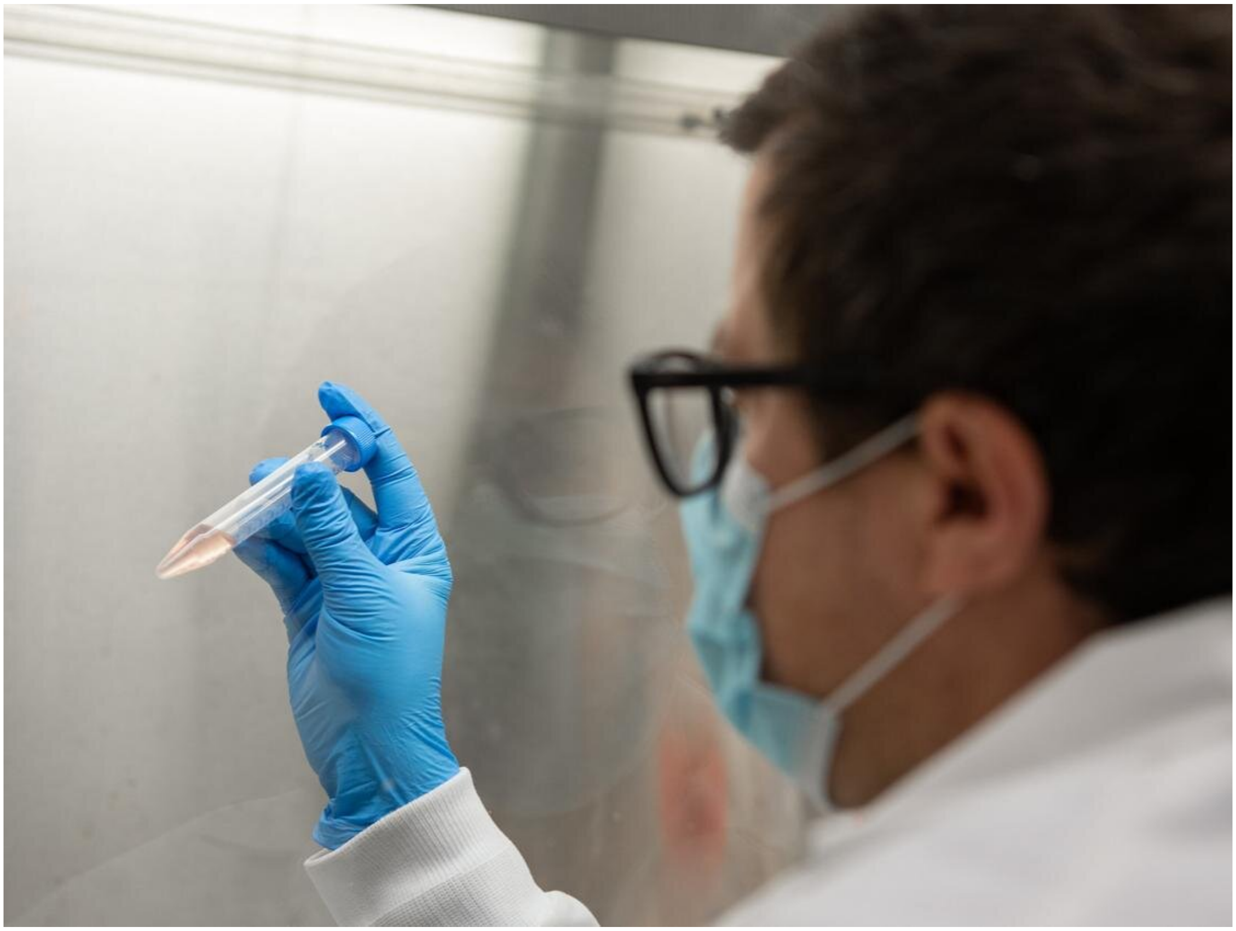
Compressing gene libraries to expand accessibility, research opportunities

