Heliocentric Certainty Against a Bottleneck of Two?

Peaceful Science

swamidass 2017-12-29 21:02:50 UTC #1

Do "Humans" arise from a single couple?

It is a deceptively simple question, with a great deal of subtlety. Some claim "no," with certainty approach that of our certainty that the sun is the center of the solar system: "**heliocentric certainty**."

This, however, is a subtle question. Most important to be careful about is the equivocation between genetics and genealogy, which does arise here (see TMRCA vs. TMR4A). More significantly, we find that all "humans" do descend each individually from many single couples (see the **Genealogical Adam[http://peacefulscience.org/genealogical-rapprochement/]**). That, however, is a conversation for another day.

Instead, here, I wanted to focus on understanding why scientists are convinced our ancestors arise as a population that never dips down to a single couple. What is the evidence behind that finding? What are its limits? How strong is that case? Most of my work takes this for granted, but it is worth pressing further into the population genetics.

On this question, there has been some very interesting discussion lately about population genetics. It deserves a deeper dive, which I want to give here. It is generally thought that the evidence against a bottleneck of one couple in our ancestors is overwhelming. In the recent past (less than 100 thousand years), this certainly seems to be the case. This question about ancient bottlenecks though.

Going forward, we should keep in mind:

1. No estimates of population size in the distance past (which are a geometric average over time windows) ever goes below a few thousand. The real question here is if the population size estimates are good enough in the distant past to detect a brief bottleneck.
2. There appears to be overwhelming evidence against a recent bottleneck less than (for example) 100 kya.
3. The case against a more ancient bottleneck is what is at question here, but even if its possible, there does not appear to be any positive evidence for an ancient bottleneck. We are discussion, instead, whether there is strong evidence against it.
4. As yet, no mathematical theory of a bottleneck has been put forward yet that explains the full range of data. The only theories offered right now validated on the data do not include a bottleneck. That could change, but that is where things stand.
5. The consensus of population genetics is solidly against any notion of a bottleneck of a single couple. Perhaps they are wrong, but the consensus is solidly against a bottleneck.

Therefore, it is correct to say the consensus is against a bottleneck of two in both our recent and distant past. This seems to be bad news for those affirm a historical Adam, ancestor of us all. Of course, this is not really important if we think Adam's line could have interbred with others, because we do not expect there to be a bottleneck in the genetic data with a genealogical Adam, ancestor of us all.

It is still worth asking how certain we should understand this scientific claim. Is it to heliocentric certainty? Or less sure?

Richard Buggs (a UK geneticist) has been pressing on this (**https://natureecoevocommunity.nature.com/channels/522-journal-club/posts/22075-adam-and-eve-a-tested-hypothesis[https://natureecoevocommunity.nature.com/channels/522-journal-club/posts/22075-adam-and-eve-a-tested-hypothesis]**). The conversation quickly became interesting. First, I pointed out that if we allow for interbreeding and keep in mind the difference between theological "human" and *Homo sapiens*, then we do see genealogical universal ancestors very early (see the **Genealogical Adam[http://peacefulscience.org/genealogical-rapprochement/]**). Second, and to the point right now, Richard Buggs raised some interesting questions about the evidence. His point is simple. We cannot know for sure, because we never tested the bottleneck hypothesis, and the evidence is not a definitive as we move past a few hundred thousand years ago.

The conversation has become technical, but I wanted to highlight on some of my thoughts here. What does the evidence tell us about a "human" bottleneck of one couple? Of course, all the same caveats regarding the definition of "human" apply. "Human" is an ambiguous term in the distant past.

Stated more precisely, what does the evidence tell us about a bottleneck in our ancestors since it seems we diverged from a common ancestor with the great apes about 6 mya?

Clarifying This Conversation's Question

In this conversation ***we are not allowing for any miracles***. For example, the bottleneck couple would not be specially created or be genetic mosaics (with different genomes in every sperm/egg). Instead, they would be biologically normal individuals, but the only ones alive at their time that produce surviving offspring till today.

Also, **we are not looking *for* evidence for a bottleneck**, but rather testing the strength of the evidence ***against*** a single-couple bottleneck. The bar here is really low. We are just asking the degree to which we really know from the evidence that there was no bottleneck in our lineage.

Finally, we are ***only*** considering bottlenecks from 100 kya to 3 mya (kya=thousands of years ago, mya=millions of years ago). We are not considering very recent bottlenecks (say 50 kya or 10 kya). Those recent bottlenecks seem to be inconsistent with the data. Instead, we are wondering about the strengths and limits of the scientific claim that "humans" arise as a population that never dips down to a single couple. The place that claim is the weakest is in the distant past, for example 500 kya ago.

So, no miracles. No positive proof. Not recent. This is, instead, a question about the limits of the scientific claim. Other questions, of course, are interesting. We, however, have a more sharply delimited concern on this thread.

To Summarize…

To Summarize this exchange, Richard Bugg's suggests this text, which is currently under revision…

> As Christian biologists, we have over the last few months reviewed the population genetic literature, asking if it is possible that all modern humans could descend from a single couple within a theistic evolution (or evolutionary creation) framework. We have assumed that humans share common ancestry with apes, and that God has not intervened with physical miracles. Our task has been difficult because the hypothesis of a bottleneck of two in the human lineage has not been directly addressed in the scientific literature using genome-wide human diversity data. Nonetheless, from those published studies of human diversity that we have reviewed, and based on our understanding of current theory, we have drawn tentative conclusions. We conclude that current human genetic diversity data does not rule out a bottleneck of two individuals in the human lineage between approximately 400,000 and 7,000,000 years ago, but neither do they show that such a bottleneck has happened. Current analyses and models suggest that a two-person bottleneck has not occurred below a threshold of approximately 400,000 years before present. More research is needed in this area, and we are open to new analyses moving this threshold up or down.
> **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/592[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/592]**

I would point out, however, that the results here do not depend on evolution, or the assumption of common descent. The conclusions here apply just as strongly to OEC and YEC models.

Guide to Contents

Claims of Heliocentric Certainty.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/3?u=swamidass] What are the scientific claims in question?

The Ecological Fallacy.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/4?u=swamidass]: *Homo sapiens* go to zero, so why couldn't they go to two?

TMRCA or Time to Most Recent 4 Alleles?[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/5?u=swamidass] TMR4A (not TMRCA) puts the bounds on a couple bottleneck.

Estimate with Median or Max?[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/6?u=swamidass] The statistically sound approach is the median.

TMR4A from Genome-Wide TMRCA.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/7?u=swamidass] An initial estimate of TMR4A.

The ArgWeaver Genome Wide Phylogenies[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/8?u=swamidass] 424 GB of data with genome-wide answers.

Genome-Wide TMR4A.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/9?u=swamidass] A better estimate of TMR4A.

ArgWeaver Does Not Assume Large Population.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/10?u=swamidass] The computed TMR4A is biased downwards, not upwards, by the prior.

The Correct Mutation Rate.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/11?u=swamidass] ArgWeager is using an experimentally confirmed mutation rate.

Correctly Weighting Coalescents.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/12?u=swamidass] An improve esitmate of TMRCA is about 500 kya.

ArgWeaver works like MAP and MrBayes.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/13?u=swamidass] Really, no assumptions of population size are made, and this is just a measure of human variation, converted to units of time.

An Estimate Robust to Correction.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/14?u=swamidass] The TMR4A estimate is exceedingly stable. AJ Roberts from *Reasons to Believe* would want a correction for the amount of genome that is not yet sequenced.

What about Recombination?[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/15?u=swamidass] The errors we see in ArgWeaver do not effect TMR4A estimates.

Trans-species variation.[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/16?u=swamidass] The evidence against an ancient bottleneck in trans-species variation is not as strong as I had thought.

Convergent Evolution or Trans-Species Variation?[https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/18?u=swamidass] A deeper look indicates convergent evolution, which violates the assumptions required for genetic clocks and undermines substantially the argument against a bottleneck using this line of evidence.

This topic will not open to general comment, as it is technically detailed, and I do not want it to get confusing. However, please comment on the companion thread: Comments on "Heliocentric Certainty Against a Bottleneck of Two"? [https://discourse.peacefulscience.org/t/comments-on-heliocentric-certainty-against-a-bottleneck-of-two/62?u=swamidass]

---

swamidass 2017-12-29 22:26:56 UTC #3

Claims of Heliocentric Certainty

This is exactly the evidence referenced in *Adam and the Genome* by Dennis Venema, when he writes (emphasis mine):

> As our methodology becomes more sophisticated and more data are examined, we will likely further refine our estimates in the future. That said, **we can be confident that finding evidence**…**that we descend from only two people just isn't going to happen**. Some ideas in science are so well supported that it is highly unlikely new evidence will substantially modify them, and these are among them. **The sun is at the center of our solar system, humans evolved, and we evolved as a population.**
>
> Put most simply, DNA evidence indicates that **humans descend from a large population because we, as a species, are so genetically diverse** in the present day that a large ancestral population is needed to transmit that diversity to us. To date, **every genetic analysis estimating ancestral population sizes has agreed that we descend from a population of thousands, not a single ancestral couple.** Even though many of these methods are independent of each other, all methods employed to date agree that the human lineage has not dipped below several thousand individuals for the last three million years or more—long before our lineage was even remotely close to what we would call "human." **Thus the hypothesis that humans descend solely from one ancestral couple in has not yet found any experimental support,— and it is therefore not one that geneticists view as viable.** **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/296?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/296?u=swamidass]**

Dennis' use of "human" here is ambiguous. He clarified to Richard Buggs…

> **The heliocentric quote, which I thought was the object of your concern, is about humans (Homo sapiens).** When I'm speaking about our lineage leading up to humans at 200KYA I use "lineage" or similar.
>
> The other two quotes remain valid. Does "it seems" sound like I'm saying this is as certain as heliocentrism? That would be quite the understatement. **That is a summary statement of all the lines of evidence in the literature to date that do not provide support for a bottleneck below ~10,000 at any time in the last 18MY (which remains the case).**
>
> All methods employed to date agree that the human lineage has not dipped below several thousand individuals for the last 3 million years or more – long before our lineage was even remotely called "human".
>
> …
>
> **So: "heliocentric certain": humans.** Pretty darn certain: lineage leading to humans over the last several hundred thousand years (say back to ~500,000 years ago). Confident but not as definitive: lineage over the last few million years. Survey of literature to date: no evidence of a bottleneck greater than thousands anywhere, regardless of time. **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/308?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/308?u=swamidass]**

Which I summarized (to his approval):

> 1. *Homo sapiens* specifically do not dip down to a single couple in 300 kya to the confidence we have in heliocentrism.
> 2. Our ancestors as a whole do not dip down to a single couple between 300 kya and 3 mya with very high confidence, but maybe not as high.
>
> **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/322?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/322?u=swamidass]**

So these are the claims in question. At least as I understand from Dennis.

---

swamidass 2017-12-29 22:44:49 UTC #4

The Ecological Fallacy

Let's start with this first claim by Dennis (rephrased by me).

> DennisVenema:
>
> Homo sapiens specifically do not dip down to a single couple in 300 kya to the confidence we have in heliocentrism.

On face value, we know that this cannot be known with certainty. Some scientists are not even sure if the remains earlier than 300 kya are fully *Homo sapien*. That means it is entirely reasonable to believe that *Homo sapiens* go to ***zero*** (less than a single couple) in this time frame.

Where is the error? The inference that (1) our ancestors never go below a few thousand, so therefore (2) Homo sapiens never go below a few thousand, is an example of the Ecological Fallacy. What applies to a group of things (the group here is "our ancestors") does not necessarily apply to all the individuals in that group (the individual here is "Homo sapiens").

To be 100% clear, this is not at all a challenge to mainstream population genetics, which makes claims about our ancestors as a whole. All the population size estimates (which are geometric averages over a time window) all include **Homo sapiens + others**, as far as I know. As far as I know, no one has found a way to figure out what the ratio is between the two population is; nor has anyone asked the question in a research study.

It is entirely possible that at sometime in the distant past, the total number of breeding Homo sapiens is precisely two, even though the total number of our ancestors at the time would be much more. Eventually, these early Homo sapiens would interbreed with others. This is an example of a hypothesis that (1) directly contradicts @DennisVenema's claim, and (2) appears entirely consistent with the evidence, as I understand it (and I'm happy to be corrected). If any such hypothesis exists (and I think I've just demonstrated one), then Dennis's claim is false.

The problem is not the findings of mainstream population genetics at all here. **The problem is rather in the inference from this finding to the claim that this means *Homo sapiens* never dip below a few thousand.** If we go back far enough their numbers go to zero. If they can go to zero without contradicting the evidence; so why can't they go to 2? (of course, the would be interbreeding with surrounding non-sapiens).

Now, I stand to be corrected if there is a body of scientific work uncovered that cleverly demonstrates how Homo sapien DNA is separable from all other DNA, and effectively estimates population sizes over the last 350 kya years. It is entirely possible this large body of work exists, and I never encountered it.

Absent that body of work, I think Dennis just misunderstood (or misspoke) the relationship between the finding of population genetics and his novel claim, not even realizing his claim was novel (see **Ecological Fallacy**). If I had done the same, I would probably just retract the mistake. Perhaps replacing it with a more precise statement that actually does have a great deal of evidence. That's just me though. I, personally, find no value in defending my own misstatements.

Everyone makes mistakes anyways. I find people trust me more when I am quick to retract mine.

---

swamidass 2017-12-29 23:15:57 UTC #5

With the first claim set aside for now, the second claim is more interesting…

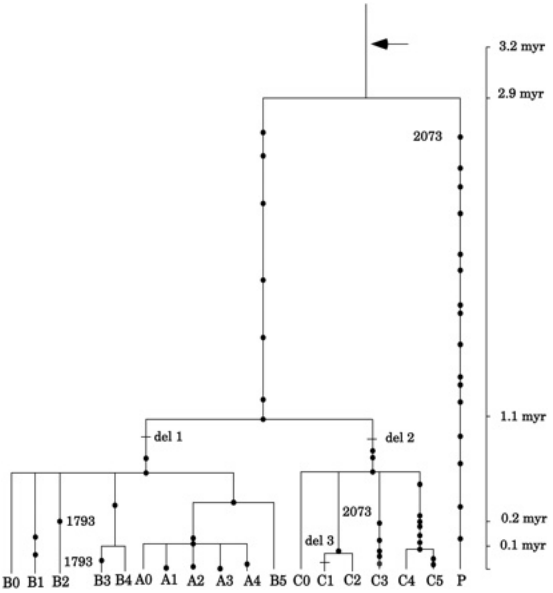
> Our ancestors as a whole do not dip down to a single couple between 300 kya and 3 mya with very high confidence, but maybe not as high.

