

CORRESPONDENCE

Open Access

# Statistical power in COVID-19 case-control host genomic study design



Yu-Chung Lin<sup>1,2</sup>, Jennifer D. Brooks<sup>1</sup>, Shelley B. Bull<sup>1,3</sup>, France Gagnon<sup>1</sup>, Celia M. T. Greenwood<sup>4,5</sup>, Rayjean J. Hung<sup>1,2</sup>, Jerald Lawless<sup>6</sup>, Andrew D. Paterson<sup>1,2</sup>, Lei Sun<sup>1,7</sup>, Lisa J. Strug<sup>1,2,7,8\*</sup>  On behalf of the Genetic Epidemiology Committee of the Canadian COVID Genomics Network (CanCOGeN) HostSeq Project

## Abstract

The identification of genetic variation that directly impacts infection susceptibility to SARS-CoV-2 and disease severity of COVID-19 is an important step towards risk stratification, personalized treatment plans, therapeutic, and vaccine development and deployment. Given the importance of study design in infectious disease genetic epidemiology, we use simulation and draw on current estimates of exposure, infectivity, and test accuracy of COVID-19 to demonstrate the feasibility of detecting host genetic factors associated with susceptibility and severity in published COVID-19 study designs. We demonstrate that limited phenotypic data and exposure/infection information in the early stages of the pandemic significantly impact the ability to detect most genetic variants with moderate effect sizes, especially when studying susceptibility to SARS-CoV-2 infection. Our insights can aid in the interpretation of genetic findings emerging in the literature and guide the design of future host genetic studies.

**Keywords:** Genetic epidemiology, Statistical genetics, Study design, Genome-wide association studies

## Background

Variability in susceptibility and disease severity in response to viral exposure is a hallmark of infectious diseases [1]. This variation is partially due to a complex interaction between the pathogen and the host genome, and it is impacted directly or indirectly by factors specific to both the pathogen (e.g., viral load, viral genotype) and the host (e.g., sex, pre-existing medical conditions, race) [2]. Identifying host genetic factors responsible for this variation can pinpoint susceptible populations, inform personalized treatment, and guide vaccine development [3]. SARS-CoV-2, the virus that causes coronavirus disease 2019 (COVID-19), appears to be no exception [4]. By Spring 2020, several international research groups

[5] were proposing or implementing studies to identify the host genetic factors that influence (i) *susceptibility to infection* with SARS-CoV-2 [5] and (ii) *disease severity* of COVID-19 once infected [5]. These include sequencing and array-based studies that investigate rare [6] and common [5, 7] genetic variants.

Numerous case-control studies of varied designs have emerged to better understand the host genetic factors associated with susceptibility to SARS-CoV-2 and severity of COVID-19. Due to the fast-paced nature of the pandemic, one major challenge is how to best utilize existing resources while accounting for the limitations of the available phenotype and genotype data. The study designs proposed to date include prospective case-control recruitment designs where array-based genotyping or sequencing is planned [6, 8], retrospective designs using existing genotyped cohorts with concurrent phenotyping for COVID-19 [9, 10], and hybrid designs using prospectively collected cases and population controls with genotype data readily available [7]. Each design has implications for

\* Correspondence: [lisa.strug@utoronto.ca](mailto:lisa.strug@utoronto.ca)

<sup>1</sup>Dalla Lana School of Public Health, University of Toronto, Room 500, 155 College St, Toronto, ON M5T3M7, Canada

<sup>2</sup>Program in Genetics and Genome Biology, The Hospital for Sick Children, Room 12.9801, 686 Bay Street, Toronto, ON M5G0A4, Canada

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

how the susceptibility and disease severity phenotypes are defined [11] and may have associated limitations that are not surmountable given the current state of the pandemic.

The field of infectious disease genetic epidemiology has long described the unique and complex implications that a requirement of exposure to the infectious agent imposes on study design [12]. Here, we draw on current estimates of exposure, infectivity, and test accuracy of COVID-19 to demonstrate the feasibility of detecting host genetic factors associated with COVID-19 susceptibility and severity with current study designs. We demonstrate why studying susceptibility to SARS-CoV-2 infection could be futile at certain stages of the pandemic. Our insights can aid in the interpretation of genetic findings emerging from the literature and guide the design of future host genetic studies.

