Minireview

Fugu: a compact vertebrate reference genome

Byrappa Venkatesh*, Patrick Gilligan, Sydney Brenner

Institute of Molecular and Cell Biology, National University of Singapore, 30 Medical Drive, Singapore 117609

Received 5 May 2000

Edited by Gunnar von Heijne

Abstract At 400 Mb, the Japanese pufferfish, *Fugu rubripes*, has the smallest vertebrate genome but has a similar gene repertoire to other vertebrates. Its genes are densely packed with short intergenic and intronic sequences devoid of repetitive elements. It likely has a mutational bias towards DNA elimination and is probably close to a 'minimal' vertebrate genome. As such it is a useful reference genome for gene discovery and gene validation in other vertebrates. Its usefulness in the discovery of conserved regulatory elements has already been demonstrated. The Fugu genome sequence is a good complement to genetic studies in other vertebrates.

Key words: Vertebrate genome; Fugu; Gene structure; Phylogenetic relationship

1. Introduction

Vertebrates are the most diverse and successful group of animals that have dominated both aquatic and terrestrial habitats. They have conserved development, anatomy and physiology and are distinguished from invertebrates by complex systems such as the skeletal, immune, nervous, endocrine and circulatory systems. An expansion in the gene repertoire through genome duplication during the transition from invertebrates to vertebrates possibly contributed to this complexity. The international Human Genome Project which aims to identify and characterize all human genes in order to facilitate the understanding of human biology has given an impetus to vertebrate genomics. By the time this article appears in print, it is expected that the first draft of the entire human genome sequence will be available in the public database.

Besides the human genome, several other vertebrate genomes are being investigated to gain insight into their biology and as models to interpret the human genome. These models include mammals (mouse and rat), a bird (chicken), frogs (*Xenopus laevis* and *Xenopus tropicalis*) and fish such as the zebrafish, medaka and Fugu (Table 1). The zebrafish, medaka, *Xenopus* and chicken are particularly useful in developmental studies, whereas the mouse and rat are used as mammalian models for genetic and physiological studies. The advent of large scale DNA sequencing revealed the potential use of a compact vertebrate genome as a reference to complex genomes to help discover novel vertebrate genes and gene regulatory elements and understand genome architecture. The Japanese pufferfish, Fugu, was proposed to fill this role [1]. Initial characterization of the Fugu genome based on random sequencing and screening of a genomic library with single copy probes revealed a 400 Mb genome containing less than 10% repetitive DNA and small introns. Since it must contain sufficient genes to specify a vertebrate, a high gene density was predicted. In this report we review the phylogenetic position of Fugu in relation to other vertebrate models and highlight the characteristics of its genome which makes it an ideal reference genome for all vertebrates.

2. Phylogenetic position

The model vertebrates listed in Table 1 are all grouped under the taxon Osteichthyes, or bony fish. Their common ancestor, a bony fish that swam the seas over 400 million years ago, diverged to give rise to lobe-finned fish (sarcopterygians), from which mammals and other tetrapods evolved, and ray-finned fish (actinopterygians), from which teleosts evolved.

The teleosts, comprising more than 24 000 species, are the most diverse and species-rich group of vertebrates. The three model fish, the Fugu (family Tetraodontidae, order Tetraodontiformes), medaka (family Adrianichthyidae, order Beloniformes) and zebrafish (family Cyprinidae, order Cypriniformes) differ in several morphological characters and are thus classified under different orders (Fig. 1). The evolutionary divergence times of these three fish lineages are not known at present due to the paucity of fossil records. However, on the basis of their classification [2,3], we can conclude that Fugu and medaka are closer to each other than either is to zebrafish. It should be noted that each of the three fish is at the same phylogenetic distance from humans, mouse or any other tetrapod, as the ray-finned ancestor of the three fish and the lobe-finned ancestor of tetrapods diverged from a common ancestor.

