Fish genomes : a powerful tool to uncover new functional elements in vertebrates

Stupka, E.

Chapter 2: Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes

Published in: Science, 2002, Vol 297, pp. 1301-1310

Abstract

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 "tora- fugu," 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.

Methods

Sequencing Methods

Inspired by Celera's success with whole-genome shotgun approach to the Drosophila (A1, A2) and human (A3) genomes, we set out to sequence the Fugu genome using a similar approach (A4). The range of contiguity and scaffolding required for useful comparisons with other genomes are determined by (i) the size of a typical Fugu gene (roughly 10 kb) and (ii) the characteristic range of syntenic contiguity between the Fugu and human genomes (approximately five genes, or 50 kb in Fugu, which corresponds to nearly 400 kb in the human genome). Fugu chromosome arms are approximately 10 to 15 Mb in length, setting the practical upper bound for sequence reconstruction. To this end, and with an eye towards efficient use of resources, we set out to generate

approximately 6X sequence coverage of the Fugu genome.

Two kb inserts were the longest that could be reliably cloned into the high copy number plasmid pUC18 and its derivatives (JGI); a 2 kb M13 library was also made and end- sequenced (Myriad). A total of 5.2 X sequence coverage was generated from these 2 kb libraries at JGI, Myriad, and Celera, as summarized in Table 1. Uniformity of clone coverage and pair-tracking fidelity was confirmed by comparing these end-sequences with previously finished cosmid and BAC sequences. A slight cloning bias was noted in some libraries, reducing the effective coverage in AT-rich regions. Over 98% of cloneend pairs were correctly tracked.

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	2.26x	1,264	3.33x
NFP*	1.97 ± 0.24	JGI	269,216	121,908	628	169	0.44x	244	0.64x
LPO	1.98 ± 0.33	JGI	164,048	67,240	498	82	0.21x	134	0.35x
XLP	1.94 ± 0.24	JGI	43,797	18,796	605	27	0.07x	38	0.10x
MYR	2.06 ± 0.28	Myriad	1,100,171	435,956	478	526	1.38x	872	2.39x
CRA*	1.97 ± 0.23	Celera	510,131	221,548	609	311	0.82x	443	1.15x
CRA2	5.36 ± 0.70	Celera	186,238	83,504	650	121	0.32x	459	1.18x
LPC	39 ± 4.6	JGI-LANL	40,509	16,114	471	19	0.05x	645	1.65x
OML	68 ± 31	JGI-LANL	26,599	12,130	561	15	0.04x	1,031	2.17x

Total	3,711,256	1,608,955	574	2,129	5.60x	5,130	12.96x

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.

To obtain intermediate-scale linking information that could span dispersed transposon-sized repeats, a 5.5 kb insert pBR322-derivative plasmid library was constructed (Celera) and end-sequenced to 1.3X clone coverage. Longer inserts up to 10 kb were attempted but could not be reliably cloned. For longer-range linkage information and assembly validation, pre-existing cosmid and BAC libraries were end-sequenced to 1.7 X and 2.7 X clone coverage, respectively. This BAC library (estimated to have insert size 85 +/- 40 kb) was the only library made from DNA of a different individual fish (G. Elgar, unpublished), and is also being fingerprinted (4.7x clone coverage), however fingerprint based maps were not available for the assembly presented here.

The net sequence from all libraries combined was 2.13 billion bases, or 5.7 X sequence coverage of a presumed 380 Mb genome. This sequence total refers to net high-quality nonvector read length of passing reads, where “high-quality” bases were determined by a quality score–based trimming protocol as described below, and passing reads had 100 or more high quality bases. Seventy-six percent of clones had passing sequence from both ends, resulting in over 1.6 million end-pair linkages.

Sequence quality trimming

A uniform trimming protocol was applied to raw sequences generated at JGI,

Celera, and Myriad to extract high-quality nonvector sequence from each read. Briefly, after initial vector screening with CrossMatch, windowed averages of Phred Q-values (A5) were calculated. Called bases with windowedverage quality less than a library- and primer-dependent threshold were discarded, and the longest stretch of continuous high quality bases retained. Reads were then further trimmed by fixed offsets from each end. Trimming parameters (minimum windowed quality score and up- and downstream end offsets) were determined for each library/sequencing batch to optimize the net length of quality sequence available to the assembler using the following protocol: (a) A sampling of reads from each library was aligned with known reference sequence from GenBank using BLAST; (b) For each set of trim parameters, the net length of aligned sequence was calculated, ignoring reads whose alignments did not extend across the entire trimmed read; (c) Trim parameters were then chosen to optimize this net length. Typically, minimum windowed Q-scores above 15-20 and offsets of 0-10 were used.

Assembly

Polymorphism rate estimation

To assess the intrinsic polymorphism rate in Fugu we used two approaches: First, all scaffolds were examined and positions at which two nucleotides had support from two or more raw sequence reads were designated as polymorphic. Assuming a Poisson distribution and making a correction for null sampling of polymorphisms, we determined variable sites to be 0.4% of the sequence, approximately five times more frequent than in the human genome. We also compared the assembled sequence to a finished cosmid, (165K09) of length 39.4 kb which should exhibit maternal/paternal variation. Positions at which two

nucleotides had support from two or more read sequences were designated as polymorphic. This procedure distinguishes true polymorphisms from sequencing errors, which occur at a comparable rate. The cosmid sequence was finished to the standard one part in 10,000 and therefore positions at which the read sequences consistently differed from the cosmid were flagged as polymorphic. We found 137 SNPs (including single base indels) and half a dozen multiple base indels ranging in size from 2 to 6 bp, which is consistent with our genome wide estimate presented.

JAZZ – a novel suite of tools for whole genome shotgun assembly

Pairwise sequence overlaps between nonrepetitive reads were calculated by means of the Malign module of JAZZ. Using a parallel hashing scheme, all read pairs sharing more than ten exact 16-mer matches were aligned using a banded Smith-Waterman method. To avoid attempting unnecessary alignments, the 16-mers that occurred frequently were not used to trigger alignments. These "unhashable" 16-mers include (A)16, (AT)8, and other common low complexity sequences whose shared occurrence in a pair of reads is not a strong predictor of likely overlap. From these unhashables a catalog of microsatellites was constructed. The computational work entailed by Malign is formally O (G d2) where G is the genome size and d is the sequence depth. These calculations can be distributed throughout the sequencing effort and are not rate limiting.

After Malign generates a set of high sequence identity pairwise alignments between (vector-screened and quality-trimmed) reads, the Graphy module of JAZZ uses this information, in conjunction with pairing relationships between clone end sequences, to create a self-consistent scaffolded layout of reads. This

calculation takes into account a wide range of information, including: the number of high quality overlaps possessed by each read relative to the expected Poisson distribution of overlaps; consistency of alignments between mutually overlapping reads, which allows isolated sequencing errors to be discounted; and repeat boundaries to be identified; increased confidence in an overlap between two reads that is "corroborated" by overlaps between their sisters, etc. Scaffolds are formed self-consistently by creating initial scaffolds using highest quality information, breaking these scaffolds based on inconsistent topology, incorporating lower quality overlaps, and iterating. This phase of the calculation is distributed and took less than one day on an 8 CPU Sun system.

Consensus sequences were generated by means of an efficient algorithm, THREE, that creates an initial tiling path across each contig, with each tile comprising a read-segment that represents those parts of the contig expected to be closer to the middle base of a read than to the middle of any other read. Master-slave alignments between these tiles and other overlapping reads are recovered from Malign, and a weighted scoring system is used to determine consensus, at the same time computing a Phrap-like consensus quality score. High-quality discrepancies with the consensus corroborated by two or more reads are flagged as putative polymorphisms. This phase of the calculation is also highly parallelized, and took less than 1 day on the 8 CPU system.

The final stage of the assembly is an attempt to close captured gaps (ie gaps internal to scaffolds). For this purpose, small Phrap based assemblies are used. For each captured gap, a weighted average of spanning clone lengths can be used to estimate the gap size. In some cases (notably those with nominally negative gap sizes), flanking contigs can be joined directly by means of weak, short,

and/or low complexity overlaps that were either not detected by Malign or can only be trusted with the additional corroboration provided by the clones spanning these captured gaps. These procedures closed 12,709 out of 45,330 captured gaps.

Repeats and assembly

Highly repetitive sequences – both the clusters of tandem repeats that are the principal component of heterochromatin, as well as the interspersed repeats that are distributed throughout the genome in both hetero- and euchromatin – are problematic for both whole- genome shotgun and BAC-by-BAC sequencing strategies (A6). These difficulties arise both from differential cloning efficiency and the complexity of faithfully assembling such genomic regions. Even deep data sets may not contain sufficient information to reconstruct long, high sequence identity repeats (especially tandemly repeated ones), and special finishing data are generally required to reconstruct these problematic genomic sequences regardless of shotgun sequencing strategy.

Major repeat classes in the Fugu genome (and a small number of low-level contaminants) were identified by culling trimmed reads with an unusually large number of high fidelity (97% nucleotide identity) sequence overlaps in initial sequencing data. These reads were clustered, and small (few thousand read) samplings of these clusters were assembled with Phrap (A5) to identify sequences that appear at high copy number in the genome. Several classes of repeats were identified, and reads corresponding to these classes were flagged and set aside for repeat-specific analyses and assemblies. In the final data set 196,050 passing reads (approximately 5.3% of the raw data) were set aside in this manner, including several families of tandem repeats (3%) and a group of

predominantly interspersed LINEs and other transposable elements (1.5%). Since different library and sequencing protocols exhibited varying representations of several repeat classes (data not shown, centromeric satellites, rRNA), indicating differential cloning or sequencing efficiencies, only approximate estimates of the coverage of the genome by these repeats can be made.

The dominant tandemly repeated element in the Fugu genome (approximately 2% of the passing reads) is a 118 nt satellite sequence (A7) presumed to be centromeric in origin (A8). A similar 118 nt repeat (57% sequence identity) has been localized to centromeres in the freshwater pufferfish Tetraodon nigroviridis (A9) which should share a similar chromosomal structure with Fugu. Over 90% of reads containing this centromeric repeat have sister reads that are also in this class, confirming the highly tandem nature of this array.

In higher vertebrate genomes, ribosomal RNA genes typically occur in tandem clusters whose repeated unit is either the 18S-5.8S- 28S rRNA operon or the 5S rRNA gene. We find this same organization in Fugu, with 0.3% of the reads matching the 18S-5.8S-128S operon and 0.6% hitting the 5S gene. The overwhelming majority of paired-sisters of these reads (85% and 73%, respectively) hit the same rRNA gene, confirming the highly tandem nature of these gene clusters. Transposable elements of various types were found in the sisters of 5S rRNA-containing reads 18 times more often than in the 18S-5.8S-128S group, indicating that transposon insertions are more prevalent within the 5S tandem repeat. The homologous Tetraodon rRNA clusters have been localized to the short arm of two chromosome pairs, confirming their tandem organization.

Long range linking information from BACs and cosmids

Approximately 3.8X clone coverage from paired cosmid and BAC-end sequences was obtained. An assembly was performed with these read pairs to order and orient the small- insert-derived scaffolds. This procedure led to substantially longer scaffolds, but also introduced an unacceptable number of large (greater than 10 kb) captured gaps spanned only by the large insert clones. This was further confounded by the large variation in BAC insert size. These are not gaps in sequence coverage, but rather in linkage. Using BAC and cosmid end linking information, 350 Mb is found in 961 scaffolds greater than 100 kb in length, with an additional 80 Mb found in 5,386 smaller scaffolds. Given the genome size, much of this apparent "excess" sequence belongs within the large captured gaps, and could be placed there with additional linking information at the 5-80 kb scale from additional 5.5 kb or cosmid-end sequence and/or other mapping information.