Was there a bottleneck among our ancestors *as a whole*, both *Homo sapiens* and non-sapiens. What is the timeline over which we are certain there was no bottleneck? 100 kya? 500 kya? 2 mya? or 13 mya? How certain should we be in this evidence?

TMRCA or TMR4A?

Population size estimates in the past from genetic sequences are fairly involved. It is not easy to explain concisely how these estimates are made, and that will be left for a future effort. However, one common way a bottleneck is argued against is using the estimated TMRCA of segments of DNA. **To be clear, this is not really the strongest evidence against a single couple bottleneck. For that we have to look elsewhere.**

Dennis, nonetheless, pointed to this paper as evidence for no bottleneck in the last 3.2 million years.



http://www.genetics.org/content/genetics/172/2/1139/F3.medium.gif[http://www.genetics.org/content/genetics/172/2/1139/F3.mediu
http://www.genetics.org/content/172/2/1139.long[http://www.genetics.org/content/172/2/1139.long]

This is a phylogeny that shows the history of a piece of DNA, which appears to have a history going back 3.3 million years. If we take the tip (at 3.2 million years), that is were we expect to find the *genetic* common ancestor. That is how some put a bound on a bottleneck.

However, if we are looking to put a minimum time on a bottleneck, the bottleneck couple certainly could have been heterozygous, carrying 4 alleles at each loci (in their autosomal genome). If that is the case, it is not the time to most recent common ancestor (TMRCA) that matters, but the time to most recent four alleles (TMR4A). That would correspond to the horizontal line which would correspond to when 4 lineages arise (which is in the much more recent past).

In this case, also, there is not enough data to precisely visualize TMR4A in the phylogeny. When there is not enough data, it is common to branch the phylogeny more than twice. Taking that uncertainty into account, the TMR4A It could easily be within 0.5 million years ago.

Estimating TMR4A

TMR4A is not regularly computed in genetics studies. Usually only TMRCA values are put forward. Is there a way to estimate TMR4A from TMRCA? One heuristic is that, on average:

```
TMR4A = TMRCA / 4
```

I put this forward in the conversation, but did not have the math worked out. It was an educated guess.
https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/248?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/248?u=swamidass]

I was honestly not comfortable with that, even after looking at some phylogenies that seem to validate the general number. So I wrote up some code to compute what the average expectation is. Assuming constant population size, we expect on average:

```
TMR4A = 0.24 * TMRCA
```

Which is almost exactly what I estimated: 1/4. Keep in mind, however, that this is a noisy estimate. And it will not hold up in practice. In a future post, I test this empirically, and find that on the ArgWeaver dataset that the actual relationship is closer to:

```
TMR4A = 0.38 * TMRCA
```

Regardless, this is only an initial heuristic. In fact, we will see that TMR4A is much more stable than TMRCA, and stays in a tighter range. The higher TMRCA is the lower the ratio TMR4A / TMRCA. So, this quite substantially reduces the time bound on a bottleneck using the TMRCA approach.

There can be a wide gap between this estimate and the true TMR4A. It has to be estimated directly from the data. For example, in the figure, the TMR4A is actually ***less than*** TMRCA / 4 (450 kya / 3.2 mya ~= 0.14, not 0.25), because this TMRCA is a high outlier. I caution strongly against "eyeballing" these graphs to get anything more than a general understanding of what we are calculating. Ultimately, we need to compute TMR4A across the whole autosomal GENOME, and understand that distribution to understand what the data shows.

The bad news is that this is very hard to do. The good news is that this is becoming easier with new markov chain inferences of recombination graphs. More on that later though.

Note: the move to TMR4A is not justified when using DNA from X-chromosome (where TMR3A would be relevant) or from mitochondrial and or Y chromosome (where TMRCA is just fine). The y-MRCA and m-MRCA do seem to put a bound a single couple bottleneck to sometime after 100 kya to 150 kya.

The Math

Here is some python code to compute the numbers (it requires numpy to run).

```
# Any large number will do, because the coalescent ages quickly converge.
expected_age = n_kingman_ages(107)

# What do we multiply TMRCA by to estimate TMR4A?
# 4th coalescent corresponds with TMR4A
print expected_age[3] / expected_age[0]
>>> 0.242990654206

# By what do we multiply the first 6 coalescents to estimate TMR4A?
print expected_age[3] / expected_age[:6]
>>> [ 0.24299065  0.49056604  0.74285714  1.          1.26213592  1.52941176]
```

> **https://repl.it/@swamidass/TMR4A[https://repl.it/@swamidass/TMR4A]**

For reference, a phylogeny with 54 diploid individuals has 54 * 2 - 1 = 107 coalescents. There 54 individuals in the Personal Genomics data used by ArgWeaver, which is why I use 107 here. Any large number, however, will do, as this summation converges very quickly.

The functions are defined as:

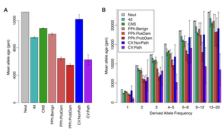
```
import numpy as np

def n_kingman_waits(n):
  """the expected wait times / Ne of the 1 through n coalescents.
  The n-th element is the (n+1) coalecent time, and is the expected time / Ne
  for n+2 alleles to coalesce to n+1. Ne is the effective population size (alleles).
  """
  k = np.arange(n) + 1
  return 4. / (k * (k+1))

def n_kingman_ages(n):
  # get the expected ages / Ne of the n-coalescents
  return np.cumsum(n_kingman_waits(n)[::-1])[::-1]
```

---

swamidass 2017-12-30 01:30:01 UTC #6

Maximum or Median TMRCA (and TMR4A)?

This discussion of TMR4A brought us to this phenomenal paper, which inferred phylogenies (and ancestral recombination) across the whole genomes of 54 individuals. This is really hard to do, and took quite a bit of methodological innovation, and a lot of computer time.



Genome-Wide Inference of Ancestral Recombination Graphs[http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004342]

Author Summary The unusual and complex correlation structure of population samples of genetic sequences presents a fundamental statistical challenge that pervades nearly all areas of population genetics. Historical recombination events produce an...

The key data the authors report is this table, which reports the max TMRCAs across the genome. Here they are, reported in units of "generations" where a generation is 25 years.

#	Chr[a]	Start	End	TMRCA[b]	Poly/kb[c]	Npoly[d]	CNV[e]	Comments
1	chr4	190590001	190600000	615775	16.6	32.8	√	Part of large intergenic region near telomere of long arm of chr 4 (see [76])
2	chr5	21560001	21570000	503311	16.2	5.1	√	Intron of *GUSBP1*
3	chr3	97930001	97940000	479803	16.4	5.3		Intergenic region in cluster of olfactory receptor genes
4	chr6	57270001	57280000	479504	13.7	28.0	√	Intron of *PRIM2*
5	chr2	223940001	223950000	449728	19.8	4.3		Intergenic region downstream of *KCNE4*
6	chr5	21550001	21560000	412679	14.2	4.4	√	Intron of *GUSBP1*
7	chr6	57230001	57230000	399887	16.2	12.8	√	Intron of *PRIM2*
8	chr6	29680001	29690000	380228	15.3	10.0		Intergenic region upstream of *HLA-F*
9	chr1	94220001	94230000	377017	8.0	4.2		Intron of *BCAR3*
10	chr8	123070001	123080000	375128	15.3	4.2		Intron of *BC052578*
11	chr11	55670001	55680000	374537	12.0	4.3		Intergenic region between *TRIM51* and *OR5W2*
12	chr6	29950001	29960000	371110	17.6	7.6		Intergenic region between *HLA-A* and *HLA-J*
13	chr17	64010001	64020000	367842	8.6	5.5		Intron of *CEP112*
14	chr6	29670001	29680000	365313	15.8	10.1		Intergenic region upstream of *HLA-F*
15	chr11	55690001	55700000	361088	11.5	4.1		Intergenic region between *OR5W2* and *OR5I1*
16	chr6	158680001	158690000	345382	10.4	4.8		Intergenic region upstream of *TULP4*
17	chr6	29720001	29730000	341797	12.4	8.0		Intergenic region between *HLA-F* and *HLA-G*
18	chr17	43790001	43800000	335647	11.2	5.0		Intron of *CRHR1*
19	chr6	8470001	8480000	325656	10.1	4.5		Intron of noncoding RNA *LOC100506207*
20	chr4	141920001	141930000	325570	12.1	3.2		Intron of *RNF150*

[a]Genomic coordinates in hg19 assembly. The genome was simply partitioned into nonoverlapping 10 kb intervals in hg19 coordinates.
[b]Posterior expected TMRCA in generations, averaged across unfiltered genomic positions in region.
[c]Number of polymorphisms in Complete Genomics dataset in region per kilobase of unfiltered sequence.
[d]Normalized polymorphism rate: number of polymorphisms per unfiltered kilobase divided first by the local mutation rate (as estimated from divergence to nonhuman primate outgroup genomes) then by the average of the same polymorphism/divergence ratio in designated neutral regions. The resulting value can be interpreted as a fold increase in the mutation-normalized polymorphism rate compared with the expectation under neutrality. The same measure was computed from the much larger 1000 Genomes Project Phase 1 data set, and was significantly elevated in these 20 high-TMRCA regions (Supplementary Figure S13).
[e]Possible copy number variant (CNV), based on Complete Genomics "hypervariable" or "invariant" labels (see Methods). Polymorphism rates in these regions may be over-estimated.
doi:10.1371/journal.pgen.1004342.t002

[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/db12047a41efa96dbf326afa9bfa3996abf12f8c.png]

Multiplying by 25, we find TMRCAs here ranging from 7.5 **million** to 15 **million** years ago (mya). Dividing by 4 (to get a TMR4A), that would seem to put a limit on a bottleneck at 3.75 mya. Is this, however, valid reasoning?

Four things are important to keep in mind:

1. These are ***maximum* TMRCAs** reported on a genomewide scan, which by definition, are not representative. They are much larger than average or median TMRCA, which is what we really want.
2. These are ***estimated* TMRCAs**, which are drawn (essentially) from a random distribution. Even if all true TMR4As were less than 1mya, for example, we would still expect to see estimated TMRCAs greater than 1mya. In fact, about 50% of estimated TMRCAs would be expected to be greater than the true TMRCA.
3. In genome wide analysis, **outlier TMRCAs are likely *erroneous***; they are the places were the assumptions required for dating are most likely violated. For example, several of these positions are subject to balancing selection (and therefore need to be dated differently) and the top positions have high copy number variation (CNV). High copy number variation is an example of a type of sequence that will be erroneously given high TMRCA just because the method used here does not take it into account.
4. As we will see, **for outlier TMRCA, our approximation of TMR4A breaks down**. For balancing selection, the gap between TMR4A and TMRCA can be arbitrarily large. The same is true for trans-species variation. We really need to measure the TMR4A directly to know what it is in these cases.

So moving from this table to limit on bottleneck time is not valid reasoning. There are methodological and statistical problems with that leap. Using the maximum TMRCA for this purpose is the definition of cherry-picking data.

What we really need is the distribution of TMRCAs (and eventually TMR4As) across the genome. The good news is that we found exactly that in the supplementary data of the paper. From there, we can get to our first plausible genomewide estimate of TMR4A.

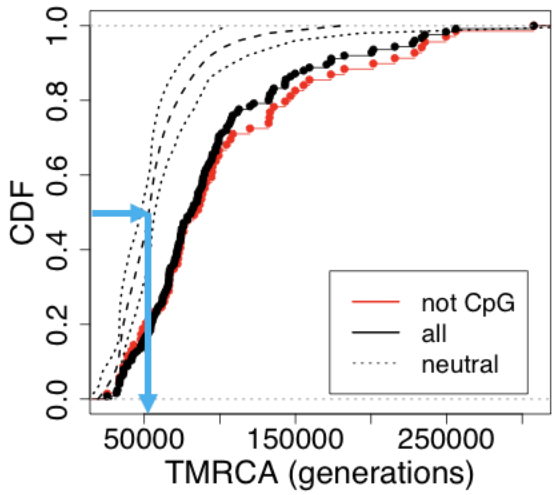
---

swamidass 2017-12-30 01:36:13 UTC #7

The Genome-Wide TMRCA Distribution

What we really need is the distribution of TMRCAs (and eventually TMR4As) across the genome. The good news is that we found exactly that in the supplementary data of the paper. From there, we can get to our first plausible genomewide estimate of TMR4A.

This figure from the argweaver paper (S17) includes a random sample of 69 neutral regions (dashed line), compared with 69 regions undergoing balancing selection and containing no CpGs (red). The black line is the 56 regions undergoing balancing selection, but with shared CpGs. **Though not the entire genome, the dashed line is going to be a good estimate of the neutral genome-wide distribution.** For the statistically untrained, this going to be a hard graph to read. It is a CDF, not a PDF (**https://en.wikipedia.org/wiki/Cumulative_distribution_function1[https://en.wikipedia.org/wiki/Cumulative_distribution_function1]**).



Several factors can conspire to increase or reduce TMRCA. Molecular clocks only work when these factors are not interfering. That is why whole genome distributions are so important. We can test the effect of different regions. For example, if we wanted, we could start to untangle how identifiably neanderthal interbreeding biases results upwards, by seeing the results on those regions separately. We can also see how balancing selection affects dates (which violates the assumptions required for dating). Some regions of the genome, also have higher mutation rates (and therefore will overestimate TMRCA).

From this, we want the best estimate of TMRCA in ***neutral*** regions of the genome (the dashed line) in a way that reduces these sources of error. This is a fairly important point, as dates can can only be reliably inferred in places that are not under balancing selection. These are the only places where a molecular clock is expected to hold. Even then, some regions will still get "lucky" and coalesce more quickly to or much more slowly. So to a first approximation, we want the the median of these values.

