Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*

Samuel Aparicio,[2,1]* Jarrod Chapman,[3] Elia Stupka,[1]* Nik Putnam,[3] Jer-ming Chia,[1] Paramvir Dehal,[3] Alan Christoffels,[1] Sam Rash,[3] Shawn Hoon,[1] Arian Smit,[4] Maarten D. Sollewijn Gelpke,[3] Jared Roach,[4] Tania Oh,[1] Isaac Y. Ho,[3] Marie Wong,[1] Chris Detter,[3] Frans Verhoef,[1] Paul Predki,[3] Alice Tay,[1] Susan Lucas,[3] Paul Richardson,[3] Sarah F. Smith,[5] Melody S. Clark,[5] Yvonne J. K. Edwards,[5] Norman Doggett,[6] Andrey Zharkikh,[7] Sean V. Tavtigian,[7] Dmitry Pruss,[7] Mary Barnstead,[8] Cheryl Evans,[8] Holly Baden,[8] Justin Powell,[9] Gustavo Glusman,[4] Lee Rowen,[4] Leroy Hood,[4] Y. H. Tan,[1] Greg Elgar,[5]* Trevor Hawkins,[3]*† Byrappa Venkatesh,[1]* Daniel Rokhsar,[3]* Sydney Brenner[1,10]*

The compact genome of *Fugu rubripes* has been sequenced to over 95% coverage, and more than 80% of the assembly is in multigene-sized scaffolds. In this 365-megabase vertebrate genome, repetitive DNA accounts for less than one-sixth of the sequence, and gene loci occupy about one-third of the genome. As with the human genome, gene loci are not evenly distributed, but are clustered into sparse and dense regions. Some "giant" genes were observed that had average coding sequence sizes but were spread over genomic lengths significantly larger than those of their human orthologs. Although three-quarters of predicted human proteins have a strong match to *Fugu*, approximately a quarter of the human proteins had highly diverged from or had no pufferfish homologs, highlighting the extent of protein evolution in the 450 million years since teleosts and mammals diverged. Conserved linkages between *Fugu* and human genes indicate the preservation of chromosomal segments from the common vertebrate ancestor, but with considerable scrambling of gene order.

Introduction

Most of the genetic information that governs how humans develop and function is encoded in the human genome sequence (*1*, *2*), but our understanding of the sequence is limited by our ability to retrieve meaning from it. Comparisons between the genomes of different animals will guide future approaches to understanding gene function and regulation. A decade ago, analysis of the compact genome of the pufferfish *Fugu rubripes* was proposed (*3*) as a cost-effective way to illuminate the human sequence through comparative analysis within the vertebrates. We report here the sequencing and initial analysis of the *Fugu* genome, the first publicly available draft vertebrate genome to be published after the human genome. By comparison with mammalian genomes the task was modest, since almost an order of magnitude less effort is needed to obtain a comparable amount of information.

Fugu rubripes, commonly known as "torafugu," is a teleost fish belonging to the Order Tetraodontiformes and Family Tetraodontidae. Its natural habitat spans the Sea of Japan, the East China Sea, and the Yellow Sea. Early work (*4*) suggested that Tetraodontiformes have low nuclear DNA content [less than 500 million base pairs (Mb) per haploid genome], which led to the conjecture that the genomes of these creatures were compact in organization. Although the *Fugu* genome is unusually small for a vertebrate, at about one-eighth the length of the human genome, it contains a comparable complement of protein-coding genes, as inferred from random genomic sampling (*3*). Subsequently, more-targeted analyses (*5*–*9*) showed that the *Fugu* genome has remarkable homologies to the human sequence. The intron-exon structure of most genes is preserved between *Fugu* and human, in some cases with conserved alternative splicing (*10*). The relative compactness of the *Fugu* genome is accounted for by the proportional reduction in the size of introns and intergenic regions, in part owing to the relative scarcity of repeated sequences like those that litter the human genome. Conservation of synteny was discovered between humans and *Fugu* (*5*, *6*), suggesting the possibility of identifying chromosomal elements from the common ancestor. Noncoding sequence comparisons detected core conserved regulatory elements in mice (*11*). This methodology has subsequently been used for identifying conserved elements in several other loci (*12*–*24*). These remarkable homologies, conserved over the 450 million years since the last common ancestor of humans and teleost fish, combined with the compact nature of the *Fugu* sequence, led to the formation of the *Fugu* Genome Consortium to sequence the pufferfish genome.

Whole-Genome Shotgun Sequencing and Assembly of the *Fugu rubripes* Genome

Sequencing and assembly. Shotgun libraries were prepared from genomic DNA that had been purified from the testis of a single animal to minimize complications due to allelic polymorphisms. These polymorphisms are estimated to occur at 0.4% of the nucleotides in our individual fish, ~fourfold as many as in human (*25*). We set out to generate ~6× genome coverage of the *Fugu* genome (Table 1). Several plasmid libraries with 2- and 5.5-kb inserts were constructed and end-sequenced by dye terminator and dye primer chemistries. The bulk of the sequence coverage resulted from 2-kb libraries (Table 1). However, the 5.5-kb library provided crucial intermediate-range linking information for assembly.

Reads passing the primary quality and vector screens ("passing reads") were assembled into scaffolds by means of JAZZ, a modular suite of tools for large shotgun assemblies that incorporates both read-overlap and read-pairing information

The 3.71 million passing reads were assembled into 12,381 scaffolds longer than 2 kb, for a total of 332.5 Mb. The scaffolding range and contiguity of the assembly are shown in table S1. A total of 745 scaffolds longer than 100 kb account for 35% of the assembled sequence (119.5 Mb); 1908 scaffolds longer than 50 kb account for 60% of the assembly (200.8 Mb); 4108 scaffolds longer than 20 kb account for 81% of the assembly (271 Mb).

[1]Institute of Molecular and Cell Biology, 30 Medical Drive, Singapore 117609. [2]University of Cambridge, Department of Oncology, Hutchison–MRC Research Centre, Cambridge CB2 2XZ, UK. [3]U.S. DoE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [4]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA. [5]MRC UK HGMP Resource Centre, Hinxton, Cambridge CB10 1SB, UK. [6]Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [7]Myriad Genetics Inc., 320 Wakara Way, Salt Lake City, UT 84108, USA. [8]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [9]Paradigm Therapeutics Ltd., Physiological Laboratory, Cambridge CB2 3EG, UK. [10]Salk Institute, 10010 North Torrey Pines Road, La Jolla, San Diego, CA 92037–1099, USA.

*To whom correspondence and requests for materials should be addressed. E-mail: saa1000@cam.ac.uk (S.A.), elia@fugu-sg.org (E.S.), gelgar@hgmp.mrc.ac.uk (G.E.), trevor.hawkins@am.amershambiosciences.com (T.H.), mcbbv@imcb.nus.edu.sg (B.V.), dsrokhsar@lbl.gov (D.R.), sbrenner@salk.edu (S.B.).

†Present address: Amersham Biosciences, 928 East Arques Avenue, Sunnyvale, CA 945085, USA.

23 August 2002

Science

Vol. 297 No. 5585
Pages 1225–1432 $9

American Association for the Advancement of Science

These scaffolds contain 45,024 contigs that total 322.5 Mb of assembled sequence. The remaining 10 Mb of scaffold sequence consists of 32,621 "captured" or "sequence-mapped" gaps (i.e., gaps flanked by contigs that are connected by spanning clones). These gaps were up to 4 kb in length, with an average size of 306 base pairs (bp). These gaps are indicated in the scaffold sequence by runs of N's whose length is the best estimate of gap size on the basis of the spanning clones; by convention, gaps projected to be shorter than 50 bp are indicated by 50 N's. Gaps account for <3% of the total scaffold length.

Five percent of the passing reads were withheld from assembly as being from high percent nucleotide identity, high–copy number repeats (*25*). About 20% of these reads have sisters placed in the assembly and therefore should contribute to filling in some captured gaps; this gap closing is ongoing. The remainder accounts for an estimated 15 Mb of unassembled, highly repetitive genomic sequence, about 10 Mb of which consist of centromeric or ribosomal RNA tandem repeats (*25*). An additional 5% of passing reads remained unassembled, accounting for an estimated 18.5 Mb of unassembled genomic sequence that is not composed of obvious high–copy number repeats but were not assembled for various reasons. Some of this sequence can be recovered by cluster assemblies and contains minor tandem repetitive genes including, for example, some small nuclear RNA arrays. Combining these unassembled sequences yields an estimated total genome size of ~365 Mb, consistent with previous estimates (*3*, *4*) and projections from sample sequencing of the freshwater pufferfish *Tetraodon nigroviridis* (*26*).