The Fugu and other pufferfish are classified under the order Tetraodontidae which is nested within the series Percomorpha. The phylogenetic relationship of pufferfish to other members of the order Tetraodontidae is not well established [4] and neither is the relationship of Tetraodontidae to other percomorphs. While some shared morphological characters associate Tetraodontids with surgeonfish (family Acanthuridae, order Perciformes), another set of characters suggests that the order Zeiformes which includes fish such as dories and parazen (not shown in Fig. 1 due to a lack of information on their DNA content) is the sister group [2]. This ambiguity might be resolved by phylogenetic analyses of appropriate molecular data. The phylogenetic placement of pufferfish

*Corresponding author. Fax: (65)-779-1117.
E-mail: mcbbv@imcb.nus.edu.sg

0014-5793/00/$20.00
PII: S0014-5793(00)01659-8

should help in tracing the evolutionary origin of Fugu and help in understanding the mechanism underlying its genome compaction.

3. Genome size

The vertebrate genome size varies greatly between, and even within, lineages. While lungfish and salamanders stand out as having huge genomes, pufferfish have an exceptionally small genome (Table 1 and Fig. 1). The mosaicism of genome size in most of the lineages (Table 1) suggests that genome expansion or compression has occurred independently in each lineage. While genome expansion appears to occur by tandem duplication of loci, duplication of chromosomes or the entire genome and the accumulation of repetitive sequences, a small genome is either the result of accumulated deletions or is the primitive state.

In Fugu, one of the main factors contributing to the small genome size is the scarcity of dispersed repeats. So far there is no evidence of major vertebrate repeat sequences such as *Alus* and SINES [1,5], although Fugu does have a small number of retrotransposon clusters and elements similar to transposons or polyproteins [5–7]. Evidently the Fugu genome is not permissive for these elements.

On the basis of the presence of duplicated copies of Hox clusters and two paralogs of several mammalian genes in zebrafish, it has been proposed that an early ray-finned fish underwent regional or whole genome duplication [8]. This implies that the small genomes, like Fugu and several other teleosts shown in Fig. 1, are the result of secondary loss of genes and/or chromosomes. Although Fugu appears to have a duplicate copy of one of the Hox clusters [8], there is no evidence for large scale duplication in the genome. Extensive searches for members of multigene families in the Fugu by polymerase chain reaction screening and library probing have not identified significant numbers of either duplicated genes or pseudogenes [9–14]. Although nine actin genes were identified in the Fugu as compared to six known so far in mammals, one of the new actin genes in the Fugu is the result of a tandem duplication and another new gene has a unique exon–intron structure suggesting that they are not the result of genome duplication [11]. If there was a genome duplication in the ancestral lineage as hypothesized, Fugu seems to have eliminated the extra copies efficiently.

If the compactness of the pufferfish genome is a derived, rather than a primitive trait, it has to be explained by accumulated deletions. Like Fugu, *Drosophila* has small introns and intergenic regions, and scarce pseudogenes and repeats. It was recently shown that the Hawaiian cricket, which has a genome 11 times larger than *Drosophila*, loses DNA 40 times more slowly [15]. Thus, a strong bias towards deletion mutations would lead to the loss of DNA that was too weakly selected to overcome the bias, or was extraneous. Such a bias would account for the small genomes of pufferfish. Investigation of insertion/deletion frequencies in the close relatives of Fugu which have larger genomes should confirm if there is indeed a deletion bias in the Fugu.

4. Gene density and synteny

Fugu genes are compressed several-fold relative to human homologs due to small introns [16–24]. Indeed, most of the Fugu introns are smaller than 300 bp. As a result, genes which contain unusually large introns in other vertebrates appear dramatically compressed in the Fugu (Table 2).