We can make our estimate. In the regions not under selection, we see a median for the TMRCA at about 50,000 generations. You can see it yourself tracing the blue line in the graph. That gives us an **estimated TMR4A of about 310 kya**, well within the timeline where some scientists think *Homo sapiens* arise. According to this one view of the data (***other data might contradict this***) it is possible that our ancestors (including the Neanderthal lines) went through a single couple bottleneck about when *Homo sapiens* arise. This is not at all our final estimate, but just what this limited view of the data shows.

To be clear, however, this is just an estimate, based on our weak estimate of TMR4A from TMRCA. What we really need to do is look at TMR4A directly, in genome-wide phylogenies themselves. We will do that next.

---

swamidass 2017-12-30 02:55:26 UTC #8

Richard Buggs writes:

> This is exactly my point. Thank you for stating it so concisely. To my mind, the way ahead would be to write a programme that computes the TMR4A for each haplotype block of the human genome, and work out a reasonable time frame using data from all blocks. Until that has been done, I do not think we can say that the bottleneck hypothesis has been rigorously tested. **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/264?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/264?u=swamidass]**

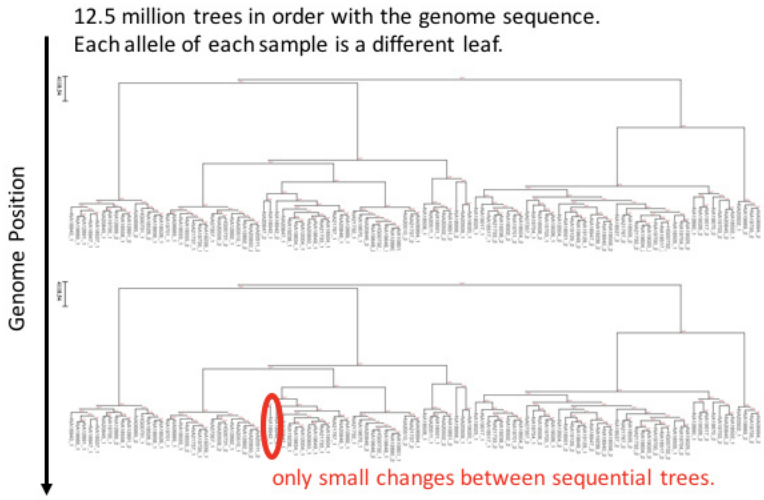
Let's give it a shot. First, however, we need some data…

Exploring Genome-Wide Phylogenies

I reached out to the authors of the paper under question. To their credit, they responded back immediately, and were very helpful. It is not as common as it should be, but Dr. Adam Seipel from Cold Springs Harbor Laboratories, was immensely helpful, answering several questions about the data.

It turns out that they had already made genome wide phylogenies available on the web, for anyone to download. **compgen.cshl.edu/ARGweaver/CG_results/download/[http://compgen.cshl.edu/ARGweaver/CG_results/download/]**. The data

looks, utlimately, like this:



There were two problems.

1. The concatenated results file is 424 GB (half a terabyte). Downloading it in on shot takes a prohibitively long time, because downloads are quickly throttled to a snail's pace.

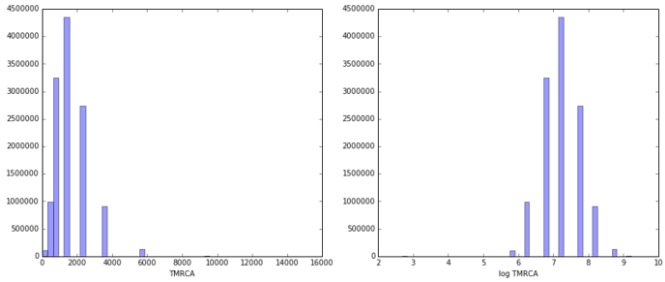
2. Processing the trees with my library of choice (dendropy, **dendropy.org/[http://dendropy.org/]**) was much slower than expected. To give a sense of the scale, there are **12,500,000 trees** covering about 200 bp each, and conditioned to be similar to one another if they are adjacent on a chromosomal arm. That is just for one sample, but there are 200 samples from the model. So this is a lot of data that is hard to process quickly.

The good news is that both these hurdles were manageable. It turns out that a single genome wide sample (about 1.5 GB) can be downloaded (I chose sample from iteration 2400) in pieces (getting past the throttle) and only about three hours of processing time is required (on a super fast compute node rented from Amazon). From here, I extracted the first 50 coalescents in to a 126 MB file.

I'm going explain the data involved, and then dive into the key results.

The Distribution of TMRCA

The first thing to do with new data is to look at its distribution. Here is the distribution (PDF, **https://en.wikipedia.org/wiki/Probability_density_function[https://en.wikipedia.org/wiki/Probability_density_function]**) of TMRCA's across the genome. This is the age of the 1st-coalescent, the way I am indexing things.



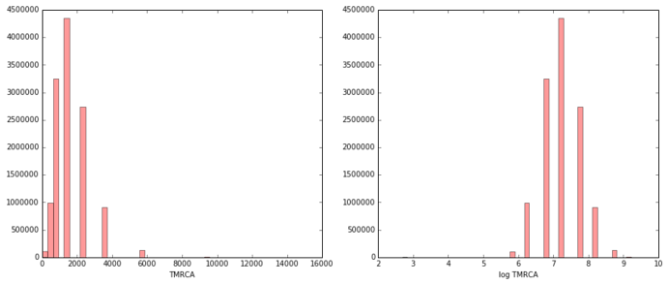
[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/3c688b0b91eb92f91b7ca83813eef3a3b48ed9fd.png]

Oddly, this is not a smooth distribution. It is not spread out, but it looks like the results are quantized. This becomes more clear when we look at the distribution for log TMRCAs. It looks like TMRCAs come in groups, spaced out by about 0.5 log units. That is exactly the case. It turns out that that the way that ArgWeaver (the program used in this study) can work so quickly is by quantizing the results. This substantially simplifies the problem. A close read would have clarified this, but I had not read the paper closely enough. The data however makes this clear.

This is a profoundly important point, as this will introduce substantial error into any individual TMRCA estimates. It highlights again the value of looking at distribution of TMRCAs, not just individual outliers.

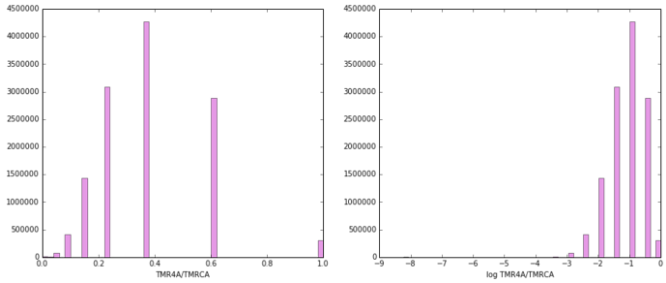
The Distribution of TMR4A

We see a similar distribution for TMR4A, which corresponds with the 4th-coalescent. The results are quantized, instead of being nicely distributed across the data. This going to present a problem as we move to estimating the TMR4A across the genome. **We have to correct for this quantization somehow.**



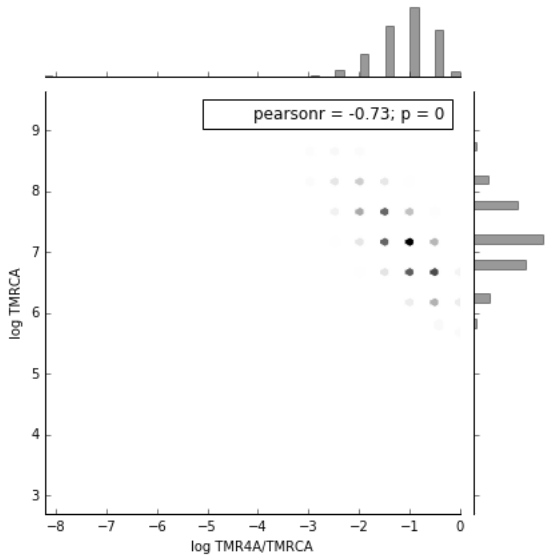
[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/d92e715014f3a275b5316ab7e5cc928216b813b2.png]

We can also ask what the relationship between TMRCA and TMR4A is, by plotting the distribution of TMR4A / TMRCA.



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/2cd6bb96efb8b9417892b538d993c2c9554722e4.png]

Once again we see same quantization, even spaced in log scale. So it seems that ArgWeaver is fixing TMR4A at some multiple of TMRCA. **In this dataset, on average, TMRCA * 0.37 = TMR4A.** Moreover, we also see a fairly strong relationship between TMR4A/TMRCA and TMRCA. This is evident on both standard and log scale (Pearson R -0.63 and -0.73, respectively), but is clearest in log scale:



The key way to understand this is that the larger TMRCA grows the smaller TMR4A is relative to TMRCA. That is an important point. If we care about TMR4A, we cannot interpret large TMRCAs easily. We have to have the underlying data directly to estimate it.

That is, however, exactly what we have, and what we can use now to make an unbiased estimate of TMRCA and TMR4A across the genome.

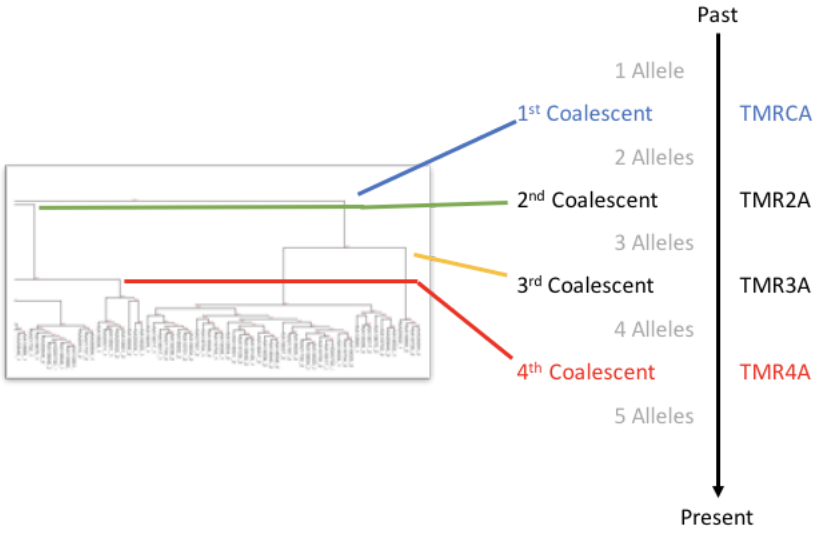
---

swamidass 2017-12-30 03:33:57 UTC #9

Richard Buggs writes:

> This is exactly my point. Thank you for stating it so concisely. To my mind, the way ahead would be to write a programme that computes the TMR4A for each haplotype block of the human genome, and work out a reasonable time frame using data from all blocks. Until that has been done, I do not think we can say that the bottleneck hypothesis has been rigorously tested. **https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/264?u=swamidass[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/264?u=swamidass]**

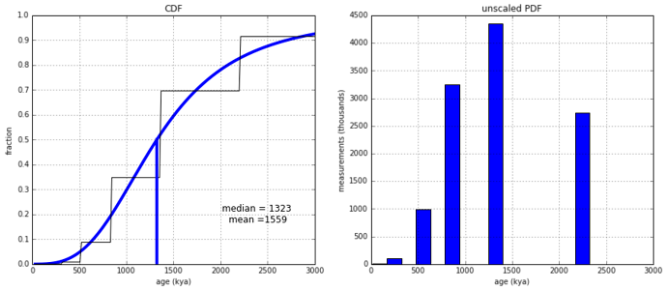
Now we are prepared to take him up on that challenge. It will be really interesting to see what the data shows. Here is a key schematic to show how this relates with the phylogenies we collected. Moreover, beware that I have reversed numbering of coalescents from standard notation (as I learned it) for clarity here.



Notice that the 4th coalescent is the TMR4A. That is what we are after. Also notice how “bottom heavy” the phylogeny is. The trunks (at the top here) are usually long, but the branches are usually short. We are after the age of fourth node in the tree, the 4th coalescent.

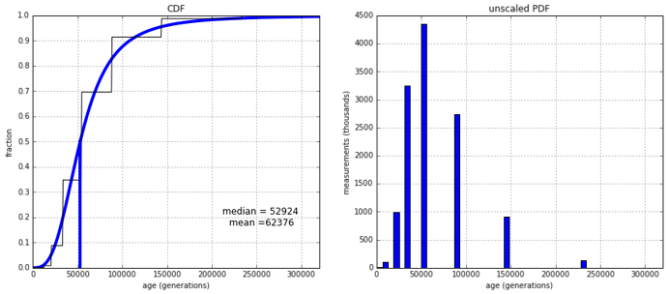
A Genome-Wide Estimate of TMRCA

So first, off, we can take a look at the TMRCA distribution. Here, the CDF is on the left and the PDF of the distribution is on the right. Time 0 is present day and time 3,000 at the end of the graph is 3 mya. The jagged black line is the CDF of the data, which is a step function because it is quantized (not smooth). The smooth colored line is a sigmoid function fit to the CDF (in log space), and the colored line dropping down shows where the median estimate is based on the curve. **Using this method, the estimated genome-wide TMRCA is 1323 kya, or 1.3 million years.**



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/68ad33fbee5d88d3d0cac26e89c238d18d359df7.png]

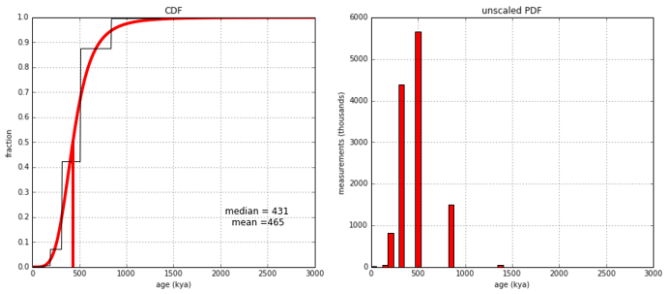
For reference, this is really close to the value computed from S17. Overlaying the two figures, it looks as if the genome-wide distribution is skewed a bit to the right at the top portion, which increases our estimates by a small amount. That is to be expected, as S17 used a small set of hand checked neutral regions. These genomewide estimate includes everything. I'm including a graph approximately scaled to S17 for reference (mind the change in units).



