PACIFIC BIOSCIENCES®

A SMRT® APPROACH FOR FINISHING PLANT AND ANIMAL GENOMES

Extraordinarily long sequencing reads and recent throughput advances are allowing scientists to affordably assemble and close larger genomes, including many plants and animals — even resolving complex repeats or extreme GC regions.

In the days of Sanger sequencing, organisms considered most scientifically important were targeted for high-quality, finished genomes. There was no question that this was an expensive and laborious goal, but it was widely accepted that a finished genome answered questions that no draft assembly ever could. As short-read sequencing technologies became available, they were embraced for their cost-effectiveness — but genome finishing, which was virtually impossible with such short reads, was a necessary casualty.

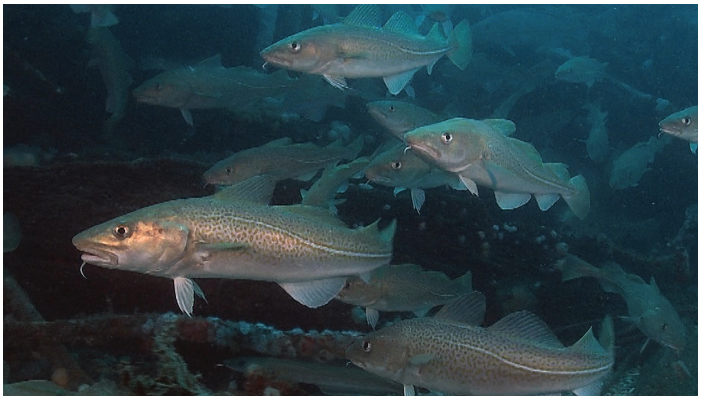
Today, scientists no longer have to compromise between cost and quality. Long-read sequencing from Pacific Biosciences offers the opportunity to upgrade draft genomes created with short reads or to perform *de novo* sequencing and assembly using only PacBio® sequencing.

Finished genomes offer critical information that cannot be gleaned from draft genomes, which may have errors in order and orientation of large contigs, as well as errors generated from repetitive regions. Closed genomes offer a permanent scientific resource for functional or comparative genomic studies, phylogenetic analysis, and a clear view of structural elements and variation.

These advantages of finished genomes are well accepted in the microbiology community, which has applied high-accuracy SMRT Sequencing for routine *de novo* sequencing of microbes, often assembling these organisms into a single contig for less than $1,000 and in less than a day. Steady advances in read length and instrument throughput have now made the PacBio RS II a practical instrument for *de novo* sequencing of large genomes, as well. Biologists focusing on plants or animals are already finding that they can tackle their favorite organisms and get high-quality, rapid, cost-effective sequencing. Here are a few examples.

Cod

Scientists at the University of Oslo's Centre for Ecological and Evolutionary Synthesis (CEES) have used multi-kilobase sequence reads from the PacBio sequencer to produce a dramatically improved genome assembly for the Atlantic cod.



North Sea cod, one of the most endangered species of cod, live in the Dutch North sea.

Lex Nederbragt, a research fellow at the University of Oslo and a member of the Norwegian High-Throughput Sequencing Centre, says that in the last decade or so, "there has been a growing interest in the genomics of this organism." Cod is the most important aquatic species in Norway and other commercial fishery nations. A good genome assembly can aid in finding those regions that influence traits important for disease resistance and growth rates, which may prove crucial for the economic success of the aquaculture industry.

In 2008, Nederbragt and his colleagues Bastiaan Star, Sissel Jentoft, Kjetill S Jakobsen, and others from the CEES-led Cod Genome Sequencing Consortium began a cod genome project using shotgun and mate-pair sequencing on the 454® platform. They mixed in some long-range information from BACs sequenced using traditional Sanger sequencing that resulted in an assembly with thousands of scaffolds and hundreds of thousands of contigs for the 830 Mb genome. Some 35 percent of the bases in the scaffolds were gaps, Nederbragt says, which of course proved quite a challenge for the Ensembl annotation team. To produce an annotated genome, the scientists had to take genes from stickleback and other fish species to reconstruct the missing pieces in cod.

