

## **ViralFlow: an automated workflow for SARS-CoV-2 genome assembly, lineage assignment, mutations and intrahost variants detection**

Filipe Zimmer Dezordi<sup>1\*</sup>, Túlio de Lima Campos<sup>2</sup>, Pedro Miguel Carneiro Jeronimo<sup>3</sup>, Cleber Furtado Aksenén<sup>4</sup>, Suzana Porto Almeida<sup>5</sup>, Gabriel Luz Wallau <sup>6\*</sup> on behalf of Fiocruz COVID-19 Genomic Surveillance Network

<sup>1</sup> Instituto Aggeu Magalhães, Departamento de Entomologia e Núcleo de Bioinformática, Fiocruz, Campus da UFPE - Av. Prof. Moraes Rego s/n, 50670-420, Recife, Brazil.

<https://orcid.org/0000-0003-3259-3456>

<sup>2</sup> Instituto Aggeu Magalhães, Núcleo de Bioinformática, Fiocruz, Campus da UFPE - Av. Prof. Moraes Rego s/n, 50670-420, Recife, Brazil. <https://orcid.org/0000-0003-0446-848X>

<sup>3</sup> Oswaldo Cruz Foundation (Fiocruz), Branch Ceará, Eusebio 61760-000, Fortaleza, Ceará, Brazil. <https://orcid.org/0000-0002-2509-5822>

<sup>4</sup> Oswaldo Cruz Foundation (Fiocruz), Branch Ceará, Eusebio 61760-000, Fortaleza, Ceará, Brazil. <https://orcid.org/0000-0002-5297-75>

<sup>5</sup> Oswaldo Cruz Foundation (Fiocruz), Branch Ceará, Eusebio 61760-000, Fortaleza, Ceará, Brazil. <https://orcid.org/0000-0002-2991-9585>

<sup>6</sup> Instituto Aggeu Magalhães, Departamento de Entomologia e Núcleo de Bioinformática, Fiocruz, Campus da UFPE - Av. Prof. Moraes Rego s/n, 50670-420, Recife, Brazil.

<https://orcid.org/0000-0002-1419-5713>

Corresponding authors:

\*zimmer.filipe@gmail.com

\*gabriel.wallau@fiocruz.br

## ABSTRACT

The COVID-19 pandemic, a disease caused by the Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), emerged in 2019 and quickly spread worldwide. Genomic surveillance has become the gold standard methodology to monitor and study this emerging virus. The current deluge of SARS-CoV-2 genomic data being generated worldwide has put additional pressure on the urgent need for streamlined bioinformatics workflows for data analysis. Here, we describe a workflow developed by our group to process and analyze large-scale SARS-CoV-2 Illumina amplicon sequencing data. This workflow automates all the steps involved in SARS-CoV-2 genomic analysis: data processing, genome assembly, PANGO lineage assignment, mutation analysis and the screening of intrahost variants. The workflow presented here (<https://github.com/dezordi/ViralFlow>) is available through Docker or Singularity images, allowing implementation in laptops for small scale analyses or in high processing capacity servers or clusters. Moreover, the low requirements for memory and CPU cores makes it a versatile tool for SARS-CoV-2 genomic analysis.

## INTRODUCTION

The emergence (Wu et al., 2020) and rapid spread of the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus which causes the Coronavirus Disease 2019 (COVID-19), and the subsequent establishment of the COVID-19 pandemic (WHO, 2020), triggered a global effort to sequence and identify the circulating SARS-CoV-2 lineages. This effort resulted in the rapid availability of more than 2 million genomes into the EpiCoV™ database hosted on GISAID in July 2021 (Shu & McCauley, 2017), representing more than 1,700 lineages described on PANGO lineages (O’Toole et al., 2021).

