Supplemental Materials for

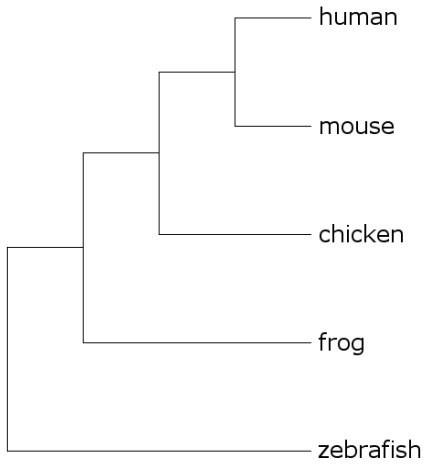
Genome-Wide Identification of Conserved Regulatory Function in Diverged Sequences

Leila Taher[1], David M. McGaughey[2], Samantha Maragh[2], Ivy Aneas[4], Seneca L. Bessling[2], Webb Miller[3], Marcelo A. Nobrega[4], Andrew S. McCallion[2,#], and Ivan Ovcharenko[1,#]

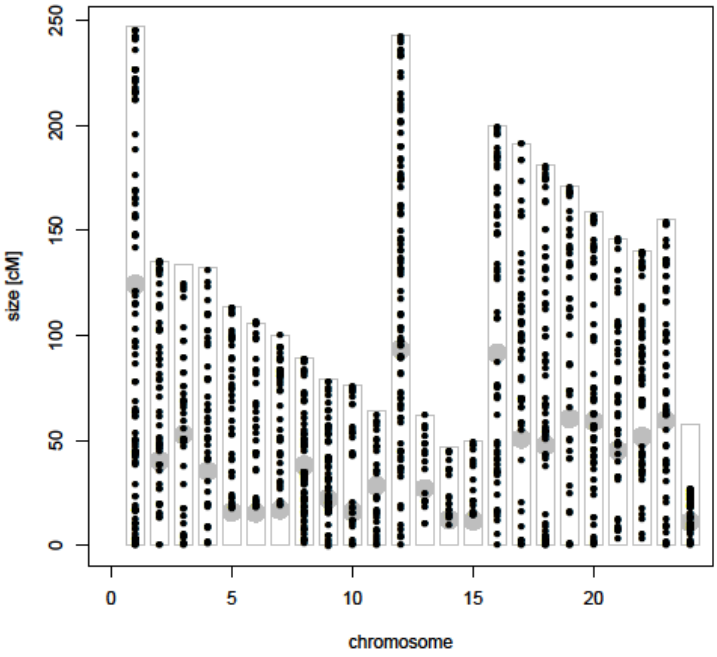
1 Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, United States of America, **2** McKusick–Nathans Institute of Genetic Medicine, Department of Molecular and Comparative Pathobiology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **3** Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, Pennsylvania 16802, United States of America, **4** Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, United States of America.

To whom correspondence should be addressed: Andrew S. McCallion (andy@jhmi.edu) and Ivan Ovcharenko (ovcharei@ncbi.nlm.nih.gov).

