Supplemental Information:

Nucleotide composition-linked divergence of vertebrate core promoter architecture

Simon J. van Heeringen[1,3], Waseem Akhtar[1,3], Ulrike G. Jacobi[1], Robert C. Akkers[1], Yutaka Suzuki[2], Gert Jan C. Veenstra[1,*]

1 Radboud University Nijmegen, Department of Molecular Biology, Faculty of Science, Nijmegen Centre for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands

2 Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

∗ E-mail: g.veenstra@ncmls.ru.nl

3 These authors contributed equally to this work

Supplemental Methods

Profiles over annotated 5' end of genes

For Supplemental Figs. S1 and S9B we used the annotated 5' end of Xtev genes (Akkers et al., 2009).

Performance of the WIC motif similarity metric

An evaluation of the performance of a motif similarity score is not a trivial problem. The choice of the benchmark dataset, either artifical or experimentally derived, can greatly influence the results. Therefore we chose to assess the performance of the WIC motif similarity score using three different datasets. The first two, used by Mahony et al. (2007) in their extensive similarity metric evaluation, are based on structural classification of motifs from two different databases, while the third dataset tests the ability of the similarity score to distinguish similar motifs favourably over dissimilar motifs. We compared the WIC score to four other similarity metrics: Pearson Correlation Coefficient (PCC), Euclidian Distance (ED), a modified version of the ED (Dist), and FISim. Both PCC and ED have been shown to perform well in benchmark studies (Mahony et al., 2007; Gupta et al., 2007), the Dist score was used by Harbison et al. (2004) in a comprehensive motif discovery pipeline in yeast and the FISim score was recently developed to improve existing similarity metrics (Garcia et al., 2009).

The pairwise comparison score of all motifs within a database is calculated, and the resulting matches are ranked on basis of their score. With a varying score threshold the fraction of positive matches (Sensitivity) is plotted against the fraction of negative matches calculated (1 - Specificity) in a receiver operating characteristic (ROC) curve (Supplemental Figs. S2B-D). Supplemental Fig. S2B shows the results for the 71 JASPAR non-Zn-finger motifs and Supplemental Fig. S2C shows the results for all 585 TRANSFAC motifs. The results for the query- and target motifs subsampled from the JASPAR database, where query motifs should score target motifs derived from the same JASPAR motif higher than other motifs (see Gupta et al. 2007) are shown in Supplemental Fig. S2D. For these three datasets the WIC score shows consistent improvement relative to the other similarity metrics.

Iterative clustering of similar motifs

An ensemble approach which incorporates different computational approaches to predict *de novo* motifs can potentially improve the quality of the motif output compared to single methods. However, a method to reduce the redundancy of largely similar motifs is needed. We chose an iterative procedure to cluster similar motifs using the WIC score, schematically depicted in Supplemental Figs. S3A and B. All pairwise similarity scores of a set of motifs are computed, using a specific metric. The two most similar motifs are merged by summing the frequency counts in the PSSM columns to preserve information about the relative motif abundance. Subsequently all pairwise scores for this new, averaged motif and the other remaining motifs are calculated. The merging of the two most similar motifs and subsequent recalculation of the pairwise similarity scores is repeated, and continues until the maximum score of the two most similar motifs is below a predefined threshold.

To evaluate this approach we clustered the JASPAR non-zinc finger motifs (Mahony et al., 2007) using different thresholds and calculated the weighted mean cluster homogeneity (Supplemental Fig. S3C). For each cluster, including the singletons, the homogeneity is defined as the highest number of motifs of a single structural family divided by the total number of motifs in the cluster. For instance, a cluster of 6 motifs, containing 3 motifs of a single family, and 1 and 2 motifs of two other families respectively, will have a homogeneity of 0.5. The mean homogeneity of all clusters is weighted by the number of motifs in the cluster. As can be seen in Supplemental Fig. S3C the WIC score gives good clustering results, reaching a higher homogeneity at a lower number of clusters than other methods.

Supplemental Fig. S3D shows the final resulting clustered motifs, starting from the 71 JASPAR non-zinc finger motifs. The clustered set includes 16 clusters and 10 singletons (not shown). Out of 16 clusters, 13 are completely homogenous, containing only motifs of the same structural family. Cluster 5, containing mostly bZIP motifs, also includes two nuclear receptor motifs (RORA_1 and RORA_2). The aligment of the motifs in this cluster clearly shows that the most informative positions of the RORalpha motifs closely resemble the bZIP motif with only one strong mismatch. The two other non-homogenous clusters consist of two reasonably similar motifs (cluster 5) and one motif clustered with 7 motifs of a single family (cluster 13).

While the clusters are highly homogenous, this iterative clustering method using WIC results in more singletons than previous approaches (3 for Mahony et al. (2007) and 8 for Garcia et al. (2009) respectively). In our *de novo* motif prediction pipeline, high homogeneity is actually preferable to the least possible number of clusters. Non-similar motifs should not be clustered together, thereby losing the information and specificity of the different motifs.

