https://doi.org/10.1038/s41592-022-01730-w

Long-read sequencing has buoyed projects in small genomics labs and large-scale projects.

METHOD OF THE YEAR: LONG-READ SEQUENCING

To large-scale projects and individual labs, long-read sequencing has delivered new vistas and long wish lists for this technology's future. **By Vivien Marx**

To the delight of scientists across the life sciences, reads, which are the output of sequencing instruments, have been getting longer. Reads might be sequenced DNA or RNA and could one day routinely be entire genomes, transcriptomes and epigenomes at high throughput and accuracy, and maybe even the amino acid sequences of proteins.

Academics have happy tales about how long-read technologies have empowered their genomics projects. A few companies have facilitated this journey, notably Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Of late, other firms presenting long-read approaches include Element Biosciences, Illumina and MGI. Ultima Genomics and others have plans in this area.

Long reads have buoyed numerous findings in individual labs. In larger ventures, among the celebrated achievements are those in the Vertebrate Genomes Project (VGP) and the Telomere-to-Telomere Consortium (T2T)[1]. A set of papers and news features related to the T2T Consortium can be found as a Nature Collection online. During the T2T project, says Adam Phillippy, a researcher at the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) who co-leads the T2T Consortium, the longest read he and his colleagues handled had one million base pairs. Long-read sequencing is being used by the Human Pangenome Reference Consortium (HPRC)[2–4]. The HPRC teams want to assemble the human genome at the T2T level of completion and capture a "better spectrum of humanity in terms of how they represent allelic diversity," says University of California Santa Cruz researcher Karen Miga, who co-leads the T2T Consortium with Phillippy and is part of the HPRC.

Population-level data from diverse populations are needed, says Heidi Rehm, who, among other appointments, is the chief genomics officer at Massachusetts General Hospital's department of medicine and medical director of the clinical research sequencing platform at the Broad Institute of MIT and Harvard. She and her colleagues have found instances in which Black people received information about risk of a heart condition without sufficient evidence on genetic variants to support it[5]. Population data had been lacking about these variants, and such data are still limited, says Rehm.

One population- and diversity-focused project underway involves 500 individuals: 50 from each of 10 different Australian Indigenous communities. Samples are taken with consent from and with involvement of these communities. From these individuals' data, one genome that represents each Indigenous community will be assembled telomere to telomere. The scientists apply whole-genome sequencing with long reads from PacBio and ONT sequencers and Illumina short-read technology. "The jury is still out on what is the best technology for telomere-to-telomere sequencing," says Hardip Patel from the National Centre for Indigenous Genomics at Australian National University, who co-coordinates the project with Ira Deveson at the Garvan Institute of Medical Research. Sequencing technology itself is rapidly changing. In 2023, the scientists will revisit their technology choices, says Patel. Technology choices have shaped long-read achievements to date and color the wish lists for the future of long-read sequencing.

It's had a big impact on the canid community to work with long-read sequencing, says NIH NHGRI researcher Elaine Ostrander given that multiple reference sequences are needed to represent different canids. Here, the Ostrander lab with companions. From left to right: Emily Dutrow, Reuben Buckley, Elaine Ostrander, Alexis Harris, Jess Hale and Dayna Dreger.

CREDIT: OSTRANDER LAB, NIH NHGRI

Achieved with long reads

Says Rockefeller University researcher Erich Jarvis, who is also a Howard Hughes Medical Institute investigator, long-read sequencing has led to fewer gaps in genome assemblies. The biological benefits of this technology for his projects include a more accurate assessment of gene duplications and their orthology, and thus a greater understanding of gene family evolution. He points to work by Constantina Theofanopoulou, a Hunter College researcher and visiting professor at Rockefeller University. Long-read sequencing helped her parse the evolutionary history of the oxytocin and vasopressin ligand and receptor families. By studying synteny — long, conserved blocks near the genes of interest — she and her team located orthologous genes across species. "It is impossible to run solid long-range synteny analysis with short reads," she says.

Long-read sequencing has had a big impact on the greater canid community, says Elaine Ostrander from NIH NHGRI, who runs a number of studies in the Dog Genome Project. This impact stems, for example, from the fact that multiple reference sequences are needed that represent different canids — wolves, coyotes and domestic dogs, among others. Given their quite dissimilar histories, different clades of domestic dogs must also be represented. Studying dogs with long reads sheds light on domestication and thus human migration, she says. Although such questions could be approached with assembled sequence from multiple types of canids from around the world and with alignment of that information to domestic dog sequences, says Ostrander, "that is intrinsically error prone when considering wild canids, or ancient canids, and does not accurately reflect history, particularly as it relates to the location and timing of domestication for many canids."

