Aquaculture Genome Technologies
Zhanjiang (John) Liu
Copyright © 2007 by Blackwell Publishing

Chapter 15
Physical Characterization of Genomes Through BAC End Sequencing

Peng Xu, Shaolin Wang, and Zhanjiang Liu

BAC (bacterial artificial chromosome) libraries are large-insert genomic libraries suitable for physical mapping, as well as serving as the basis for clone-by-clone whole genome sequencing. They are important genome resources that can be exploited for other purposes. Particularly for aquaculture genomics, where whole genome sequencing is not likely for most of the hundreds of species used in aquaculture, BAC libraries provide opportunity to gain much genomic information that would be otherwise difficult to obtain. The construction of BAC libraries and BAC-based fingerprinting for contig construction are covered in Chapter 13 and Chapter 14, respectively. In this chapter, we will briefly describe physical characterization of the genome through BAC-end sequencing analysis, and the value of BAC-end sequences for genome analysis.

Generation of BAC End Sequences

BAC-end sequences can be generated by direct sequencing of BAC clones using sequencing primers designed based on the BAC vector sequences at the border of the genomic insert. Typically, Sp6 and T7 sequencing primers can be used because these sequencing primer sequences have been incorporated into the BAC vectors. The sequencing reactions are straightforward using the dideoxy chain termination sequencing reactions (Sanger's sequencing method) except that a large number of cycles is required for cycle sequencing, usually 80–100 cycles (Xu et al. 2006) because of low copy numbers of BAC DNA.

Although BAC-end sequencing is straightforward, it is important to emphasize the significance of tracking and quality issues. As long-term genome resources, BAC-end sequences must be properly tracked with great quality assurances. Resequencing of a small fraction of the clones is generally recommended. For instance, eight clones can be resequenced from each 384-plate from positions A1, A2, B1, B2, C1, C2, D1, and D2, or better off, 16 clones can be sequenced from the diagonal positions of A1, B2, C3, D4, E5, F6, etc. Quality assessment can be performed using the raw chromatogram files directly before any trimming. The raw, untrimmed files can be processed by Phred software (Ewing and Green 1998, Ewing et al. 1998). Phred quality score cut off value is usually set at 20 for the acquisition of Q20 values. A Q20 curve can be generated to exhibit the length distribution of sequences that passed the Q20 threshold.

Sequence Processing and Routine Bioinformatic Analysis of BAC-end Sequences

BAC-end sequences need to be processed and submitted to GSS database of the GenBank. An example of the routine analysis before sequence submission is shown in Figure 15.1. The BAC-end sequences need to be trimmed of vector sequences and filtered of bacterial sequences, stored in a local Oracle database after base calling and quality assessment. We have used the Genome Project Management System, a local laboratory information management system, for large-scale DNA sequencing projects (Liu et al. 2000). Quality assessment was performed using Phred software (Ewing and Green 1998, Ewing et al. 1998) using $Q \geq 20$ as a cutoff. Repeats were masked using RepeatMasker software (http://www.repeatmasker.org) before Basic Local Alignment Search Tool (BLAST) analysis.

BLASTX searches of the repeat masked BES were conducted against Non-Redundant Protein database. A cut off value of e^{-5} was used as the significance similarity threshold for the comparison. The BLASTX result was parsed out in a tab-delimited format. In order to anchor the catfish BES to zebrafish and *Tetraodon* genomes, BLASTN searches of the repeat masked catfish BES were conducted against zebrafish

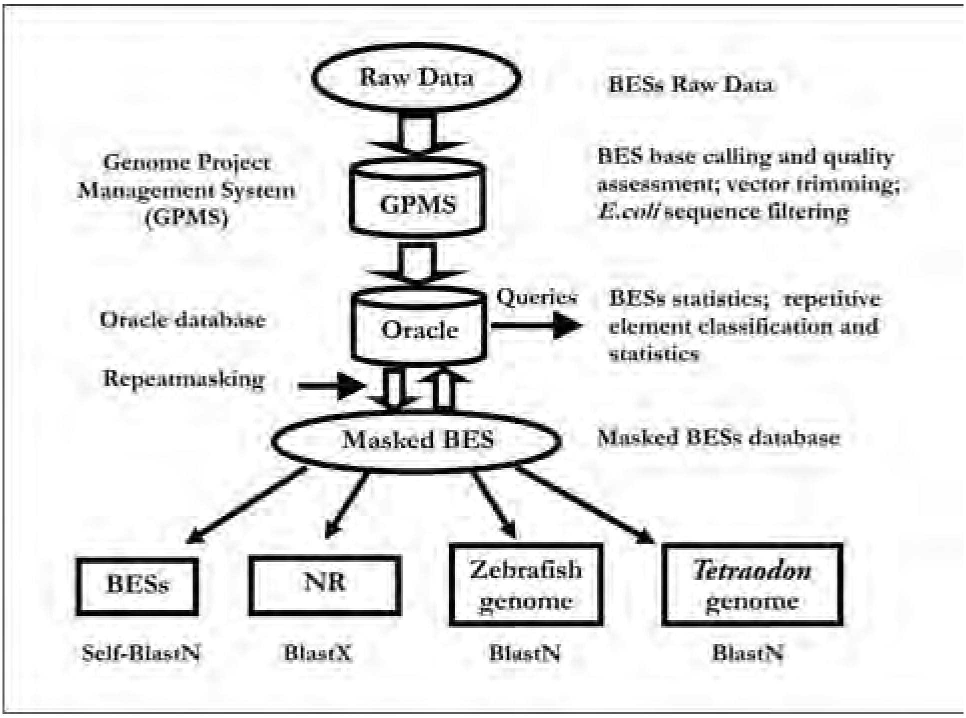
Raw Data
BESs Raw Data
Genome Project Management System (GPMS)
GPMS
BES base calling and quality assessment; vector trimming; *E.coli* sequence filtering
Oracle database
Oracle
Queries
BESs statistics; repetitive element classification and statistics
Repeatmasking
Masked BES
Masked BESs database
BESs
NR
Zebrafish genome
Tetraodon genome
Self-BlastN
BlastX
BlastN
BlastN

