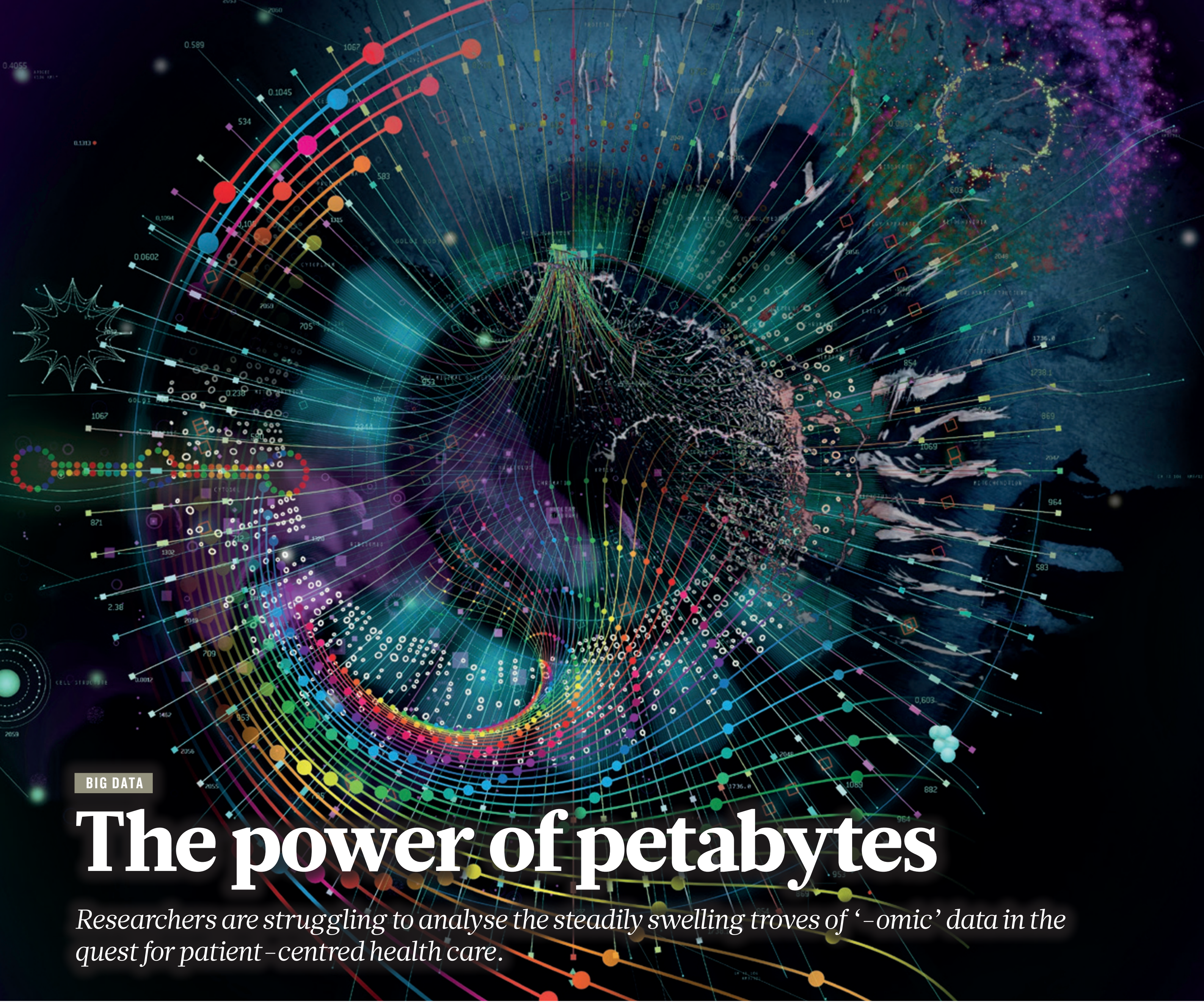


BIG DATA

The power of petabytes

Researchers are struggling to analyse the steadily swelling troves of '-omic' data in the quest for patient-centred health care.

TATIANA PLAKHOVA

BY MICHAEL EISENSTEIN

Fifteen years ago, it was a landmark achievement. Ten years ago, it was an intriguing but highly expensive research tool. Now, falling costs, soaring accuracy and a steadily expanding base of scientific knowledge have brought genome sequencing to the cusp of routine clinical care.

