ASPICov: an Automated Pipeline for Identification of SARS-Cov2 nucleotidic Variants

Valentin Tilloy[1,2], Pierre Cuzin[3], Laura Leroi[3] , Patrick Durand[3] and Sophie Alain[1]

[1]Centre National de Référence des Herpèsvirus, CHU Dupuytren, 87000, Limoges, France
[2]UF9481 Bioinformatique, CHU Dupuytren, 87000, Limoges, France
[3]IFREMER-IRSI-Service de Bioinformatique, Centre Bretagne, 29280 Plouzane, France

Corresponding Author: valentin.tilloy@unilim.fr

Introduction

Whole-genome sequencing (WGS) is used for clinical surveillance of SARS-Cov2 in order to detect emerging mutations, facilitate epidemiological studies and anticipate possible therapeutic / vaccinal escape.

After DNA extraction, SARS-Cov2 can be captured and/or amplified using different commercial kits such as Artic, Swift, Thermofisher,... depending on the viral load and nature of the sample. Obtained DNA will be then prepared into specific libraries according to sequencing technology. Raw-reads will be then generated by sequencing machines.

In this context, we have developed a pipeline dedicated to identify SARS-Cov2 mutations from reads of a broad range of samples (clinical, wastewaters,...).

Figure 1 : Steps undertaken to get raw-reads, input files for ASPICov pipeline.

Methodology

ASPICov is a **Nextflow** [1] pipeline developed to provide a rapid, reliable and complete analysis of NGS SARS-Cov2 samples.

ASPICov produces helpful results for biologic interpretation such as quality reports, VCF files, consensus sequences and various plots.

A succession of commonly used tools combined to an optimized configuration is a key for the robustness of the pipeline. Our workflow performs many steps as described in figure 2.

In order to ensure **FAIR data analysis**, the workflow follows nf-core guidelines and use **Singularity** [2] containers to wrap tool environments.

ASPICov has been optimised using a dataset of a single sample (Basa strain isolated from a patient at Limoges hospital) diluted several time (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7}) and also taken at different culture stages (passages P3, P4 and P7). This step was important for the setting up of VCF filters allowing both to identify rare mutations and exclude false positives.

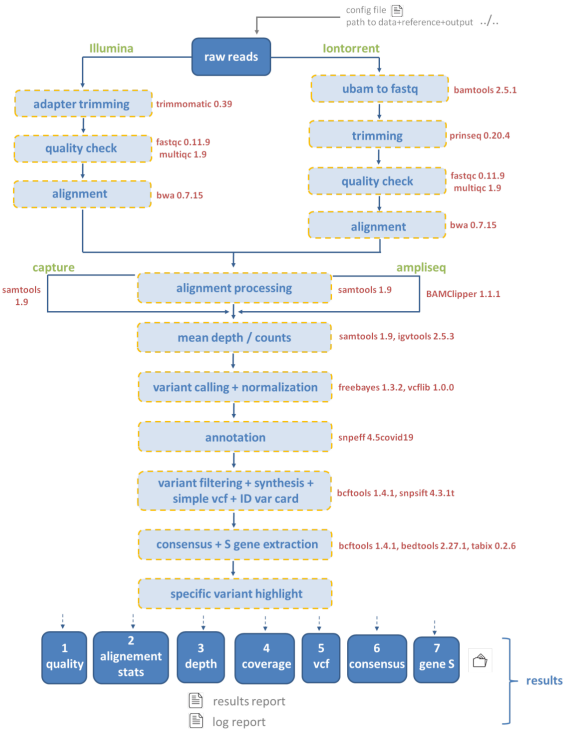

Figure 2 : Overview of ASPICov steps and software implemented.

Results

We have screened ENA and SRA public databases to get a dataset of SARS-Cov2 reads coming from different labs using different strategies and sequencing technologies. All VOC and VOI were found using ASPICov workflow demonstrating its efficiency and accuracy.

	WHO label	Pango lineage	GISAID clade/lineage	Nextstrain clade	Earliest documented samples	Notable mutation	Biosample	Information	Mutations found with ASPICov
Variants of concern	Alpha	B.1.1.7	GRY (formerly GR/501Y.V1)	20I/S:501Y.V1	UK, Sep-2020	69–70del, N501Y, P681H	SAMN18219363	amplicon COVIDSeq / Illumina NovaSeq	✓
	Beta	B.1.351	GH/501Y.V2	20H/S:501Y.V2	South Africa, May-2020	K417N, E484K, N501Y	SAMN17767551	amplicon Artic / Illumina MiSeq	✓
	Gamma	P.1	GR/501Y.V3	20J/S:501Y.V3	Brazil, Nov-2020	K417T, E484K, N501Y	SAMN18527803	? / Illumina MiSeq	✓
	Delta	B.1.617.2	G/452R.V3	21A/S:478K	India, Oct-2020	L452R, T478K, P681R	SAMN19157798	amplicon Artic / Illumina	✓
Variants of interest	Epsilon	B.1.427 / B.1.429	GH/452R.V1	20C/S.452R	USA, Mar-2020	L452R	SAMN19223936	amplicon Artic / Illumina Nextera	✓
	Zeta	P.2	GR	20B/S.484K	Brazil, Apr-2020	E484K	SAMN18606318	amplicon Artic / Illumina NovaSeq	✓
	Eta	B.1.525	G/484K.V3	20A/S484K	Multiple countries, Dec-2020	E484K, F888L	SAMEA7763427	amplicon / Illumina NovaSeq	✓
	Theta	P.3	GR	20B/S:265C	Philippines, Jan-2021	Y265L	data not found		
	Iota	B.1.526	GH	20C/S:484K	USA, Nov-2020	E484K	SAMN19009806	amplicon Artic / Illumina NovaSeq	✓
	Kappa	B.1.617.1	G/452R.V3	21A/S:154K	India, Oct-2020	L452R, E484Q, P681R	SAMN19375761	amplicon Thermo / S5 Iontorrent	✓
	Lambda	C.37	GR/452Q.V1	20D	Peru, Aug-2020	L452Q	SAMN19224075	amplicon Artic / Illumina Nextera	✓
Other variant		B.1.1.207			Nigeria, Aug-2020	P681H	SAMN17717137	? / Illumina TruSeq	✓

Table 1 : SARS-Cov2 variants of concern and interest. Some samples from public databases have been tested using ASPICov workflow to see if variants can be found. Updated on 22 june 2021. Source : https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

Perspectives

Several new features are currently under development and other will be elaborated soon:

- Global html report
- Phylogenetic analysis
- Integration of MinION and MGI sequencing technologies
- Percentage of genotypes found
- PANGO lineage determination (Pangolin)
- Nextclade / Gisaid data comparison

References

[1] Di Tommaso, P. et al, 2017, Nature Biotechnology
[2] Kurtzer G. et al, 2017, PLoS One

This pipeline (source code, documentation and full list of tool dependencies) is available at:
https://gitlab.com/vtilloy/aspicov