Says Jarvis, long-read sequencing makes it possible to measure gene network interactions across chromosomes in ways not previously possible. These reads capture G+C-rich regions, which are mainly found in gene regulatory regions. That yields, he says, "a much more complete picture within and across species of the DNA promoter regions that regulate genes." All of this also matters to the VGP, which Jarvis chairs. Chul Lee, a postdoctoral associate in the Jarvis lab, led the development of methods[6] to quantify the difference long reads make. The use of long reads squelched thousands of errors from previous genome assemblies of a number of animal species because false gene gains and losses in short-read-based assemblies were corrected.

Isidro Cortes Ciriano and his cancer-genomics-focused team at European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) develop computational tools to, for example, assess mutation patterns and genome instability in cancer. Long-read sequencing, he says, delivers ways to study repetitive and complex genomic regions such as centromeric regions, long repeats and complex structural variants. With long reads generated with nanopore sequencing on ONT instruments, they can "resolve those complex genomic aberrations in cancer that are recalcitrant to Illumina sequencing," he says.

Carolin Sauer, a postdoctoral fellow in the Cortes Ciriano lab, says that long reads are of particular interest to researchers working on cancers with copy number aberrations and unstable genomes, such as esophageal and ovarian cancers. Long-read approaches are generally better for detecting and characterizing the complex genome rearrangements and structural variation typical of many cancers.

Among the gnarly genome sections more readily tackled with long reads, says Patel, are the human genome's many types of repetitive elements: short tandem repeats of a few hundred base pairs; Alu elements, which can run around 300 base pairs; LINE1 elements, which can be up to six kilobases long; segmentally duplicated regions hundreds of kilobases

long; and the megabases of repeats within repeats such as centromeres and ribosomal DNA. These all differ in their mutational processes and regulatory roles.

Long-read sequencing has been "a huge deal" to him and his team, says Fergal Martin, who leads the EMBL-EBI's eukaryotic annotation team. The vastly higher-quality sequence helps the team to tease out structures such as genes and repetitive sequence. And with long-read RNA sequencing, researchers can describe expressed genes and find gene structures. "So it's a double win," he says.

In her microbiome and metagenomics projects, Karoline Faust, a researcher at KU Leuven, works with organisms of known genome sequences and uses ONT's MinION for "cheap in-house cross-contamination checks." Right now, to confirm the bugs in the bioreactor are the ones the lab put there, the team needs to use 16S rRNA Sanger sequencing, but that doesn't distinguish strains or identify fungi. "In my case, cheap and easy contamination checks" and organism identification are the greatest promise of long reads. Price and speed matter in such instances since spotting contamination quickly means one can quickly halt an expensive experiment.

In metagenomics, long reads "have not fully arrived and may still take time," says University of California Davis researcher C. Titus Brown. That's due to the challenges of DNA extraction for long molecules and because complex microbiomes cannot yet be sequenced at sufficient depth. Long-read sequencing successes in metagenomics, he says, mainly involve host-associated microbiomes, which are less complex and involve fewer microbial strains than microbiomes in marine environments, sediment and soil.

To tease out tough-to-find structures in the genome, long read-sequencing has been enormously helpful, say these EMBL-EBI researchers: Fergal Martin (left) leads the eukaryotic annotation team; Carolin Sauer (right) is a postdoctoral fellow in the cancer-genomics-focused lab of Isidro Cortes Ciriano (middle).

CREDIT: J. DOWLING, EMBL-EBI

Long reads past and next

ONT's technology grew out of ideas from David Deamer, then at University of California Davis; Hagan Bayley at University of Oxford; and Dan Branton from Harvard University. The company, says ONT's CEO Gordon Sanghera, sought out partners in academia from the start. The technology involves threading a string of nucleic acids through a protein nanopore. An ionic current flows across the nanopore and dips as bases pass through. A group of bases are read at a time, with around six giving the strongest signal. Thus, there are around 4,096 possibilities of what the bases could be, says Johns Hopkins University researcher Steven Salzberg. That could make it hard to use this technology, but in fact, as a hypothetical example, when the nanopore sees the sequence AGCTGA, the sequence after that only has to add only one of four letters to the last five. The next sequence is thus either GCTGAA, GCTGAC, GCTGAG or GCTGAT. "The base-calling software has to figure out which of those four possibilities is next," using the change in current, he says.

