Prediction of genes under dynamic post-transcriptional regulation from time-series epigenomic data

Supplementary Information

SUPPLEMENTARY TABLES

Supplementary Table S1. Primer sequences for real-time quantitative PCR. The utilized forward and reverse primer sequences for the RT-qPCR and ChIP-qPCR quantification of genes of interest are listed. The primer design had been performed using the primerExpress3.0 software.

Gene	**Forward (5'→3')**	**Reverse (5'→3')**
Akr1c14	TTTCAAGGAGAAGCGGATCAA	TCATGTCCTCTGAAGCCAACTG
Trim12c	GGCTGCATCACGTTGTACTTTG	CCACATGTAGATTAGGCCTCAGATT
Lamin	CGCACCGCTCTCATCAACT	CCTCATTGTCCTCAACCATGGT
Rpl13a	TGGTCCCTGCTGCTCTCAA	CCCCAGGTAAGCAAACTTTCTG
Spon2 (TSS)	TGTGTCCCCTGCACTGTGA	TGCACAGCTAGTCCATGAAGTGT
Spon2 (TES)	CAGGGTCCCCTTGAAGCAA	TCCTAGCCCTGCTCCATCAG
Akr1c14 (TSS)	GTGGTACTAAACGATGGTCACTTCA	CGGGCACAGTGGTTCCA
Akr1c14 (TES)	TCCAGCAACCCAAGAACCA	CCAAATTTCCCACAAGGAGAAC
Pou5f1 (TSS)	GTCCAGACGTCCCCAACCT	AAACTGAGGCGAGCGCTATC

SUPPLEMENTARY FIGURES

Supplementary Figure S1. Data processing and distributions. (A) Read counts of the respective data variables were normalized and $\log_2$-transformed prior to distribution-based gene filtering. The filtering is visualized in (B-E). Data distribution of log2-transformed values (assumed pseudo count = 0.001) of each of the data types is shown in form of a kernel density estimate per time point, before (B-C) and after (D-E) distribution-based data exclusion for both time course experiments. This assay- and time point-specific estimation of the data distribution was used to determine a cut-off for exclusion. Chosen cut-offs are indicated in form of grey dashed lines. Any data point (corresponding to a genomic locus with its several associated time course seq-data values) that was found below at least one of these assay- and time point-specific cut-offs, was excluded from the dataset. On basis of this, from the 43,274 genes in the original assembly, 23,572 genes were excluded for the data collection of adipocyte differentiation, and 23,334 for that of osteoblast differentiation. The units of the four displayed data types are Reads per Million (RPM) values for the H3K4me3- and H3K27ac-targeted ChIP-Seq data and H3K27ac calculated for 4 kb windows around the TSS, Reads per Kilobase Million (RPKM) for the H3K36me3-targeted ChIP-seq data as a gene length-normalized measure calculated between TSS to TES of each genomic locus, and finally Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for the RNA-seq value of each genomic locus.

Supplementary Figure S2. Variable correlation and model selection. (A-B) Pairwise correlations between RNA-seq and histone mark ChIP-seq variables in (A) adipocyte and (B)

osteoblast differentiation time course. As a representative timepoint, D0 data is shown. The density distribution of each variable is displayed in the diagonal of the figure grid and the Pearson correlation coefficient for each variable pairing is given in the upper portion, with a corresponding scatter plot found in the lower portion of the grid. (C-D) Adjusted r^2 statistic of all possible combinations of predictor variables using available data of both time course experiments. Again, D0 is shown as a representative time point. Model selection resulted in the use of H3K4me3 and H3K36me3 ChIP-seq variables as predictors of RNA-seq expression values.

Supplementary Figure S3. Time point-wise regression fits. (A-B) Time point-wise linear regression model of RNA-seq expression during adipocyte and osteoblast differentiation. Model fits of each data collection timepoint of both time courses are displayed with measured RNA-seq values over its model-predicted counterparts. Both data are given in $\log_2$-transformed FPKM values and r^2 values are indicated in the top left corner of each subplot.

Supplementary Figure S4. Genes with a high *sdR* value in the osteoblastogenesis dataset - additional examples. Data associated to three example genes is shown. Each featuring a high dynamic change of the agreement between their measured and predicted RNA-seq levels across the time course of osteoblast differentiation. Displayed are their measured (solid line) and model-predicted (dashed line) RNA-seq values (log2 FPKM) over time, during osteoblastogenesis. The respective *sdR* value is indicated in the top left corner of each subplot.

Supplementary Figure S5. Average intensity of H3K36me3 and H3K4me3 enrichments across adipocyte and osteoblast differentiation time courses. Average intensities of histone modifications H3K36me3 and H3K4me3 are plotted on the metagene - gene body normalized to 10 kb length and considering 2.5 kb regions up- and downstream of the TSS and TES, respectively. Displayed gene sets are all genes (using the genecode file release M5 of the mouse genome GRCm38.p8), genes with the top 5 % highest *sdR* values in the osteoblast differentiation, and adipocyte differentiation analysis. Each of these sets is shown across both differentiation time courses.

Supplementary Figure S6. Subsetting low *sdR* genes in two differentiation time courses. Related to Figure 2. (A-B) For this analysis, the bottom 5% of genes with the lowest *sdR* values were subset for both time course data sets (in contrast to top 5% used in Figure 2). Shown are density distributions of *sdR* values within the genes that reach above 0 log2 FPKM at least once in of the two time courses, and the bottom 5 % cut-off is visualized (C) Venn analysis of the two resulting gene lists exhibited a relatively small overlap. (D-E) Proportions of the coding and non-coding transcript types within these two lists are displayed side by side with the gene sets from the Figure 2. (F-H) Different GO terms of the types of biological process that were enriched among (F) adipogenesis- and (G) osteoblastogenesis-specific as well as (H) lineage-overlapping bottom 5% lowest *sdR* gene lists.

Supplementary Figure S7. Density distribution of 3'UTR length and miRNA binding sites. Displayed are three gene sets, including all genes for which 3'UTR information could be retrieved

(green), and such subsets of the top and bottom 5% *sdR* value genes (yellow & black, respectively), in the adipogenesis and osteoblastogenesis set (top and bottom panel). (A) Using these gene sets, the density distribution of maximum 3'UTR length is shown (in case of multiple 3'UTR sequences associated to one gene, the longest was chosen). 3'UTR information had been retrieved from the Ensembl database from the mm10 version (GRCm38.79). (B) For the same gene sets, the number of predicted miR binding sites was retrieved for each gene (from TargetScanMouse Release 7.2 August 2018, "Predicted Conserved Targets") and the resulting density distribution in each gene set is shown. Note that this analysis excluded any genes without associated 3'UTR info, while it included any of those genes for which no miRNA binding site had been predicted.

Supplementary Figure S8. RNA-seq and H3K4me3 and H3K36me3 ChIP-seq profiles of STEM based gene clusters. Related to Figure 3. The lineage-wise predicted targets of dynamic post-transcriptional regulation (the subsets with top 5% *sdR* values) were clustered by similarity in their temporal profile of model fit deviation (their residual values) using STEM (see Figure 3). The corresponding RNA-seq (A), H3K4me3 ChIP-seq (B), and H3K36me3 ChIP-seq (C) signals during adipocyte and osteoblast differentiation sets are shown. Most deviation from model fit derives from changes in RNA-seq, and not ChIP-seq signals.

