Binary Logistic Regression

The mathematical model for YSNP prediction based on YSTR signatures

By Robert Casey

Revised Version, March 7, 2017

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

A binary logistic regression model has been used to create the 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 L21 testers to around fifty major signatures under L21. With the explosion of NGS testing, keeping up this tool would have become a full time job and this tool has not been updated very much since Big Y testing started rolling around 15,000 tests.

It is believed that 90 % of L21 could now be predicted to YSNP branches in the relevant time frame. For the older YSNPs such L513, Z253, Z255, DF41, 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 1,500 to 2,500 year range. These single signature YSNPs include L226, M222, L193, L555, L371, etc. which can be accurately predicted.

During the last two years, I have shifted my analysis to mainly R-L226 where the three L226 admins have made tremendous progress in YSNP testing to complement YSTR tests. L226 now has sixty NGS tests which have revealed 28 new branches under L226 and around 400 private and equivalent YSNPs associated with L226. Individual testing of private YSNPs at YSEQ has revealed another seven branches under L226. During the last few months, a new L226 only SNP pack was rolled out which included around 40 known branches, 50 private YSNPs and 50 equivalent YSNPs under L226. This L226 SNP pack testing revealed another nine branches.

The same YSNP prediction methodology is being employed in charting of 500 67 marker L226 testers. Binary logistic regression is also being used for charting of L226. With 60 NGS tests, 60 L226 SNP pack tests, 20 Z253 SNP pack tests (when they included L226 branches) and 100 individual YSNP tests, we can now chart almost 80 % of L226 even though only 20 % of L226 has been extensively YSNP tested (tested for at least 80 % of the relevant branches). In the future, a second article will be dedicated to how binary logistic regression is used in charting below predictable YSNPs.

Statistical Review of YSNP prediction under R-L21

This article reviews the mathematics supporting the prediction of testing positive for YSNPs based on using YSTR signatures. YSNP prediction is based how well submissions match the YSTR signature associated with each YSNP. 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 after more than a couple of thousand 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 YSNPs that formed between 1,500 and 2,500 years and other methodologies must be used for other time frames. Charting with YSTRs and YSNPs involves another variation of YSNP prediction which is also based on signatures. This article primarily focuses on YSNP prediction.

Currently, the only submissions that have tested positive for R-L21 or any YSNP descendant of R-L21 are included in this analysis. 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” YSTR signatures between very old haplogroups in small numbers. This is where YSTR patterns (signatures) can overlap each other even when the common ancestor is several thousand years old. This convergence is believed to be between five and ten percent of R-L21 testers. However, this is not a random five or ten percent but found concentrated much more in certain branches under R-L21 where convergence is significant. For these YSNP branches, accuracy would be greatly degraded. Other branches have almost no convergence. Another reason for this 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 and the restriction of testing positive for R-L21 can probably be removed if the model used is expanded to include genetic distance as a second variable. It has been observed (on a very limited scope today) that convergence of signatures across haplogroup R also comes with extremely 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.

All YSNP signatures are currently based on 67 YSTR markers, so current YSNP prediction requires 67 or more YSTR markers tested. YSNP accuracy would only have very modest increase in accuracy at 111 markers since most prediction is already at 95 to 99 % accuracy. However, charting could greatly benefit from 111 marker upgrades. YSNP prediction using 37 markers submissions do not produce large enough YSTR signatures for accurate YSNP prediction across even 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 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 present a challenging obstacle). Prediction at 37 markers would cover on ten or fifteen percent of the genome with high accuracy. Prediction of YSNPs at 111 would only have a very modest increase in an already high accuracy rates but could expand coverage of the genome by several percent.

With the explosion of NGS testing that came in with wide usage of FTDNA's Big Y test, the number of predictable YSNPs increased to such a rate that the manual spreadsheet analysis in detecting new signatures became too laborious to even keep up with 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 the appropriate age of YSNPs for prediction. Yet another challenge in automating YSNP prediction is being able filter out 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 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 a five or ten years, Whole Genome Sequence tests will fall to $200 to $300 range and will the only YDNA test required which will greatly reduce the need for YSNP prediction. However, in the mean time, 10,000s of intermediate SNP packs will be an economic burden on the genetic genealogy community for a significant time frame. For very narrow scope YSNP branches, YSNP prediction would remain viable for testers belonging to these smaller scope branches.

