Minireview

Sushi gets serious: the draft genome sequence of the pufferfish *Fugu rubripes*

Martin S Taylor and Colin AM Semple

Address: MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.

Correspondence: Colin AM Semple. Email: Colin.Semple@hgu.mrc.ac.uk

Published: 28 August 2002

Genome ***Biology*** 2002, **3(9)**:reviews1025.1–1025.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/9/reviews/1025

Abstract

The publication of the *Fugu rubripes* draft genome sequence will take this fish from culinary delicacy to potent tool in deciphering the mysteries of human genome function.

Gone fishing

Fugu rubripes, also known as the Japanese pufferfish or *Torafugu*, is a teleost fish belonging to the order Tetraodontiformes and is a member of the gnathostomes (jawed vertebrates). In Japan, *F. rubripes* is eaten as a delicacy and is considered the most delicious of all fish; a teaspoonful of *F. rubripes* testes mixed with hot sake is also traditionally drunk as an aphrodisiac. Dining on *F. rubripes* elicits fear, excitement and bravado, for this is culinary Russian roulette: some parts of the fish can contain a potent neurotoxin (tetrodoxin), and improper preparation can be fatal. Interest in *F. rubripes* is not only limited to taste and toxicity, however. Following the discovery [1] that tetraodontiforms have a genome around one sixth of the size of most vertebrates, or even less, its genome has aroused considerable curiosity culminating, so far, in the publication of its draft genome sequence [2].

The business of simply sequencing a genome is fairly straightforward these days, compared to the difficulties of assembling all the pieces of sequence correctly and identifying (annotating) all the encoded genes. The central problem of genomic sequence annotation is gene prediction, and for the most part this is still done computationally. Traditionally, genome annotators use a combination of information to predict protein-coding gene structures: *ab initio* exon predictions (predictions of coding sequence made by a computer program on the basis of statistical measures of features such as codon usage, initiation signals, polyadenylation signals and splice sites), and similarity to expressed sequences and proteins. But such approaches often yield results that are far from accurate or comprehensive. This is particularly true of eukaryotic genomes, where often only a small proportion of the genome encodes protein, and phenomena such as overlapping genes and alternatively spliced exons muddy the waters further. Add the presence of non-coding and antisense RNA genes and one can appreciate why the gene-prediction literature is full of metaphors involving needles, haystacks, creeks and paddles. Luckily, it has become clear that many multicellular organisms share a large proportion of their gene repertoire with others, and that the genomic sequences conserved between them provide good indications of the locations of genes and, perhaps more importantly, of the non-coding elements controlling gene expression.

With this in mind, in 1993 Sydney Brenner and colleagues established *F. rubripes* as a model organism for genomics: here was a compact vertebrate genome of relatively modest size (approximately 365 megabases (Mb), or around one eighth the size of the human genome), a dearth of interspersed repetitive elements, but presumably a similar gene repertoire to that of other vertebrates [3]. Later work verified that there are regions of the *F. rubripes* genome where exact conservation of gene order has been maintained with mammals (during almost half a billion years of evolution) and more numerous regions where broad synteny (conservation in the order of genes along a chromosome) has been

conserved but the precise gene order and orientation has changed [4]. This work laid the groundwork for the International *F. rubripes* Genome Consortium, which was initiated in November 2000. The consortium includes the Singapore Biomedical Research Council's Institute for Molecular and Cell Biology; the Department of Energy Joint Genome Institute (JGI), Celera Genomics, Myriad Genetics, the Salk Institute and the Institute of Systems Biology in the USA; and the University of Cambridge and the MRC Human Genome Mapping Resource Centre in the UK. Moves are now afoot to begin the task of 'finishing' the *F. rubripes* genome, which could make it the first non-human vertebrate to be fully sequenced. Finished sequence - in which all the gaps are closed and misassemblies corrected - could be especially useful in the prediction of more subtle patterns of conservation, as is often seen with regulatory elements, since small regions of weak similarity are difficult to identify in the presence of gapped and/or misassembled draft sequence. In addition, finished *F. rubripes* sequence could be used to 'walk' over gapped or misassembled patches of human or mouse genome, guiding the assembly of these troublesome regions in the mammals' relatively voluminous, repeat-rich genomes.

