BMC Genomic Data

DATABASE

Open Access

HostSeq: a Canadian whole genome sequencing and clinical data resource

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

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 for health research 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

† S Yoo and E Garg contributed equally.

*Correspondence:
LJ Strug
lisa.strug@utoronto.ca
[1] The Hospital for Sick Children, Toronto, ON, Canada
[2] University of Ottawa, Ottawa, ON, Canada
[3] Simon Fraser University, Burnaby, BC, Canada
[4] University of Toronto, Toronto, ON, Canada
[5] Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON, Canada
[6] McGill University, Montreal, QC, Canada
[7] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada
[8] University of Waterloo, Waterloo, ON, Canada
[9] Sinai Health System, Toronto, ON, Canada
[10] Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada
[11] University of Calgary, Calgary, AB, Canada
[12] Alberta Children's Hospital, Calgary, AB, Canada
[13] University of British Columbia, Vancouver, BC, Canada
[14] BC Children's Hospital, Vancouver, BC, Canada
[15] St. Paul's Hospital, Vancouver, BC, Canada
[16] University Health Network, Toronto, ON, Canada
[17] The Ottawa Hospital Research Institute, Ottawa, ON, Canada
[18] Queen's University, Kingston, ON, Canada
[19] St. Michael's Hospital, Unity Health, Toronto, ON, Canada
[20] Women's College Hospital, Toronto, ON, Canada
[21] University of Alberta, Edmonton, AB, Canada

Background

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]. A growing body of work demonstrates a role for host genetics in SARS-CoV-2 [10–14]. Despite the relative novelty of the SARS-CoV-2 virus and the challenges of identifying genetic contributors in a changing environment [2], several loci contributing to infection susceptibility and illness severity have been identified [15]. Associated loci are comprised of rare and common variations and occur throughout the genome, including but not limited to chromosome X and the HLA region on chromosome 6.

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 [16, 17], 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; [18]). 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; [19]) 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.

Construction and content

HostSeq project design

HostSeq (Fig. 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 [20]. 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 for 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), as described below in the *Availability of Data and Materials* section.

