TECHNOLOGY FEATURE

THE GENOME JIGSAW

Advances in high-throughput sequencing are accelerating genomics research, but crucial gaps in data remain.

WATCHARA/SHUTTERSTOCK

BY VIVIEN MARX

To understand why high-throughput gene-sequencing technology often produces frustrating results, says Titus Brown, imagine that 1,000 copies of Charles Dickens' novel *A Tale of Two Cities* have been shredded in a woodchipper. "Your job is to put them back together into a single book," he says.

That task is relatively easy if the volumes are identical and the shreds are large, says Brown, a microbiologist and bioinformatician at Michigan State University in East Lansing. It is harder with smaller shreds, he says, "because if the sentence fragments are too small, then you can't uniquely place them in the book". There are too many ways they might fit together. "And it's harder still if the original pile of books includes multiple editions," he says.

Researchers in genetic sequencing today face a similar task. An organism's DNA — made up of four basic building blocks, or bases, denoted by the letters A, T, C and G — is chopped into short snippets, sequenced to determine the order of its bases and reassembled into what researchers hope is a good approximation of the organism's actual genome.

Today's high-throughput sequencing technology is remarkably powerful and has led to an explosion of sequencing projects in laboratories around the world, says Jay Shendure, a molecular biologist who develops sequencing methods at the University of Washington School of Medicine in Seattle. Thousands of patient tumours and more than 10,000 vertebrate species have been or are being sequenced. High-throughput sequencing is now an essential tool for basic and clinical research, with applications ranging from detection of microbial 'bio-threats' to finding better biofuels[1].

But some types of genomic DNA cannot be sequenced by high-throughput methods, leaving many frustrating gaps in data (see 'What makes a tough genome?'). For example, a genome might contain long stretches in which the sequence simply repeats — as if Dickens had filled whole pages with a word or sentence

written over and over — making that passage hard, if not impossible, to reconstruct by the usual technologies. And the widespread adoption of next-generation sequencing has meant that the quality of genome assemblies has declined significantly over the past six years, says Evan Eichler, a molecular biologist also at the University of Washington. Although "we can generate much, much more sequence, the short sequence-read data translate into more gaps, missing data and more incomplete references," he says.

Incomplete genomes make it harder for researchers to identify and interpret sequence variations. "Instead," Eichler says, "we focus only on the accessible portions, creating a biased view," which in turn hinders efforts to study the genetic basis of disease or how species have evolved. For example, the human-genome sequence, used as a reference by scientists around the world, has more than 350 gaps, says Deanna Church, a genomicist at the US National Center for Biotechnology Information. An updated reference genome is filling in much of the missing data, but "even with the release of the new assembly, there will still be gaps and regions that aren't well represented," she says. "It is definitely a work in progress."

More than 900 human genes are in regions where there is much repetition. About half of these genes are in areas so poorly understood that they are often excluded from biomedical study, says Eichler. Certain regions of chromosomes, notably those near centromeres (where the two halves of a chromosome connect) and telomeres (the ends of chromosomes) are especially incomplete in the reference genome.

UNIV. WASHINGTON

Jay Shendure is working to develop the next generation of sequencing methods.

This lack of information can have medical consequences. For example, researchers have known for more than a decade that medullary cystic kidney disease — a rare disorder that occurs in mid-life — can be caused by mutations in a gene hidden somewhere along a 2-million-base-pair stretch of chromosome 1. Early detection of the mutation is the first step towards preventative therapies, but would require a DNA test. The gene, however, lies within a region rich in sequence repeats as well as in the bases guanine (G) and cytosine (C). Such 'GC-rich' regions, like repetitions, are difficult to sequence.

Only by reverting to Sanger sequencing — a classic but more laborious approach — and combining it with special assembly methods were researchers able to decipher the DNA region involved in the disease. The results, which were published in February[2], mean that a test to screen younger members of families affected by the disorder is now a possibility.

