GenepoweRx™
Genetic Science powered by K&H

GenepoweRx®Bioinformatics for OncoRx Reporting

White Paper | Version 1.0 | 24-Feb-2023

Prepared by: K. Srinivas

Reviewed by:

Dr. Kalyan Ram Uppaluri
Managing Director

Vamsi Mohan Challa
CTO

Uppaluri K&H Personalized Medicine Clinic Pvt. Ltd. Hyderabad

Dr. Hima J Challa
Director

Dr. Kalyani P
CSO

GenepoweRx™
Genetic Science powered by K&H

Title: GenepoweRx®Bioinformatics for OncoRx Reporting

Authors: Kalyan Ram Uppaluri, Hima J. Challa, Vamsi Mohan Challa, Kalyani P, Shyam Sundar, Srinivas K

Introduction:

OncoRx is a directional test for tumor profiling enabling oncologists to predict response/resistance to targeted and immuno-therapies for cancer patients. It assists in predicting responses of FDA-approved drugs to approved biomarkers and facilitates better treatment management. **Memorial Sloan Kettering Cancer Centre (MSK),** the world's oldest and largest private cancer centre, and **GenepoweRx** (by K&H), partnered in collaboration to utilize MSK's clinical and research insights into gene mutations associated with solid tumors, along with K&H proprietary database to provide accurate recommendations for Indian population.

Structural variants (SVs) are DNA rearrangements that can profoundly affect evolution and human disease. Structural variants majorly occur in genomic variations and range from single nucleotide variants (SNVs) to more significant structural variants (SVs). The effect of the structural variants may lead to several mechanisms that can affect the protein-coding genes and cis-regulatory architecture. It affects more base pairs in the genome than SNVs (Single Nucleotide Variations), which can have a severe phenotypic impact. Some SVs are known to drive carcinogenesis, resulting in gene fusion and recurrent mutations observed in many pediatric cancers. At least 30% of cancer genomes are affected by a Structural pathogenic variation. SVs can be grouped into mutational classes, including insertion and deletion (copy number variations) and rearrangements, i.e., Inversions and Translocations.

Several tools are available for detecting copy number variations(CNVs) and structural variants from whole genome sequencing and Whole Exome Sequencing samples. But here, we use many tools to pre-process raw reads, SNPs, Indels, CNVs, and SVs detection.

Pre-requisites of raw cancer reads:

There are two types of sequencing reads Single - and paired-end reads. Single-end reads are helpful in some applications, such as small RNA sequencing, and paired-end reads are more accurate in reading alignment and detecting structural rearrangements.

Recommended Read lengths for various DNA Sequencing Applications:

- For the targeted panel, sequencing recommended read length (2 x 150 bps)
- For whole-genome and -exome sequencing, suggested read length (2 x 150 bp)
- for de novo sequencing, recommended read length (from 2 x 150 bp to 2 x 300 bp)

Sequencing Coverage:

An average number of reads that align to reference bases. The sequencing coverage level often determines whether variant discovery can be made confidently at particular base positions. At higher levels of coverage, each base is covered by a more significant number of aligned sequence reads.

Coverage is required for accurate base calling, although this can vary based on the accuracy of the sequencing platform, such as:

- For Whole Genome Sequencing, 30X to 50X coverage is required for human
- For Whole Exome Sequencing, 100X coverage is required
- For ChIP-Seq, 100X coverage is needed.
- For target sequencing, 500X or 1000X coverage is required.

NGS Read Length and Coverage:

Coverage depth refers to the average number of sequencing reads that align to, or "cover", each base in your sequenced sample.

The Lander/Waterman equation is a method for calculating coverage based on your read length(L), number of reads(N), and haploid genome length(G): C=LN/G.

For example, if we take one lane of single-read human sequences with v3 chemistry, we get:

C = (100bp) * (189x106)/(3x109) = 6.3

This tells us that each base in the genome will be sequenced between six and seven times on average.

