

## HostSeq: A Canadian Whole Genome Sequencing and Clinical Data Resource

Yoo S<sup>1,15\*</sup>, Garg E<sup>2\*</sup>, Elliott LT<sup>2</sup>, Hung RJ<sup>3,4</sup>, Halevy AR<sup>1</sup>, Brooks JD<sup>3</sup>, Bull SB<sup>3,5</sup>, Gagnon F<sup>3</sup>, Greenwood CMT<sup>6,7</sup>, Lawless JF<sup>8</sup>, Paterson AD<sup>1,3</sup>, Sun L<sup>3</sup>, Zawati MH<sup>6</sup>, Lerner-Ellis J<sup>3,5</sup>, Abraham RJS<sup>9</sup>, Birol I<sup>9</sup>, Bourque G<sup>6</sup>, Garant J-M<sup>9</sup>, Gosselin C<sup>9</sup>, Li J<sup>9</sup>, Whitney J<sup>1</sup>, Thiruvahindrapuram B<sup>1</sup>, Herbrick J-A<sup>1</sup>, Lorenti M<sup>1</sup>, Reuter MS<sup>1</sup>, Liu S<sup>1</sup>, Allen U<sup>1,3</sup>, Bernier FP<sup>10,11</sup>, Biggs CM<sup>12,13,14</sup>, Cheung AM<sup>15</sup>, Cowan J<sup>16,17</sup>, Herridge M<sup>15</sup>, Maslove DM<sup>18</sup>, Modi BP<sup>13</sup>, Mooser V<sup>6</sup>, Morris SK<sup>1,3</sup>, Ostrowski M<sup>3,19</sup>, Parekh RS<sup>1,3,20</sup>, Pfeffer G<sup>21</sup>, Suchowersky O<sup>22</sup>, Taher J<sup>3,5</sup>, Turvey SE<sup>12,13</sup>, Upton J<sup>1,3</sup>, Warren RL<sup>9</sup>, Yeung RSM<sup>1,3</sup>, Aziz N<sup>1</sup>, Knoppers BM<sup>6</sup>, Lathrop M<sup>6</sup>, Jones SJM<sup>9</sup>, Scherer SW<sup>1,3</sup>, # Strug LJ<sup>1,3</sup>

<sup>1</sup> The Hospital for Sick Children, Toronto, Ontario, Canada

<sup>2</sup> Simon Fraser University, British Columbia, Canada

<sup>3</sup> University of Toronto, Ontario, Canada

<sup>4</sup> Lunenfeld-Tanenbaum Research Institute, Sinai Health, Ontario Canada

<sup>5</sup> Sinai Health System, Ontario Canada

<sup>6</sup> McGill University, Quebec, Canada

<sup>7</sup> Lady Davis Institute for Medical Research, Jewish General Hospital, Quebec, Canada

<sup>8</sup> University of Waterloo, Ontario, Canada

<sup>9</sup> Canada's Michael Smith Genome Sciences Centre, British Columbia, Canada

<sup>10</sup> University of Calgary, Alberta, Canada

<sup>11</sup> Alberta Children's Hospital, Alberta, Canada

<sup>12</sup> University of British Columbia, British Columbia, Canada

<sup>13</sup> BC Children's Hospital, British Columbia, Canada

<sup>14</sup> St. Paul's Hospital, British Columbia, Canada

<sup>15</sup> University Health Network, Ontario, Canada

<sup>16</sup> University of Ottawa, Ontario, Canada

<sup>17</sup> The Ottawa Hospital Research Institute, Ontario, Canada

<sup>18</sup> Queen's University, Ontario, Canada

<sup>19</sup> St. Michael's Hospital, Unity Health, Ontario, Canada

<sup>20</sup> Women's College Hospital, Ontario, Canada

<sup>21</sup> Hotchkiss Brain Institute, Department of Clinical Neurosciences, and Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Alberta, Canada

<sup>22</sup> University of Alberta, Alberta, Canada

\* These authors contributed equally.

## **ABSTRACT**

HostSeq was launched in April 2020 as a national initiative to integrate whole genome sequencing data from 10,000 Canadians infected with SARS-CoV-2 with clinical information related to their disease experience. The mandate of HostSeq is to support the Canadian and international research communities in their efforts to understand the risk factors for disease and associated health outcomes and support the development of interventions such as vaccines and therapeutics. HostSeq is a collaboration among 13 independent epidemiological studies of SARS-CoV-2 across five provinces in Canada. Aggregated data collected by HostSeq are made available to the public through two data portals: a phenotype portal showing summaries of major variables and their distributions, and a variant search portal enabling queries in a genomic region. Individual-level data is available to the global research community through a Data Access Agreement and Data Access Compliance Office approval. Here we provide an overview of the collective project design along with summary level information for HostSeq. We highlight several statistical considerations for researchers using the HostSeq platform regarding data aggregation, sampling mechanism, covariate adjustment, and X chromosome analysis. In addition to serving as a rich data source, the diversity of study designs, sample sizes, and research objectives among the participating studies provides unique opportunities for the research community.

## **KEYWORDS**

SARS-CoV-2, COVID-19, host genetics, clinical databank, whole genome sequencing

# Correspondence to:

Professor Lisa J. Strug

Departments of Statistical Sciences and Computer Science

The University of Toronto

Senior Scientist

The Hospital for Sick Children

Email: [lisa.strug@utoronto.ca](mailto:lisa.strug@utoronto.ca)

## INTRODUCTION

Following exposure to SARS-CoV-2 (the virus that causes COVID-19), some individuals remain disease- or symptom-free while others develop a spectrum of symptoms from mild to severe with the potential for fatal outcomes (1). This variability in response to exposure suggests that susceptibility is mediated at least in part by host genetic factors (2). Genetic factors have been associated with acquisition and severity of other viral infections (3–7), including SARS-CoV-1 (8,9). Initial work also suggests a role for host genetics in SARS-CoV-2. However, given the novelty of the SARS-CoV-2 virus and the challenges of identifying genetic contributors in a changing environment (2), the specific variations contributing to infection susceptibility and illness severity remain largely unknown (10–13). These variations could include rare or common variations of any form including single nucleotide and structural variations.

In 2020, several countries launched efforts to identify the genetic factors affecting COVID-19 outcomes to support diagnostics, therapy and vaccine development. However, Canada was not poised to do so because, although population-based cohorts exist (14,15), a national whole genome sequencing cohort broadly consented for research and translation, and linked to rich clinical and public health data, did not exist at the onset of the global pandemic. Here we describe the development of this national platform to address pressing questions concerning COVID-19 and other health outcomes in Canada. In April 2020, as part of the Canadian pandemic response, Genome Canada (a not-for-profit organization funded by the Government of Canada) launched the Canadian COVID-19 Genomics Network (CanCOGeN; (16). CanCOGeN established a coordinated pan-Canadian network of studies in collaboration with Canada's national platform for genome sequencing and analysis (CGEn). Beginning June 2020, CGEn developed HostSeq: a national databank of independent clinical and epidemiological studies enrolling SARS-CoV-2-infected participants across Canada. The goal of HostSeq is to create a data repository with whole genome sequencing and harmonized clinical information, including comorbidities for 10,000 Canadians. With the launch of HostSeq, investigators can now begin to address questions of genetic susceptibility to SARS-CoV-2 infection and outcomes from the Canadian perspective. The approvals in place to link HostSeq to other local, provincial or national data resources expand the utility of the resource, including genetic susceptibility for future implications of SARS-CoV-2 infection. Further, summary statistics from association studies of HostSeq have been contributed and are aligned with international efforts including the COVID-19 Host Genetic Initiative (HGI; ((17) and COVID Human Genetic Effort (<https://www.covidhge.com/>)). Most importantly, we have established the research project infrastructure necessary for future pan-Canadian genome sequencing studies. In this resource

paper introducing the HostSeq Databank, we present its design characteristics, high-level analytic considerations pertaining to it, and the research opportunities this rich resource provides.

## **METHODS**

### **HostSeq Project Design**

HostSeq (Figure 1) is a project representing a consortium of investigator-initiated SARS-CoV-2-related research studies across Canada. Each partner study was required to adhere to core consent elements (Table S1), contribute blood (or in rare cases saliva) samples for whole genome sequencing, and provide clinical information using a standardized case report form (Table S2).

Within these studies, eligible participants include individuals of any age with a positive SARS-CoV-2 test performed by any Health Canada approved method. In some studies, suspected cases with clinically assessed COVID-19-related symptoms but without a positive test diagnosis were also included. Within the primary studies, each participant consented to use of their whole genome sequence for future research (18). Participants also consented to the update, linkage and collection of their data from medical records and charts, as well as from administrative databases, and the deposition of data in a cloud-based, access-controlled databank which can be shared with approved researchers including international and commercial researchers.

Additionally, participants had the option to consent to be re-contacted for updates or additional health information, or to invitations to participate in new research. Informed consent was obtained from individuals at each of the participating study sites. For the HostSeq Databank, approval was sought from the study's Research Ethics Board (REB) for inclusion in HostSeq.

The HostSeq Databank shares data with the global research community following review and approval by the HostSeq-independent Data Access Compliance Office (DACO). The DACO verifies that the proposed research has REB approval from their host institution and conforms to HostSeq's REB-approved SARS-CoV-2 and other health outcome research. DACO-approved researchers sign inter-institutional legal agreements, which outline how the shared data is to be used, stored, and privacy protected.

