SUPPLEMENTARY MATERIAL

The Shared Motif Method (SMM) Algorithm

The SMM discovers regions of local similarity between DNA sequences without respect to their order, orientation, or spacing, based on the recursive local alignment algorithm described by Waterman and Eggert (1987). We first search for the best local alignment between two sequences then mask off the resulting *alignment*. While this particular alignment match between sequence 1 and sequence 2 is not allowed in subsequent iterations, matches involving the aligned portion of either sequence with *another region of the other sequence* are allowed.

Specifically, if we define *n1* and *n2* as the length of each sequence, a matrix of *n1* □ *n2* is filled as described by Smith and Waterman (Smith and Waterman 1981). After the first local alignment is computed, the matrix is recomputed excluding the best alignment path and all matrix elements affected by this alignment following the method of Waterman and Eggert (1987). This process is iterated until an alignment with an arbitrarily small score is found (see below). While the previous (n-1) alignment path is always excluded, re-discovery of (n-2), (n-3), (n-4), ... alignment paths is possible. To avoid this, we add the additional constraint that matrix elements that are affected by the previous (n-1) alignment (that do not result from a gap) can never yield scores greater the previous alignment. Any path that leads to a score increase is therefore forbidden.

This process is repeated after inverting one of the sequences to search for inverted regions of shared similarity. Finally, all shared motifs, in both orientations, are mapped onto the original sequences and the cumulative fraction of shared motifs is calculated.

Note that overlapping regions of shared motifs are possible and are ignored when calculating the cumulative fraction of shared motifs. Shared motif *divergence* (d_{SM}) is defined as 1 – the cumulative fraction of shared motifs.

The minimum alignment score can be empircally derived by analyzing the distribution of d_{SM} for random sequence pairs with a similar nucleotide composition as the data examined. For example, using the parameters described in the Methods, analysis of 1,000 random sequence pairs of 500 bp showed that >90% of sequence pairs exhibit a $d_{SM} > 0.90$. More or less stringent parameter values may be used depending on the particular aim of the study at hand. Further details of the algorithm and its implementation can be found in the source code.

Additional SMM Positive Controls – Human/Mouse Orthologs

While the SMM is not a motif discovery algorithm, conserved blocks of sequences discovered by the method should include a large fraction of *cis*-acting elements that are experimentally known to be involved in gene regulation. We obtained all TRANSFAC Database (Matys et al. 2003) motif accession identifiers (IDs) for all genes in the Eukaryotic Promoter Database (Release 77) (Praz et al. 2002) (http://www.epd.isb-sib.ch/) which contained an NCBI RefSeq ID for human. Next, we obtained a list of all human/mouse orthologs from NCBI (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/homol_seq_pairs) and matched these with the human RefSeq IDs. Human and mouse upstream sequences, from July 2003 and October 2003 genome releases, respectively, were retrieved using the UCSC Genome Browser TableBrowser (http://genome.ucsc.edu/) and included 1kb upstream from transcription start site (TSS), if known, plus 5' UTR if known *or* 1kb 5' from translation start if the TSS was unknown. Annotated *cis*-acting DNA motifs were then obtained from TRANSFAC (Release 7.2) (http://www.gene-regulation.com/) and their prescence was verified in the upstream region of each human gene. Motifs that were not found or did not match exactly were excluded. Next, we used the SMM to estimate d_{SM} and to obtain SMM alignments using a minimum score of 48 (Methods). Finally, motifs were mapped onto SMM alignments. 62 of 79 experimentally verified motifs (78%) were found within conserved regions identified by the SMM. Details are provided in Supplementary Table 1.

Control for d_{SM} with and without 5’ UTRs

To check whether analysis of upstream sequences 5’ to transcription start versus translation start greatly affected d_{SM}, we obtained a list of all genes for which a valid transcription start site was known for *C. elegans* in the *ensembl* database (http://www.ensembl.org/Caenorhabditis_elegans/) based on 5' EST and cDNA data. Of the 2,150 pairs of duplicates in our dataset, 56 possessed an annotated transcription start site for both genes. Of these pairs, differences in d_{SM} for sequences 500 bp upstream of translation start versus 500 bp upstream of transcription start was small (mean = 0.07 ; std. error = 0.22) and there was a strong overall correlation in d_{SM} between upstream regions with and without 5' UTRs ($r_s = 0.66$, $P << 10^{-4}$, Spearman rank correlation). We conclude that the analysis of sequences 5’ to transcription start versus translation start does not greatly affect d_{SM} in *C. elegans*/*C. briggsae*.

