HOW A FIELD BUILT ON DATA SHARING BECAME A TOWER OF BABEL

The immediate and open exchange of information was key to the success of the Human Genome Project 20 years ago. Now the field is struggling to keep its data accessible. **By Kendall Powell**

In July 2000, David Haussler remembers crying as he watched the first fully assembled human genome streaming across his computer screen. He and Jim Kent, a graduate student at the time, built the first-ever web-based tool for exploring the three billion letters of the human genome. They had published the rough draft of the genome on the Internet a mere 11 days after finishing the herculean task of stitching it all together — a task assigned to them as part of the Human Genome Project (HGP), the international collaboration that had been working towards this goal for a decade. It would still be several months before the group published its analysis of the genome in the pages of *Nature*[1], but the data were ready to share.

"There it was, going out into the whole world," recalls Haussler, scientific director of the University of California Santa Cruz Genomics Institute. Soon, every person in the world could explore it — chromosome by chromosome, gene by gene, base by base — on the web.

It was a historic moment, says Haussler. Before the HGP launched in the early 1990s, "there had not been a serious discussion about data sharing in biomedical research", Haussler says. "The standard was that a successful investigator held onto their own data as long as they could."

That standard clearly wouldn't work for such a large and collaborative effort. If countries or scientists hoarded the data they were producing, it would derail the project. So in 1996, the HGP researchers got together to lay out what became known as the Bermuda Principles, with all parties agreeing to make the human genome sequences available in public databases, ideally within 24 hours — no delays, no exceptions.

Fast-forward two decades, and the field is bursting with genomic data, thanks to improved technology both for sequencing whole genomes and for genotyping them by sequencing a few million select spots to quickly capture the variation within. These efforts have produced genetic readouts for tens of millions of individuals, and they sit in data repositories around the globe. The principles laid out during the HGP, and later adopted by journals and funding agencies, meant that anyone should be able to access the data created for published genome studies and use them to power new discoveries.

If only it were that simple.

The explosion of data led governments, funding agencies, research institutes and private research consortia to develop their own custom-built databases for handling the complex and sometimes sensitive data sets. And the patchwork of repositories, with various rules for access and no standard data formatting, has led to a "Tower of Babel" situation, says Haussler.

Although some researchers are reluctant to share genome data, the field is generally viewed as generous compared with other disciplines. Still, the repositories meant to foster sharing often present barriers to those uploading and downloading data. Researchers tell tales of spending months or years tracking down data sets, only to find dead ends or unusable files. And journal editors and funding agencies struggle to monitor whether scientists are sticking to their agreements.

Many scientists are pushing for change, but it can't come fast enough.

Clinical genomicist Heidi Rehm says the field has come to recognize that big scientific advances require vast amounts of genomic data linked to disease and health-trait data. "But it isn't compatible and shareable," says Rehm, based at Massachusetts General Hospital in Boston and the Broad Institute in Cambridge. "How do we get everyone in the world — patients, clinicians and researchers — to share?"

Barriers everywhere

Sequencing the human genome made it easier to study diseases associated with mutations in a single gene — Mendelian disorders such as non-syndromic hearing loss[2] (see page 218). But identifying the genetic roots of more common complex diseases, including cardiovascular disease, cancer and other leading causes of death, required the identification of multiple genetic risk factors throughout the genome. To do this, researchers in the mid-2000s began comparing the genotypes of thousands to hundreds of thousands of individuals with and without a specific disease or condition, in an approach known as genome-wide association studies, or GWAS.

The approach proved popular — more than 10,700 GWAS have been conducted since 2005. And that has produced oceans of data, says Chiea Chuen Khor, a group leader at the Genome Institute of Singapore, who studies the genetic basis of glaucoma. A study with 10,000 people, looking at 1 million genetic markers in each, for example, says Khor, would generate a spreadsheet with 10 billion entries.

