Individual Whole Genome Mapping: From NGS Reads to Clinical Variants

Prasad Patil, BA and Peter J. Tonellato, PhD Center for Biomedical Informatics, Harvard Medical School, Boston, MA

Abstract We developed an automated, open-source pipeline that detects, annotates, and reports genetic variants in individual next-generation sequencing (NGS) whole genome assemblies using a standard, quality-driven format. We will present results from a full-scale run of our pipeline, from NGS raw short-read data to annotated and clinically relevant variants, and demonstrate an application of the annotated variants for clinical decision-making.

[1,2]**Methods** This project exclusively uses open-source technology for all alignment algorithms, variant annotations, and validation processes. Sequence alignment algorithms were compared by accuracy, speed, usage of quality metrics, and community standards to determine which algorithms to employ in our pipeline. Based on these criteria, we chose the MAQ and Bowtie alignment and variant detection software packages. Once the raw mapping data were gathered from the sequence alignment step, a BioPerl conversion script put the variant data into Human Genome Variation Society (HGVS) nomenclature. The pipeline is intended to be modular and can handle diverse forms of genomic data. Modularity was achieved through the development and use of standardized intermediate and final file formats for data and a final summary report template. Custom conversion scripts were written to transform the data into the defined standard formats. Additionally, all development was conducted in a cloud computing environment with the built-in option to scale up resources and infrastructure as the mapping alignments and variant reporting become more computationally intensive. A summary of the pipeline's architecture appears in **Figure 1**, and demonstrates how raw data can be processed by multiple tools and still be reported in a procedural, standardized format suitable for clinical decision making.

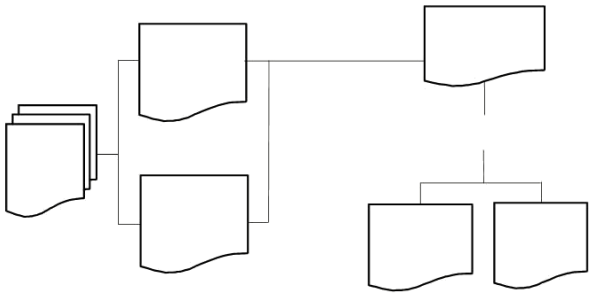

Figure 1: Summary of IWGM pipeline architecture

[3]**Results to Date** Using NGS data from an African individual's human genome, we ran the mapping and annotation processes on the Amazon Web Services cloud[45] for a total cost of approximately $1,700. We compared the resulting annotated variants to dbSNP and found 65.4% overlap (the original assembly reported 73.6% overlap), which confirms that our pipeline meets published standards for read mapping and variant detection. We were then able to pinpoint variants in genes (CYP2C9, VKORC1) known to influence the metabolism of the anticoagulant warfarin and used a predictive dosing modelto generate an individualized warfarin dosage recommendation. The predicted individualized dosage recommendation of 7.0 mg/day was 40% greater than the standard guideline dose of 5.0 mg/day, indicating a relative risk for a potential bleeding event if warfarin therapy were recommended for this patient without consideration of the patient's genetic data.

Conclusion Using the IWGM pipeline, we are able to drastically reduce the totality of genomic data for an individual into clinically relevant data suitable for incorporation into an Electronic Health Record. We have been able to demonstrate the efficiency and utility of our pipeline and the practical importance of the annotated clinical variants it produces.

Annotation

References 1. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality
scores. Genome Res. 2008;18:1851-1858. 2. Langamede B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA
sequences to the human genome. Genome Biology. 2009; 10:R25. 3. Bentley DR, Balasubramanian S, Swerdlow HP et. al. Accurate whole human genome sequencing using
reversible terminator chemistry. Nature. 2008;456:53-59. 4. Amazon Web Services [homepage on the internet]. Amazon.com, inc. c1996-2009 [updated 2009; cited
2009 October 01]. Available from: http://aws.amazon.com 5. Gage BF, Eby C, Johnson JA et. al. Use of pharmacogenetic and clinical factors to predict the therapeutic
dose of warfarin. Clin.Pharmacol.Ther. 2008;84(3):326-331.