Sanger sequencing is a painstaking process in which each type of DNA base is labelled with a different compound. The labelled DNA is then separated and the sequence is read. For the Human Genome Project, researchers combined Sanger sequencing with techniques to establish markers that locate where the sequences fit. The approach, which has been in use for decades, delivers a read accuracy and contiguity of sequence that are unmatched by current technology, Shendure says. "I couldn't do anything remotely approaching the quality of what resulted from the project." But the art of Sanger sequencing and its associated methods cannot be scaled up for the high-throughput sequencing projects done today. "We need to think about how to 'next-generation-ify' all of this," he says.

Research to do just that is well under way, with a variety of methodologies that address problems such as repetitive sequences and GC-rich regions, as well as the knotty task of assembling complete genomes for organisms that have four or even eight copies of each chromosome, for example, as opposed to humans' two.

Some of the technologies on the horizon promise to deliver longer reads and, possibly, fewer headaches for researchers trying to assemble them. But until those instruments are on bench tops, scientists are combining new and old approaches to refine sequencing.

RICH IS POOR

Some of the newer approaches aim to tackle GC-rich regions. For high-throughput sequencing, DNA is often first chopped into short fragments, which are then amplified by polymerase chain reaction (PCR). But the enzyme used in PCR "has trouble getting through" GC-rich regions, says Shendure. As a result, GC-rich stretches can end up poorly represented in the DNA sample delivered to the sequencer, thus skewing the data. Some sequencing technologies, such as those made by Illumina, based in San Diego, California, use amplification before and during the sequencing process, causing further bias against GC regions.

A number of sample-preparation approaches reduce this GC bias. The amplification step is cut out completely in platforms made by Pacific Biosciences, based in Menlo Park, California, and in a method being developed at Oxford Nanopore Technologies in Oxford, UK. And although DNA read lengths differ among platforms, the most widely used bench top sequencers — which are made by Illumina — generate short reads, of around 150 base pairs.

"The killer with short reads is that they're very sensitive to repeated content," says Brown.

What makes a tough genome?

Certain features of DNA are challenging for high-throughput sequencing.

- Long sequences of repeated bases
- Missing bases in the original sequence
- Degraded or damaged DNA
- Regions rich in guanine and cytosine

If the read length is shorter than a repeat — or, to draw on the book analogy, if the shreds of the novel are only a fraction as long as a repeated paragraph — it is hard or even impossible to uniquely place. "That's where things like long reads or other technologies can be helpful," says Shendure. Long DNA fragments can bridge repetitive regions and thus help to map them. As another way to ease assembly, researchers in Shendure's group and elsewhere are exploring different methods to tag and group DNA fragments before sequencing. "There are more on the horizon," says Shendure, but he prefers to divulge the details in research publications.

The terms 'short' and 'long' are in a state of flux in this fast-moving industry. The first generation of Illumina machines generated reads of around 25 base pairs in length; the latest ones have upped that to around 150 base pairs (see 'Extended sequence'). But it is still hard to assemble a complete genome from reads of this length.

Geoff Smith, who directs technology development at Illumina in Cambridge, UK, acknowledges the drawbacks of short-read technology for sequencing repetitive regions and various types of genomic rearrangements. He says that the company aims to address issues that crop up as researchers compare genomes they sequence to reference genomes, or sequence organisms from scratch without references.

Illumina has launched a service to allow longer reads with its current short-read technology. Last year the firm bought Moleculo, a company based in San Francisco, California, which has developed a process to create long reads by stitching together short ones through a proprietary sample-preparation and computational process. In July Illumina began offering Moleculo's process as a service for customers.

The Moleculo process first creates DNA fragments about 10,000 bases (10 kilobases) in length. The fragments are sheared and amplified, then grouped and tagged with a unique barcode that helps to identify which larger fragment they originated from and aids in assembly.

EXTENDED SEQUENCE

Illumina's first high-throughput sequencers produced very short reads.

ILLUMINA

At present, sample preparation for the Moleculo process takes around two days. Smith says that he and his team are refining the process and that by the end of the year Illumina will launch Moleculo as a stand-alone sample-preparation kit. He says that company scientists are now evaluating the kit's performance by sequencing a well-known genome, but he prefers not to say which one.

"We suspect you will be able to uncover a lot more of the genome with 10-kilobase reads versus the [150-base-pair] read length that we currently have," says Smith. He adds that the company plans to increase the fragment length to 20 kilobases. He and his team hope to "develop better molecular-biology tools to allow us to reach into these difficult-to-sequence parts of the genome but also use those tools on well-characterized genomes," he says. The team is also tuning the Illumina software to better distinguish between false and correct reads.

