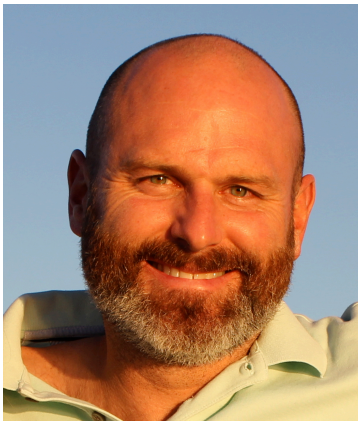
Next-Generation Mapping Enables Ideal Genome Assembly for Greater Duckweed

Scientists at Rutgers University, Washington University, and Ibis Biosciences successfully deployed Next-Generation Mapping (NGM) technology from BioNano Genomics to help produce the first complete assembly for a fast-growing aquatic plant with biofuel potential. What emerged is a clear view into a genome undergoing drastic reduction and a tool to elucidate chromosome-scale dynamics.

For a plant that is often mistaken for algae, *Spirodela polyrhiza* has received star treatment from leading genomics researchers. Years of effort finally paid off in a genome assembly that represents each of 20 chromosomes in a single scaffold, allowing scientists to study chromosome dynamics in addition to the plant's genetic and epigenetic variation.

Eric Lam, Distinguished Professor of Plant Biology at Rutgers University, and Todd Michael, Senior Principal Research Scientist at Ibis Biosciences, lead the team that achieved this feat — following an earlier attempt by Michael and collaborators that produced a more fragmented and lower coverage assembly. This broad interest in a floating aquatic plant known colloquially as greater duckweed stems from its phylogenetic position, ultra-rapid life cycle, and unusually small genome. It also has exceptional potential as a biofuel and bioremediation tool: *Spirodela* is one of the fastest-growing plants in the world and can remove pollutants such as organics and heavy metals from water.

One of the features that makes *Spirodela* so fascinating to scientists like Lam and Michael is that it's apparently in the midst of a massive genome reduction, and the new assembly offers



a high-resolution picture of the process. This invaluable insight would not have been possible without the use of NGM from BioNano Genomics. Paired with sequence data from various short-read platforms that are now widely available, these next-generation maps were the tools that finally resolved the duckweed genome — in particular its repeat elements, including transposable elements. Comparison of the sequence-based genome assembly with NGM results also provided independent validation of the genome model generated. As researchers have found with so many organisms, having a high-quality, complete genome assembly allows for a far more comprehensive understanding of genetic diversity, structural variation, and the mechanisms responsible for important biological processes.

The Duckweed Challenge

The first sequence for *Spirodela* came from a Community Sequence Project supported by the Joint Genome Institute (JGI) in 2009, which provided Michael and his team with 454 and fosmid sequencing data that resulted in a 20x assembly of the 158 Mbp genome represented in several hundred scaffolds. Using BAC-ends and cytogenetics-based techniques, they also got the first count of its chromosomes.

However, Lam and Michael knew they could do better, and aimed for an assembly that would

represent each chromosome in just one scaffold as well as providing higher fidelity via increased coverage. Having a high-quality assembly for *Spirodela*, believed to be the oldest of the duckweed genera and among the oldest monocots that diverged from the lineage that contained all the major grain crops of the world, would be a tremendous asset for the plant biology community. A team led by Lam started from scratch, generating 100x paired-end and mate-pair Illumina data and using the ALLPATHS-LG assembly protocol for a different strain of *Spirodela* from the one previously sequenced. "That assembly was as good as our 454/fosmid/BAC assembly, but it was still in hundreds of superscaffolds," Michael says. By integrating the two independent drafts, the team was able to reduce the number of superscaffolds, though they still fell short of the aim to achieve chromosome-level resolution.

Knowing that short-read data alone wasn't going to overcome the assembly challenge posed by large repeat regions, Lam's team turned to BioNano Genomics for a method to complement the sequence data. BioNano uses its Irys® System to analyze single DNA molecules, treating them with nicking enzymes and digitizing them to produce long-range information from fragments as long as 1 million base pairs. This NGM approach yields highly accurate data in a high-throughput format that can be used with information from various NGS commercial platforms to boost alignment and assembly quality, often resulting in complete *de novo* assemblies. BioNano maps are also far more cost-effective than traditional physical maps based on laborious and expensive BACs.

"BioNano did some really good work with the Lam team at Rutgers, getting high-quality DNA from *Spirodela* to generate nice maps," Michael says. They analyzed two strains of the plant, including the one that had previously been sequenced by JGI. With the BioNano maps, the team resolved the draft assemblies to fewer than 30 superscaffolds for each strain, rapidly approaching the 20 chromosomes of *Spirodela*. "It was really exciting," he says, "but we didn't want to stop there."

