

New method developed for ranking disease-causal mutations within whole genome sequences

February 7 2014

Researchers from the University of Washington and the HudsonAlpha Institute for Biotechnology have developed a new method for organizing and prioritizing genetic data. The Combined Annotation–Dependent Depletion, or CADD, method will assist scientists in their search for disease-causing mutation events in human genomes.

The new method is the subject of a paper titled "A general framework for estimating the relative pathogenicity of human genetic variants," published in *Nature Genetics*.

Current methods of organizing [human genetic variation](#) look at just one or a few factors and use only a small subset of the information available. For example, the Encyclopedia Of DNA Elements, or ENCODE, catalogs various types of functional elements in human genomes, while sequence conservation looks for similar or identical sequences that have survived across different species through hundreds of millions of years of evolution. CADD brings all of these data together, and more, into one score in order to provide a ranking that helps researchers discern which variants may be linked to disease and which ones may not.

"CADD will substantially improve our ability to identify disease-causal mutations, will continue to get better as genomic databases grow, and is an important analytical advance needed to better exploit the information content of whole-genome sequences in both clinical and research

settings," said Gregory M. Cooper, Ph.D., faculty investigator at HudsonAlpha and one of the collaborators on CADD.

The goal in developing the new approach was to take the overwhelming amount of data available and distill it down into a single score that can be more easily evaluated by a researcher or clinician. To accomplish that, CADD compares and contrasts the properties of 15 million genetic variants separating humans from chimpanzees with 15 million simulated variants. Variants observed in humans have survived natural selection, which tends to remove harmful, disease-causing variants, while simulated variants are not exposed to selection. Thus, by comparing observed to simulated variants, CADD is able to identify those properties that make a variant harmful or disease-causing. C scores have been pre-computed for all 8.6 billion possible single nucleotide variants and are freely available for researchers.

"We didn't know what to expect," Cooper said, "but we were pleasantly surprised that CADD was able not only to be applicable to mutations everywhere in the genome but in fact do a substantially better job in nearly every test that we performed than other metrics."

The CADD method is unique from other algorithms in that it assigns scores to mutations anywhere in human genomes, not just the less-than two percent that encode proteins (the "exome"). This unique attribute will be crucial as whole-genome sequencing becomes routine in both clinical and research settings.

More information: www.nature.com/ng/journal/vaop.../full/ng.2892.html

Provided by HudsonAlpha Institute for Biotechnology

Citation: New method developed for ranking disease-causal mutations within whole genome sequences (2014, February 7) retrieved 20 July 2024 from <https://medicalxpress.com/news/2014-02-method-disease-causal-mutations-genome-sequences.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.