Random sequencing [1], sequence skimming of randomly selected cosmids [5] and sequencing of select loci [25–28] all have confirmed a high gene density in the Fugu. Whereas the human genome is just 3% coding sequence [29], Fugu contains 17% [5] coding sequence. This predicts 68 000 genes of 1 kb coding length in the Fugu, at a density of one gene per 6 kb which is close to that found in the invertebrates, the nematode (5 kb per gene) [30] and fruit fly (8.5 kb per gene) [31]. Is the Fugu gene density a primitive trait retained from the ancestral duplicate invertebrate genome? This seems unparsimonious, as the pufferfish belong to a highly derived teleost lineage and the majority of teleosts that diverged earlier from the ancestral lineage have larger genomes (Fig. 1). It is more likely that a strong bias towards deletions in the Fugu has compacted the genome to a size that is close to a minimal vertebrate genome.

The short-range gene order is conserved between the Fugu and human genomes in many [5,25,27,32–34], but not all [20,26,35], loci. Because of the compressed intergenic regions, the conserved Fugu gene clusters occupy a much shorter region than their human homologues. Interestingly, in some regions with disrupted gene order, the synteny (genes on the same piece of DNA) is still conserved [20,26,34] suggesting

Table 1
Genome sizes of 'model' vertebrates

	Haploid DNA content (pg)	Haploid genome size (Mb)	No. of chromosomes (*n*)
Mammals			
Human (*Homo sapiens*)	3.5	3000	23
Mouse (*Mus musculus*)	3.5	3000	20
Rat (*Rattus norvegicus*)	3.5	3000	21
Bird			
Chicken (*Gallus gallus*)	1.25	1200	39
Amphibians			
Xenopus laevis	3.2	3100	18
Xenopus tropicalis	1.78	1700	10
Fish			
Zebrafish (*Danio rerio*)	1.8	1700	25
Medaka (*Oryzias latipes*)	1.1	1100	24
Fugu (*Fugu rubripes*)		400	22

The chicken karyotype includes 30 microchromosomes in addition to nine macrochromosomes. *X. laevis* is a tetraploid whereas *X. tropicalis* is a diploid. The DNA content of Fugu has not been determined. References: chicken [44]; *Xenopus* [45]; zebrafish [46]; medaka [47] and Fugu [1,48].

