JOE RAEDLE/GETTY

Vast stores of DNA samples and data have been produced by the increasing pace of genetic sequencing.

GENOMICS

Giant gene banks take on disease

Researchers bring together troves of DNA sequences in the hope of teasing out links between traits and genetic variants.

BY ERIKA CHECK HAYDEN

Early last year, three researchers set out to create one genetic data set to rule them all. The trio wanted to assemble the world's most comprehensive catalogue of human genetic variation, a single reference database that would be useful to researchers hunting rare disease-causing genetic variants.

Unlike past 'big data' projects, which have involved large groups of scientists, this one deliberately kept itself small, deploying just five analysts. Nearly two years in, it has identified about 50 million genetic variants — points at which one person's DNA differs from another's — in whole-genome sequence data collected by 23 other research collaborations. The group, called the Haplotype Reference Consortium, will unveil its database in San Diego, California, on 20 October, at the annual meeting of the American Society of Human Genetics.

Geneticists have not always been so willing to share data. But that seems to be changing. "It's been surprisingly easy to bring all these data sets together," says Jonathan Marchini, a statistical geneticist at the University of Oxford, UK, and one of the consortium's leaders. "There is a lot of goodwill between the people in the field; they all understand the benefits of doing this and have worked hard to make their data available."

In the past five years, there has been an explosion in rates of sequencing human genomes thanks to the falling cost of the technology. At the same time, geneticists have realized that linking genes to diseases and traits will require much bigger sample sizes than any one research centre can assemble.

It was once assumed that common diseases and traits could be traced to a few common genetic variants that would be relatively easy to find. But that has turned out not to be the case. It is now clear that thousands of different variants each play a small part in determining a person's height or risk of schizophrenia, for example. And finding those thousands of variants means looking at a daunting number of people. At the same time, the increased pace of genetic sequencing has made it possible to collect enough genomes to uncover those variants.

"Here are a bunch of data sets that individually cost millions of dollars to generate, and you have people willing to make that data available to a shared resource, which is amazing," says geneticist Daniel MacArthur of Massachusetts General Hospital in Boston.

MacArthur is part of the Exome Aggregation Consortium, another attempt to create a supersized library of human genetic variation. On 20 October, MacArthur and his colleagues plan to unveil their own public database containing the protein-coding portions, or exomes, of 63,000 human genomes originally gathered by other researchers. "We can say from looking at a very large cohort of people … this is what the distribution of rare variation looks like," says MacArthur. "And that is very powerful."

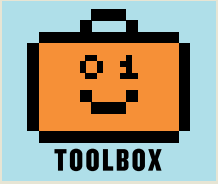
MacArthur is developing tools to comb the data for mutations that disable genes. Only some of these 'loss-of-function' mutations cause harm; predicting which are pathogenic will require knowing more about which ones regularly occur in healthy people.

Some studies are already reaping the benefits of huge data sets. On 5 October researchers published a paper on the genetics of height that included data on more than 250,000 people (A. R. Wood *et al. Nature Genet.* http://doi.org/v6k; 2014). The data had been gathered in separate genome-wide association studies, which look for correlations between genetic variants and traits or diseases, and pooled as part of the Genetic Investigation of Anthropometric Traits (GIANT) Consortium. The paper reported 697 new variants linked to height, more than tripling the previous count. Still, researchers estimate that the hundreds of common variants now identified account for only 16% of the genetic contributors to height.

Throwing even more data into the pool could reveal some of the rest, says Joel Hirschhorn, a geneticist at the Broad Institute in Cambridge, Massachusetts, and a leader of the GIANT consortium. ■

MORE ONLINE

TOOLBOX Q&A



How to publish your peer-review comments go.nature.com/rokwhe

MORE NEWS

- Experiment mimics black hole radiation in the laboratory go.nature.com/xmtnpy
- More finds on Antikythera shipwrek go.nature.com/odmwtp
- Fish show limits of mirror studies go.nature.com/eonvgs

NATURE PODCAST

Alzheimer's in 3D culture, fracking and CO_2, and why interdisciplinary science is so hard nature.com/nature/podcast