Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 26
Sequencing the Genome

Zhanjiang Liu

Genome sequencing literally means completely sequencing the entire genome so that all genome bases and their locations are known. It is the process of determining the exact order of the nucleotide bases making up the complete set of chromosomes of the genome. Because DNA sequencing is routine, genome sequencing appears intuitively very straightforward, as if 1,000 base pairs (bp) can be completely sequenced in a single run in a single lane, a genome of 1 billion bps would take 1 million lanes, and because the current automated sequencers can load 96 samples per run, to sequence the entire genome would appear to require only 10,416 runs. Although these intuitive numbers are correct for a 1× genome coverage, completely sequencing a genome is more complex than it appears to be. First, we know that a sequencing run can determine sequences from a short segment, usually approximately 1,000 bp, with shorter quality sequences. As we discussed in the sequencing technology chapter, sequencing of longer segments of DNA requires generation of nested overlapping clones, or primer walking. It is not practical to generate nested clones covering the entire genome, nor is it practical to sequence the entire genome using primer walking. Many genomic regions would be highly repetitive prohibiting direct sequencing. Second, genomic DNA cannot be handled in its entire length with its *in situ* state of chromosomes because sequencing requires a large number of molecules synchronized at the primer binding site. Genomic DNA is typically broken into tens of thousands of random segments of approximately 100–200 kilobase (kb) during DNA isolation from cells due to shearing. Third, from the prospect of project management, sequencing a whole genome also poses special challenges that demand special approaches and strategies.

Strategies for Whole Genome Sequencing

Three strategies have been used for sequencing genomes with large sizes such as those of the mammals: the whole genome shotgun (WGS), the hierarchical shotgun (the bacterial artificial chromosome [BAC] clone-by-clone approach), and the hybrid approach of the two. These strategies, however, were all developed based on the current sequencing technology using the Sanger's dideoxy chain termination method. The adoption of emerging sequencing technologies could lead to the development of new strategies for genome sequencing. We will first introduce the traditional strategies for genome sequencing, and then follow up with strategies on the horizon.

Whole Genome Shotgun (WGS)

Initially, two general strategies for sequencing a complete genome were used: the shotgun sequencing and the clone-by-clone sequencing. WGS sequencing shears

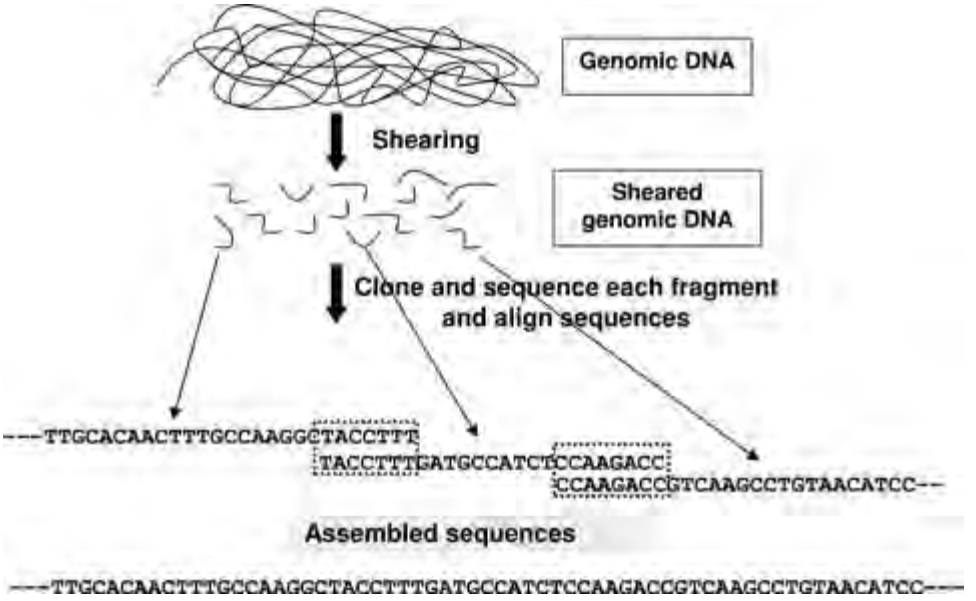

Figure 26.1. Schematic presentation of the whole genome shotgun sequencing strategy. Genomic DNA is sheared into small segments and sequenced. After sequencing, sequences are assembled into continuous sequence elements (contigs) based on overlapping. Note that usually a 6–10× genome coverage is required for the assembly of the genome, depending on the level of repetitive elements of the genome. This strategy was used by Celera for sequencing of the human genome.

genomic DNA into randomly small pieces that are cloned into plasmids and sequenced on both strands. When the sequences are obtained, they are aligned and assembled into draft sequences using bioinformatic tools (Figure 26.1).

