How to catch a fish genome with big data

April 11 2018

Scientists annotated and assembled the fish genome of *Seriola dorsalis*, AKA Californa Yellowtail, using big data and supercomputers. Closely related Seriola lalandi shown. Credit: Fishbase

If you eat fish in the U.S., chances are it once swam in another country. That's because the U.S. imports over 80 percent of its seafood, according to estimates by the United Nations. New genetic research could help

make farmed fish more palatable and bring America's wild fish species to dinner tables. Scientists have used big data and supercomputers to catch a fish genome, a first step in its sustainable aquaculture harvest.

Researchers assembled and annotated for the first time the genome—the total genetic material—of the fish species *Seriola dorsalis*. Also known as California Yellowtail, it's a fish of high value to the sashimi, or raw seafood industry. The science team formed from the Southwest Fisheries Science Center of the U.S. National Marine Fisheries Service, Iowa State University, and the Instituto Politécnico Nacional in Mexico. They published their results on January of 2018 in the journal *BMC Genomics*.

"The major findings in this publication were to characterize the Seriola dorsalis genome and its annotation, along with getting a better understanding of sex determination of this fish species," said study co-author Andrew Severin, a Scientist and Facility Manager at the Genome Informatics Facility of Iowa State University.

"We can now confidently say," added Severin, "that Seriola dorsalis has a Z-W sex determination system, and that we know the chromosome that it's contained on and the region that actually determines the sex of this fish." Z-W refers to the sex chromosomes and depends on whether the male or female is heterozygous (XX,XY or ZZ,ZW), respectively. Another way to think about this is that in Z-W sex determination, the DNA molecules of the fish ovum determine the sex of the offspring. By contrast, in the X-Y sex determination system, such is found in humans, the sperm determines sex in the offspring.

It's hard to tell the difference between a male and female yellowtail fish because they don't have any obvious phenotypical, or outwardly physically distinguishing traits. "Being able to determine sex in fish is really important because we can develop a marker that can be used to determine sex in young fish that you can't determine phenotypically,"

Severin explained. "This can be used to improve aquaculture practices." Sex identification lets fish farmers stock tanks with the right ratio of males to females and get better yield.

Assembling and annotating a genome is like building an enormous three-dimensional jigsaw puzzle. The Seriola dorsalis genome has 685 million pieces—its base pairs of DNA—to put together. "Gene annotations are locations on the genome that encode transcripts that are translated into proteins," explained Severin. "Proteins are the molecular machinery that operate all the biochemistry in the body from the digestion of your food, to the activation of your immune system to the growth of your fingernails. Even that is an oversimplification of all the regulation."

Severin and his team assembled the genome of 685 megabase (MB) pairs from thousands of smaller fragments that each gave information to form the complete picture. "We had to sequence them for quite a bit of depth in order to construct the full 685 MB genome," said study co-author Arun Seetharam. "This amounted to a lot of data," added Seetharam, who is an associate scientist at the Genome Informatics Facility of Iowa State University.

The raw DNA sequence data ran 500 gigabytes for the Seriola dorsalis genome, coming from tissue samples of a juvenile fish collected at the Hubbs SeaWorld Research Institute in San Diego. "In order to put them together," Seetharam said, "we needed a computer with a lot more RAM to put it all into the computer's memory and then put it together to construct the 685 MB genome. We needed really powerful machines."

That's when Seetharam realized that the computational resources at Iowa State University at the time weren't sufficient get the job done in a timely manner, and he turned to XSEDE, the eXtreme Science and Engineering Discovery Environment funded by the National Science Foundation. XSEDE is a single virtual system that scientists can use to

interactively share computing resources, data and expertise.

"When we first started using XSEDE resources," explained Seetharam, "there was an option for us to select for ECSS, the Extended Collaborative Support Services. We thought it would be a great help if there were someone from the XSEDE side to help us. We opted for ECSS. Our interactions with Phillip Blood of the Pittsburgh Supercomputing Center were extremely important to get us up and running with the assembly quickly on XSEDE resources," Seetharam said.

