R1b-M343 (xP312 xU106) Y-STR Report

Dirk Struve

May 21, 2017

Contents

1 Introduction

This document contains time estimates about the formation of clades and their TMRCAs (Time to Most Recent Common Ancestor) based upon Y-STR marker results reported by YFull and Family Tree DNA.

You might ask yourself, why use STRs when YFull has already published a phylogenetic tree [15] with time estimates solely based on the stable SNP mutations. There are a variety of good reasons for this:

1. Different scientific methods can yield different results. It is good to calculate time estimates using independent methods. If different methods yield the same results, we know that we can trust the methods. Otherwise we should be cautious and search for improvements.

2. Mutations happen by coincidence. There is always a statistical uncertainty. Using different methods can help to identify statistical outliers and improve the results.

3. Not all SNPs are suited for genealogical purposes. This could result in wrong time estimates. Even if YFull has carefully crafted it's method to exclude wrong SNPs [1], it is good to have a second method to verify their results.

4. STRs provide a higher time resolution than SNPs. YFull reports approximately 400 STR markers for Big Y [2] BAM files. Current results indicate that we may roughly count one mutation as 50 years. With 4500 known Y-STRs [13] we may expect further improvement in the future.

These reasons are certainly good enough to invest some time and work, but be warned: STRs are more difficult to handle than SNPs and the results should always be viewed with a certain amount of caution.

2 About Y-STR markers

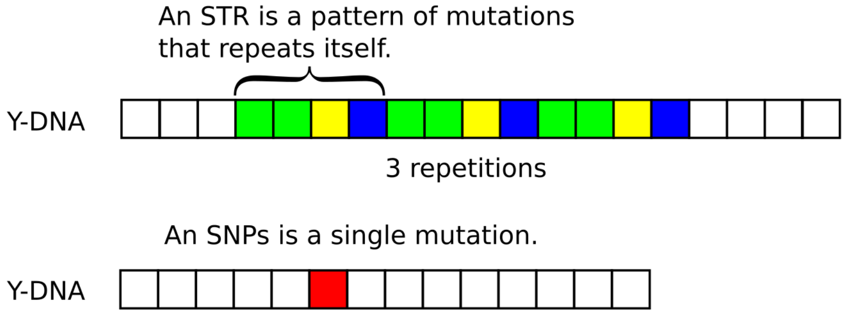
An STR is a pattern of mutations
that repeats itself.
Y-DNA
3 repetitions
An SNPs is a single mutation.
Y-DNA

Figure 1: Mutations on the human Y chromosome. An STR (Short Tandem Repeat) is a mutational pattern that repeats itself. An SNP (Single nucleotide polymorphism) is a single mutation at a specific location.

A Y-STR marker is a mutational pattern on the Y chromosome that repeats itself as illustrated in figure 1. The value of a marker is the number of mutations. For example if $DYS393 = 13$ this means the DYS393 pattern is repeated 13 times.

If a mutation occurs the number of repetitions is changed. It can increase or decrease, sometimes by several steps at once. Also a mutation does not always change the number of repetitions in the same direction. A marker can change from 13 to 14 and then back again to 13.

This makes the counting of STR mutations difficult, especially if we want to go back for a long time. We never know if a marker has stayed constant or if it has changed and changed back to the same value as before.

Counting is important. Because the mutation rates of marker sets are well known, it is possible to convert a number of mutations to a time span and thus calculate the TMRCA (Time to Most Recent Common Ancestor). For this we also have to estimate the ancestral haplotype of the ancestor (the STR values of the ancestor). It is called the modal haplotype. The method used here is described in detail in [9].

SNP mutations are mutations that occur at a single location on the chromosome. They happen only once and stay forever (with only very few exceptions). Thus they are easier to handle than STR mutations. They allow to build highly reliable phylogenetic trees and calculate a TMRCA simply by counting the number of mutations.

3 Time estimates

3.1 A word of caution

Many people take TMRCA estimates as true because they are the results of computer calculations. If several people derive different results you can often find hot discussions on the Internet that one of them must be wrong. But mutations occur by chance and there is simply no way of calculating TMRCAs with high precision for the following reasons:

1. Mutations occur by chance. YFull counts an SNP mutation that is relevant for genetic genealogy as 144 years [1]. This will always result in a high statistical uncertainty, especially within genealogical time frames, where we usually have only a few samples available.

