Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 16
Genomescape: Characterizing the Repeat Structure of the Genome

Zhanjiang Liu

Understanding genome landscape, the genomescape, is important for genome biology because repetitive elements form a major fraction of eukaryotic genomes. Specifically, characterization of repeat structures of a genome can significantly reduce the complexities involved in genome studies, facilitating linkage mapping, physical mapping, comparative mapping, and laying the groundwork for whole genome sequencing. Repetitive elements, once dismissed as mere junk DNA, are now recognized as "drivers of genome evolution" (Kazazian 2004) whose evolutionary role can be "symbiotic (rather than parasitic)" (Holmes 2002). Examples of potentially beneficial evolutionary events in which repetitive elements have been implicated include genome rearrangements (Kazazian 2004), gene-rich segmental duplications (Bailey et al. 2003), random drift to new biological function (Kidwell and Lisch 2001, Brosius 2003), and increased rates of evolution during times of stress (Capy et al. 2000, Shapiro 1999). For these reasons, the study of repeat elements and their evolution is emerging as a key area in genome biology (Zhi et al. 2006).

Based on their known functions, repetitive elements can be divided into functional repetitive gene clusters and nonfunctional repetitive elements; based on their arrangements and distribution in the genome, they can be divided into tandem repetitive elements and dispersed elements; and based on their abundance, they can be divided into high, intermediate, and low abundance repetitive elements. In this chapter, I will briefly introduce several major classes of repetitive elements, and present several selected methodologies for the characterization of genomescape.

Classical Approaches for the Characterization of Repeat Structures of Genome

Roy Britten and his colleague were the first to study the genome using genomic approaches (Britten and Kohne 1968). They studied reassociation kinetics of genomic DNA in solution using a technique termed "Cot analysis." When a solution of denatured genomic DNA is placed in an environment conducive to renaturation, the rate at which a particular sequence reassociates is correlated to the copy number of the sequence present in the genome. This principle forms the basis of Cot analysis (Britten and Kohne 1968 and for a recent review of Cot principles, see Peterson et al. 2002). Cot value is equal to the product of nucleotide concentration in moles per liter (C_0 or Co) and reassociation time in seconds (t), and, if applicable, a factor based upon the cationic concentration of the buffer. Samples of sheared genomic DNA are heat-denatured and

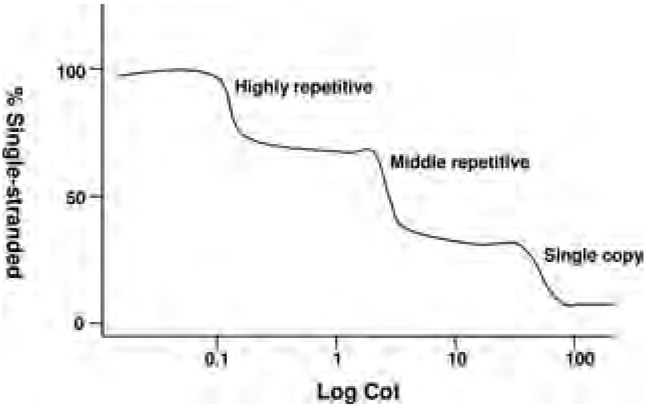


Figure 16.1. A schematic presentation of the Cot reassociation curve.

allowed to reassociate to different Cot values. For each sample, renatured DNA is separated from single-stranded DNA using hydroxyapatite (HAP) chromatography, and the percentage of the sample that has not reassociated (percentage of ssDNA) is determined. The logarithm of a sample's Cot value is plotted against its corresponding percentage of ssDNA to yield a Cot point, and a graph of Cot points ranging from the start to completion of the reassociation is called a Cot curve (Peterson et al. 1998) (see Figure 16.1). Mathematical analysis of a Cot curve permits estimation of genome size, the proportion of the genome contained in the single-copy and repetitive DNA components, and the kinetic complexity of each component Peterson et al. 2002). Cot curves (Figure 16.1) reflect the major three fractions of genomic DNA: the highly repetitive fraction that renatures rapidly; the intermediate repetitive fraction that renatures relatively slower; and the low copy number repetitive elements and the single copy genes that reassociate very slowly.

Cot analysis also allows many of the basic characteristics of genomes to be compared between organisms. Based on the reassociation kenetics, interspecific comparison of Cot data has provided considerable insight into the structure and evolution of eukaryotic genomes (e.g., Britten and Kohne 1968). Even though the Cot analysis was developed over three decades ago even before the emergence of the molecular biology era, its principles are still highly useful for genome characterizations in the genomics era.