A range of molecular biology methods have been developed to diagnose SARS-CoV-2 infections, such as RT-qPCRs, RT-LAMP, immunoassays or Sanger sequencing (Bezerra et al., 2021; da Silva et al., 2020; Nörz et al., 2021). However, only whole genome sequencing can provide enough genetic information for reliable lineage discrimination and characterization of genome wide mutation patterns that are necessary for the characterization of variants of concern (VOCs) (Lauring & Hodcroft, 2021). Amplicon-based Next generation sequencing (NGS) has become the gold standard methodology for SARS-CoV-2 genome sequencing (Charre et al., 2020), but such abundance of sequencing data from hundreds/thousands of samples also brings new challenges at bioinformatic analysis. At the moment, the Centers for Disease Control and Prevention (CDC) official git repository contains eighteen sources of bioinformatics tools to deal with different SARS-CoV-2 sequencing data (SARS-CoV-2 Sequencing Resources, 2020). However, a single validated workflow that is able to perform several key genomic analyses such as: data quality check, genome assembly, virus lineage assignment, mutation description and intrahost variants variability analysis by sample is still lacking.

In this work, we describe a workflow currently in use by the Fiocruz COVID-19 Genomic Surveillance Network (Naveca et al., 2021; Paiva et al., 2020; Resende et al., 2021). It was developed to work with paired-end Illumina amplicon sequencing reads, and is focused on both pre- and post-genomic analysis. It was designed to attend research groups with diverse computational structures such as: personal computers and multi-user servers through the containerization of the workflow with Docker (<https://www.docker.com/>) and Singularity (<https://sylabs.io/singularity/>).

## MATERIAL AND METHODS

### Workflow structure

The workflow was developed within an Ubuntu 20.04.2 LTS Docker environment ([https://hub.docker.com/\\_/ubuntu](https://hub.docker.com/_/ubuntu)) and is composed of five major steps to analyse SARS-CoV-2 Illumina paired-end amplicon sequencing data (**Figure 1a**): quality control; consensus generation; intrahost variant analysis; virus lineage assignment and mutation analysis; and Assembly metrics analytics, and can be used in different computational environments (**Figure 1b**).

The quality control step is performed with fastp v.0.20.1 tool (Chen et al., 2018), where the paired-end reads are trimmed using a minimum read quality threshold of Phred score 20. The adapters or the PCR primers and the minimum length threshold for trimmed reads should be parsed by the user. In addition to the paired-end treated data, this step provides an html with information of pre and post-treatment steps (**Figure 1c**).

The generation of the consensus genome is performed using the reference-guided assembly strategy. In this step a mapping of paired-end libraries against a reference genome is performed with BWA v.0.7.17 (Li & Durbin, 2009), we recommend the SARS-CoV-2 reference genome Wuhan-Hu-1, NCBI refseq NC\_045512.2 code. The Samtools v.1.9 (Li et al., 2009) and iVar v.1.3.1 (Grubaugh et al., 2019) tools are used to correct recovery the SNPs and indels were two consensus are generated: one with the majority allele present into every nucleotide position along the genome (ivar consensus -t 0) and another version with ambiguous nucleotide characters, in cases where majority intrahost Single Nucleotide Variants (iSNVs) encompass up to 60% of allele frequencies (ivar consensus -t 0.6). Only mapped bases with quality equal to or greater than 30 (-q 30) were used into iVar counts. Minimum depth threshold to consider a position with supported intrahost variants should be parsed by the user.

Considering that iVar did not provide an option to generate consensus harboring all iSNVs with two or more alleles found in low frequency - and the consensus genome with minor iSNVs is essential to understand the effect of intrahost variants - we developed an *in house* python script (intrahost.py) which uses the allele frequencies per position output of bamreadcount v.0.8.0 (Khanna et al., 2021) to recognize only the positions with 2 or more alleles and generate a consensus harboring all minor supported alleles. To avoid the recognition of sequencing artifacts such as intrahost variants signals, genomic positions where selected if: (i) the minor allele frequency represented at least 5% of the total allele depth; (ii) had at least 100 reads of depth; and (iii) the minor allele nucleotides were supported by reads of both senses (at least 5% of depth should come from each read sense). In the running default, to detect iSNVs present at the minimum frequency of 5% an average depth of 2000 reads is required, for analysis with lower average depth the minimum depth threshold can be parsed.