The company's initiative comes at a time of intense commercial and academic activity around long-read sequencing technology and new assembly methods. Finished genomes have taken a back seat, leaving many highly fragmented assemblies that need completing, says Jonas Korlach, chief scientific officer of sequencing manufacturer Pacific Biosciences in Menlo Park, California, whose sequencer generates read lengths of around 5 kilobases.

Korlach agrees that long reads will help to sequence repetitive regions, such as those that characterize many plant genomes, for example. They will also help with the challenge of distinguishing between copies of chromosomes, important in identifying the tiny variants that can affect biological function. Humans are diploid, meaning they have two copies of each chromosome, but "many organisms, especially plants, have even more copies, which makes resolving all the different chromosomes so much harder", Korlach says.

TOUGH NUTS

Plant sequencing, in particular, will benefit from improvements[3]. The spruce genome is a "real nightmare", says Stefan Jansson, a plant biologist at the Umeå Plant Science Centre in Sweden. Jansson led a study that generated a draft assembly of the Norway spruce genome (*Picea abies*)[4]. In addition to being the largest genome yet sequenced, it also contains many repeats, and the differences between its chromosomes are larger than in the human genome. "Sequencing diploid spruce is like mixing human and chimpanzee DNA and then trying to assemble them simultaneously," Jansson says.

Many plants have more than two copies of chromosomes. Bread wheat (*Triticum aestivum*), for example, is hexaploid, and sequencing and assembling the six sets of chromosomes to completion has proven extremely difficult. And although some strawberry species are diploid, the commercial strawberry (*Fragaria* × *ananassa*) is octoploid: it has eight sets of seven chromosomes, four sets from each parent, says Thomas Davis, a plant biologist at the University of New Hampshire in Durham. "Good thing Mendel didn't use octoploid strawberries to try to understand heredity," he says.

Davis and his colleagues have published a draft genome of the diploid woodland strawberry (*Fragaria vesca*), and now want to apply their experience to the octoploid strawberry[5]. Assembling this tough-nut genome will require high-quality reads longer than 500 base pairs, Davis says. He believes he can succeed, although he does not want to share his methodology just yet. "If anyone cracks that nut, he'll do it," says Kevin Folta, a molecular biologist at the University of Florida in Gainesville, who led the woodland-strawberry project.

The plant world has many other challenging genomes to offer. The onion genome is massive, Folta says, and sugarcane has 12 copies of each chromosome. "Those will take special techniques," he says.

Every platform has benefits and drawbacks, and scientists must weigh the costs, sample-preparation time and sequencing-error rates for each. To sequence the woodland strawberry, for example, the scientists used a combination of three platforms.

But for polyploid genomes, short-read sequencing is almost a waste of time, says Clive Brown, chief technology officer at Oxford Nanopore. "You don't know where your short read comes from, which chromosome it is from," he says. "It's very hard to piece that together." He believes that the problem will be helped by instruments, including those in development in his company, that can generate long reads without the need for special sample preparation or complex assembly. The longer the reads, the easier the assembly, because the overlapping sequences will help researchers to determine which sequence belongs to which chromosome.

Fresh approaches were needed to crack the genome of the oil palm (*Elaeis guineensis*), reported last month in *Nature*[6]. The effort was more than a decade in the making. Oil palm is an important source of food, fuel and jobs in southeast Asia, and the industry is under pressure to produce it sustainably and avoid increased rainforest logging, says study

co-leader Ravigadevi Sambanthamurthi, director of the advanced biotechnology and breeding centre at the Malaysian Palm Oil Board in Kajang, which works with the country's oil-palm industry.

With millions of repeats distributed throughout the plant's genome, short reads could fit in many possible spots in the assembled DNA sequence. "It is as if you were assembling a jigsaw puzzle in which most of the pieces are identical," says Robert Martienssen, a geneticist at Cold Spring Harbor Laboratory in New York, who co-led the project with Sambanthamurthi.

