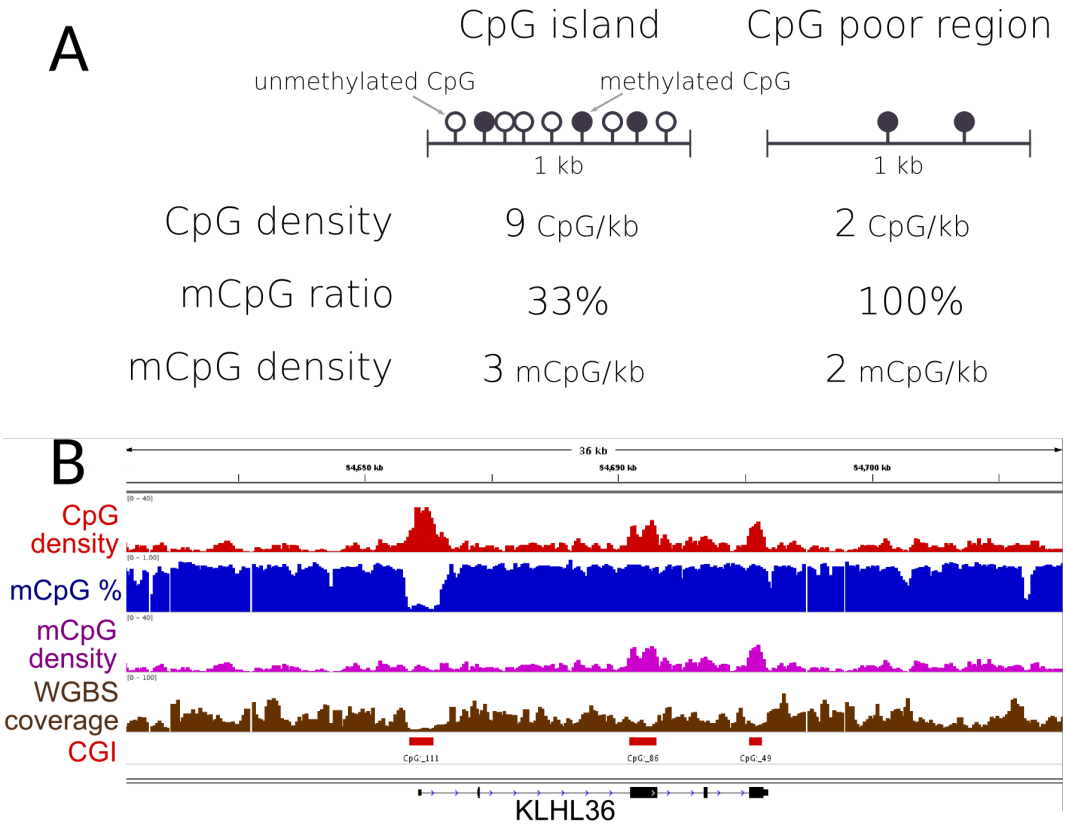
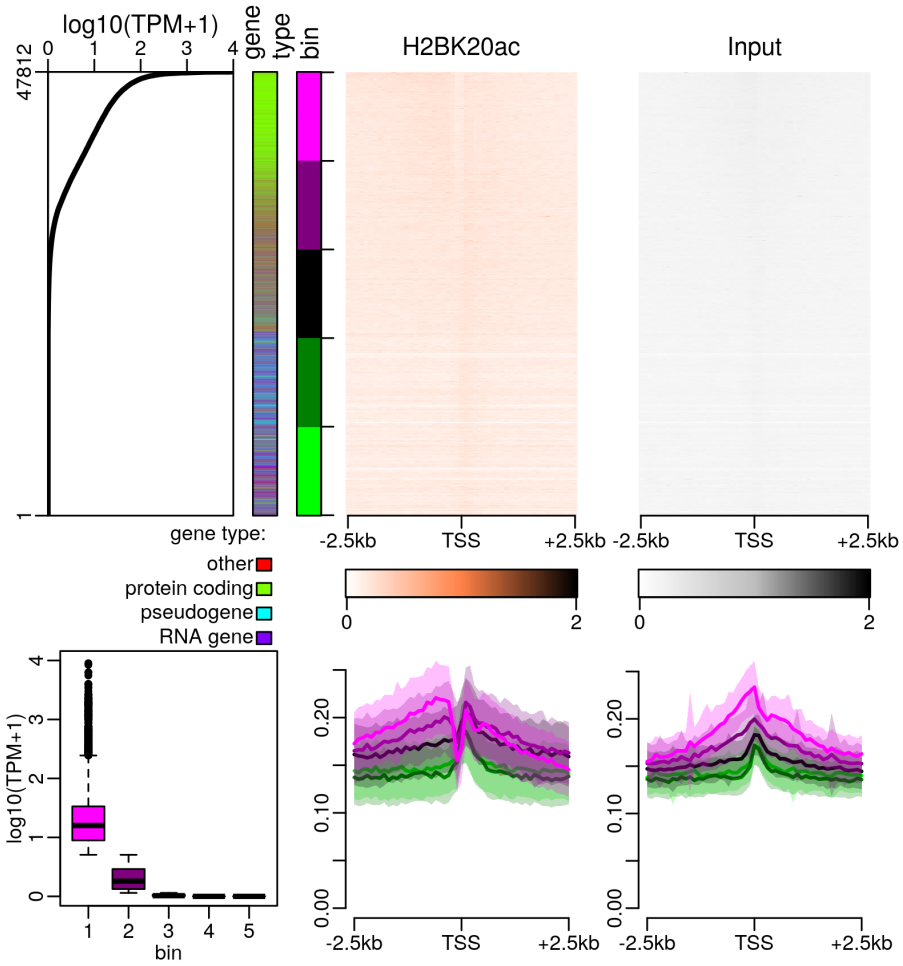
Electronic Supplementary Material (ESI) for Molecular Omics.

