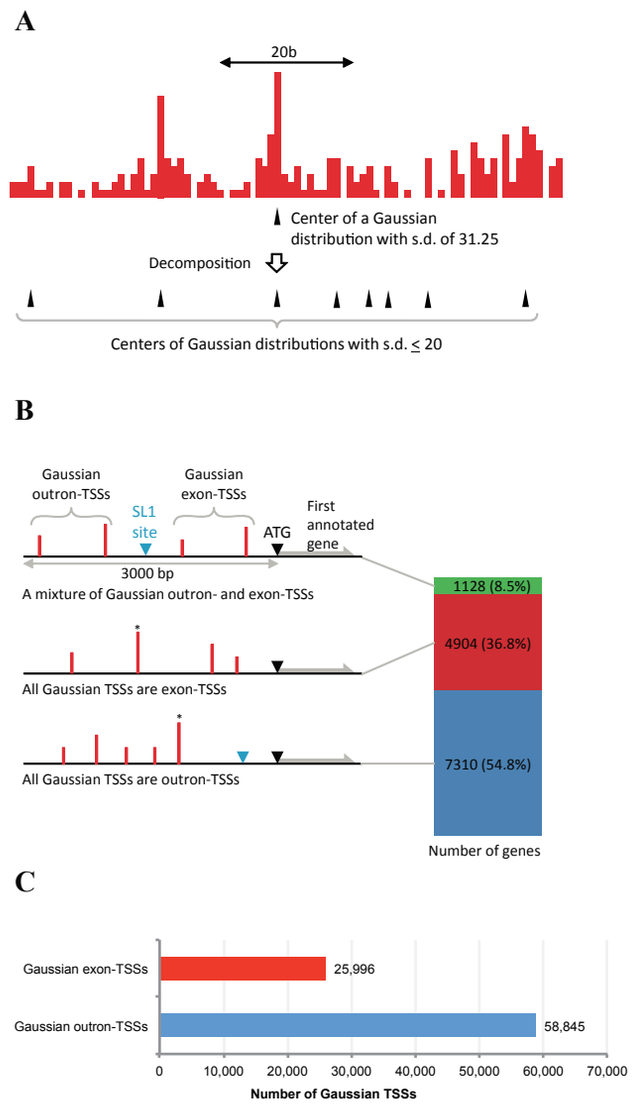
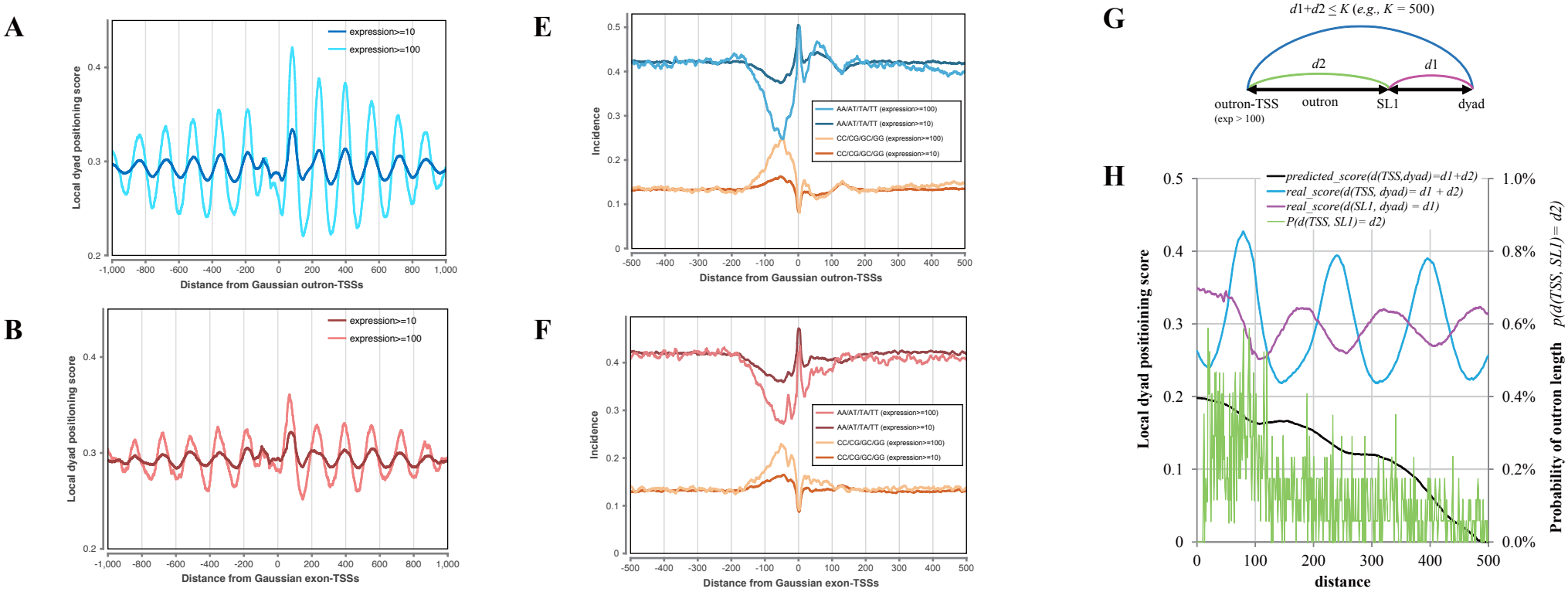
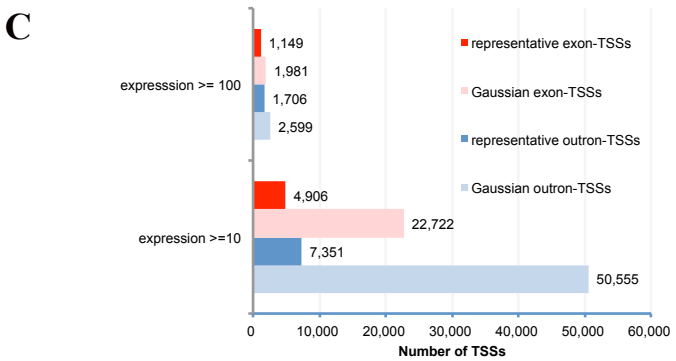