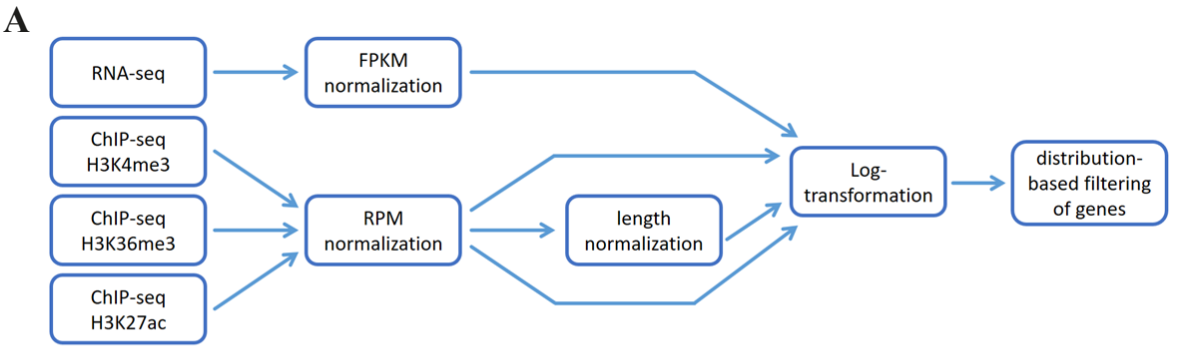
Supplementary Figure S9. *Il1rl1* locus and RT-qPCR validation of expression levels. Related to Figure 4D-E. (A) Histone modifications H3K4me3 and H3K36me3 at the genomic locus of the *Il1rl1* gene at different time points of adipocyte differentiation. (B) *Il1rl1* transcript abundance during adipogenic and osteoblastogenic differentiation as measured by RT-qPCR. Plotted on the

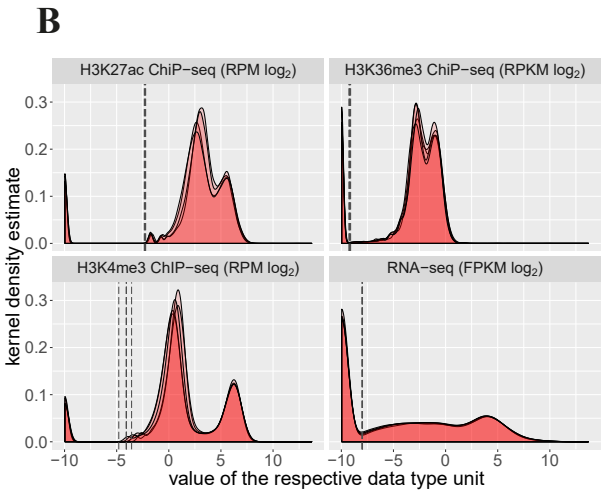
y-axis are the expression fold changes at indicated time points of differentiation relative to the undifferentiated state. Red bars and blue bars indicate adipocyte and osteoblast differentiation, respectively. The values represent the mean of three biological replicates +/- SEM. *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001 (one sample t-test).

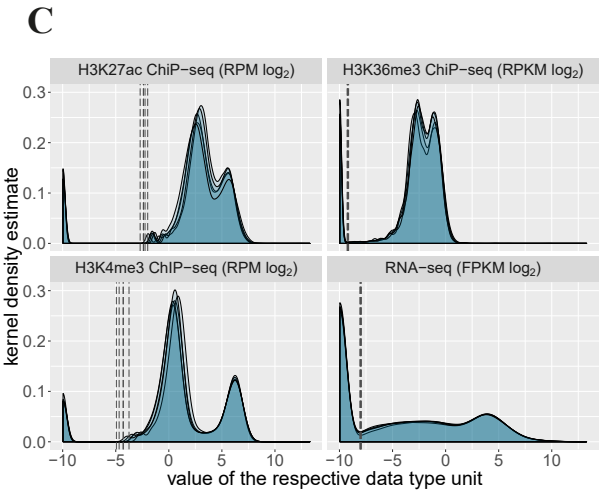
Supplementary Figure S10. ChIP-qPCR validation of H3K4me3 and H3K36me3 enrichment profiles during adipogenesis at selected loci. (A) *Akr1c14* locus-specific enrichment of H3K4me3 and H3K36me3 histone marks at TSS and TES sites, across time points of adipogenic differentiation, as validated by ChIP-qPCR. For more details, see panel D. (B) Histone modifications H3K4me3 and H3K36me3 at different time points of adipocyte differentiation at the genomic locus of the *Spon2*, a gene predicted not to be under post-transcriptional control. (C) The measured (solid line) and model-predicted (dashed line) RNA-seq values (log2 FPKM) for *Spon2* over time during adipogenesis. (D) *Spon2* locus-specific enrichment of H3K4me3 and H3K36me3 histone marks at TSS and TES sites, across time points of adipogenic differentiation, as determined by ChIP-qPCR. Plotted on the y-axis is relative enrichment at indicated time points of differentiation compared to the undifferentiated cells. The values represent the mean of three biological replicates +/- SEM. *p-value < 0.05, **p-value < 0.01 (one sample t-test). For the TSS of both *Akr1c14* and *Spon2* the displayed data of day 1 of adipogenesis (AD1) is based on two biological replicates due to an undetected value in the third sample. For the same reason, for the TES of *Akr1c14* the displayed data of day 3 of adipogenesis (AD3) is based on only one biological replicate.

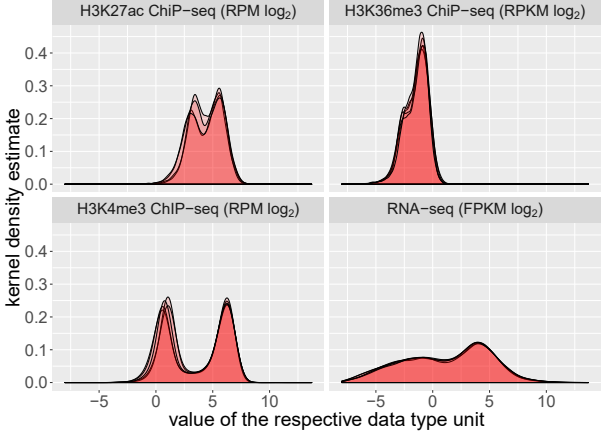
Supplementary Figure S11. *Rpl13a* gene expression upon siRNA and miRNA mimic transfection. Related to Figure 4C. Effect of transfection conditions on the *Rpl13a*, used as housekeeping gene in the RT-qPCR experiments. Displayed is the relative expression as a fraction of the respective control condition, following transfection with either si-*Lamin* (A), or miR-204 mimic (B) at 24 and 48 h post-transfection. The values represent the mean of three biological replicates +/- SEM.

