Using Simulations To Examine Temporal Discrepancy Between the First Appearance Times of Ancestors and Descendants

David Bapst

July 16, 2019

Du and Alemseged 2019: Temporal Evidence and Ancestor-Descendant Relationships

In an incomplete fossil record, the order of appearance in the fossil record may be a poor reflection of phylogeny. Additional complications arise when morphologically-delimited taxa are persistently found over geological durations, including when they 'persist through' a speciation event that produces a differentiated descendant morphotaxon. In such cases, the 'ancestral' morphotaxa will be extant at the same point in time (but perhaps not the same location) as their 'descendant' morphotaxa, and thus the incomplete record may preserve traces of the 'descendant' before its own 'ancestor'. But how often should we expect to see this in the real fossil record? How unlikely is a an age 'discrepancy' (as I will term it) of a given size?

Du and Alemseged (2019, hereafter D&A) attempt to answer this question, with specific reference to whether the *A. sediba* morphotaxon could possibly be the ancestor to *Homo*, whose first supposed fossil occurrences appear to post-date its own occurrence by a considerable duration (for the hominin fossil record). Their finding is that they overall find the observed ancestor-descendant discrepancy to be of an unlikely duration. This document is an investigation of that claim. I'm not an anthropologist, and I hold no opinion whatsoever to the actual data, so I am going to largely accept their statements about dates and taxon identity, and mainly interrogate their methods.

D&A present two lines of arguments to evaluate the supposed ancestor-descendant hypothesis, given the temporal discrepancy:

a) A model describing the probability of a discrepancy of a given magnitude, under a certain set of assumptions. (Described more below.) This model finds a very low probability of a discrepancy as large as the observed discrepancy, and consistently finds such under a range of conditions.

b) An empirical dataset of the geologic age difference between the 'first found' fossils for 28 previously-hypothesized ancestor-descendant pairs from the paleoanthropology literature. In other words, if each species was only ever a single find - the first reported collection of that find with no subsequent collections- what would the age difference be between those two species? Only one A-D pair has a discrepancy (with the descendant's first-find being geologically dated older than the first-find of the descendant), and that discrepancy is very small relative to the gap between *Homo* and *A. sediba*.

I will critique these arguments in reverse order.

The Metadata of Published Reports of Ancestor-Descendant Pairs

My initial concern upon reading the paper was that ancestor-descendant pairs as reported from the literature are often a limited subset of what might be true. As D&A describe in their discussion, this might particularly impact their conclusion because temporal information is (nearly always) taken into account when someone decides to suggest something is ancestral to something else. Thus, they use the first-finds of each taxon

instead, based on the idea that the hypothesized ancestor-descendant relationships might be based on later, roughly independent temporal information.

I rather like the attempt using first-finds to avoid the issue of a taxon's whole temporal range being non-independent of it being placed as an ancestor or descendant, but I'm not sure its enough.

First, does first-finds remove the bias? I am not entirely convinced, because when people make arguments for ancestor-descendant pairs, I think that's more likely to take into account temporal information from their earliest collections, not their latest collections. I mean, in the invert paleo world, newly published finds often change a taxon's temporal range - but that information might take decades for someone else to notice and take that into account in their interpretations. Thus, the first find and the original geochronological interpretation of that find is most influential. I would probably like to know (a) what's the geologic age difference if we restricted each pair to looking at the most recent collections from each species, and/or (b) what's the distribution of age difference look like if we took all the collections from each taxon in these ancestor-descendant pairs, and used a Monte Carlo approach to sample a single collection at random from each taxon, and look at the distribution in difference in age for all these artificial 'one-collection' taxa. (This is similar to Alroy's 1996 analysis of Cope's Rule, where he sampled subsampled species pairs from genera to examine ancestor-descendant trends in body size change.) A full examination would compare these distributions of randomized age differences, to the age differences in the actual, currently-accepted first appearance times for those taxa.

Second, while I absolutely believe that ancestor-descendant relationships is something we should talk about and consider, but I'm not convinced of the utility of previously published pairs as a baseline for what we should expect of ancestor-descendant pairs at large. For example, in my 2017 paper with Melanie Hopkins on the finely resolved record of late Cambrian pterocephaliid trilobites, we found support for up-to 16 ancestor-descendant pairs, when really only 9 pairs had been proposed previously in the literature (including several pairs where which taxon was the ancestor, and which was the descendant, was uncertain, due to uncertainty in the first appearance dates). Now, no single phylogenetic hypothesis (the method used produces samples containing hundreds of possible trees) contained all 16 pairs, but a single tree often had higher numbers of pairs than had been previously proposed. So, my expectation is that for most groups, the inclination to label specific ancestor-descendant pairs has probably been overly conservative (one exception may be ammonites or planktonic foraminifera, where systematics is still largely very traditional rather than based on quantitative analysis of character matrices, and large numbers of taxa are placed as stages along long ancestor-descendant sequences of anagenesis). Maybe hominids are like those groups, and workers have been less conservative than they should have been, but maybe not, and my expectation is that thus the ancestor-descendant pairs reported for any given group is an extremely conservative subset of those, probably over-representing those that are the most consistent with being ancestor-descendant pairs in terms of their morphological and stratigraphic relationships. Overall, this adds up to an overwhelming bias against proposing ancestors that appear later than their descendants, a bias that probably leads to us underestimating the true occurrence of ancestors occurring latter than their descendants.

Third, I would still have problems with this comparison because this is so unlike the ancestor-descendant pair we are most interested in - *A. sediba* and *Homo*. The first collection of *Homo* would be, **presumably**, us recognizing that we are extant human beings (first fossil collection would be - well, I really have no idea, but its probably pretty geologically recent). Thus, that pair has no discrepancy as evaluated by the first-find approach. One taxon is also a species known from a single very time-constrained set of collections, and the other is a genus, that becomes relatively diverse (encompassing multiple species or sub-species or proto-species or whatever you want to call them. . .). Under a standard birth-death-sampling model, we would expect that such increased lineage diversity makes sampling *Homo* later much more likely later than earlier, even if there is a long 'fuse' interval with a very incomplete record. (Note though that this dynamic, admittedly, strengthens their case for their probabilistic model.) I think really we would need to identify a specific *Homo* species, and talk about that species-pair specifically rather than the genus itself.

Du & Alemseged's Model of Range Overlap and Ancestor-Descendant Temporal Discrepancy

In my data analysis courses, I tell students that models are just sets of assumptions about the world. We may describe those assumptions as mathematical relationships, but really any model is just a bundle of assumptions, and models derive all their explanatory power from the strength and specificity of those assumptions. So what are D&A assumptions? Three of these are explicitly laid out:

1) The two taxa in question are assumed to have true ranges (the time between the true time of origination and extinction) of equal duration. In this case, the two taxa are both assumed to have 0.97 Ma (mega-annum) ranges. That's an average taken from Robinson et al - its obviously not true for *Homo*, which obviously has a longer duration than 1 Ma. It isn't clear from D&A what the choice of 0.97 Ma or inferring equal length ranges has on their findings.

2) The fossil record is assumed to have uniform sampling, across space and time, and across both the ancestor and descendant. As I already indicated above, the fact *Homo* encompasses multiple species would effectively violate this assumption, but the if that effect was taken into account, it would make it even more unlikely for the record to preserve *Homo* earlier in geologic time, and thus the bias works in agreement with D&A's findings. This is a common assumption of sampling models for the fossil record, and while it is often critized (see Steven Holland's work in particular), D&A far and away exceed the typical defense against this criticism by presenting evidence that hominin.

3) Sampling events (collections, occurrences, fossil horizons - whatever you want to call them!) are independent in time and across lineages. This is a common assumption of sampling models for the fossil record, and there aren't many cases where I would argue for the opposite.

Others are not so explicit:

4) This is a direct ancestor-descendant relationship. There are no additional, 'unknown unknown' taxa with similar sampling rates. Often, when we find ancestor-descendant pairs in the fossil record, we are probably not finding direct ancestor-descendant pairs (this morphotaxon is *directly* descended from this other morphotaxon) but rather we might be missing a few completely un-sampled, un-observed morphotaxa in between. As Foote (1996) recognized, ancestor-descendant relationships are probably relatively common in the fossil record, especially indirect ancestor-descendant relationships. Now, recognizing the possibility for 'unseen' lineages would decrease the probability of finding larger A-D discrepancies, so this bias works against D&A's test (and thus strengthens their findings).

From this, they derive the probability that any one occurrence of a descendant would occur before any one occurrence of an ancestor. Interestingly, and surprisingly to me, they find that sampling rate doesn't actually impact this probability - a well-sampled fossil record and poorly-sampled fossil record are expected to have very similar probabilities for a given ancestor-descendant discrepancy, for two taxa that have identical durations, and share the same uniform sampling rate (whatever that rate happens to be).

Some properties of this model are unclear in the draft, such as 'what if taxon range is varied?'. So let's consider scenarios under this model ourselves. In doing this, I will avoid relying on or refering to the supplemental R script provided by D&A, so that I am sure I understand all the relevant details.

This model mainly has three major variables related to intervals of various length: the duration of the range for each taxon (`T_R`, which D&A set to 0.97 Ma), the observed minimum stratigraphic overlap between the two taxa, which is a fixed variable of interest and is essentially the A-D discrepancy, or difference in age between the descendant and the ancestor (`T_d`, which D&A set to 0.8 Ma, the apparent discrepancy between *A. sediba* and *Homo*), and the true, total amount of overlap between the ancestor and the descendant (both of which, note well, have equally long ranges), and is treated as a random variable in D&A's model (`T_o`).

Several additional variables are worth noting. With respect to D&A's general notation and figures, the age of the ancestor's horizon is `H_A`, the age of the descendant's horizon is `H_D`, and the offset between an descendant's true time of origination and its time of sampling is `X_D`, while `X_A` is designated as the offset between the descendant's and ancestor's horizons, minus the minimum discrepancy (`T_d`). `X_A` and `X_D`

are thus treated as random variables in this model, with minimums at 0. For the descendant's horizon to post-date the ancestor's horizon by a gap as large as `T_d`, and for all other assumptions to hold, `T_R > T_o > T_d`, and thus putting bounds on `T_R` and `T_d` defines the range of possible values for `T_o`. Additionally, `X_A + X_D = T_o - T_d`, `H_A - H_D > T_o` and `X_D <= X_A` (D&A state greater than, but I do not comprehend any reason for why they cannot be equal and for all other conditions to still hold).

Equation 1 describes the probability of one fossil being found in a horizon within the region of overlap defined by `T_o - T_d`, and equation 2 describes the joint probability for two fossils (one for the ancestor, one for the descendant) from this limited region (referred to as `Pr(endA_endD)`). This is used in equation 3 (which is derived through a number of steps, during which the sampling rate drops out) to find the probability of `X_A > X_D`. By alternatively setting `X_D` or `X_A` equal to a random variable `t` and iterating over that variable, D&A determne that the probability of `X_A > X_D` is 1/2, regardless of whether `X_A` or `X_D`is set equal to `t`. This leads them to state in equation 4 that 1/2 is also the probability of `H_A - H_D > T_d`, when conditioned on only sampling horizons from the region of 'excess' overlap (D&A continually refer to this 'excess overlap' interval, or pair of intervals, as 'the black region(s)', in reference to their figure 2). This conditioned probability as `Pr(H_A - H_D > T_d | endA_endD)`.

In equation 5, they use the joint probability of `endA_endD`, from equation 2, to remove the conditioning from `Pr(H_A - H_D > T_d | endA_endD)`, and thus obtain the probability of `H_A - H_D > T_d`. This probability is discontinuous, with the probability always being zero when `T_d > T_o`.