2. I used a generation time of 32 years, the same as YFull. But this is a long time average. Specific lineages can have different generation times. My own paternal lineage has had a generation time of just 24 years for the last nine generations.

3. Results obey the laws of statistics and depend on the sample set. The sample sets used in genetic genealogy are usually small. Thus researchers who use different sample sets often come to different results.

If the situation is so complicated, what can be done? First, we need as many data as we can get. Second, we need to view the results with a considerable amount of caution.

For an individual to get most out of the results presented here, I recommend the following steps:

1. For time estimates below 1000 years, especially genealogical time frames, take a look at section 3.3.3. Current results indicate that genealogical time frames must be treated differently. Please also check if a generation time of 32 years is appropriate for your lineage. All time estimates presented here are based on a generation time of 32 years.

2. Compare the TMRCA results of YFull to those presented here. If they are identical this is a good clue that they are correct, because the methods of calculation are independent of each other.

3. If the results are different, search for possible statistical outliers. Does one sample in a data set show a suspiciously higher or lower number of mutations compared to the others? If this is true the result might be affected by the outlier.

4. For the results presented here, at least three samples or downstream haplotypes are needed to calculate a good modal haplotype. If this is not the case the result might be affected by upstream or parallel lineages.

5. If you are interested in the time when a clade has formed ignore the *formed* results in the trees presented here. Instead use the TMRCA estimate of the parent clade. The TMRCA estimates are averages and thus more accurate.

3.2 Method

There are several ways to calculate TMRCA estimates using STR mutations [6, 5]. The biggest drawbacks when using STRs are that STR counting fails for long time spans and that STRs can't be used to build reliable phylogenetic trees.

To get around these issues, the YFull phylogenetic tree [15] was used. The tree is correct because it is based upon SNPs. I calculated the modal haplotypes for each node of the tree and used STR mutation counting to calculate the time spans between the nodes. The method is described in detail in [9]. The calculations were performed by the Phyloage [8] program.

The method should yield better results than traditional STR mutation counting. In comparison to SNP counting it still has it's drawbacks:

1. For long time spans STR counting is inaccurate because of back mutations. For this report intermediate modal haplotypes were calculated to get more reliable counts. In many cases we have too few samples to calculate enough modal haplotypes.

2. Modal haplotypes can only be reliably calculated for a sufficiently high number of downstream samples. Again we are often missing enough samples to get the high quality we would like to have.

The sample data was taken from members of the R1b-M343 (xP312 xU106) groups [16, 3] at YFull and Family Tree DNA. For the following kit numbers the YFull results were enhanced by filling in missing values with the results from Family Tree DNA:
YF04297, YF04258, YF04369, YF03782, YF03928, YF01482,
YF04780, YF04308, YF03792, YF02716, YF02743, YF01886,
YF01913, YF01929, YF02832, YF02873, YF02895, YF03097,
YF03278, YF03317, YF03867, YF04038, YF04145, YF04721,
YF05208, YF01827, YF04793, YF05208, YF03678, YF04568,
YF04313, YF04762, YF01987, YF05543, YF04234, YF05418,
YF04303, YF04802, YF05022, YF05567, YF05749, YF05599,
YF03999, YF04142, YF04151, YF05873, YF06082, YF06130,
YF06177, YF06325, YF06342, YF06448, YF06345, YF06366,
YF06465, YF06487, YF06434, YF06620, YF06718, YF06719,
YF06757, YF07201, YF06993, YF07418, YF07451, YF07486,
YF07568, YF07893, YF08169, YF08158, YF07441, YF08067,
YF07062, YF07783, YF07766, YF07743, YF07119, YF05702,
YF06497, YF06483, YF06966, YF08241, YF08348, YF08291,
YF08292, YF08424, YF08465.

I could only merge the YFull and FTDNA results for samples that were easy to identify in both groups. If you do not find your kit number in the list and like to be included, please contact me directly or provide your ancestry information at YFull.

YFull usually reports about 400 STR marker results for Big Y samples. Thus for long time spans the merging of YFull and FTDNA results is of minor importance. Even for genealogical distances the extra markers from

FTDNA add only little more precision. In some cases the results get even worse.

