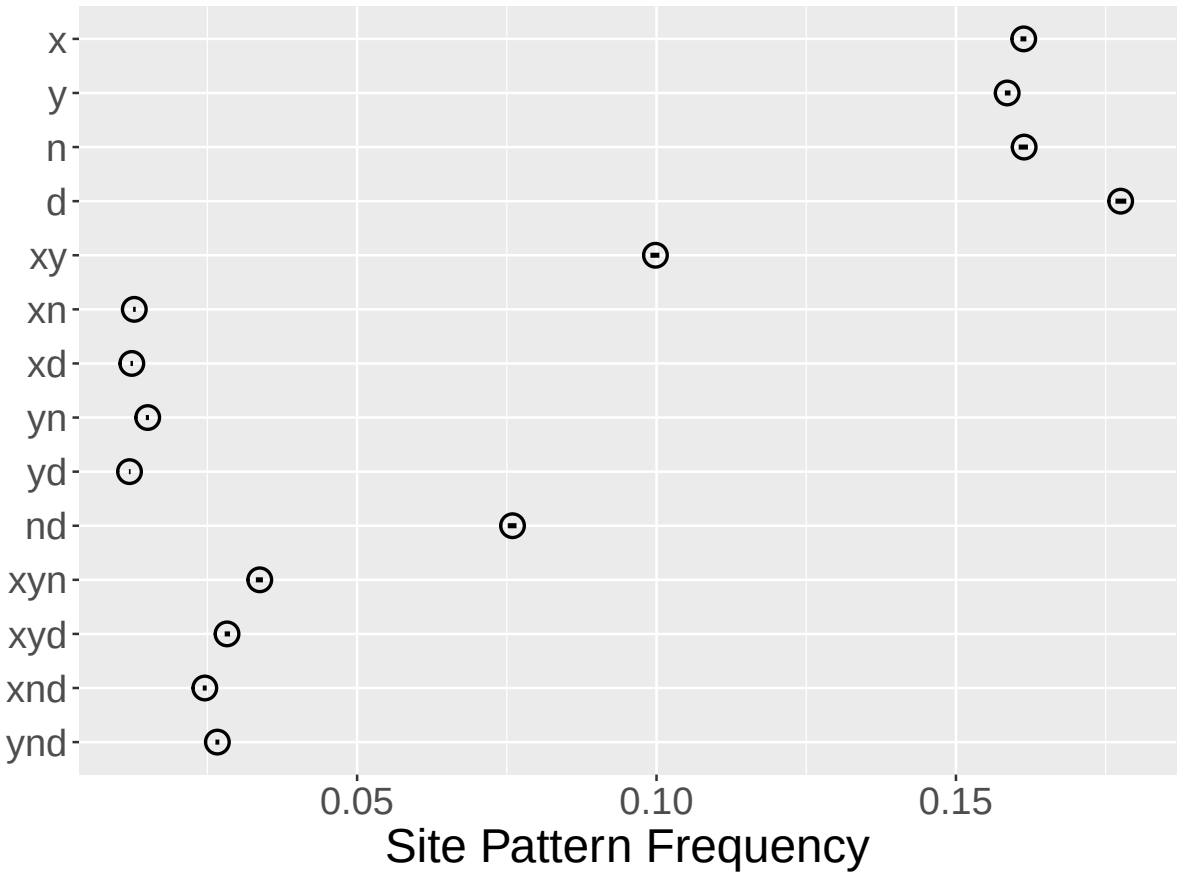
The lineages from X and Y may also coalesce w/i XY, generating site pattern *xy*.

So $xy \succ xn, yn$.

xn and *yn* should be equally common.

This is the pattern expected in the absence of gene flow.

Does incomplete lineage sorting (ILS) explain the data?



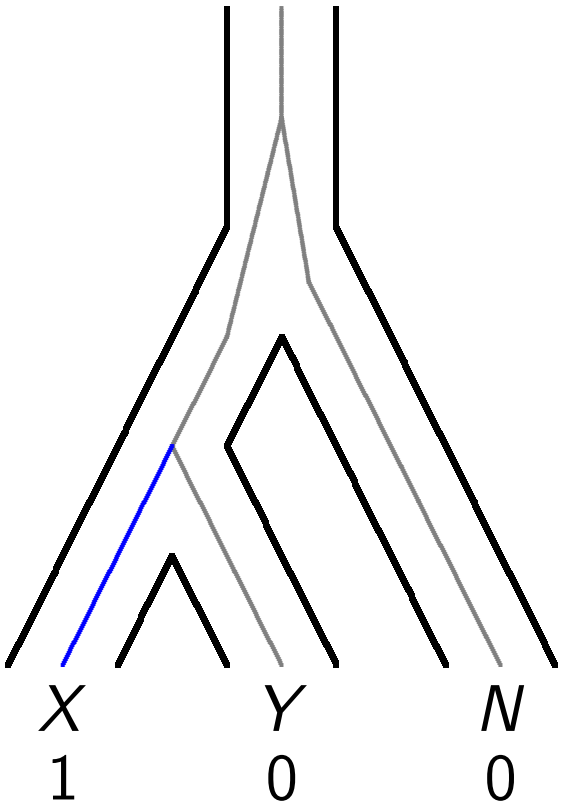
xd and *yd* have equal frequencies, as they should under ILS.