Figure S1. Gaussian distributions of TSSs. A. For reference, we decomposed Gaussian distributions with s.d. > 20bp into smaller distributions with a maximal s.d. of 20bp. An illustrative example is shown in the figure. **B.** The decomposition sometimes divided a large Gaussian distribution into multiple smaller distributions that became separated by an SL1 site, thereby increasing the number of genes that have both Gaussian outron- and exon-TSSs. Indeed, the ratio of those genes was 8.5%, which was slightly greater than 7.2% before the decomposition (Fig. 1C). **C.** The decomposition increased the numbers of Gaussian exon- and outron-TSSs in comparison with those before the decomposition (Fig. 1D).





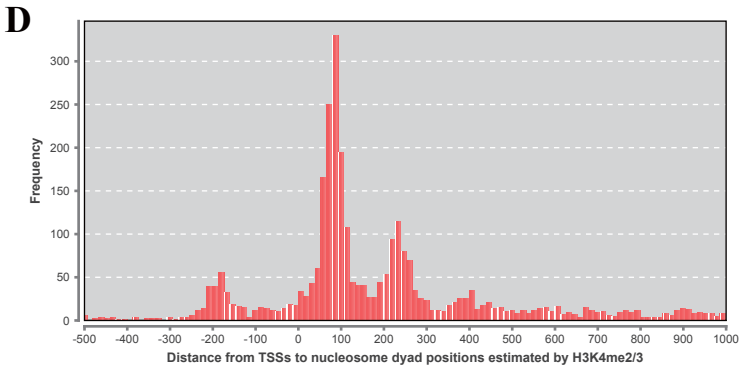


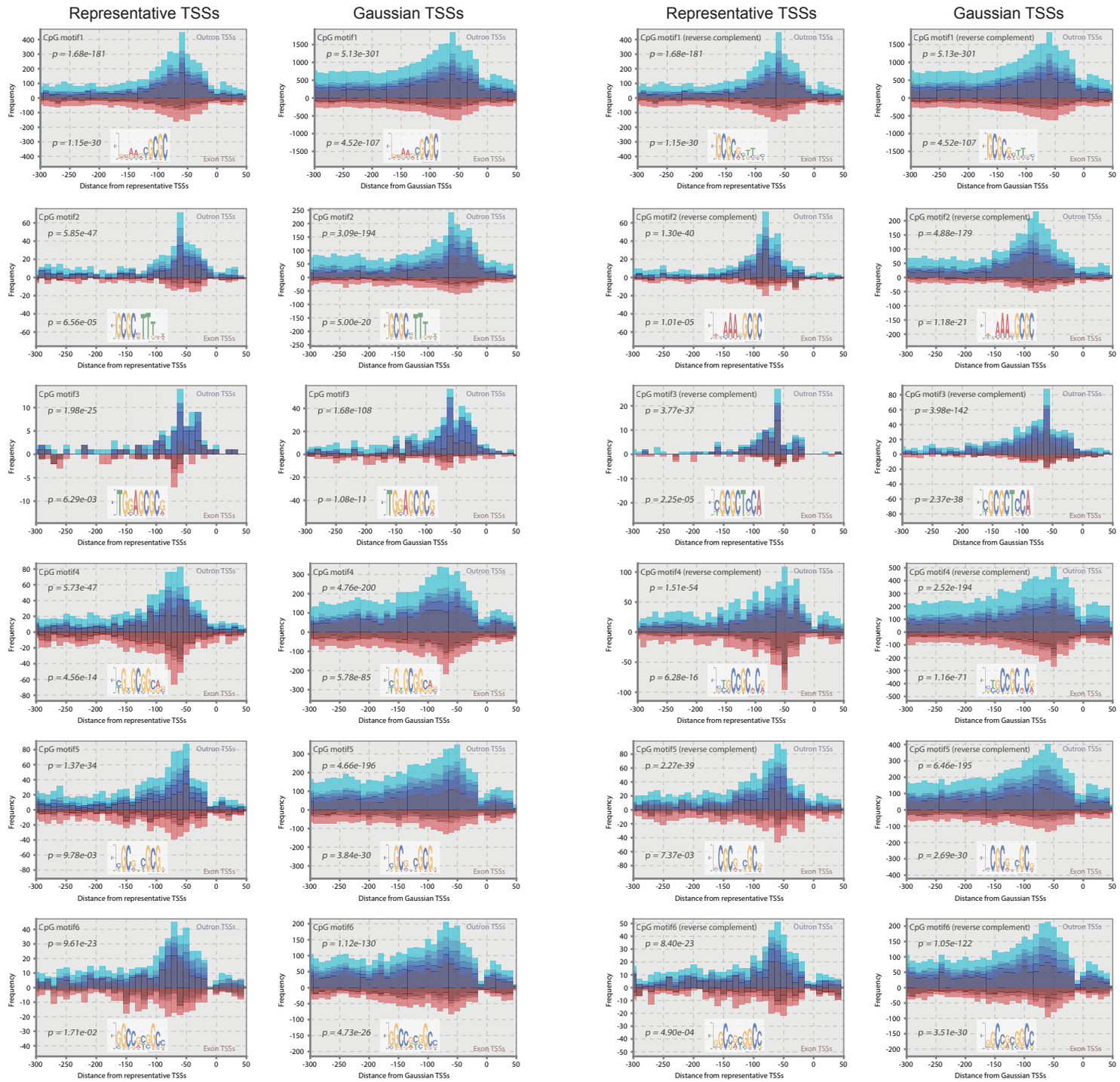
Figure S2.