Characterization of Tandem Repeats with High Genomic Copy Numbers

Tandem repetitive noncoding DNA sequences make up a large fraction of the genomes of eukaryotes. The sequence complexities of these repetitive sequences can vary from a single base pair (bp) to over two kilobase pairs (kb). Copies of the tandem repeats can vary greatly, ranging from a total size of 100 bp to more than 100 mega base pairs (Mb) (Milklos and Gill 1982). Based on the repeat lengths and array sizes, they have been divided into three classes: microsatellite, minisatellite, and satellite

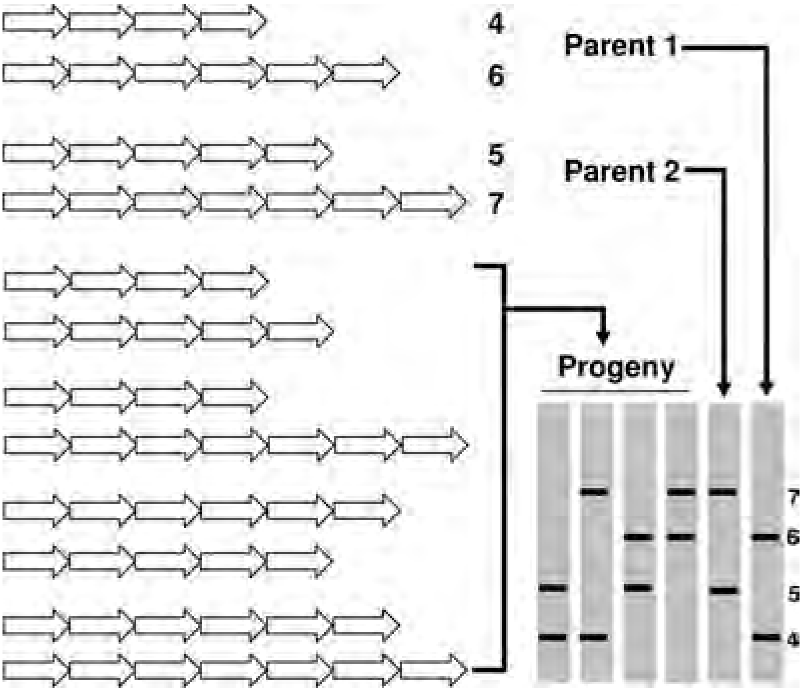


Figure 16.2. Minisatellites can be used as polymorphic markers. They are inherited as co-dominant markers.

DNAs (Levinson and Gutman 1987). As their names indicate, microsatellites include repetitive sequences with very simple sequence complexity and short arrays of the repeat (see Chapter 5). Satellite DNA exhibits more sequence complexity (generally 100 bp or longer) with long arrays of total repeat lengths. Minisatellites show intermediate features between satellites and microsatellites. The three classes of the repetitive sequences also differ in where they reside in the eukaryotic genomes. Satellite DNA is often concentrated in regions of very low recombination (Charlesworth et al. 1986) such as heterochromatic regions near centromeres and telomeres where meiotic recombination is suppressed. Minisatellite sequences are usually located in the euchromatic portions of chromosomes. Recombination at minisatellite loci appears to be higher than at satellite DNA loci (Stephan 1989, Stephan and Cho 1994). Variable number of tandem repeats (VNTR) can be regarded as minisatellites with 5–50 repeats. VNTRs are hot spots for meiotic recombination. They exhibit high levels of polymorphism, and therefore, are useful as polymorphic markers (Figure 16.2). It is believed that the high mutation rate leading to polymorphism is caused by uneven meiotic recombination. Microsatellites are tandem repeats of very short sequences (1–6 bp, as shown in Chapter 5) interspersed in various regions in chromosomes including within or near expressed genes (e.g., Gong et al. 1997, Serapion et al. 2004).