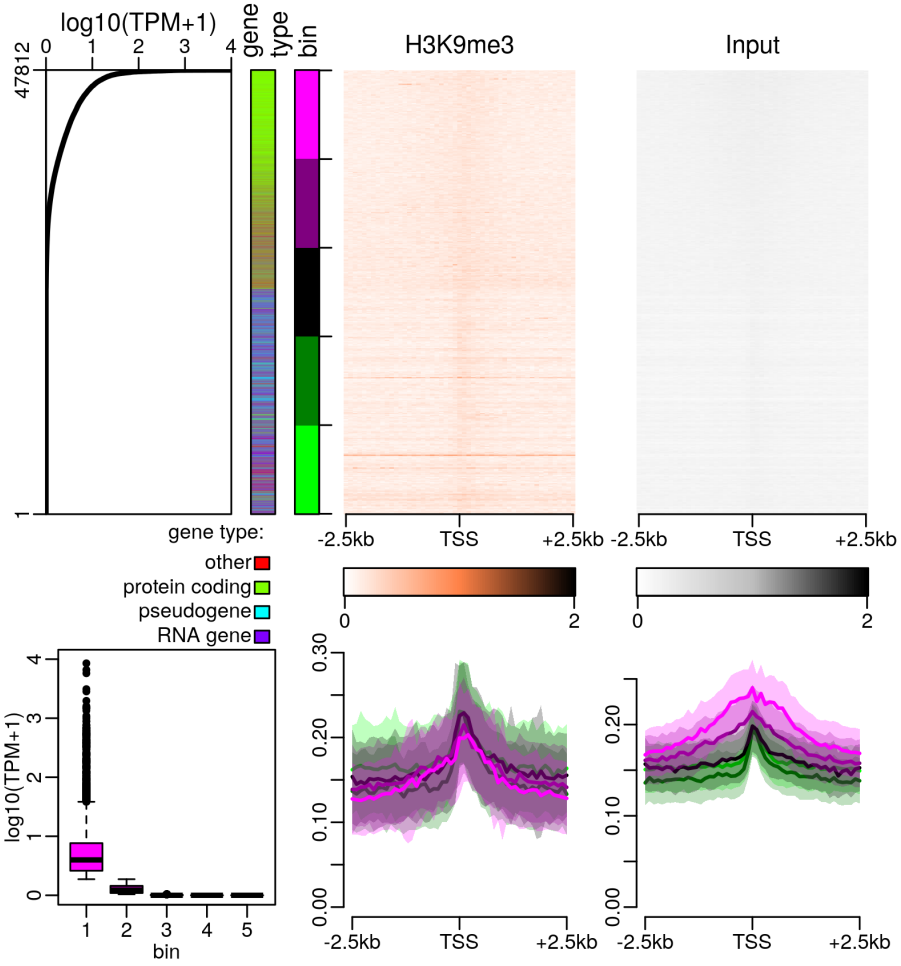
Supplementary figures



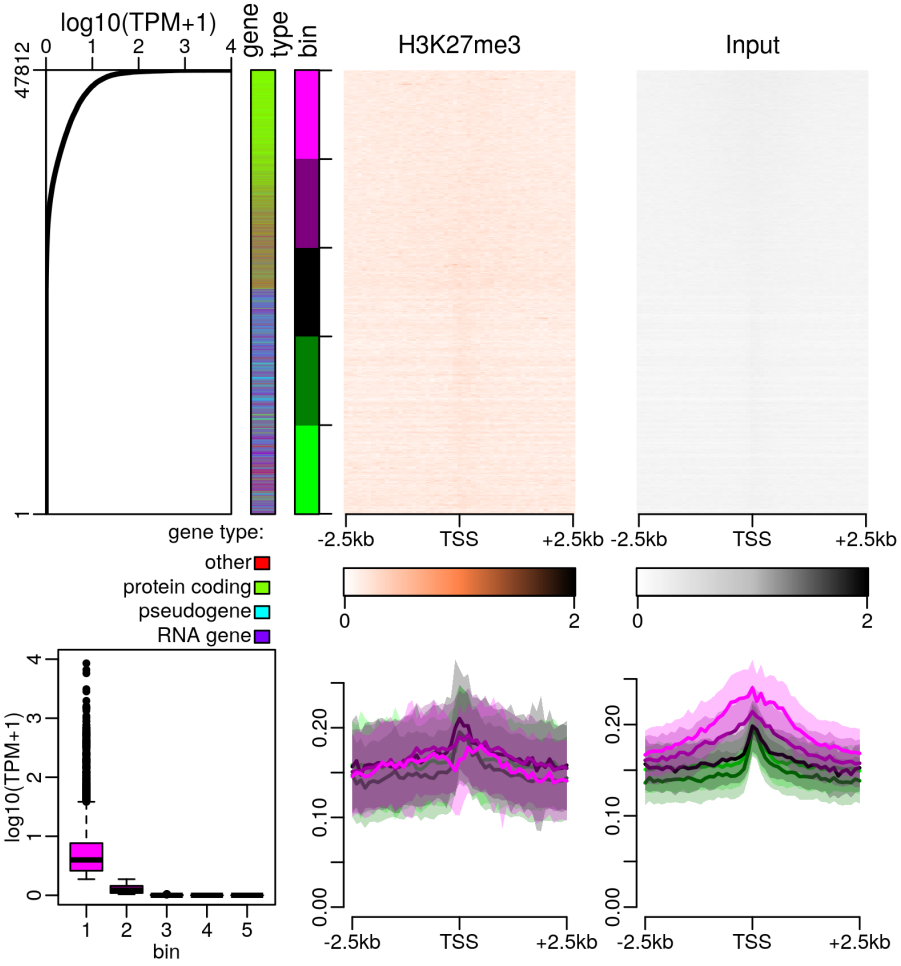
Supplementary Figure 1: WGBS analysed using CpG density, mCpG ratio and mCpG density metrics. The mCpG ratio and mCpG density metrics can behave differently while comparing regions with different CpG density. **A**. A hypothetical 1 kb genomic window, the CpG rich region (left) has an lower mCpG ratio than the CpG poor region (right), but an higher mCpG density. **B**. Four tracks obtained from WGBS data (here from the H1-hESC cell line). A 36kb region including gene KLHL36 is shown. From top to bottom: The CpG density tracks, higher at the three CpG islands (red lines) in the region. The mCpG ratio, shows high signal ($> 80\%$) everywhere but at the first CpG island. The mCpG density, with a low signal everywhere but at the second and third CpG island. Finally, the WGBS coverage is used as a control track.