A growing number of institutions are conducting genome-wide 'dragnet' searches to identify the mutations responsible for rare diseases. "The rate at which we're finding causative variants in those cases is going up," says Russ Altman, a bioinformatician at Stanford School of Medicine in California. "At some centres, it's up to 50% of cases." Genomic variants can also reveal 'driver' mutations that might reveal a tumour's therapeutic vulnerabilities, or provide clues to whether a specific individual may or may not respond to a drug — the drug's 'pharmacogenetic' properties.

The US$1,000 genome, initially conceived as a price point at which sequencing could become a component of personalized medicine, has arrived. "Our capacity for data generation relative to price has increased in a way that is almost unprecedented in science — roughly six orders of magnitude in the past seven or eight years," says Paul Flicek, a specialist in computational genomics at the European Molecular Biology Laboratory's European Bioinformatics Institute in Cambridge, UK. The HiSeq X Ten system developed by Illumina of San Diego, California, can sequence more than 18,000 human genomes per year, for example.

The biomedical research community is diving in whole-heartedly, with population-scale programmes that are intended to explore the clinical power of the genome. In 2014 the United Kingdom launched the 100,000 Genomes Project, and both the United States (under the Precision Medicine Initiative) and China (in a programme to be run by BGI of Shenzhen) have unveiled plans to analyse genomic data from one million individuals.

Many other programmes are under way that, although more regional in focus, are still 'big data' operations. A partnership between Geisinger Health System, based in Danville, Pennsylvania, and biotech firm Regeneron Pharmaceuticals of Tarrytown, New York, for instance, aims to generate sequence data for more than 250,000 people. Meanwhile, a growing number of hospitals and service providers worldwide are sequencing the genomes of people with cancers or rare hereditary disorders (see 'DNA sequencing soars').

Some researchers worry that the flood of data could overwhelm the computational pipelines needed for analysis and generate unprecedented demand for storage — one article estimated that the output from genomics may soon dwarf data heavyweights such as YouTube. Many also worry that today's big data lacks the richness to provide clinical value. "I don't know if a million genomes is the right number, but clearly we need more than we've got," says Marc Williams, director of the Geisinger Genomic Medicine Institute.

THE MEANING OF MUTATIONS

Clinical genomics today is largely focused on identifying single-nucleotide variants — individual 'typos' in the genomic code that can disrupt gene function. And rather than looking at the full genome, many centres focus instead on the exome — the subset of sequences containing protein-coding genes. This reduces the amount of data being analysed nearly 100-fold, but the average exome still contains more than 13,000 single-nucleotide variants. Roughly 2% of these are predicted to affect the composition of the resulting protein, and finding the culprit for a given disease is a daunting challenge.

For decades, biomedical researchers have dutifully deposited their discoveries of single-nucleotide variants in public resources such as the Human Gene Mutation Database, run by the Institute of Medical Genetics at Cardiff University, UK, or dbSNP, maintained by the US National Center for Biotechnology Information. However, the effects of these mutations were often determined from cell culture or animal models, or even theoretical predictions, providing insufficient guidance for clinical diagnostic tools. "In many cases, associations were made with relatively low levels of evidence," says Williams.

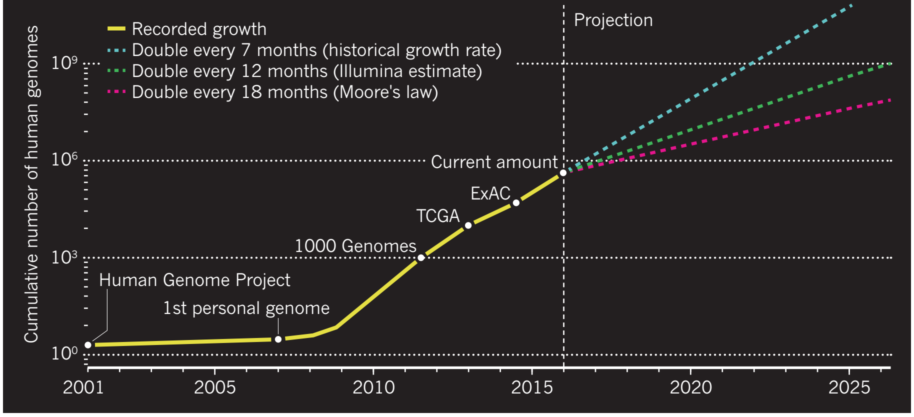
The situation is even more complicated for structural variants, such as duplicated or missing chunks of genome sequence, which are far more difficult to detect with existing sequencing technologies than single-nucleotide variants. At the whole-genome scale, each person has millions of variants. Many of these are in sequences that do not encode proteins but instead regulate gene activity, so they can still contribute to disease. However, the extent and function of these regulatory regions are poorly defined. Although capturing all this variability is desirable, it may not offer the best short-term returns for clinical sequencing. "You're shooting yourself in the foot if you're collecting data you don't know how to interpret," says Altman.