Figure 15.1. An example of routine informatic analysis of BAC-end sequences (BES) before their submission to GSS of the GenBank.

and *Tetraodon* genome sequences. The location and chromosome number of each top hit were collected from the results and parsed out in tab-delimited format.

Microsatellites and other simple sequence repeats were analyzed by using Repeatmasker as well as by using Vector NTI Suite 9.0 (Invitrogen, Carlsbad, CA) as we previously described (Serapion et al. 2004). *Microsatfinder (http://www.eusebius.mysteria.cz/microsatfinder/index.php)* is another user-friendly program for the identification of microsatellites within known sequences.

BAC End Sequences Provide an Unbiased Survey of Genomic Sequences

The average insert size of BAC libraries is usually 100–200 kilobase (kb), and such inserts are usually prepared by partial restriction digest of genomic DNA. Assuming restriction sites are randomly distributed (which would be expected to be the case except within repetitive elements, but often proved to be not), then the BAC ends represent random genomic sequences. For the sake of understanding, we will use the catfish BAC-end sequencing as an example. Channel catfish have a genome size of 1×10^9 base pairs (bp). One of its BAC libraries, CHORI 212 (http://bacpac.chori.org/catfish212.htm) has an average insert size of 160 kb, and thus every 6,250 BAC clones cover one genome. Sequencing 6,250 BAC clones from both ends would effectively generate a sequence tag per 80 kb of the genomic DNA on average. In most cases, BAC libraries have a multiple fold (at least 6–8 folds) of genome coverage to assure a near complete coverage of the entire genome. For instance, the channel catfish CHORI 212 BAC library has a 10.6× genome coverage with 72,067 clones. Sequencing all of the BAC ends from this library would generate one sequence tag per 8 kb genomic DNA, on average. Every BAC-end sequencing reaction can easily produce 500–600 bp. Sequencing all of the BAC ends of the CHORI 212 BAC library should generate 7–8% of 1× genome sequences. With exception of its expense, BAC-end sequencing can rapidly and effectively produce a reasonably unbiased survey of the genome of interest. Such a sequence survey should allow estimation of A/T (G/C) content of the genome, assessment of the repeat structure of the genome, discovery of the microsatellite markers for linkage mapping, production of genomic resources for comparative mapping, and virtually mapping genes to BACs.

Assessment of Repeat Structure of the Genome from BAC End Sequences

The repeat structure of a genome can be assessed from BAC-end sequences. Most often, such an assessment can be accomplished by answering the following two questions:

1. What is the fraction of the genomic sequence survey sharing repeat sequences with species from which the entire genome sequences are available?
2. What types of species-specific novel repeat sequences exist in the species under study?

For the first question, the answer can be derived from analysis of the BAC-end sequences using RepeatMasker (http://www.repeatmasker.org/), a program for the masking of repetitive sequences using various repeat libraries. For instance, analysis using the 11,414,601 bp channel catfish BAC-end sequences resulted in 10.86% masked using *Danio* repeat database, and 7.31% masked using *Takifugu* repeat database. Use of both *Danio* and *Takifugu* repeat databases masked 11.91% (see Chapter 16). This suggested that teleost fishes share a high level of repetitive elements and also that a significant fraction of taxa-specific repeats exists. The analysis indicated that a larger pool of repetitive elements is shared between the genomes of zebrafish and catfish than between the fugu and catfish genomes. It is also obvious that the catfish and the zebrafish genomes harbor a larger percentage of repetitive elements than the fugu genome, an expected result because the fugu genome is much more compact.

