Disease-carrying mutations target of mega-sized human genome data crunchers

[R]esearchers at Columbia and Princeton universities describe a new machine-learning algorithm for scanning massive genetic data sets to infer an individual's ancestral makeup, which is key to identifying disease-carrying genetic mutations.
TeraStructure could estimate population structure more accurately and twice as fast as current state-of-the art algorithms, the study said.

…

"We can run software on a few thousand people, but if we increase our sample size to a few hundred thousand, it can take months to infer population structure," said Kai Wang, director of clinical informatics at Columbia's Institute for Genomic Medicine…"This new tool addresses these limitations, and will be very useful for analyzing the genomes of large populations."

…

TeraStructure…samples one genetic variant at one location, and compares it to all variants in the data set at the same location across the data set…"You don't have to painstakingly go through all the points each time to update your model," said [David Blei, a professor of computer science and statistics at Columbia University].

…

[When researchers] ran TeraStructure on a simulated data set of 10,000 genomes, it was more accurate and two to three times faster at estimating population structure…The researchers also showed that TeraStructure alone could analyze data sets as large as 100,000 genomes and 1 million genomes.
The GLP aggregated and excerpted this blog/article to reflect the diversity of news, opinion, and analysis. Read full, original post: Unlocking big genetic datasets