December 13, 2019

Intro to the Y chromosome and where SNPs are found

Introduction

Meet the Y chromosome:

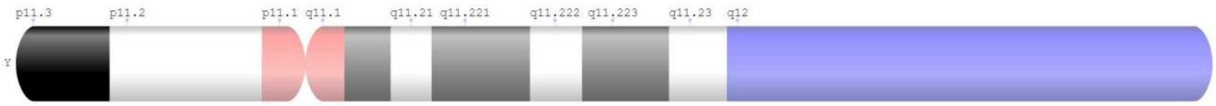


Figure 1. Typical Representation of the Human Y chromosome

Figure 1 is produced by the National Center for Biotechnology Information's Genome Decoration Page website and is a typical representation of the Y chromosome.

I'll talk about base pair numbering in a minute, but it's important for this discussion to remember that the Y like any chromosome is made up of strands of DNA and the "units" of DNA are nucleotides which contain four different bases: Thymine, Cytosine, Adenine, or Guanine, usually abbreviated T, C, A, and G. The Y chromosome can also be considered as one VERY long string of these values (so AGTCGATA…etc). Since DNA is a double helix, each base is bonded to another on the other strand and forms what's called a *base pair*. But it's easier for genetic genealogy purposes if you just consider a "base pair" as simply one position along the Y chromosome which can take on a value of T, C, A, or G.

Note that in typical pictures like Figure 1, the sections are not always uniformly represented in base pair size. The q12 region in blue for instance is actually over half of the Y chromosome in number of base pairs but looks much smaller in Figure 1.

Like every human chromosome the Y is made up of two segments known as "arms" which are named the "p" and the "q" arms. The arms are joined at the *centromere* which is usually shown as a narrow "neck" or "X" on chromosome diagrams and represents the point on the chromosome where it joins with its chromosome pair, and for autosomes (i.e. all the other chromosomes besides the X and Y) is where the pairs join during recombination. The human Y chromosome of course is only found in human males and joins with the X chromosome but does not recombine (with some exceptions that we'll discuss in a minute).

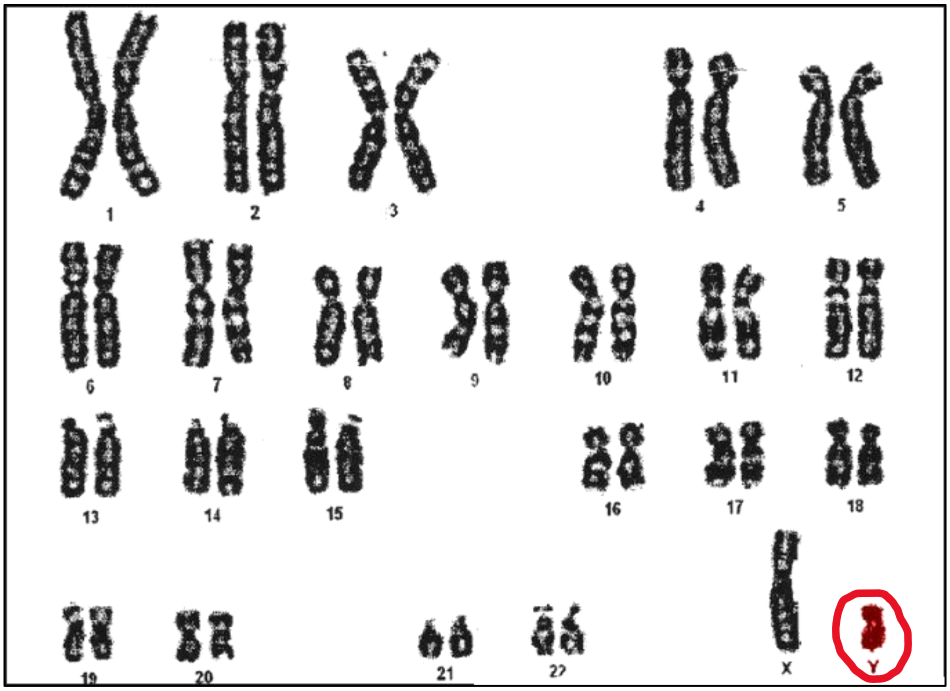


Figure 2. Typical Picture of separated chromosomes

Chromosome diagrams are usually shown split into sections called *cytogenetic bands* or *cytobands* (an example of these is shown in Figure 1 above). These really come from the staining process that is used to observe chromosomes under a microscope. The staining process creates these bands, and so the two arms of the chromosome were split into sections that map to these bands, and the sections were given names like p11.1, p11.2, etc (shown in Fig 1.).

While it's good to know that cytobands exist and that areas of the Y chromosome are often referred to by these section references, there's no reason to memorize them. These cytobands are certainly useful for genetics but it turns out they are only *roughly* the same as the sections of the Y-chromosome that are useful for genetic genealogy.

