WHEN TWO IS BETTER THAN ONE

Looking for the perfect sequencing instrument for your next experiment? As Jeffrey Perkel finds out, you might just need more than one.

Elaine Mardis and Richard Wilson, directors of the McDonnell Genome Institute at Washington University, in front of an Illumina HiSeq 2000 sequencer. Credit: Bob Boston.

Hardly a day goes by without the announcement of a newly sequenced genome. Whatever the DNA source—animal or plant, fungus or bacterium—genome sequencing has become so commonplace, the feat no longer amazes. It may shock younger readers to learn that not so long ago completed genomes merited journal covers and press conferences.

Today, the source of our groundswell of genomic data, next-generation sequencing (NGS), is regarded as largely routine. And why not? Prices have plummeted, while access has skyrocketed. Illumina's newly launched MiniSeq, priced at just \$49,500, promises 7.5 Gb of data per run—not enough to decode a human genome, but definitely enough for targeted sequencing applications (see “New instruments, more options” sidebar). At the University of Edinburgh, Scotland, undergraduate honors students can even take a course where Illumina NGS technology and advanced bioinformatics are used to decipher bacterial genomes.

Perhaps lost in all this familiarity and access is the recognition that, a decade after the first NGS instruments went online and more than a decade since the completion of the Human Genome Project (HGP), sequence assembly remains a challenge. Even the small 5.2-Mb *Bacteroides fragilis* genome—the subject of the University of Edinburgh's 2013 undergraduate course—confounds, thanks to repetitive and mobile regulatory elements called “shufflons,” and “invertible promoters.”

A recent editorial in *Nature Genetics* crystallized the problem when it likened some particularly challenging genomic regions to “the disrupted runs of shuffled playing cards from multiple decks or the compound folded layers of butter and dough in Danish pastry” (1). Yet researchers are learning there is a way to untangle those regions by blending different, orthogonal sequencing technologies.

Nurseries of evolution

In theory, bacterial genomes should resolve into single contigs. Gene-rich and relatively small, there's little extraneous genetic material available to trip up sequence assembly tools. According to Judith Risse, a bioinformatics specialist at Edinburgh Genomics who processed the data for the Edinburgh undergraduate course, researchers usually don't get that lucky; a small number of contigs is routine. Still, when the Edinburgh students turned their attention to *B. fragilis*, they got a mess.

“If you try to assemble the *B. fragilis* genome from Illumina short reads, it will not assemble into a single contig, because it has small bits of DNA that can change orientation within the genome,” Risse explains. Feed those snippets into an assembly tool, and the software effectively throws up its hands. Instead of a single contig, the students in Edinburgh's Junior Honors “Genomes and Genomics 3” course were faced with more than 20. And eukaryotic chromosomes pose a far greater challenge.

Evan Eichler is a professor of genome sciences at the University of Washington School of Medicine and a Howard Hughes Medical Institute Investigator. He studies regions of the genome often overlooked in traditional genome sequencing projects.

“I am a child of the genome era, and I study both human disease and evolution from the perspective of looking at the genome first.”

Eichler concentrates on regions of genomic instability. Flanked by long segmental duplications, these regions can misalign during meiosis, leading to either deletion of the intervening sequence, or duplication. “Think of these regions as land mines,” Eichler explains. “Once you create this architecture where these duplications are separated by lots of unique

sequence, that interspersed architecture then predisposes to higher rates of copy-number variation, creating instability.”

Eichler says there is good evidence that these regions represent nurseries of gene evolution, and they may even explain some of the biology that makes humans, well, human. The *SRGAP2* gene, for instance, is present in one copy in many higher primates, but appears in triplicate on human chromosome 1, and the two additional, truncated copies could play a role in neocortex development and spine density in the brain.

But these complex regions can also lead to disease, such as intellectual disabilities and autism. 17q21.31 is a large region that in about 20%–30% of individuals of Northern European or Middle Eastern descent is inverted relative to the reference genome. While the change correlates with increased fecundity and recombination in women carriers, children of those carriers have a much higher rate of genetic microdeletions associated with neurological deficits—a constellation of genetic defects called Koolen-DeVries Syndrome. And it’s a region the HGP, which relied on Sanger sequencing of bacterial artificial chromosome (BAC) clone libraries, and most subsequent genome-sequencing efforts, got wrong.

