

## TECHNOLOGY FEATURE

# WHEN DISEASE STRIKES FROM NOWHERE

*When healthy parents have a child with a genetic disorder, the cause is sometimes a new mutation. Tools are emerging to meet the challenge of finding such changes.*

BRAEDOSTOK / SHUTTERSTOCK



Children born with disorders not readily explained by standard tests can sometimes be diagnosed through genome sequencing and analysis.

BY VIVIEN MARX

When parents find that a child is not developing as expected, the protracted doctor visits, hospital stays and examinations only add to their distress — especially when no other family member has the condition and the standard tests on the child's blood and genes shed no light on the cause. The uncertainties, costs and anguish can be devastating to families, says Michael Friez, who directs the diagnostic laboratory at the Greenwood Genetic Center in South Carolina, a non-profit organization that analyses patients' genomes for clinicians.

Every clinical geneticist has experienced the inability to identify the cause of a child's neurodevelopmental disorder, adds Roger Stevenson, a clinical geneticist also at the centre. In the early 2000s, he began seeing a family with a toddler that had severe developmental problems, including a smaller-than-average head and intellectual disability.

It was more than a decade after their first visit before sequencing revealed that the boy had a mutation in a gene called *DYRK1A*, which is thought to have a role in brain development. The finding later helped to diagnose 16 other children in the United States and Europe who had the same symptoms — and

although the condition has no cure, Stevenson saw that identifying the gene comforted the boy's parents, as did knowing that there were other children like their son.

### NEW MUTATIONS

What was notable about this child's case was that it involved a *de novo* mutation — one that neither parent carries in their regular complement of DNA. *De novo* mutations can occur early in the development of the embryo. They can be in parents' gametes. Around 80% of *de novo* mutations seem to occur in the father's sperm and 20% in the mother's egg, says Joris Veltman, a geneticist at Radboud ►

► University Medical Center in Nijmegen, the Netherlands, who in July published a study of *de novo* mutations in people with intellectual disabilities<sup>1</sup>.

Disorder-causing *de novo* mutations are hard to detect — they have to be identified among a host of other, innocuous genetic changes. A number of software-based approaches are emerging to sift through sequenced genomes in search of such mutations.

As sequencing instruments and databases of genetic information become increasingly available, tool-builders hope that their software contributions can become part of routine medical care. But sequencing and analysis are different from, say, a blood cholesterol test — samples have to be prepared for the instruments, which churn out the genome sequence in snippets that must be assembled and aligned to a reference genome, such as that curated by the Genome Reference Consortium.

The results are not perfect. A patient's genome sequence can contain errors — caused by the machine misreading a letter of DNA, for

**“There is no single magic trick that makes our method work well.”**

example — that must be filtered out computationally. And even then, a huge number of possibilities remains. DNA bases might differ from the reference, sequences can be inserted or deleted and the number of copies of a gene can vary. Of thousands of such changes, only one might have a role in a disorder.

The child's DNA is then compared with that of the parents. Again, not all differences between their genomes connect to the child's disorder. Researchers use software that includes statistical analyses to determine which changes are most likely to have a role. And the tools add information, such as published data about the links between genes and disease. These results help to create lists of genetic changes, or variants, ranked by likelihood of being linked to a disorder. But variant analysis is still an emerging science, and the software tools are still maturing. Despite this, in some cases the approach turns up a specific genetic change that is likely to be the cause of a disorder.

### ASSORTED VARIANTS

Finding the probable genetic culprit does not mean a treatment is available. But such results help parents to cope with the situation, says Donald Conrad, a geneticist at Washington University in St Louis, Missouri. The results also inform parents about the risk that the condition might recur in their family and help them to plan future pregnancies. And some prospective parents might opt for genetic analysis as part of *in vitro* fertilization.

Most newborns carry about 60–100 *de novo* variants, says Conrad — few of which cause

any discernible problem. Software helps to sort these variants out. Conrad has developed DeNovoGear, which does statistical analysis to distinguish potentially important signals from background noise caused by experimental error<sup>2</sup>. The software also analyses the nature and frequency of sequencing errors. It then compares the genomes of parents, children and other family members to distinguish true *de novo* mutations from other types of genetic variation.

To improve the odds of finding such mutations, the analysis takes into account the frequency of known variation at a given site in the genome. It does so by drawing on data from the 1000 Genomes Project, an international research consortium that catalogues human genetic variation. “There is no single magic trick that makes our method work well,” Conrad says. “It is just the accumulation of many different attempts to squeeze out as much information as possible.”

