TECHNOLOGY FEATURE

THE DNA OF A NATION

The United Kingdom aims to sequence 100,000 human genomes by 2017. But screening them for disease-causing variants will require innovative software.

SERGEY NIVENS/SHUTTERSTOCK

BY VIVIEN MARX

Soon after the Californian twins were born, their parents grew concerned: the children were developing slowly and had floppy muscle tone. A brain scan indicated that the boy might have cerebral palsy, but doctors were puzzled over his sister's tremor and seizures. Batteries of tests failed to confirm diagnoses in either child, and treatment when the children were five with the drug L-dopa — used for people with Parkinson's disease — helped only for a while.

It was only in 2010, when the twins reached the age of 14, that whole-genome sequencing ended their diagnostic odyssey. It identified a pair of mutations in a gene that encodes the enzyme sepiapterin reductase, which is involved in production of the neurotransmitters dopamine and serotonin. Doctors modified the treatment to include serotonin; the boy's mobility improved, and the girl was no longer plagued by sudden, breath-stealing spasms[1].

Stories such as this one fuel ambitions to diagnose more quickly and accurately using genomic medicine. Indeed, tests that can probe certain disease-associated genes are increasingly becoming a diagnostic option.

But such genetic tests often fail to give a diagnosis because they are too focused on a selection of known genes on one section of the genome. In cases like that of the twins, researchers or clinicians must go further and sample a person's whole genetic sequence to find the disease-causing genes. Currently this is done only in rare cases — but a number of large-scale initiatives are poised to bring whole-genome analysis into routine medical care.

The United Kingdom has taken a giant leap into genomic medicine with the 100,000 Genomes Project, which was launched in 2012 and has been personally backed by Prime Minister David Cameron. As part of the £300-million (US$467-million) initiative, 100,000 genomes from National Health Service (NHS) patients with cancer, rare disorders and infectious diseases will be sequenced by 2017. The project's aims are to gain scientific insight by linking the disorders with precise genetic signatures; to obtain better diagnoses; to tailor treatments to individual patients; and, ultimately, to spur ▶

the development of a UK genomics industry.

The state-funded, centralized UK healthcare system is ideal for such population-based approaches in genomic medicine, says John Bell, who is a medical researcher at the University of Oxford, UK, and is also on the board of Genomics England, the NHS-owned company set up to run the project. The NHS already holds extensive clinical information on individuals, and pairing this with detailed genomic data will enable powerful insights into the links between medicine and genetics. Evidence that whole-genome interpretation can help in a wide range of disorders is mounting[2], and in the long term, Bell says, the goal is to make whole genomes part of regular NHS health records.

But before that vision can be realized, there are several hurdles that the 100,000 Genomes Project must overcome. Aside from the logistical task of extracting and sequencing DNA from thousands of individuals, there is the problem of identifying which genome variations cause disease and which are harmless — a daunting, data-heavy and time-consuming process that will require a slew of specialized companies with dedicated software.

CONSIDERABLE COHORT

Iceland was the first to launch a large-scale genomic analysis of its population. Many nations have followed suit with the explicit goal of linking health care and genomics. In the United States, the Precision Medicine Initiative plans to sequence the genomes of one million volunteers, and the Million Veteran Program is gearing up to do likewise with US military veterans. Similar projects are under way in Canada, Australia, Japan, South Korea, Singapore, Thailand, Kuwait, Qatar, Israel, Belgium, Luxembourg and Estonia.

But the 100,000 Genomes Project is the venture gaining the most steam: it has already enrolled 3,500 people with rare diseases and 2,000 individuals with cancer, and will involve roughly 75,000 people altogether (see 'The clinical genome'). People with rare diseases and their relatives will make up 50,000 of the final figure; 80% of rare diseases are inherited, so the genome of the affected person (usually a child) will be sequenced along with the genomes of two of their closest blood relatives. The remaining group of 25,000 will be composed of people with cancer, who will have their genome sequenced twice (the tumour DNA will be compared with that from a patient's normal cells), giving the grand total of 100,000 genome sequences.

