Supplemental Information

Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive

C. Xu & J. Zhang (jianzhi@umich.edu)

Supplemental Figures S1-S10
Supplemental Table S1

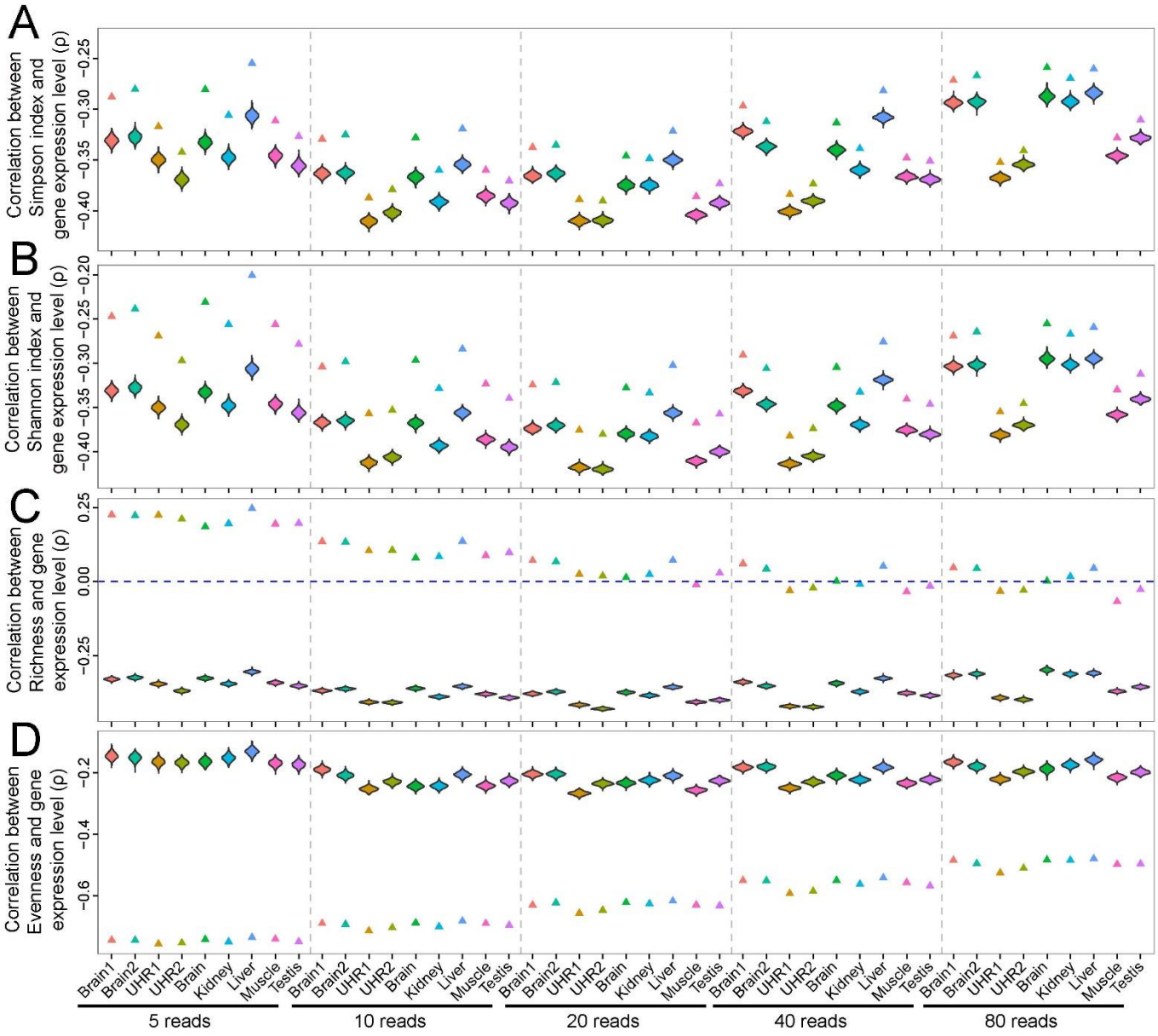

Figure S1. Relationship between human polyadenylation diversity and gene expression level investigated using different levels of down-sampling of PolyA-seq reads (from 5 to 80 reads per gene). *Related to Figure 1.* (A) Spearman's correlation between gene expression level and Simpson index of polyadenylation diversity in each of nine human tissue samples. (B) Spearman's correlation between gene expression level and Shannon index of polyadenylation diversity. (C) Spearman's correlation between gene expression level and polyadenylation site Richness. (D) Spearman's correlation between gene expression level and polyadenylation site usage Evenness. Triangles show the correlations on the basis of the original data, while the violin plots show the frequency distributions of the correlation on the basis of 1000 down-sampled data in which *m* PolyA-seq reads are randomly sampled per gene, where *m* is indicated below the tissues. For better comparison between results from the original and down-sampled data, for a given *m*, we used all genes with at least *m* reads as the original data. $P < 10^{-37}$ for all correlations in all down-sampled data.

