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

Andrea Guarracino[5], Peter Amstutz[2], Thomas Liener[3], Michael 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, USA.

[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 Biomedicine, University of Luxembourg, Luxembourg.

Project Website: http://covid19.genenetwork.org
Source Code: https://github.com/arvados/bh20-seq-resource
License: Apache 2.0

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. We followed FAIR data practices: 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, most importantly, reproducible workflows are executed after the upload in order to provide up-to-date results rapidly and in standardized data formats.

Existing data repositories for viral data include GISAID, EBI ENA and NCBI. These repositories allow for free sharing data, but do not enforce strict quality control on submitted data or metadata, and do not add value in terms of running additional analysis. In addition, some databases have a restricted license which prevents data from being used in online web services and on-the-fly computation, hindering research.

We created a prototype sequence resource within one week by leveraging existing technologies, such as the Arvados Cloud platform (http://arvados.org), Common Workflow Language (CWL) (http://commonwl.org), and the many free and open source software packages that are available for bioinformatics. Pipelines developed by several teams were combined into an omnibus pangenome analysis workflow. Computing resources for this project were generously donated by Amazon Web Services.