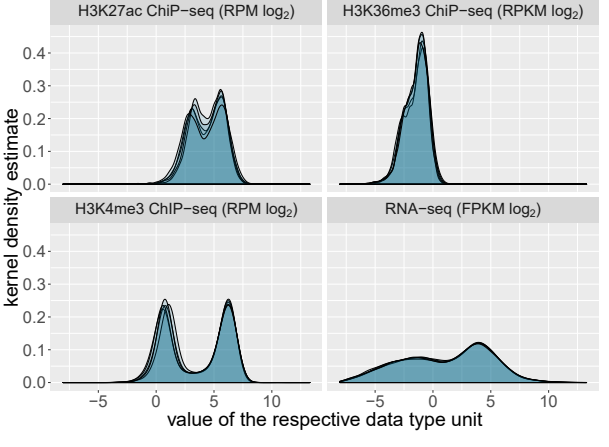
Supplementary Figure S1



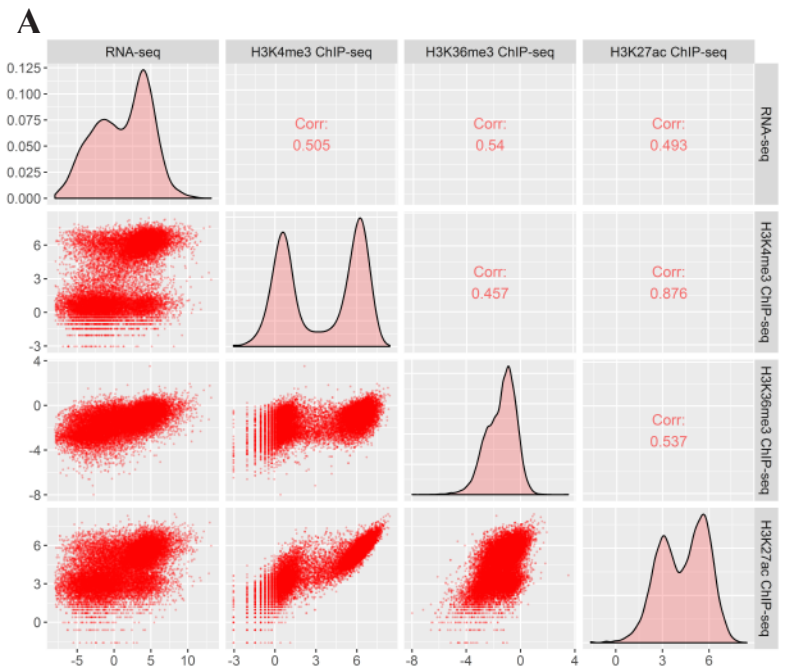


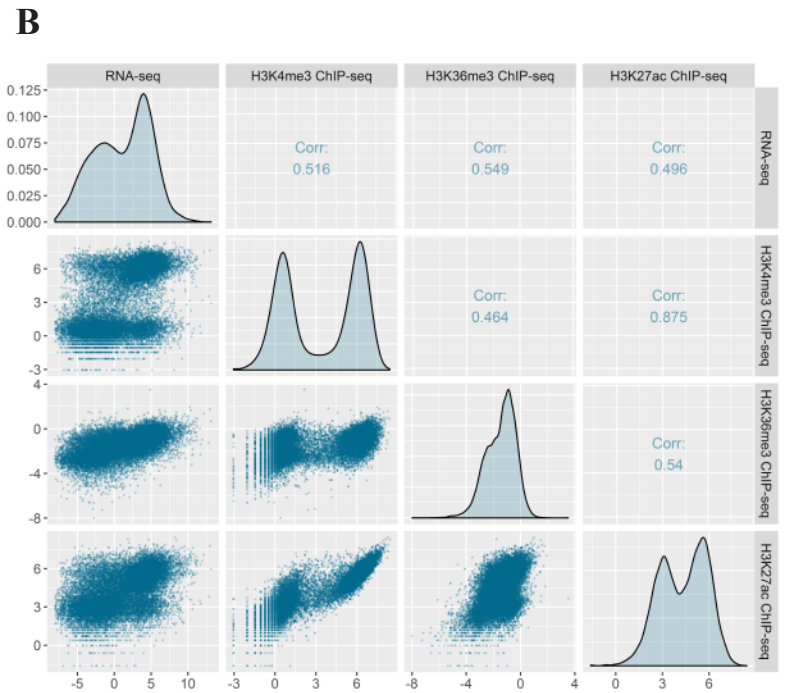


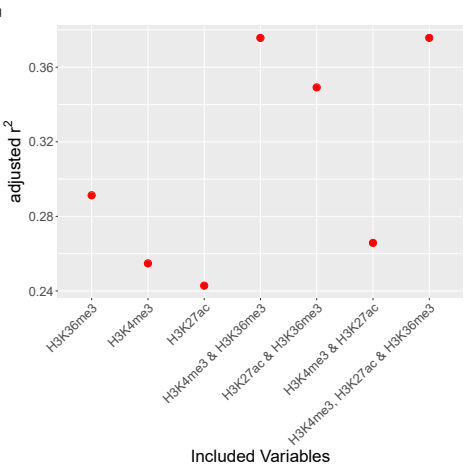