Even while the first cod genome assembly was being published, Nederbragt and his colleagues were casting about for ways to improve it. One challenge was the marked heterozygosity of the wild-caught, diploid cod being sequenced. "Besides the SNPs that you would normally expect, we see large differences over hundreds of bases — sometimes even kilobases — either missing from the other chromosome, or causing differences in regions when we align them," Nederbragt explains. The fish also had many short tandem repeats (STRs).

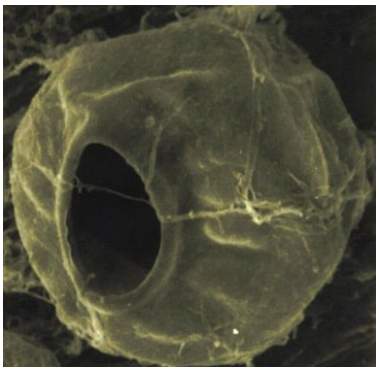
When the Oslo center acquired the PacBio *RS* in 2012, Nederbragt and his colleagues tested out the instrument by running their default cod sample. "When we looked at these PacBio reads mapping to the assembly, we saw them crossing large gaps of even multiple kilobases," he says. It was a moment the team had been anticipating for years. "I could see that the problem of STRs and heterozygosity could be addressed by this technology," Nederbragt adds.

Layering the PacBio and previous reads together, and using the highly accurate consensus, the team generated very long reads, error-corrected them using the short read data, and ran them through Celera® Assembler. "We've never seen a faster assembly," Nederbragt says; it came together in just 36 hours.

They are investigating the new genome data to see where it varies from the original stickleback-oriented assembly. They've already seen an exon in the original annotation that potentially does not exist in the all-cod assembly, Nederbragt says, noting that a full comparison of the two genome assemblies will take place in the future.

Orpinomyces

Scientists from Oklahoma State University and the University of Oklahoma teamed up with a sequencing service provider to study the genome of an anaerobic fungus found in the rumen of cows that may have implications for effective plant biomass degradation. What made this particular species so tricky to sequence were its extreme GC content — just 17 percent — and unusually high number of repeats.



Large herbivores are host to the anaerobic gut fungus, *Orpinomyces*, whose role it is to biodegrade plant material.

The study was reported in "The Genome of the Anaerobic Fungus *Orpinomyces* sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader," a paper published in the ASM journal *Applied and Environmental Microbiology*. The C1A isolate had a large genome: just over 100 Mb, with more than 16,000 genes.

Senior author Mostafa Elshaheda and his team sequenced the fungus, *Orpinomyces* sp. strain C1A, using both Illumina® and PacBio technologies. They report that the organism's extremely low GC content of just 17 percent is the lowest seen of any free-living microbe sequenced to date. Other unusual traits of the genome were its "relatively large proportion of noncoding intergenic regions," comprising some 73 percent of the sequence, and high number of simple sequence repeats, which saw "massive proliferation" in the noncoding regions. These repeats, mostly homopolyer As or Ts, made up nearly 5 percent of the genome; the authors point out that this is at least an order of magnitude higher than repeat numbers reported in other fungal genomes.

These remarkable insights were attained by a two-part attempt to sequence the organism's genome. As described in the paper, the team initially used paired-end sequencing on the Illumina platform to generate an assembly with 290-fold coverage that was "highly fragmented … with an extremely large number of contigs in the final assembly (82,325 contigs), a large proportion of the final assembly (32.4%) harbored in extremely short contigs (300 to 900 bp), and a low N50 (1,666 bp)."

So Elshaheda et al. turned to SMRT Sequencing, generating about 10-fold coverage of the C1A isolate. PacBioToCA was used to produce a hybrid assembly of the fungus that had an average quality score of 59.7. "The final assembly was a marked improvement compared to the Illumina-only assembly, as evident from the improved N50/N90 values" and other characteristics, the authors write. They note that the long PacBio reads "allowed for the identification of a large number of introns previously undetected in the Illumina assembly."

