McG COVID-19 Biobank
BQC19-JGH

Description of Biobanque Québec COVID-19 JGH Site Third Data Release and Access Procedures
2021/07/05

This document contains a brief description of the JGH Site of the BQC19, data that will be released during this third data release and the genetic and proteomic 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 1428 individuals. This data-release contains data on up to 1428 individuals for whom chart review is complete. (Note that follow-up data entry is on-going). The characteristics of this data and sample collection are below. Here we describe the data that will be released in the third 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. Future data releases will require subsequent applications for data. Each application for data will require a new data transfer agreement and fee. New ethics applications will not be required if you are continuing the same research program.

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 and proteomics that was generated for this data arose from the local collection of DNA and plasma for the BQC19 JGH site. These data types were not paid for with funds from the BQC19. 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.

Proteomic data for this program was generated by SomaLogic using their SomaScan Platform.

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.

PAV. H-413, 3755, CH. DE LA CÔTE STE-CATHERINE ROAD, MONTRÉAL (QUÉBEC) H3T 1E2
TÉL / TEL : 514-340-8222, POSTE / LOCAL : 27090 FAX : 514-340-7564
WEB : www.mcgill.ca/genepi
www.McGillClinGen.ca

McG COVID-19 Biobank
BQC19-JGH

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, day 7, day 14 and day 30 for participants remaining in the hospital. Follow-up post-discharged visits on day 30, day 180, day 365 have started and additional follow-ups on day 730 are planned. Blood samples were collected on day 90, but there is no actual follow up data for this time point.

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

1. Age, Sex, demographic information, height, weight, habits (alcohol, smoking, street drug use)
2. History of SARS-CoV-2 testing
3. Pregnancy status
4. Arrival vital signs in the Emergency Room
5. Past medical history
    a. ICD9: April 1st, 1985 to March 31st, 2006
    b. ICD10: April 1st, 2006 – present
    c. Manual chart review of all patients
    d. Strength, assistance with walking, rising from a chair, climbing stairs, and falls (SARC-F) and EQ5D with respiratory component two weeks prior to hospital arrival (asked retrospectively)
6. All laboratory results (13,000 + variables) between arrival and discharge by date and time
7. Summary labs (worst lab result over admission for core labs): Neutrophil (relative value), Lymphocyte (relative value), Monocyte (relative value), Eosinophil (relative value), Basophil (relative value), Platelet, Haemoglobin, Urea, Creatinine, NT-proBNP, BNP, Sodium Na+, Potassium K+, C-reactive protein (CRP), LDH, Creatine Phosphokinase (CPK), Albumin, AST, ALT, Procalcitonin (PCT), Troponin T hs (high sensitivity), Troponin I hs (high sensitivity), APTT (activated partial thromboplastin time), International Normalized Ratio (INR), Triglycerides, Total bilirubin, Direct bilirubin (conjugated), Glucose, Venous lactate, D-Dimer, Fibrinogen, Ferritin, IL-6, CD4, CD8.

PAV. H-413, 3755, CH. DE LA CÔTE STE-CATHERINE ROAD, MONTRÉAL (QUÉBEC) H3T 1E2
TÉL / TEL : 514-340-8222, POSTE / LOCAL : 27090 FAX : 514-340-7564
WEB : www.mcgill.ca/genepi
www.McGillClinGen.ca

8. Daily core lab values corresponding to sampling dates (see above list of core labs)
9. Daily vitals corresponding to sampling dates
10. All arterial blood gas results
11. COVID-19-specific treatments received in the hospital.
12. Level of respiratory support
13. COVID-19 complications, COVID-19 outcomes, admission to hospital, admission to the ICU, death during hospitalization or on follow-up.
14. Day 30 follow-up data on a subset of patients, including:
    a. SARC-F and EQ5D with respiratory component
    b. Current persisting symptoms
    c. Psychiatric symptoms (anxiety and depression)
    d. Detailed vaccination data (type, doses, dates, adverse events)
15. Day 180 follow-up data on a subset of patients, including:
    a. SARC-F and EQ5D with respiratory component
    b. Current persisting symptoms
    c. Psychiatric symptoms (anxiety and depression)
    d. Detailed vaccination data (type, doses, dates, adverse events)
16. Day 365 follow-up data on a subset of patients, including:
    a. SARC-F and EQ5D with respiratory component
    b. Current persisting symptoms
    c. Psychiatric symptoms (anxiety and depression)
    d. Detailed vaccination data (type, doses, dates, adverse events)
17. Genome-wide genotyping data on available individuals: ~1232 to date.
18. Whole genome sequence data on available individuals: ~1106 to date.
19. SomaLogic proteomic data generated on 800 samples, collected longitudinally on 378 individuals to date.

