Binary Logistic Regression

The mathematical model for YSNP prediction based on YSTR signatures

By Robert Casey

February 3, 2018

This article provides documentation that substantiates that binary logistic regression is the best mathematical model that represents the relationship between YSTRs and YSNPs. YSTR signatures have been used for a long time to predict YSNP haplogroups by many haplogroup administrators. This article just documents that the best statistical model that supports this empirical observations is binary logistic regression.

A binary logistic regression model was used to create the R-L21 SNP predictor tool. During the “Walk The Y” era, this tool was kept up to date since the discovery of YSNPs was at a reasonable pace to keep up with. Even today, this tool still accurately predicts around 50 % of all R-L21 testers for around fifty major signatures under R-L21. With the explosion of NGS testing, keeping up this tool updated would have become a full time job and this tool has not been updated very much since Big Y testing started rolling around over 25,000 Big Y tests.

The first version of the SNP predictor tool was solely based on one variable, the number of markers that match the haplogroup’s signature. This predictor tool required that any tester had to be confirmed R-L21 positive prior to predicting YSNPs under R-L21. This was basically a filter for convergence of YSTR markers across haplogroup R. This model works with over 95 % accuracy for those YSNP branches that have no convergence with other haplogroups under R-L21. This model had two issues: 1) it does not accurately predict the boundary condition testers which are key to research; 2) most commonly used statistical software tools (Minitab and SPSS) do not handle perfect curves as well as they should (this is an actual known defect in these software packages). In addition to these two issues, the requirement for testing R-L21 positive was a major requirement that many testers do not satisfy.

The latest iteration of the YSNP prediction tool now uses two variables: number of markers that match the signature and genetic distance from the signature. Unfortunately, this is no longer a simple X vs. Y plot as there are now two variables. However, software defect for the perfect curve defect is no longer an issue which makes using the statistical software tools much easier. The genetic distance variable acts as a genetic distance filter and removes the requirement to verify R-L21 status (now any haplogroup R will work and haplogroup R is almost always predicted by FTDNA via another YSNP prediction technology that works for much older haplogroups. The two variable model also properly predicts most of the boundary condition testers if the boundary condition testers are tested to any reasonable degree.

But there remains one minor issue with the two variable model – if the boundary condition testers with the lowest signature matches and the highest genetic distance are not YSNP tested (untested data), it is usually will predict these testers incorrectly since the constants are based only with tested data. This is usually less than one percent of testers with any reasonable sample size but does require running the statistical software regression with updated data if new boundary condition testers later test positive. When the data input run a second time, the constants will slightly change and will usually properly predict the newly tested boundary condition testers. So it is important to test boundary condition testers as much as possible in order to avoid multiple updates to the constants of the model.

It is believed that 80 to 90 % of R-L21 testers could now be predicted to YSNP branches in the relevant time frame with 99 % or higher accuracy. However, there are some haplogroups in the proper time frame that have significant convergence with other haplogroups. There also remain limitations for time frames of the haplogroups that can be predicted. To remain very high accuracy, the haplogroups must be close to the 1,200 to 2,500 year time frame. These time frames can be slightly older or slightly younger if the signature is very large. For the older YSNP branches such as L513, Z253, Z255, DF41, Z251, etc., these YSNPs are too old to accurately predict without using numerous signatures. However, 90 % of these haplogroups are ancestors of predictable YSNPs in the 1,200 to 2,500 year range. These single signature YSNPs include L226, M222, L193, L555, L371, etc. which can usually be predicted with over 99 % accuracy. R-L743 is estimated to be around 750 years old by BigTree and although YSNP prediction works fairly well, currently R-L743 is tracking one YSTR value under the signature, so accuracy could suffer significantly in this time frame.

During the last two years, I have shifted my analysis to mainly charting of R-L226 where the three R-L226 administrators and dozens of other R-L226 researchers have made tremendous progress in YSNP testing to complement YSTR tests. R-L226 now has over 100 NGS tests, over 100 SNP pack tests and over 200 individual YSNPs tested at YSEQ. This testing has revealed 72 new branches under R-L226 and over 1,000 private YSNPs and branch equivalents YSNPs associated with R-L226. With over 630 testers at 67 markers (with over 250 of these being at 111 markers), I am now able to chart 85 % of R-L226 with accuracy ranging from 60 to 95 % (accuracy is based on the size of the signature matches and genetic distance). Therefore, binary logistic regression is being extended from the predictable haplogroup time frame down to the present day testers. However, predictable haplogroups will not be fully charted as many testers have unique signatures that have not been YSNP tested to date.

During this last iteration of analysis, two variable models were created for three R-L21 haplogroups, L226, L555 and L371. Based on the current YSNP testing, all three models have 100 % accuracy for prediction. However, regression testing use curve fitting methodology that creates a model that matches the data that is being input. If no convergence is present with the existing data, it should almost always predict at 100 % accuracy. However, binary logistic regression does not use untested data, so if future boundary condition testers later become tested, the model will slip to 99 % accuracy or you can re-run the model to get new constants that usually brings accuracy back up to 100 %.

Statistical Review of YSNP prediction under R-L21

This article reviews the mathematics supporting the prediction of testing positive for YSNPs based on YSTR signatures. YSNP prediction is based how well submissions match the YSTR signature associated with each YSNP and the genetic distance from the YSTR signature. This YSNP prediction methodology is limited to more recent YSNPs that can be expressed by only one YSTR signature. Older and broader YSNPs started out with just one YSTR signature but more than 2,500 years of parallel and backwards mutations now appear as multiple signatures and YSNPs with multiple signatures will not be predicted with great accuracy. With the recent explosion of Next Generation Sequence testing, many younger YSNP branches are being discovering on a daily basis. Prediction of most of these younger YSNP will not be predicted with much accuracy as well. This is due to the fact that there just has not been enough time to generate large enough signatures to predict accurately. It appears that the sweet spot will be YSNP prediction is limited to those YSNP branches that formed between 1,200 and 2,500 years ago. Other YSNP prediction methodologies could be used for much older time frames. Charting with YSTRs and YSNPs involves another variation of YSNP prediction which is also based on signatures and genetic distance. This article primarily focuses on YSNP prediction.

Using only one variable, signature, any user of the R-L21 prediction tool must have tested positive for R-L21 or any YSNP descendant of R-L21. This methodology could be applied to other YSNPs similar to R-L21 in age and scope in any part of the genome. The reason for this requirement is due to “convergence” of YSTR signatures between many haplogroups testers under haplogroup R. This is where YSTR patterns (signatures) can overlap each other even when the common ancestor is more than two or three thousand years old. This convergence is believed to be between five and ten percent of R-L21 testers. For these YSNP branches, accuracy would be degraded but could be acceptable for research. Most YSNP branches have no convergence or very little convergence. Another reason for the R-L21 restriction was the original lack of access of data under the entire haplogroup R. In the last year or so, this issue has been eliminated. The restriction of testing positive for R-L21 can be removed if the model used is expanded to include genetic distance as a second variable which is a major improvement in YSNP prediction. It has been observed (on a very limited scope today) that convergence of signatures across haplogroup R also comes with very high genetic distance. So even though that there is some overlap of signatures of seven to twelve markers, overlap at 67 markers is much less common than originally believed due filtering by genetic distance (adding genetic distance as a second variable to the model).

