Medical Xpress
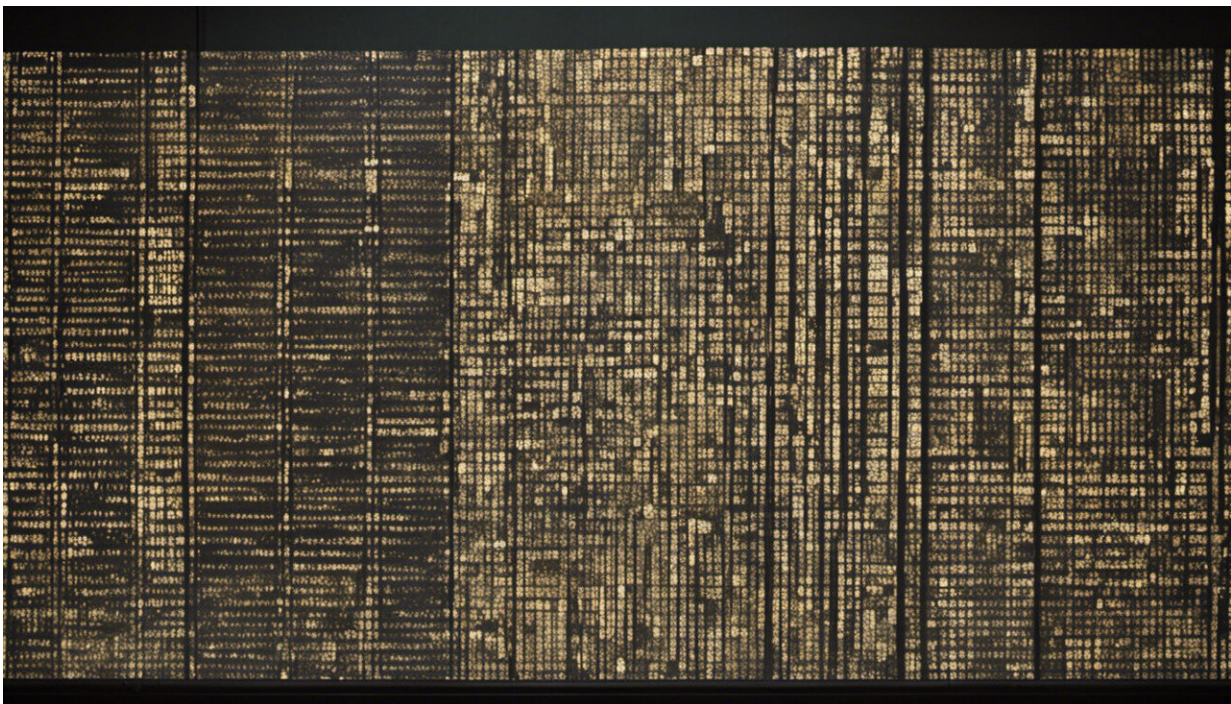
Largest collection of human exome sequence data yields unprecedented tool for diagnosing rare disease

August 18 2016

Credit: AI-generated image (disclaimer)

Based on the largest resource of its kind, members of the Exome Aggregation Consortium (ExAC) led by scientists at the Broad Institute of MIT and Harvard report scientific findings from data on the exome sequences (protein-coding portions of the genome) from 60,706 people

from diverse ethnic backgrounds. Containing over 10 million DNA variants – many very rare and most identified for the first time – the ExAC dataset is a freely available, high-resolution catalog of human genetic variation that has already made a major impact on clinical research and diagnosis of rare genetic diseases.

Featured in the August 18 issue of *Nature*, analysis of the data reveals properties of genetic variation undetectable in smaller data sets, for example, the first direct observation of mutations that arose multiple times independently among the samples – so-called "mutational recurrence." The work also uncovers a class of genes that harbor less variation than expected, representing likely disease-causing DNA variants that are rare or absent in the population because they are so detrimental to human health. With immediate utility for clinical applications, the study further shows that the ExAC database improves the ability to evaluate candidate pathogenic variants in rare disease.