Analysis of the catfish BES against the human repeat database suggested the presence of a significant fraction of similar repetitive elements between teleost fish and mammals. In addition to the expected simple repeats (2.7%) and low complexity repeats (1.1%), the RepeatMasker masked 1.08% of the channel catfish BES in the category of DNA repetitive elements, of which the vast majority (1.06%) were the MER2 type of repeats.

The identification of novel repetitive elements in the genome can be approached by bioinformatic analysis of the BAC-end sequences. Several old computer software packages are available (Devereux et al. 1984, Agarwal and States 1994, Rivals et al. 1997). However, most of the earlier programs have a limit on the maximal sequence that can be analyzed. For instance, the Repeat Finder of the GCG package (version 7.0) has a maximal sequence limit of 350,000 bp (Kurtz and Schleiermacher 1999). BLAST (Altschul et al. 1997) and MegaBLAST (Zhang et al. 2000) are quite efficient in the analysis of sequences, but are also limited by sequence size they can process. For the identification of novel repeats in the catfish genome, we have used self BLASTN to detect sequences that share a high level of similarity with other sequences, an indication of repetitive elements. The cut off e-value was set at e^{-2}, and the identify threshold was set to 90% with a minimal alignment length of 100 bp. The BLASTN results were parsed to tab-delimited format to count the redundant queries and other statistics.

Several recently developed software programs can handle large data sets up to the size of the human genome for the identification of repeats. These programs include *REPuter* (Kurtz and Schleiermacher 1999, Kurtz et al. 2000), MUMmer (Delcher et al. 1999), and FORRepeats (Lefebvre et al. 2003). The *REPuter* program (http://www.genomes.de/) is a powerful program for the analysis of repeats in large data sets. The search engine *REPfind* of *REPuter* uses an efficient and compact implementation of suffix trees in order to locate exact repeats in linear space and time. These exact repeats are used as seeds from which significant degenerate repeats are constructed allowing for mismatches, insertions, and deletions. The user has the option for defining the parameters of the search including minimum length and maximum number of errors (sequence divergence). Most often, a minimum length of 20 bp and a 10% error (90% sequence conservation) should be highly sufficient for the detection of repetitive sequences. The output is sorted by significance scores (*E*-values). In addition to finding degenerate direct repeats, *REPfind* is capable of detecting degenerate palindromic repeats (Kurtz et al. 2001). The FORRepeats is more suitable for comparison of genomes between species.

In spite of the fact that BAC-end sequences can be used to effectively estimate repeat structures, there are limitations to using BAC-end sequences for the characterization of the repeat structure of the genome, usually leading to the underestimation of the repeats in the genome. The major problem is its inability to identify the tandem repeats that had not been cloned into the BAC library. For instance, we previously described the presence of a major class of tandem repeats named *Xba* elements (Liu et al. 1998) that accounted for about 5% of the catfish genome. These elements were not detected in the BES, because they lack the *Eco*R I restriction sites necessary for insertion into BAC clones.

Identification of Microsatellites from BAC-end Sequences

BAC-end sequences are a genome resource that can be used to mine microsatellite markers. As a matter of fact, they are rich in microsatellites. For instance, during the analysis of 20,366 BAC-end sequences, it was found that 3,748 BAC-end sequences (18.4%) contain one or more stretches of microsatellite sequences. Of these, 2,365 (63%) had sufficient flanking sequences on both sides, making them potentially useful as markers for genetic mapping; 403 BES harbor microsatellite sequences at the immediate beginning of the BES, making them more difficult to be developed as markers; the remaining 980 clones had microsatellite sequences at the end of BES. It is obvious that the number of microsatellites at the end of BES was much larger than the number of microsatellites at the beginning of BES. That is because many sequencing reactions terminated due to the presence of simple sequence repeats. For this last category, additional sequencing can be conducted using primers close to the end of the BES to generate sufficient flanking sequences on both sides. With additional efforts including generation of sufficient flanking sequences and testing of polymorphism, these microsatellites should be useful for genetic linkage mapping and integration of the catfish linkage maps with the BAC-based physical map. Bioinformatic mining for microsatellites is perhaps the most productive and economic approach if the genome resources such as BES exist.

Several Web-based programs are sufficient for the identification of microsatellites within BAC-end sequences. Microsatfinder *(http://www.eusebius.mysteria.cz/microsatfinder/index.php)*, Tandem Repeat Finder (Benson 1999, http://tandem.bu.edu/trf/trf.html), and RepeatFinder (http://www.genet.sickkids.on.ca/~ali/repeatfinder.html) are all quite user-friendly. MICAS (http://210.212.212.7/MIC/index.html) is another highly user-friendly, interactive Web-based server to find nonredundant microsatellites in a given nucleotide sequence/genome sequence (Sreenu et al. 2003). The Informax Vector NTI software packages are also very efficient for the identification of microsatellites within sequences (Serapion et al. 2004), allowing establishment of a microsatellite database.

