Playing a long game

Nanopore sequencing's early adopters are pushing the limits of what can be achieved with ultra-long DNA reads, and they are also finding innovative ways to apply this technology to other biological questions.

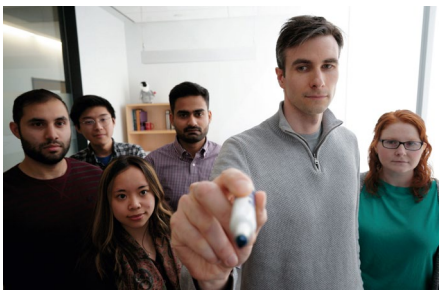
Michael Eisenstein

This past February, Adam Phillippy of the National Human Genome Research Institute showed the genomics community something it had never seen before: a complete human chromosome. It's no secret that the human genome sequence published in 2000 was merely a fragmented rough draft, and nearly 20 years later, the genome remains incomplete. Phillippy, Karen Miga of the University of California at Santa Cruz (UCSC), and their colleagues in the international Telomere-to-Telomere Consortium (T2T) now aim to rectify that—and the gapless X chromosome presented at this year's Advances in Genome Biology and Technology (AGBT) meeting is a critical first step.

That work was also a high-profile demonstration of the capabilities offered by nanopore sequencing, which can generate vast sequence reads spanning hundreds of thousands of bases—long enough to allow scientists to forge through the dense forests of repetitive sequence elements that have historically confounded assembly and analysis. "We've collectively had an interest in developing long reads to push into these 'dark regions' of the genome," says Miten Jain, a genomics researcher at UCSC and collaborator in the T2T effort who receives funding from Oxford Nanopore Technologies (ONT) through his group leader Mark Akeson.

Genomics researchers have been intrigued by the sequencing strategy developed by ONT since it first hit the market in 2014. But the technology also differs radically from other sequencing platforms, and as a relative newcomer on the market, the platform has faced stiff competition from short-read titan Illumina, as well as from long-read rival Pacific Biosciences (PacBio), which Illumina is now moving to acquire. "We used to have to convince people that nanopore sequencing works, and that it can be applied at a high throughput and at a large scale," says Matthew Loose, a developmental geneticist at the University of Nottingham.

But as ONT's platform has grown more mature and flexed its muscles in the realm of genome assembly and analysis, the system's early adopters have demonstrated that its distinctive design can be exploited not only to map uncharted chromosomal terrain, but also to obtain unprecedented insights in areas like transcriptomics and epigenomics. "I don't think people have yet fully exploited the neat aspects of this technology," says Loose.



Jared Simpson's research team at the Ontario Institute for Cancer Research. Credit: J.P. Moczulski

Error correction

Initially, the most remarkable aspect of nanopore sequencing was that it worked at all. ONT grabbed the spotlight at 2012's AGBT meeting, when chief technology officer Clive Brown introduced the MinION, a thumb-drive-sized widget priced at less than $1,000 that could generate up to 150 megabases of DNA sequence. Not only was the tiny device a far cry from existing benchtop instruments, but the underlying technology seemingly bordered on science fiction. Each MinION flow cell contains thousands of membrane-embedded protein pores; DNA strands are captured and threaded through the pore, and the sequence is deciphered on the basis of changes in electrical current across the pore produced by the passage of various combinations of nucleotides.

"My mind was blown away that you could do this," says Martin Smith, Genomic Technologies Group leader at the Garvan Institute of Medical Research in Australia. "I thought it was going to be a pipe dream." Smith was among a small cohort who got their hands on the first-generation instrument through ONT's MinION Access Program (MAP). Among these early users, the initial reaction was generally pleasant surprise—tempered by clear recognition of the system's limitations. "I don't think I'll ever forget the first time we ran a MinION sequencer and got back a read and it actually was somewhat closely related to what we were looking for!" says Loose. But the instrument's performance was also spotty and inconsistent. Wouter De Coster, a bioinformatician at the University of Antwerp, recalls spending a full day each on library preparation and sequencing, only to get reads with error rates as high as 30–40%, if they worked at all. "It was absolutely hit or miss, and more often a miss than a hit," he says.

