Genome-Level Analysis of Selective Constraint without Apparent Sequence Conservation

Olga A. Vakhrusheva[1,2,*], Georgii A. Bazykin[1,2], and Alexey S. Kondrashov[1,3]

[1]Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

[2]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

[3]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan

*Corresponding author: E-mail: olga.a.vakhrusheva@gmail.com.

Accepted: February 7, 2013

Abstract

Conservation of function can be accompanied by obvious similarity of homologous sequences which may persist for billions of years (Iyer LM, Leipe DD, Koonin EV, Aravind L. 2004. Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol. 146:11–31.). However, presumably homologous segments of noncoding DNA can also retain their ancestral function even after their sequences diverge beyond recognition (Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279.). To investigate this phenomenon at the genomic scale, we studied homologous introns in a quartet of insect species, and in a quartet of vertebrate species. Each quartet consisted of two pairs of moderately distant genomes, with a much larger evolutionary distance between the pairs. In both quartets, we found that introns that carry a regulatory segment or a conserved segment in the first pair tend to carry a conserved segment in the second pair, even though no similarity of these segments could be detected between the two pairs. Furthermore, introns from one pair that are preserved in the other pair tend to carry a conserved segment within the first pair, and be longer in the first pair, compared with the introns that were lost between pairs, even though no similarity between pairs could be detected in such preserved introns. These results indicate that selective constraint, presumably caused by conservation of the ancestral function, often persists even after the homologous DNA segments become unalignable.

Key words: conserved noncoding elements, turnover of regulatory elements, negative selection, evolution of regulatory sequences.

Introduction

Conservation of function can be accompanied by obvious similarity of homologous sequences, which may persist for billions of years. Many bacterial proteins possess more than 50% similarity to their eukaryotic orthologs. Moreover, analysis of the reconstructed genome of the LUCA revealed a history of pre-LUCA gene duplications; such duplications produced paralogous proteins whose similarity is still statistically significant (Iyer et al. 2004). Even within noncoding segments of genomes, there are ultraconservative segments that retain strong similarity within, for example, all vertebrates (Dermitzakis et al. 2003; Lowe et al. 2011).

However, conservation of the primary sequence, resulting in a meaningful sequence alignment, is not a sine qua non for conservation of other properties of the molecule. Proteins with unalignable amino acid sequences can have very similar 3D structures (Murzin and Bateman 1997). Single-stranded RNAs with dissimilar sequences can fold into identical secondary structures (Schuster et al. 1994) (e.g., AAAAAGGGTTTTT and GGGGGTTTCCCCC). At the level of functional noncoding DNA sequences, there are a number of described cases when sequences from different organisms perform similar functions, and are likely to be homologous, despite the lack of any meaningful alignment (Taher et al. 2011). For example, an enhancer of a murine gene can drive, in a transgenic assay, normal expression of its zebrafish ortholog (Fisher et al. 2006), although the murine and the zebrafish enhancers are not alignable. Another study in zebrafish has shown that the up-to-date alignment techniques are unable to detect many of the functional genomic regions (McGaughey et al. 2008). The noncoding elements that regulate homologous genes in nematodes and vertebrates are themselves alignable within both these lineages, but not between them, implying regulatory "rewiring" (Vavouri et al. 2007). However, to our knowledge, this phenomenon has not been investigated genome-wide. Here, we show that selective constraint, presumably caused by conservation of the ancestral function, often persists even after the homologous genome compartments diverge beyond alignability.

Materials and Methods

Data

We studied two quartets of species together with the two corresponding outgroup species (fig. 1). In each quartet, the number of synonymous substitutions per site Ks between the species of the two pairs is larger than 1, and therefore cannot be measured with precision. The approximate values given in figure 1 were obtained as follows. In dipterans, to roughly estimate the values of Ks between the two pairs, we scaled the protein identity-based trees by the known Ks values obtained for the more closely related species within pair 1. The first, dipteran, quartet consisted of two *Drosophila* species, *D. melanogaster* and *D. mojavensis* (pair 1: Ks_1 ~2.37; Heger and Ponting 2007), and two mosquito species, *Culex quinquefasciatus* and *Aedes aegypti* (pair 2: Ks_2 ~2.6); the estimated *Drosophila*–mosquito Ks_3 is approximately 6.5. Ks between *C. quinquefasciatus* and *Aed. aegypti,* and between *Drosophila* and mosquitoes, were estimated through calibrating *Drosophila*–mosquito tree of concatenated sequences of motor proteins (Odronitz et al. 2009) with known Ks for *D. melanogaster*–*D. mojavensis*. The second, vertebrate, quartet consisted of two mammalian species *Homo sapiens* and *Mus musculus* (pair 1: $Ks_1 \sim 0.43$; Jaillon et al. 2004), and two fish species *Tetraodon nigroviridis* and *Takifugu rubripes* (pair 2: $Ks_2 \sim 0.35$; Jaillon et al. 2004). Ks_3 is approximately 1.5 (Jaillon et al. 2004).