Prediction of melting temperature

The 11 basepair DNA melting temperature (Tm, Supplemental Fig. S8) was calculated according to SantaLucia (1998) and Owczarzy et al. (2004), also see `http://www.owczarzy.net/Tm_for_duplexes-IDT_Tech.pdf`.

EST Transcription start site selection

The genome assembly versions used for analysis of ESTs and promoter sequences are Joint Genome Institute (JGI) 4.1 (August 2005) for the *Xenopus tropicalis* genome, and hg18 (March 2006) for the human genome. BED files with the genomic positions of spliced ESTs were downloaded from the UCSC Genome Browser database (Rhead et al., 2010) (745,382 *Xenopus tropicalis* ESTs; 4,160,780 human ESTs, December 2007). The genomic locations of ESTs were intersected with the 5' exons of known genes (resp. *Xenopus* FilteredModels v1 genes and human RefSeq genes), and these collections were subjected to two different sets of criteria regarding 5' end distribution and coverage: either at least 3 ESTs overlapping in a 20bp window, or at least 6 ESTs overlapping in a 100 bp window, in both cases comprising at least 20% of the EST cluster. The start sites from both criteria were combined. A number of ESTs can only be partially aligned with genomic sequence, which in some cases is due to an exon being located in a gap of the assembly. This can lead to a contamination of the TSS collection with splice acceptor sites. To eliminate these, all identifiers of ESTs with unaligned 5' ends of 10 bp or more were marked. Clusters of TSSs were required to feature less than 25% ESTs with "flagged" unaligned 5' ends to eliminate splice acceptor site contamination of the TSS collection.

Cloning and Primer Extension

For experimentally determining the endogenous start sites 10μg of total RNA from *X. tropicalis* embryos (Nieuwkoop-Faber stage 12) was reverse transcribed with locus-specific radiolabelled primers. Genomic positions of these loci and primer sequences are given in Supplemental Table S10A. Transcription start sites were assigned only if two or more primers targeting the same RNA showed consistent results. Genomic fragments flanking the selected TSSs were amplified by PCR from *X. tropicalis* genomic DNA and cloned into pG4-BLCAT2 (Kass et al., 1997) at BamHI-XhoI sites using In-FusionTM Dry-Down PCR Cloning Kit (Clontech). Genomic coordinates of these fragments and primer sequences are given in Supplemental Table S10B. For analysis of promoter activity, 0.7 ng of each plasmid was injected into the nuclei of stage VI oocytes from *X. laevis* and RNA was isolated after overnight incubation of oocytes at 16 °C and analyzed by primer extension using the CAT30 primer (Kass et al., 1997).

References

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., Franoijs, K., Stunnenberg, H. G., and Veenstra, G. J. C., 2009. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in xenopus embryos. *Developmental Cell*, **17**(3):425–434.

Crooks, G. E., Hon, G., Chandonia, J., and Brenner, S. E., 2004. WebLogo: a sequence logo generator. *Genome Research*, **14**(6):1188–1190.

Garcia, F., Lopez, F., Cano, C., and Blanco, A., 2009. FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC Bioinformatics*, **10**(1):224.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W., 2007. Quantifying similarity between motifs. *Genome Biology*, **8**(2):R24.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett,

N. M., Tagne, J., Reynolds, D. B., Yoo, J., *et al.*, 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**(7004):99–104. PMID: 15343339.

Kass, S. U., Landsberger, N., and Wolffe, A. P., 1997. DNA methylation directs a time-dependent repression of transcription initiation. *Current Biology*, **7**(3):157–165.

Mahony, S., Auron, P. E., and Benos, P. V., 2007. DNA familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, **3**(3):e61.

Owczarzy, R., You, Y., Moreira, B. G., Manthey, J. A., Huang, L., Behlke, M. A., and Walder, J. A., 2004. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, **43**(12):3537–3554.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., *et al.*, 2010. The UCSC genome browser database: update 2010. *Nucl. Acids Res.*, **38**(suppl_1):D613–619.

SantaLucia, J., 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighborthermodynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(4):1460–1465. PMID: 9465037 PMCID: 19045.

Supplemental Figures

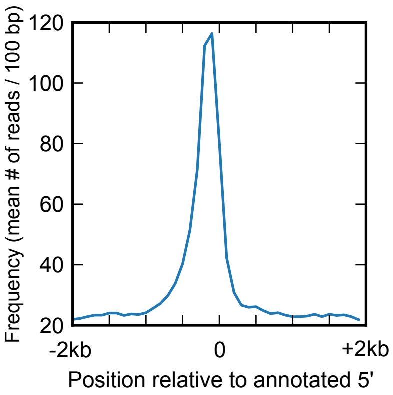


Figure S1. The distribution of all TBP ChIP-seq reads relative to the annotated 5' end of genes.