Classic sequencing methods were too laborious and expensive for the oil-palm project. So Martienssen suggested applying a technique based on a finding he had made in 1998 — that repeats in plant genomes can be distinguished from genes because the cytosine bases in the repeats usually carry methyl groups. Before fragments are sent to the sequencer, they are treated with enzymes that digest methylated DNA and thereby remove the repeats from the samples.

To complete the oil-palm project, the scientists applied this methylation-filtration technique and then sequenced the DNA regions housing genes. The technique has now been commercialized through Orion Genomics, a company based in St Louis, Missouri, which Martienssen co-founded.

The researchers used a high-throughput sequencer made by 454 Life Sciences, a company owned by Roche and based in Branford, Connecticut, that generates short reads from longer, filtered fragments. In preparing the samples, the researchers used bacteria to amplify DNA in large chunks on bacterial artificial chromosomes — an approach also used in the Human Genome Project — to pin down hard-to-map regions by retaining them next to genes with known positions to act as signposts.

Assembly of the oil-palm genome called for extensive computational resources, which crashed multiple times, the researchers say. But now, with the genome in hand, they have located a gene that encodes the shell of the palm fruit, knowledge they hope to harness to increase the plant's yield.

Sambanthamurthi says that when the researchers finally pinned down the shell gene, they popped a bottle of champagne, then celebrated with a traditional Malaysian meal served on a banana leaf.

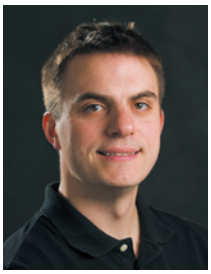
THE LONG AND THE SHORT

Bacterial genomes are smaller and less complex than those of plants and other multicellular organisms, but they, too, have regions that are tough to sequence. For example, *Bordetella pertussis*, which causes whooping cough, has hundreds of insertion sequence elements — stretches of mobile DNA inserted into various locations in the genome — each more than 1 kilobase long. Proponents of long-read technology say that spanning these regions with long reads will deliver sequencing efficiency gains.

Korlach points out that it took a team of more than 50 scientists to solve the bacterium's complete genome[7]. But long-read technology can make assembly of highly repetitive genomes faster and easier, he says. He says that he and scientists in the Netherlands were able to assemble nine whooping-cough bacterial strains in one month.

Whether a read is classified as 'long' or 'short' is in great flux. Two years ago, scientists might have said that a long read was 1 kilobase, Korlach says. "Now [Pacific Biosciences] customers are generating an average of 5,000 bases, with some reads longer than 20,000 bases — and we are working to deliver even more than that." Ultimately, a 'long read' will be as long as is needed to sequence a given genome, he says.

Korlach knows that some scientists say his company's sequencers are pricey, but he says that the newer versions have seen a significant drop in price and an increase in throughput. He says that the question of price is often raised "in the context of pure cost per sequenced base". And, he adds, if a certain sequencing technology is the only one that will work to solve a medically important question, "then there is no price tag that can be put on this medically relevant information".



MICHIGAN STATE UNIV.

Titus Brown likens high-throughput sequencing to piecing together 1,000 shredded copies of a novel.

Last year, researchers collaborating with Pacific Biosciences used the company's sequencer to distinguish the repetitive genomic region involved in fragile X syndrome, a developmental disorder that is caused by repeats in a particular region on the X chromosome, and that worsens in severity with higher numbers of repeats[8].

As technology developers get closer to instruments that produce longer reads, scientists will need longer DNA fragments at the beginning of their sequencing experiments. Several companies focus on helping researchers to prepare DNA fragments for sequencing. For example, Sage Science, based in Beverly, Massachusetts, has a platform that uses pulsed-field electrophoresis to select and sort DNA fragments of sizes ranging from 50 base pairs to 50,000 base pairs. In May, the company began marketing its instrument to accompany the Pacific Biosciences sequencing platform.

Steve Siembieda, who is responsible for business development at Advanced Analytical Technologies in Ames, Iowa, says that his company sees the trend towards longer reads as writing on the wall. The company has licensed patents from Iowa State University, also in Ames, to develop an instrument to assess the integrity, fragment length and concentration of DNA samples.