Lists of orthologous proteins for each pairwise combination of species within a quartet were downloaded from INPARANOID (Ostlund et al. 2010) database (http://inparanoid.sbc.su.se/cgi-bin/index.cgi, last accessed February 27, 2013). For each quartet, we selected unambiguous seed-ortholog pairs for each of the 6 (10 for analyses requiring an outgroup) unordered pairs chosen from the 4 (5) species. We further considered only those orthologs that comprised a four-species (five-species) clique. This procedure resulted in 5,189/3,565 and 8,179/2,522 unambiguous orthologs for the dipteran and vertebrate quartets, respectively; after exclusion of the coding sequences with internal stop codons in any of the species, the corresponding numbers were 5,183/3,541 and 8,159/2,518 for dipteran and vertebrate quartets, respectively. For sequence analysis, we used genome assembly versions identical to those given in INPARANOID to avoid orthologs misidentification due to differences in annotations between releases. Specifically, for *H. sapiens*, *M. musculus*, *C. intestinalis*, *T. nigroviridis*, *Tak. rubripes,* and *Aed. aegypti,* we used NCBI 36 (Lander et al. 2001; Wheeler et al. 2008), NCBI m37 (Waterston et al. 2002), JGI 2 (Dehal et al. 2002), TETRAODON 8.0 (Jaillon et al. 2004), FUGU 4.0 (Aparicio et al. 2002), and AaegL1 (Nene et al. 2007) assemblies, respectively, all corresponding to ENSEMBL release 52. For *D. melanogaster* and *D. mojavensis,* we used r5.13 (Adams et al. 2000) and r1.3 (Clark et al. 2007) assemblies, corresponding to ENSEMBL releases 58 and 63, respectively. For *C. quinquefasciatus*, we used CpipJ1.2 (Arensburger et al. 2010) assembly. For *A. mellifera* (Honeybee Genome Sequencing Consortium 2006), we used NCBI build 4.1. All sequence and annotation data except for data on *A. mellifera* and *C. quinquefasciatus* was fetched from ENSEMBL (Kersey et al. 2009) through usage of ENSEMBL PERL API for perl scripts. Genome sequence and annotation for *A. mellifera* and *C. quinquefasciatus* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov, last accessed February 27, 2013) (Sayers

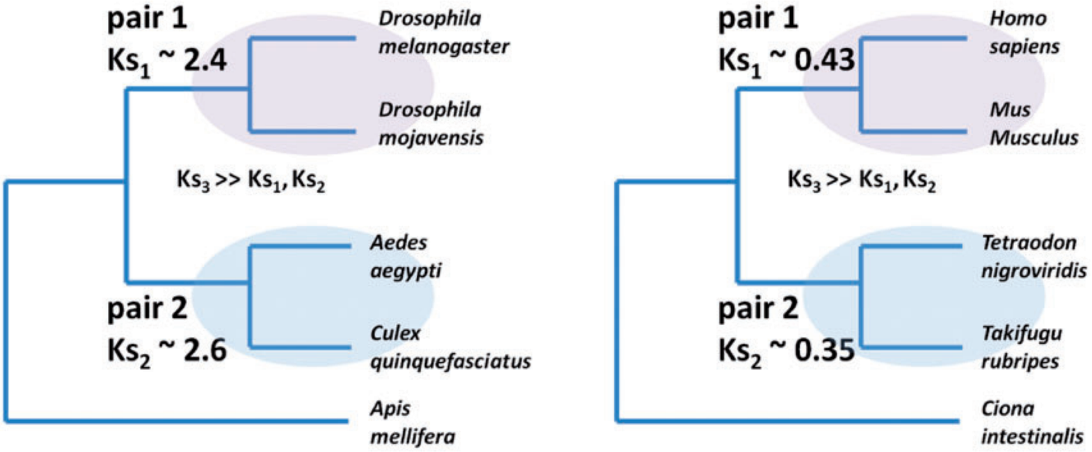

FIG. 1.—Two quartets of species used in the analysis, of 1) dipterans and 2) vertebrates, together with the corresponding outgroup species. Evolutionary distances within each pair of species, characterized by the estimated per site number of synonymous substitutions Ks, are presented.

et al. 2012) and VectorBase (http://cquinquefasciatus.vectorbase.org/, last accessed February 27, 2013) (Lawson et al. 2009), respectively. Alignments of the orthologous proteins were performed with MUSCLE (Edgar 2004) with the default parameters.

Identification and Analysis of Orthologous Introns

We selected introns orthologous in all the four species, defined as the introns in the orthologous positions of the coding sequences in the orthologous proteins, and also having the identical phase. For this purpose, coordinates of intron shadows were mapped onto protein alignments. To avoid analyzing nonorthologous introns, only introns mapping to regions of high-quality protein alignment were considered. For this purpose, we disallowed gaps in the two (for phase 0 introns) or one (for phase 1 or 2 introns) amino acid sites to which the intron mapped, and in the two immediately neighboring amino acid sites to the left and to the right of it. Furthermore, we required at least five alignment positions similar by BLOSUM62 matrix, and no more than two alignment gaps, in each of the species within 10 amino acids flanking the intronic shadow from each side. To ensure that we are studying noncoding sequences, we excluded from the analysis those introns which overlapped protein-coding exons in any known transcript for this gene.

After all filtering, we identified 5,367 and 51,844 sets of orthologous introns in quartets 1 and 2, respectively. The first 6 and the last 16 nucleotides of the intronic sequences were excluded from analyses of conservation, as they are likely to be under selective constraint due to presence of elements crucial for the correct splicing of an intron (Haddrill et al. 2005). The remaining parts of the intronic sequences of the four species were then aligned with bl2seq (Altschul et al. 1990) with anchor length set to 7 and low complexity filtering on. Alignments were performed for each pairwise combination of species from the quartet. Sets of orthologous introns with significant similarity (bl2seq *E*-value ≤ 0.0001) between sequences from different pairs were excluded from analysis.