The virus lineages signature is performed with Pangolin, the Pangolin and mutations constellation are updated at the moment of the Docker or Singularity images creation - to avoid outdated analysis - in interactive containers, the pangolin --update is strongly recommended. The consensus quality and set of mutations were evaluated with nextclade v.0.14.2. If the analyzed sample showed intrahost signals, the aforementioned analyses are performed for both consensus versions (with major and minor allele frequencies). On the other hand, the analysis will be performed only for the consensus genome with major iSNVs supported by nucleotides with major allele frequencies. In the last step, the assembly metrics such as depth and coverage were extracted with bamdst v.1.0.6 (<https://github.com/shiquan/bamdst>).

## Different environments, the same workflow

Considering the infrastructure and computational experience heterogeneity of different research groups working with SARS-CoV-2 genomic data we evaluated our workflow, and included documentation for two use cases (**Figure 1b**):

- Case I: Using a personal computer installing all dependencies; or using Docker or Singularity container services.
- Case II: Using a multi-user computational server, installing all dependencies; or using Docker or Singularity container services.

In this work, we simulate cases I and II. Case I was run on a personal laptop with the following configurations: Ubuntu 20.04.2 LTS, 02 x RAM 4Gb DDR4 2133MHz and CPU Intel® Core™ i7-6500U 2.50GHz x 4, using 2 threads in the interactive Docker container. For Case II, we used a computational server with the following configurations: A node with 191Gb of RAM DRAM 2933MHz and two CPUs Intel(R) Xeon(R) Gold 5220R CPU @ 2.20GHz totalizing 96 threads, using 24 threads for the Singularity container runs. All command-line examples can be found in the documentation at the GitHub repository. Three independent runs were performed for each case.

## Data availability

The performance of the workflow was accessed using two datasets. The first one is a public dataset of 86 Brazilian SARS-CoV-2 Illumina paired-end libraries, generated by amplicon sequencing method using the Illumina COVIDSeq protocol, available under the EMBL-EBI study accession PRJEB47823, and the second one is an artificial dataset (**Supplementary file 1**) created with ART (Huang et al., 2012) of five paired-end libraries simulating a simultaneous infection (coinfection/codetection) of different SARS-CoV-2 lineages into a single sample (samples information on **Table S1**). The code and workflow documentation is available on <https://github.com/dezordi/ViralFlow>. And the supplementary data is available on [<https://doi.org/10.6084/m9.figshare.16695262.v1>].

## RESULTS AND DISCUSSION

Ninety-one genomic sequencing samples were assembled in both environments (see Methods, Case I and Case II). The total analysis running time - artificial samples are discarded of this estimation - in Case I was 9,38 hours (**Table 1**) with a mean time by sample equals to 6,7 minutes (stdev = 1,07, for details, see **Table S2**), while for Case II it took 3,06 hours, with a mean of 2,15 minutes (stdev = 0,17) for each sample. The use of RAM memory was 0,66 Gb and 0,73 Gb in Case I and Case II, respectively, this number was similar owing to the fact that all tools used in our workflow had their computational process based on CPU and not in RAM memory. A sequencing run with 96 samples (approximately 1.04Gb of throughput and 10,9 million reads) can be fully analyzed into 3,44 hours using a server with 24 threads or into 10,72 hours into a personal computer with 2 threads.

We detected a low number of iSNV, from 0 to 2, with a mean of 0 (stdev = 0.43, for details, see **Table S3**), for the eighty-six “non-artificial” samples supporting published estimates of low intrahost variant variability of SARS-CoV-2 (Naveca et al., 2021; Shen et al., 2020). For the five coinfection/codetection artificial samples, the workflow was able to consistently detect a large amount of well supported iSNVs (47 iSNVs per sample) (**Figure 2a, Table S4**). Those results show the capacity of the workflow to rapidly detect and generate a range of useful information that are important to generate new insights such as single consensus and coinfection of different SARS-CoV-2 lineages into a single sample. Moreover, the intrahost multi-allele frequencies can also be used to detect sample contamination, in a scenario where most samples show the same intrahost pattern, the more plausible explanation is sequence contamination rather than the origin of the same iSNVs in different samples.

