Grant agreement No. 101046314

END-VOC

ENDING COVID 19 VARIANTS OF CONCERN THROUGH COHORT STUDIES: END-VOC

HORIZON-HLTH-2021-CORONA-01-02

Deliverable 3.2

Report on the results of the Bioinformatics pipeline

WP 3 – Sequencing and phylogenetics

Due date of deliverable	Month 11 – March 2023
Actual submission date	01/03/2023
Start date of project	01/05/2022
Duration	36 months
Lead beneficiary	UCL
Last editor	Damien Richard (UCL)
Contributors	François Balloux, Lucy van Dorp

Dissemination Level		
PU	Public	X
SEN	Sensitive	

END-VOC has received funding from the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

Copyright

History of the changes

Version	Date	Released by	Comments
0.2	08.01.2023	D. Richard	First draft
0.4	12.02.2023	Lucy van Dorp	Edits
0.6	28.02.2023	François Balloux	Deliverable version

Table of contents

Definitions and acronyms

Acronyms	**Definitions**
NGS	Next generation sequencing
SNP	Single nucleotide polymorphism

1. Summary of activities and research findings

Bioinformatics pipeline overview v.1.0

Multiple research groups are involved in SARS-CoV-2 surveillance and sequencing within the END-VOC consortium. As such, it is critical to effectively integrate datasets produced by distinct project partners. Datasets can be highly heterogenous in nature and their bioinformatics analysis requires expertise and involves numerous processing steps that can be time consuming to assemble within a single full analysis pipeline.

This process is largely alleviated by available all-in-one analysis workflows previously developed by multiple groups, which facilitate the analysis of heterogeneous samples without much programming skills needed. A list of such solutions was previously collated and assessed in deliverable 3.1 (WP3).

In this deliverable 3.2 "bioinformatics pipeline", we provide a 'best-practice' case study example. Given that sequence data generation and sharing is behind schedule, we compiled a highly heterogeneous sequencing data of ~850 SARS-CoV-2 samples using publicly shared SARS-CoV-2 raw reads from the NCBI Short Read Archive (SRA) repository.

We analysed this heterogeneous dataset using the nf-core/viralrecon pipeline, an all-in-one workflow, which takes raw RNA reads as input and generates consensus sequences for each sample, along with numerous detailed reports including quality metrics and lineage information.

We then perform a standard phylogenetics reconstruction to highlight a standard downstream computational analysis that can be performed on sequence alignments of consensus genomes. This second part of the report will be based on the R framework.

The chosen pipeline performed well despite the highly heterogeneous nature of the raw sequence data selected. We are confident we can perform robust medium- to high-throughput standard bioinformatics and computational genetics analyses on datasets combining genomic data produced on different sequencing platforms and sequenced at variable depths.

Detailed annotation of all the steps as well as all the scripts needed to reproduce the analysis are available on the Github platform at https://github.com/END-VOC/WP3_D3.2_SARS-CoV-2_NGS_bioinformatics_pipeline/.

2. Conclusions and future steps

The present document and associated Github page are intended to provide an example of the concomitant analysis of heterogeneous datasets, which can represent a challenge within the END-VOC consortium. Such analysis can be performed in many different ways and should be adapted according to the needs of the study.