Supplementary Figure Legends

Supplemental Figure 1. Reproducibility of CAGE libraries and correlation with RNA sequencing data.

A) Correlation plot of expression measure (log2 (tpm)) of CTSSs between two technical replicates of the CAGE library from the Prim6 stage. Pearson correlation value of 0.93 reflects high reproducibility between two replicates. **B)** Illustrative example of CAGE reproducibility between two technical replicates of CAGE library at the *ctcf* gene promoter region. **C)** Total number of CTSSs that include selected percentages of CAGE tags (top). Total number of TCs that include selected percentages of CAGE tags (middle). Lowest tpm value of TCs that include selected percentages of CAGE tags (bottom). Horizontal dashed line marks the 5 tpm threshold that corresponds to 90-95% of the CAGE tags in a majority of the stages. Developmental stages are schematized on the bottom. Color change from blue to red indicates maternal to zygotic transition of transcriptome. **D)** Distribution of number of CTSSs for varying percentages of included CAGE tags. **E)** Distribution of number of TCs for varying percentage of included CAGE tags. **F)** Correlation of CAGE TCs at promoter regions with the transcript abundance from RNA-seq suggests the quantitative nature of CAGE. Frequencies of Pearson correlation values (r) between CAGE TCs (tpm) and RNA-seq (rpkm) for all the genes for which both cage-score and Rseq-score data were available in all four stages (N=5447). Almost half of those genes (N=2475) have r>0.7 showing a robust correlation between CAGE TCs and RNA-seq data (bottom). Illustrative heatmaps showing the most negatively-correlating (top left), non-correlating (top middle) and most positively-correlating genes (top right).

Supplemental Figure 2. Shape and distribution Tag Clusters

A). Distribution of TC interquantile widths (spacing between positions of 10^{th} and 90^{th} percentile of cumulative sum of CAGE tags along TC) at different developmental stages. Dashed vertical line shows the empirically determined

boundary (10 bp) that was used for separation of sharp and broad TCs. **B)** Distribution of TC widths for distinct classes of TCs. Interquantile width was defined as spacing between positions of the 10th and 90th percentile of cumulative sum of CAGE tags along TC. Distribution is shown separately for TCs overlapping the region -1 kb to +0.5 kb relative to the annotated TSSs (5' gene end), exons, introns, and the region -5kb to -1kb upstream of annotated TSSs and intergenic regions. **C)** Distribution of stage specific utilization of alternative TCs within (-0.5 to 0.5 kb window) the promoter region of annotated Ensembl transcripts (left) and novel RNA-seq transcripts (right).

Supplemental Figure 3. Improvements to the detection of gene 5 '-ends by CAGE.

A) Distribution of CAGE tags with respect to annotated 5'-ends of Ensembl transcripts. Black vertical line indicates the annotated 5'-ends and the arrow indicates the direction of transcription. Vertical lines along the y-axis indicate the density and position of representative CTSSs with respect to annotated 5'-ends. **B)** Alignment of H3K4me3 modified histones with reference to representative CTSSs of novel intergenic TCs show enriched histone marks lie downstream of CTSSs. Y axis indicate the average H3K4me3 signal aligned to novel intergenic TSSs at the prim6 stage. **C)** Hierarchical clustering of 459 novel intergenic TCs based on their expression level at each stage. **D**) Genome browser view of genomic datasets demonstrates the dynamics of steady state RNAs on the *ctcf* gene during zebrafish embryo development. From top to bottom, RefSeq gene model (blue), H3K4me3 sequencing, RNA sequencing, CAGE sequencing and EST tracks (black) are shown. Scales are proportionate within sequencing experiments between stages. CAGE sequencing data mapping on the sense strand (red) and the antisense strand (blue) in relation to the *ctcf* gene are shown. Arrow point at EST fragment mapping to the location of CAGE evidence of transcription on the antisense strand, indicated by arrowheads. Developmental stages are schematized on the left. Arrow indicates zygotically expressed antisense transcript start sites detected by CAGE in the first intron of *ctcf* gene.

Supplemental Figure 4. Validation and promoter types of alternative promoters

A) Utilization of sharp and broad promoters among 1383 genes, which have only one alternative promoter. **B)** Schematic of embryo segmentation used in tissue specificity analysis as described in (Gehrig et al. 2009). The brain cerebellum eye and spinal cord are indicated as domains where *isl1*-enancer activity is expected, the rest of the domains (notochord, yolk and heart) are indicated as ectopic domains. **C)** Bar plot, representing the activity of the reporter in both the *isl1*-enhancer and ectopic domains, in embryos injected with combinations of enhancer-core promoter constructs. The p-values (Fisher-test) on top of each promoter type group indicate the statistical significance of the difference in activity from the control promoter fragments. Abbreviations: cereb, cerebellum

Supplemental Figure 5. Properties of non-canonical initiator containing genes.

A) Gene ontology analyses of genes with CTC-initiator motif. Only GO terms which are significantly enriched (p-value ≤ 0.05) are shown. The ribosomal GO categories are highlighted in grey. **B)** Utilization of initiator dinucleotide defined by all CTSSs within a tag cluster with dominant AA initiator dinucleotide. The values are percentage from the total tpm of the TC for each dinucleotide for a given gene.

Supplemental Figure 6. Developmental dynamics and distribution of intragenic CAGE signals.

A) CAGE tags unmapped to the genome, mapped to cDNA exon-exon junctions. Global distribution of the total number of exonic tags starting at particular position (up to -100 bp upstream) relative to exon end (position 0) is represented. First and last exons were excluded from analyses. Blue (Fertilized egg) and orange (Prim6) curves show the distribution of exonic CAGE tags, which map to the genome. Black curve represents the distribution of tags from Prim6 stage, which failed to map to the genome but aligned to cDNA. **B)** Hierarchical clustering analysis of genes with exonic tags with their promoter tags based on their expression level indicated by tpm**. C**) Pearson correlation analysis on