The WGS sequencing is generally conducted by construction of shotgun sequencing libraries with different insert sizes. Most typically, the WGS is facilitated with BAC libraries, cosmid libraries, and plasmid libraries. BAC libraries typically have large inserts of 100–200 kb. Cosmid libraries hold intermediate insert sizes of approximately 40 kb. Plasmid libraries used for WGS vary between 2 and 8 kb, but most often are around 4 kb. Sequences generated from the large insert BAC libraries serve as large anchoring islands for sequence assembly and clone orientation; sequences generated from the intermediate insert cosmid libraries serve as smaller anchoring islands for sequence assembly; and the small insert plasmid libraries provide high genome coverage allowing complete sequencing of the genome. The use of large and intermediate insert libraries is very useful for genome assembly, especially for regions with highly repetitive sequences. That is because the BAC clones used for the generation of sequences are also located on the physical map constructed by BAC-based contig construction using restriction fingerprinting. The physical map information would serve as anchoring points of where the sequences are physically located, avoiding mistaken assembly based just on sequence overlaps within the repetitive region.

The key element of whole genome shotgun sequencing is to provide more random libraries with increased coverage, representing the entire genome. In order to be sure of entire genome representation, various tricks can be applied including the above-mentioned libraries with various sizes, the use of various fragmentation methodologies such as shearing, partial restriction digestion, and cloning into high and/or low copy number vectors.

The shotgun method is faster and less expensive, as compared with the hierarchical shotgun sequencing. However, it poses greater challenges for genome sequence assembly since there is no guiding "island" of sequences as there is with the clone-by-clone approach. It is more prone to errors due to erroneous assembly of the entire genome sequence. This is particularly true in genomic regions where repetitive elements are rich, often leading to the over-condensation of the genome sequence with repeats underrepresented due to mistaken assembly. WGS strategy is therefore more adaptable to sequencing of the relatively small genomes, and those with simple repeat structures. However, due to its speed and relatively lower costs, it allows a greater level of genome coverage with the same financial resources.

Hierarchical Shotgun Sequencing (the Clone-by-Clone Approach)

The hierarchical shotgun sequencing strategy is also referred to as the clone-by-clone strategy of genome sequencing (Figure 26.2). In this approach, genomic DNA is first cleaved into segments of about 100–200 kb and inserted into BAC vectors. The BAC clones containing a high level of genome coverage of the genomic DNA are collectively called the BAC library. See Chapter 13 for its construction. BAC libraries are the basis of physical mapping, as described in Chapter 14, by construction of contigs using restriction fingerprinting. Once the contigs are constructed, the use of fingerprints, and most often also BAC-end sequences (see Chapter 15), allows the identification of a set of BAC clones containing the entire genome, but with minimal overlapping. Such a set of BAC clones is called the minimal tiling path (MTP), which is defined as a minimally overlapping set of all clones in the physical map. Each BAC DNA in the MTP is fragmented randomly into smaller pieces and each piece is cloned into a plasmid and sequenced on both strands. These sequences are aligned so that identical sequences are overlapping. These contiguous pieces are then assembled into the finished sequence of the BAC, and the physical map information and BAC-end sequences are then used to assemble the entire genome sequence. The assembly of the genome sequence usually requires 5–10× genome coverage.

The advantage to the hierarchical shotgun approach is its lower level of mistakes when assembling the shotgun sequences into contigs for the generation of genome sequence. The reason is that the chromosomal location for each BAC is known from the physical map, and there are also fewer random pieces to assemble within each BAC. Also, the complexity of repetitive elements within a single BAC should be much lower than that for the whole genome. The weakness of this method is its low speed of sequence generation and its high costs. Once produced, the genome sequences generated using the hierarchical shotgun approach are regarded highly in their qualities.