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. Around ten R-L21 YSNPs are too old since the number of parallel and backwards are hidden from testing living individuals. 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 technology was pushed to see how well prediction would work for older YSNPs. In future releases, these will be replaced with several more recent YSNPs that cover most of these older YSNPs which will increase the accuracy due having a more recent date and less time for YSTRs to produce excessive amounts of parallel and backward mutations.

The purpose of this version of this paper is to validate the original empirical approach, analyze the mathematics behind curve fitting methodology, to increase the accuracy of the YSNP prediction tool and to introduce more automation in the 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 many binary logistic regression models examined were found to have very high accuracy. However, YSNP prediction that most closely maps the classic S-Curve of the classic binary logistic regression formula:

$$y=\exp(a+b*x)/(1+\exp(a+b*x)).$$

During the last two years, the requirement of testing R-L21 was found to be unnecessary if simple filters were added for high genetic distance. This new discovery needs to be analyzed across multiple haplogroups to validate this new empirical observation. This could easily implemented by introducing a second variable on the right side of the equation or just introducing a quick and dirty filter that removes excessively high genetic distance values. This is when math conflicts with practical needs. For most signature research, EXCEL spreadsheets are used extensively for YSNP prediction with signature match values and genetic distance values being used. If more complex models are used, an equivalent EXCEL counterpart would no longer track each other.

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 empirically 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 ten, CDY markers could probably be added back. This applies to YSNP prediction even more, so the next iteration of YSNP prediction will exclude CDY markers.

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 models such Pearson, Deviance and Hosmer-Lemeshow. It also explains how simple it is to calculare the accuracy of the model via concordant pairs. By reading both books, you will also realize that there are several statistical models for YDNA analysis – not just one fits all.

Summary of original 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 submissions match the YSTR signature of the YSNP. The X axis varies from 0 markers matching the signature to all markers in the signature. The curve starts at 0 % probability and remains at 0 % for at least half the signature and then rises as some submissions start to match in the transition area and eventually stays 100 % for the last part of the curve. The curve of each YSNP varies a lot depending on the breadth and age of the YSNP. It is also affected by how genetically isolated the submissions are associated with the YSNP are as 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 more narrow breadth 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 to 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” represents the amount of shift of the entire curve along the X axis. The constant “b” represents 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

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:

1) 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. 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 the statistical package that I was using was the actual problem – not the underlying data associated with YSTR signatures. This has been a known problem with statistical software programs for many 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 packages use a new variation that adds 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 of $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 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.

2) Changing the model is another 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 or other haplogroups but would probably solve the p-value issue for the b constant for the separation issue as well.

3) 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 modified with YSTR mutation rates and would reduce the negative impact discrete variables. However, I really do 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. Eventually, upgrading to 111 or more markers could provide larger signatures as well but I believe the separation would still be present in most 111 marker signatures as well.

4) Improving randomness in testing always improves the accuracy of statistics and this is especially 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 (usually are 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 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.

It is very unlikely that this approach would work. Not only are testing candidates unwilling to take the chance to test, most admins 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. This would be making recommendations to fix a bug in statistical software packages. This is a very poor justification for testing just to make math purist happy. However, testing of the transitional is critical for the advancement of YDNA knowledge of charting of the upper portions of our haplotrees as well determining the haplotype of our ancestor vs. the approximate haplotype via signature of our ancestor. Signatures are statistical averages that work great for YSNP prediction but the haplotype of our ancestors is just not the same as the signature of our haplogroup. For most haplogroup signatures, half of the signature markers mutate just before the YSNP mutation and the other half mutate just after the YSNP mutation. Any YSTR mutation would rarely mutate with the same person that had the YSNP mutation.

Statistical packages (and models) are too lenient for p-values of the a constant

The form fitting curve methodology is too lenient with respect to missing data along the X axis in the transitional area of the S-Curve. If two values are missing along the X axis, form fitting methodology is perfectly happy in assuming one is negative and the other is positive. In reality, both could be negative or both could be positive. The constant “a” in the binary logistic regression models could change dramatically as missing data points on the X axis are later discovered. This observation is despite the fact that the statistical accuracy measurement of the “a” constant always reports high accuracy.