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

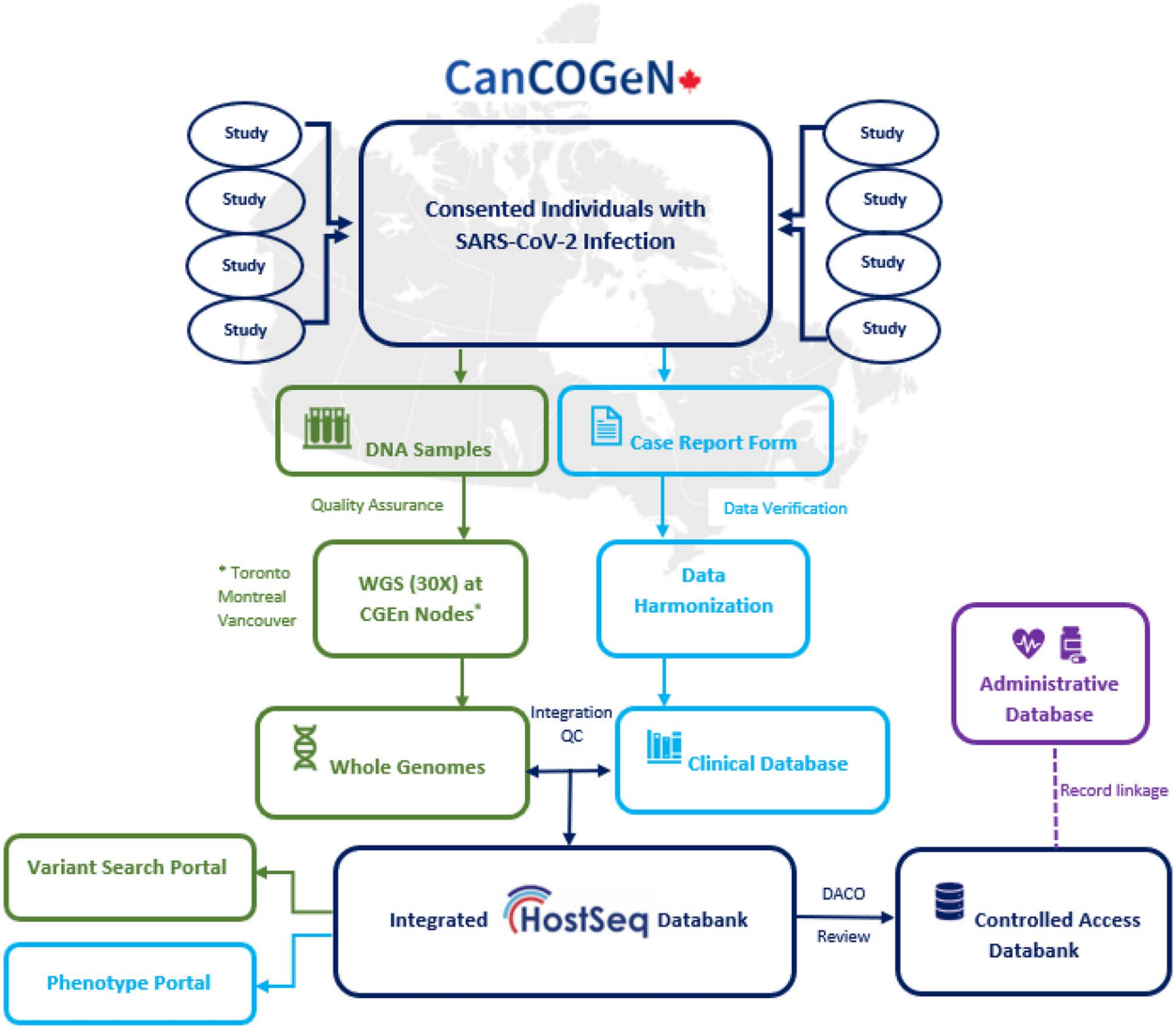

Fig. 1 Sample and data flow in HostSeq (Aspects of graphics acquired from Wikimedia Commons)

NovaSeq6000 platform at 30X depth. Prior to sequencing, quality assurance is performed at multiple stages throughout the process [21]. Concordance of the genotyping pipeline among sequencing sites is verified using the Ashkenazi trio set from the Genome in a Bottle Consortium [22].

Sequenced samples are analyzed jointly using an in-house pipeline encoded in Nextflow [23] and Snakemake [24], containerized using Docker [25]. The Genome Reference Consortium human build 38 (GRCh38 assembly version GCA_000001405.15) reference genome that includes the alternative HLA decoy genes[1] is used. Genomes are processed following the Best Practices guidelines of the Genome Analysis ToolKit (GATK v4.2.5.0). This includes alignment of sequences to the reference genome, and the genotyping of each sample individually followed by joint-calling of all genotypes together. Associated scripts can be found in a public repository (https://svn.bcgsc.ca/bitbucket/users/jmgarant). Software packages used to process and analyze the WGS data are listed in Table S3.

The in-house pipeline is as follows. Sequences are aligned to the reference genome using DRAGEN mapper (DRAGMAP v1.3.0; [26]), 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

[1] https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/

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; [27]); perform housekeeping with bcftools (v1.11) and samtools (v1.14; [28]); check for sample contamination using VerifyBamID2 (v2.0.1) [29]; check agreement between reported sex-at-birth and sex chromosome composition using PLINK software (v1.90; [30]); and predict ancestry admixture [31] and relatedness [32] using Genetic Relationship and Fingerprinting software (GRAF v2.4). We use PLINK (v2.00; [33]) and R (3.6.3; [34]) for genetic data analysis. Additionally, we compare the genetic principal components of HostSeq with the 1000 Genomes Project reference populations [35, 36] following the guidelines of plinkQC [37]. 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 provided in joint VCF format (aligned sequences can also be obtained).

