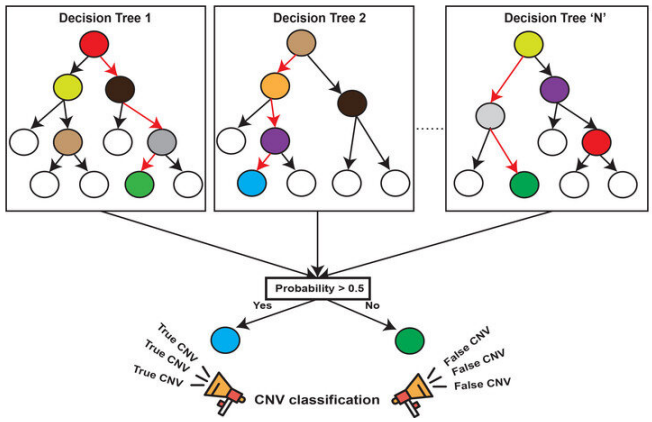
New method helping to find deletions and duplications in the human genome

10 July 2019, by Sam Sholtis



A random-forest, machine-learning method for identifying copy number variation from exome-sequencing data. A forest of hundreds of decision trees is trained on a validated set of genetic deletions and duplication, the model built from these trees can then be used to accurately identify copy number variation in sample exome-sequencing data. Credit: Girirajan Laboratory, Penn State

A new machine-learning method accurately identifies regions of the human genome that have been duplicated or deleted—known as copy number variants—that are often associated with autism and other neurodevelopmental disorders. The new method, developed by researchers at Penn State, integrates data from several algorithms that attempt to identify copy number variants from exome-sequencing data—high-throughput DNA sequencing of only the protein-coding regions of the human genome. A paper describing the method, which could help clinicians provide more accurate diagnoses for genetic diseases, appears in the July issue of the journal *Genome Research*.

"Exome sequencing is fast becoming the gold standard for identifying genetic variations in clinical settings because it is faster and less expensive that other methods," said Santhosh Girirajan, associate professor of biochemistry and molecular biology at Penn State and the lead author of the paper. "However, current algorithms for identifying copy number variation from exome sequencing data suffer from very high false-positive rates—many of the variants they identify aren't actually real. With our new method, called "CN-Learn," around 90 percent of the copy number variants we report are real."

The human genome generally contains two copies of every gene, one on each member of a chromosome pair. When one cell divides into two, the genome is replicated so that each of the daughter cells gets a full complement of genes, but occasionally errors occur during genome replication that, when present in a sperm or egg cell, can lead to an individual getting more or less than two copies of the gene.

To identify copy number variants from exome-sequencing data, researchers look at the relative amount of DNA sequences produced from each gene. If there is only one copy of a gene present in an individual, they expect to see fewer sequencing reads than if there are two copies, and three copies of a gene would lead to more reads. But it's not quite that simple, because a number of other factors can influence how many sequencing reads are produced from each gene. Researchers have therefore developed several algorithms to try to correctly identify copy number variants from exome-sequencing data. Individually, however, these algorithms are not particularly reliable.

"Generally, the high number of false positives from copy-number-variant algorithms has been dealt with by using multiple algorithms and only counting the variants identified by all the methods—like a Venn diagram," said Vijay Kumar Pounraja, a graduate student at Penn State and first author of the paper. "This approach has multiple drawbacks and limitations, so we decided to develop a new machine-learning method instead."

CN-Learn integrates data from four different copy-number-variant algorithms, and uses a small set of biologically validated deletions and duplications to learn the signatures of these genomic events. This learning process is facilitated by a machine-learning algorithm called Random Forest, which uses hundreds of decision trees to model the relationship between the genetic context of deletions and duplications and the likelihood they are validated. CN-Learn then uses this model to predict deletions and duplications in other samples without validations.

"Decisions about a patient's diagnosis and eventual treatment are made based on this information, so it's incredibly important to get them right," said Girirajan. "Because of this, we've made CN-Learn and all of the necessary supporting programs available to download in one easy package."

Provided by Pennsylvania State University