BAC-end sequences also allow an overall glance at the types and relative abundance of microsatellites in an organism. For instance, in catfish, the most abundant microsatellite type is CA/GT. Overall, AT-rich microsatellites are more abundant than GC-rich microsatellites (Figure 15.2).

The BAC-anchored microsatellites are a valuable resource for integration of genetic linkage and physical maps (Figure 15.3). Because they are identified through

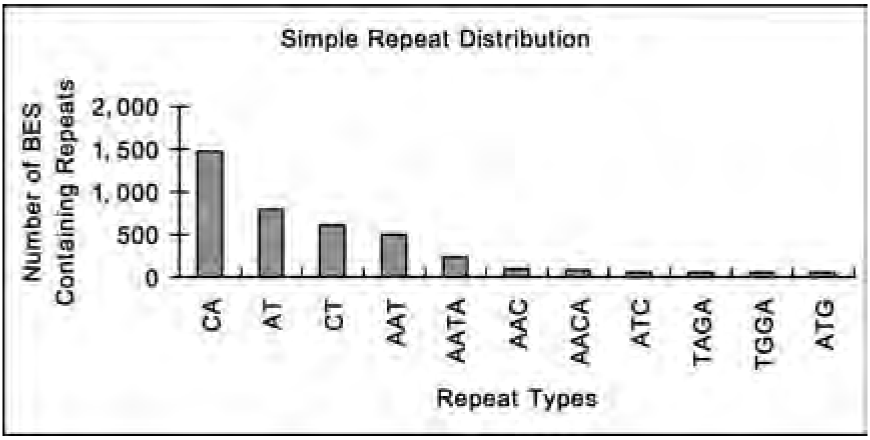

Figure 15.2. Microsatellite types and their relative abundance in the catfish genome as revealed by informatic analysis of BAC-end sequences.

BAC-end sequencing, their location on the physical map can be determined by BAC contig construction. After they are mapped to genetic linkage maps by genotyping them in a resource family, they allow alignment of the linkage and the physical maps (Figure 15.3).

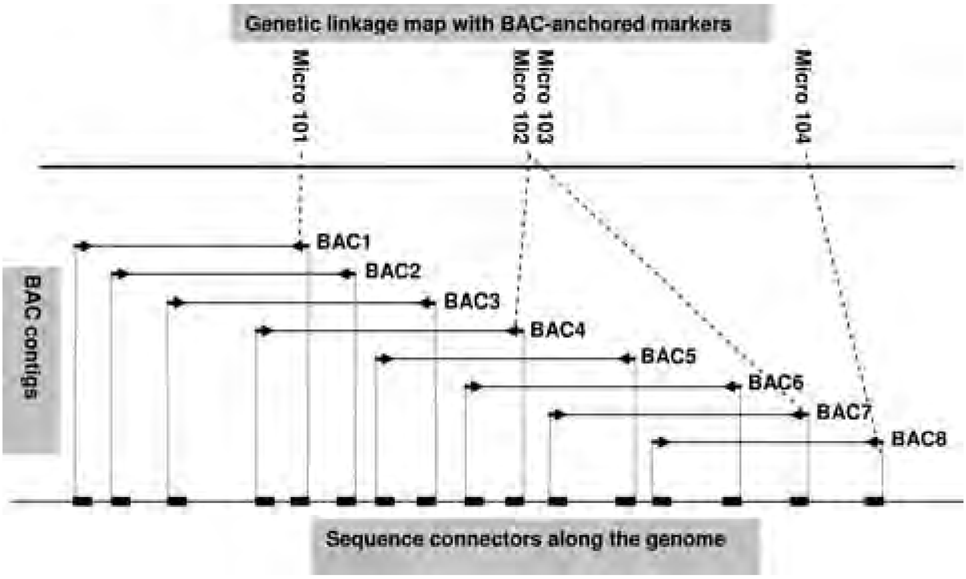
Genetic linkage map with BAC-anchored markers
Micro 101
Micro 103
Micro 102
Micro 104
BAC contigs
BAC1
BAC2
BAC3
BAC4
BAC5
BAC6
BAC7
BAC8
Sequence connectors along the genome

Figure 15.3. BAC-end sequences are produced from both ends (arrows in opposite directions on the BAC clones within the contig). After the BAC contigs are constructed, the BAC-end sequences become sequence tag connectors (shown as black rectangles along the line, bottom) along the genome as dictated by the distribution of BAC contigs. The microsatellite markers (micro 101–104) identified from BAC-end sequences allow integration of linkage map (top) and the BAC-based physical maps (middle).

BAC-End Sequencing Allows Virtual Mapping of Genes to Physical Maps

The entire business of gene mapping deals with the positions of genes on chromosomes and in the genome. Genes can be mapped in various ways including genetic linkage mapping, and mapping of genes to physical maps through hybridization, however, BAC-end sequencing is probably the most efficient way to map genes to BACs. After the BAC-based contigs are constructed, the genes located on the BACs are placed on the physical map.