## Results

### Identifying host genetic factors of susceptibility to infection with SARS-CoV-2

An important objective of host genetic studies is to determine whether there are rare or common genetic variants that modify the risk of SARS-CoV-2 infection upon exposure. Examples of the existence of such variants for other infectious diseases are sparse in the literature [3], although for SARS-CoV-2, one recent study estimated 50% heritability for infection susceptibility [4]. Case-control studies to identify such genetic contributors to SARS-CoV-2 infection have, for practical reasons, defined *cases* as individuals with laboratory confirmed SARS-CoV-2 infection and *controls* as those who tested negative [11, 13]. Given low population exposure rates at the early stage of the pandemic [14], the lack of data on an individual's exposure to SARS-CoV-2 [3], and high infectivity [15], many of these "controls" will be misclassified since it is not certain whether they were actually exposed to the pathogen and would have been infected had they been exposed. Imperfect test accuracy also contributes to misclassification. For example, concerns have been raised about false-negative reverse transcription polymerase chain reaction (RT-PCR) tests in patients showing apparent COVID-19 illness; sometimes, up to 40% of samples were deemed falsely negative [16]. False-positive tests, although less prevalent due to the high specificity of RT-PCR tests [17], lead to misclassification of cases, which can also be problematic if disease prevalence is low. For hypothesis testing in genetic studies, the result of misclassification of either cases or controls is reduced statistical power for detecting common and rare genetic variants associated with susceptibility to infection with SARS-CoV-2 (Fig. 1).

We used simulations to demonstrate the factors that impact statistical power to identify genetic variants associated with SARS-CoV-2 infection susceptibility. It

is assumed that controls are selected from persons testing negative for the virus. In addition to sample size, effect size of a variant (odds of being infected relative to being unaffected; OR) and its minor allele frequency (MAF), critical factors include rates of population-level exposure, test accuracy, and baseline infection susceptibility (Fig. 1 and Additional file 1: Supplementary Methods). We assume that the majority of the population will be infected when exposed, except those who carry protective alleles [19, 20] which lead to reduced infection susceptibility upon exposure. Therefore, the power calculation is estimated based on a range of ORs less than 1, without loss of generality.

In the simulation, the number of genetic risk variants each individual carries is determined by the MAF under Hardy-Weinberg equilibrium (HWE). Here, we specified a rare (MAF = 0.01) and common (MAF = 0.2) allele. Risk of SARS-CoV-2 infection susceptibility upon exposure varies depending on the number of risk variants a person carries and the effect size specified (OR = 0.2, 0.5, 0.7 investigated here). Population-level exposure (specified to be 5%, 20%, and 40%) and the corresponding test accuracies (80%, 90%, and 100% sensitivity) are treated as constants and used to simulate various plausible real-world scenarios. Population-level exposure, for instance, determines the proportion of simulated individuals exposed to SARS-CoV-2 and impacts the power of different study designs. Logistic regression is performed, regressing SARS-CoV-2 infection status on the number of protective alleles, and an association between the genetic variant and SARS-CoV-2 infection susceptibility is concluded in the simulation if the corresponding  $p$  value is below the genome-wide significance level ( $p < 5e-8$ ). Further simulation details are provided in Additional file 1: Supplementary Methods.

Figure 1a demonstrates that under perfect test accuracy, not only is it difficult to detect rare variants with very large protective effects (OR = 0.2; MAF = 0.01; blue), it is also extremely challenging to detect common variants with moderate protective effects (OR = 0.7; MAF = 0.2; green) using sample sizes reported in existing genetic studies (e.g., > 2000 cases [7]). Moreover, low population-level exposure (5%; solid line Fig. 1a) reduces statistical power, as many of the "controls" would have been infected if exposed to SARS-CoV-2, contributing to higher misclassification rates. Given low population exposure rates even in the hardest-hit regions [21], our chances of detecting a true association with infection susceptibility remain slim. The curation of better-defined controls through careful selection of those with a high probability of exposure to SARS-CoV-2 (such as health care workers or household members of positive cases), which is analogous to an increasing population-level exposure, leads to improved power to detect associated genetic variants.