A-B. The distributions of nucleosome positioning (local dyad positioning score) around Gaussian outron-TSSs (**A**) and exon-TSSs (**B**) of expression level ≥10 and ≥100. **C.** Numbers of representative/Gaussian exon-TSSs and outron-TSSs of expression level ≥10 and ≥100. **D**. Proximity of representative TSSs for 12,257 genes to their +1 nucleosome positions estimated by H3K4me2/3. The figure illustrates the frequency distribution of distances from the TSSs to their nucleosome positions. The peaks at 90 and 240 bp from TSS indicate the +1 and +2 nucleosome positions. **E-F.** A/T and C/G dinucleotide incidence rates around Gaussian outron-TSSs (**E**) and exon-TSSs (**F**) of expression level ≥10 and ≥100. **G.** We examined the hypothesis that the local dyad positioning score at *d*1-bp downstream of an SL1 site was merely a reflection of the outron length, *d*2, and the nucleosome positioning (local dyad positioning score) at (*d*1+*d*2)-bp downstream of the outron-TSS. We calculated the expected local dyad positioning score at *d*1-bp, which was the sum of the actual local dyad positioning score at (*d*1+*d*2)-bp downstream of the outron-TSS of expression ≥ 100 multiplied by the probability that the outron length is *d*2:

$$\sum_{0\le d2,d1+d2\le K} real\_score(d(TSS,dyad) = d1 + d2) \cdot Probability(d(TSS,SL1) = d2)$$

H. The expected score (black) is not similar to the actual score (pink). The light blue and green curves respectively represent the actual local dyad positioning score at (*d*1+*d*2)-bp downstream of the TSS and the probability that the outron length is *d*2.

Figure S3. Histograms of frequency distribution of CGCG variants around TSSs.