Genes can be placed on BACs by simple BLASTX searches of BAC-end sequences. For instance, BLASTX searches of the 20,366 channel catfish BES resulted in 2,351 BES with significant hits to 1,877 unique genes (E value $<e^{-5}$), demonstrating the efficiency. However, gene identification by BLASTX searches is complicated by the presence of erroneously annotated sequences in the GenBank, especially in cases when repetitive elements are involved. In addition, it is a challenging task to set a solid benchmark for significant hits. Because different genes may have different evolving rates, researchers have to be flexible enough to make a judgment as to what level of E-values is sufficient to putatively identify the sequence. At times, it is safe to provide the significant hits with their associated E-values, as well as the alignment length. A short alignment length may indicate only part of an exon is involved in the BAC end sequences, which even provide more concrete evidence than alignments with long sequence but scattered identities. For instance, of the 1,877 gene hits in the catfish example, 1,130 had an e-value smaller than e^{-10} in BLASTX searches with average alignment length of 73 amino acids and a range of 31–100% identity (Table 15.1). Although it is difficult to conclude what level of e-values would provide a stringent confidence on the putative gene identities of the BES, it is obvious that the lower the E-values, the more likely the BES are to be related to the putatively identified genes.

Table 15.1. Mapping of genes to BACs through BAC-end sequencing. Listed are a number of BLASTX hits of genes by BAC-end sequences, excluding redundant hits. E-values, alignment length range, average alignment length, and percentage of identities are provided as an indication for the level of similarities.

E-values	Number of hits	Alignment length (amino acid)	Average alignment length (amino acids)	(%) Identity
$<10^{-50}$	58	101–228	167	48–99
10^{-40}–10^{-50}	54	81–207	134	43–97
10^{-30}–10^{-40}	77	66–217	103	40–100
10^{-20}–10^{-30}	253	45–175	75	34–100
10^{-15}–10^{-20}	275	37–199	62	30–100
10^{-10}–10^{-15}	413	30–186	54	31–100
Subtotal	1130	30–228	73	31–100
10^{-5}–10^{-10}	747	19–193	47	23–100
Total	1,877	19–228	63	23–100

Anchoring BES to Existing Genome Sequences for Comparative Genomics

The fundamental basis for comparative genomics is the genomes are conserved in their gene sequences, as well as in the arrangement of genes and other important functional domains. The BAC-end sequences, though not continuous, provide a string of sequences that can be used to compare with genome sequences for the presence of related sequences arranged in a similar fashion. Such comparisons can be conducted by direct BLAST searches. In the case of fish, the entire genome sequences exist for zebrafish, fugu, and *Tetraodon* species. A direct comparison of BAC-end sequences and the genome sequences of the closely related species should identify chromosome-anchored conserved sequences within the species of interest. For instance, searches of the 20,366 catfish BES against *D. rerio* and *T. nigroviridis* genome sequences resulted in 3,251 (16%) and 1,670 (8.2%) significant hits (E value $<e^{-5}$), respectively. However, many BES had many hits in different genomic regions of the zebrafish genome sequence, suggesting that they are repetitive in nature. In order to obtain unique significant hits that are meaningful as resources for comparative genome analysis, the significant hits were tabulated in Excel with hit ID, chromosome information, and beginning and ends of the region of similarity. Repeated hits were then removed, resulting in 1,074 unique significant hits to the zebrafish genome sequence. Similar BLAST searches against *T. nigroviridis* genome resulted in 773 unique significant hits, suggesting that the number of genomic regions containing evolutionarily conserved sequences was greater between catfish and zebrafish than between catfish and *T. nigroviridis*, consistent with their phylogenetic relationships.

In order to understand the nature of the conserved genomic sequence blocks between catfish and zebrafish or *T. nigroviridis*, the BES with unique BLASTN significant hits were searched against the NR database using BLASTX to assess the number of genes among the conserved genomic sequences. Of the 1,074 unique significant hits to the zebrafish genome, 417 (38.8%) had significant BLASTX hits. Similarly, with *Tetraodon*, of the 773 unique significant hits, 406 had significant BLASTX hits. Clearly, of the unique BLASTN hits, the number of significant BLASTX hits was similar with the zebrafish and the *Tetraodon* genomes, suggesting that the vast majority of genes were conserved among catfish, zebrafish, and *Tetraodon*. The greater number of unique BLASTN hits to the zebrafish genome was accounted for, in the most part, by the nongene genomic sequences, reflecting the close relatedness of the catfish and zebrafish genomes.

To anchor the catfish BES to the chromosomes of the zebrafish and *Tetraodon* genomes, BLASTN and BLASTX search results were tabulated according to chromosome locations (Table 15.2). The catfish BES had significant BLASTN hits to every one of the 25 zebrafish chromosomes with a range of 16–66 hits per chromosome. Of the unique BLASTN hits, all 25 chromosomes had significant BLASTX hits to the catfish BES, with 4–30 hits per chromosome. Similar but fewer unique hits were found with the *Tetraodon* genome. The notable low number of hits was found with *Tetraodon* chromosome 20, with only a single significant hit with BLASTN or BLASTX. This was mainly because of the low sequence coverage of the *Tetraodon* chromosome 20 to date (http://www.genoscope.cns.fr/externe/tetranew/), and perhaps also because of the small size of chromosome 20.