The hope is that participants will benefit from clinical insights into their condition. But their genomes will also contribute knowledge of value to the entire patient community. One person's prostate-cancer genome, for example, might reveal specific genetic patterns that a physician can compare against the Genomics England database. The physician can then find other people with similar patterns and learn which drugs and procedures worked best for them.

INDUSTRY PARTNERSHIP

Genomics England is now selecting industry partners for each step in the process, from extracting the DNA to interpreting the genome. Sequencing-instrument manufacturer Illumina, which has its headquarters in San Diego, California, is handling the sequencing as well as the job of identifying genetic variants — known as variant calling. This is being done from Illumina's site in Little Chesterford, UK, but the company will be moving its sequencing instruments to the Wellcome Trust Genome Campus in Hinxton, UK, over the next few months as the project starts to scale up.

Illumina is processing extracted DNA using high-throughput sequencing to obtain short fragments of the strings of As, Ts, Cs and Gs that are the building blocks of DNA. These fragments are computationally assembled back into a contiguous sequence and then, using bioinformatics, scientists will compare the resulting complete genomes with the human reference genome: a representative example of the human genome that is continually updated by the international Genome Reference Consortium. The aim is to note each deviation from this reference: each genetic variant.

To identify these variants, the Illumina team will use the company's Isaac workflow, an open-source computational genome-alignment and variant-calling tool[3]. Then, to discover which of each genome's hundreds of thousands of variants have a role in the individual's disease, Genomics England will gather and analyse the sequences and variants from all 100,000 genomes at its secure data centre in Corsham, UK. But the NHS company has still to decide which software to use for this process: it will work with Illumina and academics to test, benchmark and improve the many existing variant-calling algorithms before selecting the final software for the initiative.

"The goal is to make whole genomes part of regular NHS health records."

Illumina has already sequenced more than 3,000 genomes for the project, and more are pouring in daily. For the variant analysis, the firm will collect and report different types of germline and somatic variants found at specific locations in DNA, says Peter Fromen, the company's managing director of population-sequencing initiatives. Variants can include insertions or deletions of a few nucleotides, or the substitution of one nucleotide for another. There can also be structural variants, such as changes in the copy number of a gene.

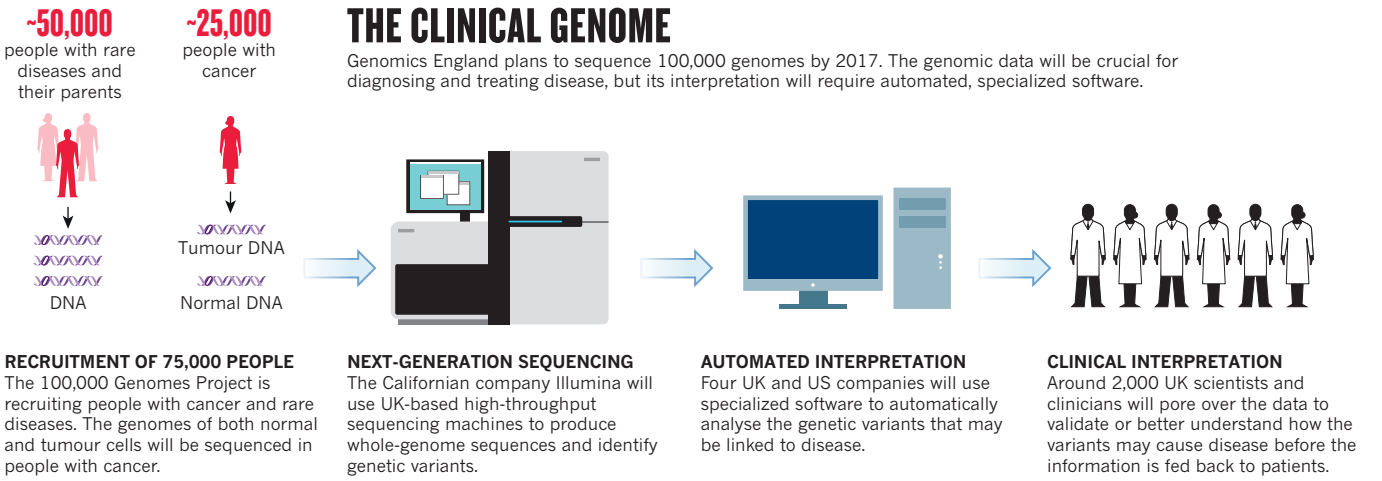
Each variant is then checked against those in existing variant databases, such as dbSNP, a short-genetic-variation database curated at the US National Institutes of Health (NIH); the 1000 Genomes Project, an international research initiative that has compiled a catalogue of human genetic variation across the world; the Exome Aggregation Consortium (ExAC) database, a collection of exome sequencing data (the exome is all DNA that codes for proteins); and ClinVar, an NIH database of variants and associated physical conditions. Making sense of the variants is the next stage, and companies have been vigorously bidding for the job. In spring 2014, Genomics England launched a 'bake off' to gauge the quality of genome interpretation expertise from around the world. Twenty-eight participating companies were asked to provide genetic-variant annotation (describing an individual gene and its protein product) and interpretation (assigning them with a function) for the genomes of 15 people with rare diseases compared with 2 samples from healthy relatives, and for 10 genome pairs of tumour DNA and undiseased DNA taken from the same individual.