The occurrence of both ends of a BAC or cosmid in the same scaffold provides an independent corroboration of assembly fidelity at the 40-100 kb scale. A total of 98.7% of cosmid ends assembled into the same small-insert-derived scaffold were placed within 35-45 kb in the proper orientation. The wide range of insert sizes in the BAC library, coupled with an extensive fingerprinting project (G.Elgar, unpublished), allowed us to further test the assembly. With a minor calibration offset, the separation of BAC-ends on the assembly was evidently in good agreement with experiment for BAC inserts ranging from 15-200 kb in size. (Note that 30 BACs had both ends assembling in the same location (inferred size zero) implying a probable insert deletion.)

Clone-end tracking

Clone end tracking is an essential requirement for successful large shotgun sequencing projects. We assessed the fidelity of these pairing relationships both before and after assembly. Before assembly, reads from clones with passing sequence at both ends were aligned against a finished cosmid sequence. For all 2 kb and 5.5 kb insert libraries, approximately 99% of such reads had sisters placed within four standard deviations of their expected location. Nearly half of the discrepancies were due to plate tracking errors, which can be identified as entire plates of incorrectly paired reads. On the basis of smaller sequencing projects at the Fugu sequencing centers, the next dominant mode of failure was chimeric inserts (i.e., two random genomic fragments that fuse and are cloned as a single insert).

Sequence accuracy

Given the high degree of similarity between Fugu proteins and those from other vertebrates, an indirect measure of sequence accuracy can be obtained by counting the number of indels introduced into exons by GeneWise (A10,A11). Since indels within coding regions introduce frameshifts, they are easily recognized as errors. We found that indels are introduced by GeneWise at a rate of one per 4,600 bp. This is likely to be a slight overestimate of the indel rate, since some small fraction of the GeneWise models may correspond to pseudogenes, but is consistent with our overall estimated error rate of 5 parts in 10,000.

Annotation methods

The method used to annotate the Fugu genome consisted of a computational pipeline which was similar to that of the human EnsEMBL project (A11,A12). We applied homology feature- based identification of genes, using BLAST to locate potential gene loci and diverse protein databases (SWISSPROT human, SWISSPROT nonhuman, EnsEMBL-mouse, EnsEMBL- human) and EST databases from a wide range of organisms. We used GeneWise and the EnsEMBL genebuilder to build and prune potential gene models. Predicted proteins were subsequently annotated with a protein pipeline designed to map Interpro domains and secondary structures onto the sequences. Our code is freely available at www.fugubase.org

As is evident from the studies of the human genome sequence, the computational determination of gene structures in vertebrate genomes is far from straightforward for several reasons. First, in genomes such as human, the ratio of coding information to noncoding means gene structures must be built across large regions of noncoding DNA sequence. Second, the data sources used to provide evidence of a predicted structure are fragmented - this creates difficulties in determining overlapping protein information. The main objective of automated annotation is to provide approximation and locations of features on a global scale which, over time, will need to be refined by further data and analysis. The gene models produced by automated prediction contain some errors, which will be eliminated over time. Refinements and additions especially to comparative features will be added to the web sites displaying live annotations (www.fugubase.org and www.jgi.doe.gov/fugu).

Fugu scaffolds from the assembly were first repeat-masked with RepeatMasker.

RepeatMasker is efficient at detecting many types of repeat, including t-RNA and ribosomal RNA sequences. A number of Fugu repeats have been discovered, and single reads that contained mostly repetitive sequences were screened out from the assembly in the early phases. Therefore the scaffold assembly is relatively depleted in certain types of repeat (eg. 118 bp minisatellite).

After repeat masking, the Fugu scaffolds were searched against a series of protein databases and similarity features were written into an EnsEMBL-like database of features. The highest matching feature over the greatest length, was used as input to GeneWise with parameters [- ext 2 -gap 12 -subs 0.0000001]. Gene models from GeneWise were subsequently pruned for redundancy using the genebuilder logic from EnsEMBL (A12).

The databases searched were:

1. Human entries from SWISSPROT and Translated EMBL (TrEMBL) version 39 (45420 entries) (SPTRhuman)
2. Nonhuman entries from SWISSPROT and TrEMBL version 39 (699219 entries) (SPTR others)
3. EnsEMBL confirmed human peptides from release 1.3.0 (28706 entries)
4. EnsEMBL confirmed mouse peptides from release 0.1 (16679 entries)
5. All Human genscan predictions from repeatmasked golden path sequence (August 6th 2001 build)

BLAST features were filtered by position so that only the best hsp (high scoring segment pair) for any given DNA position was stored. This process generated a total of approximately 2x106 similarity features from protein databases.

Searching in this fashion produces spurious hits as shadow exons in some cases

(a similarity hit in the same location but on opposite strands). In addition, the majority of short (less than 30 residue) gene models with single molecule BLAST supporting evidence were apparently low homology "dust." Finally a number of peptides with high composition of low complexity repeats, which resulted from undetected DNA sequence repeats in the genome matching other protein repeats, were observed. These were detectable when pseg was used to identify proteins of >50 residues where more than 50% of the total sequence was low complexity repeat, or <50 residues and more than 80% low complexity repeat. Each of these classes of sequence was eliminated from the final list of predicted peptides.

In order to assess the potential error in these estimates we compared the effectiveness of the genebuilding modules in our pipeline to a series of annotated Fugu sequences, which contained 209 genes. This showed that the sensitivity of pipeline detection for both exons and gene loci on this sample was 93%, while the specificity measure based on both true hits and false positives was 79%.

Translated comparison of Fugu and human genomes.

One means of enlarging the potentially homologous features for gene building is to translate and compare human and Fugu genomic DNA in all six frames. This process is computationally intensive and produces more background noise than other forms of comparison. To reduce compute time and noise, we investigated two sets of parameters for Wu-tblastx on a sample of scaffolds and made two comparisons, one with W=5, T=20000, E=1E-05, nogap and matrix=identity; the other w=4, E1=1e-05, E2=1e-05, matrix=blosum62. tblastx homology features were not used for gene building in the present dataset but for estimating how

many additional loci might be built from this method. The ratio of overlaps to nonoverlaps derived from these two parameter sets appeared to be almost linear in scaling.

Similarities were computed with Fugu ESTs from the public domain and from a small est sequencing project at the IMCB Singapore. These totalled 4000 EST sequences) Estimation of the protein domain content of the Fugu genome

Gene models taken from the aggregate of gene build methods were translated to produce conceptual proteins, which were then analysed for domain content with the following methods:

Hmmpfam (A13), HMM search of the Pfam (A14) database of protein domains.

Using FingerPRINTScan (A15) to identify PRINTS (A16) sequence motif fingerprints in the protein sequence . Pfscan to search for PROSITE (A17) motifs and sites Secondary structure prediction for helical/coiled-coil motif, low complexity regions, signal peptides and transmembrane predictions (A1820).

The threshold parameters used were the same as those in the public human genome analysis (A21).

Identification of putative conserved regions with the human genome

The classical approach to determining conservation of synteny for genes is to first assign orthologous relationships for each protein and then to examine the spatial relationships of the gene loci encoding these proteins. Assignment of orthology can be a difficult procedure to automate for some proteins because, especially in clustered gene families such as Hox proteins, the differences between family members are so few that careful alignment by hand followed by phylogenetic inference is required and in some cases accurate assignment may

remain impossible. This is not feasible for examining a whole genome. Automated assignment on the basis of protein similarity comparison alone may be in error in any given pair. The probability that multiple pairs would be misssigned to the same chromosomal segment diminishes exponentially as the size of the linkage groups increases.

We have used an alternative procedure to estimate potential regions of synteny over chromosomal segment sizes dictated by the granularity of the present assembly. Firstly, we used a reciprocal best hits method similar to INPARANOID (A22), to determine putative orthologous proteins between Fugu and human. For practicality, we used the human EnsEMBL peptide set (November2001) since human chromosome positions are easily accessible. A total of 31,059 Fugu proteins were searched against the EnsEMBL peptide dataset (28,706) and vice versa using blastp with the following parameters: BLOSUM62 matrix, expect score ≤ 1e-07 and at least 30% identity across the length of the query sequence. The reciprocal best hits (9,829) were extracted and taken as the likely orthologous proteins. In the next step, for a given human chromosome, we identified human proteins (regardless of gene order) that have orthologs linked in cis on a single Fugu scaffold. Conserved segments containing two or more genes were considered for detailed analysis. Intervening genes (i.e., nonsyntenic genes that are interspersed with the orthologous genes in a conserved segment), ranging from zero to 1,280 were calculated for each conserved segment. Both discrete and continuous intervals were examined.

Enumeration of IgSF domains.

The Pfam "ig" hidden Markov model (A14) was used to query approximations of

complete sets of proteins from D. melanogaster (The FlyBase Consortium, 2002), C. elegans (WormBase, December 2001 release), F. rubripes (genscan predicted proteins), and H. sapiens (IPI Version 2.0). IgSF domains were detected with the HMM algorithm on GeneMatcher hardware with an e-value cutoff of 1 (A23). The cutoff was set so as to detect the all the known IgSF proteins in C. elegans (A24), while minimizing false positives. Teichmann and Chothia (A24) enumerate 488 I-set IgSF domains and 64 IgSF proteins in C. elegans. Protein sets were masked with pseg prior to analysis, with parameters "25 3.0 3.3" and "45 3.4 3.75" (A25).

Detection of immune antigen receptor genes.

Antigen receptor genes were primarily detected by GeneMatcher Smith-Waterman comparisons of known genes with frame-shift–tolerant translations of the repeat masked Fugu assembly, as well as against Fugu genscan predictions. Known genes were obtained from both Genbank (NCBI) and IMGT (A26). Queries with tetrapod genes utilized BLOSUM65; other queries utilized BLOSUM30. Regions of interest were further analyzed with blastx queries (A27), PIPMaker (A28), MegAlign (DNA*, Madison, WI), and profile detection of recombination signal sequences. Once identified, elements were incorporated into query sets and used to identify additional elements.

Enumeration of GPCR proteins and cytokines

This was conducted in two stages – mining of predicted peptides and genomic sequence and then sub classification of receptor types:

Mining

The predicted Fugu peptides were matched against the pfam models for 7tms1-6 (using both the fragment and complete profiles) with a cutoff expect score of 0.001 using HMMer. The Pfam identifiers were PF00001,PF00002, PF00003, PF01461,PF01604 and PF02949. A SWISSPROT + TrEMBL seed was made automatically by running SWISSPROT+TrEMBL against 7tms1-6 (complete only) at an expect score threshold of 0.001, using HMMer.

A proprietary tblastn search of the GPCR seed against the assembly was undertaken using an expect scorecutoff of 10-10. blastp was used to identify and removesegments which were already present in the searches of predicted peptides. GeneWise was then used to build predictions using the best seed hit, using default parameters. To compare with human, GPCRs were mined from the human EnsEMBL 1.2 peptides. 7tms were extracted from this in the same way as for the SWISSPROT + TrEMBL seed above.

Classification

All GPCR protein predictions were blastp searched against the seed database. Predictions were initially classified as a member of a family based on the top BLAST hit. Clustalw was used to make multiple alignments and construct phylogenetic trees of closely related subfamilies of Fugu GPCRs (see below) and human family members taken from SWISSPROT. Proteins not clustering clearly with related family members were examined further: If the BLAST output indicated inhomogeneity (defined as there being hits to members of more than one family with a factor of 100 of the best expect score) then proteins were classified as orphan/unclassified.

The clasification scheme used was that of the GPCR database

http://www.cmbi.kun.nl/7tm/htmls/consortium.html (except that IL8 GPCR was classified as a chemokine). The groups for which trees were constructed were: 7tm1 amine receptors (i.e. histamine, serotonin etc), 7tm1 peptide receptors, 7tm1 nucleotide receptors, remaining 7tm1s, 7tm2 and 7tm3. Within these, trees were manually inspected to determine the robustness of bootstrap and proximity for clustering with known human members.

Results

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, four-fold 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); 1,908 scaffolds longer than 50 kb account for 60% of the assembly (200.8 Mb); 4,108 scaffolds longer than 20 kb account for 81% of the assembly (271 Mb).

