Correspondence between Eran Elhaik and Fernando L Mendez regarding the paper "An African American paternal lineage adds and extremely ancient root to the human Y chromosome phylogenetic tree"

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez flmendez@email.arizona.edu
date: Tue, Apr 2, 2013 at 10:50 PM
subject: questions regarding your recent paper

Hi Fernando

I have been reading your recent paper with great interest for the past couple of days and I was wondering if you can help me understand some of issues you discuss:

1. The mode you introduce to calculate L*Uy is a bit confusing.

T (63.2), the average mutations per trio, was estimated from parents who ages are 30 on average.
So, why the next component of your equation is b(1-g0/g)?
Which is the increment in mutations per year (per father age).
For someone who is 35 you will get a negative number, where in fact this person accumulated 10 more mutations since he was 30, so it should be 63+10

2. Using your model you generated many values and then divided them all by the length of autosomes. This gave you the mutation rate per year per base – are these the numbers in the second line of page 455?

3. You used the median value of Uy at age 30 and you used the range of 4.39 to 7.7 * 10-10 as lower and upper bounds, so far correct? how did you get 338kya from that?

4. All these steps had nothing to do with A00. You relied on Kong's numbers and built a tree, assuming A00 is the most ancient node. so far correct?

5. Did you estimate the divergence time between A0 and A00 to see that you get 338-202?

Thanks a lot

Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>

to: eran elhaik <eranelhaik@gmail.com>
date: Wed, Apr 3, 2013 at 3:58 AM
subject: Re: questions regarding your recent paper

Dear Eran,

Thank you for your interest in the paper

1. It was mostly a matter of lack of room for text in the supplementary material and not wanting to disrupt the flow. I attach a document with the longer explanation for the calculation of the mutation rate that we had before.

2. We used the model and the standard errors that we extracted from Kong et al. 2012 to generate the distributions from where we inferred the confidence intervals. In the second line of the second column of page 455 we report the maximal confidence interval (CI). You can see the medians and CIs as functions of the generation time in the Supplementary Figure 2.

3. The 338 kya comes from using the mutation rate per base per year, and the sequence coverage and number of mutations for each of the two branches A0 and A00 (they have different sequence coverage). The lower and upper bounds are conservative: we use the low end of the CI for the mutation rate when estimating the upper bound of the TMRCA and the high end when estimating the lower bound for TMRCA.

4. "4. All these steps had nothing to do with A00. You relied on Kong's numbers and built a tree, assuming A00 is the most ancient node. so far correct? "
I have no idea of what you mean here. Kong's numbers were only used to estimate the mutation rate per base per year. The TMRCA estimate has everything to do with both A00 and A0. A00 is not a node, but a haplogroup, and we estimate the age of the common ancestor of the haplogroups A0 and A00. Because other than A00 everyone shares a more recent common ancestor with A0, we are estimating the TMRCA of all known human Y chromosome lineages.

"5. Did you estimate the divergence time between A0 and A00 to see that you get 338-202?"
Again, I am not sure what you mean, but I assume that you are referring to the time between the most recent common ancestor of all lineages and the time of the common ancestor of A0 and the reference. The estimate of 338 ky uses information from both lineages as does the 202 ky. Because of that a direct estimate of the time between the branching events only is not expected to produce the 338 ky - 202 ky. However, the confidence interval should contain that value, even using only the median mutation rate. Given that the segment of the tree contains 18 mutations, the confidence interval for the expected number of mutations is very broad, and the corresponding interval for the time naturally contains the 338 ky-202 ky.

Do not hesitate to write me again if any of my answers is not clear.

Regards,

Fernando

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Thu, Apr 4, 2013 at 8:38 PM
subject: RE: questions regarding your recent paper

Dear Fernando,

Thank you for your detailed answers, I have some few more questions

1. Thanks for showing me the detailed steps, but my main question remains. Your first component (K+b*g0 + F) is the average #new mutations per genome. it is composed of the number of mutations you are born with (K) + those of the father times his age (b*g0) + those of the number. Why then in the second component you are adding the father again?

2. Ok

3. Maybe I am missing something, but are the regions in Table S1 supposed to be the PAR regions, they look outside of them…? Is there a way to see the exact numbers used for the calculation?