Contributing studies and data harmonization

As of December 20, 2022, 13 participating studies contributed data and biospecimens to HostSeq (Table S4). Although all 13 studies continue collecting clinical information, 6 have completed their participant recruitment. To date, we have harmonized data from all 13 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 13 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 matched with controls (see Table S4 for matching factors and control eligibility). One study—Quebec COVID-19 Biobank ("BQC19")—collected clinical data and biospecimens from 12 hospitals in Quebec [38]. 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.

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 (available through the *HostSeq Phenotype Portal* described below in *HostSeq Data Portals*). For example, all laboratory test variables are converted into predefined units; text entries in French are translated into English; and medications and complications variables are coded by timeline (prior to illness vs. during illness vs. post-discharge follow-up). Any potential data errors detected in the harmonization process are communicated to the participating study teams and resolved through follow-up.

Study-specific sample sizes currently range from 11 to 4,602. To date, in the HostSeq databank the 13 studies have contributed 9,913 clinical records and submitted 10,978 samples (Table 1). With the exception of two studies that have recruitment across multiple provinces (CANCOV, CONCOR-Donor; $n = 2{,}196$), most studies are province-specific: six studies in Ontario (GENCOV, GenOMICC, SCB, LEFT-GEN, genMARK, Understanding Immunity to Coronaviruses; $n = 3{,}114$), one in Quebec (BQC19; $n = 4{,}602$), two in Alberta (AB3C; AB-HGS $n = 262$) and two in British Columbia (GD-COVID, Host Factors; $n = 804$). Table S4 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).

Results

Clinical data summary

The results discussed in this section are based on approximately 95% 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 9,427 currently available harmonized samples, HostSeq has 54.6% females and 41.5% males (and the remaining 3.9% are missing reported sex-at-birth), with an overall mean age (at recruitment) of 47.9 years. Distributions of sex and age vary across the studies (Table S5). Apart from studies including pediatric participants (AB3C, SCB), mean age in the studies ranges from 36.9 years (genMARK) to 63.5 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

Table 1 Status of DNA sample sequencing (as of December 20, 2022)

STUDY TITLE	STUDY ACRONYM	SAMPLES	DATA
The Hospital for Sick Children's COVID-19 Biobank	SCB	566	223
Genetic Markers of Susceptibility to COVID-19	genMARK	876	738
The Canadian COVID-19 Prospective Cohort Study	CANCOV	1409	1,284
The Genetics of Mortality in Critical Care	GenOMICC	328	331
Implementation of Serological and Molecular Tools to Inform COVID-19 Patient Management	GENCOV	1,290	1,111
Convalescent Plasma for COVID-19 Research	CONCOR-Donor	787	787
Host Genetic Factors Underlying Severe COVID-19	Host Factors	11	11
The Quebec COVID-19 Biobank	BQC-19	4,602	4,323
Genomic Determinants of COVID-19: Integration of Host and Viral Genomic Data to Understand the COVID-19 Epidemiologic Triangle	GD-COVID	793	793
HostSeq—Canadian COVID-19 Human Host Genome Sequencing Ottawa	LEFT-GEN	43	43
Understanding Immunity to Coronaviruses to Develop New Vaccines and Therapies against 2019-nCoV		11	10
Alberta Childhood COVID-19 Cohort Study	AB3C	188	188
Host Genetic Susceptibility to Severe Disease from COVID-19 Infection	AB-HGS	74	71
TOTAL		**10,978**	**9,913**

SAMPLES column indicates DNA samples submitted to HostSeq for sequencing. DATA column indicates raw clinical records submitted to HostSeq. Of these, a total of 9,427 records have been harmonized

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, 54.0% were discharged home, 15.0% were transferred to other hospitals or healthcare settings (e.g., rehabilitation centers or long-term care facilities) and 11.9% were reported deceased (Table 2).