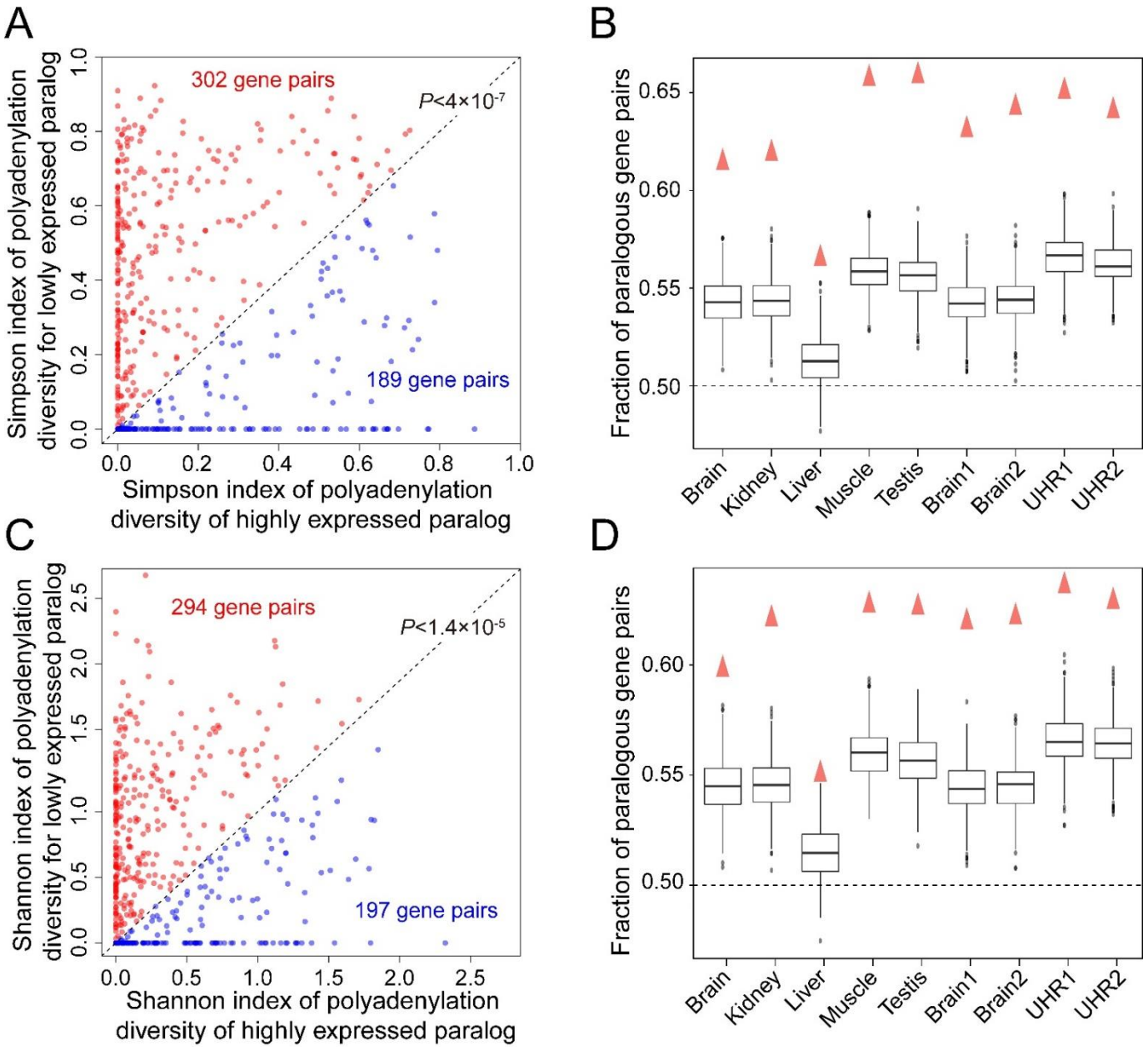

Figure S2. Within a paralogous gene pair, the relatively highly expressed paralog tends to have lower polyadenylation diversity than the relatively lowly expressed one in the human brain. *Related to Figure 1.* (A) Simpson index of polyadenylation diversity for each member of a paralogous gene pair. (B) Fraction of paralogous gene pairs for which the Simpson index of the relatively lowly expressed gene exceeds that of the relatively highly expressed one. (C) Shannon index of polyadenylation diversity for each member of a paralogous gene pair. (D) Fraction of paralogous gene pairs for which the Shannon index of the relatively lowly expressed gene exceeds that of the relatively highly expressed one. In (A) and (C), each dot represents a paralogous pair. Dots above and below the diagonal are colored red and blue, respectively. Numbers of red and blue dots are respectively indicated in the corresponding color. *P*-value is from a binomial test of the null hypothesis of equal numbers of red and blue dots. In (B) and (D), each triangle shows the result from the original data, whereas each boxplot shows the frequency distribution from 1000 down-sampled data. The bottom and top of each box are the first and third quartiles of all data points, and the band inside the box shows the median. The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box edges. The dots show outliers, which lie outside the range shown by the whiskers.