The ten best-performing companies then took part in a tender to provide annotation and interpretation services for the first 8,000 patients in the project. Genomics England has now narrowed down the pool to four bidders, subject to them passing a test phase and agreeing a contract. Congenica of Cambridge, UK, and Omicia in Oakland, California, will analyse rare-disease genomes; Nanthealth in Culver City, California, will analyse cancer genomes; and WuXi NextCODE in Cambridge, Massachusetts, will analyse both cancer and rare diseases. Cypher Genomics in San Diego, California, and its partner, the advanced technology and defence firm Lockheed Martin in Bethesda, Maryland, are reserve bidders. Most have previous experience in the disease areas for which they were chosen.

All companies will use high-performance computing to interpret genomic data and will work in Genomics England's secure data centre, where the analysis will take place. Their aim is to provide a mostly automated service — interpreting next-generation sequencing data still remains a partly manual process that can take from a couple of hours to weeks. At its peak, says Augusto Rendon, who directs bioinformatics for Genomics England, the project will be receiving up to 200 genomes a day for processing. A manual process will not do if the initiative is to come in on time and on budget — and, ultimately, if whole-genome sequencing is to serve as a widespread diagnosis tool.

THE CHOSEN FEW

Each company brings its own, disparate expertise and history to tackling the thorny interpretation task. Congenica, a spin-off of the Wellcome Trust Sanger Institute and the UK Department of Health, already does testing for NHS laboratories. It will report on inherited and acquired rare-gene mutations associated with disease. The company's Sapientia



platform has already been put to the test by analysing DNA from 12,000 children taking part in the Deciphering Developmental Disorders study, the world's largest nationwide rare-disease-sequencing programme[4]. This genome-wide study returned a diagnosis to 30–40% of participating families whose children presented with undiagnosed birth defects or learning disabilities, says Tom Weaver, chief executive of Congenica.

Omicia in the United States also has a track record of providing clinical interpretation, but using open-source, open-access tools. For the 100,000 Genomes Project, it will use gene-interpretation software called Opal to predict which variants are likely to be causing disease. The company's Phevor (phenotype driven variant ontological re-ranking tool)[5] algorithm can also take into account the patient's phenotype — the physical manifestation of disease. These highly automated tools mean that the firm can avoid having to manually compile information from a bundle of sources, says Martin Reese, the company's chief scientific officer. The algorithms statistically 'triage' variants into those that are known to cause disease, those for which disease links exist and those that need further investigation. Opal then pulls together all of the results into a report with which a physician can make treatment decisions.

