4.45  Aquaculture Genomics: A Case Study with Catfish

Z Liu, Auburn University, Auburn, AL, USA

© 2011 Elsevier B.V. All rights reserved.

4.45.1  Introduction

Aquaculture is a relatively new, but increasingly growing, sector of agriculture. It is very important not only for economic interest, but also for social and cultural significance. In many of the Asian countries, serving seafood in the dinner table is regarded as a prestigious treatment. Seafood accounts for 20% animal protein sources for the world population; the seafood industry involves a total of 144 million metric tons annual production with 44 millions fishers/fish farmers, 2.1 million vessels, contributing $166 billion dollars to the world economy, and over $25 billion dollars of international trade. Currently, aquaculture accounts for over 40% of seafood consumed in the world, and China is the only country where aquaculture produces more than 50% of consumed seafood.

Like any other agricultural sector, sustained production requires research of basic biology including growth, nutrition, reproduction, physiology, and genetics and genomics. This article focuses on recent genomic studies leading to whole genome sequencing in aquaculture species, and potential uses of genome technologies in aquaculture. Instead of reviewing progress made in all aquaculture species, the article provides a general path of research requirements to support an effective genome project in...
Aquaculture using catfish genome as an example because of the highly dynamic nature of the research area. It is likely that significant progress is being made while this article was being written.

The availability of a draft whole genome sequence significantly enhances genome research and applications of genome-based technologies for improving agricultural production and quality. As such, whole genome sequences are available or nearly available for major livestock species including cattle, swine, chicken, and horse. Whole genome sequencing used to be a nearly ‘utopia’ because of its involved huge costs. Take human genome as an example, it took the major genome centers multiple years to complete the draft genome sequencing of the first human individual, from 1986 to 2000. The direct cost involved in the human genome sequencing was at the level of hundreds of million dollars. Along with the sequencing effort, related human genome research cost much more. Thereafter, sequencing of the cattle genome cost over $60 million, much cheaper than the human genome sequencing, but yet is still at a cost level unthinkable for many aquaculture species. It was because of such huge costs, whole genome sequencing used to be regarded as a dream of lifetime achievement.

The daunting tasks of whole genome sequencing all a sudden become possible for many species including many of aquaculture species. Now with the second- and third-generation sequencing technologies, the cost of sequencing a genome with a size of 1 billion base pairs has reduced to manageable levels, most often within a few hundred thousand dollars. With such a major reduction in costs, the ‘sequencing rush’ is on the way with many species.

The rapid advances in sequencing technologies made science road maps and plans almost meaningless. Back in 2008, the USDA spent much energy developing the blueprint for USDA Efforts in Agricultural Animal Genomics 2008–17. Among many aquaculture species, the blueprint calls for the production of a draft genome sequence for catfish with only 6× genome coverage, and much lower coverage for several other aquaculture species including rainbow trout, tilapia, and shrimps. As short as 2 years later, the drastic reduction in sequencing costs recently has made it now possible to sequence the whole genome of catfish. Very recently, USDA approved funding for the generation of channel catfish whole genome sequence using the next-generation sequencing technology.

At least a dozen of aquaculture species are now being sequenced. Of the six aquatic species groups included in the US Animal Genome NRSP-8 program, whole genome sequencing projects are underway for at least four: tilapia, Atlantic salmon, catfish, and oyster. The tilapia genome project was funded by National Institutes of Health (NIH), and it is being sequenced at the Broad Institute (http://www.broad.mit.edu/science/projects/mammals-models/vertebrates-invertebrates/tilapia/tilapia-genome-sequencing-project). Organized by the Oyster Genome consortium, the oyster genome is being sequenced at the Beijing Genome Institute (http://www.intl-pag.org/17/abstracts/W45_PAGXVII_329.html), sponsored by scientists of the Oceanic Institute in Qingdao, China. Sequencing of the Atlantic salmon genome is more complex as the salmon genome is derived from a whole genome duplication approximately 20–120 million years ago [2, 75]. However, highly productive research projects, particularly those funded by Genome Canada, have prepared a solid foundation for sequencing the salmon genome. Recent communications with Dr. Willie Davidson (Simon Fraser University) and Ben Koop (University of Victoria) confirmed the initiation of salmon genome sequencing as a cooperative project between Canada, Norway, and Chile. Several countries including China, Thailand, Japan, and the US are discussing strategies for sequencing the shrimp genome. US scientists are preparing sequencing the genome of rainbow trout as well.

