Supplementary Note

Choice of X-degenerate sequences analyzed in Mendez et al. (2013).

The sequencing of A00 was performed at FamilyTreeDNA. The primers used to PCR-amplify the DNA for sequencing analysis were chosen for two reasons: to generate thousands of base pairs of new DNA sequence and to genotype a large number *known* Y chromosomes variants. This design targeted a larger region than what was eventually used. Because the assay was automated, failed PCRs were not repeated, resulting in some variation in sequenced regions across individuals. Unfortunately, this sequencing strategy also resulted in resequencing data with a bias that, if not corrected or accounted for, could have lead to biased inference. Specifically, estimated levels of variation could be upwardly biased for lineages whose known variation was used in the design of primer (see example below). The A00 haplogroup is an outgroup to all lineages used in the ascertainment of genetic variation, and thus the expected number A00-derived variants is not affected by the ascertainment scheme. Before Mendez et al., only a single study had performed extensive resequencing of an A0 chromosome (Cruciani et al. 2011), with some of the SNPs reported in that study being targeted during the primer design stage. To minimize the potential for biases, we excluded from the analysis of A0 the regions that overlap with those used in Cruciani et al. (2011), resulting in the exclusion of ~49 kb from the analysis of A0. The incomplete overlap of callable sequence between the A00 and A0 chromosomes results in further ~12 kb that are not analyzed in the A0 chromosome. As indicated above, the part that was callable in A0 but not in A00 could not be used for the analysis. Naturally, less sequence is analyzed in A0 than in A00.

Finally, we limited the sequence analyzed to non-coding regions of the X-degenerate portion of the Y chromosome. The reason for this choice is that we estimate the mutation rate with the assumption that in males the same processes govern mutation on the autosomes and on the Y chromosome. The X-degenerate portion has an ancient homology with the X chromosome, whereas the X-transposed portion has been duplicated from the X chromosome since the divergence with the chimpanzee Y chromosome, and the ampliconic region is highly repetitive (Skaletsky et al. 2003).

Example

In the main text and above we indicate that estimates of genetic variation (in the form of divergence from the ancestral sequence) assessed in genomic regions that were chosen *because* they were known to carry variants may be inflated. Here we provide an example. Assume that each of one hundred 500 bp regions were chosen as to contain a random SNP differentiating Y chromosome sequences in

haplogroups R-M269 and E-P1. If we analyzed the sequence evolution in lineages closely related to haplogroups R-M269 or E-P1, we would observe an inflation in the estimated amounts of sequence divergence since the common ancestor. The reason for this is that, regardless of how much lower the density of variants is on average on the Y chromosome, each of these regions was known *a priori* to harbor mutations separating these branches, resulting in at least one mutation per region. The corresponding estimated levels of diversity would be about twice as large as the average for the autosomes. If we now considered a Y chromosome in haplogroup I, the inflation in the estimate of sequence evolution would be smaller (but still present), and would be caused by the subset of the regions having SNPs that are shared between haplogroups I and R. However, if we sequenced a chromosome in haplogroup B, there would be no excess in the number of SNPs derived in this lineage, because, as B haplogroup branches off before the common ancestor between E and R, our choice of regions to sequence would be agnostic to variation private to haplogroup B.

Additional problems

1. “For example, Xue *et al*. sequenced approximately 10.15 Mb from two Y chromosomes of two European individuals separated by 13 generations.”

The individuals considered in Xue *et al*.[1] are Chinese, as it is explicitly stated in the second sentence of the body of that paper.

2. “This estimate is consistent with a estimates derived from human-chimpanzee Y chromosome analyses (1.5×10^{-9} - 2.1×10^{-9} substitutions per nucleotide per year) under the assumption of a divergence time range of 5-7 million years.” [2,3]

The second reference does not deal with the mutation rate of the Y chromosome. The first reference compares sequences from the X-degenerate and the ampliconic parts of the Y chromosome.

3. “... Mendez *et al* [4]... assumed the complete lack of purifying or advantageous selection on the Y chromosome.”

The statement above is false.

4. “In the Mbo samples described by Mendez *al* [4] as having A00 haplotypes, the main contribution to the variation was by genetic drift. Consequently, these haplotypes are derived from a common ancestor who lived only a few centuries ago (Mendez *et al* [4]), not 338 000 ya, when the common ancestor of A00 population was reported to live.”

These sentences make no sense; genetic drift does not contribute to genetic variation, but to its removal. Mendez *et al* [4] report a relatively recent common origin for all A00 chromosomes surveyed, and even more recent in Mbo. Mendez *et al* [4] never reported any “A00 population”, which by itself makes no sense.

5. “Less developed nations exhibit much shorter generation times (in the low 20 s)”

Elhaik *et al* [5] cite a work entitled “Mean age of mothers at first childbirth 2012”. Maybe they need to remember that Y chromosome is passed by the fathers, not the mothers, and that the mean generation time is not the same as the age at first childbirth.