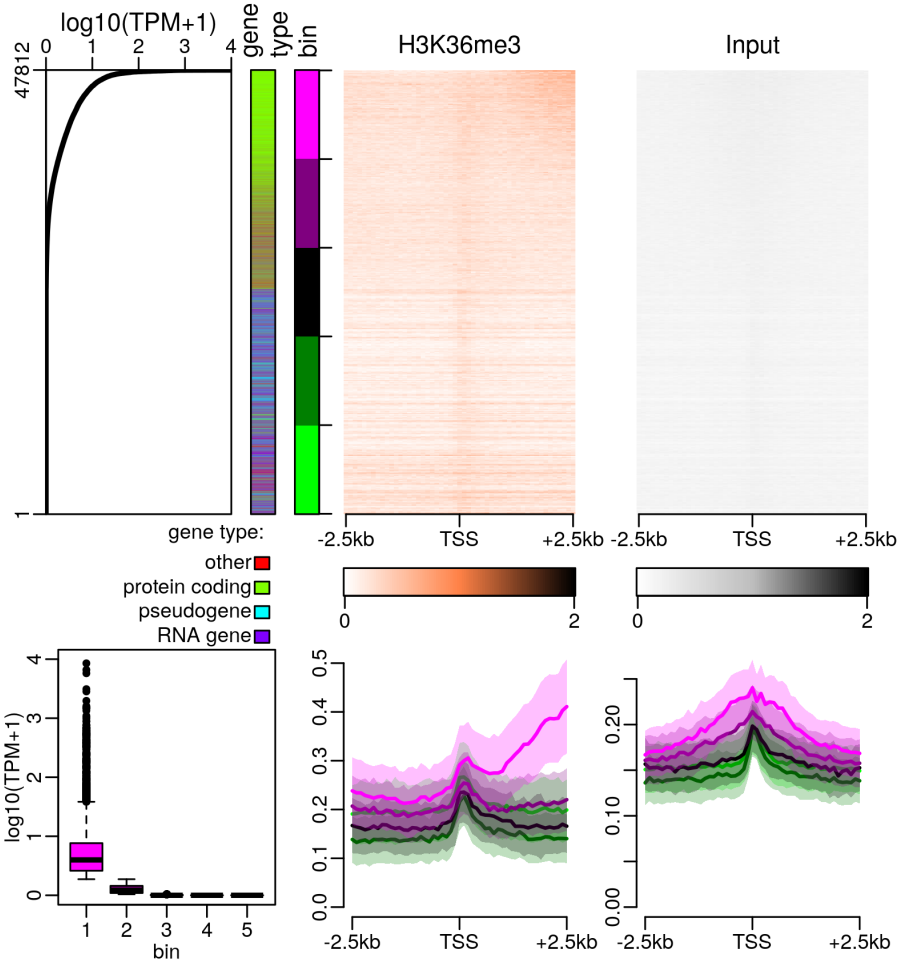
Supplementary Figure 2: H2BK20ac near the transcription start sites in H1 cell line (the only cell line where it is available in this dataset). Upper panel, from left to right: Gene expression level in all 47,812 genes annotated by Gencode. First side bar indicates the gene type (green: protein coding genes, blue: pseudogenes, purple: RNA genes, red: other types of genes), the second side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H2BK20ac ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H2BK20ac ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.



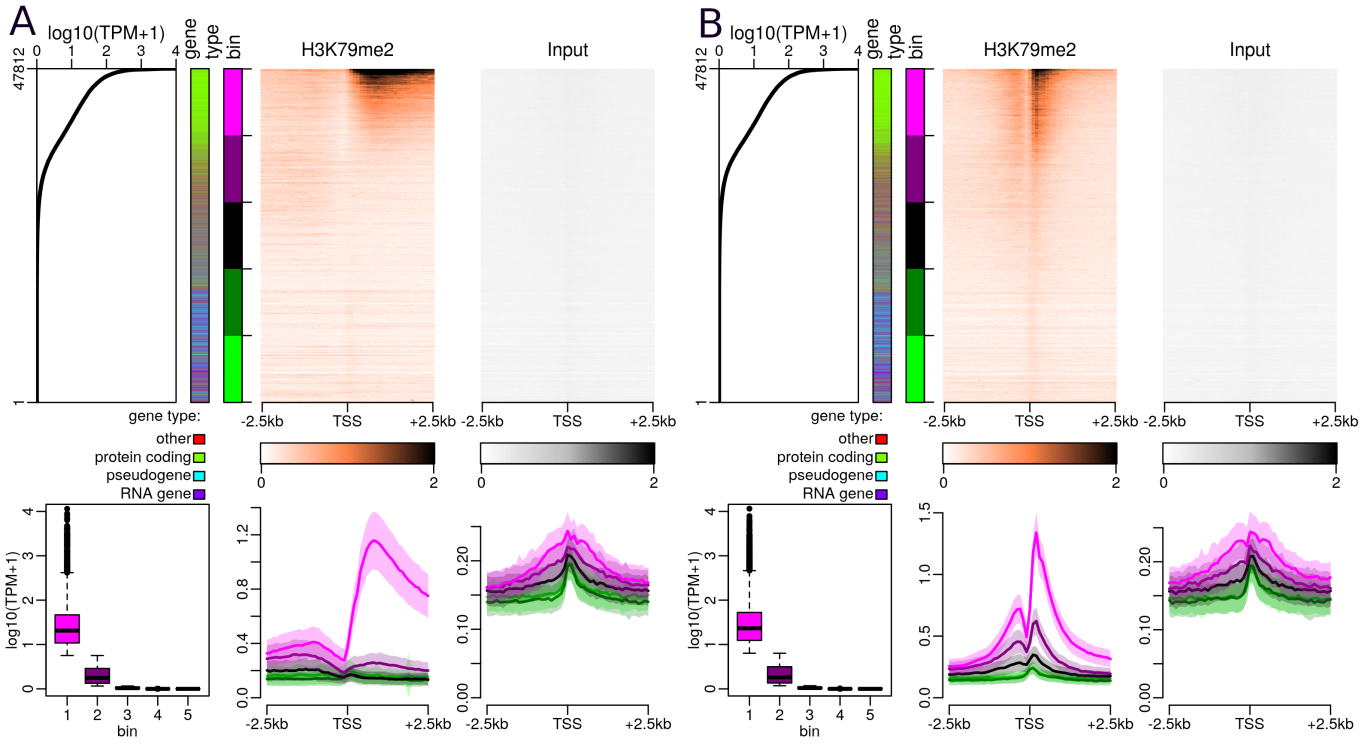
Supplementary Figure 3: H3K9me3 near the transcription start sites in the gastric tissue. Upper panel, from left to right: Gene expression level in all 47,812 genes annotated by Gencode. First side bar indicates the gene type (green: protein coding genes, blue: pseudogenes, purple: RNA genes, red: other types of genes), the second side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H3K9me3 ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H3K9me3 ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.