3.3 Results

3.3.1 111 marker TMRCA estimates

The following link shows all time estimates for 111 markers, average mutation rates. The mutation rates were taken from [7]. The tree is based upon the YFull tree 5.04 [15]. The STR mutation model is a hybrid where all markers, except the palindromic ones, are counted stepwise.

- Time estimates using 111 Y-STR markers

If the results do not appear well formatted in your browser, try to open them in a text editor and reduce the tab width.

Figure 2 compares TMRCA estimates from YFull to the STR counting method presented here. Both methods are independent of each other because the STR mutation rates were taken from traditional literature [7]. The straight line is a linear fit to the data points calculated by Gnuplot [14].

The results are similar between 1000 and 4000 years. The linear fit has an offset at $x = 0$. It indicates that the STR based TMRCA estimates within genealogical time frames are too high. This could be caused by fast mutating markers that are effected by back mutations and thus can not be counted accurately for long time spans. For longer time spans only the more stable markers add to the mutation count.

The time estimates around 5000 years get very noisy. This could be due to the long time spans and the low number of samples for many clades that make a more accurate calculation impossible.

I left out any data points for more than 8000 years because the M269 haplogroup has experienced a long time population bottleneck [15] that makes STR counting extremely unreliable.

The following table shows TMRCA estimates derived from YFull and STR counting in detail. Only clades that contain at least two subclades and/or samples with valid STR based age estimates are included in the table. Age estimates for other SNPs are too unreliable because of missing data.

As discussed before, TMRCA estimates within genealogical time frames are too high.

Comparison of TMRCA estimates in years

SNP	YFull	111 Markers	Difference
Y8447	7000	5059	-27%

FGC21060	175	617	253%
Y7771	5400	4632	-14%
Y18458	1950	2913	50%
V69	4700	3980	-15%
FGC39691	350	507	45%
Y24712	275	249	-9%
Y14051	1450	1609	11%
M269	6300	7259	16%
PF7562	5500	3661	-33%
L23	6200	7262	18%
Z2103	6100	6512	7%
Y4364	4500	4036	-10%
M12135	3100	2665	-14%
Y30217	2600	2931	13%
Y24543	750	626	-16%
Y24734	750	419	-44%
L584	4700	6486	38%
Y18781	3100	3836	24%
PH2731	1900	3239	71%
Y18441	150	433	189%
FGC14590	4700	4631	-1%
Y18687	4700	2886	-38%
Y11410	1150	1475	29%
FGC14600	650	913	41%
Y21258	375	1065	184%
Z2108	6100	4944	-18%
Y14512	3700	3500	-5%
Y16005	2600	3264	26%
Z2110	6100	4342	-28%
CTS7556	4800	4762	-0%
Y20344	2000	1789	-10%
Y20345	1300	927	-28%
Y29085	850	821	-3%
Y5592	4800	4238	-11%
CTS1450	4800	4889	2%
Y18959	4800	5405	13%
Y10789	3200	4025	26%
Y5587	4800	4206	-12%
Y5586	4400	2974	-32%
Y22219	1500	1346	-10%
V2986	1600	2442	53%

Y14306	1300	2142	65%
BY250	4500	3878	-13%
Y19469	4500	4078	-9%
Y19752	3300	2573	-22%
Y29917	1350	877	-35%
Y23493	4500	2575	-42%
Y28635	650	633	-2%
Y15982	2600	2961	14%
Y16145	2400	3237	35%
Y16143	2000	2461	24%
Y17988	2000	2108	6%
Y18455	550	762	39%
Y28634	450	546	22%
Y23386	250	762	205%
Y23387	100	168	68%
Y19137	475	1023	116%
Y19425	375	843	125%
Y19426	325	492	52%
Y23517	275	546	99%
L51	5900	5494	-6%
Z2118	5100	4353	-14%
S1161	5000	4585	-8%
FGC24138	4500	2925	-34%
L151	4800	4345	-9%
A8062	500	1197	140%
S1200	4800	5118	7%
S17624	3300	3587	9%
S21770	1800	2243	25%
A8487	1800	1907	6%
S14328	4800	4416	-8%
Y16483	2400	1923	-19%
Y28597	2800	1791	-36%
S1196	4800	5026	5%
S6868	4800	3675	-23%
S6849	1950	2011	4%
CTS7354	2900	2370	-18%
S1199	3800	3575	-5%
Y22442	2900	2814	-2%
Y23200	1150	870	-24%

