Elsevier Editorial System(tm) for Forensic Science International: Genetics
Manuscript Draft

Manuscript Number:

Title: A comment on the Paper: A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping By W. Wei, Q. Ayub, Y. Xue, C. Tyler-Smith Forensic Sci. Int. Genet. (2013) http://dx.doi.org/10.1016/j.fscigen.2013.03.014

Article Type: Letter to the Editor

Keywords: Y-STR; Y-SNP; mutation rate constant; haplotype; haplogroup; back mutations.

Corresponding Author: Prof. Anatole Alex Klyosov, PhD, D. Sc.

Corresponding Author's Institution: The Academy of DNA Genealogy

First Author: Anatole Alex Klyosov, PhD, D. Sc.

Order of Authors: Anatole Alex Klyosov, PhD, D. Sc.

Abstract:

Suggested Reviewers:

Forensic Science International: Genetics

Letter to the Editor

A comment on the Paper:

A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping

By W. Wei, Q. Ayub, Y. Xue, C. Tyler-Smith

Forensic Sci. Int. Genet. (2013)
http://dx.doi.org/10.1016/j.fscigen.2013.03.014

Anatole A. Klyosov

Newton, Massachusetts. aklyosov@comcast.net
http://aklyosov.home.comcast.net

Forensic Science International: Genetics

Letter to the Editor

A comment on the Paper:

A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping

By W. Wei, Q. Ayub, Y. Xue, C. Tyler-Smith

Forensic Sci. Int. Genet. (2013)
http://dx.doi.org/10.1016/j.fscigen.2013.03.014

Anatole A. Klyosov

Newton, Massachusetts. aklyosov@comcast.net
http://aklyosov.home.comcast.net

The cited paper, unfortunately, presents a striking example in the area of mismanagement of mutation rate constants in population genetics. Chris Tyler-Smith and his team have made at least five errors in their work on different levels of significance.

First, their selection of the SNP lineages was greatly distorted. It might be acceptable for their SNP analysis, which gave 101-115 kya for a common ancestor of the selection of A-DT lineages; however, the A and DT nodes are shown in the paper at equal level position in time. This tells us that something is wrong in their assumptions/calculations. An earlier STR-based analysis (Klyosov and Rozhanskii, 2012) gave 160,000 ya for the split, and showed a very different level for the A, BT and DT split from the main evolution Y-DNA tree.

Nonetheless, let it be 100,000 ya in this particular case. The Tyler-Smith' STR analysis does not fit, and it is strikingly incorrect. When one "averages" the STR lineages being calculated and aims at the TMRCA, they should be balanced; the dataset should be examined for different branches descending from different common ancestors. If they are mixed in the dataset, they should be separated and analyzed separately. If this is not done, the largest haplotype branch would "pull the blanket" on itself, and instead of a balanced combined haplotype dataset we would have obtained the system totally shifted to the largest branch and its common ancestor. In other words, the largest branch would play the "ancestral" (base) role, and the TMRCA would be phantom one. We cannot compare a flock of sheep and a horse, however, we can compare a sheep and a horse for a meaningful study.

In addition, the 33 haplotype dataset in the cited study contained four identical haplotypes, of a father and his three sons. All of it added to the distortion of the dataset, as described above.

Now, let us examine their STR selection for the study. Out of 33 individuals, 13 were R1b, 11 were E, three I, and the rest were singular haplogroups (their particular subclades, in fact). The dataset was greatly distorted. In fact, the paper has calculated the phantom TMRCA shifted towards R1b/E1b, which should give the TMRCA somewhere around 20-40 kya and not anything close to 100 kya. This was the greatest principal error in his study.