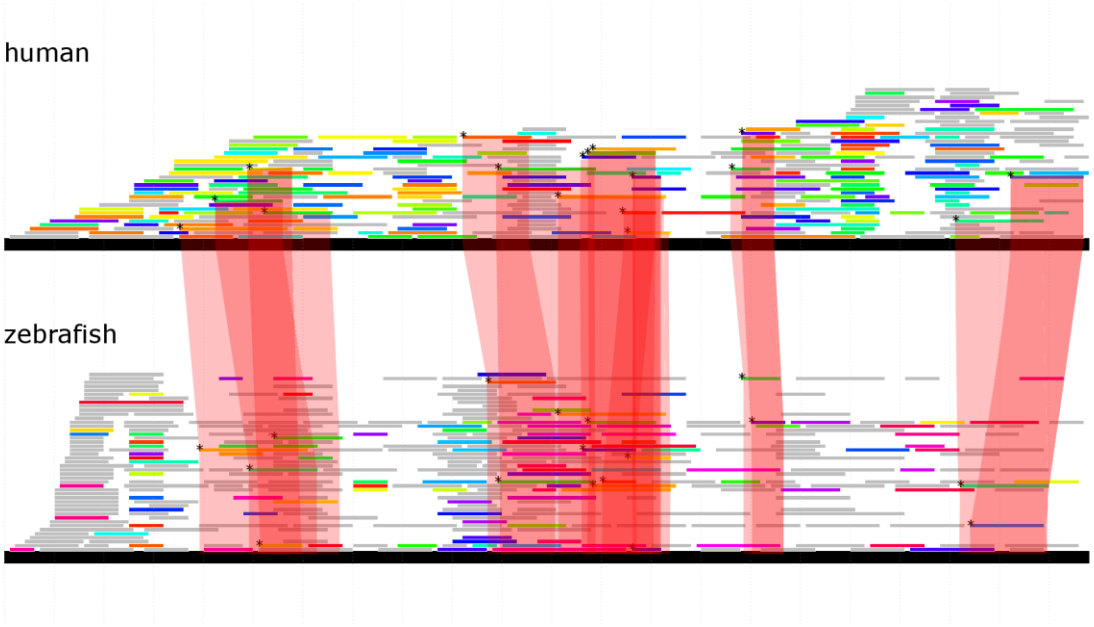
Supplemental Figures



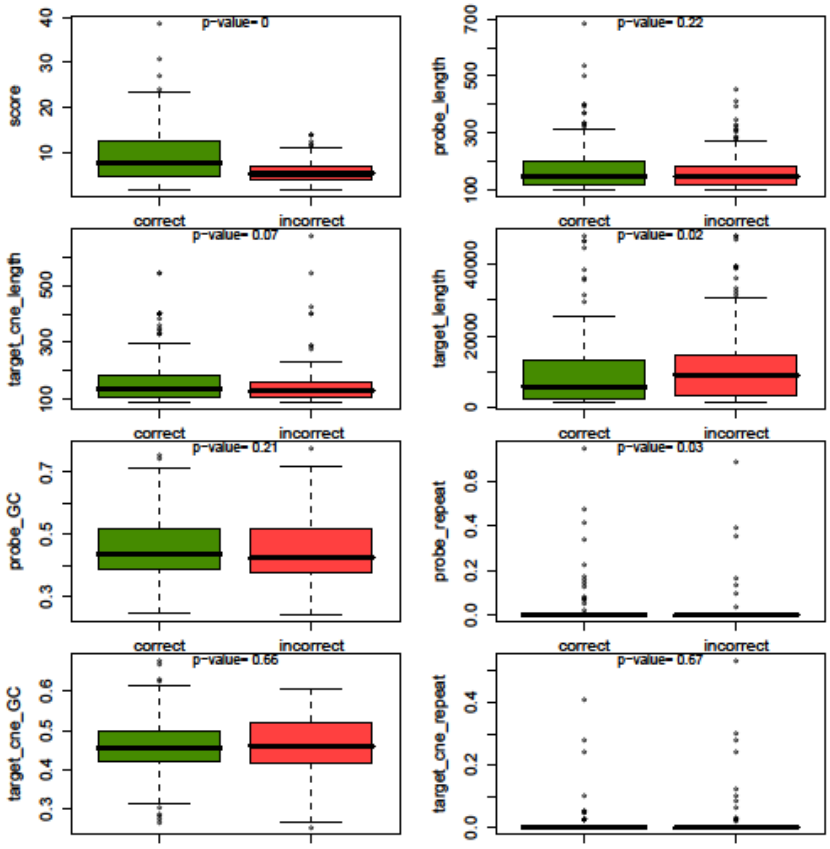
Supplemental Figure 1. Phylogenetic tree of the species analyzed in this study, adapted from from (Miller et al. 2007) and generated with the tool phyloGif (Kuhn et al. 2007).