All YSNP signatures are currently based on 67 YSTR markers, so current YSNP prediction requires 67 or more YSTR markers tested. YSNP prediction accuracy would only have very modest increase in accuracy at 111 markers since most 67 marker prediction is already at 99 % to 100 % accuracy for most haplogroups. For those haplogroups with convergence, 111 marker signatures would surely eliminate some convergence for many haplogroups. However, charting does have significant benefit from 67 to 111 marker upgrades. YSNP prediction using 37 markers submissions do not produce large enough YSTR signatures for accurate YSNP prediction across most of R-L21 but there are a small percentage of YSTR signatures that have most of the markers in the first 37 markers which could be potentially YSNP predicted as well. However, predicting only a small percentage of 37 marker signatures does not warrant the effort of implementing. For charting, 67 markers are required for comprehensive prediction and 111 markers would be useful as well (even though mixing resolutions presents challenges). Prediction at 37 markers would only cover around ten or fifteen percent of the genome with high accuracy. Prediction of YSNPs with 111 markers should be investigated when convergence is present.

With the explosion of NGS testing that came in with wide usage of FTDNA's Big Y test, the number of predictable YSNPs increased at such a fast rate that the manual spreadsheet analysis in detecting new signatures is too laborious to even keep up with the 100s of known predictable R-L21 branches. However, it is now believed that signature recognition could be automated which may allow YSNP prediction to be re-introduced across the entire genome. However, any R-L21 testing requirements would need to be removed first via introduction of genetic distance into the binary logistic model. Another challenge in automating YSNP prediction is being able to recognize and adjust for characteristics of YSTR signatures that greatly reduce accuracy. Genetically isolated YSTR signatures are easy to recognize but there are there are five to ten varieties of overlap between positive and negative testers that would also have to be automated.

The need and economics of YSNP prediction is becoming less a burden due to several factors. Currently, most individuals need to order one and sometimes two SNP packs at $100 to $120 each to verify what YSNP prediction can easily predict without testing. Also, as YSNP testing includes even more content or as prices continue to decline directly, the overall cost that YSNP prediction eliminates will continue to decline. In just five or ten years, Whole Genome Sequence tests will fall to $200 to $300 range and will become the only YDNA test required which will greatly reduce the need for YSNP prediction. However, in the mean time, ordering of 10,000s of YSNP packs will be an economic burden on the genetic genealogy community for a significant time frame. Full Genomes already offers a NGS test where all 111 markers can be extracting which eliminates the need for YSTR marker tests. Unfortunately, this test is $2,900 but the cost of this test will drop and the 1,000,000 read length is not really required. This NGS test also read around 90 % more private YSNPs than the current Big Y as well.

History of the R-L21 SNP prediction methodology

Based on curve fitting methodology, the first iteration of the R-L21 YSNP prediction tool was based on observed empirical data and analyzing the trend of matching the YSTR signatures of the YSNP by observing negative and positive tests for each YSNP in relationship with its YSTR signature. As the signature match for any YSNP decreases, the probability of testing positive declines as the YSTR values begin to overlap with other haplogroups that are not related to the YSNP being analyzed. By 2012, this empirical methodology was confirmed and replaced the statistical model of binary logistic regression. By 2014, the R-L21 tool was implemented for around fifty R-L21 YSNPs that were primarily single YSTR signatures. At this time, only around ten R-L21 YSNPs were too old to predict. The R-L21 Predictor tool was expanded to include several two signature YSNPs since there were not many YSNPs being discovered by "Walk The Y" tests and YSNP prediction technology was pushed to see how well prediction would work for older YSNPs. Each of these multiple signatures for older haplogroups should now be replaced with several more recent YSNPs that cover most of these older YSNPs. This will increase the accuracy due fitting the requirement of the 1,200 to 2,500 time frame and less time for YSTRs to produce excessive amounts of hidden mutations (independent parallel mutations followed by a dependent backward mutations).

Earlier versions of this paper validated the original empirical approach and analyzed the mathematics behind the empirical curve fitting methodology. As expected, this also increased the accuracy of the YSNP prediction tool and introduced a more automated methodology for YSNP prediction based on sound statistically mathematical models (formulas). After only a brief review of a few possible models that would be appropriate for this kind of YSNP prediction, it was quickly apparent that the best model for YSNP prediction was clearly binary logistic regression. Several binary logistic regression formulas produce very good matches for the empirical data. Additionally, the measurements of accuracy for several binary logistic regression models examined were found to have very high accuracy (usually higher than 95 %).

However, YSNP prediction that most closely maps the classic S-Curve of the classic binary logistic regression formula is the following model:

y=e**(a+b*SIG)/(1+e**(a+b*SIG))

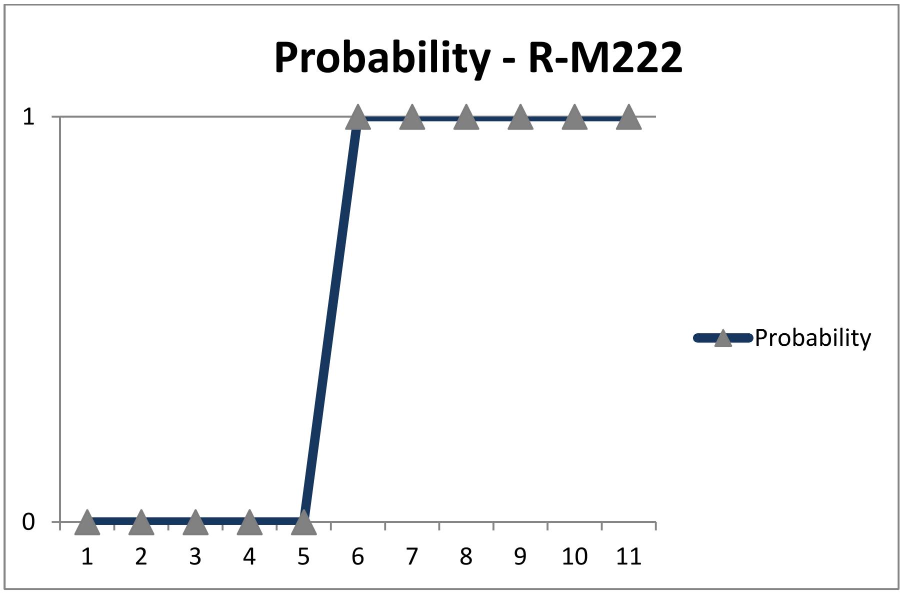
In this formula, SIG represents the number of markers that match the signature of the haplogroup. This is the independent variable of the equation and Y represents the probability of testing positive (0.0 to 1.0) for haplogroup which is the dependent variable. The terms “a” and “b” are the two constants that the regression model will calculate to make the curve fit the data. The term “**” means taking the power of. Examples are 2**2 = 4, 2**3 = 8 and 3**2 = 9. It merely means that the number on the left of this operator is multiplied by itself by the number of times by the number on the right of the operator.