For a final shot at a virtually perfect assembly, the team pulled together every asset they had on *Spirodela* — including all of the sequencing data and both BioNano maps. The Lam team integrated the original 454-based assembly to the Illumina one, using each to link superscaffolds in the other, followed by PCR validation. They then integrated the BioNano map, and reduced the assembly to 21 superscaffolds. With only one more reduction to go, the BioNano team assisted using the single molecule maps from both strains, "and we were able to make the last merge to get it down to 20," Lam says. At last, the assembly was as good as it could get. All in, it included about 250x coverage of the *Spirodela* genome.

New Discoveries

With the 20-scaffold assembly in hand, the team dug into the analysis of its gene content, methylation, genome size, and more. "With that type of assembly and accuracy, we were able to improve the annotation of the repetitive elements such as the LTRs," says Michael. Based on earlier assemblies, they expected some 12% to 15% of the genome to be transposable elements; now, they learned that fraction was closer to 25%. They also found that the plant has only 18,507 genes, about 1,000 fewer than what had been predicted based on the earlier draft.

They got their first clear view of structural variation in the plant's genome as well. Earlier work comparing *Spirodela* accessions had shown there was low single nucleotide variation between geographically distinct *Spirodela* strains. "We thought if the genome is purging or in some dynamic flux, there might not be SNP variation but there might be structural variation," Michael says. By comparing BioNano maps from two plants, they identified 96 high-confidence structural variants ranging from 100 bases to 100,000 bases in length. "We didn't find them in the mate-pair sequencing data," he adds. "But the maps made it very easy to pull out those structural variants. This is a very economical way to go right into structural variation."

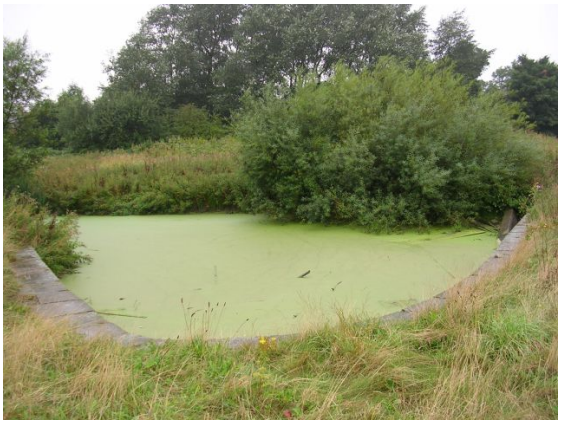
Perhaps the biggest finding was one that came from working with the BioNano team. “They noticed this interesting nicking pattern in the maps,” Michael says. “They had seen it before and thought it was a possible ribosomal repeat.” There was a region on the end of chromosome 2 with about 700 kb of unknown bases. The Rutgers team cloned one unit of the rDNA repeat region, using old-fashioned molecular biology to completely determine the sequence, which was indeed a series of ribosomal repeats. “We were able to show that *Spirodela* only has 80 ribosomal repeats in its genome, compared to *Arabidopsis* that has about 500 repeats with the same size genome,” Michael says.

This very small array of ribosomal repeats “fits exactly with how the genome works because it reproduces asexually and the life cycle is sometimes less than 24 hours, so you really don’t need many ribosomal repeats,” Michael says. “The theme of this plant is that it’s going through a genome reduction. It’s really reducing the number of genes and ribosomal repeats it has.”

‘Telomere to Telomere’

The final *Spirodela* assembly will also allow this team and others to delve into more complex questions about the plant and its relatives. “We can really start digging into evolutionary questions, such as how do genomes purge DNA, how do they bloat? What’s going on in the centromeres and telomeres in terms of shaping those chromosomes and keeping the fidelity of those chromosomes?” Michael says. “When you have the whole chromosome properly assembled you can start asking those questions, but otherwise you’re just working with scaffolds and you don’t know how they’re related and how they fit together.”

Michael and Lam are part of a growing movement within the genomics community to get back to high-quality assemblies that offer the complete picture of a genome. “I’m interested in looking at genomes from telomere to telomere,” Michael adds. “That’s the next step in genomics: understanding the complete complement of how a chromosome works, how it functions at the sequence level, and what’s going on in chromosome dynamics.” Lam also plans to leverage this high-resolution resource to deploy synthetic biology approaches for plant genome engineering in the near future.



To that end, Michael and Lam see BioNano Genomics as an easy-to-use, essential tool in the assembly pipeline. “If you’re committed to the process of doing a genome correctly — beyond the draft genome — having this piece of equipment to complement your sequencer is a high-value item,” Michael says. And he’s got the first-ever high-quality assembly of duckweed to prove it.

For more information about the Irys® System, please visit www.bionanogenomics.com.

9640 Towne Centre Dr., Ste. 100, San Diego, CA 92121 M 858.888.7600 F 858.888.7601 info@bionanogenomics.com

PN 30082 A