Mapping the Y for Genetic Genealogy

Now that the human genome has been mapped, different references are regularly produced which continually change in minor ways the total number of nucleotide positions on the Y chromosome and the reference positions for each section (as well as the known position references for each SNP and other fun impacts).

The human genome reference in use across MOST of genetic genealogy today is Genome Reference Consortium Human Build 38, usually abbreviated as GRCh38 or hg38. In hg38, the Y chromosome is 57,227.415 base pairs in length, and these are numbered starting with 0 at one end and 57227415 at the other.

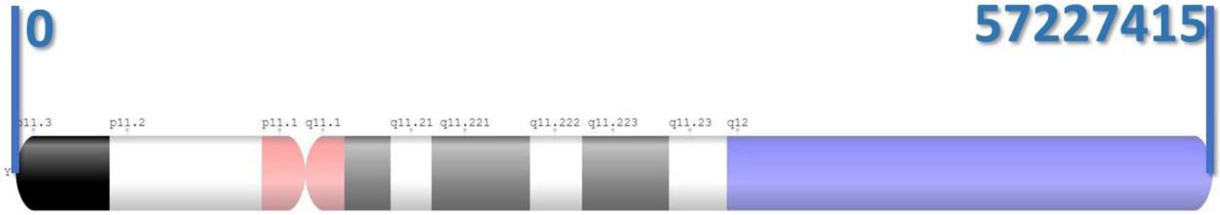


Figure 3. Base Pair Numbering on the Y chromosome (hg38)

When we say that a SNP has a position of 12345678, that simply means it's at a certain base pair location on the Y chromosome reached by counting 12345678 base pairs over from the left of Figure 3.

Now, PLEASE remember that this rather breezy overview of the Y chromosome is for genetic genealogists and vastly oversimplifies what is a true life's work for many geneticists, in fact it probably oversimplifies it so much that it would likely drive those same geneticists to rush out and drown their sorrows in a bucket of Bloody Marys. If you're truly interested in the genetics of the Y-chromosome please stop here and seek out more appropriate reference material for your interests (although if you really ARE interested in becoming a geneticist, a bucket of Bloody Marys is probably a good thing to have handy anyway).

In particular, I will only mention in passing that not ALL SNPs are single base substitutions at one base pair position like a 12345678C-G (meaning a change from C to G at base pair position 12345678). Sometimes a SNP is actually what's called an "Indel" and the base has been deleted altogether; other times it can be what's called a Multi-Nucleotide Polymorphism (MNP) meaning that it's one mutation that changed several bases in a sequence rather than just one. And the Y chromosome is also more complex than I'm covering here and geneticists have to deal with more complex region definitions like X-transposed, X-degenerate, ampliconic, and palindromic. All I'm trying to do in this discussion is cover enough explanation of the Y chromosome to clarify the approaches used in genetic genealogy without forcing any of us to become geneticists and put our entire sanity at risk.

Anyway back to the Y… the two tips of the Y chromosome are called the *telomeres*, and at the telomeres there are actually two regions that can recombine with the X chromosome. Since recombination is an activity more commonly associated with the autosomes (the non-sex-linked chromosomes), these two regions are known as the *pseudoautosomal regions* (PAR) and, since there are two of them, are helpfully named PAR1 and PAR2. PAR1 is longer at about 2.7M bps in size (I'll use the shorthand "bps" to mean "base pairs" as the unit measure of region size on the Y), and PAR2 is about 0.34M bps. These are represented in Figure 4 although the PAR2 region is so small in this diagram that it's barely noticeable.

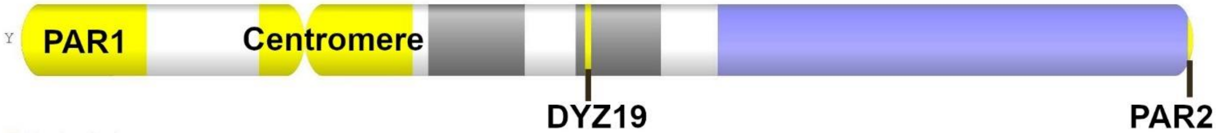


Figure 4. Some recombinant and difficult-to-read regions of the Y chromosome

Why is it important for genetic genealogy that some sections of the Y chromosome can recombine? One reason is because these regions don't have a stable ancestral reference genome value since by definition they don't have "stable" SNPs – any son might inherit the X chromosome nucleotides at positions in the PAR1 or PAR2 regions from his mother instead of the Y chromosome nucleotides from his father, so it's impossible to track back to what those nucleotides were thousands of years ago in any one man. And even if you DID identify a variant in those regions and find a man who is positive for it, his sons may not inherit that positive SNP so SNPs in these regions violate the basic principle of haplogroups that the entire Y-SNP haplotree is based on (i.e, that all descendants of an ancestor who was positive for a SNP will also be positive for that SNP).