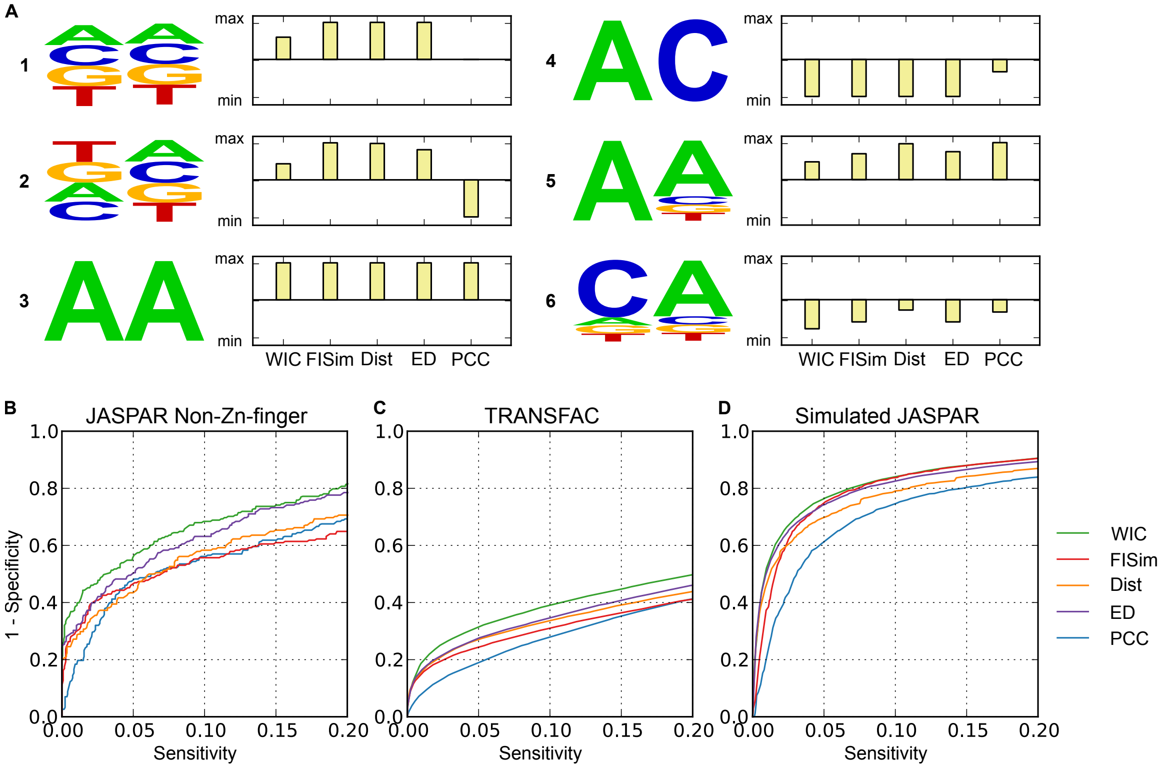
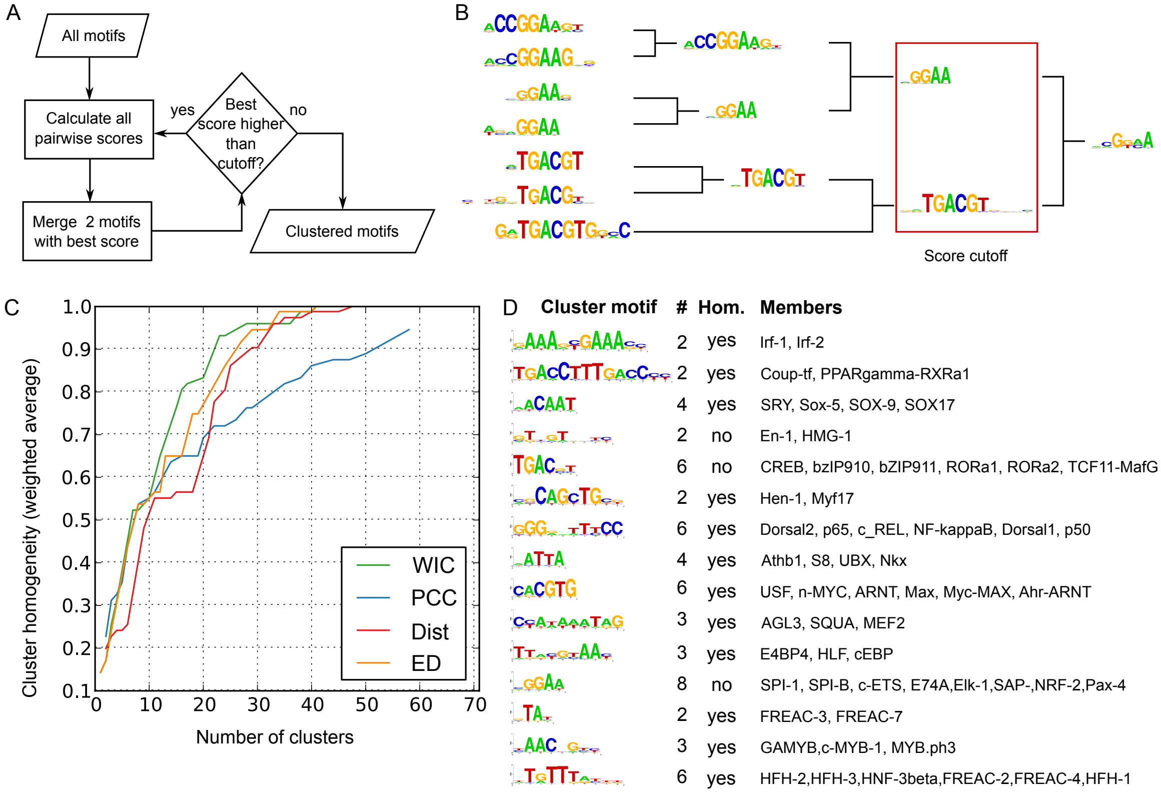


