“SAR-COV-2 WGS analysis module”

Result summary from SAR-COV-2 whole genome variant analysis

Introduction

IBDC offers end-to-end analysis services for analysing covid-19 sequence samples. It provides an easy to run analysis pipeline for the users having no prior experience of programming/command-line data analysis tools. Users can submit their sequence datasets and IBDC servers will execute the analysis 'on-the-fly' and provide results. The analysis pipeline will identify variants in the submitted samples and then compare them with known covid 19 variants to identify any any novel variant in the sample. The variants would be annotated based on the information available at https://covid-miner.ifo.gov.it/app/home GISAID [https://www.gisaid.org/]. The results consist of alignment files, variant calling files, variant screening and annotation files. All result files would be available within a compressed folder named as : ***UserID_JobID*_Results1.tar.gz.**

Analysis of NGS raw reads (fastq format)

This folder has following sub-folders and files:

Sub-Folder 1: DATAQC This folder has following two files:

File 1: '***Sample-name'*_R1_fastqc.html:** Sample QC data for the 'forward-sequence' file of the paired-end data.

File 2: ***'Sample-name'*_R2_fastqc.html** Sample QC data for the 'reverse-sequence' file of the paired-end data.

(Only one file will be present if single-end data is submitted)

***Sub-Folder 2*: FINAL_RESULTS:** This folder has following seven files:

File 1: ***'Sample-name'_CLEAN.Consensus.fasta.fa*** is the consensus sequences generated based on alignment of the NGS reads with the reference genome.

File 2 : ***'Sample-name'_VariantStats.txt*** is the final result file that will provide a list of all identified variants along with the variant status as 'Known_Variant' and 'Unknown Variant' based on the comparison with curated variant data from Covid-miner and Gear-19 database. Each variant is provided with chromosomal coordinates, reference allele. alternate allele, variant effect information, amino acid change information and gene name.

File 3: ***'Sample-name'_CLEAN.coverage.txt*** contains the clean read count, genome coverage, mean depth, mean sequence quality, mean mapping quality.

File 4: ***'Sample-name'_ CLEAN.duprem.bam*** contains the sorted and duplicate removed alignment file in bam format.

File 5: **'Sample-name'.Nextclade.lineage_report.tsv** contains the clade information of the SAR-COV-2 sample.

File 6: ***'Sample-name'.PANGO.lineage_report.csv*** contains the pangolin lineage.

*File7:****'Sample-name'*****_CLEAN_WGS-SAR-COV-2_Circos.png** contains the circos representation of the SAR-COV-2 sample.

Circos Labels

Track 1 (outer track) GC skew

Track 2 (middle track) Alignment depth [heatmap color scheme spectral-7-div]

Track 3 (inner track) Variants [scattered plot where each dot represents a variant]

Analysis of assembled sequences (fasta format)

Sub-Folder 1**: FINAL_RESULTS:** This folder has following five files:

File 1: ***'Sample-name'.fasta :*** The Assembled sequences as submitted by the user.

File 2 : ***'Sample-name'*****_*****VariantStats.txt*** is the final result file that will provide a list of all identified variants along with the variant status as 'Known_Variant' and 'Unknown Variant' based on the comparison with curated variant data from Covid-miner and Gear-19 database. Each variant is provided with chromosomal coordinates, reference allele. alternate allele, variant effect information, amino acid change information and Gene name.

File 3: **'Sample-name'.Nextclade.lineage_report.tsv** contains the clade information from nextclade.

File 4: ***'Sample-name'.PANGO.lineage_report.csv*** contains the pangolin lineage.

File 5: ***'Sample-name'*****_WGS-SAR-COV-2_Circos.png** contains the circos representation of the SAR-COV-2 sample.

Circos Labels.

Track 1 (outer track) GC skew
Track 2 (inner track) Variants [scattered plot where each dot represents a variant]

****