Thanks again
Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>
to: eran elhaik <eranelhaik@gmail.com>
date: Fri, Apr 5, 2013 at 12:32 AM
subject: Re: questions regarding your recent paper

Dear Eran,

The PAR regions are the pseudo autosomal regions, which are homologous to the X-chromosome (they recombined with it). The X-degenerate portion of the Y chromosome is the part of the Y chromosome that was originally homologous to the X chromosome but no longer recombines with it. It also contains insertions from

other parts of the genome, but evolves independently of those other parts. It is also important to distinguish the X-degenerate from the X-transposed portion of the X chromosome.
A good explanation of these different parts of the human Y chromosome regions can be found here:
http://pagelab.wi.mit.edu/pdf/2003%20-%20The%20male-specific%20region%20of%20the%20human%20Y%20chromosome%20is%20a%20mosaic%20of%20discrete%20sequence%20classes.pdf

No sequence in the study is part of the PARs.

Here is the R code used to calculate the TMRCA (only the maximum likelihood)

```
###############
#mut A0 in Mendez et al.
k1<-45
#mut rate for A0 in Mendez et al.
mRateMA0 <- 0.0001118923

#mut A00 in Mendez et al.
k2<-43
#mut rate for A00 in Mendez et al.
mRateMA00 <- 0.0001487211

################################################
#values for age
seqTMRCA<-seq(220000, 580000, by=100)

likeTMRCA<-real()
for (i in seqTMRCA){
   likeTMRCA<-c(likeTMRCA, -(mRateMA0+mRateMA00)*i+(k1+k2)
*log((mRateMA0+mRateMA00)*i))
}
########################
```

likeTMRCA contains the information on the likelihood of the TMRCA. It is graded every 100 years and starts counting at 220000 years.
The mutation rates come from the point estimate of the mutation rate and the length of the sequences (reported in the paper).

You can obtain the values for the confidence intervals when you use the high and low mutation rates (the per locus mutation rate would be proportional to the per base mutation rate), and the condition for the likelihood values at the end of the confidence intervals (you can get those from any Statistics book).

####################
In what respects to the calculation of the mutation rate:

The mutation rate per generation in the autosomes originating in males is :
the mutation rate per year times the length of the haploid genome times the generation time.

If you divide by g (the generation time) you have the equation in the text. The rest is just rearranging (you can check the calculation). In the end I have variables for which there are estimates in Kong et al. 2012.

I am not adding any strange term that I do not subtract.

Finally, let me point that K has no biological meaning. K+b*15 (assuming fertile age is reached at 15 years) is the minimum number of mutations passed by males. For all what matters K could be negative. From 15 on the number of added mutations is linear (in the range of ages that I considered).

Best,

Fernando

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Sat, Apr 13, 2013 at 4:26 PM
subject: RE: questions regarding your recent paper

Dear Fernando

Thank you so much for the explanation, I got a bit confused.
I just have few more questions

You sequenced 240kb of A00 and only 180kb of A0, aren't you concerned that there will be more mutations in the unsequenced region of A0 so the count would be higher?

When you calculated the ancestral/derived mutations between A0 A00 and the chimp, how did you address the unsequenced regions of A0 for which there are no markers to compare..?

In your code, how did you set 220000 and 580000? Where did you get these estimations…?

I ran the code and likeTMRCA ranges from 290-306… wasn't I supposed to get 338? I am trying to understand how to obtain the 338.

2. In the supplementary material you mention A000. Do you really have such sample? Why didn't you include it in the analysis?

3. In Figure 1, you wrote 125 as the split time between A0 and the ref, but in the text you wrote the Cruciani used 142k, which is correct. where did the 125 number came from?
4. You note that "the higher mutation rate produces an estimate for the common ancestor of all non-African Y chromosome haplogroups (C through T) of ~39 kya where did you get the 39k number, its not in Cruciani…?
5. how did you obtain the 209k number in Figure 1? Your numbers were obtained from a range of which you used the median and you also had upper and lower bound. Did you plug the mutation rate instead of the number you obtained from the median?
6. how did you obtain the 95% CI for the A00 rooted tree using the numbers you calculated using Cruciani's mutation rate? Your other paper (cited as 26) is not about the y-chromosome…?
7. in your STR analysis you found that the ancestor of AA and Mbo A00 lived 2-73kya, this is a huge range, any idea why if the sequences are so identical as you noted…?