Although it is extremely exciting for aquaculture geneticists to have the whole genome sequenced for many of the important aquaculture species, it is pivotally important to address some of the most important issues related to whole genome sequencing. These include issues related to effective assembly of the whole genome sequences, and issues related to effective uses of the whole genome sequences. The usefulness of the whole genome sequence relies on the efficient sequence assembly and adequate sequence annotation, which in turn depends on the availability of a number of genome resources. The article focuses on several major lines of research leading to the development of genome resources in catfish in recent years in support of the whole genome sequencing, assembly, and annotation, and provides some perspectives as to why the whole genome needs to be sequenced in the first place.

### 4.45.2 Required Resources for the Whole Genome Sequence Assembly

Assembly of a large eukaryotic genome is not an easy task. Theoretically, genome assembly works by extension of partially overlapping sequences derived from the same genomic neighborhood. The ideal situation is the minimal but sufficient overlapping sequences of long reads that allow extension as far as possible to achieve a long contig. As long as the overlapping sequences are long enough to assure that, the sequences are coming from the same genomic location, the shorter the overlapping sequences are, the greater the extension power. Of course, the longer the sequence reads, the more efficient to assemble the draft genome sequence.

### 4.45.2.1 Understanding the Genome Repeat Structures

In all eukaryotic genomes characterized to date, there are sequence repeats in the genome. Repeated sequences can be classified into various types, but structurally, two types of repeats are most important: those that are repeated in tandems and those that are repeated as dispersed elements scattered all over the place in the genome. Genomic characterization of tandem repeats can be difficult as it is extremely difficult to make assessment of the sizes of such tandem repeats, and the assembly of tandemly arranged repeats can be extremely difficult if not entirely impossible. Characterization of dispersed repeats is usually easier, but their implications for genome sequencing projects can be extremely adverse and significant.
The best-known example of the tandem repeats involves ribosomal RNA genes. They are organized in long tandem arrays containing both gene and spacer sequences. In most animal genome, there are some 100–200 ribosomal RNA (rRNA) genes and in plants the numbers can be much higher. It is not uncommon to find 5000 rRNA genes per genome in plants. Other than repeated gene families, there are various other types of tandemly arranged repeats. Most often, such repeats are discovered by restriction analysis of genomic DNA. As the distribution of restriction enzyme sites are random, digestion of genomic DNA using restriction endonuclease should result in smears. However, in most cases, some bands may appear superimposed to the smear, suggesting the presence of tandemly arranged repeats in which there are restriction endonuclease sites. These repeats are often named after the restriction nucleases names such as the Alu I elements in humans and the Hpal elements in centromeres of salmon.

Although we do not have a complete understanding of the genome repeat structure in catfish, we have characterized several major classes of repeats in the catfish genome. The most abundant repeat class is an approximately 325-bp A/T-rich Xba element with approximately 150,000 copies in the catfish genome arranged in head-to-tail tandem arrays. The catfish genome also contains approximately 9000 copies of 57-bp Mermaid and approximately 1200 copies of 64-bp Merman elements, which are short-interspersed repetitive elements found in many bony fish. We reported a highly abundant (32,000 copies) nonautonomous Tc1 element named Tipnon dispersed in the catfish genome. Through genome sequence surveys using BAC end sequences, we estimated that Tc-1/mariner DNA transposons are the major dispersed repetitive elements, representing approximately 4–5% of catfish genome sequence. Expressed sequence tag (EST) analysis indicated that the Tc-1 elements are associated with 0.6% of the catfish transcripts. Such information is crucial not only for entire genome sequencing and assembly but also for comparative genome analysis and the establishment of conserved syntenies. Because the Tc1-like elements are highly repetitive and may contain segments of other genes, special caution is needed for establishment of orthologies involving such sequences.

Repeat structure in the genome can cause major difficulties for whole genome sequencing if the repeats are long and in high copies. The long repeats cannot be sequenced through into the unique region by a single sequence run. Therefore, overall sequence depends on sequence assembly. Assembly of repeated sequences can be complicated. For many of the aquaculture species, genome repeat structures are unknown. It is, therefore, important to characterize the genome repeat structure to ensure correct whole genome sequence assembly.

**4.45.2.2 Length of Sequence Reads**

For whole genome sequence assembly, the longer the sequence reads, the better it is. Long sequence reads simplify sequence assemblies because it can directly overcome difficulties related to short repeated sequences as the long sequencing reads simply just read the repeated sequences through. Long sequence reads also have a much greater power for sequence assembly. In this regard, the traditional Sanger sequencing can produce sequences 800–1000 bp per sequencing reads, while the Pyrosequencing technologies such as the 454 sequencing generate sequences that are usually 400–500 bp. Illumina sequencing generate sequence tags that are much shorter, less than 100 bp in most platforms. The third-generation sequencing technology such as the Pacific Biosciences sequencing technology can generate kilobase-long sequences, but its wide application requires further technology maturation.