Most of these individual-level genomic data now live in 'controlled-access' databases. These were set up to deal with the sticky legal and ethical concerns that come with genomic data that have been linked to personal information — 'phenotype data' that can include health-care records, disease status or lifestyle choices. Even in anonymized data sets, it's technically possible that individuals can be reidentified. So, controlled-access databases vet the researchers seeking access and ensure that the data are used only for the purposes that participants consented to.

The US National Institutes of Health (NIH) requires its grant recipients to place GWAS data into its official repository, the Database for Genotypes and Phenotypes, or dbGaP.

ILLUSTRATION BY ANA KOVA

European researchers can deposit data into the European Genome-phenome Archive (EGA) housed at the European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK. Similarly, other large generators of genomic data, such as the for-profit company 23andMe in Sunnyvale, California, and the non-profit Genomics England in London, operate their own controlled-access databases.

But uploading data into some of these repositories often takes a long time. As a result, says Khor, the data are often "minimal and sparse", because researchers are depositing just what's required to be compliant.

Sometimes the data get stored in more than one place, and that creates other challenges. Rasika Mathias, a genetic epidemiologist at Johns Hopkins University in Baltimore, Maryland, who studies the genetics of asthma in people of African ancestry, says that decentralization is a huge problem. She is part of TOPMed, a precision-medicine programme run by the NIH's National Heart, Lung, and Blood Institute. It consists of more than 155,000 research participants across more than 80 studies and shares its data in several repositories, including dbGaP and some university-based portals.

"It's a remarkable resource," says Mathias. But it's cumbersome for an outsider to find all the pieces of available data and request access, she says. They must often provide detailed proposals and letters of support. "It's unnecessarily difficult."

Many look for workarounds. "I personally do not download dbGaP data, I just go straight to the researchers and ask if they want to collaborate," says Ruth Loos, a genetic epidemiologist at the Icahn School of Medicine at Mount Sinai in New York City. Several years ago, she tried to access a dbGaP data set, filing multiple rounds of digital paperwork, only to be rejected. "Even logging into dbGaP can be a pain. It's just not researcher-friendly," she says.

Stephen Sherry, acting director of the NIH's National Center for Biotechnology Information in Bethesda, Maryland, which runs the dbGaP, acknowledges that the processes to submit and access data are "imperfect and painful". And the complex, heterogeneous data require case-by-case review, which cannot simply be sped up by throwing "more people at the crank to turn it faster".

But, Sherry says, the NIH is investing in modernizing the system to make it more streamlined and flexible. Carrie Wolinetz, associate director for science policy at the NIH, says it is yet to be determined whether the remedy will be a dbGaP 2.0 or an alternative resource. "Do you put in a stop-gap measure, or is it time to invest in a whole bathroom renovation?" she asks.

For all the problems that controlled access causes in sharing genome data, many researchers say databases such as dbGaP and the UK BioBank, which holds genomic data on 500,000 people, are still invaluable. Mathias is fiercely protective of the participants in TOPMed and sees merit in the protection that controlled access provides. Like many, she would like to see the repositories better resourced. But, she says, "I am an advocate for the checks and balances".

And others are happy to have access, even if it is hard to obtain. "It's out of our scope to generate that amount of data," says Melanie Bahlo, who runs a statistical-genetics lab at the Walter and Eliza Hall Institute of Medical Research in Melbourne, Australia. Her lab is more than willing to wade through the digital paperwork to use the dbGaP, and has done so for more than ten projects. She also recently spent a fruitless six months chasing after a data set that was supposed to be publicly available through a research institute's data portal, but wasn't.

"Nothing is harder than getting data out of dbGaP and EGA," says Khor, "unless it's getting it from a researcher who is unwilling to share."

The sharing police

Twenty years on from the HGP, there is no specific universal policy that says research groups have to share their human-genome data, or share them in a particular format or database. That said, many journals have continued to abide by the Bermuda Principles, requiring that genomic data be shared in approved databases at the time of publication. Enforcement of these policies can be hit or miss.

> "IF YOU DON'T HAVE THE RAW DATA, YOU CAN'T LOOK AT THE QUALITY."