Completeness and accuracy. Of the 44 nonredundant *Fugu* contigs in GenBank >20 kbp (totalling 2.2 Mb), 40 were completely covered by the assembly in one or a few scaffolds. Three of the remaining four have 6- to 8.5-kb pieces missing from the scaffolds in regions that are clearly repetitive, on the basis of their depth of high-quality coverage. The fourth (GenBank accession number AH007668) contains the T cell receptor (TCR)–α locus and was matched in the assembly only within the coding sequence of the V region, suggesting a cloning or assembly problem. Similarly, all exons of a well-annotated set of 209 *Fugu* genes from GenBank could be located within the assembly, with the exception of two odorant receptor genes that were found in the unassembled reads. The single-exon odorant receptor genes are often found in tandem arrays separated by repetitive sequence, which may account for their absence in the current assembly.

The accuracy of the sequence was measured by comparing the assembly consensus with the finished sequence of cosmid 165K09 (GenBank accession number AJ010317), excluding sites that were determined to be polymorphic (*25*). The error rate was estimated to be about five errors per 10,000 nucleotides, equivalent to an overall effective Phrap quality score of $33 = -10 \log(5 \times 10^{-4})$.

Self-consistency of the assembly was confirmed by the relative placement and orientation of paired ends. For 2-kb insert clones with both ends assembled, more than 98% were found in the same scaffold within 3 standard deviations of their expected relative separation and in the appropriate (i.e., oppositely directed) orientation for each library.

To assess fidelity on a longer scale, we compared 2.2 Mb of finished *Fugu* sequence from GenBank with the assembly by means of BLAST analysis. These finished sequences were recovered in the assembly as long, continuous stretches of scaffold, further confirming the assembly over these segments. Only one discrepancy was noted: A finished bacterial artificial chromosome (BAC) differed from the shotgun assembly by a 500-bp inversion at one end of the BAC. Small cloning inversions have been noted on BAC and cosmid clone ends in previous studies and may explain this discrepancy. The JAZZ assembly at this location is supported by strong paired-end linking information; raw sequence data for the BAC itself were unavailable. As the BAC and shotgun sequences are from different individual fish, this is a possible polymorphism (*25*). Unlike the human genome, there is no chromosomal or genetic information on gene loci that requires integration, nor in this present assembly was a physical clone map integrated with the genome sequence. The scaffolds are therefore not mapped onto *Fugu* chromosomes.

Preliminary Annotation and Analysis of the *Fugu* Genome

We annotated the scaffolds with putative gene features by using a homology-based pipeline similar to that of the human EnsEMBL project (*25*, *27*, *28*). The results, as well as genome sequences, software, updated assemblies, and other information, are freely available at www.fugubase.org and www.jgi.doe.gov/fugu. *Fugu* materials are available from fugu.hgmp.mrc.ac.uk. The assembly described in this paper may also be accessed at the GenBank/EMBL (European Molecular Biology Laboratory) whole-genome shotgun divisions, accession number CAAB01000000. The whole-genome shotgun assembly of 332.5 Mb and a small database of unique unplaced reads constituting ~5% of the genome was searched.

Arrangement of gene loci. *How many gene loci?* After initial gene-building, filtering of repetitive peptides, and removing poorly supported (by BLAST match) predictions, a total of 33,609 predicted *Fugu* peptides remained (*25*). These constituted the nonredundant predicted set of *Fugu* proteins, including potential alternative predictions for

Table 1. Sequencing summary. *NFP and CRA refer to the same library, prepared at the Joint Genome Institute (JGI) but sequenced at JGI and Celera, respectively. All other libraries were prepared at the site of sequencing, with the exception of the BAC and cosmid libraries, which were prepared at the Human Genome Mapping Project (HGMP), Cambridge, UK. All DNA, with the exception of the BAC library (OML), was derived from the same individual. JGI, Celera, and JGI-LANL (Los Alamos National Laboratory) sequencing was done with dye-terminator methods; Myriad sequencing used dye primer methods. Pair-passing clones are clones with passing sequences from both ends of the insert. Fold sequence and clone coverages were calculated assuming a genome size of 380 Mb.

Library ID	Insert size (kb)	Sequenced at	No. of passing reads	Pair-passing clones	Trim read length	Total sequence (Mb)	Fold sequence cover	Clone cover (Mb)	Fold clone cover
MBF	2.00 ± 0.48	JGI	1,370,547	631,759	627	859	2.26×	1,264	3.33×
NFP*	1.97 ± 0.24	JGI	269,216	121,908	628	169	0.44×	244	0.64×
LPO	1.98 ± 0.33	JGI	164,048	67,240	498	82	0.21×	134	0.35×
XLP	1.94 ± 0.24	JGI	43,797	18,796	605	27	0.07×	38	0.10×
MYR	2.06 ± 0.28	Myriad	1,100,171	435,956	478	526	1.38×	872	2.39×
CRA*	1.97 ± 0.23	Celera	510,131	221,548	609	311	0.82×	443	1.15×
CRA2	5.36 ± 0.70	Celera	186,238	83,504	650	121	0.32×	459	1.18×
LPC	39 ± 4.6	JGI-LANL	40,509	16,114	471	19	0.05×	645	1.65×
OML	68 ± 31	JGI-LANL	26,599	12,130	561	15	0.04×	1,031	2.17×
Total			**3,711,256**	**1,608,955**	**574**	**2,129**	**5.60×**	**5,130**	**12.96×**

the same locus. These proteins are encoded by 31,059 predicted gene loci. This set of predicted proteins and loci is similar in size to the current number of confirmed human peptides from human EnsEMBL human build version 26 (29,181 gene predictions, 34,019 transcripts) (*29*) and the 31,780 nonredundant peptides in IPI 2.1 (*30*). The true number will be influenced by the fact that the present assembly is still fragmented and so some gene loci span two or more scaffolds: the residual 5% of the genome that remains to be assembled and contains some additional loci, and translated genome comparisons used to capture loci not detectable in extant protein and cDNA databases.

Because few *Fugu* cDNA sequences were available, most of our gene predictions in the present gene build rely on homology evidence from the universe of non-*Fugu* protein sequences. Figure 1 illustrates a scaffold showing BLAST similarities to protein databases, gene prediction, and tblastx hits with human sequence. The tblastx analysis provides translated comparisons of the two genome sequences. We found a total of 1,627,452 tblastx hits covering 75% of the *Fugu* gene loci, accounting for 78% of all tblastx features and giving a mean of 71.9 tblastx features per gene locus. A total of 527,902 tblastx features were outside of predicted gene loci (see, for example, Fig. 1). Assuming the false-positive level is similar for unknown and known loci, this approach would maximally add another 7331 gene loci. In reality this is certainly an upper bound because the fragmentation of the present assembly means that some loci will be represented across more than one scaffold. These considerations project the upper bound of gene loci in *Fugu* to be in the region of 38,000, excluding ribosomal and tRNA genes. We conclude that the core set of vertebrate gene loci is unlikely to exceed 40,000.

Identification of novel human putative gene loci. We searched all of the predicted *Fugu* proteins against the human EnsEMBL peptides, resulting in matches for 27,779 *Fugu* proteins with a blast expect score threshold of less than 10^{-3}. This accounted for 22,386 EnsEMBL human peptides. Of the 8761 *Fugu* proteins below this threshold, a further 1800 matched against the masked human genomic sequence when tblastn was used. Of these, a large number were short matches, which may represent missing exons from gene predictions; however, some represent potentially novel human gene loci. To establish the relation between the matching proteins and existing human gene loci, we used these putative proteins from *Fugu* gene predictions as input to attempt to build human genes through an EnsEMBL human pipeline. Predictions that overlapped with or were contained within existing loci of human EnsEMBL were eliminated, resulting in 1260 predictions that were apparently novel. After filtering for low-complexity peptides, the remainder were further searched against the National Center for Biotechnology Information (NCBI) nonredundant protein database. A total of 961 predictions remained that did not overlap with existing human proteins (*31*). About half have some nonhuman match in the NCBI nonredundant database; the remainder were not classifiable by homology. These predicted proteins represent novel putative gene loci in human.

Repetitive sequences and the Fugu genome. We derived consensus sequences for the most common interspersed repeats in *Fugu* (Table 2) (*25*). A RepeatMasker analysis of the *Fugu* assembly showed only 2.7% of the genome to match interspersed repeats. Although higher than previous estimates (*32*), this is still a significant underestimate because the *Fugu* repeat database is far from complete and repeat dense regions are underrepresented in the assembly. Despite this underestimate, the density of interspersed repeats is clearly far below the 35 to 45% observed in mammals.

