COVID-19 outcomes and the human genome

Michael F. Murray, MD [1], Eimear E. Kenny, PhD[2], Marylyn D. Ritchie, PhD[3], Daniel J. Rader, MD[3], Allen E. Bale, MD[1], Monica A. Giovanni, MS, CGC[1] and Noura S. Abul-Husn, MD, PhD[2]

BACKGROUND

In the COVID-19 pandemic, the opportunity to link host genomic factors to the highly variable clinical manifestations of SARS-CoV-2 infection has been widely recognized.[1,2] The overt motivation for this research is the clinical implementation of any new insights to improve clinical management and foster better patient outcomes.

Human infection is a complex interaction between the microbe, the environment, and the human host.[3] Variation in the human genome has only rarely been linked to complete resistance to infection by a specific microbe; far more commonly host genomic variability has been linked to complications associated with infections (see Table 1).[3–5] In this pandemic, the ability to identify host genomic factors that increase susceptibility or resistance to the complications of COVID-19 and to translate these findings to improved patient care should be the goal.

Several approaches can be taken to uncover relevant host genomic factors. Familial and population-based linkage analyses and analyses of extreme phenotypes can uncover monogenic variants contributing to COVID-19 clinical outcomes.[6] Genome-wide association studies (GWAS)[7,8] and multiomic-based approaches can be used to uncover common variants and biological networks underlying host-pathogen interactions. Likewise, data derived from genomes, such as HLA haplotypes, ABO blood groups, and polygenic risk scores (PRS),[9] can be used to understand COVID-19 susceptibility, resistance, and complications. Furthermore, biobanks linking genomic data to electronic health records (EHRs)[10] can be leveraged to investigate the impact of these genomic factors on the clinical course of SARS-CoV-2 infected patients.

Many recognize that this area of research needs to go forward in a manner that is proactively inclusive of traditionally underserved populations to both avoid the exacerbation of existing health-care disparities and to optimize discovery. Past efforts have demonstrated the value of this type of inclusion, as was seen in the extension of a CCR5-associated delta 32 correlation to HIV-1 infection in individuals with European ancestry to a promoter variant in CCR5 linked to perinatal HIV-1 transmission in individuals with African ancestry.[11–13]

As host genomic factors are discovered, new strategies supporting rapid clinical implementation should be trialed to realize improvement in outcomes for SARS-CoV-2 infected patients. Implementation will require an infrastructure to deliver relevant genomic results to infected patients and their health-care providers to guide clinical management. This commentary examines the types of genomic factors that might be identified in emerging COVID-19 discovery and implementation research, based on decades of genomic discovery, research into other human infections, and advances in genomic medicine.

PHASES OF PHENOTYPE ASCERTAINMENT IN THE COVID PANDEMIC

In this fast-moving pandemic, we believe there will be at least *two phases* to defining COVID-19 related phenotypes. Currently in the United States, we are in an *initial phase* when important limitations influence the ability of research teams to ascertain and appropriately define phenotypes of interest. These limitations include (1) the absence of widespread viral and serologic testing to accurately distinguish those who have been infected from those who have not, (2) the lack of knowledge about infection exposure at a community level, and (3) institutional limits to recruiting human subjects in a time of social distancing. Heterogeneity of testing strategies and their sensitivity, and nascent regulatory oversight may pose challenges in clear and reproducible definitions of COVID-19-related phenotypes. In the *second phase*, adequate serologic testing may allow for increased numbers and more accurate discrimination of cases and controls, as well as the ability to define additional clinical phenotypes of interest (e.g., asymptomatic seropositive individuals). The use of telemedicine, which has expanded for health-care delivery during the pandemic, in addition to

[1]Department of Genetics, Yale School of Medicine, New Haven, CT, USA; [2]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. Correspondence: Michael F. Murray (michael.murray@yale.edu)

Table 1 Sample characteristics.