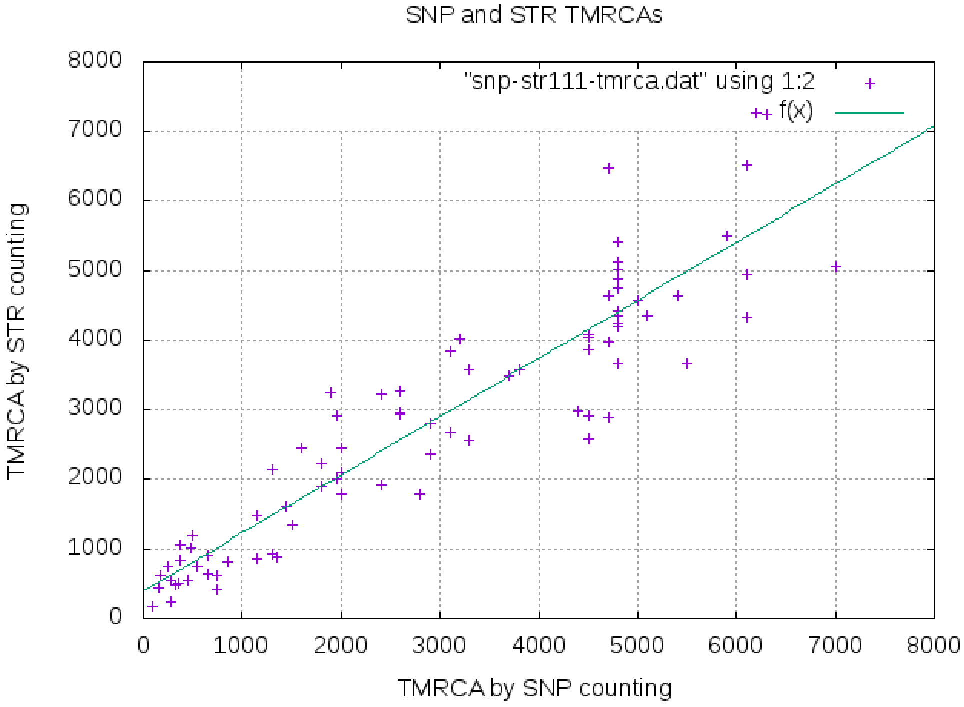

Figure 2: Comparison between TMRCA estimates derived from Y-Full's SNP based method and STR mutation counting. Both methods produce similar results between 1000 and 4000 years. The straight line is a linear fit to the data points. It does not touch the zero point indicating that STR counting produces an offset that leads to too high TMRCA values within genealogical time frames.

3.3.2 500 marker TMRCA estimates

Here is the same tree as before, but this time for 500 markers. YFull derives about 400 markers from Big Y results. The mutation rate for those markers is not yet well known.

I calibrated the mutation rate comparing to YFull's SNP estimates and got a calibration rate of 50 years/marker. Thus we may roughly count one Big Y STR mutation as 50 years. In the previous STR report I used 40 years/marker [12].

- TMRCA estimates using 500 Y-STR markers

Figure 3 again compares TMRCA estimates from YFull to the STR counting method presented here but this time 500 STR markers were used for comparison.

