Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention

Andrew L. Hufton, Susanne Mathia, Helene Braun, Udo Georgi, Hans Lehrach, Martin Vingron, Albert J. Poustka, and Georgia Panopoulou[1]

Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

Animal genomes possess highly conserved *cis*-regulatory sequences that are often found near genes that regulate transcription and development. Researchers have proposed that the strong conservation of these sequences may affect the evolution of the surrounding genome, both by repressing rearrangement, and possibly by promoting duplicate gene retention. Conflicting data, however, have made the validity of these propositions unclear. Here, we use a new computational method to identify phylogenetically conserved noncoding elements (PCNEs) in a manner that is not biased by rearrangement and duplication. This method is powerful enough to identify more than a thousand PCNEs that have been conserved between vertebrates and the basal chordate amphioxus. We test 42 of our PCNEs in transgenic zebrafish assays—including examples from vertebrates and amphioxus—and find that the majority are functional enhancers. We find that PCNEs are enriched around genes with ancient synteny conservation, and that this association is strongest for extragenic PCNEs, suggesting that *cis*-regulatory interdigitation plays a key role in repressing genome rearrangement. Next, we classify mouse and zebrafish genes according to association with PCNEs, synteny conservation, duplication history, and presence in bidirectional promoter pairs, and use these data to cluster gene functions into a series of distinct evolutionary patterns. These results demonstrate that subfunctionalization of conserved *cis*-regulation has not been the primary determinate of gene duplicate retention in vertebrates. Instead, the data support the gene balance hypothesis, which proposes that duplicate retention has been driven by selection against dosage imbalances in genes with many protein connections.

[Supplemental material is available online at http://www.genome.org. All in vivo tested elements have been deposited into the ORegAnno database [http://www.oreganno.org] under data set no. OREGDS00016.]

Researchers are increasingly recognizing the importance of *cis*-regulatory sequences in genome evolution. About 3% of vertebrate noncoding sequences are selectively constrained, and most of these sequences—often called conserved noncoding elements (CNEs)—can function as *cis*-regulators of gene expression in transgenic assays (Nobrega et al. 2003; Margulies et al. 2003; Cooper et al. 2005; Siepel et al. 2005; Woolfe et al. 2005; Drake et al. 2006; Sanges et al. 2006; The ENCODE Project Consortium 2007). The most highly conserved of these CNEs are enriched near genes that regulate transcription and development (Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2005), where the precise and complex regulation required by these genes may impose an exceptional selection against mutation within their *cis*-regulatory sequences. Researchers have proposed that the exceptional constraint on these *cis*-regulatory sequences may influence the evolution of their surrounding genome regions, both by constraining genome rearrangement (Mackenzie et al. 2004) and by promoting the retention of duplicated genes (Force et al. 1999).

Because many *cis*-regulatory sequences lie distantly from their target genes, and sometimes even in the introns of neighboring genes, intervening genomic rearrangements have the potential to disrupt their *cis*-regulatory functions. Mackenzie et al. (2004) termed this "gene interdigitation" and proposed that since many of these sequences are under strong purifying selection, they should act to conserve the surrounding genome architecture. In support of this idea, studies in vertebrates and insects have shown that regions of the genome with conserved gene order (synteny) tend to contain many CNEs (Ahituv et al. 2005; Engström et al. 2007; Kikuta et al. 2007; Navratilova et al. 2008). However, the synteny–CNE association observed in these reports could be at least partly methodological. These articles defined CNEs using algorithms that favor colinearity, and allowed CNEs to contribute to the detection of conserved syntenic regions, raising the possibility that it may simply be easier to detect CNEs in syntenic regions, and/or easier to detect conserved synteny in regions with many CNEs. In support of this possibility, Sanges et al. (2006) used more flexible algorithms and detected many CNEs that had been shuffled and rearranged in fish–mammal comparisons.

In addition to conserving genomic synteny, some investigators have proposed that *cis*-regulation may play a key role in determining which genes are retained after duplication. In general, duplicated genes should be functionally redundant, leading to rapid removal or nonfunctionalization by mutation. However, metazoan genomes have retained many functional duplicate genes (Hughes and Hughes 1993; Nadeau and Sankoff 1997; Lynch and Force 2000; Lynch et al. 2001; Blomme et al. 2006). The evolution of new beneficial functions (neofunctionalization) could act to preserve some duplicates; however, this mechanism appears to be insufficient to explain the large number of vertebrate gene duplicates, leading researchers to search for additional retention mechanisms (Lynch and Force 2000). The first of these proposed

[1]**Corresponding author.**
E-mail panopoul@molgen.mpg.de; fax 49-30-84131128.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.093237.109.

19:2036–2051 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

mechanisms was inspired by the observation that many genes possess modular *cis*-regulatory sequences that drive functions in distinct biological contexts (Force et al. 1999). After duplication, these *cis*-regulatory sequences can be subdivided by complementary degeneration, thereby requiring that both duplicates be retained in order to maintain the entire ancestral function. This model, called the duplication-degeneration-complementation (DDC) model, can act upon any gene with multiple essential subfunctions that can be separated by mutation, and while there is some evidence that this can occur at the protein-coding level for genes with domains of distinct function (e.g., Cusack and Wolfe 2007), the majority of reports supporting this model have focused on *cis*-regulation. Clearly, expression partitioning does occur frequently in retained gene duplicates (Huminiecki and Wolfe 2004; Postlethwait et al. 2004; Li et al. 2005; Duarte et al. 2006; Woolfe and Elgar 2007), and in yeast, genes with complex *cis*-regulation are over-retained after duplication (He and Zhang 2005). Within metazoan lineages, the same types of gene functions that are enriched near CNEs are also over-retained after whole-genome duplication (WGD) (Blanc and Wolfe 2004; Maere et al. 2005; Blomme et al. 2006).

Despite this evidence, an alternate model has been proposed to explain biased retention of gene duplicates. The gene balance hypothesis (GBH) postulates that selection against gene dosage imbalances will promote the retention of certain types of genes after WGD events (Veitia 2002; Papp et al. 2003; Birchler et al. 2005; Freeling and Thomas 2006; Freeling 2008). Immediately after a WGD event, genome-wide relative gene dosage is maintained, but subsequent step-wise mutation or deletion of duplicate genes can lead to deleterious dosage imbalances. Genes whose proteins have many interaction partners may be more sensitive to these dosage changes, possibly leading to an over-retention of highly connected gene functions, such as transcriptional regulators and signaling complexes (Birchler et al. 2001; Veitia 2002; Papp et al. 2003). Conversely, small-scale genomic duplications immediately disrupt relative dosage, so highly connected genes should avoid this type of duplication during evolution. This anti-correlation between gene retention after WGD and small-scale duplication is a key distinction between the GBH and the DDC models; DDC should promote the same patterns of gene retention for all types of gene duplication. In support of the GBH, vertebrate and *Arabidopsis* genes that function in transcription regulation or signal transduction are over-retained after WGD events but not after small-scale duplications (Blomme et al. 2006; Freeling 2008).

The relative merit of these two competing models remains difficult to assess. Clearly, transcription and development genes have more complex *cis*-regulation and more CNEs, which may allow greater subfunctionalization, but they are also often assumed to be highly connected and dosage sensitive. If expression partitioning via the DDC model is the main force driving duplicate retention, we would expect to see a tight correlation between gene functions that have CNEs and the gene functions that are over-retained after WGD duplication; however if retention after WGD is not driven by the DDC model, then one might expect to find key differences in these enrichments. To date, there has been no combined analysis in vertebrates of the gene functions associated with CNEs and gene duplication.

Addressing these issues requires that we are able to identify CNEs sequences deeply in evolution. Previous reports have found many CNE sequences that can be dated back to the divergence of fish and tetrapods, and in some cases all the way back to the early vertebrate WGD events (Bejerano et al. 2004; Siepel et al. 2005; McEwen et al. 2006; Stephen et al. 2008; Wang et al. 2009). Moreover, 77 CNEs were recently identified that have been conserved between humans and amphioxus, a nonvertebrate chordate (Holland et al. 2008). This report used relatively simple whole-genome blasts, leading us to suspect that more powerful methods would be able to identify many more vertebrate-amphioxus CNEs. Indeed, Sanges et al. (2006) have previously shown that CNE searches anchored to orthologous genes can be more powerful than whole-genome searches in fish–mammal comparisons.

With these goals in mind, we have developed a computational CNE identification method that relies on local similarity searches within phylogenetically defined chordate gene families. A series of simple rules allows the method to adapt to gene families of different sizes and species compositions. Random simulations are used to prove that the discovered phylogenetically conserved noncoding elements (PCNEs) are highly specific to their associated gene families. This method is powerful enough to identify PCNEs that have been conserved between vertebrates and nonvertebrates, including more than a thousand PCNEs in the basal chordate amphioxus. We test 42 of our predicted PCNEs in transgenic zebrafish assays—including examples from vertebrates and amphioxus—and find that the majority are functional enhancers. Moreover, we find a clear association between the number of PCNEs associated with a gene and the likelihood that the gene will have conserved its synteny during evolution. This trend is most apparent for extragenic PCNEs (those outside of their predicted target genes), suggesting that interdigitation of *cis*-regulatory sequences plays a key role in conserving genome architecture. Next, we use a clustering-based approach to dissect how PCNEs and other aspects of genome evolution are associated with different gene functions. These patterns are most consistent with the GBH model of gene duplicate retention and indicate that subfunctionalization of CNEs is unlikely to have been a primary determinate of gene duplicate retention in vertebrates. Further supporting the GBH, genome-wide estimates of protein connectivity suggest that genes with many interaction partners are selectively retained after WGD duplication, and avoid segmental duplication.

Results

A local-similarity-based method to identify CNEs within families of genes

In order to identify CNEs deeply within vertebrate evolution and without bias in regards to duplication or rearrangement, we have developed a CNE identification method that relies entirely on local similarity searches in the genomic regions surrounding phylogenetically defined gene families. Searching genomic regions around gene families is generally more sensitive than whole-genome searches and also allows us to identify conserved sequences in paralogous regions created by duplication events. For the analysis presented here, we have searched the genomes of four organisms: *Branchiostoma floridae* (amphioxus), *Mus musculus* (mouse), *Takifugu rubripes* (fugu), and *Danio rerio* (zebrafish). The amphioxus genome, as our best conserved chordate ancestor, forms the outgroup used to root the gene families. Mouse, fugu, and zebrafish were selected for the quality of their genome sequence, their history of use in previous CNE analyses, and their diverse genome evolution histories.

Briefly, we first build families of genes from the four genomes, where each family of genes is orthologous to a single gene in the hypothetical chordate ancestor (Table 1). Around each gene

Table 1. Orthology summary

	Genes[a]	Gene families[b]
Total	51,942	8368
Amphioxus	18,946	8368
Mouse	11,935	7443
Fugu	9612	6518
Zebrafish	11,449	6315

[a]The number of genes in our orthologous gene families.
[b]The number of gene families with at least one gene from the organism. Families are rooted by the amphioxus genes.

in a family, we extract the genomic sequence within the gene and extending out 100 kb from both the gene start and end, while masking known coding regions and repeats (Fig. 1A). The set of genomic sequences for each family is then searched for conserved regions using a two-step local similarity search. The first step uses BLASTZ, a fast but powerful local alignment algorithm, to identify a set of candidate sequences that have been conserved across species (Fig. 1B; Schwartz et al. 2003). The second step of the local similarity search fills in gaps left by the cross-species comparisons and uses a slower more sensitive local alignment algorithm to identify more distant similarities (Fig. 1C; Brudno et al. 2003). Hits are then filtered and collected into loose groups of related CNEs. A final filter removes any conserved regions that are less than 45 bp in length (Fig. 1D). A series of simple heuristic rules vary the thresholds according to the size of the gene family being searched. More details can be found in the Methods section.