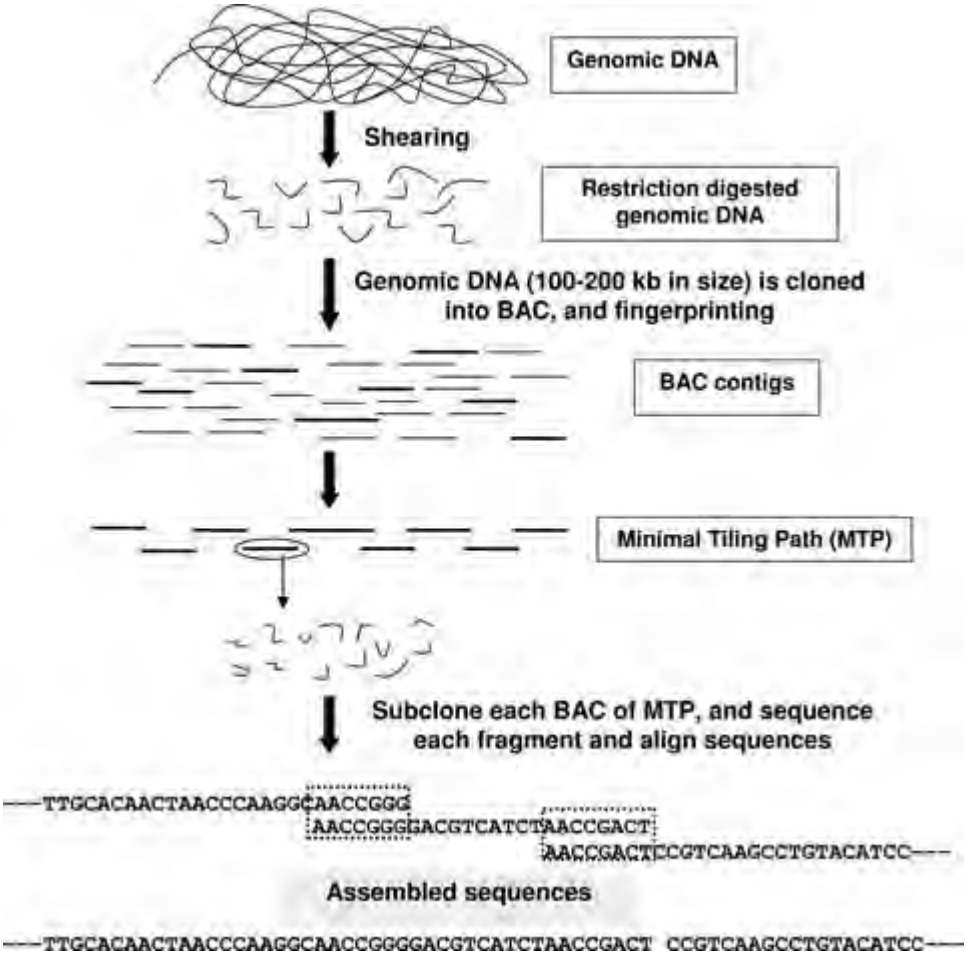

Figure 26.2. Schematic presentation of the hierarchical shotgun sequencing strategy. Genomic DNA is partially digested with restriction enzyme to produce segments of 100–200 kb that are cloned into the BAC vector for contig construction using fingerprinting. Based on the fingerprints (and often also the BAC-end sequences), a minimal tiling path (MTP) can be identified for clone-by-clone sequencing of the genome. Each selected BAC in MTP is further subcloned and shotgun sequenced for the assembly of the sequence of the BAC. The whole genome sequence is assembled from the complete BAC sequences of the MTP BAC clones. This strategy was used by the publicly funded Human Genome Project.

The Hybrid Approach

Because the whole genome shotgun and the hierarchical shotgun approaches are complementary to each other in many aspects, using one approach would alleviate the problems of the other. This gives great motivation for the use of both approaches for the production of a whole genome sequence. The relative proportion of the two approaches used in a genome project depends on the genome size, repeat structure of

the genome, the physical map quality, the availability of BAC-end sequences, and the availability of the MTP set of BAC clones, as well as financial considerations.

Selection of Minimal Tiling Path Clones

Selecting an MTP is the task of picking a set of minimally overlapping clones that span an entire contig. Many of the approaches used for picking MTP clones were developed by the Arizona Genomics Computational Laboratory (http://www.agcol.arizona.edu/), along with the fingerprinting and construction of BAC contigs software package FingerPrint Contig (FPC) (Soderlund et al. 1997, 2000). Their Web page provides a rich source for physical mapping and MTP selection information. Readers interested in the details are referred to their published papers, as well as to their Web site. I will briefly introduce the principles for the selection of MTP BAC clones.

Because of inexact coordinates in the Consensus Bands (CB) map, MTP clones cannot be selected based solely on their positions on the map. Three methods are currently used for picking MTP clones:

1. The fingerprint method, in which overlaps are determined by examination of the clone fingerprints and their map position.This method has been used for the selection of MTP clones for the *Caenorhabditis elegans* project (Coulson et al. 1986), and it was demonstrated to be effective. However, the overlapping regions among the BAC clones were large (e.g., approximately 47.5 kb overlap for the minimal tiling path picked by the International Human Genome Consortium using this approach). This is because the fingerprints were produced using 6-bp cutters of restriction enzyme that cut DNA at a rate of once every 4,096 bp. It requires a large number of overlapping fragments (usually 10–12) to be statistically significant. Recently, four-colored fluorescent fingerprinting has been adopted for physical mapping (Luo et al. 2003), which should greatly reduce the overlapping sizes in the MTP clones while maintaining statistical significance. This is because in this fingerprinting approach, four sets of fingerprints are used for contig construction.
2. The BAC-end sequence method. This approach was initially proposed in 1996 (Venter et al. 1996) using BAC-end sequences as tagged sequence connectors. In this method, BAC-end sequences are first produced (see Chapter 15) as a genome resource for the genome project. When one BAC clone is completely sequenced, its entire sequence is aligned to all BAC-end sequences to identify the clones with minimal sequence overlaps. This approach was found to be very effective in reducing the overlapping regions of MTP clones. However, false positive clones can be identified within repetitive element regions. This is because BAC-end sequences are often short (usually 400–600 bp). When coupled to the relatively low quality of BAC-end sequences in many species, the likelihood of mistaken identification of overlapping sequences is increased, especially within repetitive genomic environs.
3. The hybrid approach of the two methods. It is clear from the above discussion that the two approaches of MTP selection use different resources. The first method involves analyzing the fingerprints of a pair of overlapping clones for shared restriction fragments (bands), and verifying the integrity of the fingerprints of the potentially overlapping pair by matching bands with a spanning and two flanking clones.