But $yn \succ xn$, and the difference $>$ the confidence intervals.

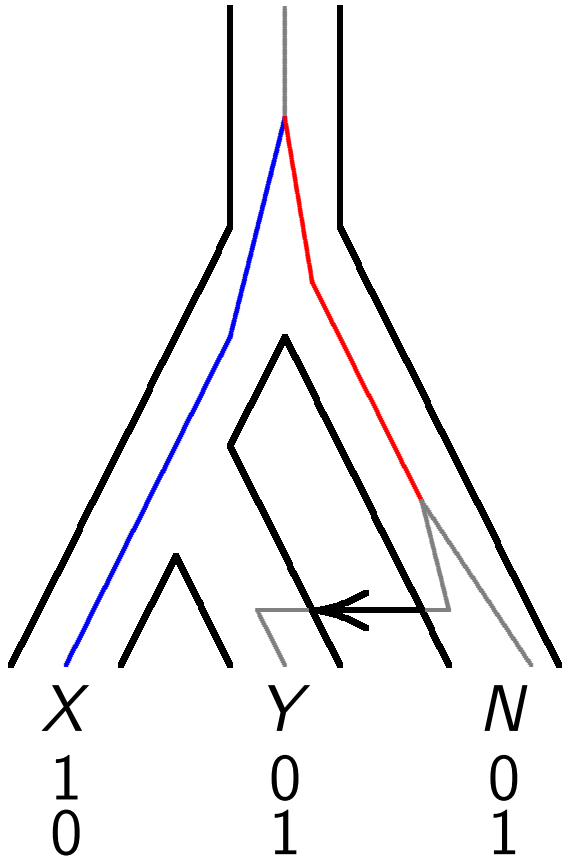
Why?