Control for Possible Effects of Errors in Gene Prediction on the Correlation

between Protein and Regulatory Divergence

Here we test the possibility that the correlation observed between d_{N} and d_{SM} is an artifact of poor gene prediction; genes that are poorly predicted may have a high d_{N} and a high d_{SM} resulting in a spurious correlation between protein and regulatory evolution. Therefore, we examined the relationship between d_{N} and d_{SM} using only those genes known to be expressed in *C. elegans* (genome-wide expression data is not yet available for *C. briggsae*). We considered a gene expressed if it was *i*) reliably detected in one or more replicate experiments of Hill *et al.* (2000) at one or more developmental time-points or *ii*) had associated EST or cDNA expression data as annotated the *ensembl* database (http://www.ensembl.org/Caenorhabditis_elegans/).

Of the 10,648 genes that met the microarray expression criterion, our dataset contained 139 *C. elegans* duplicate pairs (with both copies expressed) and 1,401 orthologous pairs (where the *C. elegans* gene was known to be expressed). Examining only these genes, we find that the weak correlation between *cis*-regulatory and protein evolution still holds for both orthologs and paralogs ($r_s = 0.17$, $P << 10^{-4}$ for orthologs and $r_s = 0.18$, $P = 0.02$ for paralogs). As in the larger dataset, a multiple regression involving d_{N}, d_{S}, and d_{SM} revealed this correlation is a result of the relationship of d_{SM} and d_{S} in duplicates (duplicate age) but not in orthologs. Similar results were found using the 56 duplicate gene pairs that had evidence of expression based on EST or cDNA data (data not shown).

Control for Tandem and Non-Tandem Duplicate Genes

Tandem duplicate genes are common in the genome of *C. elegans* and have been shown to have different genomic and evolutionary dynamics from non-tandem duplicate genes (Achaz et al. 2001; Katju and Lynch 2003; Semple and Wolfe 1999). Of the 869 pairs of duplicates in our dataset, 594 (68%) are on the same chromosome and 168 of these (28%) occur in tandem with no intervening gene— the majority of which are <1 kb apart (137 of 168). We found that rates of protein and regulatory evolution are not significantly different between tandem and non-tandem duplicate genes (Supplementary Table 2).

cDNA Microarray Analysis

To compare d_{SM} with differences in relative expression (temporal or context dependent gene expression), we compared Pearson's *r* correlation coefficient between pairs of duplicate genes across 553 experiments (Kim et al. 2001) using normalized $\log_2$ [Cy3/Cy5] ratios for genes with valid data (>30 data points with >2-fold change). 544 pairs of duplicates in our dataset met these criteria. We found no significant correlation between d_{SM} and changes in expression quality ($r_s = 0.06$, $P = 0.09$). However, cDNA microarrays unlike Affymetrix™ microarrays are sensitive to cross-hybridization (Kane et al. 2000; Castillo-Davis et al. unpublished results) and it is possible that cross-hybridization of duplicate gene transcripts obscured the relationship (if any) between d_{SM} and relative expression. Indeed, a multiple regression analysis of relative expression similarity (*r*) on d_{N}, d_{S}, d_{SM}, and percent nucleotide identity, as computed after performing an "end gap free" global alignment, revealed that the only significant predictor of similarity in relative expression was the percent nucleotide identity between duplicate pairs (Supplementary Table 3). Thus, it remains unclear whether d_{SM} is related to changes in *relative* expression. Future analysis of data from experiments not sensitive to the problem of cross-hybridization should shed light on this issue.