In the second method, draft sequence of a BAC is used to determine which additional BAC clones have the minimal overlapping with the completely sequenced BAC clone. The first method uses information that is already present in the physical map, but the overlaps can be inexact. The second method requires sequence information, but gives very exact overlaps (http://www.agcol.arizona.edu/). When both the fingerprints and the BES resources are used in the hybrid approach, the precision should be increased while the overlapping regions are reduced, leading to a reduced sequencing load for genome projects. Software such as MTP is available from Arizona Genomics Computational Laboratory using both sets of information for efficient selection of MTP (http://www.agcol.arizona.edu/software/fpc/userGuide/mtpdemo/#intro).

Emerging Sequence Technologies and Platforms

The above discussion is based on current sequence technologies. However, several promising sequencing strategies are poised to make technical advances to allow them to be effectively used for genome sequencing. A couple of these technologies, such as the 454 sequencing and the Solexa sequencing platforms, as discussed in the previous chapter on sequencing technologies, strive to demonstrate its high efficiency, high throughput, and low cost. Their application could fundamentally change how genome sequencing is conducted, not only at the level of technicalities, but also at the level of psychology and decision making. Their application for genome sequencing could have a revolutionary impact on life sciences, and bring a real opportunity to sequencing aquaculture genomes.

Sequencing the Genome

After the strategy is determined, sequencing of a genome itself is straightforward. The important aspect of large-scale sequencing is the streamlined operation with accurate tracking and recording, as well as the bioinformatics pipeline. In almost all cases, genome sequencing projects have been conducted at large genome centers or by biotechnology industries, both of which not only have the state-of-the-art sequencing equipment but also expertise gained in genome sequencing for management of large sample sets and data sets. Importantly, they are well equipped with bioinformatics capabilities.

Genome Sequence Assembly

Current sequencing technologies only allow an average read length of 1,000–1,500 bp per run, with the first approximately 500–800 bp as high-quality sequences. As discussed above, to overcome this limitation, sequencing of entire genomes is performed through either whole genome shotgun sequencing or the clone-by-clone hierarchical shotgun sequencing. In either case, DNA is sheared into smaller fragments whose

ends are then sequenced. The generated sequences need to be assembled using computer programs called assemblers. The output of assembly programs consists in a collection of contiguous sequence pieces (contigs). They are rarely, if ever, entire chromosomes reconstructed into a single contig, but many smaller pieces. Additional computer programs called the scaffolder use the information linking together sequencing reads from the ends of fragments to order and orient the contigs with respect to each other along a chromosome.

Sequence assembly is essentially a set of contigs, each contig being a multiple alignment of reads (Dear et al. 1998). The assembler relies on the basic assumption that two sequence reads that share the same string of letters originated from the same place in the genome. Mathematical modeling indicated that an 8–10 genome coverage should allow the vast majority of sequences to be assembled into large contigs (more than 200,000 bp) (Lander and Waterman 1988).

Ideally, genome sequence assembly should produce one contig for every chromosome of the genome being sequenced. In reality, however, many contigs are produced as a result of a combination of factors. Nonrandom shearing and the presence of repeats are the two greatest challenges. Even at eightfold to tenfold coverage, some portions of the genome remain unsequenced as gaps.

A number of different strategies have been proposed to deal with genome sequence assembly (Peltola et al. 1984, Huang 1996, Parsons et al. 1993, Bonfield et al. 1995, Notredame and Higgins 1996, Zhang and Wong 1997, Ewing and Green 1998, Ewing et al. 1998). Currently, there are mainly two different existing approaches for assembling sequences: (1) the iterative and (2) the All-in-One-Step approach. The first type of assembly is essentially derived from the fact that the data analysis and reconstruction approximation algorithms can be parametrized differently, ranging from very strict assembly of only the highest quality parts to very 'bold' assembly of even lowest quality stretches. An assembly starts with the most strict parameters, having the output edited manually by highly trained personnel, or by software and then the process is reiterated with less strict parameters until the assembly is finished (http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html). The second approach has been made popular by the PHRAP assembler presented by Phil Green (http://www.phrap.org/). This assembler uses low and high-quality sequence data from the start and generates a consensus by puzzling its way through an assembly using the highest quality parts as reference, giving the result to a human editor for finishing. An integrated approach of the two approaches has been developed by Chevreux and others (http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html).