The term "e" is a numerical constant that is approximately 2.71828. Just as pi (3.14159) is a numerical constant that occurs whenever the circumference of a circle is divided by its diameter, the value of "e" is found in many mathematical formulas such as those for the statistical "bell curves" or the shape of a hanging cable. The sound energy decays as it moves away from the sound source by a factor that is relative to "e." Because it occurs naturally with some frequency in the world, "e" is used as the base of natural logarithms. The Swiss mathematician Leonhard Euler was also the first to use the letter e for it in 1727 (the fact that it is the first letter of his surname is coincidental). As a result, sometimes e is called the Euler Number or the Eulerian Number. The actual mathematical formula for “e” is:

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

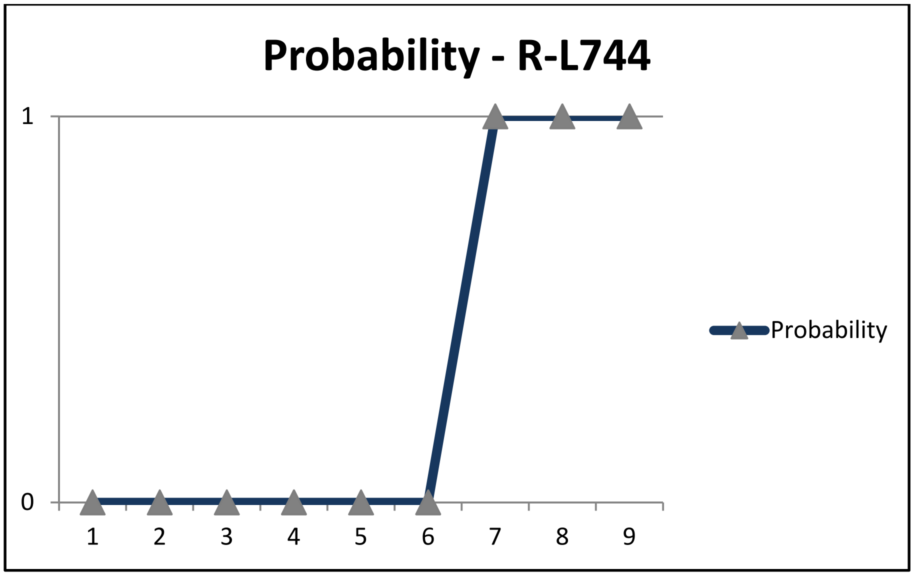
Within the spreadsheet application, EXCEL, “EXP” represents “e**” (or you can also use “2.71828^” as well). Even though this formula (model) may seem complex, the output of this formula is easy to visualize as shown on the next few pages.

Before the introduction of Big Y testing and large numbers of 67 marker testers, sample sizes were quite small, signature size was used exclusively for YSNP prediction. Around 50 % of R-L21 YNP prediction graphs were “perfect” curves. A perfect curve is where all testers test negative as the value of X rises (signature matches rising) which is immediately followed by all testers testing positive with higher values of X. The X axis is the number of markers that match the R-M222 signature and the Y axis is the probability of testing positive. This chart was created in 2012 when R-M222 only had 71 testers that had tested positive for R-M222. At this point in time, R-M222 had no descendant branches and R-L21 only had 31 descendant branches.



However, there were several issues with these curves. First, in order to greatly reduce the influence of convergence, only confirmed R-L21 testers were allowed which put a big burden on testers since R-L21 is unable to be predicted and required expensive single YSNP testing to prove this requirement (this was before the introduction of YSNP packs). Also, it was suspected that a few testers could test negative for R-M222 with 6 of 11 matches the R-M222 signature as well as a few 5 of 11 matches could test positive. Third, the statistical packages still have defects even today to handle “perfect” curves and quality of parameters was incorrectly stated as low quality when obvious observation says otherwise.

Even many smaller scope and younger YSNPs enjoyed "perfect" curves as well. In 2012, L744 was one of three branches under DF41 and only had eighteen testers that tested positive for L744. These charts are easy to understand and really made signatures a very important part of any haplogroup analysis.



Also, we suspected that much larger sample sizes of testers would eventually reveal testers that were positive when the perfect curve says they should be negative, but this would probably only result in very small acceptable error rates. However, these predictions of around 50 signatures in the R-L21 prediction tool did reveal that not all YSNP branches are predictable. We now know that this methodology only works in the 1,200 to 2,500 year range. This time range should not include the large numbers of branch equivalents being found today since we really do not care about the actual date of the YSNP mutation, we are primarily concerned only with date that includes over 95 % of the testers (the date they started to become prolific in numbers). So when picking TMRCA dates (Time to Most Recent Common Ancestor), always select the lower value for this criteria.

Another significant limiting factor was that entire collection of all haplogroup R testers initially required manually reviewing thousands of projects and manually converting 10,000s of testers into a format compatible for spreadsheets. So at the time that the R-L21 YSNP predictor tool was rolled out, data was not available for all of haplogroup R. This is again becoming a significant issue as haplogroup R now has over 54,000 67 marker testers spread across over 4,500 projects. Fortunately, this data extraction has been automated but constantly needs code changes to reflect the changing nature of the FTDNA web servers are deployed and the changing format of their YSTR and YSNP reports. Also, all relevant YSNP testing is never properly loaded into FTDNA YSNP reports and significant testing from YSEQ and Full Genomes need to be tracked as well.

However, around two years ago, I was able to finally extract all of haplogroup R for 67 markers only. This improved database provided an opportunity to remove the requirement for testing R-L21 positive as a requirement to using the R-L21 predictor tool. To my great surprise, there is significant convergence of signatures across the entire haplogroup R. However, it was also observed that this convergence was only due to the small numbers of markers found in signatures at 67 markers and it is also very obvious that the genetic distance was extremely high for these testers. So it looked pretty obvious that we needed some kind of genetic distance filter to remove these testers, so genetic distance was added to the model.

The model for the two variable YSNP prediction model is very similar to the one variable model:

y=e**(a+b*SIG+c*GD)/(1+e**(a+b*SIG+c*GD))

By adding genetic distance to the model as a second variable, the vast majority of convergence of the signatures was dramatically reduced. Unfortunately, this model has two variables which do not produce very user friendly graphs. But the addition of the second variable allowed the requirement of R-L21 to be replaced by being predicted to be Haplogroup R. Since FTDNA predicts very old haplogroups via a different YSNP prediction methodology, no prerequisite YSNP testing is required. Since the prediction of Haplogroup R is almost always available with no YSNP testing required, this is a significant improvement in end user ease of use (and reduces testing costs associated with verifying tester’s R-L21 status). The second parameter also removed the defect in the statistical software packages as the curves are no longer “perfect” curves due to having two variables. Another major improvement is that boundary condition testers are now properly predicted at significantly higher numbers which really helps explore the progression of the YSTRs under any haplogroup.

During the last two years, the requirement of testing R-L21 was found to be unnecessary by simply adding empirical genetic distance filters to remove convergence of the signatures with other haplogroups. This empirical filtering involved a second variable, genetic distance. Recently, it was decided to analyze YSNP prediction using two variables in the logistic regression model and adding genetic distance in the input stream. This was easily implemented by introducing this second variable to the regression model. For most signature research, EXCEL spreadsheets are used extensively for YSNP prediction with signature match values and genetic distance values being used. Since EXCEL spreadsheets already include signature and genetic distance, only one new column is required to add the actual binary logistic regression model formula. But generating these models does require access to a statistical software package which is too costly for wide usage by the genetic genealogy community.