We find 20,790 conserved elements across these four genomes (Table 2; Supplemental Data S2). Despite masking out known protein coding exons, 16% of these conserved elements code for peptides with significant similarity to known proteins, probably representing unannotated exons and pseudogenes. Additionally, about 7% of the conserved elements have significant similarity to known RNA species, although these estimates are probably low in fugu and amphioxus since nearly all available chordate RNA sequences are derived from mammals or zebrafish. In the two genomes that have genome-wide untranslated region (UTR) predictions—mice and zebrafish—about 9% of the elements fall within UTRs. The amphioxus conserved elements have a higher proportion of protein-coding sequences (40%), indicating that there are more unannotated coding exons in the amphioxus genome or that over this evolutionary distance *cis*-regulatory sequences become harder to identify, thereby indirectly enriching for conserved protein coding sequences. Because our UTR and RNA detection is biased toward particular species and because functional *cis*-regulatory elements may reside in UTRs, we filter out only the elements with protein similarity and remain open-minded regarding the function of the rest of the conserved elements, declaring them simply phylogenetically conserved noncoding elements (PCNEs). Overall, we find 17,511 PCNEs with a minimum length of 45 bp, including 1299 amphioxus PCNEs. These elements are found in 1281 gene families and around 6345 genes, including 786 amphioxus genes (Table 2). Because these PCNEs are defined within gene families, there is redundancy in cases where multiple gene families claim the same conserved regions. In total, 4980 (28%) of the PCNEs overlap another PCNE. After merging all overlapping PCNEs, there are 14,400 nonredundant PCNE segments, covering 2,444,564 bp across the four genomes.

Across all four organisms, 6% of the PCNEs lie within 1 kb of the transcriptional start site (TSS) of a predicted target gene and are therefore likely to be promoters, but many also lie near the 3′ end and throughout the 100 kb regions upstream and downstream of the genes (Fig. 2). The fish PCNEs are concentrated noticeably closer to their predicted target genes, in line with their more compact genomes (Supplemental Fig. S1).

Of the vertebrates, fugu has the fewest identified PCNEs (4746) (Table 2). This is at least partly due to the fact that the fugu genome has undergone a dramatic compaction (Vandepoele et al. 2004), during which it has been depleted of duplicate genes (9612 fugu genes in our gene families vs. 11449 in zebrafish) (Table 1), leading to fewer paralogous PCNE duplicates. However, there is some evidence that this reduction may have also been caused by increased PCNE loss. Of the 1299 identified amphioxus PCNEs, 432 are similar to at least one mouse PCNE, 379 are similar to at least one zebrafish PCNE, and 305 are similar to at least one fugu PCNE. These counts seem to support recent estimates that there has been increased CNE loss in the teleost—and especially pufferfish—lineages (Stephen et al. 2008; Wang et al. 2009).

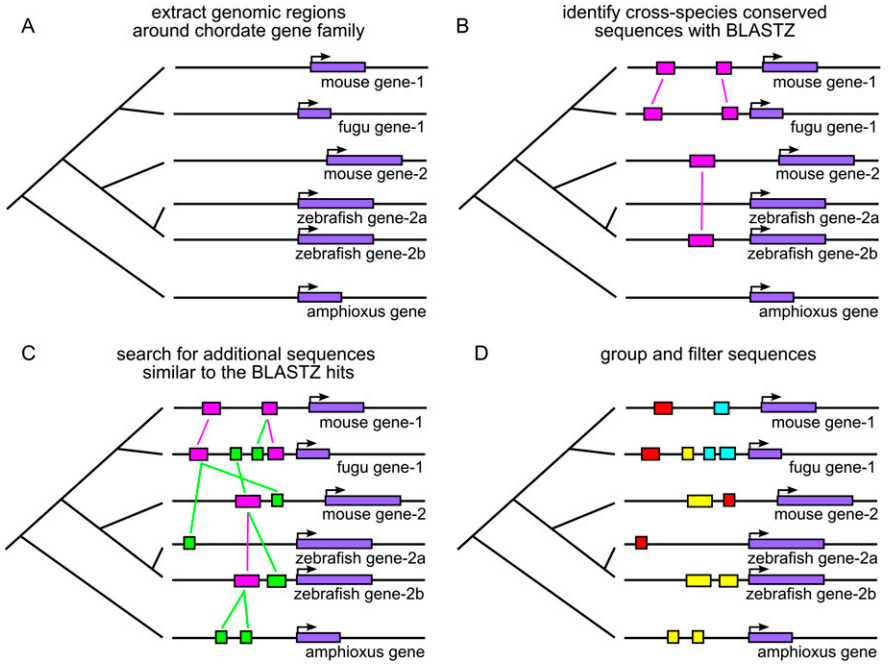

Figure 1. Overview of our method for identifying PCNEs around a hypothetical chordate gene family. (*A*) Genes related to a single ancestral chordate gene are identified, and the genomic sequences surrounding these genes are extracted, extending out 100 kb on either side of the genes. (*B*) A fast search algorithm (BLASTZ) is used to identify conserved noncoding sequences in cross-species comparisons. (*C*) A slower, more sensitive algorithm (CHAOS) searches for additional noncoding sequences similar to the BLASTZ elements. (*D*) The resulting conserved sequences are grouped into loose families and filtered to remove low-confidence sequences and any remaining coding sequences.

Table 2. Conserved noncoding element summary statistics

	Conserved elements[a]	With protein similarity	With RNA similarity	In UTR	PCNEs	Median length PCNEs	Genes with PCNEs	Nonredundant PCNE segments	Coverage (bp)
Total	20,790	3279	1511	1184	17,511	90	6345	14,400	2,444,564
Amphioxus	2174	875	184	0[b]	1299	54	786	1136	76,096
Mouse	6587	1241	393	821	5346	105	2176	4375	836,198
Fugu	5095	349	137	3[b]	4746	96	1477	4069	676,845
Zebrafish	6934	814	797	360	6120	92	1906	4820	855,425

Elements without protein similarity were declared phylogenetically conserved noncoding elements (PCNEs). Merging any overlapping PCNEs creates nonredundant PCNE segments, which are used to calculate the total genomic area covered by PCNEs.
[a]Conserved elements were detected using our gene-family-based method.
[b]Amphioxus and fugu lack genomic UTR predictions.

The thresholds used to identify these PCNEs were chosen to maximize sensitivity while maintaining an acceptably low error rate. We estimated our false-positive rate by replacing each gene in the original gene families with random substitute genes from within the same genome, thereby creating sets of genomic regions that share the same size and species characteristics as the original data. We then repeated our PCNE analysis twice using randomized gene families, finding less than 2% as much PCNE sequence as with real gene families (Table 3). The amphioxus predictions have the highest false-positive rate (10.9% and 2.2%), probably due to the larger evolutionary divergence time between amphioxus and vertebrates than between mammals and fish (>800 Myr vs. ~450 Myr) (Blair and Hedges 2005). While these tests indicate that the majority of PCNEs are the product of real sequence conservation, the primary goal of our method is to detect CNEs that are specific to a certain gene family. To assess the specificity of our predictions, we used the set of PCNEs generated from real gene families, and searched for similarity both in the original genomic sequences that generated the PCNEs and within a second randomized gene family. Across all gene families, the specificity is ~95%; i.e., 95% of the similarity hits (excluding self-hits) generated by the PCNEs lie within the sequence for their original gene family. While amphioxus had the highest false positive rate, its specificity remains quite high (95.5% and 98.4%).

To further confirm that these PCNEs represent deeply conserved sequences, we calculated the phastCons score for each mouse PCNE using existing 17-way genomic alignments (Blanchette et al. 2004; Siepel et al. 2005; Kuhn et al. 2007). PhastCons assigns a score to each aligned base pair in multigenome alignments, ranging from 0 (not conserved) to 1 (most conserved). Our mouse PCNEs have a mean score of 0.708 (95% confidence interval [CI] = 0.696–0.720), with 99.8% of the base pairs aligned, much higher than the average genome-wide conservation for all aligned blocks (0.0829, 95% CI = 0.0826–0.0831). In fact, PCNE conservation is nearly as high as mouse coding regions (0.751, 95% CI = 0.749–0.752).

PCNEs around the Sox14/21 gene family

As an example of our PCNE predictions, Figure 3 illustrates the PCNEs linked to the *Sox14/21* gene family, including a partial sequence alignment of one of the PCNE groups (Fig. 3B). Some of the fugu *Sox21* CNEs have been previously described by Woolfe et al. (2005); e.g., CNE Sox21_18 overlaps with fugu *Sox21* PCNE 7, and Sox21_19 is largely the same as fugu *Sox21* PCNE 1. Our results indicate that some of these elements can be found around both the *Sox21* and *Sox14* genes, indicating an ancient origin that probably predates the early vertebrate WGDs. PCNE order is largely conserved within the vertebrate *Sox21* and *Sox14* subfamilies, and some aspects of positioning are conserved throughout the entire family (group 5 and 1 PCNEs are always downstream, while group 2 PCNEs are always upstream). Despite conservation of PCNE order, the neighboring genes are mostly different in each genomic region, indicating that extensive gene rearrangement is possible even when PCNE order is conserved. Nonetheless, there is some evidence of ancient synteny conservation—mouse *Sox21*, fugu *Sox21*, and amphioxus *SoxB2* all lie near a heparan sulfate 6-O-sulfotransferase gene (*Hs6st3*), although it sometimes lies outside the regions shown in Figure 3A. Moreover, several genes in the family lie near a Claudin gene; however, this syntenic association

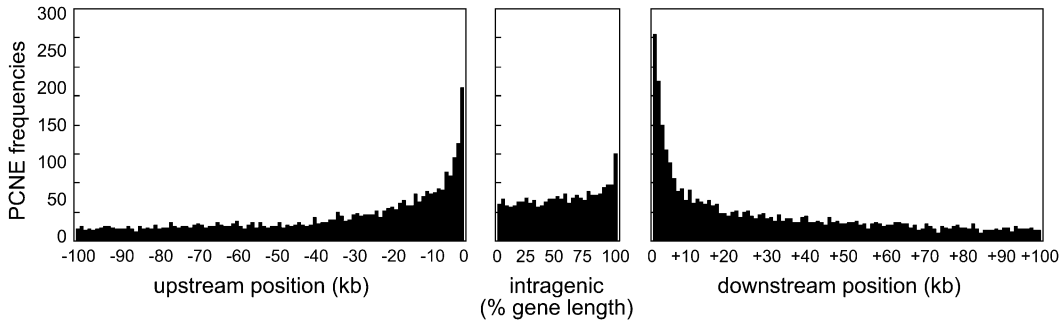

Figure 2. Histograms of PCNE locations relative to their parent genes across all four organisms studied. The *right* and *left* panels show PCNE distributions upstream and downstream (bin width is 1 kb). There is an increase in elements near the TSS and at the 3′ end, but many conserved elements continue to be found throughout the 100-kb flanking regions. In fact, 48% are found more than 10 kb from their associated gene. The *middle* panel shows the distribution of elements found within the introns and UTRs of their parent genes, using relative bin sizes. The mean size of genes with PCNEs is 31.026 kb, so the space within each gene was divided into 31 equal bins, with a mean width of ~1 kb.