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 ≥ 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

Table 2 Hospitalization and patient outcomes in HostSeq

HostSeq All Studies (*n* = 9,427)[a]		
Age[b]	Mean (SD)	47.9 (29.1)
	Median (min, Q1, Q3, max)	48.0 (0.1, 33.4, 61.9, 104.2)
Sex at birth[c]	Male	3,911 (41.5%)
	Female	5,144 (54.6%)
Hospitalization[d]	Yes	3,478 (36.9%)
	No	5,354 (56.8%)
ICU admission[e]	Yes	1,148 (12.2%)
	No	6,558 (69.6%)
Patient outcome[f,g]	Discharged alive	1,879 (54.0%)
	Transfer to another facility	521 (15.0%)
	Palliative discharge	3 (0.1%)
	Hospitalized	9 (0.3%)
	Death	414 (11.9%)

SD Standard deviation; [a] *n* = 9,427 is a subset of the expected cohort of greater than 10,000; [b] Data not available for 605 participants (6.4%); [c] Data not available for 372 participants (3.9%); [d] Data not available for 595 participants (6.3%); [e] Data not available for 1,721 participants (18.3%); [f] Denominator is 3,478 hospitalized participants; [g] Data currently in collection for 652 participants (18.7%)

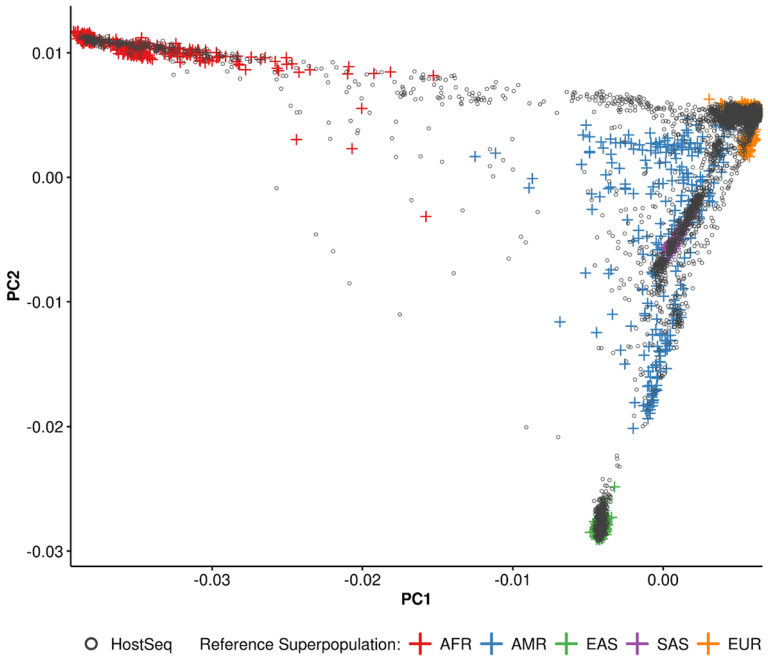

Fig. 2 PCA projection of HostSeq genomes against reference superpopulations. HostSeq genomes were merged with the 1000 Genomes reference set. The first two principal components of this merged data are shown here with HostSeq genomes in black and 1000 Genomes samples colored by their superpopulation: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

is supported by the CanDIG-server [39], a platform enabling federated querying of genomics data. Beacon APIs [40] 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 6,500 HostSeq genomes, of which 6,316 passed all quality checks (see *Methods*). Our predicted population structure covers five major ancestry groups (Figs. 2 and S2, S3; 69% European, 6% Admixed American, 8% East Asian, 8% South Asian, 6% African, and approximately 3% uncategorized) and closely matches self-reported ancestries (where available). Additionally, there are 300 and 518 pairs of first- and second-degree relationships, respectively.