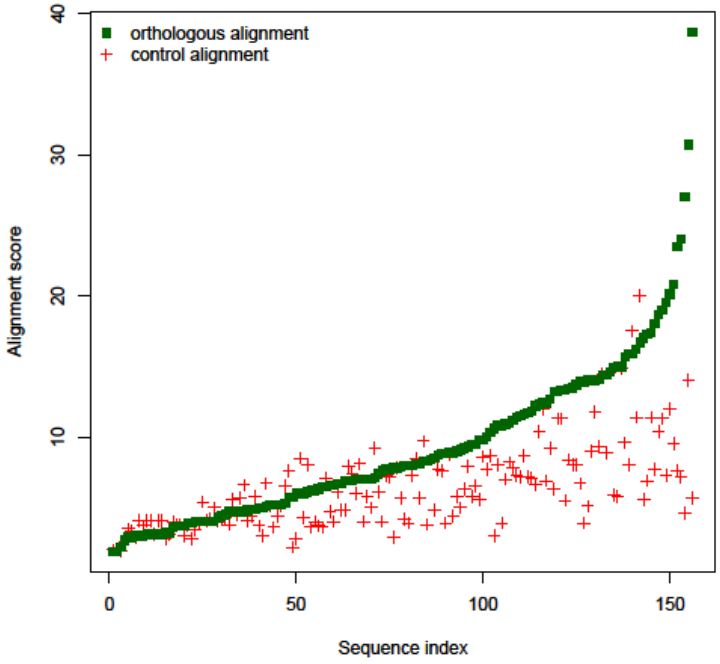
Supplemental Figure 2. Non-uniform distribution of human/zebrafish tunneled elements in the human genome.



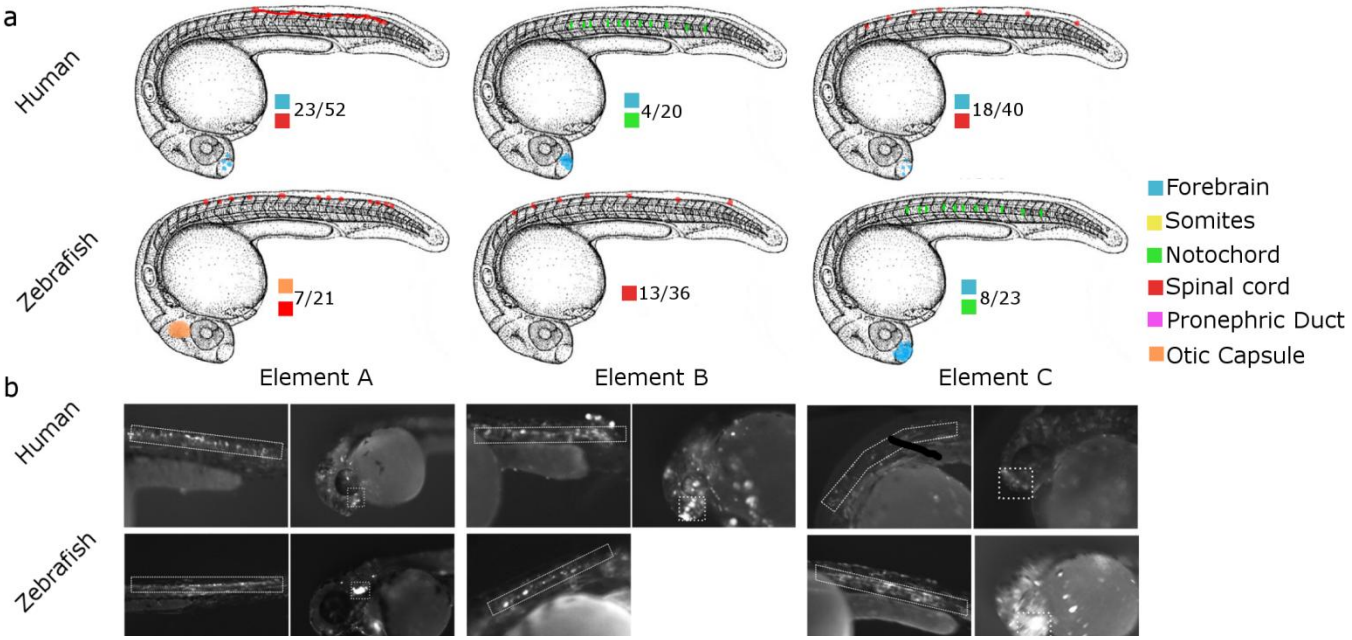
Supplemental Figure 3. Example of a TFBS-based alignment between the human and zebrafish counterparts of a tunneled element (TE). The sequences hg18 chr10:102367867-102368086 and danRer5 chr13:29087577-29087796 are both 220-bp long, respectively. TFBSs were mapped on these sequences using TRANSFAC (Matys et al. 2003), with an average density of 1.3 sites/bp. Each colored box represents a different TFBS. Gray boxes are TFBSs that have been excluded from the alignment due to their absence in the frog sequence that was used for *tunneling* between the human and the zebrafish sequence. The sites in the optimal TFBS-based alignment are marked with an asterisk; aligned TFBSs are connected with red segments.