For example, when `{r} T_R = 0.97` and `{r} T_d = 0.8` (as in D&A's analyses), and (to take a specific example), `{r} T_o = 0.9`, then the probability of `H_A - H_D > T_d` is:

```
((T_o - T_d)^2) / (2*(T_R^2))
```

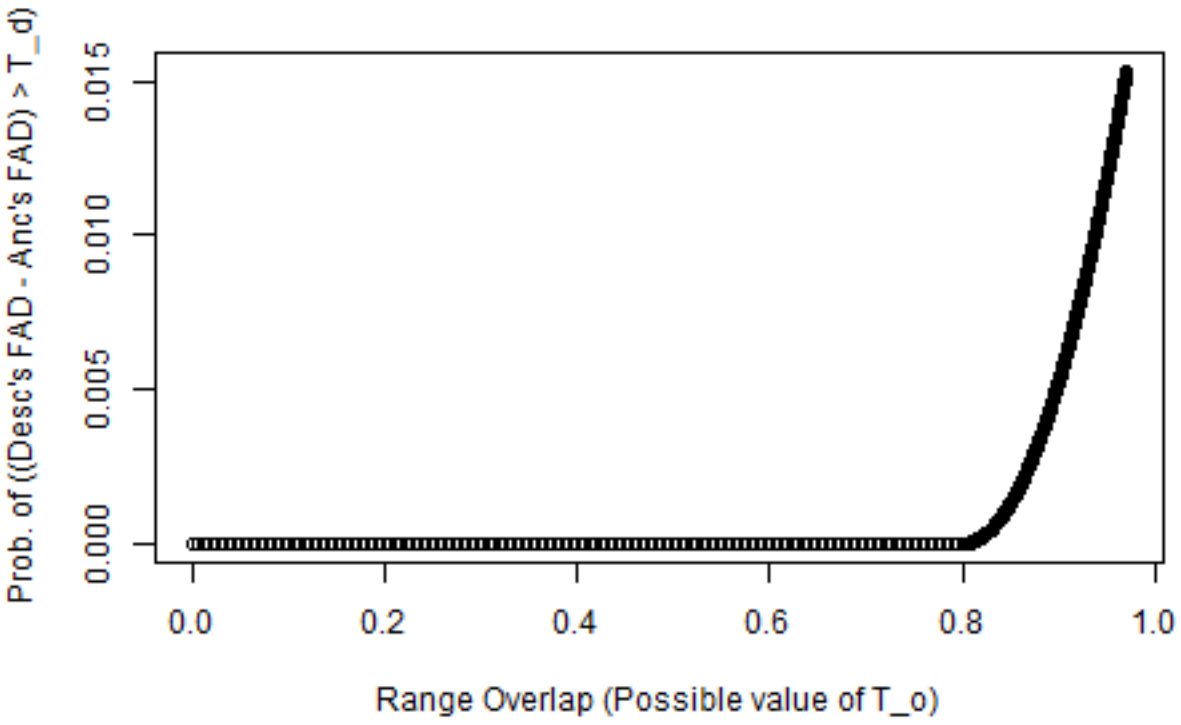
```
## [1] 0.04782655
```

Or, if `{r} T_R = 1`, `{r} T_d = 0.1`, and `{r} T_o = 0.5`, then the probability of `H_A - H_D > T_d` is:

```
((T_o - T_d)^2) / (2*(T_R^2))
```

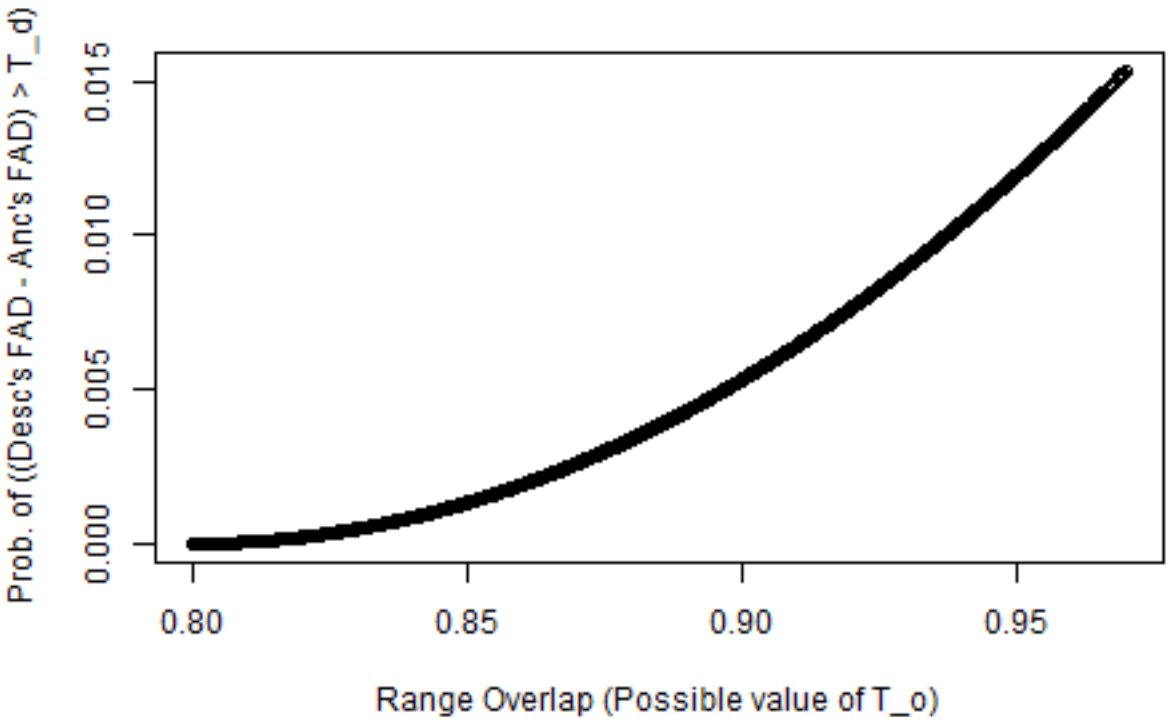
```
## [1] 0.04782655
```

We can also use this equation to recreate their figure 3:

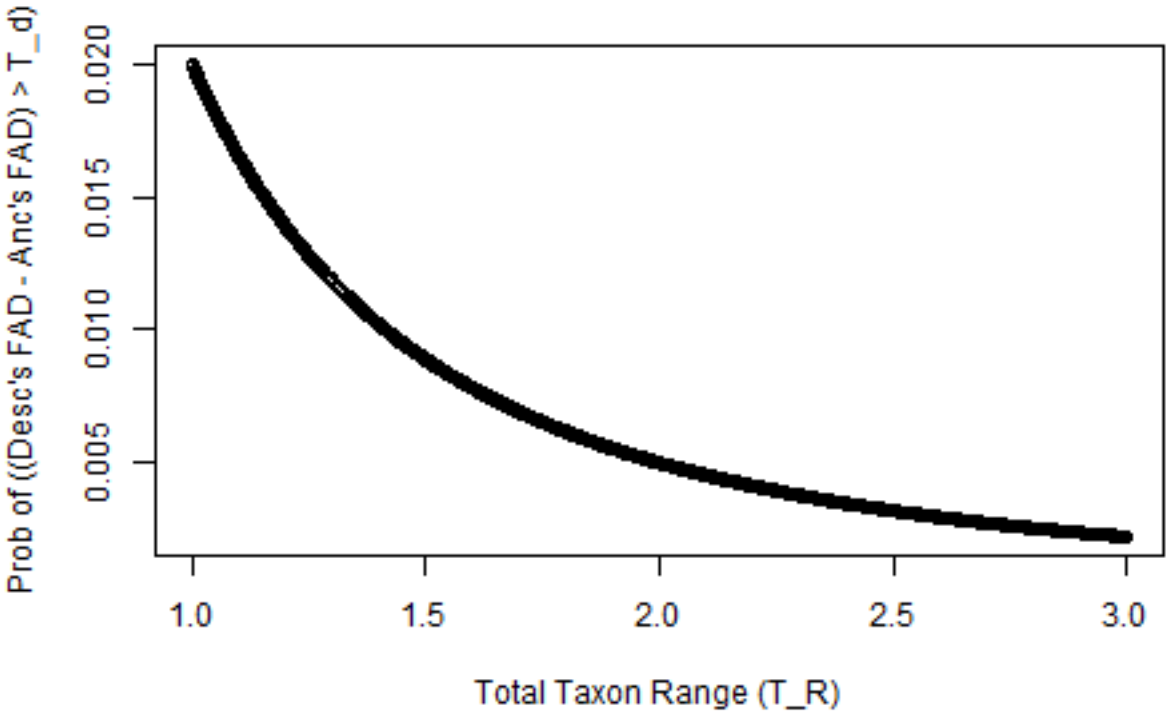


That was almost too easy. But note that most of this graph (just like D&A's Figure 3 - look at their horizontal axes) is at 0 probability, as the values of `T_o` include the range of 0 Ma to 0.8 Ma, even though `T_d = 0.8` and thus its impossible for an ancestor to occur after its descendant. Its curious why this non-informative interval with *a priori* zero probability is even shown - perhaps because they want to show the interval of range overlap from 0 Ma to 0.97 Ma because they integrate over this entire range for `T_o` when they generate their 'treating all values of overlap as equally likely' p-value using equation 6, which they report as 0.0009.

Regardless, for this examination, let's look at just the informative region that encompasses overlap values with non-zero probabilities for a gap of 0.8 Ma.



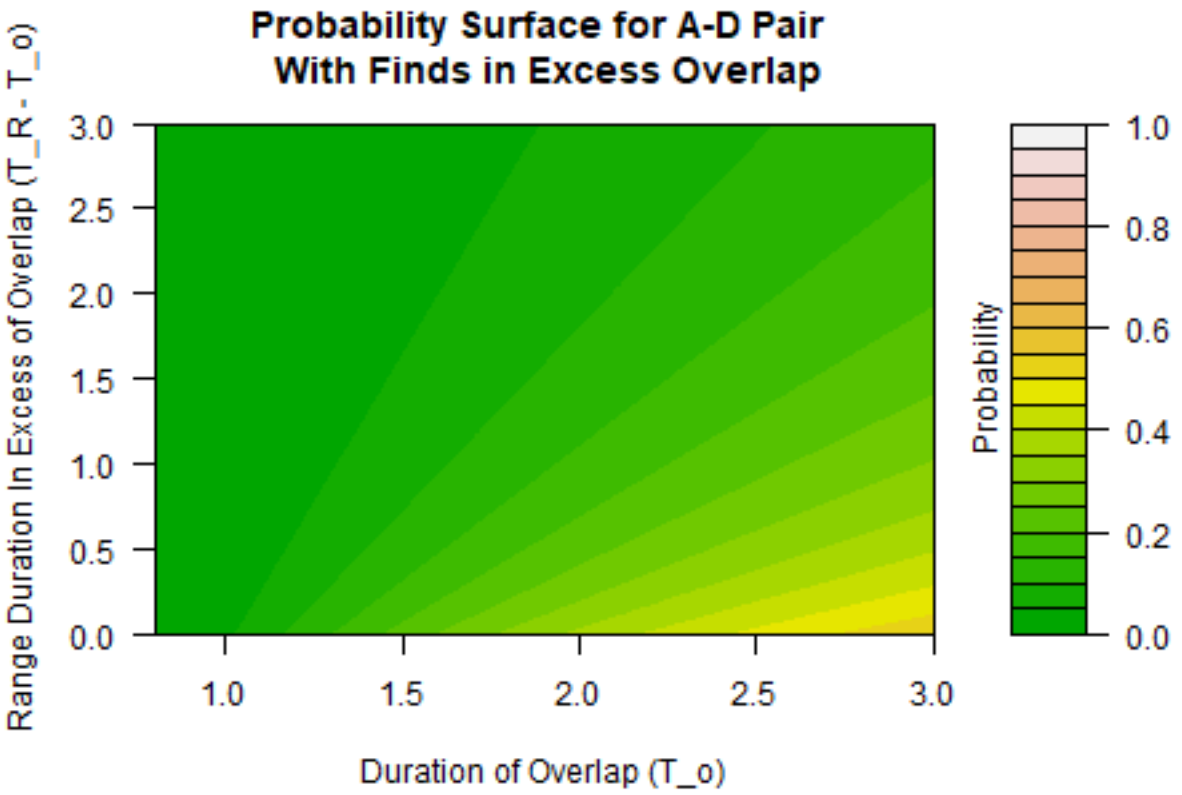
Now we can do things like allowing `T_R` to vary, instead of `T_o` - what if we assume the true overlap between *Homo* and *A. sediba* is just slightly more than the observed minimum overlap (`T_o = 1`, which would mean that *Homo* originated from *A. sediba* at most 20 Ka before *Homo* was first sampled in the fossil record), but we don't know how long the individual ranges of each taxon is, but otherwise put a hard cap on their maximum ranges at 3 million years?