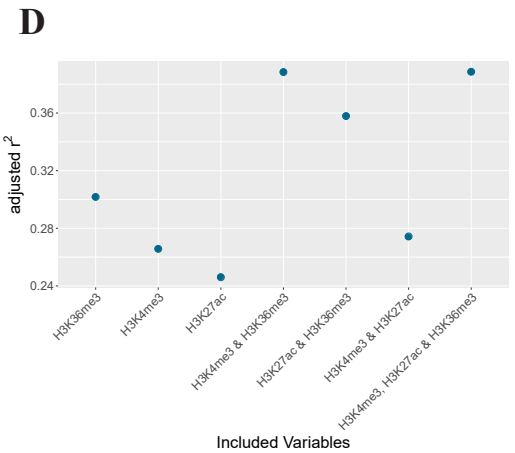


Supplementary Figure S2

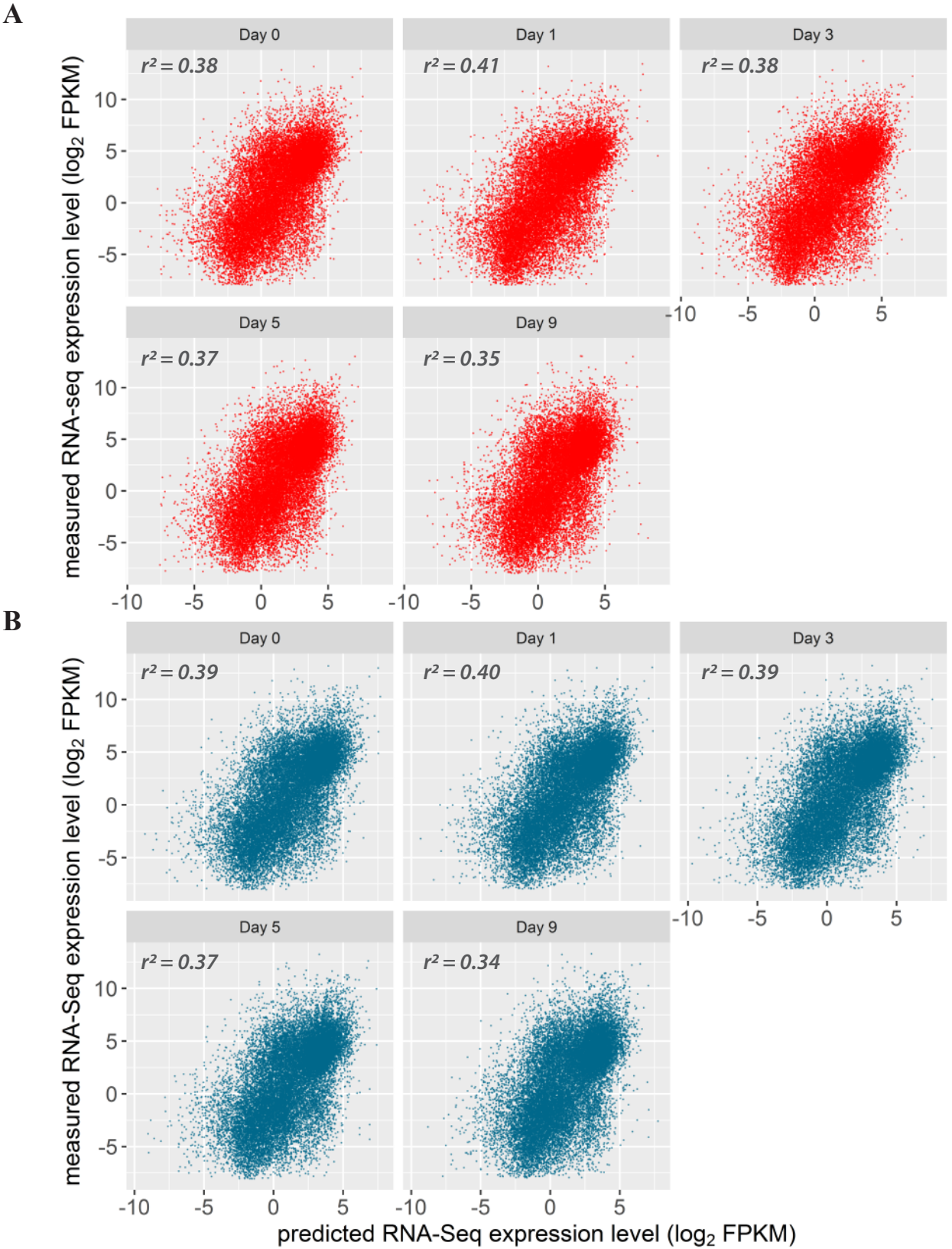




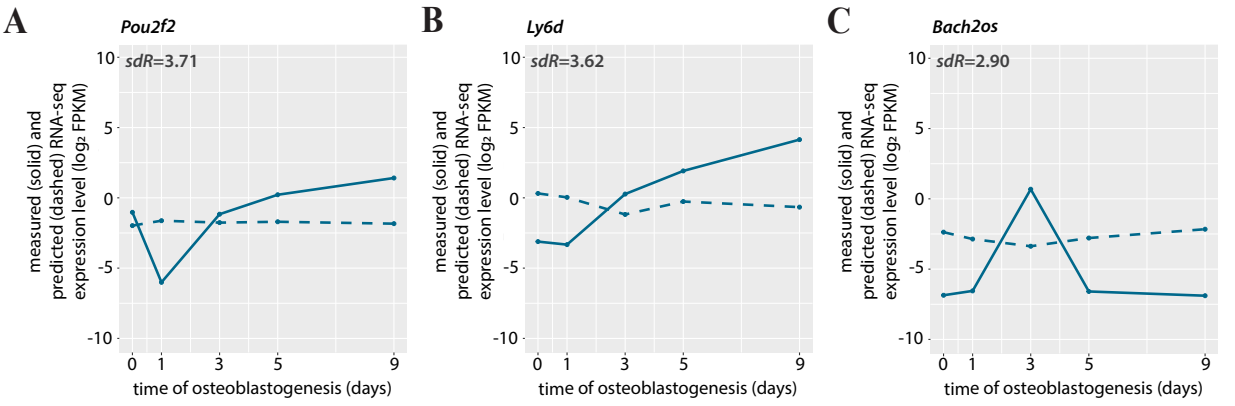




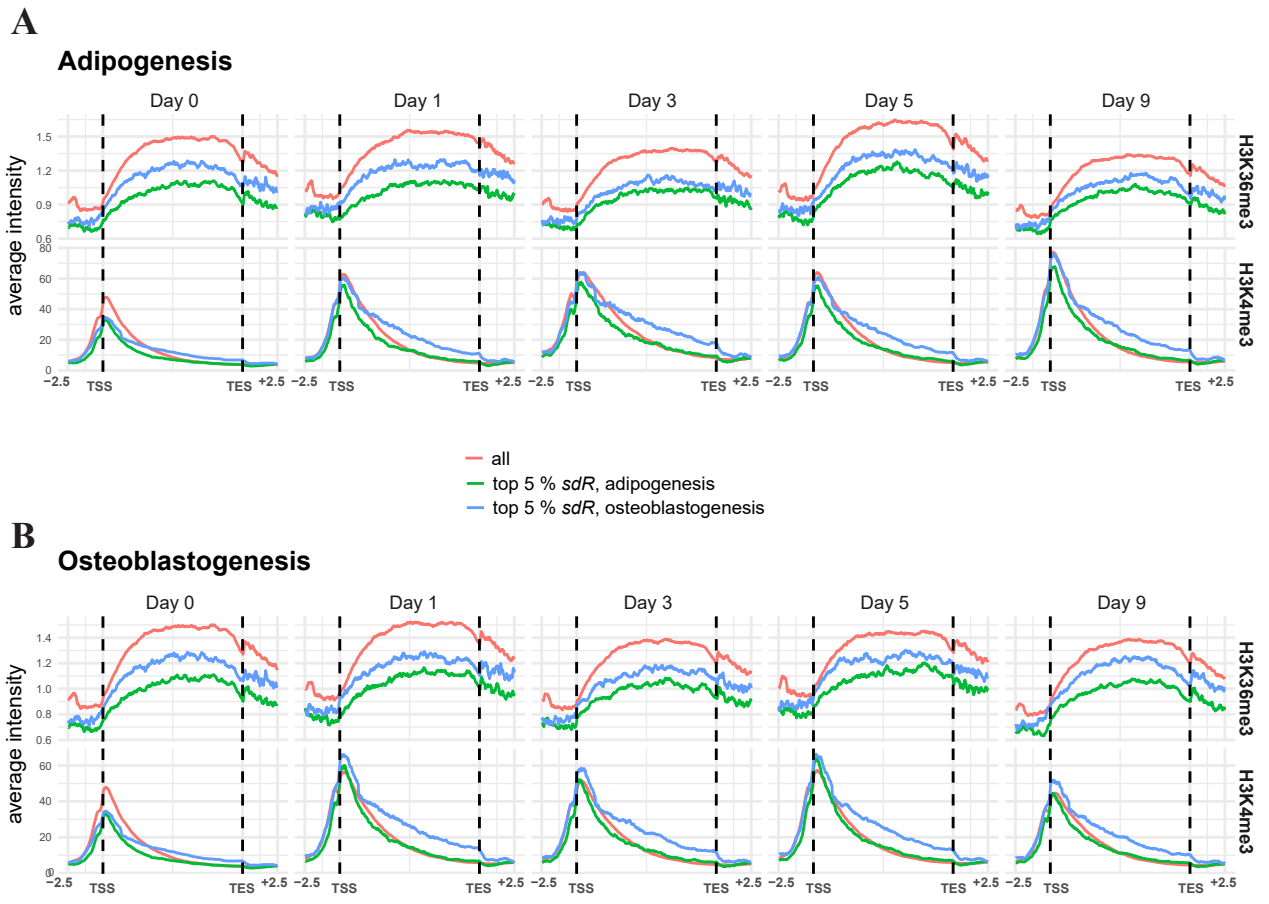