The technology's capabilities have improved considerably in the ensuing years, with multiple improvements to the pore and flow-cell chemistry. One of the biggest leaps came in 2016, when ONT substituted the error-prone pore from the early-access MinION, termed R7.3, with a newer pore, R9.4, which was engineered from the *Escherichia coli* protein CsgG. "We were reaching something like a tenfold improvement in sequencing throughput at that time," says Wigard Kloosterman, who participated in the MAP as a geneticist at UMC Utrecht and is now chief scientific officer at biotech startup Cyclomics. "And accuracy also got better, with an error rate of around 11%." Earlier this year, the company announced the launch of R10, which it claims represents an entirely novel pore structure. Early data suggest that R10 may help to overcome one of the most persistent problems with nanopore sequencing, wherein 'homopolymeric' sequences containing consecutive repeats of a particular nucleotide create a slurred signal that can be difficult to decipher.

Jean-Marc Aury, who leads a team of bioinformaticians within the Genomics Institute at the French Commission for Atomic Energy and Alternative Energies (Genoscope), is among the early users of R10 and notes that his team has observed some trade-offs. "The error rate of individual reads is higher than the R9.4, but the errors are more random—so the

Martin Smith of the Garvan Institute of Medical Research (left), with team members James Ferguson and Hasindu Gamaarachchi. Credit: K. Recsei

consensus is of higher quality," says Aury. The differences between these two pores could prove complementary if the two were used in combination, a possibility now being explored by Jared Simpson at the Ontario Institute for Cancer Research, who receives research funding from ONT. "They're going to give you different signals so that you might be able to pick something up with one pore that you can't pick up with the other," he says. "The strengths of the two could reinforce each other."

Sorting through the squiggles

These hardware advances have been paralleled by new computational tools developed both internally and by an engaged community of bioinformaticians. One of the biggest challenges early adopters faced was the fact that nanopore data looked so different from what was being produced by market leader Illumina, and required an equally distinctive toolbox. The raw output from a MinION run consists of fluctuations in current that are subsequently transformed into 'squiggle' plots, which can then be translated into a more familiar string of nucleotide sequence with specialized base-calling software.

Early base-callers were relatively error-prone, but beginning in 2017, these programs began using neural network algorithms that could boost read-level accuracy to well over 80%. Subsequent iterations of ONT's base-calling software, such as the Scrappie algorithm, have also helped to mitigate the unwanted effects of the homopolymer problem. "If you have the same base repeated multiple times, then you won't see a shift in ionic current—you'll just get this 'monotone,'" explains Jain. "This algorithm knows how long that monotone was and roughly how fast the strand is being processed, and it uses the speed and time to estimate the number of bases." The results are not perfect, but they eliminate many accidental 'deletions' that would otherwise arise from misinterpreted homopolymers.

Scrappie works by analyzing the raw data rather than the processed squiggles, and other software tools have also leveraged these unmanipulated measurements to further improve sequencing accuracy. For example, Simpson developed a tool called Nanopolish to help complete the first nanopore-only assembly of a full bacterial genome in 2015. This software uses raw current data to correct errors in the consensus produced after assembly of the various overlapping reads generated in a sequencing run. "It was really about getting this deep understanding of what affects the signal and modeling that to get the most out of the sequencer," says Simpson.

Nanopolish is still widely used, although it can be computationally intensive to run, and Simpson notes that ONT has released an alternative polishing tool called Medaka that can achieve greater accuracy with less time and effort. "You don't need a high-performance computing system—you can just do it on a laptop," he says. More generally, this consensus analysis step offers a critical opportunity to overcome read-level errors, and progress in this area has propelled nanopore sequencing onto near-equal footing with its competitors, enabling greater than 99% accuracy. "There's kind of a fixation on raw read accuracy and that's not always an important question," says Loose. "It's more important whether you can get a consensus."

Bigger and better

Even against this backdrop of technology development, nanopore sequencing was still broadly viewed as something of a 'niche' tool until a few years ago. The ultra-portable MinION proved to be a powerful tool for field applications such as tracking the Zika outbreak[1] and surveying environmental samples in the remote Antarctic[2], but clinical research and de novo genomic assembly remained the domain of Illumina and PacBio technology.

A big shift in perception occurred in April 2017, when two research groups led by Loose[3] and Kloosterman[4] independently demonstrated that nanopore can also tackle entire human genomes. This was no mean feat, however, and represented a proof of concept rather than a viable alternative to existing whole-genome sequencing strategies. "Our assembly took about 150,000 CPU hours and would have cost us about $30,000 at the time if we were to run it on Amazon Web Services," says Jain, who was first author on the study by Loose and colleagues. And although the throughput and reliability of the MinION

Matthew Loose of the University of Nottingham. Credit: University of Nottingham

were greatly improved after three years, the tiny devices were not an ideal match to a project of this scale. For example, Kloosterman estimates that his team spent half a year sequencing on 122 flow cells to achieve 16× genome coverage.