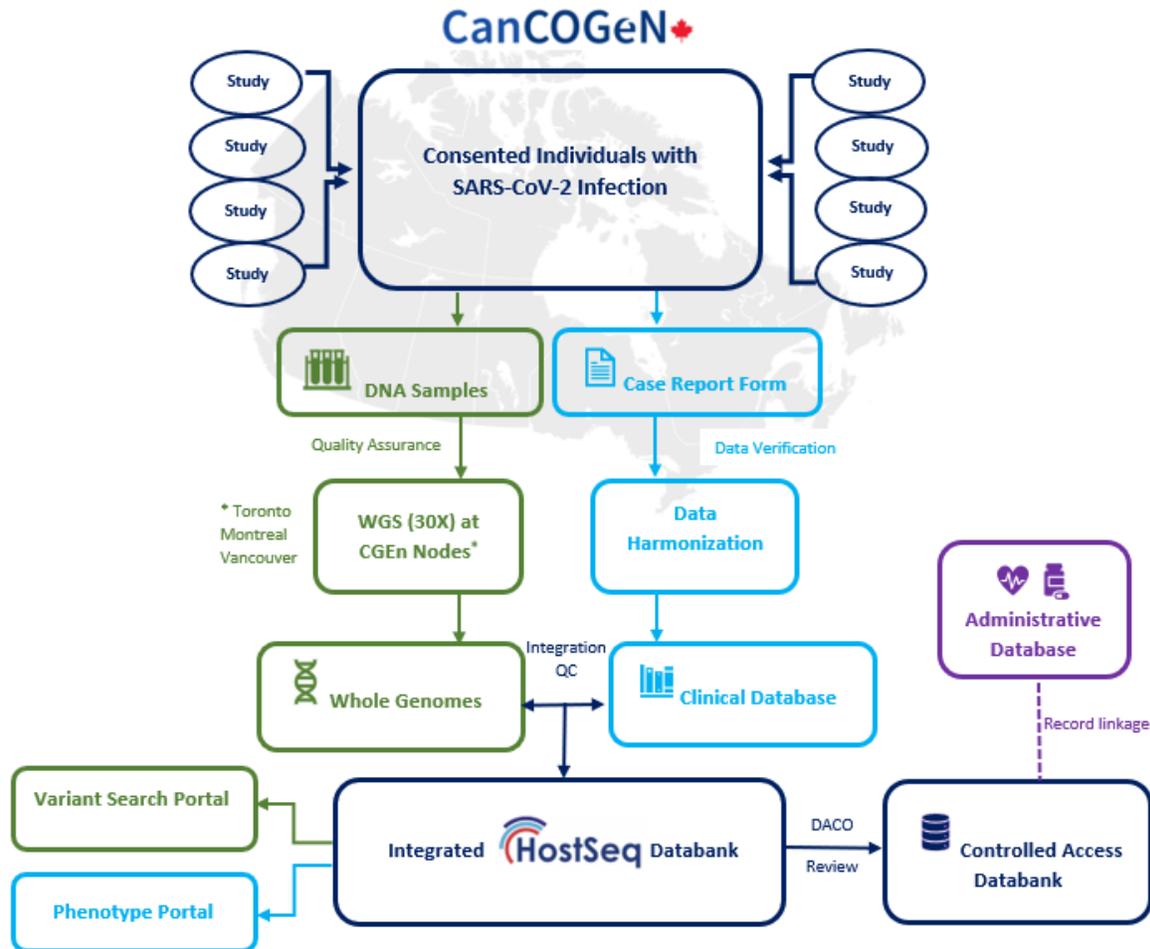


Figure 1. Sample and data flow in HostSeq<sup>1</sup>.

## Whole Genome Sequencing

All HostSeq samples undergo whole genome sequencing in a standardized fashion at one of the three CGEn nodes: Toronto (The Centre for Applied Genomics at The Hospital for Sick Children), Montréal (McGill Genome Centre at McGill University), and Vancouver (Canada's Michael Smith Genome Sciences Centre) on the Illumina NovaSeq6000 platform at 30X depth. Prior to sequencing, quality assurance is performed at multiple stages throughout the process (19). Concordance of the genotyping pipeline among sequencing sites is verified using the Ashkenazi trio set from the Genome in a Bottle Consortium (20).

<sup>1</sup> Aspects of graphics acquired from Wikimedia Commons.

Sequenced samples are analyzed jointly using an in-house pipeline encoded in Nextflow and containerized using Docker<sup>2</sup>. The Genome Reference Consortium human build 38 (GRCh38 assembly version GCA\_000001405.15) reference genome that includes the alternative HLA decoy genes<sup>3</sup> is used. Genomes are processed following the Best Practices guidelines of the Genome Analysis ToolKit (GATK v4.2.0). This includes alignment of sequences to the reference genome, and the genotyping of each sample individually followed by joint-calling of all genotypes together (21).

Sequences are aligned to the reference genome using Burrows-Wheeler Aligner (BWA v0.7.17) (22), sorted with Picard tools (v2.25.0) and bases are recalibrated using the Base Quality Score Recalibration (BQSR) of GATK. GATK HaplotypeCaller is used in Dragen mode on diploid samples for short variant discovery. Aligned sequences are thus converted to genomic Variant Calling Format (gVCF) files, which are then filtered and imported to a GATK GenomicsDB for joint-calling using the GATK GenotypeGVCFs tool. We perform HLA Class I typing using OptiType software (v1.3.1) (23); perform housekeeping with bcftools (v1.11) and samtools (v1.14) (24); check for sample contamination using VerifyBamID2 (v2.0.1) (25); check agreement between reported sex-at-birth and sex chromosome composition using PLINK software (v1.90) (26); and predict ancestry admixture (27) and relatedness (28) using Genetic Relationship and Fingerprinting software (GRAF v2.4). We use PLINK (v2.00) (29) for genetic analysis, R (3.6.3) (30) for data analysis, and snakemake (v7.2.1) (31) for internal workflow management. Additionally, we compare the genetic principal components of HostSeq with the 1000 Genomes Project reference populations (32,33) following the guidelines of plinkQC (34). Samples are excluded based on the following checks (Figure S1): (i) genotyping call rate < 95%, (ii) sex chromosome composition and reported sex-at-birth mismatch, (iii) samples identified as duplicates, (iv) possibly mislabelled samples, (v) sample contamination rate > 3%, and (vi) mean coverage < 10. The whole genome sequence data are readily available in joint VCF format (aligned sequences can be made available on request).

## Contributing Studies

As of April 6, 2022, 13 participating studies contributed data and biospecimens to HostSeq (Table S3). Although all 13 studies continue collecting clinical information, 6 have completed their participant recruitment. To date, we have harmonized data from 12 of those studies. The participating studies are predominantly prospective SARS-CoV-2 studies based in hospitals, and are seeking to identify genetic factors that contribute to varying COVID-19 outcomes. Here we summarize characteristics of the 12 harmonized studies. Three studies—genMARK, Alberta Childhood COVID-19 Cohort (“AB3C”), and Genomic Determinants of COVID-19 (“GD-COVID”)—are using a case-control design, in which laboratory-confirmed COVID-19 cases are

---

<sup>2</sup> [https://svn.bcgsc.ca/bitbucket/users/jmgarant/repos/hostseq\\_nextflow/browse](https://svn.bcgsc.ca/bitbucket/users/jmgarant/repos/hostseq_nextflow/browse)

<sup>3</sup> [https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38\\_reference\\_genome/](https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/)

matched with controls (see Table S3 for matching factors and control eligibility). One study—Quebec COVID-19 Biobank (“BQC19”)—collected clinical data and biospecimens from 12 hospitals in Quebec (35). The remaining studies are case-cohorts with patients that either have a confirmed or suspected diagnosis of COVID-19. From these studies, the HostSeq Databank includes data from study subjects on demographics, comorbidities and assessment and treatment provided for COVID-19.

Study-specific sample sizes currently range from 9 to 3,498. In total, the 13 studies have contributed 6,098 clinical records and submitted 9,097 samples (Table 1). We anticipate 1,000 more samples by Summer 2022. With the exception of two studies that have recruitment across multiple provinces (CANCOV, CONCOR-Donor; n=2,202), most studies are province-specific: six studies in Ontario (GENCOV, GenOMICC, SCB, LEFT-GEN, genMARK, Understanding Immunity to Coronaviruses; n=2,539), one in Quebec (BQC19; n=3,498), two in Alberta (AB3C; AB-HGS n=262) and two in British Columbia (GD-COVID, Host Factors; n=596). Table S3 summarizes their research objectives and study designs. Detailed information for each study is also provided on the CGen website (<https://www.cgen.ca/hostseq-studies-2>).

