

Reply to

'The 'extremely ancient' chromosome that isn't' by Elhaik et al

Fernando L Mendez^{1,2}, Krishna R Veeramah^{1,3}, Mark G Thomas⁴, Tatiana M Karafet¹ and Michael F Hammer*,¹

¹ARL Division of Biotechnology, University of Arizona, Tucson, AZ, USA;

²Department of Genetics, Stanford University, Stanford, CA, USA;

³Department of EEB, Stony Brook University, Stony Brook, NY, USA;

⁴Research Department of Genetics, Evolution and Environment, University College London, London, UK

E-mail: mfh@email.arizona.edu

In 2013, Mendez and coauthors described in the American Journal of Human Genetics a previously unknown divergent lineage of the human Y chromosome. With the inclusion of this lineage the time to the most recent common ancestor was ~65% older than without it. The analysis was strongly criticized by Elhaik and coauthors in the European Journal of Human Genetics, arguing that there had been statistical and analytical flaws. Elhaik and coauthors went as far as accusing Mendez and colleagues of data manipulation. In the following manuscript, we reject and address each and all the accusations and criticisms brought by Elhaik and coauthors.

Recently, Elhaik *et al*¹ criticized several aspects of an analysis in which Mendez *et al*² estimate the time to the most recent common ancestor (TMRCA) for the Y chromosome tree incorporating a newly identified basal branch called A00. Elhaik *et al* asserted that Mendez *et al* derived an inflated estimate of the TMRCA by applying incorrect assumptions and approximations, numerical miscalculations, and data manipulation. In particular they focused on (1) the method used to estimate the Y chromosome mutation rate, (2) the relative lengths of sequences that should be compared to estimate branch lengths in a tree, and (3) the implications of the Y chromosome TMRCA estimates reported by Mendez *et al* for human evolution. Here we show that these criticisms result from a misunderstanding of population genetic theory, as well as a misrepresentation of the methodology of Mendez *et al*. However, before addressing the various arguments put forward by Elhaik *et al*, we comment on conceptual and theoretical issues surrounding the significance of the Y chromosome TMRCA.

CONCEPTUAL FLAWS

Elhaik *et al* state in their abstract that the ‘extraordinarily early estimate’ of the TMRCA for the Y chromosome of 338 000 ‘contradicts all previous estimates in the literature and is over a 100 000 years older than the earliest fossils of anatomically modern humans’. With regard to the first point in quotations, clearly the TMRCA estimate that incorporates a newly discovered basal lineage is expected to be older than previous estimates. With regard to the second point, there are no a priori expectations why this or any other locus in the genome has to find its most recent common ancestor more recently than the emergence of anatomically modern features. The mean TMRCA values for both X-linked and autosomal loci have been estimated to be >1 Mya,³ while Elhaik *et al* themselves suggest an autosomal TMRCA of ~800 kya—much older than the estimated age (~0.2 Mya) of the earliest known anatomically modern human (AMH) fossils.⁴ In addition, the standard deviation of expected TMRCA values under a simple neutral model is extremely large such that any particular sampled genealogy can possess a TMRCA that varies wildly from the expectation. Thus, a Y chromosome TMRCA that is

older than 200 000 years is in no way incompatible with the fossil record. Moreover, the fossil record is fragmentary, both geographically and temporally. Consequently, the age of the earliest known anatomically modern fossil constitutes a lower bound for the age of AMH. Thus, as we argue in Mendez *et al*, the new Y chromosome TMRCA date estimate says very little about the origins of AMH.

Elhaik *et al* go on to conclude in the abstract that the TMRCA estimate presented in Mendez *et al*: ‘raises two astonishing possibilities, either the novel Y chromosome was inherited after ancestral humans interbred with another species, or anatomically modern *Homo sapiens* emerged earlier than previously estimated and quickly became subdivided into genetically differentiated populations’. As mentioned above, the stochastic nature of the genealogical process means that differences between TMRCA estimates for different loci are expected to be large. Only by jointly considering the age, rarity, and geographic distribution of the A00 lineage, which so far has been found only in a very restricted area of Central Africa, do Mendez *et al* speculate that the presence of A00 was influenced by a highly structured ancestral population or archaic introgression. Evidence of ancestral population structure has been obtained from the autosomes, as Mendez *et al* clearly indicate, while evidence for multiple archaic introgression events into populations of AMH has emerged based on analyses of archaic genome sequences from Eurasia.

