Analysis of L226 YFull YSTR500 data

By Robert B. Casey, February 15, 2018

This document is a preliminary review of the YSTRs reported by 22 L226 testers that submitted their data to YFull for analysis. There are two additional L226 testers that have not joined the L226 YFull project that are not included. There are now over 100 NGS tests under L226, so this analysis is intended to anticipate issues before the release of FTDNA's YSTR500. Here are major observations to date.

ACCEPTABLE CALL RATES

Here is an overall summary of YSTRs called by YFull (including FTDNA's 111 markers):

Total number markers reported	588
>= 10 % called	582
>= 25 % called	544
>= 50 % called	521
>= 75 % called	499
>= 90 % called	459
100 % called	357

It appears that the FTDNA must either use different YSTRs than reported by YFull or that a 25 % no call rate will be acceptable. However, the sample size of only 22 testers is a very small amount of data to draw many conclusions.

NAMING CONVENTION

The markers used by YFull have the same numbers with different prefixes. So it will be required to add letters to several numbers to distinguish several new markers being reported: DXYS156 to X156, DYF393 to F393, DYF385a/b to F385a/b and DYR19 to R19,

Many of the DYR markers are only one or two digits and would be more consistent length if R was added for any DYR marker with only one or two digits in the number. YFull uses ".1" and ".2" vs. "a" and "b" as suffixes for multi-copy markers. These should be converted to the FTDNA alphabetic format to reduce confusion. Some of the longer labels need to be shortened to have consistent and shorter lengths: Y-GGAAT-1B07 to 1B07, ATA71D03 to 71D03, Y-GATA-A10 to A10 and G09411 to 9411.

The vast majority of three digit markers are unique across all YSTR reported, so DYS and DYR are just omitted from three digit markers (except for the few that need another letter to distinguish them apart).

NEW THREE VALUE MULTI-COPY MARKERS

There are four multi-copy markers that consistently report three sets of values as the most common number of markers values being reported. These should be reported since they are consistently three copy markers and this will require some education of the genetic genealogy community that these three copy markers are valid. YFull reports no extra multi-copy values but with a sample size of only 22, these testers may just not have any rarer mutations. No “null” markers are reported as well.

CONVERSION OF DYS389 VALUES

Conversion the two DYS389 markers to the delta values is necessary to properly analyze YSTRs. DYS389I becomes 389-1 and 389-2 becomes DYS389II minus DYS389I.-Fortunately, multi-copy markers now have dedicated columns, so this conversion “8-9” into two columns is no longer necessary.

HOW TO HANDLE NEW YSTR VARIATIONS

YFull has dozens of new values not found in FTDNA data with “.g”, ”.a”, “.t” and “.c” being added as suffixes to integers. 459 (8.g), 497 (14.t), 167 (11.t), 170 (35.a), 171 (44.g), R20 (10.t), 241 (13.c), R45c (13.g), 629 (11.g) and 686 (22.c) are examples found. Should these suffixes just be removed or treated as new variations?

ERROR RATES WHEN COMPARED TO FTDNA

There are a total of 18 errors across 14 of the 111 FTDNA markers. This is only 0.9 % of the total number of markers across 22 testers. Even though this error rate may seem small, this is still a pretty high error rate that could have a significant negative impact on charting.

SIGNIFICANT NUMBERS OF NO CALLS

Across 22 testers, there were a total of 1,633 “no calls” out 588 markers reported which is 12.5 % of the marker values are missing. It also takes 25 % no calls to reach 499 markers which appear to be an excessive number of no calls to tolerate for YSTR500.

HIGH MUTATION RATES

Based solely on only 22 testers, it is hard to determine the mutation rates of the new markers where no academic paper has published any mutation rates. However, the number of markers (with off modal values) appear to track mutation rates to some degree and could be warning signs of possible very high mutation rates. The FTDNA rates were observed across all L226 testers (the value in parenthesis is the sample size). The YFull rates were observed across the 22 testers for L226. The mutation rates are taken primarily from the Heinilla which is the only study that covers all 111 markers and also appears to have the most accurate mutation rates (Burgarella mutations rates are also included as well).