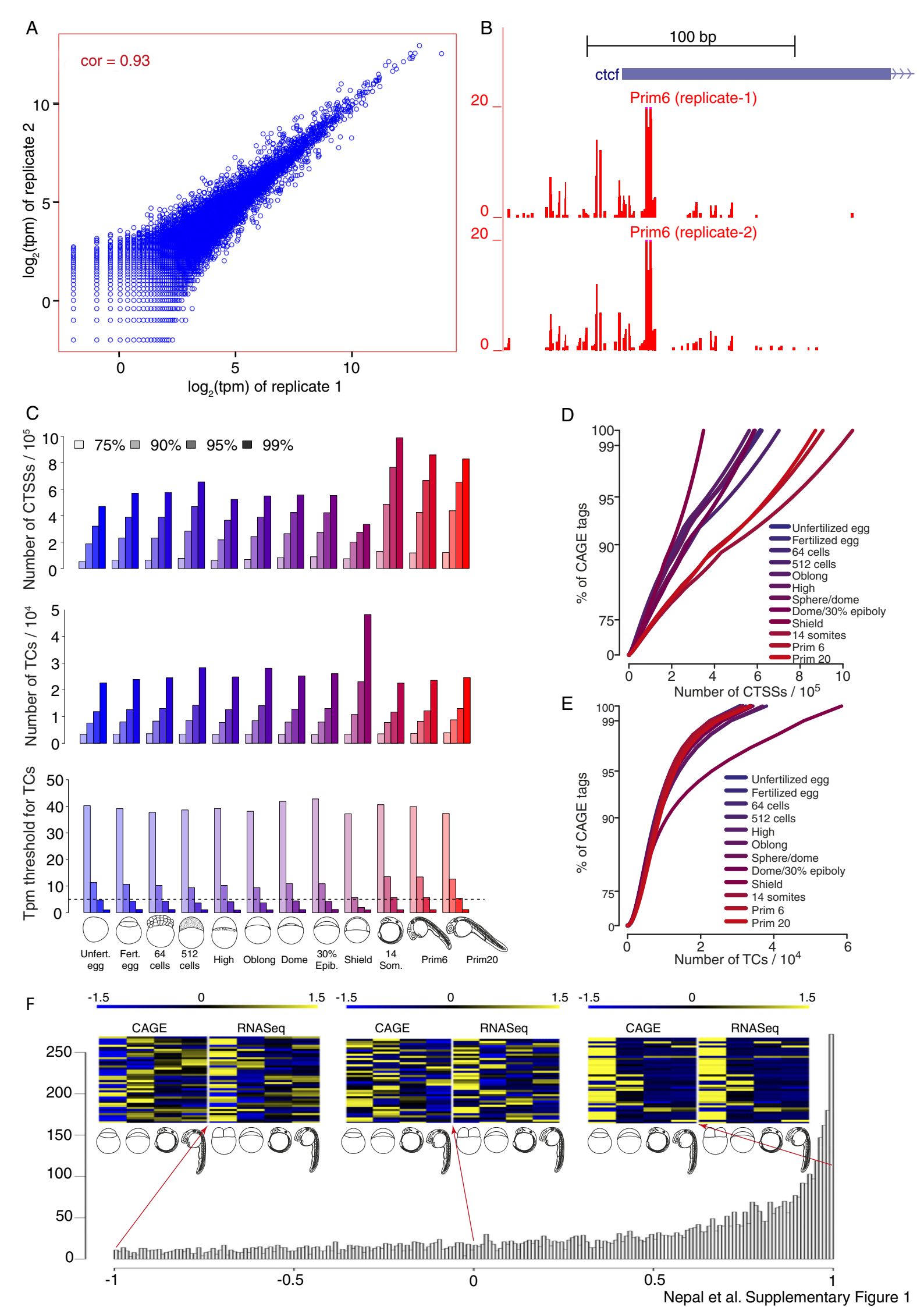
expression level of TCs, across all developmental stages, from exonic TCs versus host genes 5'-end TCs. N represents the number of genes analyzed. Enriched GO terms for genes with TCs in **D)** exons, **E)** 3'UTR exons. The heat map represents the –log (p-values) of significantly enriched GO terms, where the p-value is corrected (False Discovery Rate, FDR) for enriched terms.

Supplemental Figure 7. Developmental dynamics and distribution of intronic CAGE signals

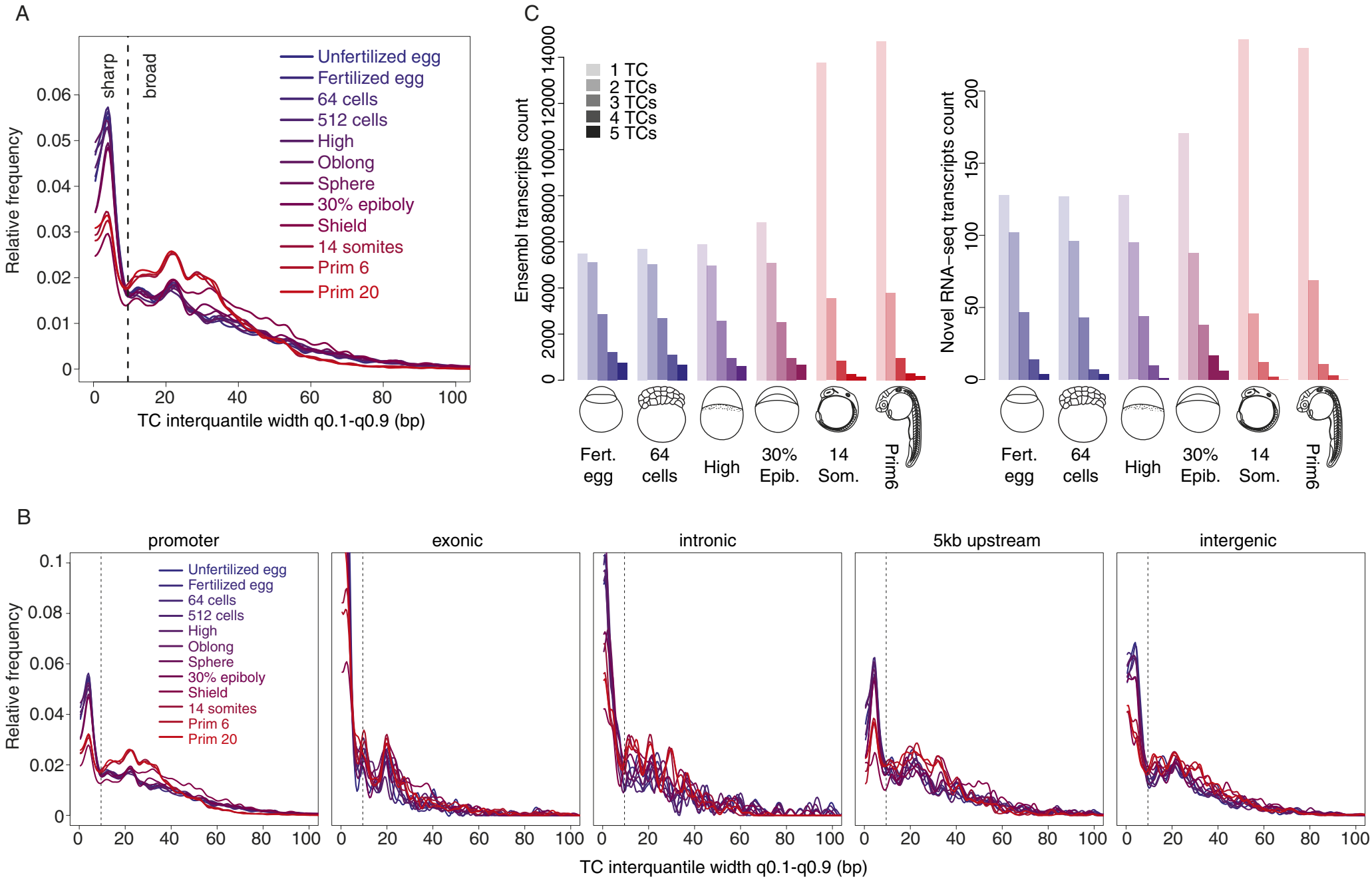
A) Illustrative examples of intra-intronic expression dynamics with respect to the host gene 5'-end activity. X-axis indicates the twelve developmental time points in increasing order of developmental time points. Y axis indicates the log2(tpm) measure of promoter and intronic expression. **B)** Pearson correlation analysis on expression level (tpm) of TCs, across all developmental stages, from introns versus the host gene 5'-end associated and exonic TCs. N represents number of genes analyzed. **C,D**) The heat map represents the –log (p-values) of significantly enriched GO terms, where p-value is corrected (FDR) for enriched terms. Enriched GO terms for genes with intra-intronic TCs **(C)** and TCs at the 3'-end of an intron **(D)**.