Trawling for sequence

A whole-genome shotgun method [5] was used to produce the draft *F. rubripes* genome sequence. Seven shotgun libraries were prepared from the genomic DNA of a single male fish, a strategy designed to minimize assembly complications due to allelic variation. Libraries were constructed with inserts of approximately 2, 5 and 40 kilobases, with the majority of sequencing reads being from the 2 kb libraries. An additional large-insert library with an average, but highly variable, insert size of 70 kb was included, from which a relatively small number of orienting sequence reads of clone ends was produced; the 70 kb insert library was prepared with DNA from a different individual from the other seven libraries, however. Clones targeted for sequencing were sequenced from both ends, with 76% of clones producing usable sequence from both ends. The opposite ends form a 'mate pair', with the inverse relative orientation of mate pair sequences and knowledge of approximate clone size being the key to producing an accurate assembly. In total, over 3.7 million sequence traces were produced that passed initial quality checks, averaging around 600 base-pairs of usable sequence per trace. This is equivalent to every base of the genome being sequenced 5.6 times (5.6x coverage), assuming, as the authors have, a total genome size of 380 Mb. These calculations are conservative, as the best estimates of *F. rubripes* genome size [1,3], as well as the results of Aparicio *et al.* [2], point to a total genome size of only approximately 365 Mb.

Assembly of sequences was carried out using a new suite of computational tools with the collective name JAZZ, which has been developed at the JGI specifically for large sequencing projects. The JAZZ procedure is a multi-step process that uses a similar approach to the Arachne [6] and Phusion [7] whole-genome shotgun assemblers. First, vector and low-quality sequence is purged and potential overlaps between sequence traces identified by looking for identical 16-mer 'words' between sequences (using a hashing algorithm). Overlapping sequences are then aligned. The tricky problem of integrating multiple pairwise alignments and incorporating mate-pair constraints was tackled by the Malign component of JAZZ. This software attempts to maximize the consistency of sequence overlaps and mate-pair constraints by the iterative building and breaking of sequence contigs (contiguous assemblies) and scaffolds and the progressive inclusion of lower-quality data. Whereas a sequence contig is a contiguous sequence without gaps, a sequence scaffold is a collection of contigs that are ordered and oriented by mate-pair information but have gaps between contigs ('captured gaps'). The final stage of the assembly process attempts to close sequence gaps within scaffolds. Gap sizes are estimated by considering all mate pairs that bridge a gap and the expected insert size of their clones. Sequences flanking the gap are then locally assembled using the program Phrap [8], which takes into account base-by-base sequence-quality data as well as mate-pair information. This assembly is locally restricted, so relatively short, low-quality or repetitive overlaps between adjacent contigs could then be permitted if supported by the mate-pair constraints. Aparicio *et al.* [2] found that they could close 28% of scaffold gaps in this manner, substantially improving the contiguity of scaffold sequences.

The current release of the draft *F. rubripes* genome (there will undoubtedly be more) is thus a pure whole-genome shotgun assembly that has not benefited from additional sources of data such as genetic maps, contigs built from bacterial artificial chromosome (BAC) fingerprints or other physical maps. Such additional sources of information can greatly add to the accurate ordering of sequence scaffolds, as illustrated by the human [9] and mouse draft genomes. If kept separate from the assembly, such datasets can provide a means of validating the assembly, and they allow the identification of systematic errors in sequence assembly. For practical reasons, no *F. rubripes* genetic map is likely to be produced, although physical mapping [10] is a possibility. A BAC fingerprint resource based on the large-insert sequencing library is being produced (Greg Elgar, personal communication), and an initial attempt to use it for assembly validation is briefly referred to by Aparicio *et al.* [2] but details of the results were not reported. This BAC resource is likely to form the basis of future efforts to finish the genome.

Without a map or fingerprint collection to provide a global means of independently validating their assembly, the consortium has been restricted to using small regions of finished *F. rubripes* genomic sequence that have already been submitted to the public databases. A total of 22 Mb of sequence from 44 non-redundant entries was aligned with

the scaffolds from the whole-genome shotgun assembly. This comparison showed that there was generally good ordering of sequence contigs within a scaffold (one 500 bp inversion was detected, but this may be a cloning artifact in the finished sequence). The scaffold coverage was also good, with only a few small gaps between scaffolds that could in most cases be attributed to highly repetitive sequences. Unfortunately, no figure was given for sequence coverage - that is, the proportion of finished bases aligned to scaffold bases. This is an important measure of the assembly, as a scaffold may contain many captured gaps which, although they contain approximate size information, do not contain sequence data. Aparicio *et al.* [2] did align a set of 209 well-annotated *F. rubripes* genes from GenBank to the assembly and found that, with the exception of two single-exon repetitive genes, every exon of every gene could be found in the assembly. This result suggests a very high degree of sequence coverage, at least in non-repetitive regions of the assembly.

