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

Elia Stupka

This work was carried out with support from the Euopean Commission Framework VI grant TRANSCODE (LSHG-CT-2004-511990) as well support from A-STAR Singapore and Temasek Life Sciences Laboratory, Singapore

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

PROEFSCHRIFT

ter verkrijging van de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 11 Mei 2011
klokke 16.15 uur

door
Elia Stupka

door

Geboren te Quartu Sant'Elena, Italy in 1977

PROMOTIE COMISSIE

Promotor

Prof. Dr. J.N. Kok

Co-promotor

Dr. Ir. F.J. Verbeek

Overige Leden

Prof. Dr. H.P. Spaink

Prof. Dr. J. Den Hertog

Dr. P. Sordino (Stazione Zoologica Anton Dohrn, Naples, Italy)

To the two shining stars in my life, Ann and Anais

To my guiding light, my grandmother Giuliana

To my grandfather Aurelio and his free spirit

Chapter 1: Introduction

Introduction

Fish as model organisms

Over the last twenty years fish have rapidly emerged as key model organisms utilized in a variety of research fields. This is owing to their position within the vertebrate subphylum, which provides them with a molecular and body make-up that shares many aspects with that of humans, combined with unparalleled capacity to perform genetic screens and visualize phenotypes, especially in the most widely studied fish species, zebrafish. The latter has enjoyed unsurpassed popularity because of its many enticing features as a model organism such as the ease of maintenance, its transparent embryos which allow powerful visualization of phenotypes, the availability of its genome, as well as a large industry which quickly developed around it to serve the needs of biologists [4-5]. Despite that the emergence of zebrafish was more by accident than by design and it is becoming quickly apparent that many other fish species are equally or even more attractive, depending on the biological question at hand [reviewed in 3]. Until recently it would have been a very large endeavour to begin work on a new model organism species, requiring the co-ordinated action of many laboratories. The development of next-generation sequencing technologies, however, makes it feasible to embark on new species, because information on the genomes, transcriptomes and proteomes can be gained with much less effor than in the past. Thus, for example, species such as *Macropodus opercularis* or *Betta splendens* (which have very compact genomes but display complex behaviour), could be investigated with greater ease, thus connecting complex phenotypes to molecular networks. Although initially great emphasis was placed on mouse and

rat as models for human disease, it is now apparent that fish can be just as good (and sometimes better) models for human disease. Zebrafish is now a well-accepted model organism for the study of complex diseases such as cancer [7], and traits such as ageing [8].

Genome sequencing and assembly

Over 40 years ago the first sequencing was achieved using the Sanger method to allow the deciphering of the sequence of a virus in the 1970s, and later allowing cloning and sequencing of human genes in subsequent years. The human genome project spurred further automation of the same process, allowing (over several years and using hundreds of millions of dollars), the sequencing of the human genome by using a BAC cloning approach (in the publicly funded project) as well as a shotgun approach (in the privately funded Celera project) using long (>500bps) high quality sequence reads. A radical step forward introduced in recent years was the development of next-generation sequencing technologies such as those from Roche 454, Illumina Solexa and ABI SOLID, which now allow a single laboratory on a single machine to obtain 300Gbs of sequence in 10 days from shorter lower quality sequence reads (up to 150bps with current Illumina technology). The data produced by this type of sequencers generates new methodological challenges in genome assembly, which, in turn, have recently pushed the development of new algorithms (discussed in depth in chapter 5 and 6).

Fish genomes

The sequencing and assembly of several fish genomes has greatly enhanced the potential of these organisms, both owing to more accurate identification of

important human orthologs and because they have enabled the discovery of other important vertebrate functional elements of the genome, beyond characterized protein-coding genes. The characteristics of fish genomes had been studied in depth long before genome sequencing was even conceivable. Extensive work by R Hinergardner (1-2) based on simple fluorometric methods had provided genome size estimates for over 200 species of fish, both teleosts and non-teleosts, providing an in-depth investigation of genome sizes throughout the evolutionary branches of this very diverse group. His studies were able to show that more evolved, specialized fishes tended to have smaller genome sizes, and that teleosts have smaller genomes than non-teleost fishes. It is based also on these results that a preliminary characterization was made by in the early 1990s by Nobel Laureate Sydney Brenner of the pufferfish genome, showing that it was likely to be one of the most compact model vertebrate genomes which could be studied [9]. Eventually five years after this initial characterization the pufferfish genome was indeed the first fish genome (and second vertebrate genome after the human genome) to be sequenced, assembled and annotated in our lab[10]. This pivotal study was followed by two more fish genomes, a very close relative of Fugu, *Tetraodon nigroviridis* [11], and a freshwater teleost, medaka (*Oryzias latipes*) [12]. With the advent of next-generation sequencing technologies dozens if not hundreds of fish genomes are now either planned for sequencing or being sequenced already.

Comparative Genomics

The ability to obtain fairly complete and accurate genome sequences for several fish species has allowed the emergence of the field of comparative genomics, i.e. the alignment and comparison of genome sequences and genome structure from

different species. The available genomes allowed comparisons on both shorter evolutionary distances (such as 20MYS between Tetraodon and Fugu), intermediate distances (such as 75MYS between Fugu and Medaka, and 100MYS between Zebrafish and Medaka) and long evolutionary distances (such as 450MYS between human and Fugu). It quickly became apparent that comparative genomics in general, and the Fugu genome in particular were a very powerful tool to detect non-genic functional elements in the genome, such as regulatory elements, which were conserved across the vertebrate lineage. This had been shown much earlier on a smaller scale in Sidney Brenner's lab [13], but the availability of full genomes brought the entire field to a new scale [reviewed in 14]. The field spurred the development of many novel bioinformatics tools, approaches and databases which further refined and optimized the basic task of aligning sequences to be able to detect and score conserved non-coding sequences to distinguish significant conservation from background noise. A variety of acronyms were created for various "classes" of conserved elements, based on the bioinformatics pipeline utilized to identify them, such as HCNEs [15] identified by using MegaBLAST between the human and Fugu genomes, and SCEs, identified using a more complex pipeline focused on shuffled elements, discussed in depth in this thesis [16]. On a larger scale the comparison of these genomes shed light on the complexities of genome duplication genome re-arrangements during vertebrate evolution, showing clearly that while large blocks of synteny are common in short distance comparisons such as those between the mouse and human genome, they are few and far apart when comparing fish to human [10-12].

Transcriptomics

While other –omics technologies such as transcriptomics using microarrays, have been pervasive in the study of human disease and in studies utilizing mouse models, these have not yet achieved their full potential in studies using fish. For the past ten years this was mainly due partly to the limited genome assembly and annotation of the zebrafish genome as well as to the scarce investment made by companies to produce accurate and complete microarray platforms for fish species. This initially lead groups to resort to cDNA arrays, such as the one we used in a study presented in this thesis [17], although these clearly suffered from incomplete coverage and technological limitations. Eventually commercial microarrays became available and started being used and a microarray-based study [18] is discussed in depth in this thesis. The advent of next-generation sequencing is completely revolutionizing the field, owing to techniques such as RNA-Seq [19], which remove the requirement of accurate a priori annotation of the transcriptome, and thus open the door to complete and highly quantitative measurement of transcripts in any species, even those for which the genome has not been sequenced. As shown in the last chapter of this thesis, combining next-generation sequencing of genomic DNA and RNA-Seq nowadays allows the genomic and transcriptomic exploration of a species for which no genome-wide information was available, such as the common carp.

Organization of the thesis

The results presented in this thesis are based on several publications in international peer-reviewed scientific journals. Below is an overview of the chapters presented in this thesis and their related publications.

Chapter 2 focuses on genome sequencing and annotation. I was privileged and honoured to be part of the team which published the first fish genome, i.e. the *Fugu rubripes* genome, and thus this chapter presents the results from that pivotal study, of which I lead the annotation effort. The chapter focuses on the main features of the Fugu genome, and the first basic comparative analyses which were conducted between the Fugu genome and the human genome. The results were published in the following paper:

- Aparicio S et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 2002;297(5585):1301-10

Chapter 3 focuses on comparative genomics. While working on the Fugu genome I was intrigued by the fact that gene order between mammals and fish had hardly been retained at all. Knowing that regulatory elements usually have even less constraints on their position and orientation I hypothesized that in order to identify a complete set of vertebrate enhancers one would have to develop a methodology that allows for shuffling during evolution to different genomic locations. Based on this hypothesis we developed a pipeline for the detection of over 20,000 SCEs (shuffled conserved elements), which we showed to be functional enhancers. The results were published in the following paper:

- Sanges R. et al. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology* 2006; 7(7):R56

Chapter 4 focuses on the use of transcriptomics technologies in fish to answer biological questions. We focused on the degradation of maternal RNA, using

microarray-based gene expression profiling, which were published in this paper:

- Ferg M. et al. The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. *EMBO J* 2007; 26(17): 3945-3956

Chapter 5 focuses on the assembly of the carp genome and transcriptome from next-generation sequencing data. This is a manuscript under preparation.

Chapter 6 provides a discussion of the results presented, proposes future directions and conclusions. In this chapter a short summary of thesis in Dutch is also provided.

Bibliography

1. Hinegardner R. Evolution of cellular DNA content in teleostean fishes. *Am Naturalist* 1968;102:517–523.
2. Hinegardner R. The cellular DNA content of sharks, rays and some other fishes. *Comp Biochem Physiol B* 1976;55:367–370.
3. Muller F. Comparative Aspects of Alternative Laboratory Fish Models. *Zebrafish* 2005;2(1):47-54
4. Zebrafish—the canonical vertebrate. *Science* 2001;294:1290–1291.
5. Grunwald DJ, Eisen JS. Headwaters of the zebrafish— emergence of a new model vertebrate. *Nat Rev Genet* 2002;3:717–724.
6. Special issue devoted to Medaka, *Mech Dev* 2004;121: 629–637.
7. Cancer genetics and drug discovery in the zebrafish. *Nat Rev Cancer* 2003;3:533–539
8. Gerhard GS, Cheng KC. A call to fins! Zebrafish as a gerontological model. Aging Cell 2002;1:104–111.45
9. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome *Nature* 1993; 366:265 - 268
10. Aparicio S et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 2002;297(5585):1301-10
11. Jaillon O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 2004; 431: 946-957
12. Kasahara M. et al. The medaka draft genome and insights into vertebrate genome evolution. Nature 2007; 447:714-719

13. Aparicio S et al. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *PNAS 1995*; 92:1684-1688
14. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 2004;5:456–465
15. Woolfe A et al. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biology* 2005; 3(1):e7
16. Sanges R. et al. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology* 2006; 7(7):R56
17. Yang Li et al. Comparative analysis of the testis and ovary transcriptomes in zebrafish by combining experimental and computational tools. *Comparative and Functional Genomics* 2004; 5:403-418
18. Ferg M. et al. The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. *EMBO J* 2007; 26(17): 3945-3956
19. Wang Z et al. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009 10(1):57-63
20. Yamamoto Y, Stock DW, Jeffery WR. Hedgehog signaling controls eye degeneration in blind cavefish. *Nature* 2004; 431:844–847
21. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 2004; 428:717-723

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.

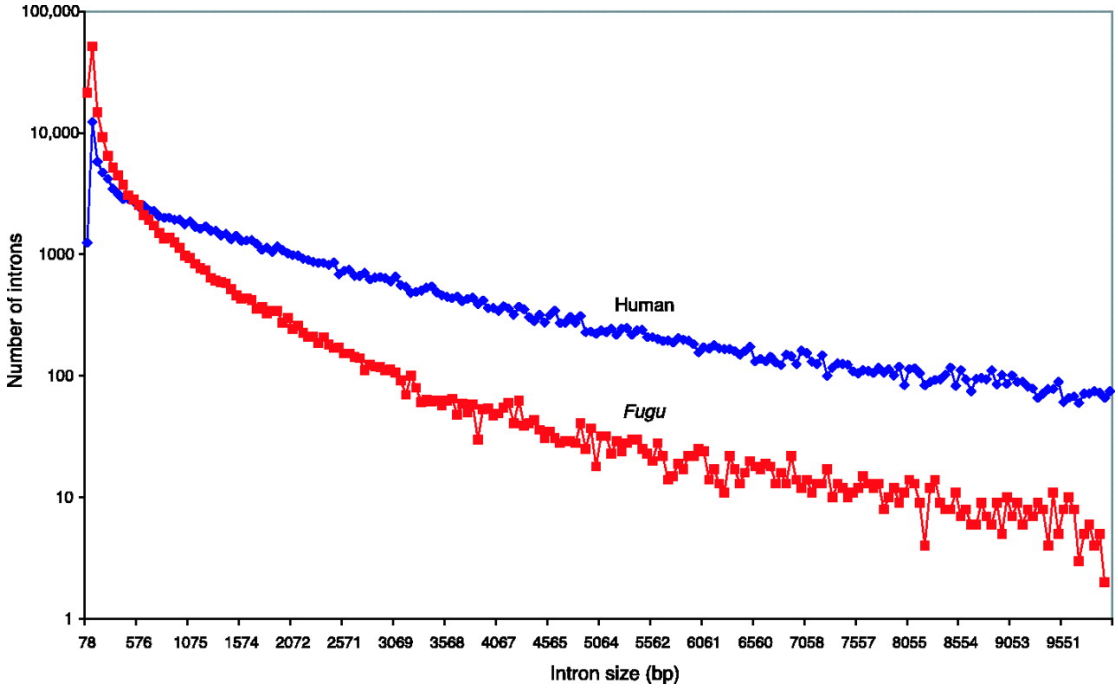


Figure 2. Comparative frequency distribution of intron sizes in Fugu and human.

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.

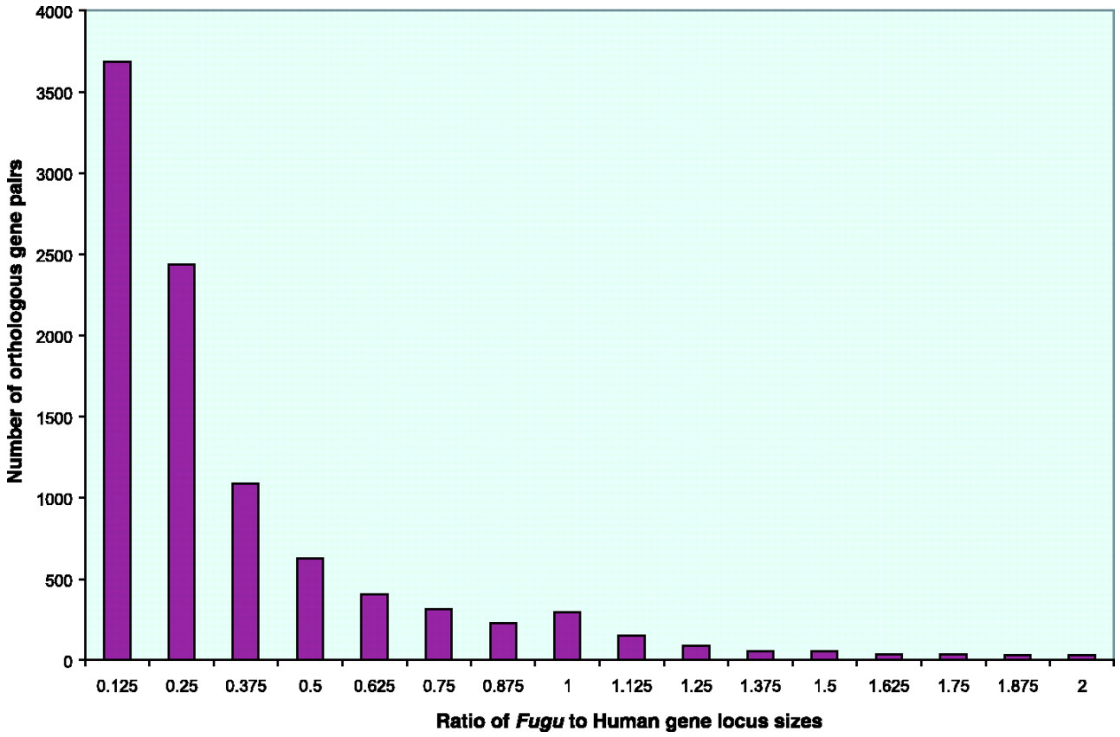


Figure 3. Distribution of ratios for gene locus sizes of putative Fugu-human orthologous pairs. Putative Fugu-human orthologous gene pairings were determined as described in supplemental methods relating to conservation of synteny.

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.

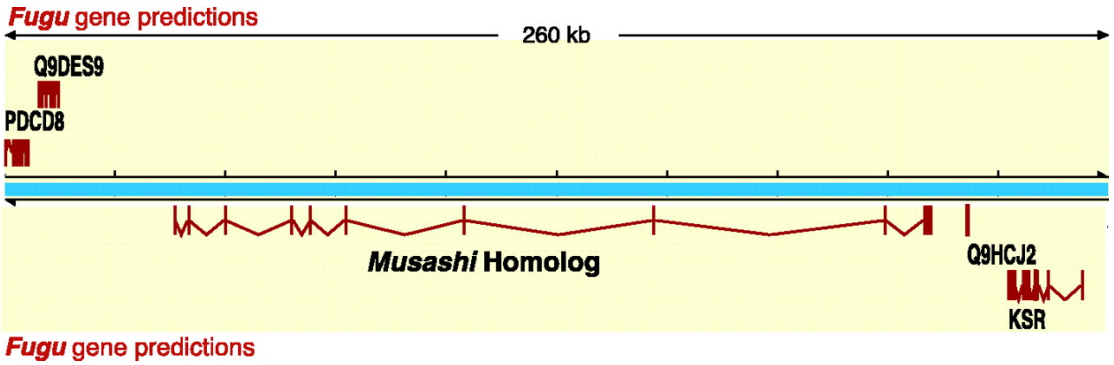


Figure 4. Organization of a giant gene locus in the compact genome of Fugu. The schematic shows a region of scaffold_1 with Fugu gene predictions shown in brown. Exons are represented by vertical brown lines. Introns are show as v-shaped brown lines between exons.

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).

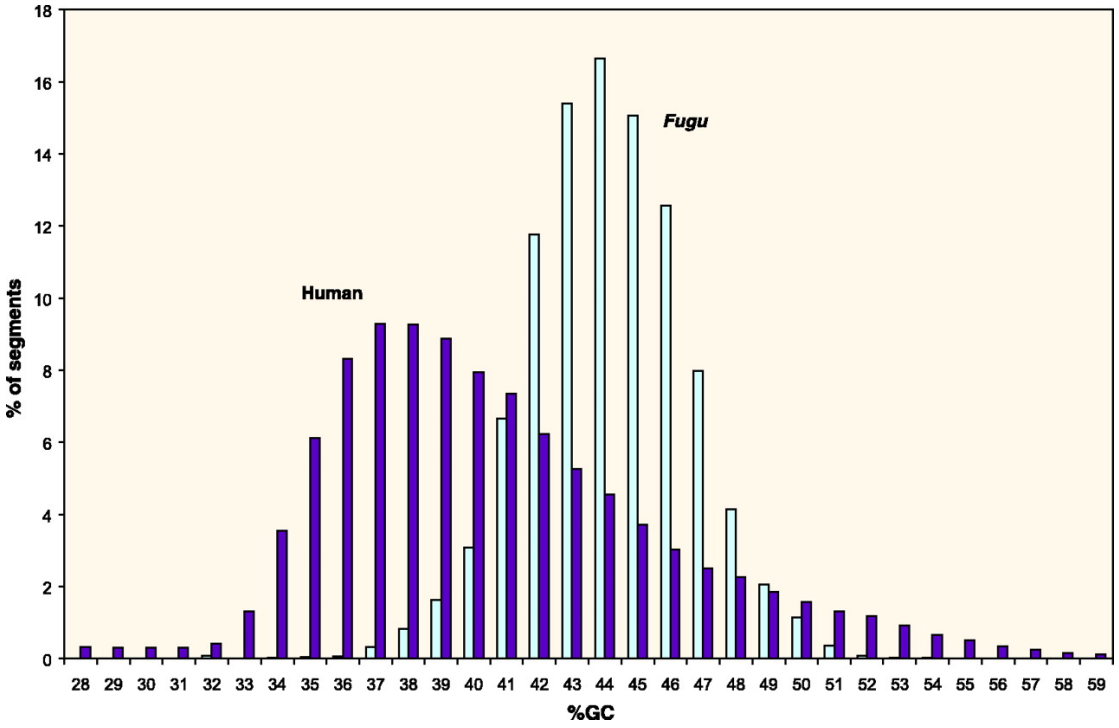


Figure 5. Distribution of GC content in the Fugu and human genomes. Sliding windows of 50 kb were used; similar conclusions were derived with windows of 25 and 100 kb (not shown).

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

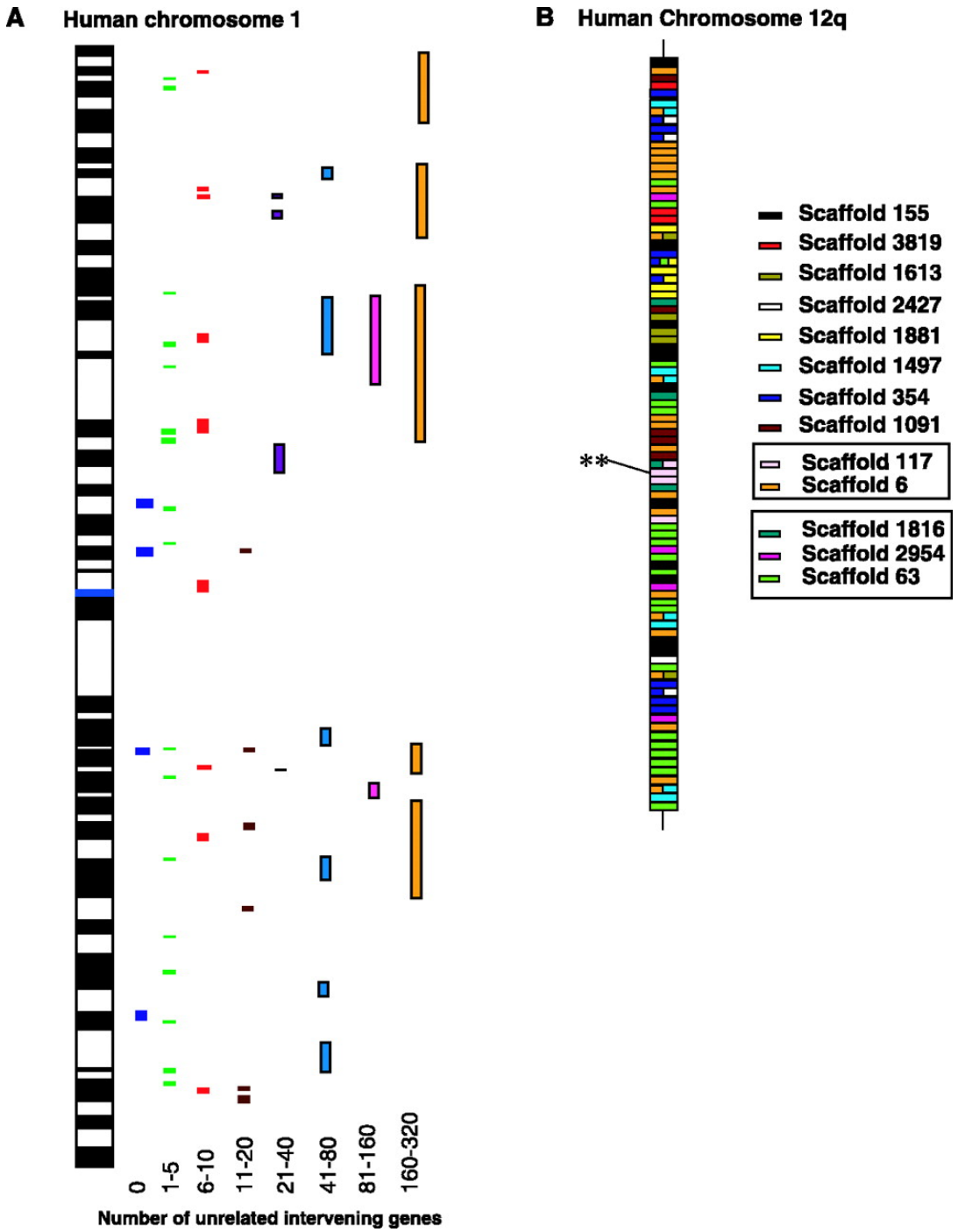


Figure 6. Conserved segments in the Fugu and human genomes. (A**) examines the distribution of conserved segments with different densities of unrelated intervening genes in discrete intervals. Each vertical track of colored boxes represents clusters of genes from *Fugu* scaffolds that map to human. The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

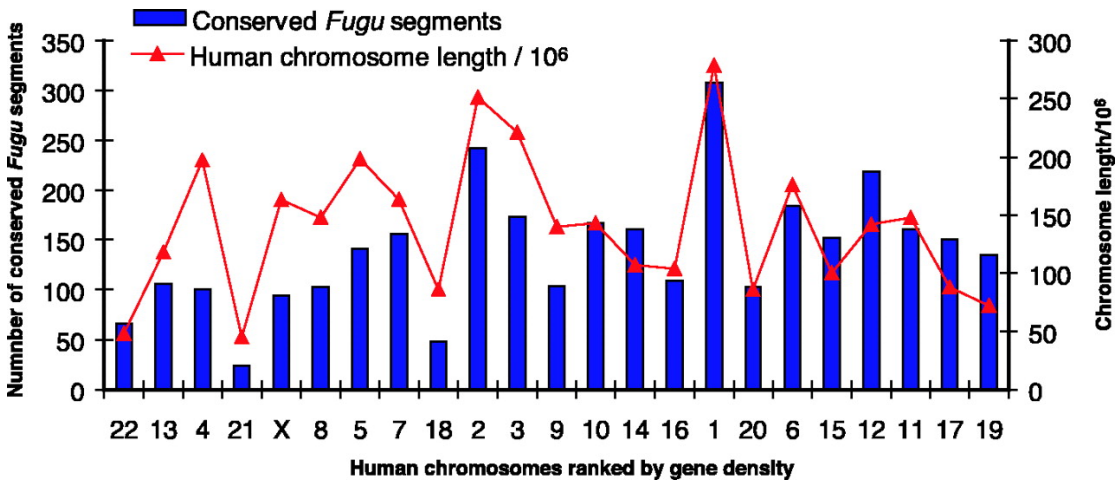


Figure 7. Distribution of conserved segments of Fugu on human chromosomes ranked by gene density. The figure shows the relation between the number of conserved segments of Fugu on human chromosomes, the length of human chromosomes, and their gene density. Chromosome 22 is the most gene poor, chromosome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

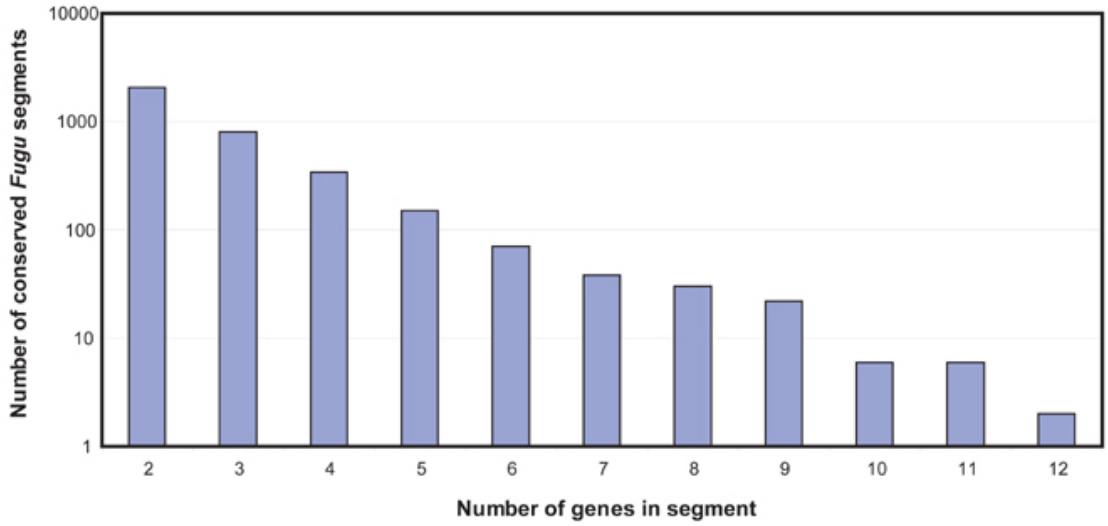


Figure S2. Relationship between the abundance of conserved segments between Fugu and human and the number of conserved genes per segment. This distribution is similar to that obtained for the comparison between human and mouse genomes (21).

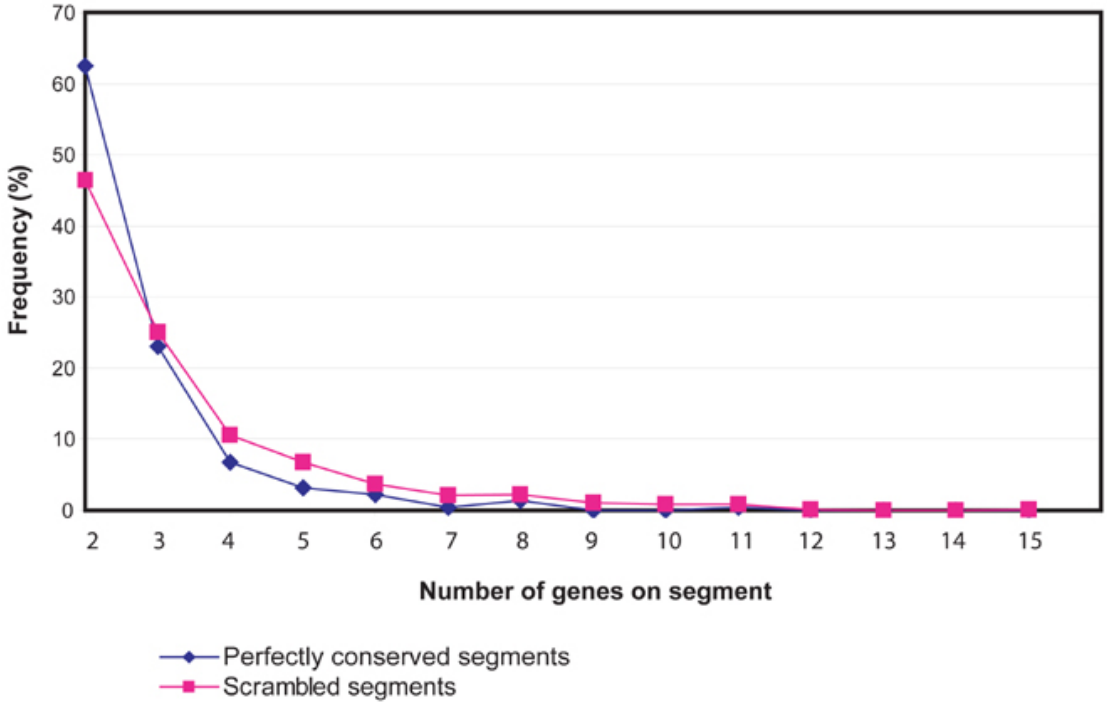


Figure S3. Relationship between the frequency of segments with a given number of genes, for perfectly conserved segments (0 intervening genes) and scrambled segments (1-n intervening genes).

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.

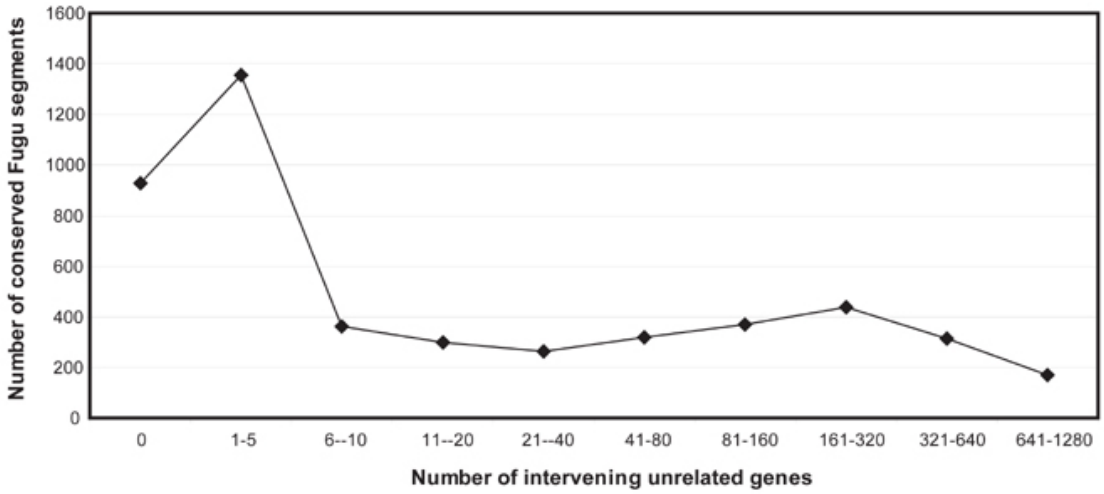


Figure S4. Relationship between the abundance of conserved Fugu-human segments and the density of unrelated intervening genes per segment in discrete intervals.

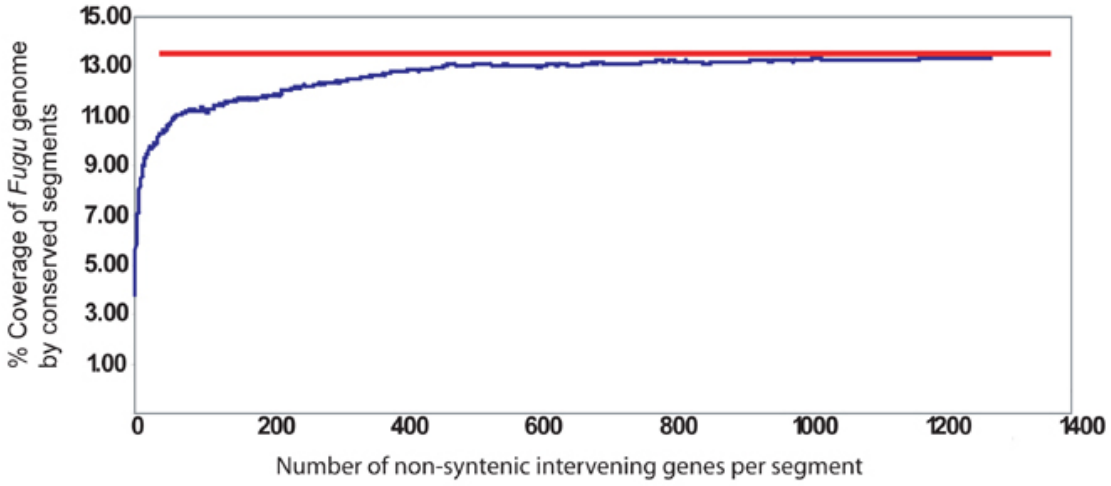


Figure S5. Coverage of the Fugu genome as a cumulative function of segments with increasing numbers of intervening genes.

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).

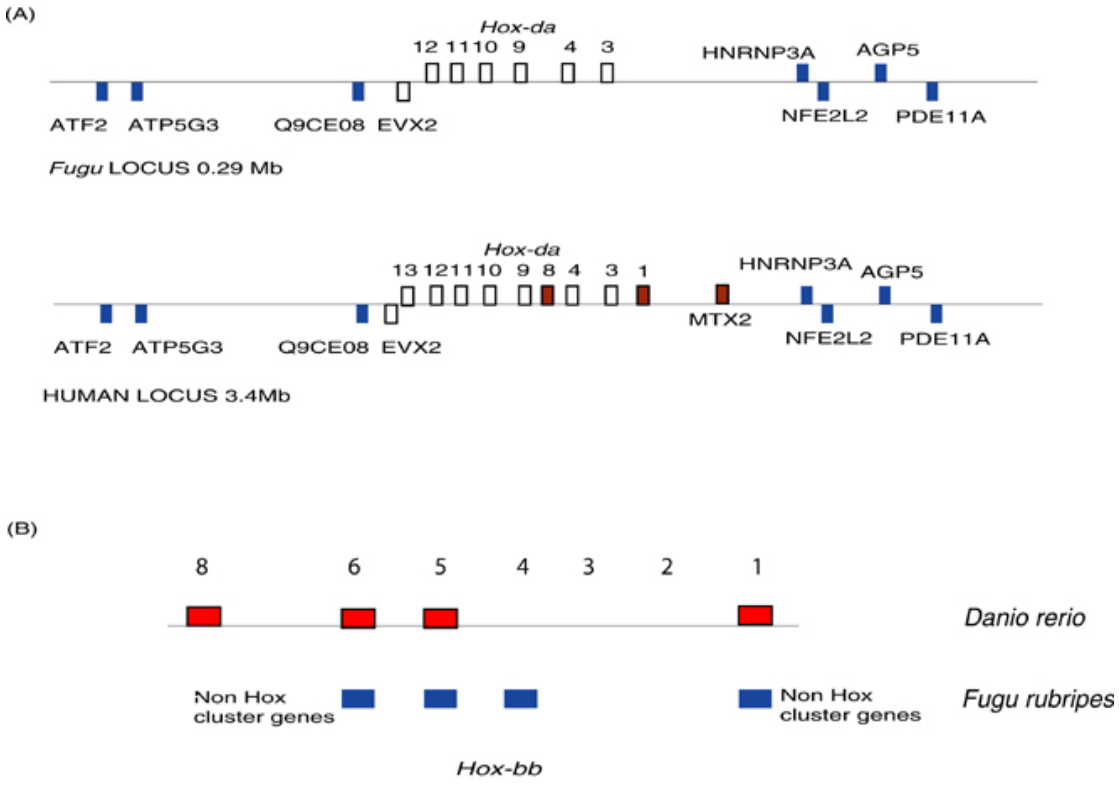


Figure S6. Hox genes. We searched for Hox genes using a homeodomain consensus motif. We were able to identify scaffolds for all of the previously described Hox complexes in Fugu, plus the complexes shown here. A number of small scaffolds contained single Hox genes, some of which could not be confidently 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).

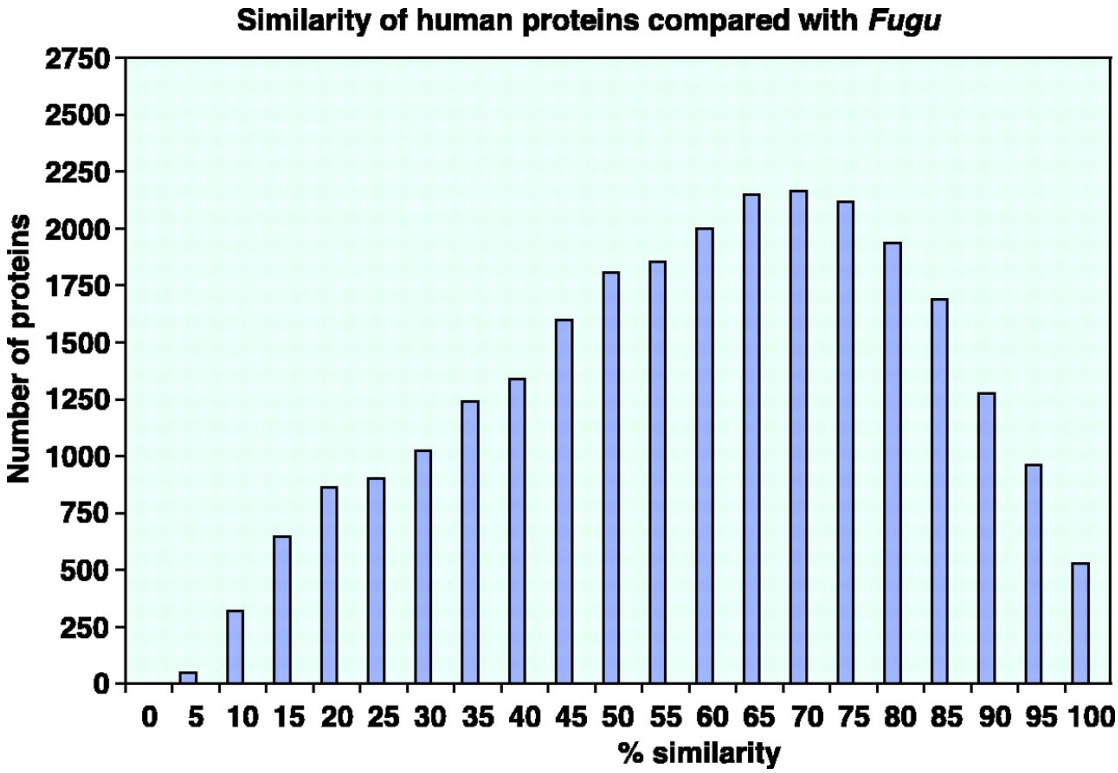


Figure 8. Distribution of protein similarities between *Fugu* and human proteomes. Global similarities were calculated as the sum of similarities in all nonoverlapping HSPs using a BLOSUM62 matrix, over the query (human) sequence length.