Supplementary Table 1. *Presence of experimentally verified DNA binding sites within shared motif blocks discovered by the SMM in orthologous human/mouse genes.*

H. sapiens[a]	*M. musculus*[a]	Number of experimentally verified motifs[b]	Motifs found in shared motif blocks[c]	d_{SM} [d]	Brief description [e]
NM_000089	NM_007743	3	3	0.27	COL1A2, collagen, type I, alpha 2
NM_000132	NM_007977	4	3	0.63	F8, coagulation factor VIII
NM_000312	NM_008934	1	0	0.89	PROC, Protein C
NM_000315	NM_020623	1	1	0.46	PTH, Parathyroid hormone
NM_000546	NM_011640	8	7	0.35	TP53, Tumor protein p53
NM_000594	NM_013693	1	1	0.24	TNF, Tumor Necrosis factor
NM_000758	NM_009969	5	5	0.31	CSF2, Colony stimulating factor 2
NM_000759	NM_009971	1	1	0.12	CSF3, Colony stimulating factor 3
NM_001547	NM_008332	2	1	0.77	IFIT2, interferon-induced protein with tetratricopeptide repeats 2
NM_001968	NM_007917	2	2	0.00	EIF4E, eukaryotic translation initiation factor 4E
NM_002133	NM_010442	1	0	0.66	HMOX1, heme oxygenase (decycling) 1
NM_002467	NM_010849	6	6	0.05	MYC, v-myc myelocytomatosis viral oncogene homolog
NM_002539	NM_013614	1	0	0.60	ODC1, ornithine decarboxylase 1
NM_002575	NM_011111	6	0	0.81	SERPINB2, serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2
NM_003094	NM_009227	2	2	0.59	SNRPE, small nuclear ribonucleoprotein polypeptide E
NM_004094	NM_026114	1	1	0.04	EIF2S1, eukaryotic translation initiation factor 2, subunit 1 alpha

NM_004462	NM_010191	2	2	0.71	FDFT1, farnesyl-diphosphate farnesyltransferase 1
NM_005252	NM_010234	20	20	0.20	FOS, v-fos FBJ murine osteosarcoma viral oncogene homolog
NM_005332	NM_010405	6	6	0.52	HBZ, hemoglobin, zeta
NM_019111	NM_010381	6	1	0.97 [f]	HLA-DRA, major histocompatibility complex, class II, DR alpha
Total		79	62	0.46 [g]	-

a: NCBI RefSeq-ID of human and mouse genes.
b: Number of experimentally known motifs in *H. sapiens* upstream region.
c: Number of motifs that are located in shared motif blocks discovered by the SMM.
d: d_{SM} was estimated from sequences ~1 kb upstream in both species.
e: Brief gene description (GenBank).
f: This orthologous pair, which exhibited almost no upstream conservation ($d_{SM} = 0.97$), belongs to the MHC gene family which has been shown to be subject to intense overdominant positive selection (Hughes and Nei 1988; Hughes and Nei 1989).
g: The mean d_{SM} across all human/mouse orthologs examined.

Supplementary Table 2. *Comparison of tandem and non-tandem duplicate genes.*

	n	d_N/d_S [a]	d_{SM}/d_S [a]	d_N/d_{SM} [a]
Tandem duplicates within *C. elegans*	168	0.29	0.93	0.29
Non-tandem duplicates within *C. elegans*	701	0.33	1.02	0.27

a: Median values are shown. The distribution of d_N/d_S, d_{SM}/d_S, and d_N/d_{SM} is not significantly different between tandem and non-tandem duplicate genes ($P > 0.1$, Wilcoxon *U*-test).

Supplementary Table 3. *Multiple regression analysis of relative expression similarity measured with cDNA microarrays on d_{SM}, d_N, d_S, and nucleotide percent identity between duplicate genes.*

Formula: expr (r) ~ $d_{SM} + d_N + d_S$ + %ID

Variable	Coefficient	Std. Error	t-value	$Pr(>\|t\|)$
d_{SM}	0.003936	0.046619	0.084	0.93274
d_N	-0.162634	0.112114	-1.451	0.14747
d_S	-0.002242	0.037980	-0.059	0.95295
%ID[a]	0.359721	0.127354	2.825	0.00491 **

a: Percent nucleotide identity (%ID) between duplicate genes.

Supplementary Figure 1. *SMM Negative Control*

Distribution of d_{SM} in *i*) control region, 1-1.5 kb upstream of translation start, *ii*) randomized upstream sequences, and *iii*) test sequences 0-500 bp upstream of translation start. The control region exhibited a distribution of d_{SM} more similar to that of randomized sequences than to that of the test sequences located 500 bp immediately upstream of translation start.