Supplementary Figure S3



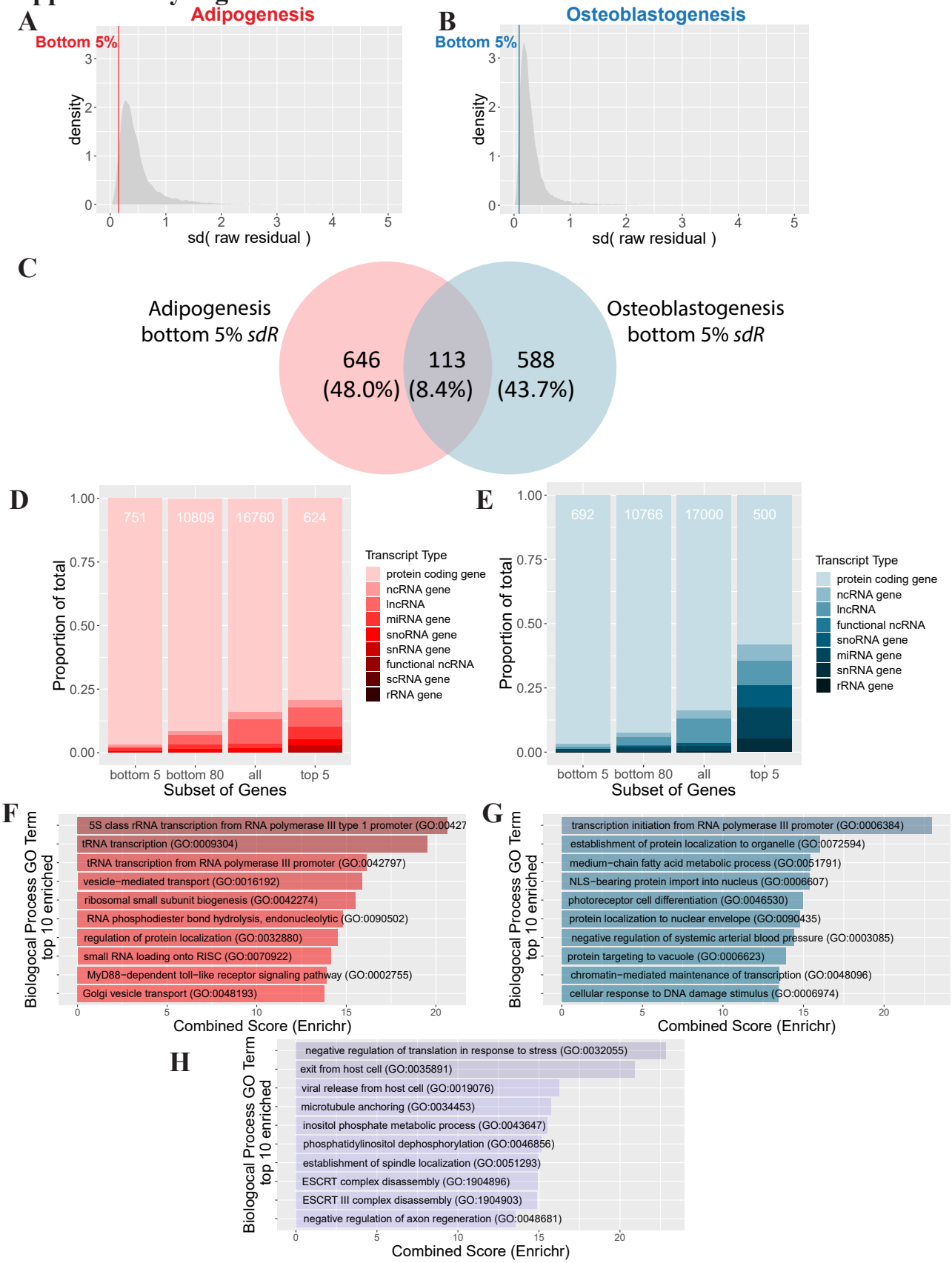
Supplementary Figure S4



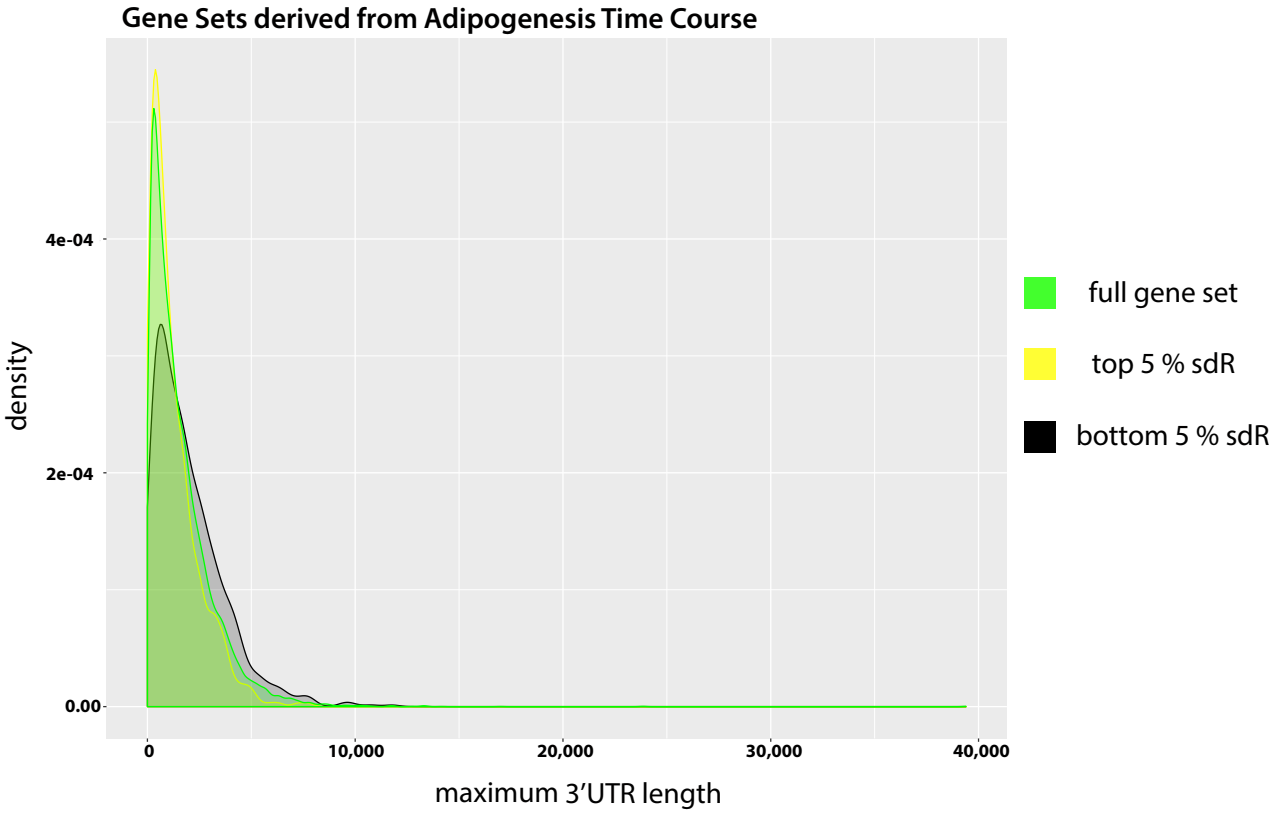
Supplementary Figure S5