Table 3. PCNE error and specificity

	Random PCNEs 1 (bp)	Random PCNEs 2 (bp)	Specificity 1	Specificity 2
Total	44,199 (1.8%)	14,992 (0.6%)	94.50%	95.70%
Amphioxus	8326 (10.9%)	1665 (2.2%)	95.50%	98.40%
Mouse	26,047 (3.1%)	6195 (0.7%)	92.00%	92.30%
Fugu	3803 (0.6%)	5300 (0.8%)	94.20%	94.60%
Zebrafish	6023 (0.7%)	1832 (0.2%)	95.80%	97.20%

PCNE predictions were repeated twice with randomized genes families, and used to quantify the error and specificity of our method.

is not present in amphioxus. Fugu *Sox14a* lies near a heparan sulfate 2-O-sulfotransferase gene (*Hs2st1*), but this gene does not appear to be closely related to the heparan sulfate 6-O-sulfotransferases, so its position may not be the product of syntenic conservation.

Many PCNEs are functional regulatory elements

We tested 42 PCNEs for in vivo enhancer activity in transgenic zebrafish embryos (Methods) and found that about half of the elements show significant activity (Table 4). We used a previously described transgenic protocol that relies on PCR amplification of the putative enhancer sequences, followed by coinjection with a GFP reporter (Müller et al. 1997; Woolfe et al. 2005). Embryos injected with an active enhancer sequence show a highly mosaic pattern of GFP expression, which can then be compiled into a composite expression pattern for each tested PCNE (Supplemental Fig. 2B–D). To determine which PCNEs have enhancer activity, we compare the number of GFP expressing embryos observed when injected with each PCNE to the number observed in injections with the GFP vector alone, and assess the significance of any difference with a conservative statistical test (Williams-corrected G-test) (Table 4). Overall, 23 out of 42 tested PCNE showed significant enhancer activity after multiple test correction (55%). In general, a higher proportion of PCNEs from the fish genomes showed significant enhancer activity (13/20, 65%), while fewer of the amphioxus elements were significantly positive (10/22, 45%); although considering the small sample sizes, these differences are not significant. Nine negative control sequence fragments of similar lengths were

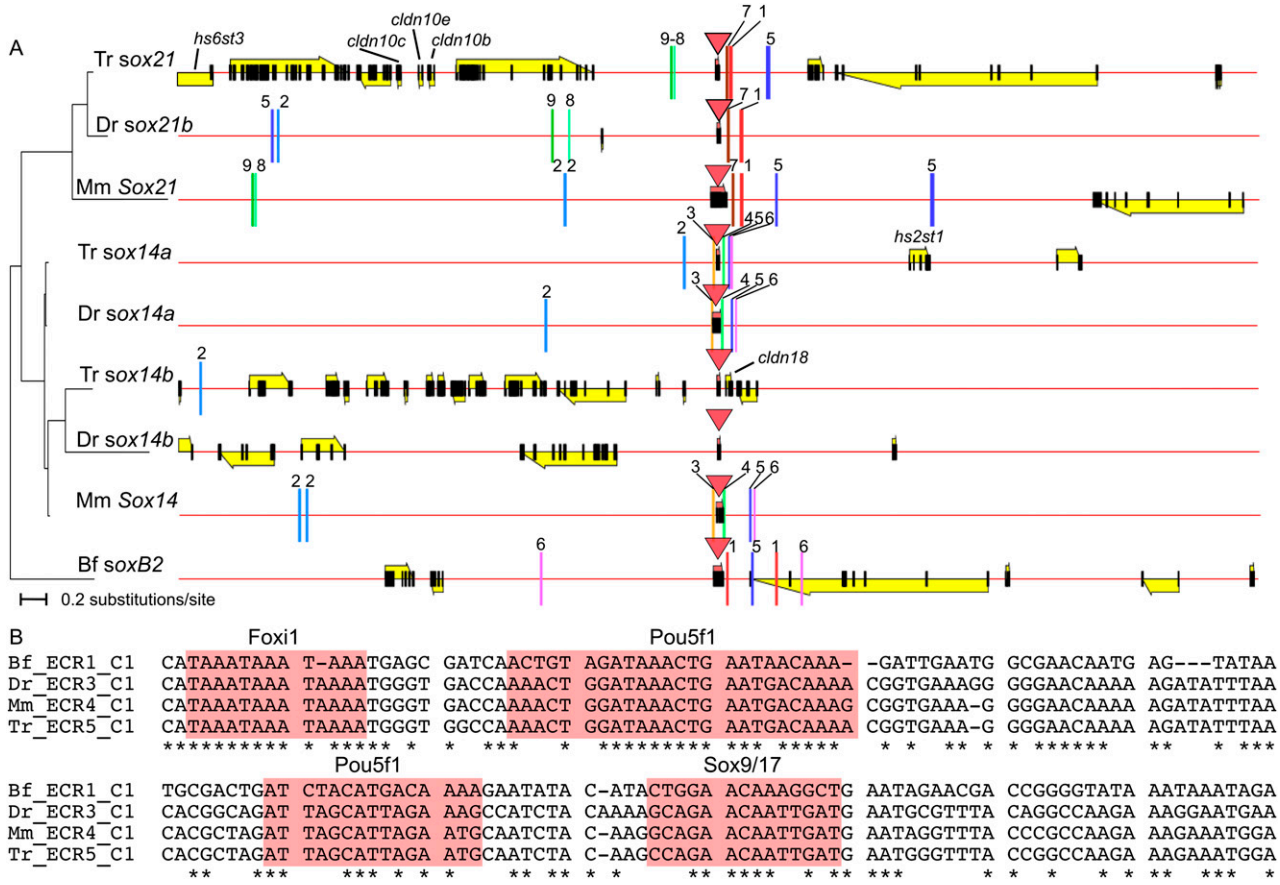

Figure 3. (*A*) PCNEs associated with the *Sox14/21* gene family. Each genomic region is 204 kb long. Genes are shown as arrows indicating the direction of transcription. Black boxes indicate exons. The small single-exon *Sox14/21* genes are labeled with red triangles. PCNEs are shown as colored lines; the colors and the numbers *above* the PCNEs indicate the group membership of each PCNE and reveal conservation of PCNE order. Neighboring genes that may represent conserved syntenic relationships are labeled; unlabeled genes are not syntenically conserved. A maximum likelihood phylogenetic tree of the Sox14/21 proteins is shown on the *right*. (*B*) A multiple alignment of a portion of the Sox14/21 group 1 PCNEs, with conserved binding motifs highlighted in pink. Bf, amphioxus; Tr, fugu; Dr, zebrafish; Mm, mouse.

Table 4. In vivo enhancer activity of tested PCNEs

PCNE	Species	Parent gene	Relative position	Length	GFP positive	Total	Percent Positive	G-test *P*-value
2031-Bf_ECR1_C2	Bf	*meis*	Upstream	150	2	187	1.07%	0.525
2791-Bf_ECR10_C1	Bf	*elav-like*	Downstream	45	5	344	1.45%	0.629
2791-Bf_ECR5_C4	Bf	*elav-like*	Upstream	51	20	154	12.99%	2.99×10^{-5}[a]
2791-Dr_ECR2_C1	Dr	*elavl4*	Downstream	391	11	56	19.64%	2.59×10^{-5}[a]
2791-Tr_ECR4_C4	Tr	*elavl4*	Intronic	144	2	115	1.74%	0.867
3062-Bf_ECR8_C3	Bf	*shh*	Upstream	57	2	243	0.82%	0.352
3352-Bf_ECR32_C5	Bf	*pax6*	Upstream	48	22	365	6.03%	0.0272[a]
3621-Bf_ECR11_C2	Bf	*six3/6*	Downstream	46	12	111	10.81%	1.79×10^{-3}[a]
3621-Tr_ECR7_C2	Tr	*six3*	Downstream	485	38	343	11.08%	1.44×10^{-5}[a]
3643-Bf_ECR1_C1	Bf	*six1/2*	Downstream	172	25	161	15.53%	7.65×10^{-7}[a]
3643-Tr_ECR4_C1	Tr	*six2*	Upstream	175	2	180	1.11%	0.551
3703-Bf_ECR2_C8	Bf	*pou3*	Upstream	67	8	140	5.71%	0.124
3958-Bf_ECR24_C1	Bf	*tfap2*	Downstream	49	6	138	4.35%	0.352
4517-Bf_ECR2_C4	Bf	*sp5*	Upstream	227	31	415	7.47%	3.01×10^{-3}[a]
4517-Tr_ECR5_C4	Tr	*sp5*	Upstream	447	48	258	18.60%	1.31×10^{-10}[a]
4881-Dr_ECR1_C11	Dr	*sox2*	Downstream	491	146	371	39.35%	1.11×10^{-33}[a]
4881-Dr_ECR1_C12	Dr	*sox2*	Upstream	232	37	96	38.54%	2.83×10^{-18}[a]
4881-Tr_ECR3_C12	Tr	*sox2*	Downstream	68	11	76	14.47%	3.03×10^{-4}[a]
4881-Tr_ECR5_C11	Tr	*sox2*	TSS	45	36	212	16.98%	1.03×10^{-8}[a]
496-Tr_ECR1_C3	Tr	*myo3b*	Upstream	477	12	259	4.63%	0.192
5157-Bf_ECR2_C2	Bf	*tob1/2*	Intronic	55	82	120	68.33%	1.52×10^{-46}[a]
5157-Tr_ECR1_C2	Tr	*tob1*	Upstream	222	10	115	8.70%	0.0134[a]
5349-Bf_ECR3_C2	Bf	*acbd6*	Upstream	72	3	126	2.38%	0.902
5596-Bf_ECR1_C1	Bf	*soxB2*	Downstream	351	113	363	31.13%	3.41×10^{-24}[a]
5596-Dr_ECR3_C1	Dr	*sox21b*	Downstream	586	7	179	3.91%	0.395
5596-Tr_ECR5_C1	Tr	*sox21*	Downstream	554	64	151	42.38%	2.69×10^{-26}[a]
5672-Tr_ECR4_C12	Tr	*lmo1*	Intronic	64	33	124	26.61%	1.27×10^{-12}[a]
5694-Bf_ECR5_C3	Bf	*barH-like*	Intronic	45	51	313	16.29%	1.96×10^{-9}[a]
5694-Bf_ECR6_C6	Bf	*barH-like*	Downstream	45	14	386	3.63%	0.386
5694-Dr_ECR1_C14	Dr	*barhl2*	Upstream	52	39	168	23.21%	3.89×10^{-12}[a]
5694-Dr_ECR1_C6	Dr	*barhl1.1*	Upstream	299	41	115	35.65%	3.35×10^{-18}[a]
5694-Tr_ECR20_C3	Tr	*barhl2*	Upstream	285	8	298	2.68%	0.733
5694-Tr_ECR6_C14	Tr	*barhl2*	Upstream	451	7	229	3.06%	0.623
5848-Bf_ECR2_C1	Bf	*msx*	Upstream	239	160	574	27.87%	1.80×10^{-23}[a]
5882-Bf_ECR12_C2	Bf	*muscleblind*	Downstream	48	1	278	0.36%	0.108
6281-Bf_ECR2_C1	Bf	*gbx*	Downstream	303	2	172	1.16%	0.585
6436-Bf_ECR2_C1	Bf	*hand*	Downstream	56	10	240	4.17%	0.314
6456-Bf_ECR2_C1	Bf	*dlx*	Downstream	54	2	150	1.33%	0.65
67-Bf_ECR4_C4	Bf	*ptpr S/D/F*	Upstream	46	3	145	2.07%	0.973
67-Dr_ECR2_C4	Dr	*ptprd*	Upstream	69	31	152	20.39%	1.05×10^{-9}[a]
67-Dr_ECR7_C2	Dr	*ptprf*	Downstream	76	3	86	3.49%	0.623
8085-Bf_ECR1_C1	Bf	*id*	Upstream	192	54	454	11.89%	1.10×10^{-6}[a]
rand_602-215234-215293	Bf	None			2	78	2.56%	0.867
rand_2286-4103-4162	Bf	None			3	218	1.38%	0.629
rand_1452-2928-2987	Bf	None			5	131	3.82%	0.476
rand_2889-297-356	Bf	None			4	498	0.80%	0.24
rand_2370-6155-6214	Bf	None			9	126	7.14%	0.0557
rand_1028-6323-6382	Bf	None			0	224	0.00%	0.0247
rand_16-46219015-46219064	Mm	None			27	293	9.22%	4.63×10^{-4}[a]
rand_12-46392970-46393019	Mm	None			0	189	0.00%	0.0353
rand_8-31645402-31645451	Dr	None			0	236	0.00%	0.0222