Figure S2. The WIC motif similarity score. (A) Columns of different information content are compared using different similarity scores. All scores are normalized relative to the highest- and lowest-scoring column. Example 1 and 2 are both columns with a distribution close to background, with a low information content. Most metrics give this a high score, while PCC even gives position with slightly different distributions, both close to background, the most negative score. Example 3 and 4 show comparison between columns of high information content. All metrics give the expected result for example 3 and 4, except PCC which does not score example 4 with the expected most negative value. Example 5 and 6 show two comparisons between columns of relatively high information content. (B-D) Five similarity metrics (WIC, ED, FISim, PCC and Dist) are compared using a ROC curve on three different benchmark datasets. The left panel (B) shows the performance on the 71 JASPAR non-Zn-finger dataset, the middle panel (C) shows the results for 585 TRANSFAC motifs and the right panel (D) is based on a dataset sampled from the JASPAR database as described in (Gupta et al., 2007).



Cluster motif	#	Hom.	Members
AAA GAAA	2	yes	Irf-1, Irf-2
TGACCTTTGACC	2	yes	Coup-tf, PPARgamma-RXRa1
ACAAT	4	yes	SRY, Sox-5, SOX-9, SOX17
T T	2	no	En-1, HMG-1
TGAC	6	no	CREB, bzIP910, bZIP911, RORa1, RORa2, TCF11-MafG
CAGCTG	2	yes	Hen-1, Myf17
GGG TTCC	6	yes	Dorsal2, p65, c_REL, NF-kappaB, Dorsal1, p50
ATTA	4	yes	Athb1, S8, UBX, Nkx
CACGTG	6	yes	USF, n-MYC, ARNT, Max, Myc-MAX, Ahr-ARNT
CCATAAATAG	3	yes	AGL3, SQUA, MEF2
TT AA	3	yes	E4BP4, HLF, cEBP
GGAA	8	no	SPI-1, SPI-B, c-ETS, E74A,Elk-1,SAP-,NRF-2,Pax-4
TA	2	yes	FREAC-3, FREAC-7
AAC G	3	yes	GAMYB,c-MYB-1, MYB.ph3
TGTTTA	6	yes	HFH-2,HFH-3,HNF-3beta,FREAC-2,FREAC-4,HFH-1

Figure S3. Iterative clustering of similar motifs. (A) A flowchart of the iterative clustering procedure. (B) A visualization of the merging steps during the iterative clustering. (C) The mean weighted homogeneity of all clusters (including singletons) is plotted against the number of clusters for four different similarity metrics (WIC, ED, PCC and Dist). The clustering procedure is repeated for different score thresholds and the homogeneity of the resulting clusters is calculated and averaged for all resulting clusters, weighting each cluster by the number of motifs present in the cluster. (D) Clustering results for the JASPAR non-Zn-finger motifs. The average motif of each cluster is visualized using weblogo (Crooks et al., 2004). The total number of motifs, as well as the name of all motifs included in each cluster, are shown.

