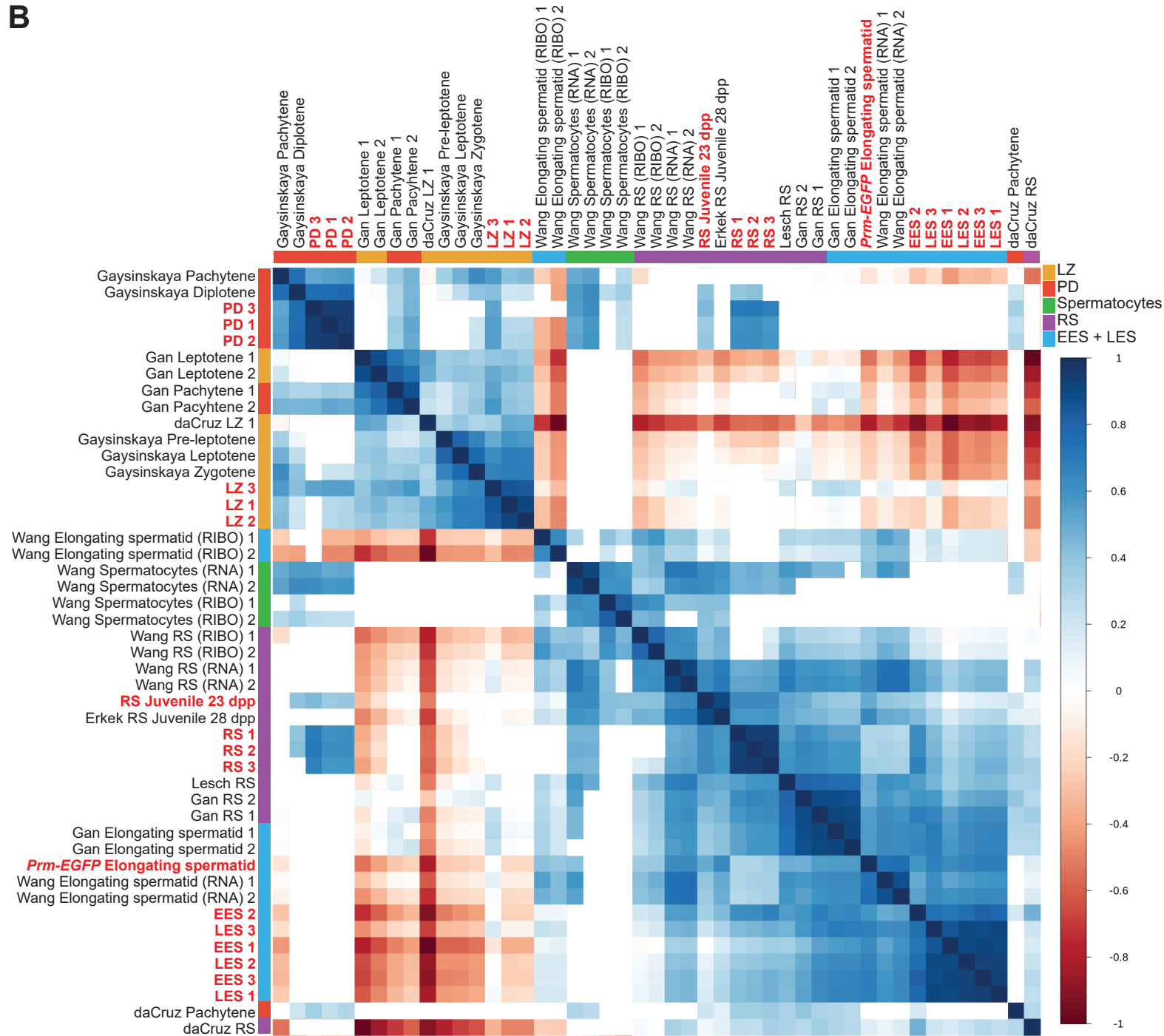


Gill Supplementary Figure S1

A

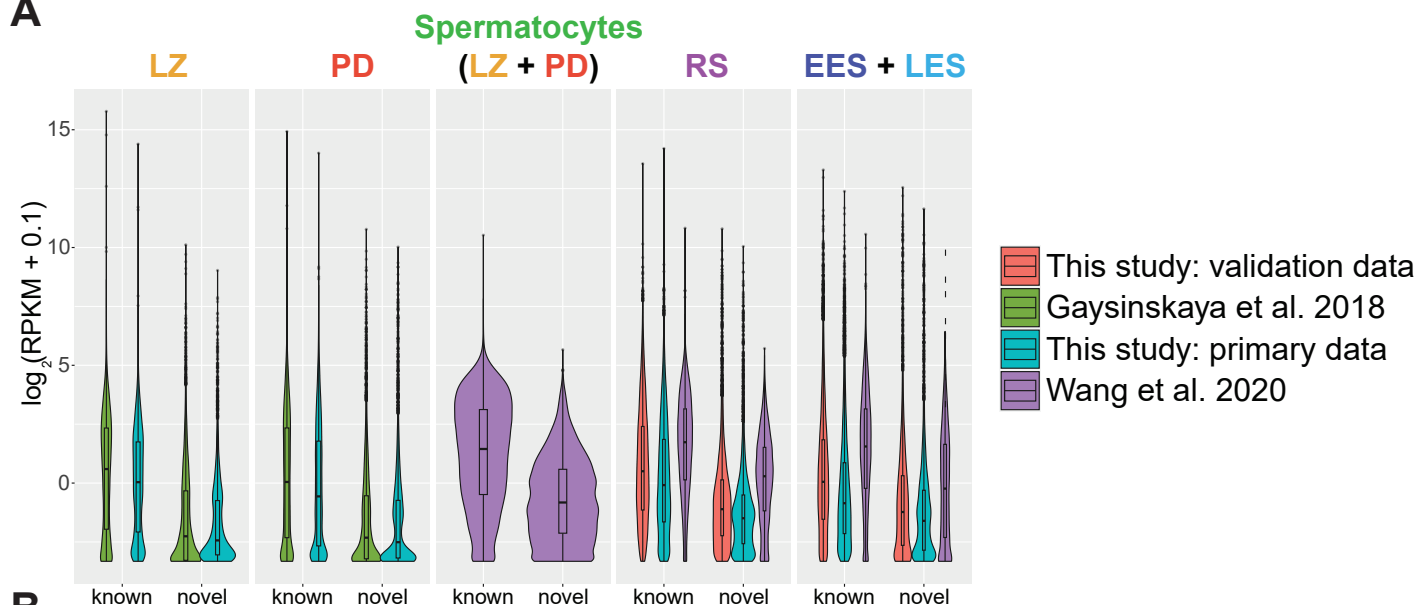
Sample	Total # reads	Total # Reads mapped	Total # De-duplicated reads	Total # Uniquely mapped	Total # properly paired
LZ1	103,107,945	96,291,002	38,797,869	36,544,639	36,374,244
LZ2	106,200,167	101,243,064	43,423,815	41,075,021	41,016,072
LZ3	106,842,536	99,174,908	38,937,410	36,606,067	36,468,756
PD1	116,157,263	110,110,950	65,120,652	62,059,172	62,019,914
PD2	127,038,629	122,396,178	87,128,111	83,310,897	83,278,672
PD3	147,279,280	137,333,598	100,844,242	96,733,086	96,677,428
RS1	114,322,331	107,913,614	43,969,439	41,094,373	41,036,836
RS2	143,446,403	137,849,519	99,898,056	95,645,031	95,602,226
RS3	142,542,696	132,531,466	92,343,342	88,201,226	88,111,180
EES1	98,800,636	90,553,570	38,582,953	37,028,938	36,974,268
EES2	105,289,928	96,930,399	62,917,794	60,402,208	60,331,582
EES3	128,353,462	121,159,712	60,701,162	58,573,030	58,549,948
LES1	105,886,195	101,139,226	50,795,901	48,903,568	48,886,140
LES2	101,750,523	97,235,159	54,378,506	52,536,698	52,527,246
LES3	118,051,244	113,843,712	55,272,528	53,210,846	53,197,484



Supplementary Figure S1. Quality control of RNA-seq (A) Table showing sequencing statistics for RNA-seq used to perform de novo transcriptome annotation. (B) Correlation plot comparing RNA-seq data generated for transcriptome annotation to validation data generated in this study and published data sets. Colored bars on axes represent cell types. Only Pearson’s correlation values with a p-value < 0.01 are shown.