Secondly, they did not introduce a proper correction for back mutations, though in the paper, they talked much about "saturation" of allele values. This is still terra incognita in population genetics, but developed well in DNA genealogy (e.g., Klyosov, 2009a; Klyosov, 2012). A quick look at paper's sample of haplotypes and a number of their mutations (removing the haplotypes of the three sons, see above) reveals that there are around 1000 mutations altogether in 30 haplotypes. It gives 1000/30/23 = 1.45 mutations per marker in the dataset. This is an impermissibly high degree of mutations

in any dataset for any meaningful calculation. The reason is simple - the haplotypes contained many "fast" markers that should have been eliminated. It is recommended for calculations of TMRCA for ancient common ancestors to calculate the "slow" 22 marker panel, or even the slower panels (Rozhanskii and Klyosov, 2011). "Fast" markers "saturate" the system, indeed, and the TMRCA is always underestimated. That is why the Tyler-Smith' paper obtained much lower STR-based TMRCA (around 20 kya) compared to the SNP-based TMRCA (around 100 kya). The underestimation was dual - the distorted, unbalanced dataset, plus the lack of proper corrections for back mutations.

Thirdly, the authors made an improper selection of markers for their haplotype (STR) study, and it is partly explained above. They should have focused on the slow markers only. There are a sufficient number of them to choose from; and they should have calibrated them (see below).

Fourth, they choose mutation rate constants uncritically. They blindly picked some markers with poorly determined mutation rate constant (in the father-son study), since there were too few mutations in them to be statistically sound. Here are examples from their choice: (a) 1 mutation in 1213 father-son pairs (it is not statistics). How one can calculate reliably the mutation rate constant from just one mutation? Well, the authors did: 8.244 x 10-4 [!]. (Notice the given accuracy). (b) 3 mutations in 403 pairs; (c) 1 mutation in 555 pairs; (d) 2 mutations in 555 pairs; (e) 2 mutations in 4565 pairs; (f) 0 mutation in 555 pairs - why choose such a marker? It was useless; (g) 5 mutations… (h) 6 mutations... (i) 6 mutations. In fact, half of the markers picked were statistically mute. What can be said on the accuracy of their calculations? In their particular case it is not really important, they killed their study anyway, and the sloppiness of the study is shocking.

Fifth, they employed the "evolution mutation rate" by Zhivotovsky et al (2004), which was a grave mistake. This "rate" was (and continues to be) a disaster for practically all population genetic studies which includes calculations for TMRCA in the last decade (see critique in, e.g., Klyosov, 2009b). The damage it brought to population genetics is beyond comprehension. The "evolution mutation rate" assigns the same rate - 0.00069 mutation per marker per 25 years - to each marker equally. We have seen above how different the markers are in terms of their mutation rate, but 0.00069 - always? No matter which markers are chosen?

Let us reexamine to see what happened. The summary mutation rate (per haplotype) in those 23 markers (actually, 21, since two markers were eliminated by the authors) [see the table below] is 0.09369 (my calculation, it is not in the paper); that is divided by 21 and yields 0.00446 mutation per marker per generation. The "evolution/Zhivotovsky" rate is 0.00069 mutation/marker/25 years, and that is 6.5 times slower. Naturally, when they used the "evolution" rate they obtained the TMRCA much higher. That is an

explanation why they obtained a "fit" between the SNP- and STR- based TMRCA when the "evolution" rate was used. They just automatically increased the TMRCA. The "recalibrated evolutionary mutation rate" was the same thing, in just one marker (out of 21) instead of the "evolution" rate of 0.00069, they changed it to 0.000351. The rest was the same 0.00069. Clearly, nothing has changed overall.

Table. A comparison of the mutation rate constants for a number of loci employed by the authors of the cited paper – from father-son pairs (Burgarella et all, 2011), from Chandler (2006), and from Zhivotovsky et al (2004). The table show how different those values are. The authors of the cited paper did not provide any support for the chosen mutation rate constants (Burgarella et al, 2011) from any other source, such the Chandler (2006) data.