With this instrument, an electric field is applied to a tiny amount of DNA so that it is pulled into a long, hair-thin capillary tube containing a gel with a fluorescent dye that binds to DNA molecules. As the DNA fragments move through the gel, they separate according to size. "Small molecules move fast, big molecules move slowly," Siembieda says. As the molecules pass by a window in the capillary, a flash of light excites the dye and a camera records the DNA fragment length (see 'Bits and pieces').

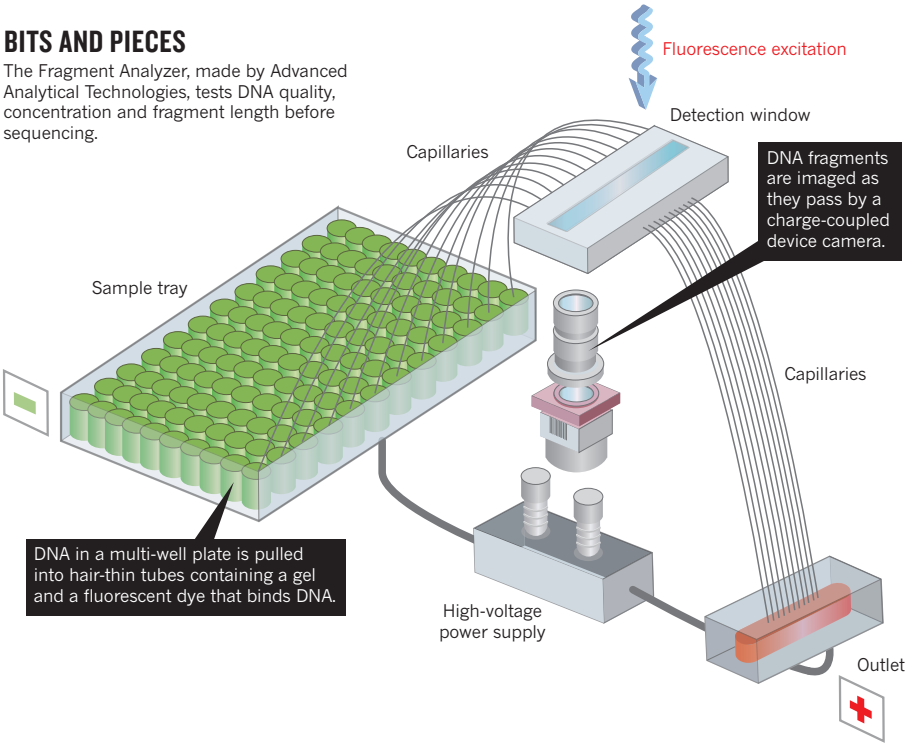
The instrument's readout tells scientists whether the size distribution of the DNA fragments is in the range needed for a given sequencing platform and whether the DNA

ROBERT MARTIENSSEN/CSHL

Oil palm has a complex and repeat-riddled genome that took more than ten years to sequence.

BITS AND PIECES

The Fragment Analyzer, made by Advanced Analytical Technologies, tests DNA quality, concentration and fragment length before sequencing.



has the right concentration. Siembieda says that skipping these measurements can be the wrong experimental shortcut — if the concentration or fragment size is off, "a sequencer may run for nine days, it will cost them thousands of dollars, plus all the time wasted to not make sure they have the appropriate material". The instrument will possibly be used in developing the Moleculo process, but negotiations between the two companies are still under way.

Technology development at Advanced Analytical is focusing increasingly on long DNA fragments, which are challenging to resolve, Siembieda says. One solution is to customize gels for different applications. At present, the company's instrument can resolve lengths of up to 20 kilobases and the company is working on resolving longer fragments, he says.

ASSEMBLY REQUIRED

Scientists are applying many methods and tricks to create longer fragments. "Unfortunately, these technology tricks create erroneous data at points, so now you're stuck with some data that may be wrong," says Michigan State's Titus Brown. He was part of an effort, published in April[9], to sequence the lamprey (*Petromyzon marinus*) genome, one-third of which is covered by long repeats. Obtaining an assembly even with Sanger sequencing, which generates 1-kilobase reads, was difficult, he says. In addition, the lamprey genome has many GC-rich regions. The team used several types of software to assemble the complete DNA sequence.