Supplemental Figure 4. Comparison of different properties between sequences that aligned correctly and sequences fail to align to their functional orthologs.

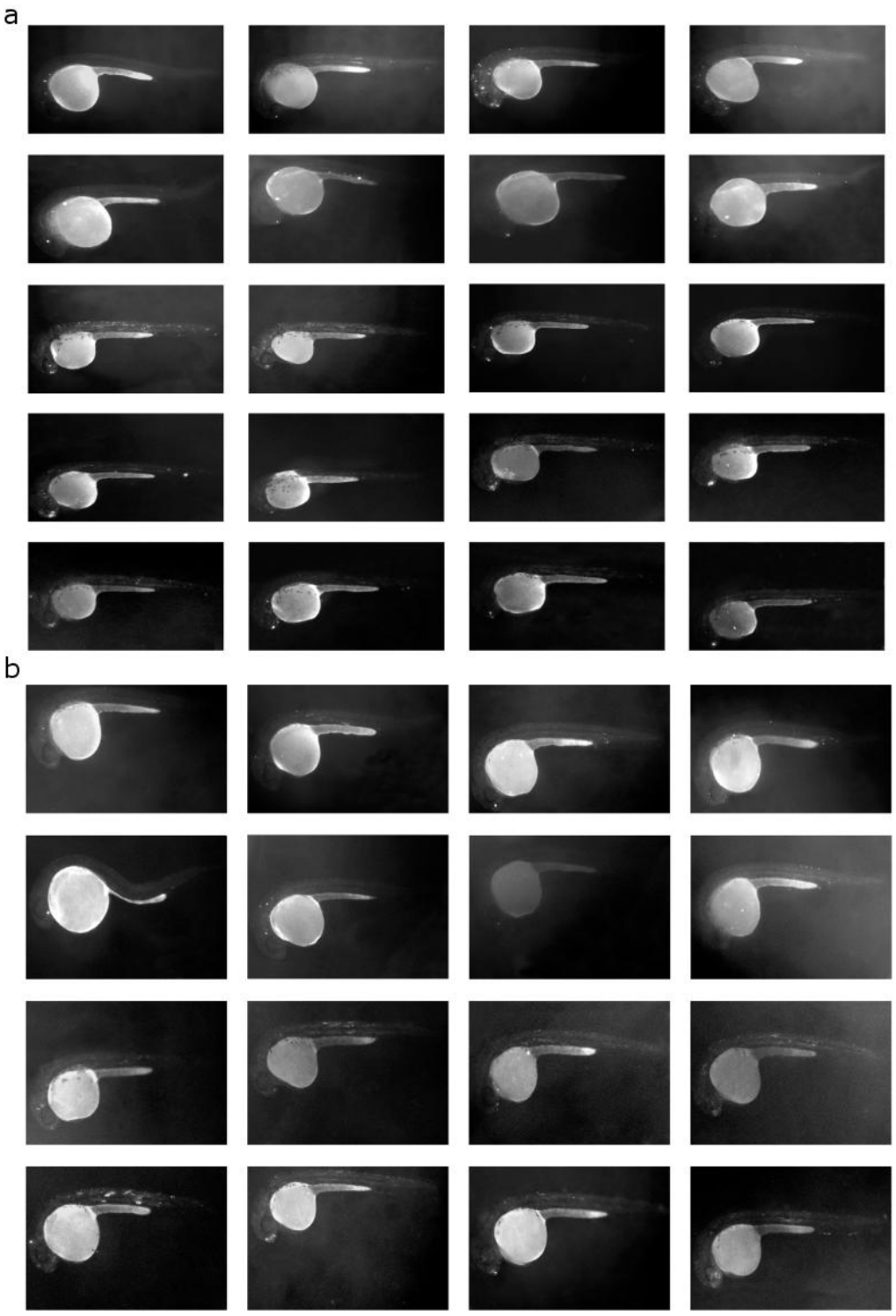


Supplemental Figure 5. Scores for the TFBS-based alignment of the human counterparts of the human/zebrafish tunneled elements to their orthologous zebrafish loci (green dots), compared to the scores of the alignment to their flanking control loci (red dots). Only the sequences that could be correctly aligned are included in the figure.