Technically, the identification of tandemly duplicated repeats is the most straightforward. These elements can be readily detected by hybridizations. For highly abundant repetitive elements, they can be observed by digesting genomic DNA with restriction

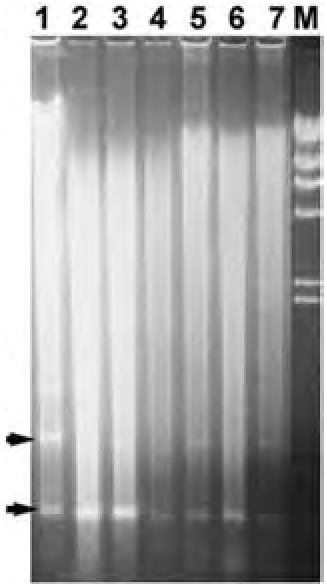
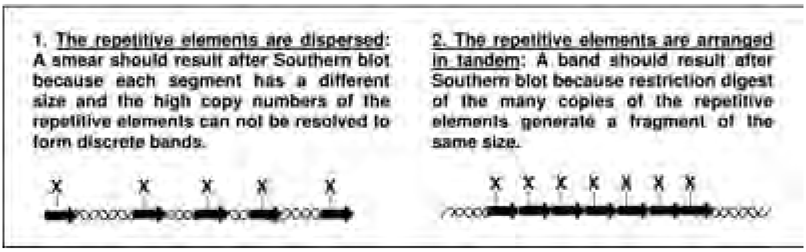


Figure 16.3. Detection of repetitive elements by direct restriction enzyme digestion. The example shown here was adopted from Liu and others (1998) from a catfish study. Genomic DNA from various strains of catfish (lanes 1–7) was digested with restriction endonuclease *Xba* I, separated on an agarose gel, and visualized by ethidium bromide staining, with molecular weight (M) on the last lane. Note the generation of discrete bands (arrows) above the background of a smear.

enzymes; direct restriction enzyme digestion of genomic DNA should produce bands above the background of a smear (Figure 16.3). Using this approach, two types of repetitive elements were found in the zebrafish genome, accounting for 5% and 0.5% of the zebrafish genome (He et al. 1992). Using a similar approach, Liu and others (1998) identified a family of A/T-rich *Xba* elements that are arranged in tandem and account for 5–6% of the catfish genome.

Differentiation of tandem repeat from interspersed repeats (see below) can be accomplished through several approaches. As discussed above, tandem repeats tend to generate discrete DNA bands (if restriction sites exist within the repeat unit). However, many interspersed repeats also generate discrete DNA bands if two or more restriction sites for the same enzyme exist within the repeat unit. One way to differentiate these two types is through the use of partial restriction digest. For tandem repeats, with limited amounts of restriction enzyme, partial digestion products will generate a band pattern exhibiting monomers, dimers, trimers, tetramers, pentamers, hexamers, etc., of the repeat unit. With incremental amounts of restriction enzyme, the higher molecular weight bands will disappear resulting in the final product, the monomer of the repeat unit. This is not the case for interspersed repeats that generate smears regardless of the amount of restriction enzyme used (Figure 16.4).

Another approach to differentiate tandem repeats from interspersed repeats is the use of fluorescent in situ hybridization. See Chapter 17. Tandem repeat types generate focused hybridization patterns with very high fluorescent signals, whereas interspersed repeats produce hybridization patterns that are scattered throughout the genome.



1. The repetitive elements are dispersed: A smear should result after Southern blot because each segment has a different size and the high copy numbers of the repetitive elements can not be resolved to form discrete bands.

2. The repetitive elements are arranged in tandem: A band should result after Southern blot because restriction digest of the many copies of the repetitive elements generate a fragment of the same size.

Figure 16.4. Two different types of arrangements and organizations of repetitive elements can be differentiated by Southern blot analysis followed by partial restriction enzyme digest.

Characterization of Intersperse Repetitive Elements

Dispersed repetitive elements included mostly long interspersed elements (LINEs), short interspersed elements (SINEs), and DNA transposons. LINEs and SINEs are retrotransposable DNA elements that duplicate in the genome via a "copy and paste" mechanism. Functional LINE elements encode a reverse transcriptase, allowing the transcribed LINE RNA to be converted into DNA, and an endonuclease for cleaving the genomic DNA at the new insertion site. SINE elements are shorter repetitive elements that require the "copy and paste" machinery of the LINEs. SINEs resemble small nuclear RNAs such as tRNA, and are transcribed by RNA polymerase III. SINEs then rely on the reverse transcriptase and endonuclease from the LINE elements to reinsert and multiply in the genome.