Another major analysis was conducted on the affect of mutation rates of YSTR markers. This was done primarily for charting but also created a change in YSNP prediction as well. Using Burgarella mutation rates, it was determined that 40 % of all 67 marker mutations were CDYa and CDYb marker mutations. Currently, it has been decided to remove CDY markers from both YSNP prediction and YDNA charting. In fact, once YSTR500 rolls out, YSNP prediction could need to have many of the faster mutating markers removed which would allow prediction of much older haplogroups. With the recent addition of 111 markers to charting, it has been found that 710 and 712 also have mutation rates that approach the rates of CDY markers. For now, 710 and 712 are being charted to see how reliable these faster mutating markers can be for charting (acceptable but challenging in the first several rounds of analysis). But for now, only CDY markers are excluded from both YSNP prediction and charting. YSNP prediction at 111 markers also seems to be unnecessary since we are getting 100 % with most models at 67 markers. However, where convergence is present with nearby haplogroups, using 111 markers for YSNP prediction may be a viable option for those 5 to 10 % of the haplogroups where significant convergence is observed.

However, it has been found that the model using two variables has one new significant issue in respect to untested data. You can either accept a one percent error rate with the first set of constants generated with a reasonable sample size or you must run new data through the regression model again if lower signature matches with higher genetic distance testers test positive for the first time. These new data points allow the regression model to usually improve the accuracy back to 100 % by updating the constants. However, empirical approaches have one major advantage over models – you can look at the upper limit of genetic distance for those that test positive and the lower limit of those that test negative which the model does not examine. It is quite obvious that further testing in between these two limits could have mixed results. If you just average the genetic distance between these two limits, it would probably produce, on the average, better prediction than the model. However, we are only talking about one percent of the testers.

Empirical models are still very effective and can work around the limitations of only using tested data for binary logistic regression models. These models can be used to determine the extreme limits of boundary condition testing. But keep in mind that if your haplogroup has a reasonable sample size, these boundary condition testers will almost always be less than one percent of the testers. But this one percent is important for analysis of the progression of YSTRs in the very early part of the haplotree but has minimal impact on the overall accuracy of the model. Below is the current empirical model for L226 (not using CDY markers):

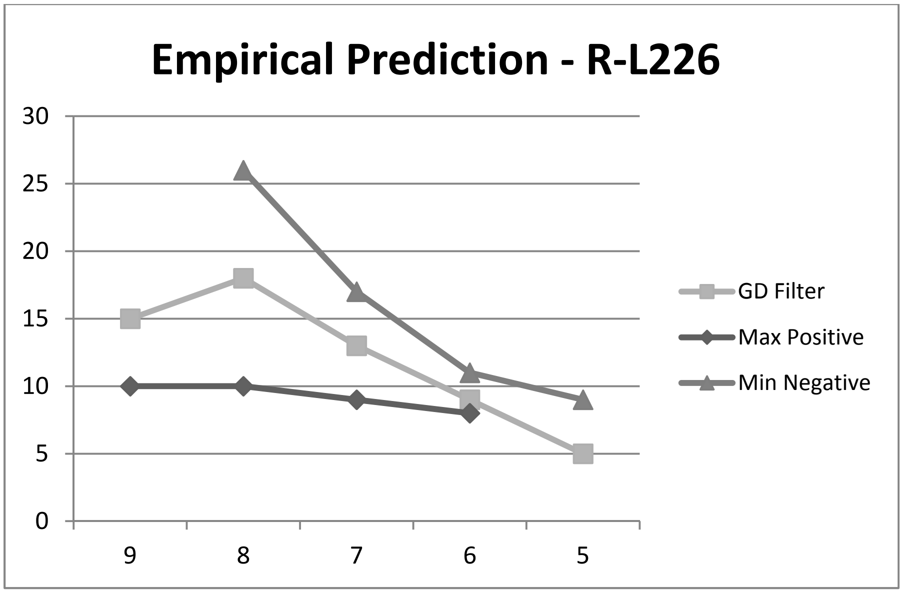
For L226, the improved empirical YSNP prediction would be:

Signature	GD	GD	GD	Untested - near
Match	Maximum	Minimum	Filter	> Max POS
	Positive	Negative	Positive	& < Min NEG
9	10	None	10 x 1.5 = 15	None
8	10	26	(10 + 26) / 2 = 18	11 (1)
7	9	17	(9 + 17) / 2 = 13	13 (2), 15 (1)
6	8	11	(8 + 11) / 2 = 9	10 (1)
5	None	9	9 x 0.5 = 5	6 (1)

Note that the model already shows SIG=8, GD=11 is predicted to be positive. The model predicted all three SIG=7 untested testers as negative. However, one of the SIG=7, GD=13 has recently tested positive. The SIG=6, GD=10 is predicted as 3.1 % probability for positive. The prediction at this low of a signature size will always be very speculative in nature. We recently had a random person order the L226 SNP pack even though they only matched four of the nine markers in the L226 signature. As expected he did test negative for R-L226 but he also tested positive for all the L226 equivalents in the YSNP pack, creating a new branch, FGC5618, father of L226. With this additional information we are now beginning to see the progression of YSTR mutations just prior to and just after the L226 YSNP mutation.

The empirical model will have three possible adjustments to genetic distance. For signature matches where no negative testers are found, you should add 50 % to the genetic distance to the highest positive tester in case new testers with higher genetic distance are tested in the future. Where there is overlap between positive and negative testers, you should take the average of the highest positive tester and the lowest negative tester. For the signature value that first has no positive results, you should take 50 % of the lowest negative genetic distance. As more haplogroups are analyzed, these adjustments could change over time.

The above table can be charted to visually see how the Genetic Distance filter is derived and observe any particular anomalies. If you extend the curve Minimum Genetic Distance Negative to include signature of 9, it appears that the values could be above 30. Also, the curve changes direction for the Genetic Distance filter due to the 50 % higher approximation. So maybe using 100 % higher would be a better adjustment. But the increase of 50 % is probably more than acceptable as this increase is only added to catch a few positive testers at genetic distance of seven that could eventually reveal higher genetic distance than ten. Another observation is that the Maximum Genetic Distance Positive curve does not show any sign of losing as much in genetic distance as the 50 % lower estimate of the Genetic Distance filter (maybe 50 % should be reduced).



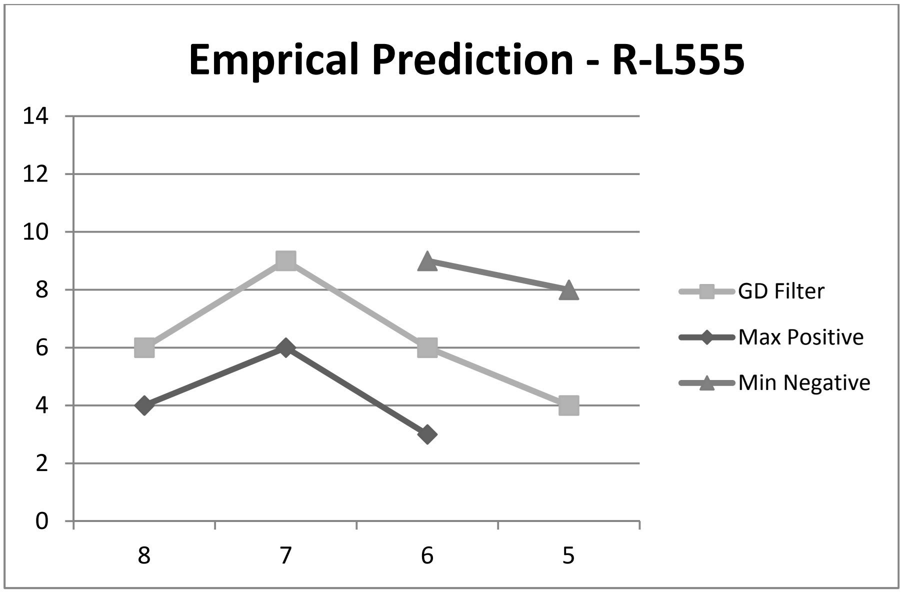
The characteristics of this chart will obviously vary between haplogroups, would be affected by the degree of convergence, the size of the signature and the age of the haplogroup. We could also create models for the maximum and minimum curves to estimate the missing values of these curves. I am less concerned about the genetic filters at higher signature matches as these are unlikely to happen. But these curves imply that you could find positive test results for much lower signature sizes and higher genetic distances that currently believed by the general public. But we are only talking about finding that last one percent of testers at the boundary conditions. Personally, I think this is very important to understanding the progression of YSTR mutations. For very low signature matches, the signature will actually become smaller and this will require manual analysis of the exceptions to determine the new signatures that would be used to find those most extreme boundary condition testers that rolled the genetic dice and got a lot of 2s and 3s and do have as many marker matches in their signatures.