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.

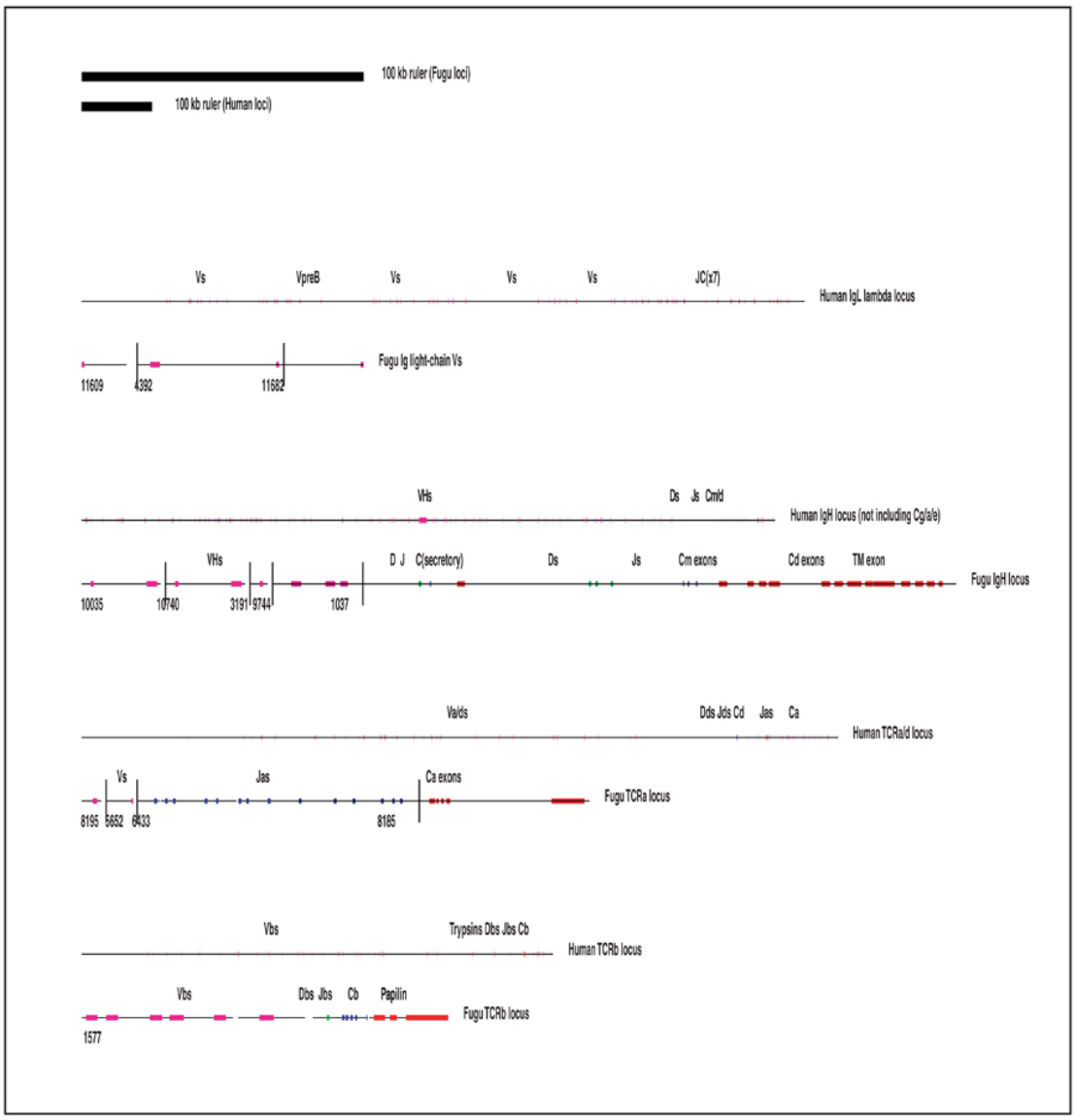


Figure S7. Schematic diagram showing organisation of Fugu and human antigen receptor loci relative to each other. Genes are shown as solid boxes along the sequence locus. Scaffolds corresponding to the encoded genes are identified. The scale bars indicate the sizes of the Fugu and human loci. The human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

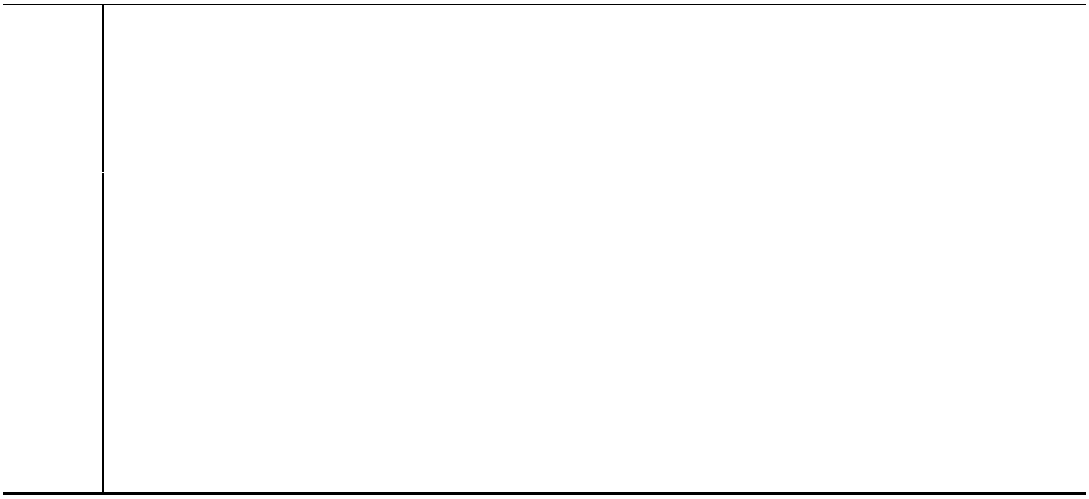


Table S4. Summary of top ranking domains in human and Fugu genes. The most populous Fugu and human protein domains are summarized. Counts represent the number of gene loci encoding a particular domain based on identification with the protein pipeline (see supplemental methods) and assigned to Interpro 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.

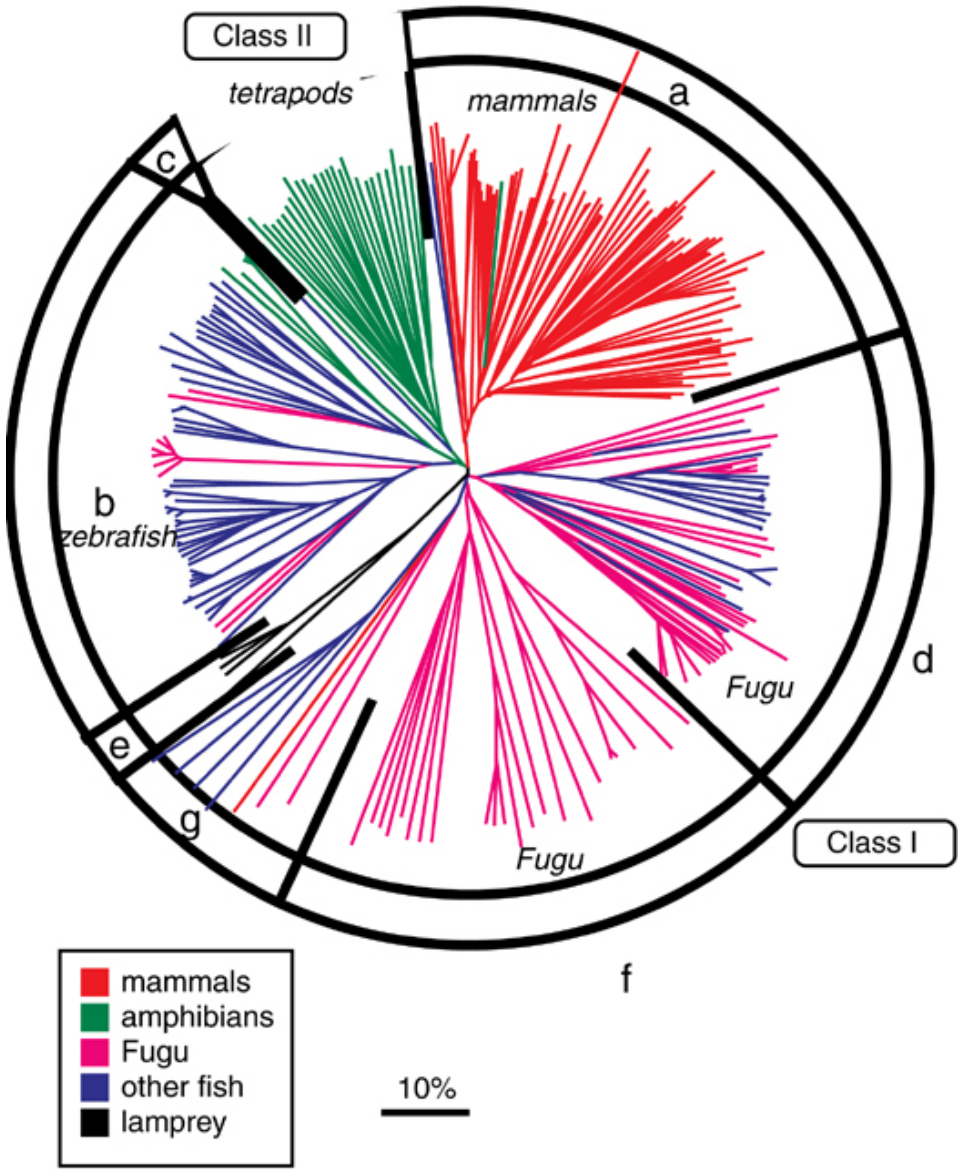


Figure S8. The figure shows a phylogenetic reconstruction of the Fugu rubripes olfactory receptors in relation to other olfactory GPCR sub families. The subgroupings are indicated with letters. We searched the Fugu genome for olfactory receptor (OR) genes as described previously (44) and in methods and 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.

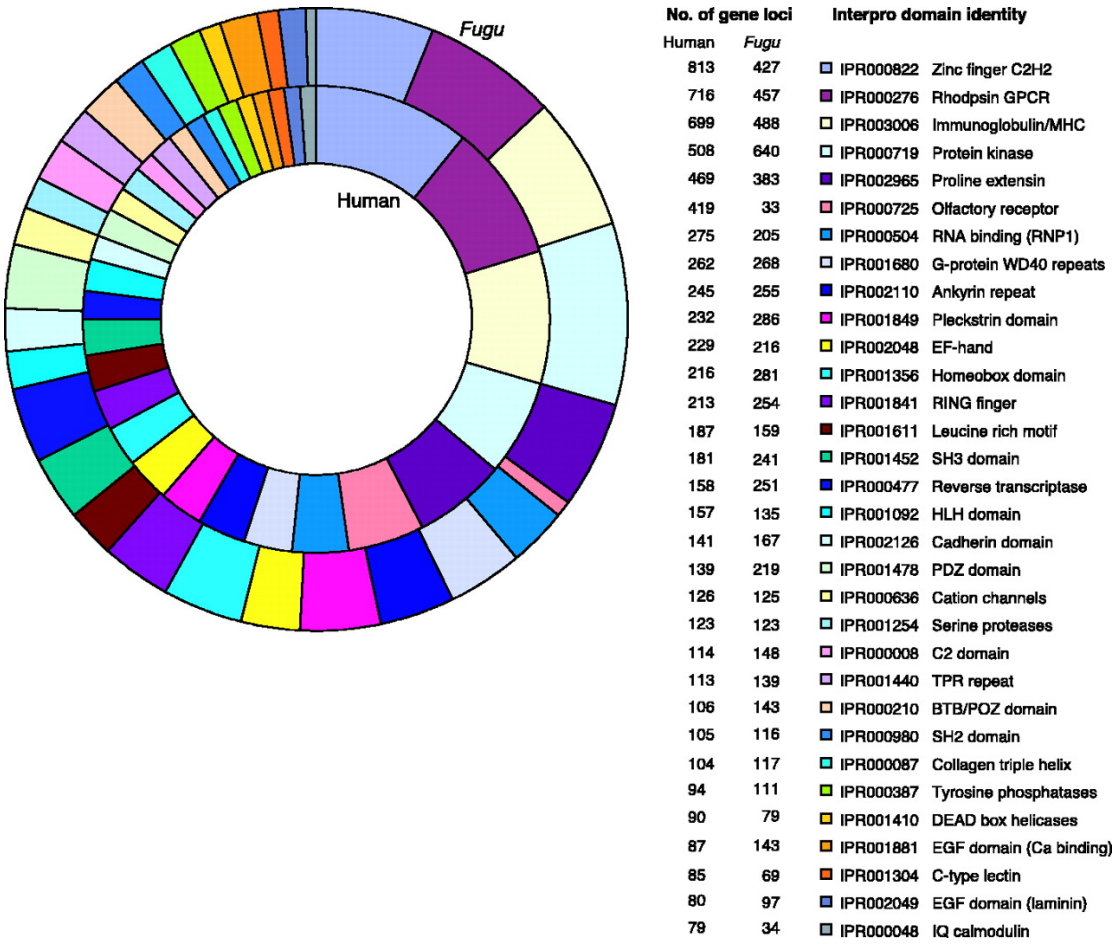


Figure 9. Protein domains of Fugu and human. The schematic shows the number of gene loci with corresponding Interpro domains in Fugu and human for the 32 most populous families.

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).

Chapter 3: Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage

Published in: Genome Biology, 2006, Vol 7:R56

Abstract

Background: All vertebrates share a remarkable degree of similarity in their development as well as in the basic functions of their cells. Despite this, attempts at unearthing genome-wide regulatory elements conserved throughout the vertebrate lineage using BLAST-like approaches have thus far detected noncoding conservation in only a few hundred genes, mostly associated with regulation of transcription and development. We used a unique combination of tools to obtain regional global-local alignments of orthologous loci. This approach takes into account shuffling of regulatory regions that are likely to occur over evolutionary distances greater than those separating mammalian genomes. This approach revealed one order of magnitude more vertebrate conserved elements than was previously reported in over 2,000 genes, including a high number of genes found in the membrane and extracellular regions. Our analysis revealed that 72% of the elements identified have undergone shuffling. We tested the ability of the elements identified to enhance transcription in zebrafish embryos and compared their activity with a set of control fragments. We found that more than 80% of the elements tested were able to enhance transcription significantly, prevalently in a tissue- restricted manner corresponding to the expression domain of the neighboring gene. Our work elucidates the importance of shuffling in the detection of cis-regulatory elements. It also elucidates how similarities across the vertebrate lineage, which go well beyond development, can be explained not only within the realm of coding genes but also in that of the sequences that ultimately govern their expression.

Introduction

Enhancers are ciscting sequences that increase the utilization and/or specificity of eukaryotic promoters, can function in either orientation, and often act in a distance and position independent manner [1]. The regulatory logic of enhancers is often conserved throughout vertebrates, and their activity relies on sequence modules containing binding sites that are crucial for transcriptional activation. However, recent studies on the cis-regulatory logic of Otx in ascidians pointed out that there can be great plasticity in the arrangement of binding sites within individual functional modules. This degeneracy, combined with the involvement of a few crucial binding sites, is sufficient to explain how the regulatory logic of an enhancer can be retained in the absence of detectable sequence conservation [2]. These observations together with the fact that we are still far from understanding fully the grammar of transcription factor binding sites and their conservation [3] make it difficult to assess the extent of conservation in vertebrate cis-regulatory elements.

Very little is known about the evolutionary mobility of enhancer and promoter elements within the genome as well as within a specific locus. Sporadic studies of selected gene families have addressed questions related to the mobility of regulatory sequences involving promoter shuffling [4] and enhancer shuffling [5]; these describe the gain or loss of individual regulatory elements exchanged between specific genes in a cassette manner [6]. These studies suggested that a wide variety of different regulatory motifs and mutational mechanisms have operated upon non-coding regions over time. These studies, however, were conducted before the advent of large-scale genome sequencing, and thus they

were performed on a scale that would not allow the authors to derive more general conclusions on the mobility and shuffling of regulatory elements.

The basic tenet of comparative genomics is that constraint on functional genomic elements has kept their sequence conserved throughout evolution. The completion of the draft sequence of several mammalian genomes has been an important milestone in the search for conserved sequence elements in noncoding DNA. It has been estimated that the proportion of small segments in the mammalian genome that is under purifying selection within intergenic regions is about 5% and that this proportion is much greater than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions, regulatory elements, non-protein-coding genes, and structural elements) that are under selection for biological functions [7-11]. In order to address this issue, sequence comparisons across longer evolutionary distances and, in particular, with the compact Fugu rubripes genome have been shown to be useful in dissecting the regulatory grammar of genes long before the advent of genome sequencing [12]. More recently, the completion of the draft sequence of several fish genomes has allowed larger scale approaches for the detection of several regulatory conserved noncoding features.

Several studies have addressed the issue of conserved non-coding sequences on a larger scale. A first study on chromosome 21 [13] revealed conserved nongenic sequences (CNGs); these were identified using local sequence alignments between the human and mouse genome of high similarity, which were shown to be untranscribed. A separate study focusing on sequences with 100% identity

[14] revealed the presence of ultraconserved elements (UCEs) on a genome-wide scale, and finally conserved noncoding elements (CNEs) [15] were found by performing local sequence comparisons between the human and fugu genomes showing enhancer activity in zebrafish co-injection assays. Although the CNG study yielded a very large number of elements dispersed across the genome, and bearing no clear relationship to the genes surrounding them, the latter studies (UCEs and CNEs) were almost exclusively associated with genes that have been termed 'trans-dev' (that is, they are involved in developmental processes and/or regulation of transcription).

One of the major drawbacks of current genome-wide studies is that they rely on methods for local alignment, such as BLAST (basic local alignment search tool) [16] and FASTA [17], which were developed when the bulk of available sequences to be aligned were coding. It has been shown that such algorithms are not as efficient in aligning noncoding sequences [18]. To tackle this issue new algorithms and strategies have been developed in order to search for conserved and/or over-represented motifs from sequence alignments, such as the motif conservation score [19], the threaded blockset aligner program [20] and the regulatory potential score [21], as well as phastCons elements and scores [22]. However, all of these rely on a BLAST-like algorithm to produce the initial sequence alignment and are thus subject to some of the sensitivity limitations of this algorithm and do not constitute a major shift in alignment strategy that would model more closely the evolution of regulatory sequences.

Two approaches were recently reported which provide novel alignment strategies: the promoterwise algorithm coupled with 'evolutionary selex' [23]

and the CHAOS (CHAins Of Scores) alignment program [24]. Whereas the former has been used to validate a set of short motifs, which have been shown to be of functional importance, the latter has not been coupled to experimental verification to estimate its potential for the discovery of conserved regulatory sequences. Unlike other fast algorithms for genomic alignment, CHAOS does not depend on long exact matches, it does not require extensive ungapped homology, and it does allow for mismatches within alignment seeds, all of which are important when comparing noncoding regions across distantly related organisms. Thus, CHAOS could be a suitable method for the identification of short conserved regions that have remained functional despite their location having changed during vertebrate evolution. The only method available that attempts to tackle the question of shuffled elements and that makes use of CHAOS is Shuffle-Lagan [25]; however, it has not been used on a genome-wide scale and its ability to detect enhancers has not been verified experimentally.

Until recently our ability to verify the function of sequence elements on a large scale within an in vivo context was strongly limited. This task was eased significantly using co-injection experiments in zebrafish embryos [26], which allows significant scale-up in the quantity of regulatory elements tested; this is fundamental when one is trying to elucidate general principles regarding regulatory elements, the grammar of which still eludes us. The co-injection technique used to test shuffled conserved regions (SCEs) for enhancer activity was previously shown to be a simple way to test cis- acting regulatory elements [15,27,28] and was shown to be an efficient way to test many elements in a relatively short period of time [15].

The analysis described herein attempts to tackle the issue of the extent, mobility, and function of conserved noncoding elements across vertebrate orthologous loci using a unique combination of tools aimed at identifying global-local regionally conserved elements. We first used orthologous loci from four mammalian genomes to extract 'regionally conserved elements' (rCNEs) using MLAGAN [29], and then used CHAOS to verity the extent of conservation of those rCNEs within their orthologous loci within fish genomes. The analysis was conducted annotating the extent of shuffling undergone by the elements identified. Finally, we investigated the activity of rearranged and shuffled elements as enhancer elements in vivo. We found that the inclusion of additional genomes, the use of a combined global-local strategy, and the deployment of a sensitive alignment algorithm such as CHAOS yields an increase of one order of magnitude in the number of potentially functional noncoding elements detected as being conserved across vertebrates. We also found that the majority of these have undergone shuffling and are likely to act as enhancers in vivo, based on the more than 80% rate of functional and tissue-restricted enhancers detected in our zebrafish co-injection study.

Results

The dataset described in this analysis is available on the internet [30] for full download, as well as a searchable site to identify SCEs belonging to individual genes.

Identification of mammalian regionally conserved elements

For each group of orthologous genes global multiple alignments among the human, mouse, rat, and dog loci were performed using MLAGAN [25]. We took

into consideration all genes for which there were predicted othologs within Ensembl [31] in the mouse genome, human genome, and any third mammalian species, which led us to analyze 9,749 groups of orthologous genes (36% of the annotated mouse genes). Most genes (about 88%) were found to be conserved in all four species considered, with only about 12% found in three out of four species (about 6% in each triplet; Figure 1). For each locus we took into account the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the preceding and following gene. This lead us to analyze 37% of the murine genome sequence overall. The alignments were parsed using VISTA (visualizing global DNA sequence alignments of arbitrary length) [32] searching for segments of minimum 100 base pairs (bp) length and 70% identity. We further selected these regions by only taking into account those regions that were found at least in mouse, human, and a third mammalian species and which overlapped by at least 50bp, which resulted in a set of 364,358 rCNEs (Table 1). These were then filtered stringently to distinguish 'genic' from 'nongenic' (see Materials and methods, below). This analysis classified 22.7% of the resulting rCNEs as 'genic', while 281,644 nongenic elements account for about 46 megabases, or 1.77%, of the murine genome.

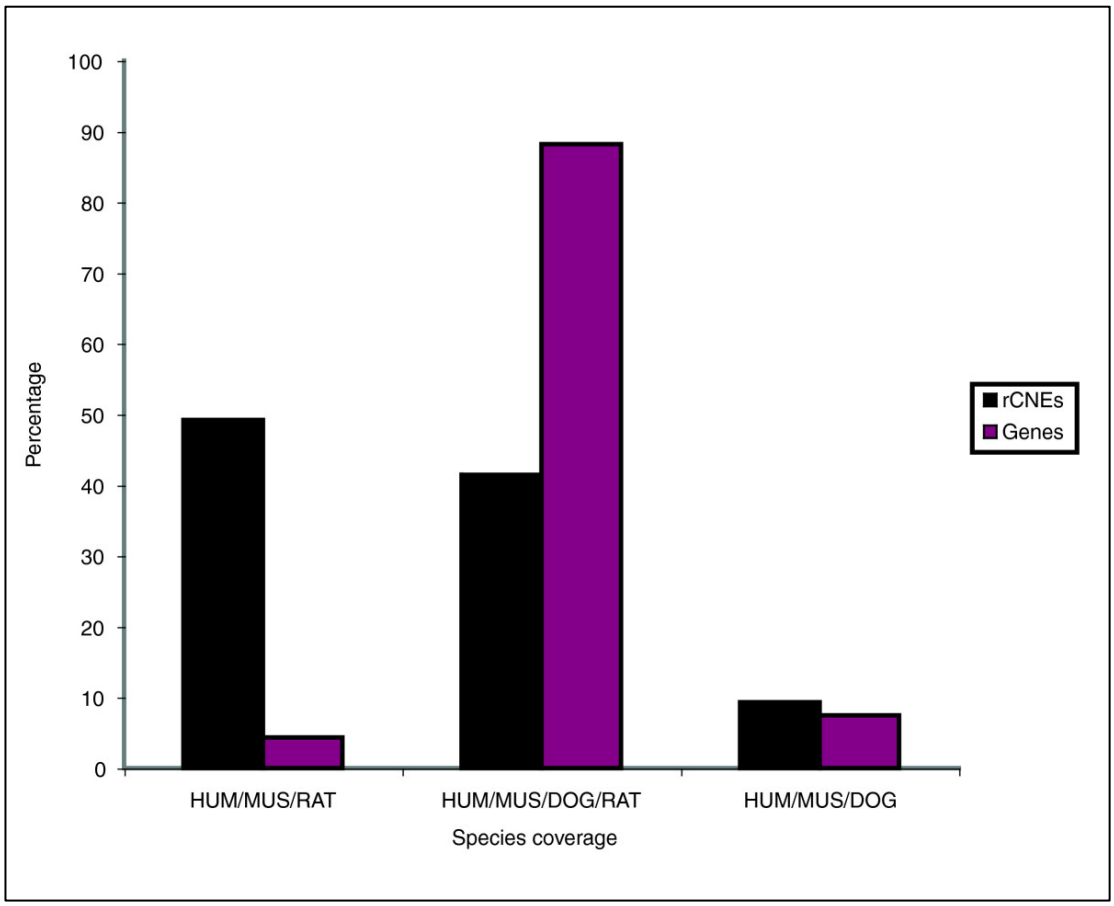


Figure 1 Number of conserved gene loci versus number of rCNEs identified in the mouse, rat, human, and dog genomes. Graph showing the number of rCNEs found conserved in the dog, rat, mouse and human genomes versus the number of genes found conserved across the same genomes. Although almost 90% of the genes can be found in all four genomes, most rCNEs can be found only in three out of four genomes. rCNE, regionally conserved element.

We further annotated mammalian rCNEs based on their position in the mouse genome with respect to the gene locus in order to define whether they were located before the annotated transcription start site (TSS; 'pre-gene'), within the intronic portion of the gene, or posterior to the transcriptional unit ('post-gene'). Approximately 54% of rCNEs were found to fall within intergenic regions, of which 37% were post-gene and 63% pre-gene (Table 1).

rCNE type[a]	Total[b]	Coding[c]	Noncoding[d]
Total[e]	364,358	82,714	281,644
Pre-gene[f]	120,001	23,832	96,169
Intronic[g]	158,722	29,002	129,720
Post-gene[h]	85,521	29,766	55,755

Table 1 Transcription potential, localization, and number of mammalian rCNEs. a)Type of conserved non-coding sequence (rCNE). B)Total number of rCNEs, including genic and nongenic. c)Number of genic rCNEs: overlapping EMBL proteins, ESTs, GenScan predictions, and Ensembl genes. d)Number of nongenic rCNEs: not overlapping EMBL proteins, ESTs, GenScan, and Ensembl genes. e)Total number of rCNEs, including pre-gene, intronic and post-gene. f)Number of pre-gene rCNEs: rCNEs localized before the translation start of the reference gene. g)Number of intronic rCNEs: rCNEs localized within the introns of the reference gene. h)Number of post-gene rCNEs: rCNEs localized after the translation end of the reference gene. EST, expressed sequence tag; rCNE, regionally conserved non-coding element.

Shuffling of conserved elements is a widespread phenomenon

We searched for conservation of rCNEs in teleost genomes using CHAOS [24], selecting regions that presented at least 60% identity over a minimum length of 40 bp as compared with the mouse sequence of the rCNEs. This method allowed us to identify regions that are reversed or moved in the fish locus with respect to the corresponding mammalian locus. For each locus in every species analyzed we took into account the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the preceding and following gene. We defined as SCEs those regions of the mouse genome that were conserved at least in the fugu orthologous locus and filtered out any sequence shorter than 20 bp as a result of the overlap analysis with zebrafish and tetraodon (see Materials and methods, below, for details). Our analysis identified 21,427 nonredundant nongenic SCEs, which were found in about 30% of the genes analyzed (2,911; Table 2). The distribution of their length and percentage identity is shown in Figure 2e,f. The median length and percentage identity (45 bp and 67%, respectively) reflect closely the cut offs provided to CHAOS in the alignment (40 bp and 60% identity), although there is

a significant number of outliers whose length is equal to or greater than 200 bp (223 elements whose maximum length is 669 bp) and whose median percentage identity is 74%. No elements were identified that were completely identical to their mouse counterpart (the maximum percentage identity found was 97%).

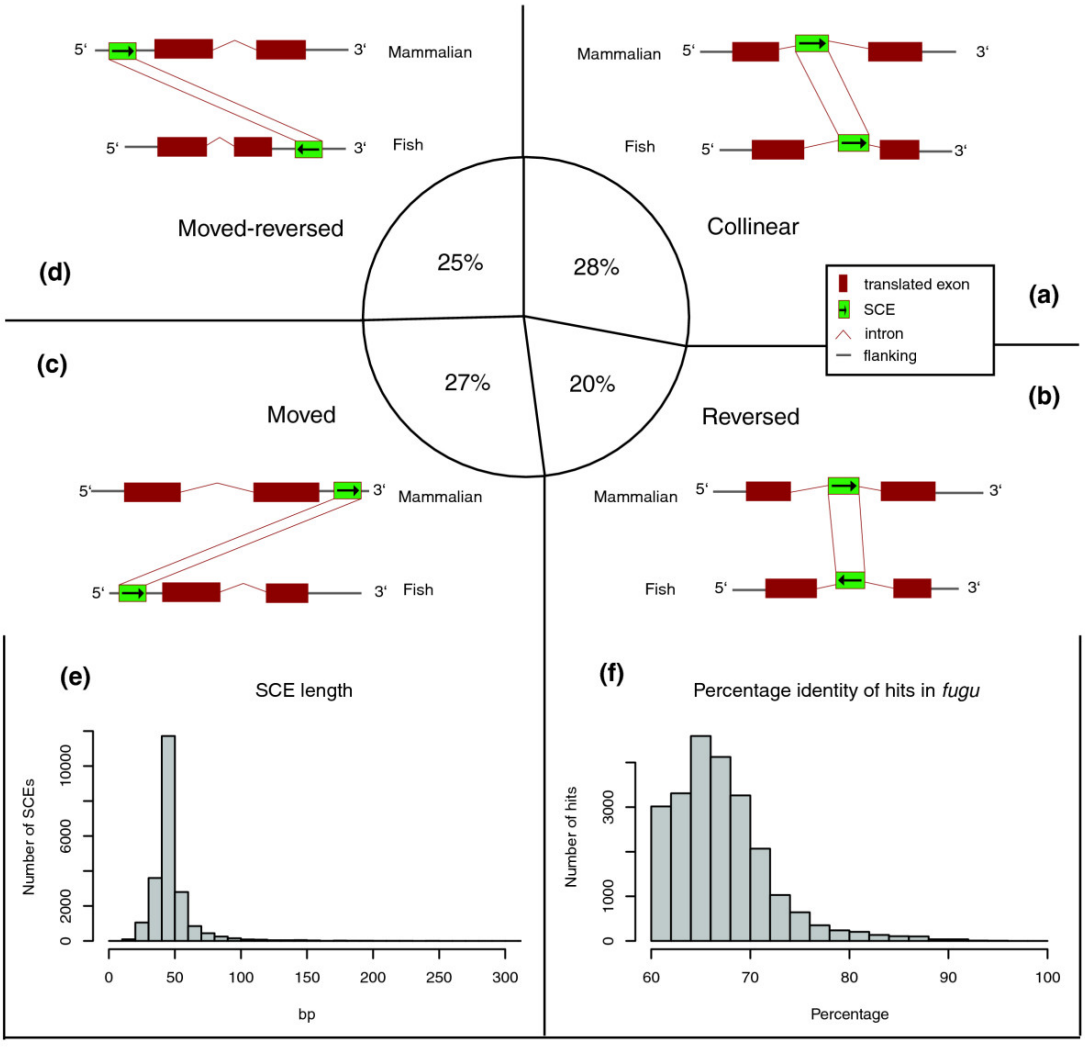


Figure 2 Distribution of length, percentage identity and shuffling categories of SCEs. SCEs were categorized based on their change in location and orientation in Fugu rubripes with respect to their location and orientation in the mouse locus. The entire locus, comprising the entire flanking sequence up to the next upstream and downstream gene was taken into consideration. Definitions of specific classes: (a) collinear SCEs (elements that have not undergone any change in location or orientation within the entire gene locus); (b) reversed SCEs (elements that have changed their orientation in the fish locus with respect to the mouse locus, but have remained in the same portion of the locus); (c) moved SCEs (elements that have moved between the pre-gene, post-gene and intronic portions of the locus); (d) Moved-reversed (elements that have undergone both of the above changes). (e) Frequency distribution of SCE length in base pairs. (f) Frequency distribution of percentage identity of SCE hits in fugu. SCE, shuffled conserved region.

SCE type[a]	Total[b]	Coding[c]	Noncoding[d]
Total[e]	27,196	5,769	21,427
Pre-gene[f]	8,387	1,363	7,024
Intron[g]	11,657	1,838	9,819
Post-gene[h]	7,152	2,568	4,584

Table 2 Transcription potential, localization, and number of vertebrate SCEs. [a]Type of SCE. [b]Total number of SCEs, including genic and nongenic. [c]Number of genic SCEs: overlapping EMBL proteins, ESTs, GenScan predictions, and Ensembl genes. [d]Number of nongenic SCEs: not overlapping EMBL proteins, ESTs, GenScan, and Ensembl genes. [e]Total number of SCEs, including pre-gene, intronic, and post-gene. [f]Number of pre-gene SCEs: SCEs localized before the translation start of the reference gene. [g]Number of intronic SCEs: SCEs localized within the introns of the reference gene. [h]Number of post-gene SCEs: SCEs localized after the translation end of the reference gene. EST, expressed sequence tag; SCE, shuffled conserved element.

We decided to investigate further the extent to which the elements identified, which are still retained within the locus analyzed, have shuffled in terms of relative position and orientation relative to the transcriptional unit, and would thus be missed by a simple regional global alignment (such as MLAGAN). The results of this revealed that only 28% of elements identified have retained the same orientation and the same position with respect to the transcriptional unit taken into account (that is to say, have remained pre-gene, intronic, or post-gene. Labeled as 'collinear'; Figure 2a), whereas others have shifted in terms of orientation ('reversed'; Figure 2b), position ('moved'; Figure 2c), or both ('moved-reversed'; Figure 2d). Thus, almost two-thirds of the SCEs identified would have been missed by a global, albeit regional, alignment approach.

A possible explanation for the large number of non-collinear elements is that they could appear shuffled owing to assembly artifacts. In order to assess whether the large number of elements identified as non-collinear were merely due to assembly artifacts, we analyzed the number of SCEs containing a single hit in fugu and not classified as collinear that also had a match in tetraodon. If the shuffling were merely due to assembly artifacts, then we would expect

approximately half of the non-collinear hits in fugu also to be non-collinear in tetraodon. The results, however, were significantly different, because more than 80% of the elements were not collinear in both species ($P < 2.2 \times e\text{-}16$ obtained by performing a χ^2 comparison between the proportion obtained and the expected 0.5/0.5 proportion). These findings emphasize that shuffling is a mechanism of particular relevance when searching for short, well conserved elements across long evolutionary distances and that its true extent can only be detected by using a sensitive global-local alignment approach, as opposed to a fast genome-wide approach [25].

Two examples of SCEs that were identified in our study are shown in Figure 3. Example A shows the locus of Sema6d, a semaphorin gene that is located in the plasma membrane and is involved in cardiac morphogenesis. This locus represents a conserved element that is found after the transcriptional unit at the 3' end of the gene in all mammals analyzed, whereas it is located upstream in fish genomes and reversed in orientation in the fugu and tetraodon genomes. Example B shows the locus of the tyrosine phosphatase receptor type G protein, a candidate tumor suppressor gene, which has a conserved element in the first intron of all mammalian loci analyzed, which is found in reversed orientation in all fish genomes, downstream of the gene in the fugu and tetraodon genomes,

and in the second intron in the zebrafish genome.

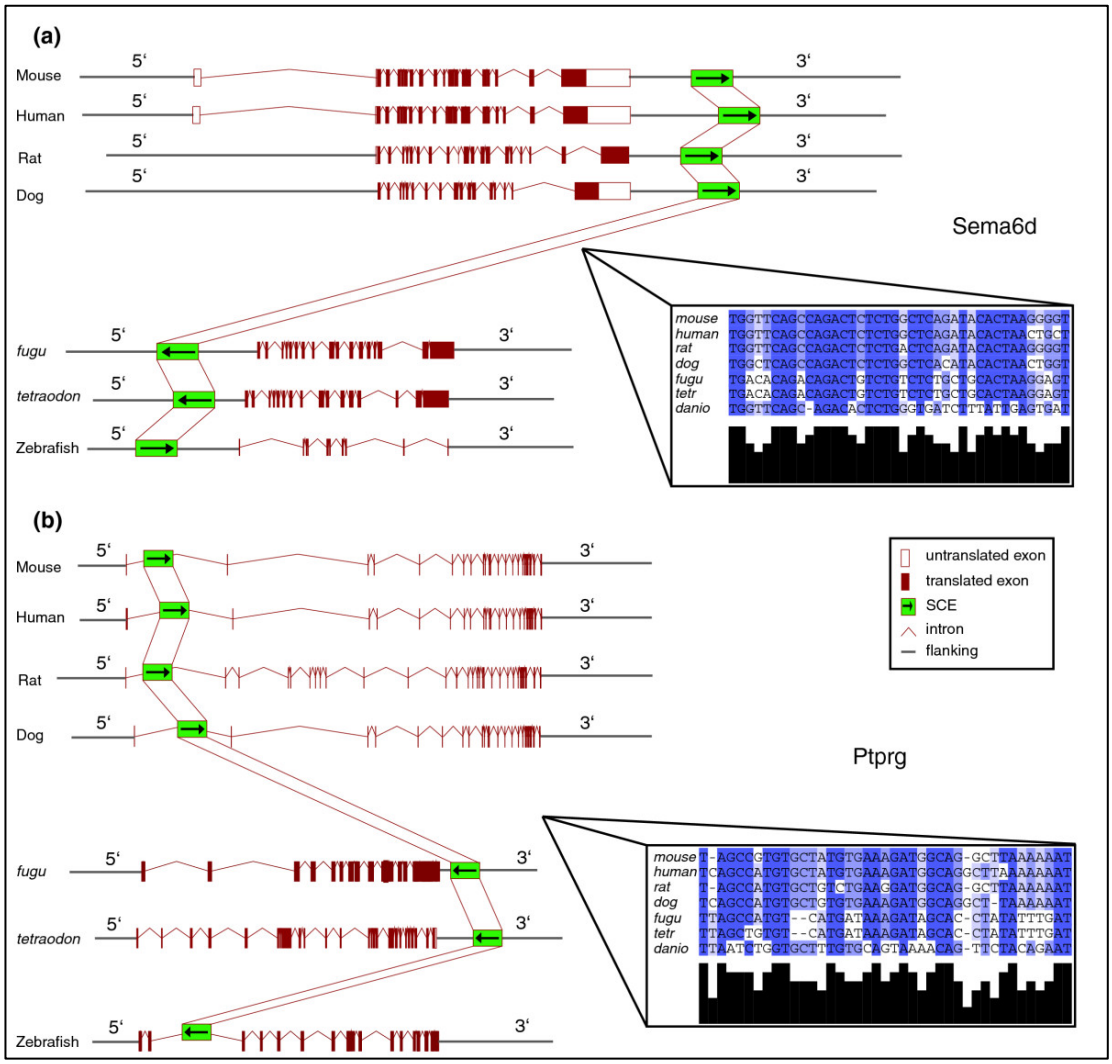


Figure 3 Examples of loci containing shuffled conserved elements. (a) The Sema6d (sema domain, transmembrane domain, and cytoplasmic domain, semaphorin 6D; MGI:2387661) locus contains a post-genic moved-reversed conserved element. The SCE is found downstream from the gene in mammalian loci and upstream of the gene in fish genomes, and in reverse orientation only in the genomes of fugu and tetraodon. (b) the Ptprg (protein tyrosine phosphatase, receptor type G; MGI:97814) locus contains an intronic moved-reversed conserved element. The SCE is found in the first intron of the Ptprg gene in mammalian genomes, downstream of the gene in reverse orientation in fugu and tetraodon, and in the second intron in reverse orientation in zebrafish. Boxes represent the multiple alignments of the SCEs identified. SCE, shuffled conserved region.

