Big Y-700 White Paper

Powering discovery in the field of paternal ancestry

Authors: Caleb Davis, Michael Sager, Göran Runfeldt, Elliott Greenspan, Arjan Bormans, Bennett Greenspan, and Connie Bormans

Last Updated: March 22, 2019

1. Introduction

A father passes down a near-exact copy of his chromosome Y (chrY) to his son. Imperfections during replication result in slight differences (e.g. single nucleotide polymorphisms; SNPs) that may be unique to the son. Genealogists use these patterns of similarities and differences to reconstruct paternal ancestry (the Y-tree). The FamilyTreeDNA team (FTDNA) released Big Y in 2013 to provide these patterns directly to customers. At that time, there were thousands of SNPs on the Y-tree connecting hundreds of men [1]. Since then, tens of thousands of men have been studied, and hundreds of thousands of SNPs have been identified. In 2018, armed with this information, updates to the genomic reference sequence [2], and expansion of the Y-tree, FTDNA decided to replace Big Y with a product that would account for updated knowledge of genealogically important regions and increase the overall coverage of chrY.

Chromosome Y has roughly 57,200,000 nucleotides [2], or base pairs (*bp*). Figure 1 shows chrY labeled according to its regions of interest. The genealogically relevant regions (white) are those that are passed intact from father to son with high fidelity. Other regions are either a) highly repetitive sequence (black), and therefore inaccessible to NGS sequencing technology, or, b) subject to recombination with chrX (grey;

pseudo autosomal regions PAR1 and PAR2) and therefore of limited utility for genealogical applications. After removing the black and grey regions from consideration we are left with approximately 23,600,000 nucleotides (23.6 Mbp) to *target* for sequencing.

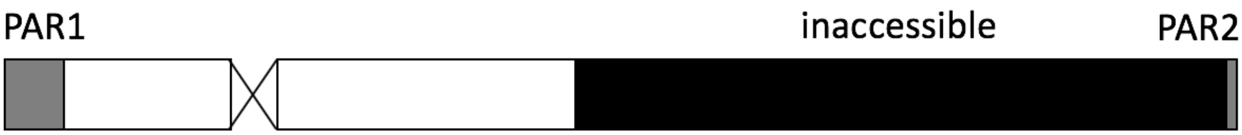

Figure 1: *Roughly 23.6 Mbp of chromosome Y (white) are genealogically relevant*

So, what's stopping Big Y from being 23.6 Mbp in size for every customer? The problem boils down to the following: DNA sequencing instruments provide a limited number of nucleotides from a random selection of DNA fragments in a sample. Therefore, it is cost-effective to sequence a sample after its fragments are enriched for regions of interest. Enrichment of chrY is at the heart of the Big Y product. To make a bigger Big Y we had to improve enrichment for the fragments of DNA most valuable for paternal ancestry…and that's exactly what we've done.

2. Results

We visualize the difference between a Big Y and a Big Y-700 with respect to enrichment of DNA in Figure 2. To understand the curve, start first at its end on the right. Very few nucleotides are spanned by 200 fragments, and so the fraction of the target covered at that depth is low as well. As we follow the curve to the left, we relax our threshold for the number of fragments; therefore, more and more of the target is spanned until at the left-hand side of the curve we have the fraction of the target spanned by a single fragment.

In most cases 10 fragments are enough to confidently call SNPs on chrY. Of course, both platforms perform well by this metric with >95% of the target spanned by at least 10 fragments. Big Y spans most of the SNPs with 50-200 fragments, 40-190 of which are already called with confidence. In contrast, Big Y-700 enriches for enough fragments to call SNPs but does so more efficiently; therefore, the sequencer's capacity is available for fragments originating from more of chrY. This example focused on just those portions of chrY harboring branch SNPs on the Y-tree. If we consider all phylogenetically useful regions of chrY, Big Y-700 routinely spans approximately 95% (22.5 Mbp) with at least one fragment.