NGS and WGS testing (Big Y, Y-Elite, WGS tests, etc) do cover and report SNPs sometimes that sit in these PAR1 and PAR2 regions, but with VERY rare exceptions they are not adopted onto Y-DNA haplotrees (the exceptions are a very few SNPs that are on the very edges of those regions and have not been affected by recombination). Sometimes several tested men within a family group will be found to have a "private" SNP from a PAR region in common, and it's tempting to use it as a branching SNP for that family group. The problem is that the SNP MAY be consistent in all the men who have tested so far but suddenly not match the actual genealogy in the next test of a member of that family. Since it's hard to know whether the results of these SNPs will match the actual genealogy, they're better left ignored.

It is certainly theoretically possible that the PAR regions will be so well-understood in the future that their recombination can be mapped and used for genealogy just like autosomal DNA. But for the foreseeable future at least they are not useful for Y-DNA genetic genealogy.

Besides the recombining regions of the Y chromosome there are also sections that are hard for the exploratory Y-DNA testing to read. In most cases this is because they are made up of highly repetitive regions which stymie the current "short-read" testing technology that low-cost NGS and WGS testing is based on. Targeted testing for specific SNPs uses a completely different method and actually CAN return results from what I'm calling the "difficult to read" regions. But that technology isn't cost-effective to apply to exploratory testing that seeks out SNPs across major sections of the Y chromosome. Y-DNA testing is expensive enough as it is and NGS/WGS testing is one of the major discovery tools that allow it to be useful for genetic genealogy, so at least for now the "short-read" NGS testing is the best compromise between affordability and discovery.

The reason some regions are more difficult is that chromosomes are made up of *euchromatic* and *heterochromatic* regions. Euchromatic regions are more "active" (i.e. have more genes), stain lighter in standard genetics research, and are easier to read using "short-read" NGS technology. Heterochromatic regions have fewer active known genes (although some studies suggest they carry backup copies of known genes), are more tightly packed (so stain darker), carry longer sequences of repetitive DNA, and are still very hard for Y-DNA testing technology to access. These heterochromatic regions are the difficult ones, and the Y is rather special in the world of chromosomes in that well over half of it is heterochromatic.

It turns out that the centromere (also marked in Figure 4) is one of these heterochromatic regions. The first linear map of the centromere was only published in 2018 using what's called "long-read" nanospore sequencing and a very readable media summary of this effort can be found here. Another media report on the challenges of reading certain areas of the Y can be found here.

It's not *impossible* to read "difficult" regions even with NGS testing, it's just difficult to "call" SNPs that are found there (i.e. identify positive or negative variants). So for instance the DYZ19 region (also marked in Figure 4) is one difficult to read region which in the early days of Y-DNA testing was the source of a lot of problematic calls. Mostly through the increases in numbers of people tested and the resulting consistency in reported results it has become easier to more consistently "call" SNPs in this region. It is shown as an example in Figure 4 but as it turns out there are many SNPs on the haplotree whose positions fall in this region. So not EVERY difficult region is a black box.

The largest heterochromatic region on the Y chromosome and one which is STILL mostly a black box is the q12 region marked in blue in Figure 4. This is a 30.3M bps region which is largely inaccessible to today's NGS technology. While Family Tree DNA (and likely other testing companies) has been successful in identifying SNPs at the edges of heterochromatic regions like q12, the bulk of it is not considered useful for now for genetic genealogy and will have to wait until testing technology is available at consumer-affordable rates that can identify variants there.

So what IS useful for genetic genealogy? After you subtract out all these hard-to-read regions, about 23.6M bps (out of 57.2M, or about 41%) of the Y chromosome is currently most useful for Y-DNA genetic genealogy.

At the time I'm writing this article the largest Y-DNA haplotree is Family Tree DNA's who have placed over 200,000 SNPs on their public Y-DNA tree. If you take all those SNPs and plot their positions on the Y chromosome, you get the orange regions in Figure 5.

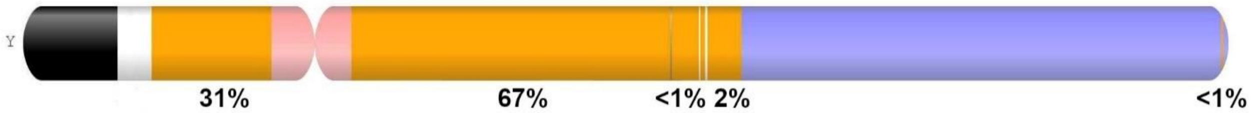


Figure 5. Region mapping over 200,000 SNPs from Family Tree DNA's Y-DNA Haplotree

Bear in mind that Figure 5 doesn't represent all the SNPs *discovered* by Family Tree DNA, just all the ones which have been mapped onto their public Y-DNA haplotree. These are (from Family Tree DNA's point of view) the highest-quality group of SNPs that mark the ancestral branching of all men.