Filleting the genome: baseline annotation

For the baseline annotation of the *F. rubripes* genome, the authors made use of, and further contributed to, the freely available Ensembl project [11,12]. Unlike the other metazoan genomes for which there are draft genome assemblies in the public domain, there are few expressed sequence tag (EST) or cDNA sequences from *F. rubripes* (currently 9,068 are submitted to the public databases, compared to over 4.6 million such sequences for human) to aid in the identification of genes, particularly the non-coding regions and transcripts. The lack of expressed sequence data has in part been compensated for in the prediction of coding sequence by using sequence similarity with the human genome to indicate candidate exons, in combination with *ab initio* gene prediction. But in the absence of good EST coverage, or better still a full-length cDNA collection along the lines of those being produced for the human and mouse [13,14], the predictions will inevitably be only approximations of the intron-exon structure. The predicted genes and hypothetically encoded protein sequences are available from, and are interactively illustrated on, the *Fugu* Genome Project website [15].

The orthodox way to discuss the data from a genome project is in terms of headline figures relating to physical size, and gene number, structure and density. The authors of this paper do not disappoint, and the usual caveats about unfinished, gapped sequence and imperfect computational predictions apply. Thus, the *F. rubripes* genome is confirmed to be around 365 Mb in size and is revealed to be made up of about one sixth repetitive sequence and one third gene loci; it does not show the marked variations in GC content seen in the human genome. As in other eukaryotic genomes, *F. rubripes* genes are not distributed evenly around the genome but instead occupy relatively gene-rich and gene-poor regions. In total there are thought to be in the region of 38,000 protein-coding genes, around a quarter of which are not found in the human (by the rather stringent criteria used). In spite of this, preliminary analysis suggests that the *F. rubripes* gene set will reveal the presence of almost 1,000 novel genes in the human genome. In general, *F. rubripes* genes are very compact, largely as a result of having short introns: 75% of *F. rubripes* introns are less than 435 bp and the same proportion of human introns are more than 2,609 bp long. Unexpectedly, given these figures, 571 'giant' *F. rubripes* genes were found, for which the *F. rubripes* locus was more than 1.3 times the size of the orthologous human locus, as a result of longer introns. The detailed anatomy of these giant loci remains to be explored.

As *F. rubripes* was sequenced primarily to provide a resource for comparative genomics, the extent of synteny between human and *F. rubripes* genomes is an important issue. It would be ideal to have an orthologous series of genes from each species, conserved in order and orientation across a genomic region of interest, but this situation is found in only a minority of cases, covering around one eighth of the *F. rubripes* genome. More often than not, given 450 million years of independent evolution, the genomic landscape in both species has been disrupted by rearrangements. Aparicio *et al.* [2] have estimated the extent of synteny by finding syntenic genomic regions sharing two or more orthologous genes and then counting the number of non-orthologous, intervening genes that disrupt the syntenic arrangement. They first identified a set of 9,829 orthologous gene pairs using a set of 28,706 Ensembl predicted human proteins; they identified 4,813 such regions, sharing 2 to 16 orthologous genes and containing 0 to 1,280 intervening genes. Of these regions, around 73% contain 2 or 3 orthologous gene pairs but only 28% contain 0 to 5 intervening genes, indicating the considerable amount of rearrangement that has taken place. Depending on the numbers of orthologous and intervening genes involved in these syntenic regions, the regions were on average between 121 and 123,167 kb long in the human genome and 16 to 330 kb in *F. rubripes*. Broader investigations of *F. rubripes* genome structure revealed little evidence for recent, segmental duplications, such as those that make up 5% or more of the human genome. On the other hand, some evidence was found for the postulated (see, for example, [16]) ancient genome duplication that gave rise to the teleost fish radiation, when species such as *F. rubripes* made their debut. Essentially, a number of human genome regions were found to be represented by six or eight co-orthologous segments of *F. rubripes* genome. A more thorough treatment of such questions awaits a *F. rubripes* genome assembly with chromosomal coordinates.