Armed with this sequence data, Elshaheda and his team performed a number of follow-up and functional studies on C1A. They found the organism to be a "remarkable biomass degrader," and in tests of several different plant materials — including switchgrass, corn stover, alfalfa, and more — C1A proved quite versatile, "able to metabolize all different types of examined plant biomass." This trait makes the organism a particularly promising candidate for use in plant bioprocessing required for the production of many biofuels, they add.

Wheat

In a paper published in the journal *Gene*, scientists from Rutgers University and King's College London reported the use of a single SMRT Cell to sequence and assemble more than 400 wheat-storage protein transcripts from 10 strains of the crop.

High throughput and long reads from SMRT Sequencing enable the study of wheat transcriptomes from 10 strains of the crop.

In "PacBio sequencing of gene families — A case study with wheat gluten genes," authors Wei Zhang, Paul Ciclitira, and Joachim Messing note that traditional studies of these cDNA sequences are so costly and labor-intensive that they have not allowed for intensive study of "the variation of each orthologous gene copy among cultivars."

That kind of study for complex traits "usually requires positional information from sequencing entire genomes," a task that would be prohibitive for this type of cross-strain interrogation. "Comparative transcriptome analysis of gene families," the scientists note, offers an alternative way to study multigenic traits "without a need to re-sequence the related genomes in their entirety."

For transcriptome sequencing, short-read technologies eliminate the cost problem, the authors add, but the short sequences "are a

critical barrier to assemble repetitive genes, which may result in inadvertently joining of different gene copies into chimeric molecules."

PacBio sequencing, on the other hand, offers not only the needed throughput but also read lengths capable of resolving long, complex genetic regions, Zhang et al. write. The paper reports a proof-of-principle study designed to determine whether SMRT Sequencing is a viable and scalable option for investigations of variation across several different crop strains.

The authors chose 10 wheat cultivars from around the world, used barcoded PCR primers for each, and pooled the samples to run on a single SMRT Cell. Sequence data had an average read length of 3,050 bp and included nearly 33,000 circular-consensus sequencing reads in the final analysis.

The scientists then compared results of one of their cultivars, a common type of wheat known as Chinese Spring, to information on the same cultivar from the NCBI protein database, finding high rates of concordance. "The accuracy of the assembly in Chinese Spring was validated with 99% identity from cDNAs obtained by conventional sequencing methods," they report. They also succeeded in sidestepping the chimera problem of short-read sequencers: "With the redundancy in sequencing coverage and the length of the sequences, our assemblies avoid chimeric joining of different gene copies."

Zhang et al. note that their method should be useful for other phylogenetic studies, as well. "We suggest our method as an efficient, low-cost method for profiling gene expression of gene families from cultivars, for which a genome has not been [sequenced] or is only available as a draft sequence," they write.

Arabidopsis

Want to check out PacBio data for yourself? We recently sequenced the Landsberg erecta ecotype (Ler-0) of *Arabidopsis thaliana*, produced a successful assembly, and made the data set publicly available. A few stats on *Arabidopsis* and the assembly using PacBio sequence data:

Chemistry: P5-C3
Genome size: 124.6 Mb
GC content: 35.91%
Raw data: 15.87 Gb
Assembly coverage: 18.74x
Polished Genome Assembly Coverage: 93.37x
Polished Contigs: 545
Max Contig Length: 13.21 Mb
N50 Contig Length: 6.36 Mb
Sum of Contig Lengths: 130.86 Mb

The data set resulting from this sequencing effort and assembly using SMRT Portal is now at http://pacbiodevnet.com/.

Thale cress (*Arabidopsis thaliana*) is one of the model organisms used for studying plant biology.

www.pacb.com/denovo

PACIFIC BIOSCIENCES®