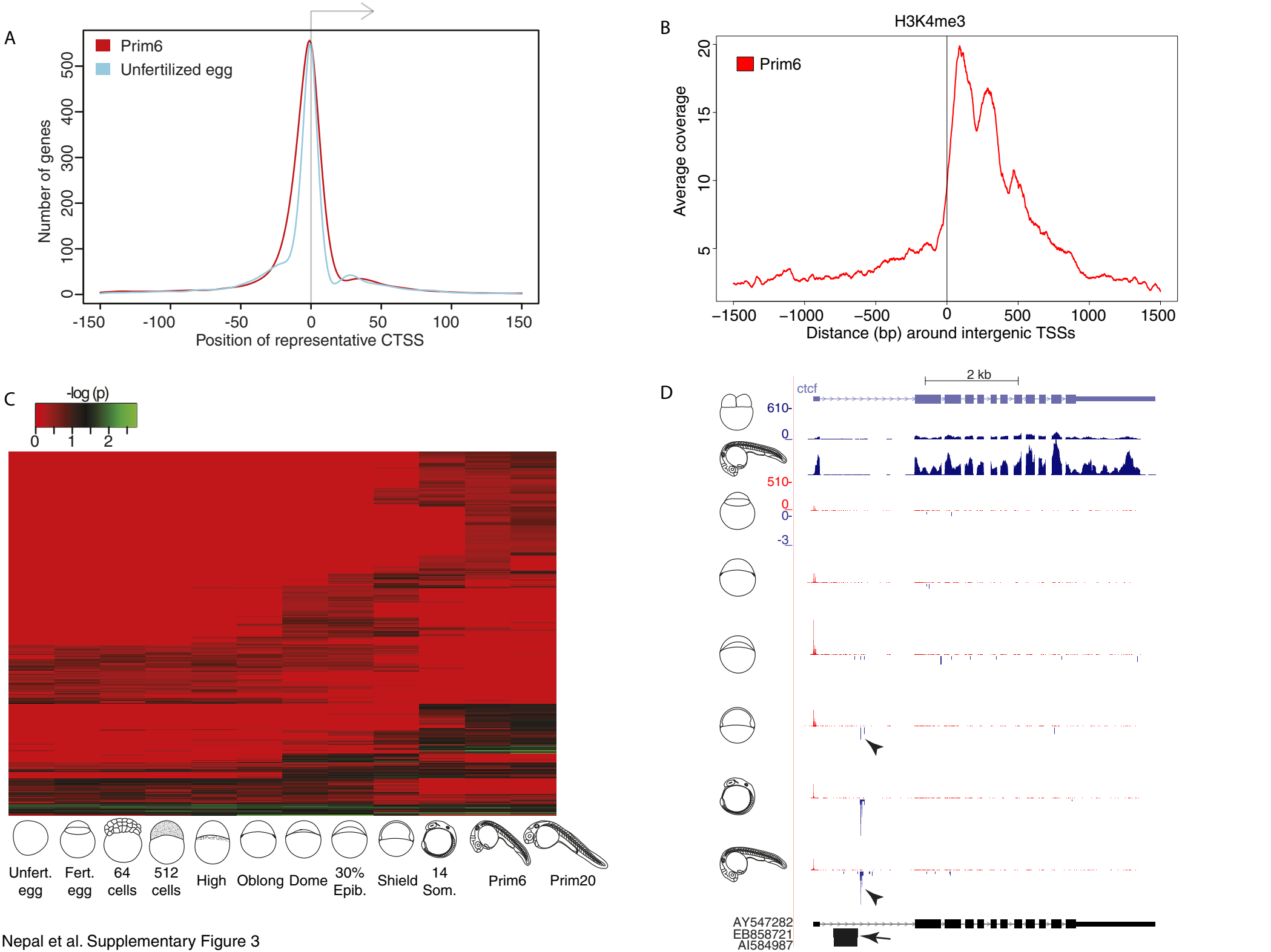
Supplemental Figure 8. Initiation site properties of intragenic CAGE signals

A) Sequence logos around the representative CTSSs of *MALAT1* intragenic TCs shows an enriched G-rich motif, which is conserved between human and zebrafish (among various stages). **B)** Sequence logos associated with exonic CTSSs, when a G at the first position of tags was either retained (right) or removed (left). This indicates no experimental G-bias of the CAGE method had an influence on the novel G-rich pattern discovered. **C)** Aggregation plot of H3K4me3 modified histones around representative CTSSs of exonic and intronic and their gene 5'-end promoter associated CAGE tags.

