Patterns

CellPress
OPEN ACCESS

Preview

COVID-19 and the differential dilemma

Sharlee Climer[1,*]
[1]Department of Computer Science, University of Missouri – St. Louis, One University Blvd, 319 ESH, St. Louis, MO 63121, USA
*Correspondence: climer@umsl.edu
https://doi.org/10.1016/j.patter.2021.100260

The conundrums of choosing candidate genes, via differential expression between treated and mock specimens, are tackled by Ghandikota et al. in this issue of *Patterns* in their efforts to tease out genetic patterns that are characteristic of coronavirus disease 2019 (COVID-19) outcomes.

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, is a heterogeneous disease exhibiting a broad spectrum of symptoms, ranging from mild (e.g., olfactory disfunction, dry cough, head or body aches, sore throat, COVID toes) to critical (e.g., cytokine storm, renal failure, cardiovascular damage, respiratory failure, lethal blood clotting, neurological disorders).[1] Intensive care units dedicated to COVID-19 cases are being confounded by divergent emergency crises, demanding a breadth of specialists and specialized equipment.[1] Although some COVID-19-positive individuals exhibit multiple symptoms, others only show one, and many are completely asymptomatic. Analyses of transcriptomics data hold potential to reveal patterns of gene expression associated with specific outcomes, thereby providing valuable foundational information for breakthrough advances, including diagnostic tools to facilitate precision treatment, seeds for generating hypotheses that decipher underlying biological mechanisms, and potential drug targets, some of which could already have effective medications that can be repurposed. However, gene expression data are noisy, and analyses are formidable. Moreover, due to the novelty of the virus, COVID-19 data are sparse. In this issue of *Patterns*, Ghandikota et al. launch into these challenges and present a multi-layered network modeling strategy to identify several biological processes that could help shed light on this enigmatic disease.[2]

Ghandikota et al. skillfully handle sparsity of COVID-19 data in two ways. First, they leverage pre-COVID data in their network analyses. This approach has been used by others, e.g., yielding the promising bradykinin storm hypothesis for COVID-19,[3] and this current work utilizes rich data from three SARS-CoV-1 infection (SARS) models, the STRING protein-protein interactions database, the Molecular Signatures Database (MSigDB), NCBI's Phenotype-Genotype Integrator (PheGenI), and NHGRI-EBI's genome-wide association studies (GWAS) catalog. Second, they integrate three separate SARS-CoV-2 infection (COVID-19) datasets for their analysis, drawing from a mouse model and human and African green monkey cell lines. They overcome the diversity of these organisms by utilizing "consensus" genes as described below.

Like many transcriptomic analyses, the study begins by determining differentially expressed genes (DEGs) with significant deviations in expression levels between the treated and mock specimens. The use of DEGs yields both obvious and subtle dilemmas. Due to the large number of statistical tests, some are likely to show significance by chance, and corrections are requisite. Balancing false positives and false negatives when choosing a multiple-testing correction method is challenging, because Bonferroni corrections tend to wipe out many significant results, and false discovery rate (FDR) tends to produce too many erroneous significant DEGs.[4] Because transcriptomic analyses are generally exploratory, FDR is commonly employed, as is done by Ghandikota et al. This approach yields 8,286 DEGs, from which they choose consensus genes that exhibit differential expression in at least two of the three datasets. This maneuver strives toward balancing the false positive/false negative quandary and produces a list of 1,467 consensus DEGs.

A more insidious issue with using DEGs is that some genes tend to be differentially expressed regardless of the phenotype being tested.[5] By using over 600 Affymetrix Human Genome U155 Plus 2.0 datasets for a wide range of phenotypes, Crow et al. ranked ~19,000 genes in a DEG "priors" list, ordered by likelihood to appear as DEGs in arbitrary transcriptional analyses. They observed 229 genes that appear in more than 10% of the DEG lists produced in the previous analyses, and one gene, *CXCL8* (aka *IL8*), included in nearly one-fifth of the studies. The data used by Ghandikota et al. were generated with Illumina NextSeq 500, and the impact of platform on rankings in the DEG priors list is currently unclear. To test whether their compendium of consensus genes is specific for COVID-19, Ghandikota et al. computed differential expression for 1,000 permutation trials in which the phenotype labels were randomly reassigned. These trials produced orders of magnitude fewer DEGs, as well as consensus DEGs, than did the unpermuted data, thereby increasing confidence in the COVID-19 specificity of the results.