	FTDNA	YFull	Heinilla	Burgarella
CDYb	0.67 (951)	0.80	0.01845	0.03512
710	0.44 (256)	0.57	0.01828	NR
712	0.50 (256)	0.59	0.01638	NR
CDYa	0.55 (951)	0.73	0.01436	0.03512
464c	0.14 (951)	0.18	0.01109	0.00126
413a	0.01 (645)	0.17	0.00893	0.0566
413b	0.02 (645)	0.00	0.00834	0.0566
390	0.15 (951)	0.27	0.00830	0.0566
481	0.33 (645)	0.27	0.00819	0.0527
714	0.18 (256)	0.59	0.00773	NR
650	0.24 (256)	0.43	0.00758	NR
464a	0.06 (951)	0.00	0.00717	0.0325
406S1	0.04 (645)	NR	0.00539	0.0226
557	0.05 (645)	0.05	0.00471	0.0226
504	0.19 (256)	0.29	0.00461	0.0318
576	0.37 (951)	0.27	0.00438	0.0202
534	0.28 (645)	0.23	0.00424	0.0202

The above chart implies that if the off modal values exceed 30 %, there would be a pretty high probability that this marker could have very high mutation rates. Below is a list of non multi-copy markers that appear to have high mutation rates (NR = Not Reported):

695	0.68	NR
F393	0.64	NR
471	0.64	NR
720	0.60	NR
R1	0.45	NR
160	0.45	NR
514	0.45	NR
627	0.41	0.01230 (Burgarella only)
159	0.40	NR
138	0.36	NR
R75	0.36	NR
R90	0.36	NR
679	0.32	NR

There are also numerous multi-copy markers with very high percentages of non-modal values. For YSNP prediction and YDNA charting below predictable YSNPs via binary logistic regression (1,200 to 2,500 years), CDYa and CDYb are very unreliable and 710 and 712 appear to be very problematic as well. Also, going from 111 markers to 500 markers would be a 4.5 times increase in sample size of markers being analyzed. With this many markers, either the YSTR sample size will have to be much larger or more fast mutating markers may have to be removed. Also, the sample size of NGS testers under L226 is only 102 testers while there are 645 testers at 67 markers and 256 testers at 111 markers.

For charting, adding 256 testers at 111 markers to the 645 testers at 67 markers has not had as significant impact as was expected. The scenarios where 111 markers really help is when a one branch has over 75 % 111 marker testers, then more branching is seen. However, 710 and 712 for these branches are very problematic for these branches. However, upgrading to 111 markers does add a modest number of new 111 only YSTR branches. Upgrading to 111 markers does help significantly in determining just how related some testers are. It is pretty common for the genetic distance to sometimes only increase 20 or 30 % while other times the genetic distance increases by 200 or 300 %. So it really helps in determining how closely people are really related much more than expected.

Based on the experience of adding 111 markers to charting, I am less hopeful that YSTR500 will have the immediate impact that the genetic genealogy community expects. However, it will be a learning experience to sort out the mutation rates of these new YSTR markers and the new characteristics of these new YSTR markers. This learning experience will benefit our ability to later better analyze YSTR600 when the read length of NGS/WGS testing exceeds 1,000 base pairs allowing NGS tests to be able to read all 600 YSTRs. When the price of these longer read length scanner tests come down in cost, YSTR testing will no longer be done separately and all future testing will be NGS/WGS testing for $200 to $300 per tester in the next five to ten years (with longer than 100 base pair read lengths).

For now, the biggest advances in the ability to accurately chart will come from YSNP testing in four flavors: 1) NGS testing to discover new private YSNPs (and is the primary vehicle to discover new branches); 2) cost effective SNP packs to place new signatures onto the haplotree (and will reveal new branches as well if they include large numbers of private YSNPs and branch equivalents); 3) testing individual YSNPs at YSEQ or FTDNA (this economical alternative is not used enough and reveals substantially more branches due to only testing private YSNPs and branch equivalent YSNPs that are rarely found in the SNP packs); 4) the future resurgence of the Nat Geo type chip array tests that could include 10,000 to 100,000 YSNPs for major haplogroups.

Our current testing priorities need to be: 1) NGS testing to discover new private YSNPs (as well as creating more branches); 2) SNP pack testing and individual YSNP testing to place YSTR signatures on the haplotree that are not YSNP tested (these also reveal many new branches as well); 3) upgrading 37 marker testers to 67 markers where accurate charting is possible and increases the sample size of 67 marker testers is very important; 4) overcoming the costly impact of restrictive privacy policies by data mining via the crude FTNDA matching system and data mining even more projects for new testers (this could easily double the number of the 67 marker testers in databases at no cost); 5) putting a higher priority on encouraging new YSTR testers where the branches are quite old, genetic distance from others is significant and surnames are varied (building your YSTR cluster).

