Introduction to the BigY

Introduction

Humans have two sex chromosomes, X and Y, the latter being responsible for dictating male lineage. The past several decades have seen rapid progress toward understanding Y polymorphisms and their applications in order to attain higher resolutions of the human evolutionary tree. Varieties of polymorphism types, and knowledge of their abundances on the Y chromosome, also led to numerous methods of ancestry analysis. Short tandem repeats (STRs) help pinpoint the relationships between individuals, while SNPs have been used for the identification of anthropological origins.

The SNPs revolution began with the advent of large scale microarray genotyping projects such as National Geographic's Genographic Project. The launch of their own microarray, the GenoChip, was meant to enable anthropological research without incidental health findings. The onset of next-generation sequencing (NGS) and high-throughput sequencing (HTS), also made possible further, rapid improvements to the Y tree of mankind.

Despite the promise of HTS, the repetitive and non-unique genetic elements of the Y chromosome present hefty challenges to current read alignment and variant discovery software methods. At the same time, Y sequencing studies continually show there is more to learn. Here, we demonstrate our own novel improvement on Y sequencing, referred to as "BigY." The methods and results we present are based on the first 1,000 BigY samples sequenced and delivered to customers.

Method

I. Assay Design

Of the approximately 60 million base assembly of the Y chromosome sequence, we focused on the male specific region (MSY). There are three distinguished categories in the MSY euchromatin (chromatin high in gene concentration and tightly packed DNA): 1) the X-transposed that confounds mapping with its similarity to the X chromosome, 2) the ampliconic regions with high internal identity, and 3) the X-degenerate regions which map most reliably.

Following this understanding of Y chromosomal structure, we designed probes to enrich the non-recombining MSY regions. The resulting targeted regions spanned approximately 20 million bases, and the 67,000 capture probes were most reliably found covering X-degenerate portions of the MSY. It also spanned roughly 85% of the “Gold Standard” regions (chromosome Y positions placed on the phylogenetic tree by the YCC, **http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2748900/#bib8**), which is approximately 10.5 million bases on Y chromosome. The visualization of our probe coverage, similar to an experiment performed by G. David Poznik et al. (2013), can be found in Figure 1.