In addition to LINEs and SINEs, another major class of interspersed elements is DNA transposons. DNA-mediated transposons move through DNA intermediates and depend on transposases. While retrotransposons are more common and abundant in vertebrates, DNA transposons are more common in bacteria, plants, and invertebrates such as fruit flies and nematodes. DNA transposons harbor inverted repeats required *in cis* for transposition. Autonomous DNA transposons also harbor a functional transposase gene to encode the transposase enzyme required for transposition *in trans.* Despite a wide range of distribution (Radice et al. 1994), DNA transposons were not discovered in vertebrates until the 1990s (Heierhorst et al. 1992, Henikoff 1992). Since the discovery of the first vertebrate DNA transposons in channel catfish by database analysis (Henikoff 1992), search for an active copy of DNA transposons has been a keen research area. Recently, DNA transposons in vertebrates have been isolated from a number of vertebrate species, but they have been best characterized in teleosts (Heierhorst et al. 1992; Henikoff 1992; Goodier and Davidson 1994; Radice et al. 1994; Izsvak et al. 1995; Ivics et al. 1996; Lam et al. 1996; Koga and Hori 2000, 2001). Members of the *Tc1*/mariner superfamily represent most of the DNA transposons discovered in vertebrates. In addition to teleosts, *Tc1*/*mariner*, *tiggers*, and other transposon-like elements have been discovered in amphibians (Lam et al. 1996) and human genomes (Oosumi et al. 1995, Smit and Riggs 1996).

Until recently *Tc1*-like elements were all identified as nonautonomous from vertebrate genomes, resulting from extensive insertions/deletions, and in-frame termination codons in their transposase genes (Lam et al. 1996; Ivics et al. 1996, 1997). Kawakami and his colleagues (2000) identified the first endogenous autonomous transposon, an *Ac*-like *Tol2* element from the Japanese *medaka* fish that can transpose both transiently in fertilized zebrafish eggs and in the zebrafish germ lineage (Koga and Hori 2000, 2001). Several additional transposases with intact open reading frames (ORF) were recently identified from the amphibian *Xenopus tropicalis* and may still be involved in active transposition (Sinzelle et al. 2005).

Characterization of dispersed repetitive elements is a great challenge. First, these elements are highly interspersed throughout the genome; second, they are present in high copy numbers, making hybridization-based approaches ineffective; third, their sequences are highly related, but differ between copies. Genomic approaches using data mining appears to be much more effective for the characterization of dispersed repetitive elements. Their abundance, composition, distribution, and nature of the sequences can all be analyzed using data mining software. REPuter (Kurtz and Schleiermacher 1999) (http://www.genomes.de/), Basic Local Alignment Search Tool (BLAST), and RepeatMasker, among many others, are frequently used to assess the repeat structures of the genome. However, the effectiveness of the software depends on the existence of repeat databases. REPuter also allows identification of new repeat types in a given organism. Often searches can be conducted by searching for the presence of repeated sequences with parameters set at, for example, a minimum length of 20 bp and a sequence variation of 10% (90% conserved). Self BLAST searches are also effective for the identification of repeated sequences in an organism upon generation of genomic sequence surveys.

RepeatMasker is highly effective for the identification of common types of repeats. For instance, the entire genome sequences of zebrafish, fugu, and *Tetraodon* are already available. Repeat masking using repeat libraries of these species allowed a good estimation of the shared repeats of catfish with these species (Xu et al. 2006), as shown in Table 16.1. However, RepeatMasker lacks the ability to detect species-specific and novel repeat types.

At the genomic level, dot plot alignments are also often used for the detection of repetitive elements (Sonnhammer and Durbin 1995). Within a genome, a dot plot illustrates duplicated sequences: repeated sequences and gene families. Between genomes, dot plots reveal levels of conservation of sequence and of gene orders.

Characterization of Gene Clusters

Many genes with related functions are arranged in clusters and are duplicated into many copies. As a rule of thumb, genes whose products are most highly demanded by cellular functions likely have multigene families, and often they are arranged together in tandem repeats. Some examples of tandemly repeated gene families include globin genes, immunoglobulin genes, rRNA genes, tRNA genes, and histone genes. Such an arrangement is beneficial in order to provide coordinated expression. For instance, the ribosomal RNA is required for the structure of ribosomes that demand a constant 1:1 ratio. The genes for rRNA are organized not only together, but also transcribed under control of the same promoter. A single transcription unit (cistron) contains the 18S, 5.8S, and the 28S rRNA genes. In addition to functional genes, pseudogenes are often

Table 16.1. Repeat composition of the channel catfish genome as assessed by RepeatMasker of the BAC-end sequences using the combined repeat database of zebrafish and *Takifugu rubripes.*

