⤤ DISEASE GENETICS

All together now for variant interpretation

Genome studies are identifying a wealth of genetic variants that are associated with human traits, particularly diseases, but identifying which of the many variants are functionally relevant remains a major challenge. A new study now presents a unifying bioinformatic package for interpreting human single-nucleotide variants (SNVs) and small insertions and deletions that provides numerous advantages over existing tools.

Various complementary bioinformatic tools are available for predicting the functional impact (which is known as deleteriousness) of genetic variants. The tools are based on distinct principles; for example, some tools analyse any site in the genome and use comparative genomics to determine whether variants at that site occur less than expected during evolution in order to infer negative selection and hence deleteriousness. By contrast, some other tools analyse only variants in coding regions to predict the biochemical consequences on protein sequences.

To combine information from different tools in a systematic and objective manner, Kircher *et al.* trained an algorithm that is based on 63 distinct types of genomic annotations (including deleteriousness estimates from existing methods) for the ability to distinguish between two sets of alleles: a set of 14.7 million computationally simulated mutations occurring along the human reference genome, and a set of 14.7 million actual mutations that have occurred since the human–chimpanzee split and that are now at or near fixation in the human population. Relative to the simulated data, the actual fixed alleles are expected to contain fewer deleterious mutations owing to natural selection.

To test the performance of their tool, called the Combined Annotation-Dependent Depletion (CADD) method, the authors looked for expected patterns in the output 'C score' estimates of deleteriousness. C scores were highest for nonsense (that is, protein-truncating) variants, particularly when they are found in essential genes, whereas scores were low for commonly occurring standing variation in the human population.

Importantly, the method was superior to existing methods at distinguishing various pathogenic variants that underlie diseases from nearby benign variants, including benign variants in the same disease-causal gene. Finally, taking a genome-scale view of human disease, CADD reported greater cumulative deleteriousness of *de novo* mutations in the exomes of patients with autism-spectrum disorders or intellectual disability than that in the exomes of unaffected controls or siblings, and it outperformed existing methods for ranking known disease-causal human mutations in the ClinVar database relative to other variants genome wide.

The authors note some limitations of CADD — such as its focus on evolutionary selection rather than clinical pathogenicity per se — but it is likely to serve as a valuable and adaptable tool for investigating the genomic underpinnings of various complex traits and diseases.

Darren J. Burgess

ORIGINAL RESEARCH PAPER Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genet.* http://dx.doi.org/10.1038/ng.2892 (2014)
FURTHER READING Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Rev. Genet.* **12**, 628–640 (2011)

Superstock