Shuffled conserved regions cast a wider net of nongenic conservation across the genome

We analyzed the type of genes that are associated with SCEs by assessing the distribution of Gene Ontology (GO) terms [33] using GOstat [34] (see Materials and methods, below). Although the results indicate significant over-representation of gene classes typical of genes harboring noncoding

conservation ('trans-dev' enrichment) as reported previously, the number of genes within our analysis containing nongenic SCEs (2,911) is approximately an order of magnitude greater than that of the number of genes containing CNEs (330). The overlap between the two datasets is 291 genes, and so almost all (>88%) genes containing SCEs also contain CNEs. A GO analysis comparing genes containing CNEs and those containing SCEs (Figure 4) revealed that there are several GO categories that are significantly under-represented in the CNE dataset as compared with ours. These categories were not seen in the previous analysis because they are not over-represented in our dataset as compared with the entire genome.

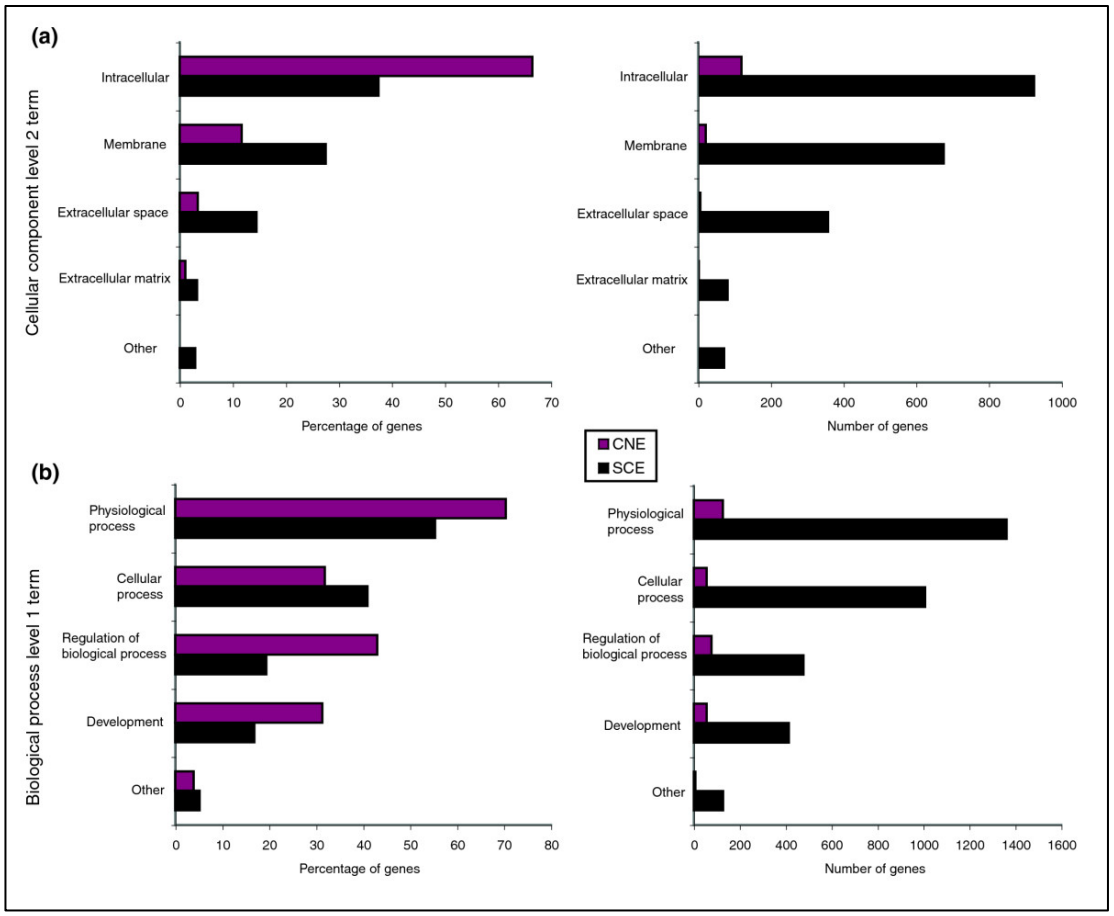


Figure 4 GO Classification of genes harboring CNEs versus genes harboring SCEs. All genes containing CNEs and/or SCEs were analyzed for GO term classification. Genes containing CNEs are shown in red and genes containing SCEs are shown in gray. Plots show differences in absolute numbers as well as

relative percentages. Classification is shown for (a) cellular component and (b) biological process categories. CNE, conserved noncoding element; GO, Gene Ontology; SCE, shuffled conserved region.

The most striking difference is found in the analysis by cellular components; there is an approximate 54-fold enrichment in genes belonging to the extracellular regions that contain SCEs as compared with genes in the same class that contain CNEs. In fact SCEs are present in more than 50% of the genes we were able to classify as belonging to the extracellular matrix and in 35% of those belonging to the extracellular space, whereas CNEs are only found in six and two such genes, respectively. These gene sets differ significantly in both extracellular regions and membrane GO cellular component categories ($P < 0.001$). Enrichments in the order of 10-fold to 13-fold are seen when comparing genes involved in physiological and cellular processes, respectively. For both of these categories our analysis was able to identify SCEs in more than 30% of the genes belonging to this class. The differences, although substantial (about sevenfold) are not as extreme when comparing 'trans-dev' genes (genes categorized as belonging to the 'regulation of biological process' and 'development' using GO) because the CNE dataset has a stronger bias for those genes ($P < 0.001$). Finally, although we identified SCEs in 40% of genes assigned to the 'behavior' class, none of the genes in this class has CNEs. The data thus suggest that there are both quantitative and qualitative differences between the two datasets.

The proximal promoter region is a shuffling 'oasis'

Because a large proportion of our dataset undergoes shuffling, we decided to investigate whether shuffling is a property that is dependent on proximity to the transcriptional unit. To address this question we divided our dataset of nongenic SCEs between collinear (as discussed above) and non-collinear (all other

categories discussed above taken together) elements, and analyzed the distribution of their distances from the TSS (pre-gene set), the intron start (intron start), the intron end (intron-end set) and the 3' end of the transcript (post-gene). This analysis demonstrated that collinear elements were distributed significantly closer to the start and the end of the transcriptional unit compared with non-collinear elements, whereas no differences were observed in terms of proximity to the intron start and intron end (Figure S1).

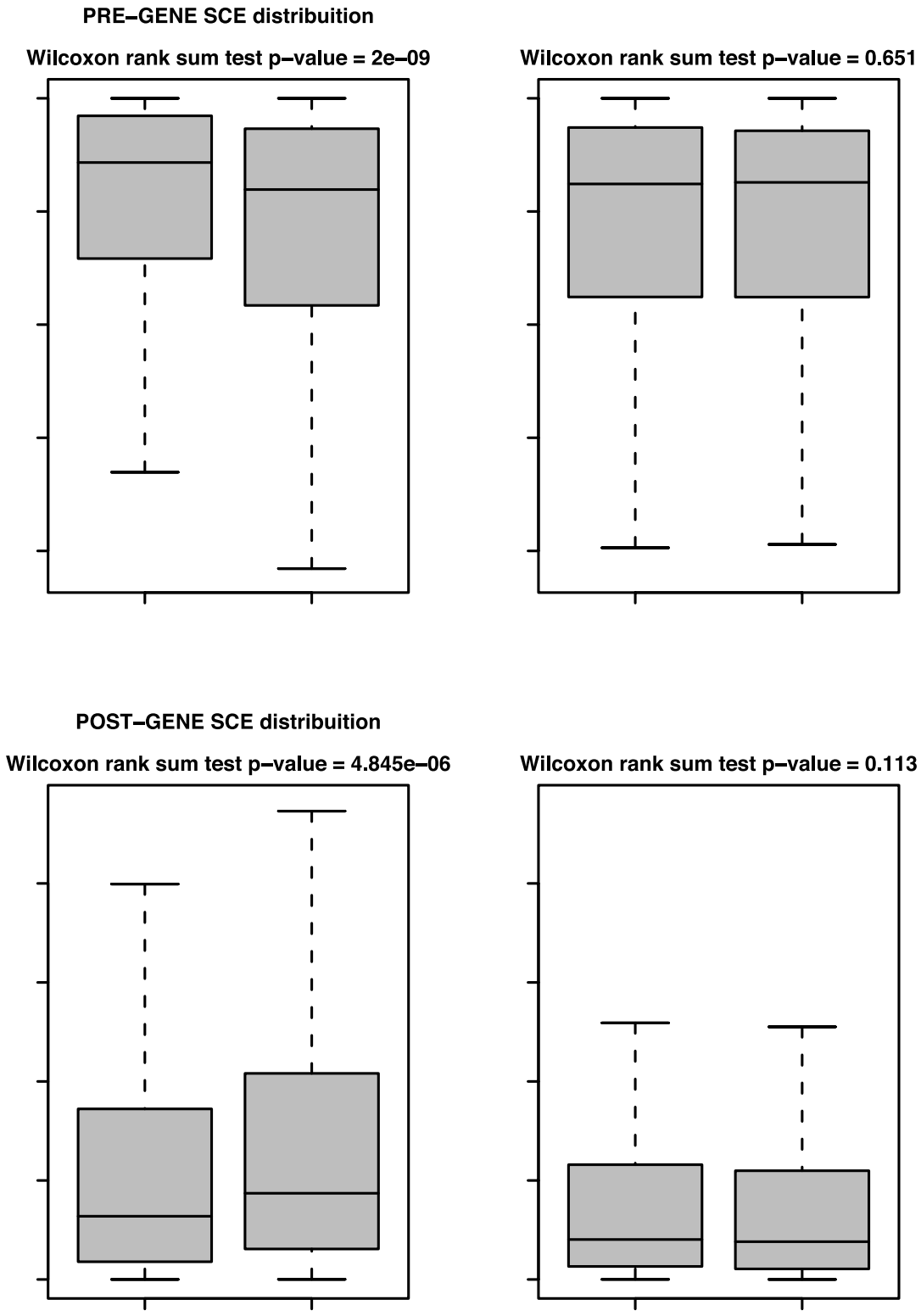


Figure S1 Boxplots comparing the distribution of the distance of collinear versus non-collinear non-genic SCEs from the transcriptional unit

In order to investigate this phenomenon at higher resolution, we subdivided all loci analyzed in our dataset into 1,000 bp windows within the areas, and verified whether the proportion of collinear versus non-collinear elements deviated significantly from the expected proportions in any of these windows (see Materials and methods, below, for details). The results of the analysis are shown in Figure 5. The only window that exhibited a high χ^2 result with significantly less shuffled elements than collinear ones (P = e-08), was the 1,000 bp window immediately upstream of the TSS. No similar results were found in any other 1,000 bp windows across the gene loci analyzed. Similar results were obtained when deploying other window sizes (data not shown). To ascertain whether the result observed was due to annotation problems, we inspected the GO classification of the genes that presented non-genic collinear elements in the 1,000 bp window discussed above and observed significant enrichment ($P < 0.001$) for 'trans-dev' genes, whereas the same test conducted on genic collinear elements in the same window revealed no significant GO enrichment.

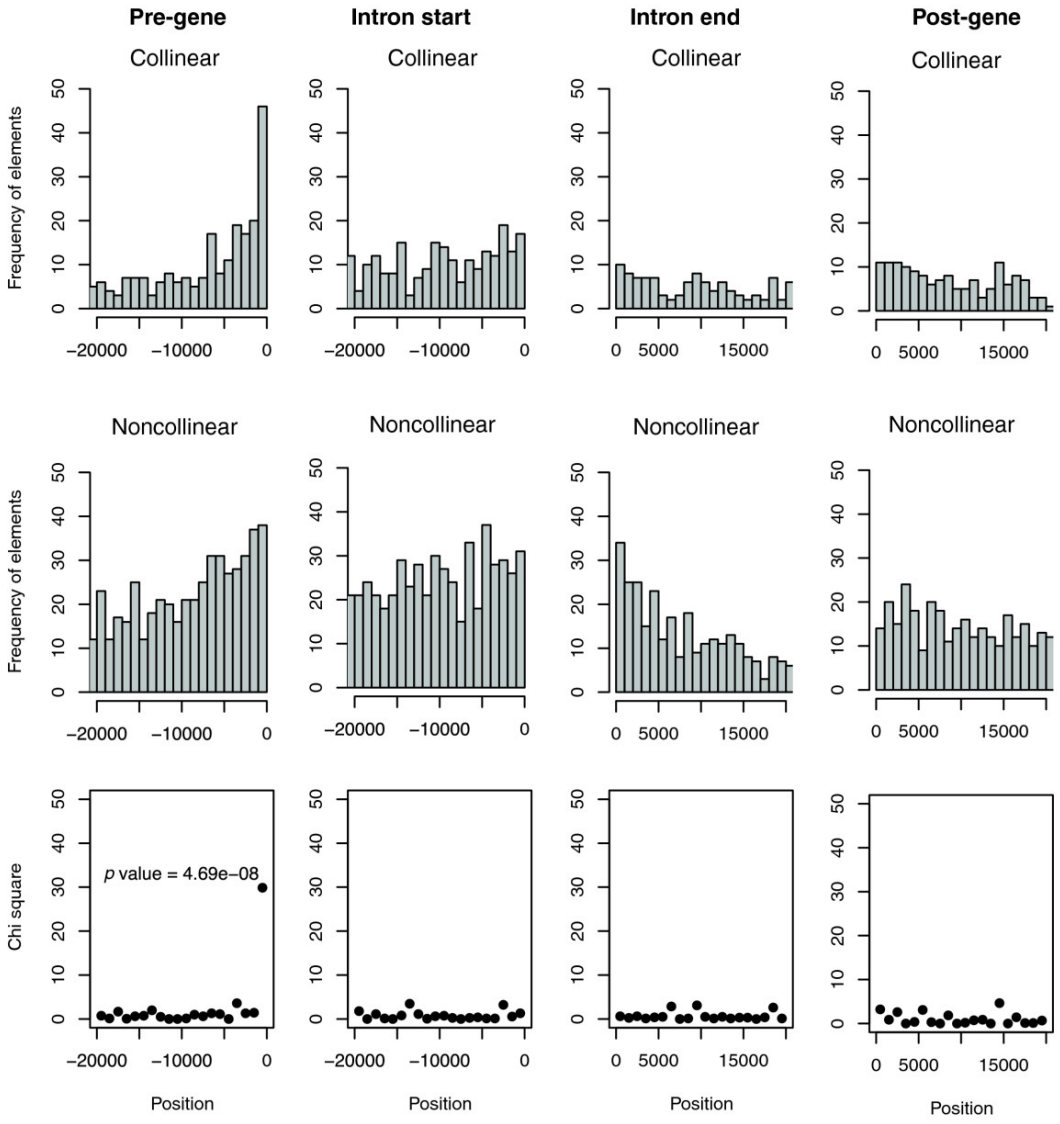


Figure 5 Analysis of SCE shuffling in 1000 bp windows. Each column in the figure shows the analysis of a locus portion (pre-gene, intron-start, intron-end and post-gene) divided into 1000 bp windows. In each column the first graph indicates the number of collinear SCEs identified, the second graph the number of noncollinear SCEs identified, and the third graph the χ2 test used to identify windows that show a significant deviation from the expected proportion of collinear to noncollinear SCEs. The P value is shown for the only window (1000 bp upstream of the transcription start site) that exhibits significant deviation from the expected proportion. bp, base pairs; SCE, shuffled conserved region.

Shuffled conserved regions are able to predict vertebrate enhancers

In order to verify the ability of SCEs to predict functional enhancer elements, we conducted an overlap analysis (see Materials and methods, below) of SCEs with 98 mouse enhancer elements deposited in Genbank. We compared the overlap of SCEs with that of two other datasets that present conservation in fish genomes, namely CNEs and UCEs. The results presented in Figure 6 show that although

CNEs and UCEs are able to detect only one and two known enhancers from our dataset, respectively, SCEs detect 18 of them successfully.

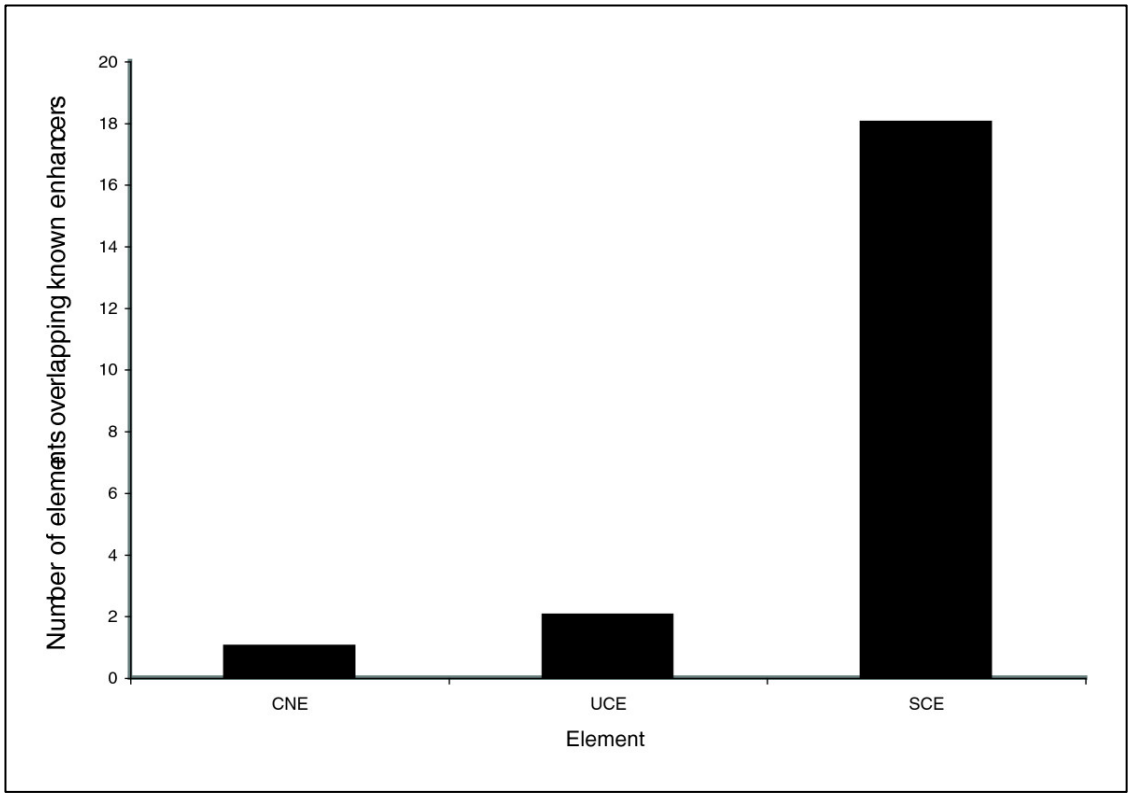


Figure 6 Overlap of known mouse enhancers with conserved elements. All mouse enhancers deposited in GenBank (94) were mapped to the genome and compared with previously published conserved elements (UCEs and CNEs) as well as our own dataset of SCEs to verify their overlap. Only one known mouse enhancer is overlapped by a CNE and two by a UCE, whereas our dataset of SCEs identifies 18 known mouse enhancers as being conserved within fish genomes. CNE, conserved noncoding element; SCE, shuffled conserved region; UCE, ultraconserved element.

Shuffled conserved regions act as enhancers in vivo

In order to validate the cis-regulatory activity of SCEs we chose a subset of SCEs to be tested for in vivo enhancer activity by amplifying them from the fugu genome and co-injecting them in zebrafish embryos with a minimal promoter-reporter construct yielding transient transgenic zebrafish embryos. Twenty-seven SCEs were tested, of which four overlapped known mouse enhancers for which activity had not previously been reported in fish, and the remaining 23 (from 12 genes, of which four were not trans-dev genes, for a total of eight fragments not associated with trans-dev genes) did not overlap any known

feature. As a control set 12 noncoding, non-repeated, and non-conserved fragments were also chosen for co-injection assays, of which nine were from the same genes from which SCEs had been picked and three were from random genes (see Materials and methods, below, for details). Owing to the mosaic expression patterns that are obtained with this technique, results were recorded in two ways: by counting the number of cells stained for X-Gal and recording, where possible, the tissue in which the LacZ-positive cells were found; and by plotting LacZ-positive cells on expression maps that represent a composite overview of the LacZ-positive cells of all the embryos tested. Results of the cell counts are shown in Table 3 and the expression maps are shown in Figure 7. The cell counts were used to define statistically which fragments exhibited tissue-restricted enhancer activity or generalized enhancer activity (see Materials and methods, below). As a positive control a published regulatory element from the shh locus, ar-C [27], was coinjected with the HSP:lacZ fragment. From a total of 27 SCEs, 22 (about 81%) were able to enhance significantly the activity of the HSP:lacZ construct in comparison with the embryos injected with HSP:lacZ only (see Materials and methods, below, for details). Of these, three out of the four tested known mouse enhancers that were found to be conserved in fish were confirmed to act as enhancers in fish. A similar percentage of positive results (82.6%) was obtained excluding these enhancers in the count. The enhancer effect in 20 out of the 22 positive SCEs was not generalized but observed in a tissue-restricted manner.

Gene	Trans dev	Name	SCE bp	SCE Class	ENH	Embryo	Cell	ce/emb	*P* value						
									Muscle	Notochord	CNS	Eye	Ear	Vessels	Other
No	NA	lacZ			Neg control	161	40	0.25							
Shh	Y	ArC			Pos control	96	242	**2.52**		**8.48E-07**					
Shh	Y	12058	45	Rev	Y	139	69	0.5	**6.86E-09**						
Otx2	Y	13988	51	Mov	Y	111	93	0.84	0.6444		**0.006269**	0.5536	0.3155		
Gata3	Y	15402	40	Mre	Y	107	103	**0.96**			0.398	0.5764	0.1906		1
Ets	Y	8744	40	Mov	Y	105	180	**1.57**			**0.002593**			**4.78E-09**	
Ets	Y	8745	46	Mov	Y	133	210	**1.58**			0.1558	0.6015	0.3619	**2.15E-06**	
Ets	Y	8726	41	Mre	Y	159	345	**2.17**			0.05534	0.6136	0.1485	**2.08E-06**	
Ets	Y	8728	48	Mre	Y	149	176	**1.18**			0.0444	0.129	0.07924	**1.31E-05**	
Pax2b	Y	31027	39	Col	Y	149	105	0.7			**0.002374**	0.06327	0.1902		
Pax6a	Y	15696	33	Mov	Y	133	122	**0.92**			**8.21E-06**	0.3343	0.01268		
Pax3	Y	24781	42	Mov	N	124	67	0.54	0.02982		0.5287	1			
Zfpm2	Y	23818	48	Col	Y	140	119	0.85			**1.49E-06**	0.01296	1		
Zfpm2	Y	23838	48	Mre	Y	131	148	**0.98**			**0.0003576**	0.04369	0.1231		
Tmeff2	N	26014	48	Mov	N	164	125	0.76			0.7654	0.02301	0.3371		0.2801
Tmeff2	N	26015	38	Mov	Y	120	159	**1.33**	**0.001035**		0.303	0.2088			
Tmeff2	N	26016	51	Mre	Y	109	148	**1.36**			**0.0006309**	0.0149	0.5862		
Jag1b	Y	16407	37	Col	N	136	98	0.72	1		0.1849	1	1		
Jag1b	Y	16408	55	Col	Y	142	109	0.86			**5.45E-08**	**0.006524**	0.3245		
Jag1b	Y	16409	44	Rev	N	106	54	0.51	1		0.5088	1	0.5058		
Mapkap1	N	17058	37	Mov	Y	143	295	**2.06**	0.6825		0.05292	0.3788	0.6065		1
Mapkap1	N	17059	39	Mov	Y	136	171	**1.26**	0.6686		**0.004037**	0.5973	0.077	0.5197	
Mab21l2	Y	23001	42	Col	Y	142	317	**2.23**			**1.24E-07**	**0.004985**	0.2339		
Mab21l2	Y	23002	37	Mre	Y	155	122	0.79			**7.85E-08**	**0.004138**			
Hmx3	Y	11669	150	Col	Y	165	136	0.82			**0.001029**	0.07062	0.01423		
Lmx1b	Y	17027	300	Col	Y	116	105	0.91			**0.00762**	0.1876	1		
3110004L20Rik	N	5803	45	Mre	N	65	16	0.25	0.2929						1
3110004L20Rik	N	5802	39	Mov	Y	122	320	**2.62**	0.1874	0.01209					
Elmo1	N	6026	45	Rev	Y	103	76	0.74	**0.007132**	0.6848					
Ets	Y	11216	NA	Ctrl	N	104	74	0.71	1						0.6954
Gata3	Y	3255	NA	Ctrl	N	174	110	0.63	0.04481		0.281	0.5739	0.02163		
1300007F04Rik	N	2797	NA	Ctrl	N	157	115	0.73							
Tmeff2	N	198	NA	Ctrl	N	145	23	0.16	0.7448		0.6597		0.3651		
Mab21l2	Y	909	NA	Ctrl	N	165	92	0.56	0.06359		1	1	1		
3110004L20Rik	N	410	NA	Ctrl	N	107	23	0.21							0.01984
Elmo1	N	10157	NA	Ctrl	N	146	38	0.26	0.287	0.8126					
Shh	Y	11271	NA	Ctrl	Y	165	83	0.5	**3.34E-07**		1	1	1		
Impact	Y	5990	NA	Ctrl	N	150	101	0.67	0.6496		0.2754		0.0622		
Ubl7	N	268	NA	Ctrl	Y	117	644	**5.5**	**0.0003325**		**7.15E-11**	0.02555	0.6197		
Lmx1b	Y	11767	NA	Ctrl	N	116	15	0.13	0.2743				0.0707		1
Irx3	Y	5945	NA	Ctrl	N	93	15	0.16	0.03938						

Table 3 Analysis of X-Gal staining in zebrafish embryos co-injected with the HSP promoter and SCEs or control fragments. For each DNA fragment tested the following information is given, from left to right: the gene locus in which the DNA fragment is found; indication about the GO classification of the gene in the 'trans-dev' class (Y = yes, N = no); the identifier given to the SCE or control fragment; the size of the SCE; the class (rev = reversed, mov = moved, mre = moved and reversed, col = collinear, Ctrl = control); summary about the potentially enhancer function of the element (Y = yes, N = no); the number of embryos injected; the total number of cells X-gal-stained; the ratio of stained cells divided by the number of embryos observed (with bold highlighting those with significant generalized enhancer activity); the *P* values for the significance of the number of cells observed in the fragment tested versus the *lacZ:HSP* control for each tissue (bold for *P* values < 0.01; see Materials and methods). See Additional data file 3 for further info on the fragments tested. CNS, central nervous system; SCE, shuffled conserved element.

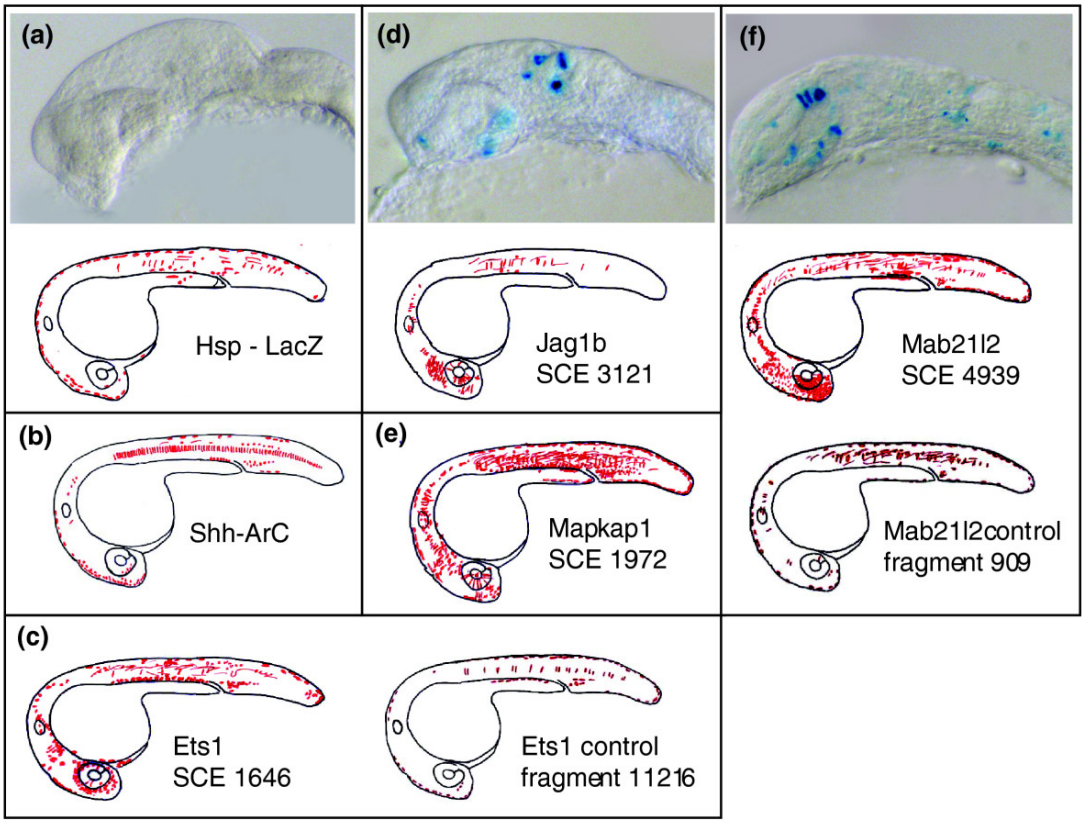


Figure 7 Expression profiles of X-Gal stained embryos. (a-f) Expression profiles of 1-day-old X-Gal stained zebrafish embryos. Each expression map represents a composite overview of the LacZ-positive cells of 65-175 embryos. Gene names and fragment/SCE id are shown. Detailed distribution of X-Gal stained cells in different tissues as well as data for all other fragments are shown in Table 3. Side view of head region of LacZ-stained embryos are shown with anterior to the left. (panel a) HSP-lacZ injected embryo. (d) Embryo co-injected with SCE 3121 associated with Jag1b gene. (f) Embryo co-injected with SCE 4939 associated with Mab21l2 gene. SCE, shuffled conserved region.

The expression patterns obtained in our experiments were compared with expression data retrieved from the Zebrafish Information Network [35,36]. Multiple SCEs found within a single gene locus gave similar tissue-restricted enhancer activity. For example, all four SCEs tested from the ets-1 locus gave expression that was highly specific to the blood precursors (SCE 1646 in Figure 7c). This result is in accordance with reported data, which showed ets-1 expression in the arterial system and venous system. Moreover, both elements tested from the zfpm2 (also described as fog2 [37]) gene gave central nervous system (CNS) specific enhancer activity, which is in accordance with a recent report showing that the expression of both fog2 paralogs is restricted to the

brain [37]. Similarly, elements tested from the mab-21-like genes gave CNS and eye specific enhancer activity (SCE 4939; Figure 7f). This pattern of expression corresponds with the patterns reported in the brain, neurons, and eye [38,39]. The SCEs that were found in the pax6a and hmx3 genes were shown to give CNS specific enhancement, which is in accordance with the reported expression of these genes in the CNS [35]. Finally, SCE 3121 from the gene jag1b gave specific expression in the CNS and in the eye (Figure 7d), which is in partial agreement with reported expression of this gene (expressed in the rostral end of the pronephric duct, nephron primordia, and the region extending from the otic vesicle to the eye [40]).

Novel enhancer functions were also detected for SCEs neighboring lmx1b1, which showed CNS specific activity, and SCEs neighboring four genes not belonging to the trans-dev category, such as mapkap1 (Figure 7e), tmeff2 and 3110004L20Rik (producing proteins integral to the membrane), and elmo1 (associated with the cytoskeleton), which exhibited strong generalized and/or tissue specific activity. No endogenous expression data are available for these genes for comparison. In contrast to the results with SCE elements, only two out of 12 (about 17%) of the genomic control fragment set derived from the same loci of the SCEs exhibited significant enhancement of LacZ activity (Table 3).

Taken together, these data demonstrate that SCEs act as bona fide enhancers that can drive tissue-restricted as well as generalized expression during embryo development.

Discussion

Widespread shuffling of cis-regulatory elements in vertebrates

In this study we demonstrate, using a unique combination of tools aimed at obtaining regional, global-local sensitive alignments applied at the genome level, that the number of conserved non-coding sequences shared between mammalian and fish genomes is at least an order of magnitude higher than was previously proposed and is spread across thousands of genes. In fact, approximately 30% of the genes analyzed presented at least one SCE. Our GO analysis results indicate a 'trans-dev' bias similar to those described in previous studies addressing genes exhibiting noncoding conservation [14,15]. On the other hand, the significant increase in the sheer number of elements identified and in the number of genes exhibiting SCEs enabled us to detect conserved nongenic elements in a third of the genes studied, indicating that conservation of cis-regulatory modules is a widespread phenomenon in vertebrates, and is not limited to a few hundred genes, as suggested by previous studies. The GO analysis also revealed that certain classes of genes, such as those located in the extracellular space and extracellular matrix, exhibit conserved non-coding sequences, which were not identified with previous approaches and indicate that non-coding elements conserved across vertebrates are present in a larger and more diverse set of genes than was previously thought. Although we also observed a larger number of genes involved in cellular and physiological processes, many of them are also assigned to 'trans-dev' categories, and so their involvement in development and regulation of transcription cannot be excluded. Indeed, it is important to note that eight out of the 23 randomly selected fragments were not associated with trans-dev genes by GO classification, and

that six of these fragments exhibited significant enhancer activity in our co-injection assays (Table 3). This confirms that conservation is not an exclusive characteristic of regulatory regions associated with trans-dev genes.

That shuffling plays an important role in the identification of conserved non-coding sequences is illustrated by the fact that 72% of our dataset was observed to be either inverted or moved, or both, in the fish locus with respect to the mouse locus. Assembly artifacts are unlikely to be an important factor in the elements identified as shuffled because they would also affect gene structures and therefore correct gene prediction and ortholog detection, which is at the basis of our dataset. We were reassured about this by our tetraodon-fugu comparison, which indicated that most elements found to be shuffled in one species were also shuffled in the other. A notable exception to the general shuffling bias in the elements found was a 1,000 bp window immediately upstream of the TSS. Taking into account that the proximal promoter region is considered to be approximately -250 bp to +100 bp from the TSS [41], and assuming that TSS annotations in the mouse genes analyzed are precise, this finding suggests that there is a class of enhancer elements that are more constrained in both position and orientation, perhaps working in tight connection to the promoter complex. The fact that the genes containing non-genic collinear elements in this window show the 'trans-dev' bias associated with our overall SCE dataset, as well as with previous analyses of noncoding conservation, reassures us that this result is not a mere product of bad annotation of the first exon in these genes. It is particularly reassuring that performing the same analysis on SCEs found in the same window but classified

as 'genic' (and thus more likely to be real evidence of annotation problems) did not exhibit this bias.

Lack of conservation can also be due to the fact that the evolution of regulatory motifs involves constant de novo creation and destruction of them over time because of their short sequences and plastic nature [42] (for review [43]). The dissection of cis-regulatory elements from different species, however, indicates clearly that there are cases in which although the same transcription factors are involved in the regulation of a gene, all sequences that are not responsible directly for the binding of transcription factors are not preserved and so overall sequence conservation is very poor [2]. Thus, the quest to identify regulatory conservation must be complemented by a more thorough understanding of the inherent grammar of regulatory sequences, which would lead to improved alignment models specifically tailored to regulatory sequences [23].

Conservation versus function

During the past few years several strategies have been deployed to perform genome-wide sequence comparisons, which in turn identified several novel functional elements in vertebrate genomes. However, they have not yet defined how far conservation of noncoding elements can be pushed to identify functional elements efficiently. The approach used to build our dataset is significantly different from previous approaches, because on the one hand it is stringent by focusing on fish-mammal comparisons and on the other hand it is more sensitive than previous approaches because of its CHAOS-based alignments and lower length cut offs. The requirement for conservation in fish genomes in the SCE dataset would thus lead to the loss of mammalian-specific enhancers, but on the

other hand it is likely to act as a stringent filter for slowly evolving DNA that may be free from any functional constraints. The differences between the SCE dataset and previously reported datasets became evident by performing an overlap analysis among them (see Materials and methods, below, for details; also see Figure S2). The partial overlap between the analyzed datasets once again emphasizes that the approach used to determine conserved non-genic elements has a notable impact on the elements identified. Approximately 50% of SCEs do not overlap any known feature, suggesting that the use of non-exact seeds for the initial local alignments has a significant impact on the analysis of noncoding DNA harboring short, well conserved elements, and that our dataset is substantially different from previous datasets both quantitatively, and qualitatively.