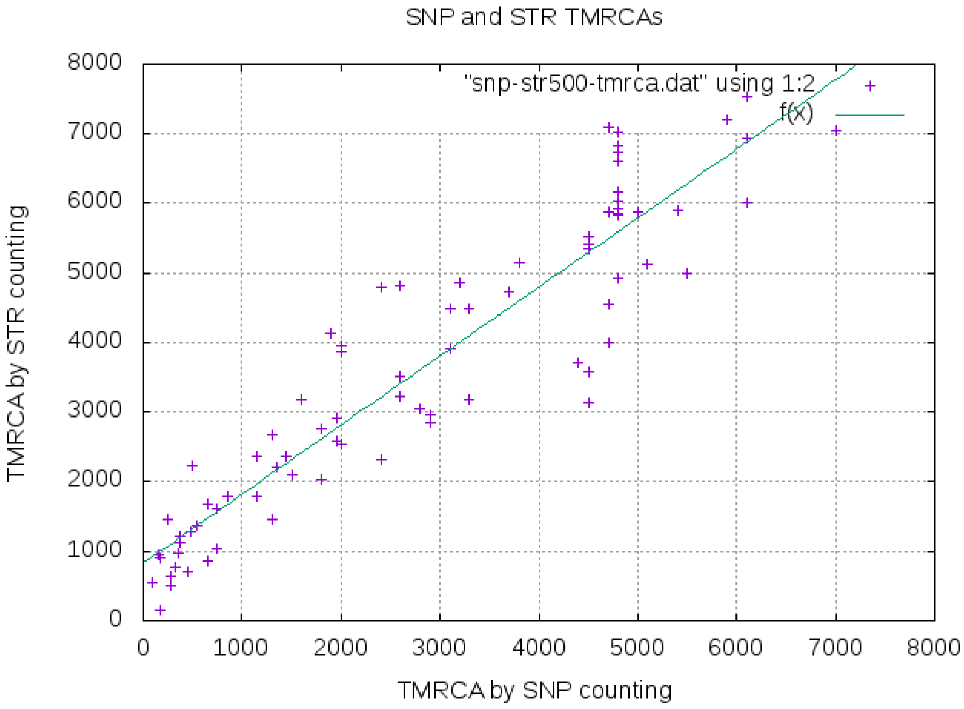

Figure 3: Comparison between TMRCA estimates derived from Y-Full's SNP based method and STR mutation counting. The straight line is a linear fit to the data points. It does not touch the zero point indicating that STR counting produces an offset that leads to too high TMRCA values within genealogical time frames. The graph looks similar to figure 2 where only 111 markers were used.

The results are similar to the results derived by using 111 markers in section 2. So I refer to the discussion there.

The following table contains the data set used for creating figure 3. Again as a minimum quality criterion only subclades that contain at least two subclades and/or samples with valid STR counts were included in the table.

Comparison of TMRCA estimates in years

SNP	YFull	500 Markers	Difference
Y8447	7000	7056	1%
FGC21060	175	917	424%
Y7771	5400	5905	10%
Y18458	1950	2908	50%
V69	4700	4551	-3%

FGC39691	350	982	181%
Y24712	275	513	87%
Y14051	1450	2361	63%
PF7562	5500	4995	-9%
Z2103	6100	7535	24%
Y4364	4500	5424	21%
M12135	3100	3919	27%
Y30217	2600	3506	35%
Y24543	750	1603	114%
Y24734	750	1042	39%
L584	4700	7102	52%
Y18781	3100	4485	45%
PH2731	1900	4141	118%
Y18441	150	952	535%
FGC14590	4700	5870	25%
L943	175	147	-16%
Y18687	4700	4011	-14%
Y11410	1150	2359	106%
FGC14600	650	1672	158%
Y21258	375	1223	227%
Z2108	6100	6944	14%
Y14512	3700	4736	29%
Y16005	2600	3228	25%
Z2110	6100	6013	-1%
CTS7556	4800	6818	43%
Y20344	2000	2532	27%
Y20345	1300	1459	13%
Y29085	850	1783	110%
Y5592	4800	5833	22%
CTS1450	4800	7025	47%
Y18959	4800	6742	41%
Y10789	3200	4866	53%
Y5587	4800	5851	22%
Y5586	4400	3710	-15%
Y22219	1500	2095	40%
V2986	1600	3190	100%
Y14306	1300	2666	106%
BY250	4500	5528	23%
Y19469	4500	5345	19%
Y19752	3300	3182	-3%
Y29917	1350	2209	64%