Sequencing Quality Score:

Sequencing quality scores measure the probability that a base is called incorrectly. While sequencing, each base in a read is assigned a quality score by a phred-like algorithm.

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Pre-process of Raw Dataset:

Step-1:

Check the quality control of the raw reads using the FastQC tool. FastQC provides a simple way to do some quality control checks on raw sequence data from high throughput sequencing. A modular set analyses whether data has any problem or not before further analysis. The main functions of QC are:

Import of data from BAM, SAM, or FastQ files (any variant)

- provides a quick overview to tell in which areas of data there may be problems
- summary graph and tables to quickly assess database
- create a result for an HTML-based permanent report
- offline operation to allow automated generation of reports without running the interactive application

Step-2:

Check for adapters and remove adapter sequences from raw reads. Cutadapt finds and removes adapter sequences, primers, poly-A tails, and other types of unwanted sequences from high-throughput sequencing reads. Cutadapt helps with these trimming tests by finding the adapter or primer sequences in an error-tolerant way. It can modify and filter single-end and paired-end reads in various ways. Cutadapt can do a lot more in adding to removing adapters such as:

- Read modification options, this includes adapter removal, quality trimming, and read name modifications.
- Filtering options are applied, such as the removal of too short or untrimmed reads. Some of the filters also allow redirecting a read to a separate output file.
- It can detect multiple adapter types. e.g., Regular 3' or 5' adapter, Non-internal 3' or 5' adapter, Anchored 3' or 5' adapter, etc.

Step-3:

Mapping the reads against the reference genome, such as the human genome using a bowtie2 aligner. Bowtie sequence aligner was originally developed by Ben Langmead at the University of Maryland in 2009. The aligner is typically used with short reads and a large reference genome or for whole genome analysis. it is the ultrafast speed with an efficient memory that works best for aligning short sequences of DNA.

It has a Burrow-wheeler transform that can be implemented for increasing the speed of alignment by reducing the amount of main memory used by the program. Bowtie is very unlikely to report the correct alignment for a read whose true point of origin spans the three variants. A similar example involves gaps, if the subject genome has a gap with respect to the reference genome, Bowtie is unlikely to report the correct alignment for a read whose true point of origin spans the gap.

Step-4:

Converting SAM format file into Bam format file using Samtools view command. It is a package of programs for interacting with high-through sequencing data. Samtools view command is the most versatile tool in the Samtools package. It allows you to convert the binary alignments in the BAM file view to text-based SAM alignments that are easy for reading and process.

- It can count the total number of alignments
- inspect the header
- capturing the FLAG (Flag field in the BAM format encodes several key pieces of information regarding how an alignment aligned to the reference genome)

Step-5:

Sort the BAM format file using the Samtools sort command. When FastQ files align with the reference sequence, alignments produced are in random order with respect to their position in the reference genome. In other words, the BAM file is in the order that the sequence occurred in the FastQ files. While calling variants or visualizing alignments in IGV, requires further manipulation of BAM. It must be sorted such that alignments occur in "genome order". That is, ordered positional based upon their alignment coordinates on each chromosome.

Step-6:

Index the sorted bam file using the Samtools index command. Indexing a genome-sorted BAM file allows one to quickly extract alignments overlapping particular genomic regions. Moreover, indexing is required by genome viewers such as IGV so that the viewer can quickly display alignment in each genomic region to which you navigate.

GATK:

It is originally developed at Broad Institute. As it applies a variety of state-of-the-art statistical methods to accurately identify differences between the reads and the reference genome that are caused either by real genetic variants or by errors. Genome Analysis Toolkit is a collection of command-line tools for analyzing high-throughput sequencing data with a primary focus on variant discovery. The tools can be used individually or chained together into complete workflows.

Picard Tool:

Picard toolkit also belongs to Genome Analysis Toolkit (GATK) which means all the program from Picard is in Genome Analysis Toolkit also. So we choose to go with a completely GATK toolkit. But in the newer version of the GATK toolkit, some of the tools have been deprecated and new tools have been launched instead of deprecative tools.