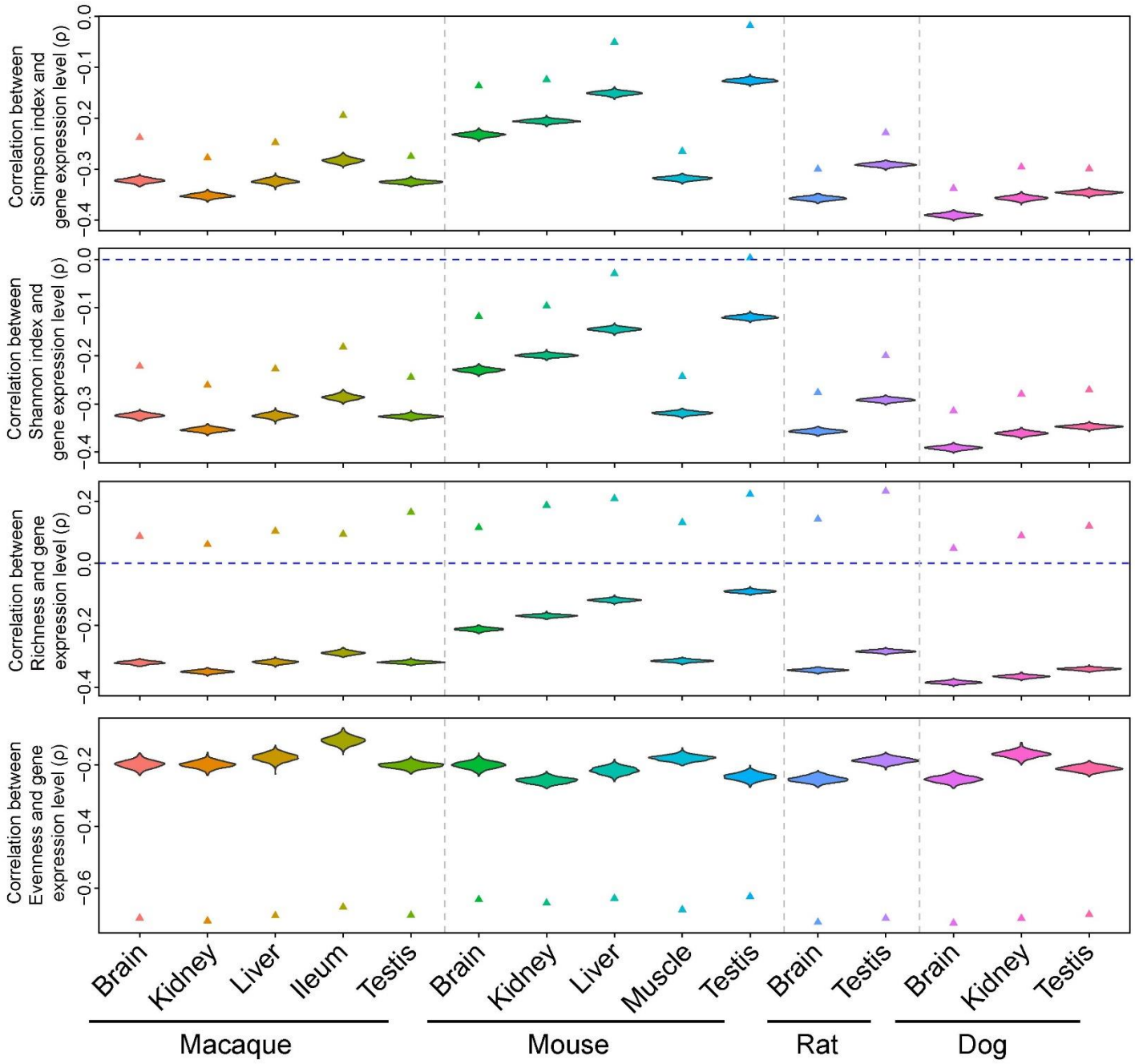

Figure S3. Correlations between various diversity indices and gene expression level in non-human mammals. *Related to Figure 1.* Triangles show results from original data whereas violin plots show frequency distributions from 1000 down-sampled data. $P < 10^{-4}$ in all cases except for Shannon index in mouse liver (P = 0.012), Simpson index in mouse testis (P = 0.093), and Shannon index is mouse testis (P = 0.78) in the original data.

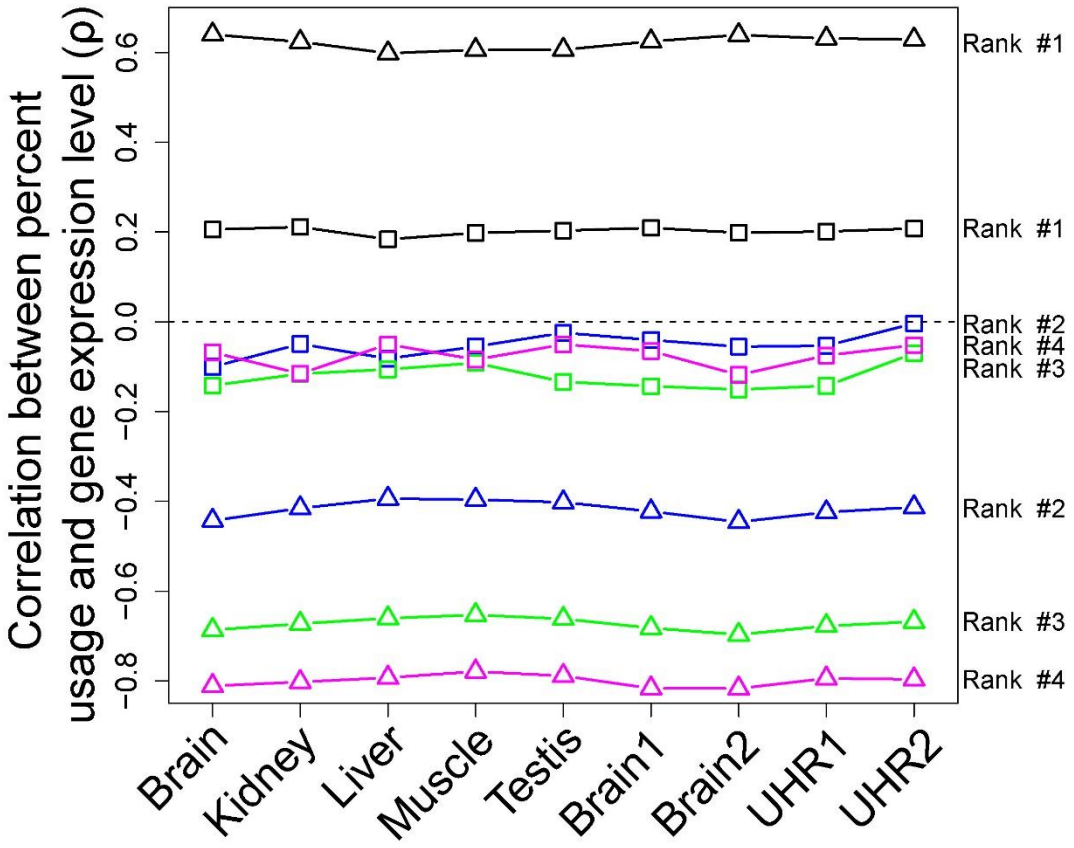

Figure S4. Spearman's correlation between gene expression level and the relative usage of each polyadenylation site in each of the nine human tissue samples examined. *Related to Figure 2.* $P < 10^{-185}$ in all cases except testis ($P = 0.22$), brain1 ($P = 0.075$), and UHR2 ($P = 0.8$) for rank #2 sites in the down-sampled data. Triangles and squares indicate the correlations on the basis of the original and down-sampled data, respectively. In each tissue sample, all correlations are based on genes with at least four polyadenylation sites found in that tissue.