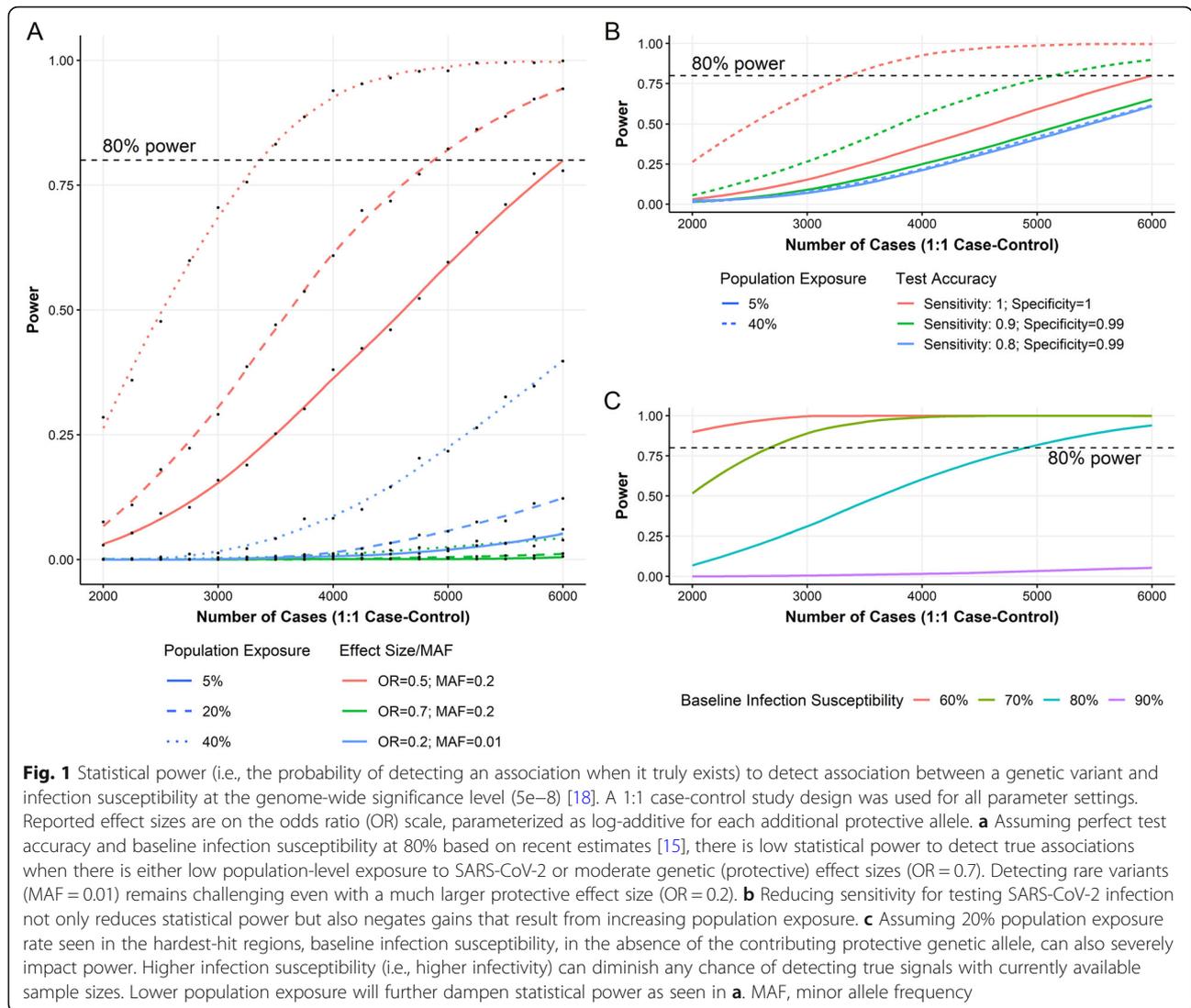


Figure 1b demonstrates that false-negative tests caused by low sensitivity of SARS-CoV-2 infection testing [16] can dramatically reduce statistical power. Increasing the number of false-negative tests (sensitivity = 0.9, green; sensitivity = 0.8, blue) leads to much reduced study power even when the genetic variant of interest is a common variant with large genetic (protective) effect (OR = 0.5, MAF = 0.2; Fig. 1b). Increasing population-level exposure or using better-defined controls can improve power (Fig. 1a), but these power gains are no longer guaranteed when coupled with an inaccurate, low-sensitivity test for SARS-CoV-2 infection. Increasing population exposure does not result in higher statistical power when sensitivity = 0.8 (blue curve; Fig. 1b) since the test misclassifies many cases as controls and offsets any power gain that would be realized by carefully selecting exposed individuals as controls. With lower test sensitivity (sensitivity = 0.7), we show that increasing population-level exposure is

detrimental to the study power for disease susceptibility (Additional file 2: Figure S1).

Figure 1c shows that, combined with low exposure rate, baseline infection susceptibility given exposure also affects power, in a manner that is inversely proportional to the baseline infection rate. For example, assuming a population-level exposure rate of 20% as seen in the hardest-hit geographic regions [22], a high baseline infection rate (e.g., 90%, the purple curve in Fig. 1c) diminishes the chance of detecting an associated variant with OR = 0.5 and MAF = 0.2 in a case-control study with 6000 cases and 6000 “controls.” This is because increasing baseline infection susceptibility results in more people infected *if exposed* to SARS-CoV-2, which, combined with low exposure rate, leads to a higher misclassification rate of controls and lower power (e.g., 5268 out of 6000 “controls” are misclassified assuming 20% population exposure and 90% baseline infection rate;

Additional file 1: Supplementary Methods). The loss in statistical power is exacerbated by the current (fortunately) low population-level exposure to the SARS-CoV-2 infection. With the sample sizes specified, detecting a true genetic association with SARS-CoV-2 infectivity appears achievable only if the infection susceptibility is low. However, recent studies have shown >80% infectivity for individuals over age 30 [15], suggesting the blue (baseline infection susceptibility 80%) or purple (baseline infection susceptibility 90%) curves in Fig. 1c more closely reflect the current reality.