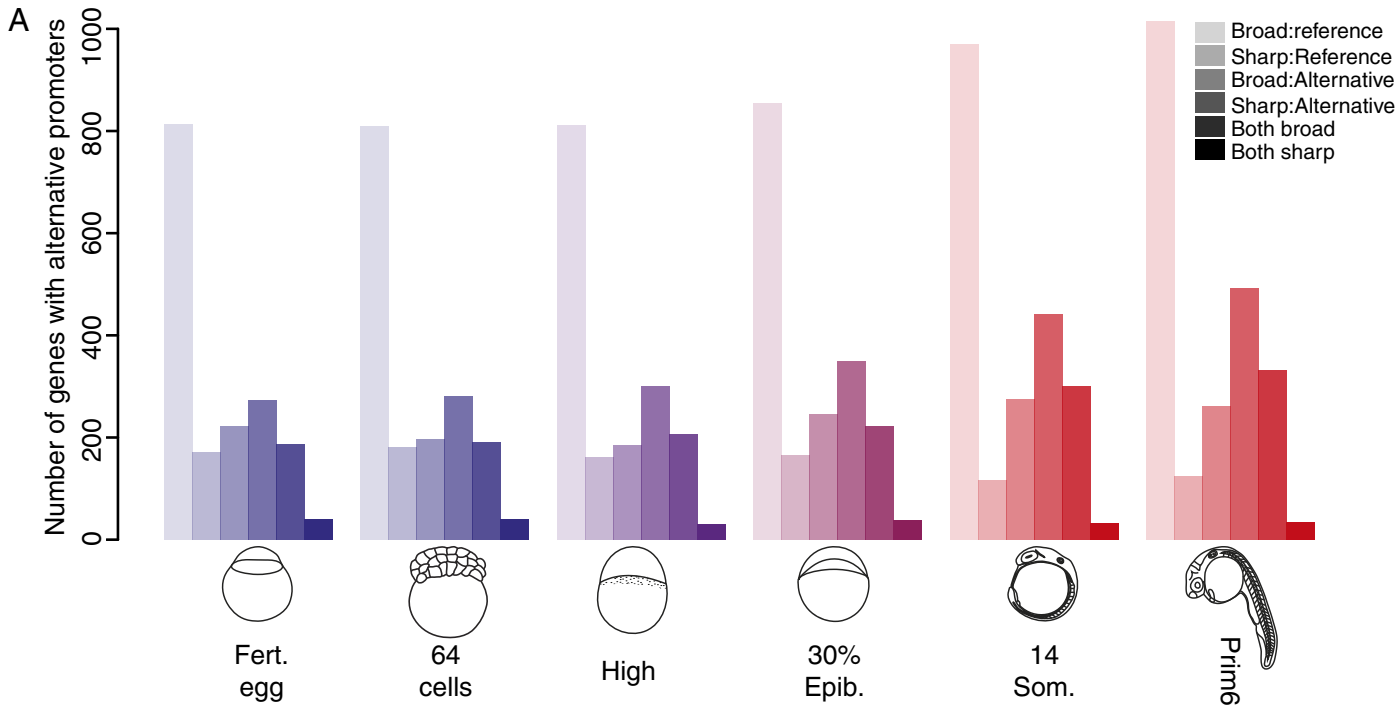


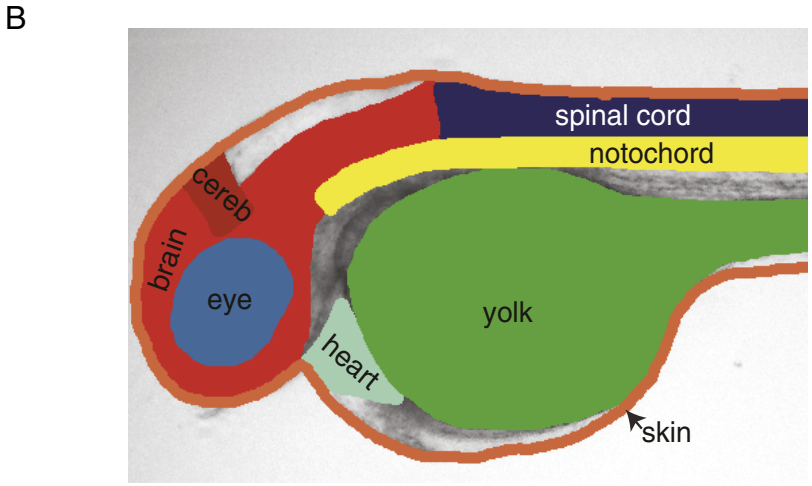
Nepal et al. Supplementary Figure 1

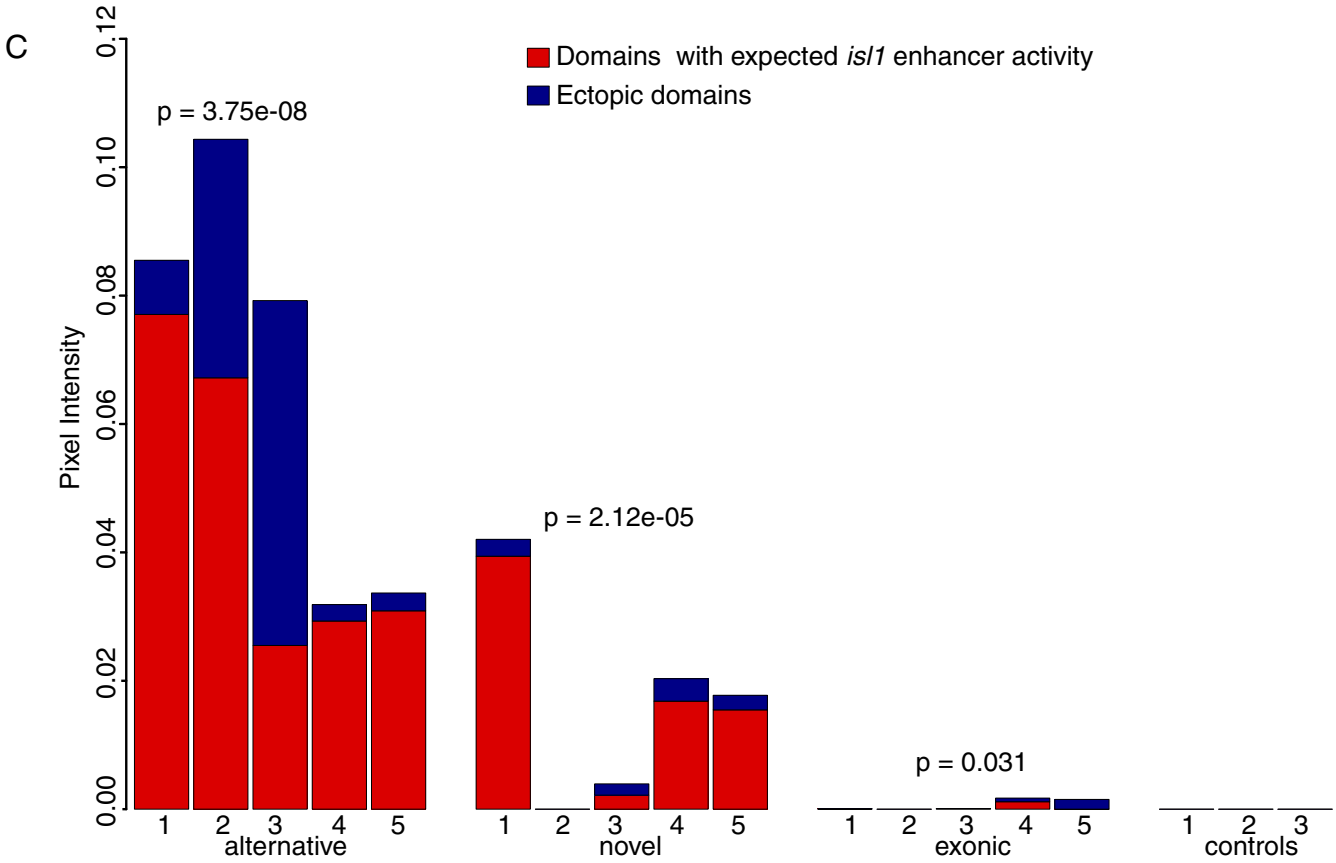

A
sharp
broad
Unfertilized egg
Fertilized egg
64 cells
512 cells
High
Oblong
Sphere
30% epiboly
Shield
14 somites
Prim 6
Prim 20
Relative frequency
TC interquantile width q0.1-q0.9 (bp)
C
1 TC
2 TCs
3 TCs
4 TCs
5 TCs
Ensembl transcripts count
Novel RNA−seq transcripts count
Fert. egg
64 cells
High
30% Epib.
14 Som.
Prim6
B
promoter
exonic
intronic
5kb upstream
intergenic
TC interquantile width q0.1-q0.9 (bp)