February 14 2022, by Gabrielle Stewart



Justin Pritchard, assistant professor of biomedical engineering and holder of the Dorothy Foehr Huck and J. Lloyd Huck Early Career Entrepreneurial Professorship, holds up a mock genetic sample. This sample, less than a tenth of an ounce (1-2 milliliters) in volume, represents the size of the "compressed" genetic library—compared to a conventional library measuring about 3 ounces (80 milliliters). Credit: elby Hochreither/Penn State

In image compression, a large file that could be cumbersome to store or share loses a small amount of visual information. This "lossiness" largely preserves the image while vastly reducing its file size—and serves as the inspiration for a new research direction in genomics, according to Justin Pritchard, assistant professor of biomedical engineering.

Pritchard and a Penn State-led team of interdisciplinary researchers developed a methodology for "compressing" extensive genetic data libraries to more manageable sizes. They published their findings in *Nature Communications* on Feb. 2.

"This idea of compression dramatically reduces the scale of the experiments, opening up possibilities for new experiments," said Pritchard, who also holds the Dorothy Foehr Huck and J. Lloyd Huck Early Career Entrepreneurial Professorship. "This can unlock biological mysteries, such as why different genes and drugs work differently together, and it allows us to unravel very complicated biology using simpler experiments."

The researchers referred to genome-scale CRISPR experiments containing data on thousands of gene effects tested in different human cell types. The effect when the gene is turned off can vary between cell types, so a large number of cells is often needed to understand the interplay between genes and phenotypes.

To predict the larger genome-scale effects from the smaller "compressed" CRISPR library, the team used a custom algorithm rooted in a common machine learning technique known as random forests. This method incorporates data provided by the researchers into a series of randomly generated decision trees that collectively produce predictions about the relationship between gene inactivation and cell growth. The model was trained on the majority of the data—leaving one data subset out—and then initially validated by testing its capacity to predict data for the omitted subset. This accuracy extended to datasets that were generated in different labs using different experimental conditions and CRISPR libraries.

This performance was possible using only a small percentage—about 1%—of the original library's information. Finally, the Penn State group performed new experiments in which they physically built these "lossy compression libraries" using synthetic biology techniques and validated the predictions in new experiments.

"A genome-scale experiment probes 18,000 genes," Pritchard said. "Using machine learning, we tunably compressed the scale of the experiment to as few as 200 genes. Despite the loss of some data in the compression, we found that a subset of 200 genes could provide surprisingly good information on the full 18,000 genes."

The technique also opens opportunities for other research, according to Pritchard. It showed transferability, meaning it could make accurate predictions matching information from entirely different datasets despite only being trained on the CRISPR data. The capacity to reduce the number of genes also enables more research on cells that can be difficult or impossible to aggregate in large amounts, such as cells within a living organism.

"We're excited about the future of this research," Pritchard said. "We can alter the composition of these lossy compression sets in real time, for different experimental questions and conditions in areas from cancer biology to biopharmaceuticals, using newer machine learning techniques. The method also helps us improve basic science by answering questions about how the genome works and encodes information on cell growth."

Boyang Zhao, Edward P. O'Brien, Luke Gilbert, Scott Leighow and Yiyun Rao from Penn State contributed to this work. Zhao contributed as first author and is also affiliated with Quantalarity Research Group in Houston. Gilbert is affiliated with the University of California San Francisco and the Helen Diller Family Comprehensive Cancer Center in San Francisco.

More information: Boyang Zhao et al, A pan-CRISPR analysis of mammalian cell specificity identifies ultra-compact sgRNA subsets for genome-scale experiments, *Nature Communications* (2022). DOI: 10.1038/s41467-022-28045-w

Provided by Pennsylvania State University

Citation: Compressing gene libraries to expand accessibility, research opportunities (2022, February 14) retrieved 16 July 2024 from https://phys.org/news/2022-02-compressing-gene-libraries-accessibility-opportunities.html