Many thanks and sorry for all the questions

Eran

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Sat, Apr 13, 2013 at 7:34 PM
subject: RE: questions regarding your recent paper

Dear Fernando

I forgot to ask, what is the pan-African dataset that was used…?
I wonder if they have autosomal data

Thanks,
Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>
to: eran elhaik <eranelhaik@gmail.com>
date: Sat, Apr 13, 2013 at 9:59 PM
subject: Re: questions regarding your recent paper

Dear Eran,

1. Part of the 240 kb in A00 that are not in A0 do contain mutations. In fact, most of those 60 kb were chosen (not by me) because they were known to contain SNPs in A0. Choosing the regions because they are known to be variable in your samples would introduce bias (overestimation of the amount of variation). That is why we could not use them in A0. However, using different amounts of sequence in each individual does not introduce a bias. You can verify in the code that I sent you that the mutation rate per genome is higher for A00 than for A0. This is entirely due to the different amounts of sequence available for each of those. Because we have the chimpanzee sequence we can identify which mutations occur in either lineage (A00 or A0).

2. We do not have any A000 sample, and we are not aware that any has been sampled and genotyped. We include the A000 sample to exemplify how the nomenclature would proceed according to our proposal.

3. As I mentioned the sequence in Cruciani et al. 2011 and our sequence for A0 do not overlap. The estimates are based on different sequences which contain different mutations, but assuming the same mutation rate (10^-9). The 125 ky comes from our data, the 142 ky from Cruciani et al. (2011). Even if the sequences had the exact mutation rate, it would be unexpected that the two estimates match exactly. However, they do have to be consistent with each other (in terms of confidence intervals), which they are.

4. The estimate is in fact in Cruciani et al. 2011. It is based on their data (we do not sequence C or R in our work). We put the citation just after the estimate. From Cruciani et al. (2011) you cannot infer that all the mutations common to C and R are also shared by E, but I verified that using one of the multiple Y chromosome genomes from E haplogroup (they are shared).

5. We have this piece of text in the second page "If we were to use the higher mutation rate (1.0*10^ 9 per base per year) rather than a realistic range derived from whole-genome sequencing (4.39*^-10 <= muY<= 7.07*10^-10), the estimated TMRCA for the tree incorporating A00 as the basal lineage would be 209 kya, which is only slightly older than current estimates of the TMRCA of mtDNA and the age of the oldest AMH fossil remains."
The 209 ky is a point estimate using a point esimate for the mutation rate: the one that has been used before, without confidence intervals, just as it has been used before.

6. The method is explained in the referred paper. The only requirement is that the sequence has to be phased, which you get if you use Y chromosome, or a homozygous sequence. I would refer you to the paper, though. You can get it from my webpage (linked at the bottom of the email). We did not apply confidence

intervals to estimates based on the fast mutation rate. Those estimates are listed just for the purpose of comparison.

7. The range you mention derives from the analysis of the sequences, not the STRs. The sequences are not identical: the Mbo do not share at least to of the mutations discovered in the AA A00 chromosome. The ranges, which are broad result from the limited amount of sequence available (only 240 kb) and the uncertainty in the mutation rate.

"In your code, how did you set 220000 and 580000? Where did you get these estimations…?" Those are not estimates, are the range of ages on which I calculate the likelihood. The ranges were chosen to include the edges of the confidence intervals, but such that there is no overlapping with the age of A0 (for the nested estimation described in the referenced paper).

The likeTMRCA function calculates the likelihood of the different ages. The maximum corresponds to 337800 (if I recall correctly). Just look for which (likeTMRCA=max(likeTMRCA)), then subtract 1, multiply the number by 100 and add 220000. That should give you the age with the MLE. For the CI it is more complex because you also have to scale the mutation rate, and look for which age gives you values are 1.92 smaller than the maximum (the function calculates the log-likelihood)

The pan-African database used belongs to Mark Thomas at UCL. It contains data from a number of populations from the countries listed. My understanding is that it does not contain autosomal information.

