Considerations for the Processing and Direct-to-Consumer Return of Exome Sequences

E.D. Harrington[1], C.Y. McLean[1], A. Shmygelska[1], A. Chowdry[1], B.T. Naughton[1]

[1]23andMe, Inc, Mountain View, CA.

Introduction

In late 2011 23andMe announced our first publicly available sequencing product: the Exome Pilot Project. To return these data directly to consumers (DTC) we implemented a processing pipeline that maximizes the value to the consumer while maintaining data quality and security at each step.

Methods

Enrollment in the pilot was limited to individuals who had been genotyped on the 23andMe platform. Samples were enriched using the Agilent SureSelect 50Mb platform[1] and sequenced to at least 80X unaligned coverage using paired-end Illumina sequencing technology[2]. Participants received their aligned raw data, variant calls, and a summary report describing relevant statistics and potential variants of interest based on a custom filtering process.

We implemented a flexible and scalable pipeline for processing sequence data (exome and whole genome) using the Broad Institute's best practices[3] (Figure 1). It employs a combination of standard tools (e.g. GATK[4], bwa[5], samtools[6], Picard[7] and snpEff[8]) and custom software to automate the tracking of samples, data distribution to and collection from compute nodes, read mapping, variant calling, report generation, and validation against our existing genotype database. To quickly meet fluctuations in demand the pipeline can be deployed either locally or on a cloud platform.

Data security is integral to DTC data delivery. Data was encrypted with keys delivered via secure messaging on the 23andMe website. The encrypted raw data for each exome was on average 6GB, making bandwidth and data integrity another concern. We delivered data via Amazon S3 and the use of encryption made errors in transfer immediately obvious.

Results

In general* the coverage vastly exceeded the goal of 80X which allowed us to call 82-92% of bases in the regions targeted by the Agilent kit (Figure 2). The resulting variant calls were of high quality and were highly concordant with the results of the 23andMe chip (Figure 3). The majority of targeted regions were called in most exomes and those that were not had the characteristics of hard-to-sequence regions (Figure 4). See figure legends for further details.

* Two samples failed at the enrichment step and one failed to achieve 80X coverage. In all cases the customer was refunded.

Acknowledgments

We thank 23andMe's customers who consented to participate in research for enabling this study. We also thank the employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

References and Resources

1. Agilent SureSelect 50Mb. http://www.chem.agilent.com/Library/datasheets/Public/5990-6319en_lo.pdf
2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–9 (2008).
3. Broad Best Practice V3. http://gatkforums.broadinstitute.org/discussion/15/retired-best-practice-variant-detection-with-the-gatk-v3
4. DePristo, M. et al. A framework for variation discovery and genotyping
5. using next-generation DNA sequencing data. Nature Genetics 43, 491-8 (2011)
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
8. Picard. http://picard.sourceforge.net
9. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92
10. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PloS ONE* **7**, e30377 (2012).
11. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, R18 (2011).