Unlike the p-values for the “b” constant, software packages and the models themselves do not seem mind at all that these values will change radically with additional test results in the transitional area. The p-values of the “a” constant will dramatically change with additional data but p-values for the a constant always get very high accuracy ratings. However, since there is so few potential testers in the transitional area, total accuracy is not affected by these additional test results. So if the slope of the curve changes a lot, yet it has only a very minor affect on the overall accuracy of the prediction of the model (concordance of pairs).

Perfect and Near-Perfect Models

The vast majority of YSNPs have a “perfect” match to the model. This scenario is where all lower signatures matches along the X axis are initially “0” and then with only one change in the X value, all remaining observed submissions are then “1”. This is known as a “perfect” model where form fitting curve methodologies can not predict the value of the constant “b” via the iterative form fitting methodology. In this case, a wide range of values of the constant “b” could be used without any affect on the accuracy of prediction. For the “perfect” match scenario, the p-value of the “b” constant is really not that relevant.

It is amazing how much time is wasted by math world on aggressively trying to introduce more complex models, substantial more data or other changes get the models fed less perfect data so the models and all their associated accuracy parameters report high values. If the model reports 100 % accuracy, why is this community so concerned about reducing accuracy to 99.9 % so that all the statistical other accuracy measurements report high accuracy as well. But 10 or 20 % of the YSNPs do have some overlap, so we really have to use statistical models for prediction of these YSNPs.

For “near perfect” models, form fitting methodology does not predict the value of “b” constant very well when there is little overlap between negative and positive results along the X axis. This is probably the most serious issue associated with YSNP prediction with form fitting curve methodologies. It usually only takes one overlapping observed values along the X axis and sometimes it takes two or three overlapping values before statistical accuracy is achieved for the “b” constant. Once there are three or four overlapping values, the statistical accuracy measurement for the “b” constant is always extremely high.

It also seems that complete separation and quasicomplete separation are about the same a perfect and near perfect models – at least for YSNP prediction. However, since both of these issues are discussed in depth, I feel obligated to include both issues.

Sample size has only minimal impact statistical accuracy

Sample size has only very indirect affect on the accuracy of YSNP prediction. For well known and prolific YSNPs, (like M222 and 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 1,000 to 1 ratio than the more narrow breadth 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. A few admins 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 affect unless admins encouraging testing in the transitional area submissions.

Admins 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 signature that well and have very high genetic distances as well. By encouraging these transitional testing candidates to test, they reduce the natural bias of these individuals not to test. By reducing the bias, more transitional area testing candidates would also reduce the issue of complete separation and quasicomplete separation. This testing not is only critical for determining the YSTR evolution of the haplogroup but also reduces testing bias which in turn makes the statistical models increase their accuracy of the p-values of the “b” constant. On the other hand you are asking testing candidates contribute our collective knowledge of the haplogroup YSTR evolution – which is probably not high on their priority list and they probably do not even understand the benefits of their testing. Most admins are not even aware of the critical importance of this kind of 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 90 to 99 % of R-L21 submissions will test negative for single signature YSNPs. Any testing of high signature matches (generally above 75 % matches) results in positive results that approach 100 % and also 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 % matches) has an impact on statistical accuracy of the “b” constant measurement. True random testing would ensure more transaction submissions would be tested but is greatly dampened by the overwhelming number of negative submissions and high signature match submissions that always test positive.

The net result is that is that only the submissions located in the transitional area of the curve have an impact on the statistical accuracy measurement of the “b” constant. Since there number of submissions found in the transitional area is routinely extremely small compared to all submissions tested under R-L21, random testing has an extremely small chance adding submissions in the transitional area along the X axis of the S-Curve. Submissions in the transitional area would account for only 0.1 % to 1.0 % of the total R-L21 testing candidates. If true random testing was enforced, the sample size required adequately test submissions in the transitional area would require between one hundred to one thousand tests.

For many YSNPs, most submissions that test positive for YSNP are somewhat isolated from other submissions that do not test positive. This isolation results in a very small number of testing candidates that are found in the transitional area along the X axis (generally between 50 and 75 % of the signature match). The number of testing submissions in the transitional area is usually much smaller than the number of testing candidates that have high probabilities of testing positive. Submissions are not tested randomly enough between 50 % and 75 % matches of the YSNP signature which creates a bias in testing due to the lack of true random testing. Since testing negative submissions have extremely small impact on statistical accuracy after only ten or twenty tests (under 50 % matches), these should really not be considered viable testing candidates and only submissions that match over 50 % should be considered viable testing candidates.