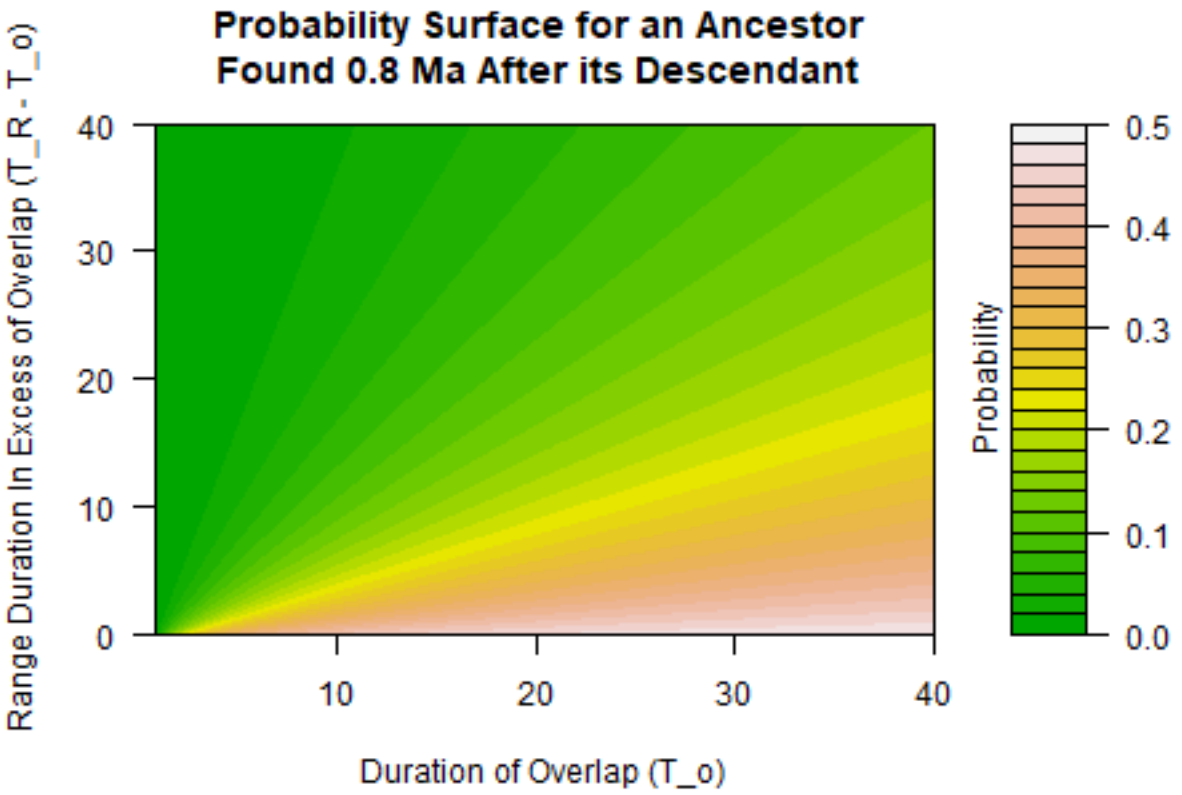
We can see that giving larger ranges relative to the overlap doesn't at all help the case for descendants to occur before their ancestors, because it means more geologic time during which only the ancestor existed, but wasn't sampled.

So, in figure 3 (or its recreation above), where only `T_o` is varied, we can see that the probability never gets better than 0.015 (1.5%), which leads us to wonder - what set of parameters would produce a high probability for the descendant to predate the ancestor? Presumably the hard cap on this probability is 0.5; this is theoretically true (in what scenario of uniform sampling could an ancestor be less likely to be found first than its descendant?), but also mathematically true as `Pr(H_A - H_D > T_d | endA_endD)` is 1/2, and thus the actual determining part of the model is the probability of fossil finds being in the 'end-caps'excess overlap' of the taxon ranges, given by equation 2 (`Pr(endA_endD)`).

So when is `Pr(endA_endD)` closest to 1? Let's hold `T_d = 0.8` constant, but vary `T_o` and `T_R`, and look at the values of `Pr(endA_endD)` we obtain. Because `T_R > T_o > T_d`, we will vary `T_o` from the value of `T_d` to 3 Ma (extremely unrealistic for our case study, as it probably would require *A. sediba* to still be extant, but this is just an exploration - calm down!), and vary the amount of time in the fossil taxon ranges that is not part of their overlap (`T_R_no`, where `T_R - T_o = T_R_no`), with `T_R_no` ranging from 0 to (again) a maximum 3 Ma (which isn't just implausible, but patently impossible in our case as the fossil record in quest isn't from six million years ago. . .).



And now we can see that the probability of sampling from those regions of excess overlap is only going to be relatively high when (a) the ranges of the ancestor and the descendant are nearly perfectly overlapping (`T_R_no ~ 0`, or `T_o ~ T_R`), and (b) when the ranges are long relative to the discrepancy between the descendant and the ancestor. And, let's expand those maxima to 80 million years just to illustrate the point (`T_d * 5oh0 = 40`) and look at a very similar plot but for the probability of observing a descendant predating its ancestor by a discrepancy as long as `T_d`, `Pr(H_A - H_D > T_d)`. We'll rescale the color scale (probability) to a range of 0 to 0.5, as we know that 0.5 is the maximum limit in this case.



Thus, under the model presented by D&A and its set of assumptions, we can see that ancestors post-dating their descendants by even what might be small geological intervals (e.g. less than a million years) is only *likely* when geologic durations are long relative to the length of the discrepancy in the ages of appearance (at least 3-5 times T_d, the discrepancy), and the descendant's range overlaps for about a quarter of its range with its ancestor.

Now, there are hundreds, or possibly thousands, of putative ancestor-descendant relationships that exist in the paleontological literature. This agrees with our theoretical understanding of diversification and the incompleteness of the fossil record, which suggests that there should sampled ancestors, and that their number may not be great but should not be negligible (Foote, 1996). However, D&A's model implies we should rarely expect ancestors to occur after their descendants. But, are these reasonable assumptions? What does a realistic birth-death-sampling simulation tell us?

Simulation Tests

Simulation Structure, Parameters, Conditions

Here, I will attempt to compare D&A's model to a birth-death-sampling simulation, obtain 10,000 simulated fossil records that meet certain conditions (i.e. having more than one taxon), and parameterize these simulation to closely as possible match the parameters D&A worked with in their study. A birth-death-sampling model is simply a model of lineages, with events occurring as if a series of independent Poisson processes, with morphotaxa 'budding' off from each other at some rate (the rate of origination, or the *birth rate*), going extinct at some rate (the rate of extinction, or *death rate*), and being sampled at particular instances in time (the *sampling rate*). These processes are presumed to be constant with uniform rates, that do not vary over time. That's almost certainly false when humanity involved, but let's just ignore that for the sake of doing the

simulations at all. A birth-death-sampling model is also known as a BDS model, a birth-death-serial-sampling model, or (most commonly) a fossilized-birth-death model (Heath et al., 2014), and is exactly the sort of model now becoming increasingly and widely applied to paleontological datasets using tip-dating approaches in Bayesian phylogenetics.

In these simulations, we'll focus on the differences in first appearance times, even though D&A's model is explicitly for taxa known from single horizons. While *A. sediba* might be known from a single horizon, that certainly isn't true for *Homo*. In my view, the question that D&A pose is about the discrepancy in first appearance times between an ancestor and its descendants. As touched briefly when discussing the single-horizon assumptions of D&A's model, though, comparing data with multiple occurrences to D&A's model will likely only favor their conclusions (because taxa with multiple sampling events are less likely to have mismatch in the order of appearances of ancestor-descendant pairs).

Handily, the extinction rate is widely known to be the inverse of the mean *true* species duration. Note the 'true' - the observed species durations will always be smaller than the true ranges, due to incomplete sampling at both ends (and thus in better sampled records, more sampling means ranges that look more like the 'true' ranges.) However, we'll ignore that effect of sampling and assume extinction rates is `1/0.97`, the inverse of the mean species duration used by D&A. (This literally means that if we had a hundred species at time A, we'd expect about half of them to be extinct 0.97 Ma later - so this is a **very** high extinction rate! The average for marine invertebrates is 0.1 per lineage, per Ma - implying an average extinction rate of 10 Ma.) Even more handily, we know that for most groups in the fossil record, the long-term origination and extinction rates are generally nearly equal (Stanley, 1976; Sepkoski, 1998; Marshall, 2016). We don't really understand why that is (its partly a mathematical artifact, but also partly *not*, and doesn't agree with estimates of speciation and extinction rates from molecular phylogenies), but we can probably safely assume that the origination rate for hominins has, in the long term, been pretty similar to the rate of extinction, given that only ones species has managed to survive to today. A per-lineage sampling rate isn't really touched upon by D&A, as the parameter does not matter to their calculations, and I would hesitate to try to estimate such myself, so we can set that also equal to the extinction rate for now.

Birth-death-sampling simulations are notorious for having wildly variable results (e.g. most that start with a single species go extinct immediately), and thus we need to set constraints on what simulation output we'll accept as a viable run or not (non-viable runs are thrown out, and the simulation is run again, until the required number of acceptable runs is produced). We will require at least 10 taxa sampled in each simulated fossil record (but no more than a 100), and at least 1 extant taxon in each output run. Most runs will likely have dozens of sampled species, even with these seemingly low limits.

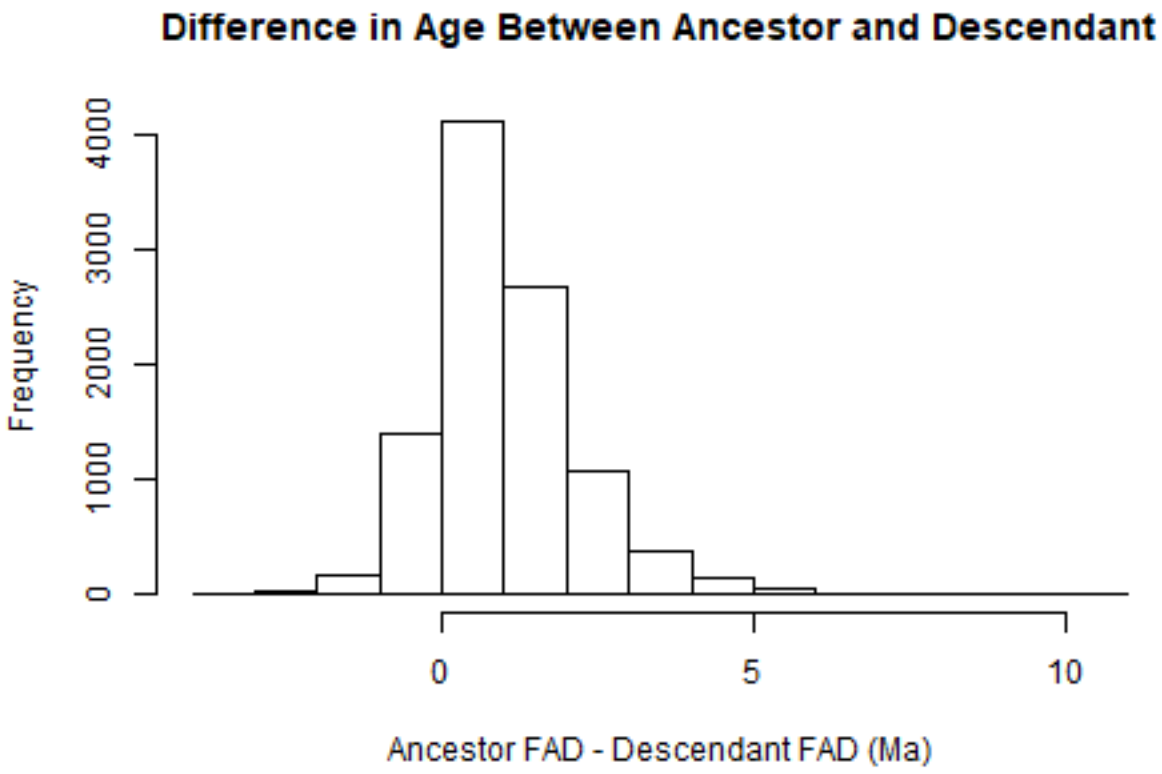
We will also not treat simply surviving to time 0 as 'being sampled' - that way, when we look at differences in sampling times, its always fossil finds, modeled as under a uniform Poisson process - exactly as expected by D&A's own model.

