NEWS

Huge data-sharing venture lays bare human genetic variation

BY **JESSICA WRIGHT**

8 SEPTEMBER 2016

A landmark collection of DNA sequences includes the code for every gene in the genomes of more than 60,000 people[1]. An analysis of this dataset was published 18 August in *Nature*, but scientists have been using the **database since its release** nearly two years ago.

The collection, called the Exome Aggregation Consortium (ExAC) browser, provides an unprecedented picture of genetic variation across the general population. It allows researchers to see whether a seemingly rare variant crops up in the general population — in which case the variant is unlikely to be harmful.

Despite studies generating troves of sequencing data, no previous database was large enough to robustly link a mutation to a particular condition. ExAC grew out of this realization, says lead researcher **Daniel MacArthur**, assistant professor of genetics at Harvard University.

“The existing databases of normal variation just weren’t big enough to actually give us a sense of the variation in the population at the extremely rare end of the spectrum,” he says.

MacArthur reached out to teams responsible for sequencing large numbers of exomes — or coding regions of the genome — and asked for their raw data. MacArthur and his team then identified variants in 60,706 genomes from 17 databases.

In an unusual move, the researchers immediately made their list of variants available online. They also posted **a draft of their results** on the preprint server bioRXiV last October.

“To give away your data two years before publishing it, who the hell does that?” says **Michael Talkowski**, associate professor of neurology at Harvard University, who was not involved in the study, applauding the researchers for their generosity.

Thanks to their efforts, the database became a go-to resource for researchers well before the study's publication, adds **Stephan Sanders**, assistant professor of psychiatry at the University of California, San Francisco, who was not involved in the study.

Sanders says he routinely checks any variant he finds in an individual with autism to see how common it is in the general population. "You know a resource is important when it just becomes part of your day-to-day life."

Ridiculous density:

MacArthur and his colleagues started with raw sequencing data from 91,000 individuals. Many of these people have no known conditions, but some have diabetes, heart disease, schizophrenia, bipolar disorder or cancer.

The researchers excluded sequencing data that didn't meet certain quality standards. They left out data from individuals who have a childhood condition such as autism or pediatric cancer, as well as from these individuals' relatives. They also looked at data only from unrelated individuals so as to avoid counting the same inherited mutation more than once.

The final sequences make up a petabyte of raw data — enough to fill up 4,000 laptops, says MacArthur. The researchers reanalyzed these sequences, identifying sites where an individual's sequence stands out from the group.

They found that these sites of variation are common. ExAC contains more than 7 million variants — one for every eight base pairs across the **exome**. Of these, nearly 200,000 are likely to abolish function of the gene's corresponding protein. "It really is a ridiculous density of variation," MacArthur says.

The database provides an essential check on results that implicate a mutation in autism. "In autism, my bent has always been that early studies on small samples are over-interpreting [which mutations are causative]," Talkowski says.

Genetic diversity:

MacArthur and his team showed that 163 of 192 variants previously thought to cause a condition are present in more than 1 percent of the population and are probably harmless. Of these 163, 126 have officially been reclassified as of December.

Many of the reclassifications relied on sequences in ExAC derived from specific populations, such as South Asians and Latinos, that are often underrepresented in genetic studies. Some variants may be common only in certain groups, making their inclusion critical to determining whether a variant is linked to a certain condition.

For example, the researchers found four Mexican individuals in the database who each carry two copies of a mutation that tracks closely with liver disease in Native American families. None of these four people have any signs of liver disease, however, suggesting that this mutation is not responsible for liver disease.

Still, the database is “lacking big swaths of human genetic diversity,” says MacArthur. It does not represent people from the Middle East, Central Asia, Oceania, Western Asia and large portions of Africa, for example. “I worry that without good representation of these other ancestries, we’ll continue to misdiagnose patients from those backgrounds,” MacArthur says.

Because the database includes sequences from individuals with certain conditions, scientists can check with the ExAC team whether their variant of interest comes from one of these individuals. (This information is not publicly available for privacy reasons.) Autism researchers can access a version of the database that does not include sequences from individuals with schizophrenia or bipolar disorder, both of which **share genetic risk factors** with autism.

Protected regions:

The database also highlights genes resistant to mutation. The researchers calculated the likelihood of a gene **acquiring a mutation** based on its length and sequence. Their algorithm revealed 3,230 genes that have significantly fewer harmful mutations than expected. This suggests that mutations in these genes are selected against, and so are particularly harmful.

This list of protected genes includes most top autism candidate genes as well as many genes associated with intellectual disability **and schizophrenia**. “It’s kind of cleared the air around the really important [autism] genes,” Sanders says.

MacArthur and his colleagues continue to expand their collection with as many sequences as they can find. They aim to have a total of 120,000 sequences in the database by October and are working on adding another 80,000.

REFERENCES:

1. Lek M. *et al. Nature* **536**, 285-291 (2016) **PubMed**