*Table 1. Status of DNA sample sequencing (as of April 6, 2022). SAMPLES column indicates DNA samples submitted to HostSeq for sequencing. DATA column indicates raw clinical records submitted to HostSeq. Of these, a total of 6,078 records have been harmonized as of April 6, 2022.*

STUDY TITLE	STUDY		
	ACRONYM	SAMPLES	DATA
The Hospital for Sick Children's COVID-19 Biobank	SCB	531	113
Genetic Markers of Susceptibility to COVID-19	genMARK	469	93
The Canadian COVID-19 Prospective Cohort Study	CANCOV	1405	809
The Genetics of Mortality in Critical Care	GenOMICC	332	331
Implementation of Serological and Molecular Tools to Inform COVID-19 Patient Management	GENCOV	1154	823
Convalescent Plasma for COVID-19 Research	CONCOR-Donor	797	625
Host Genetic Factors Underlying Severe COVID-19	Host Factors	10	10
The Quebec COVID-19 Biobank	BQC-19	3498	2559
Genomic Determinants of COVID-19: Integration of Host and Viral Genomic Data to Understand the COVID-19 Epidemiologic Triangle	GD-COVID	586	496
HostSeq - Canadian COVID-19 Human Host Genome Sequencing Ottawa	LEFT-GEN	44	44

Understanding Immunity to Coronaviruses to Develop New Vaccines and Therapies against 2019-nCoV		9	7
Alberta Childhood COVID-19 Cohort Study	AB3C	188	188
Host Genetic Susceptibility to Severe Disease from COVID-19 Infection	AB-HGS	74	0
<b>TOTAL</b>		<b>9,097</b>	<b>6,098</b>

Clinical data from the participating studies is systematically harmonized by the HostSeq team in an ongoing process. In the first stage, we verify the raw data by checking for missingness, consistency, inadmissible values, and aberrant values across the variables. In the second stage, we harmonize the data guided by a set of common definitions and rules, including application of uniform classification, coding, and measurement units specified in the HostSeq Codebook (and available through the *HostSeq Phenotype Portal* described below in *HostSeq Data Portals*). Any potential data errors detected in the harmonization process are communicated with the participating study teams and resolved by follow-up clarifications.

## RESULTS

### Clinical Data Summary

The results discussed in this section are based on approximately 60% of the total expected cohort size of 10,000 participants. Although completeness varies across studies, we have achieved over 70% completeness of key variables capturing demographics, comorbidities, healthcare use, and patient outcome. Among the 6,098 currently available samples, HostSeq has 53.4% females and 44.1% males (and the remaining 2.5% are missing reported sex-at-birth), with an overall mean age (at recruitment) of 49.4 years. Distributions of sex and age vary across the studies (Table S4). Apart from studies including pediatric participants (AB3C, SCB), mean age in the studies ranges from 36.7 years (genMARK) to 61.0 years (GenOMICC). Underlying health conditions are collected in all studies, but using a variety of collection methods (medical chart reviews, participant surveys, and patient interviews). A total of 24 comorbidity variables across cardiovascular, respiratory, immunological, neurological systems, and other pathologies are collected in HostSeq. Distributions of comorbidities across the studies are available through the *HostSeq Phenotype Portal*.

While approximately half of the HostSeq participants were hospitalized and half were assessed in outpatient or community settings, the proportion of hospitalized versus non-hospitalized patients varied substantially across the studies. In all but one study (GenOMICC), participants presented predominantly with mild or moderate symptoms and did not require admission to intensive care units or invasive ventilation support. Of the hospitalized patients, 54.2% were

discharged home, 15.6% were transferred to other hospitals or healthcare settings (e.g., rehabilitation centers or long-term care facilities) and 10.6% were reported deceased (Table 2).

Table 2. Hospitalization and patient outcomes in HostSeq. SD: standard deviation; <sup>1</sup> n=6,048 is a subset of the expected cohort of 10,000; <sup>2</sup> Data not available for 127 participants (2.1%); <sup>3</sup> Data not available for 155 participants (2.6%); <sup>4</sup> Data not available for 767 participants (12.7%); <sup>5</sup> Data not available for 1,331 participants (22.0%); <sup>6</sup> Denominator is 2,530 hospitalized participants. Data currently still under collection for 1,266 participants (20.9%).

HostSeq All Studies (n=6,048) <sup>1</sup>		
Age <sup>2</sup>	Mean (SD)	49.5 (20.6)
	Median (min, Q1, Q3, max)	50.9 (0.1, 34.4, 50.9, 94.0)
Sex at birth <sup>3</sup>	Male	2,668 (44.1%)
	Female	3,225 (53.3%)
Hospitalization <sup>4</sup>	Yes	2,530 (41.8%)
	No	2,751 (45.5%)
ICU admission <sup>5</sup>	Yes	852 (14.1%)
	No	3,865 (63.9%)
Patient outcome <sup>6,7</sup>	Discharged alive	1,371 (54.2%)
	Transfer to another facility	395 (15.6%)
	Palliative discharge	4 (0.2%)
	Hospitalized	7 (0.3%)
	Death	267 (10.6%)

## HostSeq Data Portals

HostSeq provides public access to two data portals: (1) The *Phenotype Portal* shows summaries for the major variables of the HostSeq harmonized clinical data; and (2) the *Variant Search Portal* enables queries in a genomic region to see all variants and their alleles identified in the

HostSeq genomes. Both portals are static platforms that are updated periodically when a new release version of their respective data is available.

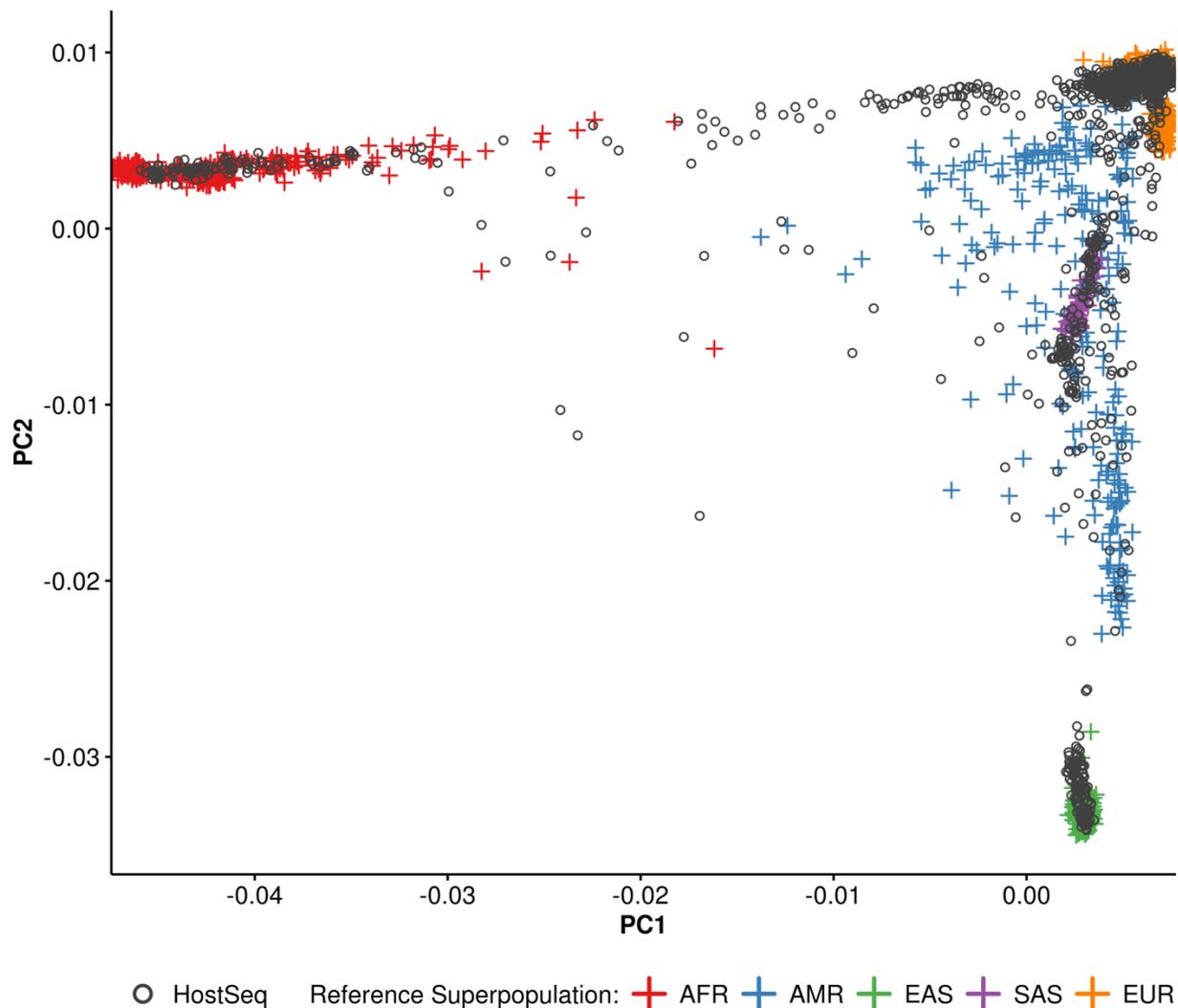
The *HostSeq Phenotype Portal* (<https://hostseq.ca/phenotypes.html>) provides information for clinical variables at aggregate and study-specific levels. Users can access variables by category (e.g., demographics, comorbidities, complications) and view their distributions (categorical variables are presented as boxplots, and numerical variables are presented as histograms and violin plots). Displays are limited to variables with  $\geq 70\%$  completeness. Researchers can also find links to the HostSeq study protocol and up-to-date data dictionaries on this portal.