Two key pieces of information are necessary for the deployment of SARS-CoV-2 outbreak control strategies: the virus lineage assignment and mutation characterization (Lauring & Hodcroft, 2021; van Oosterhout et al., 2021). The workflow generates two tabular files for each sample containing those key information: the ‘.pango.csv’ and the ‘.nextclade.csv’, which provide the virus lineage and the mutations found, respectively. The lineages identified in the 86 non-artificial samples (**Figure 3**) correspond to the set of mutations and to the expected lineages circulating at the collection date in each location (**Table S5**). For the 5 artificial samples, the virus lineages and set of mutations correspond to the allele frequencies present in iSNVs multi-allele frequencies (**Figure 2b, Table S5**). To show the precision of our workflow to detect the *indel* regions, 5 random non-artificial samples were assigned as P.1 with the deletion 11288-11297 into ORF1a and the insertion into the intergenic region at 28262 positions were manually curated with IGV (**Supplementary file 2**).

In addition to the intrahost, lineage, and mutations information generated by different tools present in our workflow, the ‘.quality.html’ generate by fastp and ‘coverage.report’ and ‘chromosomes.report’ files generated by bamdst that can be used for feedback to wet lab staff and to assess the quality of mapping and assembly steps. This information can also be crossed with the information of ‘qc.overallScore’ ‘qc.overallStatus’ columns present on ‘.nextclade.csv’ table.

Considering all the aforementioned results, the proposed workflow can be a good choice for groups that work with Illumina paired-end data and needs rapid information about both the quality of sequencing experiments, quality of consensus genome, and the lineage and mutation profiles that directly feed epidemiological reports.

## AUTHOR STATEMENTS

### Acknowledgments

We thank the LACEN-PE staff for providing the samples of SARS-CoV-2 and the Technological Platform Core for the support with their research facilities, the Fiocruz

COVID-19 Genomic Surveillance Network for testing and source of information to compare results. We also thank all the researchers around the world that are working and generating data of SARS-CoV-2 in those difficult times.

### **Author contributions**

GLW and FZD conceived the study. GLW and TLC planned and supervised the work. FZD and GLW developed the workflow. FZD, PMCJ, CFA, SA tested and made improvements into workflow code. FZD carried out the tests, analysis and wrote the manuscript. TLC and FZD developed the containerization structures. All authors agreed with the final version of the manuscript.

### **Funding**

This study was supported by the National Council for Scientific and Technological Development by the productivity research fellowship level 2 for Wallau GL (303902/2019-1) and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

### **Ethical Aspects**

This study was approved by the Aggeu Magalhaes Institute Ethical Committee—CAAE 32333120.4.0000.5190.

## **REFERENCES**

- Bezerra, M. F., Machado, L. C., De Carvalho, V. do C. V., Docena, C., Brandão-Filho, S. P., Ayres, C. F. J., Paiva, M. H. S., & Wallau, G. L. (2021). A Sanger-based approach for scaling up screening of SARS-CoV-2 variants of interest and concern. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 92, 104910.
- Charre, C., Ginevra, C., Sabatier, M., Regue, H., Destras, G., Brun, S., Burfin, G., Scholtes, C., Morfin, F., Valette, M., Lina, B., Bal, A., & Josset, L. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution*, 6(2), veaa075.
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890.
- da Silva, S. J. R., Silva, C. T. A. da, Guarines, K. M., Mendes, R. P. G., Pardee, K., Kohl, A., & Pena, L. (2020). Clinical and Laboratory Diagnosis of SARS-CoV-2, the Virus Causing COVID-19. *ACS Infectious Diseases*, 6(9), 2319–2336.
- Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Van Rompay, K. K. A., 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.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4), 593–594.

Khanna, A., Larson, D. E., Srivatsan, S. N., Mosior, M., Abbott, T. E., Kiwala, S., Ley, T. J., Duncavage, E. J., Walter, M. J., Walker, J. R., Griffith, O. L., Griffith, M., & Miller, C. A. (2021). Bam-readcount—Rapid generation of basepair-resolution sequence metrics. arXiv:2107.12817 [q-bio]. <http://arxiv.org/abs/2107.12817>

Lauring, A. S., & Hodcroft, E. B. (2021). Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA*, 325(6), 529–531.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079.