Significant enhancer activity was assessed by Williams-corrected G-test, vs. vector alone, and corrected for multiple testing by the Benjamini-Hochberg method. Randomly selected negative control elements are at the bottom of the table. Bf, amphioxus; Tr, fugu; Dr, zebrafish; Mm, mouse.
[a]Element shows significant enhancer activity versus vector alone ($P \leq 0.05$).

chosen randomly from the amphioxus, mouse, or zebrafish genomes and tested for enhancer activity—only one showed significant activity (1/9, 11%). Overall, the PCNEs had a significantly greater mean enhancer activity than the negative controls: 13.7% GFP-expressing embryos for PCNEs versus 2.8% GFP-expressing embryos for random sequences ($P = 0.031$, permutation test). To evaluate the stringency of our final PCNE length threshold, we also tested 11 conserved sequences that did not make our 45-bp cutoff—only one of these sequences showed significant enhancer activity (Supplemental Fig. 2A). Many of the tested regulatory elements drive expression in tissue specific patterns. Some examples are shown in Supplemental Figure 2, B through D, and the complete set of observed tissue patterns can be found in Supplemental Data S4.

Conserved noncoding sequences are strongly associated with gene order conservation

Our results indicate that extensive gene order rearrangement can occur in regions with highly conserved PCNE order (Fig. 3), further

suggesting that the association between PCNEs and gene order (synteny) conservation deserves rigorous testing. Here, we have compared our entirely local PCNE predictions to estimates of ancient synteny conservation, using a gene-pair-based synteny method that we have previously described (Panopoulou et al. 2003; Hufton et al. 2008). This synteny method relies entirely on the protein sequence of genes and is therefore not affected by the conservation of surrounding noncoding sequences. A gene is said to possess ancient synteny if it lies within a proximate gene pair that is also present in amphioxus (illustrated in Fig. 5A, below). This method is especially appropriate for our current question because it defines synteny on a gene-by-gene basis rather than across genomic segments, thereby allowing us to analyze the association between PCNEs and synteny conservation with high resolution. For these analyses, and all further analyses presented in this article, we have removed all olfactory receptor genes from our gene sets because their extensive tandem duplication can dominate whole-genome comparisons (for further discussion, see Methods).

We observe that PCNEs located outside of their predicted target genes (extragenic PCNEs) are strongly associated with conservation of synteny (Fig. 4A). The presence of at least one extragenic PCNE increases the probability of synteny conservation by about 10% in mice and zebrafish and 5% in fugu. Additional extragenic PCNEs smoothly increase the probability of synteny conservation, such that when genes have at least 15 extragenic PCNEs, more than 80% of mouse and zebrafish genes have evidence of anciently conserved synteny. Some concern was raised that the most distal PCNEs—near our 100-kb boundary—might be more prone to error; however, repeating the analysis presented in Figure 4A with only extragenic PCNEs that lay within 50 kb or 10 kb of their predicted target genes creates similarly strong associations between PCNE number and synteny conservation (Supplemental Fig. 3). PCNEs located within their predicted target gene's introns or UTRs have a much weaker, but also significant, correlation with synteny conservation (Fig. 4B). This trend is significantly weaker than the one observed with extragenic elements (tested by permutation of the data; see Methods).

Interestingly, gene families that have conserved their syntenic relationships in multiple places in a single vertebrate genome have more extragenic PCNEs than other syntenic genes (Fig. 5). For a gene family to have synteny in multiple locations within the genome, the genes must have been duplicated and retained during vertebrate evolution. However, gene duplication alone does not appear to explain the trend. Genes retained in duplicate after the 2R WGD events show only a weak increase in the number of linked extragenic PCNEs (Fig. 5B). Moreover, genes that are both anciently syntenic and retained WGD duplicates do not have significantly more PCNEs than do genes with only ancient synteny. The simplest explanation for this observation is that it is not the presence of duplication, per se, that increases the association between synteny and conserved elements. Rather, when a gene family has conserved its syntenic neighborhood in multiple genomic locations, there is likely to be exceptional purifying selection acting to preserve the surrounding genome architecture. These cases of exceptional synteny preservation seem to be strongly associated with genomic regions that are densely packed with PCNEs.

Genes located within segmental duplicates are also enriched for extragenic PCNEs in mice and zebrafish (for an explanation of segmental duplicate identification, see Methods) (Fig. 5B). This is surprising since segmental duplications disrupt synteny and thereby risk separating extragenic *cis*-regulatory sequences from their target genes. Nevertheless, some reports have described cases where tandemly duplicated genes have retained the same *cis*-regulation as their parental copy (e.g., Ponce and Hartl 2006). We do not observe this trend in fugu, but this may be caused by the paucity of segmentally duplicated genes in the fugu genome (155 fugu genes vs. 792 mouse genes and 1967 zebrafish genes). The olfactory receptor genes within the mouse genome are a prime example of this phenomenon of tandem duplication while maintaining PCNEs (Fig. 5B). These genes form five families representing five genes in the chordate ancestor, which have then expanded to 1127 genes in mice. These tandem arrays of olfactory genes contain 252 nonredundant PCNEs, and each olfactory gene is linked on average to one PCNE (Fig. 5B).

Different functional gene classes show distinct patterns of genome evolution

Our results indicate that genes with synteny conservation, genes retained after WGD, and segmentally duplicated genes are all

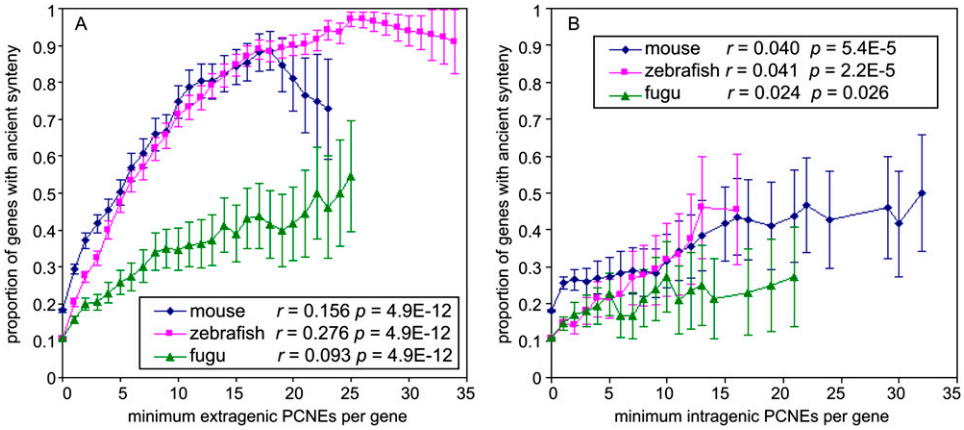

Figure 4. Genes linked to PCNEs are more likely to have conserved synteny. The association between PCNEs and synteny conservation in three vertebrate organisms for extragenic PCNEs (not in an intron or UTR of their associated gene) (*A*) and intragenic PCNEs (in an intron or UTR of their associated gene) (*B*). As genes are required to have a greater number of associated PCNEs (*x*-axis), the proportion of genes with anciently conserved synteny tends to increase (*y*-axis). Error bars, standard error of the proportion. Each data point is calculated from at least 10 genes. All of these trends represent a significant positive correlation between PCNE number and synteny conservation, where *r* is the point biserial correlation coefficient, and *p* is the Bonferroni-corrected probability that the true correlation is greater than zero.

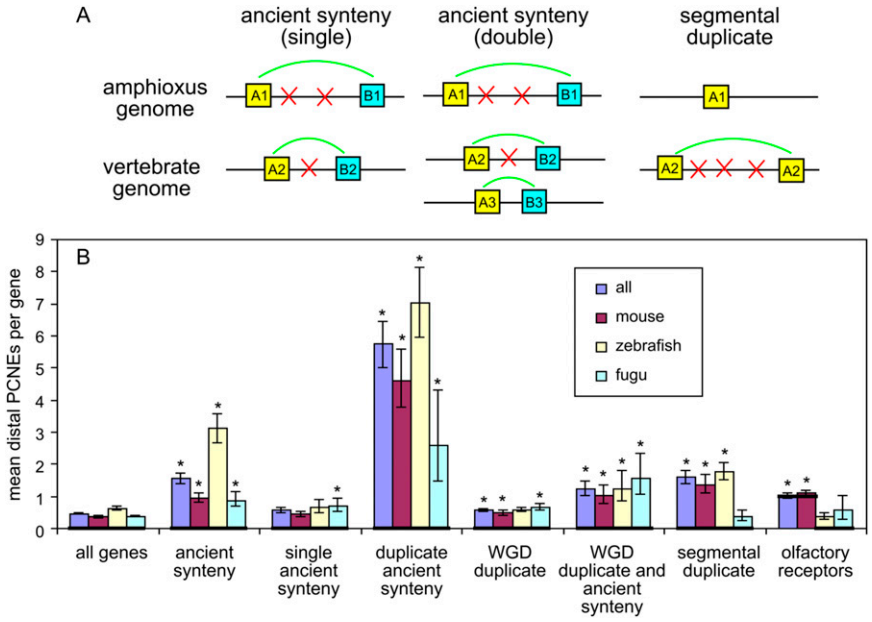

Figure 5. Genes with different synteny and duplication patterns have different PCNE amounts. (*A*) Illustration of our synteny and segmental duplication definitions. (*B*) The mean number of extragenic PCNEs linked to different synteny and duplication related gene sets. Error bars, bootstrap-based 95% confidence intervals on the mean estimates. Asterisks denote PCNE enrichments that are significant at $\alpha = 0.05$ by a Bonferroni-corrected permutation test. Bold lines show the population medians, which are zero in all cases except for the olfactory receptor genes. This serves to emphasize the fact that in all populations the majority of genes have no PCNEs, and the observed differences between the gene sets, while significant, are generally the result of changes in PCNE number around a subset of genes.

positively associated with PCNEs. Many of these processes have been previously associated with specific gene functions. For example, genes near PCNEs and genes retained after WGD tend to function in transcription and development (Bejerano et al. 2004; Woolfe et al. 2005; Blomme et al. 2006; Brunet et al. 2006), and anciently syntenic genes are enriched for metabolic, catalytic, and ribosomal functions (Hufton et al. 2008). However, a unified comparison of these functional associations has not yet been attempted. Here, we have used a clustering approach to dissect the genome evolution patterns of different gene functions, providing us with a detailed picture of how these different genome-evolution processes are interrelated.