	Discovery Categories	Approach	Yield	Non-COVID-19 Example
A	Rare Phenotypic Outliers	"Undiagnosed disease" approach, including family-based analysis when possible	Rare monogenic associations may be found in as many as 25-30% of cases[19,20]	*IRF7* deficiency[6]
B	Variants Associated with Specific Susceptibility	Cases and controls looking at risk for infection and complications of infection	Anticipate alleles with frequency 1% or above	[a]*HBB* *CFTR* *CCR5* *APOL1*
C	Polygenic Risk Scores	Genome Wide Association Studies (GWAS) across varied populations and phenotypes	Identification and aggregation of COVID-19 associated variants genome-wide	Inflammatory bowel disease[8]

[a]specific *HBB, CFTR, CCR5,* and *APOL1* alleles influencing infectious disease susceptibility are detailed in Supplemental Table 1.

community outreach efforts, can overcome barriers to recruitment in this infectious disease outbreak.

To find important genotype–phenotype correlations, there will need to be phenotypes that are ascertained in a manner that is clear, quantitative, and reproducible, and there will need to be adequate sampling from well-defined cases and controls. One rubric that can be used for phenotyping during this *initial phase* of COVID-19 host genomic research is the Ordinal Scale for Clinical Improvement proposed by the World Health Organization (WHO) in their blueprint for therapeutic trials (see Supplemental Table 1).[14] For instance, this scale can be applied across research groups and across health systems in order to allow phenotypic groupings of COVID-19 patients based on (1) need for hospitalization, (2) need for oxygen supplementation, (3) progression to respiratory failure, or (4) mortality, and these phenotypes could be readily extracted from EHRs. In the current initial phase of the COVID-19 pandemic, difficulties with the enrollment and appropriate scoring of uninfected, asymptomatic, or mildly affected patients (categories 0–2 in Supplemental Table 1) are anticipated. Specifically, asymptomatic positives will be mistakenly scored as 0 instead of 1 without either viral screening or serologic testing. In addition, patients who would be scored 0–2 are difficult to recruit and consent given the social distancing limitations that are currently in place. As serologic testing becomes more sophisticated, widespread, and robust, it is anticipated that COVID-19-related phenotyping will become more standard, facilitating reproducible and scalable COVID-19 research.

CANDIDATE GENES AND PATHWAYS

At least three lines of inquiry might inform the nomination of candidate genes for intensive interrogation with COVID-19 phenotypes: (1) what do we know about the microbial life cycle, (2) what do clinical observations in patients suggest with regard to biological pathways that are likely being triggered, and (3) what does the literature teach us about host genetics in infection that could apply to this novel infection. For example, the cellular surface receptor for SARS-CoV-2 virus is encoded by the *ACE2* gene, and critical amino acid residues in the binding interaction have been described.[15,16] This and other insights into host–pathogen interactions will elucidate specific variants, genes, and pathways underlying interindividual COVID-19 susceptibility and response. Genes and pathways related to COVID-19 could also include other viral receptor genes (e.g., *TMPRSS2*) (unpublished data: https://doi.org/10.1101/2020.03.30.20047878), inflammatory and immune response pathways (e.g., *IL-6* pathway), and genes involved in hypercoagulability and acute respiratory distress syndrome.[17] Other genes that may be of interest include genes associated with ABO blood group (e.g., *FUT2*) in light of a report on an association between blood groups and COVID-19 in China (unpublished data: https://doi.org/10.1101/2020.03.11.20031096) as well as similar associations in the past.[18] Research into the genetics of the interplay between viral infection and common diseases (e.g., diabetes and heart disease) is also of interest to many investigators. As our understanding of genes underlying SARS-CoV-2 infectivity and biological mechanisms grows, we will better elucidate their potential involvement in disease susceptibility and clinical outcomes.

GENOME-SCALE APPROACHES FOR DISCOVERY AND RISK PREDICTION

In tandem, the global scientific community has rapidly mobilized collaborative efforts to advance unbiased genome-wide COVID-19 host genomic discovery through large-scale genomic studies. For example, the COVID-19 Host Genetics Initiative is organizing analytical activities across a growing network of over 120 studies to identify genomic determinants of COVID-19 susceptibility and severity.[1] It is difficult at this stage to estimate the number of research participants needed to identify host genomic factors related to the COVID-19 novel pathogenic exposure. If we assume that the effect size and allele frequency of genetic variants important for COVID-19 susceptibility, resistance, and/or complications are as variable as other host factors in infectious conditions (i.e., Supplemental Table 1), then the number of cases and controls needed to have statistical power to identify associations could vary widely. Collaborative efforts like the COVID-19 Host Genetics Initiative should be well-powered for the unbiased discovery of novel genes and pathways. Such efforts foster data aggregation and sharing broadly among the research community and are likely to greatly impact the

speed with which COVID-19 discoveries can be made and disseminated worldwide.