So, what can one reasonably hope to learn from interrogating the *F. rubripes* genome? The data indicate that, for a human gene of interest, a biologist can expect to find a *F. rubripes* homolog of some description in around

three-quarters of cases but an ortholog in only around a third of cases. Once an orthologous *F. rubripes* gene has been discovered, there will usually be some degree of synteny around it when compared to the equivalent human region. Around three-quarters of the time this synteny will extend to include another one or two neighbouring genes but, at the same time, the syntenic region is likely to contain more than five intervening genes. The orthologous region identified in *F. rubripes* will, on average, be only one eighth of the size of the human region with fewer repetitive sequences and much shorter introns. In an alignment between any two orthologous loci the known coding sequences should be visible as among the most highly conserved regions. Additional conserved regions are predicted to be novel exons or regulatory regions, depending on whether they overlap with an open reading frame.

Finding the caviar: buried treasure in the *F. rubripes* genome

The publication of any genome sequence invariably results in an embarrassment of riches, with authors struggling to cram a multitude of new insights into their manuscript, and the *F. rubripes* sequence is no exception. Since the divergence of the two lineages that eventually gave rise to humans and *F. rubripes* there has been substantial opportunity for their shared origin to be obscured. It is a small miracle that so much similarity remains between them, despite endless rounds of selection, mutation, migration and slow genetic drift. Nevertheless, Aparicio *et al.* [2] estimate that perhaps 25% of human genes are undetectable in the *F. rubripes* genome. They believe that this is mainly attributable to the rapid evolution of such genes along with the loss or gain of certain genes from either lineage. Among the human genes undetectable in *F. rubripes* were many cell-surface receptor-ligand components of the immune system, which might be expected to evolve rapidly in response to pathogens, as well as functional classes that reflect the differences in mammal and fish physiology, such as those involved in the hematopoietic system and homoeothermic metabolism. More broadly, the human and *F. rubripes* predicted proteomes show numerical agreement, with similar complements for each protein family. Within some families, such as olfactory receptors, there are tantalizing glimpses of how our distant ancestor's genome was remodeled for a tetrapod lifestyle. It is evident that there will be much more to learn from the *F. rubripes* genome about the sets of genes that are specific to tetrapods or fish as well as those that are common to all vertebrates.

Mammalian genomes are composed of 35-45% interspersed repetitive sequence. Prior to the genome sequencing project, it had already been established that the fraction of the *F. rubripes* genome accounted for by interspersed repetitive elements was substantially less than that found in mammals. Prior to the work of Aparicio *et al.* [2] only eight *F. rubripes* interspersed repeats had been described. On the basis of detection of over-represented sequences in the pre-assembly shotgun sequence, 48 new families of *F. rubripes* interspersed repeats were identified, as well as a 118-nucleotide tandem repeat that is almost certainly the centromeric satellite repeat. A total of 5.3% of the high-quality sequence was identified as repetitive, but the potential for sampling bias precludes the use of this figure as a true estimate of genome repeat content.

Armed with consensus sequences for the newly identified *F. rubripes* and previously known eukaryotic repeats, interspersed repetitive elements of the assembled *F. rubripes* genome were annotated using RepeatMasker [17]. Nearly 3% of the current *F. rubripes* assembly was found in this way to match interspersed repeat sequences. Although this is more than had previously been estimated, it is almost certainly a gross underestimate, both because many lower copy-number repeats will not have been incorporated into the library of known repeats and because many gaps in the current assembly are likely to be due to the difficulty of assembling repeat sequences. The first step of comparative nucleotide sequence analysis is often to mask the repeat sequences. Although this is entirely appropriate in many cases, the comparative analysis of repeat content between genomes is beginning to throw light on mutational and evolutionary mechanisms that shape genomes. One of the most exciting discoveries coming directly from the draft human genome sequence was that many of the interspersed repeats in our genome have stopped actively transposing [9]: they are dying out. In contrast, the mouse genome contains several clearly active transposons and shows no sign of a decrease of transposon activity [8]. The *F. rubripes* genome, famed for its low abundance of repeats, has at least as great a diversity of major repeat families as humans - in fact, nearly every major class of transposable element that has been found in eukaryotes is represented in *F. rubripes*. Not only is there greater diversity of repeat families, but a high proportion of the repeat families - at least 40 of the 56 known and new families - show evidence of recent transposition and are likely still to be active.