Impact of statistical accuracy measurement of the “b” constant is not very significant

All statistical accuracy measurements show extremely high values for the accuracy of the model itself. The primary statistical accuracy measurement that fails to show high values are the p-values for the constant “b”. This is a known problem for binary logistic regression when the empirical data exhibits “near complete” separation of observed results and includes discrete variables (integers) which provide few values along the X axis. This impacts the accuracy of the prediction in the transitional area along the X axis (generally between 50 and 75 % matches of the YSNP signature).

The accuracy for low matches (below 50 % generally) are extremely accurate (100 %) and high matches (above 75 % generally) are extremely high (approaching 100 %). However, the statistical accuracy in the transitional area (between 50 and 75 % matches generally) cannot be predicted with high accuracy for many YSNPs. Since the number of submissions in the transitional area along the X axis is usually a very small in number, the statistical accuracy of the model still remains very high. The model produces highly accurate prediction for low signature matches and high signature matches but will many times fail to accurately predict probabilities in the transitional area of the signature matches with high accuracy. However, since so few submissions are found in the transitional area, the accuracy of the model remains very high.

The net result is that the model accurately predicts vast majority of negative results and the vast majority of positive results as well. Only in the transitional area of the X axis will prediction be less accurate. This only means that the probabilities predicted in the transitional area of the X axis should be taken with caution. For many YSNPs, this means that the YSNP predictor tool should really report “negative” and “positive” with high accuracy – but should report “maybe” within the transitional area of X axis. In this transitional area, the actual probability reported may be less accurate – but it will still be the range between 10 % and 90 % probability. A prediction of 50 % may actually be 30 % or 70 % due the low values of the p-value of the “b” constant.

This issue is somewhat correcting as well. For YSNPs that are very isolated and have little opportunity for overlapping values, the number of possible testing candidates in the transitional area does not exist or is very small. For YSNPs where overlap exists, the number of overlapping submissions will be higher which results in very high statistically accurate measurements for the p-values of the “b” constant. When high accuracy is reported for the p-value of the “b” constant, the model will be statistically accurate in all respects – including the transitional area of the X axis. If the p-value of the “b” constant reports low accuracy, this means that only the transitional area of the X axis is not as reliable for YSNP prediction. Fortunately, there are normally very few testing candidates in this transitional area which results in very high overall accuracy of YSNP prediction.

Variations in characteristics between predictable YSNPs

Some YSNPs have curves that are more the typical S-Curve and have much more gentle curves while other YSNPs have very steep slopes. Only those YSNPs that have more gradual slopes resulted in high accuracy for all statistical accuracy measurements. From a pure mathematical point of view, the steepness of the curves is directly affected by the number of overlapping negative and positive submissions along the X axis. Not only is the steepness of the curves affected but most statistical accuracy measurements are also affected by overlapping of negative and positive observed submissions. The amount of overlap is also affected by characteristics of the YSNP which creates two classes of YSNPs.

It has been determined that number of overlapping submissions 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 data: 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 degree of genetic isolation of any YSNP can be observed by examination of all testing candidates (tested and untested) of any YSNP from 75 % to 50 % of the YSNP signature. The actual transitional range for most YSNPs does vary from YSNP to YSNP. Almost every YSNP has fewer and fewer testing candidates as the signature match decreases from a 100 % match to a 75 % match. However, the number of testing candidates between 75 % to 50 % matches has two observed characteristics as the signature match decreases.

For genetically isolated YSNPs, the number of testing candidates continues to decline as the signature match declines from 75 % to 50 %. Not only do the number of testing candidates continue decline but the number of testing candidates is usually extremely small in number when compared to less genetically isolated YSNPs. From 75 % to 50 % matches, there could be a very small increase in testing candidates but these YSNPs should still be a considered genetically isolated. The more genetically isolated the YSNP, the likelihood of overlapping negative and positive results is greatly reduced which in turn decreases the statistical accuracy measurements.

Examples of genetic isolation (M222)

The broadest single signature YSNP under R-L21, M222, is known to be genetically isolated from other non-M222 submissions. The original sample size of around 100 M222 submissions did not reveal any overlapping negative and positive observations along the X axis. This sample showed the classic “perfect” model scenario where the p-value of the “b” constant was declared statistically inaccurate – even though any large range of “b” constants would produce the same model fit of 100 %. All measurements of model accuracy were perfect (Chi squared was 1.000 and concordance of values was 100 %).

