Gel electrophoresis is one technique used in genome sequencing.

TECHNOLOGY

Read the instructions

Remarkable progress in sequencing technologies and data handling is making personalized genome analysis an increasingly common feature of health care.

BY ANDREW R. SCOTT

After more than a decade of work, and at a cost of around US\$3 billion, the Human Genome Project yielded the DNA base sequence of a representative human genome in 2001. Now, some 15 years later, technological advances have created the next generation of sequencing machines, which

are capable of sequencing many genomes in a day at a cost of around \$1,000 each (see “Technological leap”).

“The sequencing is almost the easy part now,” says Cordelia Langford, senior scientific operations manager at the Sanger Institute in Hinxton, UK, and a participant in the original Human Genome Project. The technology is not perfect: inaccuracies still creep into

the sequencing data, and some regions of DNA cannot be sequenced at all. Then huge analytical effort is required to do something useful with the data generated. Nonetheless, the ability of modern technology to achieve so quickly and cheaply what once took years of enormously expensive work is making the dream of precision medicine more plausible by the day.

PHILIPPE PLAILLY/SPL

Genome sequencing reveals the exact order in which nucleotide molecules — each containing one of four bases, adenine (A), cytosine (C), guanine (G) and thymine (T) — are arranged along the strand of DNA. There are about 3 billion bases in a human genome sequence, arranged as complementary pairs that hold matching strands of the DNA double helix together, and they are distributed across 23 pairs of chromosomes.

Patients around the world are already benefiting from genome sequencing, and the cost is falling so sharply that the practice could soon become almost routine. The Sanger Institute, for example, is sequencing the genomes of patients with rare diseases and cancer as part of the 100,000 Genomes Project organized by Genomics England. Some participants already benefit from improved diagnosis and treatment, and researchers are discovering more about the genetic variations that cause disease.

Sequencing is not the only option in genetic analysis, however. A key part of the Precision Medicine Initiative, run by the US National Institutes of Health, is the more conventional, and arguably less technologically heroic, approach of genotyping. Here, the variants of specific genes that people carry are identified without knowing their full genome sequence. But genotyping requires some idea of what to look for. Sequencing is the only way to uncover everything about the DNA that governs the onset and progression of so many diseases, and to learn how our DNA keeps us healthy.

MOLECULES BY THE MILLION

To sequence a genome, you must first smash it into millions of bits. The original method used by the Human Genome Project, known as Sanger sequencing, made copies of parts of the initial fragments of DNA, each copy a single nucleotide longer than the last. These were then laboriously separated on electrophoresis gels and identified by the radioactively or fluorescently labelled nucleotides at the end of each strand. “Each of the fragments had to be sequenced one, or just a few, at a time,” explains Langford.

Sanger sequencing is still in use today, albeit in a more automated form. The technological advance that allows genomes to be sequenced in a single day is massively parallel sequencing. Billions of fragments can now be sequenced and read simultaneously, Langford says.

The Sanger Institute uses and tests several modern sequencing methods — part of its remit is to assess emerging technologies. Its main workhorse, however, and the method used most often in the 100,000 Genomes Project, is sequencing by synthesis (SBS). This is a finely choreographed cycle in which enzymes build strands of DNA that are complementary to template strands derived from the fragments of the genome being sequenced. Each new strand is built by adding the nucleotides that match the template one by one. At each

step, fluorescently labelled nucleotides bring the synthesis process to a temporary halt. An optical analysis system then scans the strands, which are held on a glass plate about the size of a microscope slide, and detects by way of coloured signals which nucleotides have been added. The chemical groups that block further synthesis can then be cut off and washed away, and another cycle of synthesis begins. In this way, nucleotide by nucleotide, base by base, new strands are synthesized as specified by the template strands, and the sequence in which the bases are added is recorded.

