NEWS

Massive sequencing database helps interpret mutations' role

BY **JESSICA WRIGHT**

23 OCTOBER 2014

R

esearchers have analyzed more than 90,000 **exomes** — the protein-coding regions of the genome — the largest such set yet, they announced Monday at the **American Society of Human Genetics Annual Meeting** in San Diego. The resource gives scientists an invaluable tool to probe the significance of specific mutations.

The Exome Aggregation Consortium, which released the data, has already made 63,538 sequences freely available online. All of these sequences are from individuals in the general population who are free of severe psychiatric conditions, and include people from all major continents.

This will allow researchers to map the prevalence and spread of variants in pretty much every gene in the human genome, says **Daniel MacArthur**, assistant professor of medicine at Harvard University.

"The motivation behind the consortium is that to really deeply understand the variation we find in any one individual's genome, we need to place it into the context of tens of thousands of other individuals," he says.

In theory, collecting these thousands of control sequences is easy, MacArthur says. Researchers across different disciplines have sequenced hundreds of thousands of exomes. But these sequences remain largely segregated in the silos of different research projects and are analyzed using different techniques.

MacArthur and his collaborators assembled and analyzed raw data from several databases, including the **Cancer Genome Atlas**, the **UK10K** project, and various autism and schizophrenia collections. There are 200,000 exomes in total worldwide and and they have access to nearly 92,000.

Genetic diversity:

The resource may be especially useful for autism research. Studying the exomes of people with autism has revealed hundreds of mutations, but it is unclear whether many of these mutations **contribute to autism risk**, says **Jeremy Willsey**, a postdoctoral fellow in **Matthew State**'s lab at the University of California, San Francisco, who was not involved in the work. This is a "huge problem," Willsey says.

Comparing against these reference exomes will help researchers make sense of the autism data, he says. In particular, exome studies tend to ignore missense mutations, which are sometimes harmless, but the new dataset may be large enough to include them in the analysis.

Previously, researchers had access to only about 6,500 control sequences from a different database. "Now we essentially have a ten-times-larger dataset, so you can imagine how much better we are at understanding what's truly rare in the population," says Willsey. "It's such an exciting dataset — especially the fact that they're making it available."

Overall, the data include more than 8 million variants, including more than 3 million missense mutations. Half of these are so rare that they are present only once among 63,000 people.

The researchers also identified more than 200,000 mutations that would abrogate the function of a protein. They also identified 20,000 variants that had been found in people with a severe disorder. The database does not provide information on symptoms or conditions, however.

"It will be critical to know if individuals with damaging mutations do — or do not — have one or another of the conditions," says **Mary-Claire King**, professor of genome sciences at the University of Washington in Seattle, who was not involved in the study. "It is of no help to know that a damaging mutation has occurred if it is not possible to learn whether it occurred in a person with autism, or colon cancer, or schizophrenia or none of these conditions."

MacArthur aims to make information on symptoms available to researchers, but says his team first has to resolve the ethical concerns inherent to sharing participant data.

Mutation map:

In the meantime, researchers can use the database to map when certain mutations arose during evolution. For example, the data reveal that a harmless variant in one gene, TH, is distributed all over the world, suggesting that it arose before humans migrated out of Africa. By contrast, another rare disease-causing mutation in the gene GLDC is present only in the Finnish population.

In a study published earlier this year, **Mark Daly** and his colleagues calculated **baseline rates of mutations** for every gene in the human genome, based on its sequence. They then used the 6,500 available exome sequences to look for genes that have fewer than the expected number of mutations — suggesting that mutations that fall in these genes will be harmful.

In another presentation Monday, the researchers showed that using ten times the number of exomes greatly increases the resolution of this analysis. Loss-of-function mutations identified in autism exome studies tend to fall into the group of genes protected from mutation, supporting their significance as autism genes, says **Kaitlin Samocha**, a graduate student in Daly's lab at Harvard Medical School who presented the work.

Using the database to see if a variant is protected in this way is "fantastically useful," says **Ivan Iossifov**, assistant professor at Cold Spring Harbor Laboratory in New York, who was not involved in the work.

With the variant information available in the new database, researchers can also break down genes into component parts and prioritize missense mutations based on where they hit the gene.

"Now that we have that data, we can treat parts of genes, or functional domains of genes, and evaluate those as opposed to treating the gene as a whole unit," says Samocha. "We would be unable to do this without the 63,000 individuals."

*For more reports from the 2014 American Society of Human Genetics Annual Meeting, please click **here**.*

***Correction:** This article has been modified from the original. The researchers have access to 90,000 exomes, not 200,000 as the previous version stated. Also, they identified 20,000 variants that have been found in people with a severe disorder, not 20,000 loss-of-function mutations in genes linked to disorders.*