M222 also had extremely few testing candidates tested in the transitional area of the X axis as well as extremely few testing candidates (tested and untested) known in the transitional area. The YSNP M222 has been “deep clade” tested for many years, so there are hundreds of negative observed values. “Deep clade” testing was later replaced by the L21 SNP pack which continued to provide many negative test results. Increasing the sample size to over 200 samples revealed one submission that tested negative with a fairly high signature match. Just adding this one additional negative submission with a fairly high signature match had a dramatic impact on the model and statistical accuracy measurements. The p-value of the “b” constant went to 0.000 (perfect) but the accuracy of the model slipped slightly (chi squared was 0.895 and concordance of values was 99.9 %). All relevant statistical accuracy measurements became very accurate with only the addition of one overlapping submission that was a fairly high signature match. Unfortunately, this person was convinced that the results was a lab error and that suspicion was later confirmed and M222 returned to a perfect curve.

A minority of the YSNPs are the second class of YSNP which are less genetically isolated. As you match the YSNP signature less and less, many more negative submissions overlap with positive submissions since the uniqueness of the signature is not isolated enough and overlaps with other submissions. For most YSNPs that are less genetically isolated, all statistical accuracy measurements were very high in accuracy due to greatly increased overlap of negative and positive submissions.

Here are examples of curves using the classic S-Curve produced by binary logistic regression:

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

The data for these curves can be extracted using my R-L21 YSNP predictor:

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

Best ways to improve the accuracy of YSNP prediction:

1) Minimize the testing bias of not testing signature in the transitional area of the S-Curve. 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 of p-value of the “b” constant.

2) Conversion of discrete variables along the X axis into numbers that exhibit more continuous characteristics. Since the mutation rate of each marker found in the YSNP signature are legitimate factors affecting the degree of matching, the mutation rate of markers found in the YSNP signature could be weighted to produce more continuous numbers along the X axis. Also, various multi-step mutations and backwards mutations from the signature could be used as well.

3) Increasing the sample size increases the chances of discovering new overlapping submissions that will improve the statistic accuracy of YSNP prediction if very large numbers of submissions are tested. This only applies if increased sample sizes are truly random in nature. For YSNP pack testing and NGS testing, all negative results need to be exhaustively reviewed to discover more negative submissions that could overlap with positive submissions.

4) Investigate using 111 marker signatures in the future. However, this would only marginally increase the number of discrete values along the X axis which would increase the possibility finding more overlapping test results along the X axis. However, this also greatly reduces the sample size of test submissions which probably has more of a negative effect than the positive effect of increase points along the X axis. Introducing a hybrid approach that includes both 67 markers and 111 markers is another valid option but inconsistent resolution would greatly complicate the analysis.

Complete and near-complete (quasi-complete) separation

Separation of negative and positive submissions along the X axis causes problems for the mathematical models used in binary logistic regression. The accuracy of YSNP prediction is most influenced by this factor than any other factor. The book, “Modern Regression Models,” by Thomas P. Ryan, 1997 has an excellent chapter on binary logistic regression and describes this issue extremely well. The mathematics behind the curve fitting mathematics depends on the overlapping negative and positive results along the X axis. Binary logistic regression can not solve the likelihood parameter estimates that would produce “perfect” prediction, but “perfect” prediction is what we should expect when there is complete separation.

So what should be done when this type of data is encountered? If the separation is considerably great, then almost all observed values would match predicted values produced by binary statistical regression models. This renders the analysis rather trivial because we essentially know what the predicted values will be before we determine those values. This will occur when any YSNP signature is very isolated from other submissions. Most YSNPs are genetically isolated from other submissions to significant degree.

Obviously, using statistical packages that add noise in the regression analysis would eliminate the low accuracy ratings of the p-values for the “b” constant. Unfortunately, almost no statistical packages have implemented this enhancement to regression analysis. Since it is now known to be a limitation of statistical software packages, I am less concerned about this accuracy rating.