Efforts are now under way to rectify this problem. The Clinical Genome Resource, which was set up by the US National Human Genome Research Institute, is a database of disease-related variants, and contains information that could guide medical responses to these variants as well as the evidence supporting those associations. Genomics England, which runs the 100,000 Genomes Project, aims to bolster progress in this area by establishing 'clinical interpretation partnerships': doctors and researchers will collaborate to establish robust models of diseases that can potentially be mapped to specific genetic alterations.

"You're shooting yourself in the foot if you're collecting data you don't know how to interpret."

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



STEPHENS, Z. D. *ET AL. PLOS BIOL.* **13,** E1002195 (2015)/CC BY 4.0 HTTP://CREATIVECOMMONS.ORG/LICENSES/BY/4.0

However, quantity is as important as quality. Mutations that offer a strong detrimental effect bring an evolutionary disadvantage, so they tend to be exceedingly rare and require large sample sizes to detect. Establishing statistically meaningful disease associations for variants with weak effects also needs large numbers of people.

In Iceland, deCODE Genetics has demonstrated the power of population-scale genomics, combining extensive genealogy and medical-history records with genome data from 150,000 people (including 15,000 whole-genome sequences). These findings have allowed deCODE to extrapolate the population-wide distribution of known genetic risk factors, including gene variants linked to breast cancer, diabetes and Alzheimer's disease.

They have also enabled studies in humans that normally require the creation of genetically modified animals. "We have established that there are about 10,000 Icelanders who have loss-of-function mutations in both copies of about 1,500 different genes," says Kári Stefánsson, the company's chief executive. "We're putting significant effort into figuring out what impact the knockout of these genes has on individuals."

This work was helped by the homogeneous nature of the Icelandic population, but other projects require a broadly representative spectrum of donors. Efforts such as the international 1000 Genomes Project have catalogued some of the world's genetic diversity, but most data are heavily skewed towards Caucasian populations, making them less useful for clinical discovery. "Because they come from the genetic mother ship, so to speak, people of African ancestry carry a lot more genetic variants than non-Africans," says Isaac Kohane, a bioinformatician at Harvard Medical School in Boston, Massachusetts. "Variants that seem unusual in Caucasians might be common in Africans, and may not actually cause disease."

Part of the problem stems from the reference genome — the yardstick sequence by which scientists identify apparent abnormalities, developed by the multinational Genome Reference Consortium. The first version was cobbled together from a few random donors of undefined ethnicity, but the latest iteration, known as GRCh38, incorporates more information about human genomic diversity.

INTO THE CLOUD

Harvesting genomes or even exomes at the population scale produces a vast amount of data, perhaps up to 40 petabytes (40 million gigabytes) each year. Nevertheless, raw storage is not the primary computational concern. "Genomicists are a tiny fraction of the people who need bigger hard drives," says Flicek. "I don't think storage is a significant problem."

A greater concern is the amount of variant data being analysed from each individual. "The computation scales linearly with respect to the number of people," says Marylyn Ritchie, a genomics researcher at Pennsylvania State University in State College. "But as you add more variables, it becomes exponential as you start to look at different combinations." This becomes particularly problematic if there are additional data related to clinical symptoms or gene expression. Processing data of this magnitude from thousands of people can paralyse tools for statistical analysis that might work adequately in a small laboratory study.

Scaling up requires improvisation, but there is no need to start from scratch. "Fields like meteorology, finance and astronomy have been integrating different types of data for a long time," says Ritchie. "I've been to meetings where I talk to people from Google and Facebook, and our 'big data' is nothing like their big data. We should talk to them, figure out how they've done it and adopt it into our field."

Unfortunately, many talented programmers with the skills to wrangle big data sets are lured away by Silicon Valley. Philip Bourne, associate director for data science at the US

National Institutes of Health (NIH), believes that this is partly due to a lack of recognition and advancement within a publication-driven system of scientific credit that leaves software creators and data managers out in the cold. "Some of these people truly want to be scholars, but they can't get the stature of faculty — that's just not right," says Bourne.

Processing power is another limiting factor. "This is not a desktop game — the real practitioners are proficient in massively parallel computation with hundreds if not thousands of CPUs, each with large memory," says Kohane. Many groups that analyse massive amounts of sequence data are moving to 'cloud'-based architectures, in which the data are deposited within a large pool of computational resources and can then be analysed with whatever processing power is required.