Y23493	4500	3143	-30%
Y28635	650	865	34%
Y15982	2600	4810	86%
Y16145	2400	4796	100%
Y16143	2000	3963	99%
Y17988	2000	3872	94%
Y18455	550	1365	149%
Y28634	450	709	58%
Y23386	250	1462	485%
Y23387	100	557	458%
Y19137	475	1272	168%
Y19425	375	1128	201%
Y19426	325	771	138%
Y23517	275	637	132%
L51	5900	7203	23%
Z2118	5100	5117	1%
S1161	5000	5880	18%
FGC24138	4500	3574	-20%
L151	4800	5921	24%
A8062	500	2228	346%
S1200	4800	6174	29%
S17624	3300	4489	37%
S21770	1800	2762	54%
A8487	1800	2025	13%
S14328	4800	6034	26%
Y16483	2400	2322	-3%
Y28597	2800	3049	9%
S1196	4800	6612	38%
S6868	4800	4930	3%
S6849	1950	2576	33%
CTS7354	2900	2968	3%
S1199	3800	5150	36%
Y22442	2900	2849	-1%
Y23200	1150	1787	56%

3.3.3 Towards genealogical time frames

Sections 3.3.1 and 3.3.2 have shown that STR counting using traditional mutation rates overestimates TMRCA values within genealogical time frames. This is likely because the mutation rates were derived from data sets that contained an unknown number of back mutations.

Genealogical time frames present a number of issues that make them difficult to handle:

1. For a small number of mutations outliers can only increase the TMRCA estimate because negative numbers of mutations are not possible. Outliers may occur due to statistical reasons (Poisson statistics) or simply because one mutation causes several changes at once (RecLOH events, multiple step mutations).

2. For genealogical time frames the number of mutations is so small that statistical uncertainties get very large.

3. Generation time may vary significantly. In this report all calculations are done using a generation time of 32 years but this simple model clearly does not fit all paper trails.

To check if the the above problems have a huge effect on calculated TMRCA estimates, I ran several statistical simulations and also took a look at the Father, Sons, Brothers project [4] to see how often unusual mutations occur.

My current conclusion is that the mentioned problems do not have a huge effect in general but they can distort the TMRCA estimates for certain subclades.

Because SNP mutations do not suffer from back mutations the YFull derived mutation rates should be valid from genealogical time frames to deep history. Section 3.3.1 has shown that traditional STR counting and SNP counting yield similar results for thousands of years. This indicates that the YFull SNP mutation rate should be close to the real one.

So I decided to derive a new mutation rate for STR counting based on YFull results within the last 1000 years. For this calculation paper trail dates were left aside. To eliminate the effects of multiple step mutations, the infinite alleles mutation model was used for mutation counting.

Figure 4 shows the result. As expected the statistical uncertainty is very high. The best fit is achieved if one STR mutation on the 500 marker scale is counted as 21 years. For the linear fit Gnuplot [14] was used.