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/c15196d8235d001acd2eee52efe1cdb6b12ca4fa.png]

A Genome-Wide Estimate of TMR4A

We can do the same thing for TMR4A. The graph here is on the same scale, for comparison to the prior figure. **Using this method, the estimated genome-wide TMR4A is 431 kya,** well within the 500 kya that Richard Buggs hypothesized as a point beyond which we may not have as much confidence in ruling out a bottleneck.



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/93dce51b3cc0740acf1f5ad9ebe448796f599bdf.png]

We want to be sure this is a robust result, not overly affected by all the quantization errors. Notice that there is higher certainty in the TMR4A estimate than the TMRCA estimate because:

1. The variance of the TMR4A is much lower than TMRCA. Visually, we see that in the reduced spread of the TMR4A PDF (right) and increased steepness in the CDF.

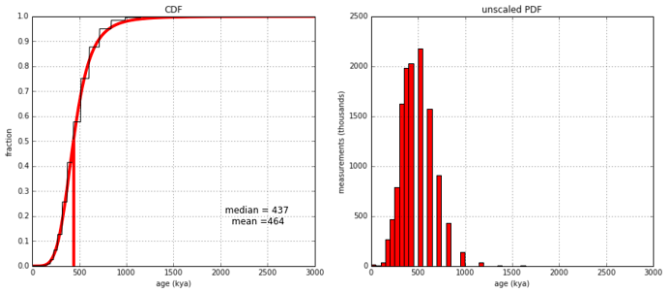
2. The closer fit between the fit curve and the underlying data.

3. Notice also that the outlier TMR4A's are much less extreme. There are fewer outliers, and the max is not too extreme. It is high, but not nearly as bad an outlier as the highest TMRCA (13 million). That is good sign, suggesting we are measuring real distribution around a real quantity of the data. We might look at these more closely later, but that is not our focus here.

Just to be sure, and raised confidence further, I computed TMR4A another way. this time averaging three estimates together. If our first estimate was good, it should be robust to these sorts of refinements.

```
Improved_TMR4A = ( TMR3A * 0.74 + TMR4A + TMR5A * 1.26 ) / 3
```

The estimates were defined earlier (**Heliocentric Certainty Against a Bottleneck of Two?** [https://discourse.peacefulscience.org/t/heliocentric-certainty-against-a-bottleneck-of-two/61/5?u=swamidass]). Here we use the 3rd and 5th coalescents to estimate the 4th coalescent, averaging them all together. Below, we see this does a lot to smooth out the distribution; but it does not succeed completely. **This improved estimate is 437 kya, nearly identical to the first estimate.** This increases our confidence that this is a reasonable estimate; it is robust to manipulations like this.



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/4d26ea37b935e1114a5ed9e7dd617d9718b0633b.png]

A large number of variations like this were attempted, and they gave similar results. That is good news. This is not the last word on TMR4A, but it is pretty good “first” word on the matter.

What is the Confidence Interval?

So we arrive at a genomewide TMR4A estimate of about 430 kya. Given all the uncertainties involved, I would say that the confidence interval here is probably +/- 100 kya. That is based on the recognition:

1. ArgWeaver is merely approximating the TMRCA, with a quantization. That certainly adds error here.
2. Mutation rate is assumed constant across 10kb windows, but is known to vary on shorter distance scales.
3. Most importantly, we expect mutation rate to be different at times in the past. For this reason, we cannot know it for sure. Especially as we are considering times in the deep past (say 300 kya) but not so deep that we can expect the central limit theorem to save us (say if we are going 5 mya).
4. This does not correct for balancing selection, by making sure we are only looking at neutral regions of the genome. This might skew the results, but it also protects against bias and makes this analysis more reproducible.
5. At the moment, all coalescents are weighted equally. This biases the averages to high recombination areas, which might be biased towards upward errors in TMRCA estimates. It would be wiser to average weighting by the length of the DNA segment to which the phylogeny applies.
6. There may be other ways to improve these estimates. Likewise, the quantization could introduce systematic bias that is not fully corrected for in our approach. Using a median of a fitted curve helps, so does averaging multiple estimates of the coalescent. It might be better, however, to find a way to refine the trees further. It’s not clear, however, how to do this. It may be infeasible with current approaches.

Based on these concerns, I can give my expert opinion, which is nonetheless only an opinion. I’d estimate that there is about 20% error one way or another, at least.

Keeping in mind that some scientists think that Homo sapiens might arise as early as 340 kya, it is plausible (***but not proven***) to imagine a bottleneck of one couple at that time. Some will debate the exact reasoning here. This, I should add, is not my view of our origins. This, however, is a plausible scientific argument.

Not All the Evidence

It is critical to point out that this is NOT all the evidence at play. For example, this is not an estimate of effective population size (which would look at the distribution of coalescents over time). It also ignores trans-species variation. Trans-species variation is shared by humans and chimps, like that we see in the MHC, and might provide the strongest evidence against a single couple bottleneck. Those parts of the genomes are some of the outliers. To full consider this hypothesis, we have to think about that data too.

We will consider these two types of evidence in time. However, it **is important not to overstate what has been shown here.** The key point I’ve put forward here is that we should not overinterpret TMRCA as evidence against a bottleneck in our lineage before 0.5 mya. A very recent bottleneck (say 50 kya) seems impossible, but a more ancient bottleneck of our ancestors (if very brief) at 500 kya might be consistent with the evidence. Sometime before 500 kya, this couple would not be *Homo sapiens*, but they might (exact dates debatable) be the common ancestor of *Homo sapiens*, *Denisovans*, and *Neanderthals*.

Moreover, data from TMRCA on mitochondria and Y-chromosomes should put a bound at least at 100 kya, after which a bottleneck is not very likely. So we still are very certain that a very recent bottleneck is inconsistent with the evidence.

---

swamidass 2018-01-21 09:27:25 UTC #10

ArgWeaver Does Not Assume Large Population

In response to the work done here, there were concerns that ArgWeaver did not allow for bottlenecks of 2, and would be biased against it. RIchard Buggs wrote…

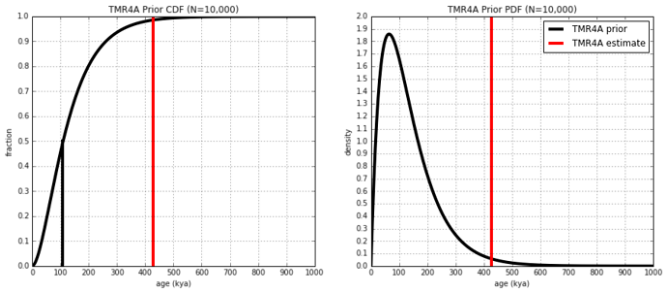
> I do think that the coalescent models used in a test of the bottleneck hypothesis would need to include the effective population size decreasing down to two as we go back in time. I realise that this would be a lot of work, but I do think that this would be necessary. Do correct me if I am missing something, Joshua.

It appears he was drawing upon an observation by Andrew Jones at the DI, who writes:

> **However, a little digging into how ARGweaver works reveals that it too assumes a constant population, and uses this assumption to assign probabilities to ancestry trees. Therefore, again, it is not clear if it is really appropriate for asking questions about Adam and Eve.** The particular reason why it is a problem is a bit technical: coalescence (branching but backwards in time) happens much more slowly in a large population. In a large population, the last few coalescents could take thousands of generations. But what if you have a small number of generations, drawing to a smaller and smaller population and terminating in a single couple? All the lineages will coalesce (down to at most four as explained above) but at a faster rate. **https://evolutionnews.org/2018/01/on-prejudiced-models-and-human-origins/[https://evolutionnews.org/2018/01/on-prejudiced-models-and-human-origins/]**

The program itself does require a population size number, and the authors used 10,000, when it is run. However, this skepticism turns out not to be the correct assessment.

ArgWeaver is using a **prior[https://en.wikipedia.org/wiki/Prior_probability]** on the trees, that is parameterized by population size (N = 10,000). The language of "assumes a large population size" is just correct. It is more accurate to say that is starts with a weak prior belief of a population size of 10,000. It is a weak prior belief, because it is designed to be quickly overcome by data. Here is what it looks like:



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/f23bb1ae787e9039e002f9d15b454792320871a6.png]

Notice that the prior median (black) is at about 100 kya, but the data shows a TMR4A of about 420 kya (as we have shown before). From here, it becomes clear why there is no reason to doubt these results:

1. As a prior, this is not an assumption, but a starting belief that is meant to be overridden by the data. The only way that the ArgWeaver program uses the population size is in computing this prior. Population size is neither simulated nor modeled in the program except for placing this weak prior on population size. **Remember, priors are not assumptions or constraints.** That is why the measured
2. The ArgWeaver output files tell us the strength of the prior vs. the data, and it is just about 5%. That means the model output is dominated 95% by the data, and not by the prior (as it is designed).
3. The prior distribution for TMR4A is at about 100 kya, but we measured the TMR4A at about 420 kya. That means the data is pulling the estimate upwards from the prior, not downwards.

This last point should end any confusion. To draw analogy, it's like we measured the weight of widgets, with the weak starting belief that the average weight of these widgets is 200 lb. After weighing several of them, and taking the prior into account, we compute the average weight is 420 lb. The fact we used a prior could be an argument that the real average is greater than 420 lb, but that is not a plausible argument that the true average is less than 420 lb. The prior, in our case is biasing the results downwards, not upwards.

With that in mind Dr. Jones was just mistaken when he writes:

> The tool used, ARGweaver, is fantastic in that it combines an enormous amount of real genetic information to model the past genetic history of humans. For this reason it gives the impression of being truly objective, and so when I first read it, I thought he had proved that there could be no bottleneck earlier than 300,000 years…However, a little digging into how ARGweaver works reveals that it too assumes a constant population, and uses this assumption to assign probabilities to ancestry trees.

Given what I have just explained, this is not a reason to doubt the results that I put forward. I do believe this data shows there could be no bottleneck earlier than 300,000 years without either miracles or our ancestors have vastly different mutation rates than us. Both those possibilities, however, are off the table right now.

Technical Details

Getting the prior required spelunking a bit into the ArgWeaver code. The key function is included below, and translated from the original C code here…

```
def prob_coal_counts( a,  b,  t,  n):
  '''The probability of going from 'a' lineages to 'b' lineages in time 't'
  with population size 'n'

  Original code was written in C and can be found here: https://github.com/mdrasmus/argweaver/blob/fe
  '''
  C = 1.0;

  for y in xrange(b):
    C *= (b+y)*(a-y)/float(a+y);
    s = np.exp(-b*(b-1)*t/2.0/n) * C;

  for k in xrange(b+1, a+1, 1):
    k1 = float(k - 1); k = float(k)
    C *= (b+k1)*(a-k1)/(a+k1)/(b-k);
    s += np.exp(-k*k1*t/2.0/n) * (2*k-1) / (k1+b) * C;

  for i in xrange(1, b, 1):
    s /= i

  return s
```

From here, we can compute the prior distribution of the TMR4A with this function call:

```
Probability_Denisty = prob_coal_counts(5,4, N_Generations, 10000 * 2)
```

The multiple of 2 is used because humans are diploid, and the time units used internally by ArgWeaver are generations. Following the paper, we also use a generation time of 25 years and a mutation rate of 1.8e-8 per generation.

Its not entirely clear why the prior comes out this low, at 100 kya. I would have thought it would be closer to 250 kya. There may be a bug in the argweaver code for computing prior, which shifts the prior downward. Even if this is a bug, this code is computing the prior used in the data we computed TMR4A, so it is ultimately what we need to understand how to interpret the impact of the prior on these results.

---

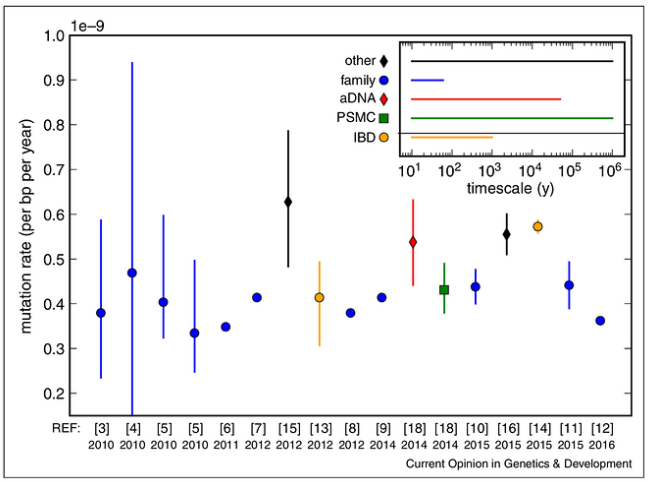
swamidass 2018-01-21 09:50:58 UTC #11

Argweaver uses a generation time of 25 years with a mutation rate of 1.26e-8 mutations per generation, which is equivalent to a mutation rate of 0.52e-9 mutations per generation.

Is this the correct number?

Directly Measured Mutation Rates

In the genomic age, with easy sequencing, we can now directly measure the mutation rate. Without getting into the details, there are several ways we can do this. More than one method. A recent review, coallates this recent data, and Figure 1 has all we need to assess the mutation rate more carefully.
https://www.sciencedirect.com/science/article/pii/S0959437X16301010[https://www.sciencedirect.com/science/article/pii/S0959437



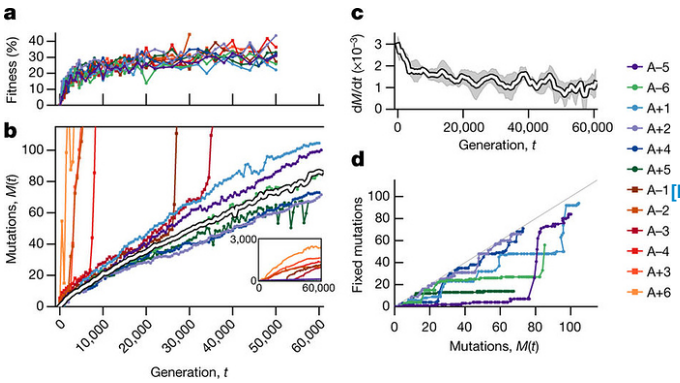
[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/f337133d49594e654900f2c4dbde8cb19eeeb721.png]

Each color is a different method, and we can see that they are pretty close, centered at about 0.5e-9. This is a result computed from several different studies, using several different methods, which all give us about the same result. This means that ArgWeaver is using an appropriate mutation rate, determined by several independent methods.