The effect of gene flow

Without admixture



With admixture

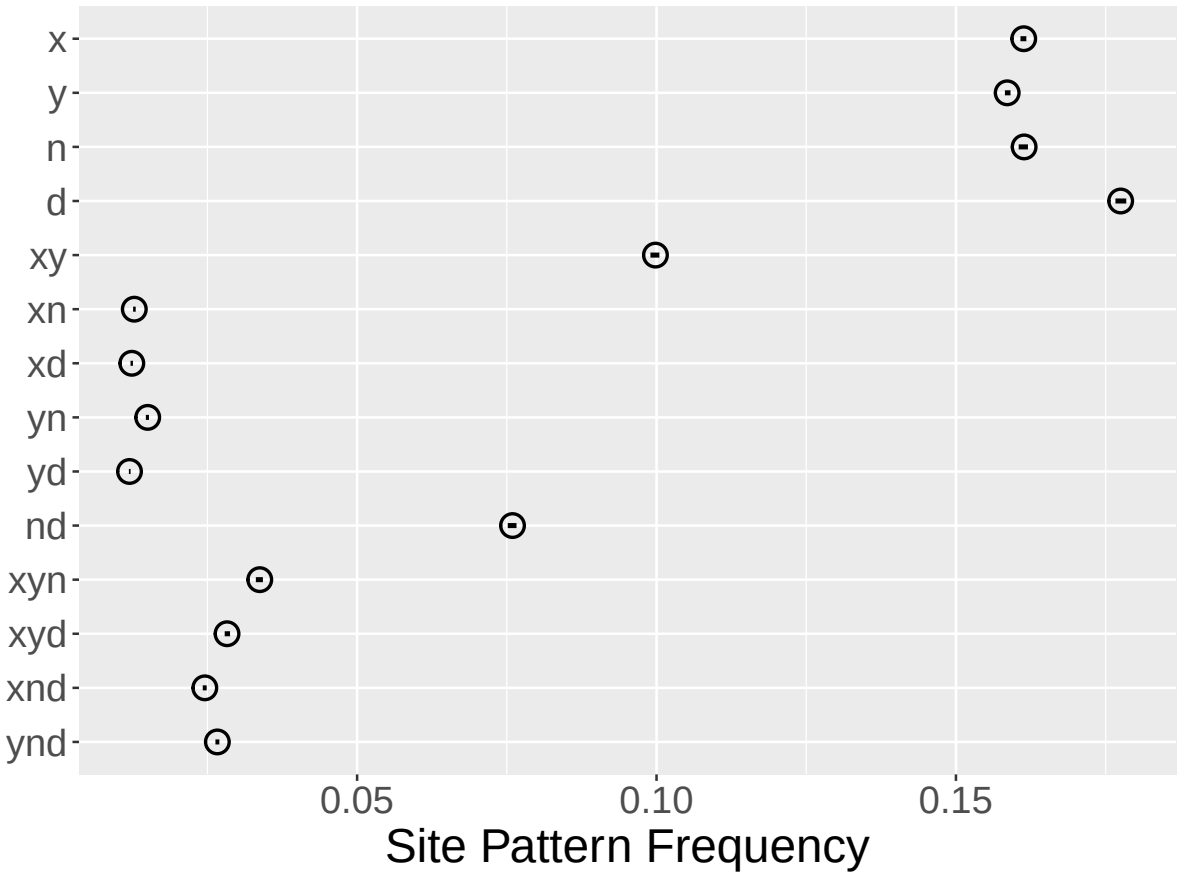


$N \rightarrow Y$ gene flow inflates the frequency of *yn*.

Also inflates frequency of *x*.

Effects are small unless the rate of gene flow is high.

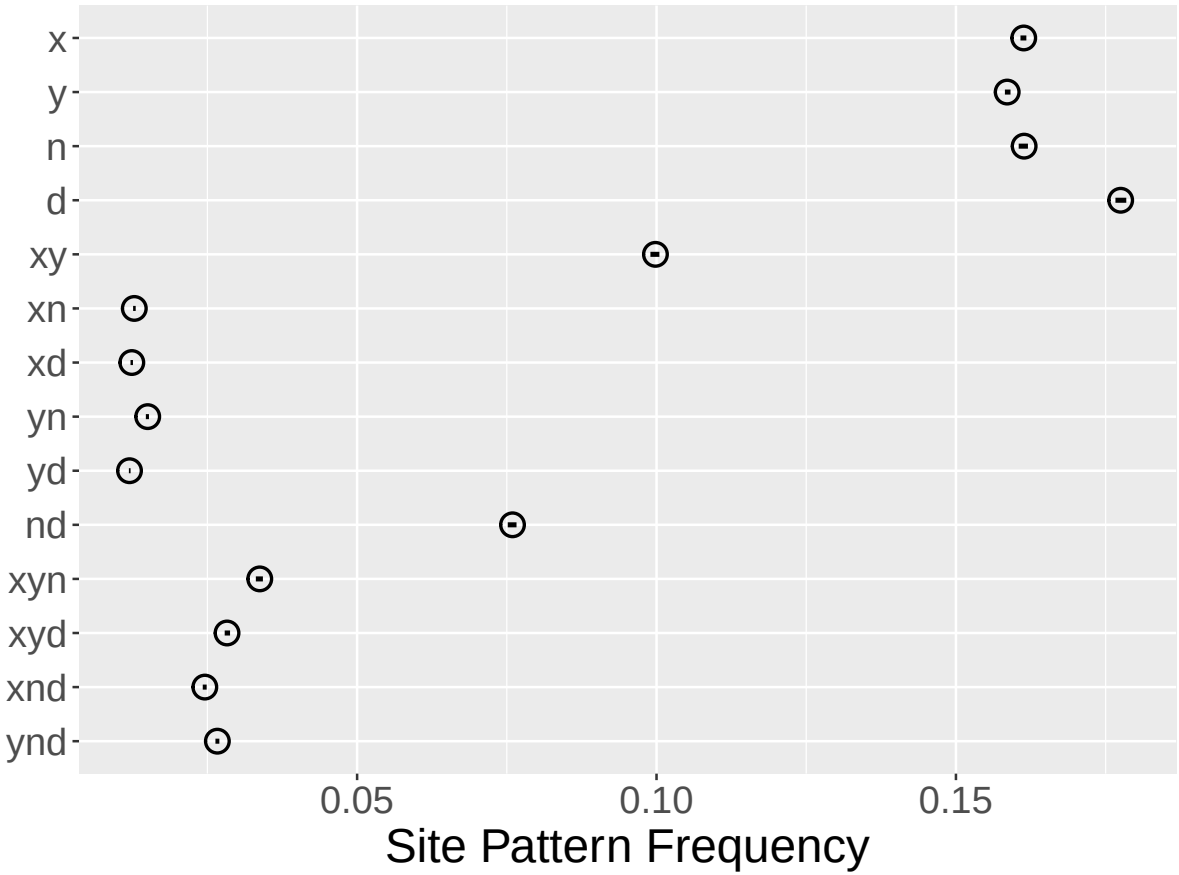
Data are consistent with $N \rightarrow Y$ gene flow