Naveca, F. G., Nascimento, V., de Souza, V. C., Corado, A. de L., Nascimento, F., Silva, G., Costa, Á., Duarte, D., Pessoa, K., Mejía, M., Brandão, M. J., Jesus, M., Gonçalves, L., da Costa, C. F., Sampaio, V., Barros, D., Silva, M., Mattos, T., Pontes, G., ... Bello, G. (2021). COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nature Medicine*, 27(7), 1230–1238.

Nörz, D., Grunwald, M., Olearo, F., Fischer, N., Aepfelbacher, M., Pfefferle, S., & Lütgehetmann, M. (2021). Evaluation of a fully automated high-throughput SARS-CoV-2 multiplex qPCR assay with built-in screening functionality for del-HV69/70- and N501Y variants such as B.1.1.7. *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, 141, 104894.

O’Toole, Á., Hill, V., Pybus, O. G., Watts, A., Bogoch, I. I., Khan, K., Messina, J. P., COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, Tegally, H., Lessells, R. R., Giandhari, J., Pillay, S., Tumedi, K. A., Nyepetsi, G., Kebabonye, M., Matsheka, M., Mine, M., ... Kraemer, M. U. G. (2021). Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. *Wellcome Open Research*, 6, 121.

Paiva, M. H. S., Guedes, D. R. D., Docena, C., Bezerra, M. F., Dezordi, F. Z., Machado, L. C., Krovovsky, L., Helvecio, E., da Silva, A. F., Vasconcelos, L. R. S., Rezende, A. M., da Silva, S. J. R., Sales, K. G. da S., de Sá, B. S. L. F., da Cruz, D. L., Cavalcanti, C. E., Neto, A. de M., da Silva, C. T. A., Mendes, R. P. G., ... Wallau, G. L. (2020). Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. *Viruses*, 12(12), 1414.



Resende, P. C., Naveca, F. G., Lins, R. D., Dezordi, F. Z., Ferraz, M. V. F., Moreira, E. G., Coêlho, D. F., Motta, F. C., Paixão, A. C. D., Appolinario, L., Lopes, R. S., da Fonseca Mendonça, A. C., da Rocha, A. S. B., Nascimento, V., Souza, V., Silva, G., Nascimento, F., Gonçalves Lima Neto, L., Riediger, I., ... Wallau, G. L. (2021). The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the Spike protein.

SARS-CoV-2 Sequencing Resources. (2020). Centers for Disease Control and Prevention. [https://github.com/CDCgov/SARS-CoV-2\\_Sequencing](https://github.com/CDCgov/SARS-CoV-2_Sequencing).

Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., Guo, L., Zhang, G., Li, H., Xu, Y., Chen, M., Gao, Z., Wang, J., Ren, L., & Li, M. (2020). Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clinical Infectious Diseases*, 71(15), 713–720.

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13).

van Oosterhout, C., Hall, N., Ly, H., & Tyler, K. M. (2021). COVID-19 evolution during the pandemic—Implications of new SARS-CoV-2 variants on disease control and public health policies. *Virulence*, 12(1), 507–508.

WHO Director-General’s opening remarks at the media briefing on COVID-19—11 March 2020. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265–269.

## FIGURES AND TABLES LEGENDS

**Table 1.** Analysis time of a dataset with 91 samples with different computational structure.

**Figure 1.** Workflow scheme. **a.** The workflow structure with the five steps. **b.** The workflow can be configured to work on diverse computational environments. **c.** Several outputs generated by the workflow per sample.

**Figure 2.** MAFs frequency sites of artificial datasets simulating co-infection events. The Black dashed line represents the expected minor iSNVs frequencies in each artificial dataset. **a.** The frequencies of 4 artificial datasets. **b.** The frequencies and associated mutations of minor and major consensus genomes.