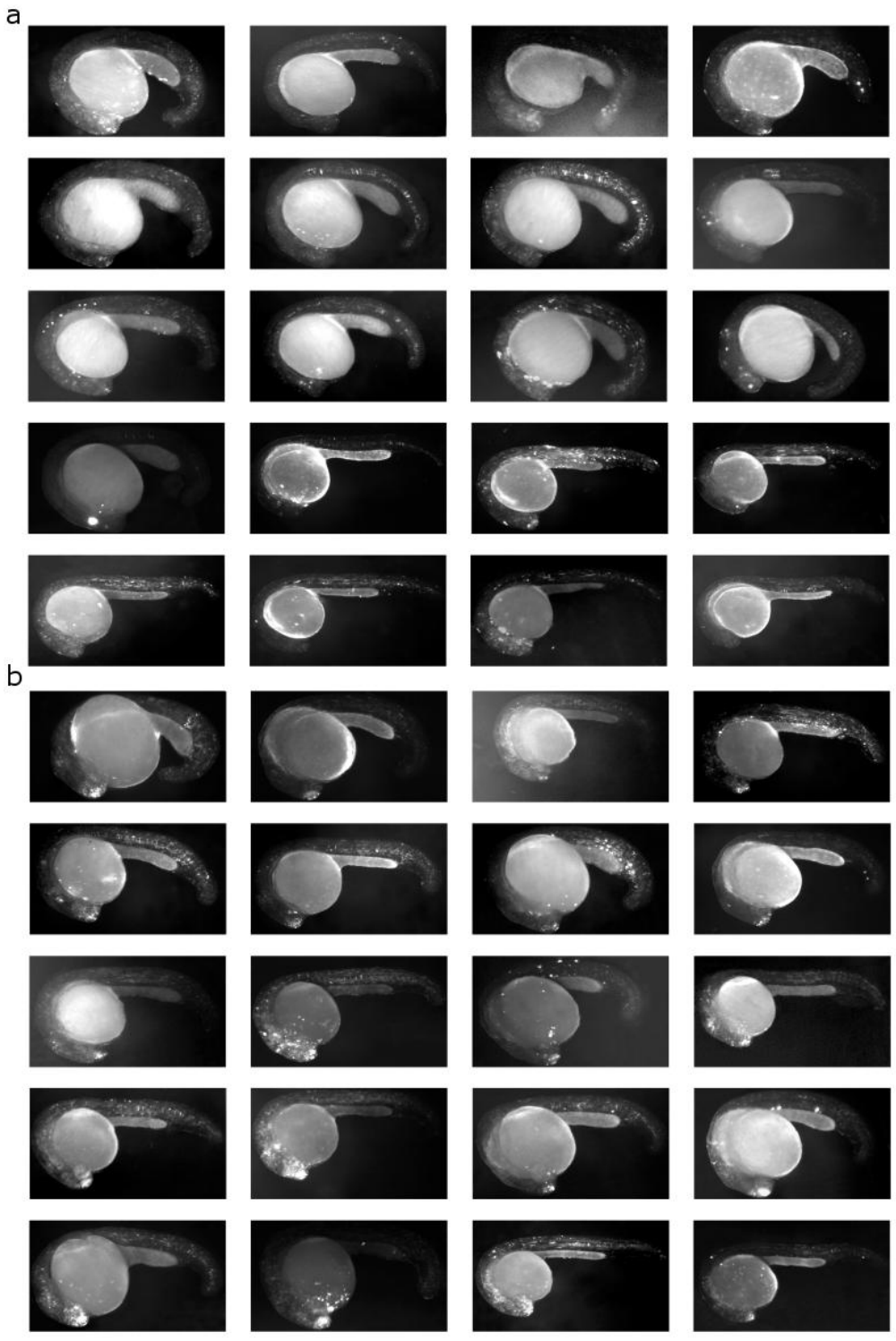


Supplemental Figure 6. Putative human and zebrafish enhancer pairs direct similar tissue specific expression (covert elements A, B, C, Table 2). (a) Composite overviews of in vivo GFP expression data from 16-20 individual zebrafish embryos per construct. The keys for the marked expression are provided next to each image. The number of fish in the set with that specific

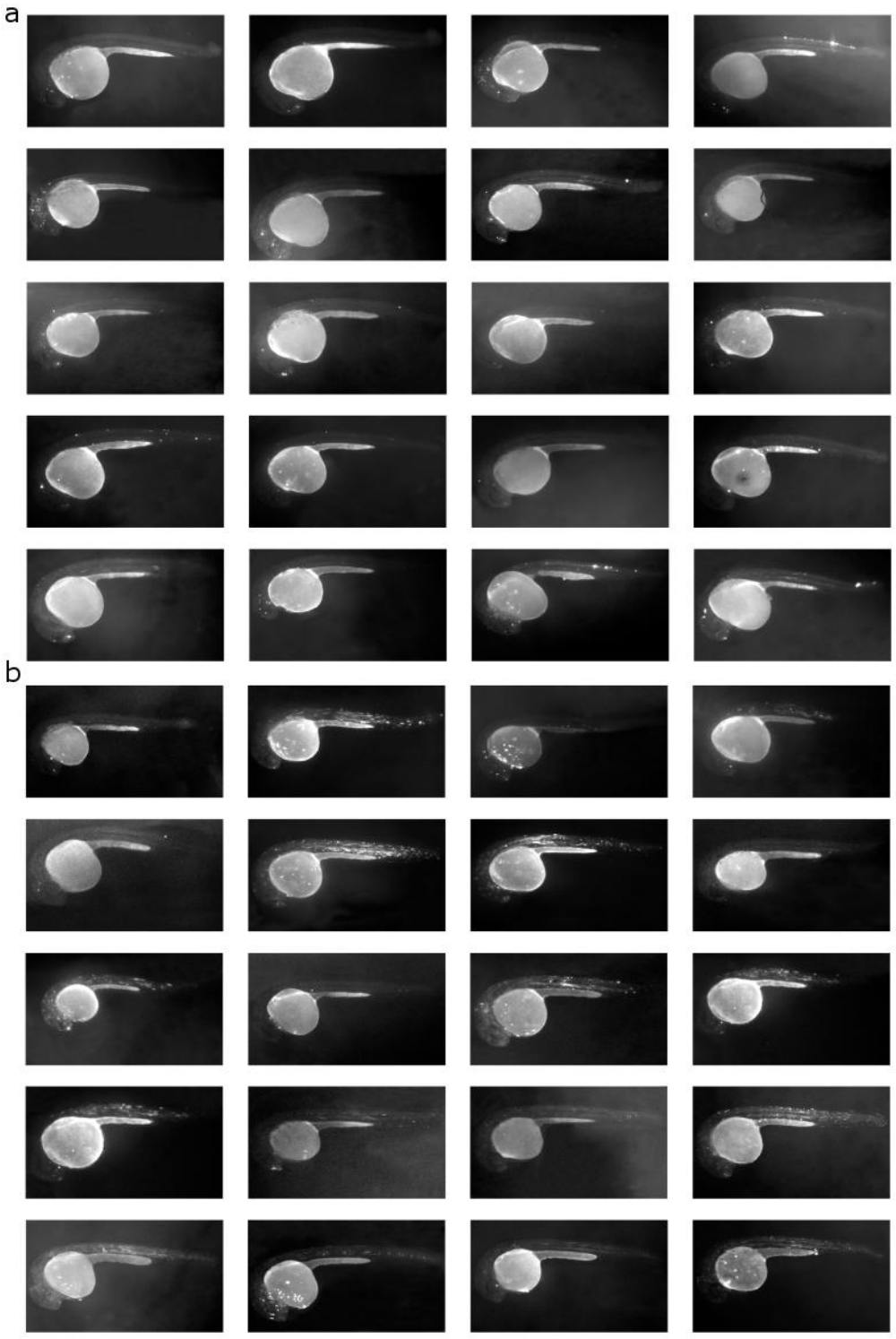
expression is given in the parentheses. (b) One representative GFP live image from each enhancer set is displayed. All zebrafish are 24 hpf oriented with anterior to the left and dorsal to the top. The dotted box demarcates the forebrain. The stacked structures of the notochord are between the dotted lines. Arrows refer to the somites. The solid line box contains the spinal cord. The pronephric duct-consistent expression is marked by the solid line ovals.