First, we identified Gene Ontology (GO) terms that were significantly associated with five different genome-evolution processes: linkage to PCNEs, WGD duplicate retention, segmental duplication, conservation of synteny, and bidirectional promoters (for the criteria used to classify genes and determine significant GO enrichments, see Methods). Bidirectional promoters were included because a previous report has shown that they act to promote synteny conservation (Li et al. 2006). For each relevant GO term, we then calculated an enrichment or depletion value for each genome-evolutionary process and then used these values to cluster the GO terms with a hierarchical method commonly used in the microarray analysis field (Methods). Figure 6 shows the mouse and zebrafish clustering results for the 50 molecular function and biological process GO terms with the most significant associations.

This clustering procedure divides the GO terms into functionally similar groups with distinct patterns of genome evolution: (1) transcription, general metabolism, and development related terms; (2) molecular transducers, specifically ion transport and kinase related terms; (3) DNA/RNA/protein metabolism; and (4) a group of catalytic terms (Fig. 6). Within the mouse, these broad groups show additional subdivision: Developmental terms cluster separately from transcription and metabolism; ion transport and kinase activity separate into distinct clusters; and a histone cluster is created. This additional resolution is at least partly due to the fact that the mouse genome enjoys more thorough functional annotation than the zebrafish genome, with approximately twice as many GO terms attributed to genes. Regardless, the overall clustering patterns are clearly similar. Since the two lineages diverged about 450 Myr ago (Blair and Hedges 2005), this suggests that these functional biases are generated by shared evolutionary mechanisms and not lineage-specific adaptations.

The transcription, metabolism, and development clusters in mouse and zebrafish account for 45% of the GO terms in the lists and appear to reflect the well-recognized association of CNEs and WGD duplicates with transcriptional and developmental processes (Fig. 6). The metabolism terms in these clusters tend to be very general, and most are parents to the more specific transcription-related terms, suggesting that transcription-related processes may play a large role in driving the evolutionary patterns observed for these functions (e.g., the term "RNA biosynthetic process" is a child of "metabolic process," "primary metabolic process," "cellular process," and "biosynthetic process"). As expected, nearly all of the gene functions in these clusters are positively associated with PCNEs and synteny conservation but negatively associated with segmental duplications. The majority also show increased retention of WGD duplicates; however, this is not consistent for all transcription and metabolism terms.

Interestingly, the differences in post-WGD gene retention divide the transcription, metabolism, and development GO terms into general and regulatory subgroups, revealing that the gene functions associated with PCNEs and those retained after WGD have important differences (Fig. 6). The general transcription and metabolism GO terms are not consistently over-retained after WGD and are enriched for bidirectional promoter pairings, while the regulatory and developmental GO terms are depleted for bidirectional promoters and have a stronger association with WGD gene retention. The boundaries between these subgroups are somewhat fuzzy; "DNA binding" is in the general group in the mouse cluster but in the regulatory subgroup in the zebrafish cluster. This is not surprising since many transcription factors are annotated with a mixture of terms from both clusters. A good example of the general subgroup genes is general transcription factor IIIC, polypeptide 4 (*Gtf3c4*), which encodes a protein involved in the production of small nuclear and cytoplasmic RNAs (Hsieh et al. 1999). This gene has 10 PCNEs and shares a bidirectional promoter with *Ddx31*, a helicase molecule presumably involved in RNA unwinding (Abdelhaleem et al. 2003). The best prototypes for the regulatory subgroup are well-known development transcription factors, like the *Hox* genes, which have anciently conserved synteny, many

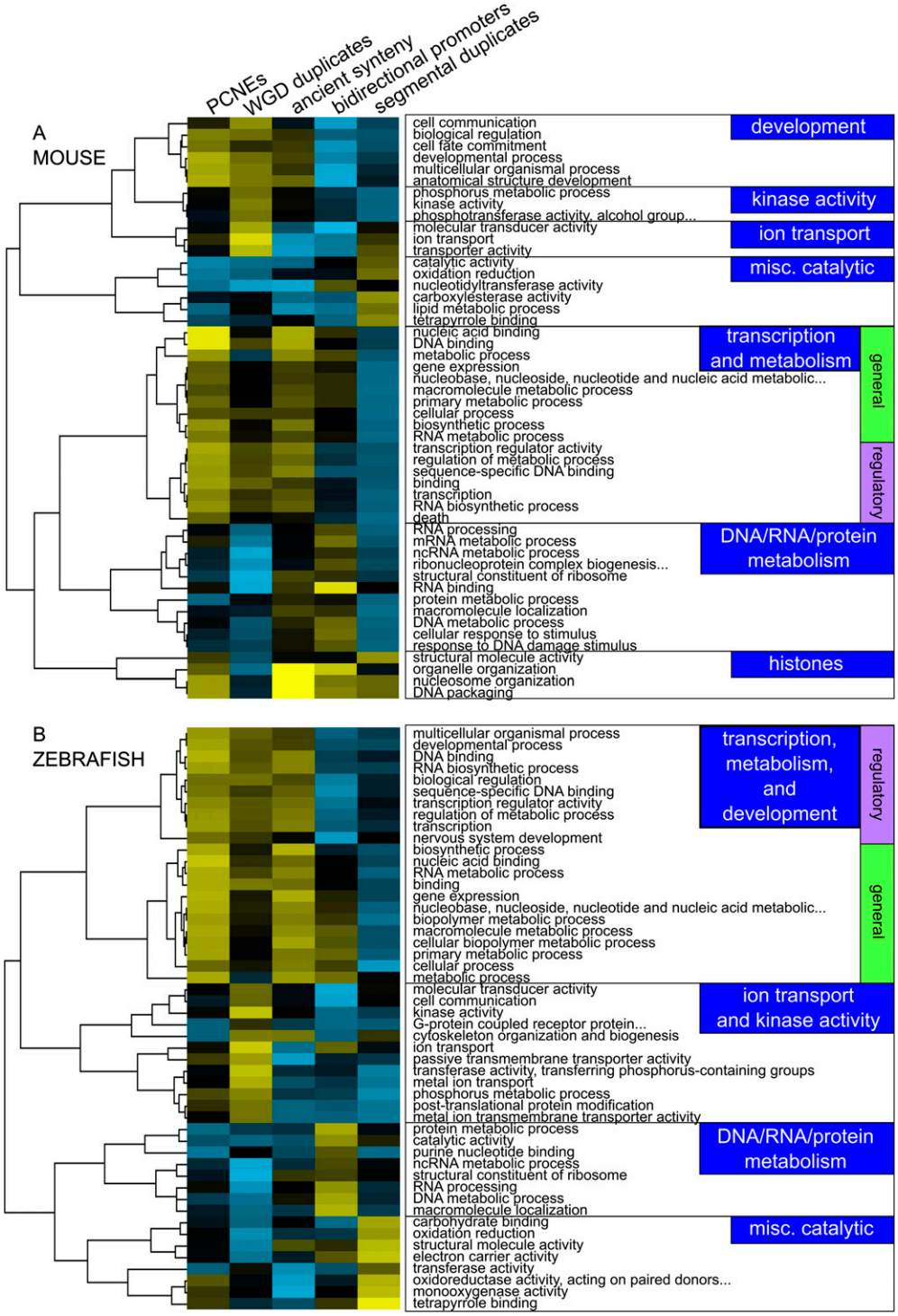

Figure 6. Different gene functions are associated with distinct patterns of genome evolution. Hierarchical clustering of gene functions (rows) and five genome evolutionary processes (columns): linkage to PCNEs, retention of duplicate genes after the early vertebrate WGD events, synteny conservation, bidirectional promoter, and segmental duplication. The gene functions are derived from the biological process and molecular function namespaces of the Gene Ontology; the 50 classes with the strongest associations are shown, using data from the mouse (*A*) and zebrafish (*B*) genomes. Yellow colors indicate that genes annotated with the GO term are more strongly associated with the genome evolutionary process than the genome-wide average; blue colors indicate less than the genome-wide average. The clustering procedure divides GO terms into groups that represent broad cellular processes (labels on the *right*), each of which is associated with a distinct pattern of genome evolution. All olfactory receptor genes were removed from the data prior to this analysis.

nearby PCNEs, and have retained their WGD duplicates. Overall, these results seem to indicate that different evolutionary pressures act on low-level transcriptional machinery and metabolic genes versus genes with more distinct regulatory functions.

Supporting the idea that PCNEs and WGD duplicate retention are not associated with identical gene sets, signaling-related gene functions are strongly favored for retention after WGD but have weak or no association with PCNEs. These terms include specific enrichments for functions related to ion transport and kinase activity, in addition to the general signaling term "molecular transducer" (Fig. 6).

The two remaining clusters—DNA/RNA/protein metabolism and miscellaneous catalytic—are generally not retained after WGD (Fig. 6). The DNA/RNA/protein metabolism cluster includes genes that are involved in building, processing, and degrading DNA, RNA, and proteins, including riboproteins, proteins involved in ubiquitin-mediated degradation (e.g., f-boxes), chaperones (e.g., prefoldin), RNA processing proteins, and DNA metabolic enzymes. These genes are enriched for bidirectional promoters but appear to avoid any form of duplication. The miscellaneous catalytic cluster includes a diverse set of terms describing catalysis-related functions. In both organisms, this cluster contains the GO terms "oxidation reduction" and "tetrapyrrole binding," indicating a common focus on redox processes. These gene functions have been favored for segmental duplication, avoid WGD retention, and generally do not retain their synteny.