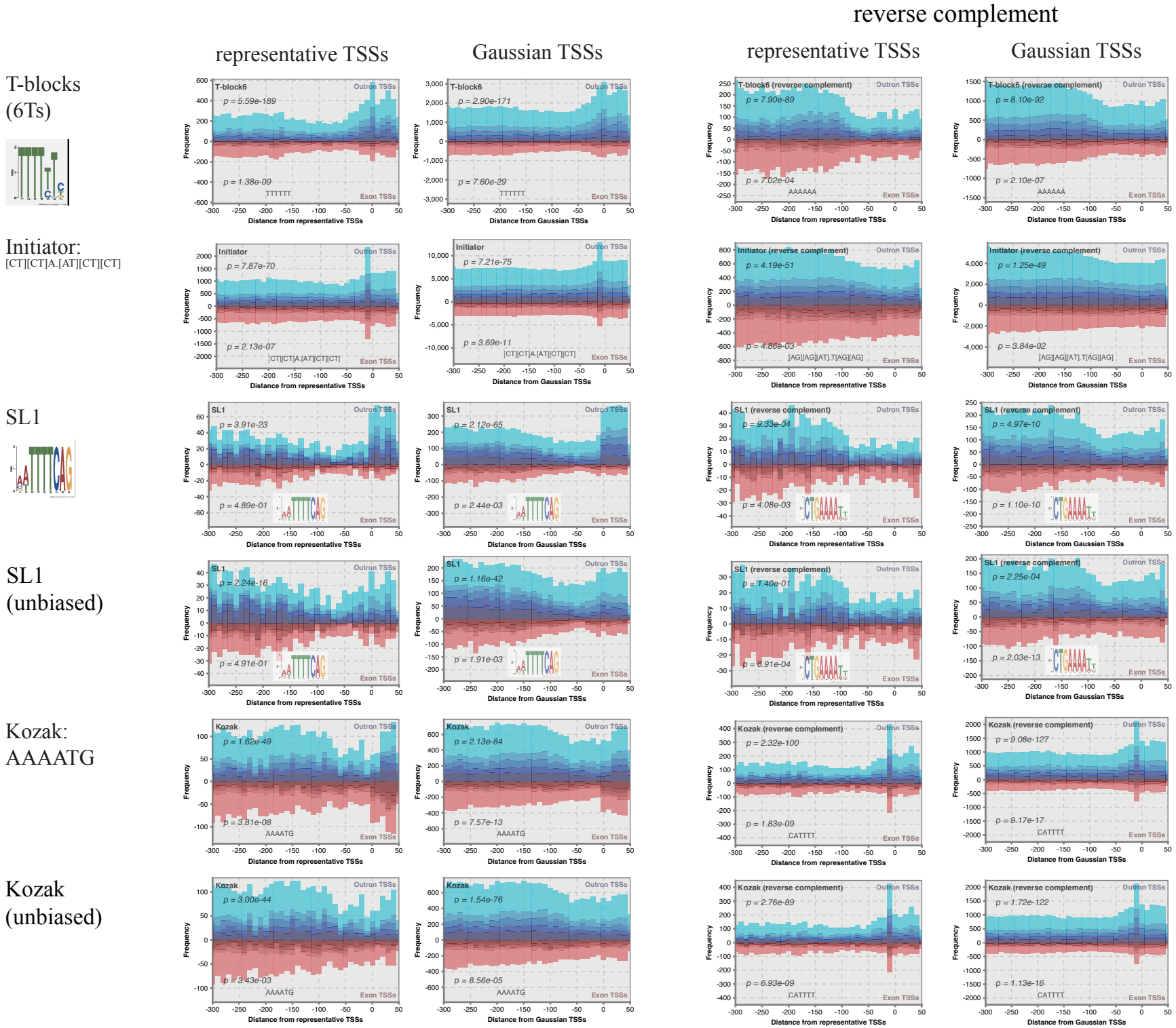


Figure S4. Histograms of frequency distribution of T-blocks (6Ts), Initiator, SL1, Kozak, and their reverse complements around TSSs. Enrichment of the SL1 motif (TTTCAG) at SL1 sites (Conrad et al. *Mol Cell Biol* **11**(4): 1921-1926, 1991) and that of a consensus translational start motif (AAAATG) at start codons (Graber et al. *RNA* **13**(9): 1409-1426, 2007) in the worm genome were previously reported (Grishkevich et al. *Genome Res* **21**(5): 707-717, 2011). Both motifs are enriched significantly around outron-TSSs, while they are not around exon-TSSs. One might be concerned that SL1 and Kozak motifs are enriched around outron-TSSs simply because these motifs respectively contain the SL1 accepter site "TTTTCAG" and the start codon "ATG" that are likely to be enriched around outron-TSSs. To cancel the effect of known SL1 sites and ATGs, we removed them from the reference genome and tested whether SL1 and Kozak motifs were still enriched in the distributions labeled with SL1 (unbiased) and Kozak (unbiased). After the removal, the *p*-values of Kozak motif around outron-TSSs remained quite low, but the *p*-value of SL1 motif around outron-TSSs increased because of the elimination of about a half of SL1 occurrences downstream of outron-TSSs.

Figure S5. Histograms of frequency distribution of known motifs (BRE, DPE, DCE, MTE and AD1) and their reverse complements around representative and Gaussian TSSs.