The major point about Figure 5 is that leaving out the q12 region shown in blue, the orange regions of Figure 5 are nearly exactly the opposite of the yellow regions in Figure 4.

Figure 5 is a good representation of the 23.6M bps which (at least for now) make up the total useful area for Y-DNA genetic genealogy. 98% of these haplotree SNPs sit in the euchromatic regions of the p and q arms, but Family Tree DNA is doing a good job of "pushing the envelope" and mining the edges of heterochromatic regions as well, though the percentages found there are still relatively small.

If you plot a density map of the SNPs across the orange regions of Figure 5, you get Figure 6, which is another representation of the Y-chromosome with the same region mapping as Figure 5 but with areas of higher SNP density shaded in darker colors.

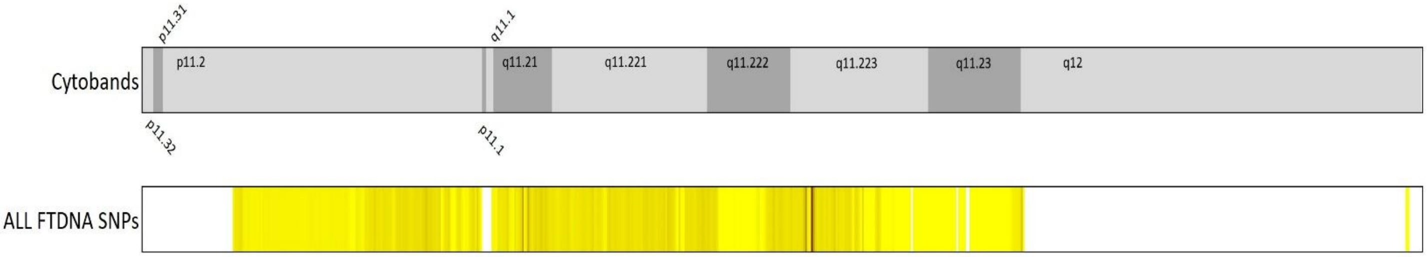


Figure 6. Density Map of over 200,000 SNPs on the Family Tree DNA public Y-DNA Haplotree (darker sections are more dense in SNPs; some blank regions like q12 have been shortened for readability)

The density map shows perhaps a little better how well-spread these 200,000+ SNPs are; while there are certainly a few areas where SNPs appear to cluster, they cover this 23.6M region very well and even push a little into both edges of the q12 region.

“Quality Regions” of the Y chromosome

People sometimes talk about “quality regions” of the Y chromosome in genetic genealogy and it’s understandable that you might think that this 23.6M area of the Y that we just talked about was the same thing. Unfortunately, it’s not.

We forget sometimes how young a field genetic genealogy really is, especially Y-DNA genetic genealogy, and how much we depend on the mapping of the human genome and more specifically of the Y chromosome which themselves are very new scientific achievements which allow us to consider the entire Y chromosome as one entity. Most genetic studies of the Y chromosome have taken a piecemeal approach and looked at either one part of it in which they were most interested, or have taken samples from select areas which were either interesting or easier to study.

In 2013, this study analyzed the SNP mutation frequencies over about 10.5M bps of the Y chromosome and found an underlying mutation rate of 8.2x10^{-10} mutations per base pair per year with about a +/-15-20% variation. The study analyzed 42 sub-regions of the Y-chromosome to develop this rate and these sub-regions are now colloquially called the “Poznik region” by genetic genealogists after the lead scientist for the study’s publication. The Poznik region (or more accurately, the 42 sub-regions) are marked in red in Figure 7.

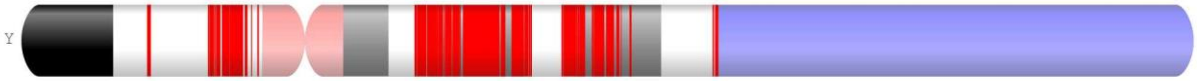


Figure 7. 10.4M bps Poznik Region (marked in red)

Unsurprisingly the Poznik subregions are all located in euchromatic regions of the Y chromosome which would have been easier for the study to analyze.

In their 2015 paper introducing their age estimating method (usually referred to as the “Adamov paper”), YFull introduced a 8.4M bps region which they called the “combBED region”; which was originally developed from the Poznik region’s overlaps with the 11.38M bps generalized original Big Y coverage region from Family Tree DNA’s 2014 white paper describing the original Big Y test, although it does include a few small areas of coverage which were not included in the Poznik region. However, YFull’s work yielded 857 sub-regions which make up the combBED and which are shown in Figure 8.

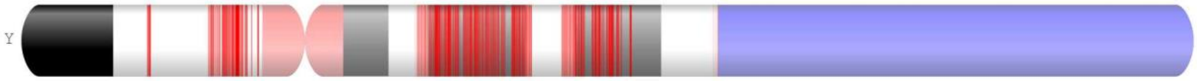


Figure 8. 8.4M bps combBED Region (marked in red)

Like the Poznik paper, the Adamov paper also calculated an underlying SNP mutation frequency of 8.2x10^{-10} mutations per base pair per year.