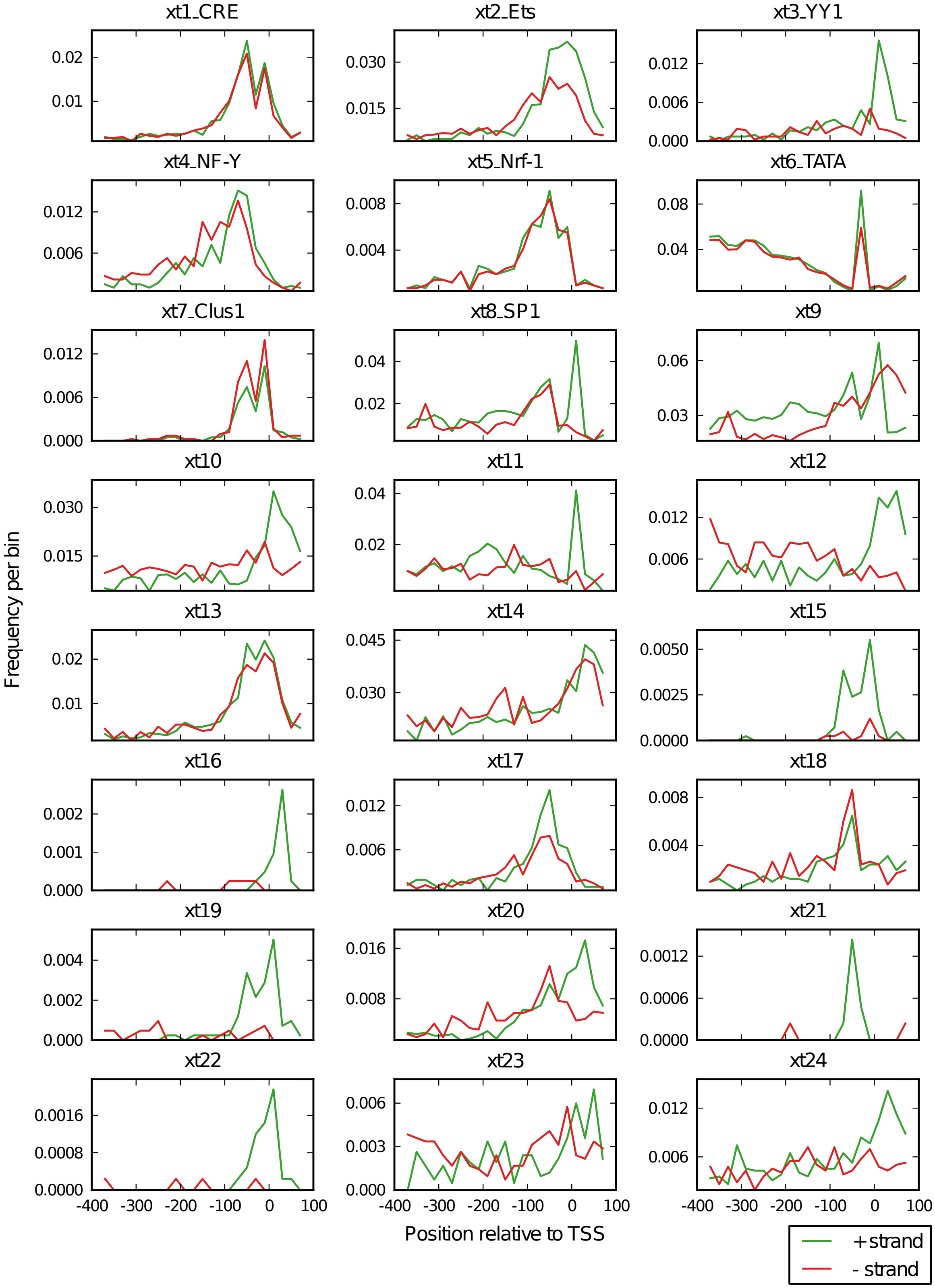


Figure S4. Differences in distribution of *Xenopus* promoter motifs (orientation). Distribution of predicted *Xenopus* promoter motifs (Fig. 3) in the + and the - orientation. The matches in the region from -400 to +100 relative to the transcription start site (TSS) are binned at 20 bp resolution for the forward (green) and the reverse (red) orientation.