Michelle Trenkmann, senior editor for genetics and genomics at *Nature* in London, says that authors are often reluctant to share, citing concerns over participant privacy, consent or national or corporate rules governing who owns the data. "What's remarkable, is that, as a field, geneticists expect the data to be shared, but sometimes they do not want to share their own data," she says. Trenkmann pushes back in such cases, and if the challenges can't be overcome, the authors must spell out their reasons directly in the paper for transparency. (*Nature*'s news team is editorially independent from its journal team.)

The journal *Genome Research*, has a 'no exceptions' policy. Executive editor Hillary Sussman explains that the journal's editors will work through data-sharing obstacles with authors on a case-by-case basis to find solutions. This can go as far as asking authors to reapply for approval from their institutional review board, going back to participants to reobtain their consent or rerunning an analysis after removing unshareable data. The journal has turned away authors who state upfront that they cannot share data. "The community and the funders demand this transparency and reproducibility," she says.

But even when authors do agree to share data, editors and reviewers have limited ability to confirm that it is being done. They might not have the time – or the access to controlled-access databases – to check data quality, formatting or completeness.

Trenkmann says funders should require researchers to have a concrete data-sharing plan from the outset of a project. This could help to shift attitudes so that researchers see sharing as a duty, she says.

An NIH-wide data-sharing policy to be implemented in January 2023 does just that. It requires all NIH grant applicants to put a Data Management and Sharing (DMS) Plan into their grant proposals and allows researchers to allocate some of their budget to the task.

This should ensure that data sharing is aligned both with ethical and privacy considerations, and with the FAIR principles – which mean that data must be findable, accessible, interoperable and reusable, says Carolyn Hutter, director of the National Human Genome Research Institute's (NHGRI) Division of Genome Sciences in Bethesda. "That does not mean, I threw my data over a wall and hope someone caught it," she says.

"The enforcement part of it is tricky," Hutter adds, "because data sharing often comes at the very end of the project." And like journal editors, grant administrators can only do spot checks of any data-sharing accession numbers that appear in annual progress reports.

Searching for solutions

There could be ways to share more simply without falling foul of proprietary or privacy issues. Many genomic stakeholders agree that an aggregated form of GWAS data, called GWAS summary statistics, can and should be shared broadly and freely. These summaries include the aggregated scores for each genetic variant found to be associated with a disease or condition across multiple genomes. They are easier for researchers to work with, and they protect participant privacy.

Many research consortia do share these on their websites or portals. But an open-access collaboration between EMBL-EBI and the NHGRI, called the GWAS Catalog[3], is working towards a centralized, standardized solution.

Starting in 2020, the GWAS Catalog gave researchers a way to submit their summary

REGENTS OF UNIV. CALIFORNIA. COURTESY OF SPECIAL COLLECTIONS, UNIV. LIBRARY, UNIV. CALIFORNIA, SANTA CRUZ. UC SANTA CRUZ PHOTOGRAPHY SERVICES PHOTOGRAPHS.

In 2000, Jim Kent, a graduate student at the University of California, Santa Cruz, helped to assemble and share the results of the decade-long Human Genome Project.

statistics along with metadata describing the study and participants. In return, researchers get a prepublication accession ID to use in preprints and submitted manuscripts.

But many researchers say that summary statistics are not sufficient for advancing genomic science. "That's a major threat to GWAS," says Chris Amos, a genetic epidemiologist who studies lung cancer at Baylor College of Medicine in Houston, Texas. Researchers need the individual-level genome data and the linked phenotypic trait data to reveal exactly how genetic variation plays out in disease. They also need the full data to check the science. "If you don't have the raw data, you can't look at the quality. That is not good enough to make a reproducible finding," Amos says.

And the owners of the data for very large cohorts, such as 23andMe and Genomics England, don't give unrestricted access to their summary statistics. They cite concerns over participants' data privacy and the wish to retain ownership of their data. In effect, they run their own controlled-access databases, with custom processes for accessing and reanalysing their data. A precondition for working with much of their data is allowing the companies to share authorship of the resulting work. Bahlo says these kinds of requirements set too high a bar for her and other bioinformaticians who wish to crunch data from Genomics England's 100,000 Genomes Project.