Another 10.9M bps "quality region" has been developed by Dr. Iain McDonald and is colloquially called the "McDonald region". It is made up of 10 sub-regions which are defined in Dr. McDonald's age analysis approach description. These 10 sub-regions are shown in Figure 9.

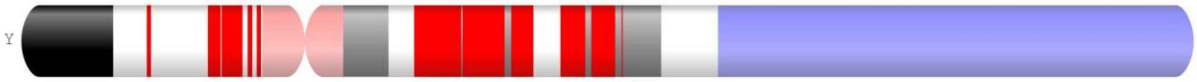


Figure 9. 10.9M bps McDonald Region (marked in red)

Each of these three regions overlaps significantly with the others but also has some unique coverage of its own. If you merge all three together, you get one large 11.6M bps region which is diagrammed in Figure 10. While no-one uses this merged "superset" of the three best-known "quality regions", it's useful to see it here and note that it still fits neatly into the larger 23.6M bps region across which haplotree SNPs are being mined.

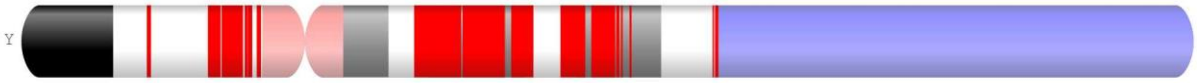


Figure 10. 11.6M bps Merger of all three regions (Poznik, combBED, Mcdonald), marked in red.

It doesn't take much effort to notice that the usual quality regions are a subset of the total 23.6M bps area represented by the spread of the 200,000+ SNPs on the haplotree. It's certainly *likely* that the whole 23.6M region can also be considered a quality region in its own, but the MAIN difference is that the 8.2x10^{-10} SNP mutation frequency has only been validated through published scientific study across the smaller regions used in the published studies. It may certainly be true that the same frequency holds across the entire 23.6M region, but that's an assumption for now.

Y Chromosome Regions and SNP Age Estimating using the Family Tree DNA Block Tree

When I say "SNP Age Estimating", I'm referring to the process, usually manual but currently automated by YFull, of estimating how long ago a particular SNP was formed. Normally people try to get the closest applicable years-per-SNP average (often "144" or "83" are quoted) and multiply the number of SNPs under a particular branch to arrive at an age estimation for the SNP at the top of that branch.

It's important first to note that the indiscriminate use of any years-per-SNP estimate without considering the underlying coverage area is incorrect, and no study or accepted practice in genetic genealogy supports it. If you think about the whole past discussion about where SNPs are discovered on the Y chromosome, it should be obvious why SNP mutation rates are tied to coverage areas: SNPs appear to occur across the entire Y chromosome, but when we test for them we're looking at less than the total 57M bps region in which they occur.

Suppose I could test a man living today AND his paternal ancestor who was born 400 years ago. Let's also say for simplicity that the underlying $8.2x10^{-10}$ SNP mutation frequency that the studies above found is true across the WHOLE Y chromosome (leave out the two PAR regions since recombination really messes things up).

The math is relatively simple: if each base pair position has a $8.2x10^{-10}$ per year chance of having a mutation, and the entire Y minus the PAR regions is 54,106,422 base pairs in length, then the base pairs in that whole region will on average have a mutation every $1/(8.2x10^{-10}$ x54,106,422) = every 22.64 years. So over 400 years there will have been 17.67 SNP mutations on average – say 18 SNPs for the sake of this example.

Note that while I'm making assumptions here about mutation rates and so on, it doesn't really matter what number I start with for the rest of this example. A man living today and his ancestor from 400 years ago will likely have SOME number of SNP differences across their entire Y chromosome, whether it's 18, 4, 25 or anything else. They're still relatively random and so are likely spread across the whole Y chromosome.

But say we do have 18 SNPs, and they fall all over the place on our Y chromosome (apart from the PAR regions). If I diagram them out, they might look something like this:

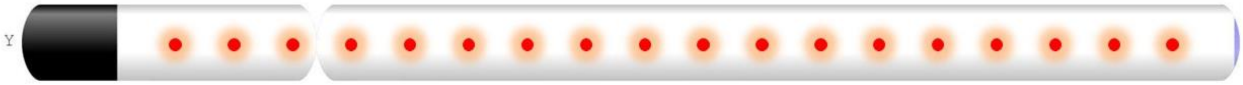


Figure 11. 18 SNPs all spread across the Y chromosome

Now we already know from the last 8 pages of my riveting explanations that we can't actually test for all 18 of these SNPs today. In fact using today's NGS testing technology, the best we can probably get is 23.6M of those 54.1M bps, but any single NGS test is only going to cover the amount that it's targeted for. Big Y700 on average finds SNPs across about 14.7M bps; other tests have different average coverage (a lot of REALLY good data on the various tests can be found on the Y-DNA Warehouse Statistics webpage).