Paradoxically, despite their low absolute abundance, transposable elements have been and probably still are very active in the *Fugu* genome. There are at least 40 different families of transposable elements in which nucleotide substitutions have accumulated to a level of <5%, reflecting a very young age and possible current activity. In contrast, the exhaustively studied but transposon-deprived human genome only contains six families of such low divergence level (*1*). We found relatively young representatives of 21 major classes of transposable elements in *Fugu*, whereas only 11 classes are known to have been active in our genome in the past ~200 million years. The neutral substitution rate in *Fugu* is not known but is likely to be higher than that in higher primates, so that >40 families in *Fugu* have been active at least as recently as the 6 families in the human genome. Despite the low overall copy number of transposon fossils, almost every class of transposable elements known in eukaryotes is represented in *Fugu* (Table 2). Thus, at least in recent times, the *Fugu* genome seems to have endured activity from more types of transposable elements than the human genome. Strong pressure against insertions and for deletions would work against transposable elements like short interspersed nuclear elements (SINEs) and many long interspersed nuclear elements (LINEs) that rely on constant creation of new copies to survive in a genome. The most common repeat, the LINE-like element Maui, has 6400 copies in the present assembly, as compared with the >1 million Alu

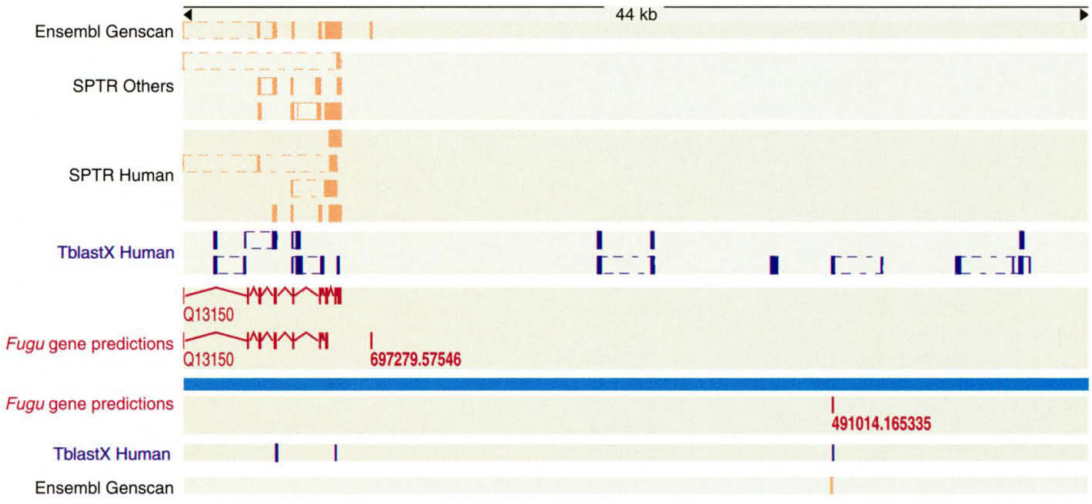


Fig. 1. The distribution of similarity features and ab initio features on *Fugu* scaffolds. Homologies of the *Fugu* sequence to entries from a variety of sequence databases are shown as solid yellow boxes for scaffold_1004. Where one database entry matches at multiple locations on *Fugu*, yellow boxes are joined by dashed lines. The source of the database is shown on the left. Tblastx translated comparisons are shown as blue boxes, again with dashed lines joining boxes from one human sequence region. Parameters used for WashU tblastx were E1 = 1×10^{-5}, E2 = 1×10^{-5}, matrix = blosum62. *Fugu* gene predictions built from database homologies are shown in brown. Exons are represented by solid vertical lines, introns by v-shaped lines between exons. All of the *Fugu* gene predictions have some overlapping homology feature, and most have matches from multiple databases. Some of the tblastx matches with human (blue boxes) have no overlapping homology features from other databases and represent potentially novel gene loci. SPTR Others, entries on genomes other than human from SWISSPROT and Translated EMBL (TrEMBL) version 39.

and >500,000 LINE1 copies in the human genome. Their relatively low copy number may be due to a high rate of deletion of junk DNA or, in some cases, higher target site specificity.

Two observations on repeats support the idea that the frequency of larger deletions relative to point mutations is much higher in the *Fugu* than in the human genome. First, the average divergence of $(CA)_n$, the most common microsatellite in both species, is 14% in human and 6.6% in *Fugu*. Unless concerted evolution of simple repeats works better in *Fugu*, this suggests that microsatellites in the *Fugu* genome are eliminated more rapidly relative to the accumulation of substitutions. Second, interspersed repeats of the same divergence level appear to have more internal deletions in *Fugu* than in human.

Thus, one aspect of the compact structure of the *Fugu* genome is the lower abundance of repeats—previously we estimated that <15% of the genome was repetitive, and this is borne out in this study. Our observations suggest that rapid deletion of nonfunctional sequences may be the predominant mechanism accounting for the repeat structure of *Fugu*.

Table 2. Repetitive DNA sequences in *Fugu* and their classification. Classification of transposable elements (*25*) that gave rise to the interspersed repeats in *Fugu*. *Gene or Genes indicates that only genes derived from this transposable class, and no interspersed repeats, are known in the human genome, indicating an ancient origin. "Many" denotes that both genes and repeat classes are present. Numbers of distinct submembers are in parentheses. A question mark indicates that the presence of a member is uncertain. In column 4, names are in bold when this report is the first to find that specific class of transposable elements in vertebrates. In the last column, we give the estimated copy number. These are still underestimates of the true number of family members because counting of elements in the unassembled, mostly repetitive 45Mb is difficult. Unclassified repeats, of which there are about 6000, constituting 0.25% of the genome, are not included in this table. For a detailed discussion of *Fugu* interspersed repeats, see supplemental text. LTR, long-term repeat.

Repeat classification	Distribution	Human members	*Fugu* members	Copy number
SINEs	Vertebrates, insects	Alu, MIR	SINE-FR (4)	5,000
Non_LTR retrotransposons				14,000
Penelope	Insects, fish	–	Bridge (2)	2,000
CRE, SLACS	*Trypanosoma*	–	–	
NeSL-1	Nematodes	–	–	
R4, Dong	Nematodes, insects	–	Rex6/DongFR (2)	1,000
R2	Arthropoda	–	–	
LINE1 group				
L1, Tx1, Ta11	Vertebrates, plants	LINE1	Tx1_FR (2)	500
DRE	*Dictyostelium*	–		
Zepp	Algae	–	–	
RTE/Bov-B group	Nematodes, vertebrates	–	Rex3/Expander (2)	2,300
CR1-group				
Tad1, CgT1	Fungi	–	–	
R1, LOA	Insects	–	–	
Jockey	Nematodes, insects	–	–	
I, ingi	Insects	–	–	
Rex-Babar	Fish	–	Rex1 (4)	2,000
L2, T1	Metazoa	LINE2	Maui (1)	6,500
CR1	Vertebrates	LINE3	–	
LTR retrotransposons				3,000
BEL/PAO	Metazoa	–	Catch (1)	35
Ty1/Copia	Eukaryotes	–	Kopi (2)	50
DIRS1	Eukaryotes	Gene*	FrDIRS1 (1)	10?
Ty3/Gypsy	Eukaryotes	Genes*	Sushi/Ronin (5)	2,500
Retroviral	Vertebrates	Many*	FERV-R (2)	100
DNA transposons				8,000
P element	Insects	Gene*	–	
MuDR/IS905	Plants	–	?	
En-Spm	Plants	–		
IS5/Harbinger	Plants, nematodes	–	Senkusha (2)	750
PiggyBack	Insects, mammals	Looper (1)	Pigibaku (1)	220
"D,D35E" transposons				
Pogo-group				
Fot1/Pot3	Fungi	–	1 (gene?)	1
Pogo	Insects, mammals	Tigger (8)	Tiggu (2)	500
Tc2	Nematodes, mammals	Tc2_Hs (2)	Tc2_FR (5)	1,800
Tc4, Tc5	Nematodes	–	?	
Tc1-Mariner-IS630				
Tc1/Impala	Metazoa	–	Tc1_FR (5)	1,400
Mariner	Metazoa, plants	Mariner (3)	–	
Hobo-Activator-Tag1				
Charlie	Mammals	Charlie(10)	Chaplin (8)	1,500
Tip100/Zaphod	Plants, mammals	Zaphod (3)	Trillian (1)	150
Classic hAT				
Tol2/Hopper	Metazoa	–	Tol2_FR (1)	1
Hobo	Insects	–		
Activator	Plants	Genes*	Furousha (2)	150
Tag1	Plants	–	–	
Restless	Fungi	–	–	

Introns in Fugu. The *Fugu* genome is compact partly because introns are shorter compared with the human genome (Fig. 2). The modal value of intron size is 79 bp, with 75% of introns <425 bp in length, whereas in human the modal value is 87 bp but with 75% of introns <2609 bp. The present annotation contains ~500 large introns that are >10 kb in size, as compared with human, where more than 12,000 introns exceed 10 kb. The total numbers of introns are roughly the same (161,536 introns in *Fugu* compared with 152,490 introns in human). Both gain and loss of introns in the *Fugu* lineage (*33*) have been observed. We examined 9874 orthologous gene pairs (*34*) and observed 456 instances of concordance between intronless *Fugu* and human genes; however, 327 human orthologs of intronless *Fugu* genes contained multiple introns and 317 *Fugu* orthologs of human intronless genes contained multiple introns.