Table 15.2. Distribution of unique BES significant BLASTN hits in the *Danio rerio* and *Tetraodon nigroviridis* genomes. The cut-off value was set at 1 × E-05.

	Zebrafish		*Tetraodon*	
Chromosome	Unique BLASTN hits	BLASTX hits of unique BLASTN hits	Unique BLASTN hits	BLASTX hits of unique BLASTN hits
1	58	22	44	32
2	39	13	55	27
3	58	28	36	19
4	36	14	13	10
5	66	27	23	13
6	40	21	15	4
7	56	22	20	11
8	36	15	20	14
9	59	24	23	12
10	42	16	32	21
11	38	15	21	11
12	37	16	24	16
13	45	16	74	17
14	54	16	14	9
15	31	9	30	25
16	53	18	23	15
17	47	15	16	9
18	28	4	21	14
19	44	20	10	8
20	52	30	1	1
21	34	10	19	14
22	28	11		
23	54	20		
24	23	7		
25	16	8		
Undesignated			239	104
Total	1,074	417	773	406

Identification of Conserved Syntenies Using BAC-end Sequences

Large-scale BAC-end sequencing is currently the most efficient strategy for building whole-genome comparatively anchored physical maps in map-poor species (Larkin et al. 2003). Conservation of genes and their locations can also be approached in a local genomic environ, allowing identification of conserved syntenies. Assuming genes and their genomic locations are conserved through evolution, it is reasonable to assume that any two given genes that are close to each other may have the same arrangement in the same genomic environ as well as in a closely related species. Based on this notion, BAC-end sequences are analyzed for the presence of genes on both ends of the same BAC clone. Given that the average insert size of BAC libraries is 100–200 kb, these genes on both ends of the same BAC clone are physically linked with a distance of 100–200 kb apart. This information can be used to determine if the same genes are arranged in the same genomic environ in species with entire genome sequences by BLAST searches. If the answer is yes, the process identifies a conserved synteny between the two species. This approach was demonstrated to be very efficient for the identification of conserved syntenies. For instance, of the 20,366 BES, 17,478 BES were mate pair sequences from 8,739 BAC clones. BLASTX searches indicated that 141 sequenced BACs harbor genes on both ends. These paired BAC-ends with genes allowed us to compare whether the same set of genes were located on the similar genomic environs in the zebrafish and *Tetraodon* genomes. Of the 141 paired BAC-ends with genes, 43 (30.5%) were located on the same chromosomes of zebrafish or *Tetraodon*, of which 23 (16.3%) appeared to exhibit a high level of conserved synteny. The number of conserved syntenies was greater between the catfish genome and the zebrafish genome than between the catfish genome and the *Tetraodon* genome. Of the 23 conserved syntenies, 21 were present between the catfish and zebrafish genomes. Additional experiments using comparative analysis and direct BAC sequencing in catfish revealed that many of the syntenies could be extended. These encouraging results suggest that comparative mapping, especially with zebrafish, will be a sound approach for future catfish genomics research. Furthermore, information gained in mapping and gene discovery projects in catfish may help to explain aspects of zebrafish and teleost genome evolution.

BAC-end Sequences and the Minimal Tiling Path for Entire Genome Sequencing

The successful sequencing of the human and mouse genomes stirred up a wave of excitement in genome biology. Currently, whole genome sequencing has been completed or is being completed for a number of vertebrate animals including important agricultural animals such as cattle, porcine, and chicken. Sequencing in fish species, however, has been limited to several model species such as the zebrafish (*Danio rerio*), *Takifugu rubripes*, *Tetraodon nigroviridis*, and medaka. As the whole genome sequences have been crucial for the study of genome expression and function, scientists are now seriously considering a genome-sequencing project for important aquaculture species. Large-scale BAC-end sequencing is not only a necessary step as a survey of the genome in the assessment of genome composition and architecture, but also a required element for the identification of minimal tiling pass for clone-by-clone based strategy of entire genome sequencing. See Chapter 26.

