Benchmarking Variant Calling Pipelines using Human Whole Exome Sequencing Data

This repository describes benchmarking of various aligners and variant callers for Whole Exome Sequencing (WES) data analysis, focusing on illumina-based platform. To find more information about the performance assessment of different variant calling pipelines click here or https://doi.org/10.1101/359109

Work Flow

We have implemented following workflows for 20 different WES pipelines in combination of five and four different aligners and variant callers respectively Figure.1

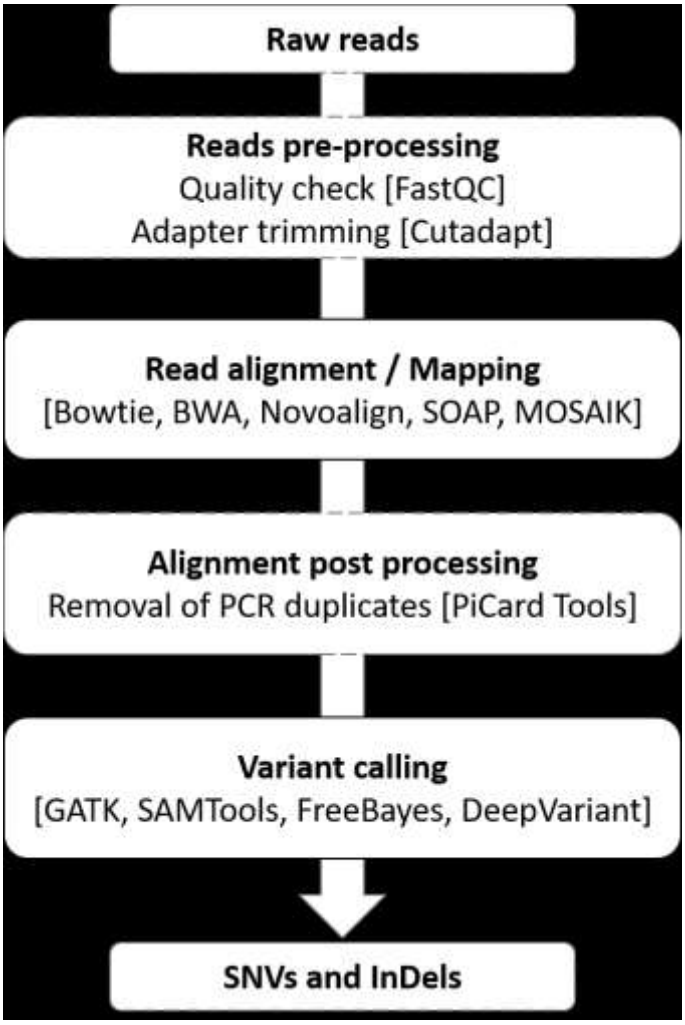

Figure.1 Schematic of NGS data analysis pipeline

The primary analysis step for all the pipelines include FastQC[1] to check the quality, Cutadapt[2] to remove adapter contaminates and low quality reads. The processed reads can be used to produce variant calls using WES shell scripts

Datasets:

FASTQ files of human exome NA12878, NA24385 and NA24631 can be downloaded from

NCBI-Sequence Read Archive (SRA- http://www.ncbi.nlm.nih.gov/sra).

The target region BED file can be downloaded from Agilent SureDesgin (http://earray.chem/agilent.com/suredesign, ELID: S0293689).

The human reference genomes GRCh37 and GRCh38 can be downloaded from the Ensembl[3] Download and extract the Reference file from here

Simulated Data:

ART toolkit [4] can be used to simulate the illumina-based paired-end read data. For example:

To generate paired-end reads of 150 bp length with the depth of 150X covering sequencing targets for Illumina HiSeq 2000 sequencing technology with 0.01 % error model for GRCh38

$ art_illumina -p Giab.sam -i seq_reference.fa -l 75 -f 20 -m 200 -s 10 -o d./outdir/dat_paired_end

High-confidence VCF and BED Download

The latest VCF and BED files with the high-confidence calls and regions can be obtained from the "latest" directory under each genome at the Genome in a Bottle FTP site:

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release

The BED file can be used to filter the highly accurate call set for WES.

Running the pipeline

$./your_pipeline.sh

The program will ask to give the location of your raw data and your reference sequence file

Example:

./BWA_DeepVaraint.sh

enter your Reference: hg37.fasta

enter the Read 1: sample1_R1.fastq.gz

enter the Read 2: sample1_R2.fastq.gz

Benchmarking Process:

The performance of variant detection by different pipelines can be compared statistically as, sensitivity = TP / (TP+FN), precision = TP / (TP+FP), false discovery rate (FDR) = FP / (TP+FP) and F-score = 2TP / (2TP+FP+FN) where, TP is true positive variant found in both GiaB validated dataset and data determined by pipeline; FP is false positive variant determined by pipeline but not validated by GiaB; FN is false negative variant, known as missing variant which is validated by GiaB but not determined by pipelines.

Using VCFtools, the statistical parameter can be calculated by giving a query VCF file for hg38 compared to GiB gold standard dataset as follows:

$vcf-compare -H Giab.vcf.gz BWA_Gatk.vcf.gz

Notice

Kindly download and install the following tools in your home directory before running the pipeline.

BWA

Bowtie2

Novoalign

SOAP

MOSAIK

PiCard Tools

GATK

SAMTools FreeBayes

Deepvariant

VCFtools

References:

1. Andrews S. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
2. Martin M (2011) Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnet Journal, 17, 10-12.
3. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, *et al*., Ensembl 2018, Nucleic Acids Research, 46(D1); D754–D761, 2018.
4. Huang W, Li L, Myers JR, Marth GT., ART: a next-generation sequencing read simulator. Bioinformatics, 28(4):593-4, 2012.