Description of Biobanque Québec COVID-19 JGH Site Second Data Release and Access Procedures
2020/11/19

This document contains a brief description of the JGH Site of the BQC19, data that will be released during this second data release and genetic data available. At the end of the document are the details for access to this data

---

The JGH Site of the Biobanque Québec COVID-19 (bqc19.ca) has collected samples and/or data from over 768 individuals. The characteristics of this data and sample collection are below. Here we describe the data that will be released in the second data release for use by approved investigators.

The JGH Site of the BQC19 has recruited individuals through its own McGill Clinical Genomics COVID-19 program, which was approved by the research ethics board of the Jewish General Hospital of McGill University.

Note that this release of data represents a proportion of the data that will be released. Nevertheless, this data is being released rapidly to enable COVID-19 research.

Note also, that this data release is governed by the data release mechanism approved by the JGH Research Ethics Board and is not a data release organized by the BQC19 itself. The whole-genome sequence data that was generated for this data arose from the local collection of DNA for the BQC19 JGH site. Nevertheless, the BQC19 Steering Board is supportive of this data release to approved investigators.

Whole genome sequencing for this program was performed at the McGill University Genome Centre.

This data release process will be managed by The McGill Interdisciplinary Initiative in Infection and Immunity (Mi4).

Applications for access to data are encouraged from investigators at academic institutions, within industry and from all countries.

Description of BQC19 at the JGH.
Inclusion criteria:

1. Individuals underwent a nasal swab for clinical reasons at the JGH hospital for SARS-CoV-2 testing
2. Cases are those who were PCR positive for SARS-CoV-2. Most patients were admitted.
3. Controls are those who were negative for SARS-CoV-2 on PCR testing on all swabs. Most were admitted to hospital for other reasons.

Subjects have been consented for use of ***all data*** captured during their clinical care (including all laboratory, medications, treatment, radiology, viral data etc…), genomics and re-contact for future studies, and sample collection. Samples were collected at day 0, day 2 and day 7 for participants remaining in the hospital. Follow-up visits on day 30 have started and additional follow-ups on days, 60, 180 and 365 are funded.

The data that will be released for this Second Data Release:

1. Age
2. Sex
3. Demographic information
4. Results of SARS-CoV-2 testing on all occasions of testing
5. Height, weight
6. Habits (alcohol, smoking, street drug use)
7. Pregnancy status
8. Date of blood tests
9. Vital signs in the Emergency Room
10. Past medical history
11. Core laboratory results, which were generated for clinical purposes.
12. COVID-19-specific treatments received in the hospital
13. COVID-19 complications
14. COVID-19 outcomes
15. Whether the patient died during hospitalization
16. Admission to the ICU
17. Whether the patient was hospitalized
18. Date of hospitalization
19. FiO2
20. PaO2
21. Whether the patient was receiving respiratory support, as reported in blood gas test results.
22. Day 30 follow-up data on a subset of patients, including
    a. Activities of daily living
    b. Symptoms
    c. Psychiatric symptoms (anxiety and depression)
23. Day 180 follow-up data on a subset of patients, including
    a. Activities of daily living
    b. Symptoms
    c. Psychiatric symptoms (anxiety and depression)
24. Genome-wide genotyping data on available individuals: ~731 to date.
25. Whole genome sequence data on available individuals: ~535 to date.

All data is detailed in the Data Dictionary (entitled JGHCOVID19Database_DataDictionary_2020-11-12.xlsx)

Future data releases will include medications received in hospital, radiographic reports.

The non-genomic data will be shared in a .csv format.
The genome-wide genotyping data will be shared in a PLINK or similar format.
The variants obtained from the aligned whole genome sequence data will be shared in a .vcf format.

Brief Description of WGS Data

1. Only .VCF files will be supplied during this release of the data.
2. Mean depth of 30x on all included individuals using Illumina's Novaseq 6000 platform (Illumina, San Diego, CA, USA).
3. Sequencing results were analyzed using the McGill Genome Center bioinformatics pipelines[1], in accordance with Genome Analysis Toolkit Best Practices (GATK)

recommendations[2]. Reads were aligned to the GRCh38 reference genome. Results from the variant quality control functions variantRecalibrator and applyVQSR functions from GATK are annotated as in the FILTER column. Further variant, sample, or genotype quality control is left at the discretion of researchers applying for the data.Sequencing results were analyzed using the McGill Genome Center bioinformatics pipelines[1], in accordance with Genome Analysis Toolkit Best Practices (GATK) recommendations[2]. Reads were aligned to the GRCh38 reference genome. Results from the variant quality control functions variantRecalibrator and applyVQSR functions from GATK are annotated as in the FILTER column. Further variant, sample, or genotype quality control is left at the discretion of researchers applying for the data.

Process for Data Access:
All applicants must fill out the BQC19-JGH data access application form, with an accompanying one-page project description. This must be submitted to bcq19imi4.med@mcgill.ca

Applications will be judged to ensure that they ask a scientific or medical question relevant to COVID-19 and that the applicant agrees to respect the data confidentiality and other conditions below.

1. The proposed project must be subject to approval and oversight by the applicant's research ethics board (REB). The applicant must provide written approval of the project by their REB, for secondary use of data. This proof does not need to accompany the data access application form, but no data will be released without REB approval.
2. An access fee of $900 CAD will be charged to academic investigators for this preliminary data release and $5000 CAD to investigators from industry.
3. Applicants should request the genetic data only if they have approval from their REB to undertake genetic analyses. The cost of access to data will be the same, whether or not the genetic data is requested.
4. The applicant must enter into a data access agreement (DTA) with the JGH. In this DTA, the following stipulations will be made:
    a. The applicant must not share any data with any investigators not listed in the application.
    b. The applicant must agree to acknowledge the use of data from the BQC19-JGH in any publication. No authorship is requested by the BQC19.
    c. The applicant must agree to return any data derived from analysis of BQC19-JGH data to BQC19-JGH for broad sharing, at the time that the investigators submit data for peer-review.
    d. The applicant must agree not to attempt to re-identify any individual within the BQC19 study.
    e. The applicant must agree to safeguard the data using the below safeguards as a minimum:
        i. All computers with access to the Information must employ logical access controls (passwords) at the device and network level.
        ii. Where the Information is held on laptops, CD-ROMs, flash memory sticks or other transportable media of any type, passwords and full encryption must be used. This applies equally to backups of the Information stored on transportable media.
        iii. The Information cannot be electronically transmitted, except as described below. This includes the transmittal of the Information by facsimile or by e-mail.

iv. Servers storing and transmitting unencrypted data, where used, must be located in a secure, controlled-access area, preferably in the same area where the Information is accessed. If located in a separate area, controls must be in place to ensure that only approved researchers can access the server. Unless the Information is encrypted continuously while outside the secure area, conduit must be used for all cabling and all cross-connect areas must be physically secured.

v. Network firewalls and access rules must be in place to prevent access to the Information, other than to approved researchers. Information may be stored on and transmitted over networks not meeting these requirements, provided that it is encrypted, except when in use by an Identified Person. Alternatively, the Information may be stored on a stand-alone computer with no external connections, or on a closed network. When a network transmits information that leaves a secure area (for example, when a series of buildings house employees within a single organization), the data must be encrypted whenever it is outside the secure area.

References:

1. Bourgey, M. *et al.* GenPipes: an open-source framework for distributed and scalable genomic analyses. *Gigascience* **8**, (2019).
2. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).