A computational effient system for improving SARS-CoV-2 sequencing surveillance

Marcello Salvi, Chiara Colantuono, Antonio Grimaldi, Francesco Panariello, Patrizia Annunziata, Andrea Ballabio, Davide Cacchiarelli

Genome Med ***14****, 90 (2022). https://doi.org/10.1186/s13073-022-01098-8*

INTRODUCTION

Genomic surveillance of SARS-CoV-2 is the only approach to rapidly monitor and tackle emerging variants of concern (VOC) of the COVID-19 pandemic. Such scrutiny is crucial to limit the spread of VOC that might escape the immune protection conferred by vaccination strategies or previous virus exposure. It is also becoming clear now that efficient genomic surveillance would require monitoring the host gene expression to identify prognostic biomarkers of disease progression.

OBJECTIVE

To guarantee a time efficient and cost-effective protocol useful for national health system we ideate an appraoch suitable for any lab with a benchtop sequencer and a limited budget to be applied, allowing an integrated genomic surveillance on premises.

METHODS

Our effort allowed us to establish novel procedures to prioritize emerging variants and identify molecular signatures associated with a viral infection, maturing a powerful tool for disease prevention and diagnosis.
From the end of December 2020 to the first week of 2022, we sequenced, uploaded to GISAID, and analyzed 17,193 SARS-CoV-2 genomes. Short time is essential to perform epidemiological surveillance and track the spread of new SARS-CoV-2 variants. To overcome the bottleneck of bioinformatic analysis, we use Nextflow to orchestrate the vast amount of data produced, store it, and analyze it in a time-effective and computationally efficient way within the GCP environment exploiting parallel computing. We set up a workflow that allows us to process more than 400 samples per day by our wet scientists and a bioinformatic pipeline to process all samples in less than 5 hours.

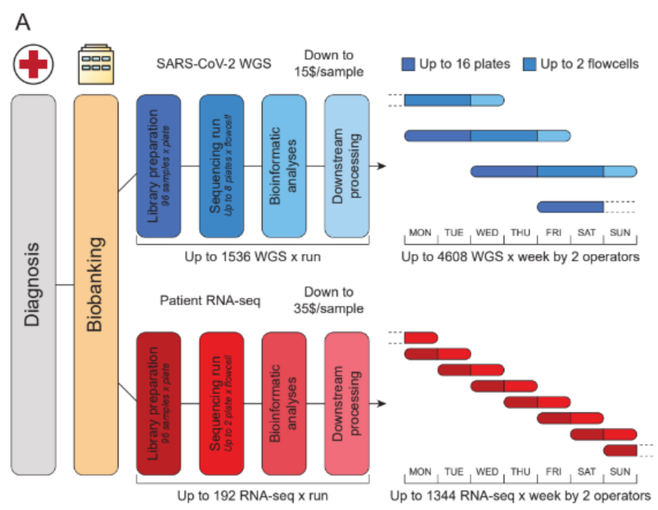

Figure 1. A systematic approach allows the generation of large and robust genomic data in short time. A) Schematic representation of the workflow set up to collect, process and analyze a considerable number of viral genomes and patient transcriptomes in one week. Healthcare centres (grey box) perform oronasopharyngeal swabs to diagnose the presence of SARS-CoV-2 genome in patients. Biobanks (yellow box) collect and store the extracted RNA that is then processed and analyzed by genomic centres to perform SARS-CoV-2 WGS (blue boxes) and RNA-seq (red boxes).

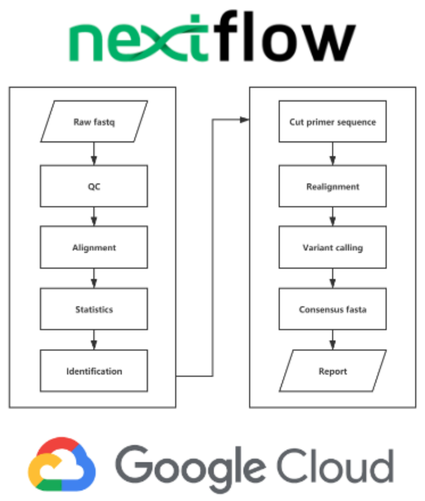

Figure 2. The implemented Nextflow workflow for SARS-CoV-2 genotyping for MGI sequencing data

RESULTS

From the end of December 2020 to the first week of 2022, we sequenced, uploaded to GISAID, and analyzed 17,193 SARS-CoV-2 genomes. Nextflow allows us to orchestrate the vast amount of data produced, store it, and analyze it in a time-effective and computationally efficient way within the GCP environment exploiting parallel computing. Our workflow has been tested throughout the Campania region, which includes the major southern Italian metropolitan areas and some of the most densely inhabited cities in Europe. Globally, in most months during 2021, we were able to sequence at least 5% of all COVID-19 positive samples, making Campania compliant with EC/ECDC recommendations.

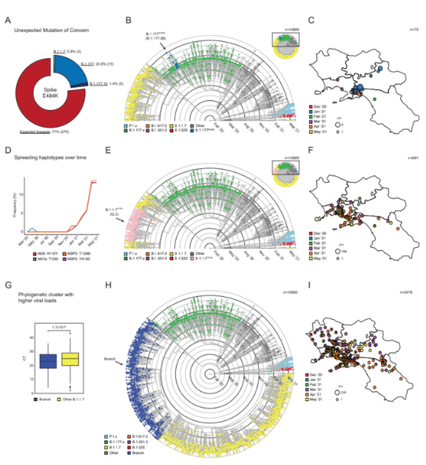

Figure 3. High-Throughput genomic surveillance allows the identification of new SARS-CoV-2 lineages. A) Donut chart representing the amount of analyzed genomes presenting the Spike E484K mutation. B) Section of the phylogenetic tree representation of the whole dataset, colored by lineages. C)Geographic distribution of genomic variants belonging to the identified lineage. D) Line plot showing the frequency trend of the selected mutations in time. E) Section of the phylogenetic tree representation of the whole dataset, colored by lineages. F) Geographic distribution of genomic variants belonging to the identified lineage. G) Boxplot showing the CT distribution in B.1.1.7 variants, divided by their phylogenetic position inside (dark blue) or outside (yellow) the identified branch. H) Phylogenetic tree representation of the whole dataset, colored by lineages. The identified branch is reported (arrow, dark blue dots). I) Geographic distribution of genomic variants belonging to the identified branch, colored by the collection date.