Scaffold length	No. scaffolds
28217	9174
54433	1484
80651	707
106867	382
133084	252
159301	145
185518	85
211735	57
237951	32
264168	22
290385	31
316602	5
342819	7
369036	5
395253	3
421470	6
447686	2
657422	4

Table S1. Ranking of scaffold sizes and cumulative length intervals in the Fugu 5.7x assembly

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 non- redundant 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)– 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 wellnnotated 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 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 non-redundant 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 7,331 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.

Figure 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.

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 8,761 Fugu proteins below this threshold, a further 1,800 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 1,260 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.

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)	-	
Hoboctivator-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	-	-	

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.

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 under-represented in the assembly. Despite this under-estimate, 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 6,400 copies in the present assembly, as compared with the 1 million Alu 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.

In depth analysis of Repeat Families in Fugu

To investigate repeats in Fugu, we derived 72 consensus sequences for interspersed repeats, representing 48 new families of transposable elements and added these to the 8 consensus sequences already present in RepBase Update (http://www.girinst.org/Repbase_Update.html). The great majority of elements could be classified (see Table 2) according to this schema.

Transposable elements traditionally are grouped in two classes, the class I elements that move by reverse transcription of an RNA intermediate (retrotransposition) and class II elements that move by a cut and paste mechanism (DNA transposition). A third class has recently been added, rolling-

circle molecules, copies of which have not yet been observed in vertebrates (A29).

Retroelements are traditionally grouped in to SINEs, and long elements with or without 'long terminal repeats' (LTRs) (retrovirus- and LINE-like elements, respectively). This classification is 'morphological' and not strictly cladistic, but the latter is probably impossible to achieve for transposable elements, because new families can arise by recombination between distantly related elements and even de novo.

The 4 SINE families we found in Fugu consist of a tRNA derived polymerase III promoter region followed by a sequence homologous to the mammalian DNA transposon family MER6, a structure currently limited to SINEs in both bony and cartilaginous fish. These types of SINEs are generally known as Mermaids after one of the first described elements (A30). Usually, the 3´ end of a SINE corresponds to the 3´ end of a local LINE-like element, on which element the SINE is dependent for transposition. However, no LINE with a MER6- type 3´ end has been described so far, and we did not find any in Fugu either. The function of the MER6 unit is currently unclear. Crollius et al. (A31) reported an absence of SINEs in both Tetraodon and Fugu, but their search with transposable element translation products did not allow detection of SINEs. These are relatively numerous elements for Fugu (~5000 copies), though certainly not compared to SINEs in other organisms.

On the basis of the relationship of the reverse transcriptases, the Penelope elements form an outlying group among retroelements, and should perhaps not be named LINE-like. Two elements found in Fugu named Bridge1 and 2

(Kapitonov and Jurka, RepBase entries 1999) have also been named Neptune and Poseidon, and Xena (=Bridge2) (GenBank entry AF355377). Penelope matches are not yet described in other vertebrates. Three basal classes of LINE-like elements contain site-specific endonucleases. We derived consensus sequences for two families of one such class. They are closely related to the element Dong in the silkworm Bombyx mori and we therefore named them Dong_FR1 and 2. A literature search revealed these to be the same as the element Rex6. Site-specific LINE-like elements had not been described in vertebrates before.

The remaining LINE-like elements appear to form a single clade, with three major branches. The LINE1 branch is widespread in vertebrates, with LINE1 in mammals, Swimmer in fish, and Tx1 in amphibia. We reconstructed two Fugu elements with two translation products closely similar to those of Xenopus Tx1 (Tx1_FR1 and 2).

The RTE branch, previously known in vertebrates from the Bov-B element in ruminants and snakes, is represented by a family named Rex3_FR or Expander (Kapitonov and Jurka, RepBase entries 1999)(A32).

The large CR1 branch is represented in vertebrates by CR1 in birds and reptiles, LINE2 in Mesozoic mammals, and four families of Rex1 (A33) in fish, including Fugu. The most widespread interspersed repeat in Fugu, Maui, belongs to the LINE2 subgroup.

All four divisions of LTR elements are present in Fugu as well. The basal BEL or PAO- group is represented by an element named Catch1 (Kapitonov and Jurka, RepBase entries 1999) or Suzu. We reconstructed two Ty1/Copia-like elements

(Kopi1 and 2), both with only a handful of copies in the current assembly. These elements, which have unusually small LTRs (204 and 243 bp), encode proteins closest related to TNT and RIR1 in rice. Vertebrate retroviruses are a distinct, verebrate-specific subgroup of the Ty3/Gypsy division of retrotransposons. There are multiple, closely related, low copy endogenous retroviruses in Fugu, of which we reconstructed two. The 5´ ends of the internal sequences match the complement of arginine tRNA, from which tRNA reverse transcription probably originates. Nomenclature generally is based on these primer tRNAs, so that we named the elements FERV-R1 and 2 (Fugu Endogenous Retrovirus R). Most LTR retrotransposon in Fugu belong to the Ty3/Gypsy class. We reconstructed 3 more families, more closely related to each other than to either Sushi or Samauri, named Ronin1-3. These elements are the most abundant LTR elements in Fugu, leaving, between them, over 2000 copies in the genome (compared with 100 Sushi and 250 Samurai elements).

There are many more matches to reverse transcriptases in the current Fugu assembly that we did not explore. We did a more exhaustive search for DNA transposon families, leading to a slight bias in the densities of each repeat as reported in Table 2 in favor of DNA transposons. Unlike the retrotransposons classes, the DNA transposon classes are not discernibly related to each other and probably have independent origins altogether. A few families of short 'foldback', 'hairpin', a.k.a. 'MITE' elements could not be classified; however, these elements consistently are found to be associated with DNA transposons (e.g. (A34)). One of these, HP_FR1, is relatively common with at least 1300 copies. Again, most classes are represented in Fugu. Absent from Fugu and any vertebrate so far are

the insect-specific P elements, and the plant-specific En-Spm and MuDR classes. One interspersed reconstructed repeat has faint similarity to a MuDR transposase (FuguRep3), but the similarity is too low to warrant a classification. We built consensus sequences for two members of the harbinger class of DNA transposons, named Senkusha1 and 2 (Japanese for harbinger). These are the first vertebrate members of this novel class of DNA transposons, which has recently been renamed PIF-IS5 class by Zhang et al 2001. In maize they have given rise to the Tourist type of nonautonomous elements.

Vertebrate members of the small PiggyBac class have been described in Xenopus (T2) and human (the relatively young elements named MER75 and MER85, and Looper). We found one representative, Pigibaku (Japanese for piggyback), with ~ 200 copies in the Fugu genome.

The large IS630-Tc1 class has transposases that are related to the integrases of retrovirus-like elements (the transposase function being ancestral). The group consists of two deep branches. In vertebrates, mariner copies and Tc1-like elements in many species represent the classical IS630-Tc1-mariner branch, though only mariner fossils have been found in the human genome. Crollius et al. (A31) found homology to mariners in Tetraodon, but not in Fugu, and we confirm here that there are no sequences in the Fugu 5.7x draft that are derived from mariner elements. We did characterize five Tc1-like elements (named Tc1_FR1 to 5), which are spread through the phylogenetic tree of the Tc1 family. Vertebrate members of the more heterogeneous pogo branch have only been described in mammals so far, including some ancient elements related to the C elegans element Tc2 (Smit, RepBase entries) and the widespread Tigger (or

MER2-) family (A34). We typified two families of elements that fit within the Tigger clade (Tiggu1 and 2) and no fewer than 6 Tc2-like elements (Tc2_FR1 to 6) that are closest related to the Mesozoic mammalian Tc2-like elements. Unlike their mammalian counterparts, many of these elements have been active very recently, and some still contain full open reading frames.

The hoboctivator-Tag1 (hAT) class of DNA transposons is about as wide spread as the Tc1-class. Whereas the Tc1-class is primarily found in Metazoa, hAT transposons have been particularly successful in plants. However, one of the three deep branches in the hAT transposon family tree, the Charlie or MER1-elements, is specific to vertebrates (A34). These elements have been the most 'successful' DNA transposons in mammals. Charlie-like interspersed repeats are also the most common in Fugu, with over1500 copies spread by at least 8 different 'Chaplin' elements. Again, the Fugu elements appear to have been active much more recently. Zaphod elements belong to a second hAT branch that has been active in ancient mammals (Smit, RepBase entries) and are most closely related to Tip100-like elements in plants. We found one Fugu element, Trillian, with a close relationship to Zaphod.

Of the 'classical' hAT elements, which tend to be the only ones mentioned in hAT transposon studies, four derived genes have been noted in the human genome (21,35), but no evidence for transposon invasions in the last 200 million year or so could be found. In vertebrates an active element, Tol2, is known from the Medaka fish (A36). A single copy of a closely related element, containing a full ORF, is present in the Fugu 5.7x draft. We found two other elements in Fugu, Furousha1 and 2 (meaning 'tramp' in Japanese), with about 90 and 50 copies

respectively, that are most similar to the exapted genes in the human genome (one of which has been named tramp).

Some of the SINE and LINE elements show the distinct subfamily patterns of elements that have vertically transmitted within the genome for tens of millions of years. Their relatively low copy number may simply be due to a high rate of deletion of nonfunctional/junk DNA. The absence of closely related sequence and the very low copy number of most other elements suggest the possibility that some of the retrovirus-like elements and probably all DNA transposon families have been introduced through horizontal transfer, rather than representing lineages that have evolved by vertical inheritance in the genome of Fugu and its ancestors. This appears to be a bold statement, though one should consider that horizontal transfer might be the norm for evolutionary survival of DNA transposons. Furthermore, horizontal transfer might be a more likely event in a marine environment than on land, as the concentration of transmitting vectors is much higher. Much more regular horizontal transfer would also explain the large number of different families and classes of transposable elements that have been active in the Fugu genome as compared to the human genome.

Edwards et al (A37) have done an extensive study on the frequency of simple repeats in the Fugu genome. Aside from an observation on the low average divergence level of these simple repeats (see below), our analysis confirmed the relative frequency of each simple repeat (data not shown). A total of 1.86% of the assembly was masked as simple repetitive DNA by RepeatMasker (this includes highly imperfect simple repeats) and another 0.56% was classified as low complexity DNA.

In comparing the T. nigroviridis and F. rubripes genomes (A31), noticed an excess of polyA runs in Tetraodon compared to Fugu. In mammals poly A regions accompany LINE1 and some SINEs like Alu, and are thus very common. The interspersed repeats in Fugu are of too low copy number to make a difference in the distribution of simple repeats; even (GATT)n, which is the tail of the most common repeat, Maui, is not over-represented.

Several observations on repeats support the hypothesis that the frequency of larger deletions relative to point mutations is much higher in the Fugu genome than in mammalian genomes. Currently, the best data comes from the average divergence of (CA)n, the most common microsatellite, which is 14% in human and 6.6% in Fugu. There are two explanations for this discrepancy: concerted evolution of simple repeats works better in Fugu, or, more plausibly, microsatellites in the Fugu genome are eliminated more rapidly compared to the accumulation of substitutions. Another supportive observation is that interspersed repeats of the same divergence level have more internal deletions in Fugu. This observation could be more informative than the simple repeat observation, but is unfortunately hard to quantify in the present assembly.

