How do rare genetic variants affect health? AI provides more accurate predictions

Whether we are predisposed to particular diseases depends to a large extent on the countless variants in our genome. However, particularly in the case of genetic variants that only rarely occur in the population, the influence on the presentation of certain pathological traits has so far been difficult to determine.

Researchers from the German Cancer Research Center (DKFZ), the European Molecular Biology Laboratory (EMBL) and the Technical University of Munich have introduced an algorithm based on deep learning that can predict the effects of rare genetic variants. The method allows persons with high risk of disease to be distinguished more precisely and facilitates the identification of genes that are involved in the development of diseases.

Every person's genome differs from that of their fellow human beings in millions of individual building blocks. These differences in the genome are known as variants. Many of these variants are associated with particular biological traits and diseases. Such correlations are usually determined using so-called genome-wide association studies.

But the influence of rare variants, which occur with a frequency of only 0.1% or less in the population, is often statistically overlooked in association studies. "Rare variants in particular often have a significantly greater influence on the presentation of a biological trait or a disease," says Brian Clarke, one of the first authors of the present study. "They can therefore help to identify those genes that play a role in the development of a disease and that can then point us in the direction of new therapeutic approaches," adds co-first author Eva Holtkamp.

In order to better predict the effects of rare variants, teams led by Oliver Stegle and Brian Clarke at the DKFZ and EMBL and Julien Gagneur at the Technical University of Munich have now developed a risk assessment tool based on machine learning. "DeepRVAT" (rare variant association testing), as the researchers named the method, is the first to use artificial intelligence (AI) in genomic association studies to decipher rare genetic variants.

The model was initially trained on the sequence data (exome sequences) of 161,000 individuals from the UK Biobank. In addition, the researchers fed in information on genetically influenced biological traits of the individual persons as well as on the genes involved in the traits. The sequences used for training comprised around 13 million variants. For each of these, detailed "annotations" are available, providing quantitative information on the possible effects that the respective variant can have on cellular processes or on the protein structure. These annotations were also a central component of the training.

After training, DeepRVAT is able to predict for each individual which genes are impaired in their function by rare variants. To do this, the algorithm uses individual variants and their annotations to calculate a numerical value that describes the extent to which a gene is impaired and its potential impact on health.

The researchers validated DeepRVAT on genome data from the UK Biobank. For 34 tested traits, i.e., disease-relevant blood test results, the testing method found 352 associations with genes involved, far outperforming all previously existing models. The results obtained with DeepRVAT proved to be very robust and better replicable in independent data than the results of alternative approaches.

Another important application of DeepRVAT is the evaluation of genetic predisposition to certain diseases. The researchers combined DeepRVAT with polygenic risk scoring based on more common genetic variants. This significantly improved the accuracy of the predictions, especially for high-risk variants. In addition, it turned out that DeepRVAT recognized genetic correlations for numerous diseases - including various cardiovascular diseases, types of cancer, metabolic and neurological diseases - that had not been found with existing tests.

"DeepRVAT has the potential to significantly advance personalized medicine. Our method functions regardless of the type of trait and can be flexibly combined with other testing methods," says physicist and data scientist Oliver Stegle. His team now wants to further test the risk assessment tool in large-scale trials as quickly as possible and bring it into application. The scientists are already in contact with the organizers of INFORM, for example. The aim of this study is to use genomic data to identify individually tailored treatments for children with cancer who suffer a relapse. DeepRVAT could help to uncover the

genetic basis of certain childhood cancers.

"I find the potential impact of DeepRVAT on rare disease applications exciting. One of the major challenges in rare disease research is the lack of large-scale, systematic data. Leveraging the power of AI and the half a million exomes in the UK Biobank, we have objectively identified which genetic variants most significantly impair gene function," says Julien Gagneur from the Technical University of Munich.

The next step is to integrate DeepRVAT into the infrastructure of the German Human Genome Phenome Archive (GHGA) in order to facilitate applications in diagnostics and basic research. Another advantage of DeepRVAT is that the method requires significantly less computing power than comparable models. DeepRVAT is available as a user-friendly software package that can either be used with the pre-trained risk assessment models or trained with researchers' own data sets for specialized purposes.

Info:

With more than 3,000 employees, the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ) is Germany's largest biomedical research institute. DKFZ scientists identify cancer risk factors, investigate how cancer progresses and develop new cancer prevention strategies. They are also developing new methods to diagnose tumors more precisely and treat cancer patients more successfully. The DKFZ's Cancer Information Service (KID) provides patients, interested citizens and experts with individual answers to questions relating to cancer.
To transfer promising approaches from cancer research to the clinic and thus improve the prognosis of cancer patients, the DKFZ cooperates with excellent research institutions and university hospitals throughout Germany:

- National Center for Tumor Diseases (NCT, 6 sites)
- German Cancer Consortium (DKTK, 8 sites)
- Hopp Children's Cancer Center (KiTZ) Heidelberg
- Helmholtz Institute for Translational Oncology (HI-TRON Mainz) - A Helmholtz Institute of the DKFZ
- DKFZ-Hector Cancer Institute at the University Medical Center Mannheim
- National Cancer Prevention Center (jointly with German Cancer Aid)

The DKFZ is 90 percent financed by the Federal Ministry of Education and Research and 10 percent by the state of Baden-Württemberg. The DKFZ is a member of the Helmholtz Association of German Research Centers.

Publication:

Brian Clarke, Eva Holtkamp, Hakime Öztürk, Marcel Mück, Magnus Wahlberg, Kayla Meyer, Felix Munzlinger, Felix Brechtmann, Florian R. Hölzlwimmer, Jonas Lindner, Zhifen Chen, Julien Gagneur, Oliver Stegle: Integration of Variant annotations using deep set networks boosts rare variant testing. Nature Genetics 2024, DOI: https://www.nature.com/articles/s41588-024-01919-z

Press release

25-Sept-2024
Source: German Cancer Research Center

Further information

- German Cancer Research Center