The *HostSeq Variant Search Portal* (<https://hostseq.ca/dashboard/variants-search>) allows for queries of the HostSeq genetic data. The primary querying functionality is supported by the CanDIG-server (36), a platform enabling federated querying of genomics data. Beacon APIs (37) from the Global Alliance for Genomics and Health (GA4GH) are also built-in to allow HostSeq to join the federated Beacon network. Users can query information about a specific allele of interest. Information about the variants that can be queried includes their position and alleles and the respective internal frequencies of the alleles (minor allele frequencies are reported if they exceed 0.1). All columns in the table can be sorted and filtered.

## Genetic Data Summary

Results reported in this section are based on an interim joint-called set of 3,063 HostSeq genomes, of which 2,992 passed all quality checks (see *Methods*) and may not be reflective of the final HostSeq Databank. Our predicted population structure covers five major ancestry groups (Figures 2 and S2-3; 73% European, 6% Latin American, 7% East Asian, 6% South Asian, 5% African, and approximately 2% uncategorized) and closely matches self-reported ancestries (where available). Additionally, there are 64 and 77 pairs of first- and second-degree relationships, respectively.

Currently HostSeq provides 114.2 million short variants consisting of single nucleotide variants and indels. We report HLA Class I haplotypes for three loci (*HLA-A*, *HLA-B* and *HLA-C*) with bi-allelic typing at 4-digit resolution (allele group with specific alleles). The numbers of unique alleles for *HLA-A*, *HLA-B* and *HLA-C* are 64, 130 and 45, respectively (the most common alleles per locus are *HLA-A*\*02:01, *HLA-B*\*07:02 and *HLA-C*\*07:01).



*Figure 2. PCA projection of HostSeq genomes against reference superpopulations. HostSeq genomes were merged with the 1000 Genomes reference set. The first two principal components of this merged data are shown here with HostSeq genomes in black and 1000 Genomes samples colored by their superpopulation: AFR=African, AMR=Admixed American, EAS=East Asian, SAS=South Asian, EUR=European.*

## DISCUSSION

HostSeq provides unique opportunities to explore the genetics among SARS-CoV-2 positive individuals in Canada and the facilitation of an organizational governance and oversight for researchers in Canada and beyond. Even though the participating studies in HostSeq are heterogeneous with different designs and objectives (Table 3 and Table S3), HostSeq is an opportunity to leverage that diversity to address research questions. Several issues need to be considered when analysing HostSeq data in a given research context. For example: (1) whether

data from different studies should be analysed separately or combined (and how to combine those data); (2) the selection strategies used by the contributing studies to recruit participants; (3) adjustment of covariates for association tests with genetic variants; and (4) the details of X chromosome analysis.

*Table 3. Aspects of participant ascertainment in HostSeq. Study participants who were not confirmed to be positive for infection either by molecular/serology test or clinical symptoms are not included in this Table.*

Desired research study eligibility	Active infection at recruitment	Past infection or disease at recruitment
Infected participants in ICU	Host Factors BQC19 CANCOV GENCOV GenOMICC SCB	SCB AB-HGS
Infected inpatient participants (i.e. hospitalized)	AB3C Host Factors BQC19 CANCOV GENCOV SCB	AB-HGS GENCOV genMARK LEFT-GEN SCB
Infected outpatient participants	AB3C BQC19 CANCOV GENCOV genMARK	genMARK GD-COVID-19 GENCOV LEFT-GEN SCB
Infected participants from community enrollment		CONCOR-Donor

### Individual or Combined Analysis

Whether an investigator's research question would be best answered by within-study comparisons or analyses including multiple studies will require careful consideration of

participant ascertainment criteria. For example, comorbidities might be analyzed within-study then combined via a meta-analysis to account for differences in study designs among the contributing studies. In contrast, for the disease severity indicated by hospitalization duration, it may be appropriate to jointly analyze the subset of studies that focus on in-patient recruitment. Table 3 provides details for the recruitment aspects that may frame such research questions. For example, to compare the genetics of hospitalized patients to non-hospitalized patients within the same study, data from AB3C, BQC19, CANCOV, GENCOV, genMARK, LEFT-GEN and SCB could be used. To compare ICU patients to non-ICU hospitalized patients, Host Factors, BQC19, CANCOV, GENCOV and SCB could contribute.

Given the heterogeneity of the studies in HostSeq, the best approach for certain outcomes may be to analyse relevant studies individually. The feasibility of combining estimates or test results from separate studies, as in meta-analyses, depends on whether the individual studies measure and estimate the same features. The appropriateness of a joint analysis of participant data from multiple studies in an overarching model (perhaps with inclusion of study effects) also depends on whether the studies measure those same features. Although the combination of study-level estimates or tests can be as efficient as joint analysis in large samples (38), meta-analysis of summary data can be less efficient in finite samples. When individual data are available, joint analysis is recommended, incorporating sparse-data methods for variants with low minor allele counts and outcomes with low prevalence (39,40). Furthermore, with study or environmental factors and other sources of heterogeneity, joint analysis can exploit gene-environment interaction (41) and give insight into sources of within- and between-study variation.

Given the dynamic nature of the COVID-19 pandemic, temporal and spatial variation within- and between-studies is another source of heterogeneity that is challenging and deserves consideration. Studies with prolonged recruitment and wide variation in dates of infection may allow such factors to be examined. When looking across the participating HostSeq studies, it may be of interest to examine changes in the profiles of recruited patients as the seropositivity rates and vaccination rates changed with time across Canada and as treatments changed and improved (for example, by combining HostSeq data with serological studies).

## **Participant Selection Mechanism**

Most of the participating studies are designed to include individuals who tested positive for SARS-CoV-2 at a participating institution or individuals who volunteered to donate blood and previously had a positive test. For such participants, it can be difficult to specify exactly what population they represent. To reduce bias and improve interpretation of results, the processes by which individuals join a given study needs to be considered (42). Here, we interpret bias relative to the effect of a variable (genetic or otherwise) in a target population. If an analysis is to involve an outcome variable (e.g., hospitalized versus not hospitalized), a genetic variable of interest and

some additional covariates, then the validity of standard statistical methods is linked to how the sample inclusion depends on the outcome. Such dependence occurs in response-selective designs in which individuals are included in a study according to the values of an outcome (43–45). Except for the simple case-control setting, weighting or conditional estimation is needed to avoid estimation bias of the genetic association. Such methods require estimation or specification of the probability of being selected for inclusion. We encourage analyses that address study sample selection mechanisms.

Methods to account explicitly for selection conditions are similar to methods used for the analysis of secondary traits in case-control studies (46,47). From a methodological standpoint, we also encourage studies of bias and Type 1 error control when standard analyses are used (such as unweighted logistic regression). When the selection mechanism is not easily described, comparison of study samples to population or administrative data may provide insights.

Finally, as HostSeq includes various ancestries, care must be taken to avoid confounding through population stratification (for example, by use of stratification, mixed models, and genetic principal components). This issue, alongside issues related to the heterogeneity of participating studies, are not unique to HostSeq, and arise in most collaborative multi-center or consortium-based research.

## **Covariate Adjustment**

The choice of adjustment covariates in tests for association of outcome with a genetic variant is context dependent and open to discussion in many settings (48,49). In testing for genetic associations with COVID-19 outcomes, one strategy would be to adjust for factors such as age and sex that may affect selection or the outcome in question but are not associated with the genetic variant (unless it is on the sex chromosomes; as mentioned below in *Sex difference and X Chromosome Analyses* below). We must also consider whether to adjust for factors such as comorbidities, which may be related both to the outcome and to the variant. This is of particular importance for severe COVID-19: in the ICU, 1-year mortality outcomes increase with each additional week spent in ICU, each decade in age, and each additional comorbid illness in the Charlson score (50). From a causal perspective, adjusting for multiple covariates without a clear conceptual framework could lead to adjustment for variables that lie on the causal pathway (51). If there is a causal link from variant to outcome that passes through such a variable, then researchers could choose to test for either direct or indirect effects of the variant. As part of the process of learning about genetic effects on COVID-19 outcomes, we encourage analyses both with and without adjusting for such factors.

For the discovery stage in genetic association studies, power considerations are important. There have been suggestions that adjusting for too many covariates decreases power (49,52), and that

two-phase strategies of genome-wide screening by simple analysis followed by targeted in-depth modelling is adequate and efficient. However, this is an area for which further study is warranted.

### **Sex Difference and X Chromosome Analyses**

COVID-19 displays sexual dimorphism with greater severity in males (53–55). In addition to environmental exposures and sex-specific autosomal genetic effects, it is reasonable to hypothesize that some X chromosomal variants play a role in COVID-19 outcomes. Indeed, one gene on the X-chromosome, the angiotensin-converting enzyme 2 (*ACE2*, Xp22.2), has been reported to be important in SARS-Cov-2 infection and genetic analysis has demonstrated association evidence with *ACE2* variants (17).

However, all published GWAS of SARS-CoV-2 susceptibility or COVID-19 severity, to the best of our knowledge, uses the traditional genotype coding (0, 1 and 2 for a female; 0 and 2 for a male) that assumes X-inactivation through a dosage compensation model (i.e., with alleles in the non-pseudo-autosomal regions being expressed exactly half of the time in genetic females; (56)). Yet, it has been reported that close to one-third of the X chromosome genes can escape X-inactivation (57,58); if so, the genotype of a male should be coded 0 and 1 by convention. To robustly deal with X-inactivation uncertainty we recommend the use of recent methods for genetic analysis of SARS-CoV-2 related research questions such as model averaging and selection (59,60) and an easy-to-implement regression model (61). Rare X-chromosome variant analysis (62,63) and X-inclusive polygenic risk scores also require careful consideration and further research.