The genome assembly work was computed at the Pittsburgh Supercomputing Center (PSC) on the Blacklight system, which at one point was the world's largest coherent shared-memory computing system. Blacklight has since been superseded by the data-centric Bridges system at PSC, which includes similar large-memory nodes of up to 12 terabytes—a thousand times more than a typical personal computer. "We ended up using Blacklight at the time because it had a lot of RAM," recalled Andrew Severin. That's because they needed to put all the raw data into the computer's random access memory (RAM) so that it could use the algorithms of the Maryland Super-Read Celera Assembler genome assembly software. "You have to be able to compare every single piece of sequence data to every other piece to figure out which pieces need to be joined together, like a giant puzzle," Severin explained.

"We also used Stampede," continued Severin, "the first Stampede, which is another XSEDE computational resource that has lots and lots of compute nodes. Each compute node you can think of as a separate computer. " The Stampede1 system at the Texas Advanced Computing Center had over 6,400 Dell PowerEdge server nodes, which later added 508 Intel Knights Landing (KNL) nodes in preparation for its current successor, Stampede2 with 4,200 KNL nodes.

"We used Stampede to do the annotation of these gene models that we identified in the genome to try and figure out what their functions are," Severin said. "That required us to perform an analysis called the Basic Local Alignment Search Tool (BLAST), and it required us to use many CPUs, over a year's worth of compute time that we ended up doing within a couple of week's worth of actual time because of the many nodes that were on Stampede."

"This experiment started with a collaboration with the Southwest Fisheries Science Center of NOAA," Severin explained. He said the project originally set out to complete a large RNA-seq project, and it turned out that there was sufficient funding to also do a genome assembly. "That resulted in a long-term collaboration with the Southwest Fisheries Science Center," Severin said. "With the recent advances in high-throughput DNA sequencing, we're now able to generate terabytes of sequencing data. This tends to be short, 100-150 base pair reads that we have to put together like a very large puzzle and figure out where all the pieces go," he added.

Severin and Seetharam's team have completed the basic picture of the genome for Seriola dorsalis, but they say there's still room for refinement. "The genome that we assembled is not perfect, in the sense that it is still in many pieces. We weren't able to fully piece together entire chromosomes," explained Seetharam. "We have many scaffolds representing each of those chromosomes, and we are missing a lot of information that is needed to fill in the gaps." Sequencing technology advancements can address these gaps, Seetharam said, through the advancement of sequencing technology that can produce longer DNA reads.

"We also hypothesized in the paper," said Severin, "of this deletion that's upstream of a gene that converts estrone into estrogen, that's part of the sex-determination pathway. That may be responsible for sex

determination. But since it's just a hypothesis based on computational methods, this needs further investigation in the lab. We could certainly follow that up with a CRISPR-like experiment to test this mutation."

Severin also mentioned data collection for a larger genome-wide association study experiment to find locations and variants in the genome associated with jaw deformities. "We're currently collecting those samples," said Severin, "but we'll be able to use that genome to provide markers to the farmers to select against fish that have these propensities for jaw deformity."

Both Severin and Seetharam are resolute in their conviction that big data can solve problems in sustainable food production. "I believe the public is going to see more of this type of big data utilization and to see why science is so important for our future," Severin said. Gene annotation, he feels, is just the tip of the iceberg. "We're going to start comparing genome assemblies with each other and start getting at what a genome is and how it works; and how for a particular genome does the presence or absence of genes or its context with regard to its three-dimensional structure, how does that make a species," Severin said.

"Big data keeps getting bigger, and we're finding answers to really interesting questions." Severin concluded. Seetharam added that "There will be more studies using big data that will have significant impactful for the general public. This level of research will foster even larger studies in the future."

The study, "Insights into teleost sex determination from the Seriola dorsalis genome assembly," was published January of 2018 in the journal *BMC Genomics*.

More information: Catherine M. Purcell et al, Insights into teleost sex determination from the Seriola dorsalis genome assembly, *BMC*

Genomics (2018). DOI: 10.1186/s12864-017-4403-1