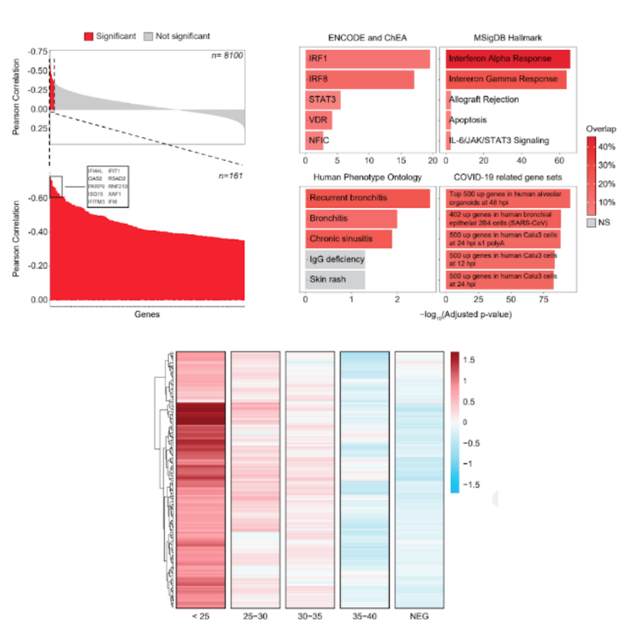

Figure 4. Transcriptional profiling of Sars-CoV-2 infected patients reveals a gene signature correlated with viral load.
A) Correlation analysis between CTs and gene expression, performed on 8100 genes, is shown as a barplot. For each gene (x axis), its correlation value (y axis) and significance (p-value <0.0001, red) is reported. Bottom: highlight of the significant results. (161 genes). The top 10 most anti-correlated genes are reported (black box). B) Pathway and gene set enrichment analysis performed for different databases using the gene signature previously identified. Each barplot shows the significance (x axis) and the percentage of overlap (fill color) between the input signature and the tested public genesets. C) Heatmap of z-scored, log2-transformed and normalized gene counts for the 161 significantly correlated genes from (A). Values have been averaged in 4 groups of samples depending on the CT (x axis) or whether they were negative.

Since the comparative analysis of our dataset with GISAID world data allowed us to retrospectively identify viral variants firstly sampled in Campania, we were interested to explore whether it was possible to unveil new viral lineages circulating in the territory. To achieve this goal, we studied SARS-CoV-2 “mutations of concern” genotyped in unexpected lineages. Interestingly, we found that the Spike E484K substitution had an unanticipated high incidence rate.

CONCLUSION

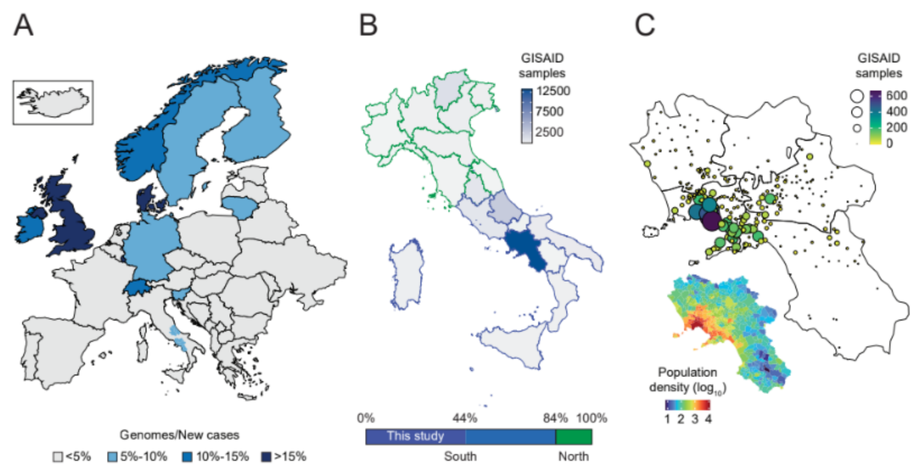

Figure 5. Characterization of SARS-CoV-2 genome evolution in the south of Italy.
A) Geographic map representing European States, colored by thenumber of 2021 months with at least 5% of viral genomes compared to new cases. B) Top: geographic map representing Italian regions, colored by the number of genomes deposited on the GISAID platform. C) Geographic distribution in Campania of the genomes analyzed in this study (top)relative to the population density (bottom).

The proposed strategy allows to scale viral genome sequencing down to 10 times less per sample. In addition this protocol minimizes the hands-on time and does not require intensive training or any particular automation. Taken altogether, these features allowed us to profile the SARS-CoV-2 pandemic in Campania (Italy) during 2020-2021. We thus identified the main variants leading each infection wave in the regional territory and discovered 3 new SARS-CoV-2 lineages specifically originated in Campania, demonstrating the potential of genomic surveillance.

ACKNOWLEDGEMENTS

We are thankful for the support of TIGEM High Content Screening and Bioinformatics Cores and Next Generation Diagnostic srl. The IZSM and Ospedale dei Colli for sampling and biobanking.