NCI CBIIT training

Exome sequencing analysis
Hands-on Tutorial

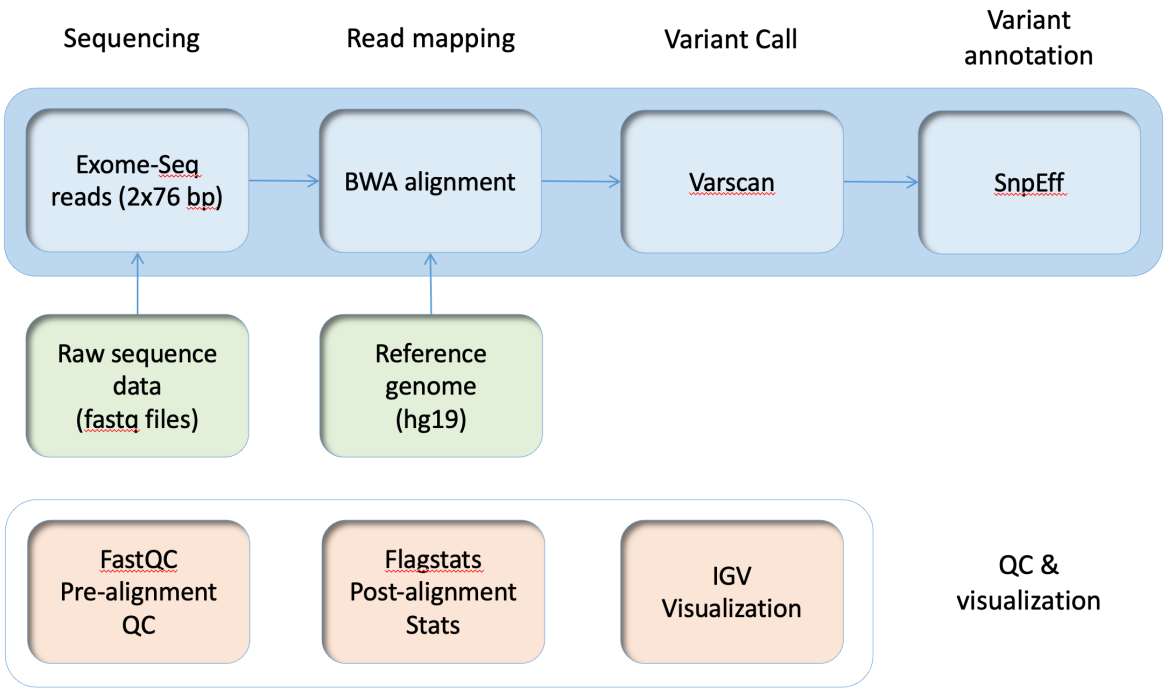
Qingrong Chen & Daoud Meerzaman

National Cancer Institute
Center of Biomedical Informatics and Information Technology

July 2023

Table of Contents

Outline of workflow



Get a Galaxy account

Galaxy is an open source, web-based platform for data intensive biomedical research (https://usegalaxy.org/).

a. Register for a galaxy account

Exome sample dataset

a. Illumina paired-end sequencing data (2x76bp)
   i. Human normal blood sample
      - demo_norm_r1.fastq (forward)
      - demo_norm_r2.fastq (reverse)
   ii. Human tumor sample
      - demo_tumor_r1.fastq (forward)
      - demo_tumor_r2.fastq (reverse)

b. Reference genome (use built-in hg19 reference genome)

Upload dataset

a. Log in Galaxy account
b. Upload exome dataset
    i. On the left panel 'Tools', select 'Get Data → Upload File'
    ii. Choose local file: go to the data folder and select all 4 '*.fastq' files
    iii. **Select Genome:** Human Feb. 2009 (GRCh37/hg19) (hg19)
    iv. **Select Type:** fastqsanger (for .fastq files)
    v. After files are uploaded, add a tag to each file
        a. Click on the dataset
        b. Click on galaxy-tags **Edit dataset tags**
        c. Add a tag starting with #normal or #tumor
        d. Check that the tag is appearing below the dataset name

Pre-alignment QC

FastQC (version 0.72) aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. We use FastQC to check sequenicng data quality.

a. Under 'NGS: QC and manipulation' and select 'FastQC'
b. Use Multiple (middle button) to select all fastq files
c. Output: Webpage & RawData

Alignment with BWA

BWA (version 0.7.17.4) is a software package for mapping sequences against a large reference genome, such as the human genome.

a. Under 'NGS: Mapping', select 'Map with BWA'
b. 'Will you select a reference genome from your history or use a built-in index?'
    i. Select 'Use a built-in genome index'
    ii. Use reference genome
        - Human (Homo sapiens) (b37): hg19
c. Select input type:
    i. Paired fastq
    ii. Select first set of reads (Forward, *_r1)
        - demo.tumor_r1.fast
        - demo.norm_r1.fastq
    iii. Select second set of reads (Reverse, *_r2)
        - demo.tumor_r2.fastq
        - demo.norm_r2.fastq

d. Output: Mapped reads in BAM format

Post-alignment Summary

Samtools flagstat (version 2.0.3) can be used to get a basic summary of an alignment for BAM dataset.

a. Under 'NGS: SAMtools' and select 'Flagstat'
b. Select all BAM files

Mark duplicates

MarkDuplicates (version 2.18.2.3) examines mapped reads in BAM files to locate duplicate reads.

a. Under 'NGS: Picard' and select 'MarkDuplicates'
b. Select all BAM files
c. If true do not write duplicates to the output file instead of writing them with appropriate flags set
    - Yes
d. Output: BAM files without duplicated reads

Variant detection

VarScan somatic (version 2.4.3.6): call germline/somatic and LOH variants from tumor-normal sample pairs.

a. Under 'NGS: Variant Analysis' and select 'Varscan'
b. Use a built-in genome
c. Reference genome:
    - Select "Human (Homo sapiens): hg19
d. Aligned reads from normal and tumor samples:
    - Select MarkDuplicates bam files for normal and tumor sample respectively
e. Run Tool
f. Output: VCF (variant call format) files

Variant annotation

SnpEff (version 4.3) is a genetic variant annotation and effect prediction toolbox.

a. Download a SnpEff database for GRCh37.75

b. Under 'NGS: Annotation' and select 'SnpEff Download'
c. Select the genome version you want to download: GRCh37.75
d. Execute

e. Under 'NGS: Annotation' and select 'SnpEff eff'
f. Sequence changes
    - Select Varscan output
g. Genome source
    - 'Custom snpEff database in your history'
    - SnpEff4.3 Genome Data
        - 'Downloaded SnpEff SnpEff4.3 GRCh37.75'

Visualize alignments and variants on IGV

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

a. Galaxy platform
    i. Select MarkDuplicate outputs under the current history
    ii. Download Dataset & Download bam_index

b. Open IGV to view the variants
    i. Launch IGV on the computer
    ii. Choose 'Human hg19' genome in the upper left corner
    iii. File -> Load from file -> Select the downloaded files to open

c. Example gene and regions
    i. TP53 gene
    ii. chr17:7,578,208 (Somatic)
    iii. chr17:7,579,472 (Germline)
    iv. chr17:7,573,057 (LOH)