Repetitive Elements	Number of Elements	Length Occupied	% of Sequence
Retroelements	1,972	349,917 bp	3.13
SINEs:	1,083	139,373 bp	1.24
Penelope	1	269 bp	0
LINEs:	643	135,520 bp	1.21
L2/CR1/Rex	525	107,885 bp	0.96
R1/LOA/Jockey	14	2,773 bp	0.02
R2/R4/NeSL	1	50 bp	0
RTE/Bov-B	37	6,137 bp	0.05
L1/CIN4	65	18,406 bp	0.16
LTR elements:	246	75,024 bp	0.67
BEL/Pao	12	3,604 bp	0.03
Gypsy/DIRS1	179	60,047 bp	0.54
Retroviral	21	3,726 bp	0.03
DNA transposons:	2,591	563,923 bp	4.57
hobo-Activator	220	23,205 bp	0.16
Tc1-IS630-Pogo	2,077	507,735 bp	4.12
En-Spm	3	160 bp	0
PiggyBac	57	6,399 bp	0.06
Tourist/Harbinger	77	7,982 bp	0.07
Unclassified:	9	751 bp	0.01
Total interspersed repeats:		914,591 bp	8.17
Satellites:	25	2,602 bp	0.02
Simple repeats:	5,798	289,211 bp	2.58
Low complexity:	3,202	127,064 bp	1.13

found in tandem clusters. Pseudogenes are defined by their possession of sequences that are related to those of the functional genes, but that cannot be translated into a functional protein. Pseudogenes are often denoted by ψ. They are formed when transcription signals such as CAAT box or TATA are abolished, or when the splicing junction is mutated so that the transcripts cannot be properly processed, or when in-frame premature termination codons are evolved. Most gene families have members that are pseudogenes constituting a minority of the family.

Histone Gene Clusters

In most organisms, five major types of histone proteins (H1, H2A, HZB, H3, and H4) combine to form a functional histone unit. Two structures are created by the dimerization of H2A/H2B and H3/H4. These structures then self-dimerize forming two tetramers (H2A-H2B-H2A-H2B and H3-H4-H3-H4) and the two tetramers join to create the cylindrical histone octamer. About 145 bp of DNA coils around this functional octamer to create a nucleosome, the basic unit of chromatin structure. Due to this highly conserved function, histone genes are among the most evolutionarily conserved genes.

Histones make up the chromatin structure of the genome and they are highly expressed to meet the high demand. High levels of histone expression are accomplished by gene duplications that form large gene families existing in tandem copies in the genome. The number of histone genes varies among species. In the yeast *Saccharomyces cerevisiae*, there are two identical copies of each of H2A, H2B, H3, and H4 (Mardian and Isenberg 1978). The H1 is present as a single copy, making a total of nine histone genes in the 12 Mb of the yeast. In higher eukaryotic organisms, two types of histone gene organization are commonly seen: (1) the histone genes are clustered and tandemly repeated, or (2) the histone genes are clustered, but no tandem duplication is seen. Sea urchins, worms, frogs, fish, and *Drosophila* all have histone genes organized in a tandemly duplicated manner. In these species, the core histone genes are in repeating units ranging from 750 to 2,000 times in the genome. Species such as chicken, human, and mouse have histone genes that are located in various clusters in the genome without a clear repeating pattern. For these species, the number of histone genes is between 50 and 200, and, while the histone genes are located in a cluster, they show little resemblance to a tandemly repeating unit.

The highly repetitive nature of histone genes makes characterization difficult. However, the highly conserved sequences offer advantages for the adoption of polymerase chain reaction (PCR). PCR primers can be designed based on gene or EST sequences of histone genes from the species of interest, or even from a closely related species. The gene sequences can then be obtained without any problem. The gene arrangement of the histone genes within the repeat unit can be determined by using PCR primers in the intergenic regions, allowing amplification across different histone genes. In rainbow trout and Atlantic salmon, the histone genes are arranged in an order of H4-H2B-H1-H2A-H3, with an external gene region of approximately 13 kb. Along with the gene region of approximately 5.1 kb, the histone cluster repeat unit in the salmonids appears to be approximately 19–20 kb (Pendas et al. 1994; Ng and Davidson, personal communications).

Determination of the number of units of a given repeat can be difficult. Most often, this can be assessed by hybridization of histone probes to bacterial artificial chromosome (BAC) filters. For instance, if 600 BACs are positive to the histone gene probes on a 10 × BAC filter, 60 BACs per genome contain histone genes. The total size of the 60 BACs can be estimated from the average size of the BAC library, minus the average overlapping regions of BACs. For instance, the channel catfish CHORI BAC library has an average insert size of 161 kb. The maximum size that can be covered by the 60 BACs is 60 × 161 kb = 9,660 kb. If the repeating unit is 20-kb long, then the maximal number of repeat units is 9,660/20 = 483. However, this estimate is on the high side because the 600 BACs have large overlapping regions. The number of repeat units can be adjusted by subtracting the average overlapping region out of the average length of the BAC. For example, if on average, 30 kb overlapping regions exist in the BAC-based contigs, then the total length covered by the 60 BAC/genome is 131 kb × 60 = 7,860 kb. Considering that each repeat unit is 20 kb long, then the number of repeat units is 7,860/20 = 393. Another approach that can provide a more accurate estimation for the number of repeat units is perhaps through the use of quantitative real time PCR.