The lack of information on individual-level exposure, coupled with low population-level exposure rates, high infectivity, and inaccurate, low-sensitivity tests, make studies designed to identify genetic variation contributing to SARS-CoV-2 infection susceptibility largely infeasible. We see that a larger study cohort, with misclassification, is not guaranteed to yield higher statistical power than a smaller dataset with well-defined cases and controls, as noted elsewhere [23]. Our simulation results indicate that careful selection of controls with a high probability of exposure (e.g., frontline workers or household members of positive cases) rather than sampling test-negative controls is a preferable strategy. Given the negative effect on power from using a low-sensitivity test, it is also preferred to include individuals who were monitored or tested multiple times in a given timeframe; such a design can potentially save costs and yield greater insights in understanding genetic susceptibility to SARS-CoV-2 infection, even with a smaller sample size.

#### Identifying host genetic factors of disease severity of COVID-19

Another major objective of host-genetic studies is to identify genetic variants associated with severity of symptoms given SARS-CoV-2 infection. Such studies may focus on severe respiratory disease [7] or alternative phenotypes identified prospectively as researchers learn more about the sequelae of infection [24]. Ideally, cases should include infected patients who have the defined phenotype and controls should be comprised of infected patients with mild or no symptoms [3, 25]. One commonly used definition of disease severity is hospitalization [11], which we use here for simulating case-control status. However, given the time, effort, and cost of recruiting and genotyping infected but non-hospitalized individuals, along with the urgent need for genetic insights, alternative designs that leverage matched controls in the general population have been proposed and implemented [7, 11]. Contrary to the low population exposure rate that drives case-control misclassification in SARS-CoV-2 infection susceptibility studies, it is the low population infection rate (population exposure rate  $\times$  baseline infection susceptibility) and the lack of information on individual-level

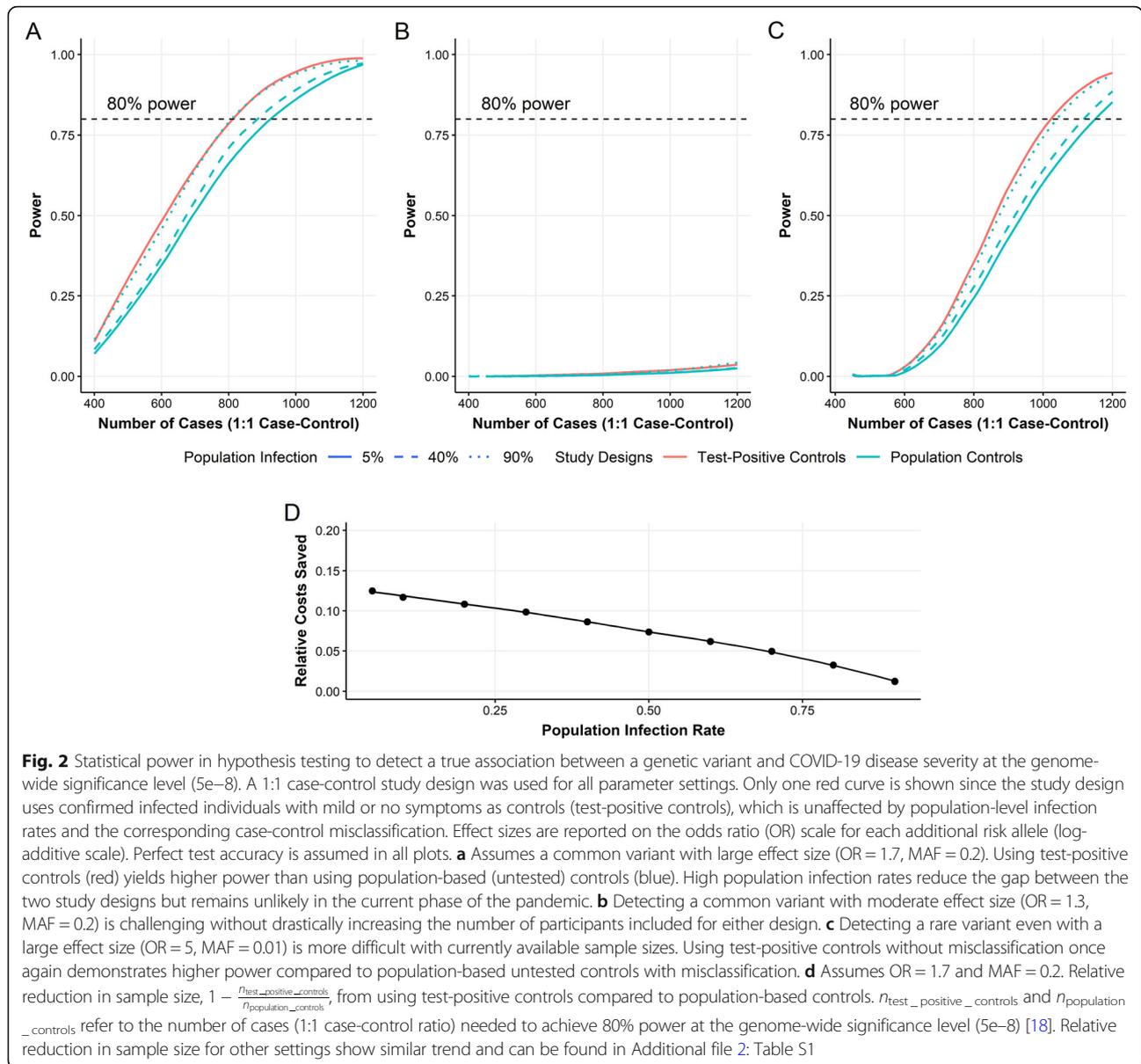
infection that lead to misclassification in disease severity studies; some of the “controls” would have developed severe symptoms had they been infected with SARS-CoV-2.