Nepal et al. Supplementary Figure 3

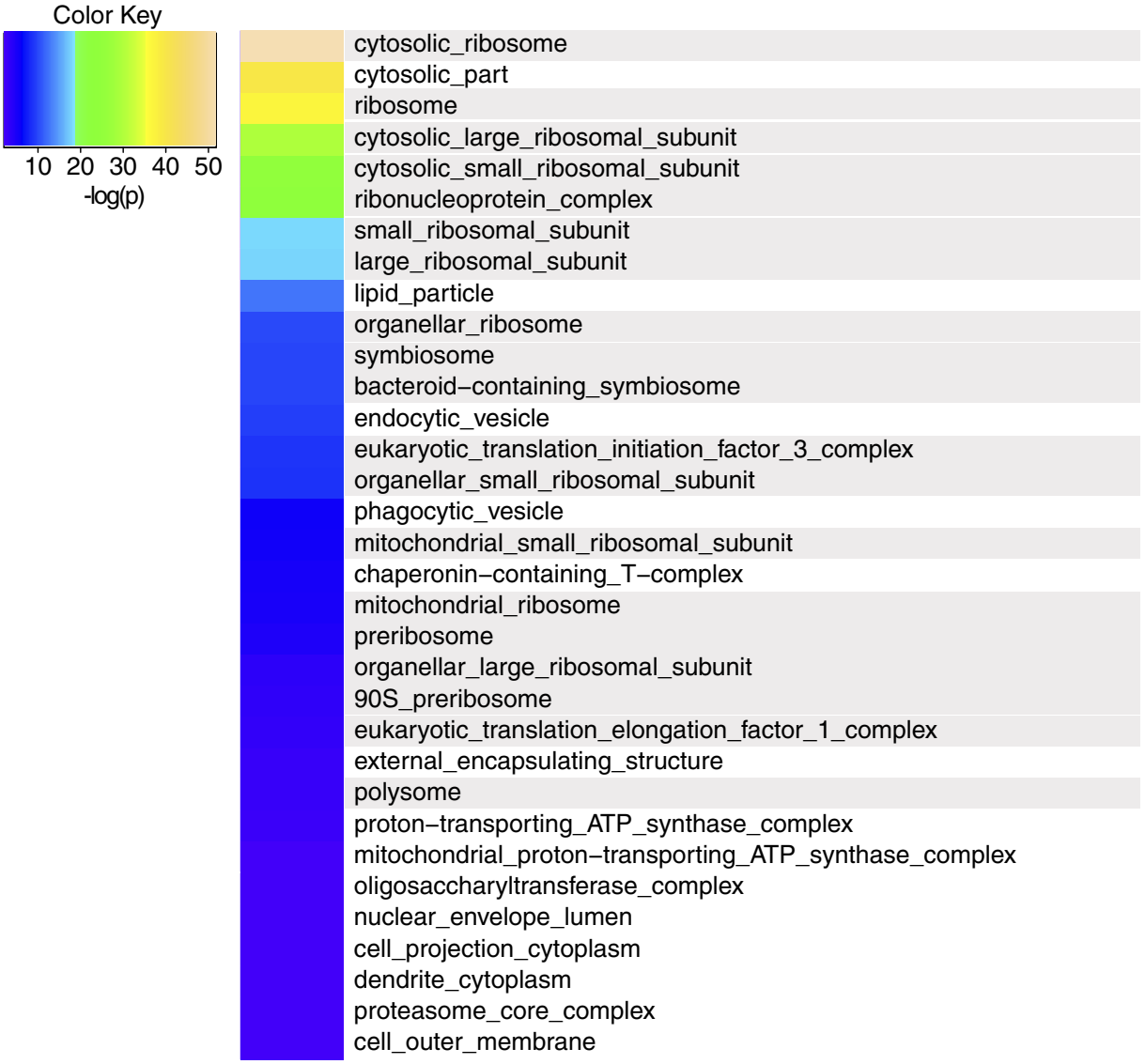




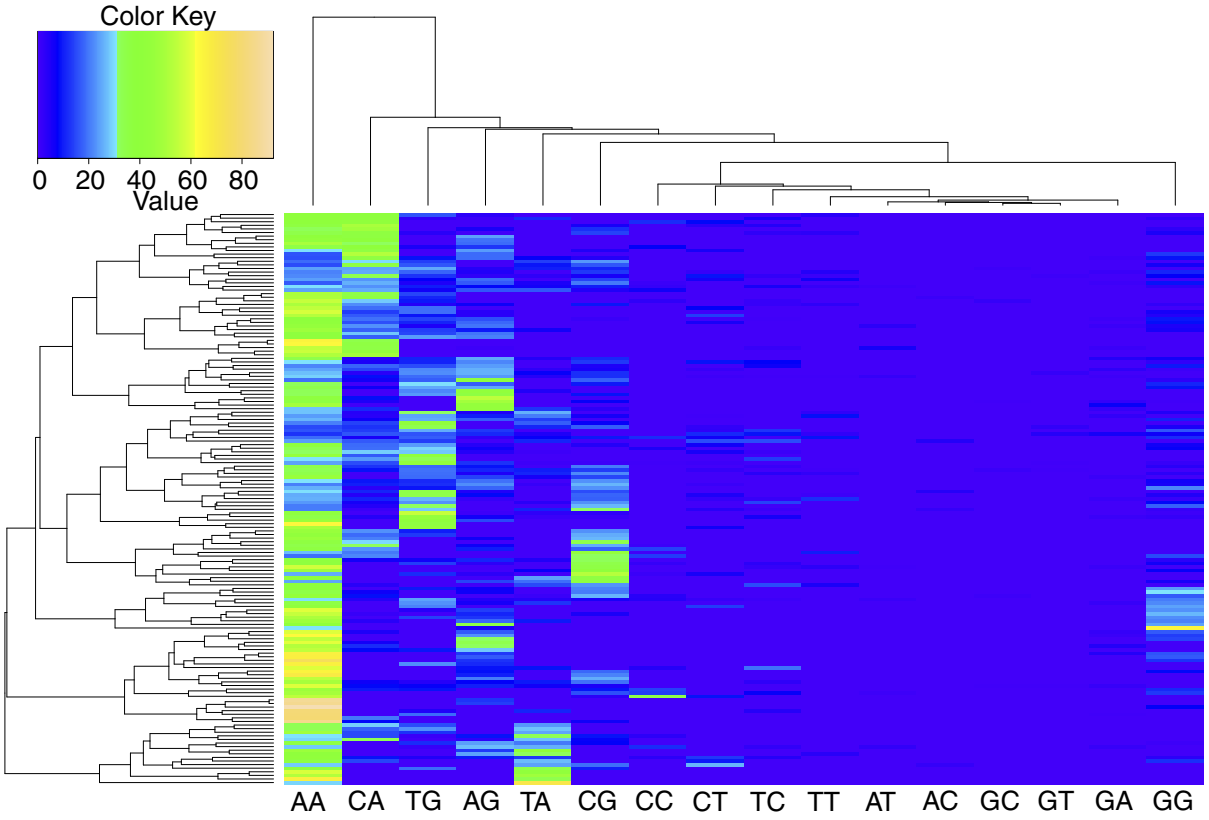