The fundamentals of genome sequence assembly are the same as assembly of short DNA sequence of a few kb. The quality of the sequences in base calling is the most important (Figure 26.3). However, the genome sequence assembly is much more complex because of the large sizes and similar sequences exist here and there in the genome by chance. In highly repetitive genomic regions, sequence assembly can be a disaster. For accurate genome assembly, the clone-by-clone hierarchical shotgun sequencing is superior to the whole genome shotgun sequencing (Figure 26.4). If nothing else, two aspects of the clone-by-clone approach make it easier for sequence assembly. First, the region under consideration is a BAC clone, most often less than 200 kb in size, as compared to the whole genome of billions of bps. Second, the repetitive regions within a single BAC can be much simpler than the situation of facing a whole genome. The coincidence of similar sequences by chance is much smaller with a

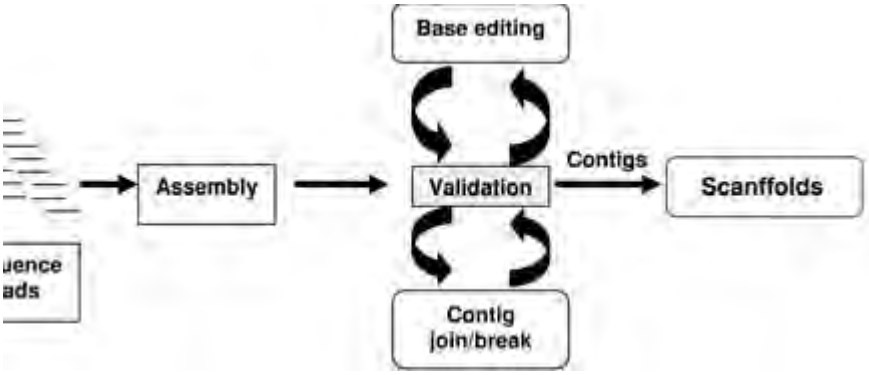

Figure 26.3. Schematic flow chart for sequence assembly processes.

single BAC than with a whole genome. In addition, gap filling is much easier for the clone-by-clone approach with known regions for higher levels of genome coverage in additional sequencing. In contrast, just increasing sequencing genome coverage may or may not easily fill the gaps.

Draft Genome Sequence and Finishing

Sequencing the bases equal to the genome size is defined as one genome coverage. Obviously, the higher the genome coverage, the greater the possibility for assembly of

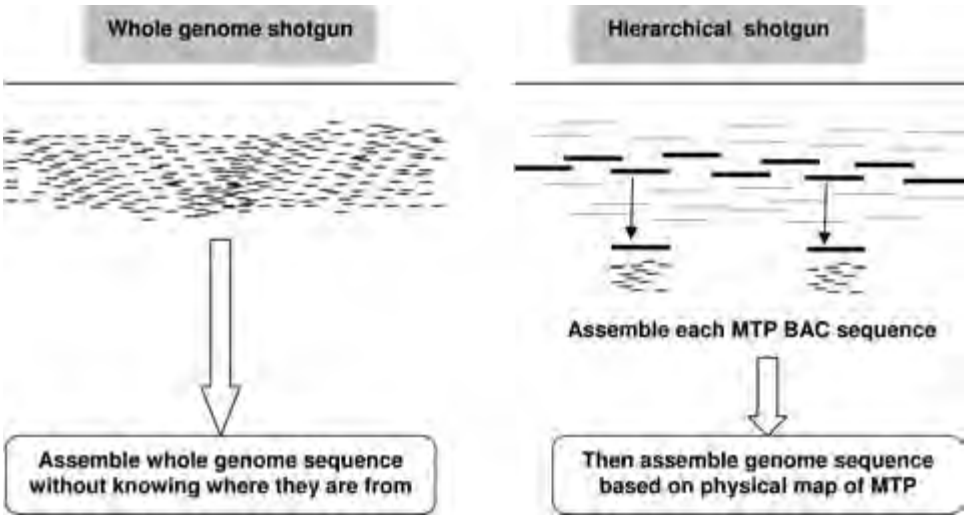

Figure 26.4. The comparison of the whole genome shotgun and the hierarchical shotgun sequencing approaches in sequence assembly. Short lines are individual sequence reads. Long lines are BAC clone inserts. Thick long lines are selected BAC clones with minimal overlapping (minimal tiling path) for sequencing.

the complete genome sequence. However, the greater the coverage is, the higher the costs. This dilemma must be balanced to provide a reasonable accurate assembly, but keep the costs as low as possible. Usually, it takes 4–5× genome coverage to assemble genome sequence into reasonably large segments (usually larger than 10,000 bp). Such an assembly is referred to as the draft genome sequence, or in other words, the sequence assembly is preliminary. Using a genome sequencing coverage of 4–5×, the human draft sequence was released in June 2000. To generate the high-quality human reference sequence, additional sequencing was needed to close gaps, reduce ambiguities that allow for only a single error every 10,000 bases. A high-quality sequence is critical for recognizing regulatory components of genes that are very important in understanding human biology and such disorders as heart disease, cancer, and diabetes. After sequencing of 8–9 genome coverage, the human genome sequence was finished in 2003.