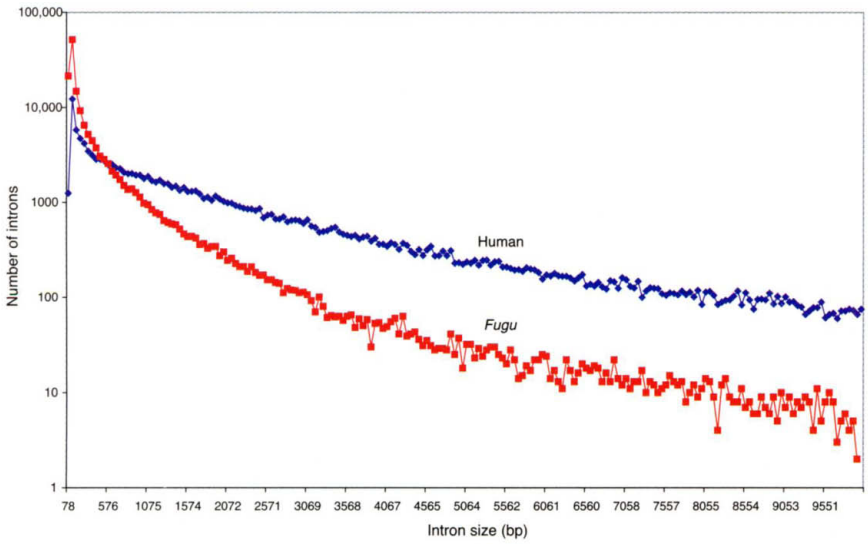


Fig. 2. Comparative frequency distribution of intron sizes in *Fugu* and human.

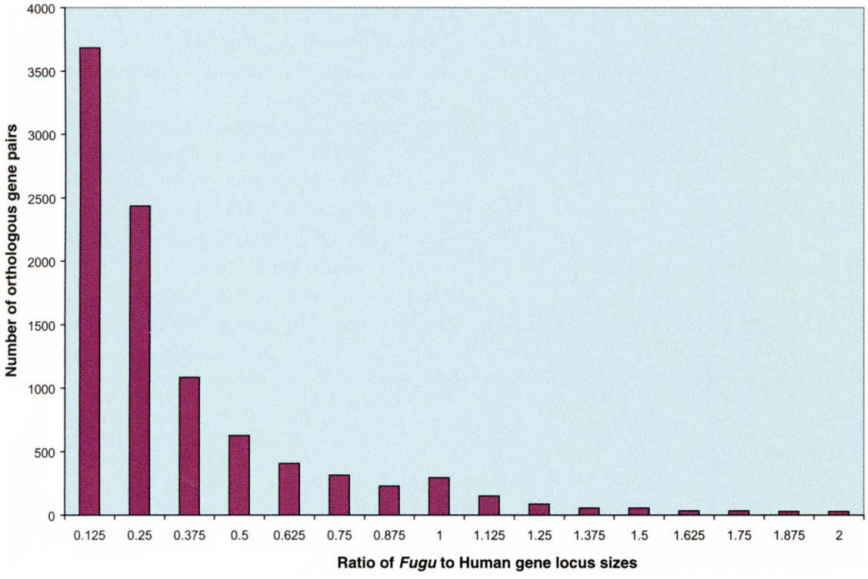


Fig. 3. Distribution of ratios for gene locus sizes of putative *Fugu*-human orthologous pairs. Putative *Fugu*-human orthologous gene pairings were determined as described in supplemental methods relating to conservation of synteny.

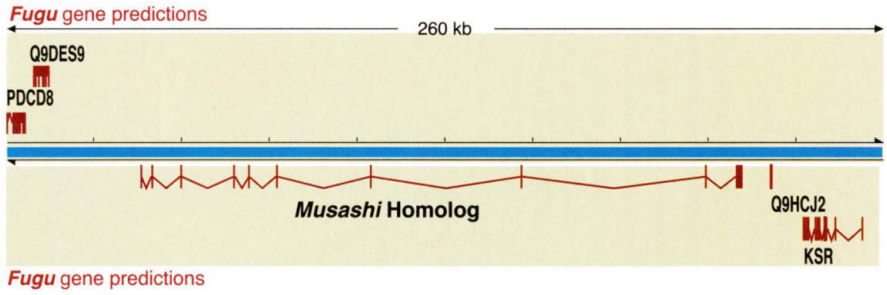


Fig. 4. Organization of a giant gene locus in the compact genome of *Fugu*. The schematic shows a region of scaffold_1 with *Fugu* gene predictions shown in brown. Exons are represented by vertical brown lines. Introns are show as v-shaped brown lines between exons.

Scaling of gene loci in a compact genome. Although the majority of *Fugu* gene loci are scaled in proportion to the compact genome size, we asked whether this was true for all *Fugu* gene loci (Fig. 3). Although the ratio of coding sequence lengths for putatively orthologous *Fugu* human gene pairs was almost unitary (*35*), we noted 571 gene loci in *Fugu* that were 1.3× or greater in size than their human counterparts. This analysis revealed a feature of *Fugu* gene loci unprecedented in previous analyses—the presence of "giant" genes with average coding sequence lengths (1 to 2 kb), but spread over genomic distances greater than those for homologs in other organisms. On Scaffold_1 (Fig. 4), we noticed a large region that was relatively bare of homology features, which on closer inspection had a predicted gene corresponding to *Fugu* transcript SINFRUT00000054697. This transcript consists of 14 putative exons predicting an RNA binding protein with similarity to proteins of the *Drosophila musashi* family (*36*–*43*). This forms part of a multigene family in humans (ENSF00000000182, heterogeneous ribonucleoprotein) with 32 members; at least 16 members can be found in *Fugu*. The most similar gene locus in human is *msi-1*, a 28-kb gene on chromosome 12. Curiously, the gene loci in human and fly are less than 50 kbp in size, whereas in *Fugu* this one locus on Scaffold_1 spans ~176 kb. The average gene density in *Fugu* is one gene locus per 10.9 kb of genomic sequence. The distribution of 1176 bp of putative coding information in 176 kb is unprecedented in *Fugu*, and the genomic organization of this gene stands in sharp contrast with that of the compact gene loci surrounding it. A paralog of *Fugu msi-1* is located on Scaffold_1927; however, the gene locus occupies only ~16 kb of genomic sequence. Exhaustive searches did not produce similarity features suggestive of undetected gene loci, and therefore we have no evidence that the introns of this particular locus might contain other embedded genes. The *Fugu msi-1* homolog on Scaffold_1 is detectable by reverse transcription–polymerase chain reaction in *Fugu* RNA.

Gene loci in the present assembly occupied ~108 Mb of the euchromatic 320 Mb, or about one-third of the genome, emphasizing the density with which they are packed in *Fugu*. However, variations in gene density

occur across the *Fugu* genome, with clustering into gene-dense and gene-bare regions, as is the case with human (Table 3) (*1*, *2*). Despite this variation in gene density, there was much lower variation in overall *Fugu* G+C content than in human (Fig. 5), regardless of gene density (Table 3). Physical methods have suggested that G+C compositional heterogeneity is less marked in poikilothermic animals (*44*), and this is confirmed by our large-scale genome sequence analysis.

Structuring of the *Fugu* Genome over Evolutionary Time

In the past, conservation of large-scale structure between genomes has been assessed by considering conservation of synteny and of gene order (*45*, *46*). Conservation of synteny means that orthologous gene loci are linked in two species, regardless of gene order or the presence of intervening genes. When evolutionary distance is large, scrambling of gene order and the presence of nonsyntenic intervening genes become frequent and so it becomes necessary to account for these features when examining conserved segments (*45*–*47*). We have examined the contiguity from *Fugu* with reference to human, looking at *Fugu* genes linked on scaffolds within the assembly whose orthologs are linked on human chromosomes.

To make *Fugu*-human comparisons, we first assigned putative orthology and then examined possible clustering of genes with respect to differing numbers of intervening nonsyntenic genes (*25*). Figure 6 shows the locations of *Fugu* gene clusters relative to human chromosomes 1 and 12 (full plot in fig. S1), allowing for varying numbers of intervening genes (table S2). Although many short conserved segments were found, considerable scrambling of gene order was observed over large distances (for example, chromosome 12 in Fig. 6). Even within short conserved segments inversions of gene order were relatively frequent (*35*).