Nepal et al. Supplementary Figure 4

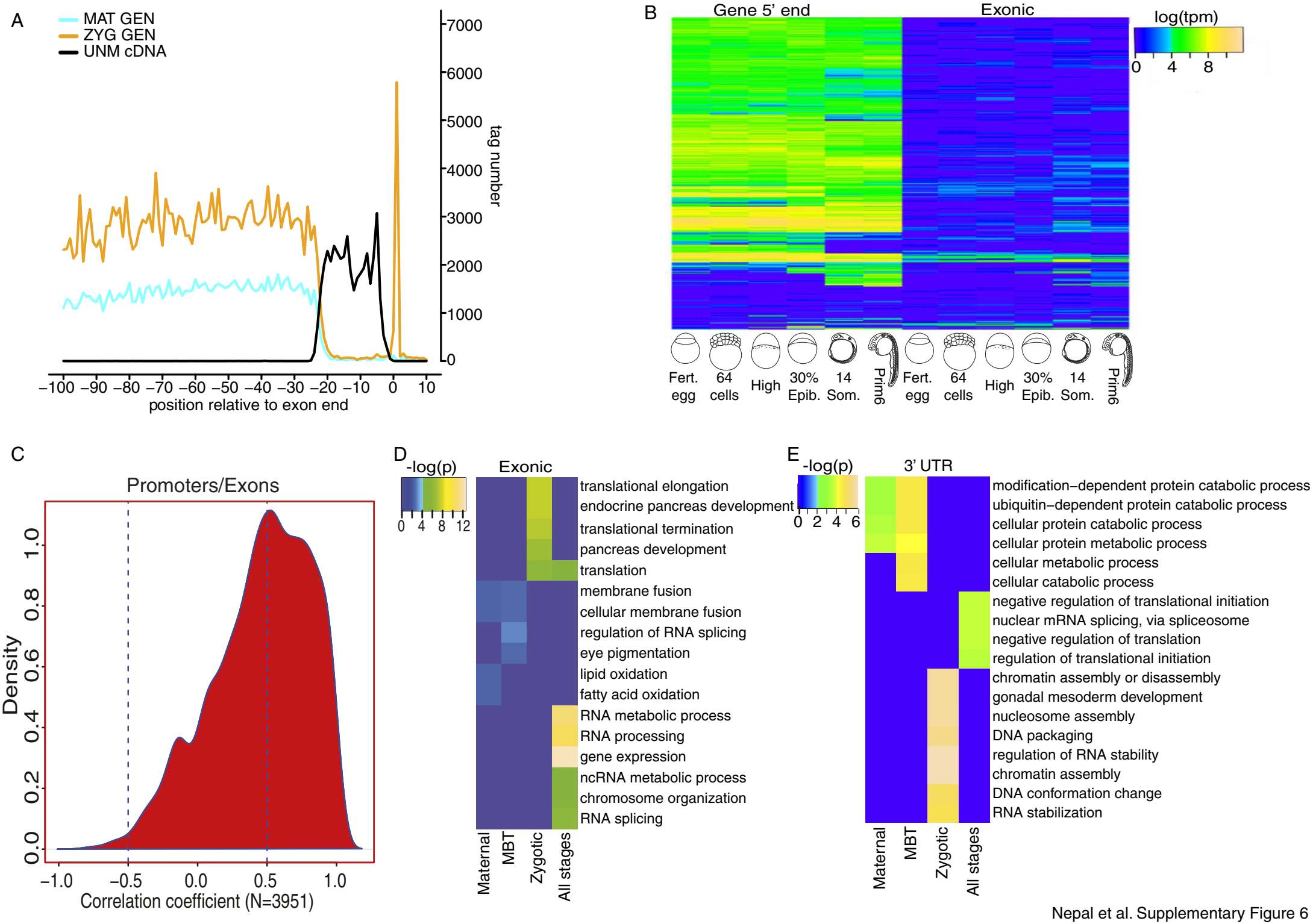
A



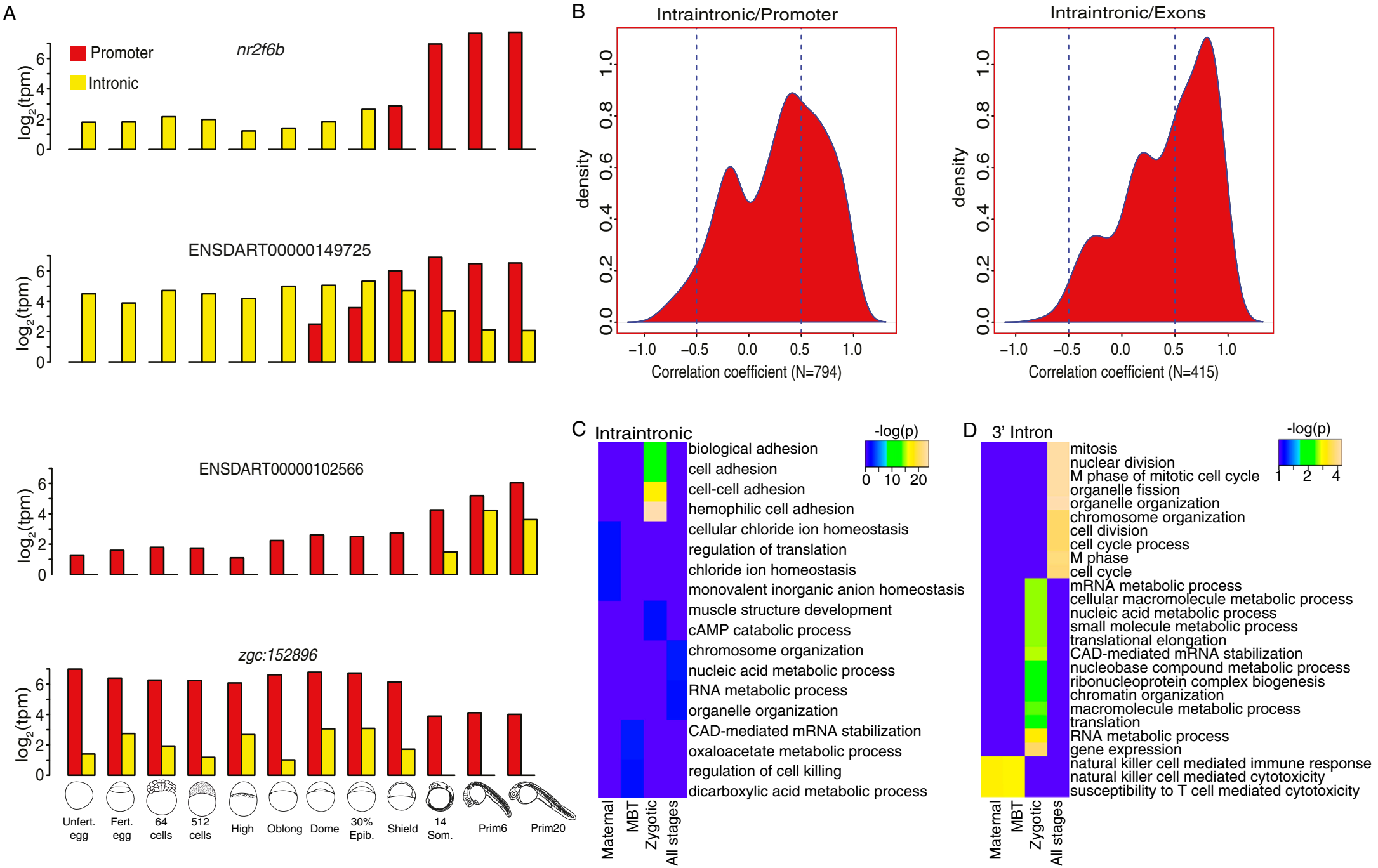