**4.45.2.3 Scaffolding Using Paired Sequence Reads**

Paired sequence reads are an important tactic to direct correct sequence assembly. For instance, if sequences are generated from a DNA fragment of 5-kb length from both ends, they would be physically linked with a spacing of approximately 5 kb. With the next-generation sequencing technologies, sequence reads are mostly short tags. Therefore, reading from both ends to produce paired sequence reads is required to enhance sequence assembly powers. Most often, DNA segments of various sizes are generated to produce paired sequence reads anchored by different physical linkage and spacing.

**4.45.2.4 Scaffolding Using Paired BAC End Sequence**

BAC end sequences are generated by single-pass sequencing of BAC clones from both ends. Although single-pass sequencing can generate a short read of 600–800 bp using Sanger sequencing, the scaffolding capacity of such reads is large. Because BAC clone inserts are large, most often over 150 kb. The two reads from the ends of the BAC insert of each BAC clone are linked with a physical separation of the BAC insert.

In catfish, over 61 Mb of BAC end sequences have been generated using Sanger sequencing (6% of catfish genome) with 103,000 BAC end sequences from two catfish BAC libraries (Quiniou, unpublished data). The total BAC end sequences represent, on average, one sequence tag per 9.7 kb of the catfish genome. The majority of BAC end sequences (88,000 sequences from 44,000 BAC clones) was mate-paired, and thus highly useful for assisting the whole genome assembly. Such BAC end sequences will be highly useful for scaffolding of whole genome sequences.

**4.45.2.5 Scaffolding Using BAC-Based Physical Maps**

A good physical map can greatly assist the assembly of whole genome sequence when coupled to the BAC end sequences. Even if gaps exist prohibiting merging of contigs, BAC-based physical maps should allow establishment of large scaffolds when the sequences are linked with BACs within contigs of the physical map.
Two channel catfish BAC libraries have been constructed and characterized [85, 108]. BAC contig-based physical maps of the channel catfish genome have been generated using high information content fingerprinting [86, 114]. Fingerprinting of 46 548 BAC clones (6.8X genome coverage) from the CCBL1 BAC library developed from an inbred meiotic gynogen permitted the assembly of 1782 contigs [86]. Fingerprinting of 34 580 BAC clones (5.6X genome coverage) from CHORI 212 BAC library developed from an outbred male permitted assembly of 3307 contigs [114]. The CCBL1 BAC library was based on Hind III partial digestion of genomic DNA, while the CHORI 212 BAC library was based on EcoR I partial digestion, so we expect many genomic regions not covered within one library could have been covered in the other. We plan to integrate these physical maps to close some gaps. Merging of the two physical maps should reduce the number of overall contig numbers.

### 4.45.2.6 Scaffolding Using Linkage Maps

Linkage maps are important because they are derived using genetic recombination. Ultimately, understanding of genomics and genomic information will need to be translated in genetic terms for genetic enhancement in aquaculture. Genetic maps are also important for the assembly of chromosome-scale sequence scaffolds. This is because no matter what sequencing strategies are used and how good a job research community can do in sequencing, many gaps will still exist within any given chromosome, resulting in many contigs. Such intrachromosomal contigs need to be arranged into linear relations to reflect the arrangements of the genome sequence contigs on the chromosome. In this regard, mapping of sequence-tagged genetic markers such as microsatellites and single nucleotide polymorphisms (SNPs) is crucially important. Their location on the genetic linkage mapping would allow them to be arranged into large scaffolds corresponding to the linkage group or the chromosome. In catfish, on the basis of the framework linkage maps constructed several years ago [52, 102], we have recently made significant progress in linkage mapping. Such progress includes placing 331 gene-associated markers to 29 linkage groups [42]. With the availability of a large number of microsatellite sequences within BAC end sequence, over 2000 microsatellites have been placed on the genetic linkage map, allowing establishment of scaffolds of the catfish genome (Peatman et al., in preparation). This genetic linkage map will serve to validate the whole genome assembly and anchor the assembly to chromosomes.

We have identified a large number of microsatellites from BAC end sequences [47, 93]. To date, we have placed over 2000 BAC end-associated microsatellites on the linkage map. Thus, almost two-thirds of the CHORI 212 BAC contigs have been placed on the linkage map to establish BAC scaffolds. These scaffolds contain large numbers of BAC end sequences, and they will be extremely useful in guiding the assembly of the reference whole genome sequence. The existing linkage and physical maps, along with the BAC end sequences, will serve as a framework for the whole genome sequence assembly, and the sequence assembly will permit additional BAC end-anchored microsatellites and SNP markers to be mapped in silico.