Gill Supplementary Figure S2

A

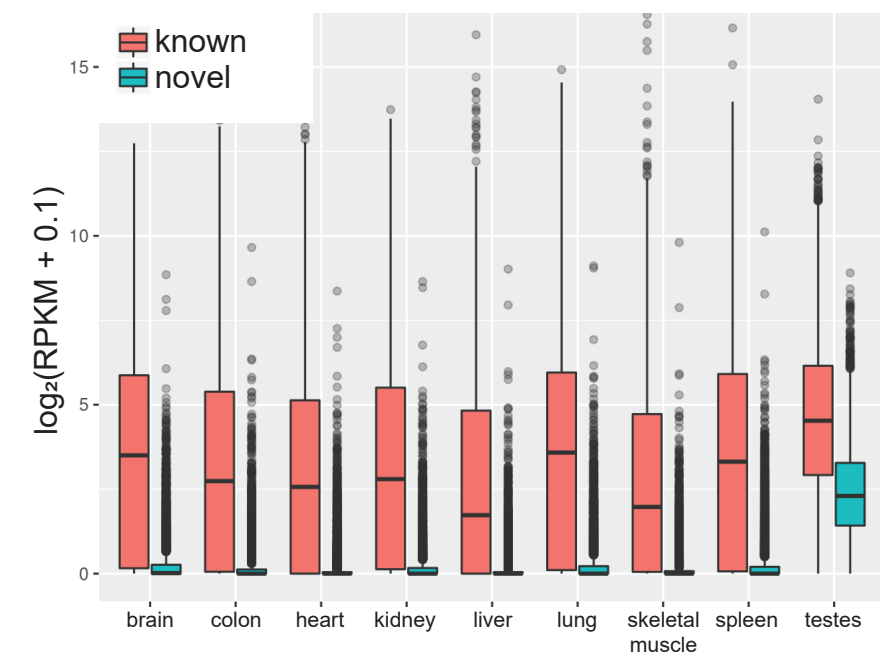


B

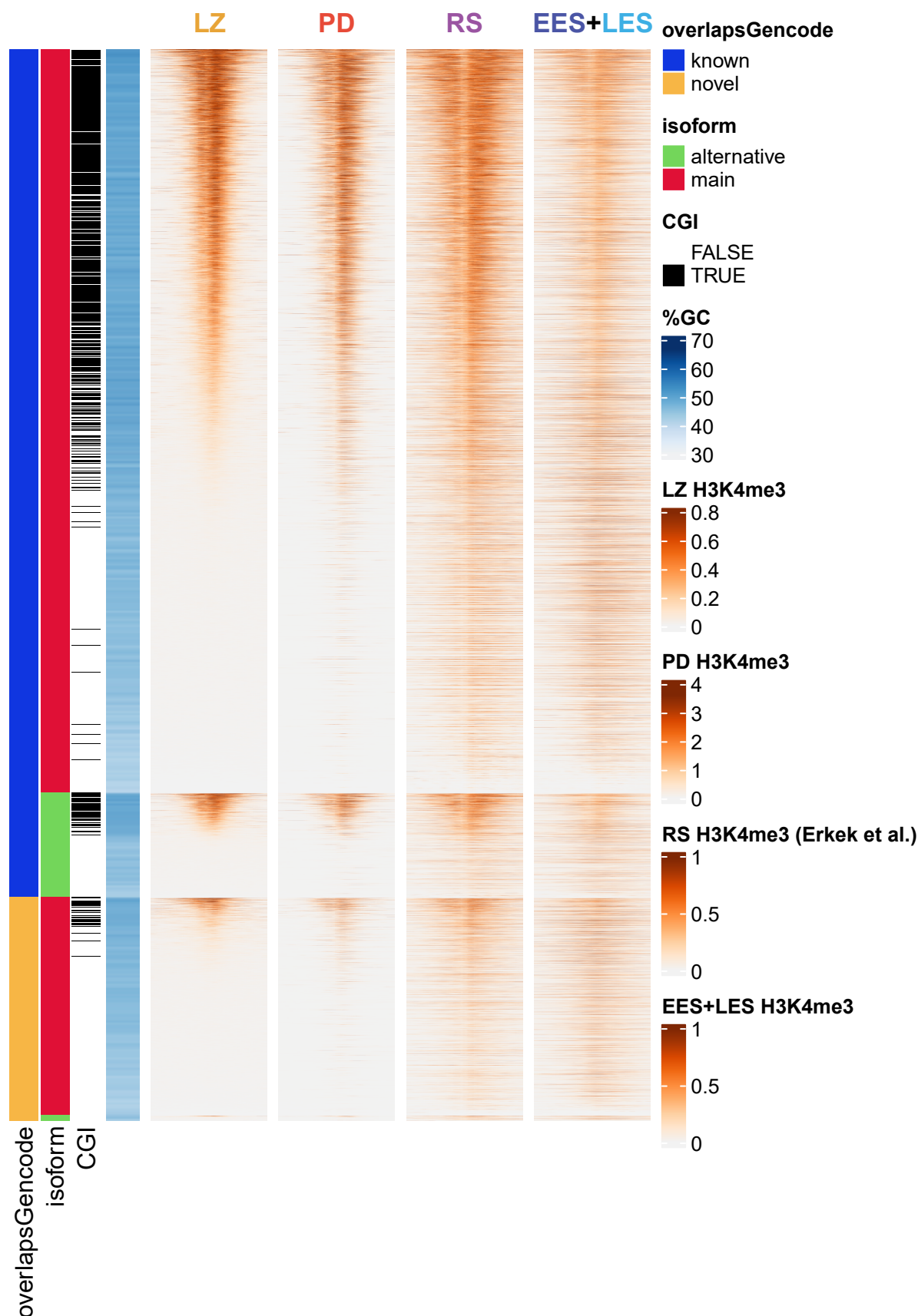


Supplementary Figure S2. Expression levels of newly identified genes (A) Violin plots showing distributions of expression values for validation data from this study and published data sets. (B) PCA plot of $\log_2(\text{RPKM})$ values for newly annotated genes in data from this study and published data sets. (C) Boxplot showing expression values for adult tissue RNA-seq for genes divided by annotation type.