The rDNA Cistron

Ribosomal RNAs are involved in structures of ribosomes for the translation machinery. There are four rRNAs: 18S, 5.8S, 28S, and 5S rRNA. The 18S, 5.8S, and 28S rRNAs are encoded as a cistronic unit in the genome. This rRNA precursor undergoes posttranscriptional cleavage to process the cistronic RNA into three functional rRNAs. The 18S rRNA becomes a part of the small ribosomal subunit whereas the 28S and 5.8S rRNA are components of the large ribosomal subunit.

Over 80% of the cellular RNA mass is composed of rRNAs. The large demand for rRNA is met in the cells by gene duplication of these genes to form the large tandem repeat. In addition, their gene products are required in a stoichiometric one to one ratio, and thus their expression must be coordinated. The evolutionary processes have produced some of the most efficient ways for this coordination by using cistrons (i.e., these genes are present together and their expression is under the control of the same promoter). Cistrons are common in prokaryotic organisms, but rare in eukaryotes. The rDNA cistrons are one of the rare examples for the use of cistrons in eukaryotic organisms.

The rDNA organization varies according to species. In humans, the 250 or so copies of the rDNA repeats are clustered on acrocentric chromosome arms of 13p, 14p, 15p, 21p, and 22p. The location of rDNA in fish and their organization is not yet well understood. In zebrafish, it is believed that 6–8 chromosomes harbor rDNA repeat clusters, but 6–12 chromosomes may harbor rDNA clusters in sturgeons (Gornung et al. 1997, Phillips and Reed 2000, Fontana et al. 2003). PCR primers can be designed for 18S, 28S, and 5.8S rRNAs based on gene sequences of these genes from the species of interest. In order to estimate the cistron size, external primers facing toward the external transcribed spacer can be used. As discussed above with the histone gene clusters, the estimation for the number of repeat units can be difficult, but BAC-based hybridization can provide a fairly accurate estimate. The Atlantic salmon genome is estimated to contain 1,000–1,500 copies of the rDNA repeats (Moran et al. 1997). Very similar approaches can be used for the characterization of tRNA genes that also exist in tandem repeats.

It is logical to think that highly duplicated genes should be subjected to more mutations due to reduced evolutionary pressure. How then can many highly duplicated genes such as rRNA or histone genes maintain their integrity? Several theories exist including coincidental evolution (concerted evolution or coevolution to explain the situation). Coincidental evolution describes the ability of two related genes to evolve together as though constituting a single locus. When members of a repetitive family are compared, greater sequence similarity is found within a species than between species, suggesting that members within a repetitive family do not evolve independently of each other. The detailed mechanisms may differ: a mutation can be removed; or the mutated copy takes over leading to both copies with a mutation; or homogenization by enzymes through strand change and editing (Liao 1999). Another interesting hypothesis is sudden correction: Every so often the entire gene cluster is replaced by a new set of copies derived from one or a few of the copies. While these models may explain the high sequence conservation of rRNA and histone genes, the truth may be that the evolutionary pressure for these genes has never been reduced as assumed by many. The demand for the gene products of highly duplicated genes may be so high that each duplicated copy is still absolutely required so that any mutation of even a single copy may lead to a detrimental impact on the organism.