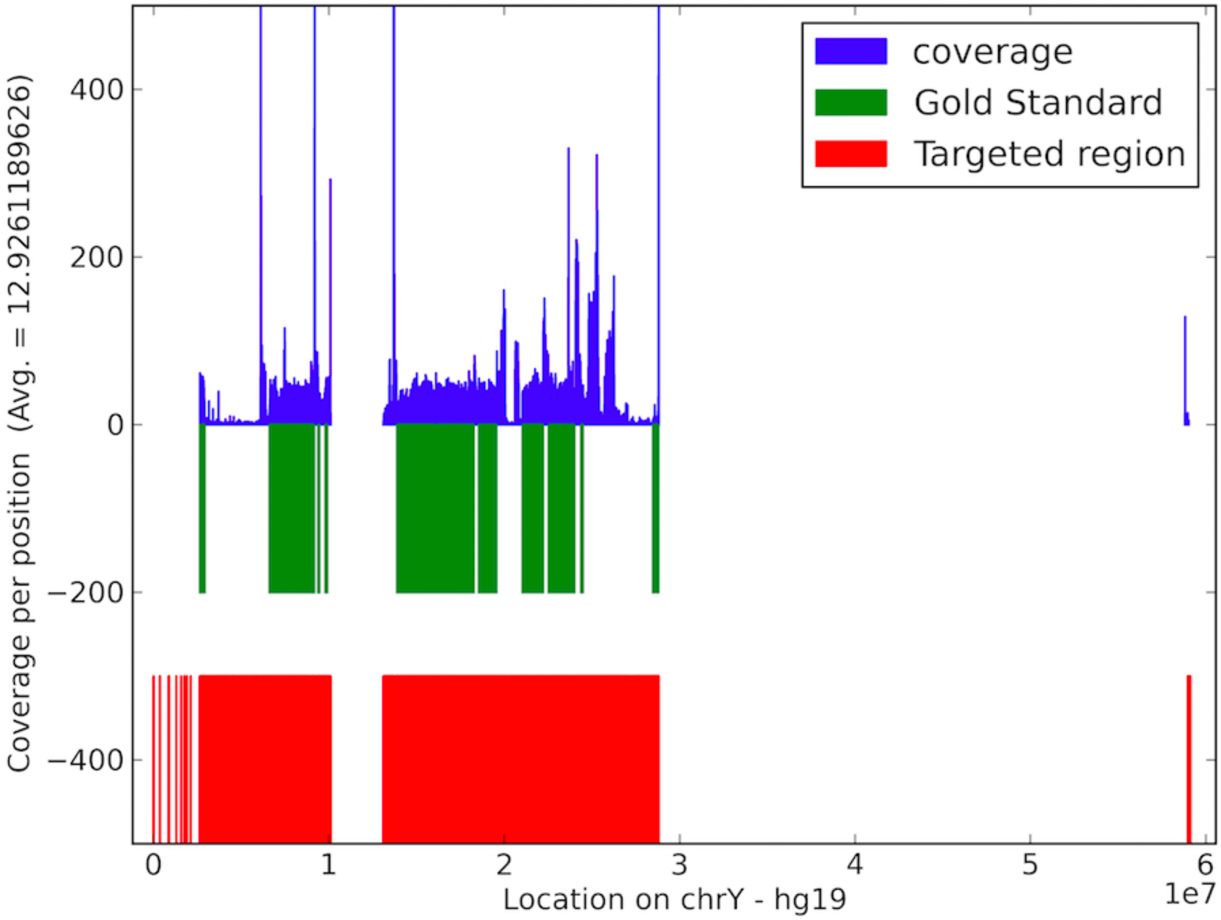

Fig1: Average coverage of 1,000 BigY samples across the Y chromosome. In this figure, the red bars indicate the regions on the Y chromosome that were targeted for sequencing. The green bars indicate the Gold Standard regions. The red and green bars are on the negative Y axis indicate presence only and are not indicative of coverage. Target regions are covered if at least one read covers the position. The blue bars indicate the sequencing coverage across the Y chromosome.

II. Sequencing

Sequencing was performed on the Illumina HiSeq platform and downstream analysis was conducted with Arpeggi genome analysis technology. Post sequencing, the reads were mapped to the hg19 version of human genome reference, followed by post-processing and variant calling, all of which were performed using Arpeggi Engine (an internal bioinformatics software for alignment and variant calling of next generation sequencing data that is now exclusive to Gene By Gene).

For each sample, high level statistics were calculated for quality control purposes, such as:

1. Total number of reads, which are the individual fragments of DNA sequences;
2. Average coverage, which is the average number of reads that span the target region of the reference it is mapped to; and,
3. Average base quality, which measures the probability that a base has been called incorrectly by the sequencer.

Confident region sizes

The Arpeggi engine variant caller calculates a genotype quality score at every position. All positions that pass the threshold (current default is 3.02) are included in the “confident regions” BED file (a file with a list of relevant positions, or ranges of positions). The typical size of the per sample confidence regions BED file ranges from 10.8 to 12.9 million bases, with the average being approximately 11.71 million bases. Figure 2a demonstrates the distribution of samples across a range of confident region sizes.