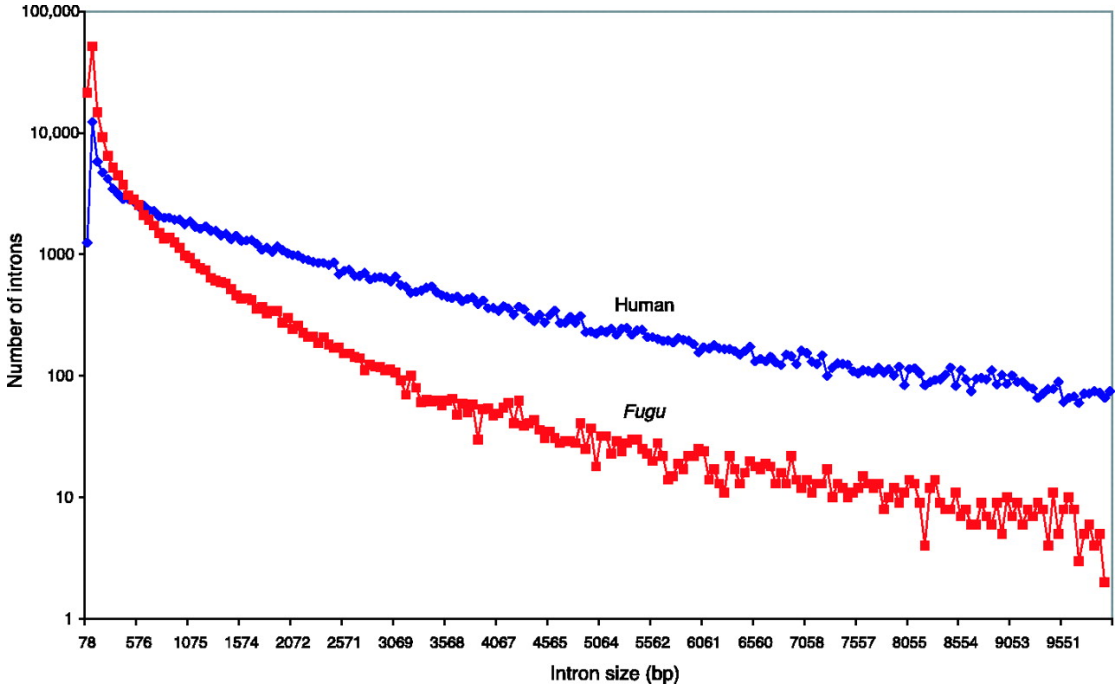

Figure 2. Comparative frequency distribution of intron sizes in Fugu and human.

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 2,609 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 (A33) have been observed. We examined 9874 orthologous gene pairs (A34) and observed 456 instances of concordance between intron-less Fugu and human genes; however, 327 human orthologs of intron-less Fugu genes contained multiple introns and 317 Fugu orthologs of human intron-less genes contained multiple introns.

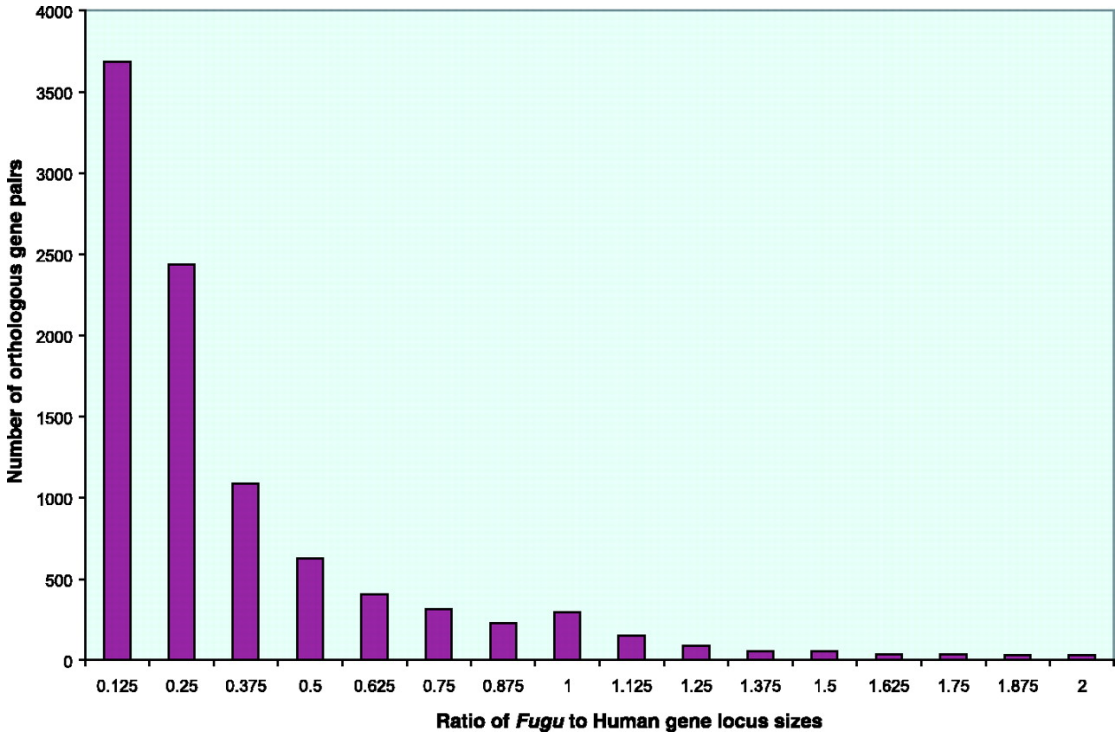

Figure 3. Distribution of ratios for gene locus sizes of putative Fugu-human orthologous pairs. Putative Fugu-human orthologous gene pairings were determined as described in supplemental methods relating to conservation of synteny.

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 1,176 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.

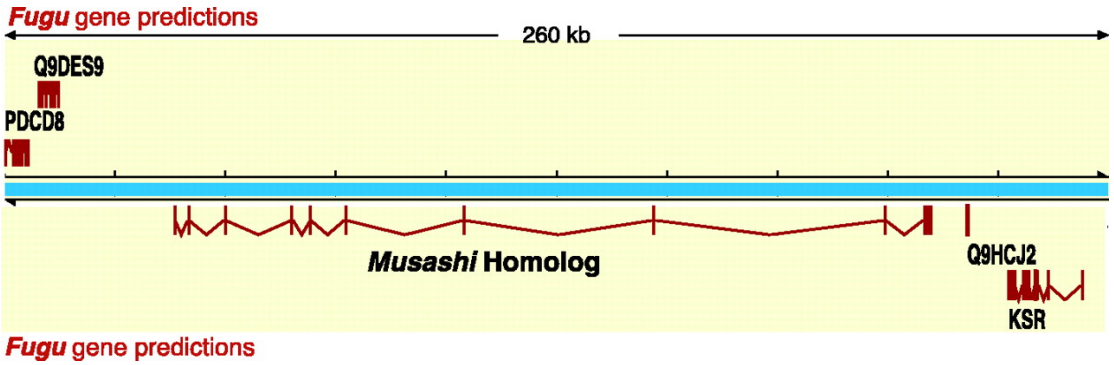

Figure 4. Organization of a giant gene locus in the compact genome of Fugu. The schematic shows a region of scaffold_1 with Fugu gene predictions shown in brown. Exons are represented by vertical brown lines. Introns are show as v-shaped brown lines between exons.

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 meth-ods 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.

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
1	8.2	43.8	24	39
2	7.5	43.7	9.8	41.7
3	8.5	44.3	4	43.7
4	7.2	44	1.3	44.4
5	7.2	43.7	0.8	48.9
6	8.1	44	0.2	51.6
7	7.3	44.2	0.1	46
8	7.1	44.3	0	0
9	5.6	44.9	0.1	53.5
10	7	44.9		
11	4.6	44.8		
12	3.7	45		
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
19	1.2	46
20	0.5	46.2
21	0.2	44
22	0.1	47.5

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).

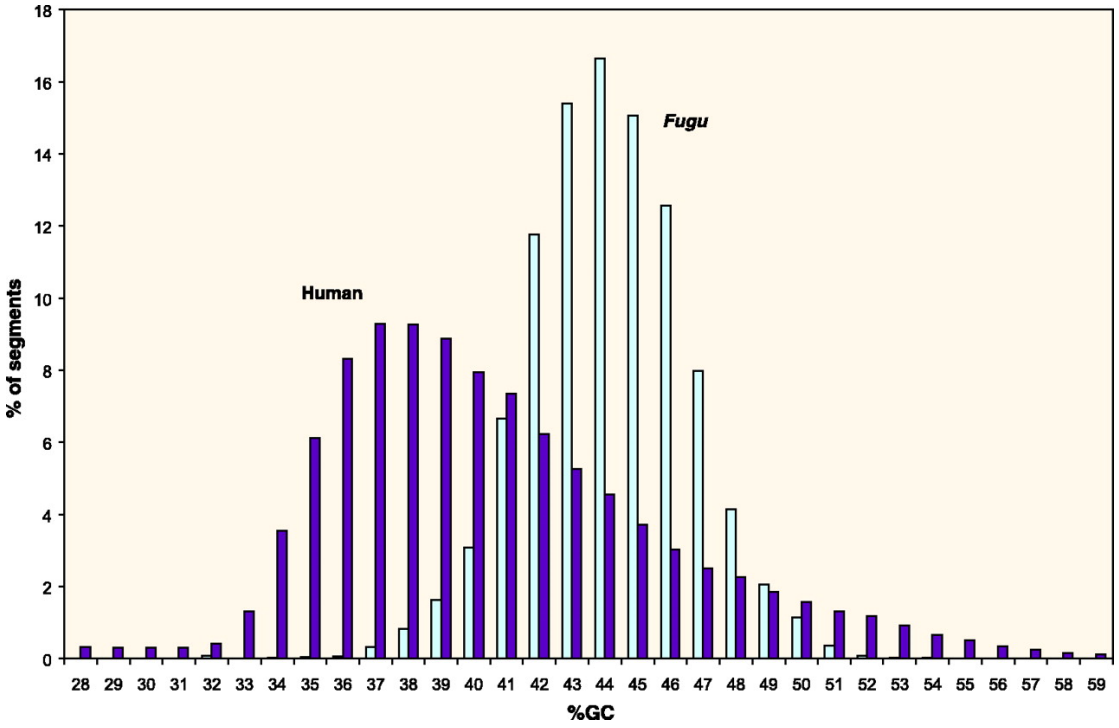

Figure 5. Distribution of GC content in the Fugu and human genomes. Sliding windows of 50 kb were used; similar conclusions were derived with windows of 25 and 100 kb (not shown).

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).

Table S2. Putatively conserved segments between Fugu and human.

Panel A. Absolute counts of conserved Fugu segments (gene pairs in clusters vs. number of intervening genes)

Gene pairs in cluster	unrestricted	0	0-5	6-10	11-20	21-40	41-80	81-160	161-320	321-640	641-1280
2	2436	765	816	143	121	90	126	117	125	89	44
3	1085	131	323	102	84	81	78	94	99	59	34
4	569	19	140	47	46	39	49	77	71	52	29
5	293	8	49	25	20	22	27	45	49	31	17
6	169	4	16	13	10	10	17	19	38	28	14
7	94	0	7	14	7	5	5	9	18	20	9
8	64	0	4	6	6	10	4	5	10	11	8
9	49	0	0	8	1	6	7	0	12	10	5
10	22	0	0	2	1	1	2	1	5	6	4
11	17	0	0	0	2	0	4	0	4	3	4
12	10	0	0	1	1	0	0	1	3	2	2
13	3	0	0	0	0	0	0	1	1	1	0
14	1	0	0	0	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	1	0	0
16	1	0	0	0	0	0	0	0	0	1	0

Panel B. Mean sizes of segments (kb) in human, corresponding to panel A

Gene pairs in cluster	unrestricted	0	0-5	6-10	11-20	21-40	41-80	81-160	161-320	321-640	641-1280
2	5912	121	307	910	1457	2484	5690	10154	24212	46262	101819
3	10085	230	474	975	1664	2307	6291	12461	24456	47799	100869
4	15744	320	533	1123	1649	3005	6055	10628	25477	52227	103174
5	17394	485	821	1218	1657	2436	5958	13712	24568	46882	88237
6	23530	629	621	1204	2069	3714	4976	13141	20954	54578	87996
7	24396	0	1354	1221	1882	4726	4234	10943	16910	51755	85662
8	26036	0	1740	1210	2669	3614	3123	11364	17107	49247	102231
9	30144	0	0	2197	4124	4303	4411	0	14436	60958	123167
10	35697	0	0	1513	4184	6351	4448	6296	11992	51462	96960
11	31301	0	0	0	4736	0	4343	0	13251	46868	77917
12	35706	0	0	1051	4910	0	0	14649	14119	61423	85626
13	33443	0	0	0	0	0	0	14742	21835	63751	0
14	33109	0	0	0	0	0	0	0	33109	0	0
15	33203	0	0	0	0	0	0	0	33203	0	0
16	41541	0	0	0	0	0	0	0	0	41541	0

Panel C. Mean sizes of segments (kb) in Fugu, corresponding to panel A.