The density of conserved segments varied between chromosomes (fig. S1). We examined the nature of this relation in terms of chromosomal gene density and chromosomal length (Fig. 7). There was no apparent correlation with gene density of human chromosomes; however, the number of conserved segments varies with human chromosomal length (Fig. 7). This suggests that the retention of conserved segments is driven largely by the probability of rearrange-

Table 3. Normalized distribution of gene densities across 100 kb windows in *Fugu* and human. The number of gene loci contained in nonoverlapping windows of 100 kb contiguous sequence was determined for *Fugu* and human, together with the mean GC content of segments in each class. The shape of the distributions was similar for 50-kb windows (not shown).

No. of genes in window	*Fugu*		Human	
	% of total windows	Mean GC (%)	% of total windows	Mean GC (%)
0	2.7	44.1	59.9	34.0
1	8.2	43.8	24.0	39.0
2	7.5	43.7	9.8	41.7
3	8.5	44.3	4.0	43.7
4	7.2	44.0	1.3	44.4
5	7.2	43.7	0.8	48.9
6	8.1	44.0	0.2	51.6
7	7.3	44.2	0.1	46.0
8	7.1	44.3	0	0
9	5.6	44.9	0.1	53.5
10	7.0	44.9		
11	4.6	44.8		
12	3.7	45.0		
13	3.3	44.9		
14	2.8	44.4		
15	2.4	44.8		
16	2.4	44.6		
17	1.4	45.1		
18	0.7	46.0		
19	1.2	46		
20	0.5	46.2		
21	0.2	44.0		
22	0.1	47.5		

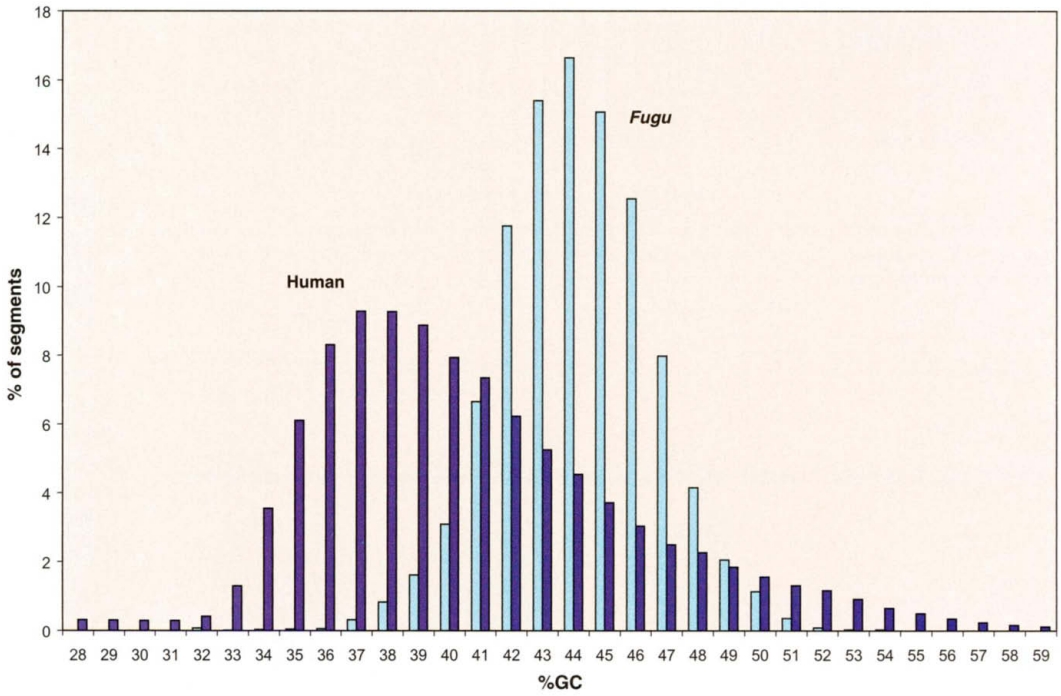


Fig. 5. Distribution of GC content in the *Fugu* and human genomes. Sliding windows of 50 kb were used; similar conclusions were derived with windows of 25 and 100 kb (not shown).

ment, which is in turn a function of chromosome length. In addition, the frequency distribution of conserved segments in relation to the number of genes per segment follows the exponential distribution noted in human-mouse comparisons (*1*) (fig. S2), for segments with identical and scrambled gene order (fig. S3). Thus, despite the separation of *Fugu* and human over 450 million years of evolution, the dominant mode of segmental conservation fits a random breakage model (*45*, *46*).

We next examined the nature of conserved segments between *Fugu* and human by looking at the frequency distribution of segments for discrete numbers of unrelated intervening genes. We noted (fig. S4) that this distribution peaks (1380 segments) at 1 to 5 unrelated intervening genes, with a much smaller peak for sparse segments of 161 to 320 intervening genes. A total of 221 of 933 segments (24%) with two or more syntenic genes show completely identical gene order. Considering the length of these segments in the *Fugu* genome, coverage rises to an exponential asymptote of 13.4% (fig. S5); however, most of the coverage is in short segments with low numbers of intervening genes—a total of 3.8% (12.6 Mb) of the genome is in segments with 0 intervening genes (perfect conservation), 5.0% (16.7 Mb) is in segments with 1 intervening gene, 7% (23.6 Mb) is in segments with up to 5 intervening genes, and 9.3% (30.7 Mb) is in segments with up to 15 intervening genes.

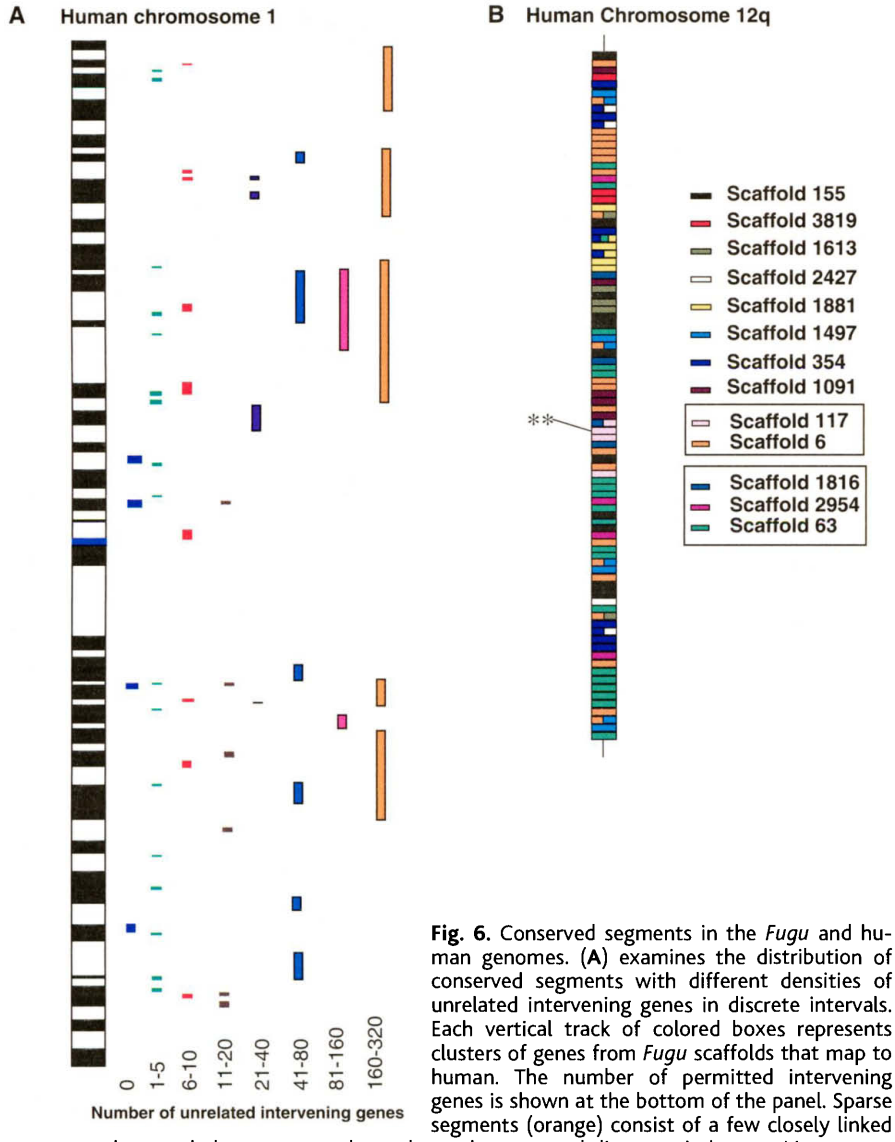


Fig. 6. Conserved segments in the *Fugu* and human genomes. (**A**) examines the distribution of conserved segments with different densities of unrelated intervening genes in discrete intervals. Each vertical track of colored boxes represents clusters of genes from *Fugu* scaffolds that map to human. The number of permitted intervening genes is shown at the bottom of the panel. Sparse segments (orange) consist of a few closely linked *Fugu* genes whose orthologs are spread over large chromosomal distances in human. Very sparse segments (>320 intervening genes) are not shown on this panel. (**B**) A detailed view of human chromosome 12q to illustrate shuffling gene order. Colored boxes represent individual genes from *Fugu* scaffolds whose orthologs on human chromosome 12q were determined through alignment by hand. The order of the orthologs along the human chromsome is shown, with the corresponding *Fugu* scaffold of origin in the key on the right. The scaffolds shown grouped together in boxes in the key are known to be linked in *Fugu*. ** indicates the position of the *Hox-c* complex on this chromosome, represented by scaffolds 117, 1327, and 1458 (the latter two are not shown in the key). Where a human gene has equally matching (co-orthologous) *Fugu* genes, this is shown as a double- or triple-colored box.