Supplemental Figure 7. GFP expression of all assayed zebrafish corresponding to the human (a) and zebrafish (b) counterparts of covert element D (Table 2) in Fig. 4. 24 hpf, anterior left, and dorsal up.



Supplemental Figure 8. GFP expression of all assayed zebrafish corresponding to the human (a) and zebrafish (b) counterparts of covert element E (Table 2) in Fig. 4. 24 hpf, anterior left, and dorsal up.



Supplemental Figure 9. GFP expression of all assayed zebrafish corresponding to the human (a) and zebrafish (b) counterparts of covert element G (Table 2) in Fig. 4. 24 hpf, anterior left, and dorsal up.

Supplemental Results

Additional properties of human/zebrafish tunneled elements (TEs)

Consistent with the assertion that TE conservation patterns differ from those of CNEs, we observed that an average of 71% of the bases are identical between TE human (TE_Hs) and frog counterparts (TE_Fs), whereas only 25% are identical between human and zebrafish (TE_Zs) counterparts (as compared to 73% and 72% in the case of human/frog and human/zebrafish CNEs, respectively). Similarly, 60% of the bases are in average identical between the TE_Fs and TE_Zs whereas 71% are expected for comparisons of frog/zebrafish CNEs. This is consistent with regions that are evolving at an accelerated rate, but show slightly slower divergence rates in the frog orthologs as compared to the human and the zebrafish.

Evidently, Frog is not the only species in which we observe local differences in the rate of evolution, but simply an example to illustrate our approach. Analogously, we determined 4,300 slowly evolving zebrafish segments and 2,200 human sequences that can be used to infer orthology between pairs of frog and human and zebrafish and frog sequences, respectively.

TEs feature evidence of a functional core

Human/zebrafish TEs are only slightly shorter than human/zebrafish CNEs (182 bp versus 183 bp on average), with nucleotides located near the more centrally located nucleotides of a TEs being generally more conserved (75% of identity between human and frog) than the outermost flanks (63%), thus arguing for the presence of a functional core within these two classes of elements (Ovcharenko et al. 2004; Prabhakar et al. 2006).

Many covert orthologs predicted in genome-wide scans are likely to be functional

The ~300 human/zebrafish putative orthologs are widely distributed in 236 loci (UCSC Known Genes and RefSeq (Hsu et al. 2006; Pruitt et al. 2007)). The human counterpart of 3% of these elements

overlap a UTR of a known human protein-coding gene, while 71% are located in introns of known genes and the rest are intergenic. Fifty-eight percent of the host genes have a zebrafish ortholog (HomoloGene release 64 (Sayers et al. 2009)). A few predicted intergenic elements (10) lie in "gene deserts", regions containing no protein-coding sequences that extend more than 1 Mb. In addition, there are 42 that are more than 100 kb away from any known gene, and 6 that are more than 500 kb away. The set of 48 annotated genes that flank intergenic elements is significantly enriched for transcription factor and protein-binding genes (p-values < 0.02 and 0.03 respectively, binomial test), suggesting that many of the associated elements may be distal enhancers of these genes. Elements that lie in introns are also often associated with transcription factor and developmental genes (p-values 0.0006 and 0.009, respectively, binomial test).