Best,

Fernando

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Sat, Apr 13, 2013 at 11:12 PM
subject: RE: questions regarding your recent paper

Dear Fernando

Thank you so much for your kind answers.

Regarding to 1) I still don't see why not sequencing the remaining 60k in the A0. You had the chimp and human reference, if there are more ancestral mutations in A0, this is informative.

Regarding the last question) you wrote:

The likeTMRCA function calculates the likelihood of the different ages. The maximum corresponds to 337800 (if I recall correctly). Just look for which (likeTMRCA=max(likeTMRCA)), then subtract 1, multiply the number by 100 and add 220000. That should give you the age with the MLE. For the CI it is more complex because you also have to scale the mutation rate, and look for which age gives you values are 1.92 smaller than the maximum (the function calculates the log-likelihood)

I followed your instructions:

```
#Mutation rate

###############
#mut A0 in Mendez et al.
k1<-45
#mut rate for A0 in Mendez et al.
mRateMA0 <- 0.0001118923

#mut A00 in Mendez et al.
k2<-43
#mut rate for A00 in Mendez et al.
mRateMA00 <- 0.0001487211

################################################
#values for age
seqTMRCA<-seq(220000, 580000, by=100)

likeTMRCA<-real()
for (i in seqTMRCA){
  likeTMRCA<-c(likeTMRCA, -
(mRateMA0+mRateMA00)*i+(k1+k2)*log((mRateMA0+mRateMA00)*i))
}
########################
likeTMRCA=max(likeTMRCA)
100*(likeTMRCA-1)+220000
```

But the results are 250k

Which MRCA model equation did you use? Any reference?

I am curious, why did you not use Walsh's approach or a similar stochastic model? Did you compare your results with a Bayesian method?

Thanks again
Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>
to: eran elhaik <eranelhaik@gmail.com>
date: Sun, Apr 14, 2013 at 12:05 AM
subject: Re: questions regarding your recent paper

Dear Eran,

I made a typo. The command should have been:
(which(likeTMRCA==max(likeTMRCA))-1)*100+220000
(I should have put the "==" instead of "=")

" I still don't see why not sequencing the remaining 60k in the A0". It was sequenced, it just was not used here. The reason stands the same. The amount of evolution on A0 would have been overestimated. Rather than correcting for the bias, which would have made the calculation fairly more complicated, I chose to remove the bias by removing the sequence. As you may imagine, not much precision is lost (the CI would have had a width that is
~ sqrt((240+180)/(240+240)) ~ 0.935 of what we report. The gain would have been marginal.

"if there are more ancestral mutations in A0, this is informative." I don't think I understand what you mean here.

"Which MRCA model equation did you use? Any reference?" The reference is 20 in the paper, another paper in AJHG. Look at the appendix. The method is explained there (for the joint likelihood, which also gives you the estimate fot the TMRCA).

"why did you not use Walsh's approach or a similar stochastic model? "
I am assuming you refer to the 2001 paper in Genetics. That model, and most if not all Bayesian models used so far make important assumptions on the demography of the population (e.g. BATWING). Some of the assumptions might not be horribly problematic (like the effective size). Other assumptions may drive the estimate (e.g. assuming panmixia when you have a highly structured population). Because we are looking at a time for which demography is poorly known, a Bayesian approach here has a considerable chance of being inadequate. Furthermore, we are not looking at the TMRCA of two randomly sampled chromosomes; these have been chosen because they are divergent. It might be the case that there is a Bayesian approach that takes into account how the samples were selected and that the demography is unknown. I am not aware of that approach, but would be happy to hear about it.

Best,

Fernando

P.S.: For the calculation of TMRCA you can always get a rough idea just doing
(45+43)/((240+180)*1000*6.17/10000000000)
which is the total number of mutations divided by the total mutation rate per year when you consider jointly the chromosomes of A00 and A0.

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Tue, Apr 16, 2013 at 12:39 AM
subject: RE: questions regarding your recent paper

Dear Fernando,

Thank you now I am getting the right numbers!

One more small question, in the first paragraph you wrote:
"Genotyping of a DNA sample that was submitted to a commercial genetic-testing facility demonstrated that the Y chromosome of this African American individual carried the ancestral state of all known Y chromosome SNPs"
However, looking at table S1, there are branches that start with A0 or A0T.
Maybe I am not reading this table very well.

