Analyze criteria for appropriateness of the predictable haplogroup.

1) The predictable haplogroup should fall into the 1500 to 2500 YBP time frame.

YFULL has the TMRCA estimate for Z3000 as 1,900 YBP and the Big Tree has the TMRCA estimate of 1,840 YBP. Based on surname diversity, these estimates are very reasonable.

2) The predictable haplogroup should have significant YSNP genetic isolation. Haplogroup should have at least 15 to 20 branch equivalents in the two highest branches of the haplogroup being analyzed.

FTDNA shows Z3000 has 27 branch equivalents which provides very good YSNP isolation. This exceeds the acceptable range for YSNP genetic isolation of 20 branch equivalents.

3) The haplogroup should have significant YSTR genetic isolation and should have seven to fourteen makers in the YSTR signature.

The signature of Z3000 includes eight markers which is acceptable. But this includes one pretty rare null marker value and one multi-step mutation as well.

4) The sample size should have at least 50 to 100 testers at Y67 or higher and should have at least 10 to 20 branches.

Z3000 has 300 testers that were confirmed to be Z3000 via YSNP testing. Another 385 testers were predicted which results in a total of 685 testers. This is now the sixth largest predictable haplogroup under haplogroup R. The number of testers far exceed the upper range of 100 testers. Z3000 has 210 YSNP branches which far exceeds upper range of 20 branches.

Conclusion – Since all four criteria are well satisfied, YSNP prediction has 99.3 % accuracy for positive prediction and prediction of negative testers also has 96.9 % accuracy (but many more remote related negative testers were not added since these were very remotely related). However, there are many boundary condition testers that are not tested for any YSNPs (which will probably require future updates to the model constants to keep the accuracy near 100 % but will not change too much over time).

Data collection beyond the Clan Colla (tracks Z3000) project was reasonably extensive. Not all predicted testers were extracted from the Clan Colla project and just under half of the surname projects were manually reviewed. This is the sixth predictable haplogroup under DF21 and these include 41.7 % of the branches under DF21:

All DF21 1,064 branches
Predictable haplogroups: Z3000 (210), CTS2147 (128), L1402 (37), S6000 (30), L362 (25) and L720 (14) – Total of 444 branches

YSNP Prediction

The YSTR signature includes eight markers at Y67 and the YSNP prediction model delivers 99.3 % accuracy for both positive prediction and 96.9 % accuracy for negative prediction.

Even though prediction accuracy is currently 99.3 %, the huge block of YSNPs associated with Z3000 could be broken up with testing of boundary condition testers. It is recommended that a few of these boundary condition testers be upgraded to Big Y700. This could assist with determining the geographic origins of this branch of mankind and could help with understanding the progression of YSTR mutations as well.

As more Z3000 YSTR matches become known for public analysis, the prediction model will definitely verify Z3000 matches properly. However, to look at YSTR matches other your own results, you must be an admin of the haplogroup project or surname project being analyzed. Privacy hides over 50 % of the confirmed testers, so having admin access is key to better accuracy. The empirical YSNP prediction done manually has 99.7 % accuracy.

The current empirical prediction model is:

Signature match	Highest POS GD	Lowest NEG GD	Empirical Model	
8	10	NA	10 x 2 = 20	with no NEG, POS x 2
7	11	NA	11 x 2 = 22	with no NEG, POS x 2
6	11	10	(11 + 10) / 2 = 11	with both, the average / 2
5	9	7	(9 + 7) / 2 = 8	with no POS, NEG / 2

But additional testing boundary condition testers could require updating of the empirical prediction model and would definitely change the constants of the YSNP prediction model.

The existing formula for YSNP prediction is automated and already done. The formulas for the Z3000 signature match, the Z3000 genetic distance from the signature and the Z3000 prediction model can be copied and pasted into new rows without any modifications required (the row number automatically updates when pasted to a new row for any new tester added or any tester that needs to be deleted).