6. “By using a lower bound of 20 years, an average of 30 years, and an upper bound of 40 years, Mendez *et al* [4] reduced the number of generations per unit time, and further inflated the TMRCA estimate.”

Figure S2 in Mendez *et al* [4] shows clearly why the above statement is incorrect: the inferred mutation rate per year increases with generation time, therefore resulting in a lower TMRCA estimate for a larger generation time.

7. “The Kong *et al* [6] data contain five data points from which to compute the maternal rate of mutations per generation....From the these five data points, Mendez *et al* [4] calculated a 'median' rate of 14.2 and a 'standard deviation' of 3.12.”

There is no apparent reason why Elhaik *et al* [5] could interpret what they state from what is written in Mendez *et al* [4]. The word “median” was used three times in Mendez *et al* [4]: 1) “... median values and generated 90% confidence intervals (CIs) for μ_Y as a function of g...”, 2) “median value of μ_Y at age 30 years”, 3) “A median-joining network ...”. Obviously, none of them refers to the number of maternal mutations (“F” in Mendez *et al* [4])

8. “Had they used the correct prediction confidence interval with 95% or 99% confidence, the number of mutations per generation in the female lineage would have been 0.52 to 27.88 and -3.78 to 32.18, respectively, where the sign – denotes 'minus'. The perplexing range -3.78 to 32.18 is due to the assumption that the number of mutations per generation follows a normal distribution. That is, the normality assumption in Mendez *et al* [4] results in the time-bending possibility that the most common ancestor of all the Y chromosomes in the world has yet to be born.”

As mentioned in the main body, Mendez *et al* [4] were estimating the mutation rate, not predicting new observations, and therefore prediction confidence intervals are not appropriate. Even with the absurd negative sign, the estimate of the mutation rate per year in the Y chromosome would be positive. Of course, the last sentence is a complete nonsense.

9. “... the authors chose to omit 60 000 bases of it because they consist of 'a large amount of mutations' (FLM personal communication).”

The statement is utterly false, and the quote is a complete fabrication of Elhaik *et al* [5].

10. “We believe that matching chromosomal regions should be compared instead of eliminating particular regions in an attempt to make the data fit a preconceived model.”

It never crossed the minds of Mendez *et al* [4] to do anything to attempt to make the data fit a preconceived model. Apparently, Elhaik *et al* [5] may considered it a viable possibility, even though they express their (sincere?) opposition to it.

11. We further speculate that omitting regions for one lineage, but including them for another may have reduced their estimated age.”

Elhaik *et al* [5] prove their speculation to be wrong in the following paragraph.

12. “... we show that the TMRCA calculation using equivalent regions of A0 and A00 yields a much lower estimate than that reported by Mendez *et al* [4].”

The comparison between values using different mutation rates is, of course, invalid (apples and oranges). When using the same mutation rate, they obtain an unsurprisingly excellent agreement: 209 500 (95 % CI = 168 000-257 400) ya when using all the sequence reported in Mendez *et al* [4] versus 208 300 (95 % CI = 163 900-260 200) ya when using only the 180 kb reported for A0. Incidentally, Elhaik *et al* [5] claim to have estimated the TMRCA using a likelihood method described in Mendez *et al* 2012a[7] ; however, the method was described in Mendez *et al*. 2012b[8], and there is no mention of it in Mendez *et al* 2012a[7]. The incorrect citation is suspicious when also one considers that the code to calculate point values and confidence intervals was directly provided by Mendez to Elhaik (correspondence in WEBPAGE).

13. “However, in this particular case, there is reason for additional consideration because not only is the TMRCA reported for human Y chromosomes by *Mendez et al* [4] significantly older than the mtDNA chromosome and the fossil age of anatomically modern humans, it is also inconsistent with population genetic theory.”

As we indicated in the main text, the age of the TMRCA estimated in Mendez *et al* [4] is not particularly surprising. Coalescence of Y chromosomes and of mtDNA are independent processes, and as it is well known in coalescent theory, a large fraction of the stochasticity of the process is associated with the last coalescent event (i.e., finding the MRCA). We also indicated that the comparison with the age of anatomically modern humans is not really meaningful. The statement that the estimate is inconsistent with population genetic theory is inconsistent with reality.

14. Under the assumption of neutrality, the effective populations size of the Y chromosome is expected to be equal to the effective population size of the mtDNA-one-quarter that of the autosomes and one-third that of the X chromosome.

In reality, the main assumption is 1:1 effective mating ratio of males to females. Even under neutrality, polygyny or polyandry would produce deviation from those expectations.