Calculation of Expected Number of Introns Carrying a Segment of Similarity within Both Pairs

We started by counting, for a particular *E*-value, the numbers of introns $N_1(E)$ and $N_2(E)$ carrying a region of local sequence similarity in pairs 1 and 2, respectively. The expected number of introns carrying local sequence similarities within both species pairs was then calculated using four different randomization procedures: 1) without accounting for any potential confounding variables; 2) accounting for intron lengths; 3) accounting for gene identity; and 4) accounting for gene identity and for whether the intron was first or subsequent.

In the first procedure, in each of the 10,000 reshuffling trials, we simply redistributed $N_1(E)$ and $N_2(E)$ introns among all introns of pairs 1 and 2 randomly, and counted the number of introns that have within-pair similarity in both species pairs in this resampled set. The resulting distributions were used to obtain the means and the confidence intervals in supplementary figure S1, Supplementary Material online. (The expected fraction of introns with similarity in both species pairs among all introns generated by this procedure is roughly equal to that obtained by simply multiplying the frequencies of introns with within-pair similarity in the first pair by this frequency in the second pair.)

As the lengths of orthologous introns are correlated between species pairs (e.g., for *D. melanogaster* and *C. quinquefasciatus*: Spearman's rho $= 0.29$, *P* value $< 2.2\text{e}{-16}$; for *H. sapiens* and *Tak. rubripes,* Spearman's rho $= 0.215$, *P* value $< 2.2\text{e}{-16}$), and longer introns are generally better conserved than the short ones (supplementary fig. S2, Supplementary Material online), the numbers of introns carrying local sequence similarities within both species could be confounded by intron lengths. To account for this effect in the second procedure, we subdivided introns into10 bins according to their length within each species pair. Bins contained roughly equal numbers of introns; precisely equally sized bins were unobtainable, due to a relatively large number of introns that fell on the bin thresholds, particularly for short introns in dipterans. The bin thresholds were as follows: *Homo–Mus* (125, 279, 499, 757, 1,077, 1,479, 2,046, 3,036, and 5,620); *Tetraodon–Takifugu* (72, 77, 82, 89, 102, 130, 192, 337, and 709); *D. melanogaster–D. mojavensis* (56, 58, 60, 62, 64, 67, 73, 124, and 573); and *Aedes–Culex* (57, 59, 61, 63, 66, 71, 125, 666, and 4,476). We then classified the $N_1(E)$ and $N_2(E)$ introns carrying regions of similarity in pairs 1 and 2 by bins of intron length in pair 1 and 2, respectively. We thus obtained the distributions of introns with regions of similarity in pairs 1 and 2 by intron length. In each of the 10,000 resampling trials, we then randomly drew, from each of these distributions, $N_1(E)$ and $N_2(E)$ introns for pairs 1 and 2, respectively, and counted the number of introns with within-pair similarity in both species pairs in this resampled set. The resulting distributions of numbers of introns with within-pair similarity in both pairs were used to obtain the means and the confidence intervals in figure 2.

The third procedure was equivalent to the first, except that we controlled for identities of the genes by only reshuffling the introns within the same gene, rather than genome-wide. Finally, the fourth procedure was equivalent to the third, except that the first introns of genes were excluded from analyses. The resulting distributions were used to produce supplementary figures S3 and S4, Supplementary Material online, respectively.

Analysis of Data on Chromatin Modifications

Data on chromatin modifications for 51,541 human and 5,367 fruit fly introns were obtained from ENCODE (ENCODE Project Consortium 2004; Ernst et al. 2011) (http://genome.ucsc.edu/ENCODE/, last accessed February 27, 2013) and modENCODE (Kharchenko et al. 2010; Roy et al. 2010) (http://www.modencode.org/, last accessed February 27, 2013) databases, respectively. ENCODE provides chromatin state segmentation and the corresponding predicted functional annotation for nine cell lines. We excluded from the analysis the two cancer cell lines (K562 and HepG2). The remaining cell lines were further subdivided into adult (GM12878, HMEC, HSMM, NHEK, and NHLF) and embryonic cell lines comprising embryonic stem cells (H1-hESC) and umbilical vein endothelial cells (HUVEC). We used segmentation tracks for Human Genome Build 36 (hg18), selecting the introns overlapping with strong enhancers (states 4 and 5), or insulators (state 8) either in all adult tissue cells or in both embryonic line cells.

For *D. melanogaster,* ModENCODE provides segmentation models for two cell lines (BG3 and S2). We used segmentation tracks for FlyBase release 5, selecting the introns overlapping with either Regulatory regions (enhancers) (state 3) or with Active introns (state 4) at least in one of the cell lines. We estimated the statistical significance of the results in 10,000 resampling trials while correcting for introns length (procedure 2 of the previous section).

Identification of Intron Losses

For the third and the fourth tests, to trace the losses of introns on the phylogeny, the tree corresponding to each species quartet was rooted with an outgroup species (fig. 1). We then selected introns present in both species from pair 1 and in the outgroup species, assuming that such introns were present in the last common ancestor of both considered species pairs. Again, orthologous introns with significant similarity (bl2seq *E*-value $\leq$ 0.0001) between sequences from different pairs were excluded from analysis. These introns were then subdivided into 1) those also present in both pair 2 species and 2) those that had been lost in at least one of the pair 2 species. (Introns absent in one of the pair 2 species were usually also absent in the other, implying loss on the branch separating pair 2 from its common ancestor with pair 1; cases of intron loss mapping to external branches were rare.) For groups of introns indicated in 1) and 2), we compared the distributions of *E*-values within pair 1, and the distributions of intron lengths within pair 1. Reciprocal tests were performed analogously.

