Executive Summary
Analysis of Z255
April 2, 2019

To Neal Downing (lead admin for Z255 project)

This is my 25th haplogroup that I have analyzed in the last six months and I have discovered that Z255 can be accurately predicted (98.2 % accuracy) and accurately charted with SAPP. In order to assist the Z255 project, I attempted to update the column F (project source) to Z255 when found there (probably missing 10 %). I found 454 positive testers of Z255 positive testers (almost all positive testers in the Z255 project are extracted). I added another 20 or 30 % from other projects (feel free to add this data to your spreadsheet or recruit from this summary). There are also 661 Z255 predicted testers as well (did not update project source when found in the Z255 project). I stopped around 450 positive testers since SAPP has a limitation of around 500 testers. I could probably find another 100 to 200 Z255 positive testers and another 300 to 400 predicted testers if I had more time for analysis of this large haplogroup.

The spreadsheet includes the binary logistic model in Column CB which is a macro which predicts Z255 with current data at 98.2 % accuracy for positive testers. New testers could be added and the macros copied and pasted to predict other possible Z255 testers. It requires column CH to CM to be copied first to update of the genetic distance). Next column K (signature match) and L (genetic distance from signature) needs to be copied second and then third copy the prediction model (column CB) which uses column K & L).

Feel free to forward this material to other Z255 admins or others that you want to share this information with. Feel free to use this data in any manner that you want and all data is from public reports from over 5,000 projects (most data is 18 months old). I then added more testers from the Z255 project and a few other projects. As new boundary condition testers are added (tested positive or negative), this would require updates to the prediction model constants or the accuracy will drop a little. The model constants were derived via AcaStat which is very easy to use program and costs only $20 to download. See the EXCEL Sheet tab labeled “AcaStat” for summary of the statistical analysis (some SAPP input is also located there as well).

If you have 20 or 30 key positive testers to add, I would be glad to make one more pass. After this last iteration is complete, this data will be used to update my L21 SNP predictor tool once I get enough haplogroups analyzed to warrant its update. My new YSNP prediction now uses two variables (signature and genetic distance from the signature). This eliminates the need test positive for L21 and now you only have to be predicted to be haplogroup R. Around 54,000 67 marker testers in haplogroup R are used for each analysis.

The SAPP tool works very well for finding surname clusters but since the chart is missing around 20 or 30 % of the confirmed Z255 testers and does not include around 1,000 possible predicted Z255 testers, the surname clusters would significantly improve with larger charting runs. The charting tool could be used to improve YSTR groupings. Also, YSNP prediction model could be used determine Z255 membership as well. I would still recruit testers with 5 to 49 % accuracy for project members but would put these in a grouping labeled boundary condition Z255 testers. Also, the model does not correctly predict eight positive testers which should remain in the project since they are YSNP confirmed.

Feel free to ask any questions and I would like some response and do not feel that you have to respond to this analysis in depth. I would be glad to make one last minor update if want me to enhance/correct any part of the analysis.

For Maurice Gleeson

I believe by charting of all of Z255, you can now use SAPP to chart you Gleeson surname cluster. Clearly Gleeson/371202 is not part of your surname cluster and Glisson/411177 is not part of your surname cluster based on YSNP data and YSTR signatures. There are just too many non-Gleeson testers that they match better than your surname cluster. The SAPP program does a wonderful job of determining surname clusters but does occasionally misplace a tester (if you feel confident that one needs to be moved to your surname cluster, use the /GENDATA to force SAPP to move any tester to your surname cluster). However, do not allow Gleeson testers to reduce their signature size matches as this would be a low probability move. If the signatures are the same (or even better), then you should force SAPP to move one or two testers to your surname cluster.

Via charting this is no convergence (overlap) of your surname cluster and other close genetic matches as YSTR signatures associated with YSNP branches eliminates this issue. However, if you delete too much data from the current input, many of the YSTR signatures will be lost. Your surname cluster should be charted with all Z255 data to filter out those that belong to your surname cluster and those that do not belong to your surname cluster. McLachlan/249281 currently appears to be the only Gleeson NPE that is part of your surname cluster (and he is a marginal call). You can add 10 to 20 additional Gleeson testing candidates (regardless of surname) to see if SAPP adds them into your surname cluster.

Your surname cluster no longer has convergence as we once thought. There is a little convergence of Z255 with other haplogroups but this is only at the two to three percent level. This convergence is not found in your surname cluster. There could be some convergence under Z255 due to the signature size is only five markers and Z255 appears to be older than BigTree or YFULL implies. There are a lot of off modal values under Z255 which implies that Z255 may be just a little too old for perfect YSNP prediction or extremely accurate charting.

The reason that we both struggled to chart your surname cluster was that the scope was too young for either YSNP prediction or charting. It is surprising that Z255 is the correct level as we thought this haplogroup was too old to be the baseline for any analysis under Z255. If you remove data and reduce the scope below Z255, I believe it will appear to have convergence but the issue is really that we have made the scope too small which is this key issue to accurate YSNP prediction and charting.