Vertebrate protein connectivity estimates support the GBH

The GBH predicts that genes with many protein–protein interactions (PPIs) should be over-retained after WGD and should avoid segmental duplication in order to conserve the relative stoichiometries of their interactions. We tested this prediction by estimating the number of PPIs for all genes in the mouse genome and then comparing the mean number of interactions for genes in different genome-evolutionary categories (Fig. 7). Genome-wide PPI data are still at an early stage in vertebrates, so we inferred PPI counts from two different databases—Human Protein Reference Database (HPRD) (Peri et al. 2003; Mathivanan et al. 2006; Mishra et al. 2006) and HomoMINT (Persico et al. 2005)—to help control for biases introduced by different database methodologies (see Methods).

As predicted by the GBH, genes that have been retained in duplicate after WGD have significantly more protein interactions than the genome mean, and genes that have undergone segmental duplication have significantly fewer interactions (Fig. 7). We also observe that genes with PCNEs are significantly overconnected, which is not surprising considering that the gene functions associated with PCNEs and WGD duplicate retention are largely—but not entirely—overlapping (Fig. 6). As usual, olfactory receptor genes were excluded from these gene sets, but like other segmental duplications, they are extremely low in protein interactions (HPRD mean = 0.007, median = 0). Genes with ancient synteny or with bidirectional promoters showed protein connectivity values similar to the genomic mean (Fig. 7).

Discussion

Identifying CNEs between distantly related species

Sequence conservation remains one of the most powerful methods available to identify *cis*-regulatory sequences, and our results indicate that these methods may be useful in more distant species comparisons than has been widely appreciated. Here we show that by using sensitive local sequence searches anchored to phylogenetically defined gene families we can identify 1299 PCNEs conserved between amphioxus and vertebrates (or 1136 when overlapping PCNEs are merged). Moreover, 45% of 22 tested amphioxus PCNEs demonstrate enhancer activity in zebrafish embryos, indicating functional conservation of both the enhancer and the transcriptional machinery that interprets these sequences, across more than 800 Myr of divergence (Blair and Hedges 2005). Overall, these findings provide a rich source of anciently conserved *cis*-regulation, possibly representing essential chordate regulatory programs.

We can detect PCNEs between extremely distant species pairs in part because we search genomic regions around orthologous gene families, rather than make whole-genome comparisons. This approach was inspired by a similar method published by Sanges et al. (2006), who showed that gene-anchored searches could greatly increase CNE detection between mammals and fish. In addition to their added power, gene-anchored methods implicitly predict a target for each identified CNE. Compared with the Sanges et al. (2006) method, our method does not rely on initial multiple alignments and uses a set of simple rules to adapt to gene families of very different sizes. Therefore, our resulting PCNEs are not biased by local genome rearrangement or gene duplication and are well suited for further evolutionary analyses.

Despite the power and flexibility of our PCNE method, we find that our ability to predict functional *cis*-regulatory sequences

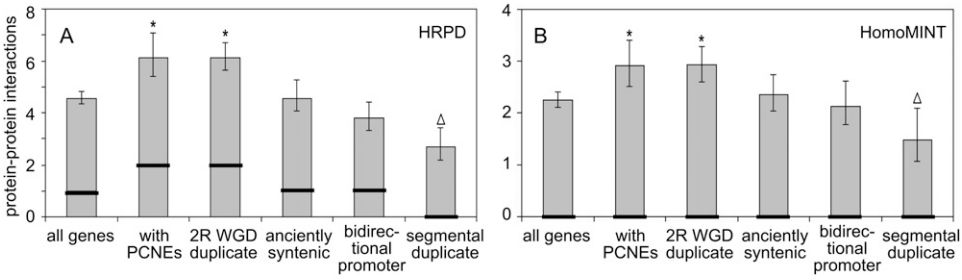

Figure 7. Retained WGD duplicates are overconnected and segmental duplicates are underconnected. The degree of protein–protein interaction (PPI) connectivity was estimated for mouse gene sets with different genome evolutionary characteristics using two different PPI databases, HPRD (*A*) and HomoMINT (*B*). Bar heights indicate the mean connectivity for each gene set. Each mean is bounded by a bootstrap-based 95% confidence interval. Median connectivity is indicated as a bold line, and the difference between the median and mean gives some indication of the right-skew of the connectivity distribution for each gene set. Genes with PCNEs or that are retained WGD duplicates have significantly more PPI connections than the genome mean (asterisks), while genes that are the result of segmental duplication have significantly fewer connections (triangles). Significance was assessed by a Bonferroni-corrected permutation test, $\alpha = 0.05$.

is comparable to previous reports. Holland et al. (2008) tested eight human CNEs identified by human–amphioxus whole-genome comparisons and found that four had enhancer function in mouse (50%). We test 22 amphioxus PCNEs and find that 10 are functional enhancers in zebrafish (45%). It should be noted that, in contrast to Holland et al. (2008), we tested the amphioxus sequences directly rather than their vertebrate orthologs. Among the 20 vertebrate PCNEs that we tested, 13 showed significant enhancer activity (65%). Previous fish–mammal CNE analyses have reported somewhat higher rates of enhancer activity: Sanges et al. (2006) reported 81% (20/27), and Woolfe et al. (2005) reported 92% (23/25). However, at least some of the disparity is due to methodological and statistical analysis differences. Here, we declare significant enhancer activity using a conservative and multiple-test corrected statistical test. Similarly, Sanges et al. (2006) reported that only 48% (13/27) of their tested elements had statistically significant general enhancer activity. Moreover, the difference between our results and Woolfe et al. (2005) seem to stem largely from the fact that we observed a higher rate of enhancer activity in our vector controls, despite using the same assay system, thereby increasing the amount of enhancer activity required to declare a PCNE significantly active (2% of control embryos expressed GFP vs. 0.2% in Woolfe et al. [2005]). Since both of these previous reports used mammalian multiple alignments that constrained the colinearity of CNEs, it is possible that constraints on colinearity may aid in the discovery of functional *cis*-regulatory sequences. Indeed, this seems plausible considering the strong correlation that we observe between synteny conservation and PCNEs (Fig. 4). Nonetheless, methods—like the one described here—that do not constrain colinearity may be able to identify functional elements in highly rearranged regions where other methods fail.

PCNEs conserve genomic synteny

Our results provide the most rigorous evidence yet that CNEs are correlated with the conservation of genome synteny. In fact, our results show that mouse or zebrafish genes linked to enough extragenic PCNEs are nearly guaranteed to have preserved some aspects of their genomic neighborhoods (>85% synteny conservation with >16 extragenic PCNEs) (Fig. 4A). The extent of this conservation is particularly striking when one considers that our definition of ancient synteny requires conservation of the proximate gene pair in both the amphioxus and vertebrate lineages. This suggests that the orthologous amphioxus genes also have complex highly conserved *cis*-regulation that is constraining their synteny, even though in many cases we cannot identify amphioxus PCNEs. In essence, complex *cis*-regulation seems to have frozen certain aspects of genome organization during both amphioxus and vertebrate evolution. Nonetheless, these tight constraints appear to have been relaxed in the pufferfish lineage, where there has also been dramatic genome compaction and accelerated loss of ancient CNEs (observed here and in Wang et al. 2009 and Stephen et al. 2008).

The MacKenzie et al. (2004) proposition that distant *cis*-regulation should act to conserve genome synteny was based on the idea of "gene interdigitation"—neighboring genes that possess spatially interlocking *cis*-regulation. Our results support this hypothesis by showing that PCNEs located outside of their target genes—those that are mostly likely to interlock with the *cis*-regulation of neighboring genes—have the greatest effect on synteny conservation. Moreover, the clear difference between extragenic and intragenic PCNEs indicates that the observed correlation between PCNEs and synteny is not caused by some passive third factor, such as a lower mutation rate within syntenically conserved regions (Fig. 4).

In addition to gene interdigitation, some of the observed synteny conservation is likely to be caused by PCNEs that regulate multiple targets, possibly explaining the increase in genome synteny conservation associated with intragenic PCNEs (Fig. 4B). In support of this mechanism, Sémon and Duret (2006) observed that neighboring genes with shared expression domains are more likely to retain their genomic synteny. We note that our gene-based PCNE prediction algorithm allows us to identify cases where PCNEs are shared by multiple genes, suggesting that investigators may be able to use methods like ours to identify candidate *cis*-regulatory sequences that control coexpressed gene clusters.

Bidirectional promoters are an additional source of synteny conservation within vertebrate genomes. Previous reports noted that genes in bidirectional promoter pairs tend to conserve their synteny (Li et al. 2006), and we observe that the majority of gene functions with bidirectional promoter enrichment are also synteny enriched (17/22 functions in mice and 11/20 in zebrafish) (Fig. 6). Moreover, the synteny conserving effects of bidirectional promoters and PCNEs appear to be largely independent. Overall, bidirectional gene pairs are not enriched for PCNEs (data not shown). Genes in bidirectional pairs but without PCNEs still show strong synteny enrichment (22% have ancient synteny, $P = 9.1 \times 10^{-4}$), as do genes with PCNEs but without bidirectional promoters (20% have ancient synteny, $P = 1.1 \times 10^{-56}$). Moreover, 43% of genes with both PCNEs and bidirectional promoters have anciently conserved synteny, an approximately additive increase in synteny conservation.

The GBH best explains the patterns of gene duplication observed in vertebrates

Our results provide compelling support for the GBH. As predicted by the GBH, we find that gene functions that are preferentially retained after WGD avoid segmental duplications, and vice versa (Fig. 6), supporting previous observations in vertebrates and plants (Blomme et al. 2006; Freeling 2008). In addition, the patterns of gene duplication that we observe seem to be better explained by the GBH than by the DDC model. We identify several transcription and metabolism-related gene functions that are enriched for PCNEs but that are not enriched for gene duplicates (Fig. 6), indicating that complex and essential *cis*-regulation is not sufficient to promote duplicate retention via subfunctionalization. Similarly, molecular transducer functions have been consistently overretained after WGD but have no PCNE enrichment (Fig. 6). These functions include kinases and ion transporters, which have been reliably favored for retention after WGD in both vertebrates and plants (Blanc and Wolfe 2004; Maere et al. 2005).

According to the GBH, genes that are sensitive to dosage imbalances will be retained after WGD and will avoid segmental duplication. Investigators have suggested that proteins with many binding partners are the most likely to be dosage sensitive (Birchler et al. 2001; Veitia 2002; Papp et al. 2003). Consistent with this idea, we observe that genes retained after WGD encode proteins with more binding partners than the genome average, while segmentally duplicated genes encode proteins with fewer partners (Fig. 7). Naturally, estimates of protein–protein connectivity must be interpreted with caution, since current vertebrate PPI databases are certain to contain a variety of biases. Developmental and signaling genes could appear overconnected because of greater research

attention, and predicted interactions may not be well-distributed among paralogous genes, possibly leading to unpredictable biases in regard to gene duplication. These resources will certainly continue to improve, but there may also be other methods available in the future for predicting genes that are likely to be dosage sensitive. Recently, investigators demonstrated that dosage sensitive proteins can be directly identified from their structures—via a property called "under-wrapping"—and in yeast, retained WGD duplicates tend to be "under-wrapped" (Liang et al. 2008).

