Restoring and Managing Historical Columbia River Basin Fish Populations with Genomics

Novel GT-Seq method using the HiSeq® System reduces genotyping costs and enables analysis of tens of thousands of fish.

Introduction

Hydropower dam construction and other environmental incursions have suppressed and sometimes decimated fish populations around the world. In the Columbia River Basin, an area approximately the size of France covering portions of Washington, Oregon, Idaho, and British Columbia, the survival of some species was imperiled, including populations of coho, sockeye, and Chinook salmon. The catastrophic loss of fish in this region represented a marked infringement of the rights of Native American tribes, which include the guaranteed right to harvest fish in all "usual and accustomed" areas from treaties signed with the US government in 1855.

Shawn Narum, PhD, is part of a team of geneticists and other researchers that are helping Native American tribes revitalize fish populations. They're providing the technical information needed by fisheries, hatcheries, hydropower operators, and Native Americans to restore Columbia River Basin salmon and other fish populations to their historical population levels.

iCommunity recently spoke with Dr. Narum about his work at the University of Idaho's Hagerman Fish Culture Experiment Station for the Columbia River Inter-Tribal Fish Commission (CRITFC). His research sparked the development of Genotyping-in-Thousands by sequencing (GT-Seq), a new next-generation sequencing (NGS) method for cost-effectively genotyping thousands of individual fish simultaneously in a single Illumina HiSeq® System lane.[1]

Dr. Shawn Narum with a Chinook salmon he caught on a recent fishing trip. NGS methods are used to identify fish and their population of origin, enabling researchers to determine the overall proportion of fish harvested from different regions.

Q: What is CRITFC and what is its mission?
Shawn Narum (SN): CRITFC is an organization that supports 4 Native American tribes in the US Pacific Northwest. That includes the Nez Perce Tribe, which is here in the State of Idaho where I work, the Yakima Nation in Washington, and the Warm Springs and Umatilla tribes in Oregon. A shared concern among these tribes is the conservation and recovery of salmon. Salmon is a subsistence food and has been an important part of their tribal cultures for thousands of years. Restoring salmon to the Columbia River Basin and being able to partake in harvesting the fish is a way of life for them. CRITFC provides the tribes with technical information to support fish conservation efforts. This includes promoting responsible management of the Columbia River Basin watershed, encouraging productive hatchery practices, and coordinating with co-managers to determine the best times and places for tribal and nontribal fisheries to harvest fish.

Q: What is your role at CRITFC?
SN: Our group performs genetic research on many of the fish species found in the Columbia River. Salmonids (salmon and trout) make up the largest focus of our work, with specific emphasis on Chinook salmon and steelhead trout. We also study coho and sockeye salmon, sturgeon, lamprey, and other resident trout species. We use genetic methods and tools to monitor harvest operations of commercial, sport, and tribal fisheries, and to study the genetic effects of hatchery practices and the genetic adaptation of fish populations to local environments.

Nathan Campbell developed the GT-Seq method, an amplicon sequencing-based approach that the Hagerman Genetics Laboratory team uses to genotype thousands of fish, such as the steelhead trout he's holding.

Q: What is the role of genetic analysis in restoring fish populations?
SN: In the last 2 decades, there has been a substantial effort to reform hatchery practices. Tribal hatcheries strive to operate "integrated" hatcheries that use local wild fish as broodstock, and other techniques to produce genetically diverse fish that best represent local wild populations. Genetic analysis enables maintenance of wild broodstock diversity and mating success evaluation of hatchery-reared offspring. We can also investigate whether domestication selection is occurring even within the short generational time frame of a hatchery rearing program. For most studies where they're using primarily broodstock of wild origin, we see that indeed the hatchery offspring are similar to their wild counterparts and provide a demographic boost to help recover native populations. This is in contrast to "segregated" hatcheries that are divergent from local stocks and have greater potential for domestication selection to accumulate over generations. These segregated stocks are intended to increase harvest opportunities, but it is very difficult to keep these fish entirely isolated from wild fish. We can use genetic tools, such as parentage analysis, to identify mating events in the natural environment and see if any negative genetic consequences are occurring from hatchery practices.

Genetic analysis enables maintenance of wild broodstock diversity and mating success evaluation of hatchery-reared offspring.