TECHNICAL ERRORS

We now focus our attention on five supposed technical issues Elhaik *et al* claim led to an overestimate of the TMRCA: (1) use of an inappropriately low mutation rate and the incorrectly assumed relationship between mutation rates on the autosomes and the Y chromosome, (2) the relationship between mutation and substitution rates, and how selection at linked sites and population sizes affect this relationship, (3) the choice of acceptable generation times, (4) the manner in which confidence intervals for the mutation rate were estimated, and (5) the comparison of sequences of unequal length to estimate the TMRCA of human Y chromosome lineages.

Use of an inappropriately low mutation rate

First, Elhaik *et al* criticize the use of a pedigree-based estimate of the autosomal mutation rate to infer the mutation rate on the X-degenerate portion of the Y chromosome rather than the use of three ‘existing’ estimates they cite in the literature. Two of the three references they cite to support this claim are irrelevant. The first reference provides only a point estimate of a mutation rate (1.24×10^{-9} mutations per site per year) based on the assumption of a known divergence time (5 million years) between human and chimpanzee sequences and the analysis of ~1.1 kb of Y chromosome sequence.⁵

Given the uncertainty in human-chimp species divergence time estimates, the noisy relationship between species divergence times and gene lineage divergence times (which are generally expected to pre-date species divergence) and especially given the short length of DNA sequence considered, the associated confidence intervals are likely to be very wide (note that no attempt was made by Elhaik *et al* to provide an estimate of the uncertainty of this mutation rate). The second reference estimates a mutation rate per year based on the observation of four mutational events in 13 meioses separating two human Y chromosomes.⁶ This results in a confidence interval (0.30×10^{-9} – 2.5×10^{-9} mutations per site per year)—so broad that it entirely brackets the confidence interval reported by Mendez *et al* (0.439×10^{-9} – 0.707×10^{-9} mutations per site per year). The third reference, a manuscript submitted more than a month after Mendez *et al* was published,⁷ reports an estimate (0.65×10^{-9} mutations per site per year) that is within 6% of the point estimate used by Mendez *et al*. Elhaik *et al* also state that 1.0×10^{-9} mutations per site per year ‘is a widely accepted estimate’, mentioning two recent references that have applied this value.^{8,9} Yet the former also acknowledges that this estimate has ‘wide confidence intervals’ and that ‘additional measurements of mutation rate are urgently needed to improve calibration’. We note that since Mendez *et al* published their paper, subsequent authors have inferred and/or used mutation rates that are lower than 1.0×10^{-9} mutations per site per year.^{7,10,11} Secondly, Elhaik *et al* contend that the mutational processes in males on the X-degenerate portion of the Y chromosome may be fundamentally different from that on the autosomes, and a linear relationship cannot be assumed. Their support for this statement relies exclusively on literature that compares the amount of divergence on the autosomes and the sex chromosomes.^{12–14} However, those comparisons do not take full account of the stochasticity of the coalescent process of Y chromosome lineages in the ancestral population (specifically in the population ancestral to human and chimpanzee). Thus, the relevance is, at best, difficult to assess without more direct estimates of chromosome specific mutation rates in humans (ie, through pedigree studies). Moreover, Pink *et al*,¹² who compare the divergence of autosomes, X chromosomes, and Y chromosomes in rodents, suggest that (1) mutation rate on the Y chromosome could be even lower than expected based on the observations in the autosomes, and (2) the linear relationship proposed by Miyata *et al*¹⁵ (and used by Mendez *et al*) is unlikely to greatly mislead in certain circumstances (ie, where replication effects dominate other external effects such as recombination rates, as also implied by Elhaik who state ‘male mutation bias may explain most of the differences in the substitution rates’) that are likely to apply in humans.