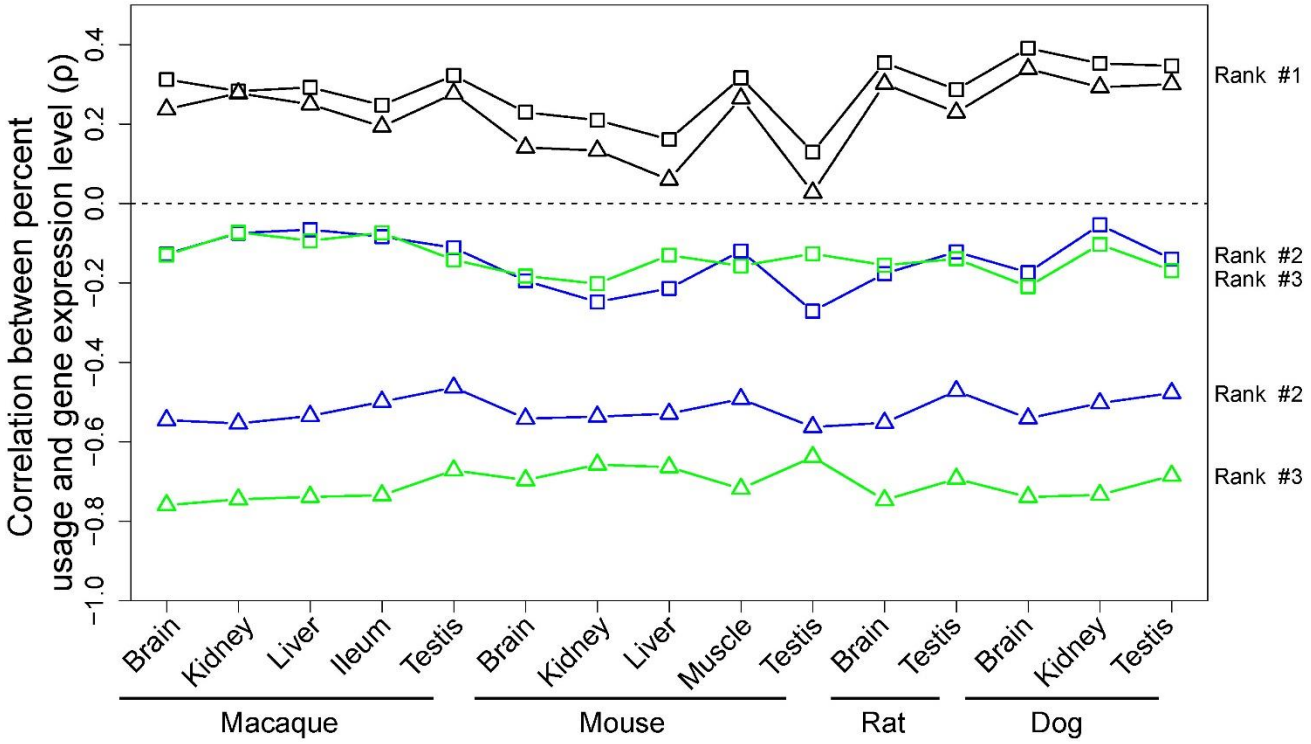

Figure S5. Spearman's correlation between gene expression level and the relative usage of each polyadenylation site in each tissue of non-human mammals. *Related to Figure 2.* Triangles and squares indicate the correlations on the basis of the original and down-sampled data, respectively. In each tissue, the correlation for polyadenylation site with a particular rank is calculated using the genes that have at least that particular number of polyadenylation sites. $P < 0.05$ in all cases.