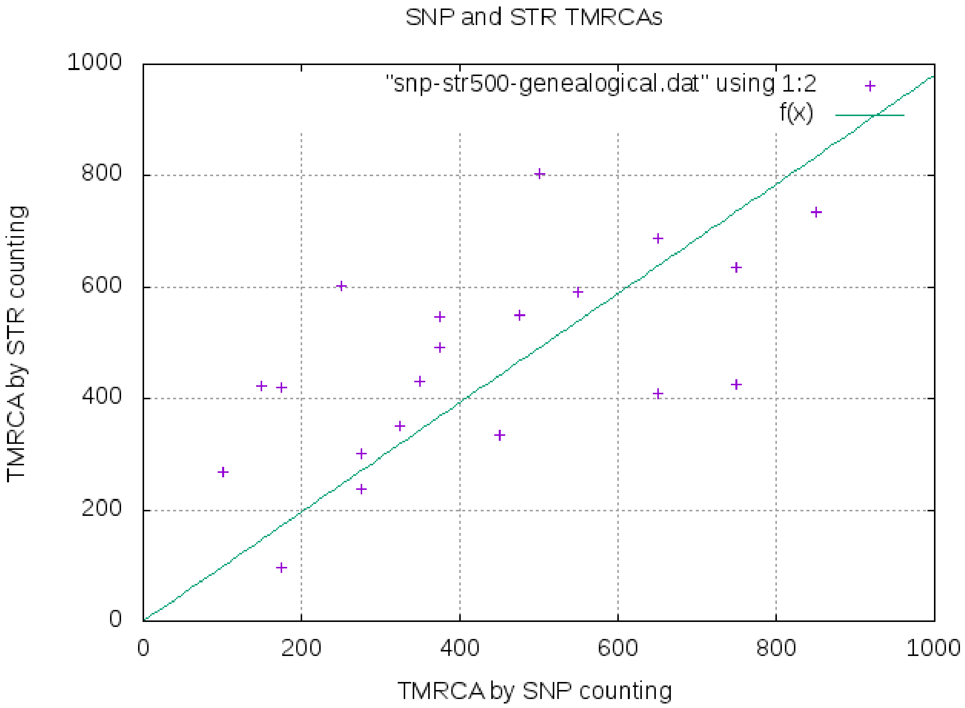

Figure 4: Comparison between YFull TMRCA estimates and STR mutation counting within the last 1000 years. For this comparison the STR mutation rate was calibrated by using the YFull dates. The straight line is a linear fit to the data points.

For comparison with YFull and paper trail dates the following table shows the data set used to create figure 4. As before each SNP must have at least two downstream SNPs and/or samples to be included in the data set.

The future will show how good these time estimates are compared to real world values. With more samples and further insights changes are to be expected.

Comparison of TMRCA estimates in years

SNP	YFull	500 Markers	Difference
FGC21060	175	420	140%
FGC39691	350	432	24%
Y24712	275	238	-13%
Y24543	750	635	-15%
Y24734	750	425	-43%
Y18441	150	422	182%

L943	175	97	-44%
FGC14600	650	688	6%
Y21258	375	547	46%
Y29085	850	735	-13%
Y28635	650	409	-37%
Y18455	550	591	8%
Y28634	450	333	-26%
Y23386	250	603	142%
Y23387	100	269	170%
Y19137	475	549	16%
Y19425	375	493	32%
Y19426	325	351	8%
Y23517	275	302	10%
A8062	500	805	62%

3.3.4 Using only YFull results

Previously I merged the STR results from FTDNA and YFull for many samples in the expectation that more markers yield better results. But is this really true?

YFull extracts STR results from next-generation sequencing data and this is often believed to be not as reliable as the specialized tests used by FTDNA. I checked how often FTDNA and YFull results are identical and it turned out that they differ in approximately 1% of all cases.

To check the effect on time estimates I repeated all calculations with exactly the same parameters as before, but this time I used only STR markers that were extracted by YFull.

The results are below. There is barely a difference. For time distances of thousands of years this was expected because a few extra markers do not add much precision in this case but even within the last 1000 years YFull extracts so many markers with reliable results that the extra results from FTDNA add only little more precision. Because of statistical uncertainties in some cases the TMRCA estimates get even worse.

Comparison of TMRCA estimates in years

SNP	YFull	500 Markers	Difference
FGC21060	175	420	140%
FGC39691	350	432	24%
Y24712	275	238	-13%
Y24543	750	636	-15%
Y24734	750	426	-43%
Y18441	150	422	182%
L943	175	97	-44%
FGC14600	650	693	7%
Y21258	375	536	43%
Y29085	850	745	-12%
Y28635	650	407	-37%
Y18455	550	625	14%
Y28634	450	346	-23%
Y23386	250	625	150%
Y23387	100	284	184%
Y19137	475	532	12%
Y19425	375	509	36%
Y19426	325	353	9%
Y23517	275	301	10%
A8062	500	811	63%