Continuous vs. Discrete Variables

Binary Logistic Regression is much more accurate when the X axis includes continuous numbers (with decimal points) vs. discrete numbers (integers). The book, “Modern Regression Models,” by Thomas P. Ryan, 1997 has an excellent chapter on binary logistic regression and describes this issue extremely well. Form fitting methodologies can yield much more accurate models with smaller sample sizes when more positions are found along the X axis. More positions found on the X axis increases the likelihood of overlapping positive and negative results along the X axis which in turn increases statistical accuracy.

The signature match could be modified by the mutation rate of each marker as well as adjustments for less common multi-step mutations and backwards mutations from the signature. This would result in the X axis becoming more continuous in nature. However, it would be very subjective to assign weighted values of mutations within the signature based on the mutation rate of the marker values. This would also introduce complexity into the analysis and complexity in understanding the YSNP prediction methodology. Any weighting factors would be subjective in nature which could affect the accuracy of YSNP prediction. However, the existing YSNP prediction methodology does not reflect the mutation rate of YSTR markers found in the YSNP signature and over-simplifies the parameters affecting YSNP prediction.

Bias

Bias reduces the accuracy of any statistical analysis and YSNP testing includes significant bias. Analysis of the tested candidates indicates that testing candidates are not well tested in the transitional area of the X axis that could yield more overlapping results. This known bias aggregates the problem of near-complete separation. Additionally, low signature matches (which always yield negative test results) are extremely under-represented except for YSNPs tested by SNP pack testing for several years. However, this bias has little influence on the accuracy of the models since so few testing candidates are found in the transitional area. Testing more transitional testing candidates appears to be the best leverage to increasing the accuracy of YSNP prediction. Additionally, testing in the transitional area is also important in determining the haplotype of the YSNP and early YSTR branching as well. The transitional area are almost always in the earlier part of the haplotree chart and will also produce many more private YSNPs and early branches in the haplotree chart.

There is also a geographic bias of testing based on the geographic origins of testing candidates. Within the R-L21 cluster of SNPs, there is a strong bias towards Irish, Scottish and English origins. The fact that Irish and Scottish surnames are clan based surnames makes Y-DNA much more appealing for these surnames as these surnames have much older origins and fewer genetic origins. Testing is also dominated by sponsors in the United States, United Kingdom and Republic of Ireland which also reflects dominant English, Scottish and Irish emigration to the United States. The continental submissions of R-L21 (France, Germany and Scandinavian countries) are probably not properly represented. Although geographical bias is a factor, it pales in comparison of the bias of not properly testing boundary condition submissions (where submissions are more likely to overlap testing negative or positive).

Sample Size

Binary statistical regression normally is more accurate with larger samples sizes. This remains somewhat true for YSNP prediction as well. For YSNP prediction, accuracy is more influenced by other factors: the degree of separation (overlap between positive and negative test results), the sampling bias found at signature matches in the transitional area of the S-Curve and the discrete number of markers found in the signature. Even where the sample size is extensive (M222 and L226), sample size is much less important than the bias of under-testing in the transitional area along the X axis.

Another form of increasing the sample size could be accomplished via using 111 marker YSNP signatures. This would help increase the accuracy by creating more discrete points along the X axis – however, this would be a marginal improvement. Any benefit of more markers in the YSNP signature would probably be offset by radically smaller sample sizes of testers at the 111 marker level. Increasing the testing of submissions in the transitional area is by far the best approach to reveal overlap between positive and negative results along the X axis. However, this takes awareness of admin focus and support of testing candidates in the transitional area.

Statistical Accuracy

The largest challenge for statistical accuracy appears to be the “near complete” separation of negative and positive observations, the bias of under-testing submissions in the transitional area of the S-Curve (where mixed results are most likely to occur) and discrete numbers being used for the X axis. Nothing can be done about “near complete” since that is the nature of YSNP signatures and exponentially smaller matches as fewer backwards and parallel mutations are found with lower signature matches. Bias could be reduced by testing more submissions in the transition area of the S-Curve. If the signature matches include weighting due mutation rates of each YSTR, then the numbers of X axis would be more continuous in nature. If testing was truly random in nature, then very large sample sizes would eventually test more submissions in the transitional area which would also improve statistical accuracy (but this really assumes the bias would be reduced which is not very likely).

The accuracy of the binary statistical regression model is always extremely high which indicates that binary logistic regression is without any doubt the proper model for YSNP prediction. The concordance of pairs always produces 95 % or higher accuracy. The concordance of pairs is the best methodology to determine the accuracy of the model since it compares the results of the model to all observed values. There is little doubt that the classic S-Curve produced by binary logistic regression is the correct model. Also, the goodness of fit (Pearson) and chi-squared values also indicate high accuracy of the model as well.

