Improved SARS-CoV-2 sequencing surveillance allows the identification of new variants and signatures in infected patients

Antonio Grimaldi[1†], Francesco Panariello[1†], Patrizia Annunziata[1,2†], Teresa Giuliano[1], Michela Daniele[1,2], Biancamaria Pierri[3], Chiara Colantuono[1,2], Marcello Salvi[1,2], Valentina Bouché[1], Anna Manfredi[1,2], Maria Concetta Cuomo[3], Denise Di Concilio[3], Claudia Tiberio[4], Mariano Fiorenza[5], Giuseppe Portella[5], Ilaria Cimmino[5,6], Antonio Sorrentino[6], Giovanna Fusco[3], Maria Rosaria Granata[6], Pellegrino Cerino[3], Antonio Limone[3], Luigi Atripaldi[4], Andrea Ballabio[1,5,7,8,9†] and Davide Cacchiarelli[1,5,9*†]

Abstract

Background: Genomic surveillance of severe acute respiratory syndrome coronavirus 2 (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 of the host gene expression to identify prognostic biomarkers of treatment efficacy and disease progression. Here we propose an integrative workflow to both generate thousands of SARS-CoV-2 genome sequences per week and analyze host gene expression upon infection.

Methods: In this study we applied an integrated workflow for RNA extracted from nasal swabs to obtain in parallel the full genome of SARS-CoV-2 and transcriptome of host respiratory epithelium. The RNA extracted from each sample was reverse transcribed and the viral genome was specifically enriched through an amplicon-based approach. The very same RNA was then used for patient transcriptome analysis. Samples were collected in the Campania region, Italy, for viral genome sequencing. Patient transcriptome analysis was performed on about 700 samples divided into two cohorts of patients, depending on the viral variant detected (B.1 or *delta*).

Results: We sequenced over 20,000 viral genomes since the beginning of the pandemic, producing the highest number of sequences in Italy. We thus reconstructed the pandemic dynamics in the regional territory from March 2020 to December 2021. In addition, we have matured and applied novel proof-of-principle approaches to prioritize possible gain-of-function mutations by leveraging patients' metadata and isolated patient-specific signatures of SARS-CoV-2 infection. This allowed us to (i) identify three new viral variants that specifically originated in the Campania

†Antonio Grimaldi, Francesco Panariello, and Patrizia Annunziata share equal contribution.

†Andrea Ballabio and Davide Cacchiarelli share equal lead contribution.

*Correspondence: d.cacchiarelli@tigem.it

[1] Telethon Institute of Genetics and Medicine (TIGEM), Armenise/Harvard Laboratory of Integrative Genomics, Pozzuoli, Italy
Full list of author information is available at the end of the article