New approach will help geneticists identify genes responsible for complex traits

17 December 2018

Alex Lipka and Liudmila Mainzer led research to improve the genome-wide association study (GWAS), a bioinformatics strategy to identify genomic regions influencing traits of interest. The study expands the capacity of GWAS to identify multiple markers at a time, as well as their two-way interactions. Credit: Alex Lipka and Liudmila Mainzer

In biomedical research, plant breeding, and countless other endeavors, geneticists are on the hunt for the specific genes responsible for disease susceptibility, yield, and other traits of interest. Essentially, they're looking for needles in the enormous haystack that is the genome of an organism.

As a frame of reference, the human genome is made up of 3.2 billion base pairs, an estimated 30,000 genes. Where do geneticists even start?

For the past 15 years, many have relied on genome-wide association studies (GWAS).

"I view a GWAS as a way to reduce the size of the haystack into genomic regions that potentially could contain causal mutations underlying a trait," says Alex Lipka, assistant professor of biometry in the Department of Crop Sciences at the University of Illinois and author of a new *Heredity* study expanding the scope of GWAS.

To run a GWAS, scientists conduct computationally intensive statistical analyses to scour the genetic code for differences. Specific variations in DNA, called markers, that exhibit the highest degree of statistical association are thought to be near genes that make biological contributions to the trait. Sometimes, these associated markers are clustered together in a particular region of the genome, narrowing the haystack.

Lipka says the approach has been used in a wide variety of organisms to identify major genes contributing to key traits, but it falls short in detecting small-effect genes or gene interactions—a phenomenon known as epistasis—that may be just as important.

"The state-of-the-art statistical approach for GWAS is to test one marker at a time for the strength of its association with the trait," he says. "If you think about the true genetic underpinnings of a trait, it's not just one gene controlling things. Multiple genes contribute to phenotypic variation in an additive manner, and are epistatically interacting with one another. What we try to do in our study is explore the use of a statistical approach that is more biologically accurate. Not only are we finding statistical models that include multiple markers at a time, we also find multiple two-way interaction effects at a time."

The researchers wanted to see if their new approach, which they call SPAEML, could accurately detect the underpinnings of simulated traits with genetic sources similar to Alzheimer's disease in humans and flower structure in corn; these traits have already been described to some extent in the scientific literature. Using custom-built software, which they have made freely available to

other researchers, and massive computers at the National Center for Supercomputing Applications, the team tested whether SPAEML could detect simulations of the traits in the dataset.

"In both the human and corn datasets, we were able to identify our simulated markers," Lipka says. "And in the human dataset we were able to distinguish between additive and interacting loci."

The finding does not reveal new information about Alzheimer's disease; remember, SPAEML was tested against existing knowledge of the trait's genetic structure. Instead, it represents proof-of-concept that advanced GWAS methods like SPAEML can detect multiple markers that contribute to the disease, even in small ways. The researchers point out that the collective contributions of such markers can result in massive changes that may lead to the disease.

Although geneticists are well aware that complex traits are rarely controlled by a single gene, until now it had been too computationally difficult to test for multiple markers or their interactions.

"The problem is the combinatorial explosion of possibilities that must be tested, because we're looking at pairs of markers," says co-author Liudmila Mainzer, technical program manager for Genomics at NCSA. "The algorithm needs to evaluate tens of thousands, hundreds of thousands, possibly millions of models in order to select the best one. It could take years in sheer computational time, which is why no one has ever done it."

It took about four years for the team to develop and refine a method that could deal with that combinatorial explosion, bringing millions of data points down to about 15,000, a number SPAEML could handle easily. Going forward, the researchers plan to unleash SPAEML on datasets with unknown genetic structures. They're already working with collaborators in the crop breeding industry and human health research to launch next steps.

"This research is really hard, but it's the right way to approach this scientific problem. With access to supercomputing resources, outstanding students, and a bit of our own youthful foolhardiness—who knows, we might just manage it," Mainzer jokes. "Based on the feedback we've had so far, it has been very rewarding,"

More information: Angela H. Chen et al, An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers, *Heredity* (2018). DOI: 10.1038/s41437-018-0162-2

Provided by University of Illinois at Urbana-Champaign