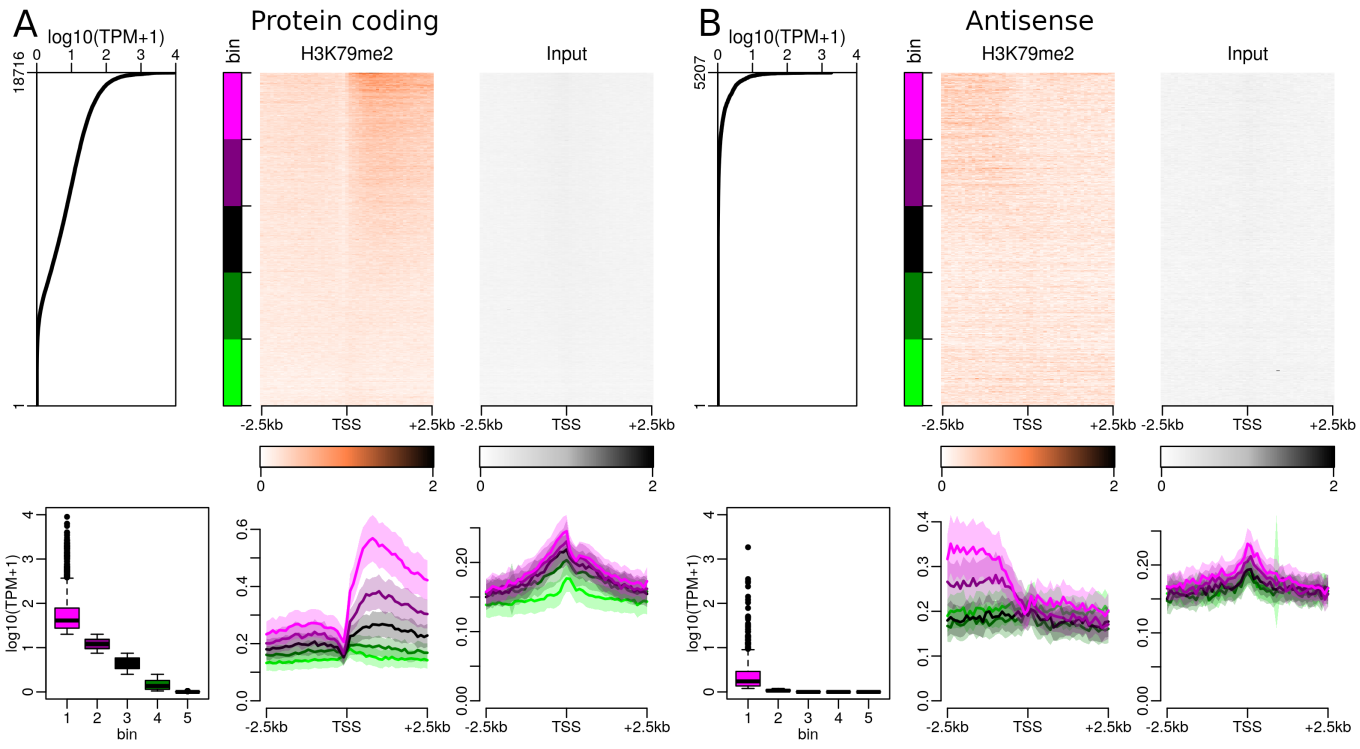
Supplementary Figure 4: H3K27me3 near the transcription start sites in the gastric tissue. Upper panel, from left to right: Gene expression level in all 47,812 genes annotated by Gencode. First side bar indicates the gene type (green: protein coding genes, blue: pseudogenes, purple: RNA genes, red: other types of genes), the second side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H3K27me3 ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H3K27me3 ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.



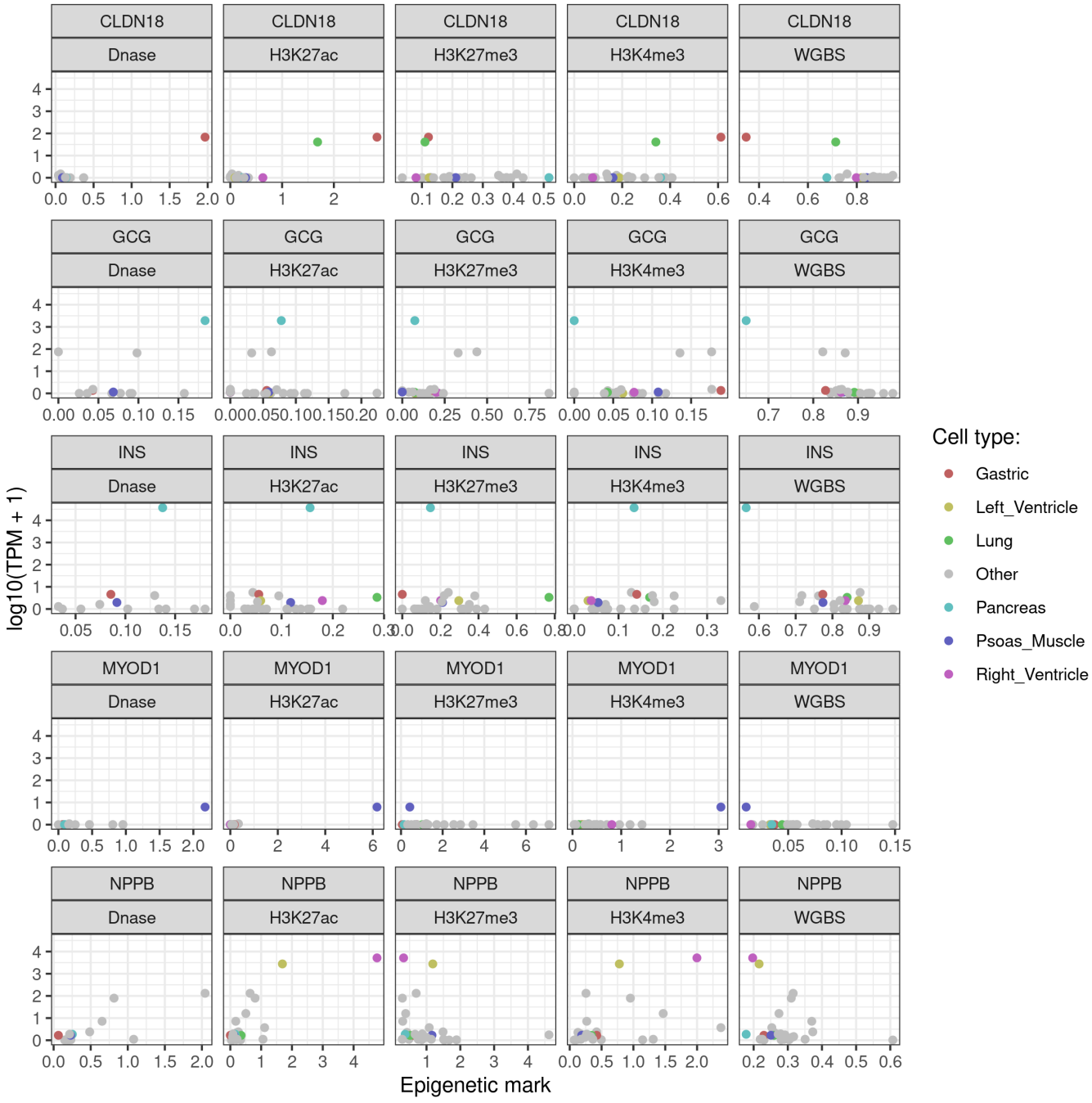
Supplementary Figure 5: H3K36me3 near the transcription start sites in the gastric tissue. Upper panel, from left to right: Gene expression level in all 47,812 genes annotated by Gencode. First side bar indicates the gene type (green: protein coding genes, blue: pseudogenes, purple: RNA genes, red: other types of genes), the second side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H3K36me3 ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H3K36me3 ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.



Supplementary Figure 6: H3K79me2 near the transcription start sites in H1-BMP4 derived mesoderm (A) or H1-BMP4 derived throphoblast (B). Upper panel, from left to right: Gene expression level in all 47,812 genes annotated by Gencode. First side bar indicates the gene type (green: protein coding genes, blue: pseudogenes, purple: RNA genes, red: other types of genes), the second side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H3K79me2 ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H3K79me2 ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.