C

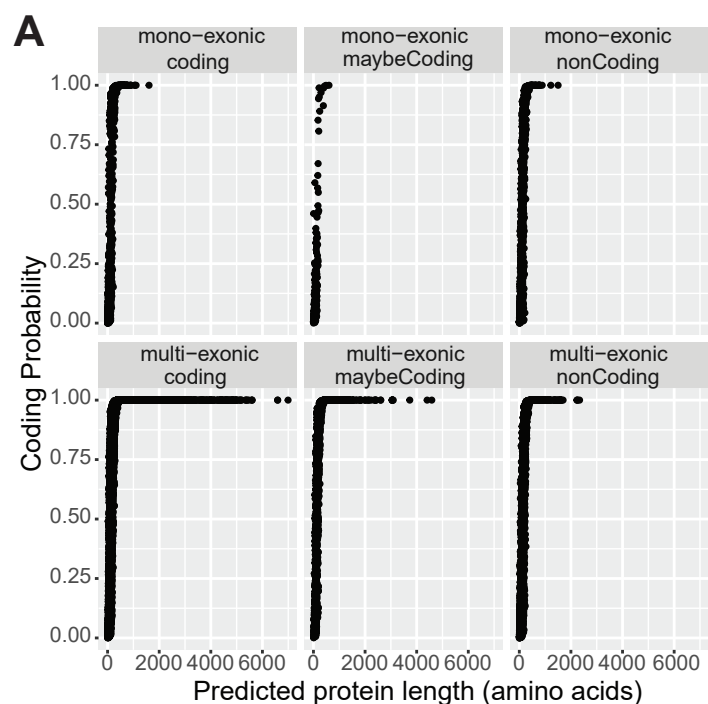


Gill Supplementary Figure S3

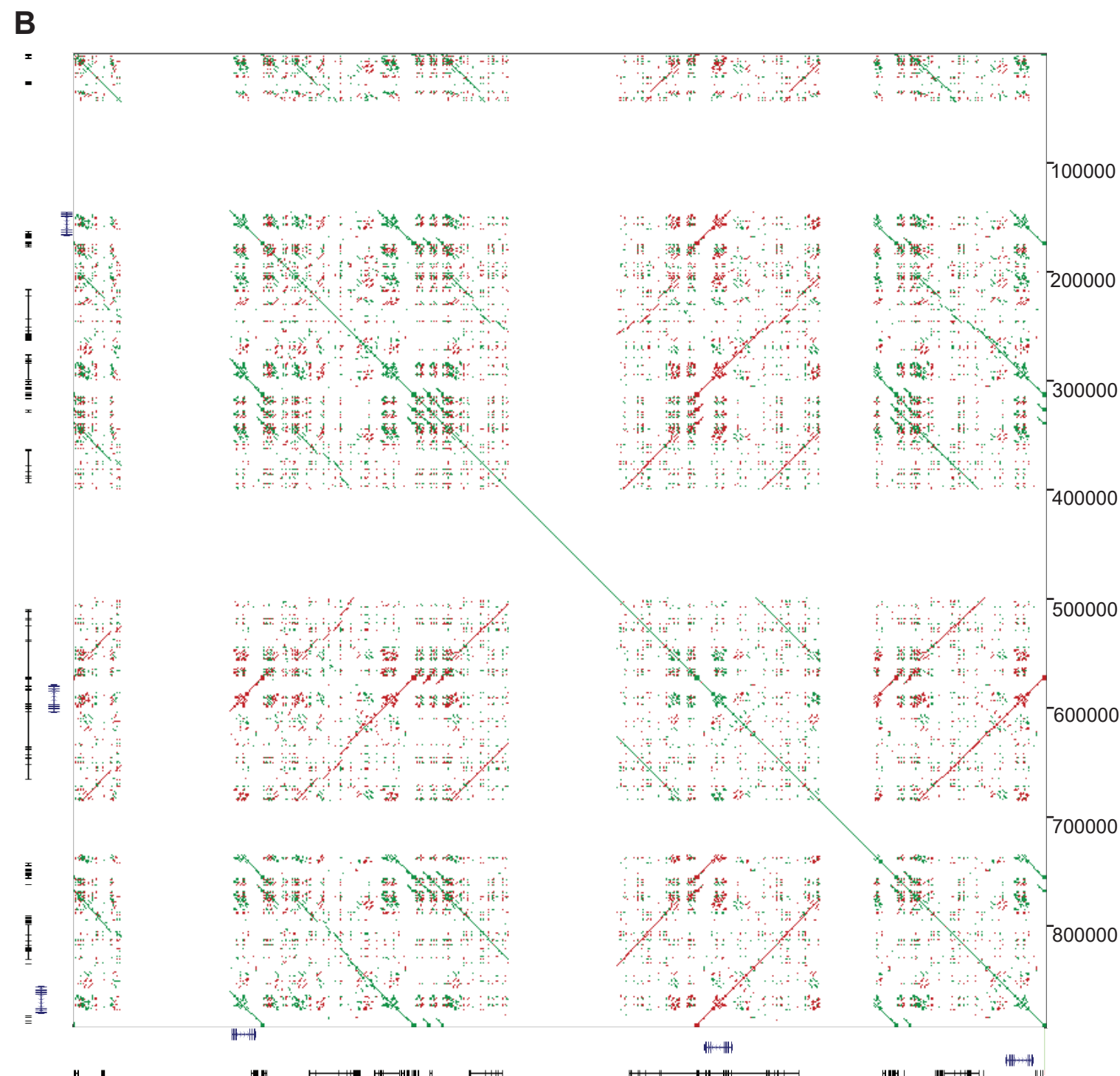


Supplementary Figure S3. Confirmation of newly identified genes by H3K4me3 ChIP-seq analysis. Heatmap showing $\log_2(\text{RPKM})$ of H3K4me3 $\pm$ 2kb from 5' end of genes. Blue genes overlap with a previously annotated transcript, while yellow represent novel genes. Within the previously annotated genes 5' ends of the highest expressed transcript are labelled in red, while alternative 5' ends are labelled in green. CGI refers to promoters overlapping with annotated CpG islands. Heatmap is ordered by H3K4me3 enrichment in LZ (within each category of genes).

Gill Supplementary Figure S4

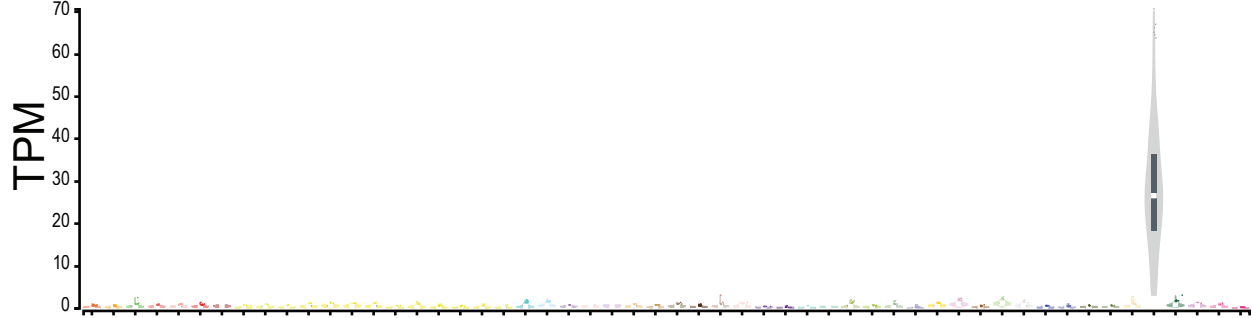


Supplementary Figure S4. Distribution of ORF length and coding probabilities in genes known to code protein. (A) Distribution of protein length (in amino acids) vs. CPAT-determined coding probabilities in GENCODE genes of known coding potential. (B) Dot plot generated with YASS (Noe and Kucherov 2005) showing genomic self-similarity in an 892 kb region on Chromosome 8 with location of newly annotated genes shown underneath. Regions with similarity in direct orientation to the genomic region are colored green, while those with similarity in the inverted orientation are colored red. Position (relative to start of region containing genes) is indicated on the left-hand access. Upper level of gene annotation (in blue) shows gene models known to be present in region (*Gm31371*, *Gm15319* & *Gm2119*), while lower level of gene annotation (in black) shows newly identified genes.