The technique was invented in the 1990s by University of Cambridge spin-out company Sol-exa, which was acquired in 2007 by Illumina, a company based in San Diego, California, that now claims a roughly 90% share of sequenced bases worldwide. “Developing the technology required the use of genetic engineering to create enzymes that will work with the modified fluorescent nucleotides,” explains Illumina’s chief scientist, David Bentley. These reactions are based on the way DNA is copied in living cells. Crucial to the advancement, Bentley says, has been the move away from natural reagents. The adoption of non-natural chemistry makes modern sequencing reactions robust and efficient enough to operate at the speeds necessary to sequence genomes in hours, rather than years, he says.

The next big challenge is one for software: analysing all the sequenced fragments and piecing them back together to form a three-billion-base genome sequence. Langford likens this to completing an incredibly complex jigsaw. But whereas a jigsaw puzzle comes with a complete picture for guidance, all the computer has to help it decide where the fragments

“The sequencing is almost the easy part now.”

should fit is the reference genome, derived from the Human Genome Project. The reference genome is a representative example of a human genome that approximates what the pieces in our individual jigsaws will create, but with slight differences that make us who we are — and these differences are central to the aims of precision medicine.

MANY WAYS TO SEQUENCE

Illumina’s SBS is one of several technologies that can read a person’s genetic code. Ion-torrent sequencing, for example, is quite similar to SBS: it also reads the sequence piece by piece from a newly synthesized strand of DNA. But rather than use a coloured marker to denote each nucleotide, the signal that distinguishes the bases comes from hydrogen ions that are released into solution when new nucleotides are added. The ions cause a detectable blip in the pH of the solution, and these blips translate into a sequence. The machine washes each nucleotide in turn through the system and monitors which one causes the ion torrent at each stage.

The length of the fragments sequenced, and therefore the complexity of piecing together the jigsaw puzzle afterwards, also varies between techniques. Some of the longest fragments are sequenced by biotech company Pacific Biosciences, based in Menlo Park, California. “Our technology delivers DNA sequence reads about one hundred times longer than the short-read technologies used in most next-generation sequencing,” says Jonas Korlach, the company’s chief scientific officer. “This makes understanding and assembling the sequence reads into complete genomes much easier.” Reading longer unbroken sections of DNA also helps to reveal complex long-range structural features, but such long-read technologies are often more

The MinION portable device was used to sequence viral genomes in West Africa’s 2014 Ebola outbreak.

expensive than other techniques.

The UK company Oxford Nanopore Technologies uses a unique system in which DNA strands are fed through tiny protein nanopores that have been inserted into a polymer membrane. Rather than requiring any DNA synthesis, the system simply notes the sequence of nucleotides passing through the nanopore, based on specific electrical signals generated by different combinations of bases. This is the technology behind the company's MinION — a portable sequencing device about the same size as a mobile phone. Clive Brown, chief technology officer at Oxford Nanopore, says that the device weighs less than 100 g; the next-smallest box on the market is 46 kg, he adds. Portability may be most important in remote areas, such as makeshift clinics set up to tackle emerging diseases in developing countries. MinION sequencing, for example, was used to sequence short viral genomes in field hospitals during the 2014 Ebola outbreak in West Africa.

Portability is simple to compare across technologies, but not all comparisons are so straightforward. Cost per sequence, for instance, depends as much on how many genomes a lab is sequencing as it does on the system being used. Accuracy can be difficult to pin down too. Manufacturers talk about accuracy of between 90% and 99.9%, often at the higher end of the range, but that still adds up to a large number of individual reads of a sequence that contain errors (R. L. Goldfeder *et al. Genome Med.* 8, 24; 2016). For this reason, genome sequencing is often repeated multiple times to achieve a truly reliable result. Practitioners talk about sequencing to differing degrees of 'depth', depending on how many times the same DNA is sequenced to increase confidence in the results. It is the accuracy of the final collated analysis that really matters.

GONE FISHING

Regardless of which sequencing technology is used, researchers and clinicians face an important decision about whether to sequence an entire genome or to take a more targeted approach. They can choose to focus on a specific region of interest in a particular chromosome. They can choose to examine only the genes that actually code for proteins or functional RNA molecules, while ignoring the vast bulk of our DNA — often misleadingly called junk DNA — that may have a crucial regulatory role or have no real function.