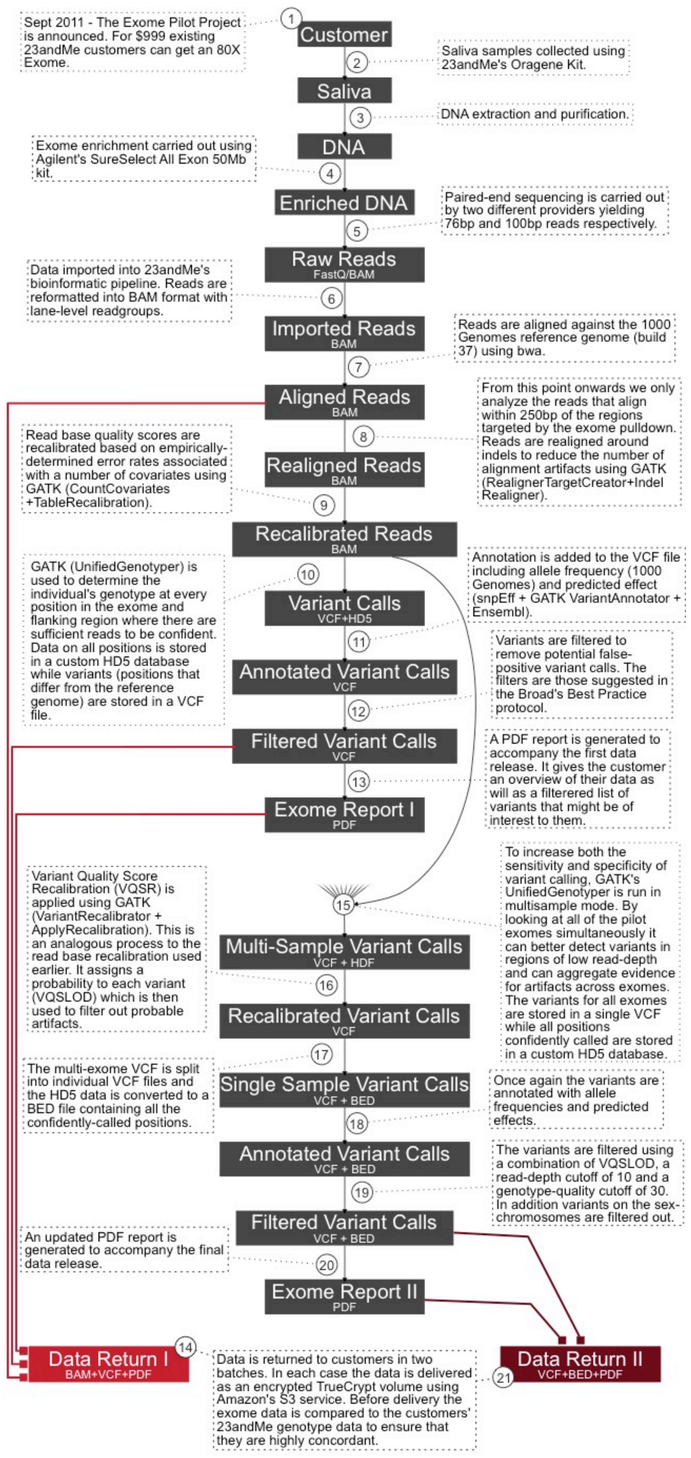

Figure 1. Exome Pilot workflow. The process can be run locally or on the cloud.

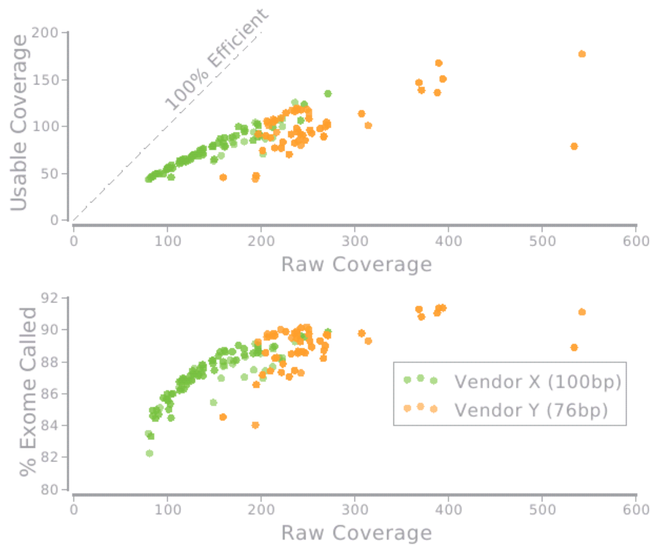

Figure 2. Coverage and proportion called for the pilot exomes. The top panel shows the relationship between raw coverage (the number of bases sequenced/total target size) and the usable coverage (the number of read bases that align unambiguously to the target region after duplicates have been removed/total target size). The shorter read length used by Vendor Y (76bp) leads to a higher proportion of ambiguously-mapped reads, however this is compensated for by the higher raw coverage. The bottom panel shows the relationship between raw coverage and the percentage of the target regions that could be confidently called.