Although the minimal length required for the putative covert elements was 100 bp, many elements are considerably longer (average size = 288 bp). In particular, the longest elements (1,130, 991, and 971 bp) lie in introns of genes TTYH3, CADPS2 and PACS2. For instance, CADPS2 is a member of the calcium-dependent activator of secretion protein family and is thought to be associated with autism susceptibility (Sadakata et al. 2007).

Predicted covert elements exhibit little natural variation in the human population. Only 107 out of 90,494 bp bases examined in the human sequence of the putative elements are at validated single nucleotide polymorphisms (SNPs) in NCBI's dbSNP database (Sherry et al. 2001). For this much DNA, we would have expected 146 validated sites, so validated SNPs are under-represented (0.73 fold enrichment compared to flanking DNA). These 90,494 bp also exhibit very few differences with the chimp genome, showing an average sequence identity of 0.974 (expected sequence identity would be 0.970). This low level of variation within the human population and in comparison with the chimp suggests that these elements are currently changing at a rate that is slower than the genome average.

Likewise, predicted human/frog and human/chicken covert elements are associated with particular functions. Putative human/frog covert elements are enriched in the neighborhood of genes with functions including organ development and transcription factor activity (all p-values <0.05, binomial test). Similarly to what we observed for human/zebrafish covert elements, 75% of the predicted human/frog covert elements overlap with peaks of p300 occupancy in forebrain, 73% overlap with p300 peaks in midbrain, and 66% overlap with p300 peaks in limb. Overall, 88% of our predicted sequences overlap with p300 peaks. Twenty-percent of the elements have H3K4me1 signatures. Analogous conclusions are valid for human/chicken elements.

Number of tunneled elements (TEs)

	hg18/xenTro2/danRer5	hg18/galGal3/xenTro2	hg18/mm9/galGal3
Raw	1,552	3,603	6,414
50kb constraint	308	997	988

Supplemental Table 1. Number of elements identified through the conservation tunneling approach. The row "Raw" indicates the number of elements we detected genome-wide, whereas "50 kb constraint" shows the number of elements encompassed by two CNEs that are at most 50 kb away from each other (as described in the main text).

Number of putative covert elements uncovered in genome-wide scans

	hg18/xenTro2/danRer5	hg18/galGal3/xenTro2	hg18/mm9/galGal3
50kb constraint	2,860	8,039	290,139
Pairs of putative covert orthologs	311	789*	3,772*

Supplemental Table 2. Summary of results for the genome-wide scans. The row "50 kb constraint" indicates the numbers of human/frog, human/chicken and human/mouse CNEs with no sequence similarity (either directly or through our conservation tunneling approach) with zebrafish, frog and chicken, respectively, for which we can define syntenic loci not exceeding 50 kb. * For practical reasons and without loss of generalization, we only computed 2,843 human/frog and 4,307 human/chicken alignments, and estimated the total number of orthologous sequence pairs based on results we obtained.

References

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC Known Genes. *Bioinformatics* **22**(9): 1036-1046.

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**(Database issue): D668-673.

Matys V, Fricke E, Geffers R, Gossling 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**(1): 374-378.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**(12): 1797-1808.

Ovcharenko I, Stubbs L, Loots GG. 2004. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* **84**(5): 890-895.

Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* **16**(7): 855-863.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**(Database issue): D61-65.

Sadakata T, Washida M, Iwayama Y, Shoji S, Sato Y, Ohkura T, Katoh-Semba R, Nakajima M, Sekine Y, Tanaka M et al. 2007. Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients. *J Clin Invest* **117**(4): 931-943.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**(Database issue): D5-15.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucl Acids Res* **29**(1): 308-311.