As part of the COVID-19 Virtual Biohackathon 2020 we formed a working group to create COVID-19 PubSeq, a Public Sequence Resource for SARS-CoV-2 virus sequences. Our goal was to create a repository that had a low barrier to entry for uploading and analyzing sequence data, without imposing any restriction on their utilization. We followed FAIR data principles: data are published with public domain (CC0) or creative commons 4.0 (CC-BY-4.0) license. structured metadata is validated against standard ontologies and, importantly, reproducible workflows are executed after the upload in order to provide up-to-date results rapidly and in standardized data formats.

Data and analysis tools together

Existing data repositories don’t enforce strict quality control on the submitted data and its metadata, and don’t add value in terms of running additional analysis. In addition, some databases have licenses that place restrictions on the data utilization.

- Data is validated for being in a supported data format, and for not being duplicated entries in the resource.
- Structured metadata is strictly validated against standard ontologies.

COVID-19 PubSeq

http://covid19.genenetwork.org

On COVID-19 PubSeq the data, metadata, and analysis tools live together, publicly and freely.

Pangeneome generation workflow

Each time someone uploads a valid sample, it is immediately combined with all the already uploaded SARS-CoV-2 genomes in order to generate an up-to-date SARS-CoV-2 pangeneome as input for Pantograph, an interactive visualization of pangeneomes.

Graphical representation of a SARS-CoV-2 pangeneome of 100 genomes realized with vgteam/odgi.

State-of-the-art standards

COVID-19 PubSeq leverages state-of-the-art standards and technologies.

- CWL is designed to meet the needs of data-intensive science, describing analysis workflows and tools
- CWL workflows are executed on the cloud after the upload in order to provide up-to-date results rapidly and in standardized data formats.
- Reproducible and scalable CWL workflows can be loaded into any RDF triple-store. The image shows the pangenome region (the first row in orange), the reference (NC_045512, the second row in green), and don't add value in terms of running.

Graphical representation of a SARS-CoV-2 pangeneome of 100 genomes realized with vgteam/odgi.

COVID-19 Integrated knowledge base

Metadata can be downloaded as Turtle RDF which can be loaded into any RDF triple-store. The Swiss Institute of Bioinformatics has included the COVID-19 PubSeq data in https://covid-19-sparql.expasy.org.

Information on a specific sample

PREFIX pubseq: <http://biohackathon.org/bh20-seq-schema#MainSchema/>
PREFIX sio: <http://semanticscience.org/resource/>
SELECT DISTINCT ?predicate ?object
{}
?sample sio:SIO_000115 "MT326090.1" .

References