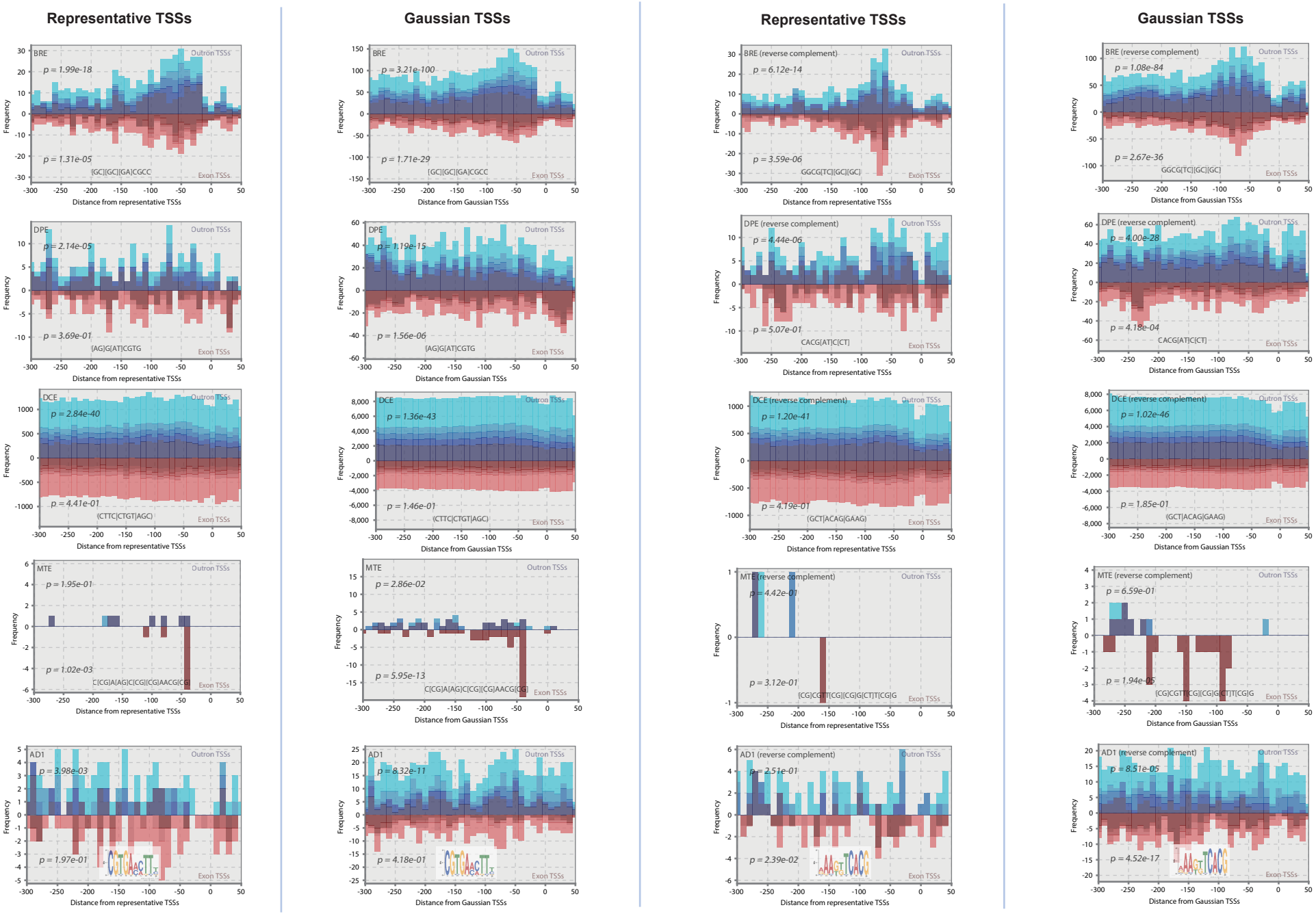
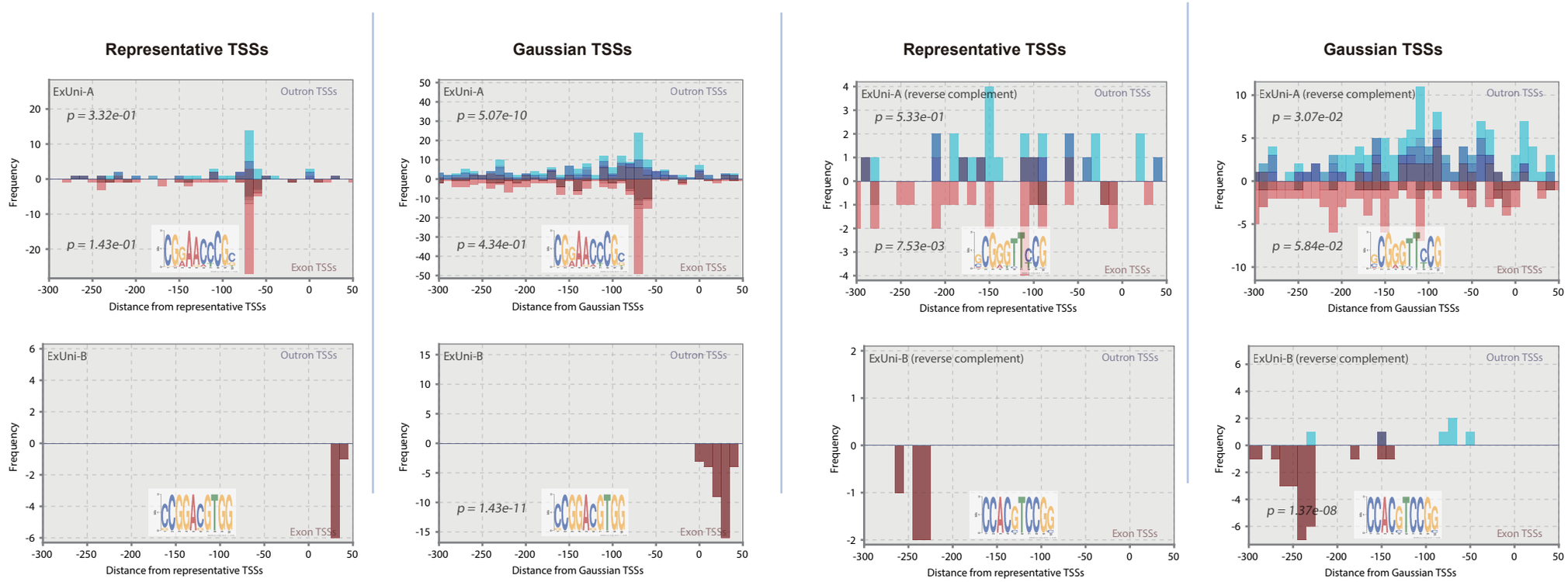


Figure S6. Histograms of frequency distribution of novel exon-TSS specific, conserved, orientation-dependent motifs.