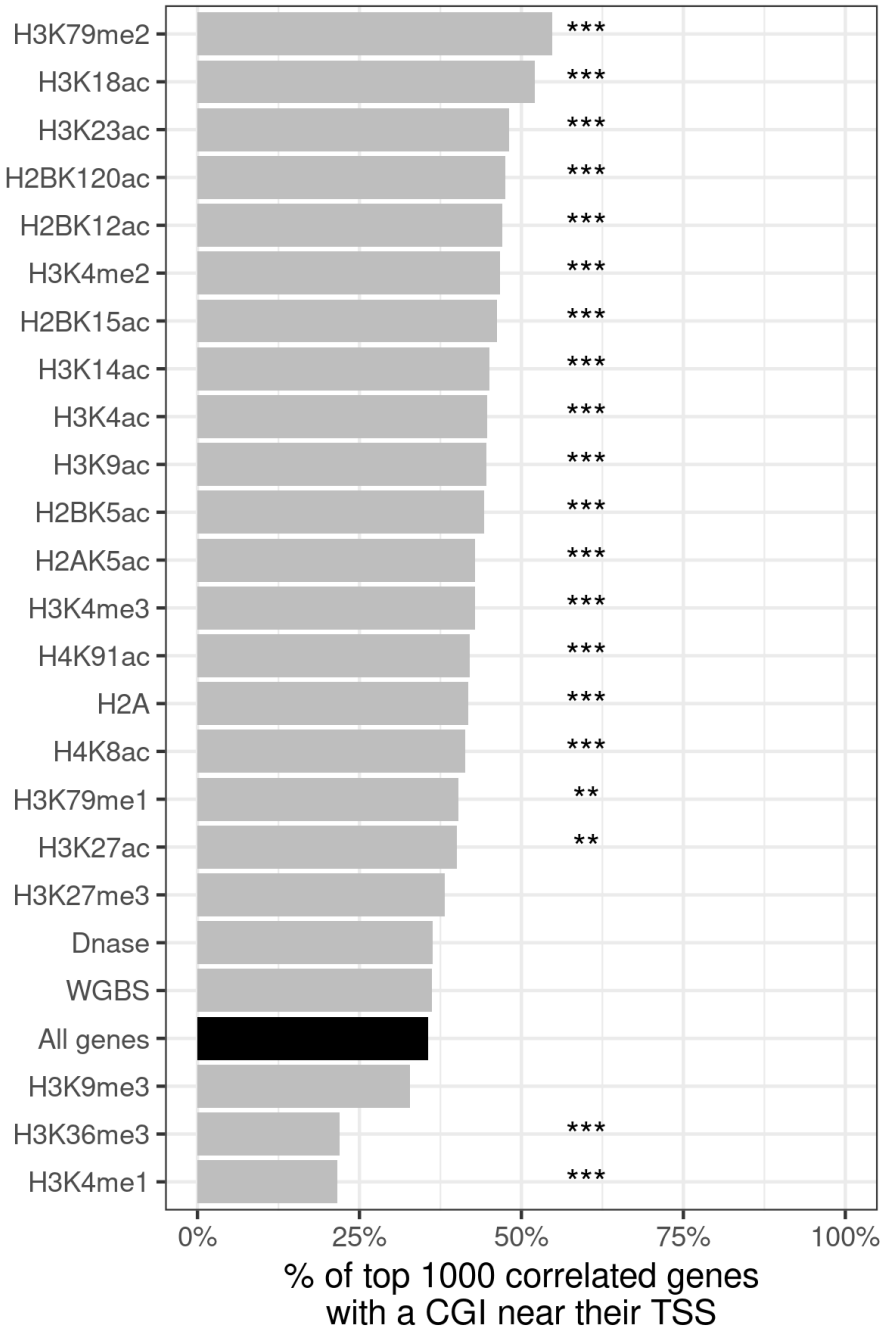


Supplementary Figure 7: H3K79me2 near the transcription start sites of protein coding genes (A) or antisense genes (B) in H1 cell line. Upper panel, from left to right: Gene expression level in all 18,716 protein coding (A) or 5,207 antisense (B) genes annotated by Gencode. The side bar indicates the genes sorted according to expression level (5 bins in total, purple: highly expressed genes, green: lowly expressed genes). Stacked profiles of H3K79me2 ChIP-seq and respective input control, sorted according to the corresponding gene expression level. Bottom panel, from left to right: Boxplot of gene expression levels in each of the 5 expression bins defined in the upper panel. Average profiles of H3K79me2 ChIP-seq and respective input control, $\pm$ SEM (Standard Error of the Mean) for each bin of promoters.