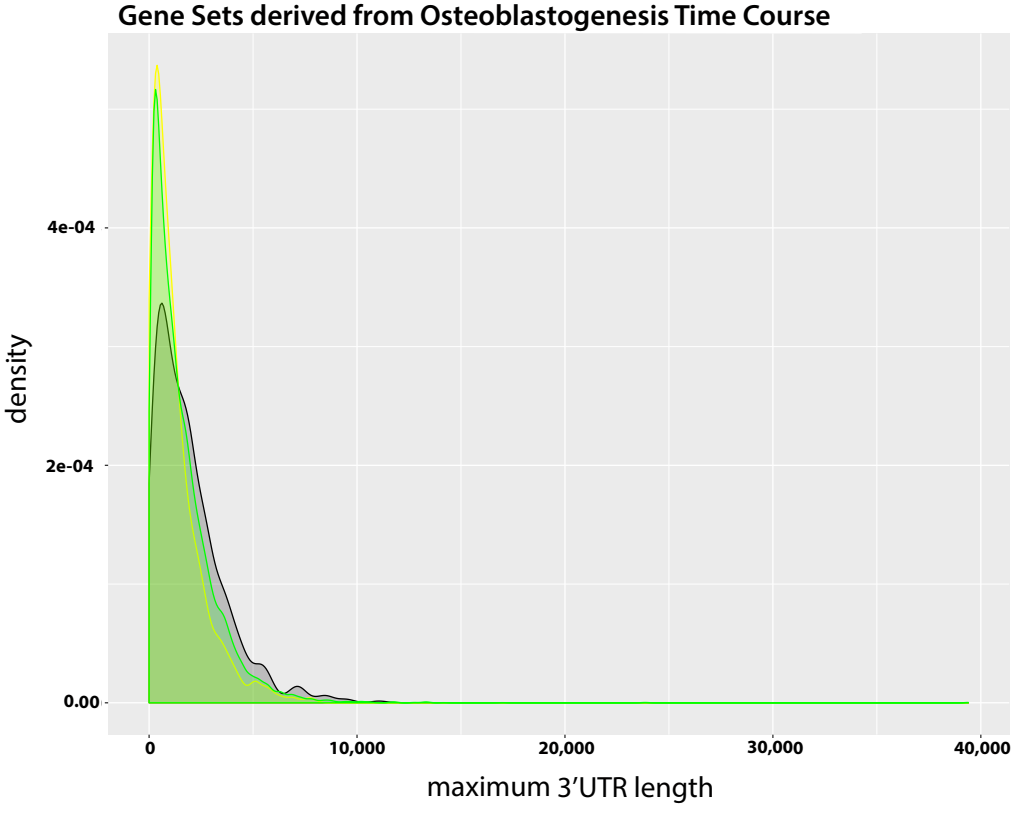


Supplementary Figure S6

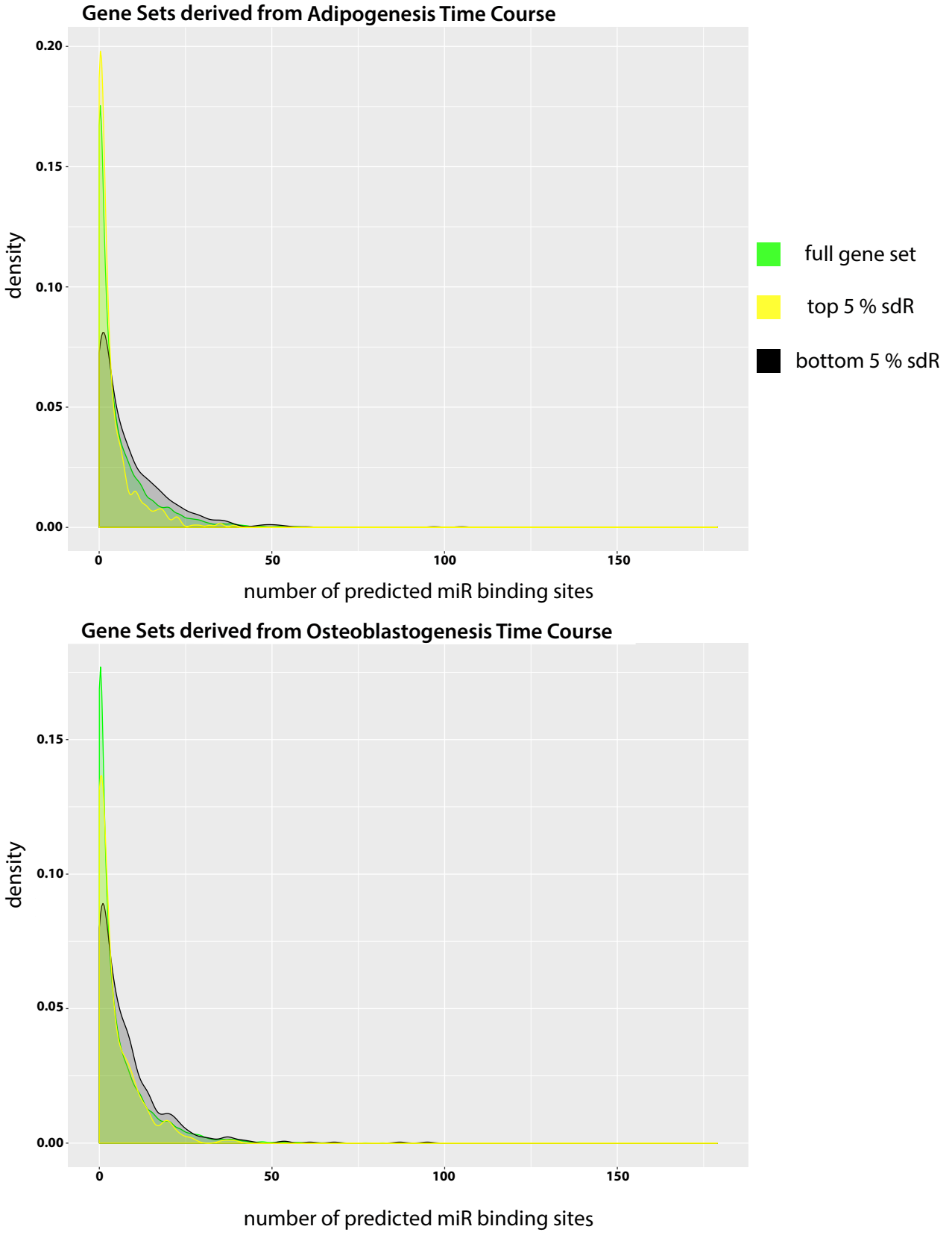


Supplementary Figure S7A

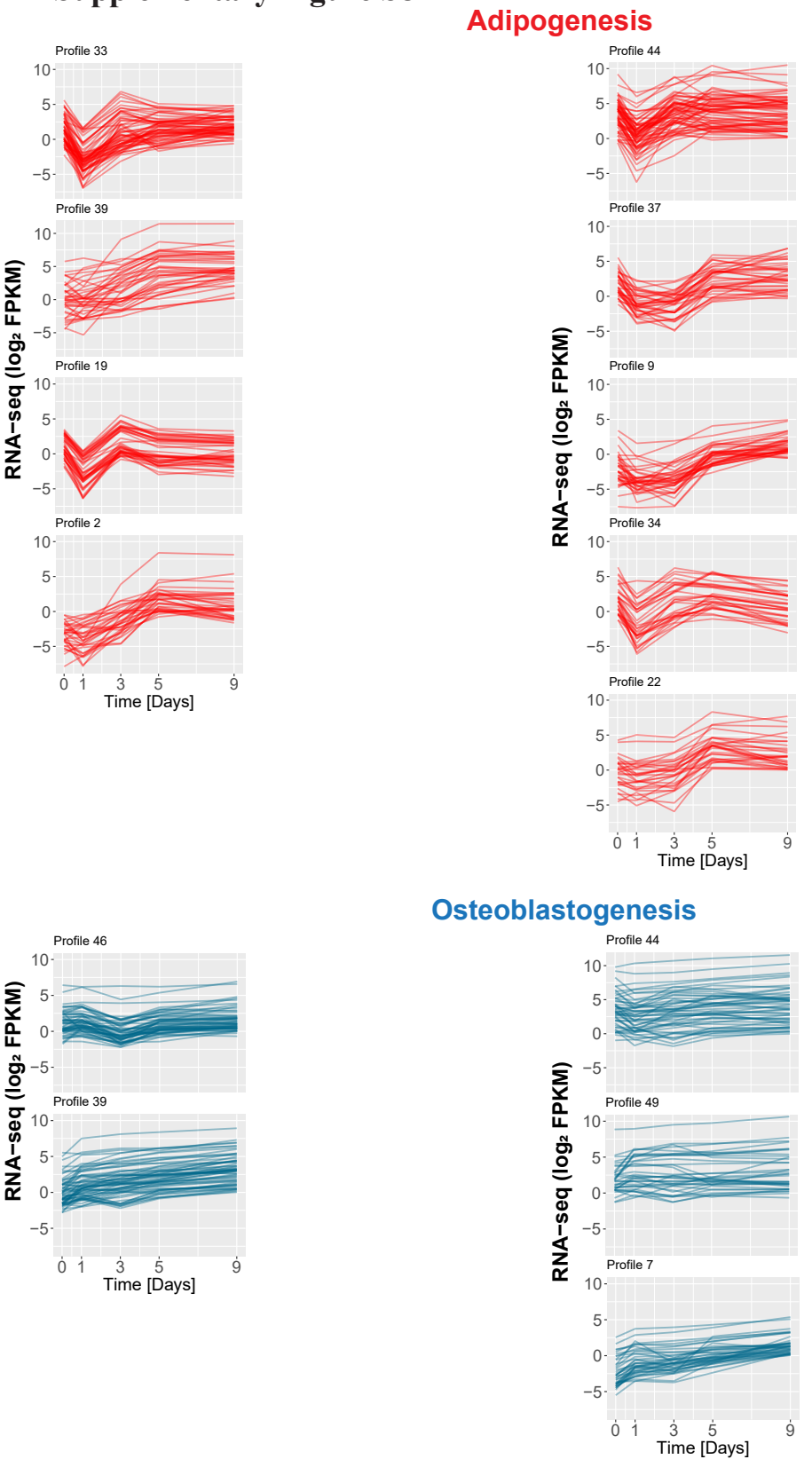




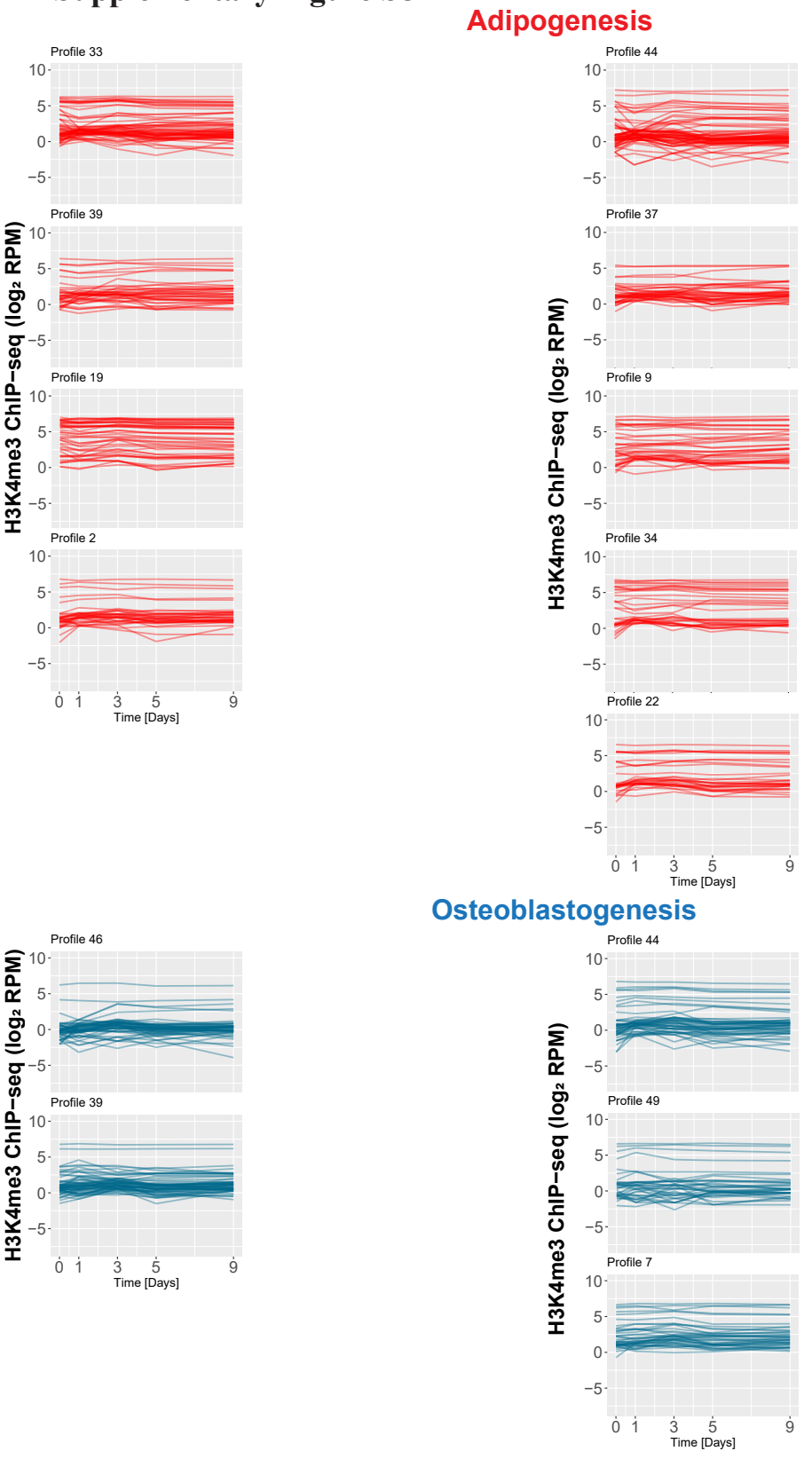
