Internet of DNA

A global network of millions of genomes could be medicine’s next great advance. Availability: 1-2 years

Noah is a six-year-old suffering from a disorder without a name. This year, his physicians will begin sending his genetic information across the Internet to see if there’s anyone, anywhere, in the world like him.

A match could make a difference. Noah is developmentally delayed, uses a walker, speaks only a few words. And he’s getting sicker. MRIs show that his cerebellum is shrinking. His DNA was analyzed by medical geneticists at the Children’s Hospital of Eastern Ontario. Somewhere in the millions of As, Gs, Cs, and Ts is a misspelling, and maybe the clue to a treatment. But unless they find a second child with the same symptoms, and a similar DNA error, his doctors can’t zero in on which mistake in Noah’s genes is the crucial one.

In January, programmers in Toronto began testing a system for trading genetic information with other hospitals. These facilities, in locations including Miami, Baltimore, and Cambridge, U.K., also treat children with so-called ­Mendelian disorders, which are caused by a rare mutation in a single gene. The system, called MatchMaker Exchange, represents something new: a way to automate the comparison of DNA from sick people around the world.

One of the people behind this project is David Haussler, a bioinformatics expert based at the University of California, Santa Cruz. The problem Haussler is grappling with now is that genome sequencing is largely detached from our greatest tool for sharing information: the Internet. That’s unfortunate because more than 200,000 people have already had their genomes sequenced, a number certain to rise into the millions in years ahead. The next era of medicine depends on large-scale comparisons of these genomes, a task for which he thinks scientists are poorly prepared. “I can use my credit card anywhere in the world, but biomedical data just isn’t on the Internet,” he says. “It’s all incomplete and locked down.” Genomes often get moved around in hard drives and delivered by FedEx trucks.

Haussler is a founder and one of the technical leaders of the Global Alliance for Genomics and Health, a nonprofit organization formed in 2013 that compares itself to the W3C, the standards organization devoted to making sure the Web functions correctly. Also known by its unwieldy acronym, GA4GH, it’s gained a large membership, including major technology companies like Google. Its products so far include protocols, application programming interfaces (APIs), and improved file formats for moving DNA around the Web. But the real problems it is solving are mostly not technical. Instead, they are sociological: scientists are reluctant to share genetic data, and because of privacy rules, it’s considered legally risky to put people’s genomes on the Internet.

The unfolding calamity in genomics is that a great deal of life-saving information, though already collected, is inaccessible.

But pressure is building to use technology to study many, many genomes at once and begin to compare that genetic information with medical records. That is because scientists think they’ll need to sort through a million genomes or more to solve cases—like Noah’s—that could involve a single rogue DNA letter, or to make discoveries about the genetics of common diseases that involve a complex combination of genes. No single academic center currently has access to information that extensive, or the financial means to assemble it.

Haussler and others at the alliance are betting that part of the solution is a peer-to-peer computer network that can unite widely dispersed data. Their standards, for instance, would permit a researcher to send queries to other hospitals, which could choose what level of information they were willing to share and with whom. This control could ease privacy concerns. Adding a new level of complexity, the APIs could also call on databases to perform calculations—say, to reanalyze the genomes they store—and return answers.

The day I met Haussler, he was wearing a faded ­Hawaiian shirt and taking meetings on a plastic lawn chair by a hotel pool in San Diego. Both of us were there to attend one of the world’s largest annual gatherings of geneticists. He told me he was worried that genomics was drifting away from the open approach that had made the genome project so powerful. If people’s DNA data is made more widely accessible, Haussler hopes, medicine may benefit from the same kind of “network effect” that’s propelled so many commercial aspects of the Web. The alternative is that this vital information will end up marooned in something like the disastrous hodgepodge of hospital record systems in the United States, few of which can share information.

One argument for quick action is that the amount of genome data is exploding. The largest labs can now sequence human genomes to a high polish at the pace of two per hour. (The first genome took about 13 years.) Back-of-the-envelope calculations suggest that fast machines for DNA sequencing will be capable of producing 85 petabytes of data this year worldwide, twice that much in 2019, and so on. For comparison, all the master copies of movies held by Netflix take up 2.6 petabytes of storage.

“This is a technical question,” says Adam Berrey, CEO of Curoverse, a Boston startup that is using the alliance’s standards in developing open-source software for hospitals. “You have what will be exabytes of data around the world that nobody wants to move. So how do you query it all together, at once? The answer is instead of moving the data around, you move the questions around. No industry does that. It’s an insanely hard problem, but it has the potential to be transformative to human life.”

Today scientists are broadly engaged in what is, in effect, a project to document every variation in every human gene and determine what the consequences of those differences are. Individual human beings differ at about three million DNA positions, or one in every 1,000 genetic letters. Most of these differences don’t matter, but the rest explain many things that do: heartbreaking disorders like Noah’s, for example, or a higher than average chance of developing glaucoma.