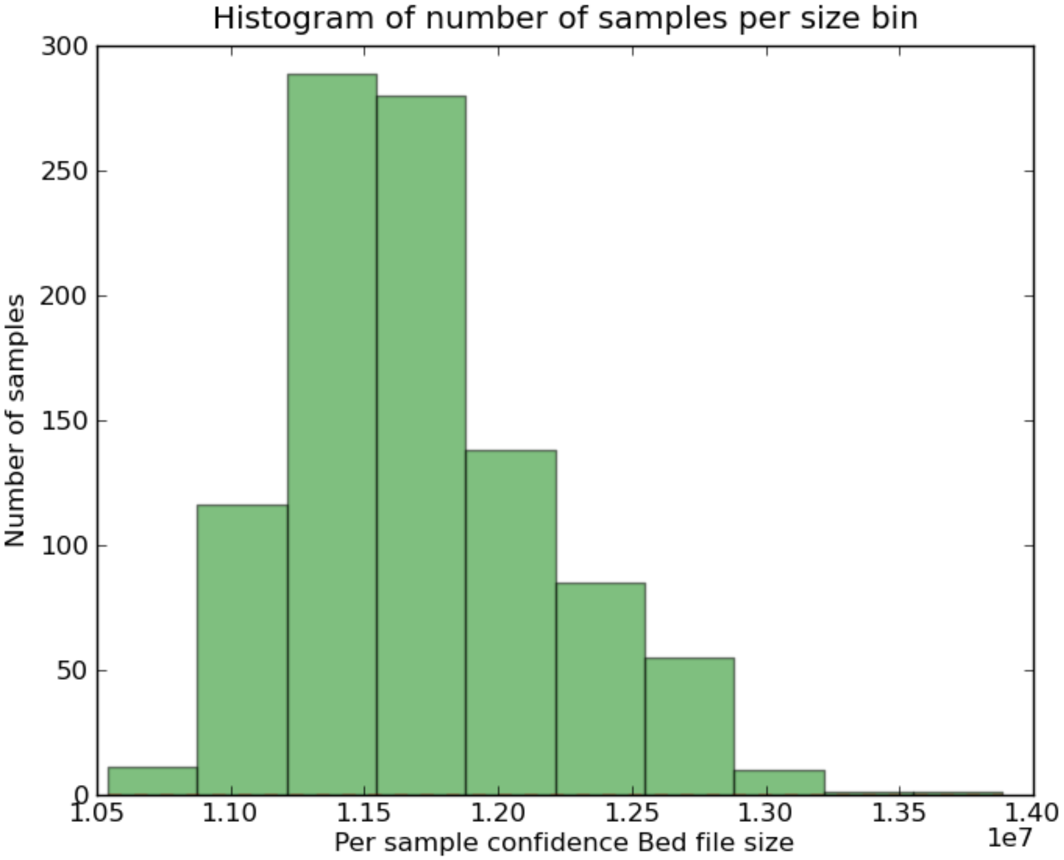

Fig 2a: The above histogram illustrates the distribution of the number of bases in the confidence regions of 1000 samples that were processed for BigY using Arpeggi Engine. The average is approximately 11.71 million bases.

Maximize potential confident region

We started by investigating the “globally covered” regions, i.e., the regions on chromosome Y for which a decent coverage was obtained, minus the regions included in the Gold Standard region. Our investigation resulted in search regions summing up to approximately 4.36 million bases. For these regions, we investigated 100 base windows (a window refers to a set of contiguous bases) at a time and slid to the next window, with an overlap of 50 bases with the previous window, to ensure we did not lose bases due to characteristics present on the edges of two windows. Based on various characteristics of each of the windows, and the individual bases in each window, we determined whether a window would be appended to the potential confident region, or, the “extended Gold Standard” regions BED file. You may access this information via this link: https://www.familytreedna.com/documents/bigy_targets.txt. The characteristics (along with default cutoffs) are as follows:

1. Fraction of allele depth (number of reads that have the allele) at position [>= 0.95]
2. Fraction of qualifying positions, i.e., positions that pass the threshold mentioned above, in a window [>=0.95]
3. Average read depth per window [>=10]
4. Average strand bias (ambiguity due to the forward and reverse strands supporting different genotypes) per window [<=0.9]
5. Fraction of non-zero coverage positions per window [>=0.95]
6. Average base quality per window [>=30]
7. Average map quality (a quantification to indicate that the read is mapped at an incorrect location) per window [>=1]

As a result of this experiment, we added approximately 3.25 million bases to the Gold Standard regions.

High confidence BED region

Post analysis, an intersection is performed between the confidence region BED file and the extended Gold Standard region BED file for each sample and labelled as ‘confidence gold’ BED file. This file is delivered to the customer as a part of the product, which also includes a VCF file (variant information file). For more information about VCF files, see **http://samtools.github.io/hts-specs/VCFv4.1.pdf**

The typical size of the per sample confidence gold BED region is approximately 10.31 million bases. Figure 2b demonstrates the distribution of samples across a range of confident gold region sizes.