$yn \succ xn$, and
$x \succ y$.

Signature of
$N \rightarrow Y$ gene flow.

Puzzling excess of *d* site pattern



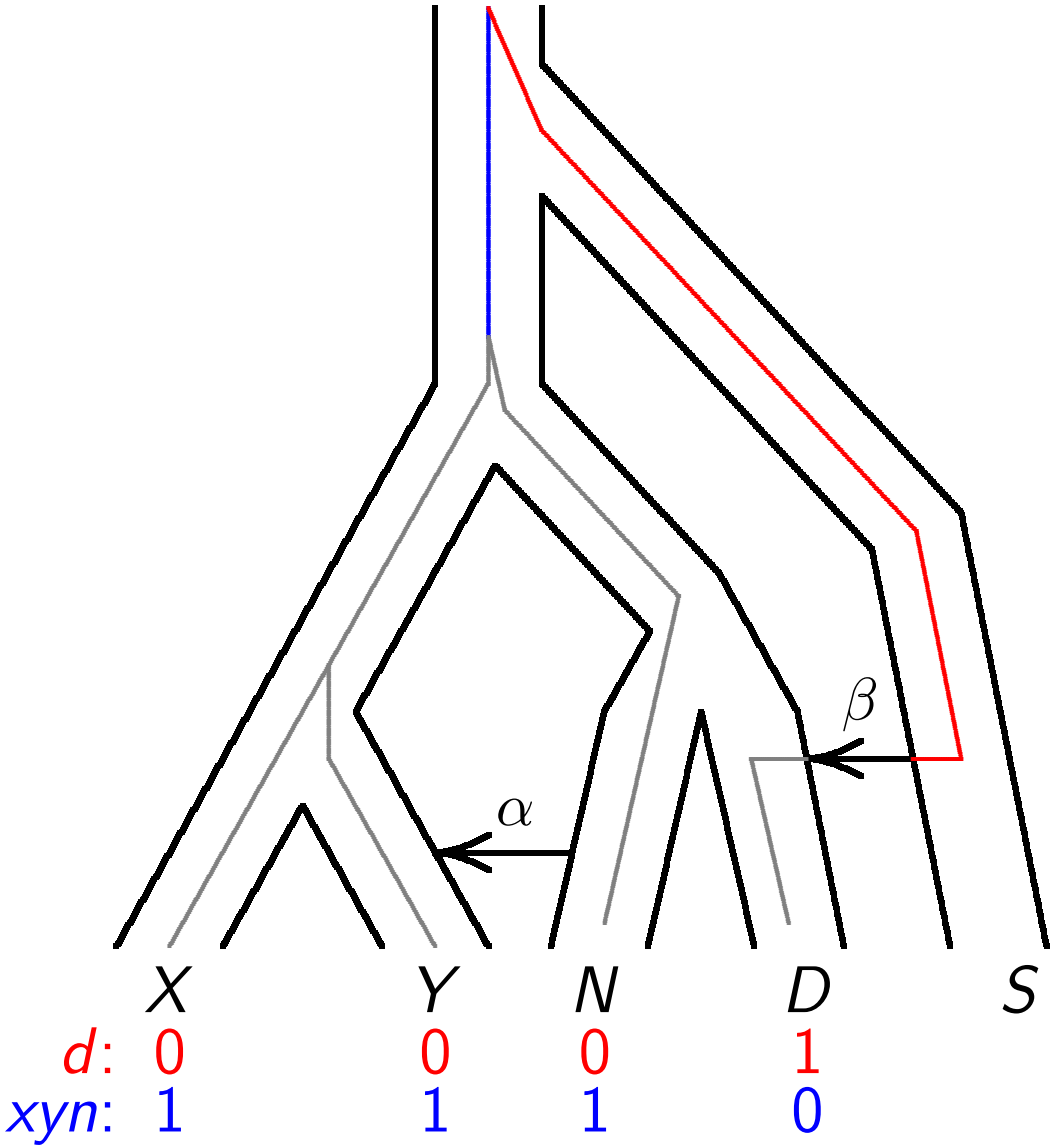
d is most common singleton

Suggests Denisovan fossil is young and *N*-*D* separation old.