Again, feel free to use this data in any manner that you want and modify the approach as well.

For Dave Vance

SAPP continues work very well across 25 haplogroups under L21. I even did one haplogroup under R1a and it works there as well. R1a is much more tree like structure and suffers some to convergence. L21 (and P312) have huge starbursts of YSNP expansion that are ideal for charting. Here is the criteria that I use to ensure maximum accuracy for both YSNP prediction and SAPP across 25 haplogroups:

1) The age of the haplogroup must be between 1,500 and 2,500 YBP (some variation is possible if all the other parameters are very good).
2) It really helps to have 10 to 20 branch equivalents for the haplogroup selected (this is probably the biggest limit for YSNP prediction and SAPP charting).
3) The signature size for the haplogroup should be at least seven markers (but this is a surprise as both YSNP prediction and charting does not seem to suffer as much as expected when this is not present).
4) The sample size is a key factor. I have tried with YSNP prediction and SAPP charting with very small numbers of testers (below 25) and this makes charting pretty problematic (YSNP prediction is less problematic). But 50 to 100 Y67 testers appear to be the minimum for sample size. The larger the sample size, the better SAPP does as it will have more data to accurately chart.

The 500 limit of YSTR testers of the SAPP tool is a major limitation. I could not add 661 predicted testers under Z255. I also stopped adding confirmed data. This analysis is missing at least 25 to 50 % of Z255 testers beyond the 1,100 known Z255 testers in the spreadsheet. Also, I miss using SAPP to sanity check my L226 haplotree which is now 790 Y67 testers. I was able to CTS4466 with analyze YSNP prediction but with 1,167 testers, it could not be charted. There are currently 470 CTS4466 positive testers, so I could chart the positive testers only.

If you have any L21 haplogroup admin who needs some assistance with SAPP, I could lend a hand in helping somebody get up to speed. I am now randomly adding new L21 haplogroups, so as long as it meets that above criteria somewhat, I would be glad to help a couple of haplogroup admins.

Rob Spencer

For those that do not know Rob, he presented at the recent FTDNA conference and dazzled the crowd. He is obviously a very important (and refreshing) heavyweight with statistics (and other math) and is an good programmer as well. Only after a little over one year in the genetic genealogy community, we need to get him up to speed as fast as possible with math required by our community: 1) data collection (without data across dozens of projects, the sample size is just too small and is missing too much key data); 2) YSNP prediction (without good YSNP prediction, you do not know what data to chart); and 3) charting (this is the ultimate goal of 90 % of testers in our genetic genealogy community). Here are a few of my YouTubes where most viewers struggle with. I try to avoid too much math detail for broad consumption. Feel free to comment on these video and have fun finding out the issues that you find.

First – why the FTDNA match reports and Tip Report are very optimistic in its assumptions. These tools assume normal distributions which are rarely present (this video needs updating and is a first pass):

https://www.youtube.com/watch?v=AD1HHb0Cwfs

Second – you need to get up to speed with the importance of charting. This is what genetic genealogy wants and SAPP delivers this very well. Also, search "David Vance SAPP" under YouTube if you want learn the details about using his tool:

https://www.youtube.com/watch?v=1W4pw9yrNRU

Third, here is my presentation on YSNP prediction (this is a pretty mature presentation):

https://www.youtube.com/watch?v=AD1HHb0Cwfs

Here is my white paper on YSNP prediction:

http://www.rcasey.net/DNA/R_L21/Math_behind_R_L21_SNP_Predictor_20180202A.pdf

These presentations are intended for people with modest backgrounds in statistics and probability theory but feel have fun finding the issues with these presentations if you want.

For Alex Williamson

Since Alex attended the FTDNA conference and participated in the discussion with Rob, Dave and myself, he might enjoy this analysis. Besides, Alex needs a break the BigTree. I used your BigTree to extract out all of the negative downstream YSNPs for the Z255 analysis. These negative downstream YSNPs greatly enhance SAPP accuracy as it restricts SAPP from moving testers down the haplotree. These negative Big Y downstream YSNPs are not loaded in the FTDNA YSNP reports. Also, FTDNA's clone of Alex's charts omits FTDNA ID numbers. Plus your charts are public and FTDNA limits access to the clones of his charts to only the projects that you belong to.

Here is our long term goal: 1) convince YFULL to publish all the public FTDNA from YSTR and YSNP reports since it is not likely that Russia will join the EU very soon or would accept stop and desist orders from the US for violating US terms and service conditions. Feed this data into the Information Warehouse for Alex's usage; 2) Automate YSNP prediction so we have more valid data to feed into SAPP. We can get our friends in Serbia (NevGen) to add what I find to their web sites and get others to clone my work. All of my newer smaller haplogroups are not in NevGen but they added all my haplogroups from my original L21 SNP predictor tool. This really needs to be automated; 3) we need to assist Dave in perfecting the SAPP tool and increase his limit from 500 to 5,000 testers so the largest predictable haplogroup, R-M222, can be charted. This would allow all three major Irish haplogroups to be charted again (M222, CTS4466 and my L226).