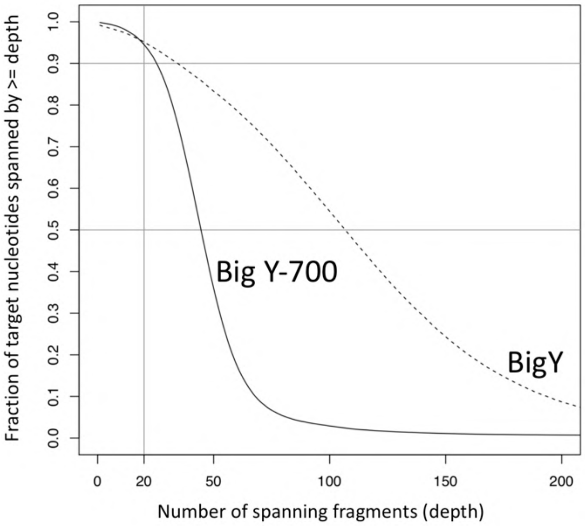

Figure 2: *Big Y-700 efficiently spans more haplotree SNPs than Big Y. A SNP is called confidently with only 10 fragments. At a depth of 10 fragments, Big Y-700 covers 98.8% of these SNPs, while Big Y covers 97.8%. Fragments that added depth to already confidently called regions in Big Y span new regions of chrY in Big Y-700.*

Knowing that Big Y-700 covers branch SNPs at least as well as Big Y, we explored regions of chrY covered by Big Y-700 but not by Big Y. We took 88 samples from a diversity of haplogroups run on both Big Y and Big Y-700 to compare the SNPs and STRs identified on each platform. For STRs, we confirmed 99.7% concordance between genotypes in Big Y and Big Y-700. For SNPs, see Figure 3. On the x-axis we have the GRCh38 coordinate of chrY in million base pairs (Mbp). Each increment represents a bin of 100 thousand base pairs (Kbp). The y-axis on the left is the percent of variants found within a bin. The total number of SNPs called in a bin is shown by the dotted black trace which uses the log scale on the right-hand axis. Some regions exhibited extremely high or low numbers of variants called, including the centromere and the 56.8 Mbp region on the distal end of the q arm (not shown). We considered just those regions with between 200 and 1,000 variants called and observed 87,816 unique SNPs. Of those, 51,417 (58.6%; grey) were identical between Big Y and Big Y-700, 32,450 (37.0%; dark grey) were unique to Big Y-700, and just 3,949 (4.5%; transparent) were unique to Big Y. This finding predicts that a customer’s Big Y-700 results will (compared with Big Y) contain 1) fewer no-calls of branch-defining SNPs, 2) more unique SNPs per sample, and 3) new branch-defining SNPs waiting to be discovered.

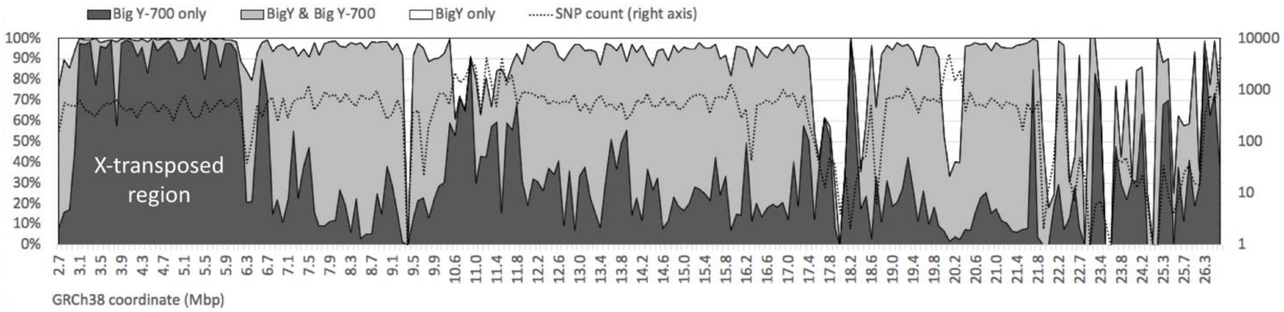

Figure 3: *Genotypes in reference population (n=88) suggest customers will receive > 50% more high-quality SNPs in Big Y-700 than in Big Y. Also, Big Y-700 recovers more of the X-transposed region than Big Y.*