The theoretical ultimate assembly of a genome could have the contig numbers to equal to the chromosome numbers. In that case, it would mean that there are absolutely no gaps along the entire chromosome to generate the continuous one sequence per chromosome. Practically, that is entirely not feasible with limitations of the budget, although it is theoretically possible. This is because with a large genome size of over 1 billion base pairs, there would be many gaps after the sequences are assembled, leading to a large number of contigs per chromosome.

The question is how many contigs per chromosome are acceptable for much of the practical operations. There is no concretely correct answer to this question because it all depends on what the practical purpose is with the whole genome sequence. In addition, when the issue is seriously dissected, it depends not on only the number of the gaps, but also the sizes of the gaps, as well as the ability to place the contigs of sequences in a linear fashion to form the scaffolds.

### 4.45.3 Required Resources for the Whole Genome Sequence Annotation

A major challenge in a whole genome project is the proper annotation of the genome sequences. Specifically, gene models and gene structures need to be supported by experimental data; exon–intron borders need to be defined; alternative splicing and differential polyadenylated transcripts need to be identified; and expression and function of the genes need to be studied. Genome resources such as full-length complementary DNA (cDNA) sequences and ESTs are particularly useful in support of genome annotation. Toward that goal, we have constructed 32 unidirectional plasmid cDNA libraries (10 libraries were normalized) from various tissues of channel catfish and blue catfish [44, 106], and generated a large set of ESTs. Work toward generating a set of full-length cDNAs are in progress.

#### 4.45.3.1 EST Resources

To date, GenBank dbEST contains 493 852 catfish ESTs based on Sanger sequencing, with 354 377 entries for channel catfish and 139 475 entries for blue catfish [14, 30, 32, 41, 44, 74, 106]. Clustering of EST sequences with CAP4 software resulted in 45 306 contigs and 66 272 singletons. All ESTs have been annotated with putative gene identification based on sequence homology [106]. Over 35% of the unique sequences had significant similarities to known genes, allowing the identification of 14 776 unique genes in catfish. Ongoing research includes a comprehensive transcriptome project that utilizes high-throughput illumina sequencing, with both pooled cDNAs from many tissues and samples from tissues important to the immune system. Such transcriptome information will also greatly facilitate annotation of the reference sequence assembly.
Very recently, USDA approved funding for the generation of a draft whole genome sequence for catfish using the next-generation sequencing technology. In this project, a doubled haploid channel catfish [103] will be used as the template, simplifying the complications caused by genome duplication. The launch of the catfish whole genome sequencing project is historic. The genome sequence will not only permit efficient identification of sequence polymorphism, but also allow their in silico mapping. The genome sequence assembly will facilitate: (1) genome-wide comparative analysis, enabling scientific insight into evolutionary biology and genome function; (2) establishment of gene orthologies with other vertebrate species, which will greatly facilitate functional genomic studies and candidate gene identification for performance traits; and (3) haplotype mapping to provide the basis for comprehensive understanding of genetic diversity.
Several performance and production traits are very important for the catfish industry including, but not limited to, growth rate, feed conversion efficiency, disease resistance, processing yield, seinability, stress tolerance, and tolerance to low dissolved oxygen. Disease problems constitute the largest single demand of even larger numbers of polymorphic markers at the level of tens of thousands or hundreds of thousands, and by the demand of automated genotyping for a large number of markers and a large number of genomes. In the absence of a whole genome sequence, we have used EST resources to mine SNPs and identified over 300,000 putative SNPs. However, a key issue of EST-derived SNPs is the possibility of sequencing errors that are subsequently interpreted as potential SNPs by data mining. To circumvent this problem, we have established quality standards for EST-derived SNPs. It appears that two factors are the most important: the contig size and the minor sequence allele frequency. Contigs with at least four sequences with the minor allele sequence being represented at least twice provided a high rate of SNP validation. Most recently, we have started sequencing of reduced representation libraries to generate additional SNPs. However, we believe that generation of whole genome sequencing using the next-generation sequencing technology with a high level of genome coverage is the most efficient way of producing the highly needed SNPs. This work is now in progress.