Here, we examine the gain in statistical power and efficiency (in terms of requiring smaller sample size) in hypothesis testing when using test-positive controls over a design that samples controls from an untested, available population sample.

The number of genetic risk variants each individual carries is again determined by the MAF specified; here, a rare (MAF = 0.01) and common (MAF = 0.2) allele investigated under Hardy-Weinberg equilibrium (HWE). Risk of COVID-19 disease severity upon infection varies depending on the number of risk variants a person carries and the effect size (the odds of severe over mild disease; OR). Population-level infection (varied to be 5%, 40%, and 90%), baseline risk of hospitalization (5%), and test accuracy (80%, 90% and 100% sensitivity) are specified as constants and varied to simulate plausible real-world scenarios. Population-level infection determines the proportion of simulated individuals infected with SARS-CoV-2, while the baseline risk of hospitalization dictates the probability of developing severe symptoms upon infection for individuals without the genetic risk variants. Logistic regression is performed, regressing hospitalization on the number of risk alleles, and an association between the genetic variant and COVID-19 disease severity is concluded if the corresponding  $p$  value is below the genome-wide significance level ( $p < 5e-8$ ). Further simulation details are provided in Additional file 3: Supplementary Methods.

Using test-positive controls (Fig. 2; red) yields higher power than using population-based (untested) controls (Fig. 2; blue) and can achieve substantial savings in genotyping costs. Power, while assuming perfect test accuracy, was investigated for a common variant with moderate effect size (OR = 1.7; MAF = 0.2; Fig. 2a), a common variant with small effect size (OR = 1.3; MAF = 0.2; Fig. 2b), and a rare variant with large effect size (OR = 5; MAF = 0.01; Fig. 2c).

The common (risk) variant effect sizes in Fig. 2a, b were chosen to reflect those recently reported in genetic studies of COVID-19 [7]. Only one red curve is shown since the study design uses test-positive individuals with mild symptoms as controls, which is not influenced by population-level infection rates. Thus, there is no misclassification attributed to infection rates. As expected, the higher red curve demonstrates that severity phenotyping in the absence of misclassification in a case-control study yields higher power than using controls with misclassification from the general population. Increases in population infection rates effectively closes the power gap between the two study designs since most population-based controls would have been infected and



hospitalized if necessary, thus reducing misclassification. However, as serological tests have shown low population infection rates (e.g., 5% in Spain by May 11, 2020) even in the hardest-hit regions [21], we expect the gap in power demonstrated in Fig. 2 to remain for the foreseeable future. This observation is most relevant for prospective studies: careful selection of individuals infected with SARS-CoV-2 as controls improves power and can reduce spending on genotyping.

The effect size of an associated risk variant can have, of course, a large impact on the power of the study design (Fig. 2a vs. b). Detecting a common variant with moderate effect size (OR = 1.3, MAF = 0.2) is challenging without drastically increasing the number of participants studied. In fact, the sample size required to reach 80%

power at the genome-wide significance level exceeds 4 times that required in Fig. 2a (Additional file 2: Figure S2). Similar patterns are seen when detecting a rare variant (Fig. 2c) as power is lower for such variants even for a large effect size (OR = 5, MAF = 0.01). To achieve the same power, the smaller sample size realized by selecting test-positive rather than population controls (red curve vs. blue curves in Fig. 2c) once again demonstrates the potential to save significant costs in sequencing given the low population infection rates in the current phase of this pandemic.

To quantify the costs saved by using controls with confirmed infection, we defined relative reduction in sample size as  $1 - \frac{n_{\text{test\_positive\_controls}}}{n_{\text{population\_controls}}}$ . Figure 2d illustrates

relative reduction in sample size with varying population infection rates when studying a common variant with a large effect size (OR = 1.7, MAF = 0.2); similar qualitative conclusions are gleaned as the parameters are varied (Additional file 2: Table S1). As expected, studies using controls with confirmed infection show the greatest benefit when disease prevalence is low. Given current levels of population infection rates at less than 10% [21], choosing controls with confirmed SARS-CoV-2 infection can save over 10% in genotyping costs. On the other hand, if genotype data for population controls are already available [9, 10, 26], ~ 1200 genotyped controls are required to detect a rare variant with very large effect (OR = 5.0, MAF = 0.01) given (currently) low population infection rate (5%). The required sample size rises to ~ 3800 for detecting a common variant with moderate effect (OR = 1.3, MAF = 0.2), and even more samples are needed if we allow the case-control ratio to decrease (Additional file 2: Figure S3). Given the sheer number of population controls available internationally, this seems like a reasonable and accessible strategy. Although beyond the scope of this article, this strategy is limited by the challenges of confounding and host-specific factors, which are often not widely available in population-based controls, leading to difficulties in interpreting genetic association results. When sample collection is planned prospectively, these results suggest a strategy of test-positive, matched controls for optimal power and interpretation.