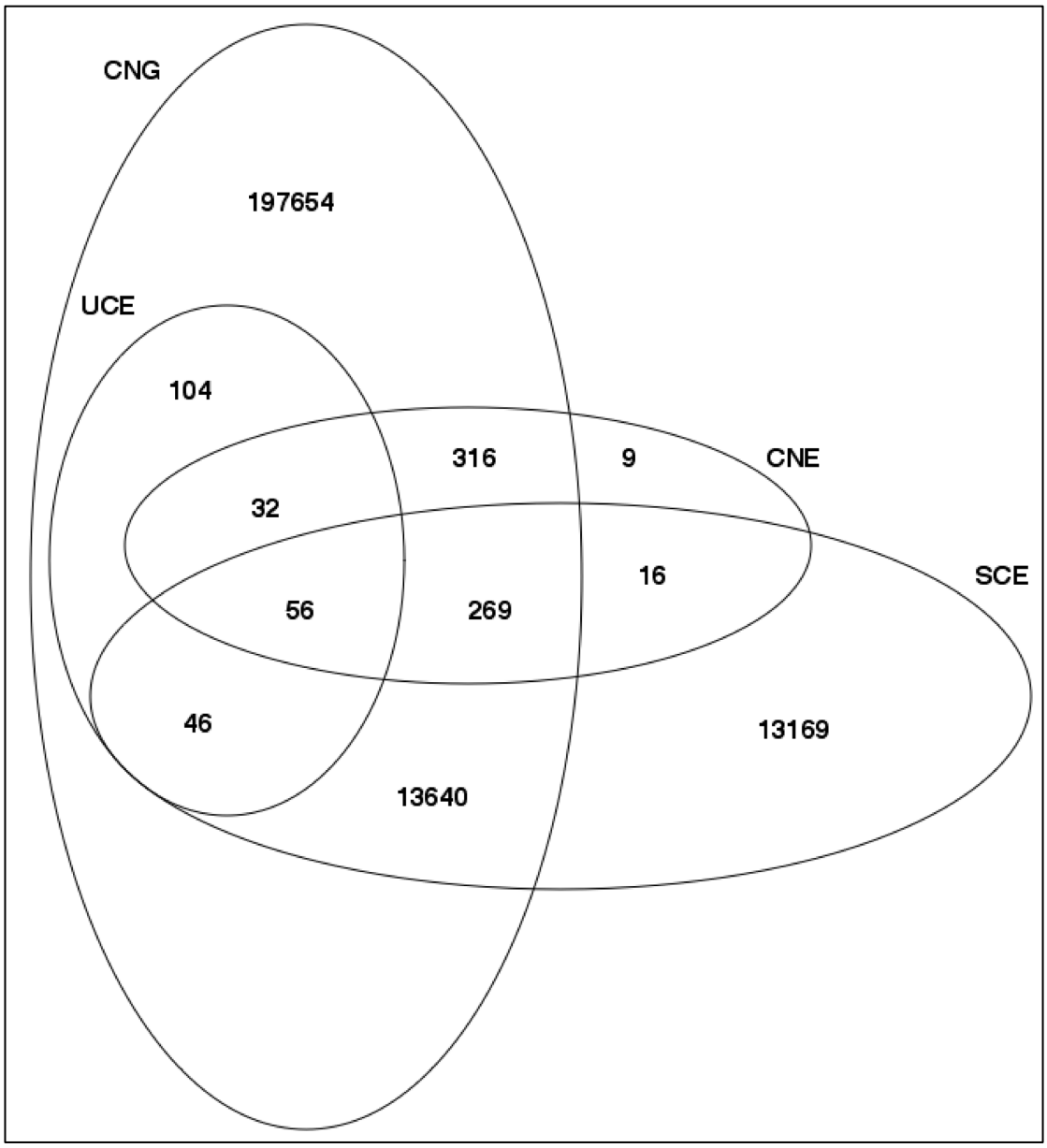


Figure S2 Venn diagram illustrating the overlap analysis of four datasets (CNGs, UCEs, CNEs and SCEs)

UCEs were detected using a whole-genome local alignment strategy between human and mouse (although they are often conserved in fish genomes as well) and selected for being 100% identical over at least 200 bp [14]. They were shown to be often located in clusters in the proximity of 'trans-dev' genes. Poulin and coworkers [44] showed that the ultraconserved Dc2 element is necessary and sufficient for brain tissue enhancer activity, and an ongoing systematic study using transgenic mice has shown enhancer activity for more than 60% of the

elements tested so far (Pennachio and coworkers, unpublished data). Our dataset overlaps only 45% of the UCE elements because of its 'regional approach', which will miss any elements that are conserved across non-orthologous loci or that are found beyond the region we took into consideration (namely, beyond the previous or next gene). Nonetheless, the results of our study indicate that the enhancer function that has so far been associated with them does not explain fully their level of conservation, because our dataset, although rich in enhancers, has much lower levels of sequence identity and length as compared with UCEs. Only one of the fragments that we tested (SCE 1973 from the mapkap1 gene) overlaps with a UCE element. The overlap is only 33 bp, and there is no further identity with the UCE in fugu, but the element nonetheless acted as a tissue-restricted enhancer in vivo. A region adjacent to the UCE in mouse (SCE 1973), although not ultraconserved, is also conserved in fish and acted as a generic enhancer in our assays, highlighting the complexity of these regions and adding to the ongoing debate regarding their function and evolution [45].

A large set of sequences, defined as CNGs, was constructed by using pair-wise local sequence comparison between the human and mouse genome on chromosome 21 (identity = 70%, length = 100 bp), and it was shown that two-thirds of them lacked transcriptional evidence in vivo [13]. The conservation of these regions in other mammalian genomes was later also confirmed [8]; however, thus far they have not been shown to represent functional regulatory elements to a satisfactory scale, and so the specificity of this method in the identification of enhancers is not known. A recent genome-wide study of

functional noncoding elements conserved in fish genomes used pair-wise local sequence comparison between the human and fugu genomes to define 1,400 highly conserved noncoding elements (length = 100) and found that these were principally associated with developmental genes [15]. The overlap analysis highlights that although CNGs are three orders of magnitude larger than UCEs and CNEs and they contain the former fully and 96% of the latter, they only overlap approximately half of the SCE dataset. This suggests that there are qualitative differences between CNGs and our dataset. Interestingly, it has been shown that megabase deletions of two-gene deserts containing thousands of CNGs in mice had no phenotypic effects [46]. The authors stated that none of the CNGs contained are conserved in fish, and when we inspected these regions we discovered only a single SCE, very close to the boundary of the deletion.

Our dataset overlaps only 51% of the CNEs within the loci analyzed, probably because of the regional approach taken, which disregards elements conserved across non-orthologous loci. On the other hand more than 88% of the genes that contain CNEs also present SCEs, thus identifying regulatory elements in the majority of those genes nonetheless. A group of CNEs were shown to act as enhancers when tested in vivo in zebrafish by co-injecting them with promoter/reporter constructs. Our data, compared with the CNE dataset, is a radical extension (of an order of magnitude) of similar conserved elements, indicating a significant quantitative difference. There is also a qualitative difference, however, because we identified elements in a very broad range of genes, including genes from the extracellular regions and membrane and many genes participating in physiological and cellular processes, which are not

transcription factors. The quantitative and qualitative differences in our dataset constitute a major departure from previously published datasets, which show conservation across vertebrates and clear evidence of involvement in enhancing gene expression, namely CNEs and UCEs. Thus, the lack of overlap between the datasets taken into consideration is probably a compounded effect of methodological differences (for example, CNEs versus SCEs), real biological differences (CNGs versus others) and a compound effect of the two differences (UCEs versus CNEs and SCEs). Our results suggest that a large portion of the noncoding genome is composed of enhancers. Although it is certain that conserved noncoding regions play other roles that we were unable to verify, either they constitute a minority or they are able to perform several functions besides that of enhancers.

Comparative genomics has been applied successfully to the study of regulatory elements in the past, using approaches based on motif libraries. Xie and coworkers [19] aligned the promoter and 3'-untranslated region sequences from four mammalian genomes by using BlastZ with a regional approach and were able to identify motifs that were over-represented in conserved regions around genes. They showed that these motifs are non-randomly distributed with respect to gene expression data but they did not identify specific instances of the motif as active copies in the genome. Thus, this study, apart from using a different methodology, focused on mammalian genomes only (as compared with our vertebrate-wide approach) and focused on proximal 5'- and 3'- untranslated region sequences, discarding introns as a negative control set based on the assumption that they contain few regulatory elements. Our study was based on

sequence alignment, focused on a broader dataset comprising several vertebrate genomes and made use of the full intergenic and intronic sequence for each locus taken into consideration.

Ettwiller and coworkers [23] proposed a novel computational method that also makes use of comparative genomics. First, they developed a novel alignment routine, called promoterwise, that models promoter evolution more closely. Then, they used an efficient method to allow direct enumeration of all possible motifs up to 12-mers, including motifs with wildcards. Finally, active instances of the motif set thus generated were confirmed by searching them in regions that were found to be conserved in the alignment routine. This work was aimed at comparing distantly related genomes, by searching for over-representation in related orthologs across mammalian and fish genomes to identify specific instances of these motifs. Moreover, they proved using experiments in Medaka that these active motifs are necessary to drive expression in vivo. This study resembles our strategy more closely because it involves a vertebrate-wide comparison, although it focused only on 5 kb promoter sequences.

Motif library based approaches are complementary to our alignment focused approach. One important difference between these approaches is that the computational requirements of motif-based approaches are very high, and so it is not feasible to execute a motif library approach over a third of the genome sequence, as was done in this work. On the other hand motif library approaches are able to pinpoint specific motifs that are at the core of the regulatory grammar, whereas our approach uncovers a dataset that is likely to contain a redundant set of regulatory motifs. It would be a natural extension of our work

to compare these datasets in order to elucidate shuffling and determine the extent to which enhancers can be represented as clusters of simpler motifs as well as to investigate shuffling of enhancers in relation to the shuffling of single motifs.

Toward improved detection of cis-regulatory elements

The fact that, despite an increase of an order of magnitude in our dataset, a similar ratio of elements was found to act as enhancers as compared with the CNE dataset suggests that the extent of sequence conservation of regulatory elements is a moving target that reflects the technique used to identify them. There is a clear need for novel methodologies to detect thus far hidden conserved elements. The algorithm Shuffle- LAGAN is an alignment program that resembles our approach, although it only aligns shuffled elements within pairwise alignments and therefore it would have not helped to bypass the initial step of selecting rCNEs found conserved in at least three mammalian genomes. A desirable extension of Shuffle-Lagan would be to add the ability to process orthologous loci from several genomes at once. More knowledge about the evolution of noncoding DNA will be needed in order to obtain better scoring schemes and thus yield not only sensitive alignments but more reliable predictions of enhancers and other regulators of gene expression [25].

An important aspect that differentiates our approach from previous BLAST-based approaches is the use of CHAOS for the alignment of mammalian loci to fish loci. In order to verify the extent to which CHAOS differs from BLAST in this particular type of search, we performed the search for SCEs from our set of rCNEs in the fugu genome, comparing NCBI BLAST and CHAOS at different word

sizes and identical length and identity cut offs. The results indicate that although CHAOS scales exponentially as word size decreases, the number of hits obtained with BLAST is almost unaltered by the difference in word size. Moreover, there is a qualitative difference in the hits obtained because the increase in number of elements identified at small word sizes using CHAOS is due in large part to shuffled elements that BLAST is unable to identify (Figure S3). This qualitative difference is most notable using word size 10, for which only about 4% of BLAST results are shuffled elements as compared with 72% of the elements identified by CHAOS.

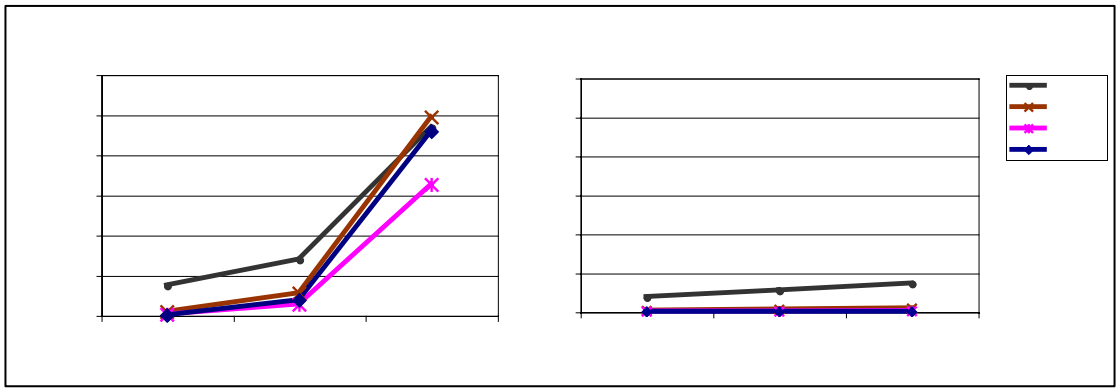


Figure S3 Number and type of conserved elements identified by CHAOS and BLAST2 in our dataset as a function of the word size used

This significant difference reiterates quite clearly that looking for sequence similarity across long stretches of identical words is not a valid approach to identifying conserved regulatory elements. At the same time, if we were to decrease word sizes to what would be biologically sensible (that is to say, word size 5-8, similar to the size of transcription factor binding sites) it would be difficult to assess whether the elements identified as conserved were the result of convergent transcription factor binding site architecture generated de novo, rather than truly conserved across vertebrate evolution. Thus, novel

methodologies need to be developed that would make use of small word sizes but include other constraints and scoring systems that would help to distinguish biological features preserved through evolution from neutrally evolving short fragments in the genome. To this extent, a well-curated resource collecting known enhancers (deposited in GenBank, for example) as well as a large set of systematically validated enhancers (such as Enhancer Browser [47]; Pennacchio LA, unpublished data) would help in building valid scoring systems and improve current methods.

In vivo transient assays

Our in vivo assays by co-injection revealed interestingly that most enhancers identified using this method were restricted in their activity to one or two tissues. Reassuringly, the expression profile of 24-hour-old embryos co-injected with the ArC positive control exhibited clear notochord enhancement (Figure 7b), as described previously [27]. The relative evolutionary closeness of fugu and zebrafish implies that expression and regulation of expression of developmentally regulated genes is probably well conserved [15,48]. Very little is known about Fugu gene expression patterns, but the availability of gene expression pattern information for many zebrafish genes provides a reliable assessment for the tissue specificity of the Fugu SCEs tested in our transient transgenic embryo assays. The functional analysis of SCEs by enhancer essays carried out in the transient transgenic zebrafish identified several new tissue restricted enhancer functions for genes where the endogenous expression pattern is not known. Future work will be required to analyze the role of these enhancers in relation to the detailed analysis of expression patterns of the genes they are associated with. In several cases the SCEs found within a locus provided

tissue specificity reminiscent of the gene expression pattern of the flanking gene, arguing strongly for a direct role of these SCEs in regulating the expression of the flanking gene. It will, however, only be possible to prove unequivocally that there is a need for these enhancers to drive the expression of the candidate gene by site-specific mutation of the SCEs in the genomic context. Two of the control fragments that do not contain detectable conservation were also shown to have significant enhancer effect, and in particular one of the two exhibited activity that was greater than that of most SCEs tested.

Mechanisms for genome-wide shuffling

Genomic rearrangements have already been reported on a large scale in a study examining gene order in regions of synteny between human and Takifugu rubripes [49]. Similar rearrangements should be seen when analyzing smaller regulatory regions that could harbor enhancers, which have strong evolutionary constraints on their sequence but frequently not on their specific localization with respect to the gene they act upon. We found that shuffling and rearrangements are not only applicable to nongenic sequences but also are a widespread phenomenon that involves 30% of the genes we analyzed.

Recently, there has been discussion on the role of cis-regulatory elements in the spatial organization of the genome and their possible role in restricting chromosomal rearrangements (see Liu and Garrard [50] and the review by Pederson [51]). The most well known examples of this are the hox clusters, although they do exhibit wider plasticity in fish genomes than in other genomes. Our work shows clearly that shuffling of cis-regulatory elements is a widespread phenomenon within orthologous loci. It would be interesting to investigate

further the extent to which shuffling occurs on a genome-wide scale. Further analysis is required to determine the real extent of this phenomenon outside orthologous loci. This is the first genome-wide study to show that regulatory elements are mobile across species; this finding should be taken into consideration when using comparative evolutionary methods to locate potential regulatory elements.

It would be useful to assess the extent of shuffling on a genome-wide basis to develop a thresholding statistic. We investigated this by searching for SCEs in fugu non-orthologous loci. Although this results in a significantly lower number of hits (23,100 hits in orthologous analysis, 9,884 in non-orthologous analysis; $P < 2.2 \times e\text{-}16$), the result shows that shuffling does occur outside of the orthologous locus. It is difficult to interpret this result without taking into account other data (for example, expression data and sequence similarity for genes considered non-orthologous or, indeed, in vivo assays on hits in non-orthologous loci) that would allow us to establish the extent to which hits in non-orthologous loci are noise and to which they represent regulatory elements in genes with similar expression patterns. Finally, we must emphasize that the fact that our mammalian rCNE dataset is built using a global alignment approach will limit the search space and will not allow us to investigate the extent of regulatory element shuffling within mammals. This data reduction step has been used in the past [52], and it was used in the present analysis based on the assumption that shuffling of regulatory elements is more likely to occur over longer evolutionary distances. Widespread shuffling of elements could act as a potential mechanism for providing new expression sites to genes that are placed in the vicinity of a

translocated enhancer. These issues can only be tackled appropriately by performing further analysis of the extent to which conserved elements shuffle beyond their locus of origin on both small and large evolutionary distances.

Conclusion

Our work shows that shuffling of cis-regulatory regions is a widespread phenomenon across the vertebrate lineage that affects approximately 70% of the conserved noncoding elements identified. The approach used allowed us to demonstrate that there is an order of magnitude more conserved elements in the vertebrate lineage than has previously been shown. Moreover, conservation of regulatory elements occurs over thousands, rather than hundreds, of genes. By casting a wider net over vertebrate noncoding conservation, we were able to demonstrate that there are hundreds of genes that do not belong to the 'trans-dev' category, such as genes found in the membrane and extracellular regions, which also contain conserved noncoding elements. Finally, our in vivo assays prove that although we cast a wider net, the catch was just as rich; more than 80% of the elements tested acted as enhancers, and the majority of them showed tissue-restricted patterns of expression in line with the neighboring gene.

Materials and methods

Selection of genes and sequences

Groups of homologous genes from the genomes of Mus musculus, Homo sapiens, Canis familiaris, Rattus norvegicus, Takifugu rubripes, Tetraodon nigroviridis, and Danio rerio were selected from the Ensembl-compara database [13] and their sequences were obtained from Ensembl database release 32 [14]. Genes were considered homologous if they were classified as best reciprocal hits in

Ensembl-compara. We analyzed all of the genes that were conserved in at least four species, of which three had to be human, mouse and fugu, and one could be either dog or rat. This selection led to 9,749 groups of homologous genes. For each gene we analyzed the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the next gene upstream and the next gene downstream. The region was extracted from Ensembl and the 5'-3' sequence of the locus was stored in a custom database (all mouse genes were stored as being in forward strand on the sequences stored). In cases in which the Ensembl gene contained multiple transcripts, the longest transcript was taken into consideration for the pre-gene, post-gene, and intron assignments of SCEs, but all exons (including those of other transcripts) were used to mask the sequence from coding regions. Similarly, if there were nested genes present in the locus, they were not taken into consideration to determine the extent of sequence to analyze, but they were taken into consideration to mask coding sequences in the region.

Identification of mammalian regionally conserved elements

Global multiple alignments among human, mouse, rat, and dog were performed on each group of homologous genes using MLAGAN [25] with default parameters. The multiple alignments thus obtained were parsed using VISTA [32] with a window of 50 bases searching for conserved segments of at least 100 bp having a percentage identity of at least 70%. From these regions we selected as rCNEs only those regions that were shared and overlapped in at least mouse, human, and a third mammalian genome (either dog or rat) with a minimum length of 50 bp. In cases in which the upstream region of an analyzed gene

coincided with the downstream region of another analyzed gene, rCNEs were counted only once.

Identification of shuffled conserved regions

Mouse rCNEs were used as query sequences against the respective fugu, zebrafish, and tetraodon homologous sequences using CHAOS [24] on both strands with the following parameters: word length 10, score cut off 10, rescoring cut off 1,000, and BLAST-like extension on. Other parameters were left as set by default including the degeneracy tolerance of 1 (allowing a single mismatch in the seed of the alignment). The hits thus obtained were filtered to retain only those with at least 60% identity and 40 bp length. Although three genomes were queried, a hit in Fugu was required to consider the result an SCE. All other hits (if any) were used to select the region of overlap as the final SCE, but only SCEs greater than 20 bp after the overlap analysis were taken into consideration.

Gene Ontology analysis

Ensembl gene IDs were converted into the corresponding RefSeq IDs before the analysis. The GOstat program [34] was used to find statistically over-represented GO IDs in the groups of genes, using the 'goa_mouse' GO gene association database as a reference. The false discovery rate and the P value cut off of 0.001 options were used. Raw output was converted in supplementary tables using a custom Perl script. The simple association of genes to GO classes presented in Figure 4 were produced using DAVID version 2 [53].

Mapping of conserved elements

rCNEs and SCEs were classified as 'genic' if they overlapped any Ensembl genes, Ensembl expressed sequence tag (EST) genes [31], ESTs [54], EMBL proteins [55], or Genscan predictions [56] from the Ensembl Mus musculus genome build release 32. Furthermore, each rCNE and SCE was classified with respect to the gene structure as 'pre-gene', 'intronic', and 'post-gene' based on its location within these three portions of the locus. According to this 'gene-centric' classification, as well as the strand of the fugu CHAOS hits (because all genes were stored in forward strand), SCEs were classified as 'collinear' (that is to say not changed in orientation and not shifted between gene portions), 'moved' (shifted between gene portions), 'reversed' (changed in orientation, but retained in the same gene portion), and 'moved-reversed' (changed in orientation as well as shifted in gene portion).

BLAST versus CHAOS comparison

A subset of about 50% of the mammalian rCNEs were used as query sequences against the corresponding fugu homologous sequences using CHAOS [24] and BLAST2 [16], using a gap penalty of 2 as was used in the CNE analysis and e-value set at infinity to ensure that no hits would be filtered because of their statistical significance, analyzing both strands. The analysis was conducted three times varying only the word length used between 20, 15, and 10. The hits thus obtained were filtered in order to take only those sharing an identity of at least 60% and a length of at least 40 bp.

Overlap analysis

Overlaps among different classes of conserved noncoding regions were defined using their genomic coordinates after having mapped all elements on the mouse loci used in this analysis. Because there is no downloadable dataset for CNGs, they were obtained by querying the GALA database [57] for conserved regions shared between human an mouse of at least 100 bp and 70% identity. CNEs [15], UCEs [14], and known enhancers were downloaded from Genbank. Enhancers were downloaded by searching for enhancer features in mouse Genbank records and then checking them manually to eliminate misnnotated entries. All the sequences thus downloaded were then mapped on the mouse loci used in our analysis by using Megablast [58] with default parameters for CNGs, UCEs and known enhancers, and with a gap penalty of 2 for mapping CNEs, in accordance with the parameters used by Woolfe and coworkers [15] in their analysis. Elements were considered mapped with 75% coverage and 75% percentage identity. Only elements that did not map to exons were taken into consideration.

Identification of control fragments

A set of control fragments to be tested in vivo was built from the same gene loci in which the tested SCEs were found, by selecting regions that were not conserved and did not present repeats, of the same length and number as the elements tested.

Zebrafish embryo injections

The enhancer activity was assayed in conjunction with the minimal promoter mHSP68, which was previously shown to have low activity in zebrafish embryos and which has allowed the detection of enhancer function from several

heterologous gene elements [28,59]. HSP68lacZ-pBS DNA plasmids containing the mouse HSP68 promoter [59] and lacZ were prepared using the Promega PureYield Plasmid Midiprep System plasmid preparation kit, digested by Promega BamHI enzyme, and DNA fragments were gel purified using the Promega Wizard SV Gel and PCR Clean-Up System kit (Promega, Madison, WI, USA). HSP:lacZ DNA fragments were resuspended in 1% phenol red containing nuclease-free water at a concentration of 25 ng/"l, as described previously [60], and were injected into the cytoplasm of zebrafish embryos at one cell stage. Wild-type embryos (Tubingen AB) were collected after fertilization and dechorionated by pronase, as described previously [61]. Fugu DNA was used for production of SCE fragments. Fragments were amplified by polymerase chain reaction, then isolated and purified using the Qiagen Qiex DNA purification kit (Qiagen, Valencia, CA, USA), and finally eluted in sterile water. For injection, phenol red was added to yield a final concentration of 50 ng/"l. Coinjection of polymerase chain reaction fragments at a concentration of 50 ng/"l reaching a range of 5 to 1 molar ratio with the HSP:lacZ fragment. Embryos were maintained at 28°C and collected at prim 6 stage [62], fixed and lacZ stained as described previously [27].

Analysis of transgene expression

LacZ stained embryos were analyzed by plotting the mosaic expression activity on expression maps, as described previously [63,64]. The co-injection experiments were repeated three times. Data from approximately 100-120 embryos were collected on a single expression map providing an expression profile. For each embryo expressing lacZ the number of expressing cells was counted and classified in muscle, notochord, CNS, eye, ear, and vessels. These

tissues were selected because they are well defined at the time of inspection [27].

Other tissues that were either difficult to determine or might have represented abnormalities (ectopic tissue growth, apoptotic mismigrating cells) were counted as 'other'. Twentythree SCEs, four SCEs overlapping known mouse enhancers, 12 control fragments, one negative control consisting only of the HSP:lacZ fragment, and the positive control ArC [65] were analyzed.

We verified the significance of the enhancement of expression over the general low level improvement of expression of co-injected fragments probably caused by carrier DNA effect (see, for example, [65]) in two ways. First, we aimed to detect tissue-restricted enhancers; second, we aimed to identify generic enhancers. To identify tissue-restricted enhancers we compared, for each fragment co-injected and for each tissue, the number of expressing cells with respect to the number of expressing cells from the embryos injected with the negative control in the respective tissues, only when the average of cells expressing lacZ in injected embryos was higher than in the control. Fisher exact tests were then used on the comparisons and a P value cut off of 0.01 was used to classify a fragment as a tissue-restricted enhancer. The identification of generic enhancers was performed by establishing the average and standard deviation of the number of expressing cells per expressing embryo in the control fragments and then classifying as enhancers fragments in which the number of expressing cells per embryo was higher than the average plus twice the standard deviation of the control fragments. In the calculation of the average and standard deviation we excluded the UBL7 control fragment because it was a clear outlier that

exhibited activity that was higher than any of the enhancers tested, including the positive control. All fragments classified as enhancers by either of the two tests were considered positive.

Acknowledgements

We appreciate the useful input from two anonymous referees and we should like to acknowledge helpful discussions with Michael Brudno, Caterina Missero, Diego Di Bernardo, Marco Sardiello, Maria Luisa Chiusano, Giovanni Colonna and Roberto di Lauro. We would also like to thank for their technical support Marco De Simone, Mario Traditi and Alessandro Davassi. A special acknowledgement also goes to the late Parvesh Mahtani, who shared our enthusiasm for this project. This work was supported by the Fondazione Telethon and the Sixth Framework Program of the European Commission (LSH-2003-1.1.0-1).

References

1. Blackwood EM, Kadonaga JT: Going the distance: a current view

of enhancer action. Science 1998, 281:60-63. 2. Oda-Ishii I, Bertrand V, Matsuo I, Lemaire P, Saiga H: Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. Development 2005, 132:1663-1674.

3. Dickmeis T, Muller F: The identification and functional characterisation of conserved regulatory elements in developmental genes. Brief Funct Genomic Proteomic 2005, 3:332-350.

4. Chuzhanova NA, Krawczak M, Nemytikova LA, Gusev VD, Cooper DN: Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. Gene 2000, 254:9-18.

5. Kermekchiev M, Pettersson M, Matthias P, Schaffner W: Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling? Gene Expr 1991, 1:71-81.

6. Surguchov A: Migration of promoter elements between genes: a role in transcriptional regulation and evolution. Biomed Sci 1991, 2:22-28.

7. Boffelli D, Nobrega MA, Rubin EM: Comparative genomics at the vertebrate extremes. Nat Rev Genet 2004, 5:456-465.

8. Dermitzakis ET, Reymond A, Antonarakis SE: Conserved non- genic sequences: an unexpected feature of mammalian genomes. Nat Rev Genet 2005, 6:151-157.

9. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA: A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet 2003, 19:119-124.

10. Sorek R, Ast G: Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 2003, 13:1631-1637.

11. Weber MJ: New human and mouse microRNA genes found by homology search. Febs J 2005, 272:59-73.

12. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S: Detecting conserved regulatory elements with the model

genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci USA 1995, 92:1684-1688.

13. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE: Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 2002, 420:578-582.

14. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: Ultraconserved elements in the human genome. Science 2004, 304:1321-1325.

15. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al.: Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 2005, 3:e7.

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

17. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988, 85:2444-2448.

18. Bergman CM, Kreitman M: Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res 2001, 11:1335-1345.

19. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 2005, 434:338-345.

20. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al.: Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004, 14:708-715.

21. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC: Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 2005, 15:1051-1060.

22. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al.: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005, 15:1034-1050.

23. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: The discovery, positioning and verification of a set of transcriptionssociated motifs in vertebrates. Genome Biol 2005, 6:R104.

24. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B: Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics 2003, 4:66.

25. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S: Glocal alignment: finding rearrangements during alignment. Bioinformatics 2003:i54-62.

26. Muller F, Blader P, Strahle U: Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements. Bioessays 2002, 24:564-572.

27. Muller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. Development 1999, 126:2103-2116.

28. Rastegar S, Albert S, Le Roux I, Fischer N, Blader P, Muller F, Strahle U: A floor plate enhancer of the zebrafish netrin1 gene requires Cyclops (Nodal) signalling and the winged helix transcription factor FoxA2. Dev Biol 2002, 252:1-14.

29. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003, 13:721-731.

30. Appendix to paper by Sanges .R et al. http://valis.tigem.it/sce.html]

31. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al.: Ensembl 2006. Nucleic Acids Res 2006, 34:D556-D561.

32. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 2000, 16:1046-1047.

33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.

34. Nat Genet 2000, 25:25-29. Beissbarth T, Speed TP: GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004, 20:1464-1465.

35. Sprague J, Clements D, Conlin T, Edwards P, Frazer K, Schaper K, Segerdell E, Song P, Sprunger B, Westerfield M: The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res 2003, 31:241-243.

36. The Zebrafish Information Network [http://zfin.org/]

37. Walton RZ, Bruce AE, Olivey HE, Najib K, Johnson V, Earley JU, Ho RK, Svensson EC: Fog1 is required for cardiac looping in zebrafish. Dev Biol 2006, 289:482-493.

38. Kudoh T, Tsang M, Hukriede NA, Chen X, Dedekian M, Clarke CJ, Kiang A, Schultz S, Epstein JA, Toyama R, Dawid IB: A gene expression screen in zebrafish embryogenesis. Genome Res 2001, 11:1979-1987.

39. Kudoh T, Dawid IB: Zebrafish mab21l2 is specifically expressed in the presumptive eye and tectum from early somitogenesis onwards. Mech Dev 2001, 109:95-98.

40. Zecchin E, Conigliaro A, Tiso N, Argenton F, Bortolussi M: Expression analysis of jagged genes in zebrafish embryos. Dev Dyn 2005, 233:638-645.

41. Smale ST, Kadonaga JT: The RNA polymerase II core promoter. Annu Rev Biochem 2003, 72:449-479.

42. Ludwig MZ, Bergman C, Patel NH, Kreitman M: Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 2000, 403:564-567.

43. Tautz D: Evolution of transcriptional regulation. Curr Opin Genet Dev 2000, 10:575-579.

44. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA: In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 2005, 85:774-781.

Adams MD: Conserved sequences and the evolution of gene regulatory signals. Curr Opin Genet Dev 2005, 15:628-633.

46. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM: Megabase deletions of gene deserts result in viable mice. Nature 2004, 431:988-993.

47. Enhancer Browser [http://enhancer.lbl.gov/]

48. Miles CG, Rankin L, Smith SI, Niksic M, Elgar G, Hastie ND: Faithful expression of a tagged Fugu WT1 protein from a genomic transgene in zebrafish: efficient splicing of pufferfish genes in zebrafish but not mice. Nucleic Acids Res 2003, 31:2795-2802.

49. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al.: Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 2002, 297:1301-1310.

50. Liu Z, Garrard WT: Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases. Mol Cell Biol 2005, 25:3220-3231.

51. Pederson T: The spatial organization of the genome in mammalian cells. Curr Opin Genet Dev 2004, 14:203-209.

52. Van Hellemont R, Monsieurs P, Thijs G, de Moor B, Van de Peer Y, Marchal K: A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol 2005, 6:R113.

53. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003, 4:P3.

54. Boguski MS, Lowe TM, Tolstoshev CM: dbEST: database for 'expressed sequence tags'. Nat Genet 1993, 4:332-333.

55. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, et al.: EMBL Nucleotide Sequence Database: developments in 2005. Nucleic Acids Res 2006, 34:D10-D15.

56. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997, 268:78-94.

57. Giardine B, Elnitski L, Riemer C, Makalowska I, Schwartz S, Miller W, Hardison RC: GALA, a database for genomic sequence alignments and annotations. Genome Res 2003, 13:732-741.

58. Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. J Comput Biol 2000, 7:203-214.

59. Kothary R, Clapoff S, Darling S, Perry MD, Moran LA, Rossant J: Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. Development 1989, 105:707-714.

60. Muller F, Lakatos L, Dantonel J, Strahle U, Tora L: TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish. Curr Biol 2001, 11:282-287.

61. Akimenko MA, Johnson SL, Westerfield M, Ekker M: Differential induction of four msx homeobox genes during fin development and regeneration in zebrafish. Development 1995, 121:347-357.

62. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF: Stages of embryonic development of the zebrafish. Dev Dyn 1995, 203:253-310.

63. Müller F, Williams DW, Kobolak J, Gauvry L, Goldspink G, Orban L, Maclean N: Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. Mol Reprod Dev 1997, 47:404-412.

64. Müller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. Development 1999, 126:2103-2116.

65. Parks RJ, Bramson JL, Wan Y, Addison CL, Graham FL: Effects of stuffer DNA on transgene expression from helper-dependent adenovirus vectors. J Virol 1999, 73:8027-8034.

Chapter 4: The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish

Published in: EMBO J, 2007, 26, 3945–3956

Abstract

Early steps of embryo development are directed by maternal gene products and trace levels of zygotic gene activity in vertebrates. A major activation of zygotic transcription occurs together with degradation of maternal mRNAs during the mid-blastula transition in several vertebrate systems. How these processes are regulated in preparation for the onset of differentiation in the vertebrate embryo is mostly unknown. Here, we studied the function of TATA-binding protein (TBP) by knock down and DNA microarray analysis of gene expression in early embryo development. We show that a subset of polymerase II-transcribed genes with ontogenic stage-dependent regulation requires TBP for their zygotic activation. TBP is also required for limiting the activation of genes during development. We reveal that TBP plays an important role in the degradation of a specific subset of maternal mRNAs during late blastulation/early gastrulation, which involves targets of the miR-430 pathway. Hence, TBP acts as a specific regulator of the key processes underlying the transition from maternal to zygotic regulation of embryogenesis. These results implicate core promoter recognition as an additional level of differential gene regulation during development.

Introduction

In most animal models, including Drosophila, Caenorhabditis elegans, zebrafish and Xenopus, the onset of zygotic gene activation is delayed until the midblastula transition (MBT). (Newport and Kirschner, 1982; Kimmel et al, 1995). Whereas there is no MBT in mammals, here zygotic gene activity is also delayed after fertilisation (Thompson et al, 1998). In the zebrafish blastula, the general delay in zygotic gene activity is followed by the sudden and broad activation of a large number of genes representing all main gene ontologies (Kane and Kimmel, 1993; Mathavan et al, 2005) leading to gastrulation. The activation of the zygotic genome is parallelled by an equally significant process, the differential degradation of maternally inherited mRNAs (Giraldez et al, 2005; Mathavan et al, 2005; De Renzis et al, 2007). Whereas little is known about the mechanisms of degradation of maternal mRNA, they are known to involve both transcription-dependent and -independent pathways (Bashirullah et al, 1999; Audic et al, 2001; Giraldez et al, 2005; Schier, 2007). Dynamic changes in expression of maternally and zygotically activated genes are observed during zygotic gene activation also in the mouse (Wang et al, 2004). Not all maternally inherited mRNAs degrade during early embryogenesis and many maternal mRNAs continue to influence embryo development until later developmental stages (Wagner et al, 2004; reviewed by Pelegri (2003)).

The initiation of zygotic transcription during MBT is believed to be regulated by a competition between chromatin and the assembly of the transcription machinery (Newport and Kirschner, 1982; Kimelman et al, 1987; Almouzni and Wolffe, 1995). The TATA-binding protein (TBP) has been implicated as a key

regulator of transcription initiation in early embryo development in vertebrates (Veenstra et al, 2000; Muller et al, 2001; Martianov et al, 2002). TBP protein levels have been shown to be limiting for transcription before MBT and are dramatically upregulated at the initiation of zygotic transcription (Prioleau et al, 1994; Veenstra et al, 1999; Bartfai et al, 2004). TBP, together with TBP associated factors (TAFs) are components of the TFIID complex, a key point at which activators can control transcription through the core promoter. Until recently, it was argued that TBP is required for the correct initiation of all RNA polymerase (Pol I, II and III)-mediated transcription in eukaryotes. However, recent reports have revealed the contrary: the composition of Pol II core promoter-binding complexes varies and is likely to represent a point of differential gene expression regulation (reviewed by Davidson (2003)). Consistently, whereas TBP is essential for early embryo development, it is not required for all Pol II transcription as demonstrated by studies on a small number of vertebrate genes (Veenstra et al, 2000; Muller et al, 2001; Martianov et al, 2002). The apparent redundancy of TBP in vertebrates is probably due to the function of TBP-like factors (TLF/TRF2) (Veenstra et al, 2000; Muller et al, 2001) and the recently described second set of TBP paralogue genes TBP2/TRF3 (Persengiev et al, 2003; Bartfai et al, 2004; Jallow et al, 2004). The functional requirement for different TBP family proteins in embryogenesis suggests specific nonoverlapping roles for these factors in regulating subsets of genes (Moore et al, 1999; Teichmann et al, 1999; Bartfai et al, 2004; Jallow et al, 2004).

Our objective was to investigate the transcriptional regulatory mechanisms that involve core promoter recognition proteins such as TBP in the whole organism.

The transition of gene activity from maternally inherited mRNAs to zygotic gene expression provides an ideal model for the analysis of the control of transcription initiation (Newport and Kirschner, 1982). By using Morpholino (MO) knockdown and microarray expression profiling, we have addressed which genes require TBP for their activity and what is the function of TBP in regulating the transition from maternal to zygotic regulation during early vertebrate embryo development. We show that TBP is preferentially required for genes that exhibit dynamic changes in their expression during ontogeny. Furthermore, we provide evidence for a previously undocumented negative regulatory role of TBP in zygotic gene activation. Importantly, we also describe a novel biological function of TBP: a role in the degradation of a subset of maternal mRNAs after MBT.

Results

TBP regulates specifically a subset of mRNAs in the dome-stage embryo

In the early embryo, the steady-state levels of mRNA result from a dynamic process of gradual degradation of maternal mRNAs and the delayed initiation of zygotic gene expression at the MBT (Figure 1A). To investigate the role of TBP in regulating genes expressed in the early zebrafish embryo, we carried out a microarray analysis of 10,501 genes at the dome stage in embryos in which TBP function was blocked using MO antisense oligonucleotides as described previously (Muller et al, 2001). The dome stage occurs 1.3 h after the start of the global initiation of zygotic transcription at the MBT (Kane and Kimmel, 1993) and any TBP-dependent changes in gene activity detected at this stage are still

expected to be mostly direct transcriptional effects. Loss of protein was confirmed by Western blot (Figure 1B). A total of 1,927 genes from the 10,501 represented on the microarray were selected, having applied stringent but commonly used criteria (FDR cut off of 0.05) to eliminate potential false positives and false negatives from the analysis (see Materials and methods). Three distinct response groups of genes were identified: downregulated genes (≤ -2-fold change), genes with low variability in expression (values between > -2 and >+2-fold-change); upregulated genes (≥ +2-fold change) (Figure 1C and Supplementary Table I). The three groups thus identified were further validated by semiquantitative RT–PCR experiments; out of a total of 39 genes representing the above groups, 37 showed comparable activity to that observed in the microarray experiments (Figure S1). The specificity of the effects detected was confirmed by microinjecting a second MO targeting TBP mRNA (TBP MO2), which resulted in comparable gene expression changes to the above TBP MO injection when analysed by RT–PCR of 28 genes (Figure S2). Furthermore, the gene expression changes caused by TBP MO injection could be reverted by injecting a form of TBP mRNA that could not be targeted by the MO (Figure S2).