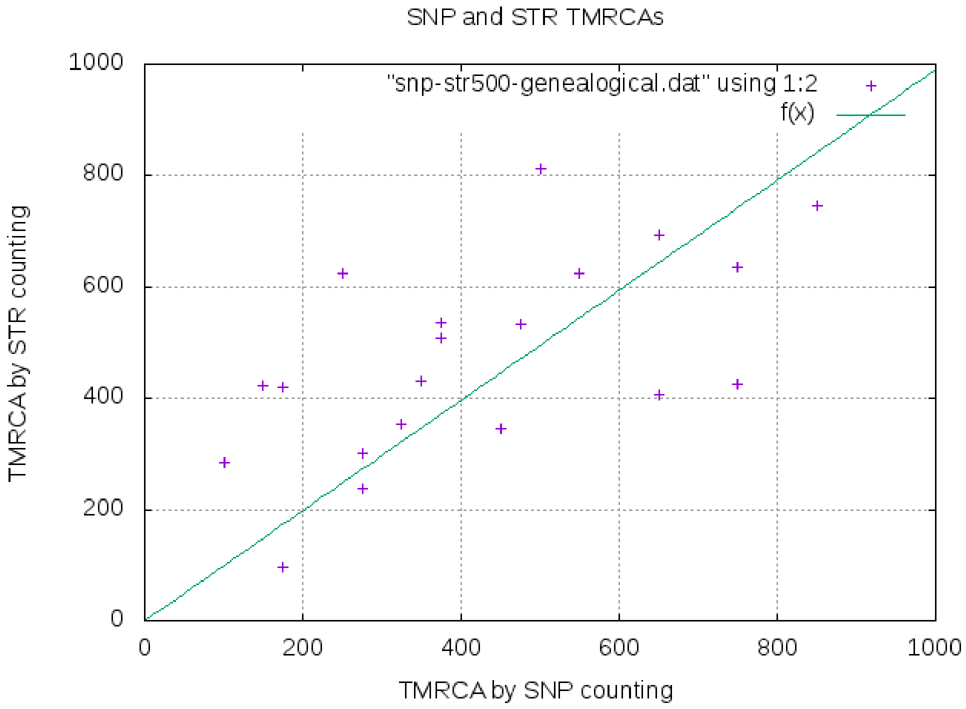

Figure 5: Comparison between YFull TMRCA estimates and STR mutation counting within the last 1000 years. For this comparison the STR mutation rate was calibrated by using the YFull dates. The straight line is a linear fit to the data points.

4 Marker statistics

A Y-STR marker statistics that was compiled from all samples using the Phylofriend [10] program is available at:

- Y-STR marker statistics

Markers with high mutation rates are expected to show only few values while markers with high mutation rates should show many values. Some markers are rarely reported and thus show only a few values even if they mutate fast.

Key points:

1. The mutation rates between different markers differ by a large amount.

2. Stable markers should be useful for long time TMRCA estimates.

3. The statistics can be used to check if some family branches exhibit rare and unique values for specific markers.

References

[1] Dmitry Adamov, Vladimir Guryanov, Sergey Karzhavin, Vladimir Tagankin, Vadim Urasin. *Defining a New Rate Constant for Y-Chromosome SNPs based on Full Sequencing Data*. The Russian Journal of Genetic Genealogy (Русская версия), Vol 6, No 2 (2014)/Vol 7, No 1 (2015).

[2] Family Tree DNA, *Introduction to the Big Y*. Family Tree DNA, 2016.

[3] Family Tree DNA, *R1b-M343 (xP312 xU106) Project*.

[4] Family Tree DNA, *Father, Sons, Brothers project*.

[5] David Hamilton, *An accurate genetic clock*, bioRxiv preprint, first posted online June 15, 2015, doi: 10.1101/020933.

[6] Anatole A. Klyosov, *DNA Genealogy, Mutation Rates, and Some Historical Evidence Written in Y-Chromosome, Part I: Basic Principles and the Method*. Journal of Genetic Genealogy, 5(2):186-216, 2009.

[7] Anatole A. Klyosov, *Ancient History of the Arbins, Bearers of Haplogroup R1b, from Central Asia to Europe, 16,000 to 1500 Years before Present*. Advances in Anthropology, Vol.2, No.2, 87-105, doi:10.4236/aa.2012.22010, 2012.

[8] Dirk Struve, *Phyloage, a program to calculate time estimates from SNP based phylogenetic trees*. GitHub, 2016.

[9] Dirk Struve, *Phyloage User Guide*. 2016.

[10] Dirk Struve, *Phylofriend, a program to calculate genetic distances*. Google Project Hosting, 2014; GitHub, 2015.

[11] Dirk Struve, *Phylogrowth, a program to calculate phylogenetic growth as a function of time*. GitHub, 2016.

[12] Dirk Struve, *R1b-M343 (xP312 xU106) Y-STR Report*. 2016-10-19.

[13] Thomas Willems, Melissa Gymrek, G. David Poznik, Chris Tyler-Smith, The 1000 Genomes Project Chromosome Y Group, Yaniv Erlich. *Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates*. The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.04.001.

[14] Thomas Williams, Colin Kelley et al., *Gnuplot*. Version 5.0.

[15] YFull, *YFull Phylogenetic Tree*. Date visited: 2017-05-18.

[16] YFull, *R1b-M343 (xP312 xU106) group*.