Nonetheless, there is ample evidence that expression subfunctionalization does occur after gene duplication (Huminiecki and Wolfe 2004; Postlethwait et al. 2004; Li et al. 2005; Duarte et al. 2006; Woolfe and Elgar 2007), and we cannot completely rule out that subfunctionalization contributes to duplicate gene retention. Here, we have reasoned that genes with the most highly conserved *cis*-regulatory elements—genes with PCNEs—would be the most likely to have duplicates retained by DDC, since weakly conserved or nonessential regulation would not force the kind of retention envisioned by the DDC model. However, it is clear that our set of PCNEs represent only a small portion of all *cis*-regulatory elements, so it remains possible that less strongly conserved *cis*-regulation may contribute to gene retention via DDC. Regardless, we must conclude that the GBH appears to better explain the complement of gene duplicates present in vertebrates.

Concluding remarks

These results reveal an intricate web of correlations between a gene's function and the evolutionary fate of its surrounding genomic sequence. The mixture of genes within any genomic region are likely to impose a variety of different constraints, dependent on their functions and regulation, that will control how the whole region is able to duplicate and rearrange. In the end, this may create distinct "genome-evolutionary microenvironments" that influence the evolution of all of the genes within a region.

Methods

Orthology and phylogenetic trees

Genes from the *M. musculus* (mouse), *T. rubripes* (fugu), and *D. rerio* (zebrafish) genomes were grouped into families by comparison to the *B. floridae* (amphioxus) genome using Inparanoid (Remm et al. 2001). This creates families of vertebrate and amphioxus genes, where each family of genes is orthologous to a single gene in the hypothetical chordate ancestor (Table 1; Supplemental Data S1). For each gene family, an automated method builds a phylogenetic tree, using a modified version of the Phylogenie method (Frickey and Lupas 2004). In our version, Phylogenie is run on a protein database consisting of the proteins within a single Inparanoid family. Phylogenie then uses each sequence in the Inparanoid family as a starting point to build a multiple alignment, from which we then select the best available multiple alignment, based on alignment length and the number of sequences included. TREE-PUZZLE is then used to infer a maximum likelihood phylogenetic tree for each family (Schmidt et al. 2002). Gene-models and genomic sequence for the vertebrates were taken from Ensembl version 42 (Flicek et al. 2008); JGI v1.0, for amphioxus (Putnam et al. 2008).

Identifying phylogenetically CNEs

Around each gene in a gene family, we extract genomic sequence within the gene and extending out 100 kb from both the gene start and end and then mask known coding regions and repeats (Fig. 1A). Our 100-kb threshold was chosen to maximize detection of real PCNEs, while controlling our false detection rate and computational running time. About 84% of the PCNEs are within 50 kb of a predicted target gene, and only 5% lie in the last 10 kb. The four organisms analyzed have genomes of very different sizes, and arguably smaller region sizes could have been used for the organisms with more condensed genomes. But the genomes have not been condensed or expanded uniformly, so it is impossible to predict appropriate species-specific region sizes for each gene family. Importantly, using a consistent region size across all species leads to a similar expectation of false-positive hits for each species.

The set of genomic sequences for each family is then searched for conserved noncoding sequences using a two-step local similarity search. The first step uses BLASTZ (Schwartz et al. 2003), a fast but powerful local alignment algorithm, to identify a set of sequences that have been conserved across species (Fig. 2B). For each gene family, all possible cross-species pairwise comparisons are made, except the zebrafish–fugu comparisons. This effectively requires all conserved sequences identified at this stage to show at least 450 Myr of conservation (Blair and Hedges 2005). The BLASTZ searches were run without chaining, so there is no preference for colinear conservation. Sequences identified by the various species comparisons were then merged into a set of nonredundant "BLASTZ-elements."

The gene families being searched can vary widely in size and species representation, and therefore if a static score threshold is used in the BLASTZ searches, the outcome tends to be dominated by random hits found within large gene families. As such, we vary the BLASTZ score threshold (K) according to the size of the sequence space searched for each species pair within each family, according to the formula $K = 126.78 \times ln(mn - 726.88)$, where m and n equal the length of the genomic sequences from each of the two species being compared (masked and ambiguous base pairs are not counted and K is required to be $\geq$1000). This formula is simply a derivation of the formula used to calculate BLAST E-values (Altschul et al. 1997). Parameters were estimated from simulations with random genomic segments from our four organisms, and provide us with a similar expectation of random hits (about three hits per species pair) regardless of the size of the gene family. While three random hits per species pair may seem like a high level of background noise, the second step of our method imposes additional constraints on each BLASTZ-element, thereby removing most false positive sequences.

The second step of the local similarity search fills in gaps left by the cross-species comparisons and uses a slower more sensitive local alignment algorithm, CHAOS, to identify more distant similarities (options: -wl 10, -co 10, -rsc 1800, -ext -v -b) (Fig. 1C; Brudno et al. 2003). In this step, each gene family's BLASTZ-elements are searched against the family's entire set of genomic regions. This allows us to identify sequences conserved between paralogous gene regions and conserved elements that may be present in tandem arrays (Fig. 1C).

After these two phases of local similarity searching we have a set of BLASTZ-elements that each possesses a collection of CHAOS-based similarity hits. At this point, we remove BLASTZ-elements that are most likely to be false positives using two thresholds. First, each is required to have generated hits in the genomic regions around at least two family genes in other species (one is likely to be the BLASTZ partner that originally generated the element). Second, every BLASTZ-element is required to have hits in at least 20% of the genes in the family (rounded down), further helping to maintain a consistent error rate across families of different sizes. After filtering, all remaining BLASTZ-elements are passed through a simple clustering and condensation routine. Any

two BLASTZ-elements are considered linked if one has a CHAOS hit that overlaps the other BLASTZ-element. Single-linkage clusters are then formed, and all BLASTZ and CHAOS hits within a cluster group are merged into a set of nonredundant elements. A final filter removes any conserved region less than 45 bp in length (Fig. 1D). This length threshold was chosen in part based on the results from our in vivo transgenic assay results. Previous reports have shown that similar small stretches of homology can be sufficient to identify functional regulatory elements in the regions around orthologous genes (Sanges et al. 2006).

All resulting elements were then screened for similarity to known protein-coding or RNA sequences. Protein-coding sequences are identified by BLASTX searching each element against all of the known protein sequences in the same genome (Altschul et al. 1997). Arguably, we could have searched each element against a larger protein set from multiple genomes. While this may have identified additional elements with similarity to proteins in other organisms, larger databases increase e-values, thereby reducing overall sensitivity. Because of these trade-offs, multiple genome searches tended to identify a similar total percentage of elements with significant protein similarity, while dramatically increasing search time. RNA searches were conducted by Infernal searching against the Rfam database (Griffiths-Jones et al. 2005). Hits with e-values below 0.05 for BLASTX, or bit scores above 20 for Infernal, were considered significant. Sequences with significant protein similarity were excluded from further analysis. Sequences and genomic locations for the final PCNE set can be found in Supplemental Data S2.

PhastCons scores were calculated for the PCNEs from 17-way vertebrate genomic alignments downloaded from the UCSC Genome Browser, for the mouse February 2006 assembly (Blanchette et al. 2004; Siepel et al. 2005; Kuhn et al. 2007). Mean scores were calculated for each nonredundant PCNE segment, and these scores were averaged to calculate the mean phastCons score for all PCNEs. CIs are Studentized bootstrap-based 95% CIs.

Functional assay for in vivo enhancer activity

We selected 42 PCNEs for functional in vivo testing, selecting PCNEs across a range of gene families while attempting to choose elements that had not been studied in previous reports. On the latter point, we were not entirely successful: 5596-Tr_ECR5_C1 was later found to overlap a previously described CNE (Sox21_19) (Woolfe et al. 2005). PCR primers for each element were designed with PRIDE (Haas et al. 1998), and the sequences can be found in Supplemental Data S3. Tested element sequences have also been deposited in the ORegAnno database as data set OREGDS00016 (Griffith et al. 2008).

Purified PCR products containing the predicted PCNE or control sequence (at a final concentration of 75 ng/μL) were coinjected with the reporter construct without ligation (at a final concentration of 25 ng/μL) and 10% phenol red as tracer, into zebrafish embryos produced from natural matings between the one to four cleavage stages, using an Eppendorf FemtoJet pressure injection system. The reporter construct, consisting of EGFP (Clontech) under the control of a minimal promoter from the human beta hemoglobin gene, was PCR-amplified from a plasmid kindly provided by G. Elgar (Queen Mary, University of London, London) (Woolfe et al. 2005). Each probe was injected in at least two different batches of embryos, each batch consisting of ~500 embryos. For every batch of injected embryos, normal uninjected embryos were raised to control for variation of the survival rate, and a positive control probe was injected in order to assess variations during the injection procedure. Injected embryos were screened for GFP expression on the second day of development (~24–26 hpf), by which time roughly 100 of the original injected embryos would be expected to have survived. Injected embryos were anesthetized in Tricaine and screened by observation under fluorescence illumination using a Zeiss Confocal Microscope (LSM510). GFP-expressing cells were classified to tissue categories similar to Woolfe et al. (2005).

The location and tissue category of each GFP-expressing cell for each embryo was recorded schematically using Adobe Photoshop software (Adobe Systems), by manually drawing color-coded schematized cells in appropriate positions onto an overlay of a camera lucida drawing of a 24-hpf embryo kindly provided by G. Elgar. GFP expression data were collected from all live normal injected embryos. Tissue counts for each tested element can be found in Supplemental Data S4.

Identifying conserved transcription factor binding sites

We used Clover (Frith et al. 2004) to identify over-represented transcription factor binding sites within the Sox14/21 group 1 elements, using all vertebrate transcription binding sites in TRANSFAC (Matys et al. 2003). Three background sequence sets were supplied to Clover: (1) the genomic regions surrounding the *Sox14/21* genes (minus coding sequence and repeats), (2) all PCNEs, and (3) all PCNEs predicted around the *Sox14/21* family.

Classifying genes according to genome evolution characteristics

For the clustering-based analyses presented in Figure 6, all mouse and zebrafish genes contained within our orthology data were classified according to five different genome-evolution characteristics. Genes were declared anciently syntenic if they were contained in a syntenic gene pair that was also found in amphioxus. Retained duplicates created by the 2R WGD events were identified by parsing maximum likelihood gene trees for each gene family, as previously described in Hufton et al. (2008). Genes with bidirectional promoters were defined as protein-coding genes on opposite strands whose TSSs lay within 1 kb of each other, a common definition used, e.g., by Li et al. (2006). Segmental duplications were identified by searching for cases where paralogous vertebrate genes (genes in the same Inparanoid family) lay near each other in the genome (with less than 10 intervening genes). All genes with at least one PCNE were considered linked to PCNEs. Olfactory genes were removed prior to analysis (for more explanation, see the GO Analysis section).

Estimating protein connectivity

Two PPI databases were used to estimate the number of connections for each gene in the mouse genome. The first source, the HPRD, is currently the largest publicly available vertebrate-centered PPI database and contains manually curated human PPIs deduced from experimental evidence in mammalian model systems (Peri et al. 2003; Mathivanan et al. 2006; Mishra et al. 2006). The second source, HomoMINT, maps interaction data from a variety of organisms onto orthologous human genes using automated methods (Persico et al. 2005). HomoMINT includes data from nonvertebrates such as yeast and drosophila, and while data from these simpler model organisms are generally much more complete, inclusion of these data created obvious biases in our protein connectivity estimates. For example, inclusion of the yeast data tends to make the most highly conserved cellular machinery seem overconnected (such as ribosome genes), while developmental genes appeared underconnected because they generally lacked yeast orthologs. As such we only used HomoMINT interactions

that had been identified in a vertebrate organism. For both databases, we then transferred predicted human PPIs onto the orthologous mouse genes using Inparanoid.