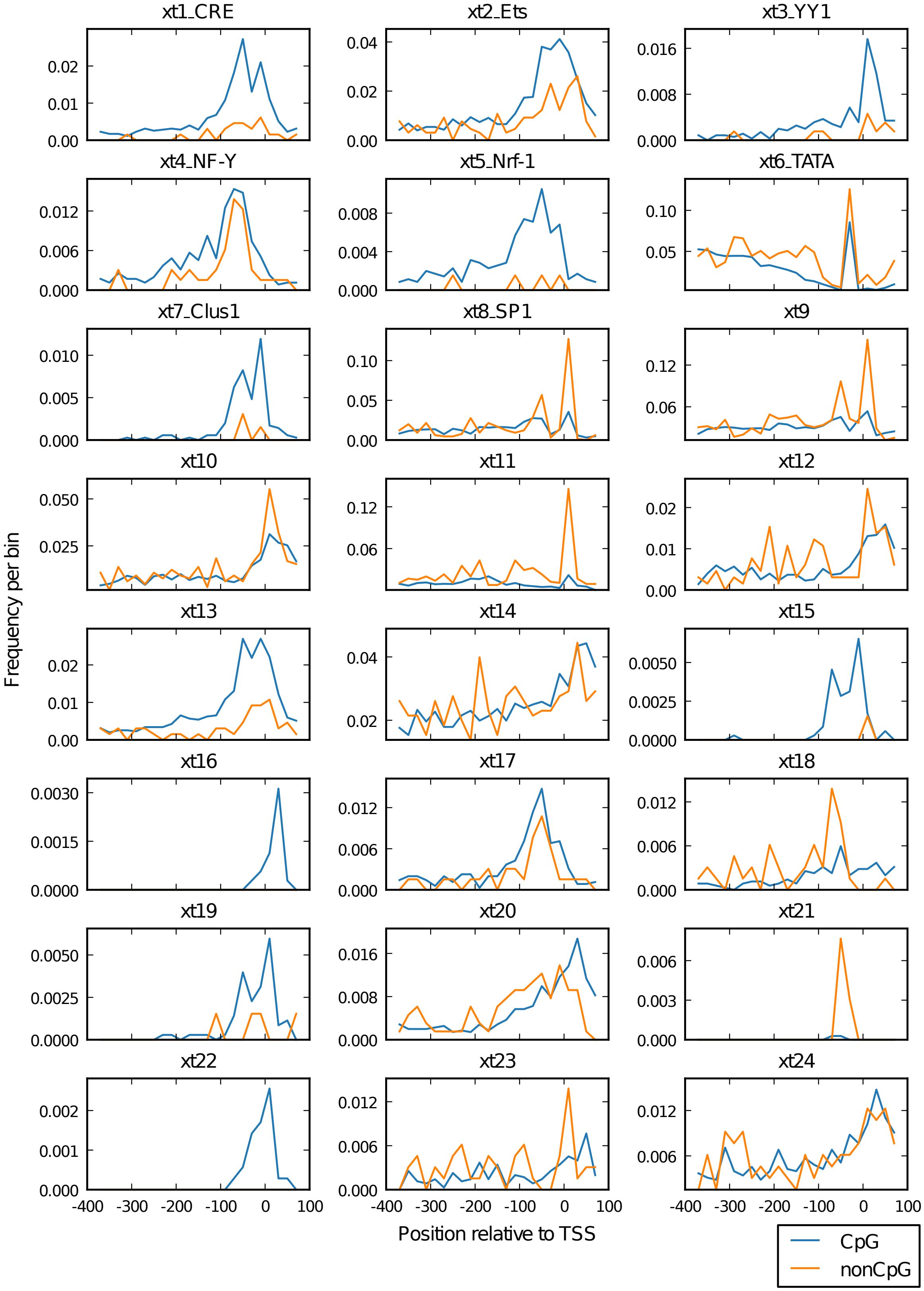


Figure S5. Differences in distribution of *Xenopus* promoter motifs (CpG-island versus non-CpG). Predicted motifs preferentially enriched in CpG-island promoters (blue) versus non-CpG promoters (orange). The matches in the region from -400 to +100 relative to the transcription start site (TSS) are binned at 20 bp resolution. Shown are all predicted motifs.

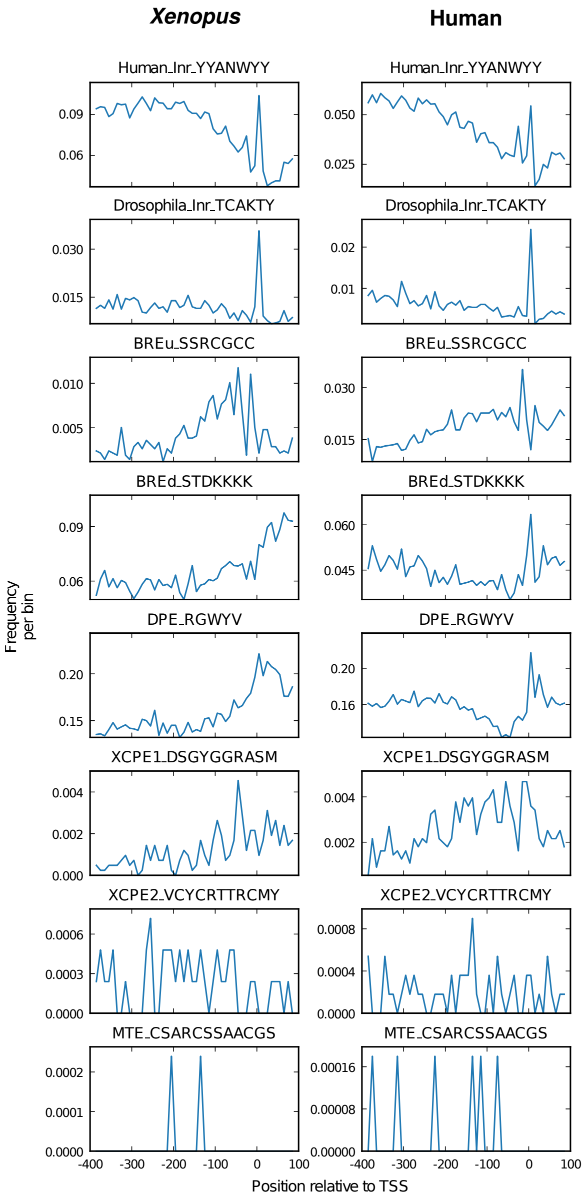


Figure S6. Distribution of known core promoter motifs in *Xenopus* and human promoters. The distribution of the positions of the motifs within a region from -400 to +100 relative to the TSS was determined by binning these positions at 10 bp resolution. The Y axis in each graph shows the frequency of the motif in each bin. The left panel shows *Xenopus* promoters, the right panel human promoters.