It should be noted that YSNP prediction of predictable haplogroups will always have very high accuracy and should be a part of any YDNA analysis. Charting in general is not near as accurate but are far superior to YSNP only charting (usually revealing 2X to 4X more branches).

SAPP Charting

There are several YSTRs with large numbers of off-modal values. SAPP has reasonable information to work with due many off modal values. Many solid surname clusters were revealed: Node 1357 Calkins (25 Calkins testers & 4 testers with other surnames); Node 962 McGuire (6 & 0); Node 1167 Peden (7 & 0); Node 1055 Carroll (15 & 1); Node 986 McNaughton (8 & 0); Node 906 Morris (8 & 1); Node 873 Adams (6 & 2); Node 792 Smith (5 & 1); Guttormsson (5 & 1); Node 961 McGuire (11 & 2); Node 771 McDonald (9 & 1); Node 1117 Biggins (10 & 1); Node 1241 McKenna (14 & 7); Node 1199 Hughes (5 & 2); Node 1188 McMahon (6 & 2); Node 954 Roderick (10 & 1); Node 1186 McGuire (6 & 2); Node 1286 Duffy (14 & 8). There are many testers that are nearby to these surname clusters that will someday merge and grow larger with sample size growth.

SAPP really likes the inclusion of negative values for sons of Big Y testers which is difficult to determine most of the time. FTDNA does not publish Big Y testing status in any publicly available report. Using the BigTree chart, many of these Big Y testers have all made their data public, so around half of the negative sons can be added. The Clan Colla admins could determine the status of many more Big Y testers via order history or Y700 YSTRs.

There are other options to potentially enhance the accuracy of the SAPP chart. A significant number of testers are tested at Y111 which could be used. Most confirmed testers are at Y111 markers but many predicted testers are still at only Y67. This could increase accuracy of charting (or could go the other way as well). Also, adding more predicted testers from numerous projects would help a lot as well.

The accuracy of charting in general is not near as good as YSNP prediction. The accuracy is highly dependent on the sample size and the amount of YSNP testing. Manual charting will always be somewhat better than SAPP but any form charting will vary some over time as more YDNA information becomes available. Testers will move around some over time – but YSNP/YSTR charting is far superior to YSNP only charting since it reveals 2X to 4X the number of branches. At 685 testers, this is largest number of testers to be charted by SAPP to date (HTML only). It would really help to have an option to display all branches on the chart which would greatly increase readability of the charts (as well as removing the > and < methodology which is confusing – only show the lower bounds as an option).

Database Issues

There are 264 known Big Y testers found in the BigTree for Z3000. 33 of these Big Y testers found in the BigTree could not be found in any public FTDNA YSTR report. Also, two Big Y testers at Y12 could not be included as charting requires a minimum of Y67. Since FTDNA does not publish Big Y testing status in any public YSTR/YSNP reports, BigTree really helps to add negative sons to Big Y tested testers. Adding these negative values for sons prevents SAPP from predicting YSNPs to lower levels. There are around probably 75,000 Y67 (or higher) testers in public STR reports for haplogroup R and probably over 150,000 testers if all data was publicly available in YSTR reports (around half of all testers are now private). Interested researchers of Z3000 have more time to locate many more relevant testers via YSTR match reports (assuming they have admin access to key projects). It is very important to get these matches to join the Clan Colla (aka Z3000) project as well as the relevant surname projects. It is also very important to have these testers enable YDNA "Opt in to sharing" so that their results will appear in the above projects and can be accessed for analysis.