Results and Discussion

We investigated, at the genomic scale, the common selective constraint (probably associated with conservation of function) in homologous, but highly divergent, noncoding sequences. We focus on the genomic segments that have diverged from their common ancestor to such an extent that they have lost all primary sequence similarity. Generally, any similarity of properties of orthologous sequences that are not alignable suggests the presence of such a selective constraint, unless this similarity can be explained otherwise. As a sample of genome segments orthologous between distant species, we used the introns of orthologous genes, because their orthology can be easily determined through flanking exons even at phylogenetic distances so large that the introns themselves are no longer alignable.

We used four tests that can provide evidence of selective constraint without sequence similarity. Each test was done on a quartet of species. A quartet consists of two pairs of species, such that the evolutionary distance at selectively neutral sites within the first (Ks_1) and the second (Ks_2) pair is sufficiently large so that the sequence conservation between two species of a pair is indicative of selective constraint, but much shorter than the distance between the two pairs (Ks_3). We studied two such quartets, of dipterans and of vertebrates (fig. 1). In each quartet, there are many introns that contain highly significant local sequence similarities within each pair, indicative of selective constraint. In contrast, there are only a few meaningful sequence similarities between introns from species that belong to different pairs within a quartet.

First, we asked whether the presence of a conserved (and, by inference, functional) segment between two species of a pair within a quartet is a significant predictor for the presence of a conserved segment in the orthologous intron between the two species of the other pair. Consider orthologous noncoding segments that have a function conserved between all four species of a quartet. Such functional conservation should lead to above-neutral sequence conservation within each pair. In addition, strong functional conservation may lead to above-expected sequence conservation even at much higher evolutionary distances that separate the two pairs. In the latter case, conservation of a sequence segment spans the entire quartet. As our focus was functional conservation without sequence conservation, we excluded from the analysis 34 and 303 introns with significant local similarities between species from different pairs in dipteran and vertebrate quartets, respectively. The remaining 5,333 introns in the dipteran quartet, and 51,541 introns in the vertebrate quartet, therefore, contained only those introns that were unalignable between the two pairs. Still, an intron that contains a significant local similarity within one pair of species contains a significant local similarity within the other pair much more often than would be expected if these segments were distributed over the introns independently in each of the pairs (supplementary fig. S1, Supplementary Material online).

However, this analysis can be confounded by differences in intron lengths. Indeed, longer introns are more likely to contain within-pair similarities (supplementary fig. S2, Supplementary Material online), in agreement with the data on their higher conservation, at least in *Drosophila* (Haddrill et al. 2005), and intron lengths are correlated between pairs of species (see Materials and Methods). This nonuniformity with respect to intron length should be controlled for.

Therefore, the expected number of introns carrying a significant local similarity within both pairs should be obtained by summation over bins of introns of different lengths, to avoid underestimating this number due to correlated intron lengths in different pairs of species within a quartet, and therefore overestimating our effect. Nevertheless, even after this correction, we observed substantially more introns with conservation in both pairs than expected (fig. 2). Therefore, among those introns that are unalignable between the two pairs of species within a quartet, an excess of introns conserved in both species pairs was observed. This excess suggests the presence of a selective constraint that did not lead to observable sequence conservation between the two pairs. The excess of conservation was stronger under more stringent similarity thresholds, that is, when the parameters were chosen in such a way that only a small fraction of introns were aligned within a pair. When this fraction dropped below 5% for the dipteran quartet, or below 3% for the vertebrate quartet, the number of introns possessing an alignment in both pairs of species exceeded the random expectation by a factor of 3; in vertebrates, when only the 0.4% introns with the most stringent conservation were used, an 8-fold excess was observed (fig. 2).

Nonrandom distribution of conservative elements among introns can be caused not only by conservation of the ancestral function but also by some common characteristics of introns not necessarily associated with their common origin. In particular, a pattern similar to that in figure 2 is expected if introns of a particular subset of genes (e.g., of highly expressed genes) are more likely to contain a conserved element, or if the first introns of genes tend to be more conserved. To test whether these common features of orthologous introns lead to the observed pattern, we reshuffled, for each species pair, the introns within each gene, and asked whether an excess of introns with coincident conservation is still observed in the data, compared with this control. This test is extremely conservative, because such reshuffling is expected to lead to false-negatives if the number of introns is low, and especially if there are many genes with only one intron. In fact, the mean number of introns with established orthology per gene in our data set was 1.95 (median = 1) for dipterans, and 7.00 (median = 5) for vertebrates. Still, even in this very conservative test, the excess of introns carrying conserved regions in both pairs remained (supplementary fig. S3, Supplementary Material online). It also remained after exclusion of the first introns of genes (supplementary fig. S4, Supplementary Material online), suggesting that it was not due to their higher conservation. Therefore, the observed coincident conservation of unalignable regions in phylogenetically remote species is not simply due to common characteristics of the orthologous introns or of the genes carrying them.

Second, we hypothesized that if some regulatory elements persist for longer than the sequence similarity of the corresponding DNA segments, we would expect introns containing a regulatory element in pair 1 also to carry a segment of similarity in pair 2 (again, correcting for the differences in intron lengths). To study this, we used genome-wide predictions of regulatory regions based on patterns of chromatin modifications (Kharchenko et al. 2010; Roy et al. 2010; Ernst et al. 2011). As in the first test, we excluded from the analysis 34 and 303 introns with significant local similarities between species from different pairs in dipteran and vertebrate quartets, respectively. Nevertheless, we found that *Drosophila melanogaster* introns, which overlap regions that are enriched in active chromatin modifications, and therefore likely to be involved in regulation, are up to approximately three times more likely to carry a segment of similarity in the mosquito species pair (fig. 3). Analogously, introns overlapping insulators or enhancers in human are respectively up to 2.6 times or 1.4 times more likely to carry a segment of similarity between the two fish species (fig. 3). In enhancers, a stronger effect was observed if only embryonic cell lines were considered (embryonic stem cells and umbilical vein cord cells, fig. 3), in line with the observations that conserved noncoding regions tend to be associated with genes involved in developmental processes (Woolfe et al. 2005).