## Discussion

Understanding how host genetic factors contribute to variation in disease susceptibility and severity can shed light on heterogeneity in the immune response and the host-pathogen interaction and facilitate the development of therapeutics and vaccines. Genetic studies have already been deployed in several countries to investigate SARS-CoV-2 infection susceptibility and COVID-19 disease severity. However, ongoing studies are hindered by a lack of information on SARS-CoV-2 exposure and infection at the individual level, low population exposure/infection rates, and inaccurate, low-sensitivity tests. In particular, given current population exposure rates, a design that uses test-negative controls will not provide adequate power to detect variants that protect from infection when exposed to SARS-CoV-2; this strategy leads to lower study power and greater difficulty for studying infection susceptibility than disease severity. A design that focuses on, for example, frontline workers such as healthcare professionals who are at high risk of exposure and are more likely to receive repeated testing would have much greater power. For case-control studies investigating the genetics of COVID-19 disease severity, collecting those who tested positive for SARS-CoV-2 infection rather than using existing population-based controls should be prioritized whenever feasible. Household

controls, which are easier to recruit and provide the added benefit of increased exposure probability, could serve as another effective selection of controls to enhance study power. Leveraging both genetically related and unrelated household controls can also disentangle contributions from genetic and environmental factors to SARS-CoV-2 infection susceptibility and COVID-19 severity. Using test-positive controls also provides the added benefit of prospective phenotypic observation to allow for future genetic studies of as-of-yet unknown phenotypic sequela of this disease, even in those who are mild or asymptomatic [24]. A prospective design also facilitates future investigation into the genetics of antibody levels and vaccine responsiveness, which is of major interest in ongoing vaccine development and deployment efforts.

Contrary to the SARS-CoV-2 infection susceptibility studies, we have assumed perfect test sensitivity and specificity for the COVID-19 severity study simulations. For study designs where controls have confirmed infection, low sensitivity (false negatives) that is characteristic of RT-PCR tests used for SARS-CoV-2 infection [16] is not a major consideration. Similarly, designs that use population-based controls without leveraging information on test results are also not affected by false-negative RT-PCR tests. Previous studies have reported near-perfect specificity for RT-PCR tests [27] which leads us to conclude that imperfect test accuracy has limited impact on host-genetic studies of COVID-19 disease severity, with the exception of meeting recruitment milestones.

Thus far, we have assumed RT-PCR tests were used to identify individuals infected with SARS-CoV-2 and that the exposure rate was constant across the sample. However, RT-PCR cannot indicate whether someone has been previously infected with SARS-CoV-2 due to its short detection period. Serology tests typically have longer detection periods and thus can be used to assess previous exposure to pathogens and contribute to a more refined selection of controls. It is worth noting that, as cases and controls are currently being selected from multiple sites for practical reasons, varying test accuracy and exposure rates over time and location would further impact study power.

Besides exposure, there are other confounders or mediators we have not addressed in this article which can further reduce power, diminishing our chances to detect variants that *directly impact* COVID-19 disease severity. That is, confounding could lead to the identification of genetic variants that impact, for example, comorbidities such as diabetes or hypertension, already known to put individuals in a higher risk category for severe disease. But these variants are likely not what would be most relevant to inform therapeutics and vaccine development. Variation associated with factors specific to the host rather than host-pathogen interactions is a

major concern for case-control studies using population-based convenience controls, rather than case-control studies in captive, highly exposed populations, such as front-line workers. One example of such confounding could be the ABO locus reported in [7], where blood donors were used as a control group and type-O blood may have been over-represented due to its universal donor status. Given the negative impact of ignoring important intrinsic variables on study power, pursuing stratification or matching on known intrinsic variables, especially confounders which are likely to be present when combining individuals from different geographical locations, is a desirable approach that should be considered in COVID-19 case-control host genetic studies moving forward [28]. Estimates of genetic effects, measured in odds ratios, are also difficult to interpret even if the unadjusted intrinsic variables are not confounders [28]. Future studies potentially through matching and/or stratification will also need to take into account the age distribution of cases and controls given the varying infection susceptibility and disease severity experienced in different age groups [15].

The identification of genetic variation that directly impacts infection susceptibility and disease severity would be an important step towards personalized treatment plans, risk stratification, therapeutic, and vaccine development and deployment. However, limited phenotypic data and exposure/infection information significantly impact our ability to detect these variants, especially our ability to detect infection susceptibility alleles. As the community moves ahead with the proposal and implementation of these studies, careful consideration of designs that prioritize clarity in the interpretation of the findings rather than expediency will prove to be more beneficial and cost-effective in the long run.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-020-00818-2>.