Minimal tiling path (MTP) is a set of minimally overlapping BAC clones in the physical map picked for use in clone-by-clone sequencing of the entire genome. For most efficient whole genome sequencing, the ideal situation is to have minimally repeated sequencing and also cover all gaps so that the entire genome sequences can be assembled (Engler et al. 2003). Three different strategies were used to generate draft sequences for the human, mouse, and rat genomes, namely, the “clone-by-clone” for human, the whole genome shotgun (WGS) for mouse, and a hybrid strategy for rat (Lander et al. 2001, Waterston et al. 2002, Rat Genome Sequencing Project Consortium 2004). With the clone-by-clone strategy, individual BAC clones are shotgun-sequenced, the sequence of each BAC clone is generated by assembling the corresponding sequencing reads, and the sequence of the whole genome is obtained by merging overlapping BAC clone sequences. To minimize sequencing the same genomic region multiple times, a set of minimally overlapping clones covering the whole genome is determined beforehand, and such a set of BAC clones are called MTP. BES serving as sequence connectors are essential for the identification of a minimum tiling path of BAC clones for whole genome sequencing (Siegel et al. 1999, Engler et al. 2003, Chen et al. 2004).

Two approaches are available traditionally for the selection of the MTP. The first is a map-based approach as used by the *Caenorhabditis elegans* project (Coulson et al. 1986) and human chromosomes 1, 6, 20, 22, and X (Bentley et al. 2001). Fingerprints of clone pairs that appear to have a minimum overlap are analyzed in the FPC Gel Image display. Viewing the gel images of neighboring clones helps identify false-positive and false-negative bands. With this method, a complete MTP can be picked before any sequencing is started, so that all clones can be sequenced in parallel. However, the amount of overlap may be large (e.g., 47.5 kb overlap for the MTP picked by the International Human Genome Consortium), because many BAC contigs were constructed with fingerprinting using restriction enzyme with 6 bp recognition sequences. Clones need to share multiple bands to have enough evidence of overlap, and on average, one band is produced every 4,096 bp. In addition, manual selection of one minimally overlapping pair is highly labor intensive.

The second approach is based on BAC-end sequences (Venter et al. 1996). In this approach, a seed clone is picked and completely sequenced. The BAC-end sequences are queried for hits to the finished sequence. The BAC clones with minimal overlapping is picked for extending sequencing. The new set of MTP clones is sequenced, and new clones are picked off the ends of this set based on minimal overlapping with BAC-end sequences. This process is repeated until the entire region is sequenced (Figure 15.4).

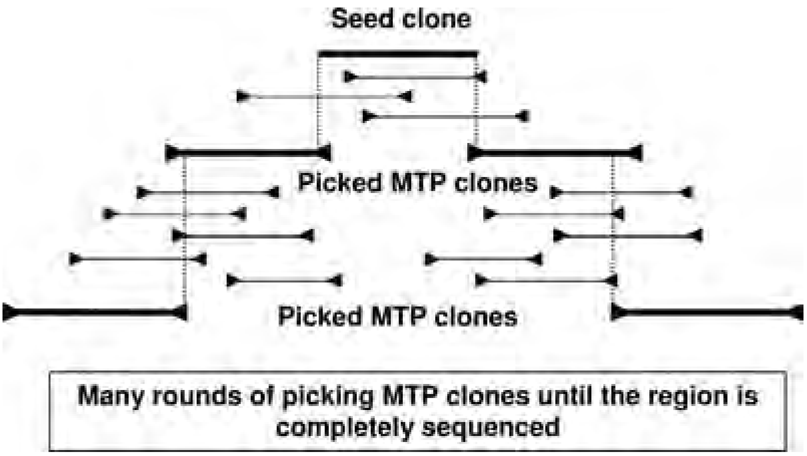
Seed clone
Picked MTP clones
Picked MTP clones
Many rounds of picking MTP clones until the region is completely sequenced

Figure 15.4. Selection of minimal tiling path (MTP) clones using BAC-end sequences (arrows). The MTP clones are selected by picking BAC clones with minimal sequence overlapping between their BAC-end sequences (triangle at the end of BAC clones) and the already sequenced seed clone. After picking the two MTP clones, they are completely sequenced and aligned to the BAC-end sequences to identify the next set of MTP clones with minimal overlap with the first round of MTP clones. This process is continued until complete sequencing of the region. Note that the BAC-end sequences can be coupled to the restriction fingerprinting information to minimize false positives and to reduce the overlapping regions.

By use of sequence information rather than the restriction fingerprints, the amount of overlap required for MTP pairs is reduced drastically. However, the risk of false positives is high, especially when considering repeats and errors in the low-quality BES. This approach, when coupled with the BAC-based physical map (i.e., the hybrid approach) has been the most effective for the selection of MTP. The MTP with minimal overlapping and maximal accuracy guarantees efficient sequencing of the genome and proper assembly of the entire genome sequences once they are sequenced.

Recently, a software pipeline CLONEPICKER was developed that integrates sequence data with BAC clone fingerprints to dynamically select a minimal overlapping clone set covering the whole genome (Chen et al. 2004). CLONEPICKER uses restriction enzyme fingerprint data, BAC-end sequence data, and sequences generated from individual BAC clones as well as whole genome shotgun sequencing reads. Incorporation of all available genomic information in this software allows accurate selection of tiling path with truly minimal overlaps.