Early on, as a University of Oxford spinout around 2015 with devices still "rudimentary in performance," says Sanghera, ONT invited scientists to try it. Around 3,000 applications quickly poured in. "It was a very broad church," he says and included a high school student in Guatemala and "some pretty powerful profs" eager to generate data at their desks in real time.

Among the envisioned applications were some on infectious disease, plant research and environmental testing. Having at one's fingertips an affordable way to generate data is transformative for the coming "genomic era," he says, in which sequencing will be applied in many ways and become more widespread. An admittedly futuristic approach, says Sanghera, is a toothbrush with a built-in ONT device. It could be used to check for "signatures" in the tiny amount of blood released when people brush their teeth. It might indicate that a cancer remains in remission or that a physician visit might be warranted. Other applications could involve detecting pathogens in food or the environment. "We have ideas and brainstorms about the kinds of products that we could bring to bear," he says. The toothbrush embodies, says Sanghera, the concept of "what we think will come."

With PacBio's High Fidelity (HiFi) sequencing, the instrument builds a consensus sequence from multiple passes around a circular molecule. HiFi sequencing grew from the company's core long-read technology — single-molecule real-time sequencing, or SMRT-sequencing[7] — that Jonas Korlach, PacBio's chief scientific officer, co-developed and that launched PacBio's path into labs.

Back in graduate school in Watt Webb's lab at Cornell University, Korlach was fascinated by macromolecular machines, one of which is the enzyme DNA polymerase (DNAP). DNAP zips along at 100 bases per second to copy the genome, such as during cell division. It's "the most powerful sequencing machine" that evolved over millions of years, he says. Korlach wanted to watch DNAP in action and worked out how to use labeled nucleotides to distinguish the four DNA bases. He explored, for example, how the enzyme could hold on to a developing DNA strand as each base is added to the complementary strand. Without a way to image a single polymerase molecule, however, the sea of labeled nucleotides would be one big blur. Webb suggested reaching out to Harold Craigshead in a neighboring Cornell lab.

There, Stephen Turner was a PhD student, and a partnership and friendship between Korlach and Turner developed that lasts to this day. Sequencing single molecules in real time intrigued Turner, and he developed zero-mode waveguides[8] that hold DNA in tiny holes embedded in a metal film. Turner founded PacBio and Korlach was employee number eight. They developed ways to make SMRT sequencing robust — for example, by attaching the label to the nucleotide in such a way that the DNAP doesn't disengage from the developing DNA strand. As PacBio and ONT

They met in graduate school and are friends and collaborators to this day. Stephen Turner (left) is PacBio's chief technical officer and founder. Jonas Korlach (right) is PacBio's chief scientific officer. He was employee number eight.

continue to advance their instruments, labs advance their wish lists.

Wish-list item: one box

Most genomics projects apply multiple technologies to wrestle with aspects such as sequence duplications, structural variation and a lack of diverse reference genomes. To build a more adequate representation of human global genomic diversity, the HPRC tested methods on HG002, an Ashkenazi man from the Human Genome Project who consented his data be used. Among the HPRC's sequencing technologies are PacBio HiFi long reads, ONT long reads, 10x Genomics linked reads, HiC linked reads, optical maps and Strand-seq data for co-scaffolding contiguous stretches.

Accuracy is one reason multiple technologies are used, says Guojie Zhang from Zhejiang University. Were long-read sequencing more accurate, labs would no longer need multiple technologies for curation and validation, he says. Says Rehm, long-read sequencing is useful for resolving some of the more complex structural variants in the human genome, and it gives researchers a way to study regions of high homology that contain genes that matter clinically. With long reads, scientists can resolve short tandem repeats better than with short-read sequencing and phase variants to determine which alleles they belong to. She works with ONT, PacBio and Illumina technology on rare diseases. In her view, long-read sequencing needs to get "definitely cheaper," she says. "And ideally, we'd get everything on one platform and not multiple."

A one-box long-read sequencing world has not yet arrived, but there are, for example, designed technology pairings. ONT and 10x Genomics have set up a protocol to use ONT PromethION devices and perform sample prep for the 10x Genomics platform to connect nanopore sequencing and 10x Genomics' single-cell and spatial assays. Users can capture sequence and, for example, isoform-specific transcript abundance and spatial transcriptomic data.