For L555, the improved empirical YSNP prediction would be:

Signature Match	GD Maximum Positive	GD Minimum Negative	GD Filter Positive	Untested - near > Max POS & < Min NEG
8	4	None	4 x 1.5 = 6	4 (4)
7	6	None	6 x 1.5 = 9	7 (1)
6	3	9	(3 + 9) / 2 = 6	7 (1)
5	None	8	8 x 0.5 = 4	7 (5)

Note that for signature match at 7, there is one untested person that is one higher than the current maximum positive tester. This person is not filtered out, so this person should be tested. Since the person is not currently tested, the existing constants of the binary logistic regression model would probably change. However, this tester is already predicted positive by the model. Note that for signature match of 6, there is one untested person that is one higher than the GD filter. This would be a speculative test for L555. The current model shows only a 1.1 % of being positive. However, even though speculative in nature, either a positive or negative result would help the accuracy of the model and would help the accuracy of the GD filter. For signature match of 5, the untested persons are significantly higher than the genetic distance filter and the current model predicts 0.0 % chance of testing positive. However, even though this tester is unlikely to be L555 positive, it would again help with the accuracy of the model and the genetic distance filter as well.

At both signature matches at 7 and 8, there are no negative testers found across 54,000 haplogroup R testers. This suggests that the signature of L555 is very genetically isolated. Also, the maximum genetic distance of all positive testers for each signature is only six (excluding the CDY markers that are not included in the analysis). Also, the signature match only tests positive down to 75 % which is high for most haplogroups, all these factors imply a much younger haplogroup. However, even though this haplogroup is much younger than most, the signature of eight is a strong signature and no there is no significant convergence present. The two variable model predicts all positive testers at 100 % and all negative testers a 0 %, so the constants associated with the current model are extremely accurate. However, boundary condition testers are not fully tested and could later require a minor correction in the constants to maintain such high accuracy.



The empirical genetic distance filter chart for L555 is quite different from L226. This is to be expected since L226 is probably 50 % older than L555. Also, it is encouraging that the characteristics of these curves may themselves could be used to create a new model with a family of curves where the models would visually easy to understand and consumed by genetic genealogy community. The accuracy and completeness of these charts thrive on fully testing the boundary condition testers in order to produce the best chart.

However, there will be a few “outliers” that will eventually test positive and even a few predicted positive testers that will test negative. This will happen as the sample size will someday get so large that there is bound to find somebody who rolled the genetic dice and got 5s and 6s for fifty generations and does not match the signature near as well as the vast majority of others.

For one particular tester that has tested positive for R-L226, we found an “outlier” due to RecLOH event. The models and empirical genetic filter charts will have trouble with this tester since he has a massive replacement of the multi-copy markers 459, 464 and CDY. Unfortunately, five of the nine markers of the L226 signature are found in 459 and 464. For more information on RecLOH:

https://en.wikipedia.org/wiki/RecLOH

It is unlikely that these outliers can predicted via any YSNP prediction. So the outliers will need manual analysis and significant YSNP testing to make up for the less reliable YSTR values.

Many people are not aware that the haplotype of the ancestor who had the R-L226 mutation will not be the R-L226 signature. Around half of the mutations in the signature happen prior to the actual person who has the R-L226 YSNP mutation and the other half will be after this person. With the recent discovery of the R-FGC5618 tester with only four matching mutations, we are now getting a better feeling of the progression of YSTR mutations that happened in the time frame of R-L226 YSNP mutation. Also, R-L226 is 90 % Irish and the R-FGC5618 tester has only French ancestors. But it is very hard to make many conclusions with only one tester.

During the last two years, the knowledge level of the genetic community leadership has really increased to match the complexity coming from the explosion of genetic YDNA data. Also, during the last two years, I continued to concentrate more and more on charting methodologies which has revealed new information that should be incorporated in YSNP prediction. It has been observed via charting that the two CDY marker values just mutate too fast to be included in charting with any reliability. If the amount of YSTR data is increased by a factor of seven (release of the YSTR500), some additional markers may need to be filtered out. Just going from 111 to 500 YSTRs will create a real increase of the number of mutations that will require substantially more testers to resolve. This applies to YSNP prediction even more, so all future iterations of YSNP prediction will only use two variables models and will exclude CDY markers and potentially more markers as the numbers of markers increase.

For those that are just starting out in statistics, I highly recommend the very easy to read "Statistics II for Dummies", by Deborah Rumsey, PhD. Be sure to get volume two which only includes twelve pages on binary logistic regression. This book can be found in most used book stores and both books are excellent introductory refreshers to Statistics – especially the models and measurements used for binary logistic regression. It quickly introduces a lot of terms that is very consumable for those that already have some math background. It introduces p-values and several measurements of accuracy such as Chi Squared values for various accuracy measurements such as Pearson, Deviance and Hosmer-Lemeshow. It also explains how simple it is to calculate the accuracy of the model via concordant pairs. By reading both books, you will also realize that there are several statistical views for YDNA analysis – not just one fits all.

Also, there is more than just one methodology to predict YSNPs based on YSTR values. NevGen uses Bayesian-Allele-Frequency Approach in predicting which haplogroup that your Y STR haplotype belongs to. This predictor tool tracks the results for prediction from Binary Logistic Regression which proves that there are always multiple mathematical approaches that work:

http://www.nevgen.org/AboutNevGen.html

From my observations for R-L21 haplogroups, this predictor tool appears to be very accurate and is highly recommended since this tool is being updated with new haplogroups on a regular basis. But this tool has the same limitations of the age of the haplogroup supported which binary logistic regression also has. Both tools do a good job of analyzing haplogroups in the 1,200 to 2,500 year ranges but both tools will not accurately predict most branches below R-M222. However, binary logistic regression is now being extended from 2,500 years to the present with great success via a modified version of binary logistic regression. Charting allows down to two marker signatures which are around 60 % in accuracy all the way up to seven marker signatures that have 95 % or higher accuracy. Also, since not all signatures have been YSNP tested to date, 15 % of R-L226 will not be charted as accuracy of one marker signatures fall significantly below 50 % which I believe is not acceptable.

Another major advantage of binary logistic regression over Bayesian methodologies is that binary logistic regression can be ported to EXCEL spreadsheets which track the empirical approaches used by dozens of haplogroup admins who use signatures and genetic distance in their analysis. Not only can they ported binary logistic regression models into EXCEL spreadsheets which are widely used by the genetic genealogy community, empirical genetic distances models can be generated that can simulate binary logistic regression. These empirical models can easily be visually presented to prospective testers in a format that is much more understandable and just as accurate.