Cypher Genomics, which is a spin-off of the Scripps Research Institute in La Jolla, California, has developed interpretation software called Mantis that ranks variants according to how likely they are to cause disease. Scientists can then follow up on the tougher cases, which can include manual analysis, says Adam Simpson, the company's chief operating officer.

The only finalist with population-wide genomics know-how under its belt is WuXi NextCODE. The firm is a spin-off of the Icelandic genetics company deCODE, which was founded in 1996 by Kári Stefánsson, who pioneered the idea of linking genealogical and genetic data from the entire Icelandic population to identify human genes associated with common diseases. When deCODE was acquired by the California-based biopharmaceutical firm Amgen, a handful of informaticians and deCODE alumni formed NextCODE, which in turn was acquired by biopharma firm WuXi PharmaTech of Shanghai, China, in January 2015.

WuXi NextCODE has created a way to provide information beyond a distilled list of possible variants. The company has built a way to handle large data sets, such that scientists can zoom in and out of a person's entire genome, and the software continuously pulls out the most up-to-date information about a variant, says Jeff Gulcher, WuXi NextCODE's chief scientific officer.

A typical question might be "I want to know if this is a real mutation or not", says Gulcher. That can mean hunting for other individuals with a rare disease and the same genomic variants. And, one day, a physician might want to find people with cancer who have similar mutations and disease course and who were treated with comparable drugs in the past ten years. The company's Genomically Ordered Relational Database is optimized for such hunts, he says.

In another case, a geneticist might want to compare 20 genomes, each with one million variants, to find out what ails a patient. Rather than compare the genomes in their entirety, WuXi NextCODE's database architecture arranges variants according to genomic position, pulling in a slice of information at a time, making the process computationally efficient. The platform has also been trained on large data sets, such as the genomic information of 300,000 Icelanders, says the company's chief operating officer, Hannes Smarason.

Nanthealth is a privately owned health-care company that is focused on using computational analysis of cancer genomic information to guide diagnosis and therapy. The firm, which did not respond to requests for information, says on its website that it has analysed more than 20,000 genomes. It was founded by physician and biomedical researcher Patrick Soon-Shiong, who also chairs a foundation that funds research and aims to erase disparities in health-care access and heads a non-profit research organization whose goal is to facilitate digital molecular diagnosis. He developed the cancer drug Abraxane (protein-bound paclitaxel), which is used to treat many types of cancers.

For Genomics England, the entire project is about providing clinical answers. But before the diagnostic readouts can reach physicians and their patients, teams of scientists and clinicians will need to pore over the data. Genomics England has organized 2,000 UK scientists with expertise in 13 rare diseases and 10 cancer types to quality check and study project results, especially where a connection to disease has not yet been well established. Using cell-biology assays and mouse models, they will explore how variants may cause or contribute to disease, knowledge that will feed back into the 100,000 Genomes Project. Some results may be quick to validate in databases, whereas others will require careful scrutiny with deeper analysis using the literature, software tools and laboratory assays.

The NHS is not known as a technological innovator, says Bell. But Genomics England will probably be transformative, he says. The 100,000 Genomes Project has generated a wave of activity that could be a powerful boost for the whole field of genomics. It will engender many commercial and academic opportunities — perhaps enabling genomic medicine to finally fulfil its promise of delivering widespread benefit to people with disease. ■

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

1. Bainbridge, M. N. *et al. Sci. Transl. Med.* **3,** 87re3 (2011).
2. Taylor, J. C. *et al. Nature Genet.* **47,** 717–726 (2015).
3. Raczy, C. *et al. Bioinformatics* **29,** 2041–2043 (2013).
4. Wright, C. F. *et al. Lancet* **385,** 1305–1314 (2015).
5. Singleton, M. V. *et al. Am. J. Hum. Genet.* **94,** 599–610 (2014).