**Additional file 1: Supplementary Methods.** Detailed simulation settings for studying SARS-CoV-2 Infection Susceptibility.

**Additional file 2: Supplementary Figures and Tables. Figure S1.** Statistical power to detect associations between genetic variants and infection susceptibility at the genome-wide significance level (5e-8) when the test sensitivity is low (sensitivity = 0.7). **Figure S2.** Statistical power to detect a true association between a genetic variant and COVID-19 disease severity at the genome-wide significance level (5e-8). **Figure S3.** Statistical power to detect a true association between a genetic variant and COVID-19 disease severity at the genome-wide significance level (5e-8) when varying the case-control ratio. **Table S1.** Relative reduction in sample size,  $1 - n_{test,positive,controls} / n_{population,controls}$ , from using test-positive controls compared to population-based controls.

**Additional file 3: Supplementary Methods.** Detailed simulation settings for studying COVID-19 disease severity.

### Acknowledgements

Genetic Epidemiology Committee of the Canadian COVID Genomics Network (CanCOGeN) HostSeq Project (<https://www.genomecanada.ca/en/>)

[news/genome-canada-leads-40-million-genomics-initiative-address-covid-19-pandemic](#)).

### Authors' contributions

YL performed all the simulations and interpreted implications for studying COVID-19 host genomic studies at different stages of the pandemic. LJS provided the initial conceptualization. Both LS and LJS provided extensive feedback throughout the study. YL and LJS wrote the original draft. All authors have edited the manuscript. All authors read and approved the final manuscript.

### Funding

Y.L. is a trainee of the CANSSI Ontario STAGE (Strategic Training for Advanced Genetic Epidemiology) program at the University of Toronto and recipient of the CANSSI COVID-19 Rapid Response Program. The senior authors represent the Genetic Epidemiology Committee of the COVID Genomics Network (CanCOGeN) HostSeq project, which is funded by Innovation, Science and Economic Development Canada through Genome Canada.

### Availability of data and materials

The work employed simulations to investigate the statistical power of current host genetic case-control studies of COVID-19. Code used for all the simulations can be accessed through GitHub [https://github.com/jerry-yclin/COVID\\_host\\_genomic\\_simulations](https://github.com/jerry-yclin/COVID_host_genomic_simulations) [29].

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Dalla Lana School of Public Health, University of Toronto, Room 500, 155 College St, Toronto, ON M5T3M7, Canada. <sup>2</sup>Program in Genetics and Genome Biology, The Hospital for Sick Children, Room 12.9801, 686 Bay Street, Toronto, ON M5G0A4, Canada. <sup>3</sup>Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada. <sup>4</sup>Gerald Bronfman Department of Oncology, Department of Epidemiology, Biostatistics & Occupational Health, Department of Human Genetics, McGill University, Montreal, QC, Canada. <sup>5</sup>Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada. <sup>6</sup>Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada. <sup>7</sup>Department of Statistical Sciences, University of Toronto, 9th Floor, Ontario Power Building 700 University Ave, Toronto, ON M5G 1Z5, Canada. <sup>8</sup>The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada.

Received: 1 September 2020 Accepted: 7 December 2020

Published online: 28 December 2020

### References

- Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? *Lancet Infect Dis*. 2006;6(10):653–63. [https://doi.org/10.1016/S1473-3099\(06\)70601-6](https://doi.org/10.1016/S1473-3099(06)70601-6).
- Kellam P, Weiss RA. Infectogenomics: insights from the host genome into infectious diseases. *Cell*. 2006;124(4):695–7. <https://doi.org/10.1016/j.cell.2006.02.003>.
- Murray MF, Kenny EE, Ritchie MD, et al. COVID-19 outcomes and the human genome. *Genet Med*. 2020;22(7):1175–7. <https://doi.org/10.1038/s41436-020-0832-3>.
- Williams FM, Freydin M, Mangino M, et al. Self-reported symptoms of COVID-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. *medRxiv*. 2020. <https://doi.org/10.1101/2020.04.22.20072124>.
- COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet*. 2020; 28(6):715–8. <https://doi.org/10.1038/s41431-020-0636-6>.