It would be a best case, says Doron Lipson, chief scientific officer of Ultima Genomics, if labs had one box for solving all, but to get "the full, complete picture, end-to-end of everything," they will likely always have to combine approaches. For now, Ultima's instruments are short-read sequencers that generate 300- to 400-base-pair reads, which they have pushed internally to 600–800, he says. The technology is a microfluidic system that uses sequencing by synthesis with nucleotides to build up a DNA strand with chemistry that needs only a small fraction of nucleotides to be labeled. Instead of pricier flow cells, Ultima uses proprietary disks. This cuts costs, especially for teams in large-scale sequencing projects, genomics centers and core facilities, he says.

Pocket-sized instruments like ONT's leverage cloud-based computing, which lowers the cost of data analysis and heightens accessibility. Beyond offering technology to help labs resolve repetitive elements in genomes, Ultima wants its instruments to generate information about genomic regions interacting across large distances. The goal overall is to make sequencing 10–50 times cheaper than now and capable of generating many different kinds of data.

Lipson sees Illumina's recently launched long-read technology as similar to approaches that have "been around for a while," he says. It involves long reads in which DNA is fragmented for sequencing and tagged to keep track of each fragment, followed by local assembly. Ultima is also exploring approaches and collaborations in this area, he says. For now, that's all the detail he wishes to disclose.

Labs might one day have "one box" that gives them all the data they need, says Lipson, "but that's still a while away." Ultimately, sequencers will become "agnostic measurement devices" with which a scientist can obtain a readout by converting a parameter of interest into something that can be captured with a DNA sequence signal. Currently, an "inflection phase" is underway that is leading an increasing number of labs to scale up sequencing for their projects. Even smaller labs can become "high-scale" sequencing facilities as both prep and analysis become easier than before. Biology's puzzles will remain, says Lipson. It's "the beauty and the curse of biology" that when measurements deliver insight on what was poorly understood, it unleashes new questions.

Wish-list item: lower cost

Says EMBL-EBI's Martin, long-read sequencing's applications would be greatly improved if it could also capture full-length sequences from genes with low expression levels and small amounts of sample. "This is really important for biodiversity initiatives, where there may be very limited samples available for a species," he says. He also wonders how PacBio and ONT can support these "essentially not-for-profit endeavors," which stand to profoundly shape planetary health.

ONT values low-cost access to DNA and RNA data through sequencing, says ONT's Sanghera. During the pandemic, its instruments were used around the world by public health labs that lacked sequencing infrastructure and sequencing experience. ONT basically gives away its devices, he says, and not just as a COVID-19-related approach. "You only buy consumables," he says, which are flow cells and reagents.

Over time, long-read sequencing has become higher throughput and easier to use, and costs have come down, says Korlach, who is involved in research projects including the VGP and the HPRC. On the computation side, says Korlach "we're piggybacking on the telecom industry." Computing gets faster as a result of advances in processors and components such as GPUs "So I think it's inevitable that long-read sequencing is going to get faster and cheaper and easier to use."

Scientists are welcome to approach PacBio about potential collaborations. "If they want to do a lot of sequencing, then we try to incentivize that," says Korlach. "If somebody wants to do something that's never been done before and we're excited about it, we're happy to support that." That can, for example, translate into a discount or other types of support so his company can help make the project happen.

For labs, both cost and accuracy are important considerations, says Patel. PacBio's HiFi platform delivers long reads with high accuracy levels, he says. But it's more expensive per base than ONT's platform.

To him, for long-read sequencing "with few exceptions, there's only one real choice, and that's Oxford Nanopore," says Salzberg. The company's small devices can easily accompany scientists into the field. To assemble a genome accurately, "HiFi technology is great,"

CREDIT: PACIFIC BIOSCIENCES

he says. "So it's good we have it." But most labs lack funds to finance using it. HiFi sequencing is brought to bear, for example in large-scale research efforts such as the NHGRI-funded HPRC. But, he says, individual labs, including his, generally do not have grants large enough to afford this technology.

Wish-list item: greater accuracy

In Patel's experience, ONT sequencers have continually improved accuracy rates. When he and his colleagues started collecting data last year for a project now being scaled up, the accuracy of ONT's platform was at 92%. "Now it's close to 99% at a base level," he says.

"We need longer and more accurate reads that are perfect or near perfect in their nucleotide accuracy, especially to get through long centromere regions," says Jarvis. Also needed: technology to scale up to produce complete and error-free genomes at scale for thousands of species per week to complete the goals of the VGP and the Earth Biogenome Project. Currently "on average I believe globally the participating labs are producing about six high-quality genomes per week," he says. That's from start of DNA isolation to finished submission to public databases. "So we have a long way to go in terms of scaling with long reads."