Supplementary Figure 2. *SMM Positive Control*

Conserved blocks of sequences discovered by the shared motif method (SMM) for experimentally well-characterized regulatory regions between species are depicted in dotplots. Experimentally characterized motifs in each species are shown as bars along each axis. Gray bars indicate the experimentally characterized motif was contained as part of a shared motif block with the SMM and black bars indicate the motif was not contained in an SMM block. Hashed bars in (c) indicate putative motifs, all of which were detected by the SMM. (a) SMM dotplot for the upstream region of the heat-shock protein F44E5.5 between *C. elegans* and *C. briggsae*. 8 of 9 experimentally verified *cis*-elements shared between the species (GuhaThakurta et al. 2002) were contained in SMM blocks. (b) SMM dotplot for the *even-skipped* locus in *Drosophila melanogaster* /

Drosophila pseudoobscura (Ludwig et al. 2000). 10 of 12 experimentally verified *cis*-elements shared between the species were contained in SMM blocks (c) SMM dotplot for upstream sequences of *Apetala-3* in *Arabidopsis thaliana* / *Brassica oleracea* (Koch et al. 2001). 10 of 10 experimentally verified and putative *cis*-elements shared between the species were contained in SMM blocks. (d) SMM dotplot for upstream sequences of *CKM* in *Homo sapiens* / *Mus musculus* (Wasserman et al. 2000). 6 of 6 experimentally verified *cis*-elements shared between the species were contained in SMM blocks.

References - Supplementary Material

Achaz, G., P. Netter, and E. Coissac. 2001. Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol* **18:** 2280-2288.

Castillo-Davis, C.I., J.M. Ranz, and D.L. Hartl. unpublished results. cross-hybridization on cDNA microarrays.

GuhaThakurta, D., L. Palomar, G.D. Stormo, P. Tedesco, T.E. Johnson, D.W. Walker, G. Lithgow, S. Kim, and C.D. Link. 2002. Identification of a novel cis-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res* **12:** 701-712.

Hill, A.A., C.P. Hunter, B.T. Tsung, G. Tucker-Kellogg, and E.L. Brown. 2000. Genomic analysis of gene expression in *C. elegans*. *Science* **290:** 809-812.

Hughes, A.L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335:** 167-170.

Hughes, A.L. and M. Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci U S A* **86:** 958-962.

Kane, M.D., T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, and S.J. Madore. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28:** 4552-4557.

Katju, V. and M. Lynch. 2003. The Structure and Early Evolution of Recently Arisen Gene Duplicates in the Caenorhabditis elegans Genome. *Genetics* **165:** 1793-1803.

Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087-2092.

Koch, M.A., B. Weisshaar, J. Kroymann, B. Haubold, and T. Mitchell-Olds. 2001. Comparative genomics and regulatory evolution: conservation and function of the Chs and Apetala3 promoters. *Mol Biol Evol* **18:** 1882-1891.

Ludwig, M.Z., C. Bergman, N.H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564-567.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374-378.

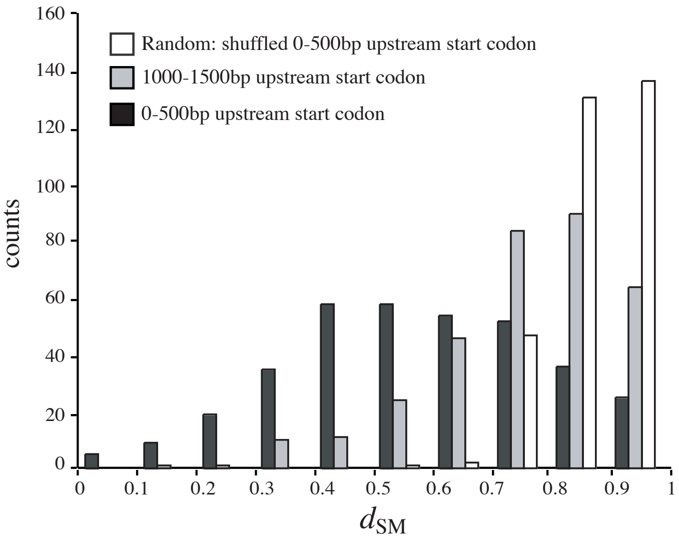
Praz, V., R. Perier, C. Bonnard, and P. Bucher. 2002. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res* **30:** 322-324.

Semple, C. and K.H. Wolfe. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J Mol Evol* **48:** 555-564.