Gene pairs in cluster	unrestricted	0	0-5	6-10	11-20	21-40	41-80	81-160	161-320	321-640	641-1280
2	24	17	27	36	29	27	32	35	20	18	17
3	31	33	29	41	28	35	26	32	35	21	23
4	39	43	39	46	41	32	41	33	47	33	40
5	54	41	65	69	52	48	33	53	53	48	63
6	65	56	58	53	83	45	72	62	72	65	70
7	73	0	78	83	65	70	90	60	64	75	77
8	83	0	114	83	124	60	88	70	76	76	94
9	97	0	0	124	175	52	75	0	84	116	116
10	127	0	0	117	288	16	71	91	116	145	142
11	117	0	0	0	258	0	93	0	82	64	145
12	134	0	0	67	330	0	0	151	98	112	139
13	153	0	0	0	0	0	0	165	125	169	0
14	151	0	0	0	0	0	0	0	151	0	0
15	165	0	0	0	0	0	0	0	165	0	0
16	143	0	0	0	0	0	0	0	0	143	0

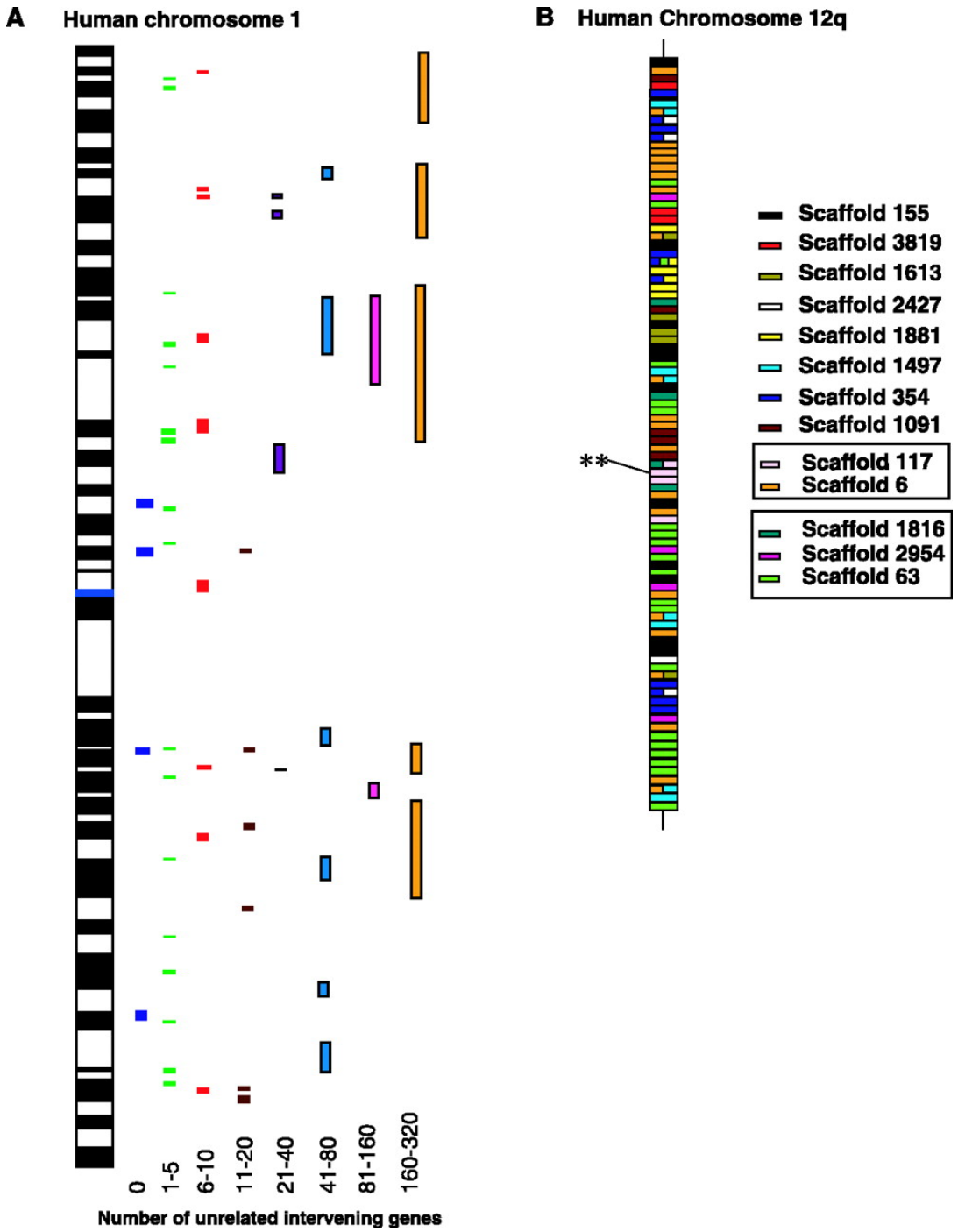

Figure 6. Conserved segments in the Fugu and human genomes. (A**) examines the distribution of conserved segments with different densities of unrelated intervening genes in discrete intervals. Each vertical track of colored boxes represents clusters of genes from *Fugu* scaffolds that map to human. The 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.**

The density of conserved segments varied between chromosomes. 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 rearrangement, 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).

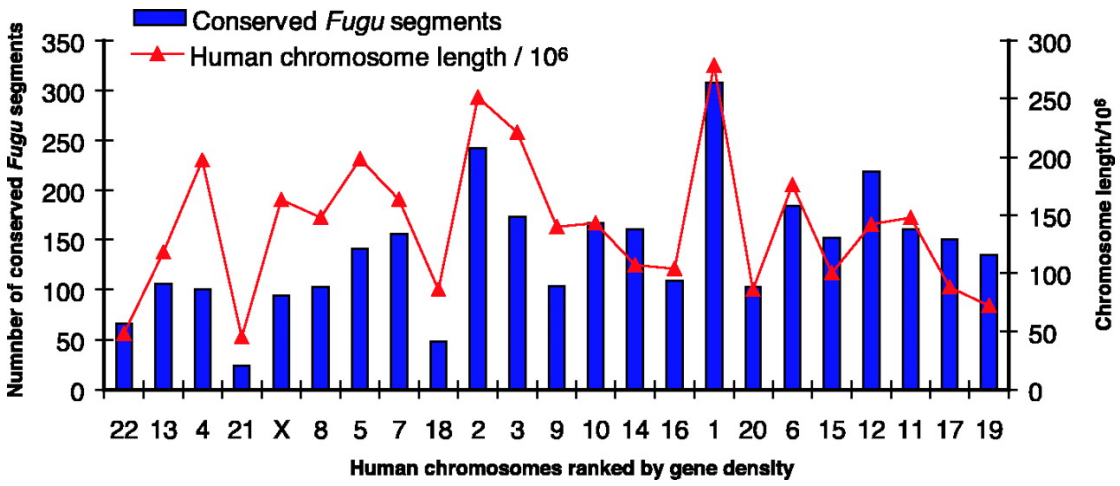

Figure 7. Distribution of conserved segments of Fugu on human chromosomes ranked by gene density. The figure shows the relation between the number of conserved segments of Fugu on human chromosomes, the length of human chromosomes, and their gene density. Chromosome 22 is the most gene poor, chromosome 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.

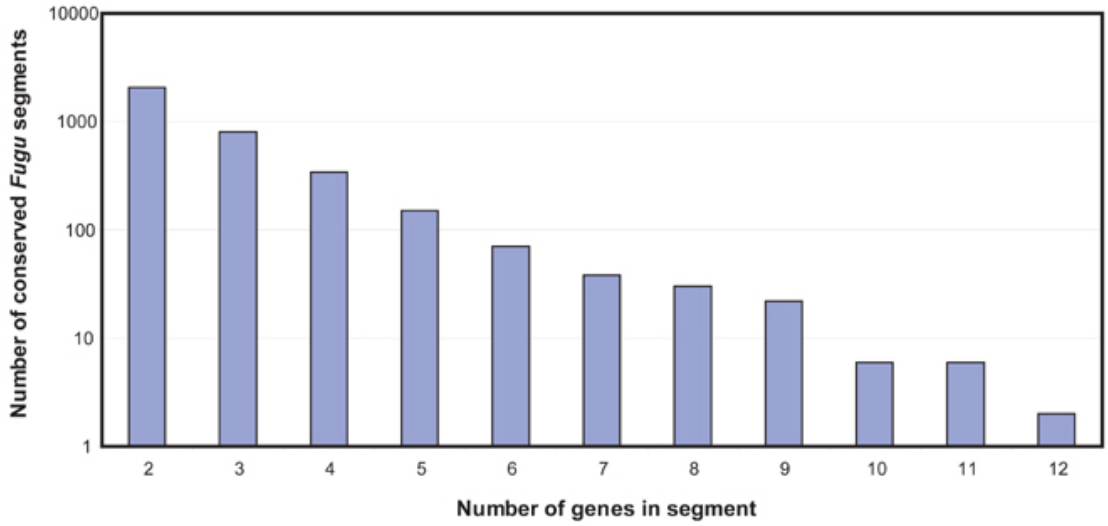

Figure S2. Relationship between the abundance of conserved segments between Fugu and human and the number of conserved genes per segment. This distribution is similar to that obtained for the comparison between human and mouse genomes (21).

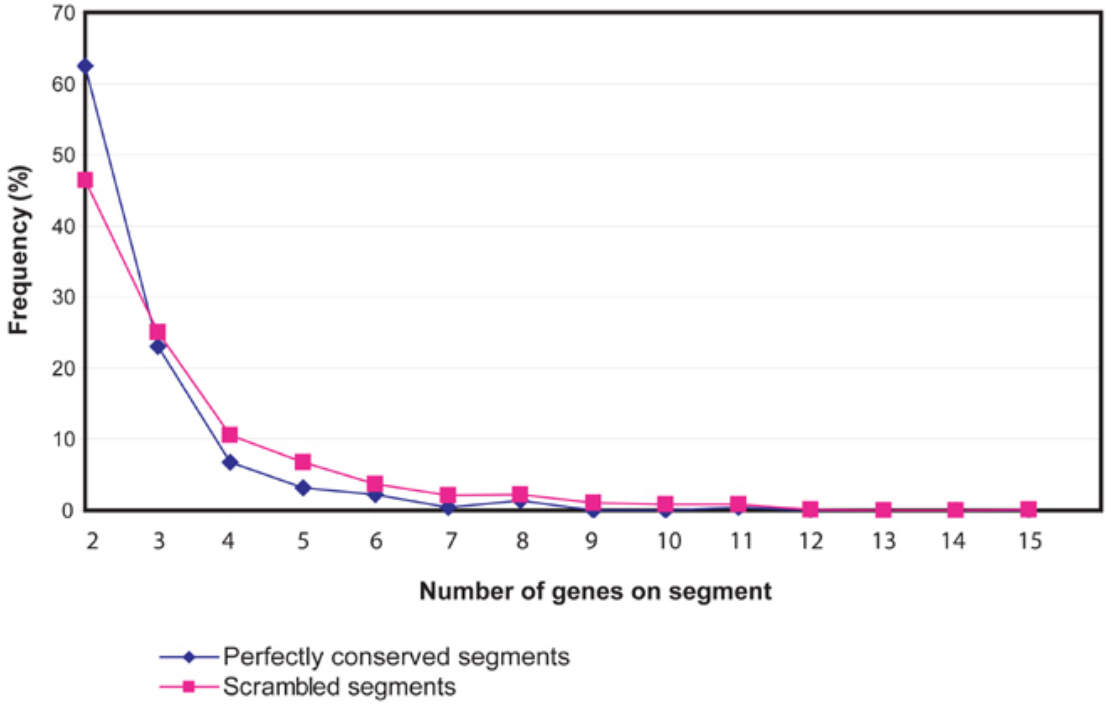

Figure S3. Relationship between the frequency of segments with a given number of genes, for perfectly conserved segments (0 intervening genes) and scrambled segments (1-n intervening genes).