Third, in addition to conservation of a sequence within an intron, presence of a functional segment may lead to a reduced rate of loss of such introns in evolution. We asked whether among the introns that were present in pair 1 species, the introns that were preserved in pair 2 species are more likely to carry a segment conserved within pair 1. Because in this analysis, we need to discriminate between intron losses and gains, we used an outgroup species to determine the ancestral state. Again, to avoid dealing with sequence similarities spanning all four species of a quartet, we excluded from the analysis 14 out of 3,073, and 54 out of 6,609, introns with significant sequence similarities between species from different pairs, in quartets 1 and 2, respectively. In the remaining set, higher prevalence of conserved, within a pair, segments within introns that were preserved between pairs would imply functional conservation (with no detectable underlying sequence conservation) spanning both pairs. Among all introns, an intron present in both *Drosophila* species and in the outgroup *(Apis mellifera)* is also present in both mosquito species in 69% of cases. Those 69% of introns were significantly more likely to carry a segment of similarity between *Drosophila* species than the remaining 31% (*P* value = 5.71e−07, Fisher's exact test; fig. 4*A*, inset). An intron present in both mammalian species and in *Ciona intestinalis* is also present in both species of fish in 98% of cases. A higher fraction of introns is preserved in the vertebrate quartet because the analyzed vertebrate species are more closely related than the dipteran species, and intron loss is less frequent in vertebrates (Putnam et al. 2007). Those 98% of introns were somewhat more likely to carry a segment of similarity between human and mouse than the remaining 2%, although the difference was not significant (*P* value = 0.167,