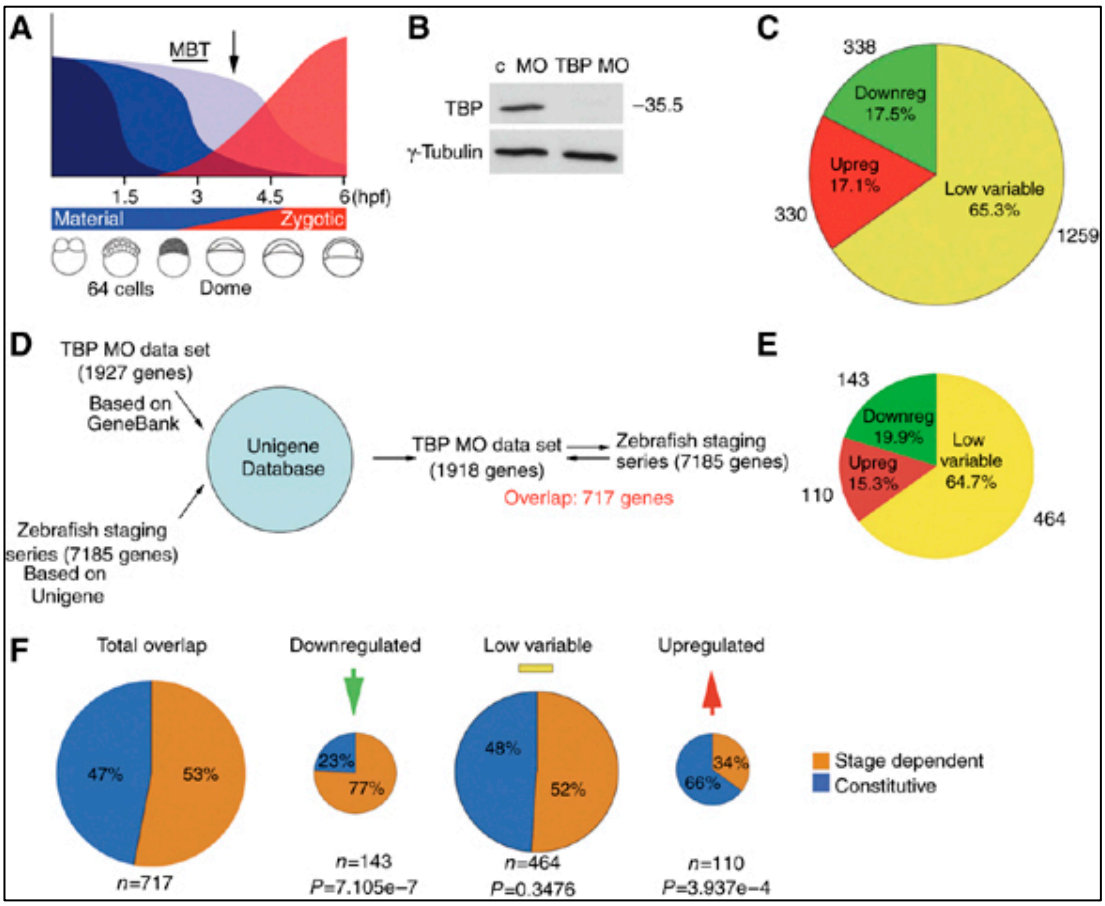


Figure 1 TBP is selectively required for both activation and repression of genes in the early zebrafish embryo. (A) Schematic diagram of the dynamics of mRNA degradation and zygotic gene activation during the MBT in the zebrafish embryo. Shades of blue indicate differential degradation of maternal mRNAs before the MBT. The red curve indicates the dynamics of zygotic gene activation. Time after fertilisation is indicated in hpf. Schematic drawing of respective stages of embryo development are shown below. The arrow indicates time point for collection of embryos for microarray analysis. (B) Western blot analysis of TBP protein levels in dome-stage zebrafish embryos injected with TBP (MOTBP MO) and control MO antisense oligonucleotides. (C) Pie chart diagram summary of expression profiling data of 1927 probes from microarray experiments carried out in dome-stage zebrafish embryos. (D) Schematic diagram of the protocol for identification of the intersection of genes analysed for TBP dependence among genes analysed for their expression dynamics during zebrafish development via the Unigene database. (E) Pie chart diagram of the proportion of genes found in the three response groups following TBP knockdown and overlap with the stage-dependent expression microarray. (F) Pie chart diagrams showing the distribution of constitutive and stage-specific genes among the total and the three response groups of genes in TBP morphant embryos. Numbers below the charts indicate the number of overlapping genes between the two data sets compared and w2 analysis of the gene distributions.

Within the 1927 genes that were used for this analysis, a large proportion of genes expressed in the dome-stage embryo (65.3%) showed no significant difference in signal strength between TBP MO- and control MO (c MO)-injected embryos, indicating that their steady-state mRNA levels are independent of TBP function. A smaller group of genes showed a significant reduction of expression

demonstrating that these genes require TBP directly or indirectly for their activation (17.5%). A similar number of genes (17.1%) showed increased levels of transcripts in TBP MO injected embryos, suggesting that TBP is required for controlling their steady-state levels by reducing their transcription and/or by enhancing mRNA degradation.

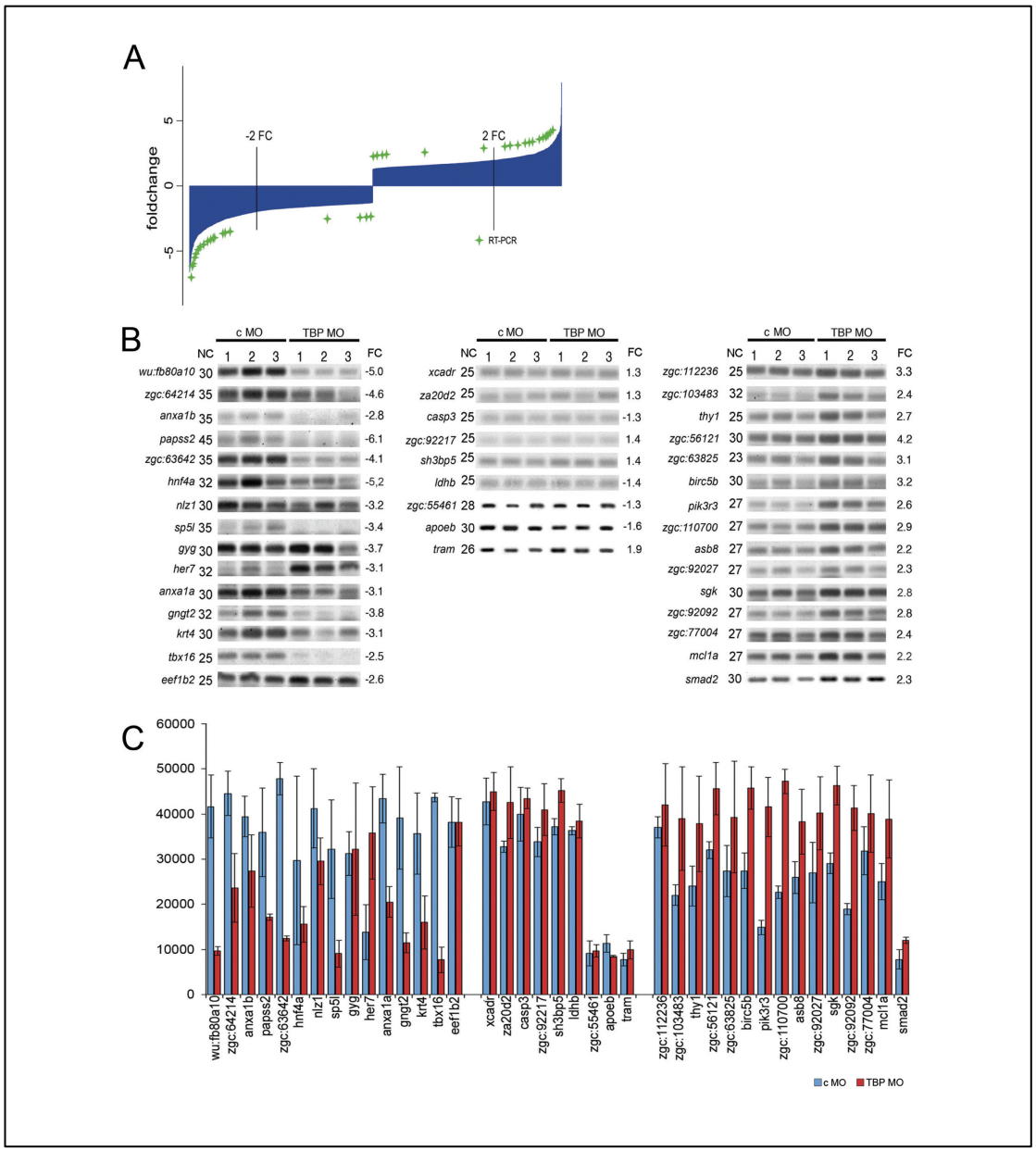


Figure S1 Validation of the microarray analysis by semi quantitative RT PCR and whole mount in situ hybridization. A, Distribution of genes chosen for validation and gene responses observed in the microarray are plotted in ascending order of fold change values. Genes chosen for validation by RT PCR are indicated by asterisk. B, RT-PCRs carried out with specific primers amplifying 39 genes from mRNA purified from embryos. MRNA was isolated from three parallel samples injected with c MO or TBP MO and used to produce cDNA for PCR analyses (see Materials and methods). The number of

cycles (NC) before saturation was determined for each PCR. The fold change (FC) values obtained from microarray analyses are also indicated. C, Quantification of the RT PCR results using the gel analysis tool of the ImageJ software, data of c MO are blue columns, TBP MO are in red. Genes zgc:55461, apoeb, tram and smad2 were analysed by using RNA from an independent set of experiments. Averages of triplicates are given with standard error. Abbreviations: c MO, TBP MO, control and TBP Morpholino antisense oligonucleotide injected embryos respectively.

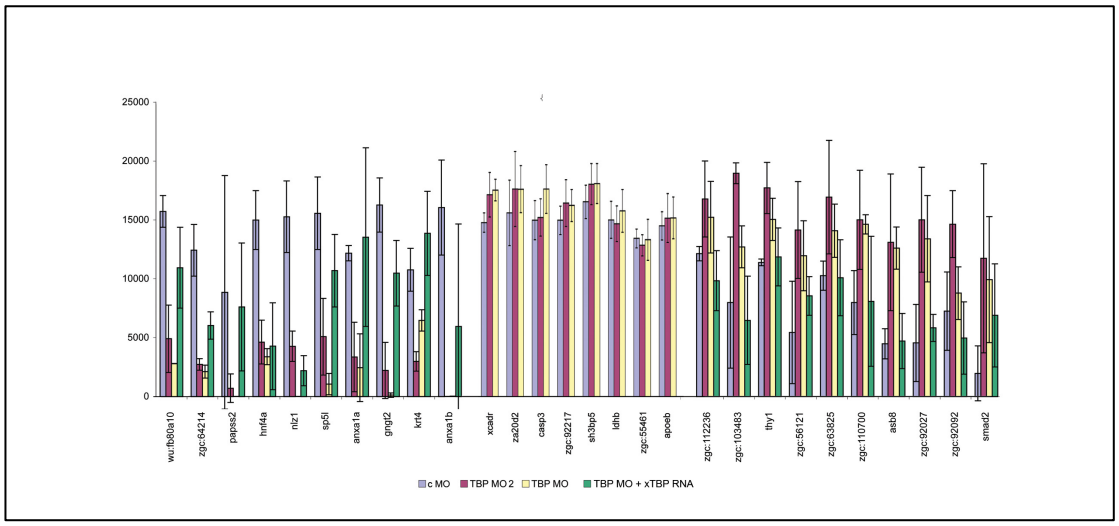


Figure S2 Quantification of semi quantitative RT-PCR to assay gene expression changes in TBP MO (yellow bars), TBP MO2 (red bars) and TBP MO + TBP mRNA (green bars) injected embryos in comparison to c MO (purple bars) injected embryos using the gel analysis tool of the ImageJ software. Averages of triplicates are given with standard error. Abbreviations as in Figure S1.

Most TBP activated genes are dynamically regulated during zebrafish ontogeny

To characterise further the genes affected by loss of TBP function, we tested whether genes in the three response groups described above showed discrete expression dynamics during zebrafish ontogeny. To this end, we compared the data set described above to an ontogenic stage-dependent expression profiling experiment on the zebrafish transcriptome (Konantz M, Otto G-W, Weller C, Saric M, Geisler R. Microarray analysis of gene expression in zebrafish development, manuscript in preparation). The two gene sets share 717 genes (Figure 1D and E) and the proportions of the three TBP morphant response groups among these 717 genes are similar to those of the total TBP microarray data set (Figure 1C and E).

Metanalysis of the ontogenic stage-dependent gene expression array was carried out to define two classes of genes. Genes showing stage specific peaks of expression activity during zebrafish ontogeny were classified as 'stage-dependent' and genes that showed no significant variation in gene expression during ontogeny were considered as constitutively active genes (Figure S3, see Materials and methods). Nearly half of the 717 genes (46.9%) that overlap between the two microarray data sets were shown to be constitutively expressed genes. The remaining genes belonged to the stage-dependent class (53.1%) showing dynamic activity during zebrafish ontogeny (Figure 1F and Figure S3).

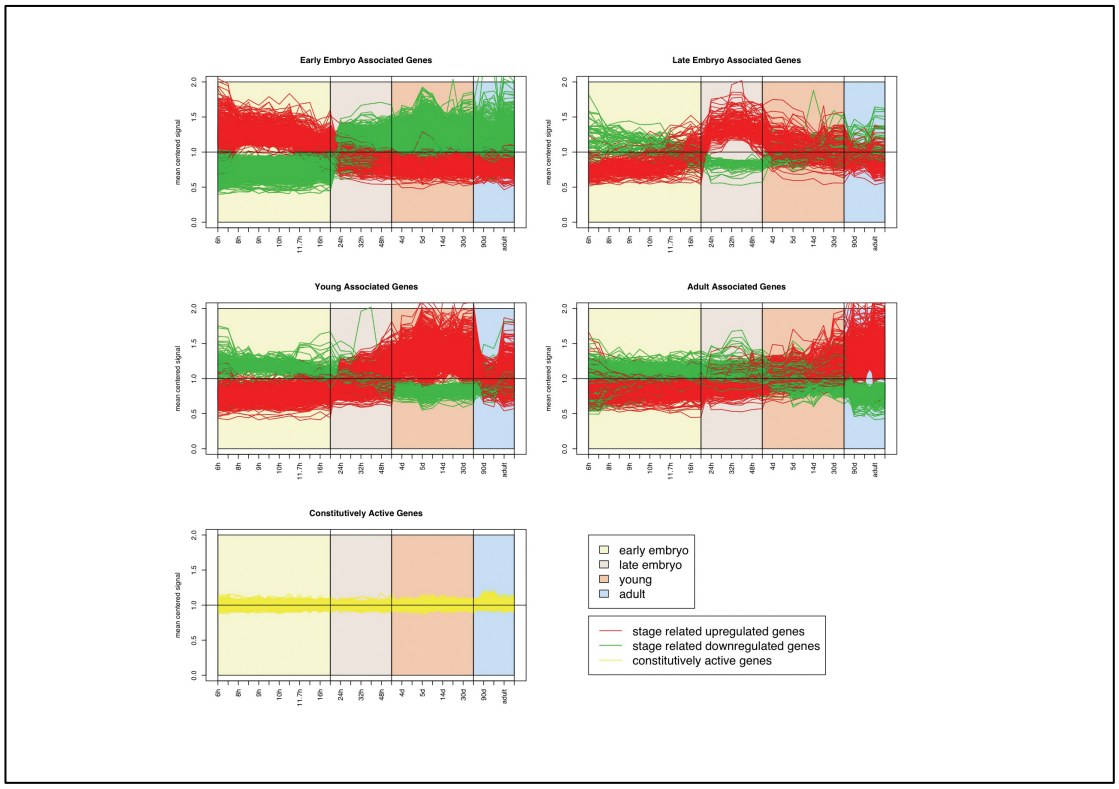


Figure S3 Association of genes with stage-specific activities during zebrafish ontogeny. Genes are grouped according to dynamic activity during ontogeny (one single peak of mRNA levels during ontogeny, top red and green lines) and constitutive activity (bottom, yellow lines). All the genes presented in the first 4 charts were grouped as ontogenic stage dependent genes. The 5th chart represents the genes that were selected for their constitutive expression throughout ontogeny. In each chart the x axis represents the time point of the sample extraction and the y axis represents the per gene mean centered signal of expression values.

Applying the 'constitutive versus stage-dependent' classification to the TBP microarray gene response groups revealed that genes that require TBP for their activation were predominantly stage-dependent (77%, Figure 1F), whereas upregulated genes in TBP morphants showed the opposite tendency. The low-variable group of genes did not show a bias to either stage-dependent or constitutively expressed genes. These results indicate that TBP-dependent activation tends to be a property of genes that show dynamic activity during ontogeny. Moreover, TBP tends to negatively regulate steady-state levels of constitutively active genes.

TBP dependence of transcription from isolated zebrafish promoters

TBP could influence steady-state levels of mRNA in zebrafish embryos both through transcriptional as well as post-transcriptional processes. To address the former, we tested 23 *gfp* constructs using promoters of zebrafish genes expressed at the sphere/dome-stage and representing various gene ontology

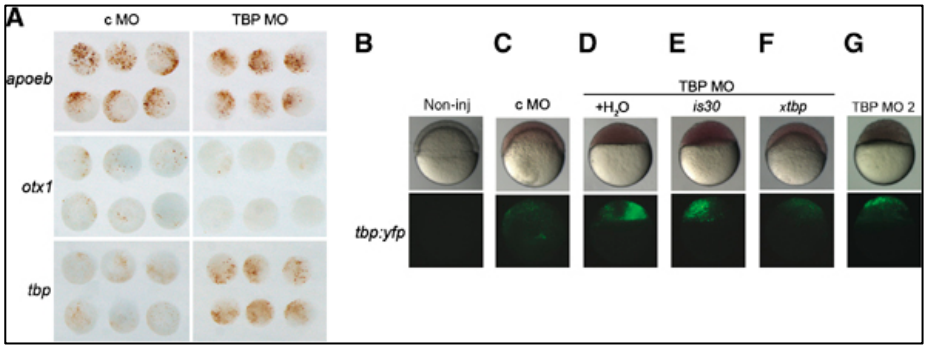


Figure 2 TBP is required for both activation as well as repression of zebrafish promoters. (A) Representative samples of whole-mount immunochemical staining of embryos injected with promoter:gfp constructs (view on animal pole) with brown staining indicating mosaic pattern of GFP activity. (B–F) Rescue of the TBP morphant phenotype by overexpression of recombinant TBP. (B) Noninjected embryos, (C–G) Injection of tbp:yfp reporter construct carried out together with MO oligonucleotides as indicated above the images. Injected embryos were split into separate batches and exposed to a subsequent injection of water, tbp or is30 mRNA, as indicated below the horizontal line (D–F). Lateral views of 7 hpf embryos in bright field (top) or fluorescence views under YFP (bottom).

classes (O'Boyle et al, 2007). TBP-dependent promoter activation was evident for seven promoters, including the *otx1* gene promoter (Figure 2A and Figure S4 and Supplementary Table II). This result is consistent with the proposed role of TBP in activating zygotic transcription of many genes during development. On

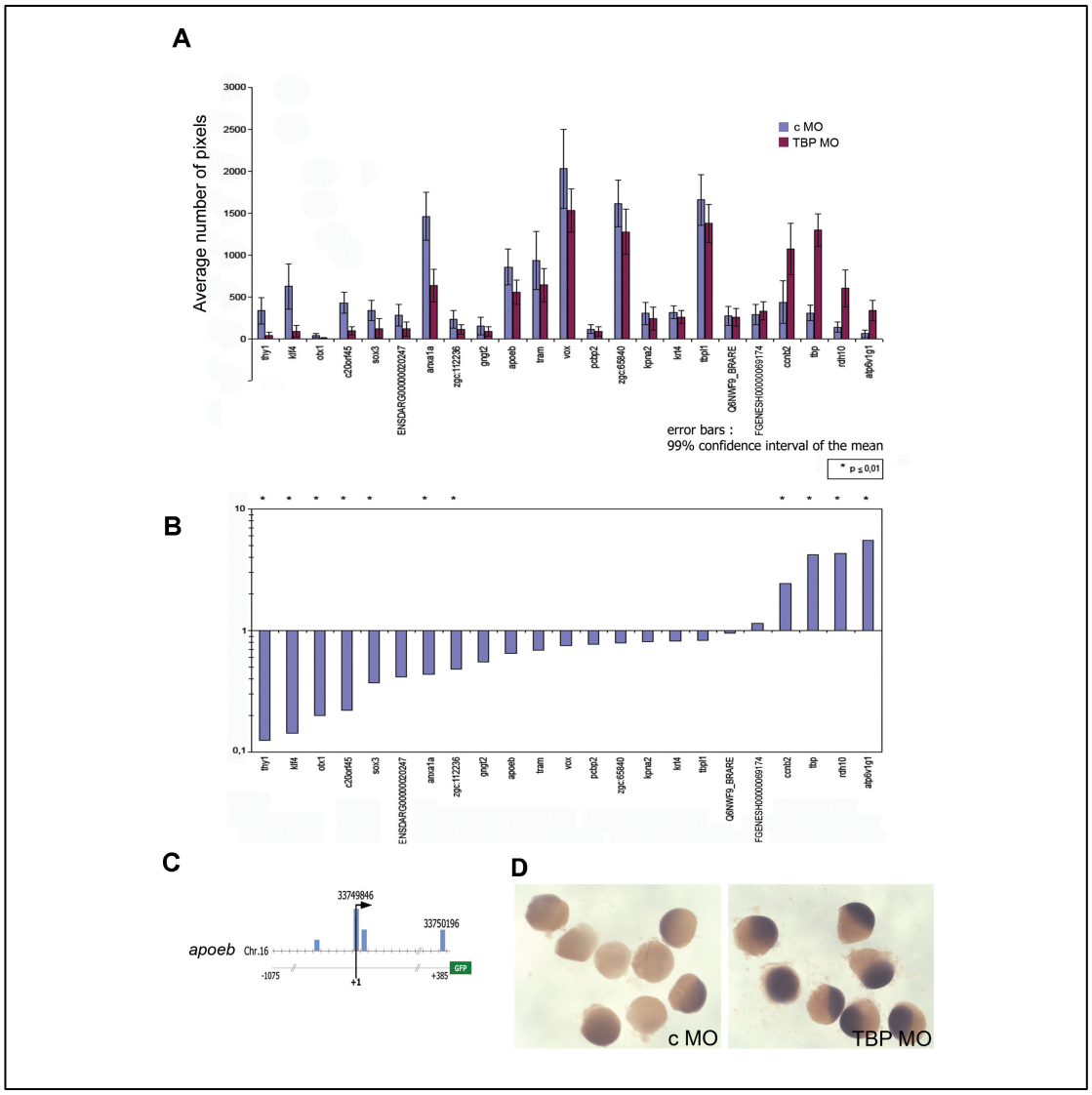


Figure S4 Analysis of dependence of the activity of a GFP transgene driven by various zebrafish promoters. A, The average number of brown pixels per embryo indicating GFP activity in immunohistochemically stained c MO (blue) and TBP MO (purple) injected zebrafish embryos. Error bars indicate the 99% confidence interval of the mean. B, The ratio of number of positive pixels between TBP MO and c MO injected embryos presented on a log scale. Asterisk indicates statistical significance by Mann-Whitney-U test and at a p-value cut-off of 0.01. C, The transcriptional start sites detected by 5'RACE in TBP MO and c MO injected embryos of the apoeb gene are compared to TSS data from public databases. Arrow indicates TSS verified by sequencing of 5'RACE products from TBP MO and c MO injected embryos. Transcriptional start sites of the apoeb gene as indicated by mapping of full length cDNAs are shown as blue bars. The shortest bar represents one incident of TSS detection. Chromosomal positions of the apoeb promoter region is indicated by numbers above and positions relative to the TSS in the promoter construct are shown below the gene depiction. D, The tbp gene is upregulated in TBP morphants. Whole mount in situ hybridization on zebrafish

embryos with a dig-labeled anti-tbp riboprobe {Bártfai, 2004 #186}. Embryos are shown in random orientation at dome stage.

the other hand, 12 promoters, including the *apoeb* gene promoter, did not show significant changes of activity upon loss of TBP function (Figure 2A and Figure S4). TBP independence of *apoeb* transcription was further confirmed by its mRNA levels (Supplementary Figure S1), and the utilisation of its TSS (data not shown) in TBP morphants. No correlation was found between known promoter motifs (such as TATA boxes, CpG islands, etc.) and TBP response (data not shown).

Several promoters (4 out of 23) showed a clear increase of promoter activity upon loss of TBP, including the 1.4-kb promoter of the *tbp* gene (Figure 2A and Figure S4). This finding suggests negative regulatory role of TBP on the *tbp* gene promoter and is in line with the inverse correlation between *tbp* mRNA and TBP protein levels at the late blastula and early gastrula stages (Bartfai et al, 2004; Figure S4D). Co-injection of a synthetic TBP MO-resistant *Xenopus* (*x*) *tbp* mRNA, but not of bacterial *IS30 transposase* control mRNA rescued the epiboly movements of the animal cap (Figure 2C–F, bright field view) and *tbp:yfp* activity (Figure 2C–F, fluorescence views). Finally, the injection of TBP MO2 resulted in comparable effects to TBP MO both in blocking epiboly movements and in the increased activity of the *tbp:yfp* promoter construct (Figure 2G). These results demonstrate that the specific loss of TBP protein is the reason for the observed upregulation of the *tbp* promoter in TBP MO-injected embryos.

TBP is required for degradation of a large number of maternal mRNAs

It is known that degradation of many maternal mRNAs involves zygotic transcription-dependent mechanisms, which may be specifically regulated by

TBP. Thus, the steady-state levels of maternal mRNAs may appear increased in TBP morphants. To test if the inhibition of the degradation of maternal mRNA occurs in TBP morphants, we searched for maternally expressed genes in the TBP morphant microarray gene sets.

We classified genes as being maternal or zygotic through another microarray experiment utilizing mRNA pre- and post- MBT; those showing a decrease of mRNA levels from pre- to post-MBT (MBT down) were classified as prevalently maternal and vice versa (MBT up) for prevalently zygotic ones (Supplementary Tables III and IV). We then compared this experiment with the TBP morphants data set, which resulted in an overlap of 131 genes (Supplementary Tables III and IV). The overlap showed that maternal mRNAs were enriched among the upregulated genes of TBP morphants (Figure 3A, MBT down) and the inverse was observed for zygotic mRNAs (Figure 3A, MBT up). A side by side hierarchical clustering analysis of gene activity fold changes in the MBT experiment versus the TBP MO experiment demonstrates further the inverse correlation between the levels of mRNAs before or after MBT as compared to mRNA levels in TBP morphants versus controls (Figure 3B).

We further verified our findings by intersecting the TBP MO microarray experiment with an independent set of 622 maternal mRNAs (Mathavan et al, 2005), which resulted in an overlap of 143 genes (Supplementary Table V). As shown in Figure 3C, maternally inherited transcripts were significantly (P-value 1.043e-11) enriched among mRNAs upregulated in TBP morphants and underrepresented in the downregulated gene set (P-value 3.483e-5). Together, these results suggest that the upregulation of genes observed in TBP morphants

could be in large part due to the specific loss of degradation of many maternal mRNAs.

Identification of TBP-dependent maternal transcripts

To validate the predicted involvement of TBP in the degradation of maternal mRNAs, we investigated the fate of individual maternal mRNAs. Zorba is a maternally expressed gene (Bally-Cuif et al, 1998), which is upregulated 2.51-fold in TBP MO embryos. We analysed the expression of zorba in wildtype and TBP-morphant embryos at regular intervals for the first 6 h of development by whole-mount in situ hybridisation (WISH). We found high levels of zorba expression in fertilised wild-type eggs and early embryos before MBT, followed by a sharp decrease soon after the MBT, followed by a slight increase at the dome stage (Figure 3D). In contrast, in TBP morphant embryos, zorba mRNA levels showed similar levels throughout early development, consistent with the assumption that degradation of maternal mRNA was impaired. We verified that the lack of degradation of zorba mRNA in TBP morphants was not due to a general delay in embryo development by observing the expression of two zygotically expressed genes: the TBP-independent gene no tail (Schulte- Merker et al, 1994), which correctly initiated transcription after the sphere stage in TBP morphants (Figure S5A); and the TBP-dependent goosecoid (Schulte-Merker et al, 1994), whose activity was lost in TBP morphants (Figure S5B). These results suggest efficient depletion of TBP at dome stage. To further verify the defect in maternal mRNA degradation, RT–PCR analysis was carried out on several maternally expressed genes that showed upregulation in TBP morphants. Zorba and smad2 (expressed both maternally and zygotically; Muller et al, 1999) showed elevated levels at dome stage in comparison to c MO-injected embryos,

suggesting loss of degradation of maternal mRNA, as opposed to the control gene bctin, which did not show a change in its steady-state levels (Figure 3E, lanes 1–4). RT–PCR analysis of zygotic genes was also carried out; the TBP-independent ntl showed no change in its mRNA levels, whereas zygotic activity of gsc dropped (data not shown) as shown previously by WISH.

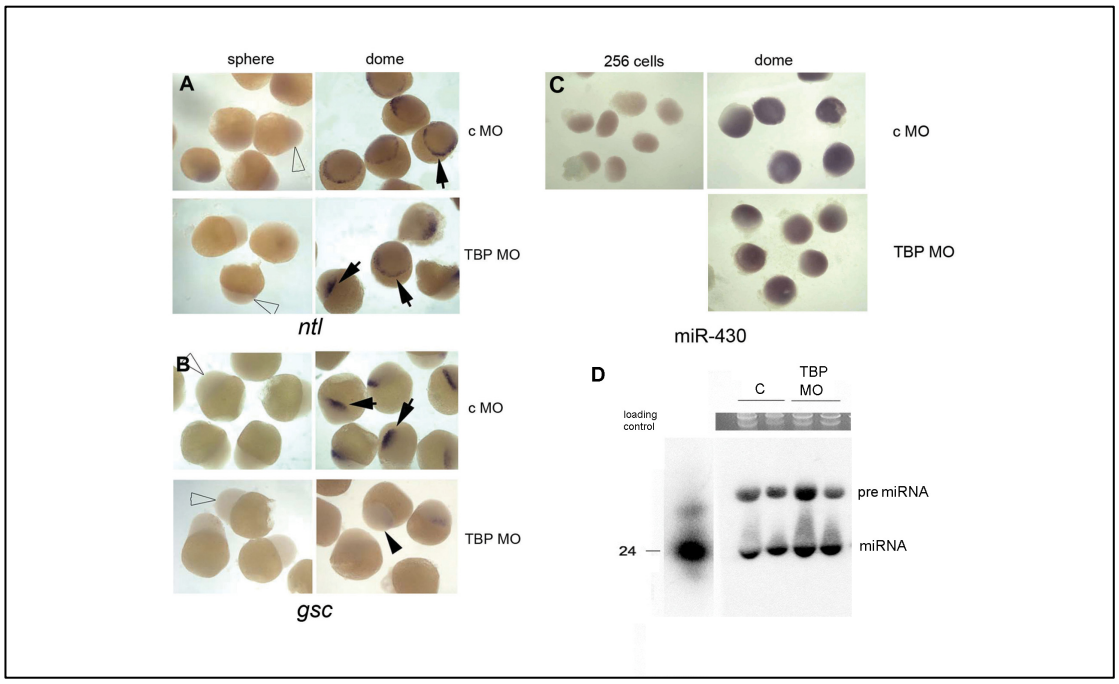


Figure S5 Degradation of maternal mRNA contributes to upregulation of genes in TBP morphants. A,B, whole mount in situ hybridisation with dig-labeled anti-ntl and gsc riboprobes. Embryos are shown at the sphere and dome stages in random orientation. Arrows indicate specific hybridization signals in post MBT embryos. Arrowheads point at embryos with no evidence of gene expression pre MBT. C, WISH analysis of miR-430 miRNA expression in zebrafish embryos using Dig labeled LNA oligonucleotide probe. Embryos are shown mostly animal pole up. D, Northern Blot to detect mature miR-430 transcripts in TBP MO injected embryos compared to wild type control. Molecular weight marker for miRNA is shown on the left.

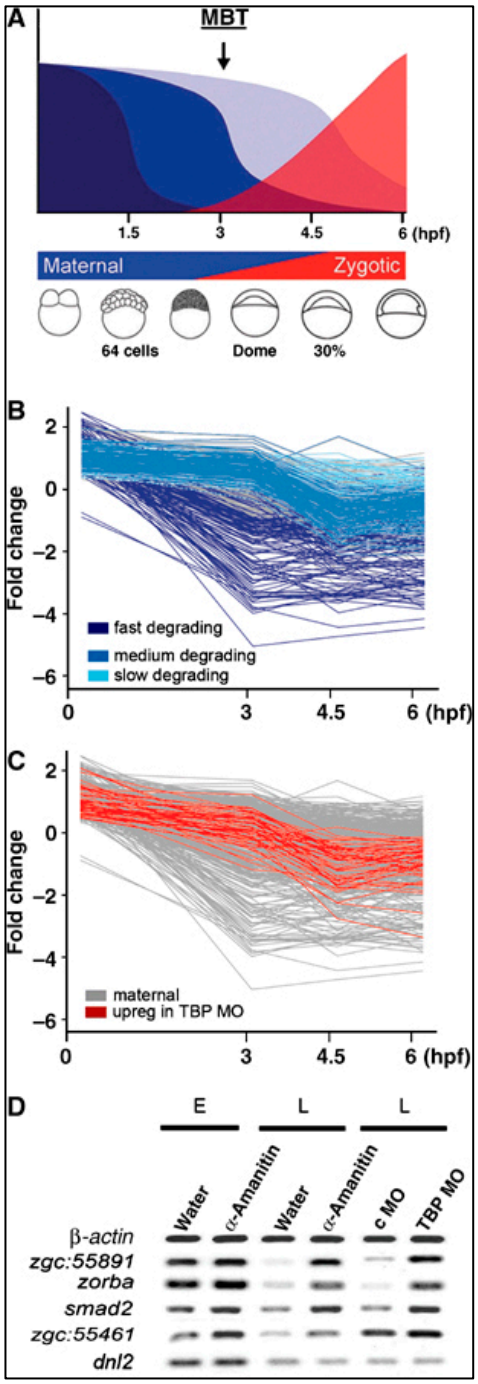
To test directly the fate of mRNAs deposited in the egg, we utilised a synthetic smad2 mRNA microinjected into the fertilised eggs (Figure 3E, smad2 (s)). This mRNA could be readily distinguished from endogenous smad2 by reducing the cycles in the RT–PCR reaction (Figure 3E, compare lanes 1–2 to 5–6 of smad2 (s)). Microinjected smad2 mRNA was more efficiently degraded in c MO- than in TBP MO-injected embryos (Figure 3E, compare lanes 6 and 8) and similar results

were obtained by WISH (data not shown). Thus, the apparent increase of smad2 mRNA levels in TBP morphants is not due to premature activation of zygotic smad2 expression, but due to the loss of degradation of smad2 mRNAs.

To verify the specificity of the maternal mRNA degradation phenotype to loss of TBP protein function, the ability of a MO-insensitive TBP mRNA to rescue the phenotype in TBP MO-injected embryos was tested. TBP MO and smad2 (s) co-injected embryos were split after injection and separate batches were injected for a second time either by xtbp mRNA or is30 mRNA. Expression of recombinant TBP resulted in increase of degradation of zorba (Figure 3E, compare lanes 7, 8 with 9, 10) as well as that of microinjected synthetic smad2 mRNA. In contrast, is30 tpase control mRNA did not result in rescue of the degradation phenotype of TBP morphants (Figure 3E, compare lanes 8 and 10). These results demonstrate that the effect of TBP MO on maternal mRNA degradation is directly attributable to the loss of TBP protein function.

TBP regulates a zygotic transcription-dependent mRNA degradation process

Little is known about the mechanisms of maternal mRNA degradation in zebrafish, however, it is likely to involve several maternal as well as zygotic transcription-dependent mechanisms. Not all maternal mRNAs were degraded in TBP morphants (Figure 3A and C). This may be due to different regulatory mechanisms acting in parallel during maternal mRNA degradation. To investigate this further, we verified the kinetics of mRNA degradation by exploiting a published microarray data set on maternal mRNAs (Mathavan et al, 2005) and compared it to our TBP morphant data set. We established three classes of mRNAs based on the time of their degradation (see Figure 4A and B



and Supplementary Table X): a 'fast' group of mRNAs, which degrade transcription independently or in a transcription dependent manner immediately after initiation of zygotic transcription; a 'medium' group which is mostly degraded after MBT by early gastrula stage; and a 'late' group degraded during neurulation and somitogenesis. The comparison of these groups with the TBP- morphant experiment (Figure 4C) showed that maternal mRNAs upregulated in TBP morphants follow the pattern of expression dynamics of the 'medium' group (P-value1/4 2.205e 06). TBP-dependent maternal mRNAs showed minimal degradation until MBT (3 h post-fertilisation (hpf)) and accelerated degradation by early gastrulation (4.5 hpf), suggesting that zygotic transcription-

Figure 4 TBP is required for the degradation of a subset of mater- nal mRNAs. (A) Schematic representation of early gene activities of the zebrafish embryo as shown in Figure 1A. (B) Degradation pattern of maternally accumulated genes during early zebrafish development until gastrulation (Mathavan et al, 2005). The 'fast' group of mRNAs, degraded before and immediately after MBT, is shown in dark blue), the 'medium' group, degraded after transcrip- tion starts at MBT, is shown in medium blue and the 'late' group, degraded during neurulation and somitogenesis, is shown in light blue. (C) Degradation pattern of genes upregulated in TBP Morphant embryos (red) in comparison to the degradation dy- namics of all maternal genes (grey). (D) Degradation of maternal RNA in control and a-amanitin-injected embryos as compared to c MO- and TBP MO-injected embryos before MBT and after MBT. Abbreviations, as in Figure 3.

dependent mechanisms are involved in their degradation.

These results suggest that TBP is only acting on a subset of mRNAs and that these mRNAs require transcription for their degradation. To test the transcriptional requirement for degradation of maternal mRNAs, we treated embryos with amanitin at a concentration that inhibits Pol II activity (Muller et al, 2001) and carried out RT–PCR analysis of gene expression. High levels of zorba, zgc:55891 and smad2 mRNAs were retained after MBT in amanitin-injected embryos similarly to TBP morphants (Figure 4D), demonstrating that these mRNAs require transcription and TBP for their degradation. These results taken together with the results obtained using synthetic smad2 mRNAs (Figure 3E) suggest that the sustained levels of maternal mRNAs in TBP morphants are due to the inhibition of maternal mRNA degradation rather than ectopic activation of zygotic transcription of the respective genes. Not all maternal mRNAs require zygotic transcription (and TBP) for their degradation as confirmed by the fact that degradation of the dnl2 gene is unaffected by amanitin and TBP MO (Figure 4D) and its degradation is primarily mediated by mechanisms acting before MBT (Figure 4C). Thus, TBP appears to function within a transcription dependent mechanism directing the degradation of a subset of maternal mRNAs eliminated in a tight time window after MBT during early gastrulation.

Degradation of maternal mRNA by the miR-430 microRNA is specifically affected in TBP morphants

Recently, a novel mechanism for maternal mRNA degradation has been described, which involves the zygotically transcribed miR-430 microRNA (Giraldez et al, 2006). Importantly, the maternally inherited zorba and smad2

transcripts have been shown to be targets of miR-430 regulation (Giraldez et al, 2006), suggesting a potential link between TBP and miR-430 function in mRNA degradation. Therefore, we explored the relationship between miR-430- and TBP- dependent mRNA degradation mechanisms.

We first addressed the question whether TBP-dependent maternally inherited transcripts represent targets of miR-430- mediated mRNA degradation in general. We compared the overlap between experimentally verified miR-430 target genes (Giraldez et al, 2006) and our TBP morphant microarray gene sets (see Supplementary Tables XI, XII and Materials and methods). As shown in Figure 5A, a significant enrichment of miR-430 target genes among the upregulated genes of TBP morphants was observed ($P = 0.002074$). The proportion of miR-430 target genes was found to be higher among TBP-upregulated genes (20%) than among maternal genes in general (14%, $P = 0.0276$). This difference is significant also after 100 randomisation experiments in which we randomly selected 100 maternal genes (15% s.d. = 3, $P = 0.0507$). This suggests that the enrichment of miR-430 targets among the upregulated genes of TBP morphants is not simply reflecting the high proportion of maternal genes among miR-430 targets and upregulated genes in TBP morphants.