Duplications and Fugu genome structure. It is widely believed that large regional or genome duplications have contributed to the structure of vertebrate genomes, and it is now well established that most teleosts contain an excess of duplicate genes in comparison with tetrapods. The mechanisms by which these have arisen are controversial but could involve tandem duplications, segmental duplications, and whole genome duplications.

Recent duplications would be expected to show a high degree of sequence conservation in coding and noncoding portions, as opposed to ancient duplications, which may show conservation in coding regions only (*1*). We used the same parameters in comparing *Fugu* to itself as were used for the human genome. With windows of 1 kb and 500 bp, we found that ~0.15 and 1.3%, respectively, of the *Fugu* genome contained duplicated segments as compared with the human genome, whereas ~5% of the genome was found duplicated in segments of >1 kb (*1*, *2*, *48*). This suggests that large, recent tandem duplications are not a contemporaneous feature of the *Fugu* genome, or if such events do occur with any frequency, they have only a short persistence and are unlikely to account for large-scale changes in structure.

The most robust evidence (*49*–*51*) for ancient duplications comes from the existence of ancient paralogous segments. Orthologous genes are related by direct descent from the last common ancestor of two species. Gene duplication complicates this by the generation of paralogs to a given locus. Where paralogs have arisen after the speciation event that separated two orthologs, they are referred to as co-orthologs or in-paralogs (*52*). Although global resolution and dating require chromosomal-scale assemblies, we have already identified some fish-specific duplications. Previously (*53*), three *Fugu Hox* complexes orthologous to tetrapod *Hox-a*, *-b*, and *-c* complexes were identified, together with a fourth complex that was subsequently observed to be orthologous to a duplicated *Hox-a* complex in zebrafish (*54*). If this arrangement was the result of an ancient, fish-specific duplication, it predicts the potential existence of additional complexes or remnants of these in the *Fugu* genome sequence. We found at least two additional complexes in *Fugu*: an ortholog of the tetrapod *Hox-d* complex (*Hox-da*) and an ortholog of the zebrafish duplicated *Hox-b* complex (*Hox-bb*) (fig. S6) (*25*).

Is there additional evidence for ancient duplications in the *Fugu* genome? In examining other regions, for example, human 12q (Fig. 6), we found co-orthologs of some genes on different scaffolds. At least 12 of 114 genes examined

in this region are represented in six co-orthologous segments of *Fugu*. With respect to human chromosome 20q12 (*35*, *55*), 19 *Fugu* scaffolds contained 64 orthologs, of which 30 appear to be co-orthologous (duplicated in *Fugu*). These are represented by at least eight co-orthologous segments, implying these were part of segmental or large-scale duplications. Similar ancient duplications were found mapping to human chromosome 16 in the region of the *polycystin-1/tsc2* locus (*35*).

Comparison of *Fugu* and Human Predicted Proteomes

We next examined the similarities and differences between the human and *Fugu* proteomes at extremes of the vertebrate radiation (Fig. 8). We selected, by inspection, a conservative threshold score of between 10^{-2} to 10^{-3} that defined distant alignments for the purposes of global comparison (*56*–*58*).

From this inspection we noticed two features: First, the majority (*59*) of peptides have some degree of match in *Fugu*; second, ~25% of predicted human proteins (8109) do not appear to have homologs in the *Fugu* genome. In a reciprocal comparison, we noted that ~6000 *Fugu* predicted proteins lacked significant homology in human. We searched the 8109 human proteins against a core set of invertebrate proteins from *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* and noted a further 429 human proteins with some degree of match in these protein sets, suggesting that these genes had been lost from *Fugu* (*60*).

We asked whether any pattern was present in the set of 8109 nonmatching human proteins. We noted that 1237 proteins were classifiable through Interpro domain classification (table S3). Of the remaining 5268 proteins, some have identifiable secondary-structure motifs such as coiled coils or transmembrane motifs; however, most of these proteins are hypothetical or are of unknown function. Among the nonmatching proteins with Interpro identities, there were many cell surface receptor–ligand system proteins of the immune system, hematopoietic system, and energy/metabolism of homeotherms.

Immune cytokines, in general, were either not detectable in *Fugu* or showed distant similarities to human proteins (table S4), even when sensitive Smith-Waterman whole-genome searches were used. These components appear to have undergone either rapid evolution of sequences or to have arisen de novo in tetrapods. Detecting short, rapidly evolving peptide ligands is always difficult, and we therefore examined in more detail potential divergence of relevant cell surface receptors (table S4). The greatest degree of similarity in cell surface receptors was for the interleukin-1 (IL-1), IL-8, and IL-6 systems, where overall identities of ~45% exist. However, receptor components could not be confidently detected for many immune cytokines. Fish have cellular immune components, and there is evidence for antiviral defences, although attempts to identify immune cytokines in fish have so far resulted only in the identification of active IL-1–like molecules of the *Toll* family and IL-8–like receptor molecules (*61*–*65*). Functional searches for other interferons and other interleukins have so far been unsuccessful. The degree of divergence in T cell–related cytokines suggests that T cell–mediated cellular immune functions have been a rapidly evolving system. This suggestion is reinforced by the apparent absence of CD4-like molecules and only a faint signature for one of the CD8 glycoprotein chains on Scaffold_119.

The general organization of TCR and immunoglobulin (Ig) loci in *Fugu* (fig. S7) (*66*) reflects organization previously described in other fishes (*67*). Unexpectedly, the *Fugu* Ig heavy-chain locus has a separate array of D- and J-gene segments followed by a single constant exon 5′ to the canonical array of D- and J-gene segments associated with δ, μ, and transmembrane exons. This partial locus duplication superficially resembles that of mammalian TCR-β. However, rather than a duplication of functionality, the single constant exon appears to be the secretory form of IgD. This observation reveals that an osteichthyan strategy for differentially producing secretory and membrane immunoglobulins relies on germ-line rearrangement, probably as an adjunct to the production of secretory forms through alternative splicing. This dual strategy contrasts sharply with the mammalian strategy of differential processing of transcripts.

Divergence was also noted in many nonreceptor systems (table S3), including compo-

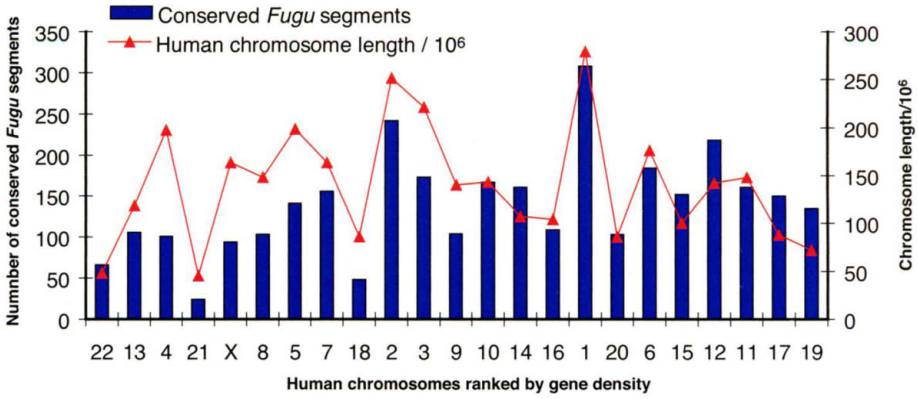


Fig. 7. Distribution of conserved segments of *Fugu* on human chromosomes ranked by gene density. The figure shows the relation between the number of conserved segments of *Fugu* on human chromosomes, the length of human chromosomes, and their gene density. Chromosome 22 is the most gene poor, chromosome 19 the most gene dense. There is no apparent relation between human chromosomal gene density and the number of segments. The distribution of conserved segments varies with human chromosomal length.