Variables of Interest from the Simulations

Once we run the simulations as described above, we can explore our 10,000 simulated fossil records. In our case, there are two central variables of interest that we can look at as summary statistics, and examine the relationships between them, to evaluate how well D&A's model holds up in these more realistic (yet still simulated) scenarios.

The first of these variables of interest is The difference in age between the first sampled appearance of the ancestor and the first sampled appearance of the descendant, for every sampled pair of ancestor-descendant taxa. For all those pairs in which descendants predate their ancestor, there will be a negative temporal offset, which is the temporal difference between descendant and ancestor labeled `T_d` by D&A. Note that this is not limited to only direct ancestor-descendants, but includes ancestor-descendants where there are unsampled intermediates as well. In order to properly avoid statistical artifacts due to accidentally including data from two seperate pairs involving the same ancestor, or artifacts brought on by ignoring taxa with more than one descendant, only a single such ancestor-descendant pair was sampled from each tree. This sampling was doing at random from all possible ancestor-descendant trees in a given simulated fossil record.

The distribution of age offsets looks, in same ways, just aswhat we would expect: dominated by cases of the ancestor preceding the descendant (positive age offset values):



Note however that a significant portion are less than zero, and thus represent ancestor-descandant discrepancy - cases where the ancestor's first appearance is younger than the descendant. How many, in fact:

```
sum(AD_diff_all>0)/length(AD_diff_all)
```

```
## [1] 0.8414
```

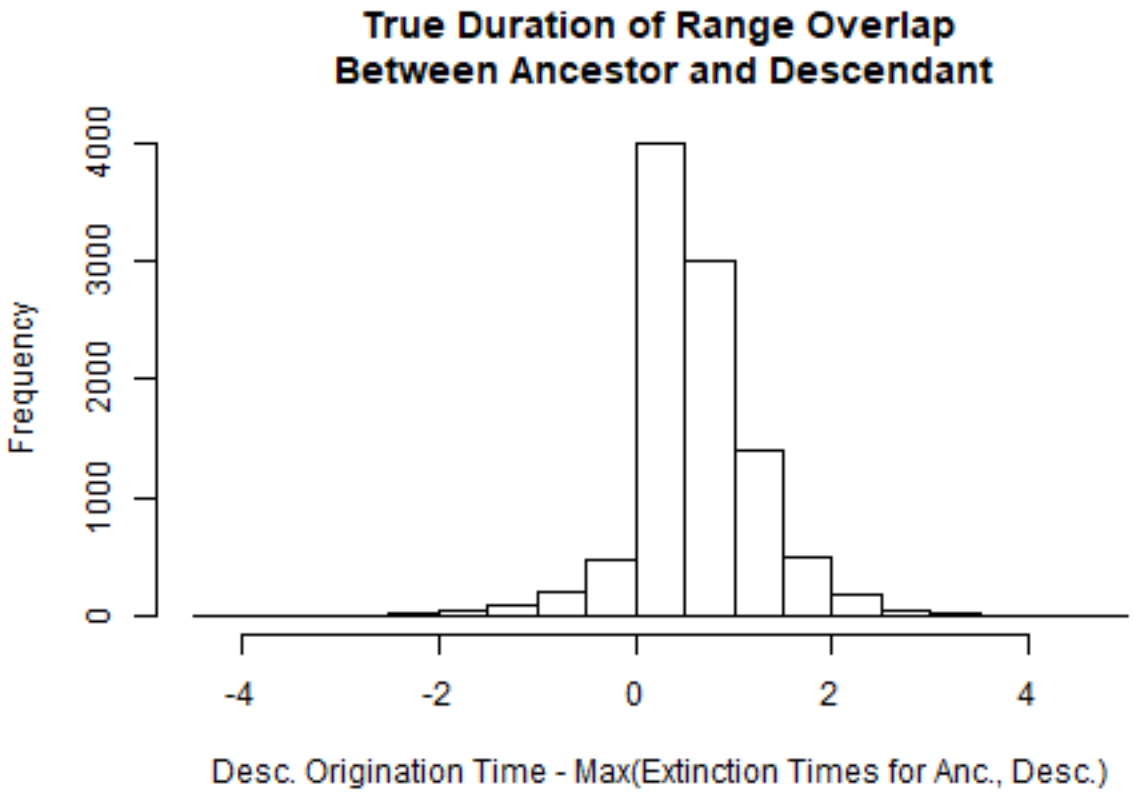
One thought that might occur is that the there might be different patterns for ancestor-descendant relationships where the ancestor or descendant are still extant. (The mathematical rules of entirely extinct taxa in birth-death-sampling models are actually usually a lot more tractable, as extant taxa forces us to consider the wonderfully complicated world of censored and truncated exponential distributions). We can record that information and look at the proportion of our returned offsets that came from those that either had extant members, as opposed to being entirely extinct:

```
sum(isExtant_all)/length(AD_diff_all)
```

```
## [1] 0.2441
```

As we go forward, we will make sure to compare the entire simulation output with what we obtain from the all-extinct and still-extant sets, to make sure there is no divergence from the whole dataset.

The second variable of interest is the amount of true overlap between the complete durations of the ancestor and the descendant (`T_o` in D&A's model), for each of the randomly-selected observed ancestor-descendant pairs we obtained temporal offsets from above. This is quite easy to calculate - note that in many cases the amount of overlap is small (and in the case of indirect ancestor-descendant relationships, may be non-existant, leading to the seemingly negative overlap values, as the two sampled taxa never were co-extant).

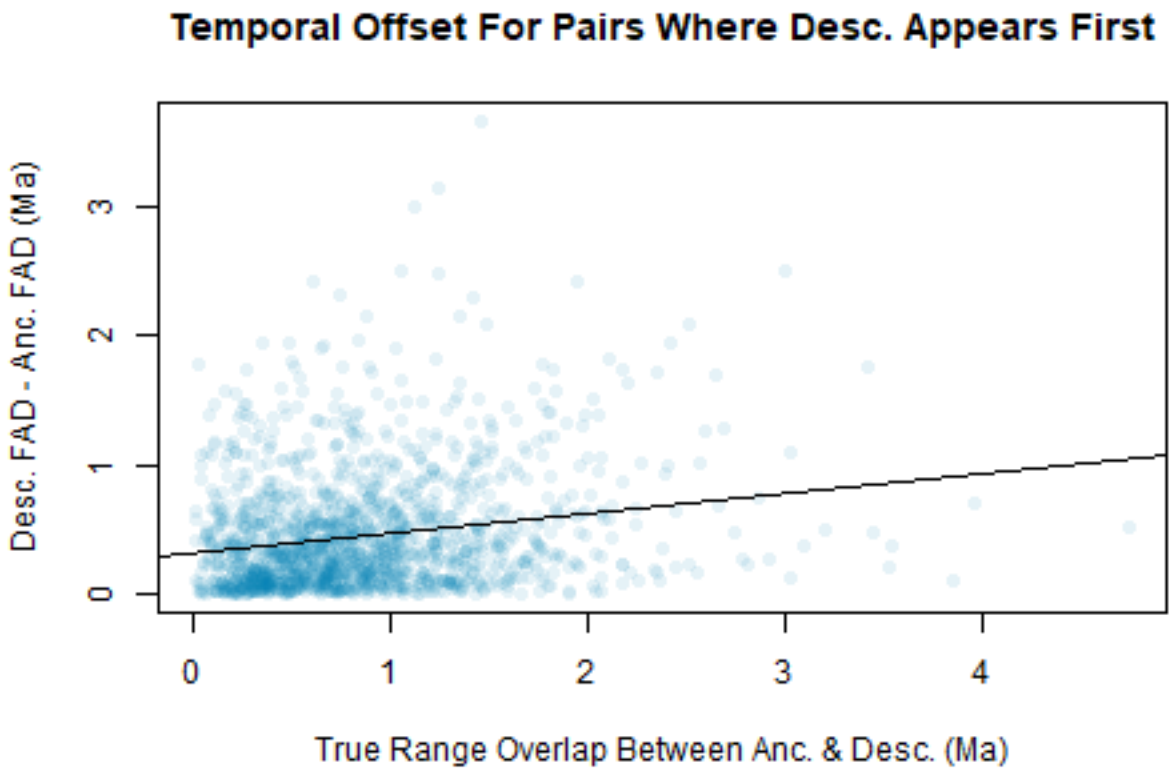


As with the temporal offset values, we may find that pairs with extant members or not might differ.

The Relationship between Anc-Desc Overlap and Temporal Offset in Birth-Death-Fossil Simulations

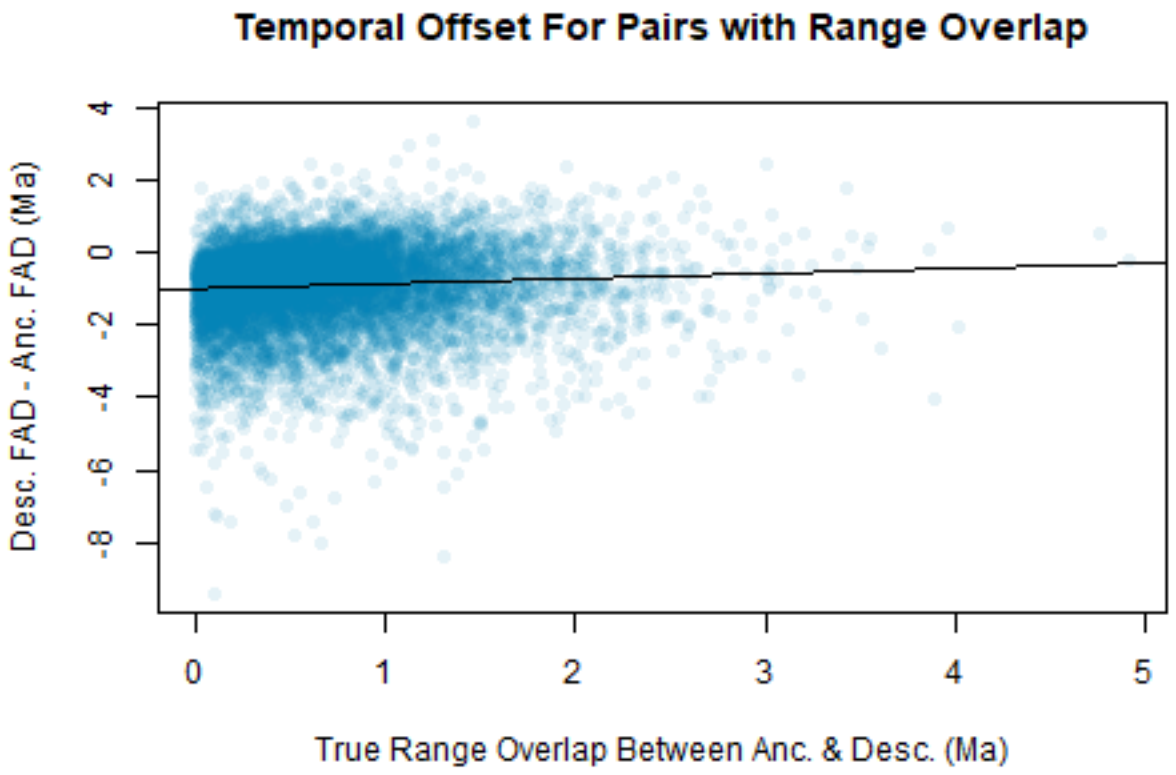
As discussed above, one of the central expectations of D&A's model is that the magnitude of ancestor-descendant age discrepancy will depend greatly on the amount of true overlap between ancestors and descendants. In either case, we will need to filter the data. For example, we could filter on those pairs that have discrepancies (the pairs where the descendant is sampled first):