### RAISING THE ODDS

The software must contend with the errors made by the sequencing instruments — reporting a ‘C’ as a ‘T’, for example. These mistakes are rare but hard to predict, says Conrad, and may explain many false-positive results in searches for *de novo* mutations. High-throughput sequencing instruments are more prone to error in some DNA regions — which also turn out to be where cells are more likely to make mistakes when copying and repairing the genome. These tricky places account for about 15% of the genome, Conrad notes, so current methods can reliably detect *de novo* mutations only in the other 85%.

Even the best software tools come up with 2–3 times as many false positives as true positives when analysing whole-genome sequence. True positives have to be teased out with follow-up experiments — for example, by using the laborious but precise Sanger sequencing method to look at the genetic region in question. “Each sequencing platform has its own idiosyncrasies,” Conrad says, and the optimal method for detecting *de novo* mutations needs to incorporate the machine's quirks into its statistical models.

Conrad is also developing statistical methods that take account of the frequency of various sequencing errors in different regions of the genome. Other software tools typically apply the same error estimates at all genomic sites. Other researchers are pursuing their own approaches.

For the study published in July<sup>1</sup>, Veltman and his colleagues sequenced and analysed the genomes of 50 people with severe intellectual disabilities (see ‘Better diagnosis’). Working with Complete Genomics (CG) in Mountain View, California, a division of the genomics giant BGI in Shenzhen, China, they identified *de novo* mutations by drawing on a number of resources. They used BGI's



Martin Reese, chief scientific officer of Omicron.

technology and software to analyse and compare genomes and whittle down the number of possible disease-causing candidates. They did not analyse the data with other tools such as DeNovoGear, so Veltman cannot compare the methods. But the advantage with BGI's analysis suite is that the software has been matched to the specifications of the sequencing technology, he says.

All the patients in the study had previously undergone extensive testing. Protein-coding regions of their genomes had been analysed, and microarrays were used to analyse variations in gene-copy number, which can occur from person to person and also in some disorders. Veltman says that the software showed high sensitivity in detecting *de novo* mutations, which enabled a more accurate diagnosis of almost half of the patients.

### INTERPRET CAREFULLY

In Veltman's view, interpreting mutations is now more possible for disorders such as intellectual disability than for diseases such as cancer or diabetes, because many cases of severe intellectual disability seem to be caused by a single mutation. But, he says, the few hundred genes that the scientific community has found to be implicated in intellectual disability form a still-incomplete list.

Veltman stresses that the sequencing quality in the study was good, but says that even the best sequencing technology can miss or misidentify *de novo* mutations. To minimize errors, researchers need to seek out the highest-quality genome sequencing, he says. Beyond that, interpreting the many genetic variations that turn up when comparing genomes — and figuring out which ones are related to a disorder — is the field's major bottleneck. Researchers also need to find better ways to analyse *de novo* mutations in the genome's non-coding regions, which are still difficult to interpret.

One suite of tools to analyse protein-coding

and non-coding genome regions is FastQForward, which integrates the software programs VAAST<sup>3</sup>, pVAAST<sup>4</sup> and Phevor<sup>5</sup>. These tools were co-developed by Mark Yandell, a computational geneticist at the University of Utah in Salt Lake City who directs software development and computational analysis related to the Utah Genome Project. That project combines family histories from the Utah Population Database with medical records, which increasingly include DNA sequence information. The project includes family histories for more than 7 million people and medical records for around 4 million of them.

Yandell and his team are using pVAAST to analyse family pedigrees in which there is a higher frequency of disease. pVAAST searches through many genomes in parallel to find alterations. The program addresses the statistical challenge presented by genomes from people who are related, he says. And it detects *de novo* mutations.

Printed out, the large family pedigrees in the Utah database can span almost 2 metres. The ones Yandell is studying include multiple family members that have mental-health issues such as schizophrenia or depression. Mental illness has a large environmental component, but he hopes that these records can help to uncover genetic factors, he says.

Studying families might offer advantages over the more typical ‘cohort analyses’ of unrelated people with similar conditions. In such cohorts, the causes of mental-health problems might be quite diverse. Yandell hopes that restricting the search to extended families will make it easier to identify gene variants involved.

VAAST uses a similar approach to that of BLAST, a widely used search tool in genetics research<sup>6</sup>. With BLAST, a scientist can take a genetic sequence and search through many genomes to find high-probability matches to it. Similarly, VAAST compares variants in a person’s genome to those collected in the 1000 Genomes Project. This comparison helps to determine the probability that a variant is causing a disease.

pVAAST extends VAAST’s capabilities to family-based sequence data. Yandell also uses Phevor, which taps into resources such as the Human Phenotype Ontology, which catalogues links between gene function and human disease symptoms.