The relationship between mutation and substitution rates

Elhaik *et al* express concern about the possibility that substitution rates (long-term rate of sequence evolution) differ from the underlying mutation rate. In principle, multiple mutational hits at the same site and purifying selection may affect the rate of evolution. Multiple hits are unlikely in the sequence and time frame that Mendez *et al* consider. Using the approach of the birthday problem,¹⁶ we estimate that the probability of no double hits in that data set is >98%. Mendez *et al* analyzed only noncoding sequences of the X-degenerate portion of the Y chromosome, making selective constraints or acceleration of sequence evolution less likely. Despite the claim of Elhaik *et al*, Mendez *et al* never ‘assumed complete lack of purifying or positive selection on the Y chromosome’. However, as shown below, this assumption would be irrelevant. Following a rather unclear and unfocused section on how deleterious mutations on the Y chromosome may affect patterns of diversity and substitution, Elhaik *et al* imply (p.4, ‘If selection is acting to reduce diversity on the Y, then the TMRCA estimates of Mendez *et al* are likely substantial underestimates’) that selection at linked sites and genetic drift will affect the rate of evolution at neutral sites. However, it seems likely that Elhaik *et al* have confused the relationship between the mutation rate and substitution rate at neutral sites (which is the same regardless of linkage to selected sites or demography) with either (a) levels of diversity in a population and mutation rates or (b) mutation rates and substitution rates at putatively selected sites. Birky and Walsh¹⁷ demonstrated over 25 years ago that ‘the fixation probability, and hence the rate of evolution, of neutral alleles is not changed by the occurrence of a linked mutation under selection.’

The choice of acceptable generation times

Elhaik *et al* suggest that the male generation times considered in Mendez *et al* might not be realistic. There are two factors that are relevant to estimating the mutation rate per year. The first factor is that the mutation rate for the Y chromosome depends weakly on, and is a monotonically increasing function of, paternal age (Supplementary Figure S2 in Mendez *et al*). The second factor is that when a paternal lineage is followed into the past, it is unlikely that its ancestors are always amongst the oldest sons. The range of ages at which males reproduce is rather broad (analogous numbers for females are presented in Fenner¹⁸). In most societies, males tend to be the older partner in a marriage. Despite what Elhaik *et al* argue, a study of age of reproduction across a wide range of hunter–gatherer and agriculturalist societies concludes that ‘projections based on Y chromosome data should use a generation interval of 31 or 32 years’.¹⁸ Considering the positive correlation between paternal age at conception and the mutation rate per year, it is not clear what Elhaik *et al* are referring to when they state ‘by using a lower bound of 20 years, an average of 30 years, and an upper bound of 40 years, Mendez *et al* reduced the

number of generations per unit time, and further inflated the TMRCA estimate.’ The choice of a range of generation times in Mendez *et al* as opposed to a single value results in wider confidence interval for the mutation rate. The point estimate is based on an average of 30 years, which is the same value used in Xue *et al*,⁶ is consistent with the value proposed based on anthropological studies. Moreover, the effect of male generation time on per year mutation rate is the opposite to what Elhaik *et al* claim (Supplementary Figure S2 in Mendez *et al*).

Confidence intervals for the mutation rate

Elhaik *et al* criticize the choice of 90% confidence intervals, rather than 95 or 99% confidence intervals for the estimate of the mutation rate. Although Mendez *et al* used 90% confidence intervals for this particular estimation, they chose the ‘worst-case scenario’ for calculating each end of the confidence interval in the subsequent estimation of the TMRCA. For instance, in the calculation of the lower bound of the TMRCA, Mendez *et al* took the value of the mutation rate at the upper bound of the 90% confidence interval (ie, the fastest) and the lower bound of the confidence interval for the expected number of mutations. In practice, this is much more conservative than the approach of Elhaik *et al* who report confidence intervals for their recalculations of TMRCA based only on a single value of the mutation rate (1.0×10^{-9} mutations per site per year). Elhaik *et al* suggest that prediction intervals should be used instead of confidence intervals because ‘calculations of Mendez *et al* involved simulation and sampling’. Elhaik *et al* do not seem to appreciate that Mendez *et al* are not interested in the range of values in future observations, but rather in estimation, and thus prediction intervals are not appropriate. Elhaik *et al* incorrectly state that Mendez *et al* assume that the number of maternal mutations is normally distributed. Mendez *et al* assumed that their mean was normally distributed, which is the asymptotic result from the Central Limit Theorem.¹⁹ In turn, Mendez *et al* estimated the standard error of the means of the total number of mutations per genome and of the number of mutations in females. Elhaik *et al* also propose to use each of five trios where paternal and maternal mutation rates are scored independently to estimate Y chromosome mutation rate. Then they report confidence intervals for the TMRCA ignoring the uncertainty in the estimate of the mutation rate in their Supplementary Table S1. We show in Supplementary Table S1 that, although ignored by Elhaik *et al*, the uncertainty in mutation rate associated to a single trio is substantial.