The exome, for example, is the part of the genome comprising only the stretches of DNA called exons that code for protein molecules. Targeting only these regions is like fishing: it requires bait. As Langford explains, an exome bait can be a collection of small sections of synthetic DNA that will bind by base-pairing to regions of DNA in a sample that identify exons. Each piece of exome bait has a corresponding magnetic bead attached to it. An external

TECHNOLOGICAL LEAP

The adoption of next-generation technologies in 2008 dramatically reduced the cost of sequencing. As the cost has fallen, the number of sequences held in publicly accessible databases such as GenBank has risen.

magnet is used to literally pull down the exon DNA, leaving everything else to be discarded. "It is an absolutely beautifully elegant technology," says Langford. Researchers can either devise their own baits for the specific parts of the genome they are interested in, or they can buy commercial bait kits that target either the whole exome or specific parts.

"Clinical applications will differ as to whether a targeted approach is enough," says Illumina's Bentley. Looking at whole genomes can detect the unexpected, such as genes that were not suspected of having a role in a disease and whose significance may be missed by a targeted approach. "For some studies exome sequencing may be okay, but it will become increasingly less sufficient as precision medicine builds," Bentley says. "There will be a moral imperative to try to fully characterize every patient and not miss anything."

Many large medical centres now have dedicated gene-sequencing centres that offer the whole gamut, from whole-genome sequencing to the precise targeting of specific genes. The Dana-Farber Cancer Institute in Boston, Massachusetts, for example, outlines the choices to patients on its website, saying: "Before starting a project, we will discuss the best sequencing strategies, experimental design, and analysis options with you." It goes on to explain that whole-genome sequencing

can discover most genomic aberrations, but that targeted sequencing is often sufficient for many clinical applications. It points out that "targeted sequencing has the advantage of sequencing larger sample sizes with lower cost and easier data analysis."

In a comprehensive review of the current state of gene-sequencing technologies, Sara Goodwin of Cold Spring Harbor Laboratory in New York discusses some of the factors that influence decisions about which technologies and methods to use (S. Goodwin *et al. Nature Rev. Genet.* 17, 333–351; 2016). Limiting the scope of an analysis can sometimes be crucial in getting fast results, she says, adding that the limiting factor for speed is often the data analysis, rather than the actual sequencing. Langford agrees, highlighting the need for "highly sophisticated software algorithms to handle the huge stream of data emerging from a modern sequencing machine".

The coming years are likely to bring more diverse applications of sequencing technology. The basic strategies that are used in DNA sequencing can also, for instance, be used to sequence RNA. Looking at RNA focuses attention on many of the parts of the genome that are most likely to have functional significance — but it may miss regions of DNA that have crucial regulatory roles, even though they are never copied into RNA. The choice of DNA or RNA sequencing, or a combination of the two, will depend on the clinical situation.

Another target, which reveals a limitation of the existing technology, is the pattern of epigenetic chemical modifications carried by some of the four bases of DNA. These modifications, such as the addition of methyl groups to specific bases, can be crucial in controlling the activity of genes — and knowing whether a gene is active can be at least as valuable as knowing which genes are present in a sequence. Bentley says that efforts to add epigenetic analysis to the sequencing toolbox are still in the research phase, but that it would provide an important additional level of information. Variations that are of crucial clinical significance may be missed by just looking at the four bases in DNA, he says, rather than by considering the effects of whatever chemical modifications they may carry.

Existing sequencing technologies are already helping many individual patients, but personalized sequencing cannot yet reveal everything that clinicians need to know to fully understand the links between DNA and disease. The technology has come a long way in the past 15 years, but "there are still many mountains ahead," says Bentley. He seems confident that solutions are within reach, however. The march towards the widespread use of personalized gene sequence analysis is well under way and is showing little sign of slowing. ■

Andrew R. Scott is a freelance science writer based in Perth, Scotland, UK.