The Bayesian approach does have one major advantage as it is a variation of the methodology to predict much older haplogroups. It is based on the methodology of YSNP prediction created in Whit Athey’s YSNP prediction tool which only predicts much older haplgroups. So the Bayesian predictor tool can predict many older haplogroups that binary logistic regression is unable to predict. However, both approaches are unable to predict most of the middle tier haplogroups such as R-P312, R-L21, DF27, Z253, DF41 and hundreds of other haplogroups which just do not have enough genetic isolation to predict with any accuracy.

Summary of original signature only model selection

There is no doubt that binary logistic models are extremely accurate in predicting if YSNPs will test positive or negative. The accuracy of most YSNPs ranges between 90 to 99 % accuracy according the statistical accuracy measurement for the model accuracy – concordance of pairs. Concordance of pairs is very easy to understand as the measurement merely compares the observed values (input to curve fitting methodologies) to what the model predicts. Any prediction over 50 % that tests negative is counted as an error and any prediction under 50 % that tests positive is counted as an error. Accuracy is a simple measurement of all correctly predicted observations divided by all observations (both correctly and incorrectly predicted).

The Y axis shows the probability of testing positive which varies from 0 to 1 (or probability between 0 % and 100%). The X axis is the number of markers that match the YSTR signature of the YSNP. The X axis varies from 0 markers matching the signature to matching all markers in the signature. The curve starts at 0 % probability and remains at 0 % probability for at least half the signature and then rises as some submissions start to test positive in the transition area. After the transition area, it stays at or near 100 % for the last part of the curve. The curve of each YSNP varies a lot depending on the age of the YSNP. It is also affected by how genetically isolated the YSTR signature markers are when compared to other submissions that do not test positive for the YSNP.

At very low signature matches (to date, it is all submissions below 50 % matches), the prediction tool will predict negative results. For much younger YSNPs, the negative trend can go much higher – sometimes up to even 80 %. The next area along the X axis is referred to as the transitional area where a mixture of negative and positive results is expected (where positive results are predicted between 10 % and 90 %). This transitional area along the X axis can be only one change in the signature (most of time) or up to four changes in the signature. The remainder of the curve (where high match signatures exist) predicts that all remaining submissions will be predicted positive. Not all YSNPs behave exactly the same due to unique factors associated with each YSNP.

All YSNPs either test negative or positive which is the binary part of the name. Regression refers to the iterative process of determining the “a” and “b” constants of the formula to match the observed values that are input into the statistical program. The statistical software package uses curve fitting methodology to determine the constants required of the binary logistic regression formulas. Each YSNP equation uses the same formula with only the two constants (“a” and “b”) changing for each YSNP formula. The constant “a” affects the amount of shift of the entire curve along the X axis. The constant “b” affects the slope of the curve – whether it only takes one change in the value of X or several values of X for the curve to transition from 0 % to 100%.

Challenges for eliminating low accuracy p-values for the b-constant using only one variable

Unfortunately, there can be challenges with using form fitting methodologies associated with binary logistic regression. This YSNP prediction methodology has statistical accuracy issues due to the nature of the data. These challenges require several caveats to be imposed on this methodology, but the overall methodology is very sound for YSNP prediction. Below is a summary of some of the challenges that have been discovered to date:

The mathematics behind form fitting binary logistic regression models can not always predict the value of the “b” constant accurately when the observed data is exhibits “complete” separation of negative and positive results. This is also called “perfect” curves. Complete separation is where all negative observed results are found with lower values of X then all subsequent higher values of X are observed with positive results. Genetically isolated YSNPs result in lower statistical accuracy measurements for the “b” constant since these YSNPs show either complete separation or near complete separation.

However, I found that almost all statistical packages have a software defect and the underlying data associated with YSTR signatures is not the problem. This has been a known problem with statistical software programs for around 25 years. Most statistical software programs fail to properly calculate the constant “b” due the iterative algorithms being used. These packages continue their search for the best b value for a very long time but do not find an answer after thousands of iterations and just stop and return very poor values for p-value ratings of the “b” constant. Several higher end packages use add-on packages that add a tiny bit of noise to the iterative process and 20 or 30 iterations later produce b constant and very high p-values for the b constant. Unfortunately, the lowest cost package with this feature is around $1,500 with yearly fees vs. the one time purchase other packages that only cost $200 or $300.

As it turns out the value of the b constant could be a wide range of values and it would have almost no affect on the accuracy of the output. The only time that this constant matters at all is when there significant amount overlap of positives and negatives – which indicates that you really chosen a haplogroup that is either too old or young for analysis. The issue of Complete and Quasicomplete separation is covered in more depth in the book “Modern Regression Methods”, by Thomas P. Ryan, 1997.

The original concept for correcting this issue was first presented by David Firth in his article “Bias Reduction of Maximum Likelihood Estimates” in 1993. This was later followed by another article “Fixing the Nonconvergence Bug in Logistic Regression with SPLUS and SAS” by Georg Heinze and Meinhard Ploner, 2003. The article “A Solution to Separation in Binary Response Models,” by Christopher Zorn, 2005 is dedicated to this issue. The fix for this problem in this article was first called “penalized likelihood model.” This article further states Firth created a routine, brlr that is part of the R library. Heinze and Ploner created the routine logistf that is part of the R library as well as the FL macro in the statistical software package SAS. The extensive 125 page article “Comparing logistic regression methods for completely separated and quasi-separated data,” by Michael Botes, 2013 is very alarming that even the high end packages SAS and SPSS still fail to handle this common scenario as well as the more economical product Minitab.

However, these limitations are now no longer relevant since I have switched to a two variable model of binary logistic regression. But there are many people out there that like the simplicity of the one variable model and they will run into this issue if they ever attempt to use statistical software packages to validate their model. By just adding a little noise to iterative regression software solves the software defect and this defect has been known for over 25 years should have been fixed by software vendors.

Possible changes to correct this flaw in statistical packages (single variable model)

1) Changing the model is the approach suggested in most of these papers. This would probably be the best workaround for the separation issue. By introducing a second parameter for genetic distance which is needed to filter out the convergence of signatures would not only remove the need for requirement of testing for R-L21 but has solved the p-value issue for the b constant for the perfect data issue as well. With the introduction of genetic distance into the binary logistic regression model the "perfect" curve or "separated data" issue has been resolved. Unfortunately, the visualization of the one variable model was very easy for the genetic genealogy community to understand.

2) Changing the data is another approach suggested in most of these papers. The mathematics used by binary logistic regression is less likely to find overlapping negative and positive results along the X axis when the X axis contains discrete variables (integers) vs. continuous variables (with decimal points). Aggravating this issue is the YSTR signatures do not produce enough points along the X axis that binary statistical regression requires for high accuracy of "b" constant. To offset this statistical issue, weighting of YSTR markers that comprise the YSTR signature could be slightly modified with YSTR mutation rates and would reduce the negative impact discrete variables. However, I really think that mutation rates play a very minor role in signatures and this approach would be really just a workaround for the statistical software packages failing to deal with such well known issue. Again, moving to two variable model solves this issue as well.