Statistical analysis of gene set associations and synteny trends

The distributions of PCNE number and protein connectivity are extremely right-skewed; the majority of genes have zero or one values. These right-skewed distributions are somewhat difficult to summarize. The mean is not a measure of central tendency in these cases, but instead tends to be proportional to the thickness of the distribution tail. Nonetheless, the median, which is resistant to tail values, can be exactly the same across distributions that have significant and biologically interesting differences (see Fig. 5, where most categories have a median of zero PCNEs). We have opted to display both the mean and the median in all cases, allowing readers to gauge the skewedness of the distribution by the difference between the two. Statistical inference is conducted with a permutation test on the mean (perm.test in the R package exactRankTests) (Röhmel 1996). Moreover, each mean is shown with a bootstrap-based 95% CI (Studentized CI), displaying the extent to which outliers may cause the means to fluctuate. These methods make no assumptions about the underlying distributions and have been shown to be appropriate for skewed populations (Ludbrook 1994; Carpenter and Bithell 2000).

The significance of the difference between the extragenic and intragenic synteny-PCNE trends was tested by permutation. For each permutation, the extragenic/intragenic labels on the PCNEs were randomly shuffled, trend lines were produced as in Figure 4, and the slope of the trend between PCNEs minimums 0–5 were calculated (the portion of the trend that is largely linear). For the real data, the slope difference (extragenic minus intragenic) was 0.041 for mouse, 0.040 for zebrafish, and 0.005 for fugu. In 1000 permutations, the mouse and zebrafish slope difference was never met or exceeded by the slope differences of their permuted data ($P < 0.001$), and the fugu slope difference was met or exceeded only 12 times with the permuted data ($P = 0.012$).

Gene Ontology analysis

Within the mouse and zebrafish genomes we identified functional gene categories that were associated with five genome-evolution characteristics: PCNEs, retained WGD duplicate, ancient synteny, bidirectional promoter, and segmental duplicate. GO annotation was provided by GOA (Barrell et al. 2009). For each genome-evolution characteristic, enriched or depleted terms were identified using the parent–child intersection method implemented by the Ontologizer software package (Grossmann et al. 2007; Bauer et al. 2008). This method accounts for the hierarchical nature of the GO when calculating enrichment *P*-values, thereby returning shorter lists of enriched terms that better capture the true points of functional enrichment. During preliminary analyses, we noticed that many of the most significantly enriched and depleted terms were related to olfactory receptor genes, of which there are more than a thousand in mice and all of which share a common and exceptional evolutionary history: extensive segmental duplication while retaining noncoding sequences and completely avoiding bidirectional promoters. While this pattern is biologically relevant, the strength of the signal created by these genes tends to drown out more subtle functional associations. Therefore, olfactory receptor genes were removed from the gene sets prior to analysis. Complete lists of significantly enriched and depleted GO terms can be found in Supplemental Data S5.

We then clustered the most enriched and depleted GO terms according to their pattern of genome evolution using a hierarchical method adopted from the microarray field (Eisen et al. 1998). For each organism, we selected the top 50 enriched or depleted molecular function and biological process terms from the olfactory-receptor free results (measured by *P*-value). For each GO term and genome-evolution characteristic, enrichment/depletion is quantified as a log-ratio, *E*, where

$$E = log_2(X/G) - log_2(T/N),$$

and X is the number of genes with the GO term and the genome-evolution feature; T is the total number of genes with the GO term; G is the total number of genes with the genome-evolution feature; and N is the total number of genes.

Positive values of E indicate an association between a GO category and genome-evolution characteristic, while negative values indicate an anti-correlation between the two. Where X equals zero, this formula produces a log-ratio of $-\infty$. In these cases, we set E equal to the smallest non-infinite E-value observed for the genome-evolution characteristic. This truncation was used in three cases for the mouse data, and two for the zebrafish data. After calculating all E-values, we have a matrix that is numerically similar to expression microarray data, where GO categories are analogous to genes, and the genome-evolution characteristics are analogous to experimental conditions. This matrix of E-values was then hierarchically clustered by Cluster3 (Eisen et al. 1998; De Hoon et al. 2004), using uncentered Pearson correlation, complete linkage, and normalization across both rows and columns (Fig. 6). Cluster visualizations were generated by Java Treeview (Saldanha 2004).

Acknowledgments

We thank G. Elgar and D. Goode for providing us with transgenic zebrafish vectors, protocols, and scoring templates. We also thank S. Hass, S. Kielbasa, S. O'Keeffe, and B. Cusack for helpful advice and code related to the handling and processing of genome annotation data. This work was supported by the Max-Planck Society (Max-Planck Gesellschaft zur Forderung der Wissenschaften e.v.).

References

Abdelhaleem M, Maltais L, Wain H. 2003. The human DDX and DHX gene families of putative RNA helicases. *Genomics* **81:** 618–622.

Ahituv N, Prabhakar S, Poulin F, Rubin EM, Couronne O. 2005. Mapping *cis*-regulatory domains in the human genome using multi-species conservation of synteny. *Hum Mol Genet* **14:** 3057–3063.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. 2009. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* **37:** D396–D403.

Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24:** 1650–1651.

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

Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol* **234:** 275–288.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: Biological implications. *Trends Genet* **21:** 219–226.

Blair JE, Hedges SB. 2005. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* **22:** 2275–2284.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16:** 1679–1691.

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

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7:** R43. doi: 10.1186/gb-2006-7-5-r43.
Brudno M, Chapman M, Göttgens B, Batzoglou S, Morgenstern B. 2003. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4:** 66. doi: 10.1186/1471-2105-4-66.
Brunet FG, Crollius HR, Paris M, Aury J-M, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23:** 1808–1816.
Carpenter J, Bithell J. 2000. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Stat Med* **19:** 1141–1164.
Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.
Cusack BP, Wolfe KH. 2007. When gene marriages don't work out: Divorce by subfunctionalization. *Trends Genet* **23:** 270–272.
De Hoon MJL, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20:** 1453–1454.
Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38:** 223–227.
Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol Biol Evol* **23:** 469–478.
Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* **95:** 14863–14868.
The ENCODE Project Consortium 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res* **17:** 1898–1908.
Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al. 2008. Ensembl 2008. *Nucleic Acids Res* **36:** D707–D714.
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn* **4:** 25–40.
Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* **16:** 805–814.
Frickey T, Lupas AN. 2004. PhyloGenie: Automated phylome generation and analysis. *Nucleic Acids Res* **32:** 5231–5238.
Frith MC, Fu Y, Yu L, Chen J-F, Hansen U, Weng Z. 2004. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32:** 1372–1381.
Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, et al. 2008. ORegAnno: An open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* **36:** D107–D113.
Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33:** D121–D124.
Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **23:** 3024–3031.
Haas S, Vingron M, Poustka A, Wiemann S. 1998. Primer design for large scale sequencing. *Nucleic Acids Res* **26:** 3006–3012.
He X, Zhang J. 2005. Gene complexity and gene duplicability. *Curr Biol* **15:** 1016–1021.
Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, et al. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18:** 1100–1111.
Hsieh YJ, Wang Z, Kovelman R, Roeder RG. 1999. Cloning and characterization of two evolutionarily conserved subunits (TFIIIC102 and TFIIIC63) of human TFIIIC and their involvement in functional interactions with TFIIIB and RNA polymerase III. *Mol Cell Biol* **19:** 4944–4952.
Hufton AL, Groth D, Vingron M, Lehrach H, Poustka AJ, Panopoulou G. 2008. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res* **18:** 1582–1591.
Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol* **10:** 1360–1369.
Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14:** 1870–1879.
Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17:** 545–555.
Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, et al. 2007. The UCSC genome browser database: Update 2007. *Nucleic Acids Res* **35:** D668–D673.
Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet* **21:** 602–607.
Li Y-Y, Yu H, Guo Z-M, Guo T-Q, Tu K, Li Y-X. 2006. Systematic analysis of head-to-head gene organization: Evolutionary conservation and potential biological relevance. *PLoS Comput Biol* **2:** e74. doi: 10.1371/journal.pcbi.0020074.
Liang H, Plazonic KR, Chen J, Li W-H, Fernández A. 2008. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* **4:** e11. doi: 10.1371/journal.pgen.0040011.
Ludbrook J. 1994. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol* **21:** 673–686.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.
Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159:** 1789–1804.
Mackenzie A, Miller KA, Collinson JM. 2004. Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks? *Bioessays* **26:** 1217–1224.
Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci* **102:** 5454–5459.
Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* **13:** 2507–2518.
Mathivanan S, Periaswamy B, Gandhi TKB, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A. 2006. An evaluation of human protein–protein interaction data in the public domain. *BMC Bioinformatics* **7:** S19. doi: 10.1186/1471-2105-7-S5-S19.
Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374–378.
McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G. 2006. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res* **16:** 451–465.
Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, et al. 2006. Human protein reference database—2006 update. *Nucleic Acids Res* **34:** D411–D414.
Müller F, Williams DW, Kobolák J, Gauvry L, Goldspink G, Orbán L, Maclean N. 1997. Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol Reprod Dev* **47:** 404–412.
Nadeau JH, Sankoff D. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147:** 1259–1266.
Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B, Becker TS. 2008. Systematic human/zebrafish comparative identification of *cis*-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol* **327:** 526–540.
Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.
Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* **13:** 1056–1066.
Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194–197.
Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363–2371.
Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. 2005. HomoMINT: An inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* **6:** S21. doi: 10.1186/1471-2105-6-S4-S21.
Ponce R, Hartl DL. 2006. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. *Gene* **376:** 174–183.

Postlethwait J, Amores A, Cresko W, Singer A, Yan Y-L. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet* **20:** 481–490.
Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453:** 1064–1071.
Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041–1052.
Röhmel J. 1996. Precision intervals for estimates of the difference in success rates for binary random variables based on the permutation principle. *Biometrical Journal* **38:** 977–993.
Saldanha AJ. 2004. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20:** 3246–3248.
Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5:** 99.
Sanges R, Kalmar E, Claudiani P, D'Amato M, Muller F, Stupka E. 2006. Shuffling of *cis*-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol* **7:** R56. doi: 10.1186/gb-2006-7-7-r56.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502–504.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human–mouse alignments with BLASTZ. *Genome Res* **13:** 103–107.
Sémon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* **23:** 1715–1723.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25:** 402–408.
Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci* **101:** 1638–1643.
Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *BioEssays* **24:** 175–184.
Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* **26:** 487–490.
Woolfe A, Elgar G. 2007. Comparative genomics using fugu reveals insights into regulatory subfunctionalization. *Genome Biol* **8:** R53. doi: 10.1186/gb-2007-8-4-r53.
Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3:** e7. doi: 10.1371/journal.pbio.0030007.

Received March 2, 2009; accepted in revised form July 29, 2009.