So imagine that in the near future, you had the bad luck to develop cancer. A doctor might order DNA tests on your tumor, knowing that every cancer is propelled by specific mutations. If it were feasible to look up the experience of everyone else who shared your tumor’s particular mutations, as well as what drugs those people took and how long they lived, that doctor might have a good idea of how to treat you. The unfolding calamity in genomics is that a great deal of this life-saving information, though already collected, is inaccessible. “The limiting factor is not the technology,” says David Shaywitz, chief medical officer of DNAnexus, a bioinformatics company that hosts several large collections of gene data. “It’s whether people are willing.”

Last summer Haussler’s alliance launched a basic search engine for DNA, which it calls Beacon. Currently, Beacon searches through about 20 databases of human genomes that were previously made public and have implemented the alliance’s protocols. Beacon offers only yes-or-no answers to a single type of question. You can ask, for instance, “Do any of your genomes have a T at position 1,520,301 on chromosome 1?” “It’s really just the most basic question there is: have you ever seen this variant?” says Haussler. “Because if you did see something new, you might want to know, is this the first patient in the world that has this?” Beacon is already able to access the DNA of thousands of people, including hundreds of genomes put online by Google.

One of the cofounders of the Global Alliance is David ­Altshuler, who is now head of science at Vertex Pharmaceuticals but until recently was deputy chief of the MIT-Harvard Broad Institute, one of the largest academic DNA-sequencing centers in the United States. The day I visited Altshuler in his Broad office, his whiteboard was covered with diagrams showing genetic inheritance in families, as well the word “Napster” written in large blue letters—a reference to the famously disruptive music-sharing service of the 1990s. Altshuler has his own reasons for wanting to connect massive amounts of genetic data. As an academic researcher, he hunted for the genetic causes of common diseases like diabetes. That work was carried out by comparing the DNA of afflicted and unafflicted people, trying to spot the differences that come up most often. After burning through countless research grants this way, geneticists realized there would be no easy answers, no common “diabetes genes” or “depression genes.” It turns out that common diseases aren’t caused by single, smoking-gun defects. Instead, a person’s risk, scientists have learned, is determined by a combination of hundreds, if not tens of thousands, of rare variations in the DNA code.

That’s created a huge statistical headache. Last July, in a report listing 300 authors, Broad looked at the genes of 36,989 people with schizophrenia. Even though schizophrenia is highly heritable, the 108 gene regions identified by the scientists explained only a small percentage of a person’s risk for the disease. Altshuler believes that big gene studies are still a good way to “crack” these illnesses, but he thinks it will probably take millions of genomes to do it.

The way the math works out, sharing data no longer looks optional, whether researchers are trying to unravel the causes of common diseases or ultra-rare ones. “There’s going to be an enormous change in how science is done, and it’s only because the signal-to-noise ratio necessitates it,” says Arthur Toga, a researcher who leads a consortium studying the science of Alzheimer’s at the University of Southern California. “You can’t get your result with just 10,000 patients—you are going to need more. Scientists will share now because they have to.”

Privacy, of course, is an obstacle to sharing. People’s DNA data is protected because it can identify them, like a fingerprint—and their medical records are private too. Some countries don’t permit personal information to be exported for research. But Haussler thinks a peer-to-peer network can sidestep some of these worries, since the data won’t move and access to it can be gated. More than half of Europeans and Americans say they’re comfortable with the idea of sharing their genomes, and some researchers believe patient consent forms should be dynamic, a bit like Facebook’s privacy controls, letting individuals decide what they’ll share and with whom—and then change their minds. “Our members want to be the ones to decide, but they aren’t that worried about privacy. They’re sick,” says Sharon Terry, head of the Genetic Alliance, a large patient advocacy organization.

The risk of not getting data sharing right is that the genome revolution could sputter. Some researchers say they are seeing signs that it’s happening already. Kym Boycott, head of the research team that sequenced Noah’s genome, says that when the group adopted sequencing as a research tool in 2010, it met with immediate success. Over two years, between 2011 and 2013, a network of Canadian geneticists uncovered the precise molecular causes of 146 conditions, solving 55 percent of their undiagnosed cases.

But the success rate appears to be tailing off, says ­Boycott. Now it’s the tougher cases like Noah’s that are left, and they are getting solved only half as often as the others. “We don’t have two patients with the same thing anymore. That’s why we need the exchange,” she says. “We need more patients and systematic sharing to get the [success rate] back up.” In late January, when I asked if MatchMaker Exchange had yielded any matches yet, she demurred, saying that it could be a matter of weeks before the software was fully operational. As for Noah, she said, “We are still waiting to sort him out. It’s important for this little guy.”