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.

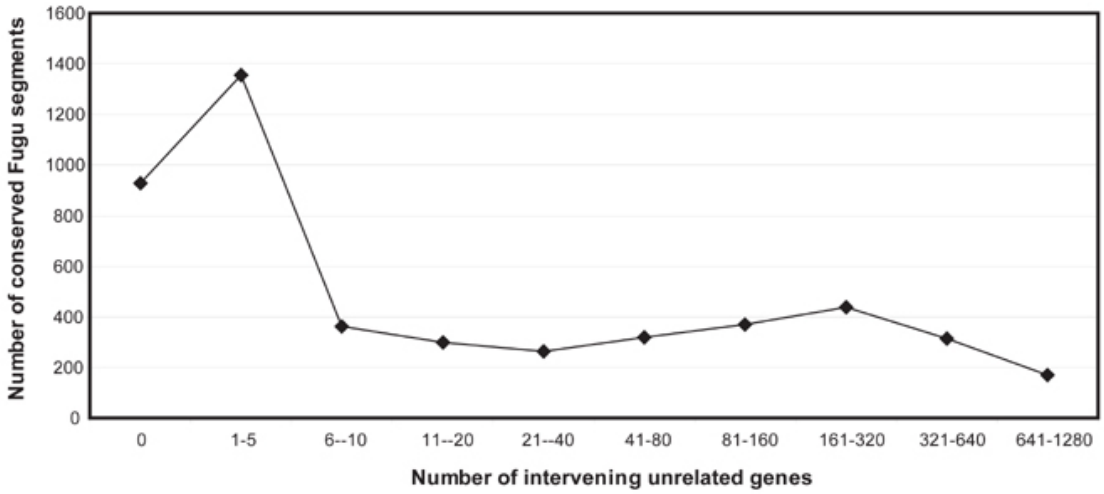

Figure S4. Relationship between the abundance of conserved Fugu-human segments and the density of unrelated intervening genes per segment in discrete intervals.

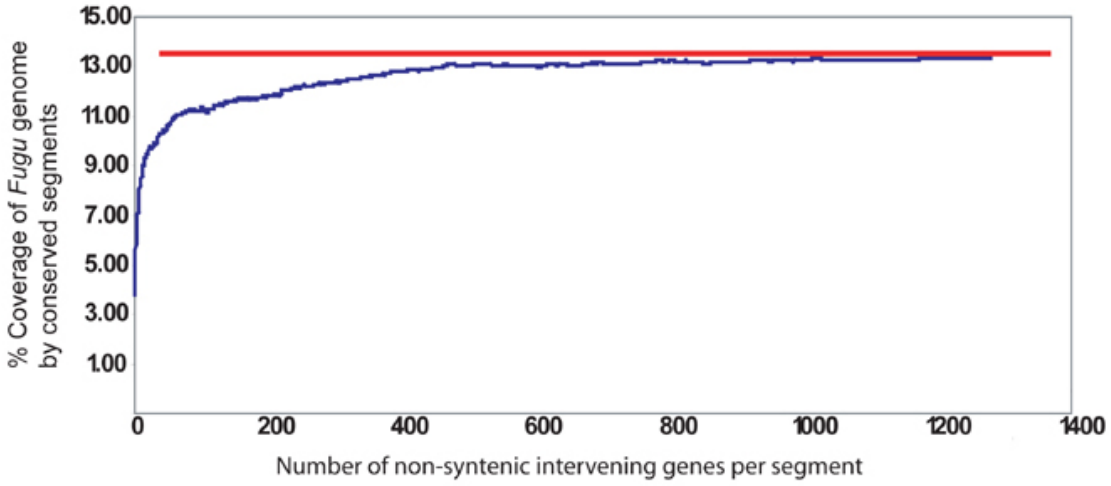

Figure S5. Coverage of the Fugu genome as a cumulative function of segments with increasing numbers of intervening genes.

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, -b, and -c complexes were identified, together with a fourth complex that was subsequently observed to be orthologous to a duplicated Hox 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).

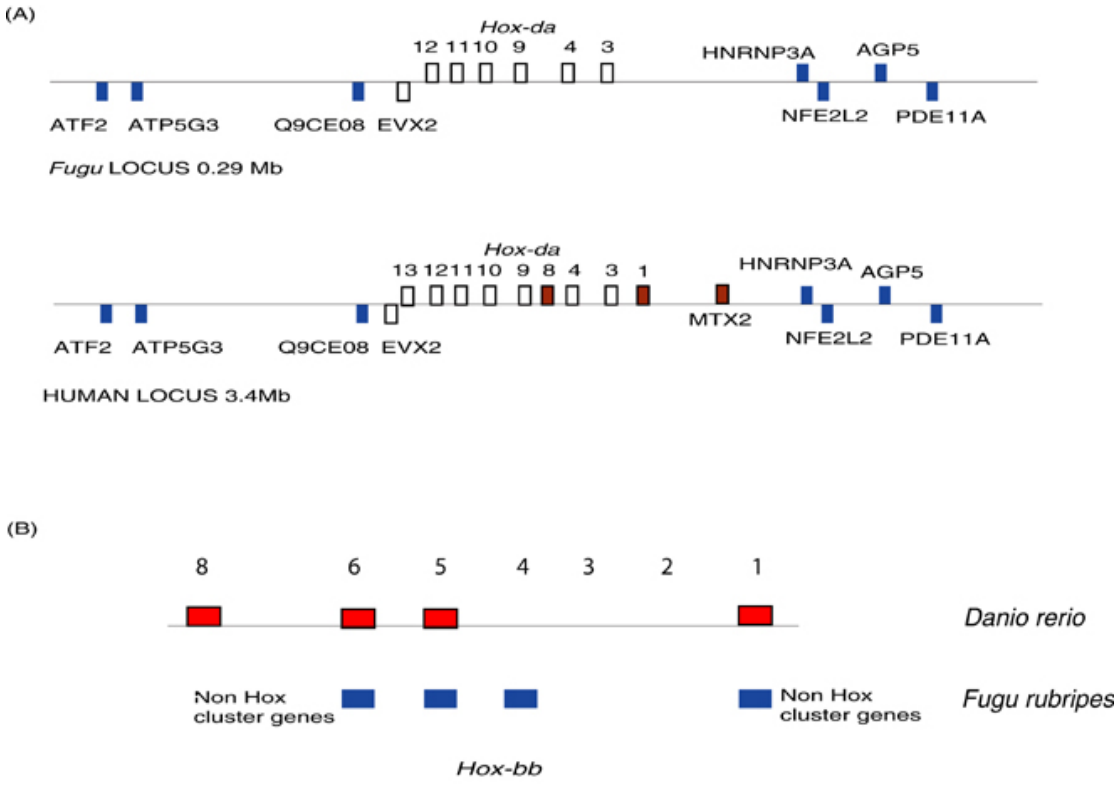

Figure S6. Hox genes. We searched for Hox genes using a homeodomain consensus motif. We were able to identify scaffolds for all of the previously described Hox complexes in Fugu, plus the complexes shown here. A number of small scaffolds contained single Hox genes, some of which could not be confidently assigned without additional linkage. The upper panel shows a schematic of a putative Hox-da complex in Fugu. Inspection of the predicted genes and comparison to other vertebrate Hox genes by alignment (not shown) revealed similarities of these proteins closest to vertebrate Hox-d. This classification is supported by the presence of an evx-2 orthologue at the 5′ end of the complex. We find no evidence of a true duplicate of Hox-d in Fugu. We have therefore classified this complex as Hox-da. When we further examined the proteins predicted in the 5′ and 3′ regions of these two Hox-da scaffolds we noted that the orthologous genes in human are also linked on chromosome 2 not only in the same region, but with the same relative spacing, over an interval of approximately 5 Mb. The Fugu region spans approximately 350 kb. The contiguity for these genes extends to the end of the scaffold sequences in each direction and could therefore be even larger than indicated here. Scaffold_183 appears to contain a truncated complex of at least four genes that clusters with the duplicate Hox-bb complex described in zebrafish. However this small cluster is evolved from the zebrafish orthologue in that a group 8 paralog has been lost in comparison with zebrafish and Fugu appears to possess a group 4 member not described in Zebrafish. Inspection of the other scaffolds shows that members of previously identified Hox genes are accounted for although not all of the complexes are contiguous in this assembly. Panel A shows the Hox-d complexes of Fugu and human compared. Hox genes are shown as open rectangles, genes present in human but absent from Fugu are shown in brown. Non-Hox genes from the locus are shown in blue. Panel B shows the zebrafish and Fugu Hox-bb complexes compared. The zebrafish genes are shown in red, paralog groups are shown above the genes.

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).

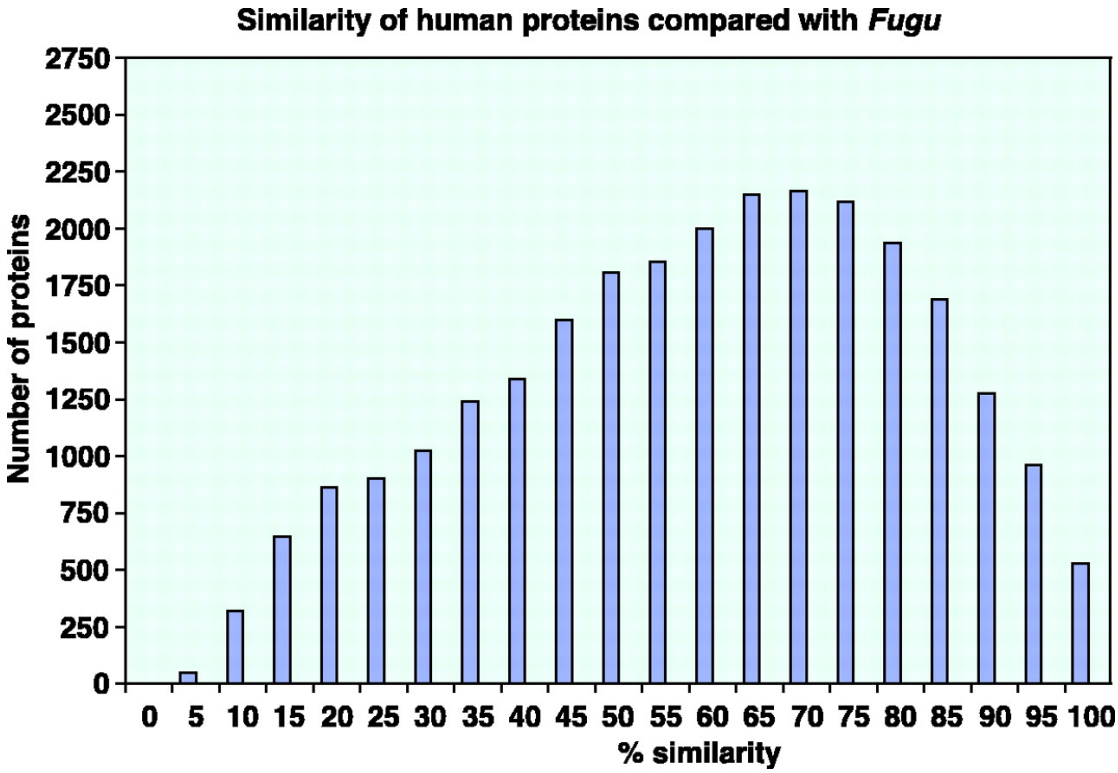

Figure 8. Distribution of protein similarities between *Fugu* and human proteomes. Global similarities were calculated as the sum of similarities in all nonoverlapping HSPs using a BLOSUM62 matrix, over the query (human) sequence length.