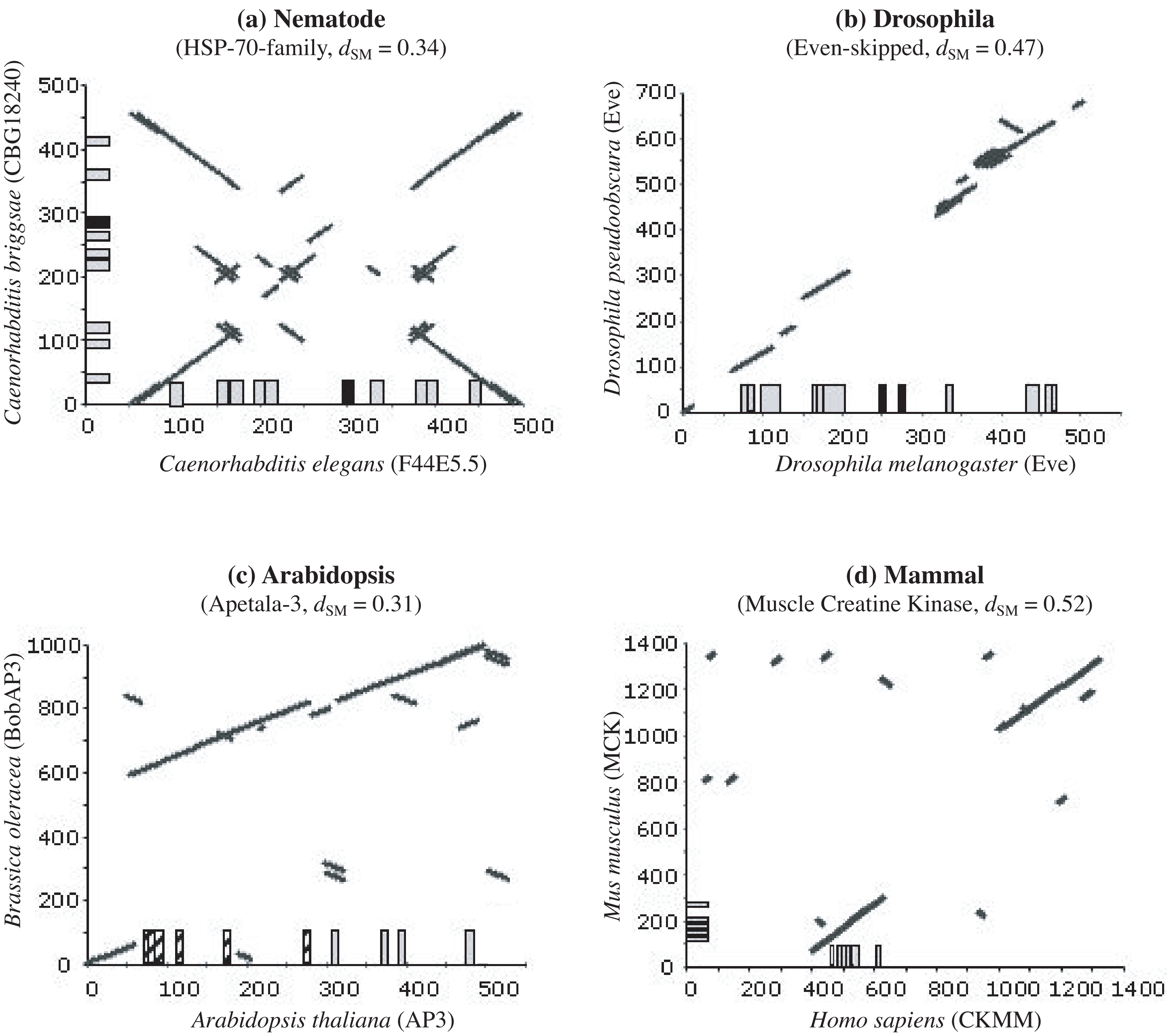
Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

Wasserman, W.W., M. Palumbo, W. Thompson, J.W. Fickett, and C.E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26:** 225-228.

Waterman, M.S. and M. Eggert. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol* **197:** 723-728.



Supplementary Figure 1



(a) Nematode
(HSP-70-family, $d_{SM} = 0.34$)

(b) Drosophila
(Even-skipped, $d_{SM} = 0.47$)

(c) Arabidopsis
(Apetala-3, $d_{SM} = 0.31$)

(d) Mammal
(Muscle Creatine Kinase, $d_{SM} = 0.52$)

Supplementary Figure 2