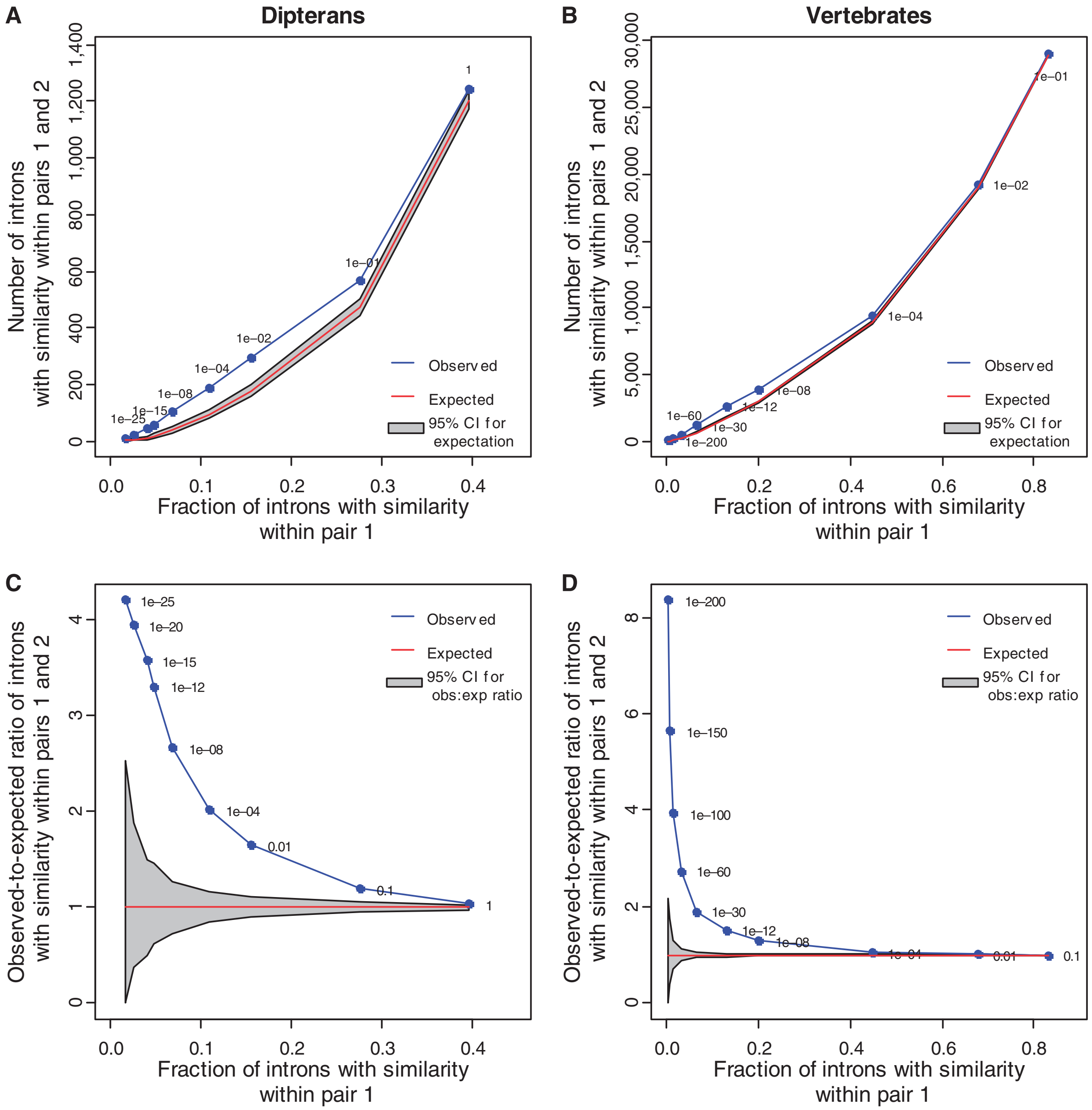

FIG. 2.—Introns that carry a segment of high similarity between species of one pair are more likely to also carry a segment of high similarity between species of the other pair within a quartet. Each blue dot corresponds to a specific BLAST *E*-value (shown next to the dot), with lower values corresponding to more stringent similarity thresholds. Each *E*-value was used to detect similar segments within orthologous introns of the two species belonging to pair 1, and of the two species belonging to pair 2. Horizontal axis, fraction of introns that carry similar segments within pair 1, among introns present in all four species. Vertical axis, number (*A, B*) or observed-to-expected ratio (*C, D*) for the number of introns that carry similar segments both within pair 1 and 2. Top, dipterans; bottom, vertebrates. Observed-to-expected ratio was defined as the ratio of the observed number of introns with similarity within both pairs to the expected number if the segments of similarity were distributed randomly over all introns, controlling for intron lengths (see text). The red line and the gray area correspond to the mean and 95% confidence intervals for the expected values calculated in 10,000 resampling trials.

Fisher's exact test; fig. 4*B*, inset), probably because nearly all the introns carry some similarity between human and mouse under the thresholds used. Moreover, among introns carrying conserved segments, this similarity was higher within introns that were preserved in both pair 2 species, compared with the remaining introns, both in the dipteran (*P* value = 0.00103) and in the vertebrate quartets (*P* value = 0.0203, Wilcoxon rank sum test with continuity correction) (fig. 4). The results of the reciprocal tests were similar (supplementary fig. S5, Supplementary Material online).

Fourth, longer introns are more likely to carry a segment of conservation than short introns (Haddrill et al. 2005).