Variation in Mutation Rates

Yes, there is variation in mutation rate.

However, the complex effects we see, for example, in the Lenski experiment (which was referenced in this conversation, https://www.nature.com/articles/nature24287[https://www.nature.com/articles/nature24287]) are not relevant to this problem. Look at figure 2 from the Lenski paper, which shows how mutation rate varies in each experiment. Look at panel B.



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/5eea1e584a74c9a92abf707833ade849ad8b4c07.jpg]

Here we can see there are two groups. One group (red and orange) are **mutators**, that at some point start rapidly mutating much more than the normal rate. One group (blue - purple) are **non-mutators** which are just mutating a a more normal rate. The key point is that there is a wild difference between these two groups.

However, this **wild of variation in mutation rates is not relevant to mammalian populations**. There is much much more constraints on mammalian germline mutation rates, and we do not see such wild swings between populations. So this is an example of an effect in the Lenski experiment that we do not need to account for when studying human DNA. Adding to that pattern, we know that much more of the human genome is non-coding than in bacteria, so it will be more clock like too.

We can measure it in different populations, and we can even detect some differences in the past. These variations, however, in humans are all relatively small. These variations, also, are not always to higher mutation rates, but also to lower mutation rates. So yes, it is likely that mutation rates were ***slightly*** higher in particular populations or points in the past (let's say within 2-fold per year), but it is also likely they were ***slightly*** lower at times too. For the most part, this just averages out over long periods when looking at the whole human population. That is not 100% true, but the law of averages is why variation in mutation rate is not going to dramatically increase our confidence interval on TMR4A by much.

Yes, we do expect some variation, but it's hard to imagine that variation being more than just 10 to 20% when averaged over 10s of thousands of years.

How Much Faster for a TMR4A of 100 kya or 6 kya.

How much faster would we have to see mutation to find a TMR4A a much lower level.

1. If TMR4A was at 100 kya, we would expect the mutation rate to be 2.1e-8 per generation, or 4x more than observed.
2. If TMR4A was at 6 kya, we would expect the mutation rate to be 32e-8 per generation, or 64x more than observed.

We can see that both these rates are *well* outside the range we see in the observed rates. It is just not likely that a bottleneck of single couple took place at 100 kya, and its entirely inconsistent with the evidence at 6 kya.

---

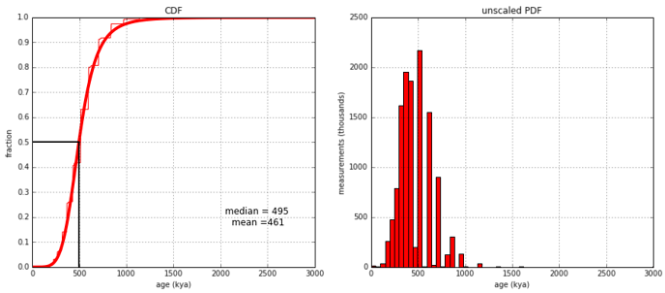
swamidass 2018-01-21 11:04:07 UTC #12

Correctly Weighting Coalescents

I've been sure for a while that there was simplification introducing error into this estimate.

> swamidass:
>
> At the moment, all coalescents are weighted equally. This biases the averages to high recombination areas, which might be biased towards upward errors in TMRCA estimates. It would be wiser to average weighting by the length of the DNA segment to which the phylogeny applies.

I finally got around to correcting this part of the code, and recomputing the TMR4A. Here is what we arrive at, a TMR4A of 495 kya, nearly 500 kya. This is a better estimate.



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/94c9420257f170b3e5f847aff3363ba3451568a2.png]

How Much Faster for a TMR4A of 100 kya or 6 kya?

How much faster would we have to see mutation to find a TMR4A a much lower level? These numbers change with this improved estimate.

1. If TMR4A was at 100 kya, we would expect the mutation rate to be 2.6e-8 per generation, or 5x more than observed.
2. If TMR4A was at 6 kya, we would expect the mutation rate to be 43e-8 per generation, or 83x more than observed.

That is why most scientists are going to find a single couple genetic origin of humans before 500 kya implausible. It just does not fit the evidence.

As I get more opportunities to refine the results, I'll continue making them and posting them here.

---

swamidass 2018-02-11 19:02:36 UTC #13

There was ongoing confusion about how ArgWeaver was using Ne (effective population size) in doing its computation. Was this an assumption that reduced our confidence in the results, or something else?

The information in **this post[https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/488?u=swamidass]** eventually cleared up the confusion for Dr. Buggs.

> Thank you for your patience with me regarding Ne and ARGweaver. I think I have misunderstood something, and I am just having more of a think about this. As I go back over your posts, I am struck by how many times you have made the same point to me, without me really taking it on board:
>
> > That means the TMR4A (and all TMRCAs) are determined primarily using the formula: D = T * R, where D is mutational distance, T is time, and R is the mutation rate. That is the key determinants of the TMR4A.

You are right, that is the key point. I'm glad we are getting chance to explain it.

> What you seem to be saying is that they are simply taking a molecular clock approach to estimating TMRCA. Time is the number of differences divided by the mutation rate. They are building phylogenetic trees and dating them.

That is exactly right. That is what they are doing, with a few bells and whistles. Essentially, this is exactly what MrBayes does (**http://mrbayes.sourceforge.net/[http://mrbayes.sourceforge.net/]**), except that unlike MrBayes, ArgWeaver can handle recombination. Technically, it is constructing ARGs (ancestral recombination graphs), not phylogenetic trees. ARGs (of the sort argweaver computes) can be represented as sequential trees along the genome. That's convenient representation that is easier for most of us to wrap our heads around, but the actual entity it is constructing is that ARG.

> The reason why I have been so preoccupied with Ne is because I thought this was a coalescent analysis, where time to coalescence is proportional to effective population size.

Except, as you are coming to see, this is not a coalescence simulation at all.

To clarify for observers, there are three types of activities/programs relevant here.

1. ***Phylogenetic tree inference***. Starting DNA sequences -> find the best fitting phylogenetic tree (or ARGs when using recombination) -> assign mutations to legs of tree (or ARG) -> use #mutations to determine length of legs. (see for example MrBayes)
2. ***Coalescence simulation***. Starting from a known population history -> simulated phylogenetic trees (or ARGs when using recombination) -> simulated DNA sequences. (see for example ms, msms, and msprime)
3. ***Demographic history inference***. Many methods are used, but one common way (see mcms) is…starting from DNA sequences -> Infer phylogenetic trees / args (task #1) -> compute the ***coalescent rate*** at time windows in the past -> Ne is the reciprocal of the coalescent rate. (see for example psmc and msmc).

It seems that there was some confusion about what ArgWeaver was doing. Some people thought it was doing #2 or #3, but it is actually just doing #1. The confusion arose because it used a fixed Ne as parameter, which seemed only to make sense if it was doing #2, and might make its results suspect if it was doing #3. However, ArgWeaver was never designed to do #2 or #3. Instead, it is doing #1.

So what is the Ne for?

One of the *features* of ArgWeaver is that it uses a prior, which is good statistical practice. They were using Ne to tune the shape of the prior, but ultimately this does not have a large effect on the trees. It's only important, in the end, when there is low amounts of data. They prior they used pushed the TMR4A downwards from what the data showed too.

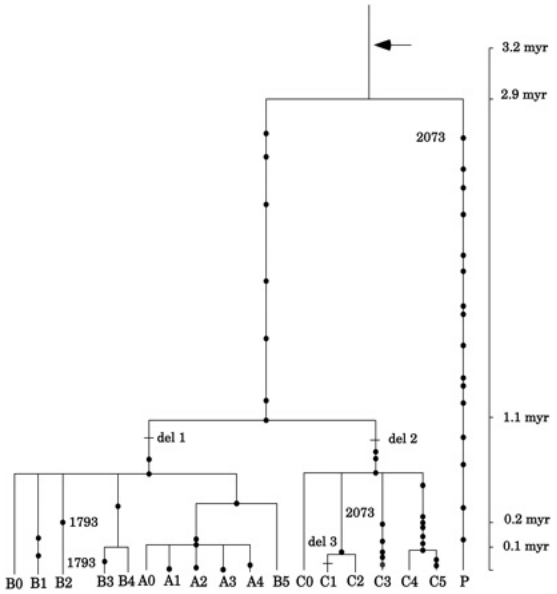
How This All Gets Confusing…

In defense of the confused, the confusing reality of population genetics is that the same quantities can be expressed in several different units. Often they are all used interchangeably without clear explanation, and its really up to the listener to sort out by context what is going on.

This is all possible because of that key equation:

$$D = R * T$$

However, it means have flexibility in the units we choose to measure the lengths legs a phylogenetic tree. To help explain, let's go back to a figure from much earlier in the conversation:



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/7a137bd8ef95f0a198251ddb8480d0ad6f8ca0d9.jpg] http://www.genetics.org/content/172/2/1139.long2[http://www.genetics.org/content/172/2/1139.long2]

.

In this figure, the dots are mutations assigned to legs in tree, the scale bar is in units of time (years in this case), and the leaves of the tree are observed DNA sequences obtained from actual humans. I've seen several units of tree length pop in this conversation and the literature…

1. Number of mutations (dots in figure, or D in my formula)
2. Years (scale bar in figure)
3. Generations (argweaver)
4. Coalescence units (number mutations / sequence length, or D in my formula)

A critical point it that in ArgWeaver the mutations are ***observed*** in the data, and the number along each leg is used to estimate the time using the formula. This is just unit conversions, provided we clarify mutation rates, the length of the sequence, and (sometimes) generation time. So these units are interconvertible if we settle one the mutation rate. If we express them as coalescence units or number of mutations, then they do not even require specifying a mutation rate, as this is a fundamental property of the data itself.

Though, as we have discussed, we have reasonable estimates of mutation rates determined by direct measurement. For example, ArgWeaver uses a generation time of 25 years / generation, and a mutation rate of 1.25e-8 / bp / generation. This is equivalent to using a mutation rate of 0.5e-9 / bp / year. Knowing these values, we can easily convert back and forth between their estimates and the number of mutations observed in the data.

Maximum Likelihood Estimation (MLE) of Lengths

One of the easiest ways to estimate a leg length is with a MLE estimate. Let's imagine we observer 10 mutations in a 10,000 bp block (or 1e4). For illustration, we can convert this to all the units we've mentioned, using the ArgWeaver values.

1. 1e-3 coalescent units (or 10 mutations / 1e4 bp).
2. 2,000,000 years (1e-3 coalescent units / 0.5e-9 mutation rate per year)
3. 800,000 generations (1e-3 coalescent units / 1.25e-8 mutation rate per generation)

In actual trees, it is a little more complex, because branch points have multiple legs. In this case, we are going to average lengths computed across the data in each leg; we are building an *ultrametric* tree (distance from tip to each leaf is the same). In this application, the *ultrametric* constraint makes a lot of sense because we all agree these alleles are related, and this gives a way to pool data together to get a higher confidence estimate that is not sensitive to population specific variation in mutation rates.

Nonetheless, these units are so trivially interchangeable, that they are not consistently used. While coalescence units is the most germaine to the data, it is also the most arcane. It is common for programs to use different units to display results more understandably. Argweaver and msprime, for example, use "generations."

Maximum A Posteriori (MAP) Length

MLE is great when we have lots of data, but it is very unstable when there is only small amounts of data. For example…

1. What if the number of bp we are looking at is really small, let's say exactly zero. In this case, what is the mutation rate? 0 mutations / 0 bp is undefined mathematically, and creates problems when taking recombination into account, some trees can end up having 0 bp spans in high recombination areas.

2. How about if the number of bp is just 100, and the observed mutations is zero. What is the mutation rate then? From the data we would say **zero**, but that's not true. We know it is low, but its not zero.

So how do we deal with these problems? **One way to solve this problem is to add a weak prior to the mutation rate computation.** There is a whole lot of math involved in analytically deriving this in a formal way (using a beta prior), but I'll show you a mathematically equivalent solution that uses something called ***pseudocounts***.

With pseudocounts we preload the estimate with some fake data, pseudo data. If the mutation rate is 0.5e-9 / year and we think this leg should be about 10,000 years long, we can use this to make our fake data. In this case, we will say the fake data is a 100 bp stretch, where we observed 0.0005 mutations (100 * 10000 * 0.5e-9). This is fake data so we can make fractional observations like this. We choose 100 bp to make this easily overwhelmed by the actual data.

Now, we estimate the mutation rate by looking at the data + pseudo data, instead of the data alone. If, for example, we are looking at no data. We would end up with a length of 10,000 years instead of the nasty undefined 0/0 we get in the MLE. Likewise, if we look at a real tree over a 2,000 bp region where 3 mutations are observed.

1. We can make a MLE estimate of the length in coalescent units, at 0.0015 (or 3 / 2000), which is equivalent to 3 million years.
2. We can also make MAP estimate of its length (using our pseudo counts), at 0.001428 (or 3.0005 / 2100, which is equivalent to 2.8 million years)

There are a few observations to make about this example.

1. These numbers can be converted into other units as discussed above. However, these are all based directly on observed data, the number of mutations in a region.

2. The MLE estimate and MAP estimate are pretty close. The more data there is, the closer they will be. They are *asymptotically* equivalent.

3. Even though our prior was 10,000 years, it's totally overwhelmed by the data in this case, to give an estimate of millions of years. We still do see a small imprint on the MAP estimate, pulling it downwards compared to MLE. By comparing the prior and the MAP estimate, we can know that the MLE estimate is higher than the MAP estimate in this case.

4. Only a few mutations is enough to increase the estimate of the length, which is why individual estimates have very high error (they will both be above and below the true value). We really need to see estimates from across the whole genome. Nonetheless, this example is not quite typical (just for illustration) and had 3 mutations in a tiny stretch of 2000 bp. That is a really high amount of mutations.

5. In the end, we want to choose a prior that will have little impact on the final results, but will help us in some of corner cases where things blow up in the MLE estimate. That is why were use a ***weak*** prior (low pseudocounts).

This is just an illustration, designed to be easy to understand without requiring statistical training. It is not precisely how argweaver works, for example, but is a very close theoretical analogy.