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

(8,109) do not appear to have homologs in the Fugu genome. In a reciprocal comparison, we noted that 6,000 Fugu predicted proteins lacked significant homology in human. We searched the 8,109 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 8,109 non-matching human proteins. We noted that 1,237 proteins were classifiable through Interpro domain classification. Of the remaining 5,268 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 non-matching 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 S3), 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 S3). 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 anti-viral 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.

Human Protein	*Fugu scaffold*	*Human protein*	*Fugu scaffold*
Interleukin-1 alpha	Not detected	Interleukin-15	Not detected
Interleukin-1 beta	5218	Interleukin-15 R alpha	?879?3846
Interleukin-1 R	7526,2663,386,111 3 others	Interleukin-16	2699
Interleukin-1 R like 1	614, 2667,6262	Interleukin-17	2739
Interleukin-1-like(toll)	398, 463, 6287	Interleukin-17 R	6141
Interleukin-2	Not detected	Interleukin-18	?5721
Interleukin-2 R alpha	Not detected	Interleukin-18 R	2663, 386
Interleukin-2 R beta	Not detected	Interleukin-19	115
Interleukin-2 R gamma	396	Interleukin-20	Not detected
Interleukin-3	Not detected	Interleukin-20 R	?3065?4345
Interleukin-3 R alpha	Not detected	Interleukin-21	Not detected
Interleukin-3 R beta	359	Interleukin-21 R alpha	?1072
Interleukin-4	Not detected	Interleukin-22	Not detected
Interleukin-4 R alpha	Not detected	Interleukin-22 R	1722
Interleukin-8	Not detected		
Interleukin-8 R alpha	1375	Interferon R ab-beta	Not detected
Interleukin-8 R beta	2580?2667	Interferon R ab-alpha	6320
Interleukin-9	Not detected	Interferon-alpha	Not detected
Interleukin-9 R alpha	Not detected	Interferon-beta	Not detected
Interleukin-10	?115	Interferon-cluster Human Chr9	Not detected
Interleukin-10 R alpha	6320	Interferon R gamma-alpha	?3065
Interleukin-11	Not detected	Inteferon R gamma-beta	?3065
Interleukin-11 R alpha	5577	CD4 beta	Not detected
Interleukin-11 R beta	5577	CD4 alpha	Not detected
Interleukin-12	2059/2697?	CD8 alpha	Scaffold_119
Interleukin-12 alpha	6141	CD8 beta	Not detected
Interleukin-12 beta	1425		
Interleukin-13	Not detected		
Interleukin-13 R alpha	316		
Interleukin-14	542		

Table S3. Interpro descriptions of human predicted peptides with distant or no significant homology in Fugu. Descriptors of human predicted peptides with distant or no matches in Fugu. The first column is the Interpro long description, the second column the number of human proteins. The source of the predicted peptides was NCBI version 26.

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.

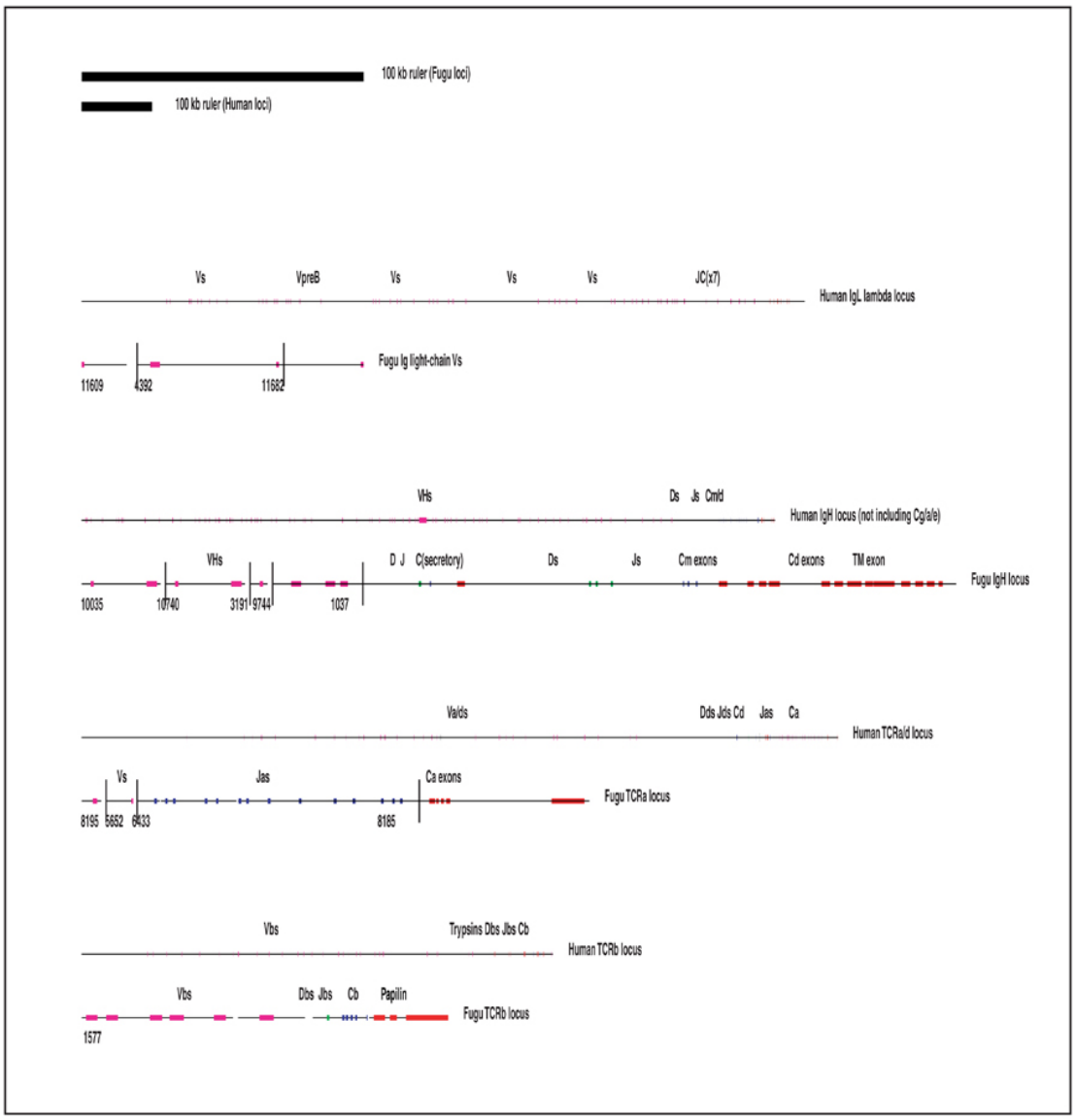

Figure S7. Schematic diagram showing organisation of Fugu and human antigen receptor loci relative to each other. Genes are shown as solid boxes along the sequence locus. Scaffolds corresponding to the encoded genes are identified. The scale bars indicate the sizes of the Fugu and human loci. The human IgL graphic is derived from Kawasaki (43).

Divergence was also noted in many non-receptor systems, including components 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 S4) 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.

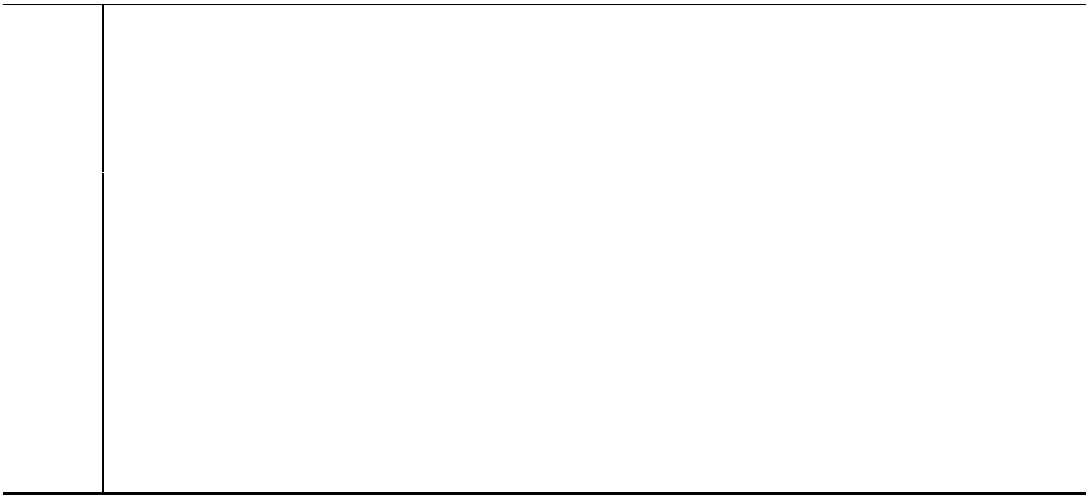

Table S4. Summary of top ranking domains in human and Fugu genes. The most populous Fugu and human protein domains are summarized. Counts represent the number of gene loci encoding a particular domain based on identification with the protein pipeline (see supplemental methods) and assigned to Interpro family numbers. Human data were taken from EnsEMBL version ncbi26. Note that some secondary structures, for example, proline-rich motifs, can be biased by the redundant nature of the signature.

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%.

[redacted]

Table S5. Classification of G-protein coupled receptors (GPCRS) in the 5.7x Fugu rubripes assembly as compared to the human genome.