Michael Schatz has developed algorithms for blending sequence data from different instruments. Credit: Cold Spring Harbor Laboratory.

Painstaking work

When researchers sequence a human genome today, they employ short-read Illumina sequencing technology for shotgun-sequencing. This strategy—in which a genome is sheared into millions of small pieces, each of which is then read out in parallel—is relatively inexpensive and ubiquitous, producing from hundreds of gigabases to even terabases per run. Yet as those data come in the form of relatively short read pairs measuring ~100 bases in length per read, specialized computer algorithms are required, both to align each read pair to the genome’s unique sequences (such as exons) and to identify variation relative to the human reference genome. What these algorithms cannot do at the moment, is assemble the genome from scratch (i.e., *de novo*). As a result, an alignment-based approach using short-read data fails to capture the large-scale structural differences (as well as many simpler, smaller-scale variants) that make each genome unique; however, there are other methods that can.

Pacific Biosciences instruments sequence DNA by watching as a single DNA polymerase molecule incorporates fluorescently labeled nucleotides in real time, while Oxford Nanopore instruments record electrical changes that occur across a membrane as a piece of DNA is fed through a pore. Though the technologies differ in mechanics, speed, price, and accuracy, both deliver reads far longer than Illumina, Ion Torrent, or even the gold standard of the HGP, Sanger sequencing. PacBio reads average more than 10-kb in length, with some reaching out to 60 kb and beyond; Oxford Nanopore’s MinION reads are shorter on average, but have the potential to extend far longer: One recent study of the *Saccharomyces cerevisiae* genome (2) recorded an average read length of 5548

Jonas Koriach, Chief Scientific Officer at Pacific Biosciences, says transcriptome analysis can require multiple sequencing approaches. Credit: J. Koriach; Pacific Biosciences.

bp and a maximum length of 191,145 bp; another study of 3 bacterial genomes reported an average read length of ~2000 bp and a maximum length of 98,366 bp (3). (Oxford Nanopore declined a request for an interview for this article.)

These technologies are relatively expensive per base compared to Illumina, have a higher “raw” (that is, uncorrected) error rate, and produce far fewer reads (up to 1 million in the case of PacBio’s newest instrument, the Sequel; tens of thousands in the case of the Oxford Nanopore MinION). As a result, researchers using these technologies typically have to do more sequencing runs and sequence more deeply than they otherwise might, says Elaine Mardis, co-director of the McDonnell Genome Institute at Washington University.

Still, if researchers can swing it, they can, by combining short- and long-read strategies, resolve regions that cannot easily be untangled with short-read technology alone.

Which brings us back to chromosome 17. According to Eichler, the original HGP assembly contained gaps. But in the case of 17q21.31, it also harbored a 900-kb inversion and many duplicated genes that were either missing or misassembled. During the HGP, the genome

was sequenced by cloning large genomic segments, 150-200 kb in size, into BACs that were then broken down into smaller sub-clones, which were sequenced, reassembled into individual BACs, and finally assembled into intact chromosomes. The HGP, he says, “became a project of 30,000 individual clones that had to be sequenced and assembled.” The problem was, some duplicated regions could foil this strategy because they were too large to accurately map. Furthermore, if such regions of the genome were structurally very different between people, such as a large inversion, teasing apart the differences became very problematic because the breakpoints mapped precisely to the regions that were too large to accurately map. Today, Eichler’s team blends Illumina and PacBio technologies to tackle those regions at a much higher throughput.

Illumina, Eichler notes, is the “currency of the realm” in terms of sequencing data. The technology has featured in more than 2100 papers, according to the Illumina web site. In this case, the team barcodes hundreds of BACs on 96- or 384-well plates and sequences them all en masse via a sequencing strategy called Nextera, producing a set of BAC “fingerprints.” The team can identify those BACs that define a tiling path across the genomic region of interest, which they then sequence and assemble more carefully with the lower-throughput and more expensive PacBio approach, using high sequence coverage (>90×) for error-correction (4).