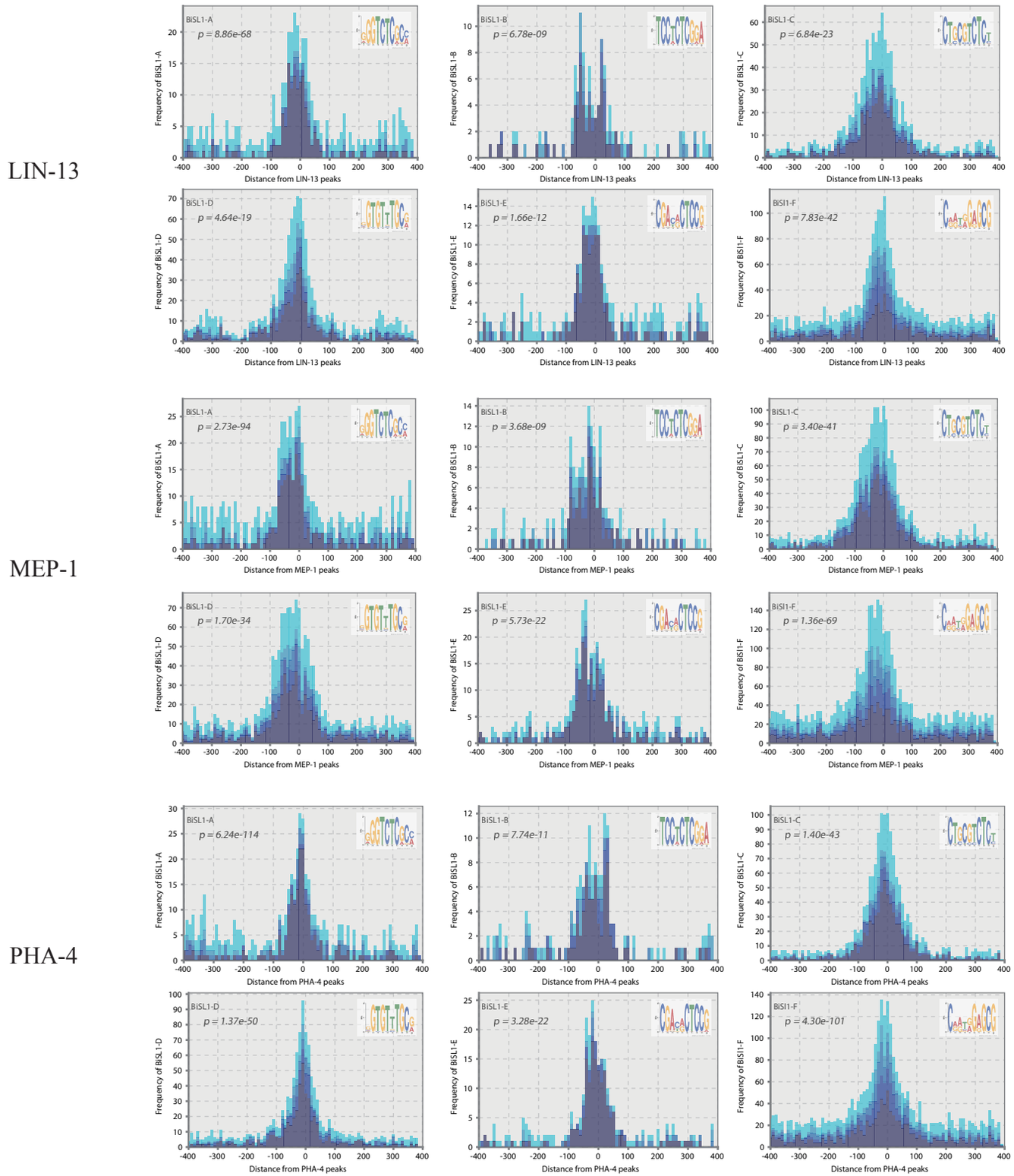


Figure S7 Frequency distribution of motif occurrences around positions associated with locally maximum binding scores of LIN-13, MEP-1 and PHA-4. The shade of color represents conservation rates (see Figure 4B and Methods). We checked the null hypothesis that two distributions of conservation rates of a motif in proximity to TF-binding sites and in the other non-promoter genomic regions were equal, by using the Wilcoxon ranksum test (see Methods). A lower p-value of a motif indicates a higher enrichment of the motif at the binding sites.