The next priority should be to continue to upgrade to 111 markers – especially if your branch of the haplotree can achieve over 75 % participation. Since many testers have a bias of fully testing their own YDNA and this does help charting, this remains a valid option. I just do not think that YSTR500 will have the impact unless well over 75 % of these testers are NGS tested in your particular part of your branch of the haplotree but it would be very hard to justify this expense. Higher resolution YSTR testing just will not help have a significant impact on creating accurate charts that most testers expect. Our primary focus on YSTR testing will be to increase the sample size of 67 marker testers over 111 upgrades or YSTR500 upgrades. YSTR upgrades to 111 markers should currently be a lower priority than YSNP testing. Since the NGS tests are the most extensive YSNP available, YSTR500 will be another minor benefit of NGS testing. Special focus should paid to upgrading 37 marker testers to 67 markers that have already extensively YSNP tested.

For L226, L226 SNP pack testing is becoming more of haplotree placement tool as the number of branches rise and the number of private YSNPs decline. With 72 L226 branches, the SNP pack test is now approaching 50 % with known branches. This results in many fewer private YSNPs and branch equivalents being included. For L226, there are now over 1,000 YSNPs that should be tested and the limit of 150 YSNPs per SNP pack is becoming a significant bottleneck. There should be a significant shift to testing individual tests at YSEQ for private YSNPs and branch equivalents. Chip array tests (similar to the old Nat Geo test) would be a much more economic technology to employ. However, the development time of these tests are more significant and could not be economically upgraded as regularly as SNP packs. The FTDNA technology using Mass Array technology is reaching its limits since only 15 % of the relevant L226 YSNPs can now be tested. The YSEQ approach of using tiered Sanger Sequencing is even more dated and is now longer used by L226 testers.

Additionally, there needs to be at least a 10 to 15 % mixture of YElite2.1 tests in any project such as L226. The extra 30 % coverage produces 30 % more branches and should be especially used in the genealogical time where an obvious surname cluster has formed. These tests should also be used in bottleneck branches where much older branches have accumulated large numbers of Big Y testers that have not broken up the branch into smaller scope branches.

EXECUTIVE SUMMARY

The increase of the resolution of YTR testing does help create new YSTR branches and does have value. However, this value is way overrated by many in the genetic genealogy community. Unfortunately, most genetic genealogists want to upgrade their own YDNA to 111 markers to be at the maximum level possible. This is an easy sell for haplogroup administrators since there is a natural bias to test your own YDNA first. However, once a person is at 111 markers and Big Y tested, there is usually still high interest and we need to encourage these testers to donate funds for the good of the project to YSNP test those that have very high or very low genetic distance from the haplogroup signature. The L226 project has raised funds to SNP pack test those testers that belong to a common signature but no YSNP testing has been completed to date (to be able to place these testers on the haplotree). With only around $1,000 raised, this testing has resulted in raising the percentage of testers charted from 80 to 85 %. By the end of 2018, we hope that this targeted mini-project will get the project to 90 % charted. You should always offer three options: 1) for the potential tester to pay for their own testing cost to be able to test more signatures; 2) present partial funding (50 %) of the test to those with only two or three testers belonging to signatures; 3) present full funding (100 %) to those that have four or more testers belonging to the same signature.

The YSTR500 will have even less impact than the 111 marker upgrades due much smaller sample sizes. However, for surname clusters, upgrades to 111 markers and YSTR500 could help significantly. But the current cost of NGS testing is quite expensive for this kind of testing. Once your NGS testing only reveals three or four new private YSNPs, testing private YSNPs at YSEQ would be much more economical to reveal new branches. Also, a mixture of 30 % more coverage from YElite2.1 should be considered for surname cluster research to discover branches from YSNPs not tested by FTDNA’s Big Y.

The highest priorities should be extensive testing of YSNPs (but not those that are predicted at higher accuracy). There needs to be 20 to 25 percent testing with NGS testing (and 10 to 15 % of these tests should be YElite2.1 testing for surname clusters and bottleneck branches). This should also include another 20 to 25 percent testing with SNP pack testing and another 10 to 15 percent of testing of individual YSNPs at YSEQ (which is a challenge). There currently is a strong incentive for discovering (NGS) or revealing (SNP packs and individual YSNP testing) new branches which really helps create more total branches under any haplogroup.

The second priority should be growing the sample size of 67 marker testers. Existing administrators should consider creating a new administrator position whose primary job using the FTDNA matching system to discover new 67 marker testers since around 50 % of these testers are not known to haplogroup administrators. Also, data mining of smaller projects and regular review of known projects also remains a good source of increasing the sample size of 67 marker testers. Upgrading 37 marker testers to 67 marker testers is much more preferable to upgrading 67 marker testers to 111 marker testers. Encouraging 67 marker testing of different sons and grandsons of oldest proven ancestors is another option (as long as they are born in the 1830s or earlier). Encouraging speculative testing of lines that others believe could be related is another good option for growing the 67 marker sample size.