In aggregate, knowledge of host genomic factors could lead to improved care for patients with COVID-19, through risk stratification, as well as targeted prevention and treatment options. For example, GWAS discovery efforts could yield PRS for COVID-19 clinical outcomes, which could be used in the context of other clinical data to risk stratify patients early in the disease course. Host genomic factors could be linked to variability in the protective immune response and have implications for vaccination strategies, or could be used to optimally select patients for novel therapeutic treatments and trials. However, as it can take many years for genomic discoveries to directly benefit patients,[10] in parallel we need to prepare our health systems with infrastructure to rapidly integrate high quality, clinically relevant COVID-19 host genomic findings into the care of individuals with SARS-CoV-2 infection.

CONCLUSIONS

The COVID-19 pandemic currently threatens to overwhelm health-care systems and undermine economies. There is no proven therapeutic and no vaccine for the novel coronavirus causing this pandemic. In this moment, we emphasize the sentiments voiced by the COVID-19 Host Genetics Initiative, namely that "[i]nsights into how to better understand and treat COVID-19 are desperately needed. Given the importance and urgency in obtaining these insights, it is critical for the scientific community to come together around this shared purpose."[1]

As the community works together to develop a COVID-19 host genomics research engine, we are poised for novel discovery and advances in genomic medicine. A model to understand human genomic variants linked to COVID-19 outcomes can be conceived as a continuum from ultrarare to common. We offer Supplemental Table 2 as a way to think about findings that can be expected from this research.[19,20] It is imperative that the research community prioritize high-quality and reproducible findings, even under the pressure for expediency, and be mindful of ethical, legal, or social issues that could emerge related to the COVID-19 impact among different groups within society.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-020-0832-3) contains supplementary material, which is available to authorized users.

DISCLOSURE

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

REFERENCES

1. The COVID-19 Host Genetics Initiative. https://www.covid19hg.org. Accessed 15 April 2020.
2. The COVID Human Genetic Effort. https://www.covidhge.com. Accessed 15 April 2020.
3. Murray MF. Susceptibility and response to infection. In: Rimoin DL, Connor JM, Pyeritz RE, et al., editors. Emery and Rimoin's principles and practice of medical genetics. 6th ed. London: Churchill Livingston; 2013.
4. Gabriel SE, Brigman KN, Koller BH, et al. Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science. 1994;266:107–109.
5. Ahuja SK, He W. Double-edged genetic swords and immunity: lesson from CCR5 and beyond. J Infect Dis. 2010;201:171–174.
6. Ciancanelli MJ, Huang SX, Luthra P, et al. Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. Science. 2015;348:448–453.
7. European Bioinformatics Institute. The NHGRI-EBI catalog of published genome-wide association studies (GWAS). https://www.ebi.ac.uk/gwas/. Accessed 15 April 2020.
8. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
9. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50:1219–1224.
10. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. Cell. 2019;177:58–69.
11. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature. 1996;382:722–725.
12. Liu R, Paxton WA, Choe S, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell. 1996;86:367–377.
13. Kostrikis LG, Neumann AU, Thomson B, et al. A polymorphism in the regulatory region of the CC-chemokine receptor 5 gene influences perinatal transmission of human immunodeficiency virus type 1 to African-American infants. J Virol. 1999;73(Dec):10264–10271. PMID: 1055934
14. World Health Organization. R&D blueprint novel coronavirus COVID-19 therapeutic trial synopsis.18 February 2020. https://www.who.int/blueprint/priority-diseases/key-action/COVID-19_Treatment_Trial_Design_Master_Protocol_synopsis_Final_18022020.pdf. Accessed 15 April 2020.
15. Yan R, Zhang Y, Li Y, et al. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367:1444–1448.
16. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020 Mar 30; https://doi.org/10.1038/s41586-020-2179-y [Epub ahead of print].
17. Hernández-Beeftink T, Guillen-Guio B, Villar J, Flores C. Genomics and the acute respiratory distress syndrome: current and future directions. Int J Mol Sci. 2019;20:E4004.
18. Cooling L. Blood groups in infection and host susceptibility. Clin Microbiol Rev. 2015;28:801–870.
19. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014;312:1870–1879.
20. Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med. 2015;17:578–586.