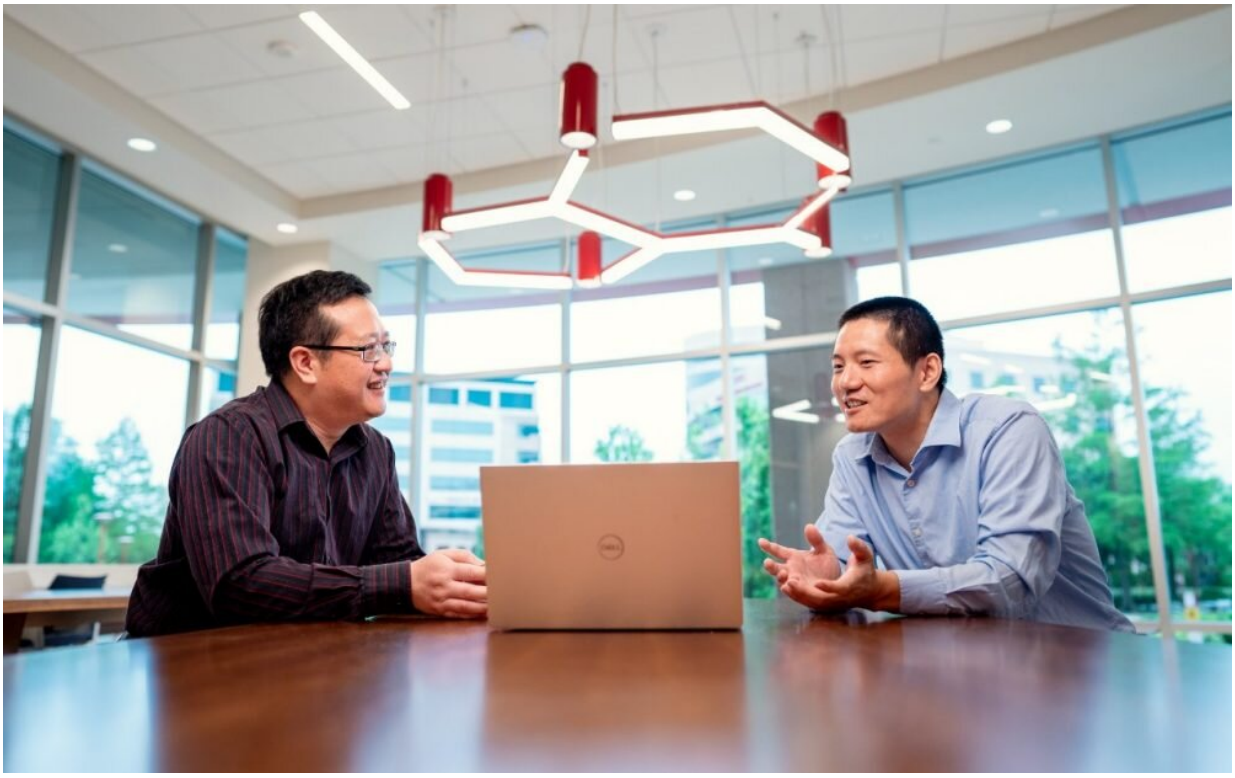
A tool to unlock the 'numbers game' of big data in rare disease research

May 11 2022



Gang Wu, Ph.D., and Wenan Chen, Ph.D., Center for Applied Bioinformatics. Credit: St. Jude Children's Research Hospital

Computational scientists at St. Jude Children's Research Hospital have created a tool to find genes and genetic variants that predispose people to rare diseases. Finding genes and genetic variants that contribute to rare

diseases is difficult due to confounding factors introduced when using large public datasets. The tool offers a solution to account for the confounders and identify statistically significant results. The evidence was published today in *Nature Communications*.

Identifying the genetic variants that contribute to rare diseases may increase understanding of disease development and offers potential new avenues for therapy. Researchers use large amounts of data from ever-improving sequencing methods to find these connections. The lack of a systematic method to find statistically sound results has made it challenging in finding rare disease related genes.

"This is all a numbers game," said corresponding author Gang Wu, Ph.D., St. Jude Center for Applied Bioinformatics director. "Traditionally, if you have a small cohort study of 20 to 50 unrelated individuals with a very rare disease, you have almost no way to find a novel gene variant that reaches statistical significance in its contribution to the disease without prior knowledge of candidate genes. Now we have an approach that can potentially help find novel disease predisposition genes."

"For diseases like amyotrophic lateral sclerosis (ALS), or pediatric brain tumors, we probably know that up to 20% of the patients can be explained by germline predisposition to the disease," he said. "Our tools will help find the remaining unexplained heritability that can be contributing to those diseases."

CoCoRV: Taming uncontrolled data

To address these issues, the scientists created a tool called the consistent summary counts based rare variant burden test (CoCoRV). The team was able to show that CoCoRV could find known genetic variants contributing to a variety of rare diseases, including multiple cancers and

ALS. In addition, for each ailment, the researchers identified previously unknown genetic variants that may represent a predisposition to that disease.

"CoCoRV is built upon the experience we have from working with many, many sets of sequencing data at St. Jude," said first author Wenan Chen, Ph.D., Center for Applied Bioinformatics. "We often determine whether something is a real signal or a technical artifact. When you have a large amount of data, you can use this knowledge to derive rules that systematically categorize what is a true signal versus which are bad quality in other datasets. We built that experience into a tool that would be helpful for others to use."

Rarity invites statistical precarity

Rare diseases are, by their nature, uncommon. Data on individuals is correspondingly sparse. In addition, few studies collect enough data from healthy individuals that are similar enough to a population of patients with a rare disease to reach statistical significance, a concept called reaching statistical power.

The lack of matched healthy controls creates a challenge for scientists searching for the genetic variants related to developing rare diseases. By using publicly available databases of human genomes, scientists can use advanced statistical methods to create "synthetic control groups." These groups can then highlight the contrast between many healthy individuals' genes and the genes from a small group with a particular disease, to reach statistical power.

But public databases are often assembled with different methods, present information in different ways and are difficult to compare to each other. The field lacked a method to consistently integrate the information from public databases and calculate these advanced statistics

to create synthetic control groups. The St. Jude scientists created CoCoRV as a solution.

"Our tool provides a consistent and systematic way to maximize the power of an analysis and minimize the risk of finding false positives. Users can therefore confidently scan for potential pathogenic variants or try to identify risk genes for a rare disease," Chen said.

More information: Wenan Chen et al, A rare variant analysis framework using public genotype summary counts to prioritize disease-predisposition genes, *Nature Communications* (2022). DOI: 10.1038/s41467-022-30248-0

Provided by St. Jude Children's Research Hospital