ConVarCal: A Reliable and Robust Platform for Next-Generation Sequencing Variants Identification

Yonglan Zheng¹, Toshio F Yoshimatsu¹, Alex Rodriguez², Ravi K Madduri², Paul J Dave², Olufunmilayo I Olopade¹

1. Center for Clinical Cancer Genetics, Department of Medicine, The University of Chicago; 2. Computation Institute and Argonne National Laboratory, The University of Chicago

Background

In the era of next-generation sequencing (NGS), enormous volume of single nucleotide variants (SNVs), insertions and deletions (Indels), and structural variants (SVs) can be identified at both individual and population levels. To minimize the false positive/negative detection rates, to gain better variant annotation, and to curtail the bottleneck inherent to processing of Big Data, we developed ConVarCal (Confident Variant Calling), a platform for NGS variants identification.

Methods

ConVarCal integrates multiple NGS reads aligners, variant callers and annotator in a malleable manner using the elastic computing capability of Globus Genomics built upon Amazon Web Services. Raw data (e.g. FASTQ) can be safely transferred via Globus Online and automatically fed into ConVarCal for BWA-MEM or Bowtie2 alignment. The resulting BAM files can be processed by highly parallelized variant calling pipelines, including Atlas2, FreeBayes, GATK HaplotypeCaller, Platypus, and SAMtools mpileup (Figure1). To realize dynamic parallelization of some callers, wrapper scripts using Swift language are implemented. The output multi-sample VCF files can then be normalized before employing the Consensus Genotyper to select a set of highly confident SNVs and Indels. SVs can also be detected using tools such as Delly and CONTRA. Variants can be further annotated on ANNOVAR. Here we elucidate the performance of ConVarCal by analyzing germline BROCA cancer risk panel targeted sequencing data (Illumina 2500 paired-end, 1.3Mbp targeted regions, average depth of 260x) in a test dataset of 200 Nigerian breast cancer patients with family history and/or early age at onset.

Results

FASTQ files (average 1.65 GB per sample) were transferred across continental USA in about 3 minutes per sample at each interval. The entire analysis of 200 samples on ConVarCal were completed in one to two days, depending on the configuration and resource availability of cloud computing.

The total numbers of SNVs and Indels called by 5 callers are: Atlas2: 11,558; FreeBayes: 10,842; GATK: 11,948; Platypus: 11,482; and SAMtools: 12,236. 9,925 variants were found in agreement with all five callers, Atlas2: 85.5%; FreeBayes: 10,842; GATK: 11,948; Platypus: 11,482; and SAMtools: 12,236. The total numbers of SNVs and Indels called by 5 callers are: Atlas2: 11,558; FreeBayes: 10,842; GATK: 11,948; Platypus: 11,482; and SAMtools: 12,236. Among them, ConVarCal confidently identified 14 deleterious SNVs and 11 damaging Indels (BRCA1: 15, BRCA2: 4, PALB2: 2, BRIP1: 1, CHEK2: 1, NBN: 1, TP53: 1) in 29 subjects. All mutations detected have been confirmed by Sanger sequencing.

ConVarCal takes full advantage of Globus Genomics for NGS variant detection in a reliable and robust manner. It is scalable for larger tasks, and modular to adapt more tools such as FermiKit and VarScan2. New tools like BreakDancer, BreakSeq2, Lumpy, Pindel and MetaSV are currently being optimized to further enhance our platform.

Funding

John and Editha Kapoor Charitable Foundation, Susan G Komen for the Cure, and NCI U01CA161032