Around 2010, error rates with early PacBio instruments ran around 15–20%, says Phillippy, but this has changed dramatically. HiFi reads, which are used in large-scale research efforts such as the HPRC, involve circular consensus sequencing, in which multiple passes around a circular molecule are made. HiFi sequencing essentially performs error correction on previous passes, says Salzberg. Accuracy reaches 99% accuracy and higher.

From these multiple passes around a circular molecule, the instrument builds a consensus sequence with a hidden Markov model. PacBio collaborated with Google to explore how its tools for genomic analysis, machine learning and algorithm development might further enhance HiFi read accuracy, increase data yield and increase computation speed. Together the teams developed a deep-learning-based approach called Deep Consensus, which has a transformer architecture. Transformers played a role in, for example, developing DeepMind's AlphaFold 2 protein structure production platform. Deep Consensus helps to correct errors, especially in the genome sections that are more difficult to sequence and assemble, and it's now built into the PacBio platform.

The company basically gives away its devices. "You only buy consumables," says Gordon Sanghera, CEO of Oxford Nanopore Technologies. Users buy flow cells and reagents.

For the past decade, Salzberg and his team have been assembling tree genomes, and to do so they combine Illumina short-read sequencing and ONT nanopore sequencing. "That's currently our recipe," he says. ONT was once an error-prone technology that has been tremendously improved, says Phillippy. It's now, in his view, an instrument "taken seriously" for structural variant detection, single-nucleotide variant calls and disease diagnostics, too. Salzberg agrees on the rise in accuracy levels. The technology is well-suited to the tree genomes he works on, which can be ten times the size of the human genome and contain within these perhaps 30 gigabases, repeats of different lengths.

Not so long ago, human genome assemblies were built from reads of 55–75 base pairs. "They're terrible assemblies," says Salzberg. To his recollection, the Solexa machine generated reads of 25 base pairs. The first human genome assembled with Illumina technology, which acquired Solexa, was fragmented and based on 54-base-pair reads. To get to the current human genome assembly, labs have been wrestling with long repetitive regions, attacking gaps at the telomeres and around centromeres, which has millions of repetitions and a complex structure. "That's why the HiFi reads are necessary," and why it's used, for example, by the T2T Consortium.

Wish-list item: save time

Some ONT users wish the company would change its products and especially its software less frequently. Says Sanghera, the rule book for disruptors is to "release software, hardware regularly." But, alongside the "innovators" and early adopters in genomics who crave the newest tools and rapid changes, as ONT's platform has matured, additional, applications-focused customers have emerged. To address this group, the company is setting up the Q-Line, for which things do not change as frequently, he says. This two-pronged product development approach is still "a work in progress."

Readily producing an error-free assembled genome that runs "telomere to telomere" is a "big dream" for many biologists, says Zhang. And it's made possible with long reads that ease genome assembly. Long-read sequencing, he says, "also allows us to capture the full set of transcripts with all alternative splicing forms."

Yet assembling "a T2T" genome remains time-consuming. For instance, it means manually correcting artificial structural variants that may have been generated. It's also computationally expensive. For long-read sequencing, faster, longer and cheaper are on his wish list, says Zhang.

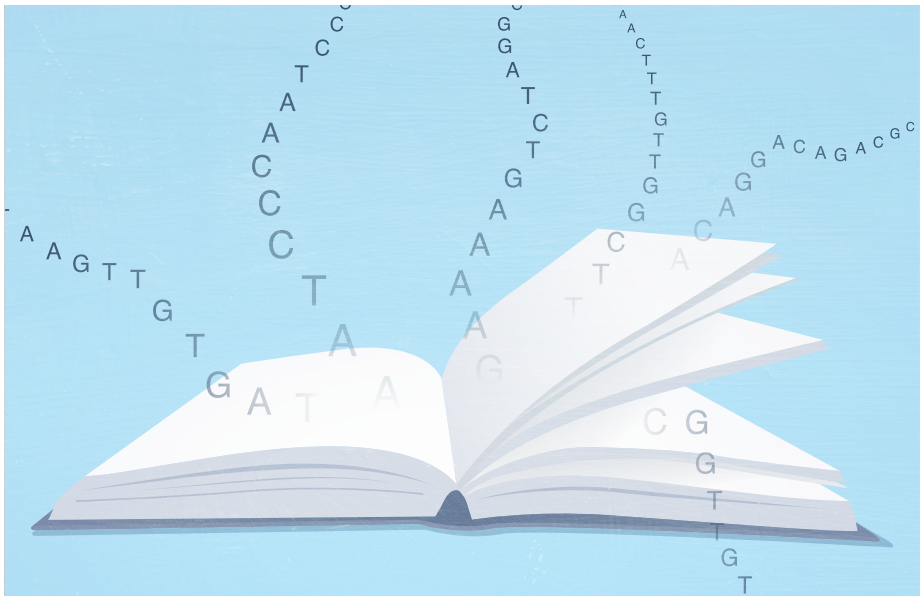
New methods, says Korlach, and not just those in sequencing, are generally more expensive than when they become more widespread and commoditized. He is glad to see how much of biology is being revealed because labs can now use long reads to resolve genomic regions they could not previously. One important advance among many, he says, has been the ability to resolve the six gigabases of the human genome such that one can separate haplotypes and study maternal and paternal alleles.