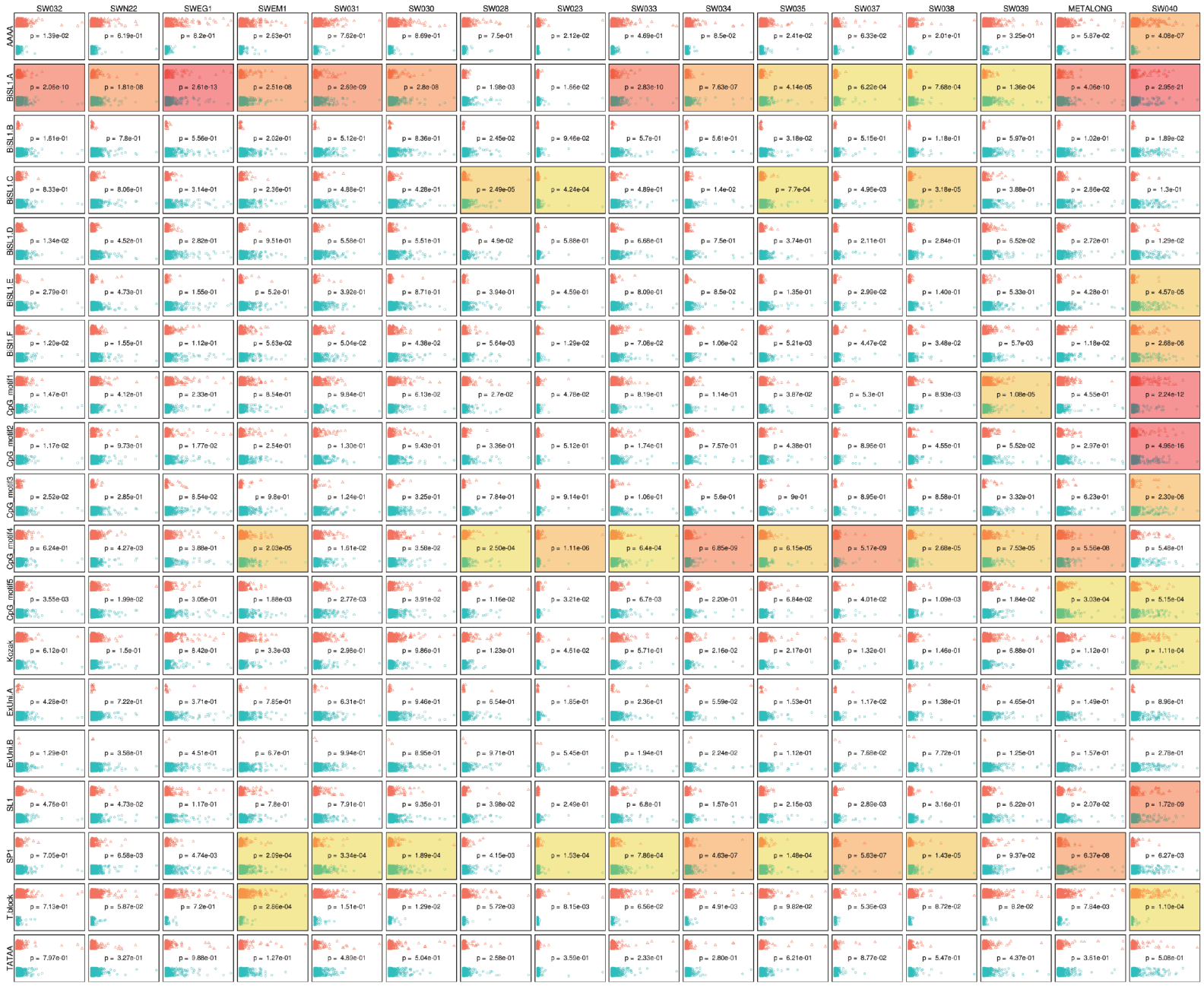
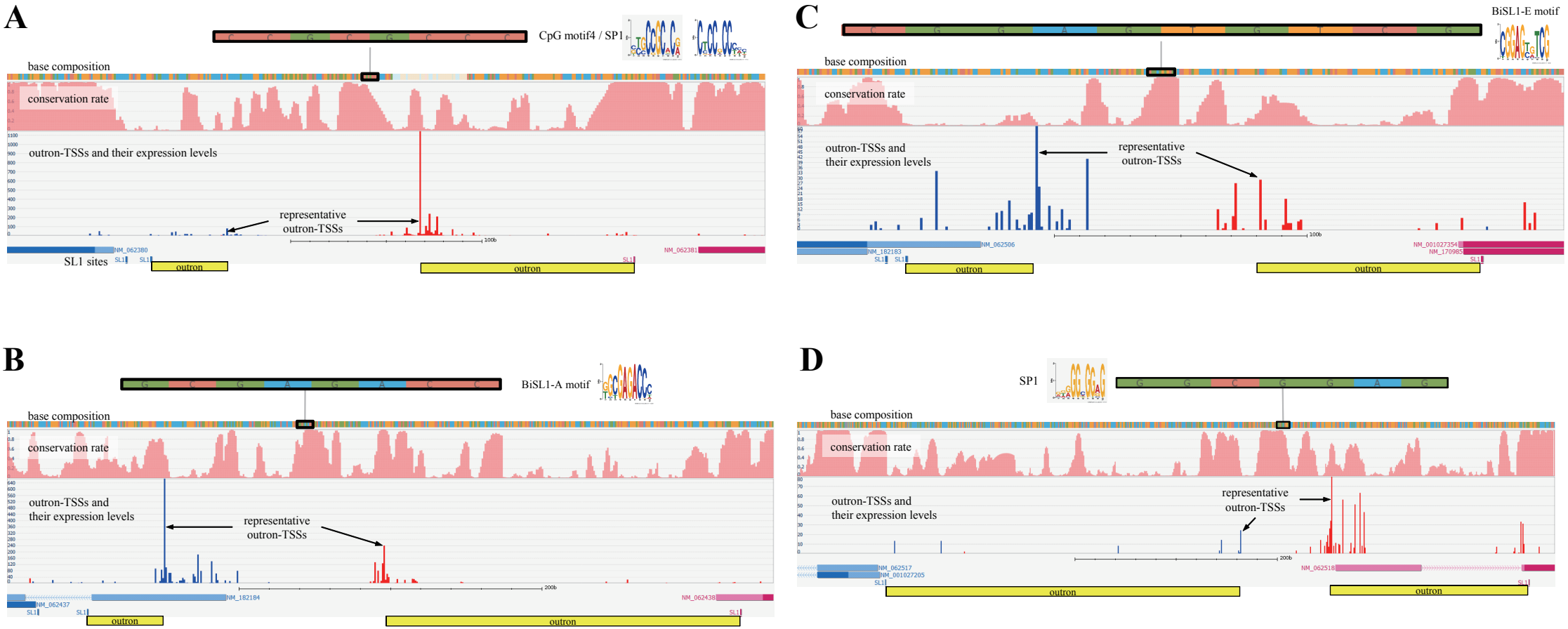