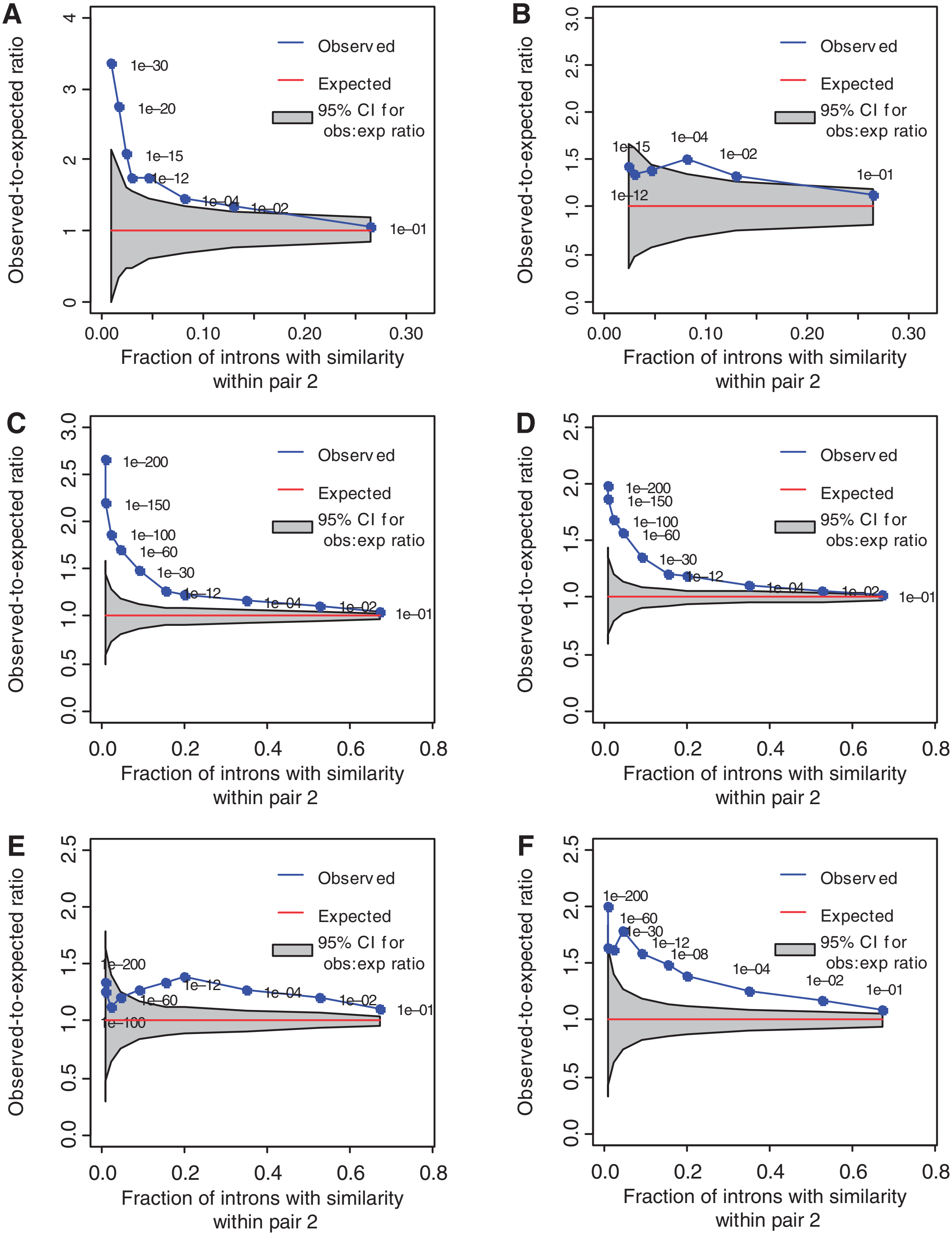

FIG. 3.—Introns that carry a segment of similarity in pair 2 are more likely to overlap regulation-associated elements within pair 1. Each blue dot corresponds to a specific BLAST *E*-value (shown next to the dot). Each *E*-value was used to detect similar segments within orthologous introns of the two species belonging to pair 2. Horizontal axis, fraction of introns that carry similar segments within pair 2, among introns present in all four species. Vertical axis, observed-to-expected ratio for the number of introns that carry a regulation-associated element within pair 1, according to modENCODE (Kharchenko et al. 2010; Roy et al. 2010) (*A*, *B*) or ENCODE (ENCODE Project Consortium 2004; Ernst et al. 2011) (*C*–*F*) data, and also carry a segment of similarity within pair 2. (*A*, *B*) Dipterans; (*C*–*F*) vertebrates. (*A*) Active introns, any cell line; (*B*) enhancers, any cell line; (*C*) insulators, all adult cell lines; (*D*) insulators, both embryonic cell lines; (*E*) strong enhancers, all adult cell lines; (*F*) strong enhancers, both embryonic cell lines. Observed-to-expected ratio was defined as the ratio of the observed number of introns with a regulation-associated element within pair 1 and similarity within pair 2 to the same number expected if the regulation-associated elements and the segments of similarity were distributed randomly over all introns, controlling for intron lengths. The red line and the gray area correspond to the mean and 95% confidence intervals for the expected values calculated in 10,000 resampling trials.

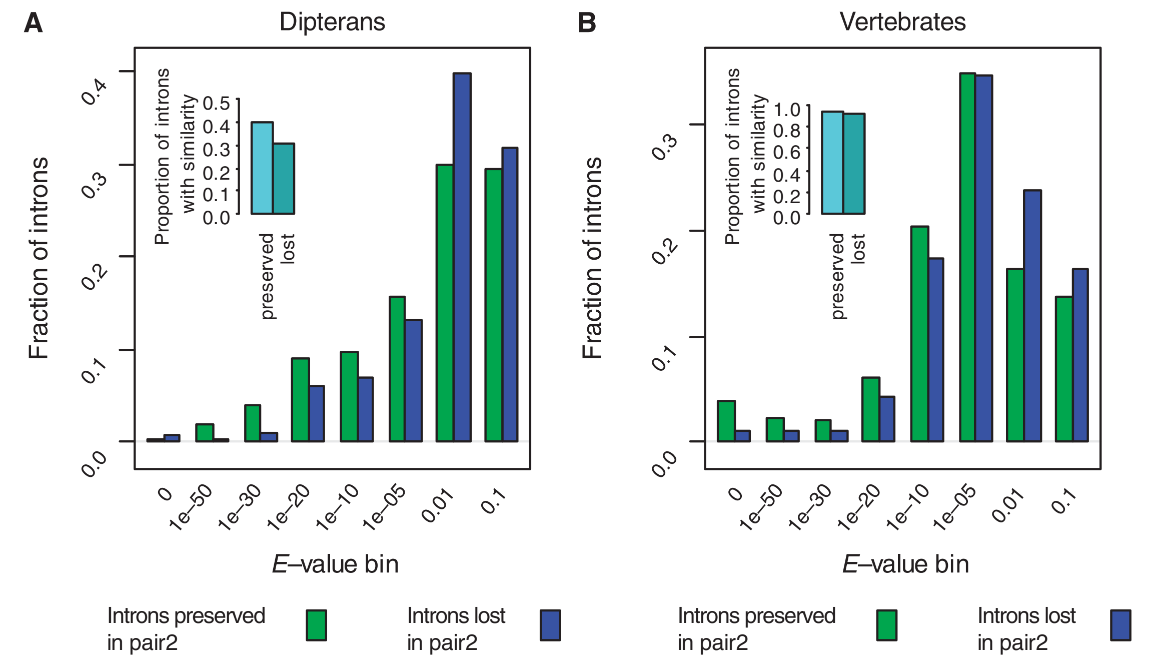

Fig. 4.—Introns preserved in both pair 2 species are more likely to carry a conserved segment within pair 1. For each quartet, the distribution of *E*-values within pair 1 are shown for introns preserved in both pair 2 species (green), and for introns lost in at least one of the pair 2 species (blue). *E*-values indicated below the horizontal axis correspond to the lower *E*-value threshold. Insets show the fraction of introns with at least a marginal (*E*-value ≤ 1) similarity observed, for the same two groups.