The result was a nearly 1.3-Mb assembly built from 8 BACs, 99.994% identical to the previously generated Sanger alignment but costing significantly less—about \$625 per BAC compared with \$4,000–\$5,000 using Sanger sequencing.

Eichler—who previously served on PacBio’s scientific advisory board—says this strategy has the capacity to resolve “huge gaps in our understanding of human genetic variation and as a result add more than a piece or two to the puzzle of missing heritability.” Applying the

Sequel, the latest generation sequencing instrument from Pacific Biosciences, was launched in September 2015. Credit: Pacific Biosciences.

New instruments, more options

Pacific Biosciences and Illumina have both launched new sequencers in recent months. The PacBio Sequel, which debuted in September 2015, addresses three key limitations of the company’s older PacBio RS II, says Chief Scientific Officer Jonas Korlach—size, price, and throughput. The Sequel is about one-third as large and half as expensive as the RS II, yet it offers seven times the throughput. Whereas the RS II offered 150,000 “zero-mode waveguides” (i.e., reaction wells), Sequel has 1 million.

In January, Illumina added the \$49,500 MiniSeq to its sequencer lineup. The MiniSeq, says Diane Lince, associate director of product marketing, “is an easy-to-use option for targeted sequencing applications that is packaged into one nice, affordable box.”

With an output half that of the MiSeq (up to 25 million 2×150 bp reads, for 7.5 Gb total), the MiniSeq represents the new entry point into Illumina’s sequencer portfolio. “We’ve designed this instrument to really reduce the barrier to entry for customers that are looking to enter the next-generation DNA or RNA sequencing space,” Lince says. Instruments are slated to begin shipping during the first quarter of 2016.

—J.P.

method to a complex region of chimpanzee chromosome 17 produced a 766-kb assembly that added 241 kb to the chimpanzee reference genome and anchored an additional 525 kb that had been previously assembled into contigs but not mapped to a chromosome.

“It’s painstaking work,” says Mardis, whose center has assisted in some of these studies. “It’s worth it for solving the region, so that people can better understand the intervening genes in the context of genetic studies. But it’s a tough go, requiring unique skill sets, and it’s not fun, which is why there are so

few groups that really pursue it as avidly as we have.”

Data blending

Though Eichler’s team uses both Illumina and PacBio sequencing, they don’t actually merge the data produced by the two instruments. Rather, they assemble their PacBio reads alone, then use the Illumina data in a “polishing step” to correct for errors. Others, though, have developed computational approaches that blend the data more directly.

Michael Schatz, associate professor of quantitative biology at the Cold Spring Harbor Laboratory, and his colleague Richard McCombie were among the earliest users of Oxford Nanopore’s miniature MinION sequencer. “It’s very exciting to hold a DNA sequencer in your hand. And it is relatively easy to run,” Schatz says. Though suffering some early growing pains, the researchers have seen consistent improvements in the evolving technology. But the raw error rate is still too high. In a paper published in October 2015, McCombie and Schatz used the MinION to sequence the *S. cerevisiae* genome (2). The average read in 2D mode—the instrument’s high-accuracy mode in which both strands of the molecule are sequenced—was just 75% identical to the *S. cerevisiae* reference genome, though “many” reads exceeded 80%.

To circumvent that problem, the team developed an algorithm called Nanocorr,

INDUSTRY LEADING SMALL RNA LIBRARY PREP KIT

GEL-FREE OR LOW INPUT,
HERD TOGETHER.

Randomized Adapters Reduce Sequence Bias

With the industry leading NEXTflex™ Small RNA-Seq Kit v3, you can construct Illumina-compatible small RNA-Seq libraries from as little as 5 ng of total RNA, or go completely gel-free with 100 ng. With the incorporation of randomized adapters, ligation bias is reduced, resulting in more accurate data.

Detect more miRNAs per sequencing run than with competitors' library prep kits.

Request a NEXTflex™ Small RNA-Seq Kit v3 today and increase the accuracy of your small RNA analysis.