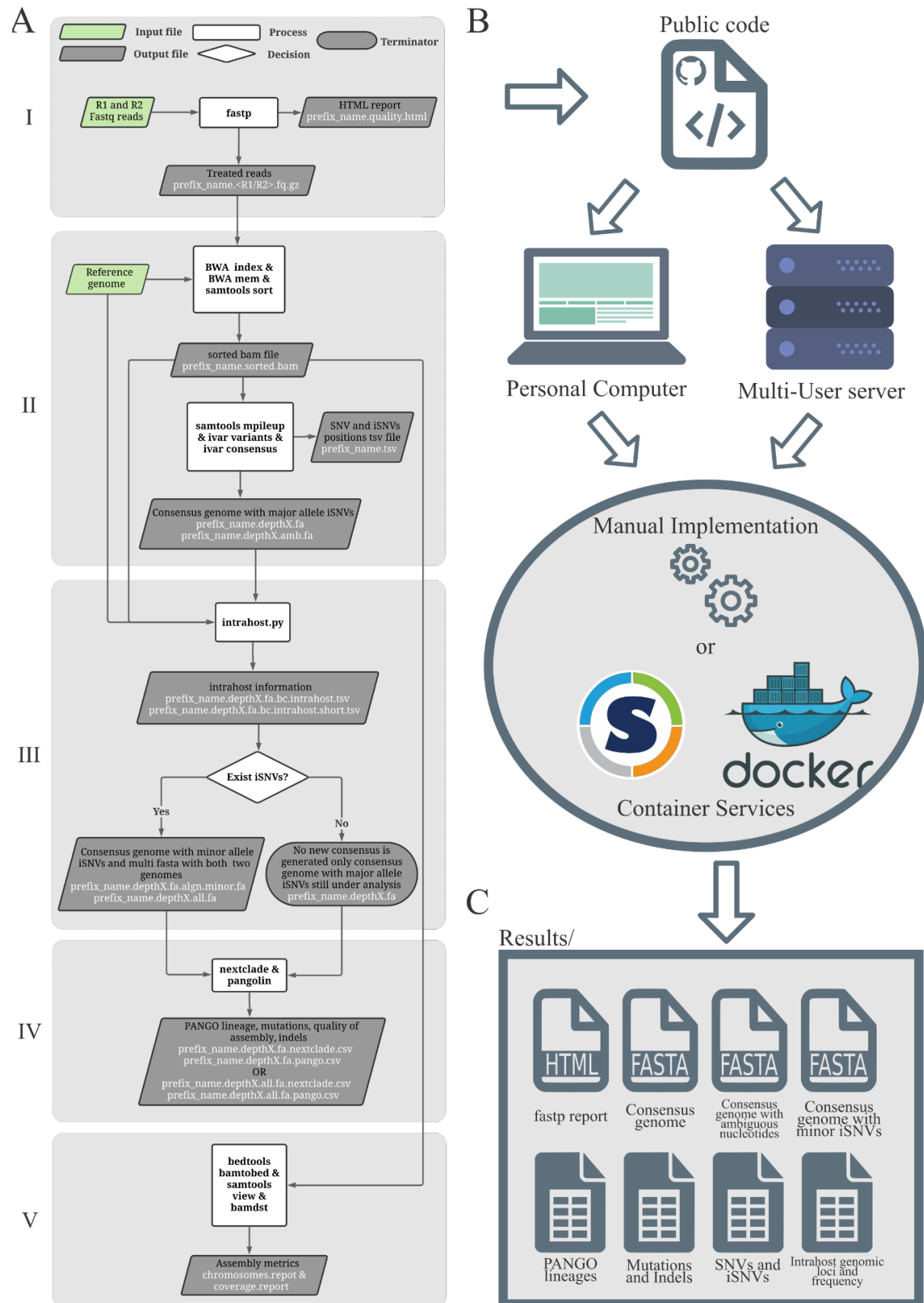
**Figure 3.** Lineages identified in consensus genomes with lineages found in the 86 samples dataset from Pernambuco state. Data recovered from GISAID at [seplag.pe.gov.br](http://seplag.pe.gov.br) at 30 Aug. 2021.

**Table 1.**

<b>Cases</b>	<b>RAM usage</b>	<b>Time per sample*</b>	<b>Total time*</b>
<b>Case I**</b>	0.66	2.15296 (0.17)	183.60319 (0.022)
<b>Case II***</b>	0.73	6.734235 (1.07)	563.98072 (7.11)

\* Mean time and standard deviation in minutes; \*\* Personal computer; \*\*\* Computational server.

**Figure 1.**





**Figure 3.**