B



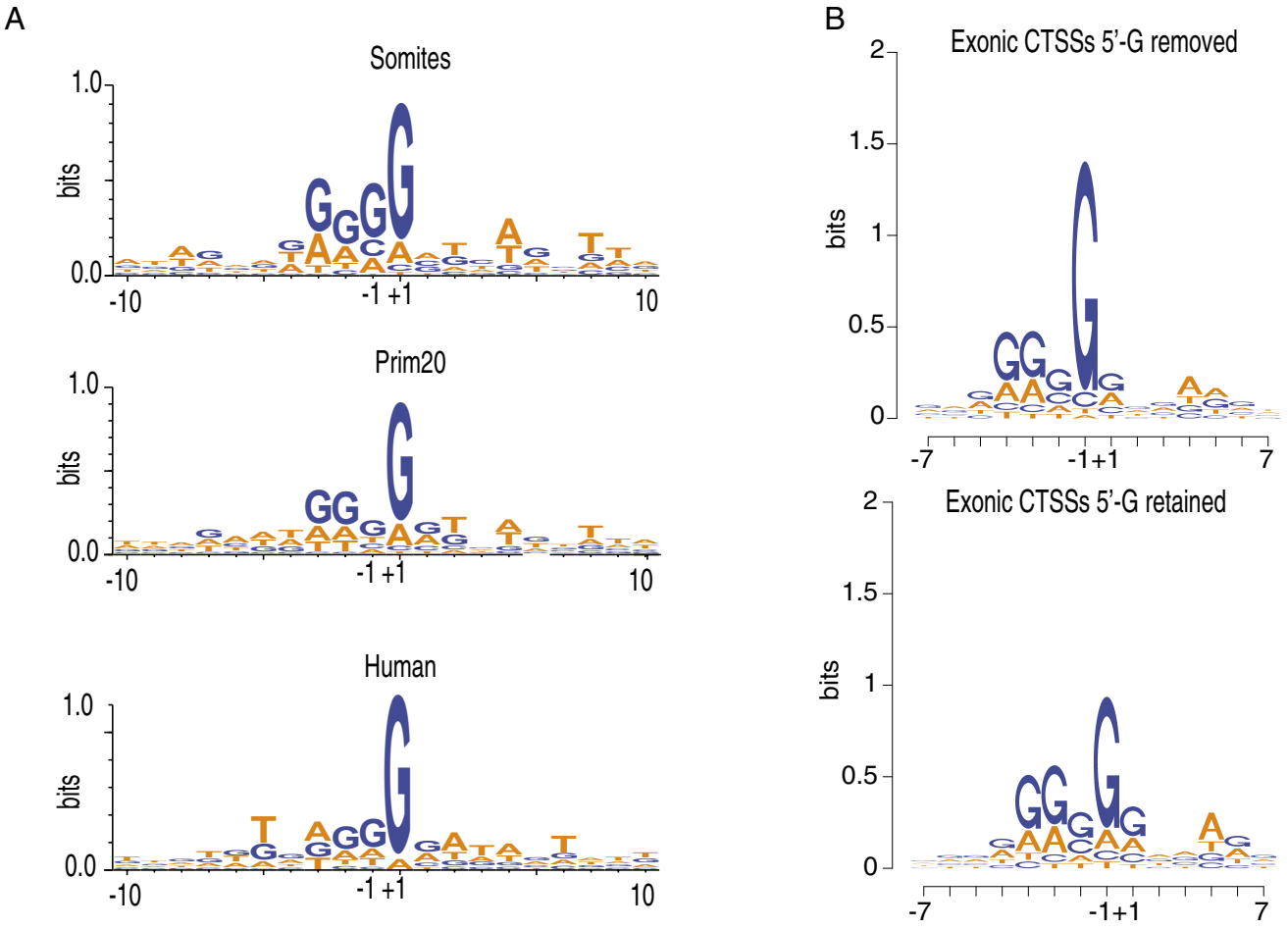
Nepal et al. Supplementary Figure 5

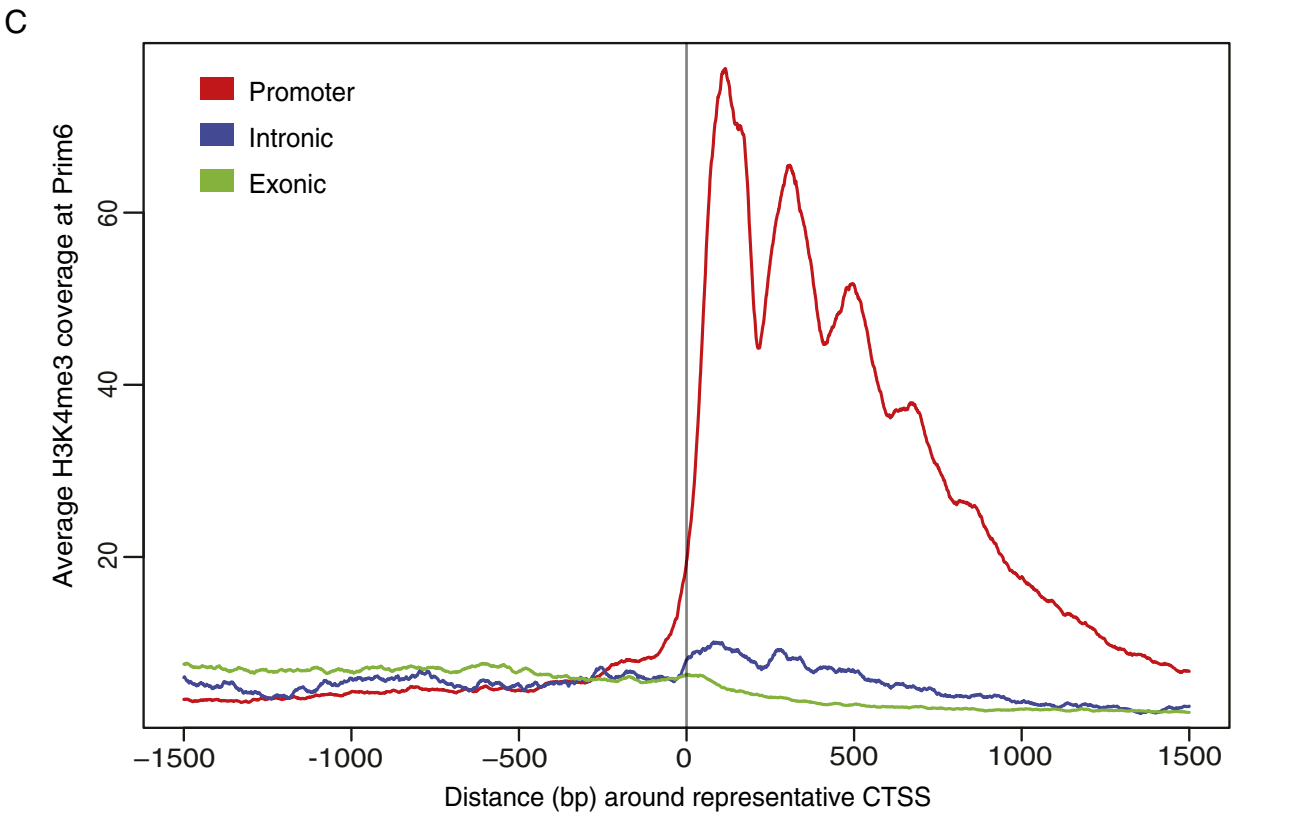


Nepal et al. Supplementary Figure 6



Nepal et al. Supplementary Figure 7





Nepal et al. Supplementary Figure 8