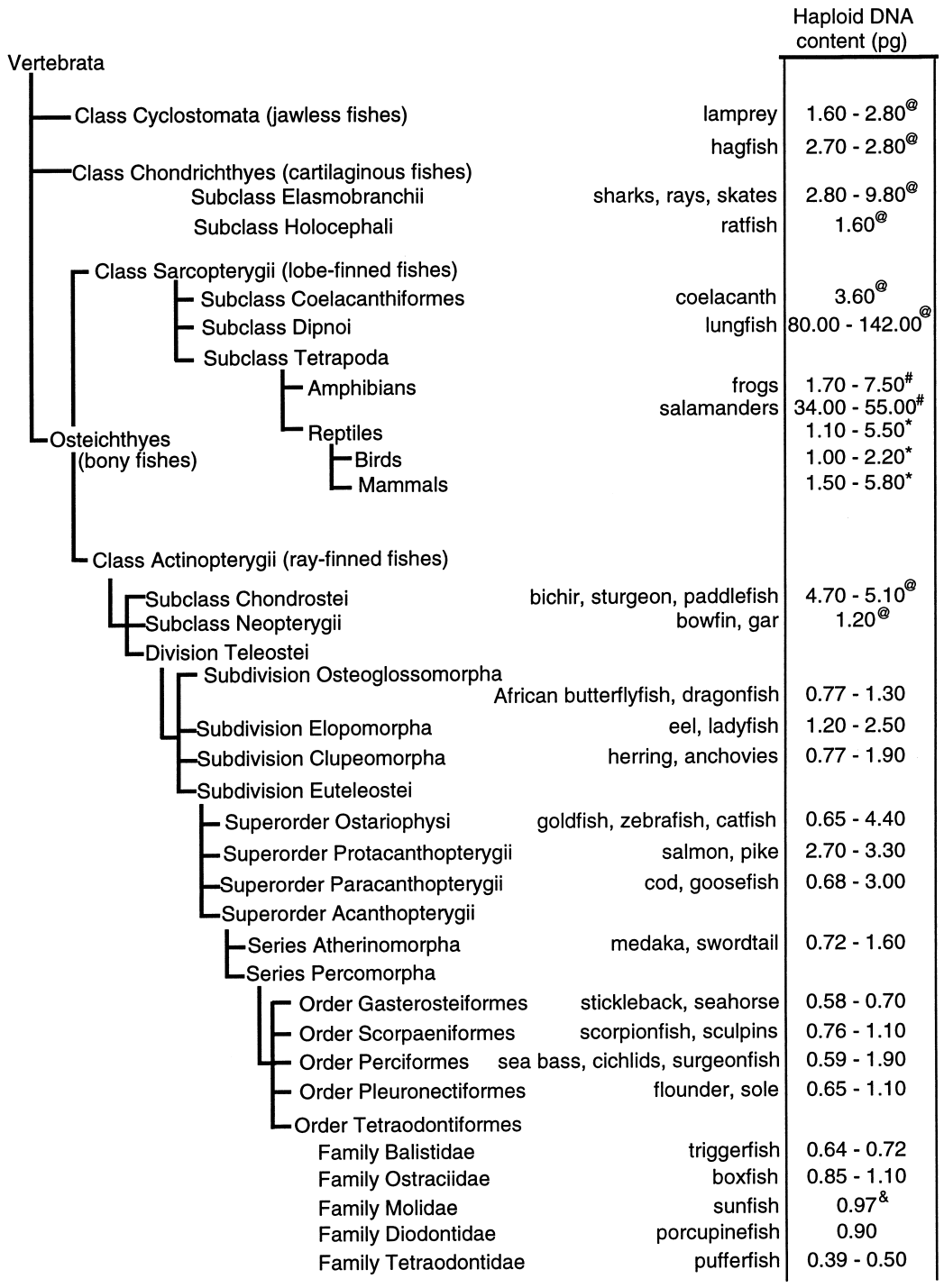

Fig. 1. Cellular DNA content of vertebrates. The phylogenetic relationships shown here are based on morphological characters [2]. Common names of only some members of each group are given. (References: @[40]; *[41]; #[42]; &Venkatesh and Brenner, unpublished; others [43]).

that the rearrangements have occurred intrachromosomally. Such intrachromosomal inversions are presumably mediated by transposons. At this stage it is not known if such inversions are typical of the Fugu lineage or widespread in other teleosts.

The loci with conserved gene order between Fugu and other vertebrates give an indication of the distribution of evolutionary breakpoints in the chromosome and should help in understanding the architecture and evolution of chromosomes. Gene linkage information in the Fugu will be useful for identifying the exact orthologs in other vertebrates, particularly for genes from multigene families which show high sequence homology between paralogs and orthologs. The linkage information can accelerate the identification and cloning of candidate genes associated with genetic diseases or mutant phenotypes in other vertebrates where the genomic sequence is not available. Recently the gene for *ferroportin1*, which transports iron from maternally-derived yolk stores to the circulation, was localized in zebrafish mutant *Weissherbst* and rapidly cloned based on its position in the Fugu genome [36].

Table 2
Comparison of large mammalian introns with their Fugu homologs

Gene name	Size of gene (kb)		Intron	Size of intron (bp)		References
	human	Fugu		human	Fugu	
HD	170	23	1	11 926	537	[16]
			2	12 286	137	
			6	7 879	230	
			29	11 927	92	
			40	10 641	131	
HisRS	18[a]	3.5	2	7 500	150	[17]
C9	90	2.9	1	22 500	77	[19]
			4	10 000	135	
			5	7 500	80	
			6	8 000	69	
			9	22 000	164	
			10	13 200	99	
NF1	351	27	1	> 100 000	2575	[20]
			7b	> 45 000	3942	
TUPLE/HIRA	100	9	1	20 600	786	[21]
			5	7 700	73	
			13	9 600	821	
			15	13 700	98	
			24	19 800	226	
APP	300	10	1	54 916	2200	[22]
			2	21 907	75	
			3	36 594	92	
			5	28 957	106	
			6	21 658	159	
			10	19 313	218	
			12	8 613	343	
			16	9 952	83	
SS	54	4.7	1	26 322	825	[23]
			7	5 337	309	
			9	7 373	110	

APP, beta-amyloid precursor protein gene; HD, Huntington's disease gene; C9, complement component 9 gene; NF1, neurofibromatosis Type 1 gene; HisRS, histidyl-tRNA synthetase gene; SS, spermine synthase gene.
[a]Hamster gene.