Our work with coho salmon provides an interesting example of how we use genetic analysis. Coho salmon were considered functionally extinct throughout much of the Columbia River by the 1980s or earlier. They were reintroduced successfully by Native American tribes into multiple river systems, and large spawning runs have returned in recent years (approximately 280,000 coho salmon in 2014). In coho salmon and other fish species, we can study genetic adaptation of fish populations to their local environments. In particular, we have used NGS methods such as Restriction site Associated DNA sequencing (RAD-Seq) and RNA-Seq to detect signals of evolutionary adaptation to environmental stressors such as warmer conditions. These methods enable us to separate our analyses of adaptive variation affecting functional genes from neutral variation appearing in noncoding parts of the genome. Neutral variations are better indicators of interbreeding and genetic relationships between different populations.

Another role for genetic analysis is genetic tagging and monitoring of stocks. There are several distinct salmonid stocks in the Columbia River Basin. Some of these stocks are fairly abundant, but there are others that are listed under the Endangered Species Act. At multiple points in their lifetimes, these healthy and endangered stocks intermingle. Using genetic tagging and monitoring, we can determine which stocks are prevalent in various harvest locations and at what times. This information is useful in fishery management to assist with coordinating harvest seasons and locations.

Q: How do you obtain fish DNA samples for analysis?
SN: We need only approximately 10 nanograms of DNA, which we can obtain from a very tiny piece of fin tissue. This sampling is easy to do and is nonlethal. For example, as adult salmon return from the ocean to their natal streams, the first dam they encounter is Bonneville Dam. We can divert some of those fish into a trap where we take a tissue sample and release them back into the water quickly.

Q: What are some of the methods you have used for genetic analysis of fish?
SN: When I started in 2002, we were using capillary gel electrophoresis to look at about a dozen microsatellite markers. In 2008, we started performing more SNP discovery and made the transition to using whole-genome information. At that point, we were still using low-throughput technologies. It wasn't until 2010 that we started using NGS. It became clear to us that some of the RAD-Seq methods being presented at the time could be very useful for our studies.[2] We performed the first few runs on the Illumina Genome Analyzer® II sequencing system, and did some more runs at a core lab with a HiSeq® 2000 System. It was about 3 years ago when we purchased the HiSeq 1500 System for our own lab. Since then, we've been using the HiSeq System extensively for RAD-Seq, RNA-Seq, and now GT-Seq.

Q: Is RAD-Seq a productive approach for your studies?
SN: Yes, it's very productive. RAD-Seq enables us to sequence around certain restriction enzyme sites to obtain a reduced representation of the genome. In salmonids, RAD-Seq enables us to genotype somewhere around 10,000 SNPs for about 50 individuals in a single lane of a HiSeq run. After we began using RAD-Seq, we realized that genotyping by sequencing was widely applicable for our work.

GT-Seq enables us to genotype thousands of fish at relatively low cost.

Q: What is GT-Seq and how did you develop it on the HiSeq platform?
SN: I have to give credit to Nathan Campbell, a scientist here in the Hagerman Genetics Lab and first author on our paper describing the technique. He came up with the idea to use amplicon sequencing with Illumina technology. He developed the subsequent lab protocols and bioinformatic pipelines and deserves the recognition for developing the GT-Seq method.

As Illumina NGS system users know, the HiSeq flow cell has a large capacity for sequencing DNA fragments. In our RAD-Seq studies, we were looking at 10,000 SNPs for 50 individual fish. We didn't need that many SNPs for high-throughput screening, and instead used 5' exonuclease fluorescence allelic discrimination assays to genotype about 200 SNPs on 96 fish at a time. We were satisfied with the quality of the data, but we realized that this approach wouldn't be viable from a cost standpoint for the high-throughput types of studies we envisioned. There are hatcheries

throughout the Columbia River Basin, and from Alaska to California. If we could genotype all the parents, we could efficiently tag the millions of offspring they produce. However, that means we'd need to genotype tens to hundreds of thousands of individuals. In developing the GT-Seq method, this was the throughput we were looking to achieve within a realistic cost scenario.

One way to understand GT-Seq is that it is an inversion of the RAD-Seq ratio between the number of individuals and the number of sequenced loci. Instead of sequencing a high number of RAD loci on a low number of individuals, GT-Seq uses the HiSeq System flow cell capacity to sequence a reduced number of SNP loci (approximately 100–500) on thousands of individuals per lane.

> The HiSeq System has enabled us to move from genetic research using just a handful of markers, to genomics where we're covering much more of the genome.