Imagining the future

Around 2010, he began working with long-read sequencing with early PacBio instruments, says Ohio State University researcher Kin Fai Au, who has been vice chair of research in the department of biomedical informatics and is moving to the department of computational medicine and bioinformatics at the University of Michigan. As a postdoctoral fellow at Stanford University, he was near PacBio's headquarters in neighboring Menlo Park. He also had access to ONT's early access program, when the instruments' error rate ran as high as 30–40%, he says, and the reads could not be aligned. These days, he says, accuracy rates are 95–96% and rising.

As a graduate student, Au had worked on software for Illumina short-read-based RNA sequencing to predict gene isoforms of a transcript, but it didn't work well. When

CREDIT: OXFORD NANOPORE TECHNOLOGIES



Long reads are a way for many labs to tackle gnarly genome sections such as complex genome rearrangements and the many types of repetitive elements.

the team applied their method to PacBio SMRT-sequencing reads, they detected thousands of isoforms. At the time, he recalls, PacBio's SMRT-sequencing error rate was high, which has changed.

For a long time, he says, RNA sequencing meant one needs to reverse transcribe the RNA to cDNA to sequence it. ONT's technology was the first to enable direct RNA sequencing[9]. RNA expression data hold great value because of the way they can capture differences across cell and tissue types. Beyond RNA abundance, he and others are using long reads to assess the epitranscriptome, RNA modifications and RNA structure. For such tasks it's helpful to have direct RNA sequencing, says Au.

Unlike DNA, RNA has no amplification method, and the sample sizes for RNA sequencing are a challenge. As the technology continues to change, so too will requirements for input material, says Au. Labs are also working on protocols to maintain DNA molecules longer than one million base pairs. The longer the read, the easier the assembly.

Beyond DNA and RNA sequence, says Au, labs want to capture many other facets: long-range and three-dimensional interactions, repetitive sequences of many sizes, transposable elements, epigenetic alterations such as 5-methylcytosine, histone modifications. They seek chromatin accessibility data to assess which transcription factor binds where.

Many '-seq' named methods have emerged and, says Au, and more will emerge related to long reads as "one of the future efforts in the community." Long-read RNA-seq, long-read ATAC-seq and many others will lead to new opportunities in bioinformatics given the new types of information and data they will yield.

Au is excited about the technology to come. "Science always proceeds in some ways that we never can imagine," he says. He is intrigued by Illumina's long-read technology, formerly called Infinity and renamed to Complete Long-Read technology. According to the company, it generates reads longer than 30 kilobases and can work with 50 nanograms of DNA. Both ONT and PacBio cover the long-read sequencing market, says Au, and PacBio is more expensive for users than ONT. In his view, the two companies and others might not take over all of Illumina's market share, but they will keep making inroads because, in his opinion, "long-read sequencing is the future."

Vivien Marx ✉
Nature Methods.
✉e-mail: v.marx@us.nature.com

Published online: 12 January 2023

References

1. Nurk, S. et al. *Science* **376**, 44–53 (2022).
2. Wang, T. et al. *Nature* **604**, 437–446 (2022).
3. Jarvis, E. D. et al. *Nature* https://doi.org/10.1038/s41586-022-05325-5 (2022).
4. Liao, W.-W. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.09.499321 (2022).
5. Manrai, A. et al. *N. Engl. J. Med.* **375**, 655–665 (2016).
6. Ko, B. J. et al. *Genome Biol* **23**, 205 (2022).
7. Levene, M. J. et al. *Science* **299**, 682–686 (2003).
8. Eid, J. et al. *Science* **323**, 133–138 (2009).
9. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. *Nat. Biotechnol.* **39**, 1348–1365 (2021).