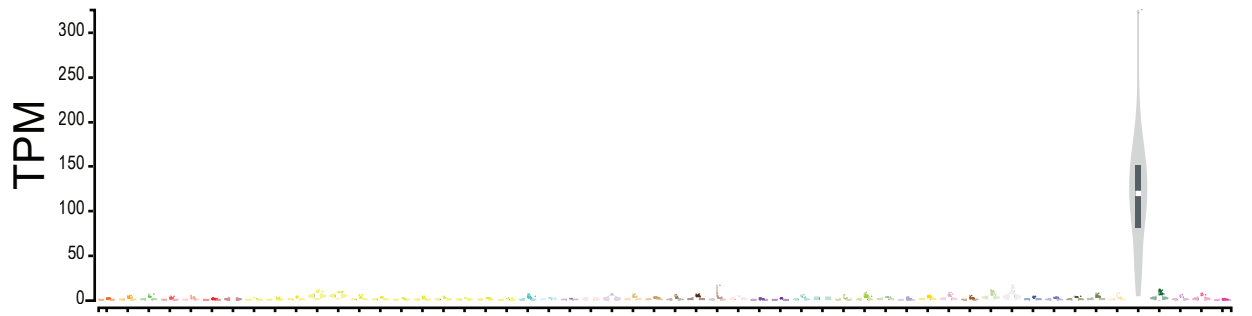


Gill Supplementary Figure S5

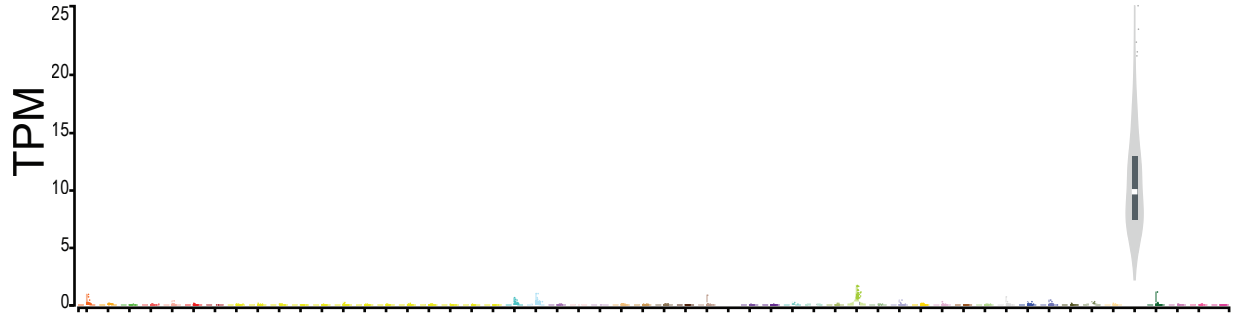
A *C2ORF16*



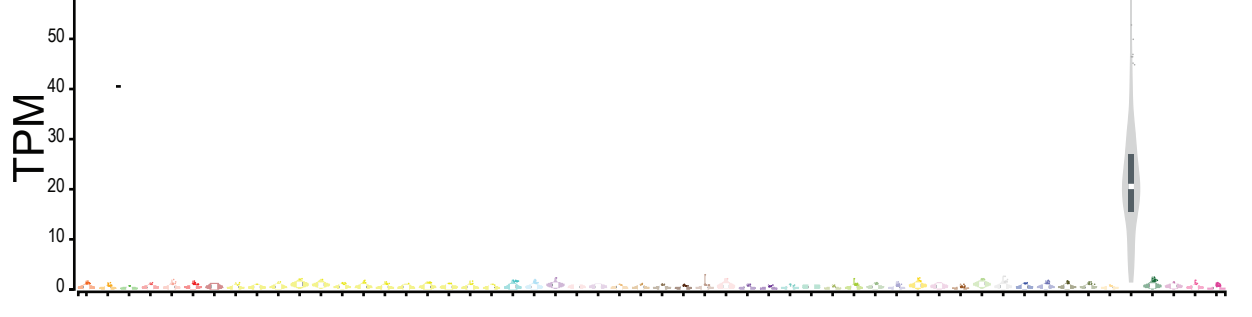
SRRM5



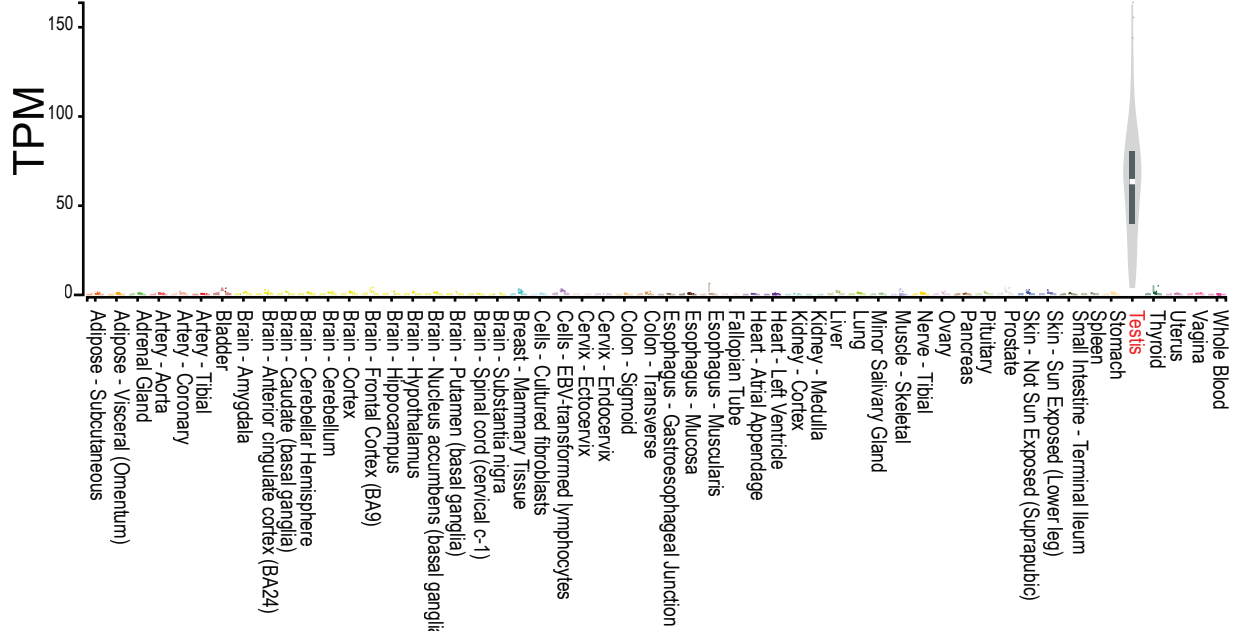
C18ORF63

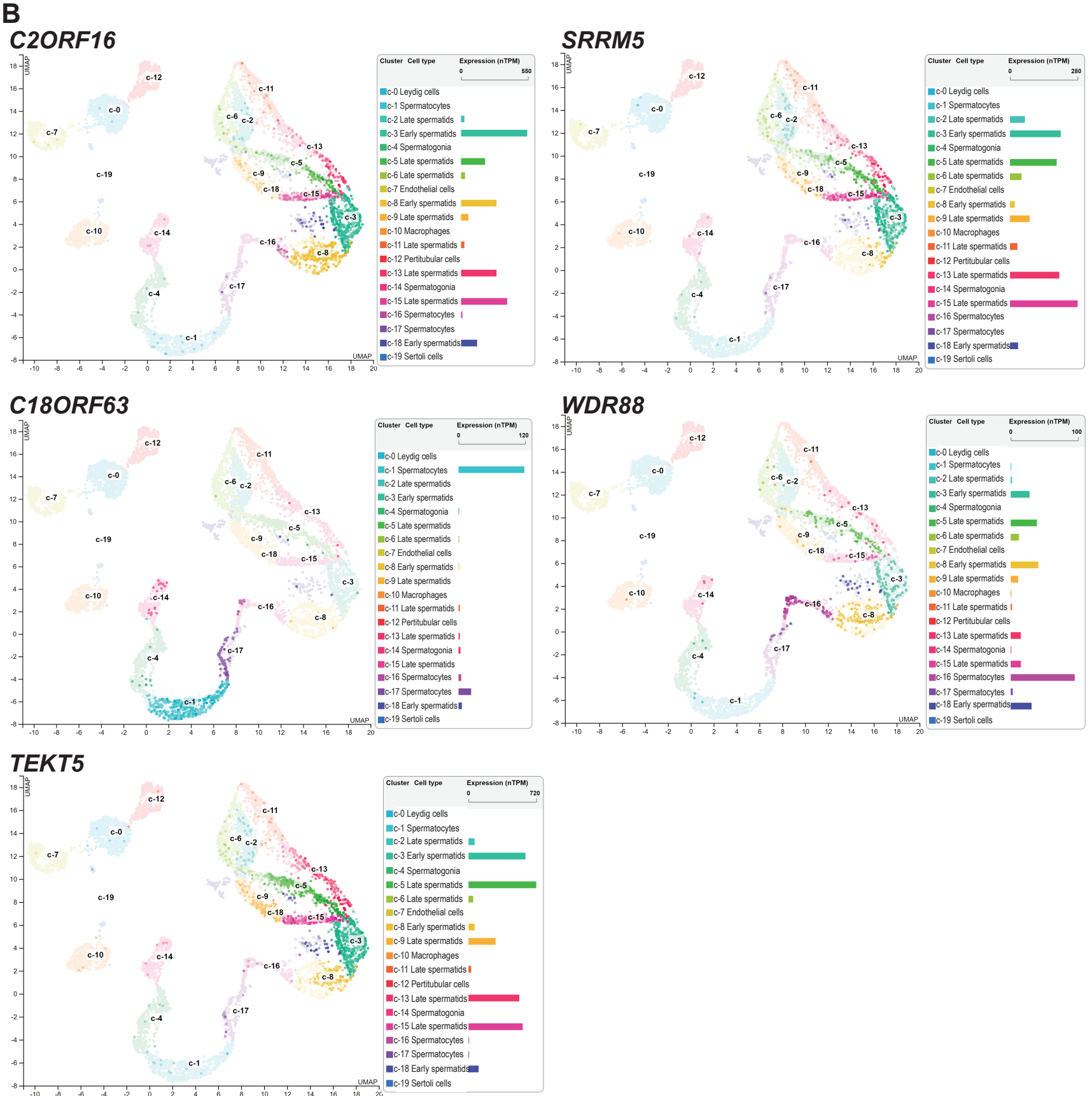


WDR88



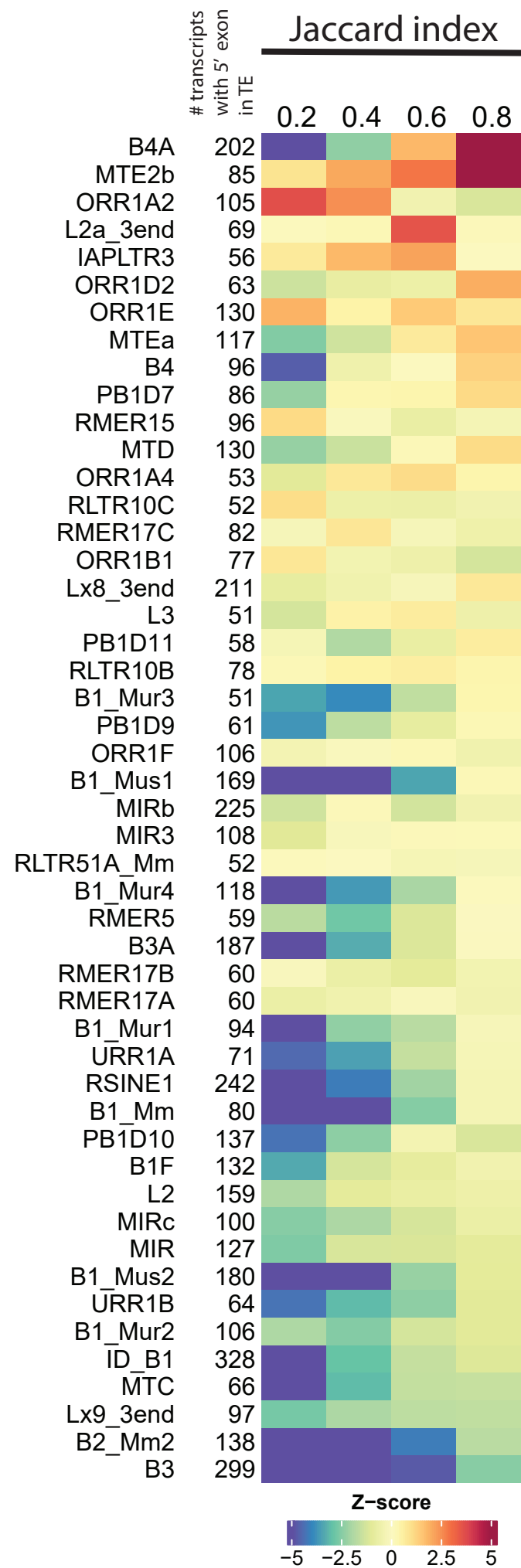
TEKT5





Supplementary Figure S5. Expression of human homologs of novel mouse protein coding genes. (A) Violin plots of TPM values for human tissues profiled in the GTEx project for genes homologous to the novel mouse protein-coding genes identified in this study. Image credit: GTEx consortium. (B) UMAP plot of expression of homologs in single-cell RNA-seq data from human testis. Image credit: Human protein atlas.

Gill Supplementary Figure S6



Supplementary Figure S6. Expansion of repeat overlaps with first exons of novel genes. Heat-map showing Z-score of Jaccard Indices comparing 5' exons to repNames at defined levels compared to randomized distributions. Figure 5C shows the repNames with the top 10 Z-scores.