

HumanBase Sees Connections Between COVID-19 and Diabetes

Center for Computational Biology

When Chandra Theesfeld, a research scientist at the Lewis-Sigler Institute for Integrative Genomics at Princeton University, was starting her career as a biologist, she knew that studying the human genome came with significant challenges. At the time, high-quality curated databases could help those scientists studying smaller organisms, like yeasts and worms, but nothing remotely comparable existed for humans. As a result, human genomic studies tended to be piecemeal and limited in scope, narrowly focused on just one or a few human genes. Scientists often relied heavily on what they remembered from the literature to make meaningful inferences from the available data.

The genomics revolution changed all that. A flurry of advancements in genomic analysis, research computing and data storage yielded a tremendous influx of new data from the human genome, and with it came the promise of a more comprehensive understanding of human biology. But a new problem soon emerged: The data were too vast for individual scientists to dig through. "The traditional forms of analysis in

humans just weren't possible anymore," Theesfeld says.

In need of a resource that could help researchers properly sift through human genomic datasets at scale and find the new insights hidden there, scientists at the Center for Computational Biology (CCB) at the Flatiron Institute created HumanBase, an interactive software platform that allows Theesfeld and other biologists to access results of tens of thousands of experiments in one place and make connections in a systematic way that springboard human biological discovery.

Launched in 2018, HumanBase brings robust computing power and advanced algorithms to bear on thousands of genomics datasets, enabling scientists to make connections across genes in ways that are impossible using traditional methods. HumanBase uses machine learning to reach into published and publicly available datasets from tens of thousands of genomics experiments and make predictions about how genes from specific tissues of the body interact with each other. Machine learning is uniquely suited to finding nuggets of biological gold in these large, diverse datasets: A biological signal that is faint in any one dataset may stand out when many datasets are integrated to generate one large network.

"HumanBase is designed around connections, like gene-to-gene, gene-to-disease and mutation-to-disease

connections, and draws these connections in a data-driven way, one that can't be replicated by individual scientists mining the literature," says Aaron Wong, a data scientist and project leader at the CCB. Olga Troyanskaya, CCB deputy director for genomics and a professor of computer science at the Lewis-Sigler Institute, describes the power of the networks assembled by HumanBase in a strategy called functional module detection. "By looking at the genes in the context of a network, we can discover the pathways impacted by a disease," she says.

To use functional module detection, scientists first input a list of up to 4,000 genes. HumanBase will take the list and generate a network showing how the genes work together in a particular cell type, such as a kidney cell. The results are displayed in weblike maps showing how each gene is associated with others in the database. Clusters in the maps known as modules contain genes with common functions. For example, one module might consist of genes that promote viral replication. Functional module detection can also suggest a function, or set of functions, for a previously uncharacterized gene. "The networks connect genes that are working together in the same pathway, and the module detection in turn reveals higher-order processes and pathways," says Theesfeld.

In 2020, scientists at the Flatiron Institute and the University of Michigan used HumanBase's functional module detection to examine the mechanics of COVID-19 infection in kidney cells. In work published October 7, 2020, in *Kidney International*, whose initial findings were shared in May on the medRxiv preprint server, the authors investigated why individuals with diabetes are more susceptible to COVID-19. SARS-CoV-2, the virus that causes COVID-19, enters the cell by locking onto a protein on the cell surface called ACE2. The scientists sought to learn what is different in kidney cells that express ACE2 and, specifically, to understand what it is about these cells in people with diabetes that makes them especially susceptible to infection.

Credit: Lucy Reading-Ikkanda/Simons Foundation

In ACE2-expressing kidney cells both from patients with diabetic kidney disease (DKD) and from patients with COVID-19, the scientists found thousands of genes that showed increased expression. They used HumanBase to construct modules from these genes, revealing that in both groups of patients, these modules relate to viral entry, replication and immunity. "HumanBase showed us that [the DKD] cells with expressed ACE2 have a cellular program already activated which makes them exceptionally vulnerable to the virus," says Matthias Kretzler, a nephrologist and professor of medicine at the University of

Michigan and co-corresponding author on the study.

Troyanskaya adds, "Without any input from the virus, the cells of the diabetic kidney already look similar to cells of patients who have the virus. The diabetic kidney is essentially primed for SARS-CoV-2."

The connections between ACE2 expression and biological processes relating to immunity and viral activity, and between programs in the cells of patients with DKD and COVID-19, could not have been made without HumanBase, says Theesfeld. "You have a list of thousands of genes — you can't make sense of that. You need the sophisticated connection-drawing power of networks and machine learning to find the common threads," she says. Further research is needed to determine definitively if the activation of viral infection pathways in DKD is responsible for the increased susceptibility of patients to COVID-19, and if COVID-19 infection in patients with DKD results in cumulative kidney damage.

The study also showed that common medications for hypertension and DKD do not increase the levels of ACE2, despite initial concerns, and thus patients could continue safely taking these medications. "This was a critical piece of knowledge we could share quickly with our global kidney doctor community, and we could add a mechanistic explanation for why," says Kretzler.

Importantly, the functional modules uncovered by HumanBase are a starting point for exploring new therapeutic avenues to treat COVID-19. Scientists will examine how kidney tissue grown in a lab responds to drugs that target genes and processes that HumanBase shows are activated in cells expressing ACE2. The processes identified by HumanBase suggest roles for particular structures of the cell, such as the ribosomes and cell membrane, during SARS-CoV-2 infection. Theesfeld describes two recent studies in which scientists showed experimentally how SARS-CoV-2 infection upsets the functioning of these very structures, the ribosomes and cell membranes, in kidney cells. "We see different omics [high-throughput molecular analysis] approaches that validate our predictions," she says. "A next step would be to look for drug targets in those pathways."

The COVID-19 study has implications for virus biology in general, too, says Theesfeld. Some viruses use receptors other than ACE2 to enter the cell. If HumanBase were applied to study another virus, "would we find the same processes upregulated in cells that use a different receptor?" she asks. The results could shed light on which viral processes are universal and which might be unique to coronaviruses.

At present, researchers are using HumanBase to reveal the

cellular processes activated in lung cells during SARS-CoV-2 infection after treatment with Moderna's vaccine, Troyanskaya notes. "These processes are complex," she says. "The pathways involved and their connections are only revealed at the network level, showing the biological coherence behind large sets of genes." Theesfeld adds, "The HumanBase approach is a powerful and general way to reveal the network effects of dysregulation in human disease."

The long-term support of the Simons Foundation in developing HumanBase was critical to this work, says Wong. "A key mission of the Flatiron Institute is to develop cutting-edge algorithms and make them broadly available, not just to computational people but also to biologists and biomedical scientists," he says. Troyanskaya agrees, emphasizing that HumanBase lets biomedical and clinical scientists make connections formerly in the domain of computer scientists. "The critical connection is between a biologist's insight, the data and advanced computational analysis: HumanBase allows this loop to work without an advanced computer scientist in it."