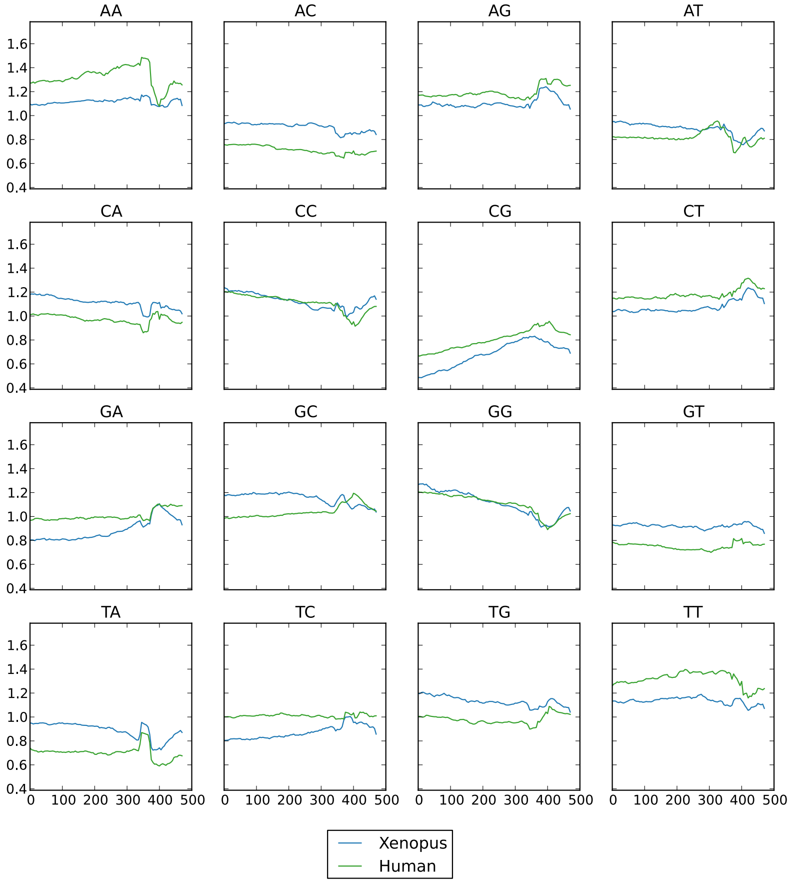


Figure S7. Normalized dinucleotide frequencies of *Xenopus* versus human promoters. The normalized dinucleotide frequency (mean observed dinucleotide frequency in a 30 bp window, divided by the expected dinucleotide frequency) is plotted for *Xenopus* (blue) and human (green) promoters.

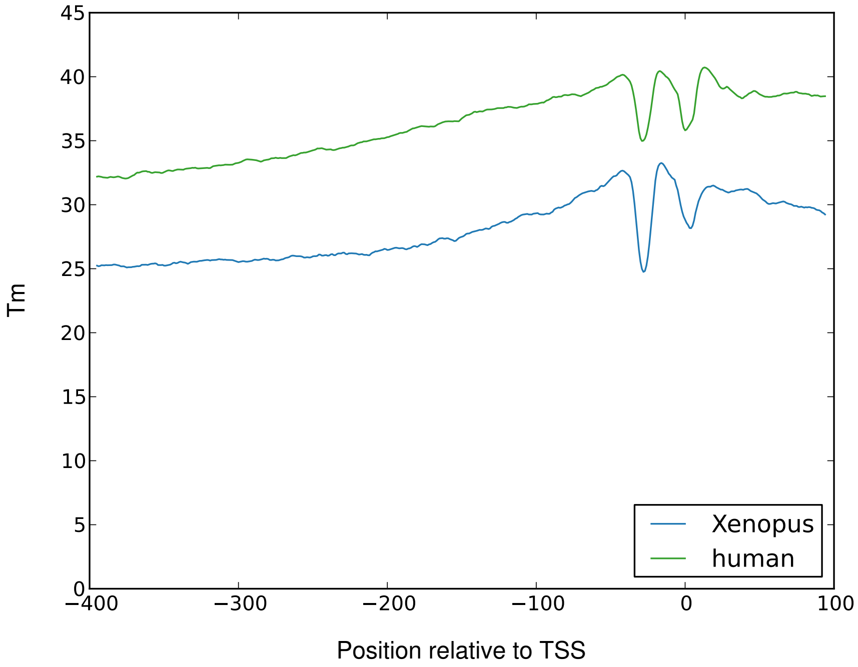


Figure S8. Melting temperature of *Xenopus* and human promoters. The predicted 11 nucleotide melting temperature (Tm) is shown for *Xenopus* (blue) and human (green) promoters.