Supplementary Figure S7B



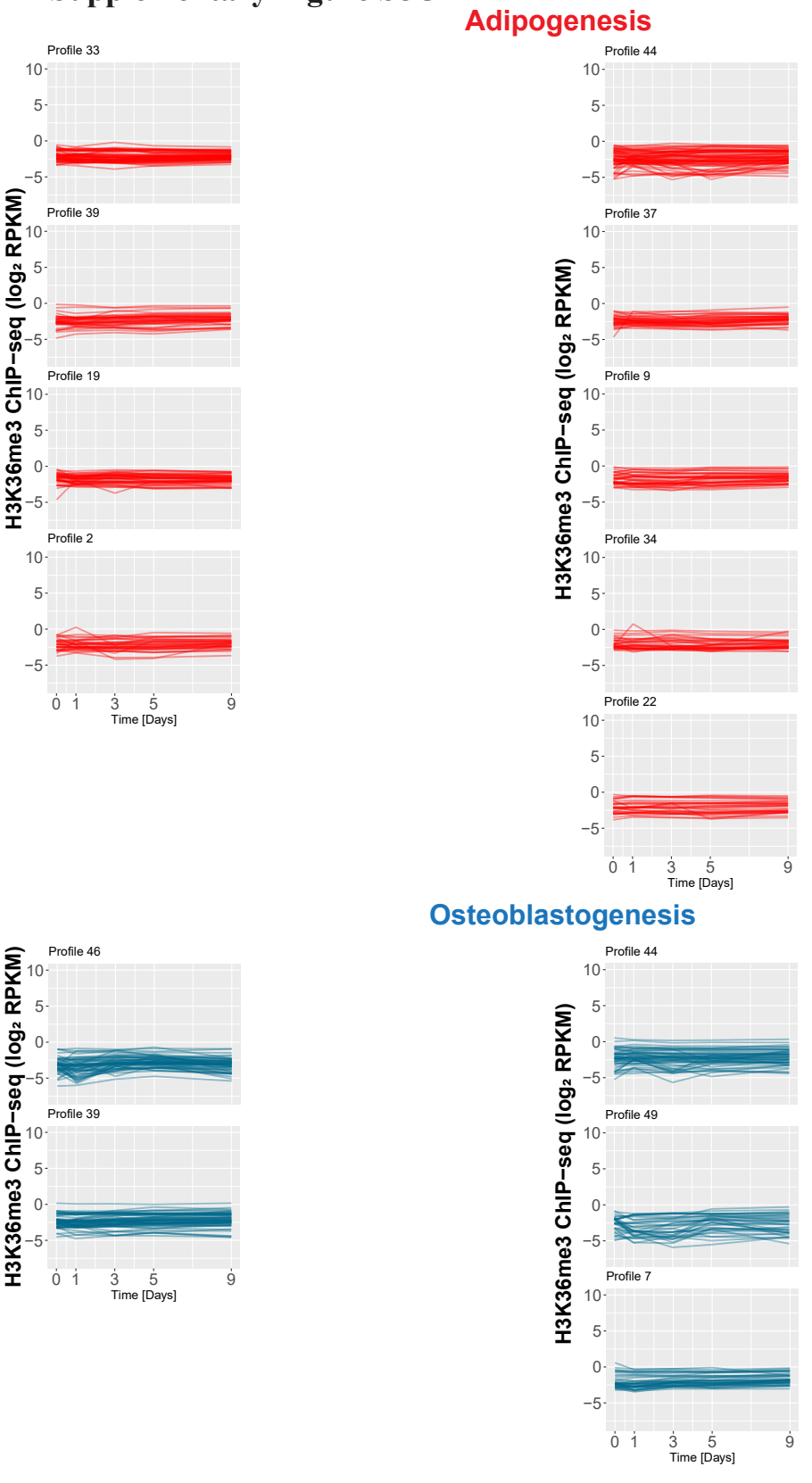
Supplementary Figure S8A



Supplementary Figure S8B

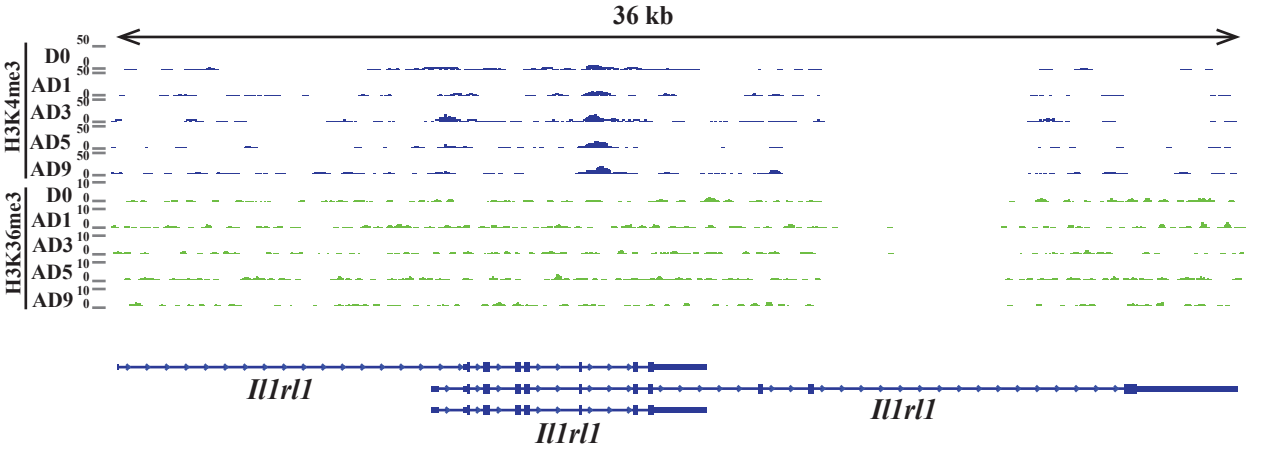


