SARS-CoV-2 genome sequencing and analysis: enabling population level management of COVID-19

Rand Dubis[1], Suzanne Drury[1], Nick Lench[1], Alistair Johnson[1], Piero Dalle Pezze[1], Tautvydas Petkus[2], Andrius Daunoravičius[2], Andrew Bayliss[1], Daniel Griffiths[1], Matthew Nelson[1] and Anthony Rogers[1]

1. Congenica Ltd, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton, CB10 1DR, UK

2. Telesoftas, Savanoriu pr. 178, Kaunas, Republic of Lithuania

Introduction

Since its emergence in December 2019, cases of coronavirus disease (COVID-19) have surged around the world, having major public health ramifications in over 200 countries. In addition to testing populations for the presence of the SARS-CoV-2 virus, the World Health Organization recommends viral genome sequencing and international data sharing to help identify and track potentially more virulent strains.

Congenica was approached to provide expert assistance to rapidly scale a national SARS-CoV-2 sequencing programme. Whilst the partner had implemented a test and trace system, bioinformatics and next generation sequencing (NGS) support was required to monitor the evolution and transmission of the virus.

Objective

Implement scalable NGS workflows for SARS-Cov-2 genome sequencing, integrated with an analytical pipeline, to 1) enable identification of strains and evolution, 2) identify and monitor infection outbreaks and 3) determine local transmission levels in comparison with cases associated with migration activity.

Methods

RNA extracted from SARS-CoV-2 positive cases was reverse-transcribed using New England Biolabs LunaScript RT SuperMix (NEB). 400nt amplicons targeting the SARS-CoV-2 genome were produced using ARTIC v3 nCov-2019 primers. Libraries were prepared using NEBNext Ultra II DNA Library Prep (NEB) and sequenced on Illumina MiSeq.

A Nextflow pipeline was built to automate data analysis using the nCoV-2019 coronavirus bioinformatics protocol developed by the ARTIC network. The data was assigned a lineage using the Phylogenetic Assignment of Named Global Outbreak LINeages (Pangolin) COVID-19 lineage assigner. The data is then shared with the Nextstrain pipeline for real-time tracking of pathogen evolution.

Results

An optimised standard operating procedure was established for the rapid scale-up of viral genome sequencing. Congenica also built an integrated analytical pipeline to allow automatic processing from sequence data to report. Data is recorded in a database to enable chronological reporting and downstream analysis. Initial runs of the pipeline allowed rapid sequencing of over 100 samples,

identifying many different viral strains originating from around the world and established a baseline SARS-CoV-2 population for comparison.

Conclusion

We have developed a scalable, integrated laboratory and bioinformatics workflow and analytical pipeline to identify existing and new strains of SARS-Cov-2 that can be used at a population level to monitor infection and transmission. Furthermore, the analysis pipeline provides accurate strain distribution data that could help identify geographic clusters of strains indicative of growing local outbreaks. This system can be readily deployed to countries lacking in available resources that require a rapid solution for SARS-Cov-2 genome sequencing and analysis.