So I'll never find ALL 18 SNPs through NGS testing. Assuming these SNPs are relatively evenly spread across the Y chromosome, I'll probably find 4-5 of them if I can test 25% of the Y, 6 of them if I can test 33% of the Y, 9 of them if I can test 50% of the Y, and so on.

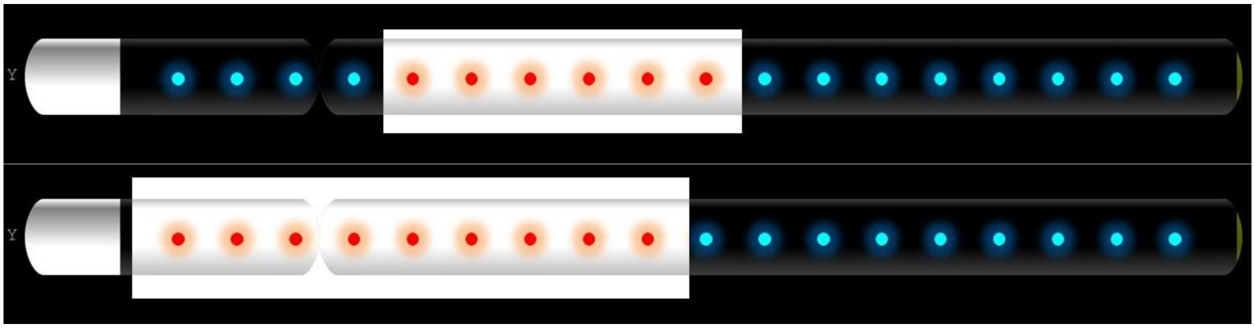


Figure 12. Depending on how much area my test covers, I'll count a different number of SNPs

Say I tested 33% of the Y and found 6 SNPs. From testing another descendant of my 400-year-old ancestor, I know that these 6 SNPs occurred in the last 400 years, and that gives me 400/6 = 66.67 years per SNP.

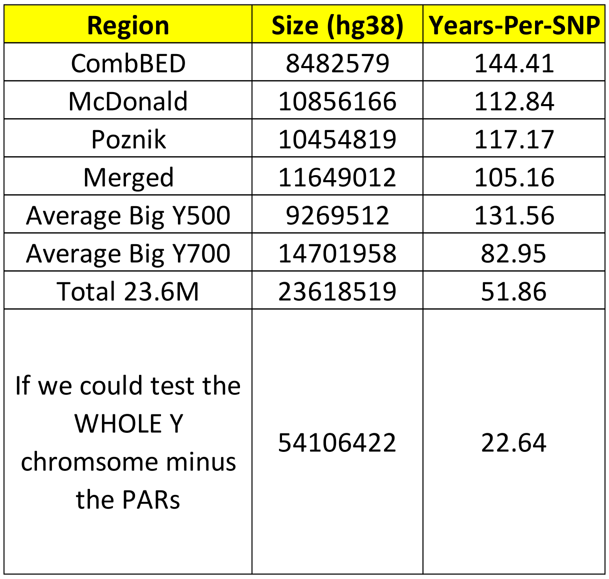
Say I tested 25% of the Y and found 4 SNPs. From testing another descendant of my 400-year-old ancestor, I know that these 4 SNPs occurred in the last 400 years, and that gives me 400/4 = 100 years per SNP.

See how the coverage area of the Y-DNA test that I took affects my years-per-SNP estimate, even though *I started with a single mutation rate to begin with*?

Generally speaking, a smaller test region size will always yield a higher years-per-SNP number because you're counting fewer SNPs. A larger test region size will always yield a lower years-per-SNP number because you're counting more SNPs.

Again this makes sense if you think about working backwards to calculate the SNP age estimates. If my test found 4 SNPs back to a common ancestor and I was using 100 years per SNP, I'd get 4x100 = 400 years. If I took a more advanced test that covered more of the Y chromosome and I found 6 SNPs and I was using 66.67 years per SNP, I'd get 6x66.67 = 400 years. Both will get me the same answer *as long as I use the years-per-SNP estimate that is appropriate for the test coverage that I'm looking at*.

Just for comparison's sake, here are the years-per-SNP estimates for all the regions that I've been talking about here (assuming again that the $8.2x10^{-10}$ underlying rate holds true across all these areas):



Region	Size (hg38)	Years-Per-SNP
CombBED	8482579	144.41
McDonald	10856166	112.84
Poznik	10454819	117.17
Merged	11649012	105.16
Average Big Y500	9269512	131.56
Average Big Y700	14701958	82.95
Total 23.6M	23618519	51.86
If we could test the WHOLE Y chromsome minus the PARs	54106422	22.64

Figure 13. Years-per-SNP figures for the various coverage areas (Big Y figures sourced from Y-DNA Warehouse Statistics).