Subsequently, we have checked the degradation patterns of miR-430 target genes by analysing the overlap between the maternal genes with known degradation kinetics (Mathavan et al, 2005) and miR-430 target genes (Supplementary Table VI). The degradation kinetics of miR-430 target genes largely but not exclusively overlap with that of the maternal genes degraded by

TBP-dependent mechanisms (Figure 5B). Taken together, these results indicate a correlation between miR-430- and TBP-dependent mRNA degradation.

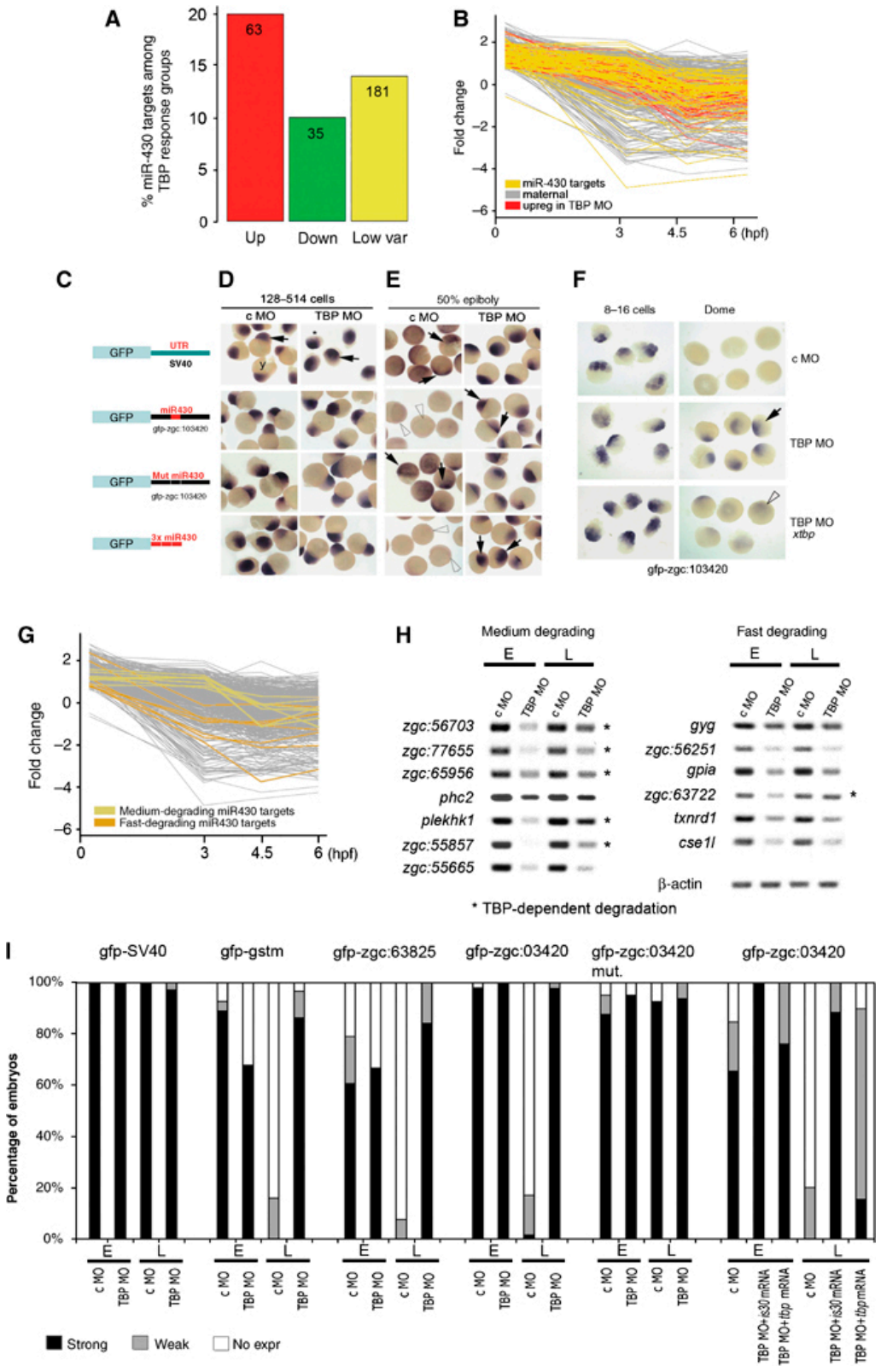


Figure 5 TBP is required for miR-430-dependent maternal mRNA degradation. (A) Distribution of miR-430 target genes among the different response groups of genes regulated in TBP Morphant embryos is shown as percentage of the respective TBP MO response groups. The number of overlapping genes between the two microarray data sets compared is indicated in the columns. (B) Degradation dynamics of miR-430 target maternal mRNAs during early zebrafish development (yellow) in comparison to the maternal mRNAs upregulated in TBP Morphants (red) and all maternal mRNAs (grey). (C–F) MiR-430 target mRNAs are degraded by a TBP-dependent mechanism. Synthetic mRNAs injected into zebrafish embryos are shown (A). Microinjected GFP mRNAs are detected by WISH (arrows) in randomly oriented representative groups of early embryos before MBT (D) and several hours after MBT). (F) Injection of TBP mRNA rescues the mRNA degradation phenotype. Detection of microinjected synthetic gfp-zgc:103420 mRNA by WISH using a gfp antisense probe. Embryos are shown at the indicated stages in random orientation. (G) Degradation dynamics of maternal miR-430 target genes. Representative examples of 'fast'- and 'medium'-degrading mRNAs used in RT–PCR analyses (H) are shown. (H) RT–PCR analysis of the mRNAs at 16-cell stage and at 30% epiboly stage in embryos injected with c MO or TBP MO. (I) Degradation of synthetic mRNAs injected in zebrafish embryos. Embryos were injected with mRNAs indicated above the bar chart. Percentage of embryos with different signal levels after in situ hybridisation using a gfp probe are shown. The numbers of embryos injected are shown in Supplementary Table III. Abbreviations are as in Figure 3.

Redundant and specific function of TBP in the activation of subsets of genes at MBT

We next tested whether miR-430-dependent mRNA degradation requires TBP function. Embryos were injected at the zygote stage with a combination of synthetic mRNAs containing UTR sequences with or without miR-430 target sites (Giraldez et al, 2006) together with TBP MO or c MO, and mRNA distribution was detected by WISH before and after the MBT. A synthetic gfp mRNA containing the 30 UTR region from SV40 that lacks miR-430 target sequences was not degraded until the 50% epiboly stage (Figure 5D, E and I and Supplementary Table VIII), suggesting that they are not degraded by the miR-430 pathway. In contrast, gfp mRNA linked to the UTR from the zgc:103420 gene containing a miR- 430 target site is degraded by the 50% epiboly stage in c MO injected embryos, but not in TBP MO-injected embryos (Figure 5C–E and I). Similar results were obtained with gfp mRNA fused to the 30 UTR sequences of the zgc:63825 and gstm genes containing miR-430 target sites or when only the miR430 target site sequences were added to gfp (Figure 5C–E and I). These results suggest that TBP is required for miR-430- dependent degradation of

several mRNAs. Upon injection of a transcript containing a mutated 30 UTR sequence of zgc:103420 lacking the miR-430 target site, the mRNA became insensitive to degradation until the 50% epiboly stage (Figure 5C–E and I), indicating that the degradation of these synthetic mRNAs are indeed miR-430-dependent. The defects in miR- 430-dependent degradation of mRNA in TBP morphants was also rescued by overexpression of recombinant TBP. Coinjection of synthetic tbp mRNA with *gfp-zgc:103420* and TBP MO resulted in reversal of the mRNA degradation phenotype of TBP morphants, indicating that the mRNA degradation effects relate directly to loss of TBP (Figure 5F and I and Supplementary Table VIII). As miRNA genes are known to be transcribed by polymerase II (Lee et al, 2004), miR-430 may be a candidate target of TBP-dependent transcription regulation. However, miR-430 expression in TBP morphants is unaffected (Supplementary Figure S5D), suggesting that TBP functions downstream of miR-430 production in the mRNA degradation process. We investigated further whether TBP is involved in general microRNA function or the mir-430 pathway specifically. Thus, we co-injected miR-430 and miR-1 with their respective target mRNAs (Giraldez et al, 2006) with TBP and c MO.

Discussion

In summary, we have demonstrated a differential requirement for TBP in the regulation of mRNA levels in the early zebrafish embryo and pinpointed three levels of regulatory activities associated with TBP function. The approach to block TBP function by MO oligonucleotides resulted in the efficient depletion of TBP in the embryo before the MBT, and thus allowed the detection of the earliest effects of loss of TBP protein on zygotic transcription and associated maternal

degradation processes. The extensive use of a second TBP targeting MO and rescue experiments with a MO-insensitive recombinant TBP in this study provided strong experimental verification of the specificity of our key findings to loss of TBP function.

Redundant and specific function of TBP in the activation of subsets of genes at MBT

We next tested whether miR-430-dependent mRNA degradation requires TBP function. Embryos were injected at the zygote stage with a combination of synthetic mRNAs containing UTR sequences with or without miR-430 target sites (Giraldez et al, 2006) together with TBP MO or c MO, and mRNA distribution was detected by WISH before and after the MBT. A synthetic *gfp* mRNA containing the 30 UTR region from SV40 that lacks miR-430 target sequences was not degraded until the 50% epiboly stage (Figure 5D, E and I and Supplementary Table VIII), suggesting that they are not degraded by the miR-430 pathway. In contrast, *gfp* mRNA linked to the UTR from the *zgc:103420* gene containing a miR- 430 target site is degraded by the 50% epiboly stage in c MO injected embryos, but not in TBP MO-injected embryos (Figure 5C–E and I). Similar results were obtained with *gfp* mRNA fused to the 30 UTR sequences of the zgc:63825 and gstm genes containing miR-430 target sites or when only the miR430 target site sequences were added to *gfp* (Figure 5C–E and I). These results suggest that TBP is required for miR-430- dependent degradation of several mRNAs. Upon injection of a transcript containing a mutated 30 UTR sequence of zgc:103420 lacking the miR-430 target site, the mRNA became insensitive to degradation until the 50% epiboly stage (Figure 5C–E and I), indicating that the degradation of these synthetic mRNAs are indeed miR-430-

dependent. The defects in miR- 430-dependent degradation of mRNA in TBP morphants was also rescued by overexpression of recombinant TBP. Coinjection of synthetic tbp mRNA with *gfp-zgc:103420* and TBP MO resulted in reversal of the mRNA degradation phenotype of TBP morphants, indicating that the mRNA degradation effects relate directly to loss of TBP (Figure 5F and I and Supplementary Table VIII). As miRNA genes are known to be transcribed by polymerase II (Lee et al, 2004), miR-430 may be a candidate target of TBP-dependent transcription regulation. However, miR-430 expression in TBP morphants is unaffected (Supplementary Figure S5D), suggesting that TBP functions downstream of miR-430 production in the mRNA degradation process. We investigated further whether TBP is involved in general microRNA function or the mir-430 pathway specifically. Thus, we co-injected miR-430 and miR-1 with their respective target mRNAs (Giraldez et al, 2006) with TBP and c MO. miR-430-mediated mRNA degradation was blocked in TBP morphants, as opposed to that by miR-1, confirming the specificity of TBP function to a subset of miRNA-dependent processes (Figure S6).

Given the tight temporal control of TBP-dependent mRNA degradation, we hypothesised that those miR-430 target mRNAs are degraded in a TBP-dependent manner, which are eliminated at a 'medium' rate during late blastulation/ early gastrulation. To test this hypothesis, we utilised the overlap between miR-430 target genes and maternal genes and identified those, which show either medium or fast degradation (Figure 5G). Then we tested both fastand medium-degrading miR430 target genes for TBP dependence of their degradation. The results indicate that medium-degrading mRNAs are more likely

to be TBP-dependent (increased accumulation in TBP MO injected embryos) than fast-degrading mRNAs (5 of 7 versus 1 of 6, respectively, Figure 5H). Taken together, our results indicate that TBP is specifically required for the degradation of a subset of miR- 430-dependent mRNAs that are eliminated at late blastula/ early gastrula stages.

In this study of the transcriptome of the early zebrafish embryo, we have found that the expression of most genes remains weakly or not affected in TBP morphants. It is tempting to speculate that the TBP paralogue TBP2/TRF3, which has similar DNA-binding properties to TBP and is expressed at high levels in ovaries (Bartfai et al, 2004), may contribute to the control of steady-state levels of mRNAs before gastrulation, thus complementing TBP function. However, due to the presence of maternally inherited TBP2 protein in the early zebrafish embryo, MO knockdown of TBP2 is inefficient before gastrulation (Bartfai et al, 2004) and new ways of interfering with TBP2 function in the oocyte will have to be developed to address TBP2 function in the early zebrafish embryo. Another TBP-related factor (TLF/ TLP/TRF2) has also been shown to affect transcription regulation (Veenstra et al, 2000; Muller et al, 2001). Together, these alternative transcription initiation mechanisms may explain the large number of unaffected gene activities in TBP morphants and emphasise the need to define the boundaries of TBP-regulatory mechanisms.

The differential regulation of genes during early development by TBP raises the question of which genes are specifically regulated by TBP and what promoter properties do they possess. Among genes requiring TBP for their activation, there was an enrichment for genes with stage-dependent activity during

ontogeny. This observation suggests that TBP is required for genes that are expressed in a tightly regulated manner and demonstrates an important regulatory

role for TBP as a core promoter-binding factor during ontogeny. Does the correlation between TBP function and the type of genes regulated by TBP mean a correlation between TBP dependence and core promoter motif composition? The recent genome-wide analysis of mouse and human promoters revealed that the binding site for TBP, known as the TATA box, is more likely to be present in genes associated with tightly regulated transcription and tissue specificity (Carninci et al, 2006). Our results, which show that ontogenic stage- dependent genes preferentially require TBP, are in line with the observations in mammals. However, our failure to detect a direct correlation between the presence of a TATA box and TBP dependence of gene activation in zebrafish is possibly due to the small set of genes available for this analysis. The small number of promoters analysed combined with the very low number of TATA boxes encountered in zebrafish promoters (for less than 10% of genes verified, unpublished data) is consistent with findings in mammals (Carninci et al, 2006) and hampers the establishment of statistically significant correlations.

TBP limits certain gene expression activities in the zebrafish embryo

The striking presence of upregulated genes in TBP morphants could be attributed to two distinct mechanisms: the direct or indirect negative regulatory role of TBP on zygotic transcription as well as the block of maternal mRNA degradation. In this paper, we show evidence for both mechanisms, thus

revealing the complexity of regulatory roles played by TBP during the maternal to zygotic transition in zebrafish.

The observed negative regulatory role of TBP in zygotic gene activation is consistent with recent observations made in cultured cells. Inverse correlation between TBP occupancy and gene activity has been reported recently in Drosophila cells (Lebedeva et al, 2005). Moreover, TBP was found to inhibit transcription of the NF1 promoter in transfection experiments (Chong et al, 2005). Thus, published biochemical evidence (Chong et al, 2005) together with our observations on the biological roles of TBP together make it feasible to hypothesise that TBP may repress promoters directly during embryogenesis. To test this possibility, as well as to assess the direct effects of TBP in early development, will require the establishment of assays of promoter occupancy by TBP in vivo during zebrafish embryonic development.

The mRNA degradation machinery active during maternal to zygotic transition requires TBP function

The temporally regulated degradation of maternal mRNAs at and after the MBT is part of a general mechanism to regulate the maternal to zygotic transition and is required for normal development of the zebrafish embryo (Giraldez et al, 2006; Schier, 2007). We showed here that TBP is required for a zygotic developmental programme, which facilitates the regulated degradation of a subset of maternally deposited mRNAs during late blastula/early gastrula stages in the zebrafish. This finding is one of the few currently available explanations for the interplay between zygotic transcription and mRNA degradation during early embryogenesis (De Renzis et al, 2007).

Where does TBP function in maternal mRNA degradation? The most likely scenario is that TBP is required for the expression of a zygotic gene product, which acts in the mRNA degradation mechanism. The components of the mRNA degradation pathway(s) in the embryo that may be affected by TBP function are largely unknown. However, our analysis of the temporal regulation of maternal mRNA degradation allowed us to characterise the set of mRNAs that require TBP for their elimination and provides useful tools for future research into the molecular mechanisms involved in the control of mRNA degradation by transcription-dependent mechanisms. We concluded that TBP-dependent mRNAs are specifically degraded after the MBT but before the end of gastrulation in a transcription-dependent manner. The zygotically expressed miR-430 becomes active at the MBT and was recently shown to regulate the degradation of many mRNAs in zebrafish (Giraldez et al, 2006). In our study, TBP was found to affect miR-430 target mRNAs but only a subset of them. Importantly, we were able to pinpoint the set of miR- 430-dependent maternal mRNAs degraded during late blastula early gastrula that require TBP. Thus, TBP may act on yet unidentified components of a pathway feeding into or functioning genetically downstream to the miR-430-dependent maternal mRNA degradation pathway. The pursuit for other factors in this pathway, including other potential TBP-dependent miRNAs, remains an urgent subject for future investigation. It will be important to address whether the restricted role TBP plays in the degradation of a subset of mRNAs can be associated with a particular developmental programme such as the various aspects of spatial and temporal control of gene activities during gastrulation. Another interesting question is whether TBP-dependent degradation of maternal mRNAs is a general feature of

vertebrate development. Whereas it is premature to make direct comparisons with other vertebrates, there are indications for similarities in mRNA degradation dynamics. For example, smad2 is a maternally expressed gene in the mouse with a similar early degradation pattern to that described here for zebrafish (Wang et al, 2004).

The role of TBP in regulating specific subsets of genes during early embryo development demonstrates that promoter recognition proteins, once considered as constitutive components of transcription initiation, play specialised roles in differential gene expression regulation and may explain the diversity of core promoters (Butler and Kadonaga, 2002). Our findings support the model that promoter recognition by general transcription factors represents an additional regulatory level in transcription, which may provide flexibility for the cells of the animal to respond to signals at the core promoter level during ontogeny.

Materials and methods

Embryo injection experiments

Wild-type embryos (Tubingen AB) were collected after fertilisation and dechorionated by pronase treatment as described previously (Westerfield, 1995). The TBP-specific MO (TBP MO) used was described previously (Muller et al, 2001). TBP2 MO was CAAAAGACGTAAACGATAATTCGCA. Embryos were injected also with a c MO with five mismatches relative to the TBP-specific MO (GACGTACGCTGTTCTTCTCCTCGAT) or a standard c MO (CCTCTTACCTCAGTTACAATTTATA) provided by Gene-tools LLC (Phylomath, USA). MOs were injected into the cytoplasm of zebrafish embryos at the one cell stage. A total of 1200 embryos were collected for microarray analysis at the

dome stage. Embryos were co-injected with combinations of MOs and mRNAs for analysis of gene expression. For synthetic *tbp* mRNA production, the *pCS2+xcTBP* was used containing as a *XhoI–XbaI* fragment of the biologically active core domain (aa 104–297) (Schmidt et al, 2003) of the Xenopus laevis TBP (Veenstra et al, 1999). The *pFOL876* plasmid in a *pCS2+* vector with an *Escherichia coli IS30 transposase* gene was used to make control mRNA (Szabo et al, 2003). Both mRNAs were injected at 100ng/ml concentration. mRNAs were produced by in vitro transcription of linearised plasmids using the mMESSAGE mMACHINE Kit (Ambion, UK). Plasmids for the production of synthetic *gfp* mRNAs with various miR-430 target sequences were described by Giraldez et al (2006). In co-injection experiments, MOs were co-injected with 20ng/ml *gfp*-UTR mRNAs followed by injection with tbp or is30 tpase mRNAs.

Twenty-three promoter constructs were made which contained 1–2 kb upstream and 30–300 bp downstream sequence around the predicted (9 promoters) or 50 RACE verified (14 promoters) transcriptional start site (TSS) linked to a *GFP* reporter. Promoter fragments were amplified from genomic DNA using TripleMaster Enzyme Mix (Eppendorf, Germany) and cloned into the pGlow-TOPO vector (Invitrogen, Germany). The promoter constructs were injected into the cytoplasm of one-cell stage embryos. At 50% epiboly, embryos were fixed overnight in 4% PFA. TBP-dependence analysis of promoters was carried out by co-injecting MOs in a concentration of 1mM with 100ng/ml of mRNA. For recombinant TBP rescue experiments, the tbp promoter was PCR amplified and subcloned into the *pCS2 + yfp* replacing the CMV promoter by using SalI and NcoI sites. Yellow fluorescence protein was visualised by using a Leica MZ16F

fluorescent stereo microscope and digital image recording. Western blotting with the 3G3 antibody against TBP was carried out as described previously (Muller et al, 2001).

Whole-mount in situ hybridisation and immunostaining

Fixed embryos were incubated overnight at 41C with wild-type *GFP* (1:500) (Torrey Pines Biolabs, USA) or cycle3 *GFP* antibodies (1:200) (Invitrogen, Germany). Embryos were incubated with goat- anti-rabbit HRP secondary antibody in PBT (1:500) (DakoCytomation, Denmark) for detection by diaminobenzidine solution according to manufacturer's instruction (Vector, USA). An LNA oligonucleotide probe was used in WISH to detect miR-430 miRNA as described previously (Giraldez et al, 2006). WISHs were carried out with standard protocols (Westerfield, 1995).

RT–PCR analysis of maternal mRNA degradation

Analysis of a selected set of genes by RT–PCR was carried out using 50 embryos for each treatment group. Microinjected and wild-type embryos were collected at the 512-cell stage and shield stage. Total RNA was extracted using Trizol (Invitrogen) following manufacturer instructions. A 1mg portion of total RNA was reverse transcribed (M-MLV Reverse Transcriptase, Promega, Germany) and PCR was carried out from several targets using the oligonucleotide primers specified in Supplementary Table IX. PCR products were separated in a 2% agarose gel.

Gene identification and statistical analysis of EST microarray data

Three developmental stages around MBT were analysed on microarrays containing the primary set of 16177 oligos (16k set). RNA was obtained from

freshly laid eggs (zygotic target), 1k-stage embryos (1k target) and 30% epiboly embryos (30% epiboly target). The following hybridisations were performed with dye swap each: zygotic versus 1k, zygotic versus 30% epiboly and 1k versus 30% epiboly. Microarray chips representing 10501 non- redundant Genbank ESTs of the primary 16416 set of 65-mer oligonucleotides were used in the subsequent hybridisations. Two independent hybridisations were carried out for three biological repeats of treatment groups (TBP MO- versus c MO-injected embryos), resulting in a maximum of 12 data points per gene. The reliability of the fold changes was assessed using a regularised t-test (Baldi and Long, 2001) and the adjustment of P-values to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995) selecting genes with FDR smaller than 0.05 (minimum 5 data points). This cut-off value of FDR is the minimal widely used (Shi et al, 2006) and results in a good balance of quality versus quantity in the selected gene lists.

Annotation of ESTs of the TBP microarray, in relation to the stage-dependence array and to the zebrafish genome

Several metanalyses of different, unrelated microarray experiments were carried out as has been described previously (Cavallo et al, 2005). Gene sets of the TBP morphant microarray were compared with existing gene sets (Mathavan et al, 2005) generated on the same platform (Compugen microarray), with the experiment GSE4201, http://www.ncbi.nlm.nih.gov/projects/geo (Giraldez et al, 2006) and the experiment E-TABM-33 from http://www. ebi.ac.uk/arrayexpress (Konantz M, Otto G-W, Weller C, Saric M, Geisler R. Microarray analysis of gene expression in zebrafish development, manuscript in preparation) executed on the Affymetrix platform (See Supplementary data for details). The outdated gene

annotation by Compugene (Mathavan et al, 2005) was updated by converting all Entrez Nucleotide identifiers of the probes to their respective Unigene identifiers, using a custom Perl script and release 91 of Unigene. The heatmap and hierarchical clustering of microarray experiments was created using the R programming language version 2.2.0. The clustering of fold changes is based on the 'complete linkage' method provided by R. See http://cran.r-project.org/ for details.

Degradation pattern of maternal transcripts

Maternal genes present in the TBP-knockdown and miR-430 targets gene sets (Giraldez et al, 2006) were identified by utilizing an existing data set of transcripts accumulated in the unfertilised egg (Mathavan et al, 2005). The variation of steady-state mRNA levels over developmental time was determined by comparing fold changes as described in (Mathavan et al, 2005) and was visualised by plotting their values in the unfertilised egg and at 3, 4.5 and 6 hpf). For a clustering of maternal mRNAs into fast-, medium- and slow-degrading subgroups the following criteria were used: fast- degrading, >200% decrease of fold change from 0 to 3hpf; medium–degrading, <100% from 0 to 3hpf, >100% from 3 to 4.5 hpf; slow-degrading, <50% decrease of fold change till 4.5 hpf).

Identification of miR-430 targets among the genes of the TBP microarray

We downloaded the raw data of the experiments (Giraldez et al, 2006; wild type, MZ Dicer and MZ Dicer miR-430-injected) from the GEO database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc1/4GSE4201). The probe summaries were generated by using RMA from the BioConductor collection of packages. The Limma package was used to perform the statistical analysis and

select differentially expressed genes. The following comparisons were made: MZ Dicer versus wild type; MZ Dicer versus MZ Dicer miR-430–injected; wild type versus MZ Dicer miR-430-injected. We selected as differentially expressed genes all the genes resulting in at least one comparison with a corrected P-value p0.05. All the genes significantly upregulated in MZ DICER compared to both wild type and MZ DICER miR-430-injected were considered as potential miR- 430 target (1650 probes).

Acknowledgements

We thank L Tora for TBP antibodies, discussions and critical reading of this manuscript. We also thank NS Foulkes and M-E Torres Padilla for critical comments and R Caogero and F McNish for their advice. We acknowledge S Schindler and N Borel for technical assistance and T Dalmay and T Rathjen for their help in detecting miR-430 in embryos. We thank A Giraldez and A Schier for sharing GFP constructs containing miR-430 targets, and L Byrnes, R Bree, S Goyle and R Geisler for sharing unpublished data. This work was supported by funds from the DFG (MU 1768/2) to FM and by the EU (contract 511990) to FM and ES and the BMBF to FM and FO and an ESF travel grant to MF.

References

Almouzni G, Wolffe AP (1995) Constraints on transcriptional activator function contribute to transcriptional quiescence during early Xenopus embryogenesis. EMBO J 14: 1752–1765

Audic Y, Anderson C, Bhatty R, Hartley RS (2001) Zygotic regulation of maternal cyclin A1 and B2 mRNAs. Mol Cell Biol 21: 1662–1671

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17: 509–519

Bally-Cuif L, Schatz WJ, Ho RK (1998) Characterization of the zebrafish Orb/CPEB-related RNA binding protein and localization of maternal components in the zebrafish oocyte. Mech Dev 77: 31–47

Bartfai R, Balduf C, Hilton T, Rathmann Y, Hadzhiev Y, Tora L, Orban L, Muller F (2004) TBP2, a vertebrate-specific member of the TBP family is required in embryonic development of zebrafish. Curr Biol 14: 593–598

Bashirullah A, Halsell SR, Cooperstock RL, Kloc M, Karaiskakis A, Fisher WW, Fu W, Hamilton JK, Etkin LD, Lipshitz HD (1999) Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster. EMBO J 18: 2610–2620

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57: 136–143

Butler JE, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev 16: 2583–2592

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y et al (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635

Cavallo F, Astolfi A, Iezzi M, Cordero F, Lollini PL, Forni G, Calogero R (2005) An integrated approach of immunogenomics and bioinformatics to identify new tumor associated antigens (TAA) for mammary cancer immunological prevention. BMC Bioinformatics 6 (Suppl 4): S7

Chong JA, Moran MM, Teichmann M, Kaczmarek JS, Roeder R, Clapham DE (2005) TATA-binding protein (TBP)-like factor (TLF) is a functional regulator of transcription: reciprocal regulation of the neurofibromatosis type 1 and c-fos genes by TLF/TRF2 and TBP. Mol Cell Biol 25: 2632–2643

Davidson I (2003) The genetics of TBP and TBP-related factors. Trends Biochem Sci 28: 391–398

De Renzis S, Elemento O, Tavazoie S, Wieschaus EF (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol 5: e117

Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, Schier AF (2005) MicroRNAs regulate brain morphogenesis in zebrafish. Science 308: 833–838

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science 312: 75–79

Jallow Z, Jacobi UG, Weeks DL, Dawid IB, Veenstra GJ (2004) Specialized and redundant roles of TBP and a vertebrate-specific TBP paralog in embryonic gene regulation in Xenopus. Proc Natl Acad Sci USA 101: 13525–13530

Kane DA, Kimmel CB (1993) The zebrafish midblastula transition. Development 119: 447–456

Kimelman D, Kirschner M, Scherson T (1987) The events of the midblastula transition in Xenopus are regulated by changes in the cell-cycle. Cell 48: 399–407

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF (1995) Stages of embryonic development of the zebrafish. Dev Dyn 203: 253–310

Lebedeva LA, Nabirochkina EN, Kurshakova MM, Robert F, Krasnov AN, Evgen'ev MB, Kadonaga JT, Georgieva SG, Tora L (2005) Occupancy of the Drosophila hsp70 promoter by a subset of basal transcription factors diminishes upon transcriptional activation. Proc Natl Acad Sci USA 102: 18087–18092

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23: 4051–4060

Martianov I, Viville S, Davidson I (2002) RNA polymerase II transcription in murine cells lacking the TATA binding protein. Science 298: 1036–1039

Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H, Ruan Y, Korzh V, Gong Z, Liu ET, Lufkin T (2005) Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet 1: 260–276

Moore PA, Ozer J, Salunek M, Jan G, Zerby D, Campbell S, Lieberman PM (1999) A human TATA binding protein-related protein with altered DNA binding specificity inhibits transcription from multiple promoters and activators. Mol Cell Biol 19: 7610–7620

Muller F, Blader P, Rastegar S, Fischer N, Knochel W, Strahle U (1999) Characterization of zebrafish smad1, smad2 and smad5: the amino-terminus of smad1 and smad5 is required for specific function in the embryo. Mech Dev 88: 73–88

MullerF, Lakatos L, Dantonel J, Strahle U, Tora L (2001) TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish. Curr Biol 11: 282–287

Newport J, Kirschner M (1982) A major developmental transition in early Xenopus embryos: II. Control of the onset of transcription. Cell 30: 687–696

O'Boyle S, Bree RT, McLoughlin S, Grealy M, Byrnes L (2007) Identification of zygotic genes expressed at the midblastula transition in zebrafish. Biochem Biophys Res Commun 358: 462–468

Pelegri F (2003) Maternal factors in zebrafish development. Dev Dyn 228: 535–554

Persengiev SP, Zhu X, Dixit BL, Maston GA, Kittler EL, Green MR (2003) TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed. Proc Natl Acad Sci USA 100: 14887–14891

Prioleau MN, Huet J, Sentenac A, Mechali M (1994) Competition between chromatin and transcription complex assembly regulates gene expression during early development. Cell 77: 439–449

Schier AF (2007) The maternal-zygotic transition: death and birth of RNAs. Science 316: 406–407

Schmidt EE, Bondareva AA, Radke JR, Capecchi MR (2003) Fundamental cellular processes do not require vertebrate-specific sequences within the TATA-binding protein. J Biol Chem 278: 6168–6174

Schulte-Merker S, Hammerschmidt M, Beuchle D, Cho KW, De Robertis EM, Nusslein-Volhard C (1994) Expression of zebrafish goosecoid and no tail gene products in wild-type and mutant no tail embryos. Development 120: 843–852
Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM et al (2006) The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24: 1151–1161

Szabo M, Muller F, Kiss J, Balduf C, Strahle U, Olasz F (2003) Transposition and targeting of the prokaryotic mobile element IS30 in zebrafish. FEBS Lett 550: 46–50

Teichmann M, Wang Z, Martinez E, Tjernberg A, Zhang D, Vollmer F, Chait BT, Roeder RG (1999) Human TATA-binding protein- related factor-2 (hTRF2) stably associates with hTFIIA in HeLa cells. Proc Natl Acad Sci USA 96: 13720–13725

Thompson EM, Legouy E, Renard JP (1998) Mouse embryos do not wait for the MBT: chromatin and RNA polymerase remodeling in genome activation at the onset of development. Dev Genet 22: 31–42

Veenstra GJ, Destree OH, Wolffe AP (1999) Translation of maternal TATA-binding protein mRNA potentiates basal but not activated transcription in Xenopus embryos at the midblastula transition. Mol Cell Biol 19: 7972–7982

Veenstra GJ, Weeks DL, Wolffe AP (2000) Distinct roles for TBP and TBP-like factor in early embryonic gene transcription in Xenopus. Science 290: 2312–2315

Wagner DS, Dosch R, Mintzer KA, Wiemelt AP, Mullins MC (2004) Maternal control of development at the midblastula transition and beyond: mutants from the zebrafish II. Dev Cell 6: 781–790

Wang QT, Piotrowska K, Ciemerych MA, Milenkovic L, Scott MP, Davis RW, Zernicka-Goetz M (2004) A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. Dev Cell 6: 133–144

Westerfield M (1995) The zebrafish book. Eugene: University of Oregon Press

Chapter 5: Assembly of the carp genome

Elia Stupka[1,2] , Yanju Zhang[3], Chrstian Henkel[4], Hans Jansen[4], Geert Wiegertjes[5], Maria Forlenza[5], Ron Dirks[4], Herman P Spaink[6], Fons J Verbeek[3]

1 UCL Cancer Institute, University College London, Gower Street, London, WC1E 6BT, United Kingdom

2 Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, 4 Newark Street, Whitechapel, London, E1 2AT, United Kingdom

3 Leiden Institute of Advanced Computer Science (LIACS), Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

4 ZF-Screens BV, Bio Partner Center, Niels Bohrweg 11, 2333 CA Leiden, The Netherlands

5 Cell Biology and Immunology Group, Wageningen Institute of Animal Sciences, Wageningen University, Zodiac, Marijkeweg 40, 6700 AH Wageningen, The Netherlands

6 Department of Molecular Cell Biology, Universiteit Leiden, Einthovenweg 20, 2333 ZC, Leiden, The Netherlands

In preparation for publication

Abstract

In this study we present the assembly of the common carp (*Cyprinus carpio*) genome. We utilized only next-generation sequencing technologies applied to a combination of standard short DNA libraries as well as mate-pair libraries. We assessed several de novo assemblers (Abyss, SOAPdenovo, CLCBio) and parameters. Our final assembly was obtained by using CLCBio for contig assembly, and SOAPdenovo for scaffold assembly. We were thus able to assemble a genome of 1.647Gb, well in line with the estimated genome size for this organism, of which ~300Mbs which are found in gaps. The final scaffold N50 produced is of ~8Kb, thus presenting many scaffolds with complete gene structures. The largest scaffolds present very good collinearity with the zebrafish genome, and the mitochondrial genome has been completely covered in a single scaffold. Utilizing over 2,000 carp sequences available in Genbank (mostly gene fragments) we were able to show that the majority of sequences had 100% coverage in the current assembly, and most of them were recovered in a single scaffold. Based on Genbank carp sequences the assembly shos 99.5% coverage of existing data. Using our own contig obtained by assembling RNA-Seq data we were able to confirm that the assembly provides very good coverage of carp gene content (RNA-Seq contigs had median coverage of 98.7% and average coverage of 92.47%). However this data, which presents more complete gene structures than current Genbank datasets, indicates fragmentation of gene models in the current assembly (median number of hits=3), suggesting we ought to obtain further mate-pair library data to improve scaffold assembly and lead to less fragmentation.

Introduction

Cyprinus carpio (common carp) is one of the most important freshwater cultured fish species. It has been widely used in fish biology research[1]. A single female is capable of producing up to a few hundred thousand eggs that can be efficiently fertilized in vitro, which enables hundreds of thousands of pharmaceutical drug candidates to be tested with less genetic diversity. Thus, common carp is a relevant model system for high throughput screens of pharmaceutical compound libraries.

Microarray technologies have been widely used and have been remarkably successful in identifying genome-wide gene expression patterns. However, there are a number of shortcomings e.g. low sensitivity and specificity and low consistency across platforms, and, above all, they rely on a very accurate definition of the transcriptome for their design. Next generation sequencing is a high-throughput sequencing technology which can produce millions of sequence reads from DNA and cDNA in a few days at a low cost, without the need for a priori knowledge of the full transcriptome. In terms of expression profiling, in comparison to microarrays, NGS has much higher dynamic range, base-level resolution, richer splicing information and ability to detect previously unknown genes, as long as adequate sequencing depth is obtained.

The aim of this project is to obtain an assembly of the carp genome using de novo sequencing of carp DNA and an assessment of the transcriptome by deep sequencing of cDNA as well using existing data from other species (e.g. zebrafish). This chapter is structured as follows:

- Introduction

- Results:
    - Assessment of a preliminary assembly obtained from pseudo-tetraploid DNA comparing ABYSS [4] and CLCBio [5]
    - In-depth evaluation of different assembly strategies using sequence from haploid DNA, comparing SOAPdenovo and CLCBio, as well as testing several parameters for SOAPdenovo assembly.
    - Quality assessment of the assembly produced, based on alignment to the assembly of carp BAC sequences, carp Genbank sequences, and carp RNA-Seq contigs and analysis of percentage of query sequence aligned as well as number of hits obtained.
    -

Results

Initial Dataset: pseudo-tetraploid material

Initially a partially inbred carp strain, pseudo-tetraploid, was available and used to produce and sequence the following Illumina genomic libraries:

- Standard DNA library of 200bp DNA fragments, sequenced with 10 lanes of Illumina GAIIx 51bp paired-end reads and 2 lanes of single-end 51bp reads
- Mate-Pair 5Kb library sequenced using 7 lanes of Illumina GAIIx 36bp paired-end reads

The following additional data was used to QC and improve the assembly was downloaded from Genbank:

- 2,136 Genbank DNA records including the mitochondrial genome
- Zebrafish genome mapping

Furthermore two Illumina GAIIx 51bp paired-end lanes were used to sequence RNA samples using RNA-Seq, and the reads were de novo assembled into 10,197 contigs of length greater than 500bp, used to identify potential genes (and/or gene fragments) in genomic data. See [2] for more details of RNA-Seq assembly and initial contig assembly of genomic data.

Preliminary Genome Assembly

Using ABYSS and a varying K parameter from 20 to 30, several genome assemblies were generated. Owing to the low coverage. Using the N50 (i.e. 50% of the contigs are at least N50 long) we could assess the genome assembly that was most contiguous. Given the low coverage of the initial dataset (approximately 10X) we did not expect a very contiguous assembly. Indeed, as shown in Figure 1, the best N50 was obtained using K=23, and was 231bp, i.e. not significantly longer than the original DNA fragments, indicating that a large part of the genome was not assembled beyond the original DNA fragments. Given the low coverage obtained, increasing the N50 also has a significant effect in the overall portion of the genome that is found within the assembly, i.e. significant removal of redundant fragments in sub-optimal assemblies at lower K parameters. As shown in Figure 1 the overall size of the genome assembled into contigs of at least 100bps decreases significantly as K (and N50) increase.