Comparison of sequences of unequal length

Elhaik *et al* speculate that estimating the TMRCA using sequences of unequal length for A00 and A0 is biased. Contradicting their own statement, Elhaik *et al* use their fast mutation rate estimate (with no

uncertainty) to estimate the TMRCA and obtain almost identical point estimates when they use the original data compared with when they restrict the analysis to only the region that was sequenced for A0. Likewise, the confidence interval for the first case is a subset of the confidence interval for the second set (209 500 ya; 95% CI: 168 000–257 400 ya, and 208 300 ya; 95% CI: 163 900–260 200 ya, respectively). The method to estimate the TMRCA is based on computing the likelihood for the TMRCA using the observed number of mutations in each of the branches. Mutations in each branch are independent and so the likelihoods are multiplicative. The length of sequence used only affects the linear relationship between the expected number of mutations and the TMRCA. See the Supplementary Note for a detailed explanation for why Mendez *et al* reported 240 kb of the A00 chromosome, but only 180 kb of the A0 chromosome. Briefly, with outgroup (chimpanzee) sequence available, mutations that are derived in A00 may be identified by comparing the A00 sequence with that of the outgroup and either A0 or the reference sequence. On the other hand, identifying mutations that are derived in all A0-T lineages requires the use of A00 sequences. Therefore, analyzing mutations along the A0 branch requires coverage in A00, but analyzing mutations in the A00 does not require coverage in A0. Thus, the sequence considered for A0 has to be a subset of that used for A00. This is a direct consequence of the reference sequence being more closely related to A0 than to A00. Mendez *et al* chose to perform the analysis on the number of derived mutations on A0 instead of in the reference sequence to provide a robust estimate (independent of the mutation rate) of how much older the TMRCA of Y chromosome haplogroups is relative to the TMRCA of the A0-T haplogroups.

CONCLUSION

After detailed examination of the criticisms presented by Elhaik *et al*, we show that there are both technical and conceptual flaws that undermine their claims. While not central to the arguments of Elhaik *et al*, there are multiple additional problems in their manuscript, some of which we discuss in the Supplementary Note. However, we do wish to point out that the supposed quotation, cited as personal communication FLM, was entirely fabricated, and we include the full set of email correspondence between FL Mendez and E Elhaik as a supplementary file.

REFERENCES

- 1 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.; 22:1111-1116
- 2 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.
- 3 Blum MG, Jakobsson M: Deep divergences of human gene trees and models of human origins. *Mol Biol Evol* 2011; 28: 889–898.
- 4 McDougall I, Brown FH, Fleagle JG: Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 2005; 433: 733–736.
- 5 Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW: Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci USA* 2000; 97: 7360–7365.
- 6 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.
- 7 Francalacci P, Morelli L, Angius A *et al*: Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* 2013; 341: 565–569.
- 8 Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R: A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am J Hum Genet* 2011; 88: 814–818.
- 9 Wei W, Ayub Q, Chen Y *et al*: A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res* 2013; 23: 388–395.
- 10 Poznik GD, Henn BM, Yee MC *et al*: Sequencing Y chromosomes resolves discrepancy in time to

common ancestor of males versus females. *Science* 2013; 341: 562–565.

11 Scozzari R, Massaia A, Trombetta B *et al*: An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. *Genome Res* 2014; 24: 535–544.

12 Pink CJ, Swaminathan SK, Dunham I, Rogers J, Ward A, Hurst LD: Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biol Evol* 2009; 1: 13–22.

13 Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD: Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol* 2006; 23: 565–573.

14 Wilson Sayres MA, Venditti C, Pagel M, Makova KD: Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution* 2011; 65: 2800–2815.

15 Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T: Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* 1987; 52: 863–867.

16 Durrett R: *DNA Sequence Evolution*, 2nd edn. Springer, 2008.

17 Birky Jr CW, Walsh JB: Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci USA* 1988; 85: 6414–6418.

18 Fenner JN: Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 2005; 128:415–423.

19 Casella G, Berger RL: *Statistical inference*, 2nd edn. Brooks/Cole, 2001.