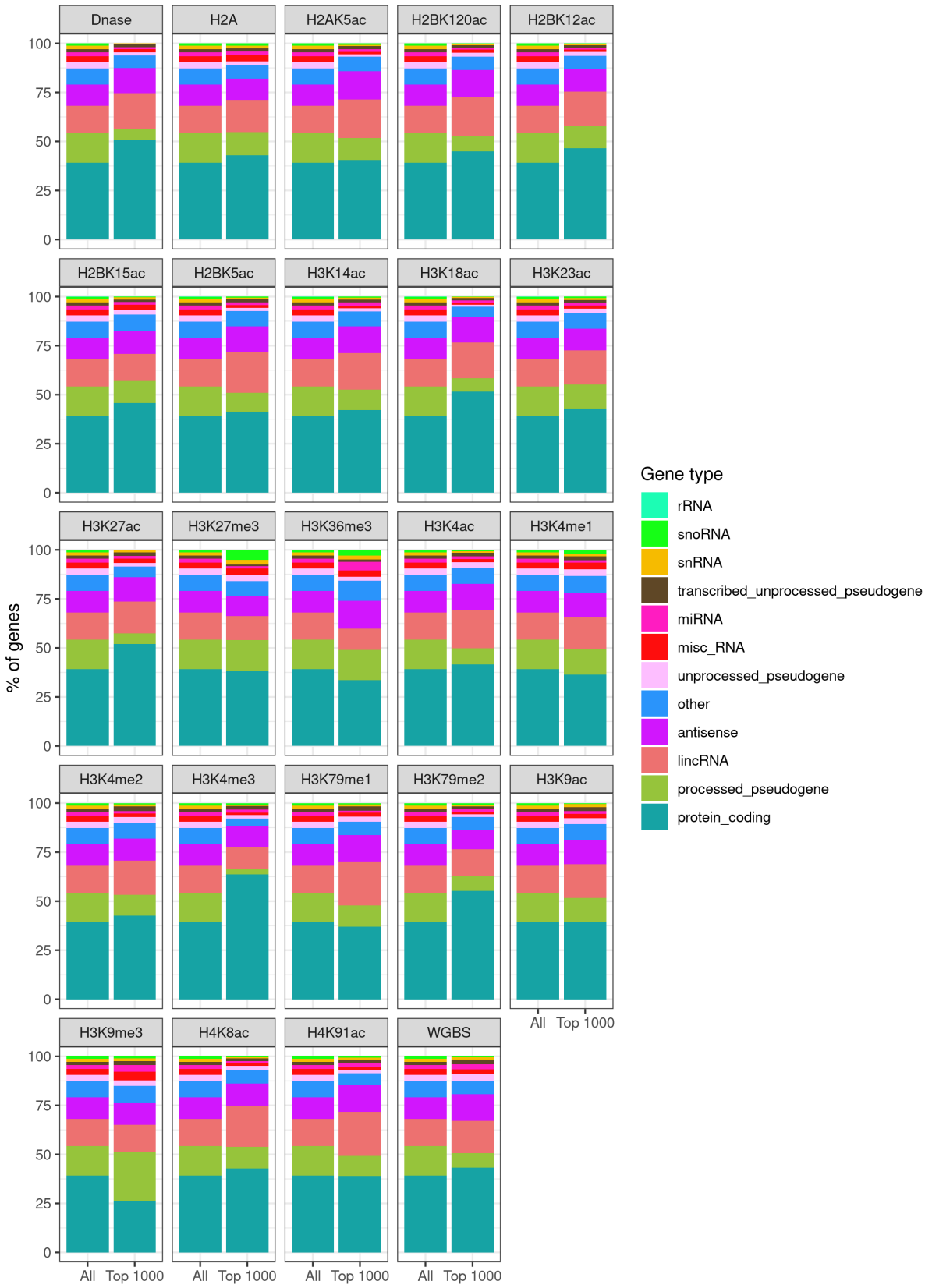


Supplementary Figure 8: associations between epigenetic marks near the TSS and expression levels of 5 tissue-specific genes. Each dot represents value for a cell or tissue type in the datasets.

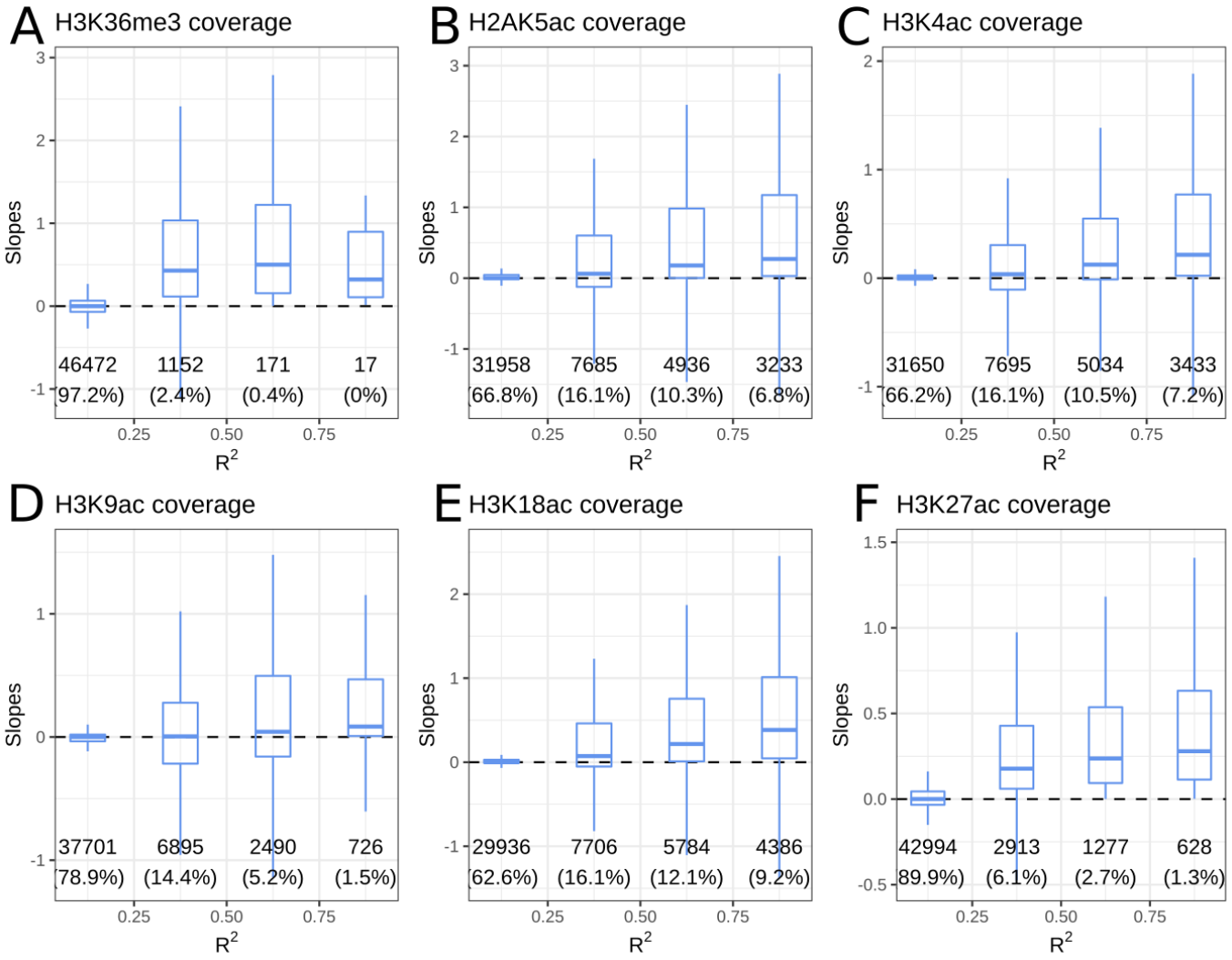
Supplementary Figure 9: Gene ontology analysis of the top 1000 genes with the highest R^2 values of correlation between gene expression levels and each epigenetic marks at *pm* 500 bp from their TSS.