In July, scientists published a comparison of software programs used to assemble sequence reads[10]. The researchers found that different assemblers give different results — even when fed the same sequence reads. Brown says that biologists should never forget that assemblies are not certainties. Every new sequencing technology — from how the DNA sample is prepared to the sequencing chemistry — has the potential for error and bias. "If you have short reads, or bad biology, you're going to have a very hard time getting a good assembly, even in theory," he says.

Ideally, a genome assembly should deliver end-to-end chromosomal sequences, says Shendure. What worries him more than the discordance among assemblers in the comparison study is that all of the assemblies were very fragmented. "That's not a fault of the assemblers, that's a fault of the data that we're putting into the assemblers and the fact that we're not capturing contiguity at these longer scales," he says. "The algorithms can only make do with the ingredients that they are provided by the technologies."

Brown is hopeful about the potential impact of longer-read technology. If Pacific Biosciences or Oxford Nanopore "deliver on many inexpensive long reads — more than 10 kilobases, I'd say — regardless of accuracy, you would end up revolutionizing the genome-assembly field, because it would give you so much more information to work with", he says. However, he adds, assembly software has to be compatible with each sequencing method. "So we're continually playing catch up, where new sequencing technologies lead to new sequence-analysis approaches a year or three later."

Eichler agrees that sequencing and assembly must continue to improve. Read lengths longer than 200 kilobases and with 99.9% accuracy rates will be needed to unpick repeats and other complications, he says. He says that the Pacific Biosciences instrument and what he knows of Moleculo "fall short of this, but are on the right track". All read-length requirements depend on the genome and complexity, he adds. For many bacterial genomes, current read lengths and accuracy are already sufficient, he says.

THE NEXT TELESCOPE

Oxford Nanopore plans to launch its new sequencing technology in the near future, but no date has been given. The technology expands on findings by researchers at Harvard University in Cambridge, Massachusetts, the University of Oxford, UK, and the University of California in Santa Cruz to harness the abilities of pore-forming proteins for DNA-sequencing devices.

One of the weaknesses of current high-throughput sequencing technology is amplification chemistry, says Oxford Nanopore's Clive Brown. Although DNA is made up of four bases, it is possible that more than those canonical four — such as bases that are methylated — should be detected, he says.

And in some sections of genomes, bases are naturally missing. But current sequencers do not capture such variations — instead, says Brown, they produce the equivalent of a four-colour photocopy of a picture with many more colours. "A lot of the detail is lost immediately, as soon as you make a four-colour copy," he says. Ideally, "you take a chromosome and run it through the sequencer. You can't quite do that yet." He, too, says that the next crucial phase of sequencing technology will be about long reads.

Brown says that to his mind, sequencers are just opening the door to characterizing the genome. People can get "very cosy about what they can see", with scientific instruments, he says. He likens today's sequencers to the first telescopes, which offered a view of the Moon's features and exploration of the visible spectrum. "It gets you a long way, you can count the stars, see the planets," he says. But the telescope does not show other celestial phenomena — such as dark matter or galactic movement.

Like astronomers with their telescopes, genome researchers will get a clearer picture of the genome as the sequencing technologies improve, he says. And, inspired by that picture, they will strive to see even more. ■

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

1. Shendure, J. & Aiden, E. L. *Nature Biotechnol.* **30,** 1084–1094 (2012).
2. Kirby, A. *et al. Nature Genet.* **45,** 299–303 (2013).
3. Schatz, M. C., Witkowski, J. & McCombie, W. R. *Genome Biol.* **13,** 243 (2012).
4. Nystedt, B. *et al. Nature* **497,** 579–584 (2013).
5. Shulaev, V. *et al. Nature Genet.* **43,** 109–116 (2011).
6. Singh, R. *et al. Nature* **500,** 335–339 (2013).
7. Parkhill, J. *et al. Nature Genet.* **35,** 32–40 (2003).
8. Loomis, E. W. *et al. Genome Res.* **23,** 121–128 (2013).
9. Smith, J. J. *et al. Nature Genet.* **45,** 415–421 (2013).
10. Bradnam, K. R. *et al. GigaScience* **2,** 10 (2013).