6. Casanova JL, Su HC, Human Genetic Effort COVID. A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell*. 2020;181(6):1194–9. <https://doi.org/10.1016/j.cell.2020.05.016>.
7. Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide association study of severe COVID-19 with respiratory failure [published online ahead of print, 2020 Jun 17]. *N Engl J Med*. 2020;NEJMoa2020283. <https://doi.org/10.1056/NEJMoa2020283>.
8. CGEn. Canada's national platform for genome sequencing and analysis. CGEn's response to COVID [Press release]. May 10<sup>th</sup>, 2020. Available at: <http://www.cgen.ca/covid-19-response>. Accessed 8 Aug 2020.
9. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203–9. <https://doi.org/10.1038/s41586-018-0579-z>.
10. CLSA. Canadian Longitudinal Study of Aging. CLSA response to COVID-19. [Press release]. June 29<sup>th</sup>, 2020. Available at: <https://www.clsa-elcv.ca/coronavirus>. Accessed 10 Aug 2020.
11. The COVID-19 Host Genetics Initiative. Phenotype definitions for analyses 2.0. [Online]. April 27<sup>th</sup>, 2020. Available at: [https://docs.google.com/document/d/1okamrqYmJfa35CLvCt\\_vEe4PkrTwggHq7T3jbeyCI/edit](https://docs.google.com/document/d/1okamrqYmJfa35CLvCt_vEe4PkrTwggHq7T3jbeyCI/edit). Accessed 15 June 2020.
12. Mozzi A, Pontremoli C, Sironi M. Genetic susceptibility to infectious diseases: current status and future perspectives from genome-wide approaches. *Infect Genet Evol*. 2018;66:286–307. <https://doi.org/10.1016/j.meegid.2017.09.028>.
13. Fonds de recherche du Québec – Santé. Creation of the Québec COVID Biobank [Press release]. March 26<sup>th</sup>, 2020. Available at: <http://www.frqs.gouv.qc.ca/en/espace-presse/nouvelles-et-communiqués/nouvelle/creation-of-the-quebec-covid-biobank-tk9qemah1585225878700>. Accessed 8 Aug 2020.
14. Flaxman S, Mishra S, Gandy A, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*. 2020; 584(7820):257–61. <https://doi.org/10.1038/s41586-020-2405-7>.
15. Davies NG, Klepac P, Liu Y, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med*. 2020;26(8):1205–11. <https://doi.org/10.1038/s41591-020-0962-9>.
16. Woloshin S, Patel N, Kesselheim AS. False negative tests for SARS-CoV-2 infection - challenges and implications. *N Engl J Med*. 2020;383(6):e38. <https://doi.org/10.1056/NEJMp2015897>.
17. Watson J, Whiting PF, Brush JE. Interpreting a COVID-19 test result. *BMJ*. 2020;369:m1808. <https://doi.org/10.1136/bmj.m1808>.
18. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*. 2008;32(3):227–34. <https://doi.org/10.1002/gepi.20297>.
19. Debnath M, Banerjee M, Berk M. Genetic gateways to COVID-19 infection: implications for risk, severity, and outcomes [published online ahead of print, 2020 Jun 11]. *FASEB J*. 2020. <https://doi.org/10.1096/fj.202001115R>.
20. Cao Y, Li L, Feng Z, et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov*. 2020;6:11. Published 2020 Feb 24. <https://doi.org/10.1038/s41421-020-0147-1>.
21. Pollán M, Pérez-Gómez B, Pastor-Barriuso R, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *Lancet*. 2020;396(10250):535–44. [https://doi.org/10.1016/S0140-6736\(20\)31483-5](https://doi.org/10.1016/S0140-6736(20)31483-5).
22. Sutton D, Fuchs K, D'Alton M, Goffman D. Universal screening for SARS-CoV-2 in women admitted for delivery. *N Engl J Med*. 2020;382(22):2163–4. <https://doi.org/10.1056/NEJMc2009316>.
23. Zheng G, Tian X. The impact of diagnostic error on testing genetic association in case-control studies. *Stat Med*. 2005;24(6):869–82. <https://doi.org/10.1002/sim.1976>.
24. Meng H, Xiong R, He R, et al. CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia at admission in Wuhan, China. *J Infect*. 2020;81(1):e33–9. <https://doi.org/10.1016/j.jinf.2020.04.004>.
25. World Health Organization. WHO R&D blueprint: novel coronavirus: outline of trial designs for experimental therapeutics. Geneva: World Health Organization; 2020. <https://apps.who.int/iris/handle/10665/330694>.
26. CanPath. Canadian Partnership for Tomorrow's Health. CanPath Awarded \$2.1 million CIHR Grant for SUPPORT-Canada COVID-19 Initiative [Press release]. June 25<sup>th</sup>, 2020. Available at: <https://canpath.ca/2020/06/canpath-awarded-2-1-million-cihr-grant/>. Accessed 10 Aug 2020.
27. Nörz D, Fischer N, Schultze A, et al. Clinical evaluation of a SARS-CoV-2 RT-PCR assay on a fully automated system for rapid on-demand testing in the hospital setting. *J Clin Virol*. 2020;128:104390. <https://doi.org/10.1016/j.jcv.2020.104390>.
28. Keogh RH, Cox DR. Case-control studies. Cambridge: Cambridge University Press; 2014.
29. Lin Y, Brooks JD, Bull SB, Gagnon F, Greenwood CMT, Hung RJ, Lawless J, Paterson AD, Sun L, Strug L. Statistical power in COVID-19 case-control host genomic study design. GitHub. 2020. [https://github.com/jerry-yclin/COVID\\_host\\_genomic\\_simulations](https://github.com/jerry-yclin/COVID_host_genomic_simulations). Accessed 7 Dec 2020.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.