COD GENOME ASSEMBLY:
LONG READS OFFER UNIQUE INSIGHT

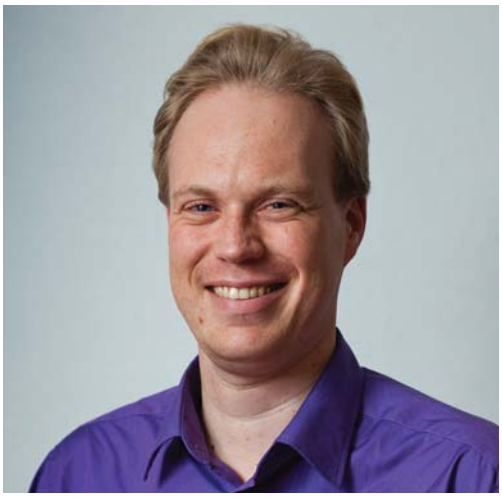
Scientists at the University of Oslo's Centre for Ecological and Evolutionary Synthesis (CEES) applied long PacBio® reads to a genome that was proving particularly difficult to assemble. Today, sequencing problems associated with the Atlantic cod genome are a thing of the past — and researchers are using their new assembly as the foundation for a major resequencing effort that's just getting started.

Recent work using multi-kilobase sequence reads generated from Single Molecule, Real-Time (SMRT®) technology is enabling a dramatically improved genome assembly for cod, an economically important fish species. In many ways the cod genome seemed like a puzzle that might never be fully solved, but the Pacific Biosciences® sequencing platform made significant inroads — and just in time, as the team of researchers working on cod recently received funding to resequence 1,000 more of them. Being able to base these new efforts on a reliable genome assembly will make future results far more meaningful.

Lex Nederbragt, a research fellow at the University of Oslo and a member of the Norwegian High-Throughput Sequencing Centre, regularly contemplates the broader importance of cod. There is interest in domesticating cod for aquaculture, and the 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 this industry. Cod is the most important aquatic species in Norway and other commercial fishery nations. Moreover, cod has an interesting population ecology; that is, some populations do exceedingly well, whereas others get depleted through fishing and never recover to historic abundances. In the last decade or so, "there has been a growing interest in the genomics of this organism," Nederbragt says.

> "Having a really good reference genome will make a big difference."

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 having 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. "They managed to produce a meaningful annotated genome by taking well-known genes from stickleback and other fishes to try to put together the missing pieces in cod," he adds. In generating an assembly and annotation, the project was a success; but scientists knew that for certain regions of the genome, the genes would not accurately represent the cod genome.



Lex Nederbragt, a research fellow at the University of Oslo, uses the PacBio Sequencing System to sequence the cod genome.

Still, having the draft genome assembly was a real step forward and enabled the first big biological finding: nowhere in the cod's 22,000 genes could scientists find the genes necessary for functionality of the MHC II pathway, a critical component of the white blood cell-mediating major histocompatibility complex that exists in all jawed vertebrates. "The immune system of cod is really different from what we're used to seeing," Nederbragt says. "Some key genes associated with that pathway are completely gone from the genome. This has never been seen before, and people are really surprised that this was at all possible." Knowing the pitfalls of the genome assembly,

the research team validated the finding by adding more data and mining information from various sources, including other cod-like fish.

Naturally, CEES scientists are following up on this oddity of the cod genome, looking into how the immune response of cod functions compared to other fish species. Also, the cod genome indicates an expansion of the MHC I complex, so some are investigating whether that has enabled MHC I to compensate for the potential depletion of MHC II in this organism.

Assembly Improvement

Even while the first cod genome assembly was being published, Nederbragt and his colleagues were casting about for ways to improve it. The team's initial attempt at this extracted DNA from the same cod used for the original study and ran mate-pair sequencing with the Illumina® platform. That data helped explain why the genome was so fragmented, though it could not fix the problem. The cod sequenced was from the wild population, a normal diploid fish whose marked heterozygosity — sequence differences between maternal and paternal chromosomes — seemed to be causing issues during assembly. "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. "This confuses assembly programs."

Another problem for the assembly was the presence of many short tandem repeats (STRs). "They're so long that they're longer than the Illumina reads," Nederbragt says, noting that these regions are challenging for Sanger and 454 sequencing as well. "We estimate that 10 to 20 percent of the gaps are flanked, and probably spanned, by those sequences."

What the cod team really needed was sequence data long enough to span these regions of heterozygosity and

In the most Northern part of the Dutch North Sea live big schools of North Sea cod, one of the most endangered species.

"When we looked at these PacBio reads mapping to the assembly, we saw them crossing large gaps of even multiple kilobases. I could see that the problem of STRs and heterozygosity could be addressed by this technology."

STRs. Their big break came in 2012 when the Oslo center acquired the PacBio *RS*.

Building a Better Reference

As they tested out the new instrument, Nederbragt and his colleagues ran their default cod sample to get a sense of the PacBio performance with DNA they already knew very well. "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.

"We've never seen a faster assembly," Nederbragt says; it came together in just 36 hours.

Indeed, the multi-kilobase reads from the PacBio *RS* confirmed what the team had suspected all along: that these short tandem repeats were preventing other sequencing technology from getting through gaps, and that the proliferation of these regions was causing a fragmented assembly. For the first time, the team actually had data indicating that its theory about the heterozygosity problem was correct: the reads showed long stretches of different sequence flanked by sequences that matched each other, indicating heterozygous regions.

This development altered the course of the genome project. "It made us change our sequencing plans," Nederbragt says. "We decided to use the remaining funds on generating about 8x coverage of PacBio reads." Since they already had so much 454 and Illumina data, the team opted to add in the new PacBio reads to improve the existing assembly. "Together with PhD student Ole Kristian Tørresen from our group, we

are currently trying to make use of all the data we have in the best possible way," he says.

Layering these 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.

As Nederbragt and his colleagues sifted through the new assembly, they realized that the assembler was splitting haplotypes rather than merging them, so the heterozygous regions were being run as linear sections of the genome, rather than alternates of the same section. "The sequencing problem is now gone; it looks like we have the whole genome present in PacBio reads," Nederbragt says. "Now it has become a bioinformatics challenge." He and his team are currently working to quantify the regions they believe should be split into haplotypes and to figure out the differences between them.

As they determine the best bioinformatic solution to the assembly, they are starting to investigate 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. For now, they are focused on getting this new assembly into its 23 pseudochromosomes, which can then be shipped off for annotation. "The goal is to get the annotators an assembly good enough that they don't need to retrofit it with information from other organisms," he says.

Implications for Future Initiatives

The interest in generating an improved cod genome assembly took a critical turn when the Norwegian Research Council funded a large grant to resequence 1,000 cod. The four-year Aqua Genome Project, as it is known, will produce truly meaningful and useful results if this resequencing can be done with a high-quality draft of the actual fish being studied, rather than relying on the old cod assembly with information from other fish genes woven into it. "The goal is to do deep cataloging of genomic variation," a process that will also include studies of transcriptome, methylation, and structural variation data on the PacBio platform, Nederbragt says. "Having a really good reference genome will make a big difference."

> "If you want to have long-range information that you can trust, PacBio reads are very useful."

The Aqua Genome Project seeks to characterize the variation between wild-caught and farmed fish in an effort to increase the success of aquaculture. This is no trivial matter: "Sustainable aquaculture could contribute to solving the world's food problems," Nederbragt says.

As he and his colleagues get started on that new project, they'll be relying on their PacBio sequencer. "If you want to have long-range information that you can trust, PacBio reads are very useful," Nederbragt says.

PN: CS106-062713