Phevor helped clinicians to diagnose a 12-year-old boy who had life-threatening diarrhoea and intestinal inflammation. Genetic analysis with VAAST had come up empty. By combining the analysis with Phevor, the researchers traced the boy’s illness to a *de novo* mutation in *STAT1*, a gene involved in many intestinal disorders. The finding, which was confirmed with Sanger sequencing, enabled the boy’s doctors to properly treat and stabilize his condition.

Yandell hopes that genetic analysis will soon

be a routine part of clinical diagnosis. Towards that aim, he and Martin Reese, a co-developer of VAAST, developed Opal, a platform that helps clinicians to interpret and use the results from software-based genetic analyses. Reese is chief scientific officer of Omicia, a company in Oakland, California, that offers genetic analysis using several tools, including Opal, VAAST, pVAAST and Phevor.

Reese says that his company tries to fill the gap between tools developed in academia and the needs of clinicians. The VAAST algorithm does the hard-core maths to analyse the matches, score their probabilities and create a ranking, he says. The Opal software then searches for clinical and biological data about the candidate genes — added information that can help to determine which candidates are more likely to be causing the disease.

In June, Omicia began working with Laboratory Corporation of America, a large medical-testing company based in Burlington, North Carolina. Omicia will interpret genomic data as part of clinical trials.

Data analysis is Omicia’s specialty. Unlike many other companies in the field, it does not do sequencing. “We’re slicing and dicing the genome based on your clinical question,” says Reese. His team first assesses the sequence quality and filters out typical sequencing errors before hunting for changes such as *de novo* mutations.

#### FUTURE MEDICINE

Eventually, clinical standards in this area will emerge, but for now service providers use the approaches they deem to be best for these complex analyses. Reese believes that many diseases, if not all of them, have contributions from *de novo* mutations. These contributions are hard to identify, he says, but whole-genome

analysis raises the probability of finding them, as Veltman’s study shows.

Conrad says that detection of *de novo* mutations can be a standard medical test only when the genetic complexities of diseases they cause are better understood and tool developers have found ways to address them, and after the technical issues related to high-throughput sequencing have been resolved.

Between 20% and 90% of the *de novo* mutations detected by software and with the help of whole-genome sequencing can be false positives. “Researchers can accommodate this with extensive follow-up validation experiments, but this is just simply not practical for a routine diagnostic test,” Conrad says.

Better approaches are also needed for the tough-to-sequence regions of the genome, and the software has to cover the spectrum of mutations, from single-base changes to insertions or deletions. And researchers need to better understand changes such as large copy-number variations, regions of repetitive sequence and other types of DNA rearrangements, says Conrad.

Greenwood Genetic Center, which uses genetic analysis to diagnose patients, does its own analysis and uses commercial services. Scientists and companies doing genetic analysis will soon have access to many of the same shared resources, and Friez says that he looks forward to seeing how that will help patients with neurodevelopmental disabilities. For now, patients, their families and clinicians all face the same issue: researchers’ ability to identify mutations associated with disorders is not always matched by a medical understanding of these mutations, and therapies that might arise from knowing about them are far in the future.

But genetics does deliver some answers for these patients and families, says Veltman. “From what I hear from my clinical colleagues, these families are very happy to finally get an answer — it often means closure for them, they can give the disorder in their child a place and better accept it,” he says. “In regards to therapy and treatment, unfortunately options are still quite limited, but progress is being made.” ■

**Vivien Marx** is technology editor for *Nature* and *Nature Methods*.

1. Gilissen, C. *et al.* *Nature* **511**, 344–347 (2014).
2. Ramu, A. *et al.* *Nature Methods* **10**, 985–987 (2013).
3. Yandell, M. *et al.* *Genome Res.* **21**, 1529–1542 (2011).
4. Hu, H. *et al.* *Nature Biotechnol.* **32**, 663–669 (2014).
5. Singleton, M. V. *et al.* *Am. J. Hum. Genet.* **94**, 599–610 (2014).
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215**, 403–410 (1990).

#### CASE STUDY

##### *Better diagnosis*

In a study published in July<sup>1</sup>, a team of researchers in the United States and the Netherlands analysed the genomes of 50 people who had intellectual disabilities that could not be explained through standard tests. DNA sequencing showed that each person had more than 4 million single-base variations and more than 250 copy-number variations. Computational analysis and comparison of the patients’ genomes with those of their parents whittled the number down to some 60 *de novo* mutations per person, of which 1–2 were in genes identified by the software as having a connection to intellectual disability. The genomic analysis led to a more accurate diagnosis for 20 of the 50 patients.