All data is detailed in the Data Dictionaries entitled:

- **all_labs_dictionary_17.JUN.2021.csv**
- **BQC19_DataDictionary_2021-06-18.csv**
- **icd.9.10_codes_dictionary_17.JUN.2021.xlsm**

Note that the only additional data type included in this release, compared to the 2nd data release is the proteomic data.

However, this third data release contains clinical information on 1428 individuals. The second data release contained information on 768 individuals.

McG CLINICAL GENOMICS
GENOMICS TO IMPROVE CARE

McG COVID-19 Biobank
BQC19-JGH

Hôpital général juif
Jewish General Hospital

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

The non-genomic data will be shared in a .csv format. We strongly suggest not to use Excel to view or manipulate this data.
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 genome-wide genotyping data

1. PLINK format files of genotype (bqc19_jgh_gen_v#.bed/bim/fam) and imputation data (bqc19_jgh_imp_chr##_v#.bgen) for 1,207 individuals will be supplied during this release of the data.
2. For array genotyped variants, genomic positions were reported in reference to human genome build GRCh37/hg19. In addition to this, we also provide 308 million genetic variants imputed from the TOPMed reference panel1. For imputed variants, genomic positions were reported in reference to human genome build GRCh38/hg38.
3. Genotype data release includes all individuals who passed sample QC (removing (1) outliers of heterozygosity and missing rates, (2) sex aneuploidy and (3) those with discordant self-reported sex and genotype-inferred sex). Variant QC was not performed in this dataset.
4. Before imputation, genotype QC ((1) batch effect quality control, (2) discordant variant calls across control replicates and (3) other basic quality controls [removing MAF < 0.001, call rate < 98% and Departure from Hardy-Weinberg equilibrium (HWE) (p value < 1 x 10-6)] was performed using individuals of genotype-defined European ancestry.

Brief Description of WGS Data

1. A total of 1106 variant call format (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 pipelines2, in accordance with Genome Analysis Toolkit Best Practices (GATK) recommendations3. 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. Please note that some annotations were performed on a larger sample than what is currently released (for e.g. due to patient withdrawals), and annotations may therefore not exactly match what is released to researchers.
4. Further variant, sample, or genotype quality control is left at the discretion of researchers applying for the data.

PAV. H-413, 3755, CH. DE LA CÔTE STE-CATHERINE ROAD, MONTRÉAL (QUÉBEC) H3T 1E2
TÉL / TEL : 514-340-8222, POSTE / LOCAL : 27090 FAX : 514-340-7564
WEB : www.mcgill.ca/genepi
www.McGillClinGen.ca

5. Recommended tools for the analysis of sequencing data include bcftools (samtools), vcftools, plink, and hail.
6. Researchers are encouraged to seek expert support in genome sequencing data if they want to use it for their analysis.

Brief Description of Proteomics Data

1. Data was generated using the SomaLogic SomaScan on citrated plasma samples (collected in ACD tubes).
2. This includes 5,284 SOMAmer reagents targeting 4,742 proteins.
3. Sample linking file includes the date and time of sampling and freezing.
4. Details of the protein targeted can be found in bqc19_jgh_prt_soma_list.xlsx; other related information can be found in bqc19_jgh_prt_dict.pdf and bqc19_jgh_prt_tech_note.pdf.
5. A description of the demographics and clinical data on the samples and individuals that have undergone proteomics profiling can be found in this article: https://www.nature.com/articles/s41591-021-01281-1

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 $3000 CAD will be charged to academic investigators for this third data release and $8000 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. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021).
2. Bourgey, M. et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. Gigascience 8, (2019).
3. 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).