### **Health Research in the Canadian Context**

People living in Canada are insured under single-payer health care systems administered at the provincial or territorial level. These systems broadly cover physician and hospital services, as well as procedures. This provides a unique opportunity to conduct passive follow-up to understand the short-term and long-term outcomes related to SARS-CoV-2 infection. Administrative health data are generated through patient contact with the health care systems and maintained in multiple databases that, with the appropriate approvals, can be linked using a unique encoded identifier to study specific, patient-level data (including genetic data). These data are administrative or procedural (e.g., surgeries, emergency department visits, hospital visits, comorbidities, routine medical exams), clinical (e.g., prescription medications, cancer screening), laboratory (e.g., blood measurements), social (e.g., education, income), and environmental (e.g., rurality, walkability, food insecurity, exposure to air pollution). The participant informed consent used by HostSeq allows for linkage to these data, transforming the HostSeq dataset into a longitudinal study. Specifically, linkage to administrative provincial data will provide: 1) a

retrospective, longitudinal account of medical histories, health system utilization and diagnoses; and 2) prospective, longitudinal follow-up tracking the natural history of SARS-CoV-2 infection including multisystem inflammatory syndrome in children (MIS-C) and Long COVID, identifying new diagnoses (e.g., diabetes, cancer), long-term health outcomes (e.g., premature mortality), and health resource utilization. Linkage of the HostSeq study samples to provincial administrative data offers opportunities to collect additional data on risk factors and longitudinal outcomes, and opportunities to extend genetic association analyses. Administrative data can also facilitate evaluation of the representativeness of study samples and inform future study design.

The limitations of HostSeq data for investigation of specific scientific questions depend on limitations of the relevant participant studies. In addition, investigations that involve combining data or results from separate participant studies may require assumptions about comparability or heterogeneity; such assumptions should be scrutinized.

## **CONCLUSION**

Through the HostSeq initiative, Canada has built research infrastructure to investigate health effects of SARS-CoV-2 infection and COVID-19, and their association with genetic variants. This infrastructure can also be used for future epidemics. The unique features of the HostSeq project highlighted here present novel opportunities to develop, evaluate, and apply statistical methods that contribute to the understanding of genetic associations with COVID-19-related morbidity and mortality, as well as other phenotypes. The augmentation and linkage of the HostSeq questionnaire and genetic databank with other data resources is made possible by broad and flexible consent and will generate a dynamic population-based resource. This will allow for study of a broad range of research questions and sustained productivity over the years to come.

## **ACKNOWLEDGEMENTS**

We are very thankful for the support we received from Genome Canada, Innovation, Science and Economic Development Canada, SickKids Foundation, Toronto Academic Health Science Network (TAHSN), CFI, CIHR, COVID-19 Immunity Task Force (CITF), Ministry of Colleges and Universities COVID-19 Rapid Research Fund, Michael Smith Health Research BC, PHCRI Early Career Award, Genome Alberta and the Alberta Children's Hospital Foundation, and internal support from the Hotchkiss Brain Institute and Department of Clinical Neurosciences (University of Calgary). We also wish to express gratitude to all HostSeq project participant studies and the individual participants within these studies for their contribution.

## REFERENCES

1. Government of Canada. COVID-19 signs, symptoms and severity of disease: A clinician guide. 2021 [Accessed Summer 2022]. Available from: <https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/guidance-documents/signs-symptoms-severity.html>
2. Lin YC, Brooks J, Bull S, Gagnon F, Greenwood C, Hung R, et al. Statistical power in COVID-19 case-control host genomic study design. *Genome Med.* 2020;12(1):115.
3. Allers K, Schneider T. CCR5 $\Delta$ 32 mutation and HIV infection: Basis for curative HIV therapy. *Curr Opin Virol.* 2015;14:24–9.
4. Nordgren J, Svensson L. Genetic susceptibility to human norovirus infection: An Update. *Viruses.* 2019;11(3):226.
5. Coppola N, Marrone A, Pisaturo M, Starace M, Signoriello G, Gentile I, et al. Role of interleukin 28-B in the spontaneous and treatment-related clearance of HCV infection in patients with chronic HBV/HCV dual infection. *Eur J Clin Microbiol Infect Dis.* 2014;33(4):559–67.
6. Trandem K, Anghelina D, Zhao J, Perlman S. Regulatory T cells inhibit T cell proliferation and decrease demyelination in mice chronically infected with a coronavirus. *J Immunol.* 2010;184(8):4391–400.
7. Mahallawi W, Khabour O, Zhang Q, Makhdoum H, Suliman B. MERS-CoV infection in humans is associated with a pro-inflammatory Th1 and Th17 cytokine profile. *Cytokine.* 2018;104:8–13.
8. Ng M, Lau KM, Li L, Cheng SH, Chan W, Hui P, et al. Association of human-leukocyte-antigen class I (B\*0703) and class II (DRB1\*0301) genotypes with susceptibility and resistance to the development of severe acute respiratory syndrome. *J Infect Dis.* 2004;190(3):515–8.
9. Lin M, Tseng HK, Trejaut J, Lee HL, Loo J, Chu CC, et al. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. *BMC Med Genet.* 2003;4(1):1–7.
10. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick A, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. *Nature.* 2021;591(7848):92–8.
11. Kousathanas A, Pairo-Castineira E, Rawlik K, Stuckey A, Odhams C, Walker S, et al. Whole genome sequencing reveals host factors underlying critical COVID-19. *Nature.* 2022;in press.
12. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature.* 2021;600(7889):472–7.
13. Zhang Q, Bastard P, COVID Human Genetic Effort, Cobat A, Casanova JL. Human genetic and immunological determinants of critical COVID-19 pneumonia. *Nature.* 2022;603(7902):587–98.
14. Raina P, Wolfson C, Kirkland S, Griffith L, Oremus M, Patterson C, et al. The Canadian Longitudinal Study on Aging (CLSA). *Can J Aging Rev Can Vieil.* 2009;28(3):221–9.
15. Dummer T, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *Can Med Assoc J.* 2018;190(23):E710–17.

16. Song L, Liu H, Brinkman F, Gill E, Griffiths E, Hsiao W, et al. Addressing privacy concerns in sharing viral sequences and minimum contextual data in a public repository during the COVID-19 pandemic. *Front Genet.* 2022;12:716541.
17. COVID-19 Host Genetics Initiative, Ganna A. Mapping the human genetic architecture of COVID-19: An update. *medrxiv.* 2021;21265944.
18. Knoppers B, Beauvais M, Joly Y, Zawati M, Rousseau S, Chasse M, et al. Modeling consent in the time of COVID-19. *J Law Biosci.* 2020;7(1):1–6.
19. Corbett R, Eveleigh R, Whitney J, Barai N, Bourgey M, Chuah E, et al. A distributed whole genome sequencing benchmark study. *Front Genet.* 2020;1524.
20. Zook J, Catoe D, McDaniel J. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3(1):1–26.
21. Van der Auwera G, O’Connor B. *Genomics in the cloud: Using Docker, GATK, and WDL in Terra.* 1st ed. O’Reilly Media; 2020.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
23. Szolek A, Schubert B, Mohr C, Sturm M, Kohlbacher O. OptiType: Precision HLA typing from next-generation sequencing data. *Bioinforma Oxf Engl.* 2014;30(23):3310–6.
24. Danecek P, Bonfield J. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2).
25. Zhang F, Flickinger M, Gagliano Taliun S, InPSYght Psychiatric Genetics Consortium, Abecasis G, Scott L, et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genomic Res.* 2020;30(2):185–94.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D. Plink: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–75.
27. Jin Y, Schaffer A, Feolo M, Holmes J, Kattman B. GRAF-pop: A fast distance-based method to infer subject ancestry from multiple genotype datasets without principal components analysis. *G3 Bethesda Md.* 2019;9(8):2447–61.
28. Jin Y, Schaffer A, Sherry S, Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. *PLOS ONE.* 2017;12(6):e0179106.
29. Chang C, Chow C, Tellier L, Vattikuti S, Purcell S, Lee J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4(7):s13742–15.
30. R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria; 2020.
31. Molder F, Jablonski K, Letcher B, Hall M, Tomkins-Tinch C, Sochat V, et al. Sustainable data analysis with Snakemake. *F1000Res.* 2021;10(33).
32. Roslin N, Weili L, Paterson A, Strug L. Quality control analysis of the 1000 Genome Project Omni2,5 genotypes. *bioRxiv.* 2016;078600.

33. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
34. Meyer H. meyer-lab-cshl/plinkQC: plinkQC 0.3.2. 2020.
35. Tremblay K, Rousseau S, Zawati M, Auld D, Chasse M, Coderre D, et al. The Biobanque quebecoise de la COVID-19 (BQC19) - A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PLOS ONE*. 2021;16(5).
36. Dursi L, Bozoky Z, de Borja R, Li H, Lipski A, Brudno M. Federated network across Canada for multi-omic and health data discovery and analysis. *Cell Genomics*. 2021;1(2):100033.
37. Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke S, et al. Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol*. 2019;37(3):220–4.
38. Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol*. 2009;33(3):256–65.
39. Ma C, Blackwell T, Boehnke M, Scott L. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol*. 2013;37(6):539–50.
40. Chen DG, Liu D, Min X, Zhang H. Relative efficiency of using summary versus individual data in random-effects meta-analysis. *Biometrics*. 2020;76(4):1319–29.
41. Kraft P, Yen YC, Stram D, Morrison J, Gauderman W. Exploiting gene-environment interactions to detect genetic associations. *Hum Hered*. 2007;63(2):111–9.
42. Griffith G. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. 2020;11(1):1–12.
43. Tao R, Zeng D, Franceschini N, North K, Boerwinkle E, Lin DY. Analysis of sequence data under multivariate trait-dependent sampling. *J Am Stat Assoc*. 2015;110(510):560–72.
44. Lawless J, Kalbfleisch J, Wild C. Semiparametric methods for response-selective and missing data problems in regression. *Stat Methodol Ser B*. 1999;61(2):413–38.
45. Huang B, Lin D. Efficient association mapping of quantitative trait loci with selective genotyping. *Am J Hum Genet*. 2007;80(3):567–76.
46. Monsees G, Tamimi R, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol*. 2009;33(8):717–28.
47. Tounkara F, Lefebvre G, Greenwood C, Oualkacha K. A flexible copula-based approach for the analysis of secondary phenotypes in ascertained samples. *Stat Med*. 2020;39(5):517–43.
48. Gail M, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. *Biometrika*. 1984;71(3):431–44.
49. Pirinen M, Donnelly P, Spencer C. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet*. 2012;44(8):848–51.

50. Herridge M, Cheung A, Tansey C, Matte-Martyn A, Diaz-Granados N, Al-Saidi F, et al. One-year outcomes in survivors of the Acute Respiratory Distress Syndrome. *N Engl J Med*. 2003;348:683–93.
51. Lederer D, Bell S, Branson R, Chalmers J, Marshall R, Maslove D, et al. Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Ann Am Thorac Soc*. 2019;16(1):22–8.
52. Aschard H, Vilhjalmsson B, Joshi A, Price A, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet*. 2015;96(2):329–39.
53. Peckham H, de Grujter N, Raine C, Radzisweska A, Ciurtin C, Wedderburn L. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nat Commun*. 2020;11(1):1–10.
54. Vahidy F, Pan A, Ahnstedt H, Munshi Y, Choi H, Tiruneh Y, et al. Sex differences in susceptibility, severity, and outcomes of coronavirus disease 2019: Cross-sectional analysis from a diverse US metropolitan area. *PLOS ONE*. 2021;16(1):e0245556.
55. Pradhan A, Olsson PE. Sex differences in severity and mortality from COVID-19: Are males more vulnerable? *BMC*. 2020;11(53).
56. Song Y, Biernacka J, Winham S. Testing and estimation of X-chromosome SNP effects: Impact of model assumptions. *Genet Epidemiol*. 2021;45(6):577–92.
57. Tukiainen T, Villani AC, Yen A, Rivas M, Marshall J, Satija R, et al. Landscape of X chromosome inactivation across human tissues. *Nature*. 2017;550(7675):244–8.
58. Lee S, Wu M, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762–75.
59. Wang J, Talluri R, Shete S. Selection of X-chromosome inactivation model. *Cancer Inform*. 2017;16:1–8.
60. Chen B, Craiu R, Sun L. Bayesian model averaging for the X-chromosome inactivation dilemma in genetic association study. *Biostatistics*. 2020;21(2):319–35.
61. Chen B, Craiu R, Strug L, Sun L. The X factor: A robust and powerful approach to X-chromosome-inclusive whole-genome association studies. *Genet Epidemiol*. 2021;45(7):694–709.
62. Derkach A, Lawless J, Sun L. Pooled association tests for rare genetic variants: A review and some new results. *Stat Sci*. 2014;29(2):302–21.
63. Lee S, Abecasis G, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5–23.

## APPENDIX

*Table S1. HostSeq Core Consent Elements. In order to deposit datasets in HostSeq COVID-19 controlled-access Databank, all the elements in this table must be obtained in the research consent.*

ELEMENTS	DETAILS
<b>Research data</b>	Whole genome sequencing of the sample and the ongoing collection of clinical data from participant's medical records/chart, administrative databases, etc.
<b>International sharing</b>	International sharing of genetic and clinical data
<b>Future research use</b>	Future health research on COVID-19 and other health outcomes
<b>Commercial use</b>	Use of genetic and clinical data for commercial purposes
<b>Controlled access</b>	Sharing of genetic and clinical data through a controlled-access mechanism
<b>Storage on cloud servers in Canada</b>	Storage of genetic and clinical data in the HostSeq Databank, on centralized Canadian cloud servers
<b>Duration of storage</b>	Indefinite storage of genetic and clinical data
<b>Data withdrawal</b>	Not possible to withdraw data that has already been distributed and used
<b>Re-identification</b>	Low risk that the participant could be re-identified in the future
<b>Re-contact (optional)</b>	Optional re-contact to update personal information, provide new health information, or to be invited to participate in new research projects.

*Table S2. HostSeq Case Report Form.*

IDENTIFICATION	COMORBIDITIES	LAB RESULTS
Host Hospital	<b>Comorbidities: Immune system</b>	Haemoglobin
Source study participant ID	HIV	WBC count

HostSeq ID	Immunocompromised status	Lymphocyte count
DNA ID	Onset of COVID-19 to immunocompromisation	Neutrophil count
Sequencing site	Organ transplant	Haematocrit
Was informed consent obtained?	Organ type of transplant	Platelets
Study enrollment date	Autoimmune or rheumatologic disease	APTT/APTR
DOB	Diabetes	APTT
Age	Type I diabetes	PT
Laboratory COVID-19 test result	Type II diabetes	INR
Suspected to be COVID-19 positive?	<b>Comorbidities: Respiratory system</b>	ALT/SGPT
Ambulatory	Asthma	Total Bilirubin
Hospitalized State	COPD	AST/SGOT
Date of admission	Cystic fibrosis	Glucose
Sex at birth	Sleep apnea	Blood Urea Nitrogen (urea)
Gender	Do you use a home CPAP device?	Lactate
Ancestry	<b>Comorbidities: Genitourinary/Metabolic</b>	Creatinine
Height (cm)	Chronic kidney disease	Sodium
Weight (Kg)	Liver disease	Potassium
<b>DEMOGRAPHICS</b>	Gallbladder	Procalcitonin
Country of birth	Pancreas	CRP
Education	<b>Comorbidities: Cardiovascular system</b>	CT chest
Employment	Coronary intervention	Chest X-Ray performed?
Type of residence	Coronary artery bypass	ECG
Household composition	Congestive heart failure	POCUS
<b>PREGNANCY</b>	Hypertension	Echocardiogram
Pregnancy - currently pregnant?	Myocardial infarction	LDH
<b>RISK FACTORS</b>	Myocardial infarction Type I	D-Dimer
Cigarettes	Myocardial infarction Type II	Fibrinogen
How many cigarettes daily?	Peripheral vascular disease	Ferritin
Vaping	Stroke	Triglycerides
Cannabis	Arrhythmias	IL-6
<b>AT ADMISSION/ASSESSMENT</b>	<b>Comorbidities: Neurological</b>	CD4
COVID-19 test date	Dementia	CD8
Date of diagnosis	Neurological or neuropsychiatric disease	CD4/CD8 ratio
Serology	<b>Comorbidities: Cancer</b>	NT-proBNP
Serology positive test date	Currently diagnosed with cancer?	BNP
Date of last negative test	Patient age at diagnosis	Troponin

PCR	Leukemia	<b>TREATMENT</b>
Commercial serology test kit name	Lymphoma	ICU or High Dependency Unit admission?
Onset date of first/earliest symptom	Sarcoma	If yes, date of ICU admission
ER triage date at this facility	Carcinoma	If yes, date of ICU discharge
Transfer from another facility?	Myeloma	Prone ventilation
ABO blood type	Mixed types	Inhaled nitric oxide
Rh factor	Cancer location	Tracheostomy inserted
Home medications	Cancer treatment in the past 12 months	Extracorporeal support
BCG vaccine	Other comorbidities	RRT or dialysis
COVID-19 Vaccine?	<b>PATHOGEN TESTING</b>	Inotropes/vasopressors
Name of first dose vaccine	Pathogen testing done during this illness	OTHER intervention or procedure
Name of second dose vaccine	Influenza	<b>MEDICATION</b>
<b>SYMPTOMS</b>	Coronavirus	Antiviral agent
Cough	RSV	Azithromycin “Zithromax”
Days with cough	Adenovirus	Any other antibiotic
Difficulty breathing	Enterovirus	Corticosteroid
Fever	Bacteria	Antifungal agent
Days with fever	Other infectious respiratory diagnosis	Colchicine
Heart rate	Was there a physician diagnosis of pneumonia?	Chloroquine “Aralen”
Highest respiratory rate	Suspected non-infective	Hydroxychloroquine
Systolic blood pressure	<b>COMPLICATIONS</b>	Tocilizumab “Actemra”
Diastolic blood pressure	Viral pneumonitis	Kineret “Anakinra”
Oxygen saturation	Bacterial pneumonia	IVIG
Fatigue	Acute Respiratory Distress Syndrome	Plasma
Myalgia (general aches and pain)	Pneumothorax	Other COVID therapy
Runny nose	Pleural effusion	<b>OUTCOME</b>
Sore throat	Cryptogenic organizing pneumonia (COP)	Outcome
Loss of taste/smell sense	Bronchiolitis	Outcome date
Nosebleed	Meningitis / Encephalitis	Self-care ability versus before illness
Ear pain	Seizure	Repeat hospital visit within 30 days?
Wheezing	Stroke / Cerebrovascular accident	Date and reason
Chest pain	Congestive heart failure	
Joint pain	Cardiac inflammation (mark all that apply)	
Headache	Cardiac arrhythmia	