Currently HostSeq provides 174.5 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* in 4436 genomes are 73, 145 and 49, respectively (the most common alleles per locus are *HLA-A**02:01, *HLA-B**07:02 and *HLA-C**07:01).

Utility and 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 heterogenous with different designs and objectives (Table 3 and Table S4), 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.

Individual or combined analysis

Whether an investigator's research question would be best answered by within-study comparisons or analyses

Table 3 Aspects of participant ascertainment in HostSeq

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

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

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 be used.

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 [41], meta-analysis of summary data can be less efficient in smaller 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 [42, 43]. Furthermore, with study or environmental factors and other sources of heterogeneity, joint analysis can exploit gene-environment interaction [44] 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 [45]. 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 [46–48]. 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 [49, 50]. 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 [51, 52]. 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 [53]. 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 [54]. 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 [52, 55], 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 [56–58]. 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 [19].

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 [59]). Yet, it has been reported that close to one-third of the X chromosome genes can escape X-inactivation [60, 61]; 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 [62, 63] and an easy-to-implement regression model [64]. Rare X-chromosome variant analysis [65, 66] 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.

Conclusions

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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-023-01128-3.

Additional file 1: 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. **Table S2.** HostSeq Case Report Form. **Table S3.** Software used for processing WGS data. **Table S4.** List of HostSeq participating studies as described in respective protocols. **Table S5.** Distribution of sex and age across HostSeq studies ($n = 9,427$). SD: Standard deviation; IQR: interquartile range. **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, and genomes classified into 8 ancestry groups using GRAF-pop. They are further categorized into 5 superpopulations: AFR - African and African-American, AMR - Latin American Asian and Latin American African, EAS - Asian-Pacific Islander and East Asian, SAS - South Asian, and EUR - European. 3% of genomes remain uncategorized. **Figure S3.** Genetic distances score of HostSeq genomes. The four genetic distances (GD1-4) scores from GRAF-pop represent the 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.

Acknowledgements

We wish to express gratitude to all HostSeq project participant studies and the individual participants within these studies for their contribution.

Authors' contributions

LJS led the study design and implementation. Data harmonization: SY; genomic data analysis: EG. Genetic epidemiology: LJS, JDB, SBB, LTE, FG, CMTG, RJH, JFL, ADP, LS. Writing: SY, EG, LTE, RJH, ARH, JDB, SBB, FG, CMTG, JFL, ADP, LS, LJS. Data processing and sharing: SY, EG, LTE, MLo, ARH, RJSA, IB, GB, J-MG, CG, JL, JW, BT, MSR, J-AH, OOA, SL, MHZ; SJMJ led the data processing and sharing. Study contributors: JL-E, UA, FPB, CMB, AMC, JC, MH, DMM, BPM, VM, SKM, MO, RSP, GP, OS, JT, SET, JU, RLW, RSMY. NA coordinated the three CGEn nodes; SET led study site recruitment; BMK designed the consent and data access process; MLa led the Quebec site; SJ led sequence informatics and the variant portal; NA, SWS and LJS provided overall study oversight. The author(s) read and approved the final manuscript.