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'lty' will be disregarded
```



The dashed line here represents the trend line of a linear regression (not that the relationship between these two variables is very linear). Note that for offset, we are here looking at the age difference between the descendant FAD and the ancestor FAD, so that the ancestor

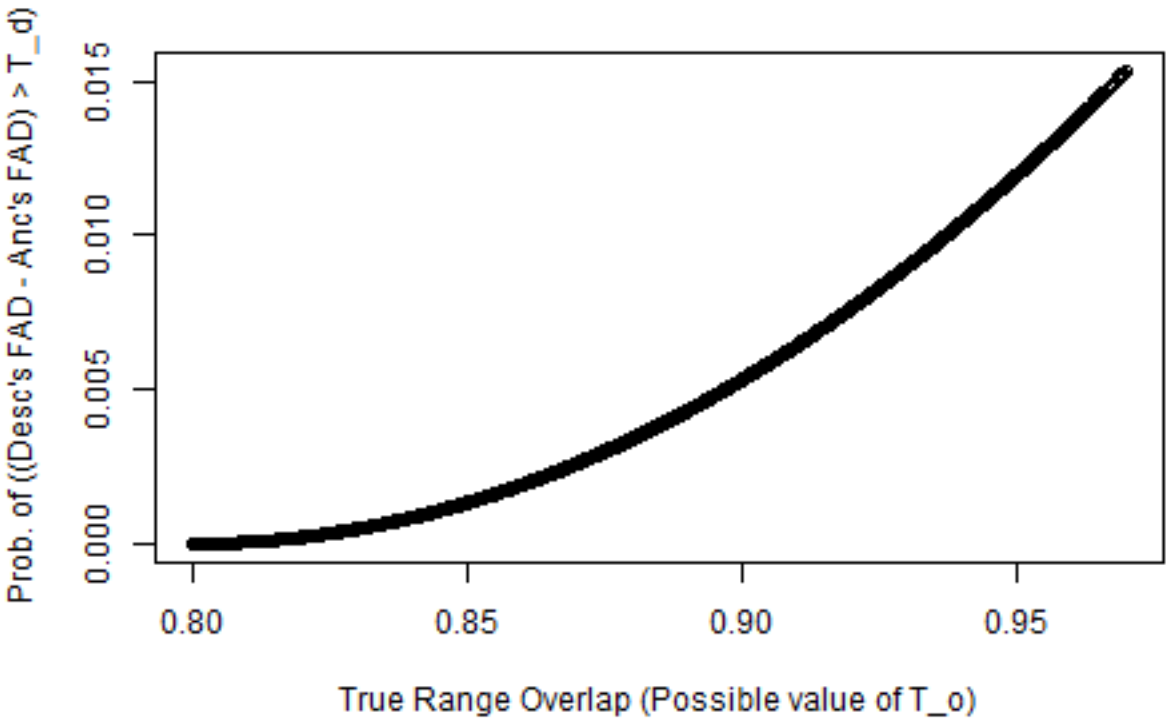
Or, we could examine the temporal offset for the pairs (again, negated) where there is actual overlap between the ancestor and the descendant:



In both cases, although the relationships are very noisy, we can see that increased overlap leads to larger discrepancies between the ancestor and the descendant. But we'd loosely always expect that relationship from first principles.

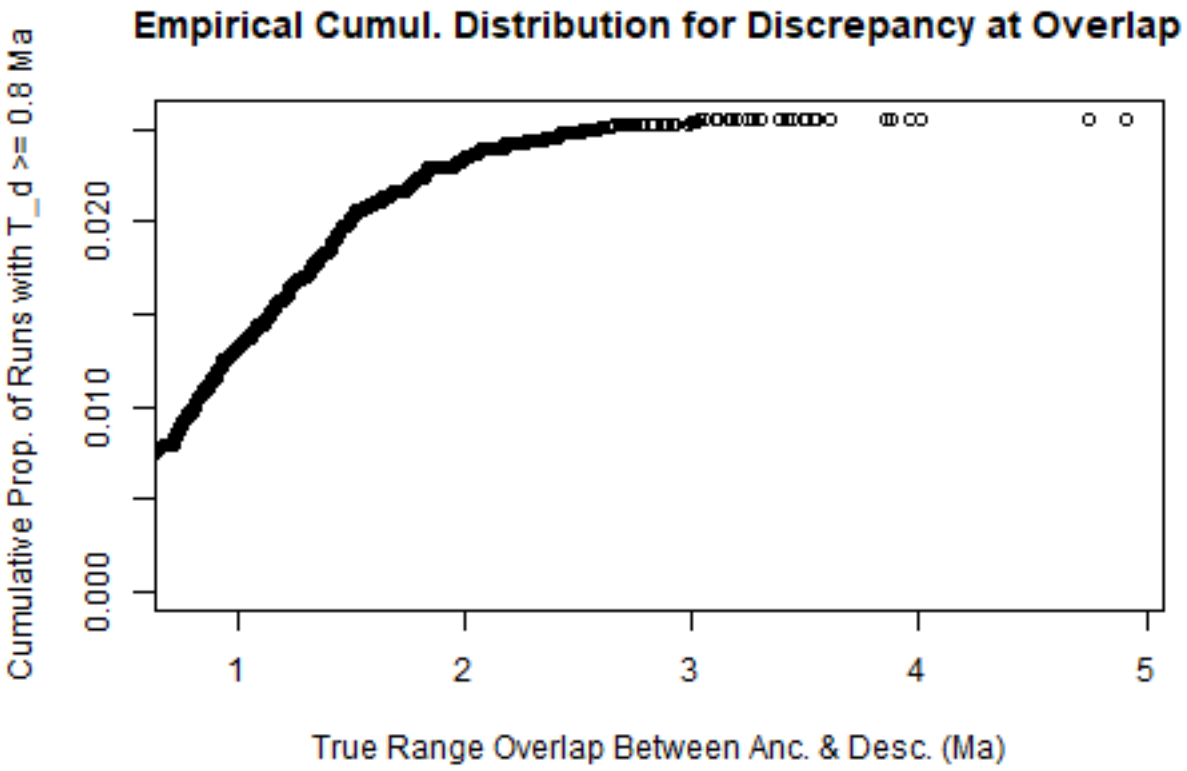
Comparing Simulation Results to Du and Alemseged's Expected Probabilities

So, how much does this match what D&A's model tells us to expect? For example, remember D&A's figure 3 that we recreated, using their equation 5b:



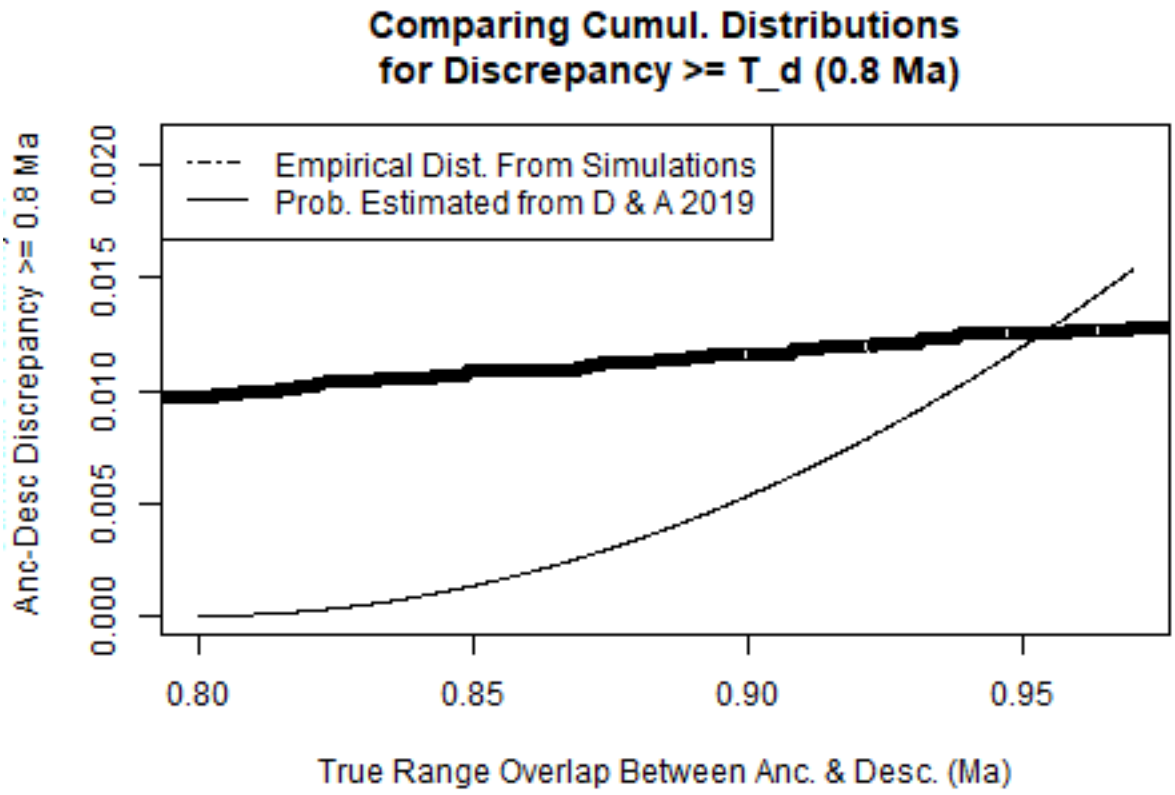
If we simply take the value of `T_d = 0.8 Ma` used to generate this figure, we can look at our simulated datasets, and for different values of ancestor-descendant range overlap (`T_o`), we can ask how many simulations have a discrepancy of 0.8 Ma or more.

(NOTE: The following probability is thus cumulative across range overlap, representing the joint probability that (a) a pair has an overlap this large or larger *and* (b) that the pair has a discrepancy greater than 0.8 Ma. That is my interpretation of what equation 5 in D&A represents. It occurs to me that it might instead by a density, given how D&A integrate over overlap in equation 6 - but then other aspects of their derivation don't make sense to me. Regardless, if I'm wrong about my interpretation of equation 5, just skip down to the section **So just how likely is a discrepancy of 0.8 Ma?** and ignore everything in between. None of this will impact my final assessment anyway!)

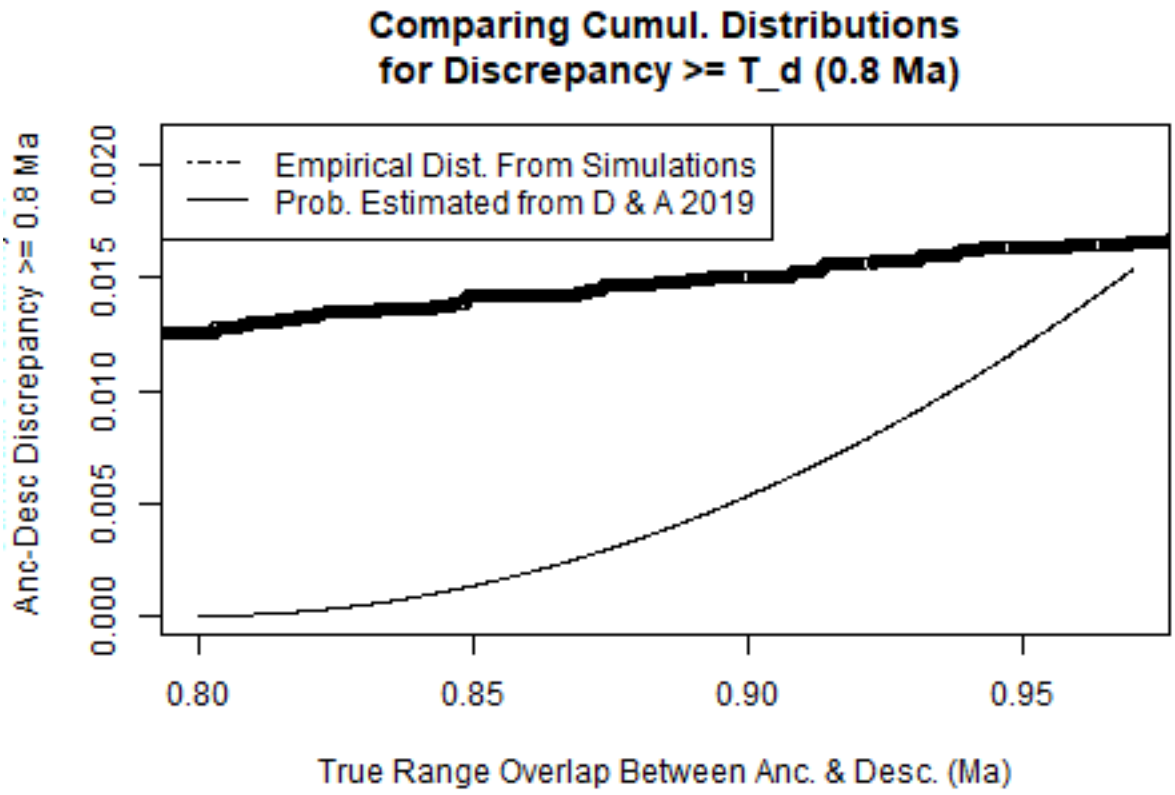


The above is plotted such that we only see the range overlap values greater than `T_d` (and thus that range where a discrepancy of 0.8 Ma or greater is even possible.)

Note that the difference in the horizontal axes on this plot and the previous plot - as D&A's model only considers ranges of 0.97 Ma, the maximum overlap cannot exceed that value. D&A's model doesn't even take into account the possibility for overlaps as great as 2 Ma, or the discrepancies possible at that size of overlap. Thus, to compare the two, it may be more appropriate to restrict our observations to the range of overlaps between 0.8 Ma and 0.97 Ma, and then superimposing our empirically-derived cumulative distribution of observing a discrepancy greater than `T_d` with the probability estimates from D&A within that range.



However, because discrepancies cannot be larger than overlaps, the simulated distribution at a particular overlap value (e.g. 0.9 Ma) may be underestimated relative to D&A's model expectations, as the cumulative distribution is the proportion of simulation runs showing a value equal to or greater than the value of interest, divided by the number of simulation runs. The numerator is unaffected within our restricted range of overlap values, because discrepancies cannot be larger than overlaps, the cumulative distribution estimated from simulations at a particular overlap value (e.g. 0.9 Ma) doesn't take into account discrepancies larger than 0.97 Ma within the plotted range. However, the denominator for that cumulative probability considers many scenarios where overlap is much greater than 0.97 Ma. Obviously, we have relaxed many assumptions from D&A's model in building this simulation comparison, and many of the simulation runs are not possible under D&A's very specific model assumptions. However, we can at least say that D&A's model is not considering scenarios with overlap that high, and thus remove that proportion from the denominator, and see how much that impacts our comparison.