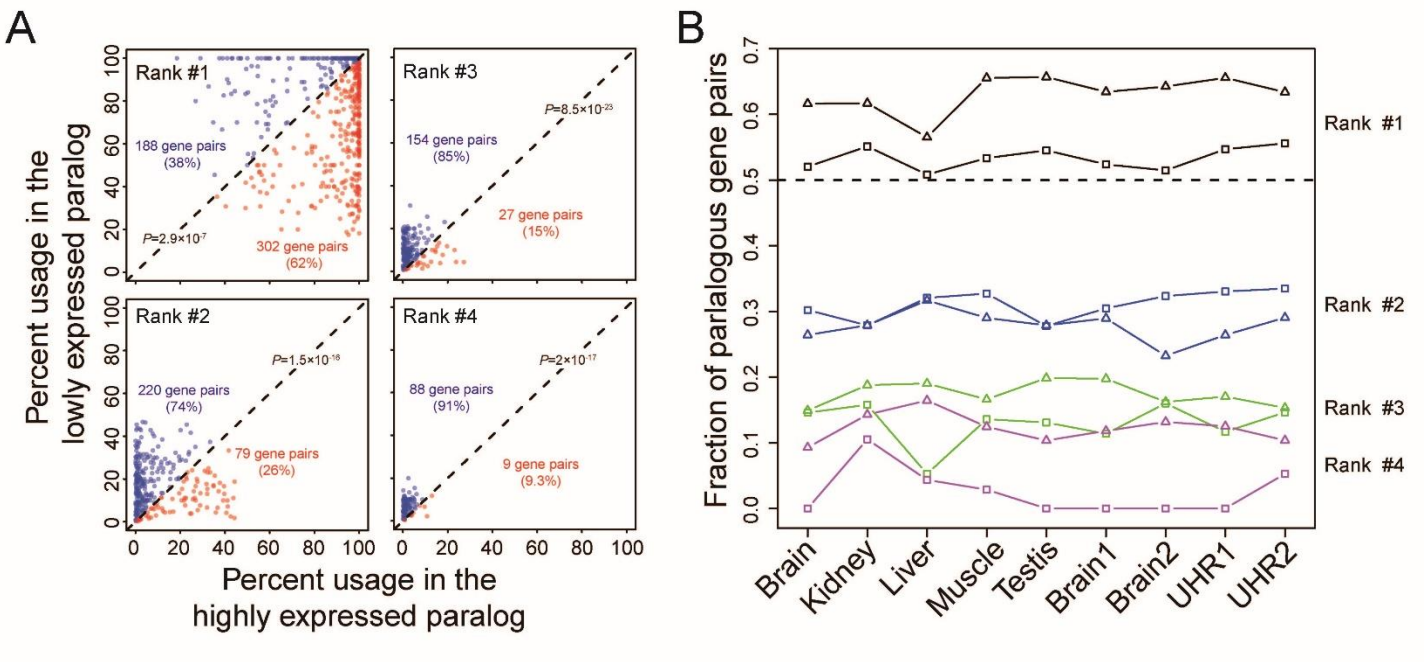

Figure S6. Relative uses of ranked polyadenylation sites in human paralogous genes. *Related to Figure 2.* (A) Relative uses of rank #1, #2, #3, or #4 polyadenylation sites in the relatively lowly expressed gene and relatively highly expressed gene of each paralogous gene pair in the human brain. Each dot represents a paralogous gene pair. Dots above and below the diagonal are colored blue and red, respectively. Numbers of blue and red dots are respectively indicated. *P*-value is from a binomial test of the null hypothesis of equal numbers of blue and red dots. (B) Fraction of paralogous gene pairs for which the relative usage of a ranked polyadenylation site is higher in the more highly expressed paralog. Triangles and squares show results from original and down-sampled data, respectively. $P < 0.05$ in all cases except brain ($P = 0.39$), liver ($P = 0.75$), muscle ($P = 0.11$), brain1 ($P = 0.25$), and brain2 ($P = 0.51$) for rank #1 sites in the down-sampled data.