3) Improving randomness in testing always improves the accuracy of statistics and this is also true for YSNP prediction. The testing of YSNP submissions are not selected at random as required for statistical accuracy. Most sponsors of YSNP testing either test very high signature matches or randomly test via SNP packs (which are usually extremely low signature matches). This creates a bias of under testing the critical submissions in the transitional area. The transitional area of signature matches is very important to the statistical accuracy measurement of the "b" constant. Aggravating this issue is that the transitional signature matches are the not plentiful of testers as well (matches generally between 50 % and 75 % of the YSNP signature for most YSNPs). This bias in lack of testing transitional values of X produces fewer overlapping negative and positive results across the X axis which reduces statistical accuracy measurement of the "b" constant. It only takes one or two overlapping results before the statistical software packages will report good p-values for the b constant. Testing candidates are generally unwilling to take the chance to take speculative tests but most haplogroup administrators are also not aware of the importance testing in the transitional area. It is highly unlikely that we can change the behavior of YDNA testing. Again, this issue is almost eliminated with the two variable model.

Increasing the sample size only has a minimal affect on statistical accuracy

Sample size has only very indirect affect on the accuracy of YSNP prediction. For well known and prolific YSNPs, (like R-M222 and R-L226), there are multiple layers of YSNP packs that include these broad YSNP branches. There are hundreds of negative observations as well as many positive submissions. The number negative observations will always far exceed positive observations due the economics of YDNA testing and the technology that encourage massive testing once vs. numerous smaller tests. Testing of well known YSNPs and prolific branches outnumber less known branches 100 to 1 ratio for the more and even up to a 10,000 to 1 ratio than the more narrow scope YSNPs. Increasing the sample size below 50 % of the signature match has extremely small impact on the accuracy. Testing high signature matches that will always test positive also has minimal impact on the accuracy of YSNP prediction. Only extensive testing transitional area of the S-Curve impacts the statistical accuracy measurements.

The primary influence of total sample size is that boundary condition testing submissions will sometimes track the number of total submissions tested. For YSNPs that have been included in broad tests for some time, the increased sample size will have more impact as it is more likely that some negative submissions with high signature matches could be revealed. The amount of testing of boundary condition testing submissions seems to vary dramatically from YSNP to YSNP. Few haplogroup administrators understand the value of testing these boundary condition testing candidates while other YSNPs sponsors only test specific surnames or test candidates with only very high probability of testing positive. There is bias of under-testing of boundary condition submissions (those in the transitional area) which varies dramatically from YSNP to YSNP. Due to this bias, merely increasing sample size has minimal impact on accuracy.

Haplogroup administrators need strongly advocate testing in the transitional area primarily to understand the early evolution of YSTR in the highest part of their haplotree. This focused testing would reduce bias since these individuals rarely test since they do not match the haplogroup signature very well and they have higher genetic distances as well. By encouraging these transitional testing candidates to test, they reduce the natural bias of these individuals not to test. On the other hand you are asking testing candidates contribute to our collective knowledge of the haplogroup YSTR evolution – this is probably not high on their priority list and they probably do not even understand the benefits of this testing. Most haplogroup administrators may have sponsor or partially sponsor these boundary condition testers that see little genealogical relevance of this kind of YSNP testing.

Random testing has almost no impact on statistical accuracy

Random testing within R-L21 has very minimal impact on YSNP prediction accuracy for its numerous branches under R-L21. Random testing tends to test low signature matches (below 50 %) which are always negative. This is due to the fact that between 95 to 99 % of R-L21 submissions will test negative for single signature YSNPs. Any testing of high signature matches (generally above 75 %) results usually in 100 % positive results and have only a minor impact on the statistical accuracy. Both low signature matches and high signature matches do not have a major impact on the accuracy of YSNP prediction. Only testing of the transitional values along the X axis (generally between 50 and 75 % signature matches with lower genetic distances) has any impact on statistical accuracy.

Random testing would rarely help due to the nature of technology currently being used. YSNP testing is currently from three primary sources: 1) NGS testing (primarily Big Y). These generate 10,000s of negative results for YSNPs in areas that are not even closely related. This data is so extensive that FTDNA and most administrators do not either bother to track YSNP mutations that are distantly related; 2) YSNP pack testing is rarely ordered by people who are not are predicted at 99 % or higher accuracy. Also, these tests are just too limited in numbers of YSNPs as they only include only around 150 YSNPs and do not extensively test for broader and older haplogroups; 3) individual YSNP testing at YSEQ is almost always limited to testing YSNPs that are known to associated with the haplogroup under investigation.

However, the National Geographic test did test 10,000s of YSNPs at a very economical cost. Unfortunately, this test and other similar tests have not been properly upgraded to keep up with the explosion of YSNP branches being discovered. These tests could be upgraded to include 10,000 to 100,000 YSNPs which could be targeted to different parts of the haplotree of mankind. This test would have a significant impact as it would be randomly tested by many testers in transitional areas. So if this kind of test is updated in the near future, you would have a test where random testing could help significantly.

Variations in characteristics between predictable YSNPs

It has been determined that amount of convergence is directly dependent on how genetically isolated the submissions of any YSNP is from submissions that test negative for that YSNP. The degree of genetic isolation produces two classes of YSNP prediction: 1) genetically isolated YSNPs and 2) YSNPs that are less isolated which have significant overlap with submissions that do not test positive for the YSNP. The size of signature greatly improves the probability of genetic isolation. For those with significant convergence, prediction at 111 markers may be alternative.

The actual transitional area for YSNPs does vary from YSNP to YSNP. This is primarily affected by the age of the haplogroup being analyzed. Younger haplogroups tend to require much higher percentages of signature matches to test positive while older haplogroups generally have positive results at much lower percentages of signature matches.

The numbers of testing candidates vary significantly in the transitional area from haplogroup to haplogroup. This is due to genetic isolation but can be quantified to determine the degree of genetic isolation being observed. The slope of the curves for testing candidates is another factor that can be quantified and analyzed. This type of analysis could be important to the automation of the recognition of haplogroup signatures. This would make YSNP prediction much less tedious to maintain with the fast growing number of predictable haplogroups being discovered every week.

Several examples of models using binary logistic regression

Here are examples of several curves using the classic S-Curve produced by binary logistic regression (using only signature as the only dependent variable):

http://www.rcasey.net/DNA/R_L21/stats/M222_Prob_Curve_20120410A.xls

Here is a link to my R-L21 YSNP predictor tool based only signatures (note only around 50 signatures were analyzed and this tool has not been updated in several years). This tool is currently based on signatures as the only variable:

http://www.rcasey.net/DNA/R-L21_SNP_Predictor_Intro.html

Here is a link to an EXCEL spreadsheet that has the two variable model used for R-L226 using the package SPSS. The macros in this spreadsheet can be altered to be used for any haplogroup. Column CB includes the binary logistic regression model based on Signature (column K) and Genetic Distance (column L). The L226 spreadsheet is updated weekly with several new YDNA tests during the previous week:

http://www.rcasey.net/DNA/R_L226/Haplotrees/L226_Signatures.xlsx

With the assistance of James Irvine, the two variable model was ported to R-L555 with the constants being updated with SPSS (based on January, 2018 data):

http://www.rcasey.net/DNA/Temp/L555_CIAStudydata2018-1-23abb.xlsx

The two variable model was also ported to R-L371 with the constants being updated with SPSS (based on May, 2017 data):

http://www.rcasey.net/DNA/Temp/HG_R_Master_L371_20170630D.xlsx

With only a two week trial of SPSS, I was only able to analyze three haplogroups under R-L21. Unfortunately, most of the time was spent with preparing data for input into SPSS plus getting up to speed with the SPSS product as well. However, with conversion to the two variable model, the perfect curve is no longer an issue and most entry statistical packages can now be used.

