MedicalXpress

# New statistical method for genetic studies could cut computation time from years to hours

March 18 2010

In the ongoing quest to identify the genetic factors involved in disease, scientists have increasingly turned to genome-wide association studies, or GWAS, which enable the scanning of up to a million genetic markers in thousands of individuals.

These studies generally compare the frequency of genetic variants between two groups — those with a particular disease and healthy individuals. Differences in the frequency of a given variant suggest the variant may be involved in the disease.

Over the last few years, such studies have successfully implicated hundreds of genes in human disease, and the research has been used to identify risk and protective factors for asthma, cancer, diabetes, heart disease, mental illness and other conditions.

But genome-wide association studies aren't perfect. In fact, the genealogy of study participants can sometimes prove a stumbling block to accurate findings.

"Unfortunately, differences in frequencies can arise for reasons unrelated to the disease if the individuals collected have ancestry from different regions of the world," said Eleazar Eskin, associate professor of computer science at the UCLA Henry Samueli School of Engineering and Applied Science, who holds a joint appointment in the department

of human genetics at the David Geffen School of Medicine at UCLA.

"This problem, called 'population structure,' has led to many apparent discoveries of genes involved in disease which later turned out to be artifacts," he said.

In a new study to be published in the April edition of the journal Nature Genetics, Eskin and his research group unveil a new computational strategy for GWAS that corrects for population structure and is both faster and easier to use.

One of the basic assumptions in typical GWAS is that participating individuals are "unrelated," and investigators typically perform screening procedures to ensure that pairs of individuals are not close relatives. However, due to the complex history of the human population, none of the individual pairs are perfectly unrelated, and each individual pair is somewhat distantly related to various degrees. This is referred to as "pairwise relatedness."

"Such a variety in degrees of relatedness — which we call 'sample structure' — can be manifested into two different forms: population structure and hidden relatedness. While typical statistical methods for GWAS handle only either of the two forms, our method can handle both aspects of sample structure simultaneously in a computationally efficient manner," said Hyun Min Kang, an assistant research professor in biostatistics at the University of Michigan and an author of the study.

"Moreover, if the samples come from a very homogeneous population, it is possible that some of the subjects are, in fact, distantly related," said Chiara Sabatti, professor of human genetics and statistics at UCLA and a corresponding author of the study. "In the analysis of GWAS, it is necessary to correct for such sample structure, which can lead to spurious association signals. The methods presented in our paper allow

researchers to do this in a manner that is both fast and effective."

Eskin's team worked with a data set of 5,000 people from Finland who were born in the same year, tracked over an extensive amount of time, and had a large amount of population relatedness.

The 5,000 people produced a data set of 300,000 variants. From these 300,000 points of variation, the group examined pairwise relatedness between individuals, which means they compared the number of mutations each shared. From the mutations, Eskin's group could estimate how related individuals were to each other.

"It was very interesting to see how much these pairwise relations explained of the trait," Eskin said. "So what we did in this paper is we proposed a statistical method that also allowed us to correct for a wide range of sample structure by explicitly accounting for pairwise relatedness between individuals using high-density markers in modeling the distribution of observable traits."

This variance component in the new strategy, called EMMAX (Efficient Mixed Model Association Expedited), would capture the complex mixture of both population structure and hidden relatedness, direct byproducts of genealogy, and correct for these relationships when performing genetic mapping.

"Capitalizing on the characteristics of complex traits in humans, we made a few simplifying assumptions that allowed us to dramatically increase the speed of computations, making our approach readily applicable to genome-wide association studies with tens of thousands of samples," Eskin said.

"Our variance component model is actually a widely known classical model for genetic mapping," Kang said. "However it was too

computationally costly to be applied to the current scale of GWAS involving thousands of individuals with hundreds of thousands of genetic variants because even the fastest method — which we previously developed — took years of computational time to analyze the data once. We further expedited the method by capitalizing the characteristics of most human association studies, reducing the computational time from years to hours."

According to Eskin, their method will also have a large impact on admixed populations, which are basically samples of individuals who have ancestry from multiple regions around the world. Studies on Los Angeles, for example, would benefit from this method greatly, as people in the city are very ethnically diverse and it's difficult to obtain very accurate estimates of people's ancestry.

Provided by University of California - Los Angeles