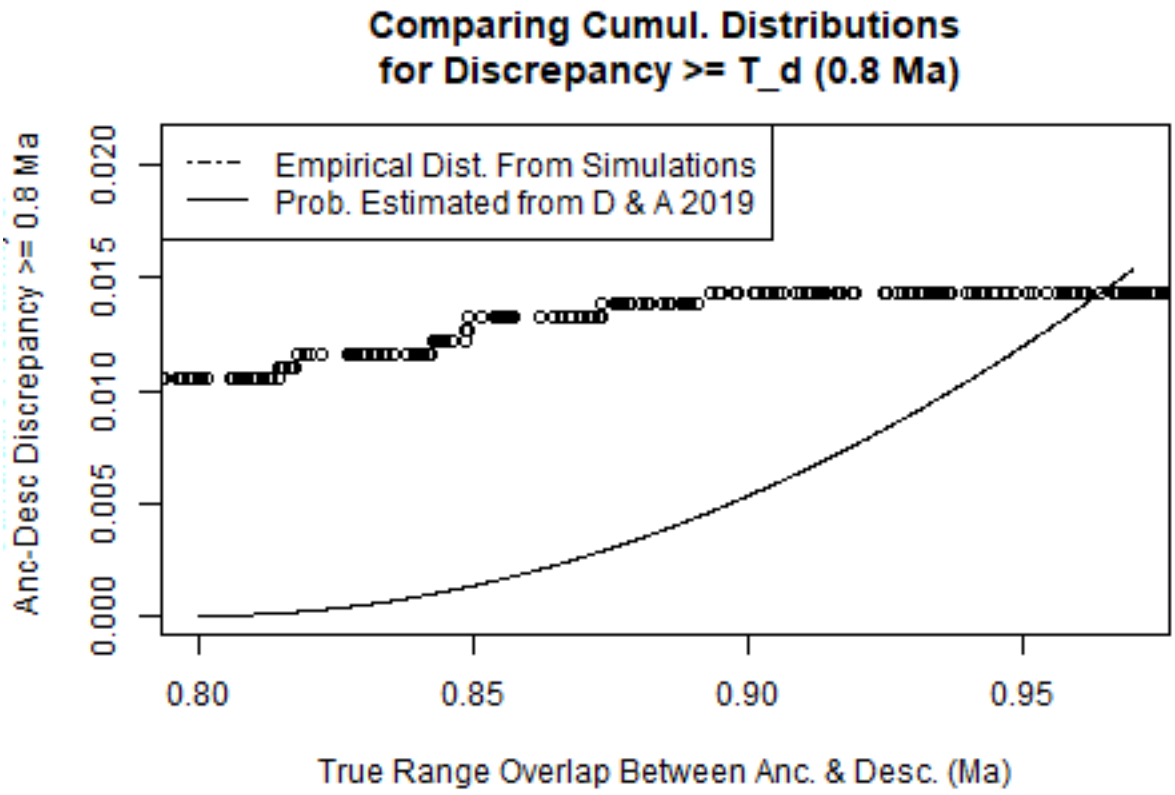


This did not really have a very sizeable effect, as the number of ancestor-descendant pairs with overlaps that large a small minority of our simulation results.

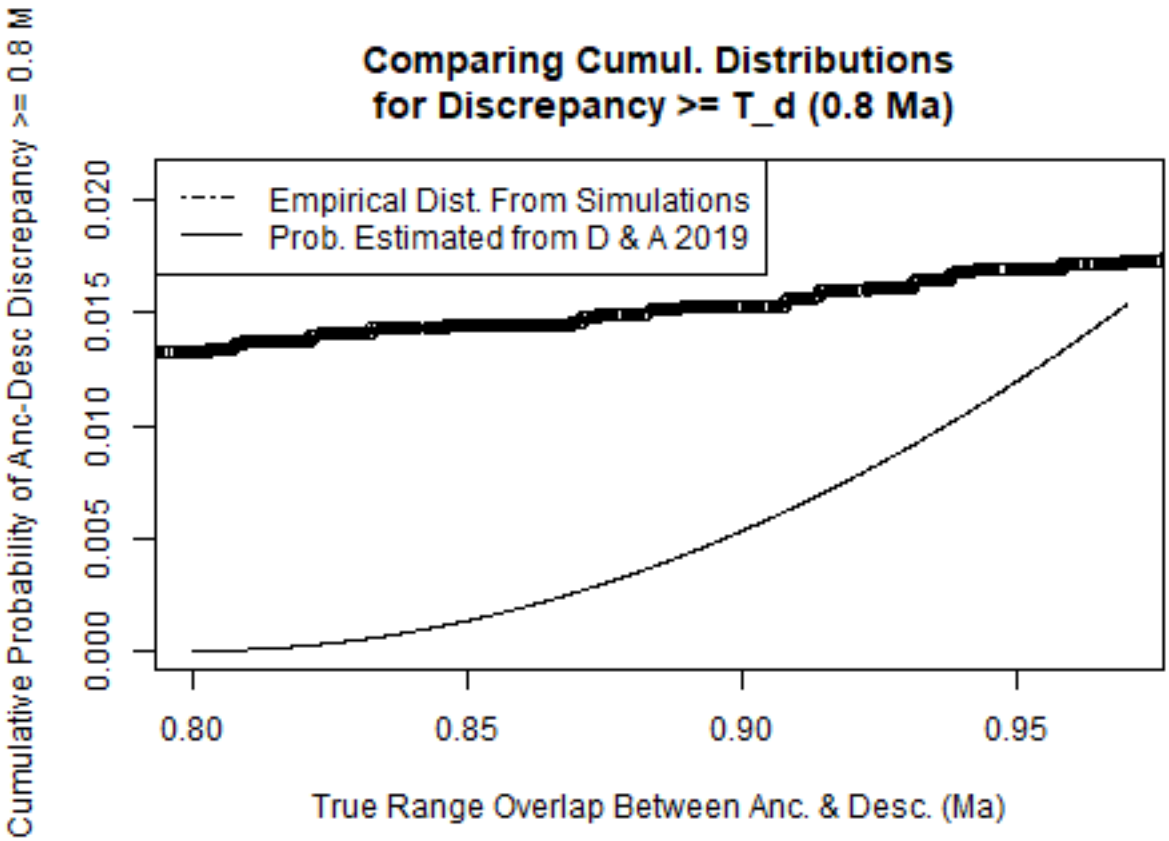
Overall, from the last two superimposed plots, we can say that while the two have *extremely* different shapes. This may owe to the fact we are considering both direct and indirect ancestor-descendant pairs, or possibly having something to do with our relaxation of D&A's strict assumption of descendant and ancestors having similar taxon durations in the much more realistic assumptions of our siulations. However, the values are remarkably similar. Overall, one has to assume that even if some of D&A's assumptions are strenuous, the probabilities they obtain are very close to what would be expected from this particular system.

Does it differ if we condition on extant taxa, or on extinct-only taxa?

Repeating the last plot, but conditioning our data on whether one or both taxa in an AD pair were still extant...



. . . or all-extinct.

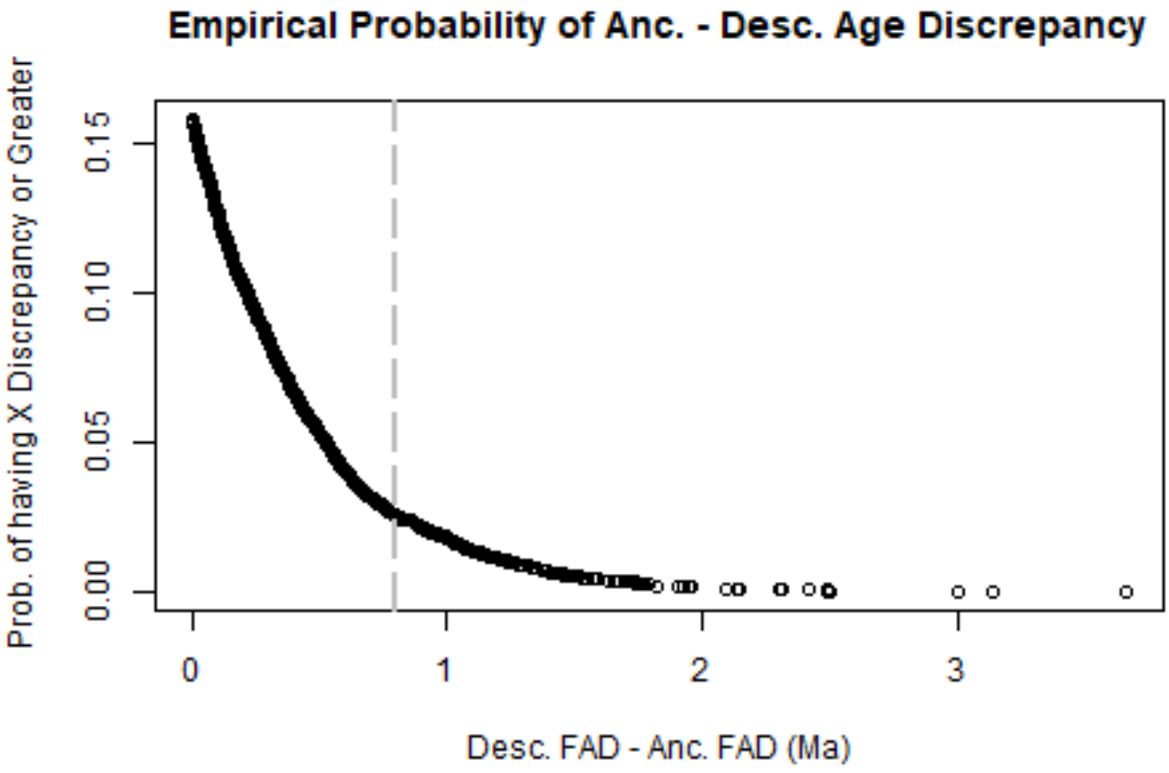


In both cases the lines get less smooth (less data to calculate the empirical cumulative distributions from), but both show fairly similar values and shapes to the plot using the full dataset.

So just how likely is a discrepancy of 0.8 Ma?

An alternative way of looking at the results of these simulations might be to ask how often we would expect to see a discrepancy of 0.8 Ma or greater in a fossil record with this set of generating parameters, without tying the probability value to a specific overlap value. D&A do this in their paper with their equation 6c, which integrates over 'all possible' values of `T_o` (i.e. the range of 0 to 0.97 Ma, the maximum of which is`T_R`), to find a *summary* p-value of 0.0009.

But what would be the comparable probability from our BDS simulations? To do this, first we'll need the cumulative distribution for the temporal offset itself, rather than for achieving a particular discrepancy at different values of range overlap (as D&A approached the issue).



Note that age offset is plotted so that the pairs with earlier appearing descendants have *positive* offsets, not negative. While this is more intuitive, this also means the curve is actually a reverse cumulative distribution - each point's vertical scale represents the proportion of simulations that have a discrepancy of the discrepancy duration shown on the horizontal axis, *or larger*, and thus the cumulative probability increases to the left. The vertical line is our discrepancy value of interest, 0.8 Ma.

One item to note is that discrepancies themselves are not rare in these simulated datasets, as we've already seen, but that the majority are quite small, and probably insignificant in duration in a real fossil record (we would never realistically even see such small discrepancies given chronostratigraphic uncertainty, correlation issues, etc).

The closest observed discrepancy value to 0.8 Ma, for which we have a cumulative probability from our empirical curve, is:

```
# the value of offset closest to 0.8 Ma
AD_discrepancy_uniq[closeDiff_T_d]
```

```
## [1] 0.8014188
```

Ah, that's pretty close. So what is its cumulative probability?

```
# the cumulative probability of observing a discrepancy at least as large as the above
cumulAbove[closeDiff_T_d]
```

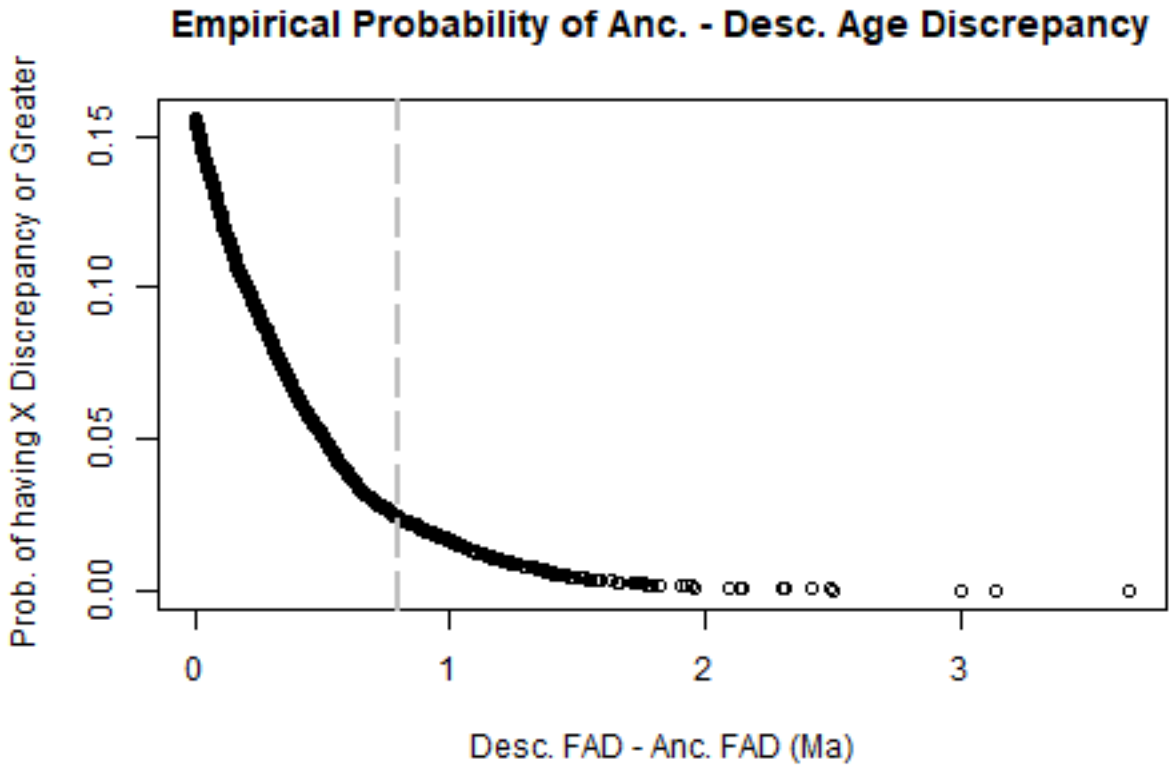
```
## [1] 0.0256
```

This number is *much* higher than the 0.0009 probability reported for D&A - more than two orders of magnitude higher. Note that this value is calculated from our simulations rather simply - it is the number of simulated, independent ancestor-descendant pairs with a discrepancy greater or equal to 0.8 Ma, divided by the total number of simulated pairs. Thus this value applies to all possible values of range overlap, just as

the 0.0009 reported by D&A does. Instead, the cumulative probability from the simulations is more similar to the cumulative probabilities of a given discrepancy value for different overlap values, as seen seen above. This value is higher than the 0.01-0.02 overlap probabilities discussed above, but only slightly.

Okay - so why do we get values that are so much larger here, but not for the probabilities at different values of overlap? Well, one possibility is that equation 5 represents a density, not a probability, in which case my comparison in the previous section was wrong (but doesn't impact my calculation of the probability of a discrepancy of a given size in this section). But let's consider more possibilities. We certainly obtain slightly higher probabilities for 0.08 Ma at specific overlap values than what D&A consider, but not high enough to explain our much higher probablility for the gap, independent of overlap. As discussed above, it might have something to do with our consideration of taxa with multiple occurrences (instead of just one find each), or our consideration of both direct and indirect ancestor-descendant relationships, but (as discussed earlier in this document) both of those should reduce this probability, not inflate it. The difference probably has more to do with the fact that taxon ranges are not forced to be a single value, but can differ and have all sorts of values, and that overlap can be across a very large range of time. Afterall, this includes some pairs that overlapped as much as four million years.

So, much like we restricted our examination of the probability of a gap at different overlap values to a restricted range of overlap values, what if we examine just those taxa with an overlap of less than 2 million years, adjusting both numerator and denominator in our calculation? (2 million years as an upper limit on the amount of overlap was one alternative scenario considered by D&A.)



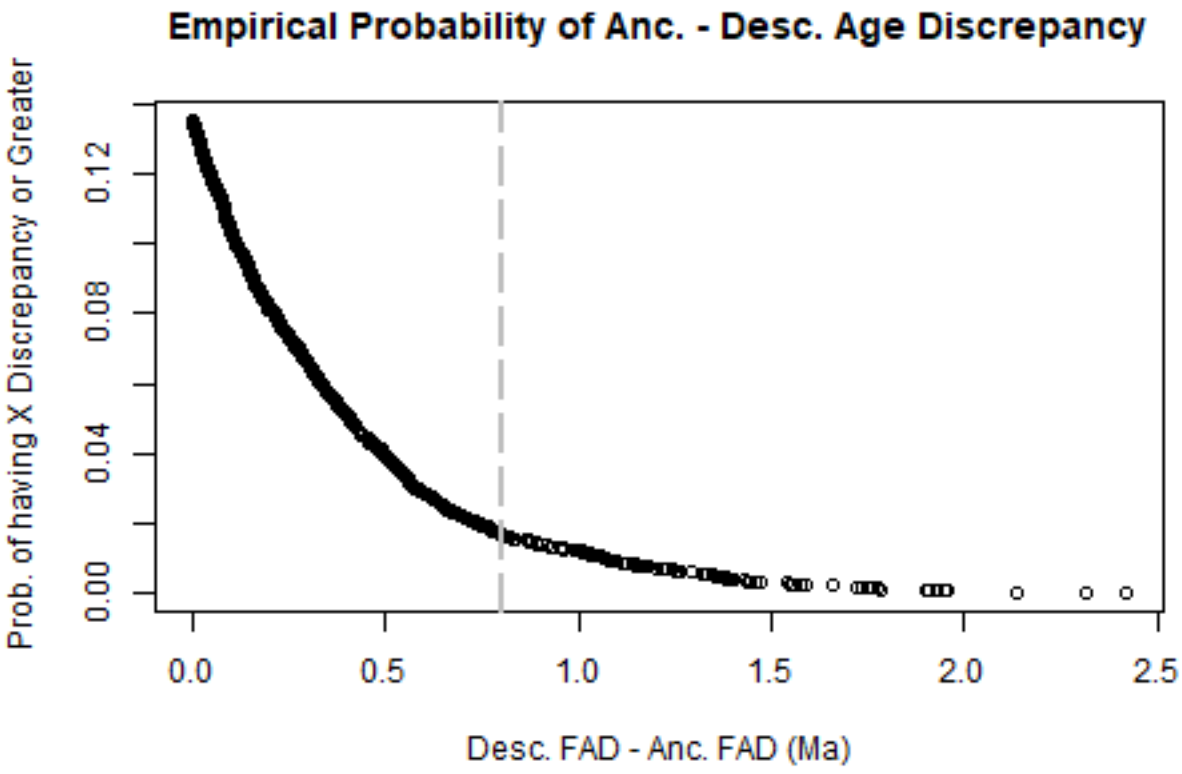
Well, that doesn't look different at all. What is the cumulative probability for `T_d >= 0.8`?

```
cumulAbove[closeDiff_T_d]
```

```
## [1] 0.02401724
```

Its a nearly imperceptible decrease in the probability. We could go further - what if we restrict ourselves to

pairs that overlapped for less than 0.97 Ma, the same upper threshold of overlap considered by D&A?



Okay, now we are beginning to see some slight visual change - but what is the new cumulative probability for `T_d >= 0.8`?

```
cumulAbove[closeDiff_T_d]
```

```
## [1] 0.0166277
```

A greater decrease, but we are still orders of magnitude beyond D&A's estimate. We will move forward with this probability, even if a maximum overlap of 0.97 is overly restrictive.

So this suggests that our increased probability value from simulations isn't due to including overlaps beyond what might be reasonable for *A. sediba* and *Homo*. Instead, it would seem that allowing ranges to vary and be independent of each other may have more to do with the birth-death-sampling simulations finding a higher overall probability for a given magnitude of discrepancy.

So what do these probabilities mean?

Statisticians and those who do a lot of statistics can read a probability value, just like a geologist can read the age '340 Ma' and immediately be able to put that value into context. This probability value may not look like much to one without that perception, so a more useful way of discussing a probability is to convert it to an odds ratio, which (for whatever reason), tends to be an easier way for most people to conceptualize probability.

```
# the odds of observing such are thus, 1 in... (rounded)
signif(1 / cumulAbove[closeDiff_T_d], digits=0)
```

```
## [1] 60
```

Compare this to the odds ratio for D&A's 0.0009 probability:

```
# the odds of observing such are thus, 1 in... (rounded)
signif(1 / 0.0009, digits=0)
```

```
## [1] 1000
```

So, if we had a fossil record that was as we described with our stated generating parameters - had this much turnover (extinction & origination rates), was of similar size (number of taxa), similar sampling (set equal to extinction rate, which is a very high sampling rate) - then we should expect to see a discrepancy of 0.8 Ma about one out of sixty times among those taxa that overlapped less than 1 Ma.

Now, 1/60 is definitely less than the generally accepted threshold for alpha probabilities from frequentist tests of 0.05 (1/20 odds), but 1/60 is not a *rare* thing. A complication with approaching this gap of 0.8 Ma with a frequentist statistical approach is that discrepancy in questions is at least partially challenged due to the magnitude of its temporal discrepancy. If we select an item among a larger set due to it being an apparent outlier, then we should not be surprised if we find that it is, indeed, an outlier (as D&A find in their metadata analysis of previously reported ancestor-descendant pairs). Testing for outliers *post-hoc* is always a statistically complicated task for this very reason. I would argue that the probabilities found by the birth-death-sampling simulation suggest 0.8 Ma would be an outlier, but not an incredible one. In a random, noisy universe, we should always expect to see outliers and coincidences - we should investigate each one carefully, but that doesn't mean that the mere fact something is an outlier does not mean it isn't accurate data.

Consider one way to interpret these odds: if we imagine that in the hominin fossil record there might be 60 such ancestor-descendant pairs (both known and unknown to workers, both direct and indirect ancestor-descendant relationships), then we **should** expect on average one discrepancy of that size or larger. Sometimes, there will be none, sometimes, there will be more than one.

D&A themselves report on 28 putative ancestor-descendant relationships, not including *A. sediba* and *Homo*. How often should we expect to see a at least one discrepancy of 0.8 Ma among 29 ancestor-descendant pairs, under the generating parameters of these simulations?

```
1 - ((1 - cumulAbove[closeDiff_T_d]) ^ 29)
```

```
## [1] 0.3850763
```

So, 40% of the time, we'd expect to see at least one gap as large as 0.8 Ma among 29 ancestor-descendant pairs, under these generating parameters. And that's presuming there aren't additional ancestor-descendant pairs beyond the literature.

Note that considering the possibility of a gap over 29 pairs does not really change D&A's interpretation, if we use the integrated probability of 0.0009 of D&A.