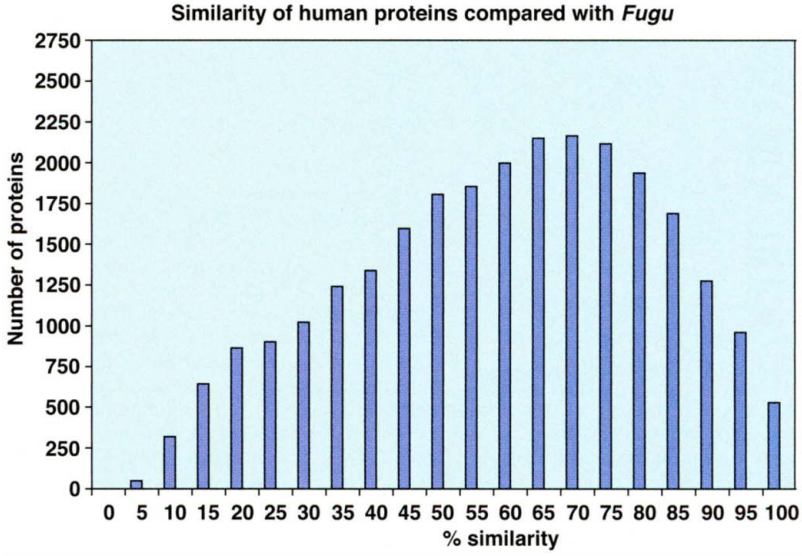


Fig. 8. Distribution of protein similarities between *Fugu* and human proteomes. Global similarities were calculated as the sum of similarities in all nonoverlapping HSPs using a BLOSUM62 matrix, over the query (human) sequence length.

nents of the cell cycle, apoptosis-related proteins, and gametogenesis proteins. The spectrum of these differences reflects the differentially evolved physiologies of mammals and fish.

The comparative classification of predicted proteins by domains (Fig. 9, table S6) principally indicates numerical concordance between *Fugu* and human. Notable exceptions include potassium channel subunits and kinases, which appear in excess in *Fugu*, whereas C2H2 zinc finger proteins are more numerous in human. We examined G-protein–coupled receptors (table S5) in detail to explore the evolution of subfamilies. Some variability in subfamily sizes was detected (for example, adrenoreceptors are more numerous in *Fugu* than in human); however, most subfamilies were of similar size. Olfactory receptors (fig. S8) show a clear expansion of different subfamilies in *Fugu*. Likewise, the absence of type I subclass A olfactory receptors suggests that these may be the result of a tetrapod-specific expansion. Even where simple numerical concordance in family sizes was noted, it does not necessarily reflect an underlying similarity in the evolution of the proteins. For example, the CxC2 cytokines (table S5) has 21 members in human and 23 in *Fugu*. However, when compared directly (*35*), only nine of the *Fugu* family members could be assigned orthologs, and most had global sequence similarity of less than 35%.

Conclusions

The feasibility of assembling a repeat-dense mammalian genome with whole-genome shotgun methodology is currently a matter of debate (*68*, *69*). However, using this approach, we have been able to sequence and assemble *Fugu* to a level suitable for preliminary long-range genome comparisons. Was it efficient to obtain the sequence of an entire vertebrate genome in this way? We have estimated the expenditure of the consortium to have been around $12 million (U.S.), including the salaries of the people involved and the expenses of obtaining the single *Fugu* specimen used to derive the DNA for this work. This is probably two orders of magnitude less than the cost of obtaining the human genome sequence. It suggests that even in the absence of mapping information, many vertebrate genomes could now be efficiently sequenced and assembled to levels sufficient for in-depth analysis.

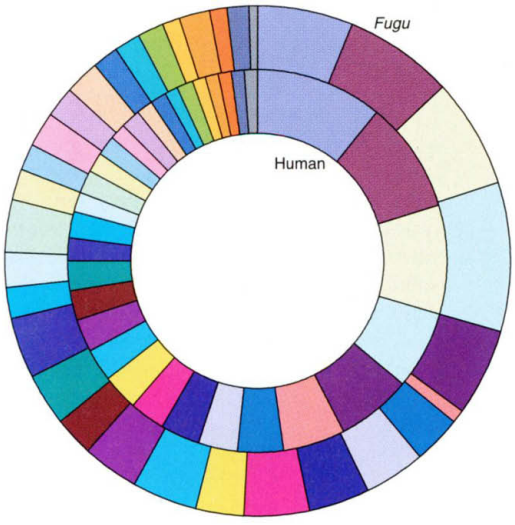
The gene-containing fraction of this vertebrate genome is a mere 108 Mb. Despite the overall eightfold size difference between *Fugu* and human, "gene deserts" are also present in *Fugu*, although these regions are scaled in proportion to the genome size. "Giant" gene loci, with a low ratio of coding to noncoding DNA, occur in *Fugu*, in sharp contrast to the compactness of genes around them. In flies (*70*), large introns occur preferentially in regions of low recombination, and this has led to the suggestion that large introns are selected against. If intron sizes were simply scaled as genome size, we would not expect to find extreme outlier genes without some evidence of their existence in other species. Further study of these extreme examples may illuminate the balance of gain and loss of DNA in genomes during evolution. The presence of large intron structures in *Fugu* implies that, despite general evolution of *Fugu* toward compactness, *Fugu* splicing machinery is still able to recognize and process large introns correctly.

The other key feature of the compactness of *Fugu* is the low abundance of repetitive DNA. Paradoxically, there is evidence of recent activity of transposon elements and far more diversity of repeat families in *Fugu* than in human. It is unclear why this should be the case and how this relates to the low abundance of repeats. However, the most parsimonious hypothesis is that sequences are deleted more frequently than inserted.

The number of gene loci in *Fugu* is similar to that observable in human. Our predictions are of course limited by the nature of automated gene-building pipelines, and we do not yet incorporate gene structures built from *Fugu* expressed sequence tags or from translation comparisons of *Fugu* and human genomic sequences. Nevertheless, we find no evidence for a core vertebrate gene locus set of more than 40,000 members. Simply comparing the present *Fugu* gene builds and prediction features with those of human also enabled us to discover almost 1000 human putative genes that have so far not been described in public annotation databases. This emphasizes that comparisons of vertebrate genomes will continue to inform the annotation of gene loci in the human genome.

There are certainly more similarities than differences between the *Fugu* and human proteomes; however, we have shown that a large fraction, perhaps as much as 25% of the human proteome, is not easily identifiable in *Fugu*. This set of proteins could represent evolution of proteins between two vertebrates so that they are no longer mutually recognizable at the sequence level, loss of genes common to other vertebrates in *Fugu*, or gain of genes specific to tetrapod or mammalian orders, or erroneous human gene predictions. We believe that rapid evolution of proteins may account for most of the observable differences. Regardless of the mechanism, the large set of human and *Fugu* proteins that are not mutually recognizable helps to define a set of previously unannotated human genes that may be at the core of differences between tetrapods and teleosts. Comparisons with other completed genomes will refine these sets to reveal the elements unique to each taxon.

Finally, in examining conservation of synteny, we have shown that a substantial fraction, about one-eighth of the *Fugu* genome, shows conserved linkages of two or more genes with the human genome. Over chromosomal scales, it is clear that the order of genes has been extensively shuffled, with many nonsyntenic intervening genes breaking up the segmental relation of *Fugu* and human chromosomes. Nevertheless, more than 900 segments of two or more genes show conserved linkage. In tackling the challenges of deciphering complex genomes, the enumeration of conserved segments between



No. of gene loci		Interpro domain identity	
Human	*Fugu*		
813	427	IPR000822	Zinc finger C2H2
716	457	IPR000276	Rhodpsin GPCR
699	488	IPR003006	Immunoglobulin/MHC
508	640	IPR000719	Protein kinase
469	383	IPR002965	Proline extensin
419	33	IPR000725	Olfactory receptor
275	205	IPR000504	RNA binding (RNP1)
262	268	IPR001680	G-protein WD40 repeats
245	255	IPR002110	Ankyrin repeat
232	286	IPR001849	Pleckstrin domain
229	216	IPR002048	EF-hand
216	281	IPR001356	Homeobox domain
213	254	IPR001841	RING finger
187	159	IPR001611	Leucine rich motif
181	241	IPR001452	SH3 domain
158	251	IPR000477	Reverse transcriptase
157	135	IPR001092	HLH domain
141	167	IPR002126	Cadherin domain
139	219	IPR001478	PDZ domain
126	125	IPR000636	Cation channels
123	123	IPR001254	Serine proteases
114	148	IPR000008	C2 domain
113	139	IPR001440	TPR repeat
106	143	IPR000210	BTB/POZ domain
105	116	IPR000980	SH2 domain
104	117	IPR000087	Collagen triple helix
94	111	IPR000387	Tyrosine phosphatases
90	79	IPR001410	DEAD box helicases
87	143	IPR001881	EGF domain (Ca binding)
85	69	IPR001304	C-type lectin
80	97	IPR002049	EGF domain (laminin)
79	34	IPR000048	IQ calmodulin

Fig. 9. Protein domains of *Fugu* and human. The schematic shows the number of gene loci with corresponding Interpro domains in *Fugu* and human for the 32 most populous families.

