Research News

Genetic testing has a data problem. New software can help

Computer scientists help find a way out of a genetic data maze

Computer scientists have found a new way to store genetic testing data.

Credit and Larger Version (/discoveries/disc_images.jsp?cntn_id=298521&org=NSF)

May 10, 2019

In recent years, the market for direct-to-consumer genetic testing has exploded. The number of people who used at-home DNA tests more than doubled in 2017, most of them in the U.S. Some 1 in 25 American adults now know their ancestry.

As the tests become more popular, companies are grappling with how to store the accumulating data and how to process results quickly. A new tool called TeraPCA, created by computer scientists at Purdue University, is now available to help. The results were published in the journal *Bioinformatics* (/cgi-bin/good-bye?https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz157/5430929). The research is funded by NSF's Directorate for Computer and Information Science and Engineering, supporting new insights in big data analytics.

Despite our many physical differences, any two humans are about 99 percent the same genetically. The most common genetic variations, which contribute to the 1% that makes us different, are called single nucleotide polymorphisms, or SNPs (pronounced "snips").

There are 4 to 5 million SNPs in every person's genome. Those are a lot of data to keep track of on even one person; doing the same for thousands or millions of people is quite a feat.

For the largest genetic testing companies, storing and processing data is not only expensive and technologically challenging, but comes with privacy concerns. These companies have a responsibility to protect personal health data. Storing it on their hard drives could make them attractive targets for hackers.

TeraPCA was designed with these challenges in mind: processing data too large to fit on a computer's main memory at one time. It makes sense of large datasets by reading small chunks at a time. In the future, TeraPCA may be useful not only as you learn about your ancestry, but it could also make new genetics research possible, shedding light on disease risks and possible cures.

-- NSF Public Affairs, (703) 292-8070 media@nsf.gov (mailto:media@nsf.gov)