ArgWeaver Works Like MAP

ArgWeaver works very close to a MAP estimate. Our median TMR4A estimate is very much like a MAP estimate of TMR4A. What are the differences, however, with how ArgWeaver works from MAP…

1. The MAP and MLE estimates are designed to find the ***mean*** of the distribution, but we are using the ***median*** instead. The median, however, is more stable than the mode, and there are also some theoretical reasons for choosing this here too (it's an unbiased estimator in this case). Also, as we will see, using the median improves our estimate of error tolerance.

2. ArgWeaver is not making a single MLE or MAP estimate (as described above). Instead, it is sampling ARGs based on fit to the data (likelihood) and the prior. This called Markov Chain Monte Carlo (MCMC) and is closely related to a MAP estimate when a prior is used in sampling (as it is here).

3. ArgWeaver prior is not implemented using pseudocounts, instead they are using an explicit prior distribution. Using a prior distribution (rather than psuedocounts) is the preferred way of doing this, as it is less *ad hoc*, more flexible, has clear theoretical justification, and clarifies upfront the starting point of the algorithm.

4. The ArgWeaver prior does not use a fixed time (we used 10,000 years above), but a range of times. This is how the Ne comes in. They use the distribution of times expected from a fixed population of 11,534. I have no idea why the chose such a specific number, and note that I misreported this as 10,000 years earlier.

5. The ArgWeaver prior is on the time of coalescence, not the length of a leg in the tree. This is subtle distinction, but the TMR4A is actually the weighted sum of several legs in the tree. The prior ArgWeaver uses says that we expect (not having looked at data) for that TMR4A time (which is a sum of leg lengths in the tree) to be at about 100 kya. As implemented, it's a weak prior, and is overwhelmed by the data. Ultimately, the tree lengths computed by ArgWeaver are not strongly influenced by the prior.

6. Though I have explained this as actions on trees, ArgWeaver is applying this to branch lengths on the ARGs (the ancestral recombination graph). This is important because ARGs end up using more information (larger lengths of sequences) to estimate the length than naively trying to estimate phylogenetic branch lengths independently for each tree. The trees we have been using are an alternative representation of an ARG that is less efficient, but easier to use for many purposes (like estimating TMR4A).

In the end, to ease interpretation, ArgWeaver reports results in "generations" but its converting using the equations I've already given. So we can easily convert back and forth into any of these units. Most importantly, at its core, we are just using the fundamental formula:

```
D = R * T
```

Mutational distance is the product of mutational rate and time. That's all that is here. That is what enables the conversions.

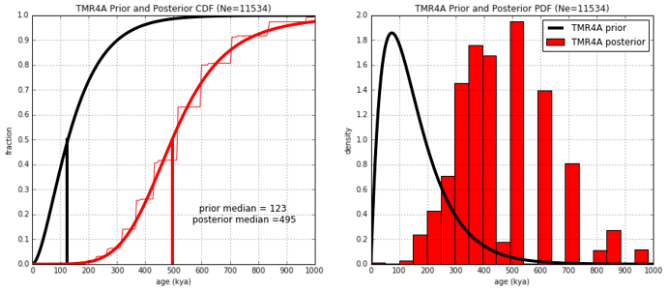
The fact that ArgWeaver makes the surprising decision to use Ne to parameterize its weak prior is just a non issue. As I have explained, the prior it uses for TMR4A is lower than TMR4A, so it's just pulling the estimate down any ways. Getting rid of it will only increase the estimate (only a small amount). MAP estimates, also, are considered vastly superior to MLE estimates, so it just makes no sense to doubt this result for using a better statistical technique.

A Prior Is Not an Assumption

It should be clear now why it is just incorrect (despite that footnote in the paper) to call a prior an assumption. It is also incorrect to say that argweaver is "simulating" a large population. All it is doing is using a *weak* prior on the tree lengths, and that is a good thing for it to do that makes the results more stable.

The language of prior and posterior is chosen intentionally. The terms are defined in relation to taking the data into account. In Bayesian analysis, the prior is updated by the data into the posterior. Then, the posterior becomes the new prior. We can then look at new data, to update it again. So priors, by definition, are not assumptions. They are starting beliefs that are updated and improved as we look at more data. It is just an error to call them assumptions.

With that in mind, we can visualize how the prior (black) for TMR4A is updated by the data into a posterior (red). Notice how the prior is lower than the posterior. That is how we know that using an MLE estimate would be higher than this MAP estimate, likely by just a small amount. Also notice that the Ne used here matches that used in ArgWeaver, which brings the median prior TMR4A to about 120 kya (not 100 kya as stated before).



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/9455c58eb63483733713bcd4cffc18b147799f14.png]

---

swamidass 2018-02-11 21:53:45 UTC #14

Systematically Correcting for Errors

One of my good friends in this conversation is **Dr. AJ Roberts[https://reasons.org/about/anjeanette-roberts]** from *Reasons to Believe*. She is a cautious and careful voice, and though we have our disagreements I pay close attention to her thoughts and questions.


[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/948e197145da85fad8df0438515435fb2879f2b2.jpg]

She raises an important point, which I will step through here. The reason is to show that the median TMR4A estimate is robust to concerns like raised by AJ.

What about the Whole Genome?

AJ makes the important point that this is not the whole genome. Most obviously, it does not include the Y-Chromosome and Mitochondrial DNA. Less obvious to those outside the field, “whole genome sequences” (WGS) are not actually the entire genome. About 10% of the genome (give or take) is not included in WGS. Why is that?

Why isn’t WGS the whole genome?

Some might be concerned about cherry picking: perhaps scientists are excluding data that does not fit some conclusion. That is not the case, however. The real reason is technical. Its very hard to sequence and assemble about 10% of the genome. Some of this DNA is very tightly bound to protein, and we cannot easily separate it. Some of this DNA is very repetitive, which makes assembling a final sequence very difficult. So the problem is not intentional cherry picking, but a real technical challenge.

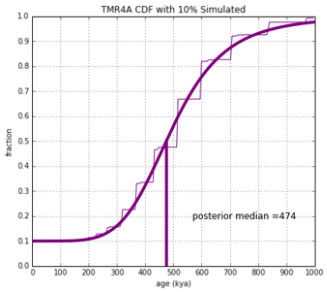
Still, its a valid question to ask how much adding this data could spoil our results. Maybe if we saw those missing DNA sequences it would change the results. How much could the results shift?

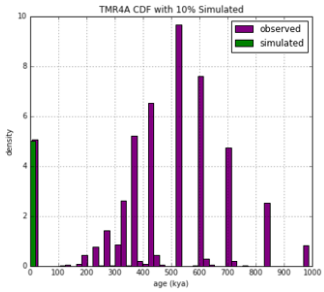
Adding in 10% Simulated Data

One reason I chose to use a median estimate (rather than mean or mode) is that it is remarkably stable, and also eases error analysis for questions just like this.

Let’s say there is 10% of the genome missing, and we can add that in someday. How much would the estimate be *expected* to change? Well, if the processes involved on 90% of the genome also apply to this 10%, we would expect to have nearly the same result. That, however, makes a guess about the data we have. Honestly, we know this DNA is different, so it is possible it is changing under different rules

A better way to frame the question is to ask if we got substantially lower TMR4A estimates from the hidden 10% of the genome, what is the ***maximum*** the median TMR4A estimate could possibly change? That turns out to be easy to answer. We know that the extended data (green simulated, and purple real), in this case, would just shift a little bit to the left the TMR4A median. We even simulate a case where the 10% extra DNA has a TMR4A of 1 generation (unrealistically low)…





[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/944b39ffab417deec32ad07ef296b50a45c377f8.png]

This means, ***at most***, adding in 10% more of the genome (green) would decrease the median TMR4A estimate by 20 kya, or about 2 kya per 1% of the genome we add. Essentially, this is a very very robust estimate that we do not expect to change much be adding in data like this.

What about the Y chromosome and Mitochondrial DNA?

These types of DNA, are different for 3 reasons.

1. They have no recombination, and for this reason do not require ArgWeaver to estimate the tree.
2. Because they have no recombination, are single locus measurements that only produce a single estimate, instead of the millions of trees that ArgWeaver gives us. Adding in two more estimates to this distribution has zero impact on the median TMR4A.
3. A Couple will only have ***one*** copy of each these pieces of DNA. So we do not want to look at TMR4A, but at TMRCA here. They provide an independent test, and end up being well over 100 kya. However, as single locus, the error on these estimates will aways be much higher than that which we get from genome wide TMR4A.

For these reasons, its best to use Y-chromosomes and mitochondrial DNA as an independent check for sanity, but not to refine these estimates. Likewise, these results on TMR4A, because they represent 12.5 million locations in the genome, are much higher confidence than the Y-chromosomes and mitochondrial DNA estimates.

Adjustment for Functional DNA

There are many other adjustments we can discuss. One example that will be interesting to AJ will be the effect of correcting for the percentage of the genome that is “functional”.

If the precise sequence of a part of the genome is important (because it is important to gene regulation or because it encodes a gene), then we will not see as much variation at those points. For this reason, we expect the TMRCAs at these regions to be much lower. One way to correct for this is by estimating the percentage of the genome with low TMRCA we should exclude from the analysis.

For example, we could say about 10% of the genome has enough function that it should be excluded from the analysis. Using our shortcut formula above, that would ***increase*** our TMR4A estimate from 495 kya to 515 kya. Not too much of a change, which is good news. I did not make this adjustment, though, because it’s not fair to only make adjustments in one direction. The fact is we can make adjustments in both directions, and they are going to (for the most part) just cancel out.

Of course those who think that more than 10% of the genome is dense with function with critical sequences, we might rule out even more of the genome than 10%. That could end up increasing the TMR4A estimates more.

An Easy Estimate to Adjust

For this reason, our estimate of 495 kya TMR4A is easy to refine. We can use this approximate formula:

```
495 kya + 2 kya per 1% genome adjustment
```

This gives us the ***maximum*** the model can change with new data. In fact, the change may be much much less. As we can see, this estimate is very stable to excluding large amounts of the genome, or adding more data in. We demonstrated this by computing the maximum change in median TMR4A from adding in 10% more of the genome, and applied it to compute how much functional parts of the genome could be biasing results downward.

The good news is that none of these effects are large. Tweaks here are not going to change our estimate much.

---

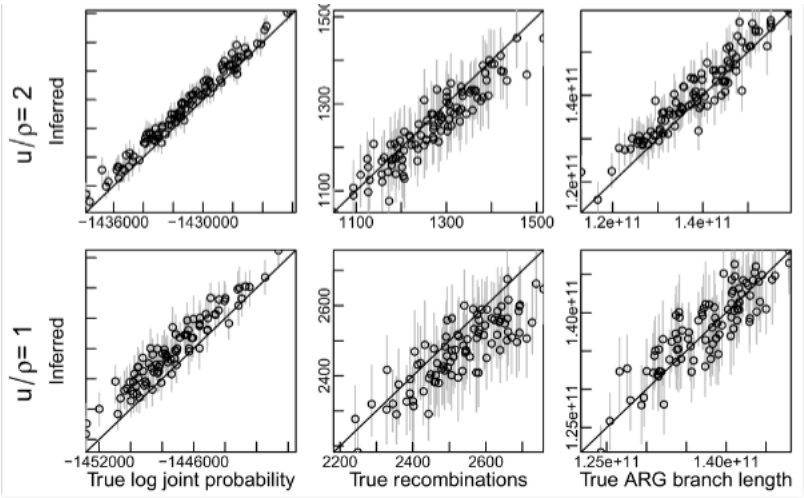
swamidass 2018-02-11 22:16:11 UTC #15

There were concerns offered by Richard about errors in inferring recombination.

> I think that the lack of detection of other recombination events will have an effect of the TMRCA, causing it to be overestimated.

What About Recombination?

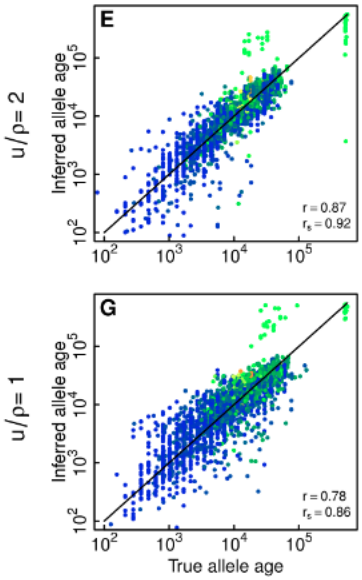
Sometimes ArgWeaver has a hard time identifying recombinations, but these do not have a strong effect on TMRCA estimates. Case in point is S4:



These figures show that when u/p (the ratio of mutation to recombination rate) is near one, some of the of the recombinations are missed (middle bottom). However, look at the figure (right bottom) where it shows the true ARG branch length. We see that the variance of the estimate increases but **there is no systematic error that is shifting estimates upwards from the true value**. This is a critical point. Remember we are taking the median of the TMR4A, which only depends on where the distribution is centered, not its spread. That is why we chose median, not mean. **This graph is very strong evidence that the recombination inference errors are NOT increasing the TMR4A.**

In fact, even when recombinations are underidentified (bottom row, middle column), ARG length is still measured about the same, but with higher variance (bottom row, right column). We chose an estimator that does not depend on variance, however, so this has zero impact on TMR4A.

We see the exact same pattern in S7.



The correlation between the true and estimated TMRCA drops a little, from 0.87 to 0.78, but if there is systematic error, it is very low. We cannot tell precisely this graph, *but we can from the prior graph*. **The median of each distribution is going to be very close to each other.** Taken with the prior figure, this is evidence that the under inference of recombination is not a major source of error. In fact, these data points show that recombination inference mistakes do not change the average/median TMRCA estimates. Also, TMRCA has much higher variance than TMR4A, and TMR4A will be even less susceptible to these types of errors.

> The clarify this last point, the recombinations that are detectable are those that will reduce the TMRCA substantially. Those that do not reduce the TMRCA are much more difficult to detect, so they are missed.

> It is also worth noting from Figure S8 that at mu/rho = 1 only about 60% of tree topologies are correct in ARGweaver. My guess is that incorrect trees are likely to be less parsimonious than correct trees, and so would elevate the TMRCA.