```
1 - ((1 - 0.0009) ^ 29)
```

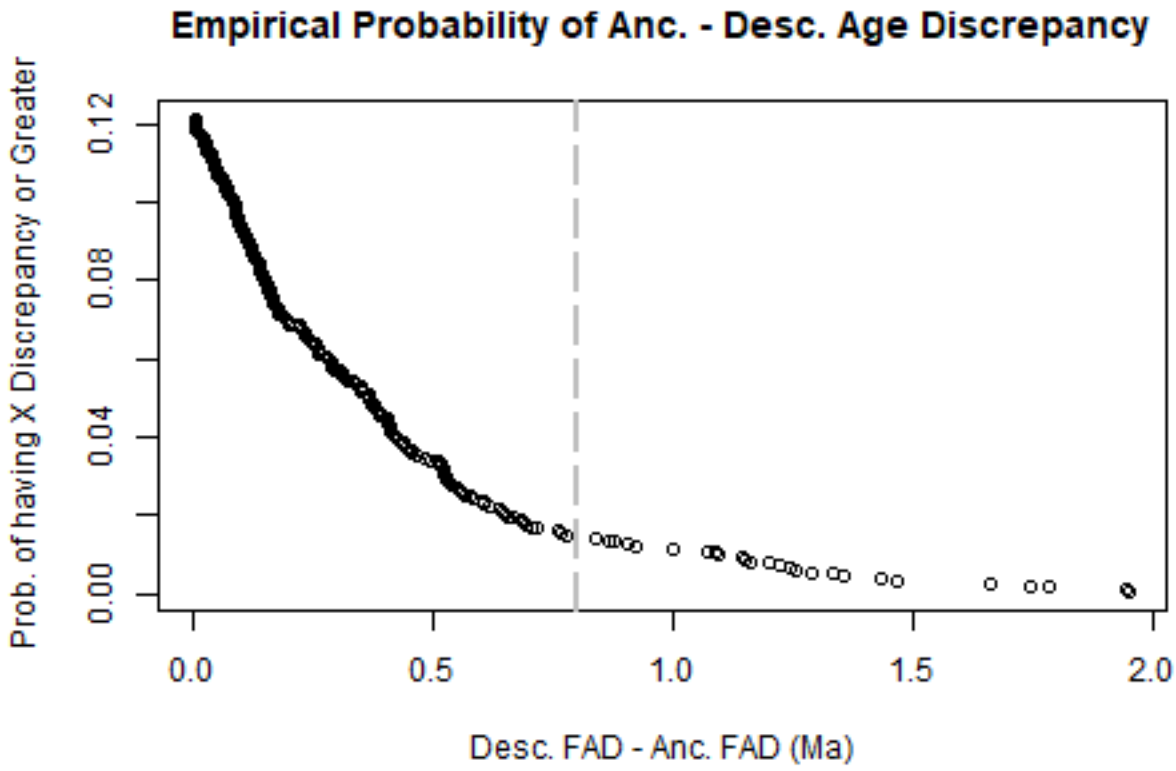
```
## [1] 0.02577379
```

So that's just 2.5% of the time we'd expect such under D&A's probability value. Thus, it is the difference in the probabilities estimated by an idealized model (such as D&A's) and simulations that make an enormous difference in whether a 0.8 Ma gap is unlikely or not.

Overall, given the overall probabilities found by these simulations, I would not interpret the temporal evidence as being strongly against *Homo* being a descendant of *A. sediba* due to their 0.8 Ma discrepancy between their first appearance datums from the known fossil record.

Does it differ if we condition on extant taxa, or on extinct-only taxa?

Again, we should check if the extant / extinct distinction has any bearing on this finding. Repeating the last plot, but conditioning our data on whether one or both taxa in an AD pair were still extant...



Which gives a cumulative probability for `T_d = 0.8`, conditioned on at least one taxon being extant, as:

```
cumulAbove[closeDiff_T_d]
```

```
## [1] 0.01435671
```

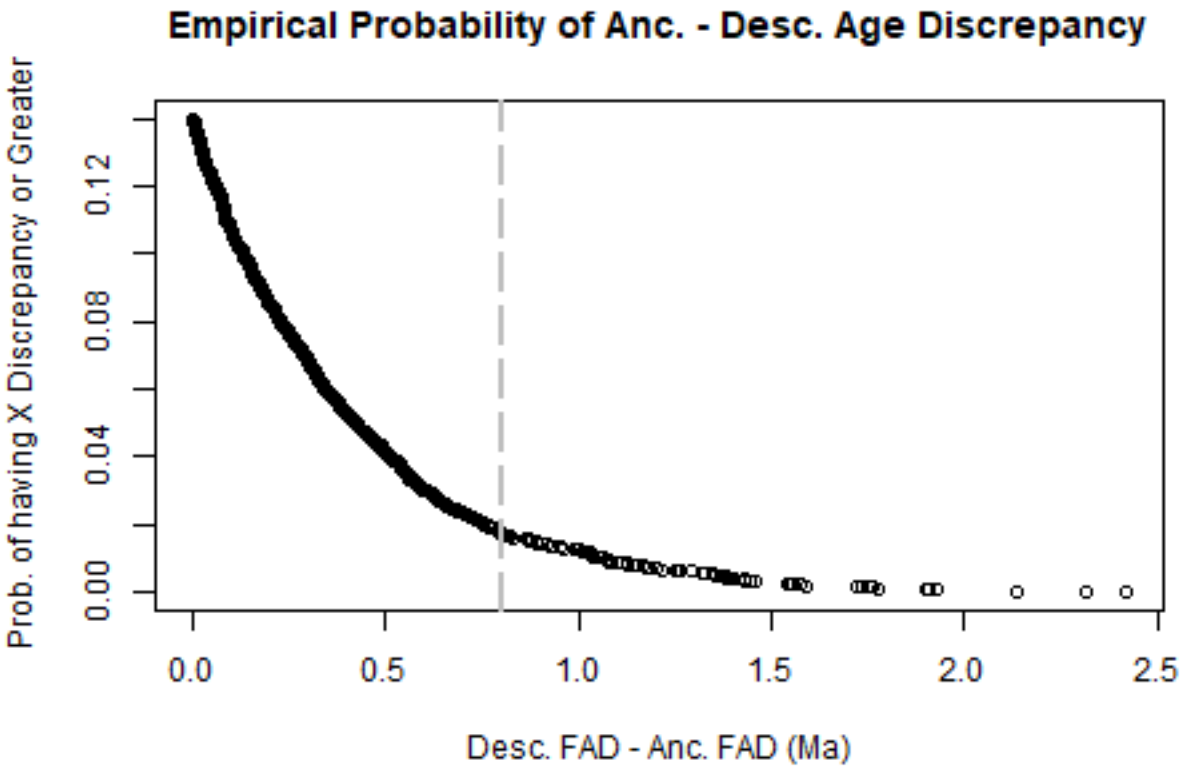
And an odds ratio of: (rounded)

```
signif(1 / cumulAbove[closeDiff_T_d], digits=0)
```

```
## [1] 70
```

So, a little less likely, but pretty similar to the odds ratio above.

And now, let's do the same for those pairs were all taxa were extinct:



Which gives a cumulative probability for `T_d = 0.8`, conditioned on at least one taxon being extant, as:

```
cumulAbove[closeDiff_T_d]
```

```
## [1] 0.01732631
```

And an odds ratio of: (rounded)

```
signif(1 / cumulAbove[closeDiff_T_d], digits=0)
```

```
## [1] 60
```

So, both extinct and extant subsets of the simulated data consistently give similar probabilities of observing a discrepancy as large as 0.8 Ma.

Summary and Remaining Questions

Let's summarize what we've discussed:

1) D&A's argument relies on two main pieces of evidence - a meta-analysis of published ancestor-descendant relationships, and the age difference between their first-found sampled fossil occurrence. I am not very convinced, as I think there is a deep bias in most areas of paleontology against treating temporally-incongruous taxa as ancestor-descendant pairs, and I have concerns about such data ever leading to reasonable information about what sort of temporal discordence we should expect. I think their 'first-find' approach is not sufficient, and other algorithms for picking a single find should be considered (as well as simply looking at the first appearance times).

2) D&A's central argument is based on an idealized model giving the probability of a direct ancestor-descendant pair having a certain age discrepancy between the fossil horizons for an ancestor and its descendant, for various values of possible duration overlap. This model makes many assumptions, some of which they defend exceptionally well (this is one of the best examples of a paper testing and showing support for homogeneity of sampling - I will be using this in classes as an example of such!), some of which work probably against their argument (e.g. considering only direct ancestor-descendant pairs probably increases the odds of a discrepancy), while others have uncertain effects (assuming every ancestor-descendant pair has the same fixed stratigraphic range). Their model predicts that gaps on the order of 0.8 Ma would have a low probability, below 2%, across the range of range overlap values.

3) Birth-death-sampling simulations testing this idealized model, which violate several of D&A's assumptions such as range constancy and direct ancestor-descendant relations, still finds probabilities similar to those depicted by D&A's model for the probability of an ancestor-descendant discrepancy of 0.8 Ma. However, the probabilities are slightly higher and have a very different shape compared to that returned by D&A's model.

4) Overall, simulation-based probabilities for a gap of 0.8 Ma, independent of overlap or only conditioned on overlap being less than 2 or 1 Ma, are on the order of 1.6 - 2.5%, which is 1/60 - 1/40 odds, respectively. Similar probabilities are obtained regardless of whether we look at pairs that include one or more extant taxa, or pairs that are entirely extinct. This probability is two orders of magnitude greater than the probability of 0.0009 reported by D&A (0.0009 = 0.09 %). This suggests that allowing taxon ranges to vary (as under a birth-death model) has a heavy impact on our assessment of the probability of a given discrepancy.

5) Given that examination of this putative ancestor-descendant pair is somewhat driven by the magitude of the temporal discrepancy, it seems inappropriate to reject the potential for an ancestor-descendant relationship with a 0.8 age discrepancy under a frequentist interpretation of the probability as a *p*-value. When we account for the number of ancestor-descendant already reported from the literature, as reported by Du & Alemseged, the probabilities as obtained from the simulation go from suggesting the *A. sediba-Homo* ancestor-descendant discrepancy is a suspicious outlier, to it simply being an **expected** outlier.

In conclusion, I do not think the arguments of D&A are supported when their arguments are carefully considered, and the results of their idealized probabilistic model is tested against a more realistic birth-death-sampling model. That said, there are additional routes forward.

First, D&A suggest that under their very specific assumptions, sampling rate does not affect the results. The simulations run here use a sampling rate that is very high relative to similar estimates from other groups in the fossil record, equal to the value used for origination and extinction rate. It is possible that the probabilities found here for an Anc-Desc pair having a discrepancy as large as 0.8 Ma depends on sampling rate, or that this very high sampling rate is unrealistic. However, reducing the sampling rate should, if anything, lead to a more incomplete record, one where discrepancy should be more common. Comparability of records with different sampling rates is an issue, however, and so this is not attempted here.

Second, as I stated, I think there are issues with comparability with the metadata regarding published ancestor-descendant relationships, but I do believe that some additional steps beyond the first-find approach, such as a random-find approach, and comparisons to the currently known FADs might be very productive avenues to explore. Additionally, this data could gives us a glimpse into the process of paleontological systematics itself. The history of taxon stratigraphic and phylogenetic interpretation contains a lot of valuable information about how we, as a culture of scientists interested in evolutionary relationships, decide to list a set of taxa as ancestor-descendant pairs, how that evolves over time with our stratigraphic understanding, the degree to which current chronostratigraphic understanding lags or laps ahead of statements about ancestor-descendant relationships, etc. These are things that should be considered further, even if they are hard to seperate.

Overall, D&A do an incredible service to the literature: the field of understanding, analysizing, and testing ancestor-descendant relationships has been long considered a fantastical back-water that is considered

unquantifiable, and thus, a form of paleontological quackery. Temporal information absolutely holds the power to affect our decision to support putative ancestor-descendant relationships. However, from temporal information alone, and using many of the same assumptions made by D&A, I do not find evidence from temporal data alone to support rejecting the hypothesis that *A. sediba* is an ancestor of *Homo*.