### 4.45.5 Aquaculture Genomics and Performance Traits

#### 4.45.5.1 DNA Markers

Most recently, the effort of marker development was shifted to the identification of SNPs. This shift was driven by the demand of even larger numbers of polymorphic markers at the level of tens of thousands or hundreds of thousands, and by the demand of automated genotyping for a large number of markers and a large number of genomes. In the absence of a whole genome sequence, we have used EST resources to mine SNPs and identified over 300,000 putative SNPs. However, a key issue of EST-derived SNPs is the possibility of sequencing errors that are subsequently interpreted as potential SNPs by data mining. To circumvent this problem, we have established quality standards for EST-derived SNPs. It appears that two factors are the most important: the contig size and the minor sequence allele frequency. Contigs with at least four sequences with the minor allele sequence being represented at least twice provided a high rate of SNP validation. Most recently, we have started sequencing of reduced representation libraries to generate additional SNPs. However, we believe that generation of whole genome sequencing using the next-generation sequencing technology with a high level of genome coverage is the most efficient way of producing the highly needed SNPs. This work is now in progress.

### 4.45.5.2 Important Catfish Traits and the Need for Genome-Based Tools for Selection

Several performance and production traits are very important for the catfish industry including, but not limited to, growth rate, feed conversion efficiency, disease resistance, processing yield, seinability, stress tolerance, and tolerance to low dissolved oxygen. Growth rate is a primary trait for production because producers need to complete the production cycle in approximately 18 months to allow aquaculture ponds to be used with 2 years for a single crop. As feed accounts for over 50% of all variable costs, enhanced feed conversion efficiency will help to reduce feed costs and increase profit margins. Disease problems constitute the largest single cause of economic losses in aquaculture. Diseases cause major losses, up to one-third of the industry every year. Enteric septicemia of catfish (ESC) caused by Edwardsiella ictaluri and columnaris disease caused by Flavobacterium columnare are the major causes of losses. In addition, channel catfish virus disease, the protozoan parasite Ichthyophthirius multifiliis, and motile aeromonad septicemia caused by Aeromonas hydrophila can also cause major economic losses. Processing yield is important because most catfish are processed into fillet for marketing. Seinability is a very important trait because any fish that is left in the pond after seining can significantly reduce the feed conversion efficiency. Stress tolerance and low dissolved oxygen tolerance are important because these traits affect survival rates and affect energy consumption due to the use of aerators. Stress tolerance is also important immediately before processing of catfish during transportation because stress of catfish can lead to the so-called red flesh disease, a consequence of high levels of stress during harvesting and transportation immediately before processing. Although most of these traits can be
Genetic selection holds great potential to improve production efficiency in farm-raised catfish. Although current breeding stocks have retained a considerable amount of genetic variation as measured by allelic heterozygosity, outbred catfish stocks are readily accessible in freshwater systems. Molecular markers linked to favorable genetic variation can be used for broadstock selection for traits that have low heritability, or are difficult, expensive, or lethal to measure (e.g., resistance to pathogen infection and carcass yield). Marker-assisted selection depends on the identification of marker alleles that are in linkage disequilibrium with alleles that positively influence production phenotypes. Such markers can serve as proxies for desired haplotypes, but become useless if linkage disequilibrium is disrupted or if the marker allele is not present in the population of interest. These caveats are less likely as marker density increases throughout the genome, ultimately to the level of complete genome sequence. Because catfish stocks are relatively outbred compared with traditional livestock species, haplotype blocks will likely be smaller. Therefore, high marker densities are required throughout the genome in order to effectively select for beneficial haplotypes.

Although channel catfish is the major species used in aquaculture [98, 99], blue catfish possess several superior traits including high resistance against ESC disease, greater processing yield due to its uniform body shape, and high seinability due to its higher position in the water column than channel catfish. The interspecific hybrid catfish are fertile, and, as a matter of fact, higher generations of hybrid catfish have been produced. Therefore, interspecific introgression is a viable breeding strategy to introgress genes involved in the superior traits harbored by blue catfish. Toward this direction, genetic linkage mapping has been conducted in the interspecific resource families [42, 52].

In conclusion, more than two decades of animal genomics research have demonstrated that whole genome sequence is the most significant essential genome tool and resource, lack of which severely limits genetic improvements using genome-based technologies. However, genome resources in support of whole genome sequence assembly and annotation are essential in order to make the whole genome sequence highly useful. To this end, we have generated many of the needed genome resources for the assembly and annotation of the whole genome sequence in catfish. However, generation of the whole genome sequence only opens the real first step of the long march toward genetic enhancement, but by no means draws a period of it. The research community needs to be focused on aquaculture production and performance traits, take advantage of the unprecedented genome information and technology, and make real progress toward genetic improvements of aquaculture brood stocks.

References