My current haplogroup R database has grown from 51,000 testers to 66,500 testers over the last eighteen months. Around 20,000 testers have had updates to their terminal YSNP as well as another 15,000 have upgraded from Y67 to Y111. Also, around another 10,000 have updated their surname as well. The current database includes over 95 % of the public testers from October, 2019 for all testers with IDs higher than 270,000 and all other testers whose ID starts with a letter. I continue to work down from 270,000 for the 2019 pull and manually extract many testers from the haplogroups that I analyze (around 5 to 10 % are now 2020 and 2021 manual extractions). The 2019 pull included around 6,000 of the largest FTDNA projects associated with haplogroup R. Unfortunately, many projects are now turning off their public reports and many new testers have not changed their default of private to public. With around half of the data now being missing, the accuracy of any charting will be lower (but YSNP prediction does not need very large sample sizes to become very stable and remain accurate).

I continue to work on automating and improving the extraction of the data. The actual pull of YSTR reports is now quite automated but these queries require maintenance. The separation of the multi-copy markers into separate columns and conversion of 389-2 to the delta format is now automated. However, every pull requires the manual conversion of extra multi-copy markers to normalized values which is labor intensive and would be extremely difficult to automate with EXCEL formulas. This database also requires continual manual analysis of derived columns: 1) surnames are extracted for the male Earliest Known Ancestor and donor surname fields; 2) the most labor intensive step is the conversion of the terminal YSNP into a string of relevant YSNPs for analysis. I am also now adding the source columns for donor surname, male Earliest Known Ancestor and terminal YSNP branch to the master EXCEL spreadsheet. I now have formulas that detect changes in these three fields which invokes the need for a manual analysis of these key sources of data that get updated. I also automatically add the derived column testing resolution (Y67 or Y111). But these new columns are not currently completed across the entire database, so significant manual analysis is still required.

Any tester that is confirmed or predicted to be Z3000 should join the Clan Colla and relevant surname projects. Also, privacy is always a significant issue, so turning on YDNA Opt in to sharing (making it public) is highly recommended as well.

Additional observations

Discovering that Clan Colla tracks Z3000 was a major discovery for me (but reviewing the Clan Colla FTDNA web page, this is well documented). I highly recommend changing the name of the project to Z3000 to elevate the status to a predictable haplogroup project (still keep the Clan Colla origins in the description as well). This project does not show up the lists for haplogroup projects which should be updated. This project is very extensive in scope and is well organized like haplgroup projects should be.

I would like to add another 25 to 50 predicted Z3000 testers, however, this would exceed the limits of SAPP. At 685 testers, only the HTML version can be generated (PNG fails to generate). An earlier version with 656 testers did generate a PNG file but was very fuzzy with my default image program. I added more testers to explore the improved upper limits of SAPP. I also only added half of the negative branches for Big Y testers to hopefully push the limits a little further.

SAPP is doing an excellent job of revealing an amazing number of surname clusters with not that many YSTR mutations to work with (Z3000 is a relatively younger predictable haplogroup that does not allow enough time for larger signatures for surname clusters). There are many testers nearby to surname clusters that have the same surname – so over time these surname clusters will grow in size (more extensive Big Y testing will assist this consolidation as well).

The Clan Colla refers to 425 = 0 and 511 = 9 as being predictors of Z3000. These are by far the strongest indicators of Z3000 but this has lower accuracy than the binary logistic regression prediction model:

425 = 0 (one confirmed tester is not a match and six predicted testers are not a match)
511 = 9 (14 confirmed testers are not matches and 13 predicted testers are not a match)

This method fails properly predict 34 times which is not very reliable. The mathematical model only has two false calls for predicting positive results. This method is much more complicated to analyze but all the formulas for Z3000 are already done. The empirical prediction model misses only one tester and is not very difficult to understand or calculate. Predictable haplogroups with large numbers of testers reveal the best results for charting due to large sample sizes that most predictable haplogroups do not have. Manual charting of Z3000 would be a great addition to the analysis of Z3000 – but this is a very laborious commitment of time (I manually chart L226 and this predictable haplogroup is now almost 1,000 testers at Y67 or higher). The Clan Colla project does an excellent job for its groupings and also attempts to assign YSTR signatures where possible.