Looking beyond the work presented by Ghandikota et al., a pressing challenge for future analyses involving DEGs is to capture genes that do not signal differential expression when examined in isolation but exhibit significance when examined as a group containing additional genes (Figure 1). A single gene can yield multiple protein species due to genetic polymorphisms and via regulatory mechanisms such as alternative splicing and post-translational modifications. Furthermore, more than 500 proteins are currently known to moonlight and perform diverse tasks while using a single specific amino acid sequence.[6] These gene multi-tasking operations deepen the intricacies of differential assessments. The toy

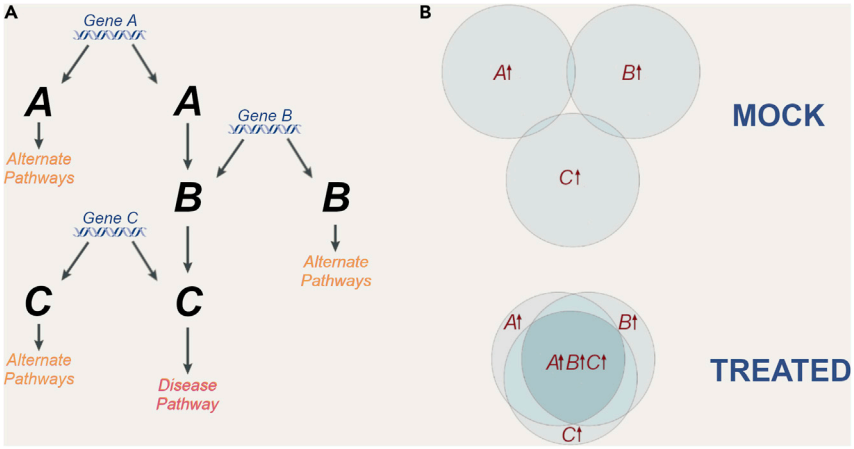

Figure 1. Combinatorial differential expression example

(A) Toy example of a disease pathway including genes *A*, *B*, and *C*, each of which show low marginal effects between treated and mock specimens due to genetic activities in alternate processes.

(B) Venn diagrams for high expression of genes *A*, *B*, and *C*. In this toy example, none of the mock specimens have simultaneous high expression for the three genes, whereas most of the treated exhibit this synchronized expression.

example in Figure 1 portrays an epistatic interaction in which all three genes are required for the disease pathway. It should be noted that a similar situation could arise for additive interactions in which multiple contributing genes must be considered in unison to observe a signal. In general, a collection of genes interacting in a disease-associated process could exhibit strong differences between treated and mock specimens when tested as a whole, yet each involved gene could show low marginal effects.

Given the accumulations of mutations that SARS-CoV-2 has sustained to date, the arms race between virus and vaccines is likely to extend into the foreseeable future.[7] The prevalence of so-called long COVID[8] and the emergence of evidence of long-term neurological and psychiatric outcomes[9] further emphasize the criticality of diagnosing and treating the heterogeneous sequalae presented. Continued generation of COVID-19 omics datasets and focused development of tactical strategies to extricate knowledge from these data are invaluable for treating individuals afflicted by this baffling disease.

ACKNOWLEDGMENTS

This research is supported in part by the National Institutes of Health grants 1RF1AG053303-01 and 3RF1AG053303-01S2.

REFERENCES

1. Wadman, M., Couzin-Frankel, J., Kaiser, J., and Matacic, C. (2020). A rampage through the body. Science *368*, 356–360.
2. Ghandikota, S., Sharma, M., and Jegga, A.G. (2021). Secondary analysis of transcriptomes of SARS-CoV-2 infection models to characterize COVID-19. Patterns *2*, this issue, 100247.
3. Garvin, M.R., Alvarez, C., Miller, J.I., Prates, E.T., Walker, A.M., Amos, B.K., Mast, A.E., Justice, A., Aronow, B., and Jacobson, D. (2020). A mechanistic model and therapeutic interventions for COVID-19 involving a RAS-mediated bradykinin storm. eLife *9*, 1–16.
4. Glickman, M.E., Rao, S.R., and Schultz, M.R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J. Clin. Epidemiol. *67*, 850–857.
5. Crow, M., Lim, N., Ballouz, S., Pavlidis, P., and Gillis, J. (2019). Predictability of human differential gene expression. Proc. Natl. Acad. Sci. USA *116*, 6491–6500.
6. Chen, C., Liu, H., Zabad, S., Rivera, N., Rowin, E., Hassan, M., Gomez De Jesus, S.M., Llinás Santos, P.S., Kravchenko, K., Mikhova, M., et al. (2021). MoonProt 3.0: an update of the moonlighting proteins database. Nucleic Acids Res. *49* (D1), D368–D372.
7. Rubin, R. (2021). COVID-19 Vaccines vs Variants-Determining How Much Immunity Is Enough. JAMA *325*, 1241–1243.
8. The Lancet Neurology (2021). Long COVID: understanding the neurological effects. Lancet Neurol. *20*, 247.
9. Taquet, M., Geddes, J.R., Husain, M., Luciano, S., and Harrison, P.J. (2021). 6-month neurological and psychiatric outcomes in 236 379 survivors of COVID-19: a retrospective cohort study using electronic health records. Lancet Psychiatry *0*, S2215-0366(21)00084-5.