Scaling up has gotten simpler since then. After two years of early-access testing, ONT has released the PromethION, an instrument for high-throughput sequencing. "We were able to do six human genomes across two flow cells each," says Loose. "And we got coverage ranging from 40 to 85 times in four days of sequencing." Early users have been impressed but note that the outcome of an experiment is very much dependent on the quality of sample preparation. "With a good sample and a good flow cell, 100 gigabases per run is definitely feasible," says De Coster. "But if your DNA quality is very bad, it's going to be 30 gigabases or less." Current versions of the instrument can run either 24 or 48 flow cells in a single experiment, enabling users to collect several terabases per experiment on a fully loaded instrument.

This puts PromethION in the same ballpark as the other leading sequencing platforms in terms of throughput, although the competition remains fierce. For example, market leader Illumina reports that its top-of-the-line NovaSeq 6000 instrument can routinely generate up to six terabases of sequence data from two flow cells over the course of two days, with the output comprising short paired reads spanning 100–150 bases each. And on the long-read front, PacBio has stated that its Sequel II instrument can generate up to 320 gigabases per sequencing cell within 30 hours, generating reads spanning tens or hundreds of kilobases and with a mean accuracy of more than 99% per read.

Nanopore users have also benefited from the surge in development of efficient genome-assembly software for PacBio instruments, which have become a popular choice for de novo genome assembly. Many

Jean-Marc Aury of the Genoscope lab at the French Commission for Atomic Energy and Alternative Energies. Credit: A. Couloux

of the most popular tools for fitting long reads into larger contigs, such as MiniMap2[5] and Canu[6], are essentially platform agnostic, and can be configured to deliver best results based on the characteristics of the data generated by the different systems. "I would say that the long-read toolbox is pretty unified right now," says Winston Timp, an engineer specializing in sequencing technologies at Johns Hopkins University.

In principle, read lengths are limited only by the size of the DNA fragments that can be delivered intact to the pore. This gives nanopore technology a major edge in terms of building ultra-long-range sequence assemblies without the gaps associated with short-read contig building. "We've been able to sequence entire yeast chromosomes—that's around 200 to 300 kilobases," says Aury. Jain notes that fragments of this scale proved invaluable in terms of boosting the quality of the nanopore human genome sequence, essentially doubling the contiguity of their group's assembly. Today, there is friendly competition among users to see who can achieve the longest single read. Smith's lab was the first to cross the one-megabase mark in late 2017, and Loose and his collaborator Nicholas Loman at the University of Birmingham have received funding from the Wellcome Trust for a 'long-read club' to develop strategies for pushing the outer limits of length.

Sequencing at this scale is no mean feat. There are kits for isolating large DNA fragments—for example, Smith and colleagues used a technique developed for a genomic mapping platform from BioNano Genomics, which was designed to precisely position sequence reads relative to each other even over considerable distances. However, these long strands behave differently than other preps. "It's so viscous—almost like a gel plug—that just getting it on the flow cell is probably the trickiest part," says Smith. Nanopore sequencing is also very sample-hungry and finicky about the quality of the preparation—problems that become exacerbated when one is deliberately hunting 'whales', as ultra-long reads are colloquially known. But the results continue to astonish—in late 2018, Loose described a 2.3-megabase sequence[7] that was so long that the base-caller mistakenly divided it into 11 reads, and users keep vying for new records. "We're seeing really impressive results from people on Twitter," says Loose.

Fill in the blanks

Nanopore is particularly well suited for exploring structural variations in complex genomes. "A great example are these LINE retrotransposon elements in the human genome," says Kloosterman. "These are around 6 to 8 kilobases, and if you have 20-kilobase reads you can see them from beginning to end." These would be nearly impossible to reconstruct with 250-base reads, and top-notch base-level accuracy is not essential for this sort of mapping.

De Coster and colleagues have been using the PromethION to systematically identify repetitive elements and other sources of structural variability in the human genome[8], with an eye toward detecting risk factors for neurological disorders. "We have seen that we can expect around 27,000 structural variants larger than 50 nucleotides in the human genomes, and that they contribute more than single-nucleotide polymorphisms to the variation between humans," says De Coster. In Kloosterman's experience, nanopore can achieve near-perfect sensitivity for large rearrangements, such as the chromosomal abnormalities typically seen in cancer genomes, but falls short with small insertions or deletions and is still not ideal for single-nucleotide variants.