Step7:

run GATK's MarkDuplicates tool to tag duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. These Duplicates can arise during sample preparation. This tool works by comparing sequences in the 5' prime positions of both reads and read-pairs in a SAM/BAM file. Tool main output is in SAM/BAM file. MarkDuplicates tool also produces a metrics file indicating the numbers of duplicates for both single- and paired-end reads.

Step8:

run GATK's AddOrReplaceReadGroups to assign all the reads in a file to a single new read-group. It assumes the presence of at least one RG tag, defining 'read-group' to which each read can be assigned. This tool accepts INPUT BAM and SAM files

Step9:

run 'samtools sort' to sort the readGroups bam file

Step10:

run 'samtools index' to index the readGroups-sorted file

Step11:

run GATK's BaseRecalibrator tool for the First pass of the base quality score recalibration. It generates a recalibration table based on various covariates. The default covariates are the read group. Here -ip stands for the amount of padding (in bp) to add each interval.

Step12:

run GATK's ApplyBQSR tool for the Second pass in a two-stage process called Base Quality Score Recalibration (BQSR). It recalibrates the base qualities of the input reads based on the recalibration table produced by the BaseRecalibrator tool.

Step13:

HAPLOTYPECALLER TOOL: run GATK's HaplotypeCaller tool to call the SNPs and indels simultaneously via local denovo assembly of haplotypes in an active region. If the program encounters a region showing signs of variation. It discards the existing mapping info and reassembles the reads in that region.

Step14:

run GATK's FilterVCF tool to apply one or more hard filters to a VCF file to filter out genotypes and variants.

Step 15:

MUTECT2 TOOL : run GATK's Mutect2 (TUMOR Mode only) runs on a single type of sample e.g. the tumor or the normal. Mutect2 uses a Bayesian somatic genotyping model that differs from the original MuTect and also uses the assembly-based machinery of HaplotypeCaller. Mutect2 also generates a stats file names a output.vcf.stats.

Step 16:

run Gatk's FilterMutectCalls to apply filters to Filter variants in Mutect2 VCF calls.

Step 17:

Variants annotation using OncoKB API

Annotates variants in MAF with OncoKB annotation. Supports both python2 and python3.

MafAnnotator.py: A Mutation Annotation Format (MAF) file is a tab-delimited text file that lists mutations.

MAFAnnotator annotated genes from MAF file by OncoKB Level of Evidences rules.

FusionAnnotator.py: Annotate fusions by OncoKB Level of Evidences rules.

CnaAnnotator.py: Annotate copy number alterations by OncoKB Level of Evidences rules.

ClinicalDataAnnotator.py: Annotate clinical data by OncoKB™ Level of Evidences rules.

OncoKBPlots.py: Draw OncoKB Actionability genes graph

References:

1. Van Belzen, IAEM, Schönhuth, A., Kemmeren, P. et al. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. npj Precis. Onc. 5, 15 (2021). https://doi.org/10.1038/s41698-021-00155-6
2. Collins, R.L., Brand, H., Karczewski, K.J. et al. A structural variation reference for medical and population genetics. Nature 581, 444–451 (2020). https://doi.org/10.1038/s41586-020-2287-8
3. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores.Genome Res. 2008;18:1851-8.
4. https://sapac.illumina.com/science/technology/next-generation-sequencing/plan-experiments.html.
5. Meldrum, C., Doyle, M. A., & Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. The Clinical biochemist. Reviews, 32(4), 177–195.
6. https://sapac.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_coverage_calculation.pdf

7. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988 Apr;2(3):231-9. doi: 10.1016/0888-7543(88)90007-9. PMID: 3294162.
8. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
9. http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf
10. MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. Available at: <http://journal.embnet.org/index.php/embnetjournal/article/view/200>. Date accessed: 08 july 2021. doi:https://doi.org/10.14806/ej.17.1.200.
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/pdf/nihms-366740.pdf