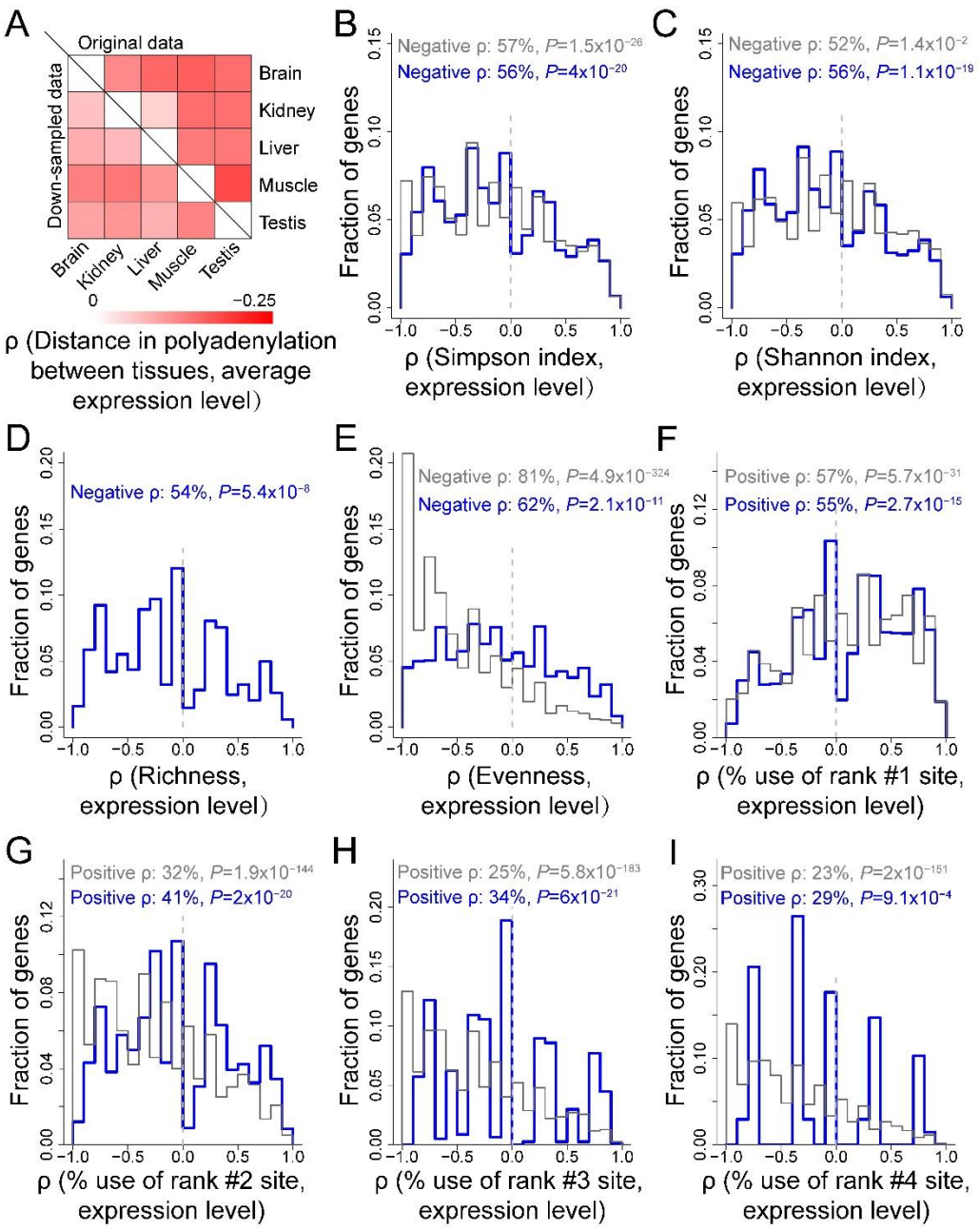

Figure S7. Among-tissue comparisons of polyadenylation in humans. *Related to STAR Methods.* (A) Spearman's correlations between the mean expression level of a gene in two tissues and its between-tissue distance in APA. Above and below the diagonal are results obtained using the original and down-sampled data, respectively. All correlations are statistically significant ($P < 10^{-23}$). (B-E) Frequency distribution of Spearman's correlation between the Simpson index of polyadenylation diversity (B), Shannon index of polyadenylation diversity (C), polyadenylation site Richness (D), or polyadenylation site use Evenness (E) of a gene in a tissue and its expression level in the tissue across the five human tissues. (F-I) Frequency distribution of Spearman's correlation between the percent usage of the rank #1 (F), #2 (G), #3 (H), or #4 (I) polyadenylation site of a gene in a tissue and the gene expression level in the tissue across human tissues. In (B)-(I), *P*-values are from binomial tests. Grey and blue colors show results from the original and down-sampled data, respectively.