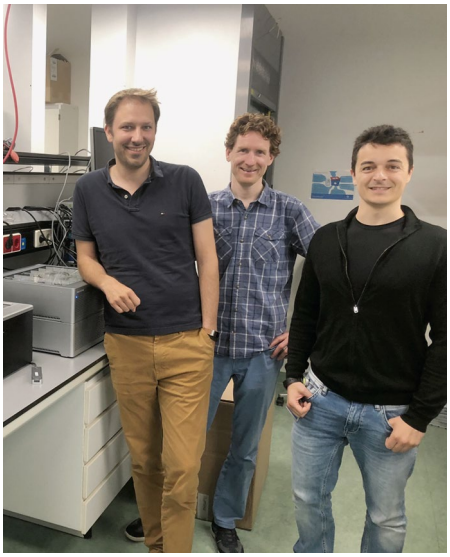
Most researchers interested in reconstructing full genomes therefore combine nanopore with other technologies that can further improve assembly contiguity and accuracy. For example, Aury's lab has found nanopore to be a good fit for studying complex and often highly polyploid plant genomes, but it cannot do the whole job alone. "It was still insufficient to get the chromosome-scale organization, and so we have been using BioNano optical mapping," says Aury. "You also still need Illumina data to polish the consensus."

The T2T team has likewise employed a multipronged approach. In an initial pilot effort last year, Miga and colleagues used a set of 200-kilobase reads to build a sequence scaffold for the never-before-mapped centromere of the human Y chromosome[9]. "We were able to assemble what ended up being a 315-kilobase centromere," says Jain, who was lead author on the publication. "And it's worth noting that when we started doing this tiling, we did not know what length it would be." But once the nanopore foundation was laid, they used Illumina short-read data to polish the final assembly. Moving forward, the T2T initiative will combine data from PromethION, PacBio, and Illumina, as well as from long-range mapping technologies like BioNano's, as demonstrated with their recently completed X chromosome sequence.

Such all-out genomic assaults are not practical for routine clinical use, but a few groups are exploring targeted sequencing methods that can leverage nanopore technology for high-accuracy sequencing of single-nucleotide variants. Kloosterman's company Cyclomics, which he cofounded with his long-time collaborator Jeroen de Ridder, has developed a strategy for capturing and circularizing short DNA fragments and then enzymatically replicating them over and over to yield long strings of repeats. These can then be sequenced to obtain a highly accurate consensus. By implementing this with the MinION, Kloosterman hopes to deliver a low-cost, portable technique for performing 'liquid biopsies' for cancer based on mutation detection in circulating tumor DNA. Other targeted sequencing strategies can potentially be applied directly to raw DNA samples. For example, Timp and colleagues have used the genome-editing enzyme Cas9 to achieve selective cleavage at genomic sites of interest[10]. These cleaved ends can then be 'marked' for preferential sequencing, which allows for several-hundred-fold enrichment relative to results from noncleaved sequences in the same sample.

Opening the floodgates

As nanopore grows more competitive for DNA-sequence analysis, researchers also are finding that these tiny holes are equally well suited for studying a variety of other biomolecules. "The nanopore doesn't care about what you're putting into it," says Timp. For example, he and Simpson have used nanopores to map the epigenetic marks associated with DNA methylation. In the



Cyclomics founders Jeroen de Ridder, Wigard Kloosterman, and Alessio Marcozzi. Credit: E. Kuijk, UMC Utrecht

early days of MinION, modified DNA bases such as 5-methylcytosine (5-mC) had a confounding effect on current readings that confused the base-calling software. But this noise can be turned into news if the software can recognize the patterns that arise from modification and discriminate them from normal bases. Simpson and Timp manufactured a wide variety of DNA sequences that incorporated 5-mC in different positions and sequence contexts, and then they trained Nanopolish so that it could consistently discriminate these same patterns in the wild[11].

They are continuing to collaborate on the identification of other naturally occurring DNA modifications, and the resulting data could ultimately be incorporated into future base-calling software for use in routine sequencing experiments. However, this feature of nanopore sequencing can also be exploited to study other features of chromosomal biology. For example, Timp's team has treated DNA samples with a methyltransferase enzyme that preferentially marks sequences in relatively open stretches of chromatin—generally associated with actively transcribed genes—and then detected these modification patterns via nanopore sequencing[12]. "We found that we can phase chromatin states and methylation on individual molecules and identify imprinted genes with allele specificity," says Timp.