Best ways to improve the accuracy of YSNP prediction (for two variable model):

1) Minimize the testing bias of not YSNP testing of good testing candidates in the transitional area. Reducing this bias should reveal the maximum number of overlapping test results. Reducing this bias of under-testing in the transitional area has much more of an affect than increasing sample size. Reducing the bias of under-tested boundary condition submissions would result in most significant impact on improving statistical accuracy any model for YSNP prediction. This will only happen with proactive focus by haplogroup administrators.

2) Increasing the sample size of YSTR testers will always increase the chances of discovering new overlapping submissions that will improve the statistic accuracy of YSNP prediction. However, we are talking about factors (2X to 4X) of increase vs. a 20 to 40 % increase in testers. Smaller increases only have minor impact on accuracy. Also, increasing the amount of YSNP testing helps as well. Again, unless this increase in sample size increases exposure to more boundary condition testers, minimal change in accuracy should be expected.

3) Investigate using 111 marker signatures in the future for haplogroups that have significant convergence. However, this also greatly reduces the sample size of test submissions which probably has a negative effect. Unless convergence is present, developing models at 111 markers will not seem to improve the accuracy of the 67 model since the 67 marker prediction is already at 100 % accuracy. A larger signature could catch a few more boundary condition testers which may be worth the effort in identifying and testing the best boundary condition testing candidates.

4) For many haplogroups that have large percentages of non-English speaking countries, reducing the geographical bias of testing will remain a challenge. There are several geographies that are rapidly improving but the economics of testing and the laws in several countries are a major impediment to reducing geographical testing bias. Testing is dominated by American testers and other former British colonies. The current laws in France are a major impediment and the French genetic genealogy community should lobby for more liberal policies.

5) The fear of privacy issues is another major impediment as this radically reduces the sample size as many surname projects are closed to the public view and are not well analyzed by haplogroup administrators as they have no access to this information. Many projects do not display either the YSNP, YSTR or both reports. Also, FTDNA has over-reacted to privacy by changing the default privacy setting of all new testers to private (Project only). Many individual testers also have the right to make their research private which is a form of “lurking” as they can see public reports but their information is not displayed in any FTDNA reports. You can help with this privacy issue by lobbying for more access to projects and for newer testers to update their privacy setting to “public” (Anyone). Also, not all testers even join projects and not every small project is being looked at as well. It is believed these issues could double the sample size.

CONCLUSIONS (two variable model)

The accuracy of two variable models are so high that there are only minor issues that remain and will almost always keep accuracy above 99 % when little convergence is present. Untested data that is in the transitional area will eventually result in less than 100 % accuracy but the vast majority of time running another regression model with new data usually changes the constants just enough to get accuracy back to 100 %.

The transitional area (lower signature matches with slightly higher genetic distance) remain a challenge for getting tested and analyzed. These testing candidates are not only marginal signature matches but are also marginal genetic distance matches. They have few or no matches and really do not see the advantage of testing to better understand the origins of their haplotree. It is really up to haplogroup administrators to determine these boundary condition testers and encourage or sponsor their YSNP testing.

Increasing the resolution of the signatures to 111 markers does not seem to be a viable solution of increasing the accuracy in the near future as the reduction of sample sizes currently greatly offsets the benefits of the modest in increase in the size of the signatures. However, for charting these 111 marker upgrades will reveal many new YSTR branches but not very much unless the ratio of 111 to 67 markers is very high. But we also need to be preparing for YSTR500 with any experience that we can gain from 111 markers. YFULL is already providing YSTR500 in raw data format and FTDNA recently announced its intention to release YSTR500 for their Big Y testers in the near future as well.

However, the number of fast mutating markers would surely greatly increase as well – putting more pressure to remove the faster mutation rate markers from signatures and charting. It has been determined that the faster mutating markers, CDYa and CDYb are just too volatile for YSNP prediction or charting and the 68 to 111 markers, 710 and 715 markers are very fast mutating markers as well. The mutation rates published in existing documentation vary radically between articles and are missing the vast majority of YSTR500 markers. It is very critical that FTDNA releases the criteria for inclusion of the YSTRs in YSTR500 and that will hopefully include information on mutation rates of the markers that get included.

Lastly, YSNP prediction is still a very useful tool for haplogroup administrators but there is a shift to charting which has a larger impact on the genealogical community. Initial analysis of several R-L21 haplogroups show that charting has a very bright future. The same binary logistic regression methodology is also employed in charting as well. The methodology for charting is much more complex than YSNP prediction and will be covered in a future paper that is dedicated to charting with signatures.

Extending YSNP prediction down to younger YSNPs via charting

During the last two years, Next Generation Sequencing (NGS) has greatly expanded the number of branches detected below predictable YSNPs. Also, the introduction of comprehensive YSNP pack testing for predictable YSNP branches has also greatly increased the amount of YSTR and YSNP data below predictable YSNP branches. Individual testing of private YSNPs at YSEQ has also provided much needed information than more costly NGS testing options. With the inclusion of private YSNPs in SNP packs and testing of individual YSNPs at YSEQ, many additional branches have been discovered without NGS testing.

The same prediction methodology can be extended to prediction of testers being assigned to genetic clusters below the predictable YSNP. There are a small percentage of signatures under predictable YSNPs that have very large signatures which yield accuracy levels just as high of prediction of its ancestral YSNPs. However, the vast majority of signatures below any predictable YSNP are much smaller than the seven to twelve markers found in most predictable YSNPs. The accuracy of prediction can quickly go down to 60 % for this level of prediction.

Even with accuracy in the 60 % to 95 % range, charting is extremely valuable in reducing testing costs based on reasonably accurate prediction. With reasonably accurate charting, it becomes very obvious when NGS testing is best option or when individual testing of private YSNPs is a best option. The more economical SNP packs provide almost as much information as NGS testers since they can test 80 to 90 % of the known branches under any predictable YSNP branch. However, SNP packs will never discover new private YSNP which will eventually revealed as future branches. But from a charting viewpoint, the SNP pack tests are incredibly useful for charting below the predictable YSNP haplogroups.

Unfortunately, most current charting tools are based on network joining software methodology which is not the correct math to chart YDNA data. The network joining methodology ignores the mathematical issue that parallel mutations must greatly outnumber backwards mutations by a factor of ten to forty times for faster mutating markers. The same also holds true for the ratio of parallel mutations to multiple mutations of the same marker along one path. The network joining methodology produce charts based on backwards and parallel mutations being equal probability which result in substantial numbers of backwards mutations that probability theory does not allow. Recently, bias has been added to these charting tools but the networking joining methodology still has more ground to cover to emulate what signature methodology provides. Also, most network joining methodology usually will assign every tester to some part of the haplotree even when accuracy is way falls to ten to twenty percent.

Charting of over 600 67 marker submissions under R-L226 has revealed that parallel mutations are extremely common for the faster mutating markers. There are as many as 20 to 40 parallel mutations of these more faster mutating marker values. Signature methodology stops at 60 % accuracy and currently leaves the last 15 % of L226 testers uncharted. However, network joining methodology uses one marker signatures of faster mutating markers for placement on the haplotree charts with low predicted accuracy between 10 and 20 %.