Conclusion

BAC-end sequencing is simple enough not to refer to it as a genome technology. However, BAC-end sequences are extremely rich in genome information. The analysis of BAC-end sequences allows an unbiased sampling of the genome as to its composition and architecture. Bioinformatic mining of BAC-end sequences allows the uncovering of the repeat structure of the genome and the identification of polymorphic microsatellites. Many markers discovered from BAC-end sequences can be used to integrate the genetic linkage maps and the physical map as they can be placed on both maps. BAC-end sequencing is one of the most efficient approaches to locate genes to physical maps. BAC-end sequences can be exploited for comparative genome analysis including characterization of evolutionarily conserved syntenies. For many aquaculture species, the assessment of their genomes by BAC-end sequencing is probably as good as they can get as their genomes may never be sequenced.

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Agarwal P and DJ States. 1994. The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. Proc Int Conf Intell Syst Mol Biol, 2, pp. 1–9.

Altschul SF, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, pp. 3389–3402.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 27, pp. 573–580.

Bentley DR, P Deloukas, A Dunham, L French, SG Gregory, SJ Humphray, AJ Mungall, MT Ross, NP Carter, and I Dunham. 2001. The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. Nature, 409, pp. 942–943.

Chen R, E Sodergren, GM Weinstock, and RA Gibbs. 2004. Dynamic building of a BAC clone tiling path for the Rat Genome Sequencing Project. Genome Res, 14, pp. 679–684.

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, pp. 7821–7825.

Delcher AL, S Kasif, RD Fleischmann, J Peterson, O White, and SL Salzberg. 1999. Alignment of whole genomes. Nucleic Acids Res, 27, pp. 2369–2376.

Devereux J, P Haeberli, and O Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res, 12, pp. 387–395.

Engler FW, J Hatfield, W Nelson, and CA Soderlund. 2003. Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res, 13, pp. 2152–2163.

Ewing B and P Green. 1998. Base-calling 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. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res, 8, pp. 175–185.

Kurtz S, JV Choudhuri, E Ohlebusch, C Schleiermacher, J Stoye, and R Giegerich. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res, 29, pp. 4633–4642.

Kurtz S, E Ohlebusch, C Schleiermacher, J Stoye, and R Giegerich. 2000. Computation and visualization of degenerate repeats in complete genomes. Proc Int Conf Intell Syst Mol Biol, 8, pp. 228–238.

Lander ES, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, et al. 2001. Initial sequencing and analysis of the human genome. Nature, 409, pp. 860–921.

Larkin DM, A Everts-van der Wind, M Rebeiz, PA Schweitzer, S Bachman, C Green, CL Wright, EJ Campos, LD Benson, J Edwards, L Liu, K Osoegawa, JE Womack, PJ de Jong, and HA Lewin. 2003. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. Genome Res, 13, pp. 1966–1972.

Lefebvre A, T Lecroq, H Dauchel, and J Alexandre. 2003. FORRepeats: detects repeats on entire chromosomes and between genomes. Bioinformatics, 19, pp. 319–326.

Liu L, L Roinishvili, X Pan, Z Liu, and C Kumar. 2000. GPMS: A web based genome project management system. The Proceeding of 4th World Multiconference on Systematics, Cybernectics, and Informatics SCI 2000, pp. 62–67.

Liu ZJ, P Li, and R Dunham. 1998. Characterization of an A/T-rich family of sequences from the channel catfish (*Ictalurus punctatus*). Mol Mar Biol Biotechnol, 7, pp. 232–239.

Rivals E, O Delgrange, JP Delahaye, M Dauchet, MO Delorme, A Henaut, and E Ollivier. 1997. Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. Comput Appl Biosci, 13, pp. 131–136.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of Type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol, 6, pp. 364–377.

Siegel AF, B Trask, JC Roach, GG Mahairas, L Hood, and G van den Engh. 1999. Analysis of sequence-tagged-connector strategies for DNA sequencing. Genome Res, 9, pp. 297–307.

Sreenu VB, G Ranjitkumar, S Swaminathan, S Priya, B Bose, MN Pavan, G Thanu, J Nagaraju, and HA Nagarajaram. 2003. MICAS: A fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. Applied Bioinformatics, 2, pp. 165–168.

The International Human Genome Mapping Consortium. 2001. A physical map of the human genome. Nature, 409, pp. 934–941.

Venter JC, HO Smith, and L Hood. 1996. A new strategy for genome sequencing. Nature, 381, pp. 364–366.

Waterston RH, K Lindblad-Toh, E Birney, J Rogers, JF Abril, P Agarwal, R Agarwala, R Ainscough, M Alexandersson, P An, SE Antonarakis, J Attwood, R Baertsch, J Bailey, K Barlow, S Beck, E Berry, B Birren, T Bloom, P Bork, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420, pp. 520–562.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with other fish species. Anim Genet, 37, pp. 321–326.

Zhang Z, S Schwartz, L Wagner, and W Miller. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol, 7, pp. 203–214.