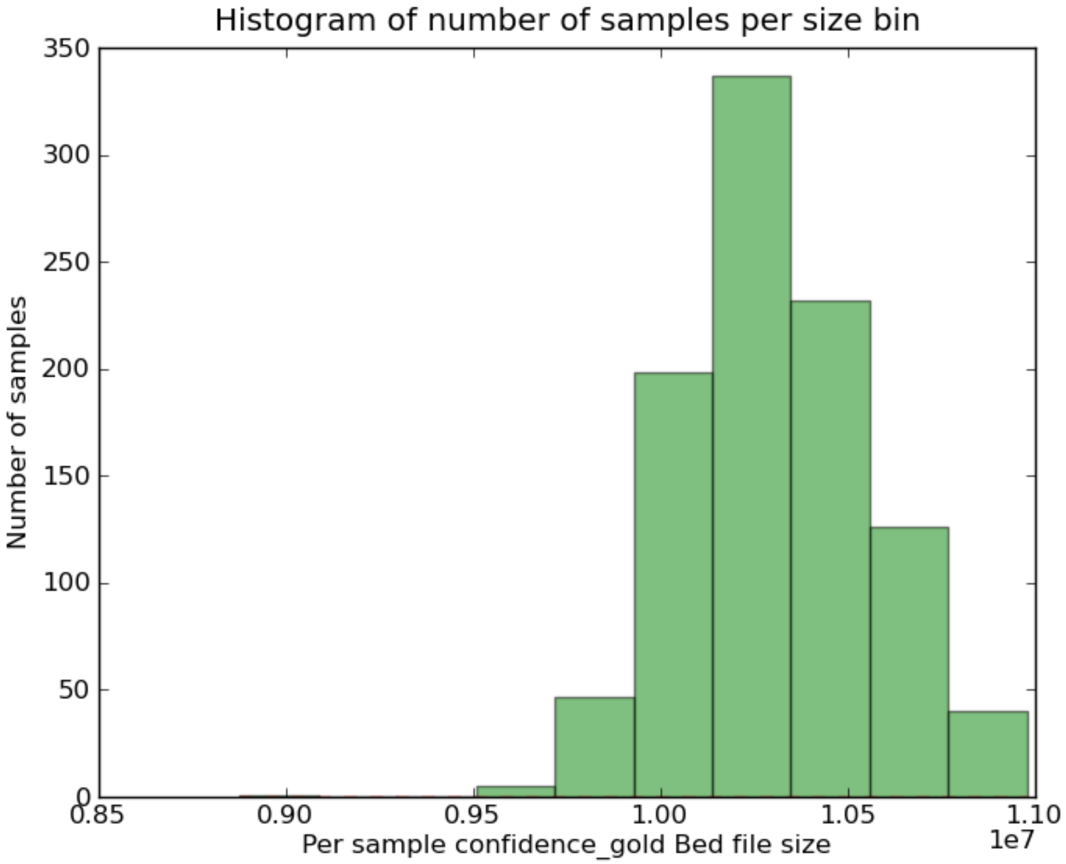

Fig 2b: The above histogram illustrates the distribution of the number of bases in the confidence Gold regions of 1,000 samples that were processed for BigY using Arpeggi Engine.

III. Variants analysis

Common SNPs

For the samples sequenced with BigY, the confident regions were checked against the known positions in FTDNA's internal database of ~40k SNPs 40k data (our in-house microarray Chip data), SNP Chip data (third party microarray chip data) and ISOGG data (http://www.isogg.org/tree/ISOGG_YDNA_SNP_Index.html). Table 1 below illustrates the average number of bases in common between the confident regions of each sample and the above sources:

Common with FTDNA (~40K)	**Common with SNP Chip (~11K)**	**Common with ISOGG (~13.3K)**
29,426	10,080	10,043

Table1: In the above study, we computed an average number of common regions found between the positions in the per sample confident regions and each of FTDNA 40k, SNP Chip and ISOGG data

Concordance with orthogonal data

We compared calls for five samples for which we had 1000 Genomes data. We observed, out of all common positions between BigY and 1000 Genomes, i.e., for regions which had adequate coverage for both 1000 Genomes and BigY, 100% of calls were concordant. For the positions called in 1000 Genomes and not called in BigY, they either had a read depth of less than 2, or were cases where our variant caller would have made a heterozygous or reference genotype call. Similarly, the calls made by BigY and missed by 1000 Genomes was due to the lack of adequate coverage in those specific regions in the 1000 Genomes samples. The following table (Table 2) shows the actual numbers for this comparison.

Sample	BigY calls	1000 Genomes Calls	Concordant calls	Mis-matched calls	Calls not in BigY (Total)	Calls not in BigY (read depth < 2)	Calls not in 1000 Genomes
GRC11075053	486	115	54	0	61	13	432
GRC11075074	1785	967	815	0	152	80	970
GRC11075082	1701	915	716	0	199	123	985
GRC11075100	485	99	63	0	36	11	422
GRC12076445	1845	839	680	0	159	123	1165

Table 2: This table shows the high level comparison of all the calls made by BigY and 1000 Genomes, along with concordance analysis.

Conclusion

BigY covers approximately 20 million bases of the non-combining MSY regions of the Y chromosome, which overlaps with 85% of the Gold standard region and >75% of other SNP Chip based genotyping technologies. After using our cost effective NGS instruments, HiSeq 2000 and 2500, to sequence and then performing in-depth downstream analysis with Arpeggi Engine, it is possible to provide genotypes in approximately 10.31 million bases on average with confidence. The calls made by Arpeggi Engine have proven to be comparable to calls made by others, such as the 1000 Genomes project.

Discovery of novel variants in samples will eventually assist refinement of the human evolutionary tree, thus improving our understanding of human genealogy.