# 

# DNA Sequencing Analysis Pipeline - BioCanRx

Requester: Sumedha Sudhaman ([sumedha.sudhaman@sickkids.ca](mailto:sumedha.sudhaman@sickkids.ca))

PI: Dr. Trevor Pugh (PMCC, UHN)

## 1. Pipeline Overview

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **ANALYSIS CATEGORY** | **TOOL** | **DESCRIPTION** | **INPUT** | **OUTPUT** |
| ✔ | Pre-processing | **BWA + Cocleaning** | PMGC Pipeline | FASTQ files | BAM files |
| ✔ | QC | **BAMQC** | PMGC Pipeline | BAM files | PDF reports |
| ✔ | QC | **Picard** |  | BAM files | TXT file |
| ✔ | Germline variants | **HaplotypeCaller** |  | BAM files | VCF |
| ✔ | Somatic SNV/Indel | **Mutect2** |  | Paired Tumour/Normal BAM files | VCF, MAF, and coverage file |
| ✔ | Somatic SNV/Indel | **Strelka** |  | Paired Tumour/Normal BAM files | VCF |
| ✔ | Somatic SNV/Indel | **Varscan2** |  | Mpileup output from Paired Tumour/Normal BAM files | VCF and .snp and .indel files |
| ✔ | Somatic SNV | **Mutect** |  | Paired Tumour/Normal BAM files | VCF, MAF, and coverage file |
| ✔ | Somatic Indel | **Indelgenotyper** |  | Paired Tumour/Normal BAM files | VCF |
| ✔ | Copy Number Analysis | **Varscan2** |  | Mpileup output from Paired Tumour/Normal BAM files | SEG file |
| ✔ | Variant Annotation | **Variant Effect Predictor** |  | VCF output from variant callers | VCF |

## 2. Environment Variables

|  |  |
| --- | --- |
|  | **Location** |
| **Genome ($REF)** | igenome-human/hg38 |
| **Interval File ($BED)** | **IMPORTANT: PLEASE use this BED/Interval file for the variant calling software as specified**  /cluster/projects/pughlab/references/S04380110\_Covered.headless\_hg38\_liftover\_without\_alt.bed |

## 3. Other Project Specific Notes

The **WES paired fastq data** is in the **Project raw data directory on h4h**.

The **Tumour\_normal\_matched\_pair\_list.xlsx** details the tumour and matched normal pairs.

**Notes:**

1. **For the tumour\_id\*.fastq that have paired end data from more than 1 lane, please concatenate all the R1s and R2s into a common R1 and R2 fastq pair.**
2. **Some of the tumours have the same matched\_normals.**
3. **One tumour sample (MD1452T2**) **does not have a matched normal. Please only call variants using haplotypecaller for this sample only, and vep annotate that sample.**
4. **Some fastqs are merged read pairs as .fq files.**

**Project raw data directory on h4h:** /cluster/projects/pughlab/external\_data/biocanrx/WES/fastq/ hg38\_batch2/

**For Tumours:** /cluster/projects/pughlab/external\_data/biocanrx/WES/fastq/hg38\_batch2/tumour/

**Matched\_normals:** /cluster/projects/pughlab/external\_data/biocanrx/WES/fastq/hg38\_batch2/normal/

**Project analysis directory on h4h:** /cluster/projects/pughlab/projects/biocanrx/WES/hg38\_batch2/

**4. Requested Directory Structure**

├── project\_dir

│   ├── haplotypecaller

│   ├── indelgenotyper

│   ├── mutect

│   ├── mutect2

│   ├── strelka

│   ├── varscan2

│   │   ├── copynumber

│   │   └── somatic

│   ├── VEP

| ├── ANNOVAR

│   ├── bam\_files

│   │   ├── merged\_processed\_bam (*include canonical chrs only*)

│   │   └── alt \_bams (*include all other bams*)

| | |--QC

## 5. Analysis Parameters

### 5.1 Germline Variant Calling

HaplotypeCaller (GATK v3.8)

|  |  |
| --- | --- |
| **Manual** | <https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php> |
| **Argument**  -T | HaplotypeCaller |
| -R | $REF |
| -stand\_call\_conf | 30 |
| -I | Sample.bam |
| -o | Sample\_name.raw.snps.indels.vcf |

### 5.2 Somatic Variant Calling

#### Mutect2 (GATK v3.8)

|  |  |
| --- | --- |
| **Argument**  -T | MuTect2 |
| -I:tumor | Tumour.bam |
| -I:normal | Normal.bam |
| -R | $REF |
| -o | ./MuTect2/ *Tumour\_Normal\_names*.MuTect2.vcf |
| --max\_alt\_allele\_in\_normal\_fraction | Default (0.3) |
| --max\_alt\_alleles\_in\_normal\_count | Default (2) |
| --max\_alt\_alleles\_in\_normal\_qscore\_sum | Default (20) |
|  |  |

#### Strelka (v1.0.14)

|  |  |
| --- | --- |
| **Manual** | <https://sites.google.com/site/strelkasomaticvariantcaller/> |
| **Argument**  --normalBam | Normal.bam |
| --tumorBam | Tumour.bam |
| --referenceFasta | $REF |
| --config | strelka\_config\_bwa.ini |
| --output-dir | ./Strelka/*Sample\_name*/myAnalysis |
|  |  |