One can also thread RNA strands through the same nanopores, which allows for direct analysis of intact transcripts without the need for enzymatic conversion to cDNA, a process that can potentially introduce biases into the transcriptomic data. "You get a native measure of the full-length RNA, which means you get all the splice junctions," says Jain, who recently collaborated on the transcriptome-scale nanopore sequencing of mRNA from a human cell line[13]. "We got 10 million RNA reads, where the longest was 22 kilobases long and spans 116 exons." Timp, who led that study, notes that these sequences also include the full-length poly(A) tail that terminates every strand—a structure with an important regulatory role in mRNA stability and translation that is typically lost in cDNA-based transcriptomic methods. Exact quantification can be difficult, given the platform's problems with homopolymers, but base callers that track transit time through the pore can mitigate this problem.

The generation of transcriptome-scale nanopore data is more labor intensive than with cDNA-based short-read RNA-seq protocols. "The input material requirements are very high, so you need a lot of RNA," says Aury, "and the throughput is still very low." The error rate also remains higher for individual RNA reads than for cDNA sequences obtained with Illumina or PacBio. And as with DNA, this is further exacerbated by chemical modifications that can befuddle base callers, although the problem is far more dire for RNA. "For DNA, there's only a dozen or so modifications," says Smith. "But RNA—especially ribosomal RNA or tRNA—is known to have hundreds of modifications."

This latter problem is also an opportunity, however, if base callers can be trained to recognize and interpret these modifications. A recent preprint from Eva Maria Novoa, at the Center for Genomic Regulation in Barcelona, and colleagues demonstrated the feasibility of this for N^6-methyladenosine, one of the commonest modifications in mRNA[14]. "We looked at consistent base-calling errors from these modifications, and it seemed to work quite well," says Smith, who collaborated on the study, "but it's still a challenge." This challenge will only grow more profound as researchers attempt to train software that can recognize the error profiles not just from individual, distinct modifications but from a daunting myriad of combinations of altered bases. "In our group, we call this a 'ten-year endeavor'," says Jain.

And although ONT has developed a formidable head start in the implementation of nanopore-based biomolecular analysis, other companies and academic researchers are also exploring the technology's potential. For example, Roche has been quietly developing a protein nanopore-based technology acquired from startup Genia as a potential tool for clinical diagnostics, and Ontera is working on a handheld device that uses solid-state nanopores that can potentially identify nucleic acids, proteins, and even pathogens present in a given sample. And at the University of Washington, Jens Gundlach's team has been using nanopore proteins derived from the microbe *Mycobacterium smegmatis* to study the dynamic interplay between nucleic acids and various 'motor proteins', such as the helicase enzymes that unwind DNA[15].

For the early users who helped nanopore sequencing to find its footing, these various 'next-wave' applications are injecting new excitement into the field—and sparking the imagination as to what opportunities might lie further down the nanoscale rabbit hole. Protein sequencing is at the top of Timp's agenda, and he notes that a handful of academic studies have already begun clearing a path in this direction. "I'm not saying that amino acids would be easy, but think about how painful it is to do mass spectrometry," he says. "If this is something a nanopore could accomplish, that would be amazing." ❐

Michael Eisenstein
Science writer, Philadelphia, PA, USA.
e-mail: michael@eisensteinium.com

Published online: 30 July 2019
https://doi.org/10.1038/s41592-019-0507-7

References

1. Faria, N. R. et al. *Genome Med.* **8**, 97 (2016).
2. Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y. & Tighe, S. W. *J. Biomol. Tech.* **28**, 2–7 (2017).
3. Jain, M. et al. *Nat. Biotechnol.* **36**, 338–345 (2018).
4. Cretu-Stancu, M. et al. *Nat. Commun.* **8**, 1326 (2017).
5. Li, H. *Bioinformatics* **34**, 3094–3100 (2018).
6. Koren, S. et al. *Genome Res.* **27**, 722–736 (2017).
7. Payne, A., Holmes, N., Rakyan, V. & Loose, M. *Bioinformatics* **35**, 2193–2198 (2019).
8. De Coster, W. et al. *Genome Res.* **29**, 1178–1187 (2019).
9. Jain, M. et al. *Nat. Biotechnol.* **36**, 321–323 (2018).
10. Gilpatrick, T. et al. Preprint at https://www.biorxiv.org/content/10.1101/604173v2 (2019).
11. Simpson, J. T. et al. *Nat. Methods* **14**, 407–410 (2017).
12. Lee, I. et al. Preprint at https://www.biorxiv.org/content/10.1101/504993v2 (2019).
13. Workman, R. E. et al. Preprint at https://www.biorxiv.org/content/10.1101/459529v1 (2018).
14. Liu, H. et al. Preprint at https://www.biorxiv.org/content/10.1101/525741v1 (2019).
15. Craig, J. M. et al. *Nucleic Acids Res.* **47**, 2506–2513 (2019).