Analysis of DF73 and S933 (R-L21>Z253>DF73>S933)

Step 1)

Determining if these YSNP branches meet the criteria for a haplogroup that is both predictable and that can be charted.

a) The TMRCA must be between 1500 and 2500 YBP. According to YFULL:

DF73 – 2500 YBP (has five branch equivalents)
S933/Z2188 (son of DF73) – 2500 YBP (has four branch equivalents)
Z17259/Z2203 (grandson of DF73) – 2500 YBP (has one branch equivalent)

With ten branch equivalents between these three branches, the TMRCA should not be the same, so some of the higher branches must be older and/or the some of the lower branches must be younger.

According to the Big Tree:

DF73 – 2500 YBP (has three branch equivalents)
S933 – 2140 YBP (has six branch equivalents)
Z17259 – No TMRCA (has four branch equivalents)

b) For genetic isolation, it is preferable to have ten to fifteen branch equivalents. This increases the accuracy of both YSNP prediction and charting. Both YFULL and Big Tree fall way short of having a sufficient number of branch equivalents required for high accuracy.

c) The sample size of confirmed and predicted testers should be between 50 and 100 testers. With actual sample size being only 23 testers, this is yet another cautionary flag for accuracy.

d) The minimum acceptable signature size must be six markers or more. For DF73, all five confirmed testers have between only 2 markers to four markers. This eliminates any chance of YSNP prediction and reliable charting. There are seven markers in the signature at S933. Only two testers matches all seven markers but six testers match six of seven markers. There are also one tester that has only five matches and another tester that has only four matches. This suggests that the accuracy of YSNP prediction will be significantly below the normal 95 to 99 accuracy.

Step 2)

Collect data based on the seven marker signature of S933. This data collection revealed only eleven confirmed testers for S933 or lower YSNP branches and only found only ten to twenty possible predictable testers based on reason signature matches and reasonable genetic distance. The data collection also involves collecting testers that are confirmed not to be S933 via YSNP testing (these are required to train the YSNP model).

a) This data collection started with around 54,000 Y67 haplogroup R testers that were collected from over 6,000 projects around three years ago. This master database has been partially updated since this date with the review of around thirty haplogroups.

b) Since the number of testers found was quite small, I next examined projects with common surnames and geographic origins to update the YSNP testing and add many new testers that are confirmed to be DF73 or S933. With this updated version, I created the first YSNP model.

c) I next used another pull of 3,000 projects that were pulled around one year ago. This resulted in several updates to YSNP testing and several new testers as well. After this iteration, I added a lot of negative testers with the model which did change the constants of the model but did not change the prediction to any significance.

d) People reviewing this haplogroup are free to pull new data and update the spreadsheet for new projection models (or just adding more to the chart based on the current model). I would be glad to assist others in making an update to the data (if done in the near future).

e) I highly encourage others to encourage or sponsor YSNP testing of the “boundary condition” testers. Also, upgrading good Y37 matches to Y111 is another way to increase the sample size. If you results are included in this analysis, I would be glad to assist you testing recommendations.

Step 3)

Used AcaStat (statistical software tool available for only $10) to determine the constants of the binary logistic regression model and to determine the accuracy of the model.

Estimated Model

Results = -12.284264 + 3.063993(SIG) + -0.75916(GD)

Classification Table (0.50 cutpoint)

	Predicted Y		
Observed Y	0	1	% Correct
0	105	0	100.00
1	2	8	80.00
		Total	98.26

The accuracy is only 80 % accurate (partially due to only ten confirmed testers). This is with one tester being removed (which would have been a false negative). However, one tester has a RecLOH mutation which is a massive single mutation. Mathematical models are expected to reject these samples from the model as they have inconsistent mutation rates.

Step 4)

General Observations of analysis

a) There are a lot of untested “boundary condition” testers for signature match of six:

	Positive	Negative	Untested
GD = 8	1	0	0
GD = 9	0	1	2
GD = 10	0	2	6
GD = 11	1	0	2
GD = 12	0	2	8

If the many untested testers later turn out to be much more positive than negative, the prediction model could change significantly.

b) There as dozens of negative testers with better matches than the DF73 confirmed testers. This supports the speculation that the TMRCA of DF73 must be much older than 2500 YBP.

c) Even the prediction model is only 80 %, only two testers are not properly predicted. This will probably improve over time as the sample size increases (both new YSTR testers and more YSNP testing of existing YSTR only testers). It should be expected that the charting will have issues that track the poor accuracy of YSNP prediction of S933.

d) This YSNP has very unusual geographical origins than most Z253 testers. It has a lot of Mediterranean surnames for most of the testers: Spain, Portugal, Puerto Rico, Columbia, Mexico and El Salvador. It also includes the surname of Kong and one tester whose origin states Iran. Another tester has no geographic origins but is found in a Greek geographic project. But there are also several testers from Ireland and United Kingdom as well which is much more typical for Z253. This also suggests that lower testing rates from such diversity of countries shown.

Step 5)

Charting of confirmed and predicted testers via SAPP tool

a) Charting works pretty well for many testers that share several YSTR mutations. However, some of the testers have huge signatures not shared with others (these may not be S933). So the poor performance of YSNP prediction may be lowering the accuracy of charting.

b) There are three testers that match all seven markers of the S933 signature but have the highest genetic distance to be predicted to be S933. All four are a genetic distance of ten to twelve and all four have English surnames. There are currently no signature matches of seven that are confirmed to not belong to S933. So the only solution for these “boundary condition” testers is to be YSNP tested to confirm their S933 status.

c) Many testers have many private YSTR mutations which is a yellow flag for accuracy. This is partially due to the smaller sample size but those with a lot private YSTRs may not be S933 or may not be properly placed in the chart as well.

d) With the huge diversity and number of unique surnames involved, it is obvious that this haplogroup is not well tested to date. Most charts of predictable haplogroups have clusters of testers where a surname cluster has formed or shows some sign of forming. There is not one surname with two testers in the current chart.

e) Even with all the issues of charting (and YSNP prediction), the S933 chart does have value which will definitely improve as the sample sizes of this haplogroup increases.