"The success of ExAC was made possible by the willingness of our colleagues in many large, disease-focused consortia to openly share sequencing data," said Daniel MacArthur, senior author of the study, co-director of the Program in Medical and Population Genetics at the Broad Institute, and an assistant professor at Massachusetts General Hospital and Harvard Medical School. Previous resources contained far fewer exomes without much ancestral diversity, so they were inadequate for studies of rare disease variants. "The scale and diversity of the ExAC resource is invaluable," MacArthur added. "It gives us the ability to discover extremely rare variants and offers an unparalleled window into the roots of rare genetic diseases."

After collecting the raw data from tens of thousands of human exomes from research collaborators around the globe, the consortium relied upon the analytical and computational capabilities of the Broad Institute's Genomics Platform and Data Sciences and Data Engineering group to

produce a catalog of [human genetic variation](#) of unprecedented resolution – roughly one variant every eight bases, or letters, of DNA. Many of these variants had never been reported and most are very rare, occurring in less than 1 in 10,000 people.

With a patient's genome sequence in hand, a clinician can compare any rare mutations found in his or her genome with those in the ExAC database, shedding light on the genes and proteins that may underlie a patient's disorder. A variant found in a patient's DNA sequence that is extremely rare in ExAC, especially one that is predicted to disrupt the function of the resulting protein, then becomes a key suspect in causing the rare disease. Since its release to the scientific community in October 2014 the ExAC resource has had more than five million page views online, and has allowed clinicians to provide more accurate genetic diagnoses for thousands of rare disease patients.

"The ExAC resource gives us incredible insight when evaluating a patient's genome sequence in the clinic," said Heidi Rehm, medical director of the Broad's Clinical Research Sequencing Platform and chief laboratory director of the Laboratory for Molecular Medicine at Partners Personalized Medicine. In clinical sequencing, many DNA variants are rare or understudied, so it is unclear if they have any effect on disease risk and whether they should be taken into consideration when diagnosing and treating patients. By looking at the frequencies of a patient's variants in the ExAC database, Rehm and her team can rule out those that are relatively common, allowing them to more quickly home in on the true disease-causing variants and avoid costly follow-up on benign ones.

The resource has also been used by researchers to identify dozens of new rare genetic disorders. "In our own research, using the ExAC resource has allowed us to apply novel statistical methods to identify several new severe developmental disorders," said Matthew Hurles, a researcher at

the Wellcome Trust Sanger Institute and frequent user of the ExAC database. "Resources such as ExAC exemplify the benefits that can be achieved for families coping with rare genetic diseases, as a result of the mass altruism of many research participants who allow their data to be aggregated and shared."

The ExAC database is also being used by researchers exploring the more fundamental effects of genetic variation, for example, looking at variation in transcription factor proteins and its impact on protein-protein interaction networks.

Interestingly, variation that was expected, but not found, in the data offered new insight. Some genes were found to have less than the expected number of missense mutations, which change the protein sequence, or "loss-of-function" mutations, which obliterate protein function. With such a large sample size, the researchers were able to quantify the deficit of these types of mutation per gene, identifying a few thousand "highly constrained" genes for which natural selection has weeded out these mutations because their effects are so detrimental. With no knowledge about the diseases they cause and often no actual instances of these mutations in the ExAC database, the "missing variation" indicates that these highly constrained genes are likely to cause severe disease. If clinical or research sequencing reveals a loss-of-function or missense mutation in one of these genes in a patient's genome, it becomes a strong candidate for causing his or her rare disease.

The ExAC data also revealed more than 100 previously reported disease-causing mutations to actually be benign, reducing the number of these false positive findings in databases widely used by clinical labs. This finding demonstrates the value of the ExAC database in assessing claims that specific mutations cause disease.

"With its large sample size and high resolution across many populations, the ExAC database provides much greater power to interpret rare disease-causing variants than ever before, even for common diseases," said Jose Florez, an institute member at the Broad, chief of the Diabetes Unit at the Massachusetts General Hospital, and an associate professor at Harvard Medical School.

More information: Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. August 18, 2016. [DOI: 10.1101/030338](#)

Provided by Broad Institute of MIT and Harvard