5. Conserved regulatory elements

A key step in understanding gene regulation is the identification of regulatory elements and is traditionally addressed by promoter deletion. This is a laborious and tedious procedure, typically hampered by massive intergenic regions in mammals. However, as pufferfish are 400 million years distant from mammals, irrelevant DNA in the promoter region will be randomized, so the conserved sequence is likely to be functional. Thus, the compact intergenic regions of the Fugu should allow swift and unambiguous detection of putative regulatory elements ('phylogenetic footprinting') by comparison with mammalian sequence.

Sequence comparison between Fugu and the mouse has identified several conserved non-coding sequences [26,32,37–39]. Such a conserved sequence in the Fugu and mouse Hoxb-4 was able to limit the expression of a reporter gene precisely to the anterior boundary of rhombomere 6/7 in transgenic mice similar to the mouse sequence [37]. Similarly, the relative contributions of several conserved elements in the Fugu and mouse Otx2 genes to the full expression pattern were identified by deleting the elements from the Fugu construct and assaying in transgenic mice [38]. In another experiment, the Fugu isotocin gene (the teleost homolog of oxytocin) introduced into transgenic rats expressed specifically in those magnocellular neurons in the SON and PVN of the hypothalamus that express rat oxytocin gene [26]. Furthermore, the Fugu gene responded to osmotic stress in the same manner as the rat gene. These experiments demonstrate stringent conservation of the DNA binding specificity of many transcription factors across very large evolutionary distances.

Since the Fugu genes and promoters are small, it is easy to manipulate them in vitro and in vivo. Fugu cosmids typically contain several genes together with the elements required for precise expression and regulation. They are ideally suited for gene regulation studies in cell lines as well as in transgenic systems.

6. Discordant introns

Several Fugu genes contain extra introns compared to their mammalian homologs, an apparent paradox for a fish with a small genome. However, we have shown that these introns are present in many other teleosts, and are a result of 'intron gain' by the teleost lineage at various stages of evolution [3]. Contrary to these genes, the rhodopsin gene in Fugu lacks introns whereas its tetrapod homolog contains several introns. This has been shown to be the result of 'intron loss' in a basal teleost lineage [3]. On the basis of the genes characterized so far, it appears that the teleost lineage is more susceptible to changes in genomic structure than the tetrapod lineage.

The loss or gain of nuclear introns are rare events occurring at different stages in the evolution of a lineage and are not easily reversible. As such, the presence or absence of introns in different lineages can be used as synapomorphs (shared derived characters) to identify evolutionary branchpoints and to trace the phylogenetic relationships. A comparison of genomic structures of Fugu and human genes will allow the

identification of discordant introns within these lineages. Tracing the origins of such introns should identify evolutionary branchpoints in vertebrate lineages and resolve the phylogenetic relationships of various groups.

7. Concluding remarks

The compact genome of the Fugu, with a high gene density, short introns and intergenic regions, is conceptually a normalized cDNA library that also contains the elements involved in gene regulation and the maintenance of chromosome architecture. Because of the compact size and lack of repetitive sequences, it can be sequenced at a much lower cost than other vertebrate genomes. As Fugu contains a similar repertoire of genes to other vertebrates, it is a powerful tool as a reference genome for comparative genomics.