This hypothesis seems false. We can see from figure S4 and S7 that about 50% are less parsimonious than the correct tree (higher TMRCA) and about 50% are more parsimonious (lower TMRCA). Remember, we do not expect the trees to be precisely correct. There are just estimates, and we hope (with good reason) that the errors one way are largely cancelled by the errors the other way when we aggregate lots of estimates.

Finally, we are aggregating a lot of estimates together to compute the TMR4A across the whole genome. This is important, because by aggregating across 12.5 million trees, we reduce the error. While our estimate in a specific part of the genome might have high error, that error cancels out when we measure across the 12.5 million trees. This is a critical point.

The statistics here substantially increases our confidence in these numbers.

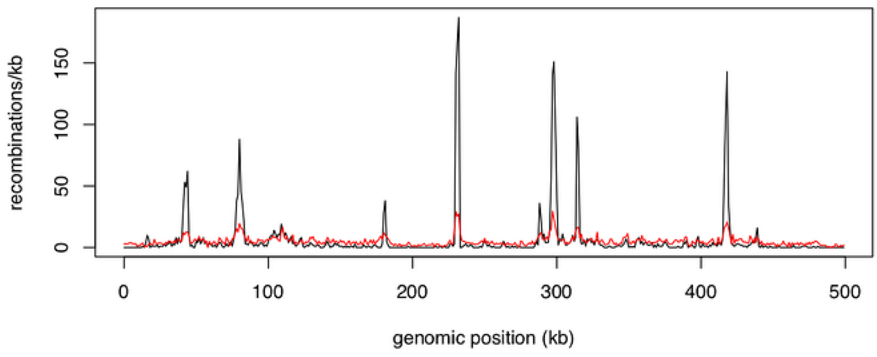
Just about any source of error we can identify will push some of the TMRCA estimates up and some of them down. However, because we are looking at the median of all these estimates, this increase in variance does not affect the accuracy much. A great example of this is mutation rates.

A Closer Look

> In the end, these figures validate pretty clearly my alternate hypothesis, which I formed based on knowledge of how this algorithm works.

I should also add that the referenced supplementary figures (S4 and S7) appear to be using only 20 sequences. Accuracy improves dramatically as more sequences are added. For the data we used, there were 108 sequences, so we expect better accuracy than the figures shown.

Also, S6 is an important figure, that shows the inferred vs true recombination rates for simulation using the known distribution of recombination rates across a stretch of the genome.



A few things to note about this.

1. For most of the genome, recombination rate is low (corresponding to high u/p), but only jumps up at recombination hotspots (the places where u/p is low).
2. The program is good at inferring recombination in most of the genome, but sometimes misses recombinations that make no difference to TMRCA in recombination hotspots.
3. The model picks up some of the recombinations, but not all of them. Most of the recombination inference errors are there, in the recombination hotspots, which are confined to a very small proportion of the genome.
4. At recombination hotspots, the trees will span shorter amounts of the genome than the rest of the genome. Trees with low bp span are signature for high recombination rate.
5. That means that most of the genome has a high u/p and is being estimated accurately, but there is only difficulty at recombination hotspots where u/p is low.
6. By weighting trees by the number of base pairs they cover, we can dramatically reduce any error that might be introduced by recombination inferences. That's because recombination hotspots are were the vast majority of the errors are, and these hotpots are just a few percent of the genome.

And that is exactly what I did. Rather than reducing the TMR4A estimate, downweighting the error prone recombination hotspots (by weighting by bp span of trees) ***increases*** the TMR4A estimate **(from about 420 kya ot 500 kya)**.

> > At the moment, all coalescents are weighted equally. This biases the averages to high recombination areas, which might be biased towards upward errors in TMRCA estimates. It would be wiser to average weighting by the length of the DNA segment to which the phylogeny applies.
>
> I finally got around to correcting this part of the code, and recomputing the TMR4A. Here is what we arrive at, a TMR4A of 495 kya, nearly 500 kya. This is a better estimate.

I'm going back over all this to point out I was already thinking about the effect of recombination and correcting for it in a plausible way. There are always sources of error in any measurement. This is no exception. The fact there is error however, does not mean the error is large. Clearly, we are only computing an estimate, but this is a good estimate of TMR4A.

Of note, correcting for recombination errors by downweighting recombination hotspots ***increases*** the TMR4A estimate (from about 420 kya about 500 kya). It does not decrease it. Thats because for trees spanning only short segments of the genome, they will be more influenced by the prior. That's because in short segments of the genome, there is not enough data/evidence to overwhelm the prior, so it takes over. On longer genome segments, the data is strong enough to disagree with the prior. As we have seen, the prior pull the TMR4A estimates downwards on real data. So in the end, reducing the effect of recombination hotspots just increases the TMR4A estimate. This is appropriate, because we want the TMR4A least dependent on the prior.

This may seem surprising, and in conflict with the the S4 and S7 data. It is not. In the S4 and S7 experiments, the prior matched the simulation, and did not pull the results up or down. In the real data, the prior pull the TMR4A estimates down, and pulls them down most in recombination hotspots because their bp spans are smallest. So this counterintuitive effect makes sense as an interaction with the prior and recombination hotspots. This error is important to understand, because **unlike most types of errors**:

1. it is biased in one direction (towards artificially lowering TMR4A)
2. its impact is large (about 70 kya, or about 15% relative effect)

Note, also, that I identified this source of error and corrected for it several weeks ago. Even in my first estimate, I disclosed it was going to be an issue.

> At the moment, all coalescents are weighted equally. **This biases the averages to high recombination areas, which might be biased towards upward errors in TMRCA estimates**. It would be wiser to average weighting by the length of the DNA segment to which the phylogeny applies.

Before I looked at the prior, however, I guessed wrong on the direction of the effect. I cannot identify any other sources of error likely to have this large an effect. Also, this adjustment was within my +/-20 confidence interval, which shows even my original estimate was not overstated.

Moreover, I have at this point corrected for it. A better correction might take this further, by just excluding the trees with small bp lengths, thereby excluding all regions where recombination rate is high. This refinement, will certainly increase the TMR4A estimate. I'm more inclined to improve this estimate with a different program first. That would likely have more value in the long run.

swamidass 2018-02-18 01:22:32 UTC #16

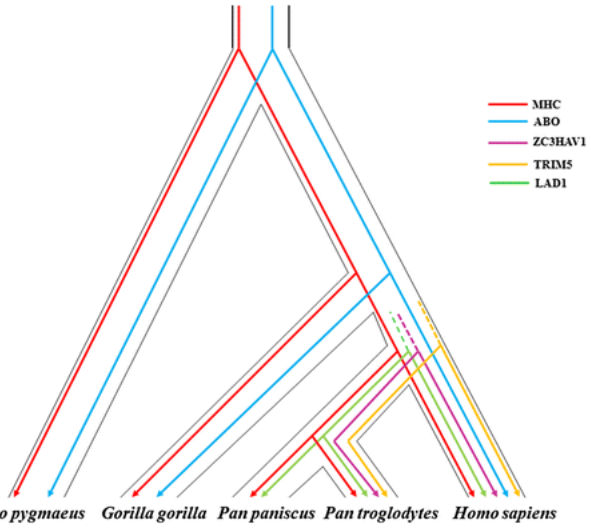
As we have seen, there is a limit how far back the evidence from **Human Variation** gives us confidence against a single couple bottleneck. Before about 500 kya, it is possible that such a bottleneck, if brief, would be undetected in by current population genetics models. The specific number may be adjusted upwards by further analysis, but it's a good starting point for now.

However, this is not actually the strongest argument put forward against a single couple bottleneck since we diverged from chimpanzees. For that, we have to look more closely at **Trans-Species Variation**.

Trans-Species Variation

Human Variation and Trans-Species Variation are related but different. To measure human variation, we look at a large number of human sequences. To measure trans-species variation, we look at a large number of both human and non-human sequences, usually chimpanzee. From looking at this data, we might find evidence of alleles that appears ***both*** in chimpanzees (for example) and humans.

This figure illustrates what appears to be happening in trans-species variation:



https://humgenomics.biomedcentral.com/articles/10.1186/s40246-015-0043-1[https://humgenomics.biomedcentral.com/articles/10.1186/s40246-015-0043-1]

The key point is that along each of the colored lines, *several* lineages are being shared between different species at a single place in the genome. Normally, there would be just one lineage on these time scales, but balancing selection maintains *multiple* lineages of alleles. By counting the number of allele lineages shared between humans and others, we can put a hard-stop lower bound on a bottleneck going back before humans and chimps diverge. Whatever bottlenecks there are they have to be big enough to include all the trans-species lineages.

Molecular Clock Not Valid

One tempting argument, which is not quite right, is to just estimate the TMRCA (or TMR4A) of these alleles, the same as we did across the genome, and use this as an estimate of a bottleneck time. This however, is an error, because the assumptions required for the molecular clock are not held in balancing selection.

Something called "balancing selection" is critical for enabling variation to last long enough to be shared this long between humans and other species, and this usually happens in proteins important for our immune response. So we see trans-species in only a few regions of the genome.

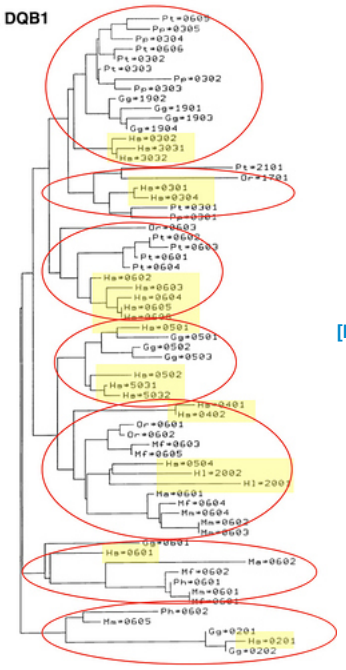
However, balancing selection violates the conditions required to accurately date variation in DNA. We cannot use our formula D = R * T here, because, in this case, we do not have a valid way of estimating R over these time frames. While in neutral regions of the genome, the average mutation rate works in our favor, at times we expect balancing selection to be increasing the rate of change in unpredictable and untestable ways. This can happen very rapidly as balancing selection can even select for increased mutation rates within this region.

Ayala's Argument Against a Bottleneck

The argument here is two part. First, from effective population size estimates, and second from trans-species variation. I'm not going to engage the argument about effective population size, because it appears to be incorrect. Very tight bottleneck can still have high effective population size, and it seems Ayala missed this point. But this just takes us back to the TMR4A work.

This is where **trans-species** variation becomes important. It gives an independent way of dating alleles. If an allele in humans is closer to non-human alleles, it appears that it existed before those two species diverged, and was maintained by balancing selection to this day.

This study by Francisco Ayala was the first, to my knowledge, to make the case against a bottleneck by studying trans species variation in HLA alleles.
https://www.sciencedirect.com/science/article/pii/S1055790396900135[https://www.sciencedirect.com/science/article/pii/S1055790



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/27c8a98f8d062548e2133c5cbdef85044bcaee7f.jpg]

This figure from Ayala shows human alleles with other primate alleles joined by ***similarity***, not phylogenetic analysis that respects nested clades. I've highlighted the human alleles in this figure, and drawn red circles around 7 clusters of alleles which appear to be shared between human and other species. Remember, we can only put 4 alleles at each position in the genome of a couple, so this seems (at least on face value) to demonstrate there must have been at least 4 individuals in the tightest bottleneck of our ancestors.

Ayala's summary is:

> Figure 4 is a genealogy of the HLA alleles obtained by the UPGMA method, which assumes constant rates of evolution and thus aligns all 19 alleles at the zero- distance point that corresponds to the present. The ge- nealogy suggests that **8 allele lineages were already in existence 15 Myr ago**, at the time of the divergence of the orangutan from the lineage of African apes and hu- mans; and that **12 allele lineages were in existence 6 Myr ago, at the time of divergence of humans, chimps, and gorillas**.

The difference between his numbers and mine in how we determine lineages. There is some ambiguity in how we determine the cutoffs. Still, as long as we see more than 4 lineages with trans-species variation, its seems like evidence against a single couple bottleneck. From this, he argues,

> There is, however, no evidence supporting the claim that ex- treme bottlenecks of just a few individuals, such as postulated by some speciation models (Mayr, 1963; Car- son, 1968, 1986), have occurred in association with hominid speciation events, or with major morphological changes, at any time over that last several million years.

This is probably correct, in that there is no evidence *for* a bottleneck that I can see. But he means here to mean that a bottleneck has not happened: i.e. there is evidence *against* a bottleneck in the last several million years. That may be incorrect.

Some Technical Asterix

Generally speaking, this work has been shown to definitely discount any notion of a single couple bottleneck. On face value, that is certainly true. However, there are some big caveats.

1. The molecular clock based dates computed in these studies, it does not appear can be trusted.

2. We do not really know the confidence on any of these clusters, because Ayala did not estimate them using modern bayesian methods.
3. He also used a similarity based method to build the trees, rather than a true phylogenetic reconstruction. This is important, because it can produce different clusters.
4. It does not appear convergent evolution was accounted for in this analysis. Convergent evolution, at this level, can create the appearance of shared history when there is none.
5. His population simulation used a bottleneck lasting 10 generations (e.g. 10 individuals for 10 generations), which is much longer than the bottlenecks we are considering (e.g. 2, to 10, to 500, to 2500, to 12500).

While these are interesting results, at some point, this analysis needs to be done with better methods to really determine how many lineages are persistent over the last 6 mya. Moreover, effort to correct for convergent evolution is important here too. On the simulation size, a brief bottleneck needs to be considered, rather than just those of 10 generations.

A Finding Not Replicated

Ayala focused his work on HLA-DBQ1 (one of the MHC genes), but similar work has shown trans-species variation at other locations in the genome. However, I could not uncover a single other study that shows more than 4 lineages with tran-species variation.

