CROSS TALK

How to mine treasure troves of genetic data

BY **JESSICA WRIGHT**

5 APRIL 2016

Scientists are sequencing whole genomes of people with autism at an unprecedented scale, generating massive amounts of data. As autism researchers grapple with this shared wealth, collaboration and transparency will be crucial, they say.

The list of genes tied to autism has grown dramatically over the past few years, thanks in large part to **two massive projects**. Each of the projects ventured to sequence the **exomes** — regions of the genome that code for proteins — of thousands of people with autism and their family members. This approach can identify mutations in genes, but skips over the non-coding sections in between. Other methods capture **large structural variations** in the DNA of people with autism, but miss small, subtle irregularities.

Until now, scanning through the entire genome has not been possible at a large scale. But a lower cost of sequencing, along with advances in technology, has made whole-genome sequencing for thousands of genomes suddenly possible. This breakthrough, researchers say, is poised to propel our understanding of the genetics of autism by identifying many more and new types of genetic variants.

Autism Speaks is spearheading **an effort called MSSNG** (a name that is symbolically 'missing' vowels) to **sequence the whole genomes** of 10,000 people with autism or their family members. Half of the sequences are already available to researchers on **an open platform** hosted by Google.

Meanwhile, the Simons Foundation (*Spectrum*'s parent organization) is funding the **New York Genome Center** to **sequence the whole genomes** of more than 2,600 families in the **Simons Simplex Collection** (SSC). These families each have one child with autism, unaffected parents and one unaffected sibling. The first stage of the project — 2,000 sequences from 500 families — is nearly complete, with 1,982 sequences already available to scientists.

With funds from the National Institutes of Health, the center is also sequencing another 500 of these families and expects to have 2,000 more sequences available later this year. The ultimate

aim is to sequence genomes from all 2,664 families.

Together, these two efforts are generating massive amounts of data that researchers can scour for clues to the genetic basis of autism. We asked experts what they hope to learn from this treasure trove of sequences and how best to work with the communal resources.

Steve Scherer

Director, Centre for Applied Genomics,Hospital for Sick Children in Toronto, Canada

Two separate whole-genome sequencing studies may uncover different aspects of autism's genetic architecture.

Different datasets paint complementary pictures

I think it is important to sequence a well-characterized cohort like the SSC, as there are a lot more variants to be found. As our team has shown, whole-genome sequencing **uncovers new genetic variants** that other technologies miss. And it can find variants that reside in the 98 percent of DNA that does not code for genes. With these new genetic variants, and by looking in the entire genome, I'm confident we will uncover interesting new forms of autism genetic risk.

The goal of our study, called MSSNG, is similar: to sequence the whole genomes of autism families. The families included in this venture — a collaboration between Autism Speaks, Google and the Hospital for Sick Children — probably capture a broader cross-section of people with autism than does the SSC, which used stringent criteria to collect families that have only one child with autism. These 'simplex' families may have a unique genetic architecture, because they may be enriched in mutations that arise spontaneously, instead of ones that are inherited.

Through MSSNG, we have sequenced the genomes of children with autism and their parents, called trios, as well as 'quads,' which include two affected children from families that have at least two children with autism. In some special cases, we have sequenced extended pedigrees from these families going back at least three generations.

So far, we have sequenced more than 7,000 genomes, with 5,000 of them available to researchers. Nearly 100 investigators from 37 institutions are signed into our system and are using the data.

We've worked hard make the portal easily accessible, so that the medical genetics community can also use the data. An increasing number of diagnostic laboratories are performing genetic testing for autism and related disorders. They need to compare their testing data against all available

sources of information to see what it might mean. To make this comparison, clinicians need to be able to quickly search the massive whole-genome sequence datasets. We're working to set up the portal to make this search as simple as possible.

Michael Ronemus

Research Assistant Professor,Cold Spring Harbor Laboratory

Having a complete picture of autism's genetic variants will help answer many important outstanding questions in autism research.

Best answers yet to autism questions

Having spent most of the past decade working with the SSC, I can't deny that it will be incredibly interesting to see the results of the whole-genome sequencing project. Having the sequence of the entire genome is the holy grail for understanding the molecular genetics of autism spectrum disorder.

I think we've become fairly proficient at finding large DNA duplications and deletions, which are called **copy number variants** (CNVs), and changes to single nucleotides in the coding sequence in people with autism. But these are only a fraction of all the possible variants. We need the entire sequences to provide more accurate estimates of the frequencies of these types of mutations.

We will finally be able to start looking for other types of genetic variants in individuals with autism, such as noncoding RNAs, structural variants that **flip or swap segments of DNA** and small CNVs not detectable by previous methods.

A second, but very important consideration is that we will have to figure out how to deal with having so much data. We will need to develop new algorithms. Having such a rich set of data available will be a huge boon to this discovery process.

It's also going to take a truly coordinated effort to build the computational infrastructure to store and delve through all of it. Despite whatever good intentions we have going in, there will undoubtedly be unforeseen issues. The importance of whole-genome sequencing data overall is only going to increase, and I believe that this project will create a template for how to conduct large, family-based, whole-genome sequencing studies.

There are still many unanswered questions in autism. For instance, what is the basis of the **female protective effect**, in which it takes more severe mutations to lead to autism in women than it does in men? Do unaffected mothers of children with autism **carry autism genes**? What is the contribution to autism risk of variants that are **common in the general population** versus those

that are **rare**?
These are just some of the most obvious questions, and whole-genome sequencing is likely to provide the best answers yet. But I would not be surprised if having such comprehensive data from such a large set of families yields lots of interesting findings that have nothing directly to do with autism. Exploring the sequences of this collection has already been fascinating, and will continue to become even more so.

Lucia Peixoto

Assistant Professor of Biomedical Sciences,Washington State University Spokane

To spread the wealth from the new genetic data, researchers should make their methods transparent.

Tools and analysis must accompany raw data

I'm excited to do a lot of things with the new autism whole-genome sequencing data. But there is a challenge with open-access data that I've seen a lot in many different fields: We need to think about how to make the data truly accessible to as many people as possible.
Once data are released, it's assumed that they will be easy to use. But that's not really true. Depending on the type of data, how they're accessed, and whether they're raw or processed data, the ability to take advantage of their power could be limited to the researchers who have the expertise needed to generate the data to begin with. What we really want is for all researchers to be able to use the public data. People may want to take the new data and compare them with their own, or just to search the data for the three or four genes they're interested in. If all kinds of researchers are looking at a dataset from different points of view, that's how we can really generate new knowledge.
To achieve this goal, it is important that researchers make every step they took to analyze the data publicly available, not just the raw data and the results of the analysis. It is important to make the tools used to analyze the data available as well. Otherwise, researchers might spend time and money reproducing tools that already exist, which is just a poor return on investment.
There's also the risk that researchers might end up with different results from the same dataset without knowing what led to the different conclusions. I know it's hard work to share methodology and analysis in a way that's understandable to the non-expert, but it helps everybody to go the extra mile to do it. We are all trying to solve a problem that we care about. We all care about autism. Let's have a platform that we can all use to help each other out.

People used to think that the researchers who generated data and those who analyzed the data were somehow fundamentally different. There was even the impression that researchers who relied on publicly available datasets were not generating new knowledge. But this view is changing. We're realizing that there is so much data that one person or research group can't possibly do all of the analysis on their own.