The ultimate goal of any sequencing project is to determine every single bp of the original set of chromosomes. However, it is rare for an assembly program to be able to reconstruct a single piece of DNA per chromosome, leading to gaps in the reconstruction of the genome. These gaps are filled in through directed sequencing in a process called finishing or gap closure. Gap filling allows merge of sequence contigs into supercontigs and scaffolds (Figure 26.3). Finishing is the process whereby additional sequences are obtained to achieve full coverage and continuity over very long distances by filling in the gaps. Only the Human Genome Project is undertaking finishing. Gap filling sounds easier than it is, but requires great effort and effective strategies.

Gap Filling Using Expressed Sequence Tag (EST) Resources

In most cases, a large EST database exists for the genome being sequenced. After initial sequencing for the generation of draft sequence, many smaller segments of sequence reads are left unassembled because there are no overlapping sequences, or in other words, there are gaps. One of the strategies for gap filling is to use the leftover unassembled sequences to search the EST database of the species, allowing identification of exon segments. The exon sequences can be used to design the primer for direct BAC sequencing.

Gap Filling by Creating Libraries with Variable Sizes

In many cases, the existence of gaps could be attributed to the methods used for the construction of genome sequencing libraries, either because of the restriction enzymes used, or because of the selection of certain insert sizes of the libraries. By changing the restriction enzymes used for the construction of sequencing plasmid libraries, or increasing the average insert sizes, it is possible to close the gaps, allowing continuous assembly of the genome sequence. Therefore, planning the use of several libraries with various size distributions is beneficial.

Gap Filling by Primer Walking Sequencing or by Transposon Insertion into BACs

Primer walking sequencing is highly useful for gap filling, especially for sequences generated with the clone-by-clone approach. In the approach, the regions with obvious gaps are known as BAC clones, and primers can be designed immediately following the known sequences to extend sequencing.

The BAC clones can also be further sequenced by creating transposon insertion sublibraries. In this case, transposon sequences are randomly inserted into the BAC where gaps exist, creating a sublibrary. The transposon sequences provide primer-binding sites for sequencing the neighbor regions. Kits for making transposon insertion libraries are available (e.g., the Tn5 transposon insertion system of EPICENTRE (http://www.epibio.com/a_simple_invitro_transposition_reaction.asp). The transposon insertion clone is particularly useful for genomic regions containing highly repetitive or simple sequences prohibiting the design of sequencing primers. Obviously, such regions can also be dealt with PCR amplification followed by cloning into a plasmid vector and sequencing.

Scaffolding

Scaffolding is the contig merging processes using not only sequencing information, but also the mate pair sequencing reads (Figure 26.5). The contigs produced by an assembly program can be ordered and oriented along a chromosome using mate pair sequencing reads. Mate pair sequencing reads are sequencing reads obtained from both ends of a single clone. For instance, sequencing of a BAC from both ends would generate mate pair reads that are physically linking them together by the distance of the BAC insert size. Similarly, mate pair reads are obtained by sequencing both ends

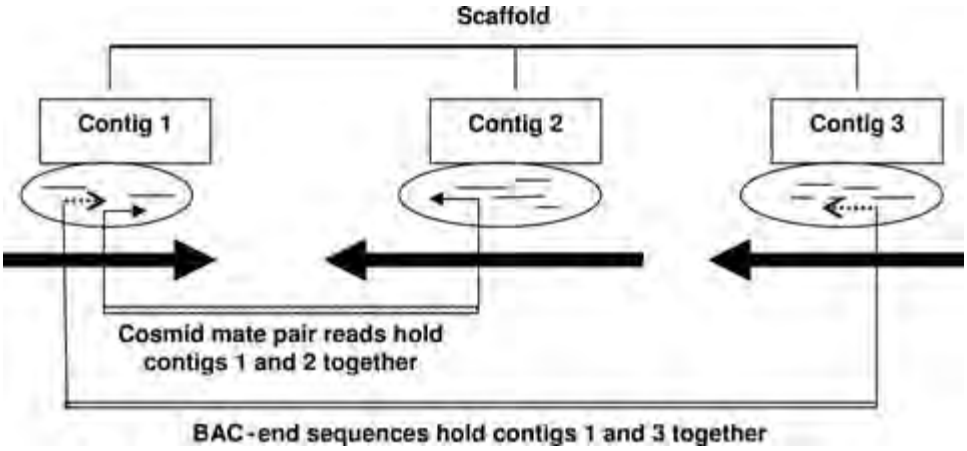

Figure 26.5. Schematic presentation of scaffolding using mate pair sequences. Shown are examples of mate pair sequences from a BAC and a Cosmid clone.

of cosmid libraries or plasmid libraries. Within the assembly, the paired end reads must be placed at a distance consistent with the size of the library from which they originate and must be oriented toward each other. Within an assembly, each read is assigned an orientation corresponding to the DNA strand from which the read was generated. The constraints provided by mate pairs lead to constraints on the relative order and orientation of the contigs. In such a process, the mate pair reads are used to order and orient the contigs along a chromosome, combining them into scaffold.