So at this point you're probably saying to yourself… ok net this out please. I'm on the Family Tree DNA Block Tree trying to figure out how old a branching point is – what years-per-SNP rate should I be using?

I'll get to that in a couple of paragraphs but I need to point out some things first from Figure 13: starting with the fact that the years-per-SNP figure across the *whole* 23.6M bps region that Family Tree DNA is using for the 200,000+ SNPs that are on their haplotree is 51.86. But… even a Big Y700 only on average covers 14.7M bps, not 23.6M. So the repeated application of Big Y testing has resulted in SNPs being found across up to 23.6M bps, but each individual test has covered between 9.2M (Big Y500) and 14.7M (Big Y700) bps.

The net effect of this is that the years-per-SNP on the Family Tree DNA Block Tree goes *down* as you get higher up in the haplotree. This should make intuitive sense as well – if you look up the Block Tree at the blocks of older SNPs, they'll include some SNPs in those older blocks that weren't actually covered in *your* test but they were found in enough other tests that Family Tree DNA knows where to place them. So everyone gets the benefit on the haplotree of SNPs that were found in regions that their own test didn't even cover. At the highest levels of the haplotree, this frequency SHOULD approach 51.86 years per SNP. But down at the individual test levels close to present day, the frequency is only as good as whatever specific testing has been done.

I validated this recently by conducting some analysis of the haplotree under L513, which is a 3800-year old SNP in R1b. At the time there were 422 Big Y tests which had been taken in L513, fairly evenly split between Y500 and Y700, leading to about 413 SNPs on that portion of the haplotree.

The resulting spread of those 413 SNPs shows that they do make up a portion, but not all, of the full 23.6M region (compare Figure 14 to Figure 5 above):

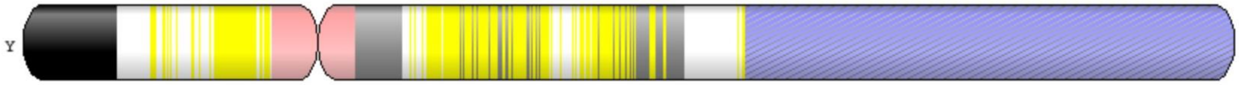
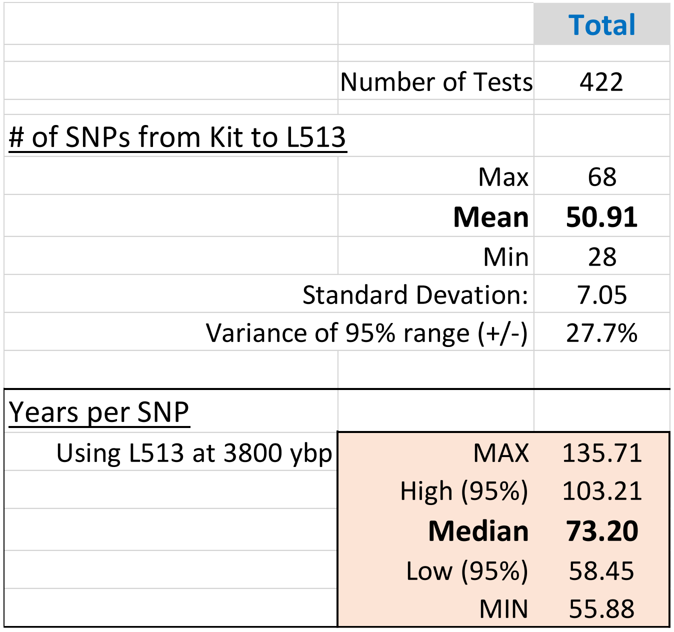


Figure 14. SNP Distribution for 413 SNPs under R1b-L513 on the FTDNA Haplotree

And the analysis of SNP branches under L513 gave an average years-per-SNP estimate of 73.20 years on average. Note that the variance is +/- 27.7% - I don't talk about variance in SNP mutation rates in this discussion but bear in mind it is VERY large. **Even SNP-based estimates can vary by nearly 30% - that's 100 years on EITHER side for every 330 years back in time.**



		Total
	Number of Tests	422
# of SNPs from Kit to L513		
	Max	68
	Mean	**50.91**
	Min	28
	Standard Devation:	7.05
	Variance of 95% range (+/-)	27.7%
Years per SNP		
Using L513 at 3800 ybp	MAX	135.71
	High (95%)	103.21
	Median	**73.20**
	Low (95%)	58.45
	MIN	55.88

Figure 15. Figures derived from just the R1b-L513 portion of the FTDNA haplotree

So back to "your" question: what years-per-SNP estimate should I use?

Well you have two other options of course: first, wait for Family Tree DNA to release their own age estimation process which we've heard they will do "shortly" but hasn't yet been given a release date. When they do, it'll apply all the underlying test information necessary to calculate the ages at least as closely as is statistically possible.