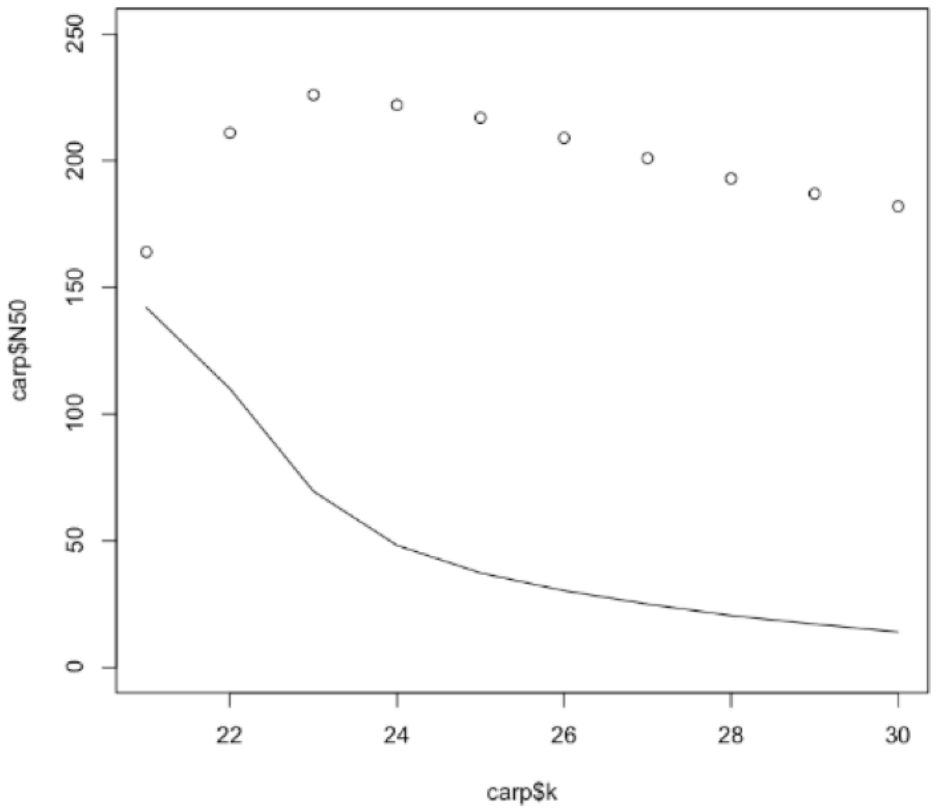


Figure 1 Preliminary assembly of the carp genome. Continuous line indicates the N50 of the assembly. Circles indicate overall assembly size (i.e. sum of the contigs longer than 100bp)

Subsequently further ABYSS assemblies were performed on this dataset introducing the scaffolding options available in ABYSS, but due to the low coverage, no significant differences were found in the final assemblies.

Haploid material assembly

Given the poor results obtained from the initial assembly, we then moved onto working from new material, i.e. deriving sequence reads from haploid DNA. Also, we changed assembly strategy, evaluating the commercial CLC Bio assembler, for contig assembly, and the SOAPdenovo tools (for both contig assembly and scaffolding). We also applied some pre-processing of the data prior to CLC Bio assembly (based on adaptor removal, trimming of low quality bases, and bridging of paired-end reads). The utilization of the CLC Bio assembler yielded a

much improved contig assembly, with an N50 of 1,409bp. The additional pre-filtering steps improved the contig assembly further to an N50 of 2,260bp, as shown below.

Assembly	n	n>100bp	n>50bp	median	mean	N50	Max	SUM
CLC Bio	1637271	1635617	250045	384	735	1409	17597	1.20E+09
CLC Bio + pre-filtering	1086163	1083124	159656	587	1135	2260	26293	1.23E+09

Assembly strategy

In order to obtain a final carp genome assembly which would be both of good contiguity (i.e. with a high N50) as well as of good representation of the genome (i.e. with a high coverage of currently known carp sequences) we tested several strategies, briefly summarized below:

- Different algorithms: ABYSS (see above), CLCBio de novo assembler, SOAPdenovo, as well as combination of CLCBio and SOAP
- Different SOAPdenovo K paramters: K tested from 33 to 40
- Different SOAPDenovo L (length of contig used for scaffolding) parameter (from 70 to 400)
- Trimming of 200bp sequence reads as well as 5Kb mate-pair reads

The chapters below detail the results obtained when varying each of these parameters.

Varying the K parameter in SOAPdenovo

De novo assembly algorithms implement an efficient strategy for a task that is computationally very demanding, i.e. the comparison of all short sequence reads

against each other. This is performed in order to identify which reads align to each other. This data in turn forms the basis of a graph that is used to identify the optimal path to reconstruct the full genome. Taking into account that for a genome such as the carp genome one is usually starting with 10^9 reads, these tools have to make 10^9 x 10^9 comparisons, i.e. 10^{18} comparisons. If traditional alignment approaches were to be used, the task would be computationally impossible on standard servers. As an example, if each alignment took 1 CPU second, it would require 10^{10} CPU years to obtain all the alignments. In order to drastically reduce the CPU time required, therefore, de novo assembly algorithms implement a K-mer based approach, i.e. before comparing sequences to each other, the sequences are rapidly scanned for all possible K-mers of a user-defined length. All subsequent steps are then performed on a K-mer representation of the original sequences, reducing drastically computational time required. This is a gross approximation of a full alignment strategy, which works well because of the sheer amount of sequences involved in the assembly process. The optimal size of K-mers to be used, however, cannot be easily determined a priori, since each specific length will leads to quite different, and unpredictable, results. Generally speaking K-mers which are approximately half the size of the sequence read tend to produce good results, but specific K-mer lengths (represented as the K parameter in all algorithms) have to be tested. Moreover it is important to note that the "optimal" K might be different depending on what is taken as a measure of success. As shown in our work a specific K-mer length might lead to better N50 but slightly lower overall sequence quality and viceversa.

In the CLCBio package until recently the user was not able to specify the K parameter to be used, and the package did not report the K parameters chosen. In recent releases this has been modified.

Thus, in order to test the effect of the K parameter on the final assembly, we deployed separate SOAPdenovo assemblies for K-mer length ranging from 33 to 40 (since the reads are 76bp, and thus K=38 should be the optimal K). As shown in Figure 2, a K = 35 leads to the most contiguous scaffold assembly (N50 = 5,510bp), while K=39 lead to the most contiguous contig assembly (N50 = 689). Notably, while the scaffold N50 is better than the N50 obtained with the CLC Bio software, the best Contig N50 is still well below that obtained by CLC Bio.

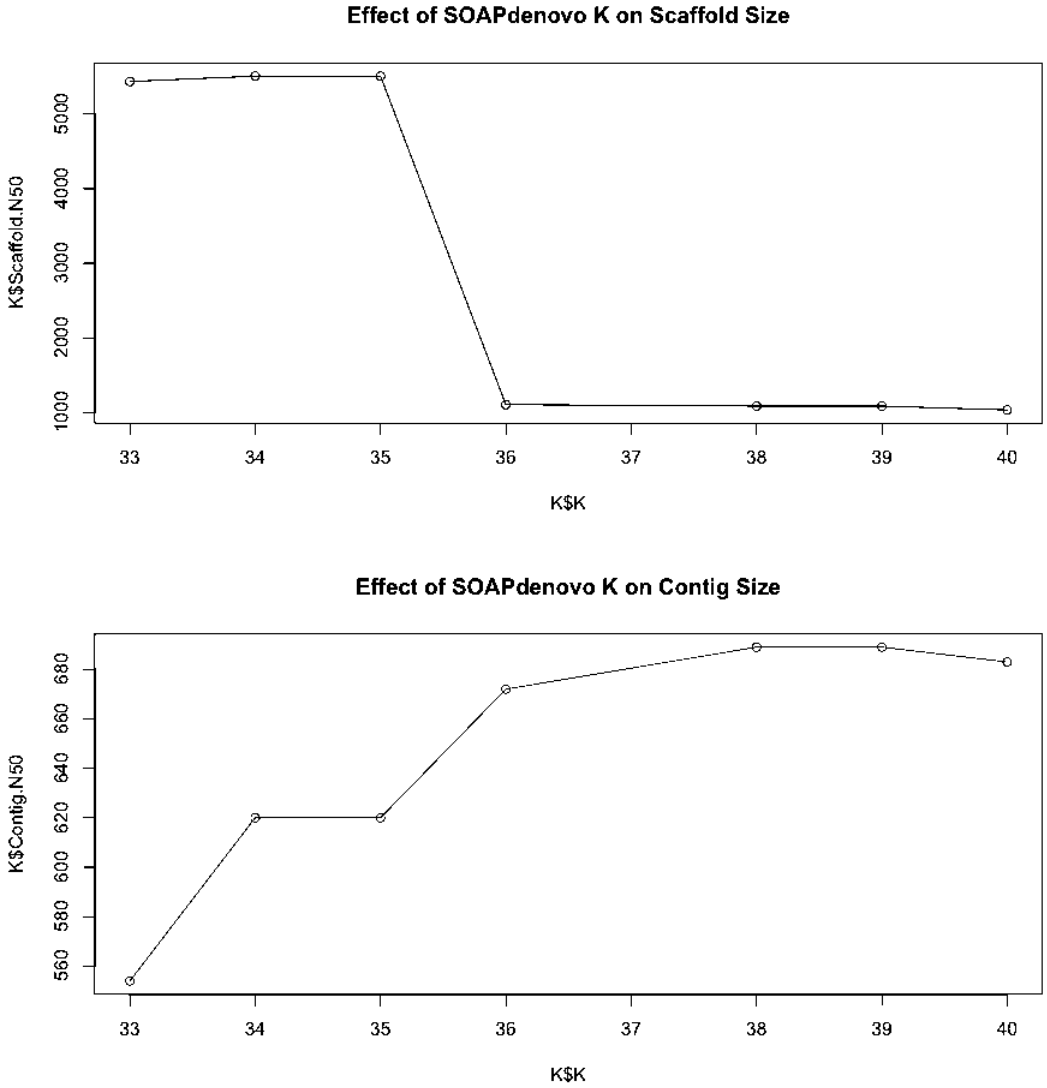


Figure 2 Assembly contiguity (top=scaffold, bottom=contig, based on the N50 statistics) as a function of the K parameters utilized for obtaining the assembly

Varying the L parameter in SOAPdenovo

One of the parameters which can be modified in SOAPdenovo is also the minimum length of contigs which are used for scaffolding. We thus tested from the minimum (2*k, i.e. in our case 70) to 400, and found that in terms of final N50 the optimal L is equal to the minimum, i.e. 70.

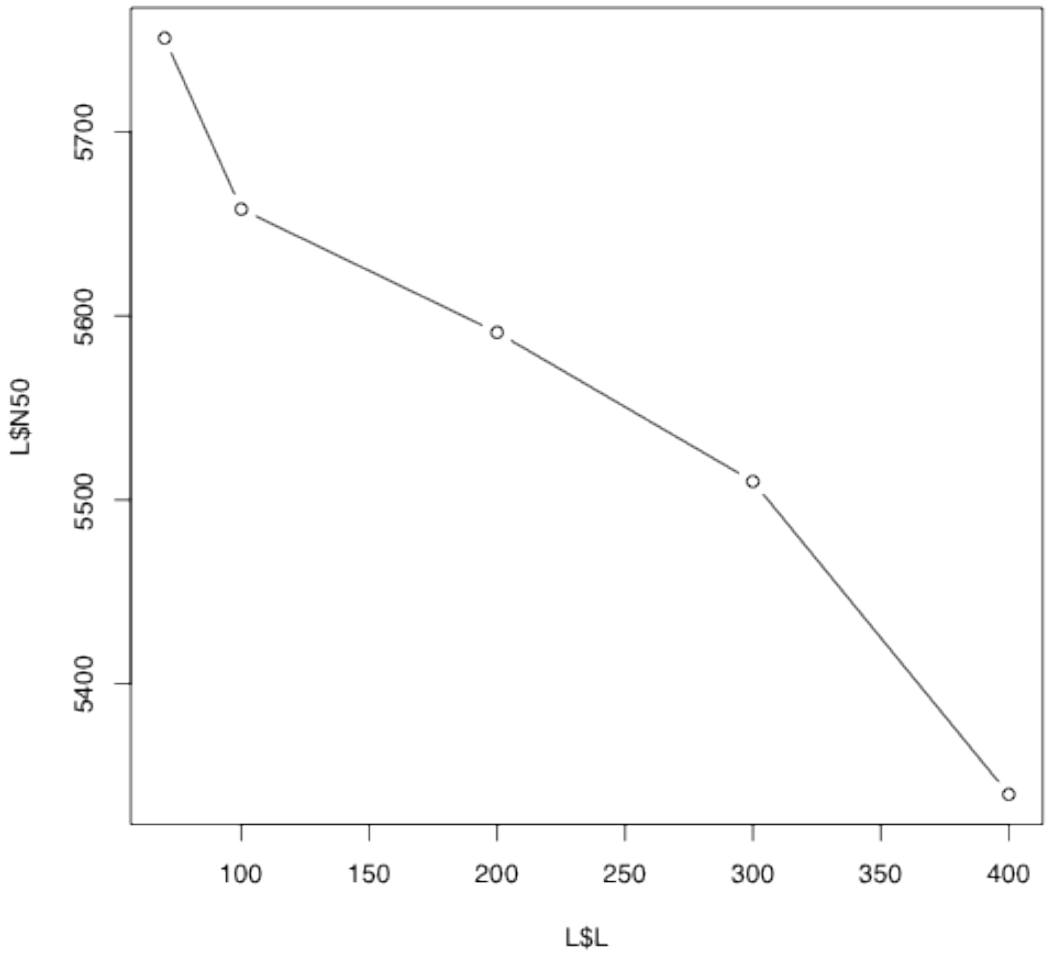


Figure 3 Assembly contiguity (based on the N50 of the scaffold length) as a function of the L parameter utilized for obtaining the assembly, tested for default L=70, as well as for L=100,200,300,400.

Testing read trimming strategies

Illumina sequencing reads have a quality profile which decreases as a function of read length, especially in the final part of the read. It is therefore often common to trim sequencing read ends in order to improve overall sequence quality and decrease. We therefore tested whether "hard trimming" (i.e. removal of a certain number of bases from the end of the sequence, regardless of its quality) would improve the assembly. We tested trimming of all reads (200bp library and 5Kb library) to 74bp and 72bp. As shown in Figure 4 the trimming of the last 2 and 4 bases had a marginally negative effect on the final N50 (N50 was 10bp less for the 74bp trimmed reads, and 50bp less for 72bp trimmed reads), and we thus did not utilize trimming of all reads for the final assembly. This reflects the fact that the majority of sequence reads were of good quality and thus "hard trimming" (i.e. trimming of all reads to a certain length, regardless of quality) leads to an overall loss of information content because the number of bases of good quality outweighs those of poor quality, which would be beneficial to remove. A better approach would be to trim sequence reads to a variable length, by trimming only low quality base pairs. Variable length reads, however, are not compatible with some softwares and aligners, and thus are not an ideal choice for the overall protocol.

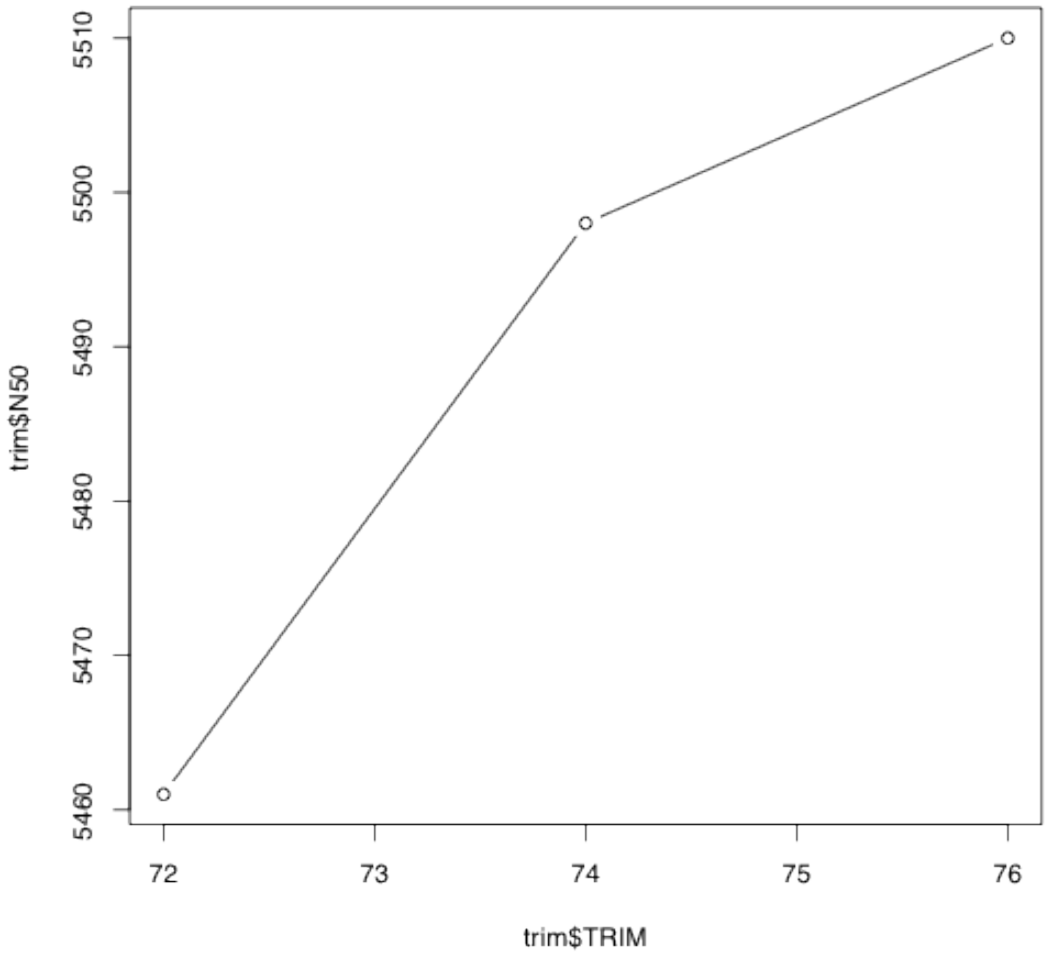


Figure 4 Assembly contiguity (based on the N50 of the scaffold length) as a function of the length to which reads were trimmed (full read = 76bp) Linear fit or Exp fit more realistic ? include datapoints

Mate-pair libraries often present a different issue which affects the final quality of the assembly, referred to as the "read-through" problem, i.e. that some of the fragments generated from the mate pair library might be very proximal to the junction of the mates (i.e. the point where the circuralization of the 5Kb fragment occurs), and thus, when sequencing at longer read lengths (as in our case 76bp), the sequence might "read through" into the opposite mate, which would then cause issues when mapping. We therefore tested the trimming of the 5Kb library reads to 35bps. When K is smaller than the trimmed length (i.e. in our case K=33, shorter than 35bp length) trimming has a minor negative effect (from an N50 of

5,438 to an N50 of 5424). When K >= 35, it actually produces a much poorer assembly, since K is as long (or longer) than the read itself and thus fails to produce correct alignments. Overall the results suggest a similar conclusion to minor trimming of all reads above, i.e. that due to limited read-through issues, the disadvantages of trimming outweight the advantages.

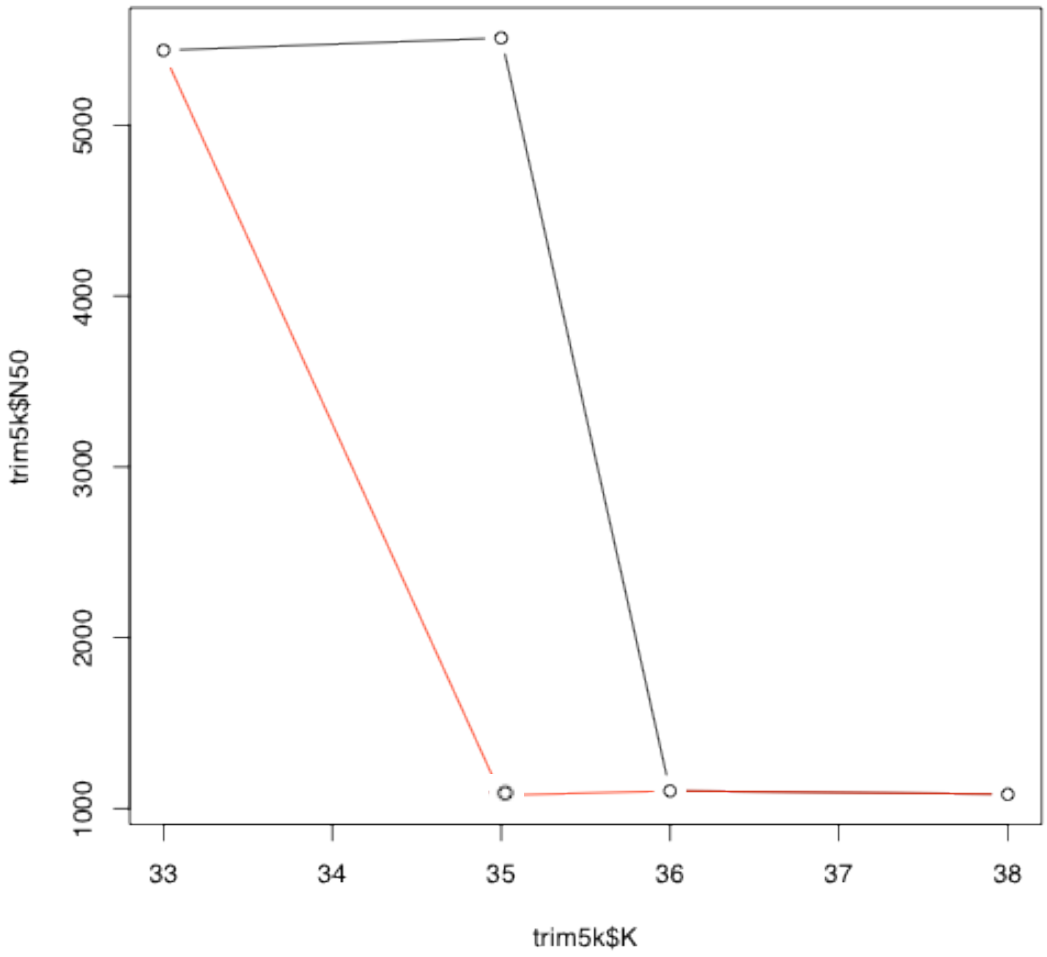


Figure 5 Comparison of assembly scaffold N50 at several K parameters, without trimming reads (black) and trimming 5Kb library reads to 35bps (red)

Testing combination of assembly softwares

CLC Bio yielded a better contig N50 than all the SOAP approaches, but it provides no scaffolding capability, therefore we decided to test a combined approach utilizing the SOAP scaffolding tools on the contigs generated using the CLC Bio software. This yielded the assembly with the best Scaffold N50 (8,043bp), far superior to all other approaches utilized previously. In order to do this we had to contact the developers of SOAPdenovo at BGI to obtain a separate tool which allows to prepare existing contigs for a particular K parameter for the subsequent mapping and scaffolding steps peformed by SOAPdenovo. This was performed with a compiled program named *prepare* kindly provided directly by the SOAPdenovo team, not currently released in the public domain.

Adding BAC end reads

Recently a research group in China published a small number of BAC end sequences (2,688) for the carp genome [2]. We contacted the authors and obtained the FASTA sequences for these BAC end reads. We then converted them to the appropriate format required by SOAPdenovo and tested whether the addition of these BAC end reads would improve the assembly. However the final assembly obtained by adding the BAC end sequences was actually marginally worse, with an N50 of 7,803 bp. This is probably due to the very low number of BAC end sequences utilized. A further dataset is being produced by the same laboratory of more than 80,000 sequences, which in the future could significantly aid the assembly.

Assembly Statistics

Following all the above attempts we opted for our final assembly for the following strategy:

1. Contigs were obtained by pre-processing reads (checking for overlaps between paired reads, so that they can be merged and discarding low quality nucleotides) and then using the CLC Bio software for de novo assembly, which yielded a set of contigs with N50 of 2,262bp.
2. These contigs were then prepared for SOAPdenovo assembly using a "prepare" script kindly contributed by the BGI SOAPdenovo team.
3. Scaffolding was then performed using SOAPdenovo by using both the 200bp reads and the 5kb reads without trimming, using a K=35, and default L and G parameters (L=K*2, G=50), the R flag (,i.e. use reads to solve tiny repeats) and M=3 (maximum strength for merging similar sequences)

The statistics of the assembly thus obtained are as follows:

Total number of sequences: 779,686

Total Size: 1,647,732,536 (just below the average of the estimated genome sizes published in genomesize.com, which is 1.71Gb)

Contig Statistics:

Minimum: 100

Maximum: 26327

Average: 1137.73

N50: 2262

Scaffold statistics:

Minimum: 100

Maximum: 115,091

Average: 2113.33

N50: 8,043

A comparison with the CLC Bio contig assembly shows that we thus produced a significantly lower number of total sequences, with a total size that is significantly more in line with the estimated genome size for the carp genome, and with a 4x improvement in terms of final N50:

Number of sequences: 1,086,163

Total Size: 1,235,762,671

Sequence lengths:

Maximum: 26,327

Average: 1,137.73

N50: 2,262

The assembly displays a clear log normal distribution of scaffold sizes, with a small increase of scaffolds of length similar to the overall assembly N50 size, i.e. between 5kb and 8kb.

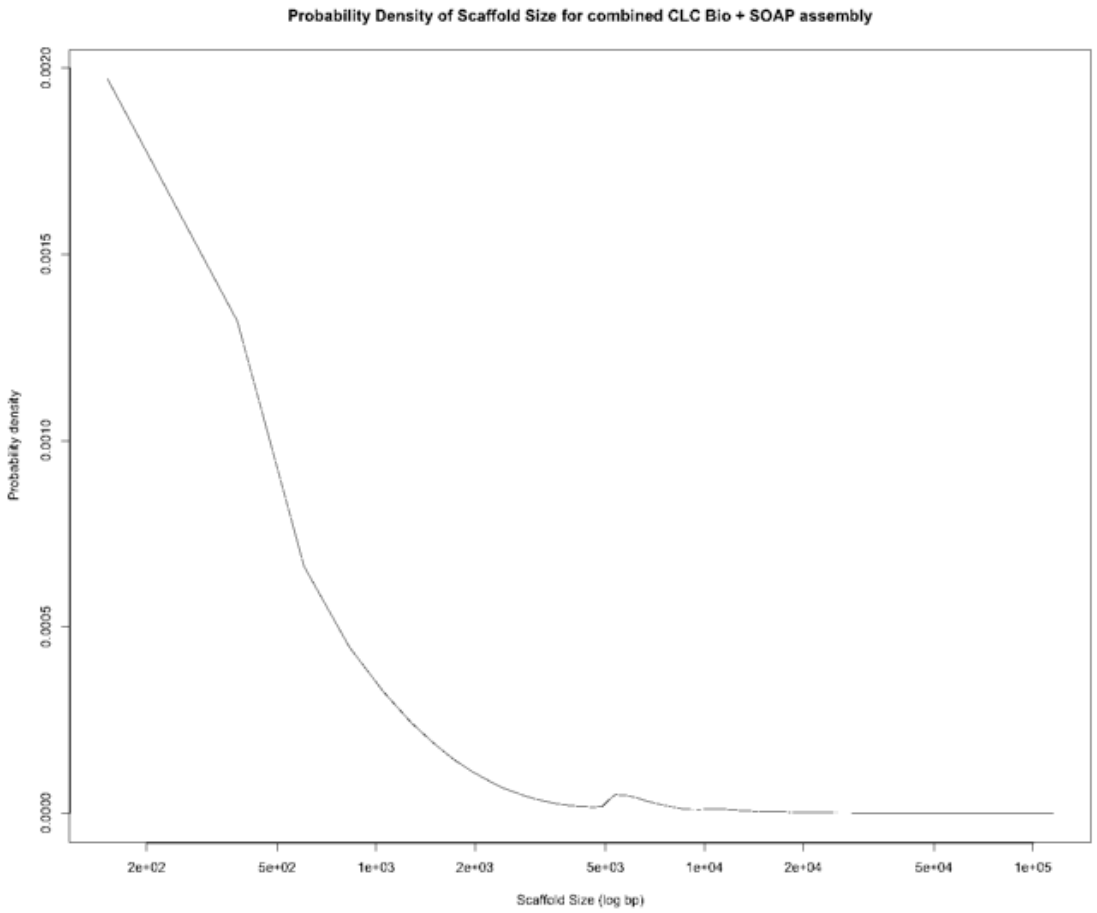


Figure 6 Density distribution of length of scaffolds in final assembly (X axis: length in base pairs). The figure shows that there is a slight preference for scaffolds of approximately 5Kb, i.e. scaffolds of N50 size.

Largest scaffolds

As a preliminary overview of the quality of the final assembly, we used BLAT [6] on the Ensembl genome browser [7], to map the largest scaffolds produced on the zebrafish genome. Owing to the nucleotide BLAT-based search only high similarity hits were found across the scaffold matching to exons and conserved regions within the genome. Reassuringly, all scaffolds were found to match a single location on the zebrafish genome in a collinear manner (see Table below, as well as Figures 7-9), indicating good synteny between these two genomes, and also showing that these largers scaffolds are not due to major assembly artifacts generated during the assembly process. Interestingly, one of the largest scaffolds

(scaffold24544) contains the two largest known vertebrate genes, i.e. Titin A and part of Titin B.

Scaffold ID	Length	Zebrafish top match	Gene Names	Gene descriptions
scaffold1889	115,091	chr16:26.44Mb-26.55Mb	cadm4	cell adhesion molecule 4
scaffold8659	106,795	chr16:35.20Mb-35.30Mb	me1	cytosolic malic enzyme 1
scaffold24544	100,347	chr9:43.8Mb-44.01Mb	ttnb,ttna	titin b, titin a

Table 1 Mapping of carp scaffolds larger than 100Kb to the zebrafish genome, indicating zebrafish chromosomal location for main hit, as well as names and descriptions of genes contained in the zebrafish locus.

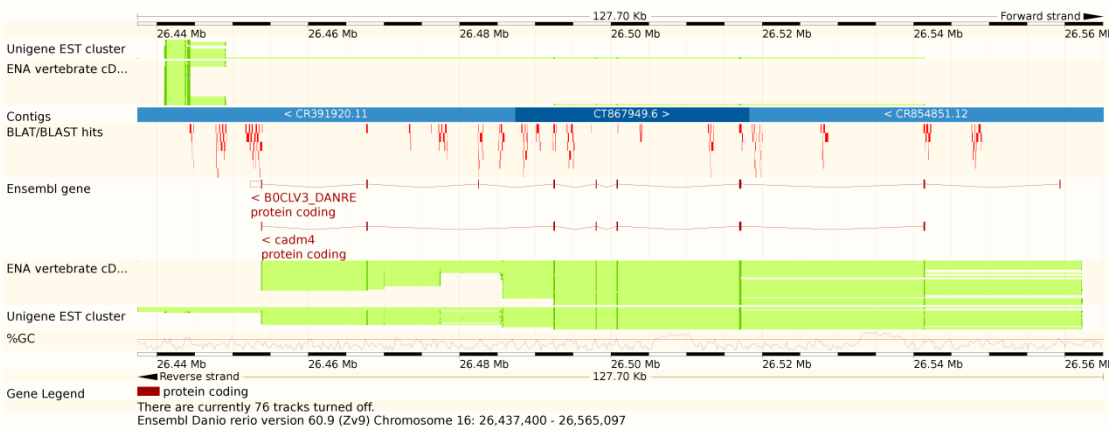


Figure 7: Scaffold1889 mapping to zebrafish chromosome 16

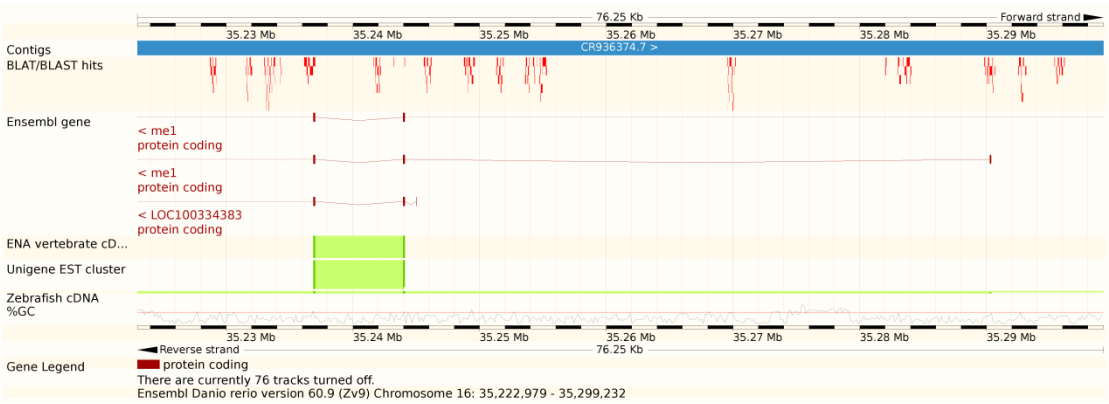


Figure 8 Scaffold8659 mapping to chromosome 16

Figure 9 Scaffold24544 mapping to zebrafish chromosome 9 containing the Titin loci

Quality Assessment

In order to assess the quality of the assemblies produced, in terms of their usefulness and representation of known carp genes and BAC sequences, we downloaded all available carp nucleotide sequences from GenBank. This comprises over 2,000 sequences, mostly of known partial or full carp genes, as well the whole mitochondrial genome and two BAC clones. We then proceded to map stringently (using Blat) all these sequences to each assembly generated.

Coverage of existing BAC clones

The Genbank deposited sequences include two longer BAC sequences, which can be useful to obtain an initial (although biased) measure of quality. The statistics obtained are shown in the table below. While BAC Clone BX571725 shows little variability across assemblies, BAC clone BX571686 indicates clearly that the highest coverage of the clone is achieved with the combined approach CLC + SOAP, which reaches 81% coverage of the clone, a remarkable result considering that 98% identity cutoff was used for this comparison. The coverage increases to 87% when using a 95% percentage identity.

Assembly	BAC BX571686.2 coverage	BAC BX571725.1 coverage
k35_r_m3_l70_g50_BAC_fake6	76.11653646	74.31533749
k35_r_m3_l300_g200	77.36002604	74.59810441
k35_r_m3_l70_g50	75.13020833	73.37278107
k33_r_m3_l300_g200	74.93489583	72.03225638
k39_r_m3_l300_g200	76.12955729	73.82049537
CLC	78.88346354	72.69466408
CLC + SOAP K35	**81.5625**	74.26297324

Table 2 Coverage of two carp BAC clone sequences obtained from different genome assemblies attempted

Coverage of all carp Genbank sequences

While the BAC clones provide interesting evidence of good coverage of existing carp sequence data for longer sequences, this is very anecdotal in nature, because only two clones are available. We thus proceeded to assess the coverage of the entire Genbank dataset containing over 2,000 sequences. As shown below, the average coverage of the query carp DNA sequences increases with improved assembly strategy. Both assemblies based on the CLCBio contigs produce the best coverage possible, with very small differences between the combined CLC Bio + SOAP approach and the CLC Bio approach alone. This is expected since the additional SOAP step is providing mostly "bridging" information across existing sequence, rather than novel sequence information. Among the SOAPdenovo assemblies the one made using K=39 has better coverage than others. Although this assembly has poor scaffold N50 statistics, it has the best Contig N50 statistics, indicating possibly that these are a better measure of final coverage of existing data (at significant expense of scaffold contiguity). In other words owing to the higher Contig N50 it is likely to better represent actual contiguous sequence which affects the final coverage mapping.

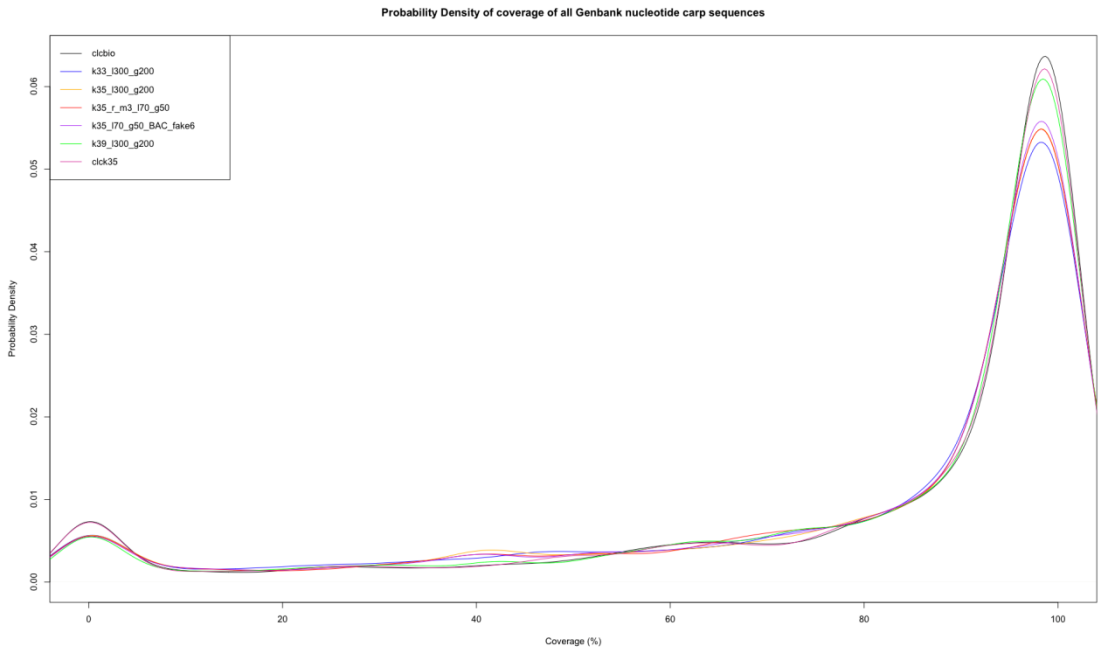


Figure 100 Probability density of coverage of carp Genbank nucleotide sequences for each assembly analyzed. The two assemblies based on the initial CLC Bio Contig assembly (clcbio and clck35 which indicates the CLCBio + SOAP assembly) have the highest fraction at high coverage.

In order to assess fragmentation (i.e. to what extent known carp sequences are fragmented on multiple scaffolds), we verified the number of hits found in each assembly for the entire set of carp nucleotide sequences found in Genbank. The graph in Figure 10 indicates clearly that the CLC Bio based assemblies have lower fragmentation (i.e. the DNA sequences used present fewer hits, despite the increased coverage), and the combination of SOAP and CLC Bio (clck35) has the lowest fragmentation of hits, as expected given its higher N50. Importantly the majority of known carp sequences have a single hit in the assembly, indicating that the majority of the genes (or gene fragments) searched can be found in a single scaffold sequence.

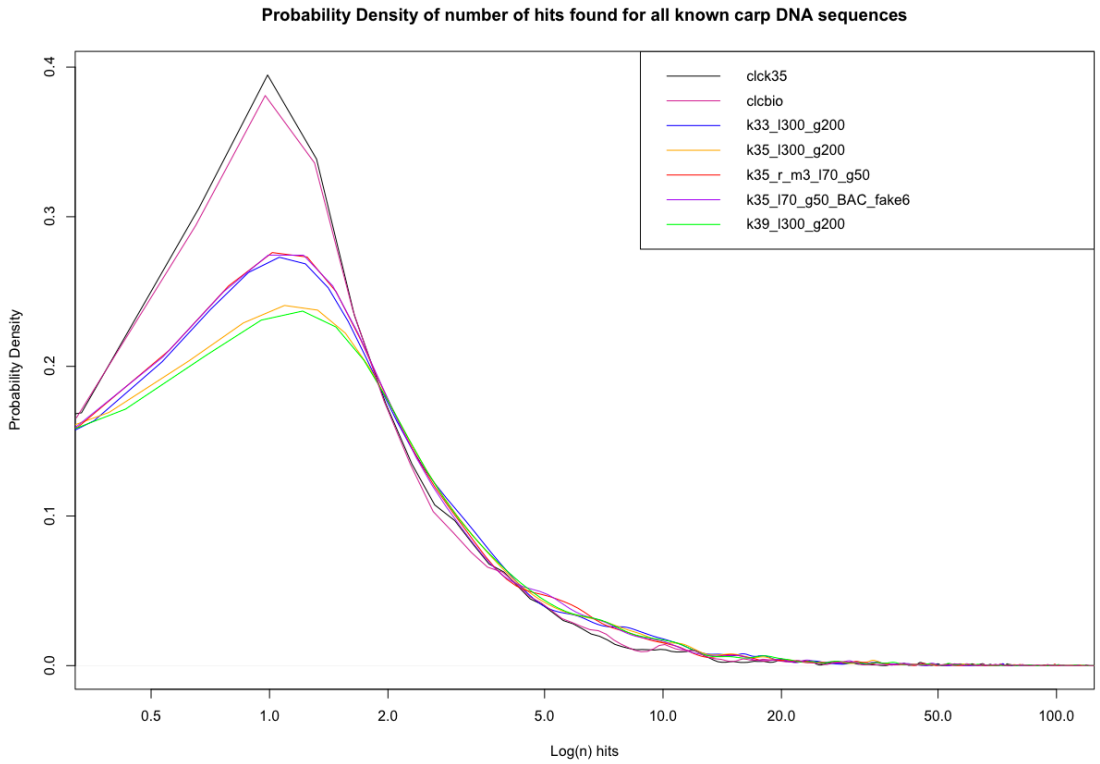


Figure 11 Probability density of number of hits obtained for each carp Genbank query sequence. The combined CLC Bio + SOAPdenovo assembly shows the lowest number of hits per query, with the majority of query sequences having a single hit.