BiooScientific.com/SmallRNA

BIOO SCIENTIFIC®
THE NGS EXPERTS™

which aligns short Illumina MiSeq reads to the long Oxford Nanopore sequences, and then computes the consensus. It was a strategy they first had developed to blend PacBio and Illumina reads (5). Applying that approach to their yeast genomic data, accuracy increased from 67% to >97%. On *E. coli* genomic data, the approach produced a single 4.6-Mb contig with >99.99% accuracy, compared to “hundreds of contigs” using Illumina data alone. “There’s a dramatic decrease in the error rate through these algorithms that can, in a computational way, polish the sequences into near perfection,” Schatz says.

Bridging the gap

There are other applications where blended sequence strategies are finding use. Researchers interested in determining which genes are being expressed in the transcriptome can use short-read technology, says Jonas Korfach, Chief Scientific Officer at PacBio. But there is an increasing interest in probing the specific transcript variants of a given cell type, and for that, researchers require a different approach.

“PacBio is really good at illuminating and resolving transcript structure but maybe not so good at [mRNA] counting,” Korfach explains. “And Illumina, or other short-read technologies, are the opposite: They’re really good at counting, and they’re really not that good about resolving the transcript structure.”

In one recent study, Colleen Nelson and colleagues at Queensland University of Technology in Australia, used PacBio sequencing—specifically, a high-accuracy strategy called CCS (circular consensus sequencing)—to identify a novel gene fusion associated with prostate cancer, which they validated using the Illumina reads (6).

Others blend the technologies to address questions of sequence context, says Zamin Iqbal, a group leader at the Wellcome Trust Centre for Human Genetics in Oxford, UK. With some Gram-negative bacteria, for instance, it’s not enough to know that a given sequence is present, researchers also need to know where in the genome (or extrachromosomal plasmid) it is located—information that only can be determined using longer reads.

In the case of *B. fragilis*, combining Illumina and Oxford Nanopore data allowed Edinburgh’s Risse, working with

Mick Watson of the Roslin Institute at the University of Edinburgh, to complete a complex bacterial genome rife with mobile elements, deconvolving the genome from 21 contigs averaging 523-kb in size to a chromosome-sized assembly when adding 8.5× nanopore data into the mix.

Of course, not every application requires long reads, Risse notes. “If you’re only interested in genes, those genic regions will always assemble into long-enough contigs to be useful for expression analysis [and] SNP detection.”

But that may nonetheless be where the field is heading. According to Eichler, most genome sequencing projects capture single-nucleotide variant data for only 95% of the genome, and an even smaller fraction of the structural variation. In 2015, Eichler’s lab, working with Pacific Biosciences, assembled the complete genome of a human haploid hydatidiform mole using only long reads averaging 5.8 kb (7). More than a decade after the “completion” of the HGP, that effort closed 50 existing gaps in the reference genome, identified tens of thousands of new structural variants, and added nearly a megabase of sequence to the genome—data that should prove helpful for future geneticists. “The power of genetics and genomics comes from the fact that we’re comprehensive,” Eichler says. “The minute we stop being comprehensive in terms of our understanding of genetic variation, that’s when we fall short.”

References

1. “Whole genome?” *Nature Genetics*, 47:963, 2015.
2. Goodwin, S., et al., “Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome,” *Genome Res*, 25:1750–1756, 2015.
3. Laver, T., et al., “Assessing the performance of the Oxford Nanopore Technologies MinION,” *Biomolecular Detection and Quantification*, 3:1–8, 2015.
4. Huddleston, J., et al., “Reconstructing complex regions of genomes using long-read sequencing technology,” *Genome Res*, 24:688–696, 2014.
5. Koren, S., et al., “Hybrid error correction and de novo assembly of single-molecule sequencing reads,” *Nat Biotechnol*, 30:693–700, 2012.
6. Tevz, G., et al., “Identification of a novel fusion transcript between human relaxin-1 (RLN1) and human relaxin-2 (RLN2) in prostate cancer,” *Mol Cell Endocrinol*, 420:159–168, 2016.
7. Chaisson, M.J.P., et al., “Resolving the complexity of the human genome using single-molecule sequencing,” *Nature*, 517:608–611, 2015.

Written by Jeffrey Perkel, Ph.D.

BioTechniques 60:56–60 (February 2016)
doi: 10.2144/000114376

www.BioTechniques.com