Genomic Data Analysis

OVERVIEW

In addition to providing sequence data generation, Broad Genomic Services also supports analytical activities for both the Broad community and as a service externally. Utilizing the Broad developed FireCloud platform to run version controlled analysis workflows, we collaborate with a wide variety of scientific groups from both the germline and somatic disease communities to establish leading best practices in analysis capabilities.

GERMLINE AND OTHER ANALYSIS SERVICES

Coverage Analysis

- This workflow allows for the identification of under-covered regions (defined as <20x coverage in 20% or more samples) and is compatible with hg19 or hg38 data. This tool is useful for characterizing and optimizing panels, exomes, and genomes.

Germline Rearrangement Detection

- Manta (version 1.3.1) is used to call structural variants in germline sequencing data.

Genotype Concordance Assessments

- Genotype concordance compares variants called in a callset (typically VCF) to a truth dataset. The truth dataset is currently NIST hg19 NA12878, but this can be modified if needed, and the input is most efficient as a VCF, but BAM and CRAM are also accepted.

ULP-WGS ichorCNA Purity/Ploidy Analysis

- This tool calculates the purity and ploidy of an ultra-low pass whole genome.

SmartSeq2 Single Cell RNA expression and QC

- This pipeline uses HISAT2 to align reads sequenced from each cell to the reference and produce a BAM file, RSEM to estimate expression values, and Picard to obtain QC metrics such as alignment metrics, GC bias, insert size, quality by cycle, and RNA coverage metrics. Future plans will include quality metrics across the samples in a plate.

SOMATIC VARIANT ANALYSIS SERVICES

Somatic SNV, Indel and CNV Calling

- **Tumor-Only or Matched Tumor-Normal Analysis -** With your choice of either GATK3 or GATK4 versions of Mutect2, and the GATK4 version of the CNV caller, this service provides somatic SNV, insertion, deletion, and copy number calls with or without the use of a matched normal. This service generates both coding and non-coding mutational load on a sample or sample set, and quality control metrics such as the percentage of somatically callable bases, cross-individual contamination, and mutational spectrums with lego plots. When a matched normal is available, the analysis workflow provides germline calls with HaplotypeCaller. A panel of normals is used as a noise model to improve the specificity of calls, and is used in lieu of a matched normal.
- **Somatic Variant Add-ons**
  - Annotation with snpEff
  - Consolidation of MAF files from multiple individuals into one
  - Conversion of CNV output to older ReCapSeg format
  - Generation of PASS-only MAFs
  - Venn diagram comparison of variant call sets

Somatic Structural Variation Detection

- **Tumor-Only or Matched Tumor-Normal Structural Variation Detection -** SvABA, formerly known as Snowman, is a method for detecting structural variants in sequencing data using genome-wide local assembly.

Mutational Burden Calculation

- This tool calculates both coding and non-coding mutational load on a sample or sample set. It can be run on the whole genome, the whole exome, a list of cfDNA genes, or a custom defined gene list. The required input is a MAF.

Somatic Performance Assessments

- Used to calculate false positive rate, sensitivity for calling variants by allele fraction and read depth, and repeatability of somatic variant calling using the Jaccard similarity index. Replicates of NA12878 (recommended 6) are required for false positive rate calculation. Replicates of the 5, 10, and 20plex HapMap Cell Line DNA pools (required >3 each) are required for sensitivity and repeatability calculation. These tools may be used for assessing process changes or characterizing new panels. Mutect1 + Indelocator or Mutect 2 tools are used depending on the analysis pipeline.

FOR MORE INFORMATION

Web: genomics.broadinstitute.org
Email: genomics@broadinstitute.org