Supplementary Figure S8C

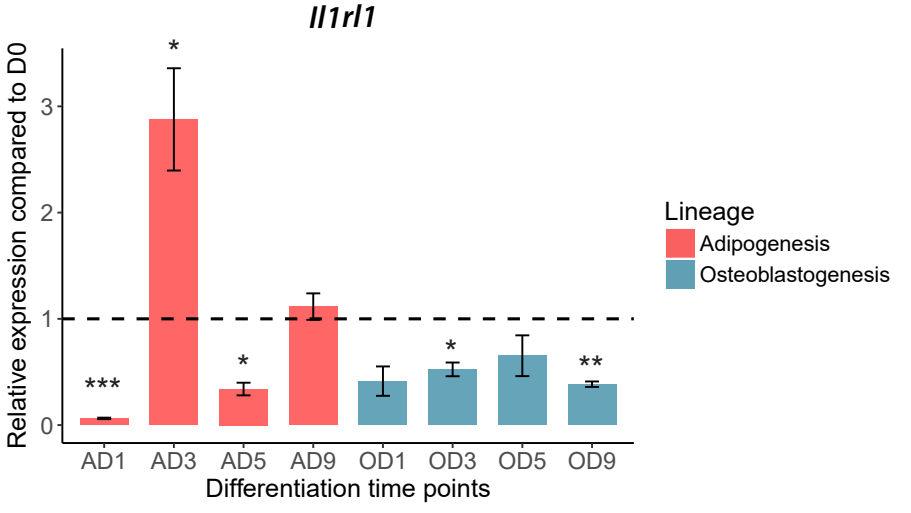


Supplementary Figure S9

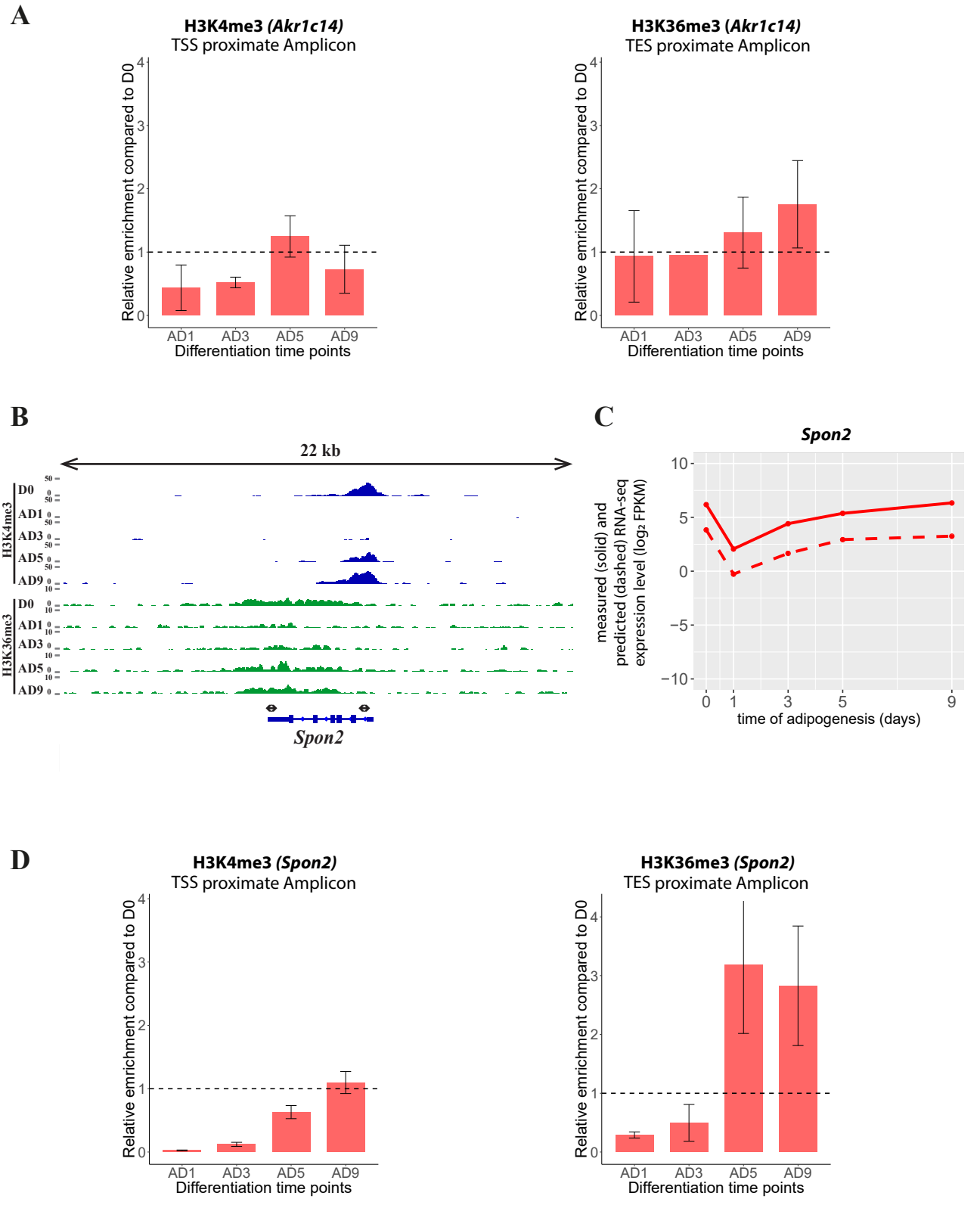
A



B



Supplementary Figure S10



Supplementary Figure S11