The Fugu sequences will be particularly useful for validating computer-predicted hypothetical genes in the human genome, and revealing novel genes in the human genome that do not contain any known domains. A comparison of non-coding sequences between the evolutionarily distant Fugu and human genomes provides the maximum stringency for detecting conserved regulatory elements. The function of such conserved elements can be tested in cell lines or transgenic systems. Fugu cosmids typically contain genomic information equivalent to that in a large mammalian BAC or even a YAC clone and are preferred for expression studies in cell lines and transgenic animals as they are easy to manipulate and less susceptible to recombination.

The genetic studies in other teleosts such as zebrafish and medaka, which are evolutionarily closer to Fugu than to tetrapods, can also benefit from the Fugu genome sequence. The sequences and their linkage in Fugu can accelerate positional cloning of candidate genes associated with various zebrafish and medaka mutants. Fugu cosmids containing the orthologs of candidate genes should carry all the regulatory elements required to faithfully recapitulate the spatial and temporal expression pattern of the candidate and, thus, efficiently rescue mutant phenotypes.

References

[1] Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S. (1993) Nature 366, 265–268.
[2] Nelson, J.S. (1994) Fishes of the World, John Wiley, New York.
[3] Venkatesh, B., Ning, Y. and Brenner, S. (1999) Proc. Natl. Acad. Sci. USA 96, 10267–10271.
[4] Leis, J.M. (1983) in: Ontogeny and Systematics of Fishes (Moser, H.G., Ed.), Special Publication No. 1, pp. 459–463. Am. Soc. Ichthyol. Herpetol.
[5] Elgar, G., Clark, M.S., Meek, S., Smith, S., Warner, S., Edwards, Y.J.K., Bouchireb, N., Cottage, A., Yeo, G.S.H., Umrania, Y., Williams, G. and Brenner, S. (1999) Genome Res. 9, 960–971.
[6] Poulter, R. and Butler, M. (1998) Gene 215, 241–249.
[7] Poulter, R., Butler, M. and Ormandy, J. (1999) Gene 227, 169–179.
[8] Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M. and Postlethwait, J.H. (1998) Science 282, 1711–1714.
[9] Macrae, A.D. and Brenner, S. (1995) Genomics 25, 436–446.
[10] Sarwal, M.M., Sontag, J.M., Hoang, L., Brenner, S. and Wilkie, T.M. (1996) Genome Res. 6, 1207–1215.
[11] Venkatesh, B., Tay, B.H., Elgar, G. and Brenner, S. (1996) J. Mol. Biol. 259, 655–665.
[12] Yamaguchi, F., Macrae, A.D. and Brenner, S. (1996) Genomics 35, 603–605.
[13] Yamaguchi, F. and Brenner, S. (1997) Gene 191, 219–223.
[14] Naito, T., Saito, Y., Yamamoto, J., Nozaki, Y., Tomura, K., Hazama, M., Nakanishi, S. and Brenner, S. (1998) Proc. Natl. Acad. Sci. USA 95, 5178–5181.
[15] Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L. and Shaw, K.L. (2000) Science 287, 1060–1062.
[16] Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S. and Beck, S. (1995) Nat. Genet. 10, 67–76.
[17] Brenner, S. and Corrochano, L.M. (1996) Proc. Natl. Acad. Sci. USA 93, 8485–8489.
[18] Schofield, J.P., Elgar, G., Greystrong, J., Lye, G., Deadman, R., Micklem, G., King, A., Brenner, S. and Vaudin, M. (1997) Genomics 45, 158–167.