#### Samtools mpileup (v0.1.18)

|  |  |
| --- | --- |
| **Manual** | <http://samtools.sourceforge.net/> |
| **Argument**  -q | 1 |
| -f | $REF |
| -I | $BED |
| *slot 3* | *Normal*.bam |
| *slot 4* | *Tumour*.bam |
| > | ./mpileup/*Tumour\_Normal*.mpileup |

#### Varscan2 Somatic (v2.4.2)

|  |  |
| --- | --- |
| **Manual** | <http://varscan.sourceforge.net/using-varscan.html> |
| **Arguments(to get default output)**  *Slot 1* (analysis type) | somatic |
| *Slot 2* (Input file) | Mpileup output from paired T/N |
| *Slot 3* (Output file) | ./Varscan2\_Somatic/Samplename.varscanSomatic |
| --mpileup | 1 |
| --min-var-freq | 0.02 |
| **Arguments(to get vcf output)**  *Slot 1* (analysis type) | somatic |
| *Slot 2* (Input file) | Mpileup output from paired T/N |
| *Slot 3* (Output file) | ./Varscan2\_Somatic/Samplename.varscanSomatic |
| --mpileup | 1 |
| --min-var-freq | 0.02 |
| --output-vcf | 1 |

#### Mutect (v1.1.5)

|  |  |
| --- | --- |
| **Manual** | <https://www.broadinstitute.org/cancer/cga/mutect_run> |
| **Argument**  --analysis\_type | MuTect |
| --reference\_sequence | $REF |
| --input\_file:normal | *Normal*.bam |
| --input\_file:tumor | *Tumour*.bam |
| --coverage\_file | ./Mutect/*Sample\_name*.wig.txt |
| --intervals | /cluster/projects/pughlab/references/S04380110\_Covered.headless\_hg38\_liftover\_without\_alt.bed |
| --vcf | ./Mutect/*Sample\_name*.vcf |
| --downsampling\_type | NONE |

#### Indelgenotyper (**GATK v3.0-0**)

|  |  |
| --- | --- |
| **Argument**  -T | IndelGenotyperV2 |
| -I:normal | Normal.bam |
| -I:tumor | Tumour.bam |
| -o | *Sample\_name*.indelgenotyper.vcf |
| -R | $REF |
| -metrics | *Sample\_name*.indelgenotyper.outmetric |
| -ws | 300 |
| --somatic |  |

### 5.3 Copy Number Analysis

#### Varscan2 Copynumber (v2.4.2)

|  |  |
| --- | --- |
| **Manual** | <http://varscan.sourceforge.net/using-varscan.html> |
| **Arguments(to get default output)**  *Slot 1* (analysis type) | copynumber |
| *Slot 2* (Input file) | Mpileup output from paired T/N |
| *Slot 3* (Output file) | ./Varscan2\_Somatic/Samplename.varscanCopyNumber |
| --mpileup | 1 |

### 5.4 Variant Annotation

#### Variant Effect Predictor (VEP v92) on perl/5.18.1

|  |  |
| --- | --- |
| **Manual** | <http://useast.ensembl.org/info/docs/tools/vep/script/index.html> |
| **variant\_effect\_predictor.pl** |  |
| **Argument**  **--fork** | 4 |
| --species | homo\_sapiens |
| **--offline** |  |
| --everything |  |
| --shift\_hgvs | 1 |
| --check\_existing |  |
| --total\_length |  |
| --allele\_number |  |
| --no\_escape |  |
| --xref\_refseq |  |
| --buffer\_size | 256 |
| --dir | /cluster/projects/pughlab/references/VEP\_fasta/92\_GRCh38 |
| --fasta | /cluster/tools/data/genomes/human/hg38/iGenomes/Sequence/WholeGenomeFasta/genome.fa |
| --input\_file | HaplotypeCaller|Mutect|Strelka|Varscan|Indelgenotyper.vcf |
| --force\_overwrite |  |
| --custom | /cluster/projects/pughlab/references/gnomad/gnomad\_v2.0.2/Exome/gnomad.exomes.r2.0.2.sites.vcf.gz,gnomAD,vcf,exact,0,AF\_POPMAX,AF\_AFR,AF\_AMR,AF\_ASJ,AF\_EAS,AF\_FIN,AF\_NFE,AF\_OTH,AF\_SAS |
| --vcf |  |
| --output\_file | ./VEP/*Sample\_name*.Caller.vep.vcf |

#### ANNOVAR (annovar/20150617)

|  |  |
| --- | --- |
| **Manual** | <http://annovar.openbioinformatics.org/en/latest/> |
| Usage | table\_annovar.pl *sample.vcf* humandb/ |
| **Argument**  -buildver | hg38 |
| -out | annovar\_sample\_name |
| -remove |  |
| -protocol | refGene,cosmic70,exac03, avsnp147, clinvar\_20160302, dbnsfp30a |
| -operation | g,f,f,f,f,f |
| -nastring | . |
| -vcfinput |  |
|  |  |