Q: How does GT-Seq reduce the cost of genotyping thousands of fish?
SN: GT-Seq is an amplicon sequencing method that uses 2 rounds of tailed PCR before the individual samples are mixed. In our initial development, we used 192 SNP loci to perform genotyping. The first round of PCR amplifies all 192 SNP sites simultaneously and adds the sequencing primer sites. The second round of PCR adds 2 unique barcode sequences and the Illumina capture sites. The first barcode maps back to 1 of the 96 individual wells in a microplate, and the second barcode identifies which microplate is holding that well. At that point, we can mix the amplification products of all the individuals together for loading on the HiSeq System, because we'll be able to read the barcodes and tie all the resulting sequence reads back to the individuals they came from. In addition to enabling us to genotype thousands of fish in a single HiSeq lane, we realized even greater economy from the dual barcode, well-and-plate identification system. The number of individual barcodes we need to order is only the sum of 96 and the number of plates we're going to use, rather than a separate barcode for each 1 of the thousands of individuals.

Q: What is the principal benefit of GT-Seq?
SN: The main benefits are economy and throughput. Many of our population studies require analysis of many individuals. Relative to array genotyping or the 5' exonuclease assays we used previously, GT-Seq enables us to genotype thousands of fish at relatively low cost. We wouldn't be able to carry out these increasingly large population studies without GT-Seq.

Q: What were the technical challenges you and Nathan Campbell encountered in developing GT-Seq?
SN: In a GT-Seq experiment, the sequencing objective is to cover a panel of 100–500 custom amplicons containing targeted SNPs. Initially, some of the loci were amplified more efficiently than others. Without a way of balancing out amplicon production from all the targeted loci, we would have seen a lot of sequencing coverage for the most abundant amplicons, and inadequate coverage of the less abundant ones. For our steelhead trout genotyping experiment, we eventually came up with a protocol that would amplify our 192 SNP genotyping loci relatively evenly. The panel is well-suited for high-throughput screening applications like the ones we're performing at CRITFC.

After the amplification steps, there are thousands of individual samples in microplate wells, each with 192 amplification products. This is another step where the amount of DNA needs to be balanced out. In this case, the objective is to prevent over- or under-representation of individuals (rather than amplicons) in the HiSeq flow cell and the resulting sequencing data. We looked at several methods to make sure that we would be loading a similar amount of amplicon DNA from every individual. We found a commercial product, SequalPrep, that was compatible with our throughput requirements and normalized the amount of PCR product DNA across all the individuals.

Q: Have you tried GT-Seq in species other than steelhead trout?
SN: We've used the technique with steelhead trout and have already developed panels for Chinook, coho, and sockeye salmon, as well as Pacific lamprey. We're genotyping large numbers of individuals of those 4 salmonids from various populations in the Columbia River. Our lamprey study is now underway as well.

Q: Would GT-Seq be useful for other species and applications?
SN: There is intense interest in performing amplicon sequencing in fish and many other organisms as well. GT-Seq changes the cost structure of genotyping by sequencing, where you want to look at relatively few SNP markers across many individuals. Usually there are a few hundred SNPs that are high targets of interest or candidates for a particular trait. If you're talking about cows, it might be milk production. If you're talking about aquaculture production of Atlantic salmon, scientists are looking for a fast-growing fish with high-quality filets. These are all scenarios where people might want to use a high-throughput, low-cost genotyping platform.

> From our perspective, the number of sequence reads per run is probably the greatest benefit of the Illumina HiSeq platform.

Q: What role did the HiSeq System play in your research?
SN: The HiSeq System has enabled us to move from genetic research using just a handful of markers, to genomics where we're covering much more of the genome. We use the HiSeq System extensively for RAD-Seq and also for RNA-Seq, which allows us to look at gene expression across the transcriptome in thermal adaptation studies. We're now using the HiSeq System even more extensively for high-throughput genotyping with the GT-Seq method. From our perspective, the number of sequence reads per run is probably the greatest benefit of the Illumina HiSeq platform. Two-hundred million reads per lane is a tremendous number of reads that can be used to obtain the coverage needed for genotyping by sequencing applications.

AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAG

References

1. Campbell NR, Harmon SA, Narum SR. Genotyping-in-Thousands by sequencing (GT-Seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Mol Ecol Resour. 2014; doi: 10.1111/1755-0998.12357. [Epub ahead of print]
2. RAD-Seq Genotypes Less, But Offers More, www.illumina.com/content/dam/illumina-marketing/documents/icommunity/article_2011_10_Floragenex_RADseq.pdf

Learn more about the Illumina system mentioned in this article:

- HiSeq System, www.illumina.com/systems/hiseq_2500_1500.html

Follow the Illumina Agrigenomics Community

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®