Sequence Annotation

After the initial generation of A, C, G, Ts, the raw sequence data provide very little biological insight. Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, gene signals such as promoters, etc. To use any sequence it must be interpreted in the context of other biological knowledge. This is the process of annotation, the task of adding explanatory notations to the sequence text. We define an annotation as the biological evaluation and explanation of a specific region on a nucleic acid sequence that includes any feature that can be anchored to the sequence, such as gene transcripts, an exon, a promoter, a transposable element, a regulatory region, mutations, cut sites, start and stop signals, transcription factor binding sites, probe or primer binding sites, or a CpG island (Lewis et al. 2002).

Sequence annotations are performed using powerful computer programs. Many programs were developed for the annotation of DNA sequences. Some examples include Genotator (Harris 1997), Artemis (Rutherford et al. 2000), AceDB (http:// www.acedb.org), Apollo (Lewis et al. 2002), Genamics Expression (http://genamics.com/expression/annotating.htm), EAnnot (Ding et al. 2004), the distributed annotation system (DAS) (http://stein.cshl.org/das/), etc. For the aquaculture genome community, much can be learned from the human genome research community and the recent sequencing projects of livestock animals such as chicken, bovine, and swine sequencing projects. In most cases, however, genome annotation is a complex task requiring assistance from large genome sequencing centers, or bioinformatics experts. The biological information, however, most often comes from the research community.

What is Next with A, C, G, Ts?

After the generation of genome sequence, it is not the end of genome biology. As described at the U.S. Department of Energy's (DOE) Genome Project Web page, "The words of Winston Churchill, spoken in 1942 after 3 years of war, capture well the HGP era: 'Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.' The avalanche of genome data grows daily. The new challenge will be to use this vast reservoir of data to explore how genes are expressed and their networking, and interactions with one another and with the environment to create complex, dynamic living systems. Systematic studies of function on a grand scale, i.e., functional genomics, will be the focus of biological explorations in

this century and beyond. Deriving meaningful knowledge from DNA sequence will define biological research through the coming decades and require the expertise and creativity of teams of biologists, chemists, engineers, and computational scientists, among others.

Ethical, Legal, and Social Issues (ELSI) of Genome

Sequencing coming into the genomics era also brings with it many ethical, legal, and social issues (ELSI). This is sometimes also referred to as the GE3LS (genomics, ethics, environment, economics, law, and society). The ELSI is serious enough that the National Institutes of Health (NIH) and DOE spend 3–5% of their genome research budget to address such issues. Many questions arise as a result of genome sequence availability. Although we may not have answers for many of these questions, some examples of the questions posted on DOE's Genome Project Web site (http://doegenomes.org/) are relevant to many: *Who should have access to personal genetic information, and how will it be used? Who owns and controls genetic information? How does personal genetic information affect an individual and society's perceptions of that individual? How does genomic information affect members of minority communities? Do healthcare personnel properly counsel parents about the risks and limitations of genetic technology? How reliable and useful is fetal genetic testing? What are the larger societal issues raised by new reproductive technologies? How will genetic tests be evaluated and regulated for accuracy, reliability, and utility? (Currently, there is little regulation at the federal level). How do we prepare healthcare professionals for the new genetics? How do we prepare the public to make informed choices? How do we as a society balance current scientific limitations and social risk with long-term benefits? Should genetic testing be performed when no treatment is available? Should parents have the right to have their minor children tested for adult-onset diseases? Are genetic tests reliable and interpretable by the medical community? Do people's genes make them behave in a particular way? Can people always control their behavior? What is considered acceptable diversity? Where is the line between medical treatment and enhancement? Are GM foods and other products safe to humans and the environment? How will these technologies affect developing nations' dependence on the West? Who owns genes and other pieces of DNA? Will patenting DNA sequences limit their accessibility and development into useful products? etc., etc.*

Unfortunately, we do not have any genomes sequenced today from aquaculture species, but fortunately, the aquaculture genome sequence will not involve an equal number of challenges as the human genome sequence in relation to the ethical, legal, and social issues. However, many of the ethical, legal, and social issues will be similar to the questions above facing the human genome situation. By the time the genomes are sequenced from aquaculture species, hopefully answers will be found for many of the above questions. On the other hand, because the aquatic environments and many of the unique characteristics of aquaculture species, and their close relationship with the environment, the generation of aquaculture species genome sequence likely will pose challenging questions related to ethical, environmental, economic, legal, and social issues.