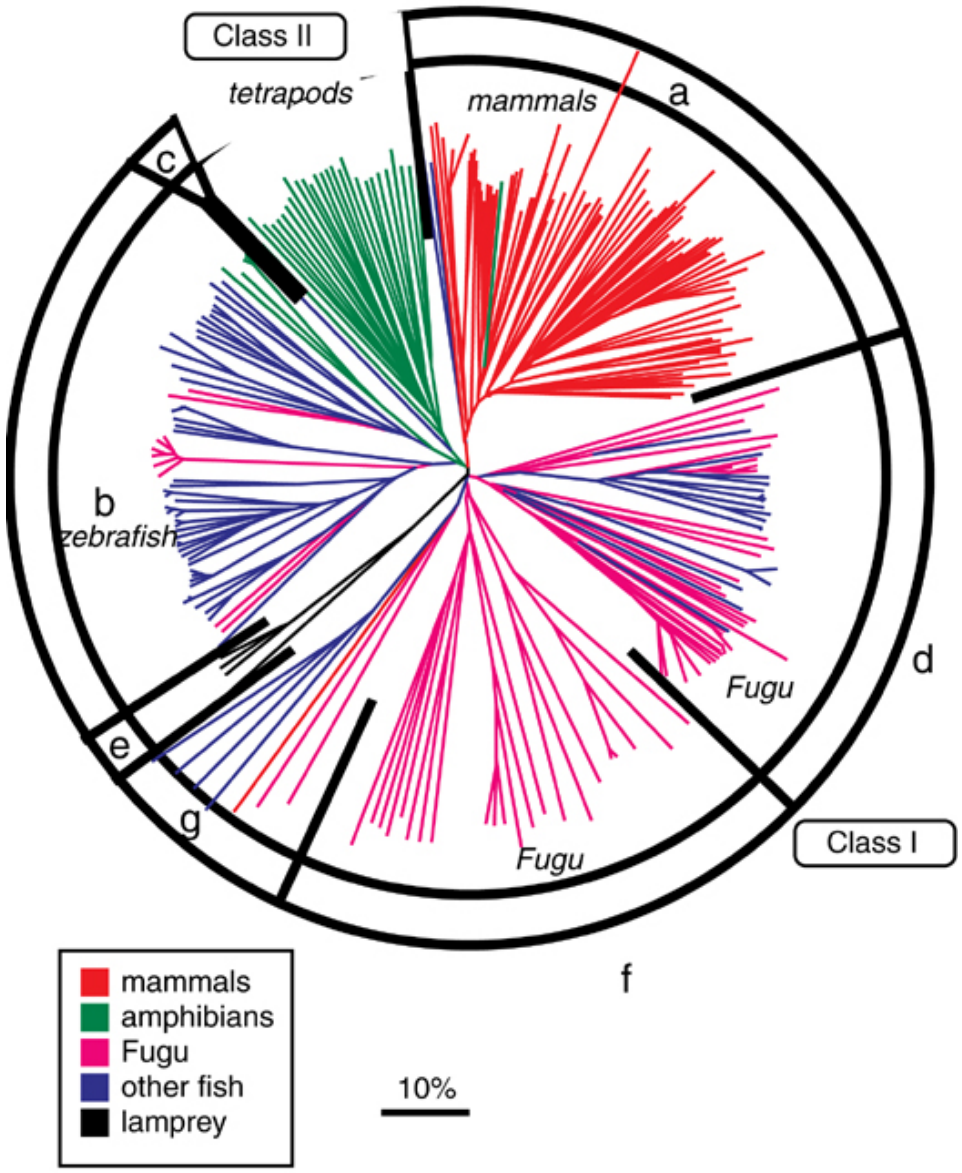

Figure S8. The figure shows a phylogenetic reconstruction of the Fugu rubripes olfactory receptors in relation to other olfactory GPCR sub families. The subgroupings are indicated with letters. We searched the Fugu genome for olfactory receptor (OR) genes as described previously (44) and in methods and recognised 62 OR-like sequences. We performed a phylogenetic reconstruction with an array of previously published ORs from several fish species (catfish, zebrafish, medaka fish, goldfish, salmon), some relevant lamprey sequences, ORs from amphibians (including Xenopus), and all the mammalian class I ("fish-like") ORs detected in mouse, rat and human. The total set includes 265 sequences, the vast majority of which are class I ORs from many species, and a minority are class II ORs from Xenopus. Surprisingly, the human genome has more apparently functional "fish-like" OR genes than the Fugu. The results clearly indicate differential expansion of this gene super family in the different vertebrate lineages. The Fugu genome includes representatives of all subclasses of class I ORs except for that which expanded in tetrapods (subclass a). Two new subclasses (f and g) were defined, one of which (f) appears to be specific to Fugu. These are putative GPCR genes that deviate quite strongly from the typical OR consensus, yet they are most closely related to ORs than to anything else. It is possible that these GPCRs have attained a new function not related to olfaction. Both new subclasses (f and g) appear to be more diverged than the rest, which would be compatible with positive evolution for a new specialization. When studying the mammalian expansion of subclass (a), it was unclear whether this was a lineage-specific expansion from one,or a few, members. Alternatively this expansion could have predated the divergence from fish. Their absence from the Fugu genome shows that this expansion is tetrapod-specific.

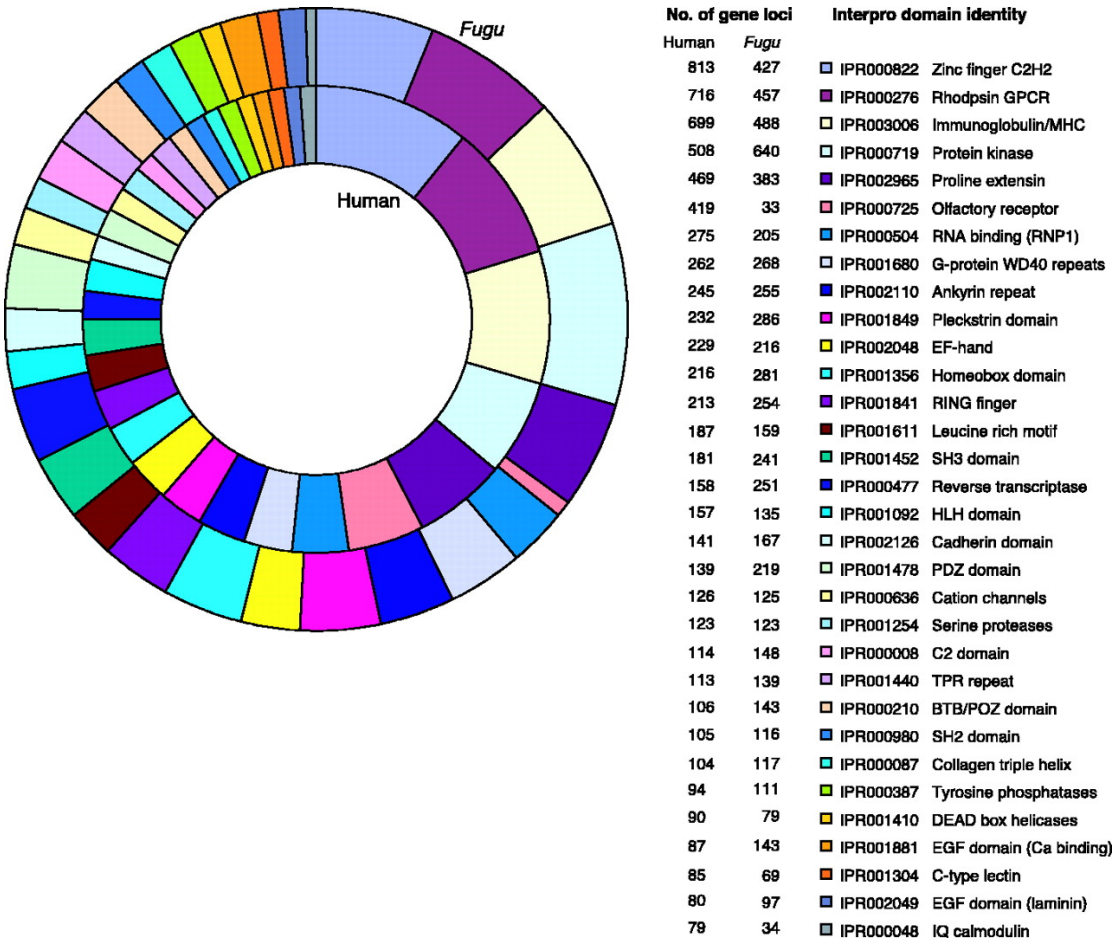

Figure 9. Protein domains of Fugu and human. The schematic shows the number of gene loci with corresponding Interpro domains in Fugu and human for the 32 most populous families.

Ig receptor loci in Fugu

Immunoglobulins have been extensively studied in fish. Completion of the *Fugu* genome sequence provides a view of this system in a completed osteicthyean genome. Previous studies in *Fugu* have shown the presence of Ig-receptor arrays (A3840). The immunoglobulingene superfamily (IgSF) consists of all proteins containing a domain with an immunoglobulin (Ig) fold. These domains were first described in antibodies, but are now known to be spread throughout metazoan proteins. IgSF homologs in bacteria have also been described (A41). The number of IgSF proteins in genomes and the number of different function they serve has increased in step with the organizational complexity of metazoans (table S6). In particular, the number of IgSF

domains and proteins in *H. sapiens* is almost twice that of *F. rubripes*, despite a similar total number of proteins. IgSF domains are highly involved in the development of the nervous system; increased nervous system complexity may explain a significant fraction of the difference in IgSF domains between fish and humans.

Organism	Total Proteins	IgSF Domains	IgSF Proteins
D. melanogaster	13054	128	309
C. elegans	20354	83	423
F. rubripes	34216	488	1336
H. sapiens	35962	970	1899

Table S6. Abundance of Ig domains in metazoa. Immunoglobulin domains found in the proteomes of representative metazoans. Each IgSF protein has one or more IgSF domains. See supplemental text for methods.

Both humans and *Fugu* possess heavy and light-chain Ig loci as well as T-cell receptor (TCR) alpha and beta loci. *Fugu* also possesses multiple "novel immune-type receptor" (NITR) genes, which are not present in humans (A42). We identified NITRs on at least four short scaffolds, suggesting that they may form a tandem array, as has been observed in other fish. We were unable to identify orthologs for TCR gamma or delta sequences in *Fugu*. We used high-sensitivity similarity searches with both variable and constant regions from all sequences present in IMGT. These searches were sensitive enough to produce hits to other immune receptor loci, as well as to MHC chains and other IgSF domains, but did not hit any loci identifiable as TCR gamma or delta.

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 eight-fold 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 non-coding 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 1,000 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 unnnotated 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 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. B. Venkatesh, S. Brenner, 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).

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

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.

Additional References

A1. M. D. Adams et al., Science 287, 2185 (2000).

A2. E. W. Myers et al., Science 287, 2196 (2000).

A3. J. C. Venter et al., Science 291, 1304 (2001).

A4. J. C. Roach, C. Boysen, K. Wang, L. Hood, Genomics 26, 345 (1995).

A5. B. Ewing, P. Green, Genome Res 8, 186 (1998).

A6. J. L. Weber, E. W. Myers, Genome Res 7, 401 (1997).

A7. S. Brenner et al., Nature 366, 265 (1993).

A8. G. Elgar et al., Genome Res 9, 960 (1999).

A9. H. Roest Crollius et al., Nat Genet 25, 235 (2000).

A10. R. Guigo, P. Agarwal, J. F. Abril, M. Burset, J. W. Fickett, Genome Res 10, 1631 (2000).

A11. E. Birney, R. Durbin, Genome Res 10, 547 (2000).

A12. T. Hubbard et al., Nucleic Acids Res 30, 38 (2002).

A13. S. R. Eddy, Bioinformatics 14, 755 (1998).

A14. A. Bateman et al., Nucleic Acids Res 30, 276 (2002).

A15. P. Scordis, D. R. Flower, T. K. Attwood, Bioinformatics 15, 799 (1999).

A16. T. K. Attwood et al., Nucleic Acids Res 28, 225 (2000).

A17. L. Falquet et al., Nucleic Acids Res 30, 235 (2002).

A18. A. Lupas, M. Van Dyke, J. Stock, Science 252, 1162 (1991).

A19. H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Protein Eng 10, 1 (1997).

A20. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, J Mol Biol 305, 567 (2001).

A21. E. S. Lander et al., Nature 409, 860 (2001).

A22. M. Remm, C. E. Storm, E. L. Sonnhammer, J Mol Biol 314, 1041 (2001).

A23. E. G. Shpaer et al., Genomics 38, 179 (1996).

A24. S. A. Teichmann, C. Chothia, J Mol Biol 296, 1367 (2000).

A25. J. C. Wootton, S. Federhen, Methods Enzymol 266, 554 (1996).

A26. M. P. Lefranc, Nucleic Acids Res 29, 207 (2001).

A27. S. F. Altschul et al., Nucleic Acids Res 25, 3389 (1997).

A28. S. Schwartz et al., Genome Res 10, 577 (2000).

A29. V. V. Kapitonov, J. Jurka, Proc Natl Acad Sci U S A 98, 8714 (2001).

A30. N. Shimoda et al., Biochem Biophys Res Commun 220, 226 (1996).

A31. H. R. Crollius et al., Genome Res 10, 939 (2000).

A32. J. N. Volff, C. Korting, K. Sweeney, M. Schartl, Mol Biol Evol 16, 1427 (1999).

A33. J. N. Volff, C. Korting, M. Schartl, Mol Biol Evol 17, 1673 (2000).

A34. A. F. Smit, A. D. Riggs, Proc Natl Acad Sci U S A 93, 1443 (1996).

A35. A. F. Smit, Curr Opin Genet Dev 9, 657 (1999).

A36. K. Kawakami, A. Shima, N. Kawakami, Proc Natl Acad Sci U S A 97, 11403 (2000).

A37. Y. J. Edwards, G. Elgar, M. S. Clark, M. J. Bishop, J Mol Biol 278, 843 (1998).

A38. B. R. Peixoto, S. Brenner, Immunogenetics 51, 443 (2000).

A39. K. Wang et al., Immunogenetics 53, 31 (2001).

A40. N. Miller et al., Immunol Rev 166, 187 (1998).

A41. A. Bateman, S. R. Eddy, C. Chothia, Protein Sci 5, 1939 (1996).

A42. N. A. Hawke, J. A. Yoder, G. W. Litman, Immunogenetics 50, 124 (1999).

A43. K. Kawasaki et al., Genome Res 7, 250 (1997).

A44. G. Glusman et al., Mamm Genome 11, 1016 (2000).