GENOMICS

Scoring all human mutations

Combining 63 annotations provides a unified score for the potential deleteriousness of every possible human mutation.

When it comes to interpreting mutations in a genome of interest, choosing the best annotation tool can be daunting. A number of high-quality tools exist, but, as Gregory Cooper from the HudsonAlpha Institute observed, many score only a small number of annotations, such as protein-coding genes.

Cooper, together with Jay Shendure from the University of Washington, sought to distill the output of all existing annotation tools into a single score and provide such a score for every position in the human genome.

For their framework, named CADD (combined annotation–dependent depletion), the team used machine learning to distinguish benign from potentially deleterious mutations. First they compared 15 million single-nucleotide variants (SNVs) that are fixed in the human lineage (i.e., they appear in all humans and can thus be classified as benign) to 15 million simulated mutations that had not undergone any kind of selection and were therefore expected to contain a certain percentage of deleterious mutations. Mutations that no longer appear in the fixed alleles are likely to be deleterious because they were selected against in the ~10 million years since the split of the human lineage from our common human-chimpanzee ancestor.

A good illustration of this is stop codons. Only up to 200 stop codons are fixed in the human lineage; but in the simulations, the researchers found 8,000—a discrepancy indicating that there is strong selection to eliminate stop codons. The same principle holds for other annotations. "Pick your favorite annotation; to some extent they all capture deleteriousness," Cooper concludes.

The scientists then used the comparative data based on 63 different annotations to train a support vector machine and ran the resulting model to score all 8.6 billion possible SNVs of the human genome. "We precomputed the scores," says Cooper. "So you don't have to run the algorithm: you can just look up the score."

CADD scores will help interpret the genomes of patients with Mendelian diseases caused by high-penetrance mutations and also prioritize low-penetrance variants found in genome-wide association studies. The researchers found a substantial number of SNVs with high CADD scores in noncoding variants, supporting the hypothesis that mutations in regulatory regions contribute to many diseases. The team is committed to updating CADD regularly; the scores will improve further as new annotations become available.

Nicole Rusk

RESEARCH PAPERS

Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genetics* **46**, 310–315 (2014).