Although Aparicio *et al.* [2] found evidence for a perhaps surprising level of transposon activity and diversity, by mammalian standards the copy number of even the most abundant repeat (a LINE-like repeat with approximately 6,400 copies) is relatively low. The low overall copy number of transposons, in spite of evidence of continuing transposition, and a relative dearth of older transposon 'fossils', suggests that there may be a high rate of deletion of selectively neutral DNA. To investigate this possibility further, Aparicio *et al.* [2] assessed the rate of nucleotide deletion events between interspersed repeats of the same divergence level, and found an excess of deletions in *F. rubripes* relative to human. Extrapolating from this result suggests that "the frequency of larger deletions relative to point mutations is much higher in the *F. rubripes* genome than in mammalian

genomes" [2], and provides a mechanism for the acquisition and maintenance of a compacted genome. If the rate of deletions is high in the *F. rubripes* genome, its repeats could have adapted by proportionally increasing their rate of transposition, in a classic 'arms race'. This could explain the rarity of transposon fossils and the continuing activity of extant elements, but is difficult to reconcile with the maintenance of wide repeat diversity. Aparicio *et al.* [2] propose that many of the interspersed repeats in the *F. rubripes* genome have been introduced through horizontal transfer. By the authors' own admission this is a bold statement that is backed up by little evidence, but it could explain both the diversity and high proportion of active transposons relative to mammalian genomes. Could it be that spawning externally increases the exposure of the fish germline to transposon 'infection'?

Other fish to fry

In hindsight, *F. rubripes* was rather an odd choice for a model vertebrate genome to sequence. The credentials of its genome are not in doubt, and there is an ample supply of *F. rubripes* (several hundred tonnes are sold annually), but it is a large aggressive fish that is not readily bred or grown in a laboratory environment. Another pufferfish, *Tetraodon nigroviridis*, has similar genome parameters to *F. rubripes*, but it can be readily catered for in a laboratory setting. A *T. nigroviridis* whole-genome shotgun project has been conducted in parallel to the *F. rubripes* project, as a collaboration between Genoscope in France and the Whitehead Institute in the USA, and a whole-genome assembly has been produced and recently released [18]. Preliminary sequence information from *T. nigroviridis* has already been used to identify evolutionarily conserved regions of the human genome and to predict gene number [19]. The small evolutionary distance separating *F. rubripes* from *T. nigroviridis* relative to their distance from mammals suggests that comparison of mammalian sequence to both of these genomes is unlikely to reveal more information than comparison to just one. But these genomes are not completely redundant: comparison of *F. rubripes* and *T. nigroviridis* could provide significant insights into the mechanisms of vertebrate gene and genome evolution.

As well as the difficulties we have discussed in the definition of gene structures from genome sequence, the other major obstacle on the path to an informative genome sequence is the functional annotation of genes. Once we have discovered a new protein-coding gene we want to know whether the protein contains conserved domains or motifs, how its structure looks, which complexes or pathways it participates in and how it affects the anatomy, physiology and behavior of the organism. Ultimately, to squeeze the maximum meaning and medical application from sequence data, we would like to know every point at which a gene product significantly impacts upon the phenotype. Only a very modest amount of information can be inferred computationally at the moment, so the emphasis over the next decade or two must turn to novel investigations of gene function in relatively well-understood model organisms. And herein lie some difficulties for *F. rubripes*.

Although there is an abundant supply of *F. rubripes*, and *T. nigroviridis* can be grown in the laboratory, for the bench biologist there are already other more user-friendly fish in the sea: both zebrafish (*Danio rerio*) [20] and medaka (*Oryzias latipes*) [21] appear seductive alternatives. Both can be happily cultured in laboratories and boast extensive collections of ESTs, genetic maps, the availability of mutant strains and the use of morpholinos to knock-down gene expression. A preliminary, draft assembly of the zebrafish genome has already been made available by the Wellcome Trust Sanger Institute [22], based upon whole-genome shotgun data, and a higher-quality draft, based upon BAC sequence, is gradually emerging. *F. rubripes* might therefore be said to be the bioinformaticist's friend but merely a passing acquaintance of the bench scientist. But it would be wrong to say that *F. rubripes* is unimportant in functional investigations. Recently it has been shown that promoter elements may be closely conserved between orthologous genes in *F. rubripes*, mouse and human, and that the shorter promoters of *F. rubripes* may in fact act to drive appropriate tissue-specific expression in transgenic mice [23]. Thus, the compact *F. rubripes* genome may be seen as a densely packed library of minimal vertebrate regulatory elements, providing invaluable short-cuts to functional studies of such elements in larger, repeat-rich genomes like our own.