Conclusion

Although no genomes have been sequenced from aquaculture species, it is likely that some important aquaculture species will soon be subjected to genome sequencing. (After this chapter was written, NIH made a committment to sequence the tilapia genome.) The major reasons for this assessment include their economic importance, their close relationship with the environment, their roles in the study of genome evolution, and many unique characteristics of aquaculture species. In addition, given the declining world fisheries, aquaculture needs to be developed and sustained, which demands genome-based technologies. Many biological and production problems are unique to aquaculture species, and sequencing of related model species will not provide relevant information. For instance, the enteric septicemia of catfish is unique in catfish, and the zebrafish genome information may not help in elucidation of the disease resistance genes in catfish.

However, the low research funding, the relatively small research community, and the large numbers of species involved in aquaculture make the initial funding of genome sequencing projects difficult. There are no limitations in technology, biology, or readiness for genome sequencing, because basic genome resources are well prepared from several important aquaculture species including the salmonids, catfish, tilapia, shrimp, and the oysters. The limiting factor is financial. The hard part is to obtain the first pot of money. Once the first pot of funds is obtained, international collaboration is expected to meet the remaining demands.

The emerging sequencing technology such as the 454 sequencing platform brings greater hope to the aquaculture species for genome sequencing, as sequencing with a genome coverage of 5× cost several hundred thousand dollars that may be justified with funding agencies. The problem is with the ability of sequence assembly with the short sequencing reads of the 454 system. Let's hope the technology will improve soon. On the other hand, many "sequence islands" as produced by the 454 sequencing platform, when anchored by well-developed physical maps and genetic maps, may provide sufficient tools for studies of agricultural related issues such as performance and production traits.

References

Bonfield JK, KF Smith, and R Staden. 1995. A new DNA sequence assembly program. Nucleic Acids Res, 23, pp. 4992–4999.

Coulson A, J Sulston, S Brenner, and J Karn. 1986. Towards a physical map of the genome of the nematode *C. elegans.* Proc Natl Acad Sci, 83, 7821–7825.

Dear S, R Durbin, L Hilloier, G Marth, J Thierry-Mieg, and R Mott. 1998. Sequence Assembly with CAFTOOLS. Genome Res, 8, pp. 260–267.

Ding L, A Sabo, N Berkowicz, RR Meyer, Y Shotland, MR Johnson, KH Pepin, RK Wilson, and J Spieth. 2004. EAnnot: a genome annotation tool using experimental evidence. Genome Res, 14, pp. 2503–2509.

Ewing B and P Green. 1998. Basecalling of automated sequencer traces using Phred. II. Error probabilities. Genome Res, 8, pp. 186–194.

Ewing B, L Hillier, MC Wendl, and P Green. 1998. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res, 8, pp. 175–185.

Harris NL. 1997. Genotator: a workbench for sequence annotation. Genome Res, 7, pp. 754–762.
Huang X. 1996. An improved sequence assembly program. Genomics, 33, pp. 21–31.
Lander ES and MS Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2, pp. 231–239.
Lewis SE, SM Searle, N Harris, M Gibson, V Lyer, J Richter, C Wiel, L Bayraktaroglir, E Birney, MA Crosby, JS Kaminker, BB Matthews, SE Prochnik, CD Smithy, JL Tupy, GM Rubin, S Misra, CJ Mungall, and ME Clamp. 2002. Apollo: a sequence annotation editor. Genome Biol, 3, RESEARCH0082.
Luo MC, C Thomas, FM You, J Hsiao, S Ouyang, CR Buell, M Malandro, PE McGuire, OD Anderson, and J Dvorak. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics, 82, pp. 378–389.
Notredame C and DG Higgins. 1996. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res, 24, pp. 1515–1524.
Parsons R, S Forrest, and C Burks. 1993. Genetic algorithms for DNA sequence assembly. ISMB, pp. 310–318.
Peltola H, H Soderlund, and E Ukkonen. 1984. SEQAID: a DNA sequence assembling program based on a mathematical model. Nucleic Acids Res, 12, pp. 307–321.
Rutherford K, J Parkhill, J Crook, T Horsnell, P Rice, MA Rajandream, and B Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics, 16, pp. 944–945.
Soderlund C, S Humphray, A Dunham, and L French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res, 10, pp. 1772–1787.
Soderlund C, I Longden, and R Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci, 13, pp. 523–535.
Venter JC, HO Smith, and L Hood. 1996. A new strategy for genome sequencing. Nature, 381, pp. 364–366.
Xu P, SL Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. Animal Genetics, 37, pp. 321–326.
Zhang C and AK Wong. 1997. A genetic algorithm for multiple molecular sequence alignment. Comput Appl Biosci, 13, pp. 565–581.