"There's been a gradual evolution towards this idea that you bring your algorithms to the data," says Tim Hubbard, head of bioinformatics at Genomics England. For Genomics England, this architecture is contained in a secure government facility, with strict control over external access. Other research groups are turning to commercial cloud systems, such as those provided by Amazon or Google.

PRIVACY PROTECTION

In principle, cloud-based hosting can encourage sharing and collaboration on data sets. But regulations on patient consent and privacy rights surrounding highly sensitive clinical information pose tricky ethical and legal issues.

In the European Union, collaboration is impeded by member states having different rules on data handling. Sharing with non-EU nations relies on cumbersome mechanisms to establish adequacy of data protection, or restrictive bilateral agreements with individual organizations. To help solve this problem, a multinational coalition, the Global Alliance for Genomics and Health, developed the Framework for Responsible Sharing of Genomic and Health-Related Data. The Framework includes guidelines on privacy and consent, as well as on accountability and legal consequences for those who break the rules.

"In data-transfer agreements, you could save yourself pages and pages of rules if the institution, researcher and funder agree to follow the Framework," says Bartha Knoppers, a bioethicist at McGill University in Montreal, Canada, who chairs the Alliance's regulatory and ethics working group. The Framework also calls for 'safe havens' that allow the research community to analyse centralized banks of genomic data that have been identity-masked but not fully 'de-identified', so they remain useful. "We want to link it to clinical data and to medical records, because we're never going to get to precision medicine otherwise, so we're going to have to use coded data," explains Knoppers.

Integrating genomics into electronic health records is becoming increasingly important for many European nations. "Our objective is to put this into the standard National Health Service," says Hubbard. The UK 100,000 Genomes Project may be the furthest along at the moment, but other countries are following. Belgium recently announced an initiative to explore medical genomics, for example.

LEN RUBENSTEIN/THE BROAD INSTITUTE

Rapid advances in technology are transforming genomics research.

All these nations benefit from having centralized, government-run health-care systems. In the United States, the situation is more fragmented, with different providers relying on distinct health-record systems, supplied by different vendors, that are generally not designed to handle complex genomic data. The NIH launched the Electronic Medical Records and Genomics (eMERGE) Network in 2007 to define best practices.

FROM DATA TO DIAGNOSIS

The immediate goal of genomically enriched health records is to explain the implications of gene variants to physicians, and one of its earliest implementations is pharmacogenetics. The Clinical Pharmacogenetics Implementation Consortium has translated known drug–gene interactions reported in PharmGKB (a database run by Altman and his colleagues) for clinical use. For example, people with certain variants may respond poorly to particular anticoagulants, leading to increased risk of heart attack. "The issue there is, how do you take a practitioner who has 12 minutes per patient and about 45 seconds of time allocated for prescribing drugs, and influence their practice in a meaningful way?" says Altman.

As long as deciding how to adapt care to genetic findings remains a job for humans, this process will remain time- and labour-intensive. Nevertheless, combining genotype and phenotype information is proving fruitful from a research perspective. Most clinically relevant gene variants were identified through genome-wide association studies, in which large populations of people with a given disease were examined to identify closely associated genetic signatures. Researchers can now work backwards from health records to determine what clinical manifestations are prevalent among individuals with a given genetic variant.

And the genome is only part of the story — other '-omes' may also be useful barometers of health. In July, Jun Wang stepped down as chief executive of BGI to start up an organization to analyse BGI's planned million-genome cohort alongside equivalent data sets from the proteome, transcriptome and metabolome. "I will be initiating a new institution to focus on using artificial intelligence to explore this kind of big data," he says.

IT TAKES PATIENTS

As researchers strive to integrate data from health records and clinical trials with genomic and other physiological data, patients are starting to contribute. "When we're focused on things like behaviour, nutrition, exercise, smoking and alcohol, you can't get better data than what patients report," says Ritchie.

Wearable devices, such as smartphones and FitBits, are collecting data on exercise and heart rate, and the volume of such data is soaring (see 'page S12) as it can be gathered with minimal effort on the wearer's part.

Each patient may become a big-data producer. "The data we generate at home or in the wild will vastly exceed what we accumulate in clinical care," says Kohane. "We're trying to create these big collages of different data modalities — from the genomic to the environmental to the clinical — and link them back to the patient." As these developments materialize, they could create computational crunches that will make today's 'big data' struggles seem like pocket-calculator problems. And as scientists find ways to crunch the data, patients will be the ultimate winners. ■

Michael Eisenstein *is a freelance science writer based in Philadelphia, Pennsylvania.*