Figure S8. Motifs associated with gene expression counts from tissue-enriched cell populations. The rows present sequence motifs, and the columns show tissue types in the LogeSAGE data (Goszczynski and McGhee; Wong et al.; McKay et al. 2003; Blacque et al. 2005). Note that the matrix in Figure 6 is transposed here because of the space limitation. The table presents p-values of Wilcoxon rank-sum test for assessing the null hypothesis that two distributions of expression levels of genes associated with the presence or the absence of the upstream motif were equal in each tissue. To highlight significantly low p-values, we colored p-values $\leq 10^{-10}$ red, and p-values ranging from 10^{-10} to 10^{-3} orange. In each cell, the horizontal axis represents the expression levels of genes, upper orange points indicate genes associated with the presence of the upstream motif in the row, and lower light blue points show other genes with the absence of the motif.

Figure S9. Examples of potential bidirectional promoters.



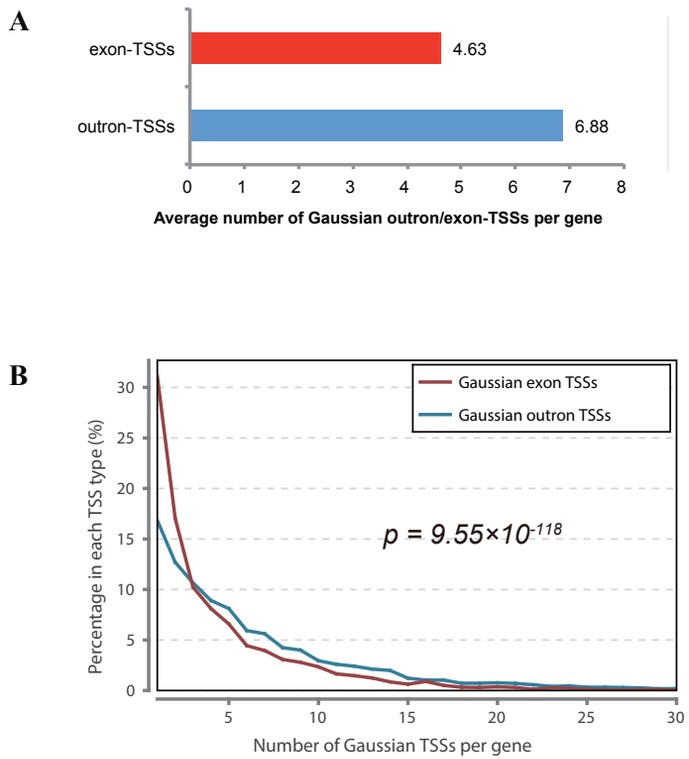


Figure S10. Numbers of Gaussian outron/exon-TSSs. **A.** Average numbers of Gaussian exon-TSS peaks and outron-TSS peaks for trans-spliced and non-trans-spliced genes. The outron-TSS average is significantly larger than the exon-TSS average according to the z-test (p-value $<10^{-116}$). **B.** Frequency distributions of numbers of exon-TSS and outron-TSS Gaussian peaks per gene. The outron-TSS distribution was significantly greater than the exon-TSS distribution according to the Wilcoxon ranksum test (p-value $< 10^{-117}$).