Fugu and human may form an important starting point for detecting conserved regulatory elements. We have also identified several sparse conserved segments for most human chromosomes. These segments are tightly linked in *Fugu* but dispersed over whole chromosomes in human. Tracing the fate of such segments in other species may allow us to reconstruct some of the evolutionary history of vertebrate chromosomes.

References and Notes

1. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. S. Brenner *et al.*, *Nature* **366**, 265 (1993).
4. R. Hinegardner, *Am. Nat.* **102**, 517 (1968).
5. M. K. Trower *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1366 (1996).
6. K. Gellner, S. Brenner, *Genome Res.* **9**, 251 (1999).
7. S. Baxendale *et al.*, *Nature Genet.* **10**, 67 (1995).
8. B. Venkatesh, S. Brenner, *Gene* **211**, 169 (1998).
9. ———, *Gene* **187**, 211 (1997).
10. O. Coutelle *et al.*, *Gene* **208**, 7 (1998).
11. S. Aparicio *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1684 (1995).
12. B. Venkatesh *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12462 (1997).
13. J. Flint *et al.*, *Hum. Mol. Genet.* **10**, 371 (2001).
14. P. L. Pfeffer *et al.*, *Development* **129**, 307 (2002).
15. W. P. Yu *et al.*, *Oncogene* **20**, 5554 (2001).
16. J. M. Wentworth *et al.*, *Gene* **236**, 315 (1999).
17. D. H. Rowitch *et al.*, *Development* **125**, 2735 (1998).
18. H. Marshall *et al.*, *Nature* **370**, 567 (1994).
19. H. Popperl *et al.*, *Cell* **81**, 1031 (1995).
20. S. Nonchev *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9339 (1996).
21. B. Kammandel *et al.*, *Dev. Biol.* **205**, 79 (1999).
22. L. M. Barton *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6747 (2001).
23. S. Bagheri-Fam *et al.*, *Genomics* **78**, 73 (2001).
24. S. Brenner *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2936 (2002).
25. Supplemental methods and data are available on *Science* Online.
26. C. Fischer *et al.*, *Cytogenet. Cell Genet.* **88**, 50 (2000).
27. T. Hubbard *et al.*, *Nucleic Acids Res.* **30**, 38 (2002).
28. E. Birney, R. Durbin, *Genome Res.* **10**, 547 (2000).
29. EnsEMBL human databases can be accessed at www.ensembl.org.
30. IPI maintains a nonredundant and updated set of human proteins, which can be accessed at www.ebi.ac.uk/IPI.
31. The sequences of these predicted human proteins are available from the project Web sites
32. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
33. B. Venkatesh, Y. Ning, S. Brenner, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10267 (1999).
34. These pairings were from the comparative linkage analysis, described in the supplemental material, estimating conserved synteny, which can be accessed at *Science* Online.
35. S. Aparicio *et al.*, data not shown.
36. M. Okabe *et al.*, *Nature* **411**, 94 (2001).
37. A. Yoda, H. Sawa, H. Okano, *Genes Cells* **5**, 885 (2000).
38. W. Wang *et al.*, *Mol. Biol. Evol.* **17**, 1294 (2000).
39. Y. Hirota *et al.*, *Mech. Dev.* **87**, 93 (1999).
40. S. Sakakibara, H. Okano, *J. Neurosci.* **17**, 8300 (1997).
41. M. Okabe *et al.*, *Dev. Neurosci.* **19**, 9 (1997).
42. S. Sakakibara *et al.*, *Dev. Biol.* **176**, 230 (1996).
43. M. Nakamura, H. Okano, J. A. Blendy, C. Montell, *Neuron* **13**, 67 (1994).
44. G. Bernardi, *Gene* **241**, 3 (2000).
45. J. H. Nadeau, D. Sankoff, *Mamm. Genome* **9**, 491 (1998).
46. J. H. Nadeau, B. A. Taylor, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 814 (1984).
47. S. Aparicio, *Nature Genet.* **18**, 301 (1998).
48. J. A. Bailey *et al.*, *Am. J. Hum. Genet.* **70**, 83 (2002).
49. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
50. J. H. Postlethwait *et al.*, *Nature Genet.* **18**, 345 (1998).
51. L. G. Lundin, *Genomics* **16**, 1 (1993).
52. M. Remm *et al.*, *J. Mol. Biol.* **314**, 1041 (2001).
53. S. Aparicio *et al.*, *Nature Genet.* **16**, 79 (1997).
54. A. Amores *et al.*, *Science* **282**, 1711 (1998).
55. S. F. Smith *et al.*, *Genome Res.* **12**, 776 (2002).
56. C. Chothia, A. M. Lesk, *EMBO J.* **5**, 823 (1986).
57. B. Rost, *Protein Eng.* **12**, 85 (1999).
58. We examined the best local identity BLASTP matches from comparing the human proteome with *Fugu*. An expect score threshold of 10^{-2} to 10^{-3} rejects most alignments of <25 to 30% distant protein alignments. It has been previously shown by Chothia, Lesk, Rost, and others that 90% of alignments at or below this "twilight zone" of similarity are unlikely to represent true structural homologies.
59. We found 26,390 of 34,019 matches comparing human peptides with *Fugu* peptides, and a further 687 human peptides that matched *Fugu* assembled sequence or sequence fragments.
60. The accession numbers of these proteins can be accessed at the *Fugu* project Web sites.
61. E. Y. Lee, H. H. Park, Y. T. Kim, T. J. Choi, *Gene* **274**, 237 (2001).
62. A. M. Najakshin, L. V. Mechetina, B. Y. Alabyev, A. V. Taranin, *Eur. J. Immunol.* **29**, 375 (1999).
63. D. B. Lehane, N. McKie, R. G. Russell, I. W. Henderson, *Gen. Comp. Endocrinol.* **114**, 80 (1999).
64. N. Miller *et al.*, *Immunol. Rev.* **166**, 187 (1998).
65. J. L. Grondel, E. G. Harmsen, *Immunology* **52**, 477 (1984).
66. B. R. Peixoto, S. Brenner, *Immunogenetics* **51**, 443 (2000).
67. J. Stenvik, T.O. Jorgensen, *Immunogenetics* **51**, 452 (2000).
68. R. H. Waterston, E. S. Lander, J. E. Sulston, *Proc. Natl. Acad. Sci. U.S.A.* **5**, 5 (2002).
69. E. W. Myers, G. G. Sutton, H. O. Smith, M. D. Adams, J. C. Venter, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4145 (2002).
70. A. B. Carvalho, A. G. Clark, *Nature* **401**, 344 (1999).

71. Supported by the Agency for Science, Technology and Research, Singapore; the U.S. Department of Energy; and the Molecular Sciences Institute, Berkeley, California. We thank many colleagues and members of our labs for comments on earlier versions of the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/297/5585/1301/DC1
Sequencing Methods
Supplemental Data
Tables S1 to S7
Figs. S1 to S8
References

21 March 2002; accepted 14 June 2002

Tracing Black Hole Mergers Through Radio Lobe Morphology

David Merritt[1]* and R. D. Ekers[2,3]

Binary supermassive black holes are produced by galactic mergers as the black holes from the two galaxies fall to the center of the merged system and form a bound pair. The two black holes will eventually coalesce in an enormous burst of gravitational radiation. Here we show that the orientation of a black hole's spin axis would change dramatically even in a minor merger, leading to a sudden flip in the direction of any associated jet. We identify the winged or X-type radio sources with galaxies in which this has occurred. The inferred coalescence rate is similar to the overall galaxy merger rate, implying that of the order of one merger event per year could be detected by gravitational wave interferometers.

The detection of gravitational radiation from coalescing supermassive black holes (SBHs) would constitute a rigorous test of general relativity in the strong-field limit (*1*). However, the expected event rate is uncertain because the emission of gravitational waves is negligible until the separation between the SBHs falls below $\sim 10^{-3}$ to 10^{-2} pc. By contrast, simulations of binary SBHs at the centers of galaxies suggest that binary decay may stall at separations of ~ 1 pc, too great for the efficient emission of gravitational waves (*2*). It is unclear whether stellar- or gas-dynamical processes are capable of bridging this gap in a time shorter than the mean time between galaxy mergers. Here we consider whether black hole coalescence can alter the spin axis of the larger black hole and yield a detectable geometric signature in the radio observations of merging galaxies.

We estimate the effect of binary black hole coalescence on the spin of the resulting black hole using angular momentum conservation

$$\mathbf{S}_1 + \mathbf{S}_2 + \mathbf{L}_{\rm orb} = \mathbf{S} + \mathbf{J}_{\rm rad} \qquad (1)$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are the spin angular momenta of the two SBHs just before the final plunge,

[1]Department of Physics and Astronomy, Rutgers University, New Brunswick, NJ, 08903, USA. [2]Australia Telescope National Facility, CSIRO, Post Office Box 76, Epping, NSW 2121, Australia. [3]Radio Astronomy Laboratory, 623 Campbell Hall, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: merritt@physics.rutgers.edu