The second option is that you could get all the SNP positions from present day (including private SNPs) back to the SNP that you want the age of, and check which ones are within the combBED region. That would allow you to estimate the age of the SNP using the 144.41 years-per-SNP figure that's appropriate to the combBED. Or, simply download your test BAM file and pay the $49 to YFull to do that analysis for you (yes, I know people sometimes have a problem with sharing raw data. But we wouldn't even have genetic genealogy if everyone felt that way).

If you DO want to just estimate an age using the SNPs on the Block Tree, I would suggest the following. Bear in mind please that in suggesting a simple approach I'm approximating wildly, but since there's nearly a 30% error margin on SNP-related age estimates anyway the approximations don't add much additional error themselves.

1. If you're including SNPs in any blocks older than 3000 years ago, use 60 years-per-SNP for those higher branches;
2. If you're including SNPs in any blocks between 1000 and 3000 years ago, use 75 years-per-SNP for those branches;
3. Within the last 1000 years you need to know how many of the tests are Y500 versus Y700.
    a. For blocks of SNPs with only Y700s under them or Y500s AND more than one Y700 test under them, use 83 years per SNP;
    b. For blocks of SNPs with only Y500 or Y500s with only one Y700 test under them, use 130 years-per-SNP.

This takes more work than just using one single years-per-SNP figure. But it's MUCH more likely to be closer to the mark.

I will reiterate again that the error ranges of SNP mutation are large enough that you will have problems applying even this approach very close to present day. As an example, this is the Block Tree in one area of the haplotree close to present day:

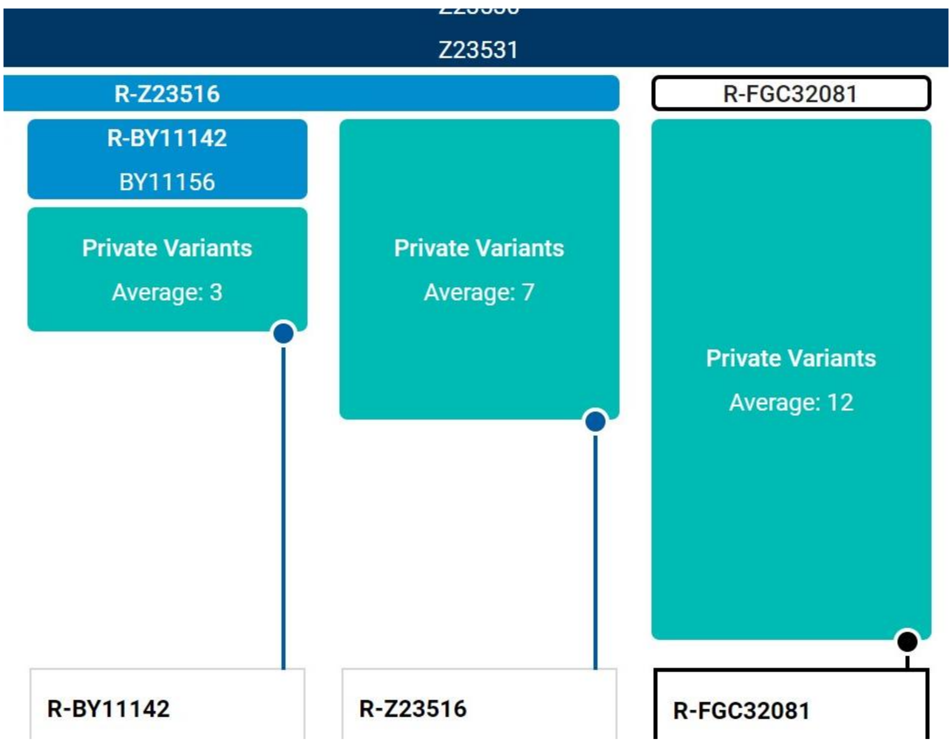


Figure 16. One example of varying SNP mutation rates

The SNP block that ends with Z23531 spawned off 3 branches, one with a total of 6 SNPs (3 shared, 3 average privates), one with 8 SNPs (1 shared, 7 average privates), and one with 13 SNPs (1 shared, 12 average privates).

So over the same approximate amount of time, one branch had 6 mutations, one had 8, and one had 13. It's impossible to get a single years-per-SNP rate that applies to all three of those branches. Clearly the 13 SNP branch had somewhere near double the mutation rate that the 6 SNP branch did. You can average them and get 9 SNPs, but the average doesn't really apply to any of the branches, it's only an average.

In these cases though I DO average them and (because I know these are all Y700 tests) I would apply the 83 figure and get an average age estimate (rounded) of 750 years ago for when these branches all split off from each other. But when I apply the additional knowledge that they all carry the same surname today I realize that even that estimate is probably too old. Based on other information, I've placed this age as more likely about 500 years old or so. But that's the variances you live with in SNP-based age estimation.