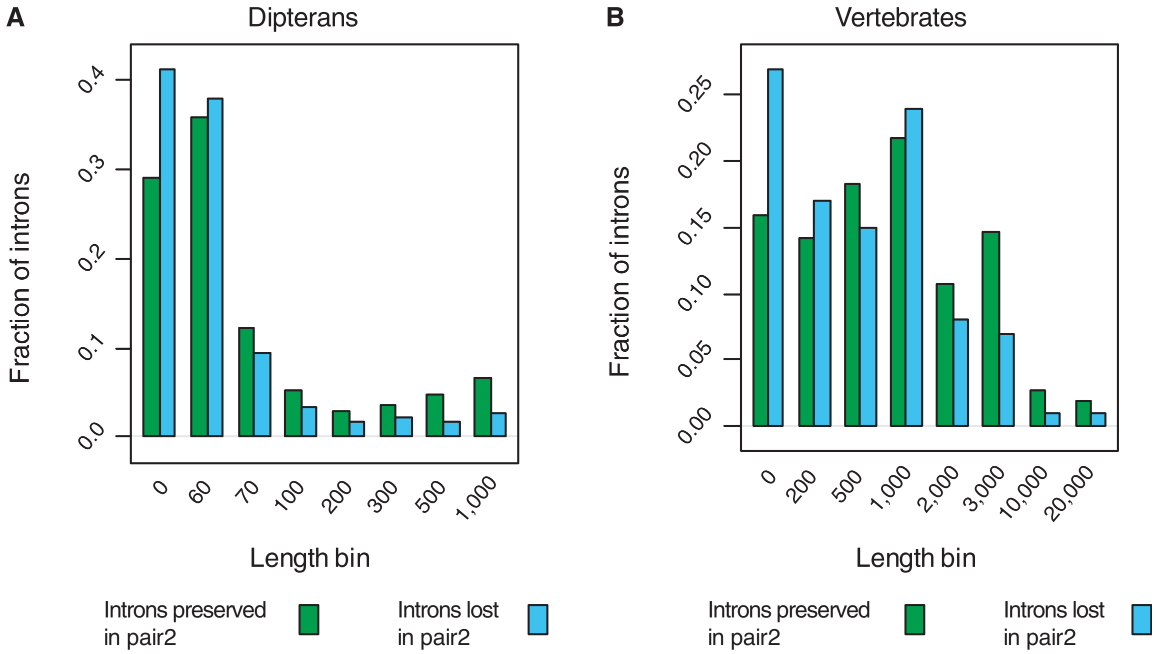

Fig. 5.—Introns preserved in both pair 2 species tend to be longer within pair 1 species. For each quartet, the distribution of lengths within pair 1 are shown for introns preserved in both pair 2 species (green), and for introns lost in at least one of the pair 2 species (blue). Lengths indicated below the horizontal axis correspond to the lower length threshold. The length of the shorter of the two orthologous introns in pair 1 was used.

Therefore, a function conserved in all four species of a quartet may be associated with a larger length in pair 1 of introns preserved in pair 2, compared with the introns lost in pair 2. Indeed, among the 3,059 (6,555) dipteran (vertebrate) introns found in both pair 1 species and in the outgroup, the introns also present in both pair 2 species tend to be longer in pair 1 species than introns lost in pair 2, in both quartets (dipterans: $P = 2.64e{-}18$; vertebrates: $P = 0.00111$; Wilcoxon rank sum test with continuity correction; fig. 5). Again, for both quartets, the results of the two reciprocal tests were similar (supplementary fig. S6, Supplementary Material online). Therefore, we observe a higher preservation of introns possessing long orthologs in a phylogenetically remote species, and therefore presumably most likely to carry a functional DNA segment in those species, compared with introns with short orthologs.

Thus, all the four analyses provide evidence for selective constraint which keeps operating even after diverging orthologous introns became unalignable, in the evolution of dipterans and vertebrates. Apparently, long-living functional elements in orthologous genomic compartments, which persist longer than the conservation of the primary sequence, are common. The excess of introns with conserved segments (fig. 2) suggests that such elements reside within approximately 5% of introns in dipterans, and approximately 3% in vertebrates.

Selective constraint acting on homologous, unalignable DNA segments is also likely to be common within intergenic regions, which carry numerous regulatory elements (Heintzman et al. 2009). However, it is difficult to subdivide unalignable intergenic regions into orthologous compartments, which are the core of our analysis. The precise nature of constraint imposed by the conserved function on the evolution of homologous DNA segments which are no longer alignable also remains a mystery.

Supplementary Material

Supplementary figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Ministry of Education and Science of the Russian Federation (grant nos 11.G34.31.0008 and 8814), by the Russian Foundation for Basic Research (grant no. 12-04-33202 mol_a_ved), and by the Molecular and Cellular Biology Program of the Russian Academy of Sciences. We wish to thank Dr Peter Kharchenko for help with ENCODE data and two anonymous reviewers for their suggestions.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

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

Aparicio S, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310.

Arensburger P, et al. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 330: 86–88.

Clark AG, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Dehal P, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167.

Dermitzakis ET, et al. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302:1033–1035.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640.

Ernst J, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17: 1837–1849.

Heintzman ND, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459:108–112.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949.

Iyer LM, Leipe DD, Koonin EV, Aravind L. 2004. Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol. 146: 11–31.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957.

Kersey PJ, et al. 2009. Ensembl genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res. 38:D563–D569.

Kharchenko PV, et al. 2010. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature 471:480–485.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lawson D, et al. 2009. VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res. 37:D583–D587.

Lowe CB, et al. 2011. Three periods of regulatory innovation during vertebrate evolution. Science 333:1019–1024.

McGaughey DM, et al. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. 18:252–260.

Murzin AG, Bateman A. 1997. Distant homology recognition using structural classification of proteins. Proteins (1 Suppl):105–112.

Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723.

Odronitz F, Becker S, Kollmar M. 2009. Reconstructing the phylogeny of 21 completely sequenced arthropod species based on their motor proteins. BMC Genomics 10:173.

Ostlund G, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.

Putnam NH, et al. 2007. Sea anemone genome reveals ancestral Eumetazoan gene repertoire and genomic organization. Science 317:86–94.

Roy S, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797.

Sayers EW, et al. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 40:D13–D25.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci. 255:279–284.

Taher L, et al. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. Genome Res. 21:1139–1149.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. 8:R15.

Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Wheeler DL, et al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36: D13–D21.

Woolfe A, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3:e7.

Associate editor: Martin Embley