Determination of Copy Numbers of Repetitive Elements

It is difficult to determine the exact copy numbers of repetitive elements. However, their copy numbers in the genome can be assessed by hybridization, or in some cases by real time PCR (e.g., Chen et al. 2006). The choice of approaches depends on the nature of the repetitive elements. If the repetitive elements or genes have exactly the same or highly conserved sequences, then real time quantitative PCR can be used to determine the copy numbers. However, in most cases, the repetitive elements are highly related in sequence, but carry many base substitutions, or exist as remnants with only part of the entire sequences. In this later case, hybridization is an effective strategy to assess the copy numbers. The copy number of the repetitive elements of interest can be assessed by comparing the hybridization signals of the genomic sample with that of a serially diluted sample with known copy number of the same target molecules. The major problem of a hybridization-based approach for the assessment of copy numbers is the saturation of signals (the after-black-is-black nature of signals). Therefore, serial dilutions should be made with genomic DNA, as well as the control DNA such as a plasmid containing the target sequence. Each copy of plasmid contains one copy of the target sequence. For instance, the hybridization signal of 2 μg genomic DNA is equivalent to that of 8 nanogram (ng) plasmid containing one copy of the target sequence. There are 1.4×10^9 molecules (copies) of plasmids in the 8 ng plasmid DNA (moles of plasmid $= 8 \times 10^{-9}$ gram/molecular weight [5,000 bp × 660/bp], number of molecules = moles $\times 6.023 \times 10^{23}$ per mole). There are 2,000,000 genomes in 2 μg of genomic DNA (the catfish genome size is 1 pg DNA per cell). Thus, the copy number of the repetitive element $= 1.4 \times 10^9$ copies/2,000,000 = 700 copies. When determining copy numbers of repetitive elements, one should carefully select the hybridization probes so that they cover the typical region of the repetitive elements under consideration.

Conclusion

Repetitive elements comprise a major fraction of eukaryotic genomes. In humans, over 50% of their 3 billion bp genome is composed of repetitive elements. The proportion of repetitive elements in the genomes of aquaculture species is yet to be discovered. Once regarded as junk DNA, repetitive elements are once again gaining their popularity not only because of their abundance, but also because of their potentially unrealized biological functions. Understanding of repetitive elements and their organizations is by no means an easy task. Many of the current approaches involve bioinformatic analysis, though the role of solid experimental approaches cannot be neglected. Recent studies also indicate that regulation of biological functions can also be achieved through gene copy numbers, thus accurate determination of gene families and their copy numbers may prove to be important. Understanding of the repeat structure in aquaculture species will facilitate studies in linkage mapping, physical mapping, and lay the foundation for entire genome sequencing. Even after the entire genome is sequenced, correct genome assembly will rely on the understanding and comprehension of repeat structures of the genome.

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

Bailey JA, G Liu, and EE Eichler. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet, 73, pp. 823–834.

Britten RJ and DE Kohne. 1968. Repeated sequences in DNA. Science, 161, pp. 529–540.

Brosius J. 2003. How significant is 98.5% "junk" in mammalian genomes. Bioinformatics, 19(Suppl 2), II35.

Capy P, G Gasperi, C Biemont, and C Bazin. 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity, 85, pp. 101–106.

Charlesworth B, C Langley, and W Stephan. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics, 112, pp. 947–962.

Chen Q, M Book, X Fang, A Hoeft, and F Stuber. 2006. Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR. J Immunol Methods, 308, pp. 231–240.

Fontana F, M Lanfredi, L Congiu, M Leis, M Chicca, and R Rossi. 2003. Chromosomal mapping of 18S–28S and 5S rRNA genes by two-colour fluorescent *in situ* hybridization in six sturgeon species. Genome, 46, pp. 473–477.

Gong Z, T Yan, J Liao, SE Lee, J He, and CL Hew. 1997. Rapid identification and isolation of zebrafish cDNA clones. Gene, 201, pp. 87–98.

Goodier JL and WS Davidson. 1994. *Tc1* transposon-like sequences are widely distributed in salmonids. J Mol Biol, 241, pp. 26–34.

Gornung E, I Gabrielli, S Cataudella, and L Sola. 1997. CMA3-banding pattern and fluorescence *in situ* hybridization with 18S rRNA genes in zebrafish chromosomes. Chromosome Res, 5, pp. 40–46.

He L, Z Zhu, AJ Faras, KS Guise, PB Hackett, and AR Kapuscinski. 1992. Characterization of *Alu* I repeats of zebrafish (*Brachydanio rerio*). Mol Marine Biol Biotechnol, 1, pp. 125–135.

Heierhorst J, K Lederis, and D Richter. 1992. Presence of a member of the *Tc1*-like transposon family from nematodes and *Drosophila* within the vasotocin gene of a primitive vertebrate, the Pacific hagfish *Eptatretus stouti.* Proc Natl Acad Sci USA, 89, pp. 6798–6802.

Henikoff S. 1992. Detection of Caenorhabditis transposon homologs in diverse organisms. New Biol, 4, pp. 382–388.

Holmes I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. Genome Res, 12, pp. 1152–1155.

Hong SH, JS Kim, SY Lee, YH In, SS Choi, J-K Rih, CH Kim, H Jeong, CG Hur, and JJ Kim. 2004. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens.* Nat Biotechnol, 22, pp. 1275–1281.