Hutter acknowledges that not all the current growing pains of genomic data sharing can be fixed simply through improvements to the dbGaP or by sharing summary statistics in the GWAS Catalog. "The dbGaP wasn't positioned to evolve and handle every new type of data," she says. For example, the cost of storing data from whole genomes is very different from that for GWAS data. As such, the NHGRI has created a cloud-based infrastructure known as the Analysis, Visualization, and Informatics Lab-space (AnVIL), where researchers can share and analyse across large genomic data sets, including whole genome and exome sequences.

Another NIH initiative is the Researcher Auth Service (RAS), which would authorize researchers to access AnVIL, the dbGaP and several other data resources. "The vision is that we'd push this out like a visa stamp," says Sherry, allowing researchers to ultimately merge and analyse data at will in cloud-based systems. "We're building one of the first systems of library cards for researchers," says Sherry.

Haussler and some other big-data wranglers also have ideas. As data-sharing frustrations were mounting in 2013, Haussler, along with David Altshuler, Eric Lander and other international colleagues laid the groundwork for the Global Alliance for Genomics and Health, or GA4GH (see go.nature.com/3app3xr). It started with the same ideals as the HGP. "We'd get the world to share data on one big database, and we'd all agree on how we'd use that data, and Kumbaya," says Haussler. "Very quickly, it became evident that that was utterly impossible."

Instead, the GA4GH now focuses on creating standards for the multitude of genomic databases around the world. Its working hypothesis is that it will be technically possible to harmonize data (like the GWAS Catalog on a grander scale) and to federate, or loosely link, the disparate data warehouses.

GA4GH chief executive Peter Goodhand uses the analogy of global mobile-phone communications. There's huge competition between mobile-phone makers and service providers, but at the end of the day, they all have to work on the same network. "For true interoperability to take place, there have to be working relationships between the providers," says Goodhand. "You can set up the systems that permit the sharing and make it easier."

Scientists used a GA4GH standard to create the Matchmaker Exchange, for example. This service lets clinicians and researchers working on the rarest of rare diseases search a single federated network of eight international databases to find individuals with a similar genotype or phenotype to a case they're working on. If a match is returned, both parties are connected in a way that protects both patient confidentiality and research ownership and authorship. The NIH's RAS will also use a GA4GH standard, called the Data Repository Service, a software interface that helps different repositories to communicate.

Bahlo and others say that data federation efforts become even more important as the field pivots to digging deeper into phenotype data, which have grown in scope and complexity. "That data comes in all sorts of forms — environmental exposures, smoking status, medical imaging data," says Bahlo.

She and others see data federation as a great opportunity to inject global equity into genomic data sharing. Researchers from developing countries could access and work with data sets without needing to generate their own data or have their own supercomputing resources. And better data sharing should also improve representation of non-white, non-European global ancestries. Under-representation is especially stark for continental African ancestries, which make up less than 0.5% of all GWAS participants[4] (see pages 209 and 220).

Haussler thinks that positive peer pressure should convince scientists to share in better ways. The need is only growing. Twenty years after releasing the first human genome to the Internet, his team has built a way for anyone to explore the SARS-CoV-2 viral genome[5].

"Data should be a living thing," says Haussler. "I want to click on it and play with it immediately. That should be the motivation. If you don't share your data, you can't do that."

Kendall Powell is a freelance science journalist in Lafayette, Colorado.

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Chong, J. X. *et al. Am. J. Hum. Genet.* **97**, 199–215 (2015).
3. Buniello, A. *et al. Nucl. Acids Res.* **47**, D1005–D1012 (2019).
4. Mills, M. C. & Rahal, C. *Commun. Biol.* **2**, 9 (2019).
5. Fernandes, J. D. *et al. Nature Genet.* **52**, 991–998 (2020).