seizures	Cardiac ischaemia	
Altered consciousness/confusion	Cardiac arrest	
Abdominal pain	Coagulation disorder / Disseminated Intravascular Coagulation	
Diarrhea	Anemia	
Nausea/vomiting	Rhabdomyolysis / Myositis	
Conjunctivitis	Acute renal injury/ Acute renal failure	
Skin rash	Gastrointestinal haemorrhage	
Other symptoms	Pancreatitis	
Asymptomatic	Liver dysfunction	
	Hyperglycemia	
	Hypoglycemia	
	Inflammatory syndrome/Kawasaki Disease like	
	Other, please specify	

Table S3. List of HostSeq participating studies as described in respective protocols.

Study, PI	Objectives	Eligibility
GENCOV Jordan Lerner-Ellis Jennifer Taher	To identify the characteristics of the antibody response that result in maintained immune response and better patient outcomes; to determine impact of genetic differences on COVID-19 infection severity and immune response; to determine impact of different viral strains on antibody response and patient outcomes.	All outpatients (seen in ER and assessment centres) with mild symptoms as well as hospitalized patients with severe symptoms, recruited from six hospitals in Ontario.
GenOMICC David Maslove	To identify genetic determinants of severe, life-threatening COVID-19 infections.	All patients with confirmed COVID-19 admitted to ICU
CANCOV Angela Cheung Margaret Herridge	To evaluate early to one-year outcomes in patients with COVID-19 and their family caregivers.	Hospitalized patients (>16 years and positive COVID-19 test) and their caregivers.  Anticipated deaths or withdrawal of life-sustaining treatment within 48 hours, catastrophic neurological injury, patients unlikely to comply with follow-up are excluded.

<p>genMARK</p> <p>Upton Allen</p>	<p><b>Primary.</b> To identify genetic variation associated with:</p> <p>i). Susceptibility to COVID-19 and SARS-CoV infection among study subjects vs controls</p> <p>ii). Likelihood of severe disease among study subjects.</p> <p><b>Secondary.</b> To assess immune responses, including T cell function and cytokine/chemokine profiles among subjects with different severity of COVID-19.</p>	<p>Individuals without a known immunodeficiency who experienced illness due to confirmed COVID-19 infection.</p> <p>Controls: individuals exposed to COVID-19 but are symptom-free as far as COVID-19 is concerned. Controls are tested for subclinical infection. Controls are 1:1 matched by age groups from within the same home, using the most distantly related household member.</p>
<p>LEFT-GEN</p> <p>Juthaporn Cowan</p>	<p>To understand immunopathogenesis of persistent symptoms</p>	<p>All patients with confirmed COVID-19.</p>
<p>BQC19</p> <p>Vincent Mooser</p>	<p>To support the institutions of the health and social services network (HSSN) by making the unique biological material and data accessible for research on COVID-19 and related diseases.</p>	<p>All patients who tested for COVID-19 at participating institutions</p>
<p>CONCOR-Donor</p> <p>Rulan Parekh</p>	<p>To determine if the titers of SARS-CoV-2 specific neutralizing antibody and serum ELISA antibodies are correlated with clinical and demographic factors; to characterize duration of immunity over 1 year and determine if the titers are modified by clinical and demographic factors; to identify genetic predictors of SARS-CoV-2-specific neutralizing antibody and serum ELISA antibody titers at baseline and over time</p>	<p>Individuals <math>\geq 17</math> years, who recovered from COVID-19 infection and volunteer to donate convalescent plasma</p>
<p>AB3C</p> <p>Francois Bernier</p>	<p>To establish prospective cohort study of all children in Alberta tested for or diagnosed with confirmed or probable COVID-19 infection; to conduct a detailed multiomic precision-medicine evaluation of some children in Alberta with confirmed or probable COVID-19 infection and some healthy controls; to evaluate adaptive immune response to SARS-CoV-2 virus in children measuring the development of immunity against COVID-19 in children with and without clinically apparent confirmed or probable COVID-19 infection</p>	<p>All children in Alberta, aged 0-17 years, who have a respiratory sample tested for COVID-19 in 2020; diagnosed with probable COVID-19 (as defined by Alberta Health), but not tested.</p> <p>Controls: 25 healthy children in Calgary who are matched on age and gender. The controls will be identified from previous research conducted by the ACH Infectious Diseases Epidemiology &amp; Vaccine Evaluation or enrolled in All Our Families cohort or in the Healthy Infants and Children Clinical Research Program.</p>

<p>Host Genetic Factors Underlying Severe COVID-19</p> <p>Catherine Biggs</p>	<p>To determine what proportion of patients with severe COVID-19 have an underlying IEI and characterize the role of genetic factors in COVID-19 disease; to explore molecularly-targeted treatments e.g., if patients with hyperinflammation have variants dysregulating JAK-STAT signaling, JAK inhibitors</p>	<p>Individuals aged &lt;70 years with documented COVID-19 infection, who are previously healthy with severe COVID-19. (Severe disease is defined as any of the following: requiring ICU admission, cardiorespiratory support such as non-invasive or mechanical ventilation, extracorporeal membrane oxygenation, or unusual complications such as multisystem inflammatory syndrome in children (MIS-C) associated with COVID-19, encephalitis, other inflammatory).</p>
<p>GD-COVID</p> <p>Stuart Turvey</p>	<p>To identify host and viral genomic factors associated with COVID-19 disease susceptibility and health outcomes</p>	<p>Individuals of any age with documented SARS-CoV-2 infection, who are a resident of BC; have viral genome sequencing data available through VirusSeq; provide permission to contact Health Authority, CTCR or BCCDC COVID-19 survey</p>
<p>Understanding Immunity to Coronaviruses</p> <p>Mario Ostrowski</p>	<p>To determine B cell and antibody immunity to coronaviruses; to identify candidate targets of the virus for vaccine development; to identify novel neutralizing antibodies to 2019-nCoV that can be used in therapeutics; to understand the innate immune response against 2019-nCoV; to identify host factors that play critical roles in developing strong immunity to coronaviruses.</p>	<p>Individuals 18 years old or older. Those with blood samples hemoglobin &lt; 100 g/L or pregnant or with hemophilia or platelet count &lt; 50,000 / ul are excluded.</p>
<p>SickKids COVID-19 Biobank</p> <p>Rae Yeung</p>	<p>To establish a central biobank of biological samples and related core patient level data that will serve as a single point of access that supports a unified and collaborative approach to clinical and translational research related to the current COVID-19 pandemic and future works.</p>	<p>All patients presenting to the Hospital for Sick Children with a suspected or confirmed diagnosis of COVID-19, and family members of participants as applicable.</p>
<p>AB-HGS</p> <p>Gerald Pfeffer</p>	<p>Identify host genetic susceptibility factors associated with requirement for hospitalisation due to COVID-19. Family members who did not require hospitalisation were recruited as controls.</p>	<p>Confirmed SARS-CoV-2 infection (by provincial nucleic acid test) Required admission to hospital (either medical units or critical care) Age &lt;60 years Unvaccinated participants only Omicron variant cases were excluded</p>

Table S4. Distribution of sex and age across HostSeq studies (n=6,048). SD: Standard deviation; IQR: interquartile range.