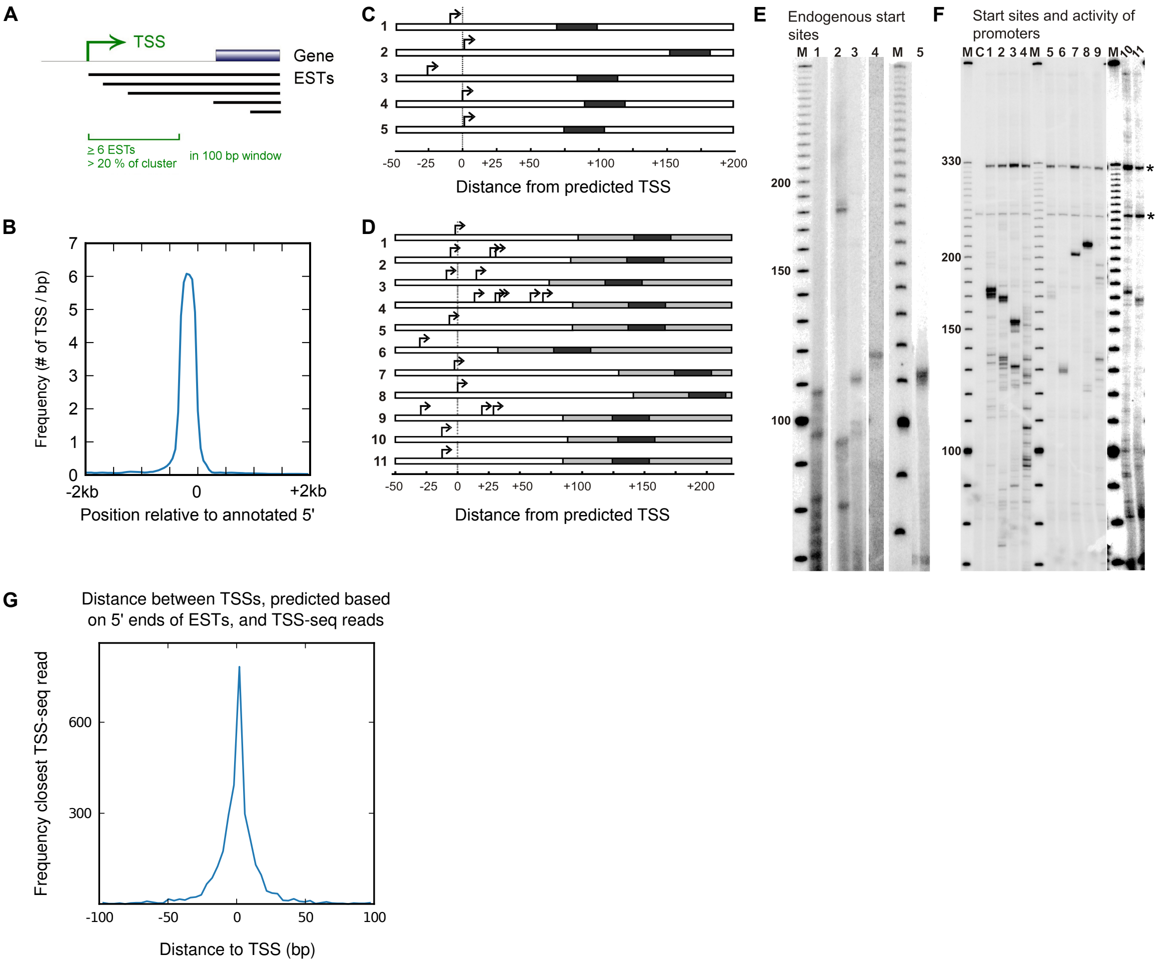


Figure S9. Experimental validation of predicted transcription start sites. (A) Illustration of TSS selection on basis of clusters of spliced ESTs. The 5' ends of a minimum number of 6 spliced ESTs must lie within 100 bp of the 5' end of the most 5' extended EST. The number of ESTs within that window must comprise at least 20% of the total number of ESTs belonging to the EST cluster. (B) The distribution of all predicted TSSs relative to the annotated 5' end of genes. (C) Comparison of experimentally determined start sites with five predicted TSSs by primer extension analysis with gene-specific primers. Each locus is represented by a horizontal bar and the shaded part on this bar shows the position of the primer used (see Fig. S9E for details). For each locus two or more primers were used to check the specificity of the primers, only promoters for which consistent results were obtained are shown. (D) Primer extension analysis of cloned promoters (for details see Fig. S9F) injected into *X. laevis* oocytes. The white portion of each bar represents the promoter sequence, the lightly shaded part depicts the plasmid DNA whereas darkly shaded part shows the position of the primers. (E) Primer extension with gene-specific primers against mRNAs corresponding to five selected TSSs (lane 1-5, for details see Supplemental Table S10A), lane M: 10 bp ladder. For each locus two or more primers were used to check the specificity of the primers, only promoters for which consistent results were obtained are shown. (F) Primer extension analysis of cloned promoters (lane 1-9, for details see Supplemental Table S10B) injected into *X. laevis* oocytes, lane C: control RNA from uninjected oocytes, *: transcripts arising from internal controls. (G) Histogram of the distance between TSSs, predicted based on the 5' ends of ESTs, and the closest TSS-seq read.

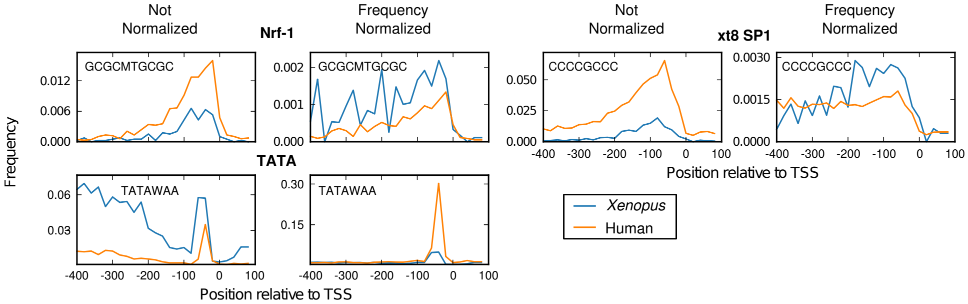


Figure S10. Distribution of motifs that have a different species-preferential enrichment before and after nucleotide frequency normalization in the EST TSS datasets.