In the end, what the *F. rubripes* sequence can tell us underlines what it cannot. The large phylogenetic distance between humans and *F. rubripes* (around 450 million years) is the reason that clear patterns of conservation may be detected in functional regions. Assuming a ballpark neutral average of 0.1-0.5% sequence divergence per million years between vertebrates [24], a high signal-to-noise ratio is to be expected. Unfortunately this extensive divergence must also be reflected in the presence or absence of functional regions - with many coding and noncoding sequences gained, lost, or no longer detectable as homologous between *F. rubripes* and humans. Many comparative studies have used mouse and human sequences (which diverged around 110 million years ago), where one might expect something in the region of 55-89% similarity in non-coding regions. This background 'noise' will obscure many conserved regions, particularly where similarity is modest, as is often seen in regulatory regions. Three-way comparisons between orthologous *F. rubripes*, mouse and human sequences will at least provide a clear picture of the genomic regions conserved across all three, and by inference across vertebrates. In addition, such comparisons can be expected to uncover regions conserved between human and mouse but not in *F. rubripes*, providing a first glimpse of the coding regions and regulatory elements that define mammals and distinguish us from other vertebrates.

References

1. Hinegardner R: **Evolution of cellular DNA content in teleostean fishes.** *Am Nat* 1968, **102:**517-523.
2. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit AF, *et al*: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science,* 25 July 2002 (10.1126/science.1072104)
3. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S: **Characterization of the pufferfish (*F. rubripes*) genome as a compact model vertebrate genome.** *Nature* 1993, **366:**265-268.
4. Elgar G, Clark MS, Meek S, Smith S, Warner S, Edwards YJ, Bouchireb N, Cottage A, Yeo GS, Umrania Y, Williams G, Brenner S: **Generation and analysis of 25 Mb of genomic DNA from the pufferfish *F. rubripes* rubripes by sequence scanning.** *Genome Res* 1999, **9:**960-971.
5. Roach JC, Boysen C, Wang K, Hood L: **Pairwise end sequencing: a unified approach to genomic mapping and sequencing.** *Genomics* 1995, **26:**345-353.
6. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12:**177-189.
7. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases.** *Genome Res* 2001, **11:**1725-1729.
8. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8:**195-202.
9. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:** 860-921.
10. Dear PH, Cook PR: **Happy mapping: linkage mapping using a physical analogue of meiosis.** *Nucleic Acids Res* 1993, **21:**13-20.
11. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
12. **Ensembl** [http://www.ensembl.org/]
13. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H *et al.*: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.
14. Strausberg RL, Feingold EA, Klausner RD, Collins FS: **The mammalian gene collection.** *Science* 2001, **286:**455-457.
15. **The *Fugu* Genome Project** [http://www.fugu-sg.org]
16. Taylor JS, Van de Peer Y, Braasch I, Meyer A**: Comparative genomics provides evidence for an ancient genome duplication event in fish.** *Philos Trans R Soc Lond B Biol Sci* 2001, **356:**1661-1679.
17. Smit AFA, Green P: **RepeatMasker** [http://ftp.genome.washington.edu/RM/RepeatMasker.html]
18. ***Tetraodon nigroviridis* Database** [http://www-genome.wi.mit.edu/annotation/tetraodon/]
19. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F, *et al.***: Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence.** *Nat Genet* 2000, **25:**235-238.
20. **The Zebrafish Information Network** [http://www.zfin.org]
21. **The Medakafish Homepage** [http://biol1.bio.nagoya-u.ac.jp:8000/]
22. **Ensembl Zebrafish Genome Browser** [http://www.ensembl.org/Danio_rerio/]
23. Brenner S, Venkatesh B, Yap WH, Chou CF, Tay A, Ponniah S, Wang Y, Tan YH**: Conserved regulation of the lymphocyte-specific expression of *lck* in the *F. rubripes* and mammals.** *Proc Natl Acad Sci USA* 2002, **99:**2936-2941.
24. Tautz D: **Evolution of transcriptional regulation.** *Curr Opin Genet Dev* 2000, **10:**575-579.