I cannot do a full review here, but we can see the balancing at other genes, with fewer lineages in the end. For example…

https://www.ncbi.nlm.nih.gov/pubmed/10866107[https://www.ncbi.nlm.nih.gov/pubmed/10866107]

This figure is fairly typical of findings…

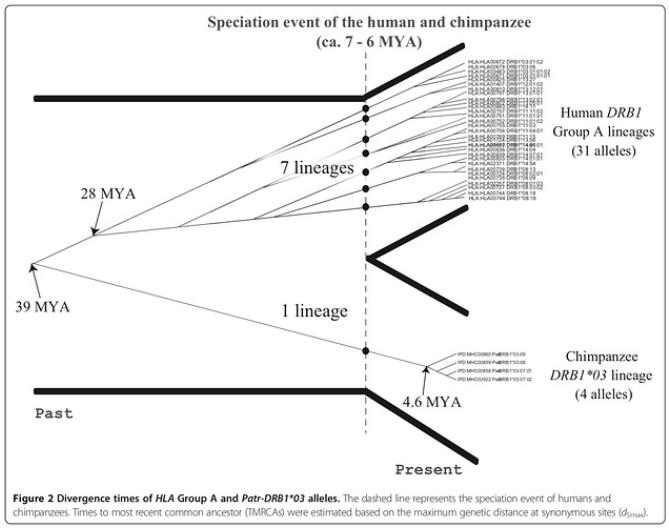
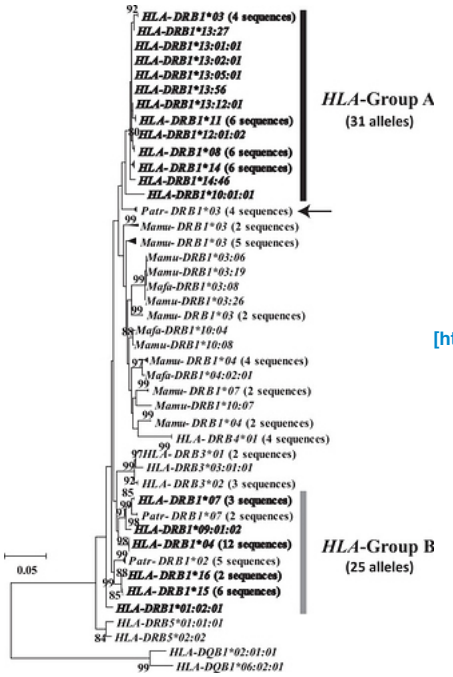


Figure 2 Divergence times of *HLA* Group A and *Patr-DRB1*03* alleles. The dashed line represents the speciation event of humans and chimpanzees. Times to most recent common ancestor (TMRCAs) were estimated based on the maximum genetic distance at synonymous sites (d_{synad}).

[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/9f2be821e13c62e325dbf21521bcb30a48729cbf.jpg]
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072476/[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072476/]

This figure shows a molecular clock based estimate (which do not appear well-calibrated) of 7 lineages at 6 mya, however, less than four lineages (only 0 in this case) are shared with chimpanzee. **Reviewing several papers, I cannot find replication of Ayala's findings of more than 4 lineages being shared between humans and other species.**

We can see this pattern in this figure too…

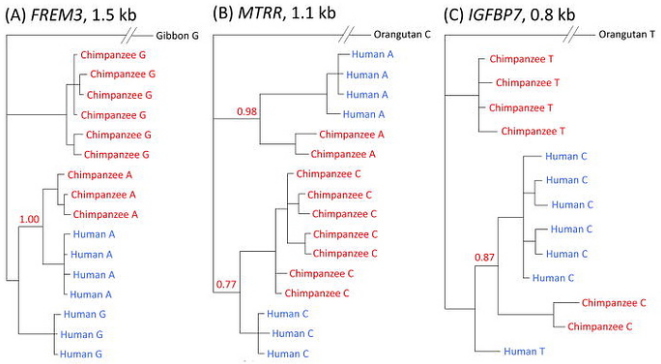


[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/d37694609c01353024a43336476adc5e6aacb8f7.jpg]

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072476[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072476]

Here, the bold leaves are human sequences. Notice the difference between this figure and Ayala's. There is numbers on the edges (which indicate confidence) and we just do not see nearly as many lineages in common. The authors here conclude there is just ***one*** lineage in common.

Here is another typical results figure:



[https://discourse-cdn-sjc2.com/standard17/uploads/peacefulscience/original/1X/88a45a888243e77cd2933c817f67eedc990fdd94.jpg]

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3612375/[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3612375/]

Each tree is a different region of the genome. The numbers on the tree indicate confidence in the tree at that place (unlike Ayala's figure). Notice, again, that there does not appear to be more than 4 clusters at a single gene with both human + chimpanzee alleles. Usually, we just see 1 or 2.

While Ayala is an established scientist, his work was done in 1996, well before modern sequencing efforts, and modern bayesian analysis of phylogenetic trees. While no one has published on DBQ1 since he did in 1996, it is very surprising that no one else has replicated his result in the last 22 years on another locus, or even the one he looked. **Of course, if someone can find a study that does, please let me know!**

The apparent failure to replicate this finding (with (1) much more data, which should make it easier, and (2) improved methods), discounts substantially my confidence in his findings. We just know much more about how to analyze DNA sequences, and we have so many more of

them than. It is not surprising that our understanding might advance from a paper published in 1996, before the genomic age.

One Line of Evidence? One Paper?

At the moment, the Ayala paper appears to be the ***only*** study which shows more than 4 allele lineages with trans-species variation. His analysis, however, did not estimate confidence nor did it use phylogenetics to determine lineages (it used sequence similarity instead). In 22 years, I cannot find a paper that replicates his finding with better methods (or any methods). Certainly, trans-species variation has been observed, but not more than 4 lineages, as far as I can tell.

This is not enough evidence by which to make a confident claim against a single generation bottleneck.

The Way Forward

The right way forward, then, is to study trans-species variation with the data we have now, but better methods than did Ayala. This takes some difficult work, however. I'm not 100% sure if we will give it a try here, but we might. This, also, is the most likely place a future study might uncover evidence against a single couple bottleneck.

Until that happens, however, I am not sure this is strong evidence against a brief bottleneck. I stand to be corrected, however, if someone can produce a study that shows this. If you find one, please send it to me.

---

swamidass 2018-02-19 04:38:18 UTC #18

I wanted to further expand on this deficiency in Ayala's study.

> swamidass:
>
> It does not appear convergent evolution was accounted for in this analysis. Convergent evolution, at this level, can create the appearance of shared history when there is none.

Convergent Evolution or Trans-Species Variation?

Convergent evolution, rather than shared history, is an alternate explanation of Trans-Specific variation. If we see several alleles in both humans and chimps clustered together, there are two possible explanations:

1. this could be because the allele lineaged existed in the common ancestor of the two species,
2. or it could be because of convergent evolution.

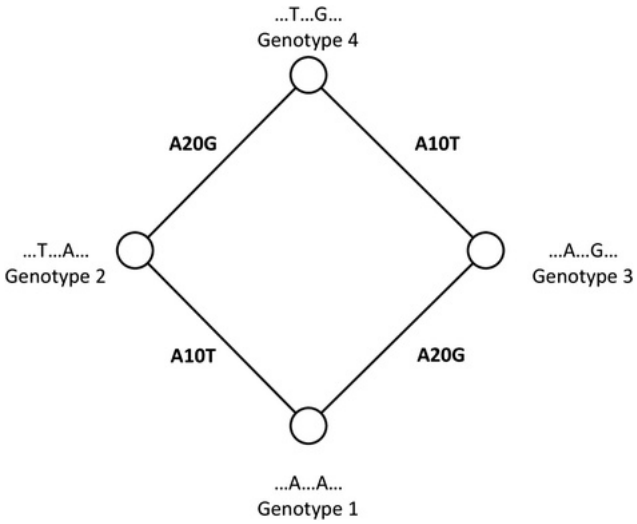
Ayala never considered or tested for this possibility. This is a critically important point, and why his use of similarity in phylogenetic analysis substantially undermines his point. In order to trust the tree, we need to know how many discordant mutations there are in the tree. However, he used similarity (not nested clades) to build the tree. He had no way to know if the data actually made sense as tree recapitulating common descent or not.

A basic feature of scientific thinking is to test hypotheses. Ayala's paper did not rule out the hypothesis of convergent evolution.

Testing for Convergent Evolution

The good news is that convergent evolution leaves a tell tale sign. We should see a large number of mutations that cannot fit into a tree like structure if convergent evolution is at play. It turns out that several groups have been studying convergent evolution on a genomewide scale, and HLA types regularly are outliers in these analyses.

For example, take a look at this study. **https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-016-0722-0[https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-016-0722-0]** It builds "allelic graphs", which I won't explain here in detail, except to say that when ever we see a cylcle, like this square, we know that it cannot fit into a tree structure:



https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-016-0722-0[https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-016-0722-0]

We expect a few of these in neutral evolution, but not many. If the data fit a tree, we would only see a single path from the top genotype to the bottom one. However, if we see both paths, we know that different alleles are taking different paths, which means that they do not actually share history here. It is a type of homoplasy, and it is a signature of convergent evolution. Notably, this signature arises to convergent evolution, is not likely caused by Trans-Species variation.

If we see a large number of these squares in the genetic diversity of a particular part of the genome, that is evidence that the similarity we see between sequences is not actually a signature of common descent. Rather, in these cases, another hypothesis is favored: convergent evolution.

So what do the authors find?

Well, HLA genes have a massive excess of squares, a clear sign of pervasive convergent evolution. Ayala's gene HLA-DBQ1 is not mentioned in the text, but we find it in the supplementary data as one of the genes with clear evidence of convergent evolution. In this case, it is not consistent with recombination or balancing selection alone.

Another gene, HLA-DRB1 is the most variable HLA gene. It is notable for having over 500 squares in the DNA of about merely 1,000 individuals, compared with an expected number of less than 10. That means if we had tried to put the DNA into a tree, we would see ***at least*** 500 mutations discordant with a phylogenetic tree. This is just a ***stunning*** result, because it means that HLA-DRB1 alleles are just not well described as a tree. The variation we see is evolving and re-evolving over and over again. Amazing.

It also validates my methodological concern about Ayala's work:

> swamidass:
>
> He also used a similarity based method to build the trees, rather than a true phylogenetic reconstruction. This is important, because it can produce different clusters.

It is just not an accurate view of the data to present HLA-DBQ1 in a tree based on a similarity matrix. One needs to test first to see if it actually fits a tree. We cannot even correctly determine ancestral history among human alleles themselves, let alone between species. The data seems to look more like convergent evolution than standard common descent, i.e. Trans-Species variation.

This is not exactly a new result, back in 2000, a test of Ayala's hypothesis was done on HLA-DBQ1. They also found strong evidence of convergent evolution. **https://www.semanticscholar.org/paper/Convergent-evolution-of-major-histocompatibility-c-Kriener-O'huigin/cf9f45169d245b7ab883a5a461f3a16fec62b751[https://www.semanticscholar.org/paper/Convergent-evolution-of-major-histocompatibility-c-Kriener-O'huigin/cf9f45169d245b7ab883a5a461f3a16fec62b751]** However, the allele graph makes clear how much this affects the data. Perhaps more importantly, this *Nature Genetics* study from 1998 directly disputes Ayala's paper, arguing that this is rapid convergent evolution: **https://www.nature.com/articles/ng0398-237[https://www.nature.com/articles/ng0398-237]**.

Remember, Ayala did not even consider convergent evolution. He did not test for it. This seems to a valid alternative hypothesis, which also seems to better explain the data.

Moreover it is not really accurate to present trans-species variation as a settled finding of genomic science. At best, it is one competing hypothesis among many. However, it might even be accurate to say that it is the disfavored hypothesis. There are many more papers

disputing Ayala's findings than supporting it. No one should present this as as indisputable and settled evidence against a sharp bottleneck.

Perhaps the data will bear out Ayala's initial hypothesis, but a lot of work needs to be done to demonstrate this to be the case.

What About Common Descent?

Everyone believes these alleles share common descent (at least back to 4 alleles). However, this is good reminder that genetic data can pick up signatures that erase the nested clade signature we usually see in DNA. Homoplasy is a real feature of the data, and expected even when there is common descent.

This is a great example of how there are rules in biology (e.g. DNA falls into nested clades), but there are exceptions (convergent evolution), that are very important to understanding this data.

Moreover, the next time someone points to mutations that do not fit the tree pattern in species, remember two things.

1. We expect a few discordant mutations, even in neutral evolution. That is not evidence against common descent.
2. Convergent evolution, also, can produce discordant mutations. Not usually ever as much as we see in HLA genes, but more than we expect from neutral evolution.
3. We observe homoplasy and convergent evolution in cancer (called recurrent mutations).
4. We observe homoplasy and convergent evolution in human variation (which everyone agrees shares common ancestry).

In case #2, we still expect to see a signature of common descent in most cases. However, it is such a pervasive pattern in HLA-DRB1 that it appears that the signature of common descent is erased, even though we all agree these alleles share common ancestry. And #3 and #4 are direct empirical evidence that convergent evolution is expected at a DNA level (#3) and that we homoplasy is observable in DNA everyone agrees shares common ancestry (#4).

Once again, the rule is that **most (but not all)** DNA fits into nested clades (a tree), but some does not. Neutral evolution produces nearly nested clade data, but positive selection (and balancing selection) can also lead to convergent evolution. Homoplasy (violations of nested trees) are ***expected*** in some DNA.

The Median TMR4A Estimate Unaffected

It's important to understand how these findings interact with the TMR4A esitmates.

The convergent evolution creates homoplasy that will artificially increase TMRCA estimates upwards. Because a tree is a bad fit for the data, it will be impossible to find a parsimonious tree. This will inflate the TMRCA values substantially. This reinforces what I've said from the beginning. The molecular clock, in these regions, is not well calibrated.

Another indicator of this is that a much larger fraction of mutations in this region are non-synonymous (i.e. not neutral). This is an indicator that positive selection is driving most of the changes a far more rapid rate than neutral evolution. The end result of this is artificially inflated TMRCA estimates. Remember, that D = T * R only in regions where dynamics like this are not taking place.

This does not, however, create a problem for our estimate of a bottleneck limit. Remember that we used the ***median*** of TMR4A over the whole genome. So, this estimate is not really influenced much by a small portion of the genome in error. The estimate shifts only about 2 kya per 1% of the genome in error. That is the reason we used the median in the first place, it makes the estimate remarkably stable to errors like this.

Convergent evolution is really the exception to the rule in human variation. It is not accounted for by most phylogenomic methods, but that does not matter in our genome wide analysis. Our final estimate is not strongly influenced by this problem.

---

swamidass 2018-05-27 20:04:55 UTC #19

---

swamidass 2018-07-11 11:48:18 UTC #20