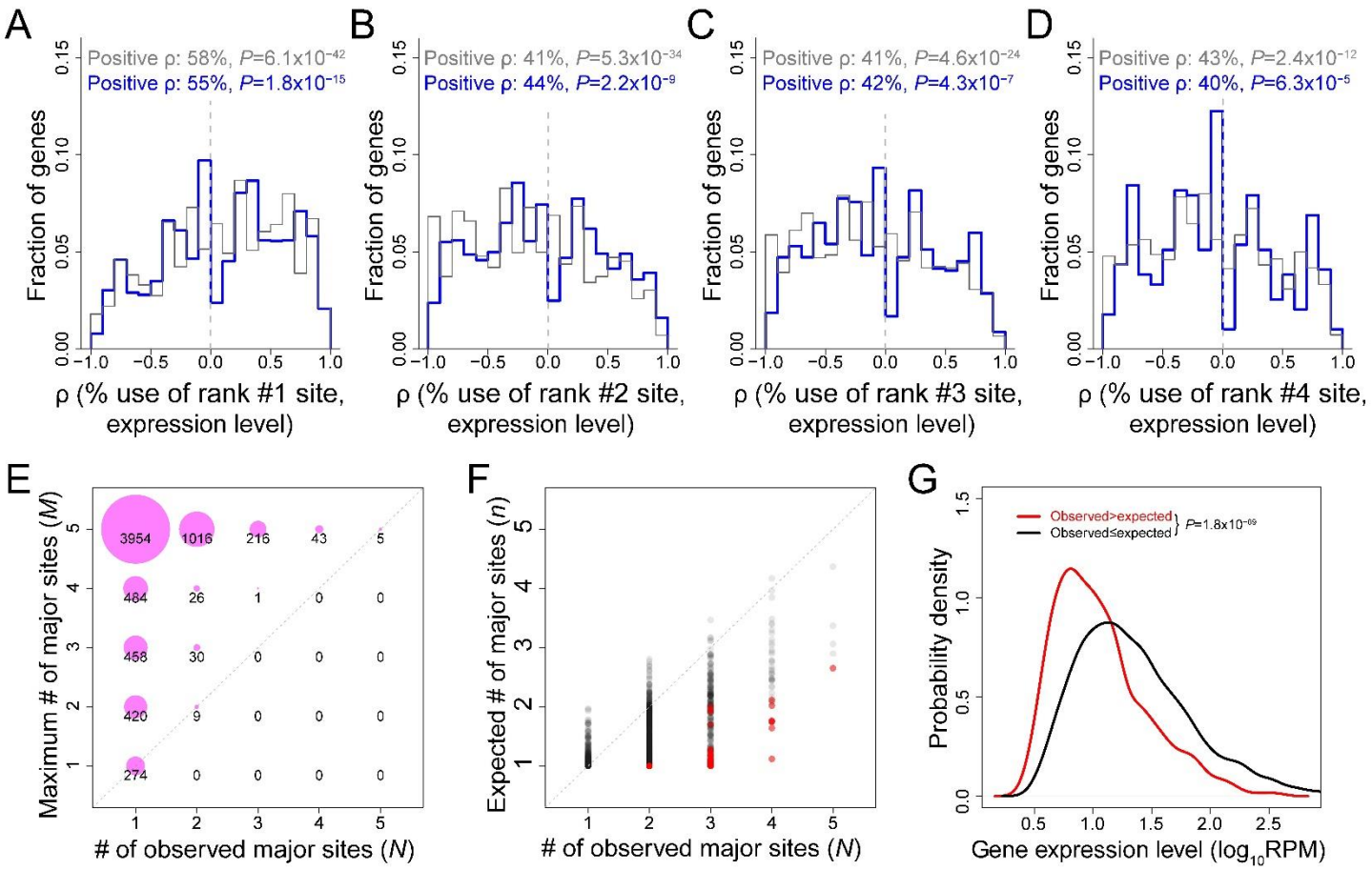

Figure S8. A small fraction of human genes show different major polyadenylation sites in different tissues. *Related to STAR Methods.* (A-D) Frequency distribution of Spearman's correlation between the relative usage of the global rank #1 (A), #2 (B), #3 (C), or #4 (D) polyadenylation site in a tissue for a gene and the gene expression level in the tissue across human tissues. The global rank of a polyadenylation site is determined using all reads from the five human tissues. *P*-values are from binomial tests. Grey and blue colors show results from the original and down-sampled data, respectively. (E) The maximum number of different major polyadenylation sites that a gene can have (given its observed polyadenylation sites) in the five human tissues (*M*) is much greater than the observed number of different major sites (*N*) for almost all genes with $M \geq 2$. The area of a circle is proportional to the number of genes in the circle. (F) The expected number of different major sites in the five human tissues under the hypothesis of no difference in polyadenylation site usage between tissues (*n*) is significantly smaller than the actual number (*N*) for a minority of human genes. Each dot represents a gene, with red dots denote genes whose *N* exceeds *n* significantly ($Q < 0.05$). No gene has a significantly lower *N* than *n* ($Q < 0.05$). (G) The probability densities of expression level for genes with larger *N* than *n* (not necessarily significantly; red) and the rest of the genes (black). In this panel, *n* is estimated using down-sampled data to equalize the sampling error among genes.

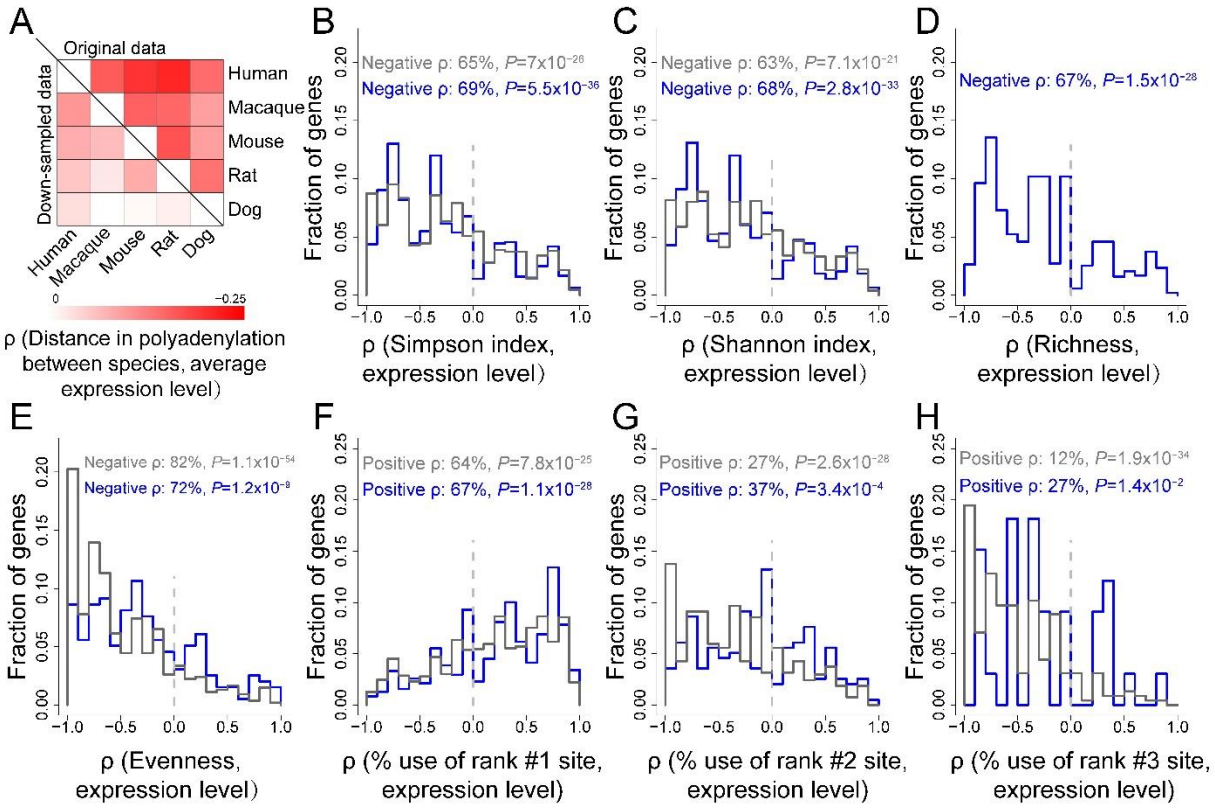

Figure S9. Among-species comparison of polyadenylation in the brain. *Related to STAR Methods.* (A) Spearman's correlations between the mean expression level of a gene in two species and its between-species distance in APA. Above and below the diagonal are results obtained using the original and down-sampled data, respectively. All correlations are statistically significant ($P < 0.05$) except for the comparisons between rat and macaque ($P = 0.17$), dog and macaque ($P = 0.20$), dog and mouse ($P = 0.74$), and dog and rat ($P = 0.33$) in down-sampled data. (B-E) Frequency distribution of Spearman's correlation between the expression level of a gene and its Simpson index of polyadenylation diversity (B), Shannon index of polyadenylation diversity (C), polyadenylation site Richness (D), or polyadenylation site usage Evenness (E) across five mammals. (F-H) Frequency distribution of Spearman's correlation between the expression level of a gene in a species and the percent usage of rank #1 (F), #2 (G), or #3 (H) polyadenylation site of the gene in the species across five mammals. Rank #4 sites are not examined because there are too few genes to allow an informative analysis. In (B)-(H), *P*-values are from binomial tests. Grey and blue colors show results from the original and down-sampled data, respectively.