	Number of father-son pairs	Number of mutations	Mutation rate constant k x 10^5 per marker per generation	Chandler' constants (2006) k x 10^5 per marker generation	Zhivotovsky' "Population", or "evolution" mutation rate constant (2004) k x 10^5 per marker generation
DYS576	555	9	1622	1022	69
DYS389 I	7,864	20	254	226	69
DYS448	1,213	1	82	135	69
DYS389 b	7,842	28	357*	242	69
DYS19	9,840	23	234	151	69
DYS391	9,279	25	269	265	69
DYS481	403	3	744		69
DYS549	555	1	180		69
DYS533	555	2	360		69
DYS438	4,565	2	44	55	69
DYS437	4,381	6	137	99	69

DYS570	555	7	1261	790	69
DYS635	1,920	12	625		69
DYS390	9,340	22	236	311	69
DYS439	4,542	28	616	477	69
DYS392	9,264	5	54	52	69
DYS643	555	0	0		69
DYS393	7,835	6	77	76	69
DYS458	1,243	13	1046	814	69
DYS456	1,243	8	643	735	69
Y-GATA-H4	2,083	11	528	208	69

As a result, the paper is a total embarrassment. Other minor things: I could not find in the paper the mutation rate constant they employed for SNPs (1×10^{-9} per nucleotide per year? Any other figure?). I also could not find how the authors converted generations (from father-son studies) into years.

References

Burgarella, C., Navascues, M. (2011). Mutation rate estimates for 110 Y chromosome STRs combining population and father-son pair data. *European Journal of Human Genetics, 19*, 70-75.

Chandler, J. F. (2006). Estimating per-locus mutation rates. *Journal of Genetic Genealogy, 2*, 27-33.

Klyosov, A.A. (2009a) DNA Genealogy, mutation rates, and some historical evidences written in Y-chromosome. I. Basic principles and the method. J. Genetic Genealogy, 5, 186-216.

Klyosov, A.A. (2009b) A comment on the paper: Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish Priesthood. Human Genetics, 126, 719-724.

Klyosov, A.A., Rozhanskii, I.L. (2012) Re-examining the "Out of Africa" theory and the origin of Europeoids (Caucasoids) in light of DNA genealogy. Adv. Anthropol. 2, 80-86.

Klyosov, A.A. (2012) Ancient history of the Arbins, bearers of haplogroup R1b, from Central Asia to Europe, 16,000 to 1500 years before present. Advances in Anthropology, 2, 87-105.

Rozhanskii, I.L., Klyosov, A.A. (2011) Mutation rate constants in DNA genealogy (Y chromosome). Adv. Anthropol. Vol. 1, No. 2, 26-34.
Zhivotovsky, L. A., Underhill, P. A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T. et al. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population divergence time. The Amer. J. Hum. Genetics, 74, 50-61.

Table. A comparison of the mutation rate constants for a number of loci employed by the authors of the cited paper – from father-son pairs (Burgarella et all, 2011), from Chandler (2006), and from Zhivotovsky et al (2004). The table show how different those values are. The authors of the cited paper did not provide any support for the chosen mutation rate constants (Burgarella et al, 2011) from any other source, such the Chandler (2006) data.

	Number of father-son pairs	Number of mutations	Mutation rate constant k x 10^5 per marker per generation	Chandler' constants (2006) k x 10^5 per marker generation	Zhivotovsky' "Population", or "evolution" mutation rate constant (2004) k x 10^5 per marker generation
DYS576	555	9	1622	1022	69
DYS389 I	7,864	20	254	226	69
DYS448	1,213	1	82	135	69
DYS389 b	7,842	28	357*	242	69
DYS19	9,840	23	234	151	69
DYS391	9,279	25	269	265	69
DYS481	403	3	744		69
DYS549	555	1	180		69
DYS533	555	2	360		69
DYS438	4,565	2	44	55	69
DYS437	4,381	6	137	99	69
DYS570	555	7	1261	790	69
DYS635	1,920	12	625		69
DYS390	9,340	22	236	311	69
DYS439	4,542	28	616	477	69

DYS392	9,264	5	54	52	69
DYS643	555	0	0		69
DYS393	7,835	6	77	76	69
DYS458	1,243	13	1046	814	69
DYS456	1,243	8	643	735	69
Y-GATA-H4	2,083	11	528	208	69