Ivics Z, PB Hackett, RH Plasterk, and Z Izsvak. 1997. Molecular reconstruction of *Sleeping Beauty,* a *Tc1*-like transposon from fish, and its transposition in human cells. Cell, 91, pp. 501–510.

Ivics Z, Z Izsvak, A Minter, and PB Hackett. 1996. Identification of functional domains and evolution of *Tc1*-like transposable elements. Proc Natl Acad Sci USA, 93, pp. 5008–5013.

Izsvak Z, Z Ivics, and PB Hackett. 1995. Characterization of a *Tc1*-like transposable element in zebrafish (*Danio rerio*). Mol Gen Genet, 247, pp. 312–322.

Kawakami K, A Shima, and N Kawakami. 2000. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. Proc Natl Acad Sci USA, 97, pp. 11403–11408.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. *Science,* 303, pp. 1626–1632.

Kidwell MG and DR Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution, 55, pp. 1–24.

Koga A and H Hori. 2000. Detection of *de novo* insertion of the medaka fish transposable element Tol2. Genetics, 156, pp. 1243–1247.

Koga A and H Hori. 2001. The Tol2 transposable element of the medaka fish: an active DNA-based element naturally occurring in a vertebrate genome. Genes Genet Syst, 76, pp. 1–8.

Kurtz S and C Schleiermacher. 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics, 15, pp. 426–427.

Lam WL, P Seo, K Robison, S Virk, and W Gilbert. 1996. Discovery of amphibian *Tc1*-like transposon families. J Mol Biol, 257, pp. 359–366.

Levinson G and GA Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol, 4, pp. 203–221.

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet, 64, pp. 24–30.

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

Mardian JK and I Isenberg. 1978. Yeast inner histones and the evolutionary conservation of histone-histone interactions. Biochemistry, 17, pp. 3825–3833.

Milklos G and A Gill. 1982. Nucleotide sequences of highly repeated DNAs: compilation and comments. Genet Res, 39, pp. 1–30.

Moran P, KM Reed, J Perez, TH Oakley, RB Phillips, E Garcia-Vazquez, and AM Pendas. 1997. Physical localization and characterization of the *Bgl* I element in the genomes of Atlantic salmon (*Salmo salar* L.) and brown trout (*S. trutta* L.). Gene, 194, pp. 9–18.

Oosumi T, WR Belknap, and B Garlick. 1995. Mariner transposons in humans. Nature, 378, p. 672.

Pendas AM, P Moran, and E Garcia-Vazquez. 1994. Organization and chromosomal location of the major histone cluster in brown trout, Atlantic salmon and rainbow trout. Chromosoma, 103, pp. 147–52.

Peterson DG, WR Pearson, and SM Stack. 1998. Characterization of the tomato (*Lycopersicon esculentum*) genome using *in vitro* and *in situ* DNA reassociation. Genome, 41, pp. 346–356.

Peterson DG, SR Schulze, EB Sciara, SA Lee, JE Bowers, A Nagel, N Jiang, DC Tibbitts, SR Wessler, and AH Paterson. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res, 12, pp. 795–807.

Phillips RB and KM Reed. 2000. Localization of repetitive DNAs to zebrafish (*Danio rerio*) chromosomes by fluorescence *in situ* hybridization (FISH). Chromosome Res, 8, pp. 27–35.

Radice AD, B Bugaj, DH Fitch, and SW Emmons. 1994. Widespread occurrence of the *Tc1* transposon family: *Tc1*-like transposons from teleost fish. Mol Gen Genet, 244, pp. 606–612.

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.

Shapiro JA. 1999. Transposable elements as the key to a 21st century view of evolution. Genetica, 107, pp. 171–179.

Sinzelle L, N Pollet, Y Bigot, and A Mazabraud. 2005. Characterization of multiple lineages of *Tc1*-like elements within the genome of the amphibian *Xenopus tropicalis.* Gene, 349, pp. 187–196.

Smit AF and AD Riggs. 1996. *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA, 93, pp. 1443–1448.

Sonnhammer EL and R Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene, 167, pp. GC1–10.

Stephan W. 1989. Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol, 6, pp. 198–212.

Stephan W and S Cho. 1994. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics, 136, pp. 333–341.

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 fish model species. Anim Genet, 37, pp. 321–326.

Zhi D, BJ Raphael, AL Price, H Tang, and PA Pevzner. 2006. Identifying repeat domains in large genomes. Genome Biol, 7, p. R7.