What is the value of A00 in the case of:
6991670

6991670

A0

L1073

T->C

The A00 has a total of 43 mutations, and the A0 45 mutations.
Do you mean to say that all the 43 are ancestral?
Can I say that all the 45 mutations of A0 are ancestral?

Thanks again,
Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>
to: eran elhaik <eranelhaik@gmail.com>
date: Tue, Apr 16, 2013 at 1:19 AM
subject: Re: questions regarding your recent paper

Dear Eran,

The "ancestral state of all known Y chromosome SNPs" refers to the SNPs known at the time the sample was first genotyped. The 45 mutations in A0 are mutations at which A0 is derived since its common ancestor with A00. The 43 mutations are mutations at which A00 is derived since its common ancestor with A0 (and with all known human Y chromosomes that are not A00). A0-T indicates mutations at which A00 is the only lineage that is ancestral. Those are mutations that happened in the lineage leading both to A0 and the remaining Y chromosome lineages (i.e., A1-T). Mutations indicated as A0 are mutations that are observed only in chromosomes that are in the A0 haplogroup, although some of them might not be shared by all A0 Y chromosomes. The haplogroup A0 is known to be variable. See, for example: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0049170;jsessionid=276061790215941B887BA36030CD1C15

In the specific case of the mutation that you mention (L1073). A0 has the derived allele C, and A00 shares the ancestral T with chimpanzee and the reference sequence.

I don't know what you mean here "Do you mean to say that all the 43 are ancestral? Can I say that all the 45 mutations of A0 are ancestral?"
but I hope that my previous explanation clarified the issue of ancestral and derived mutations for these lineages.

Best,

Fernando

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Tue, Apr 16, 2013 at 4:59 PM
subject: RE: questions regarding your recent paper

Dear Fernando

Thank you again, this is most helpful.

1. Regarding the: The "ancestral state of all known Y chromosome SNPs" refers to the SNPs known at the time the sample was first genotyped.

Did you find new SNPs after the genotyping in which A00 was not the ancestral?

2. How can that statement be true, if there are 43 mutations at which A00 is derived?

3. If I understand you correctly, of the 45 mutation at which A0 is derived, only in 18 (A0-T) , A00 is the only lineage that is ancestral. In the other 27, A00 is ancestral but there are SNPs in other chromosomes of the A haplogroup that are ancestral. Is that correct?

Many thanks,
Eran

---

from: Fernando L Mendez <flmendez@email.arizona.edu>
to: eran elhaik <eranelhaik@gmail.com>
date: Tue, Apr 16, 2013 at 11:10 PM
subject: Re: questions regarding your recent paper

Dear Eran,

1. All the mutations that are indicated as A00 are derived in the A00 sample, and were found after the chromosome could not be assigned to any known haplogroups back them. The SNPs were discovered by sequencing. Think of the sentence that you quoted as referring to SNPs known before this Y chromosome was ever sampled.

2. Again, the statement is referring to before the Y chromosome was ever looked at.

3.
"If I understand you correctly, of the 45 mutation at which A0 is derived, only in 18 (A0-T) , A00 is the only lineage that is ancestral."
CORRECT

" In the other 27, A00 is ancestral but there are SNPs in other chromosomes of the A haplogroup that are ancestral. Is that correct? "

For the other 27 SNPs, the A0 sample is derived, the A00 is ancestral. If anything else than A0 has a mutation in that position, it is inferred to have occurred independently.

Best,

Fernando

---

from: eran elhaik <eranelhaik@gmail.com>
to: Fernando L Mendez <flmendez@email.arizona.edu>
date: Wed, Apr 17, 2013 at 4:55 AM
subject: RE: questions regarding your recent paper

Dear Fernando

1) I don't understand, the statement is from the first line of the abstract and clearly refers to the Y chromosome AFTER it was sequenced.

What do you mean "All the mutations that are indicated as A00 are derived in the A00 sample?" so none of them were ancestral?

I believe I understand what you mean, but then again there are A00 derived mutations for which A0 or some other A lineage have the ancestral mutation – which do define the basal portion of the Y chromosome. Is that correct?

Maybe you can rephrase the sentence to read before and after the A00 was sequenced?

Thanks a lot
Eran