But our 2017 analysis of this hypothesis led to absurd result: Denisovan fossil only 4000 y old.

Something was missing from our model

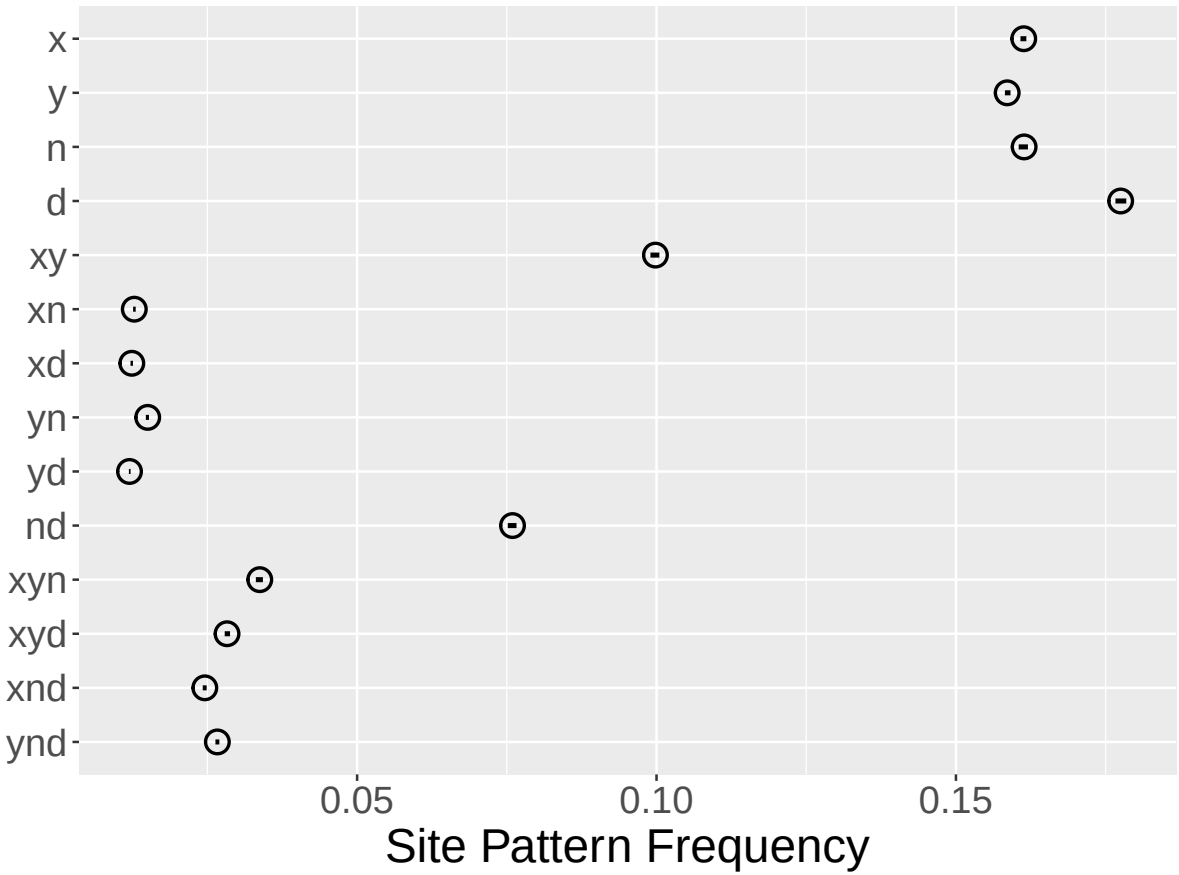
Admixture from superarchaics into Denisovans



S is a "superarchaic" hominin, distantly related to all others.

$S \to D$ gene flow inflates frequency of d and xyn.

Superarchaic gene flow into D



d most common singleton

xyn most common tripleton

Signature of $S \rightarrow D$ gene flow (Prüfer et al 2014).

What we learned, just by staring at the data

1. Europeans and Africans are close relatives.
2. So were Neanderthals and Denisovans.
3. European-African separation more recent than Neanderthal-Denisovan.
4. Neanderthals contributed genes to Europeans
5. Superarchaics contributed genes to Denisovans.

This analysis has been exploratory. Legofit extends these ideas to estimate parameters and test hypotheses about history.