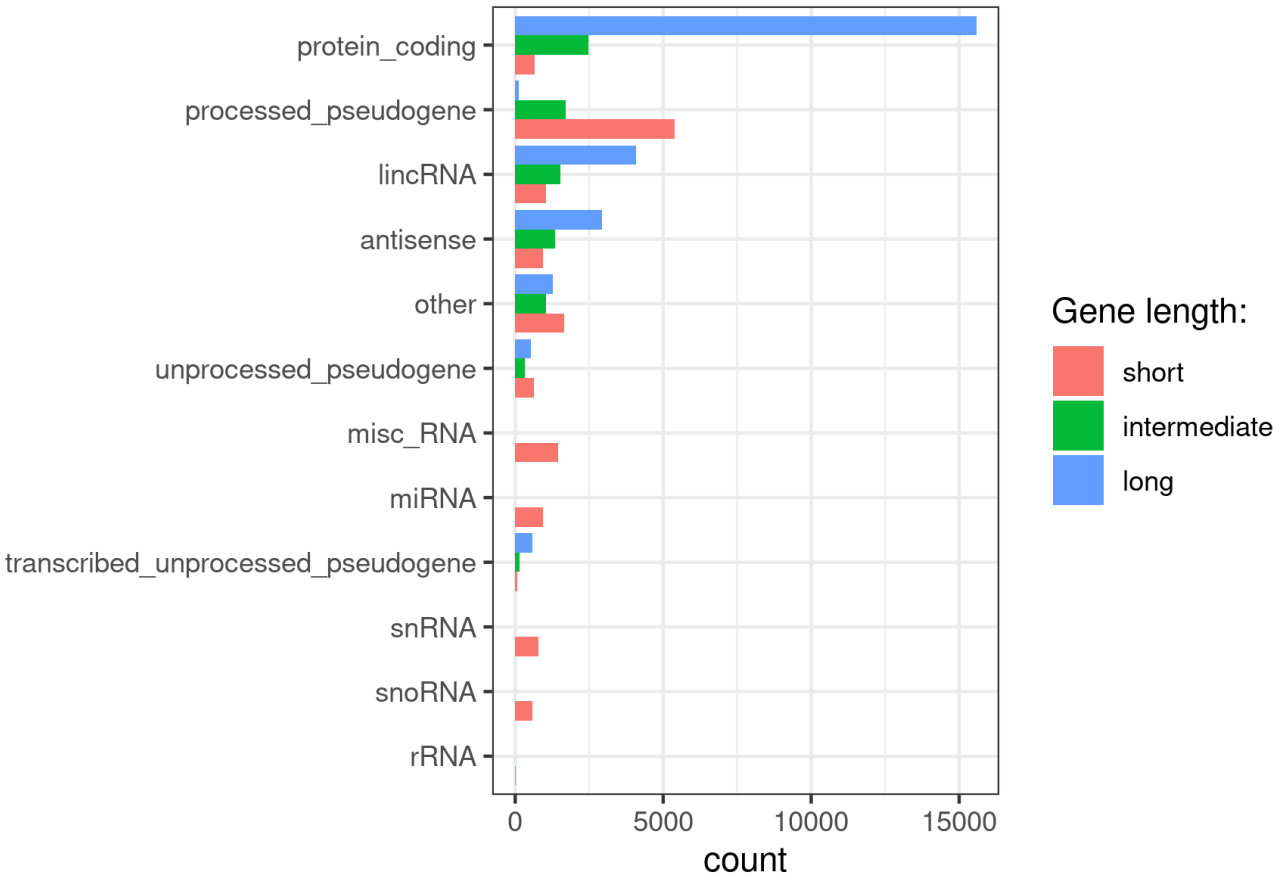
Supplementary Figure 10: Percentage of genes with a CpG island near their promoter along the top 1000 genes with the highest R^2 values of correlation between gene expression levels and each epigenetic marks at *pm* 500 bp from their TSS. Stars reflects p-values from a proportion tests against all genes. **: p-values < 0.01. ***: p-values < 0.001.



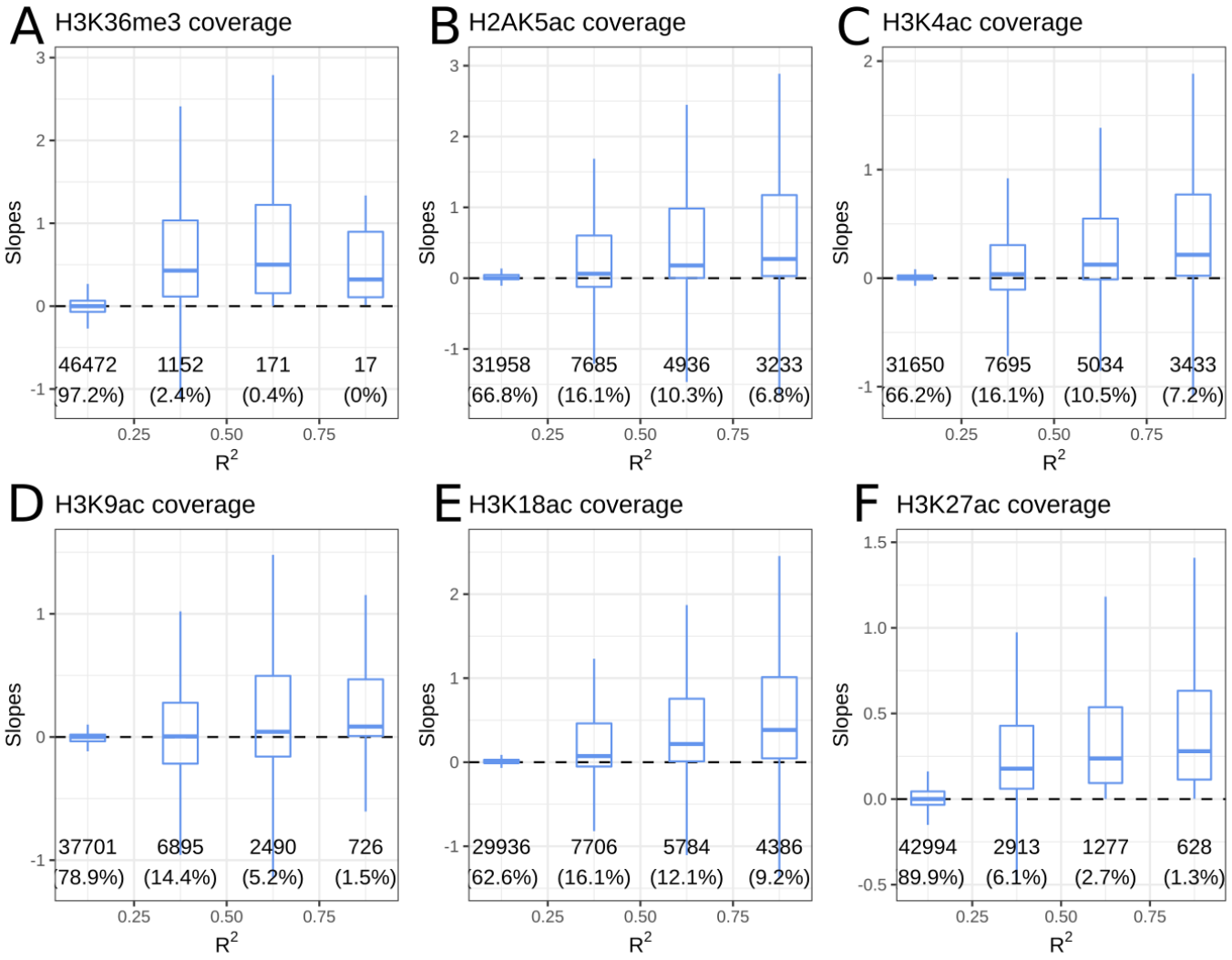
Supplementary Figure 11: Percentage of genes of each gene types along all genes (left bars) or the top 1000 genes with the highest R^2 values of correlation between gene expression levels and each epigenetic marks at pm 500 bp from their TSS (right bars).