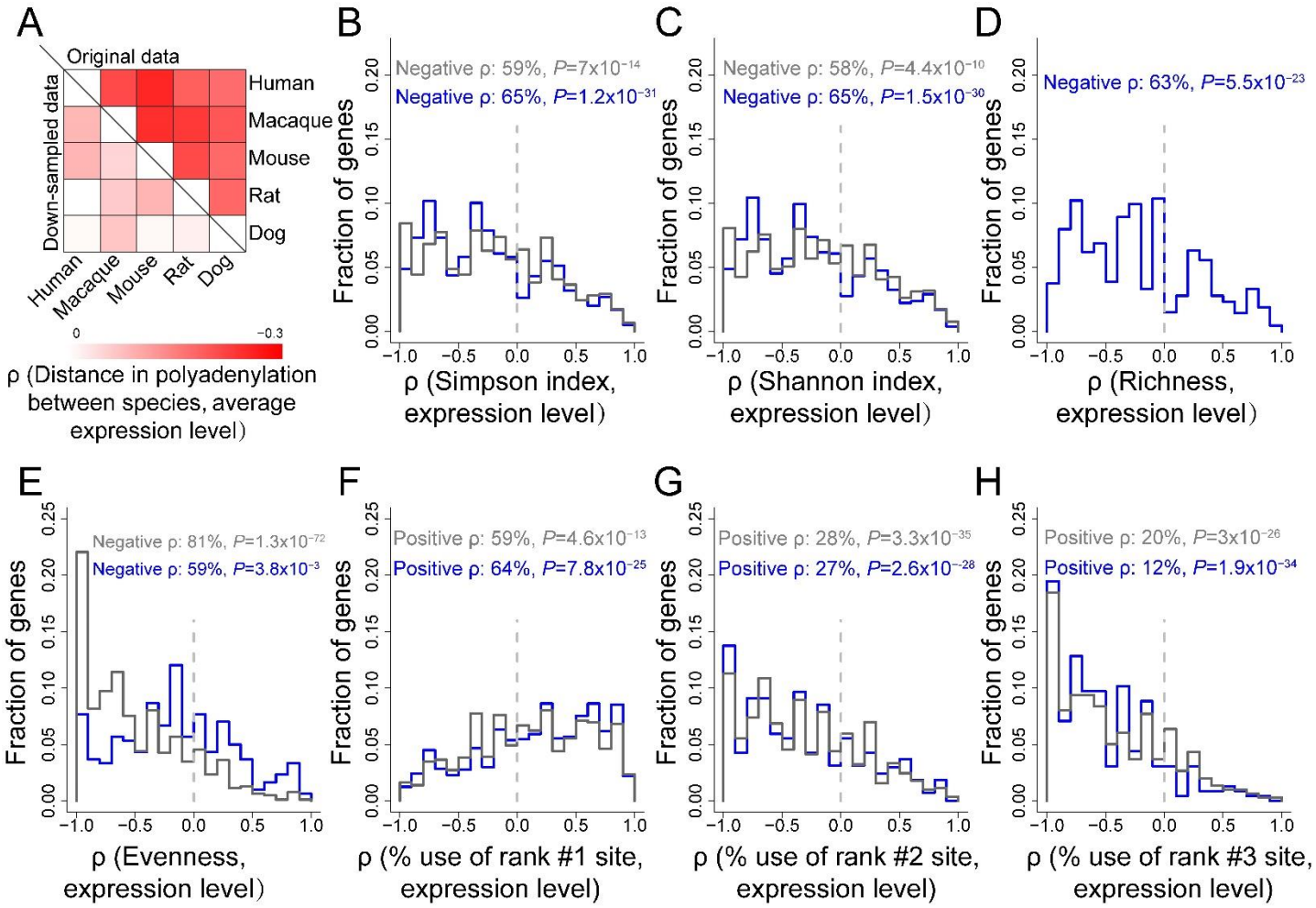

Figure S10. Among-species comparison of alternative polyadenylation in the testis. *Related to STAR Methods.* (A) Spearman's correlations between the mean expression level of a gene in two species and its between-species distance in APA. Above and below the diagonal are results obtained using the original and down-sampled data, respectively. All correlations are statistically significant ($P < 0.05$) except for the comparisons between rat and human ($P = 0.21$), dog and macaque ($P = 0.57$), dog and mouse ($P = 0.57$), and dog and rat ($P = 0.24$) in down-sampled data. (B-E) Frequency distribution of Spearman's correlation between the expression level of a gene and its Simpson index of polyadenylation diversity (B), Shannon index of polyadenylation diversity (C), polyadenylation site Richness (D), or polyadenylation site use Evenness (E) across five mammals. (F-H) Frequency distribution of Spearman's correlation between the expression level of a gene in a species and the percent usage of the rank #1 (F), #2 (G), or #3 (H) polyadenylation site in the species across five mammals. Rank #4 sites are not examined because there are too few genes to allow an informative analysis. In (B)-(H), *P*-values are from binomial tests. Grey and blue colors show results from the original and down-sampled data, respectively.

Table S1. Comparison between divergences and polymorphisms at major, minor, and pseudo-PASs.
Related to Figure 4.

Types of PASs	# of substitutions	# of SNPs	# of substitution / # of SNPs	*P*-value (Fisher's exact test)	
Major PASs	418	645	0.65	6.3×10^{-3} (Major vs. Minor)	
Minor PASs	704	871	0.81		
Pseudo-PASs	3125	3589	0.87	1.2×10^{-5} (Major vs. Pseudo)	0.19 (Minor vs. Pseudo)