We confirmed Big Y-700 reveals novel branch-defining SNPs by examining from among the 88 individuals, a cluster of 11 samples from haplogroup J-ZS1716 (Time to most recent common ancestor; TMRCA 1,000-1,500 years ago). While Big Y revealed 16 shared SNPs and 7 branches in this clade, a few relationships remained unresolved. We proceeded with Big Y-700 SNPs called in a minimum of 2 samples. We used BY101* sample, S12 (see Figure 4), as an outgroup to eliminate all variants at the BY101 level and above. Big Y-700 retained all of the 16 shared SNPs from Big Y, and produced 8 new variants for placement on the tree.

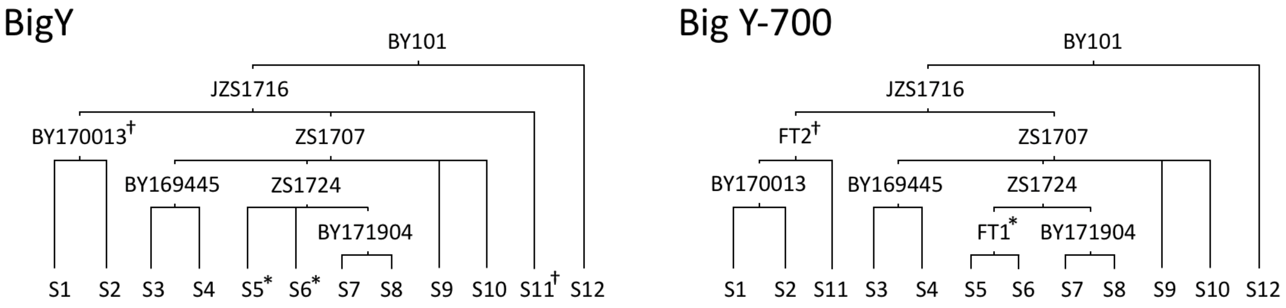

Figure 4: *Novel branch SNPs FT1 and FT2 identified in clade of 11 samples (TMRCA 1-1.5 kya). During placement of the novel branch SNPs, samples S5 and S6 (*) in panel Big Y reorganized under FT1 as shown in panel Big Y-700. Similarly, sample S11 and branch SNP BY170013 (†) reorganized under FT2.*

While 6 of these 8 new variants proved to be equivalents to known branches, two are branch SNPs reported here for the first time. The first SNP resolved relationships between J-ZS1724* samples S5 and S6, known by genealogy to be each other’s closest match. Big Y did not produce a unique SNP shared between them; however, Big Y-700 did, thus creating a new branch for these two samples (see Figure 4; J-FT1;

11172919 G>A). The second SNP resolved relationships among samples positive for J-ZS1716. Sample S11 appeared to be a lone member of a third branch of J-ZS1716 along with J-ZS1707 and J-BY170013. Big Y-700 uncovered a new variant present in S11 and J-BY170013 but absent from members of J-ZS1707, forming a new branch upstream of J-BY170013 (see Figure 4; J-FT2; 13286886 G>A).

3. Conclusions

Big Y-700 raises the bar for paternal ancestry products. It is our biggest and most informative Big Y yet. Given fragments of DNA originating from more of chrY, Big Y-700 customers can expect to receive 40% more STRs and 50% more high-quality SNPs than they did with Big Y. These novel SNPs will answer long-standing questions about relationships between tens of thousands of men from around the world. With the largest Y-tree and the most chrY NGS matches in the industry, FTDNA, project administrators, and Big Y-700 customers are poised to resolve male migration patterns and paternal genealogy with unprecedented granularity.

References

[1] Wei et al. (2013). A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Research*, 23(2), 388-95.

[2] Genome Reference Consortium (GRCh38.p12).