Funding

PI, Affiliations	Study name	Grants (funder and details)
Stephen Scherer, Lisa Strug, The Hospital for Sick Children, Toronto, ON	**CGEn HostSeq—Canadian COVID-19 Human Host Genome Sequencing Databank**	Genome Canada, Innovation, Science and Economic Development Canada
Vincent Mooser, CGEn-Montreal, QC	**Biobanque Quebec COVID-19**	PHAC 2021-HQ-000051 FRQ-S MSSS
Rae Yeung, The Hospital for Sick Children, Toronto, ON	**SickKids COVID-19 Biobank**	CFI cost center # 6,220,200,122 (Proposal ID HSC0005268) CIHR/COVID-19 Immunity Task Force: CIHR GA4-177,739 CIHR MM1-181,123
Angela Cheung and Margaret Herridge, University Health Network, Toronto, ON	**The Canadian COVID-19 Prospective Cohort Study (CanCOV)**	Canadian Institutes of Health Research (CIHR), COVID-19 Rapid Research Funding Opportunity—Clinical Management and Health System CIHR/COVID-19 Immunity Task Force: Grant number: 447643
Jordan Lerner-Ellis, Jennifer Taher, Sinai Health, Toronto, ON	**Implementation of serological and molecular tools to inform COVID-19 patient management (GENCOV)**	CIHR #VR4-172,753 CIHR sub-awards: # 461,170 and #461,304
Rulan Parekh, The Hospital for Sick Children, Toronto, ON	**Adaptive Immunity and Outcomes of Convalescent Plasma**	Ministry of Colleges and Universities (Ontario COVID-19 Rapid Research Fund)
Francois Bernier, University of Calgary, Calgary, AB	**Alberta Childhood COVID-19 Cohort (ABCCC)**	Genome Alberta (RRP2) Alberta Children's Hospital Mitogen DX
Upton Allen, The Hospital for Sick Children, Toronto, ON	**COVID-19 genMARK study**	SickKids Foundation University of Toronto # 508,791

PI, Affiliations	Study name	Grants (funder and details)
Stuart Turvey, BC Children's Hospital, Vancouver, BC	**Genomic determinants of COVID-19**	Genome British Columbia COV199
David Maslove, Queens University, Kingston, ON	**Genetics of Mortality in critical care (GenOMICC)**	Ontario Innovation Fund Innovation Grant administered by the Southeastern Ontario Academic Medical Organization (SEAMO)
Catherine Biggs, Stuart Turvey, BC Children's Hospital, Vancouver, BC	**Improving outcomes through precision medicine for adults with primary immunodeficiencies**	Providence Healthcare Research Institute
Mario Ostrowski, St. Michael's Hospital, Unity Health, Toronto, ON	**Understanding Immunity to Coronaviruses to Develop New Vaccines and Therapies against 2019-nCoV**	CIHR VR1-172,711
Gerald Pfeffer, University of Calgary, Calgary, AB	**Host Genetic Susceptibility to Severe Disease from COVID-19 Infection**	Hotchkiss Brain Institute, University of Calgary Cumming School of Medicine, University of Calgary

Availability of data and materials

The datasets generated and analysed during the current study are made available to researchers worldwide through a Data Access Agreement and Data Access Compliance Office (DACO) approval (https://www.cgen.ca/daco-main). The datasets are deposited in the HostSeq Databank, which is a data repository that facilitates data access controls that are suitable for hosting sensitive health data. Access to this repository is granted to any researcher with DACO approval. The DACO verifies that the proposed research has REB approval from their host institution and conforms to HostSeq's REB-approved SARS-CoV-2 or 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.

Aggregated data are publicly available through two data portals: a phenotype portal showing summaries of major variables (https://hostseq.ca/phenotypes.html) and their distributions, and a variant search portal enabling queries in a genomic region (https://hostseq.ca/dashboard/variants-search). Access to the variant search portal requires a login (any researcher can register for a login to the variant search portal).

The code used for processing the WGS data can be found in a publicly accessible repository: https://svn.bcgsc.ca/bitbucket/users/jmgarant.

Declarations

Ethics approval and consent to participate

HostSeq was approved by the Research Ethics Board of the Hospital for Sick Children (lead site) (#1000070720 from 2020-present). Written informed consent was obtained from all participants or parents/guardians/substitute decision makers prior to inclusion in the study.