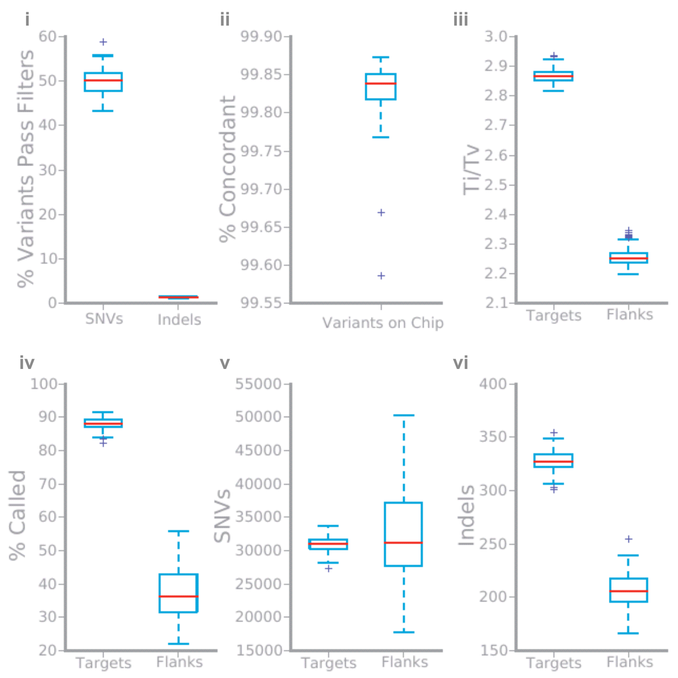

Figure 3. Summary of the data returned to customers. i) The filtering steps described in Figure 1 removed ~50% of SNVs and almost 99% of indels. ii) The remaining variants are on average 99.85% concordant with the 23andMe genotyping chip . iii) The transition to transversion ratios for targets and flanks (250bp either side of the targeted regions) are in the expected range for coding and non-coding sequences respectively. iv) The vast majority of the targeted bases and a significant proportion of the flanking regions could be called. v+vi) Each customer carried ~30,000 SNVs and several hundred indels in the targeted regions.

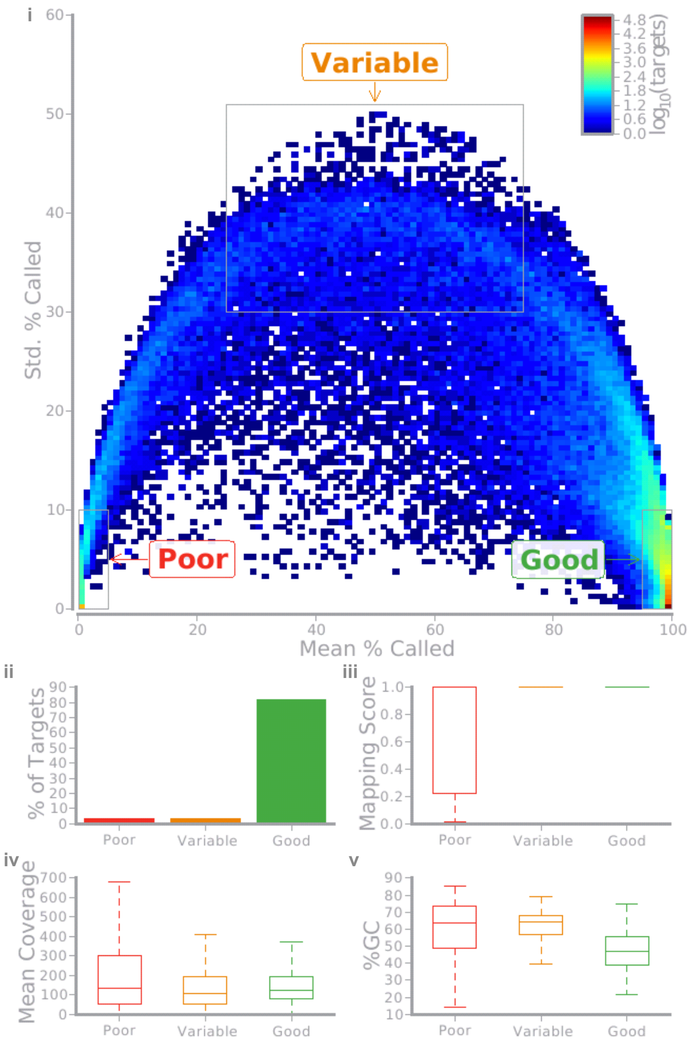

Figure 4. Assessing target performance across exomes. i) A 2-dimensional histogram of target regions. For each target in each exome the percentage of bases that can be called was calculated. Each target was binned by its mean (x-axis) and standard deviation (y-axis) of this value across all exomes. Three sets of targets were selected for further investigation – Poor: mostly uncalled across all exomes; Variable: a highly variable call rate; and Good) consistently well-called across all exomes. ii) The total number of targets per region. iii) The Poor region is characterized by much lower mappability scores[9] than the other regions, suggesting that it is enriched for repetitive and paralogous sequences. iv+v) The high GC-content and low read-depth of the Variable region suggest that the variability in call rate may be due to a previously-described PCR bias[10].