COVID-19 PubSeq: Public SARS-CoV-2 Sequence Resource

Andrea Guarracino[5], Peter Amstutz[2], Thomas Liener[3], Michael R. Crusoe[4], Adam Novak[6], Erik Garrison[6], Tazro Ohta[7], Bonface Munyoki[1], Danielle Welter[8], Sarah Zaranek[2], Alexander (Sasha) Wait Zaranek[2], Pjotr Prins[1]

[1]Department of Genetics, Genomics and Informatics, The University of Tennessee Health Science Center, Memphis, TN, USA., [2]Curii Corporation Boston, MA, [3]Independent, [4]Department of Computer Science, Faculty of Sciences, Vrije Universiteit Amsterdam, The Netherlands, [5]Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy, [6]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA, [7]Database Center for Life Sciences, Tokyo, Japan, [8]Luxembourg Centre for Systems Biomedecine, University of Luxembourg, Luxembourg

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**.

Reproducible and **scalable** CWL workflows are executed on the cloud **after the upload** in order to provide **up-to-date results** rapidly and in **standardized** data formats.

http://covid19.genenetwork.org

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

State-of-the-art standards

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

Findable Accessible Interoperable Reusable

The FAIR Guiding Principles[1]

Ontobee — http://www.ontobee.org

COMMON WORKFLOW LANGUAGE — http://commonwl.org

creative commons — https://creativecommons.org

OLS — Ontology Lookup Service

GitHub — https://github.com

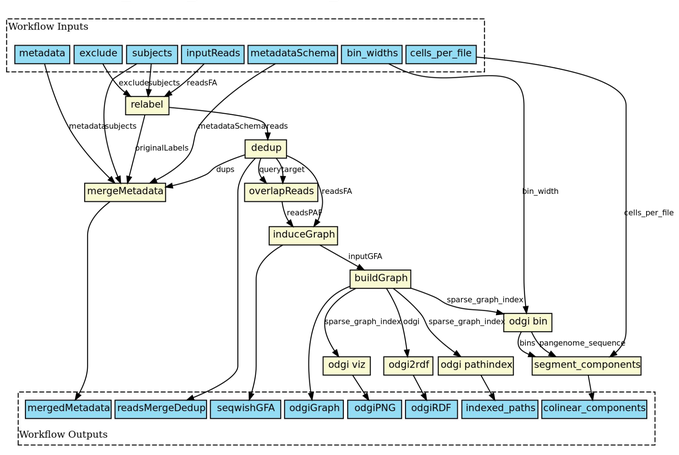
COVID-19 PubSeq is running on Arvados, a cloud open source platform for managing, processing, and sharing scientific and biomedical data.

Arvados™ — http://arvados.org

Pangenome generation workflow

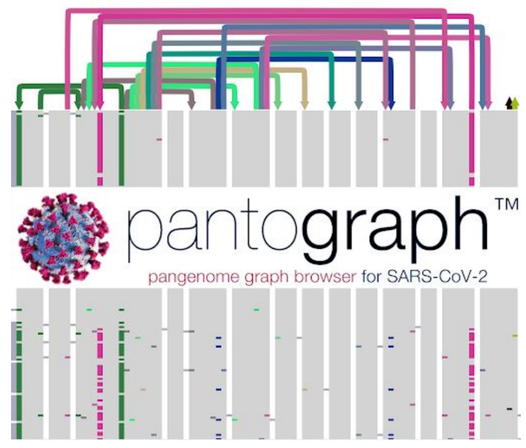
Each time someone uploads a valid sample, it is immediately combined with all the already uploaded SARS-CoV-2 genomes in order to generate an up-to-date SARS-CoV-2 **pangenome** as input for **Pantograph**, an interactive visualization of pangenomes.

`pangenome-generate.cwl`



- A graphical pangenome[2] models the **full set of genomic elements** in a given species or clade.
- Pangenomes naturally express **genome rearrangements**, therefore Pantograph allows researchers to browse the entire genetic diversity in a SARS-CoV-2 population that would otherwise be underestimated.

https://graph-genome.github.io



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

COVID-19 Integrated knowledge base

Swiss Institute of Bioinformatics

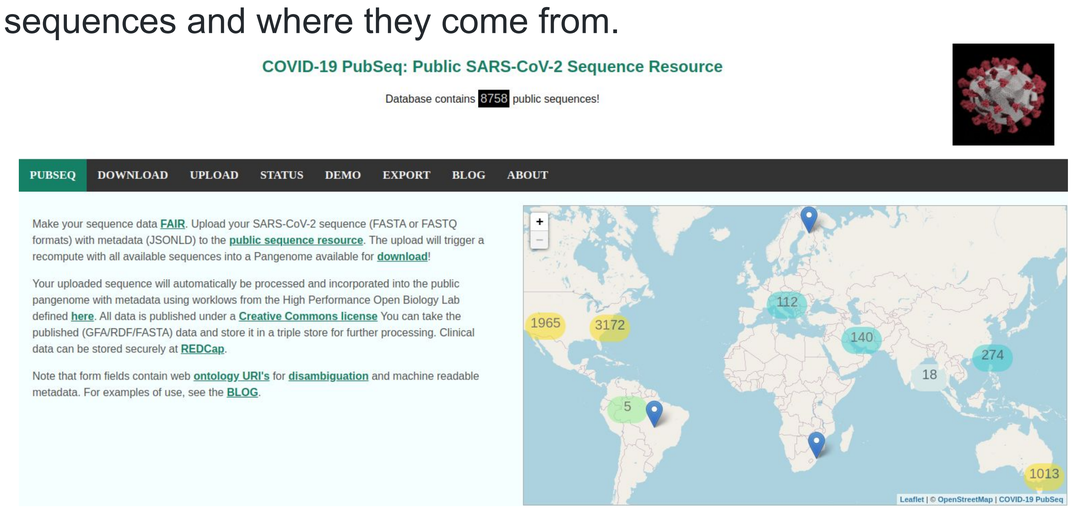
Metadata can be downloaded as **Turtle RDF** which can be loaded into any RDF triple-store. The **Swiss Institute of Bioinformatics** has included the COVID-19 PubSeq data in https://covid-19-sparql.expasy.org.

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" .
   ?sample ?predicate ?object .
}
```

COVID-19 PubSeq

COVID-19 PubSeq home page showing the number of submitted sequences and where they come from.



References

1. Wilkinson et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3, 160018.
2. Eizenga et al. (2020). Pangenome graphs. *Annual Reviews of Genomics and Human Genetics*. 21.

Acknowledgments

curii — aws — THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER. — E/S/R Science for Communities