As shown in Table 2 below only 22 out of 2,100 known carp DNA sequences analyzed could not be located in the assembly, and the majority of them are either haplotype specific, or within regions known to present high variability and difficult assembly such as the olfactory receptors and the MHC complex, which have required years of specific dedicated work in much larger projects such as the human genome project. Only 7 carp gene fragments cannot be located, which, based on existing data in Genbank, would indicate we are missing less than 0.5% of the gene content.

Sequence	Type	Type
FJ198033.1	Microsatellite	Repeat
FJ490421.1	3-beta hydroxysteroid dehydrogenase partial mRNA, 117bp	Gene fragment
FJ490420.1	cytochrome P450 21-hydroxylase partial mRNA, 144bp	Gene fragment
FJ655360.1	haplotype HcII3 cytochrome oxidase subunit II (COII) mitochondrial gene	Haplotype-specific sequence
FJ655287.1	haplotype Hc2 cytochrome b mitochondrial gene	Haplotype-specific sequence
FJ655286.1	haplotype Hc1 cytochrome b gene, partial cds; mitochondrial.	Haplotype-specific sequence
FJ655355.1	haplotype Hd51 tRNA-Pro gene and control region, partial sequence; mitochondrial	Haplotype-specific sequence
EU203669.1	MHC class II antigen beta chain (Cyca-DAB1) gene, Cyca-DAB1*05 allele, exon 2 and partial cds	MHC Complex
X95436.1	mRNA for MHC class II beta chain, D(clc)B	MHC Complex
Z47730.1	Cyca-DXA2*01 gene for MHC class II alpha chain	MHC Complex
EF042096.1	enolase mRNA, partial cds	Gene fragment
EF042095.1	enolase mRNA, partial cds	Gene fragment
BD262014.1	Method of identifying organism by comparative gene analysis and primer and hybridization probe for effecting the method.	Patent sequence
AY505343.1	uncoupling protein 3 (UCP3) mRNA, partial cds	Gene fragment
AB194212.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR35	Olfactory Receptor cluster
AB194205.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR25	Olfactory Receptor cluster
AB194203.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR23	Olfactory Receptor cluster
AB194202.2	OFRE mRNA for olfactory receptor, partial cds, clone: CCOR21	Olfactory Receptor cluster
AJ628728.1	partial mRNA for sialic-acid binding protein-4	Gene fragment
AX084639.1	Sequence 171 from Patent WO0055361	Patent sequence
AF008558.1	brain-derived neurotrophic factor (BDNF) gene, partial cds	Gene fragment

Table 2 List of carp Genbank sequences which could not be mapped to the final genome assembly, indicating Genbank ID, description and type of sequence. Most sequences are either haplotype specific or belong to variable regions of the genome.

Gap Filling

In order to further improve the final assembly generated using the combined CLC Bio + SOAP strategy the GapCloser algorithm (part of the SOAPdenovo package) was used, which aims at filling gaps found within scaffolds. The software was able to remove ~25% of the Ns found in the original assembly, as detailed below:

- Number of Ns before Gap Filling: 421,965,634
- Number of Ns after Gap Filling: 309,918,033

Since gap filling can sometimes lead to lower quality sequences within the gaps, QC was performed again on the new gap-filled assembly by mapping all carp Genbank records to it. The results, shown below, indicate that the gap filled assembly is of higher quality than the assembly produced without gap filling.

- Before Gap Filling:
    - Average Coverage: 91.26%
    - STDEV Coverage: 17.93%
    - Median coverage: 98.06%
- After Gap Filling:
    - Average Coverage: 92.33%
    - STDEV Coverage: 17.38%
    - Median coverage: 98.31%

Mitochondrial genome

A very good example of a high quality scaffold found in the assembly is the mitochondrial genome, which would be found in Scaffold C2172197 and which contains the entire carp mitochondrial genome (as shown from BAC Clone AP009047.1). The alignment indicates complete coverage and only one sequence segment inverted with respect to the BAC clone deposited, as shown in Figure 12.

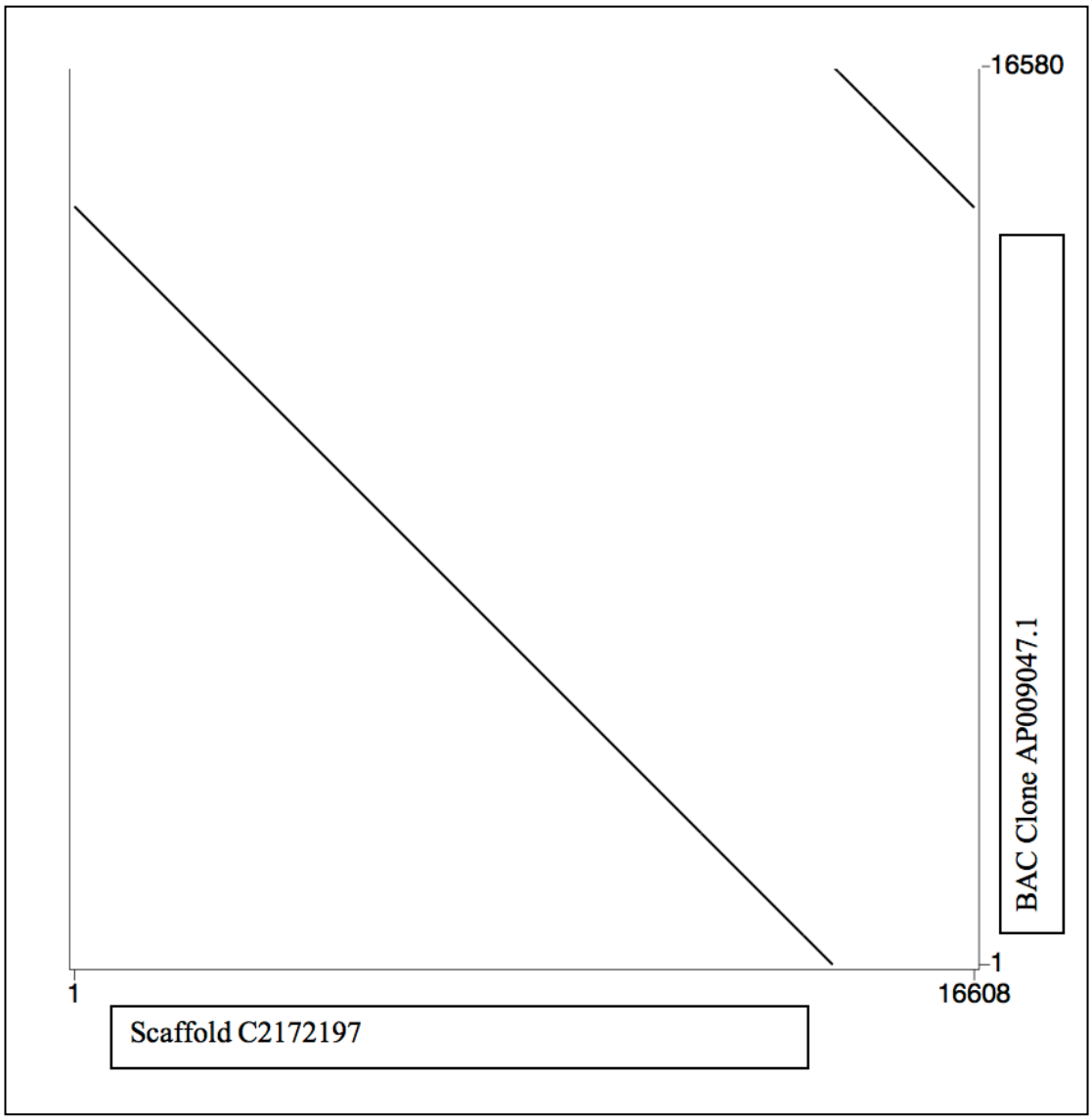


Figure 12 Figure showing the base by base alignment of Scaffold C2172197 and Clone of the mitochondrial genome (BAC Clone AP009047)

RNA-Seq Analysis

The RNA-Seq data, assembled into contigs using CLC Bio, was also mapped to the final genome assembly. In line with the data obtained using Carp Genbank entries, the overall coverage was very high for most contigs (see Figure 13, median coverage = 98.7%, average coverage = 92.47%), although usually spread over a few different scaffolds (see Figure 14, median number of hits = 3, mean number of hits = 9.71)

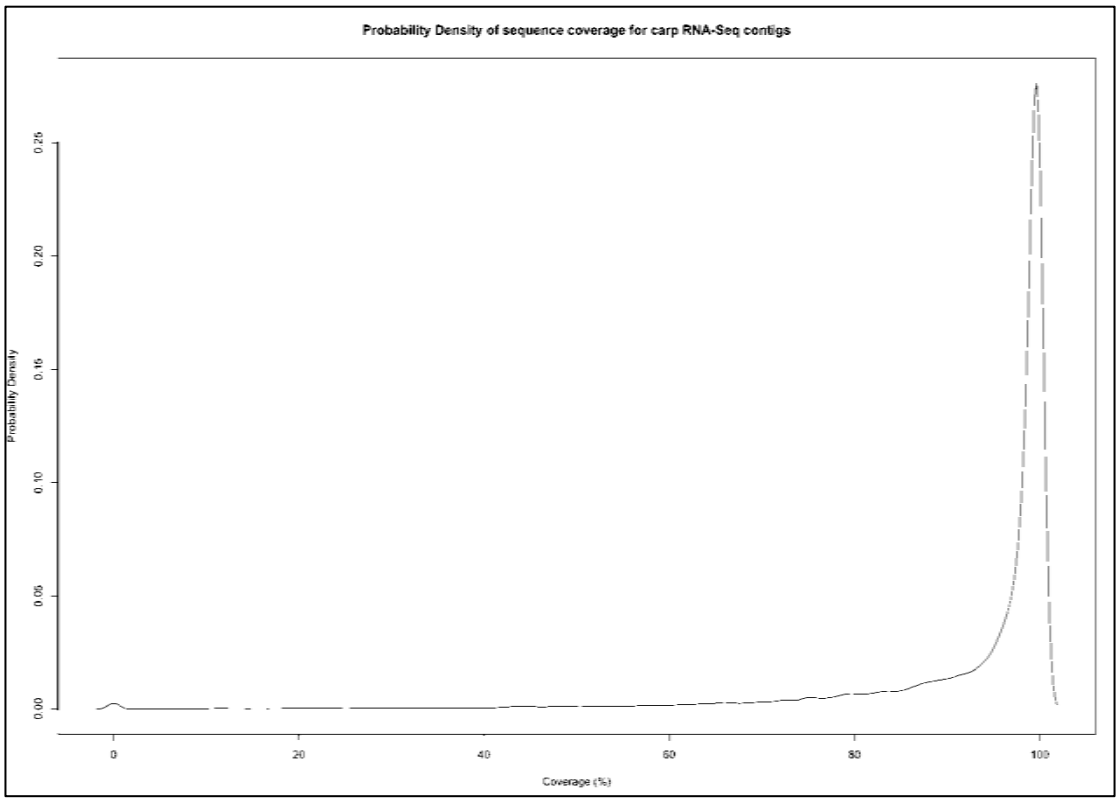


Figure 13 Proability Density of Sequence Coverage (%) for Carp RNA-Seq Contigs.

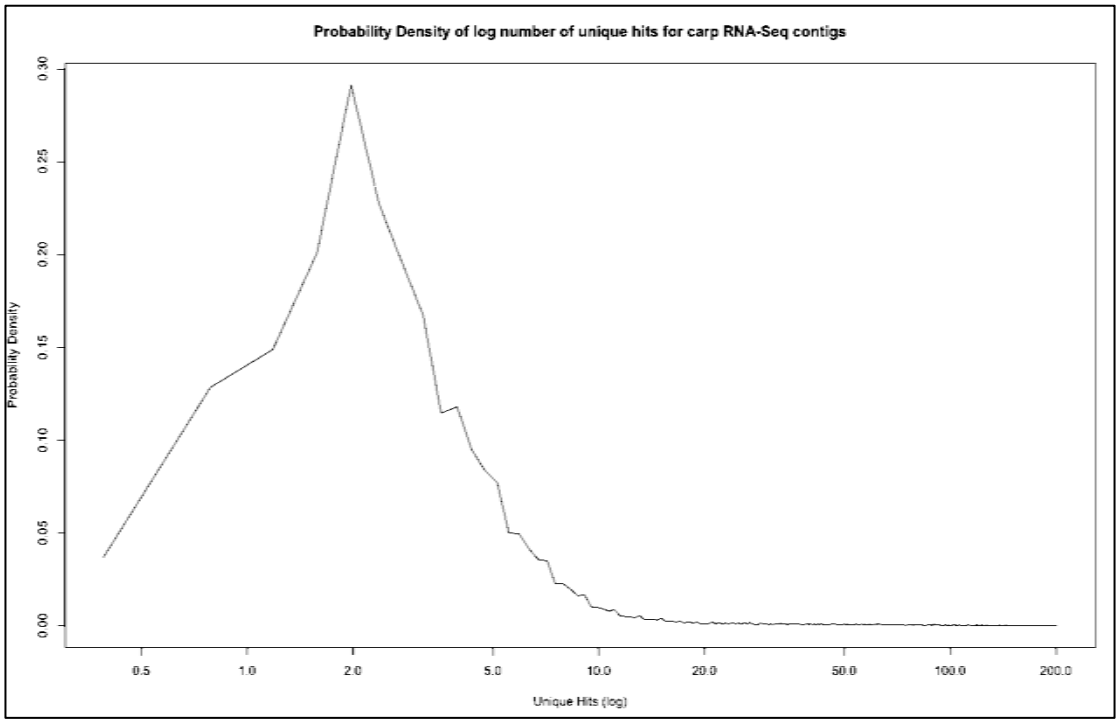


Figure 14 Probability Density of number of unique hits (log scale) for Carp RNA-Seq Contigs

As an example, one of the longest RNA-Seq contigs, contig_1067 (6,347bp) mapped fully to scaffold50890. This scaffold and RNA-Seq contig have their best mapping to chromosome 19 in the zebrafish genome (as well as a few other lower identity copies in other chromosomes, such as chr16, chr14 and chr25). All of the mappings are unspliced and are not annotated. A further analysis, using BLASTX on the NR database indicates that this is likely to be a processed pseudogene, as it contains an RT_LTR region, i.e. a reverse transcriptase domain.

Methods

Genome Assembly

In this project, initially the ABYSS de novo assembler was used due to lower memory requirements as compared to the very first algorithm published, Velvet (40-50Gbs of memory needed by ABYSS as compared to several hundred gigabytes required by Velvet). In a second phase most of the assemblies were performed using SOAPdenovo which not only uses similar amounts of memory but supports multi-threading, thus allowing rapid assembly of large datasets on a multi-CPU machine. The assemblies were performed on a 32 CPU Opteron server with 512GB RAM, allowing several assemblies to be run in parallel, exploring parameter space (such as modifying the K parameter to identify the optimal K and modifying other SOAPdenovo parameters such as L, G and read trimming options described). ABYSS was run in paired-end mode (i.e. using the command abyss-pe) and the following parameters: k=n (with n varying from 20 to 30) l=51 c=2 n=10 name=fish lib='lib200 lib5kb' lib200=pe200.fa lib5kb=pe5kb.fa se=se.fa. The first runs were performed as above. Subsequently ABYSS, from version 1.0.15, supported scaffolding, so the following option was added, to scaffold the contigs generated, including scaffolds which are likely to contain repeats: OVERLAP= --scaffold --mask_repeats.

SOAPdenovo was run with the following parameters: SOAPdenovo all -K n (ranging from 33 to 39) -p 24 (number of CPUs to use) -R -M3 -s soap.config (config file for run) -o name_of_assembly. The SOAPdenovo config file varied depending on the approach taken, the following is an example in which we used the 200bp reads, 5kb reads and BAC end reads, for both building contigs and scaffolds in successive steps:

```
#maximal read length
max_rd_len=76
[LIB]
#average insert size
avg_ins=200
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used, 1=contig, 2=scaffold,3=both
asm_flags=3
#in which order the reads are used while scaffolding
rank=1
#fastq file for read 1
q1=/data_n1/stupka/carp/CARPDATA/newdata/soap_elia/all_1.txt
#fastq file for read 2 always follows fastq file for read 1
q2=/data_n1/stupka/carp/CARPDATA/newdata/soap_elia/all_2.txt
[LIB]
avg_ins=5000
asm_flags=2
rank=2
reverse_seq=1
q1=/data_n1/stupka/carp/CARPDATA/5kb_reads/mp5k_1_sequence.txt
q2=/data_n1/stupka/carp/CARPDATA/5kb_reads/mp5k_2_sequence.txt
```

Gap Filling was performed using the GapCloser algorithm part of the SOAPdenovo package and was run as follows: GapCloser -o name (name of filled assembly) -b soap.config (config file similar to the above indicating where reads are located) -a name.scafSeq (file to be used for Gap Filling) -t 24 (number of threads to use).

QC Analysis

The QC analysis was performed using the server/client version of BLAT version 34, which is called gfServer and gfClient respectively. The binaries were downloaded from the UCSC website. Each assembly produced was transformed into 2Bit format (a compressed sequence format used by BLAT), and then a server was loaded with each assembly as follows:

```
gfServer start localhost port name_of_assembly.2bit
```

Each assembly was loaded on a different local port as required by the server.

The gfClient program was then used to search all Genbank nucleotide records specifying the assembly to be searched and the minimum identity cutoff. Furthermore we chose the "pseudoblast" output format, which is easily parsable using BioPerl. Thus the final command-line was as follows:

```
gfClient localhost port assembly_location query.fasta
output_name -minIdentity=95 -out=blast
```

The output was then parsed with a Perl script which uses the BioPerl BLAST parser Bio::SearchIO to parse the results. All hits with e-value < 1e-05 and score < 100 were removed. The number of hits was recorded. Then, since many small, overlapping hits are found, we used a temporary MySQL database to load all hits and quickly find the minimal subset of features which encompass all hits with no repetitions (this is in a module name Bio::MCE::Range which was kindly provided by Remo Sanges). This set of features was used to calculate final coverage of the query sequence.

Graphical Reporting

All graphical reporting was produced using R statistical functions. The following functions were among those most commonly used:

- read.table was used to import TAB delimited from files generated from Perl scripts or from the command-line into R datasets

- The density function was used to obtain probability densities where needed

- The plot and lines functions were used to produce final plots

Discussion

Initial pseudo-tetraploid ABYSS based assembly

The initial assembly of the pseudo-tetraploid carp genome was performed at a time when genome assembly had just started becoming a possibility for smaller labs as compared to large genome centres, but library making strategies, sequencing protocols and assembly algorithms were heavily under development. The first assembly obtained from the pseudo-tetraploid genome, although clearly not a good basis for future work was very useful to learn some of the key aspects that enable a successful genome assembly, summarized below:

- Obtaining very good coverage (e.g. 30x or more) of standard (e.g.200bp) genomic DNA libraries with long (e.g. 100bp) paired-end sequencing is fundamental to obtain a reliable initial contig dataset where each fragment is ideally sequenced completely with overlapping paired-end reads

- Obtaining mate-pair libraries sequenced at high depth with short reads has a significant effect on the ability to scaffold. This is in order to avoid "read-through", i.e. sequencing through to the other side of the mate-pair.
- Preparing mate-pair libraries of different insert sizes (e.g. 600bp, 5Kb, 10Kb) provides a range of "scaffolding opportunities" (e.g. 600bp for bridging over most repeat elements, 5Kb and 10Kb to provide long-range bridging) and also has significant effect on final scaffolding ability
- Availability of a BAC library for BAC end sequencing is of great benefit for obtaining very long range scaffolding, i.e. in the range of 100s of Kbs
- Recent chemistry upgrades made by Illumina (in particular v5 or TruSeq chemistry) have a significant impact on assembly of large genomes, due to overall higher quality base calling as well as longer overall read length

Evaluation of ABYSS

While clearly ABYSS provided the first "affordable" approach to de novo assembly, by utilizing on average 40Gb of memory for each carp assembly, it still had several caveats, which impacted the final results:

- Since ABYSS does not yet support multi-threading it took several days for each assembly to complete, limiting the range of options we could test
- Despite trying in several ways, the scaffolding options did not yield any results, i.e. no improvement to the scaffold N50, no clear scaffolding.

Haploid DNA CLC Bio and SOAP de novo based assembly

The next dataset of 200bp sequence reads obtained from a haploid genome, combined with an improved assembly strategy, allowed us to produce a radically

improved assembly, on which most of the optimization and analysis work was performed.

CLC Bio Contig Assembly

Although CLC Bio was not initially our choice of software due to its commercial nature and costs involved, evaluations of its de novo assembly algorithm lead us to decide to make use of it, since it produced far better contigs than any other approach tested. The assemblies produced "out of the box", i.e. without tweaking the data or parameters, were already superior to those we had produced. Once the data was pre-processed and merged, it lead to even better contig assembly, which was subsequently used for scaffolding with SOAPdenovo. The limitations of using the CLC Bio software are:

- It is commercial software and, unlike other bioinformatics commercial software there is no academic free availability in any form, and a specific license (not the one for the CLC Genomics Workbench) needs to be purchased for the command-line tools
- It is a "black box" approach, i.e. we do not have any details of how the assembly process works, and we cannot understand why it is so much superior to other approaches
- It does not support scaffolding, which was the main reason why obtained the best final assembly by combining this tool with SOAPdenovo

Based on all the QC performed we are quite confident that the superior N50 obtained does not come at a cost in terms of sequence quality, since the assemblies based on CLC Bio contigs consistently show the best coverage of existing sequences.

The K parameter

Using SOAPdenovo we evaluated extensively the best K parameter to use. The analysis of the contiguity of the assemblies highlighted an interesting aspect: the K parameter which produces the best Contig N50 (K=39) is not the same which produces the best Scaffold N50 (K=35). In fact the Scaffold N50 drops radically when using K=39, which is why for the final assembly (which is based on pre-made CLCBio contigs) we decided to use K=35. We discussed this aspect with the authors of SOAPdenovo, and they recommended not varying the K parameter within the different steps of the SOAPdenovo assembly (e.g. using K=39 for contig assembly and K=35 for scaffold assembly), because this would lead to overall decreased sequence quality and inconsistencies in the assembly. From a point of view of QC, the assembly produced with K=39 actually presents slightly higher coverage of carp DNA sequences, indicating that in terms of sequence quality contig N50 is potentially a better measure, at a very significant cost in terms of contiguity. Taking into consideration that we perform a final Gap Filling step in the scaffolds produced, we prefer to have higher scaffold contiguity (i.e. many gaps to fill) if the sequence quality is not overly compromised. The issues above highlight the limits of K-mer based approaches, which strongly limit "comprehensive" assembly of sequences.

Other SOAPdenovo parameters

Although we explored quite extensively other SOAPdenovo parameters, any modifications to defaults seemed to impact negatively the final N50. The minimum length of the contigs used for scaffolding did not improve the final N50, although there were reports recently in the Phallusia genome project

(Patrick Lemaire, personal communication) that increasing L lead to a better assembly. This could be due to the specific nature of our assembly.

Often trimming reads leads to improved results in the context of a variety of next-generation sequencing projects, due to the usually poorer quality of the last bases of each sequence. Our trimming attempts, however, did not improve the N50. This indicates that although base qualities decrease towards the end of the read, overall those positions have a higher ratio of information vs. noise for the overall assembly process. The ideal approach is to perform a variable length quality-based trimming (i.e. remove bases below a certain quality) but since many tools/aligners do not deal with variable length FASTQ sequences well, we did not try that approach. Another approach is simply to replace low quality bases with Ns, and this was done in the pre-processing of the reads for the CLC Bio based contig assembly. In fact read trimming can produce a radical artifact, as was shown in Figure 5, i.e. that if the K parameter is equal or larger than the length of the trimmed reads, the assembly is reduced drastically, because SOAPdenovo is unable to compare K-mers appropriately in the assembly process.

BAC end reads

BAC end Sanger reads are clearly extremely beneficial for obtaining very long-range scaffolding. We obtained a very limited dataset of 2,200 BAC end reads. This set, unfortunately, was not large enough to see any improvement in the assembly. This is probably also because there is no option to assign a "weight" in SOAPdenovo to a piece of evidence. Thus even though the BAC end reads are from Sanger sequencing and are thus likely to provide stronger evidence than a

single read from a FASTQ file, there is no way to encode this in the process. As a test it was tried to simply repeat 6 times the same BAC end reads, thus "faking" a stronger weight for the BAC ends and indeed, some longer scaffolds were produced, because of additional linking information. This assembly was not used further, however, because it could contain potentially some erroneous assemblies due to the forced duplication that was introduced. The laboratory which has produced this first set of 2,200 reads has communicated to us that a further, much larger set of approximately 80,000 reads will be available soon and thus we will able to attempt another assembly with this new dataset when it is available.

Assembly Assessment and QC

Given the initial data obtained the final assembly obtained is of very good quality, as assessed through a variety of approaches:

- Mapping the largest scaffolds produced to the zebrafish genome identifies collinear mappings of similar size, indicating that no major artifacts were produced in the process of assembling larger scaffolds
- The coverage of the only two available BAC clones is reasonably high (81% and 74%) indicating good coverage of existing genomic sequences
- The coverage of known Carp nucleotide sequences is remarkably good. The median coverage is 98%, and the median number of hits is just 1, indicating that the vast majority of known carp nucleotide sequences is very well covered and is usually found in a single scaffold
- Very few sequences are not mapped at all in the current assembly, and the majority of those (15 out of 22) are unlikely to be mapped in our genome

assembly because they belong to highly variable regions, recombinant regions and haplotype-specific regions.

- The mitochondrial genome was found completely in a single Scaffold, with only one potential segment in the wrong assembly location
- The RNA-Seq Contigs have very good coverage in the current assembly (median = 98%), confirming that most genes should be found in this assembly.

On the other hand there are clearly some caveats:

- The majority of the sequences deposited in Genbank are fairly short and it is thus unsurprising that they are mapped in one or very few scaffolds (is this referring to DR or CC)
- The BAC coverage is not as good as the coverage of genomic sequences. While two BAC clones are probably too few to make any conclusive statements the assembly probably lacks good coverage of repeat-rich regions.
- While the contiguity is good taking into account the starting data, it is still not sufficient to recover many multi-gene loci, thus strongly limiting any studies aimed at understanding synteny among fishes and/or vertebrates, as well as the search for promoter regions, enhancer regions, etc. which all require long contiguous assemblies spanning several gene loci
- As highlighted by the mapping of the RNA-Seq contigs, on average genes are still fragmented over several scaffolds, thus it is paramount to improve the contiguity of the scaffold assembly with further sequencing of longer libraries and hopefully more BAC end data.

References

1. Hulata G. A review of genetic improvement of the common carp (*Cyprinus carpio L.*) and other cyprinids by crossbreeding, hybridization and selection. Aquaculture. 1995;129:143-155. doi: 10.1016/0044-8486(94)00244-I.
2. Yanju Zhang, Functional annotation of microRNAs and de novo genome sequences through heterogeneous data analysis, PhD Thesis, Leiden University, The Netherlands (2011) In Press.
3. Yan Li, Peng Xu, Zixia Zhao, Jian Wang, Yan Zhang and Xiao-Wen Sun Construction and Characterization of the BAC Library for Common Carp Cyprinus Carpio L. And Establishment of Microsynteny with Zebrafish Danio Rerio *Marine Biotechnology* DOI: 10.1007/s10126-010-9332-9
4. Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones and İnanç Birol, ABySS: A parallel assembler for short read sequence data, *Genome Research* (2009) 19: 1117-1123
5. http://www.clcbio.com/
6. W. James Kent BLAT—The BLAST-Like Alignment Tool, *Genome Research* 2002. 12: 656-664
7. P.Flicek et al, Ensembl 2011, NAR (2011) 39 (suppl 1): D800-D806.

Chapter 6: Discussion

Impact of next-generation sequencing on genome research

This thesis spans across several key genomics fields, reflecting the development of the discipline: from genome sequencing and assembly, to comparative genomics and transcriptomics. These fields have been impacted heavily by the emergence of next-generation sequencing, which has provided faster, more affordable tools to obtain genomes, transcriptomes, and "regulomes". In this discussion I aim to contextualize the results obtained during the thesis in the backdrop of these new technologies.

As noted by Lincoln Stein in his recent paper on cloud computing [1], while the cost of sequencing a base has fallen by half about every five months, the cost of storing each byte of data is dropping, but not as fast (every 14 months). This shifts completely the "data paradigm": while in the past obtaining data (e.g. the sequence of a gene) was a major achievement that would be closely guarded until publication, now obtaining data is easy, and the bottleneck becomes the bioinformatics analysis. Immediate, open, public release of data should thus be encouraged further, to enhance the data analysis potential and the biological conclusions that can be derived. This also has strong implications with regards to the type of biological questions that can be asked and the way in which they are asked. One obvious example is the genetics of outbred populations, unthinkable until recently, and now a reality [2].

The sequencing and assembly of entire genomes has been "commoditized" owing to next-generation sequencing. The effort involved in sequencing, assembling and annotating the Fugu genome in terms of money (approximately 10M

dollars), time (approximately 3 years) and people (approximately 20) involved would now require probably 2 orders of magnitude less effort (100K dollars, 3 months, 2 people). This reduction in costs/effort has brought the possibility of sequencing a genome to being a standard research project of a standard research lab. As an example, our lab is now actively involved in the genome sequencing of 4 species, and in the planning phases for approximately 25 more species, while a center such as the BGI in Shenzhen has recently set out to sequence 10,000 animal genomes.

The advances in data generation have pushed even further the key bottleneck towards the analysis of the data, i.e. the ability to use bioinformatics and biostatistics approaches to derive meaning from these large biological datasets. Moreover the data is often so rich that more than one person/team can utilize the same dataset and derive different, complementary, biological conclusions. One example is RNA-Seq, where the same dataset can be used to study several different biological aspects such as quantification of gene expression, discovery of novel genes, identification of alternative splicing. Each application requires different algorithmic approaches, dedicated bioinformatics effort and extensive validation. Thus the emphasis shifts away from being able to generate data to that of being able to analyze, obtain and validate novel biology reliably and extensively.

Searching for regulatory elements

The identification of regulatory elements genome-wide was, until very recently, confined to methods employing comparative genomics, such as the one looking we employed to identify shuffled conserved elements present in fish genomes,

presented in chapter 2. In this field, yet again, next-generation sequencing has had a major impact. The possibility to obtain genome-wide mapping of immunoprecipitated chromatin using Chip-Seq [3] has enabled, especially in mammalian systems, to obtain quickly very comprehensive signature of the regulatory code of the genome, including several histone marks, POL2 occupancy, and the regions bound by key de-acetylases such as p300, CBP [4]. Specifically in relation to enhancers a comprehensive studies conducted by Len Pennacchio showed clearly that a p300 Chip-Seq based approach [5] recovered with much higher success rates true functional enhancers than the older-fashioned sequence conservation based approach [6,7].

While Chip-Seq based approaches are clearly very promising, they are not directly and easily applicable to a large variety of species, as demonstrated by the fact that in non-mammalian vertebrate species, e.g. fish, so far there is no published extensive catalogue of enhancers. This is mainly due to the fact that the technologies need to be adapted to each specific species: the identification of antibodies that work effectively is often not trivial (easier for histone marks which are well conserved over longer evolutionary distances, but less straight-forward for DNA-binding proteins), the immunoprecipitation protocol needs to be adapted and optimized, and, last but not least, these techniques rely on large numbers of cells, which are often not available (and cell lines are often also not available). Finally, comparative genomics provides a large, unbiased, picture of regions of the genome that are under evolutionary constraint, regardless of their function. In order to identify all these regions via Chip-Seq approaches one would have to combine a vast number of Chip-Seq protocols, and might still miss

some with novel functions. Thus, for the time being, comparative genomics will still provide useful information on functional elements of the genome, which is complementary to Chip-Seq approaches.

Transcriptomics

In our study of MBT transition we had to resort to the technological platforms which were widely available at the time, i.e. microarrays. Microarrays have proven a fairly reliable measure of gene expression (for genes which are expressed at reasonable levels) in organisms such as the mouse and human genome. These organisms have benefited early on from a fairly complete genome sequence and assembly, very extensive biological sequence collections (ESTs, cDNAs, CAGE data, etc) and therefore gene annotation is very mature. This in turn has allowed microarray manufacturers to produce reliable oligonucleotide probes. Furthermore the very large market, usage and competition has forced continuous improvements of microarray platforms. The same cannot be said for other species. While many model organisms are not catered for at all by mainstream microarray manufacturers, for others, like *Danio rerio*, microarrays are available but far from ideal due to the poor (until recently) genome sequence and assembly and poor (until recently) gene models available. This is why in our analysis of MBT transition in zebrafish we could work only on approximately 10,000 genes, from which less than 2,000 were then usable for the final analysis.

RNA-Seq, on the other hand, provides a species-independent, unbiased, quantitative assessment of the transcriptome, which allows any lab, working on any species, to sequence the cDNA obtained from any RNA sample of interest.

Besides freeing the researcher from the need of a supported dedicated platform for the species of interest, it also captures a wider dynamic range of transcription, from very poorly expressed transcripts, to very highly expressed transcripts, without the limits imposed by the optical read-out of microarrays [8]. Moreover, RNA-Seq can be used effectively to study not only quantification of transcripts, but also alternative splicing [9] and novel gene prediction [10].

RNA-Seq on the other hand, as for many next-generation sequencing techniques, provides novel and difficult challenges from the bioinformatics analysis point of view. Mapping of reads to the genome is more complex due to the presence of spliced reads, which map across distant regions in the genome. Several algorithms have been developed in recent times to account for this aspect, such as, for example, TopHat [11] and SplitSeek [12], but these only aid in improving the quality of the mapping, without providing a complete solution for gene prediction or alternative splicing prediction. Newer algorithms such as Cufflinks [10] and many under development as part of the RGASP competition, such as mGene (developed by the group of Gunnar Rätsch at the Friedrich Miescher Institute) provide a much more sophisticated usage of RNA-Seq data and genome sequence to model accurately splice junctions, gene models and alternative splicing, for both coding and non-coding genes.

Genome Assembly

In genome assembly probably more than in any other genomics field the impact of next-generation sequencing has been radical. The commoditization of sequencing, coupled with the improvement of algorithmic tools and the commoditization of servers with large memory and CPU power has enabled the

average laboratory to undertake independently a whole genome sequencing, de novo assembly and annotation project, which until recently was confined to large sequencing centres. As shown in the last chapter of the thesis, the field is shifting rapidly, and while work was being conducted on the chapter new tools were being developed which assisted us in the de novo assembly of the carp genome. Although the work still needs to be complemented by further sequencing to improve contiguity we have shown convincingly that we were able to produce an assembly which is likely to contain a significantly large portion of the carp transcriptome (probably more than 90%) as assessed on the basis of both known carp DNA sequences as well as our own RNA-Seq dataset.

Similarly annotation of a genome was a heavy undertaking which involved comparative genomics as well as very expensive Sanger-sequencing based EST sequqencing projects. It can now be completed in a few weeks with a few Illumina lanes of RNA-Seq material, providing a good baseline for a preliminary annotation. In both the RNA-Seq and the genome assembly approach it is clear that the length of the sequences is still a limiting factor. Obtaining truly complete gene models from RNA-Seq requires very high depth. This, in turn, requires to obtain RNA from a range of tissues, or to utilize normalization protocols, since usually highly expressed genes will be well assembled, while genes with lower expression will have lower coverage and thus will not be assembled well. Similarly, while the genome assembly is satisfactory for preliminary identification of genes, mapping to other genomes, etc. it does not provide good multi-genic contiguity and is thus greatly limited in terms of more in-depth analysis. To achieve this either very high depth is required or some

complementary data, e.g. BAC end Sanger reads. As the cost of next-generation sequencing keeps dropping and sequence length increases, the cost/benefit ratio of using complementary Sanger based datasets will change. As shown with the publication of the Panda Genome [13], a complete de novo assembly with Scaffold N50 of over 1GB from Illumina sequencing only is now possible, as long as one can afford very high depth sequencing (in their case over 100X of the genome).

References

22. Stein LD The case for cloud computing in genome informatics. *Genome Biology* 2010; 1(5):207. Epub 2010 May 5
23. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 2008; 3(10): e3376. doi:10.1371/journal.pone.0003376
24. Valouev, A et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods 2008;* 5:829–834
25. Barski A, Cuddapah S, Cui K, Roh T, Schones DE, Wang Z, Wei G, Chepelev I Zhao K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 2007; 129: 823–837
26. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA ChIP-seq accurately predicts tissue-specific activity of enhancers *Nature* 2009; 457:854-859
27. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. 2006 *Nature*; 444:499–502
28. Visel, A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. 2008 *Nature Genet*ics; 40:158–160
29. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B Mapping and quantifying mammalian transcriptomes by RNA-Seq 2008 *Nature Methods*; 5(7):621-628
30. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB Alternative isoform regulation in human tissue transcriptomes 2008 *Nature*; 456:470-476
31. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation 2010 *Nature Biotechnology*; 28(5):511-515
32. Trapnell C, Pachter L, Salzberg SL TopHat: discovering splice junctions with RNA-Seq 2009 *Bioinformatics*; 25(9):1105-1111
33. Ameur A, Wetterbom A, Feuk L, Gyllensten U Global and unbiased detection of splice junctions from RNA-seq data 2010 *Genome Biology*

11:R34
34. Ruiqiang Li et al. The sequence and de novo assembly of the giant panda genome 2009 *Nature*; 463:311-317

Curriculum Vitae

Name: Elia Stupka
Date of birth: 09-05-1977
Place of birth: Quartu Sant'Elena (CA), Italy
Institute at which PhD was conducted: Leiden Institute of Advanced Computer Science

Education

University of York, UK – M.Res. in Biomolecular Science
Created a human mutation database at the European Bioinformatics Institute, in Cambridge, UK (Human Mutation, 2000)

University of York, UK – B.Sc. in Biology
Final year project based on models of game theory entitled "Evolution of co-operation by indirect reciprocity: a spatial model", (Shell Award for best ecological project of the year)

United World College of the Adriatic, IT – International Baccalaureate
IB Diploma with highest mark for experimental extended essay "Mapping transgenic *per* lines of *Drosophila Melanogaster*"

Employment

2009-ongoing: Scientific Director, UCL Genomics
Senior Lecturer, Bioinformatics, Joint Appointment between University College London and Barts and the London School of Medicine and Dentistry and

2006-2009: Bioinformatics Program Manager at CBM-Cluster in Molecular Biomedicine,Trieste, Italy

2003-2006, Bioinformatics Program Manager at Telethon Institute of Genetics and Medicine, Naples, Italy

2003-2006, joint appointment as faculty at the European School of Molecular Medicine, Italy

2003-2004, Bioinformatics Program Manager at Temasek Life Sciences Laboratory, Singapore

2001-2003, Project Manager, Institute of Molecular and Cell Biology, Singapore

1999-2001, Ensembl Scientific Programmer at the European Bioinformatics Institute Cambridge,UK