[19] Yeo, G.S., Elgar, G., Sandford, R. and Brenner, S. (1997) Gene 200, 203–211.
[20] Kehrer-Sawatzki, H., Maier, C., Moschgath, E., Elgar, G. and Krone, W. (1998) Gene 222, 145–153.
[21] Llevadot, R., Estivill, X., Scambler, P. and Pritchard, M. (1998) Gene 208, 279–283.
[22] Villard, L., Tassone, F., Crnogorac-Jurcevic, T., Clancy, K. and Gardiner, K. (1998) Gene 210, 17–24.
[23] Boeddrich, A., Burgtorf, C., Crollius, H.R., Hennig, S., Bernot, A., Clark, M., Reinhardt, R., Lehrach, H. and Francis, F. (1999) Genomics 57, 164–168.
[24] Reboul, J., Gardiner, K., Monneron, D., Uze, G. and Lutfalla, G. (1999) Genome Res. 9, 242–250.
[25] Trower, M.K., Orton, S.M., Purvis, I.J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C.G., Elgar, G., Sherrington, R., Rogaev, E.I., St George-Hyslop, P., Brenner, S. and Dykes, C.W. (1996) Proc. Natl. Acad. Sci. USA 93, 1366–1369.
[26] Venkatesh, B., Si-Hoe, S.L., Murphy, D. and Brenner, S. (1997) Proc. Natl. Acad. Sci. USA 94, 12462–12466.
[27] Miles, C., Elgar, G., Coles, E., Kleinjan, D.J., van Heyningen, V. and Hastie, N. (1998) Proc. Natl. Acad. Sci. USA 95, 13068–13072.
[28] Gellner, K. and Brenner, S. (1999) Genome Res. 9, 251–258.
[29] Dunham, I., Shimizu, N., Roe, B.A. and Chissoe, S. et al. (1999) Nature 402, 489–495.
[30] The Worm Consortium, (1998) Science 282, 2012–2018.
[31] Adams, M.D. et al. (2000) Science 287, 2185–2195.
[32] How, G.F., Venkatesh, B. and Brenner, S. (1996) Genome Res. 12, 1185–1191.
[33] Aparicio, S., Hawker, K., Cottage, A., Mikawa, Y., Zuo, L., Venkatesh, B., Chen, E., Krumlauf, R. and Brenner, S. (1997) Nat. Genet. 16, 79–83.
[34] Brunner, B., Todt, T., Lenzner, S., Stout, K., Schulz, U., Ropers, H.H. and Kalscheuer, V.M. (1999) Genome Res. 9, 437–448.
[35] Gilley, J. and Fried, M. (1999) Hum. Mol. Genet. 8, 1313–1320.
[36] Donovan, A., Brownlie, A., Zhou, Y., Shepard, J., Pratt, S.J. and Moynihan, J. et al. (2000) Nature 403, 776–781.
[37] Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R. and Brenner, S. (1995) Proc. Natl. Acad. Sci. USA 92, 1684–1688.
[38] Kimura, C., Takeda, N., Suzuki, M., Oshimura, M., Aizawa, S. and Matsuo, I. (1997) Development 124, 3929–3941.
[39] Rowitch, D.H., Echelard, Y., Danielian, P.S., Gellner, K., Brenner, S. and McMahon, A.P. (1998) Development 125, 2735–2746.
[40] Hinegardner, R. (1976) Comp. Biochem. Physiol. 55B, 367–370.
[41] Olmo, E., Capriglione, T. and Odierna, G. (1989) Comp. Biochem. Physiol. 92B, 447–453.
[42] Tiersch, T.R., Chandler, R.W., Wachtel, S.S. and Elias, S. (1989) Cytometry 10, 706–710.
[43] Hinegardner, R. and Rosen, D.E. (1972) Am. Nat. 106, 621–644.
[44] Bloom, S.E., Delany, M.E. and Muscarella, D.E. (1993) in: Manipulation of the Avian Genome (Etches, R.J. and Gibbons, A.M.V., Eds.), pp. 39–59. CRC Press, Boca Raton, FL.
[45] Amaya, E., Offield, M.F. and Grainger, R.M. (1998) Trends Genet. 14, 253–255.
[46] Beir, D.R. (1998) Genome Res. 8, 9–17.
[47] Ozato, K., Wakamatsu, Y. and Inoue, K. (1992) Mol. Mar. Biol. Biotech. 1, 346–354.
[48] Miyaki, K., Tabeta, O. and Kayano, H. (1995) Fish. Sci. 61, 594–598.