15. “Current observations of the TMRCA across genomic regions (Table 1) are incompatible with the high Y chromosome TMRCA computed using the derived Y chromosome mutation rate, but are consistent with a Y chromosome TMRCA calculated using the mutation rate estimated from a Y-pedigree”.

There is no such a thing as an observed TMRCA for any locus in the human genome. The only hominin DNA sequenced which corresponds to remains older than 100 000 years is the mitochondrial DNA of an individual from Sima de los Huesos,[9] which does not cluster within the modern human mtDNA tree. "Current observations of the TMRCA across genomic regions" are consistent with anything, because they do not exist.

16. "If selection is acting to reduce diversity on the Y, then the TMRCA estimates of Mendez *et al* [4] are likely substantial underestimates, putting them even more at odds with estimates of the TMCA on the mtDNA, X and autosomes."

As explained before, selection does not change the rate of evolution of neutral sequences.

NOTE:
Elhaik et al [5] also present in Table 1 different estimates of average TMRCA across autosomes, X chromosome, and mtDNA, which they label as "Observed TMRCA". While for the autosomes they report an estimate based on whole genome sequence data, their X chromosome estimates are based on two loci (one locus per cited work). The comparison that ensues makes no sense. Elhaik *et al* [5] take point estimates for the TMRCA of the Y chromosome under two different mutation rates (0.617×10^{-9} and 1.0×10^{-9} substitutions per nucleotide per year), multiply them times four and times three, and compare the result with point estimates for the autosomes and the X chromosome. There are three fundamental flaws in this approach: 1) the estimates for the TMRCA of autosomes, X chromosome and mtDNA are based on the assumption that the divergence times between human and chimpanzee sequences are known, and therefore cannot be used to assess the validity of estimates using a different set of assumptions (apples and oranges again). 2) Elhaik *et al* [5] report no intervals at all for the prediction on TMRCA for the autosomes, X chromosome and mtDNA based on the assumption of panmixia. Due to the large stochastic variance associated to the coalescent of Y chromosome sequences, those confidence intervals would be very large. 3) The stochastic variance in the last coalescent event (and thus the width of the prediction confidence intervals for the TMRCAs of autosomes, X chromosome and mtDNA) would be much larger if there was ancient population structure.

17. "In the Middle Paleolithic (~100–200 kya), AMH like the Omo (195±5 kya) and the *Homo sapiens*

idaltu (160–154 kya) evolved from these archaic *Homo sapiens* and persisted alongside modern humans."

As the name would suggest anatomically moder humans (AMHs) were/are modern humans.

18. "The question whether and to what extent AMH interbred with their archaic predecessors is one of the most fascinating questions in anthropology."

If Elhaik *et al* [5] are referring to AMH interbreeding with people that were dead (the archaic predecessors of AMH) that interbreeding would be extremely unlikely to produce descendants.

19. "We have shown that consistently throughout their examination, *Mendez et al* [4] have chosen the assumptions, approximations, numerical miscalculations and data manipulation that inflated the final TMRCA estimate."

We have shown this statement to be utterly false.

20. "However, we argue that the autosomally-derived Y substitution rate lacks support, and show that the TMRCA estimate from sequence data should be 208 300 (95% CI = 163 900–260 200 ya), which is within the time frame of the emergence of AMH, excluding the possibility of introgression with more ancient hominin taxa."

We have addressed already the criticisms on the estimate of the mutation rate. Elhaik *et al* [5] assume that the mutation rate is known without error: the width of their confidence interval is merely a consequence of the limited length of Y chromosome analyzed. Finally, it is clear that Elhaik *et al* [5] have NOT excluded the possibility of introgression from archaic hominin taxa.

References

1. Xue Y, Wang Q, Long Q *et al*: Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol* 2009; **19**: 1453-1457.
2. Kuroki Y, Toyoda A, Noguchi H *et al*: Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet* 2006; **38**: 158-167.
3. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ *et al*: The male-specific region of the human Y

chromosome is a mosaic of discrete sequence classes. *Nature* 2003; **423**: 825-837.

4. Mendez FL, Krahn T, Schrack B *et al*: An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am J Hum Genet* 2013; **92**: 454-459.
5. Elhaik E, Tatarinova TV, Klyosov AA, Graur D: The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. *Eur J Hum Genet* 2014. NO PU INFORMATION
6. Kong A, Frigge ML, Masson G*et al*: Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012; **488**: 471-475.
7. Mendez FL, Watkins JC, Hammer MF: Global genetic variation at *OAS1* provides evidence of archaic admixture in Melanesian populations. *Mol Biol Evol* 2012; **29**:1513-1520.
8. Mendez FL, Watkins JC, Hammer MF: A haplotype at *STAT2* Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am J Hum Genet* 2012; **91**:265-274.
9. Meyer M, Fu Q, Aximu-Petri A, Glocke I *et al*: A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 2014; **505**: 403-406.