Study	Sex		Age	
	Male	Female	Mean (SD)	Median (IQR)
GENCOV	359 (43.6%)	451 (54.8%)	45.2 (14.6)	45.0 (24.0)
GenOMICC	151 (53.5%)	127 (45.0%)	61.0 (14.9)	62.5 (20.0)
CANCOV	268 (34.1%)	375 (47.6%)	51.2 (16.6)	51.0 (24.5)
CONCOR-Donor	194 (31.0%)	431 (69.0%)	43.8 (13.7)	43.0 (23.0)
BQC19	1,257 (49.1%)	1,294 (50.6%)	55.4 (20.8)	56.1 (30.5)
genMARK	39 (42.4%)	53 (57.6%)	36.7 (17.7)	37.0 (29.0)
LEFT-GEN	23 (52.3%)	21 (47.7%)	50.5 (16.3)	48.0 (24.3)
GD-COVID-19	192 (38.7%)	304 (61.3%)	45.6 (19.0)	46.0 (28.0)
AB3C	92 (48.9%)	95 (50.5%)	11.5 (4.5)	12.0 (7.0)
SCB	69 (61.1%)	44 (38.9%)	9.3 (7.6)	7.0 (12.0)

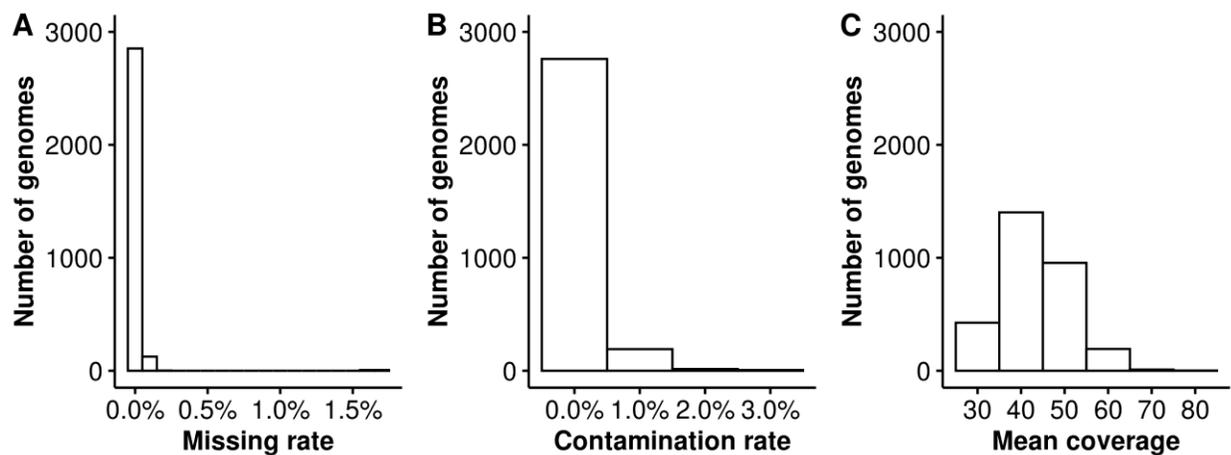


Figure S1. Quality of HostSeq genomes. (A) Missing rate < 5% (B) Contamination rate < 3% (C) Mean coverage > 10.

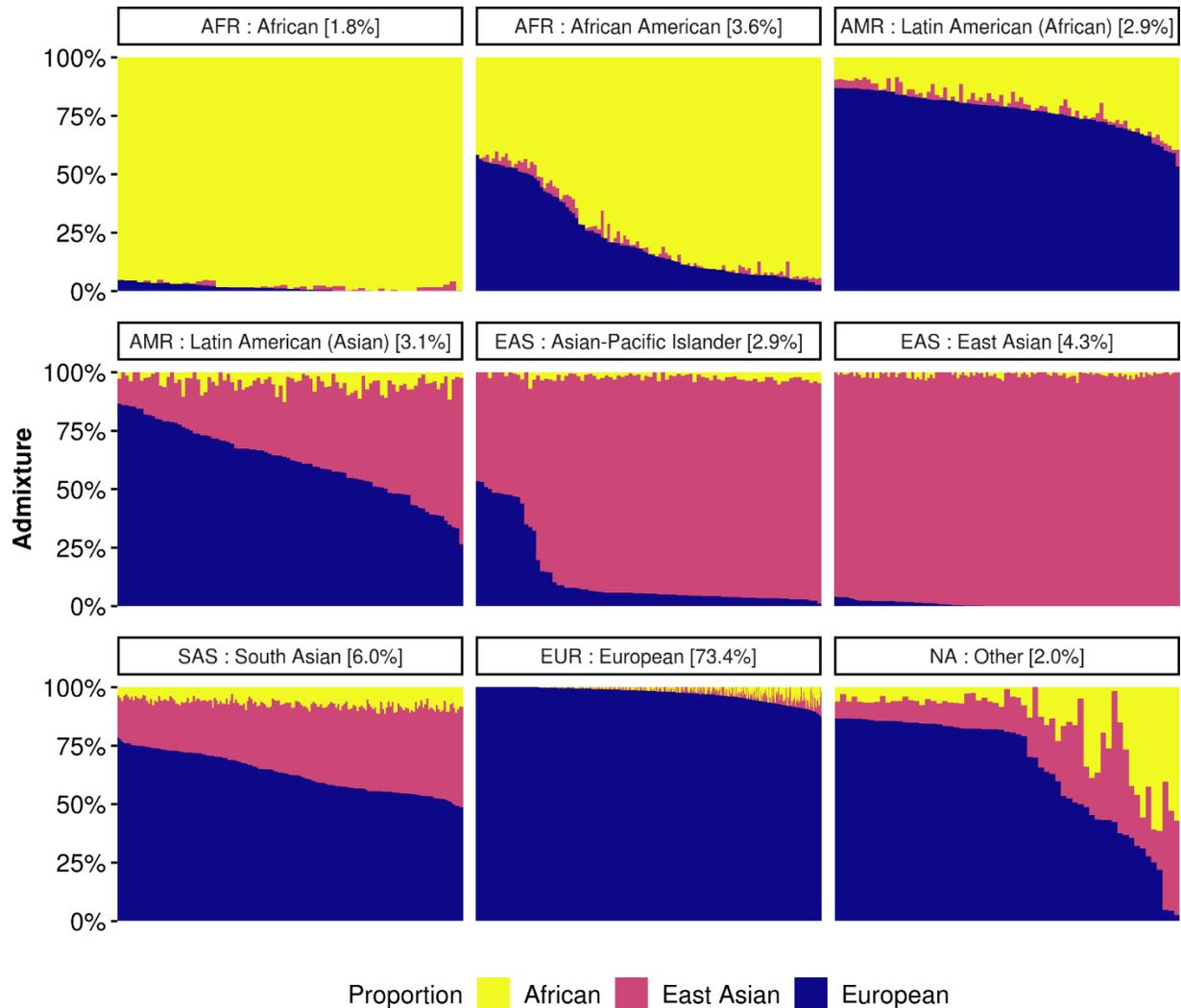
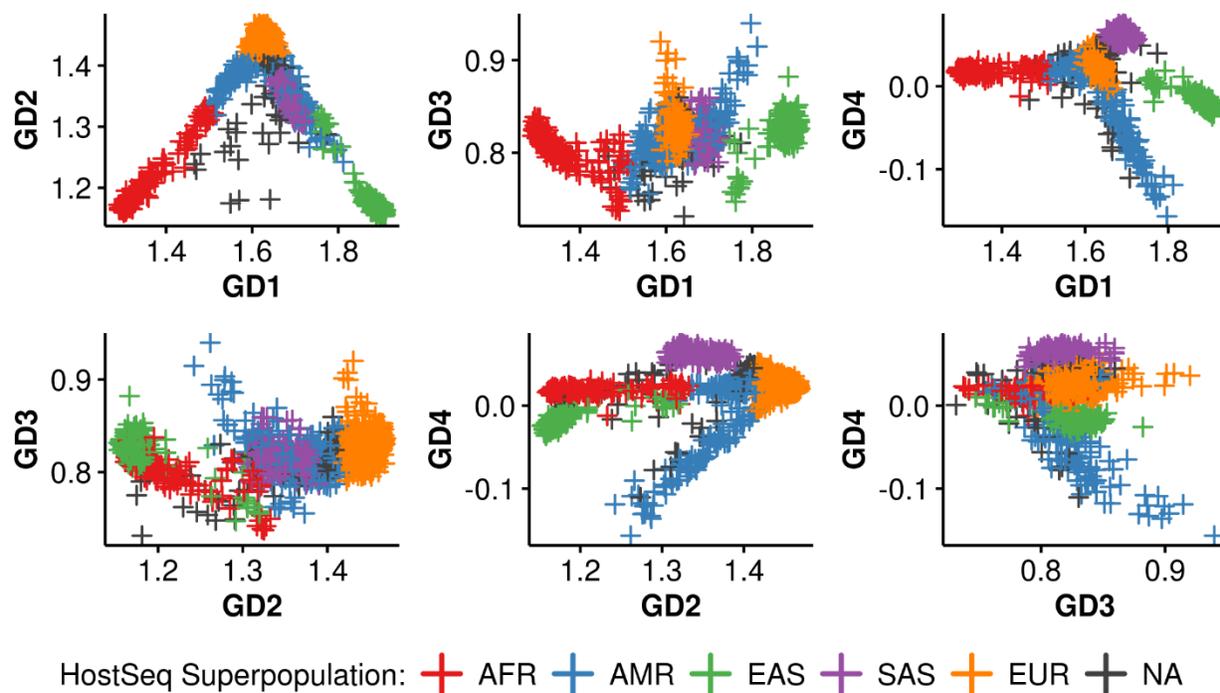


Figure S2. Predicted population admixture and ancestry classification in HostSeq genomes. Each bar represents a genome. Proportion of African, East Asian and European ancestries is determined (according to their definitions in GRAF), and genomes classified into 8 ancestry groups using GRAF-pop (see Methods). They are further categorized into 5 superpopulations: (i) AFR - African and African-American, (ii) AMR - Latin American Asian and Latin American African, (iii) EAS - Asian-Pacific Islander and East Asian, (iv) SAS - South Asian, and (v) EUR - European. 2% of genomes remain uncategorized.



*Figure S3. Genetic distances score of HostSeq genomes. The four genetic distances (GD1-4) scores from GRAF-pop (see Methods) represent distance of each genome from several reference populations and are used to predict ancestry. Barycentric coordinates of GD1 and GD2 are used to predict admixture proportion of African, East Asian and European ancestries.*