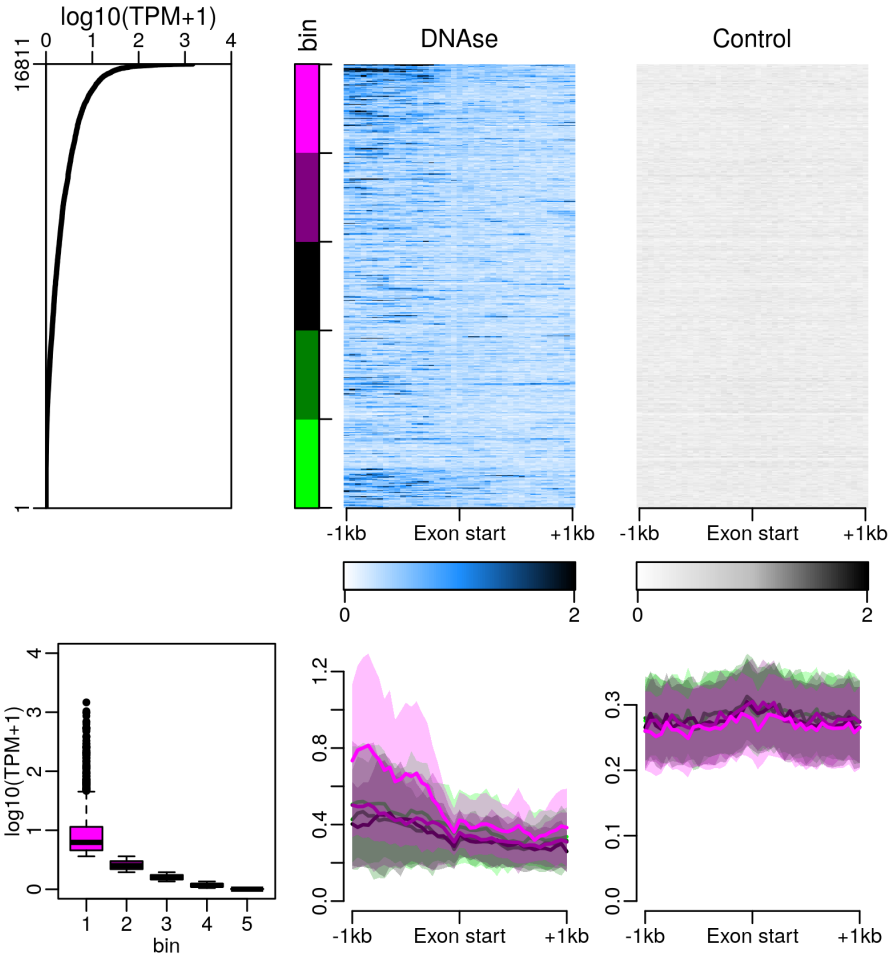
Supplementary Figure 12: Regressions between gene expression levels and epigenetic marks near the TSS of genes. One linear regression was performed for each individual genes, and the slope values of all the regressions are displayed according to their corresponding correlation coefficient R^2).



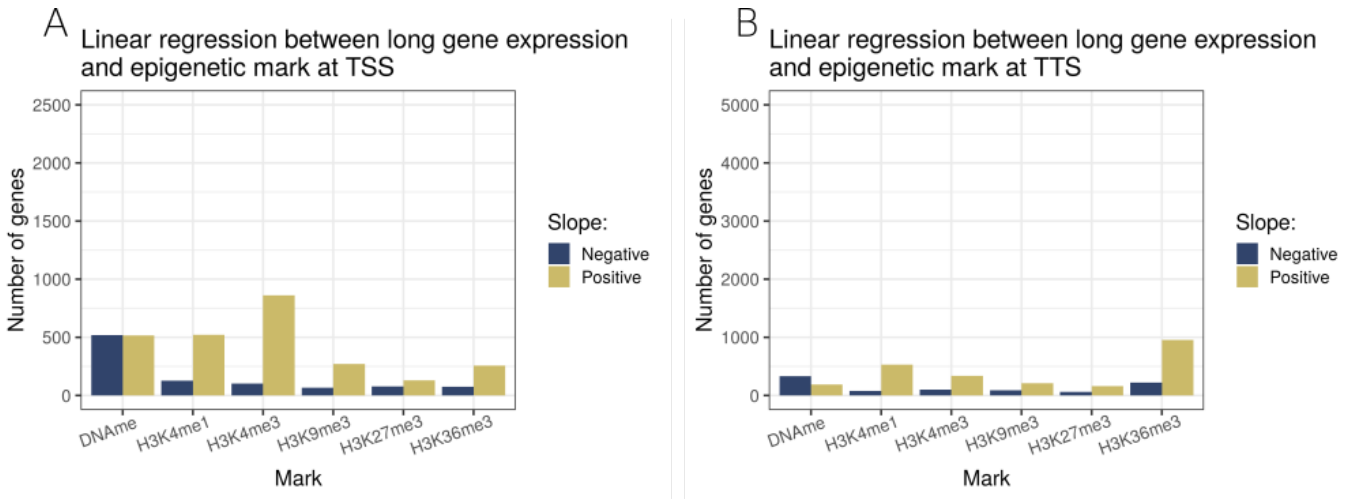
Supplementary Figure 13: Number of short ($\leq 1kb$), long ($> 3kb$), and intermediary genes depending on the GENCODE gene type.



Supplementary Figure 14: Regressions between gene expression levels and epigenetic marks near the TTS of long genes. One linear regression was performed for each individual genes, and the slope values of all the regressions are displayed according to their corresponding correlation coefficient R^2).



Supplementary Figure 15: DNAse1 profile around middle exon start sites in gastric tissue show a narrow decrease in DNAse accessibility near splicing acceptor sites. Upper panel, from left to right: middle exon expression levels in 16,811 middle exons. The side bar indicates 5 bins used in the figure S3 bottom panels (purple: highly expressed exons, green: lowly expressed exons). Then: Stacked profiles of DNAse1 profile and control, sorted according to the exon expression levels. Bottom panel, from left to right: Boxplot of exon expression levels in each of the 5 bins defined in the upper part. Then, average DNAse profiles and control $\pm$ SEM (Standard Error of the Mean) for each bin of exons.



Supplementary Figure 16: Linear regression models including six epigenetic modifications characterised in 27 cell types, focusing on the 25,068 long genes (> 3kb, amongst 47,812 analysed genes). Each bar represents the number of genes with a statistically significant slope ($p <= 0.01$), either positive (golden) or negative (deep blue). **A.** Linear regression model of gene expression level and the levels of 6 epigenetic modifications near their respective TSS (±500 bp) for long genes only. **B.** Linear regression model of gene expression level and the levels of 6 epigenetic modifications near their respective TTS (±500 bp) for long genes only.