The constant “a” directly affects the shift of the entire S-Curve along the X axis. Even when there are missing data points along the X axis, the p-factor of the “a” constant always remains high even though the accuracy of the “a” constant (and predicted values) could change radically if missing X values are later added to the analysis. The statistical accuracy measurement of p-factor for the constant “a” does not appear to be as reliable as the statistical packages reports.

The constant “b” directly affects the slope of the S-Curve. Higher values of the constant “b” result in steeper curves and lower values produce more gradual changes in the curve. This is the most challenging mathematical issue that YSNP prediction faces. However, this also a well documented bug in almost all statistical packages. Many YSNPs have little or no overlap of positive and negative results along the X axis which will result lower in accuracy of the p-value of the constant “b”. However, even though the form fitting methodology can not accurately predict “b”, this limitation makes no difference on the accuracy of the model (equation) since it is already a near-perfect model. Since the values of the constant “b” only affect the transitional submissions and these represent a very small percentage of the sample size, the accuracy of the model will always remain very high. Usually it only takes one overlapping submission (sometimes two or three overlapping submissions) and then p-value of “b” shows very high accuracy.

CONCLUSIONS

The accuracy of the p-values for the “b” constant requires from one to three overlapping values along the X axis. However, the lower accuracy of the p-values of the “b” constant is known to be a software bug in almost all statistical packages approach to regression analysis. True accuracy for observed data is primarily dependent on testing as many submissions as possible in the transitional area of the S-Curve (even over-testing). Sample size has extremely small impact on accuracy since the transitional area along the X axis contains very few testing candidates. Reducing the bias for under-testing submissions in the transitional area of the S-Curve seems to be the most significant action item in order to improve the statistical accuracy of the “b” constant which improves YSNP prediction in the transitional area of the S-Curve.

Only the transitional area along the X axis really needs to be tested since the outcome of this testing can not be predicted with extremely high accuracy. For most YSNPs, there are usually very few submissions found in the transitional area and this is the only area where accuracy is less than desired. This could result in three testing options: 1) do not test since there is virtually no chance of testing positive; 2) do not test since there is extremely little chance of testing negative; 3) test all transitional submissions since more knowledge is gained about the YSNP with these tests. Unfortunately, it is human nature to want to validate that your YSNP will test positive even the odds approach 100 % that they will test positive.

Increasing the width 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. Of course, as full genome sequencing testing become available in the next few years, YSTR signatures could be increased to 400 to 500 YSTRs. However, the number of fast mutating markers would surely greatly increase as well – putting more pressure to include the mutation rate of each marker as part of the signature match measurement and radically larger sample sizes to lessen the impact of inclusion of faster mutating marker values. It has been determined that the faster mutating markers, CDYa and CDYb are just too volatile for YSNP prediction and these markers really are not useful for YSNP prediction and will be eliminated in all future YSNP prediction.

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, with the introduction of comprehensive YSNP pack testing for predictable YSNP branches, this has also greatly increased the amount of YSTR and YSNP data below predictable YSNP branches. Individual testing of private YSNPs has also provided much needed information that more costly NGS testing provides. With the inclusion of private YSNPs in SNP packs and testing of individual YSNPs, 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 or when individual testing of private YSNPs is a better 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 prediction / charting below the predictable YSNP branch.

However, most charting tools are based on network joining software methodology which is not the correct math to chart YSNP/YSTR data. The network joining methodology ignores the biological issue that parallel mutations must greatly outnumber backwards mutations by a factor of ten to thirty times. The network joining methodology produce charts based on backwards and parallel mutations being equal which result in substantial numbers of backwards mutations that DNA does not allow. Also, network joining methodology usually will assign every tester to some part of the haplotree even when accuracy is way below five percent.

Charting of the 500 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 30 parallel mutations of these more faster mutating marker values. Signature methodology stops at 60 % accuracy and currently leaves 20 % of L226 testers uncharted. However, network joining methodology uses one marker signatures of faster mutating markers for placement on the haplotree charts. When there are 20 parallel mutations of a particular marker, this would result in a five percent chance of accuracy which I believe would do more harm than good.