Additional REB information from the participating PIs:

Site	PI	REB
The Hospital for Sick Children	Stephen Scherer	1000070720
The Hospital for Sick Children	Rae Yeung	1000070060
The Hospital for Sick Children	Upton Allen	1000069580
University Health Network	Angela Cheung	CTO ID 2157
Queens University	David Maslove	CTO ID 3209
Mount Sinai Hospital	Jordan Lerner-Ellis	CTO ID 3302
The Hospital for Sick Children	Rulan Parekh	1000070462
BC Children's Hospital	Catherine Biggs	H20-01667
University of Montreal Health Centre		19.389 (internal) MP-02–2020-8929 (multicentre)
BC Children;s Hospital	Stuart Turvey	H21-00054
The Ottawa Hospital	Juthaporn Cowan	20210119-01H
Unity Health	Mario Ostrowski	20–044
University of Calgary	Gerald Pfeffer	20-0574_MOD3
University of Calgary	Francois Bernier	20-0480_MOD6

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 June 2022 Accepted: 22 February 2023
Published online: 02 May 2023

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. CCR5Δ32 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 in hibit 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;607(7917):97–103.
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. COVID-19 Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. Nature. 2022;608(7921):E1-E10.
15. Niemi MEK, Daly MJ, Ganna A. The human genetic epidemiology of COVID-19. Nat Rev Genet. 2022;23(5):533–46.
16. 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.
17. 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):E710–7.
18. 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.
19. COVID-19 Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. Nature. 2022;608(7921):97–103.
20. 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.
21. Corbett R, Eveleigh R, Whitney J, Barai N, Bourgey M, Chuah E, et al. A distributed whole genome sequencing benchmark study. Front Genet. 2020;11:612515.
22. Zook J, Catoe D, McDaniel J. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3(1):1–26.
23. Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9.
24. Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. F1000Research. 2021;10:33.
25. Van der Auwera G, O'Connor B. Genomics in the cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media; 2020.
26. Illumina, Inc. DRAGMAP. 2019. [Accessed Summer 2022]. Available from: https://github.com/Illumina/DRAGMAP.
27. 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.
28. Danecek P, Bonfield J. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):008.
29. 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.
30. 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.
31. Jin Y, Schaffer A, Feolo M, Holmes J, Kattman B. GRAF-pop: A fast distance-based method to infer subject ancetry from multiple genotype datasets without principal components analysis. G3 Bethesda Md. 2019;9(8):2447–61.
32. 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.
33. 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):13742–815.
34. R Core Team. R: A language and environment for statistical computing. 2022. Available from: https://www.r-project.org/.
35. Roslin N, Weili L, Paterson A, Strug L. Quality control analysis of the 1000 Genome Project Omni2,5 genotypes. bioRxiv. 2016. https://doi.org/10.1101/078600v1.
36. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
37. Meyer HV. meyer-lab-cshl/plinkQC: plinkQC 0.3.2. 2020. Available from: https://meyer-lab-cshl.github.io/plinkQC/.
38. 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):e0245031.
39. 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.
40. 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.
41. Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol. 2009;33(3):256–65.
42. 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.
43. 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.
44. 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.
45. Griffith G. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun. 2020;11(1):1–12.
46. 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.
47. 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.
48. Huang B, Lin D. Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet. 2007;80:567–76.
49. Monsees G, Tamimi R, Kraft P. Genome-wide association scans for secondary traits using case-control samples. Genet Epidemiol. 2009;33(8):717–28.
50. 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.
51. 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.
52. 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.
53. 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(8):683–93.
54. 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.
55. 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.
56. Peckham H, de Gruijter N, Raine C, Radziswska 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.
57. 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.
58. Pradhan A, Olasson PE. Sex differences in severity and mortality from COVID-19: Are males more vulnerable? Biol Sex Differ. 2020;11:53.
59. 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.

60. 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.
61. Lee S, Wu M, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4):762–75.
62. Wang J, Talluri R, Shete S. Selection of X-chromosome inactivation model. Cancer Inform. 2017;16:1–8.
63. 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.
64. 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.
65. 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.
66. 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.

Publisher's Note