JUL 10, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.4r3l2244xl1y/v1

Protocol Citation: Ana C Reis, Daniela Pinto, Mónica V. Cunha 2023. Bioinformatic workflow for the analysis of SARS-CoV-2 Spike sequencing raw data from wastewater. **protocols.io** https://dx.doi.org/10.17504/protocols.io.4r3l2244xl1y/v1

Protocol status: Working
We use this protocol and it's working

Created: Jul 10, 2023

Last Modified: Jul 10, 2023

PROTOCOL integer ID:
84767

Bioinformatic workflow for the analysis of SARS-CoV-2 Spike sequencing raw data from wastewater

Ana C Reis[1], Daniela Pinto[1], Mónica V. Cunha[1]

[1]Centre for Ecology, Evolution and Environmental Changes (cE3c) & CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

Daniela Pinto
Centre for Ecology, Evolution and Environmental Changes (cE3...

ABSTRACT

This protocol describes the bioinformatics procedure to analyse SARS-CoV-2 Spike Illumina sequencing data from wastewater. This workflow can be applied to other sequencing datasets to generate high-quality map to reference results.

Create work directory and organize raw data

1 Import and organize the fastq.gz raw data files.

Examine read quality

2 FastQC is a quality control tool for high throughput sequence data which assesses multiple metrics and provides a QC report. FastQC (Galaxy version 0.11.9) is used for preliminary read quality assessment.

Download: https://github.com/s-andrews/FastQC
Reference: Andrews, S. (n.d.). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimming of read sequences and re-examine read quality

3 Trimmomatic is a flexible read trimming tool for Illumina NGS data. Trimmomatic (Galaxy version 0.38) is used to remove adapter sequences (-ILLUMINACLIP) and to remove poor quality nucleotides (-SLIDINGWINDOW:4:30). FastQC (Galaxy version 0.11.9) is used again to check if read metrics have been improved.

Download: https://github.com/usadellab/Trimmomatic
Reference: Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, 30(15):2114-20. doi: 10.1093/bioinformatics/btu170

Align reads with reference genome

4 Burrow-Wheeler aligner (BWA) is a software package for mapping DNA sequences against a large reference genome, and has three different algorithms that can be applied to perform the alignment.
The trimmed reads are aligned with SARS-CoV-2 reference genome (NCBI NC_045512.2) using BWA with MEM algorithm (BWA-MEM) (Galaxy version 0.7.17), with the default algorithm parameters.

Download: https://github.com/bwa-mem2/bwa-mem2
Reference: Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2

Filtering BAM file

5 BAM filter (Galaxy version 0.5.9) is used to clean the resulting BAM files by removing reads smaller than 35 base pairs.

Calling variants

6 iVar variants is a bioinformatics tool used to call variants - single nucleotide variants (SNVs) and INDELs (insertions and deletions) from aligned BAM files. This tool is able to identify codons and translate variants into amino acids. Two parameters can be user-defined: (i) the minimum quality, which is the minimum quality for a base to be counted towards the ungapped depth to calculate the variant frequency at a given position; and (ii) minimum frequency, which is the minimum frequency required for a SNV or INDEL to be reported.

The iVAR variants (Galaxy version 1.4.2) is used to call SNVs and INDELs, with the minimum quality score threshold to count a base set to 20, and the minimum frequency threshold set to 0.03. The output is converted to a VCF file.

Download: https://github.com/andersen-lab/ivar
Reference: Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., Jesus, J. G. D., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Rompay, K. K. A. V., Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., & Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, *20*(1):8. doi: 10.1186/s13059-018-1618-7

Variants annotation

7 SnpEff is a variant annotation and effect prediction tool, therefore it annotates and predicts the effects of genetic variants (such as amino acid changes). The SARS-CoV-2 variants are annotated with SnpEff (Galaxy version 4.5covid). The database with genome SARS-CoV-2 (NCBI NC_045512.2) as reference genome is used.

Download: https://pcingola.github.io/SnpEff/se_introduction/
Reference: Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 6(2), 80–92. doi: 10.4161/fly.19695

Assessing alignment quality metrics

8 Quality metrics of the alignment (number of reads, coverage, GC-content, etc.) is assessed with QualiMap Bam QC (Galaxy version 2.2.2).

Download: http://qualimap.conesalab.org/
Reference: Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics, 32(2):292-

4. doi: 10.1093/bioinformatics/btv566
García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics, 28(20), 2678–9. doi: 10.1093/bioinformatics/bts503

Result analysis

9 The following criteria are implemented to validate SNPs and INDELs for posterior analysis:

- Major variants (>50% alteration): minimum of reads set to 10
- Minor variants (between 0.3 and 49% alteration): minimum of reads set to 20, and the minimum of reads with alteration set to 3

SARS-CoV-2 lineage determination

10 The identification of SARS-CoV-2 lineages and the determination of their relative abundance is performed with Freyja, a bioinformatics tool designed to work with mixed SARS-CoV-2 samples. The file with reads aligned to reference genome, obtained after step four of this workflow, is used as input. Freyja measures the SNV frequency and sequencing depth at each position in the genome, and then returns an estimate of the true lineage abundances in the sample, based on the outbreak.info curated lineage metadata file that summarizes lineages by WHO designation.

Download: https://github.com/andersen-lab/Freyja

SARS-CoV-2 mutation identification

11 Mutations associated to variants of concern (VOC) other than Omicron, mutations shared by several VOC, mutations with frequencies inferior to 50%, and mutations with no variant and/or lineage association are highlighted.
The occurrence of these mutations and their association with SARS-CoV-2 variants and lineages may be confirmed at outebreak.info (https://outbreak.info/). Outebreak.info congregates a series of tools designed to explore COVID-19 and SARS-CoV-2 data with variant surveillance reports, data on cases and deaths, and a standardized, searchable research library. Outebreak.info uses SARS-CoV-2 sequences deposited at GISAID Initiative for the generation of mutation reports, however it removes from its database sequences with collection dates specifying only the year or dates in the future; sequences with lengths ≤ 20,000 base pairs; sequences with more than 5,000 nucleotide substitutions; and sequences with a contiguous deletion spanning > 500 base pairs.