Supplemental figures

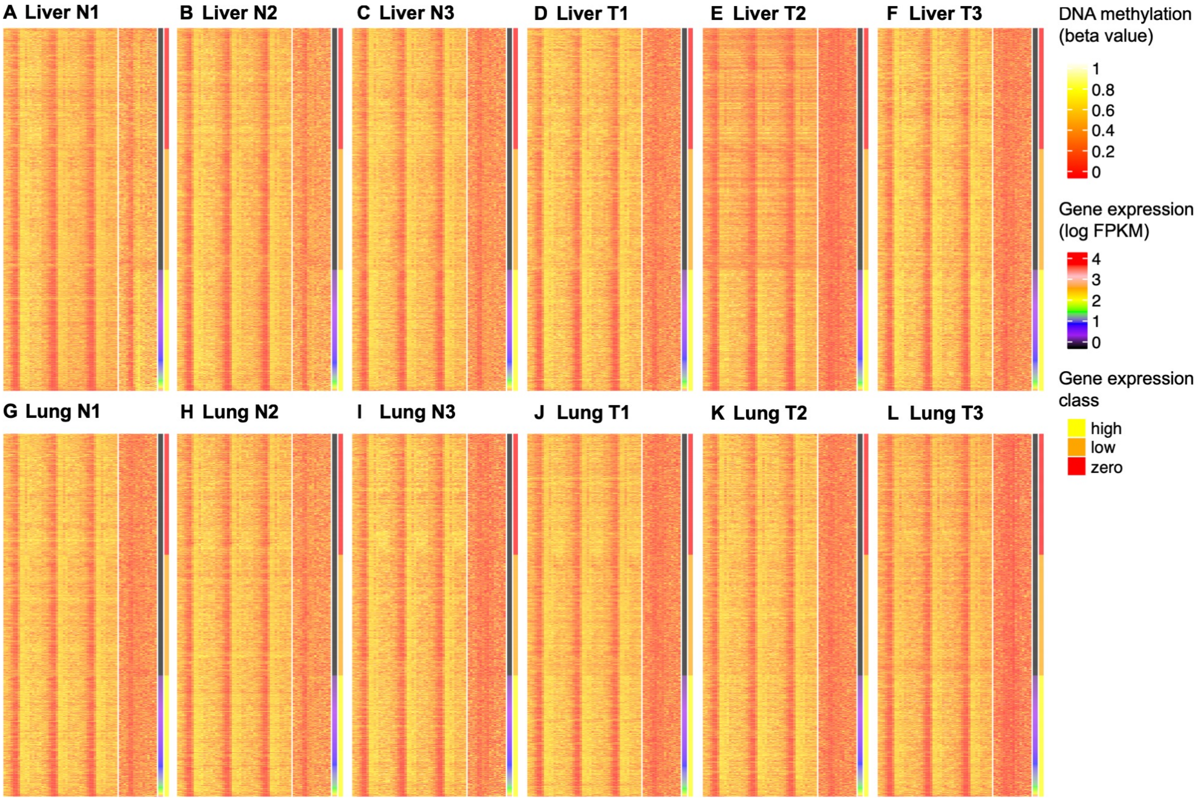

Figure S1: Heat maps of the large data sets of the different samples. In each panel, each row represents a transcript and the transcripts are sorted in ascending order according to their expression levels. The first four blocks of columns represent methylation levels based on WGBS, oxWGBS, 5mC and 5hmC, respectively. Within each block, the different columns are respectively Up5-Up1, FirstEx, FirstIn, IntEx, IntIn, LastEx, LastIn and Down1-Down5. After the four methylation blocks, the last two columns show the log expression level and expression class, respectively. The different panels correspond to normal livers 1-3 (**A**-**C**), liver tumors 1-3 (**D**-**F**), normal lungs 1-3 (**G**-**I**) and lung tumors 1-3 (**J**-**L**).

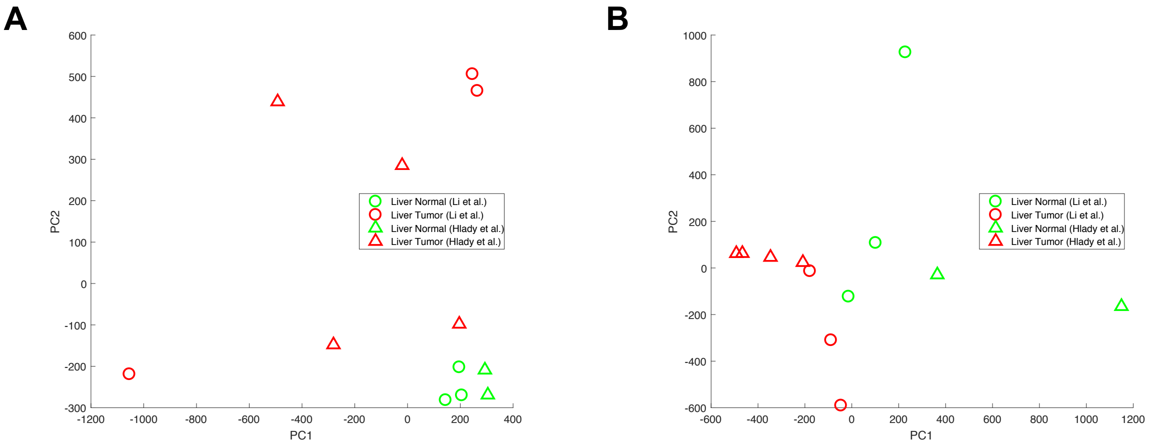

Figure S2: Projection of the liver tumor and normal liver samples from Li et al. (2016) and Hlady et al. (2019) onto the first two principal components based on 5mC (**A**) and 5hmC (**B**) beta values. For the data from Hlady et al. (2019), 5hmC beta values were computed from TAB-RRBS data and 5mC beta values were computed by subtracting the corresponding 5hmC beta values from the RRBS beta values, set to zero if negative. The CpG sites covered by both data sources were then collected and projected onto the space orthogonal to a vector that indicates the data source. Principal component analyses were then performed on the resulting data.

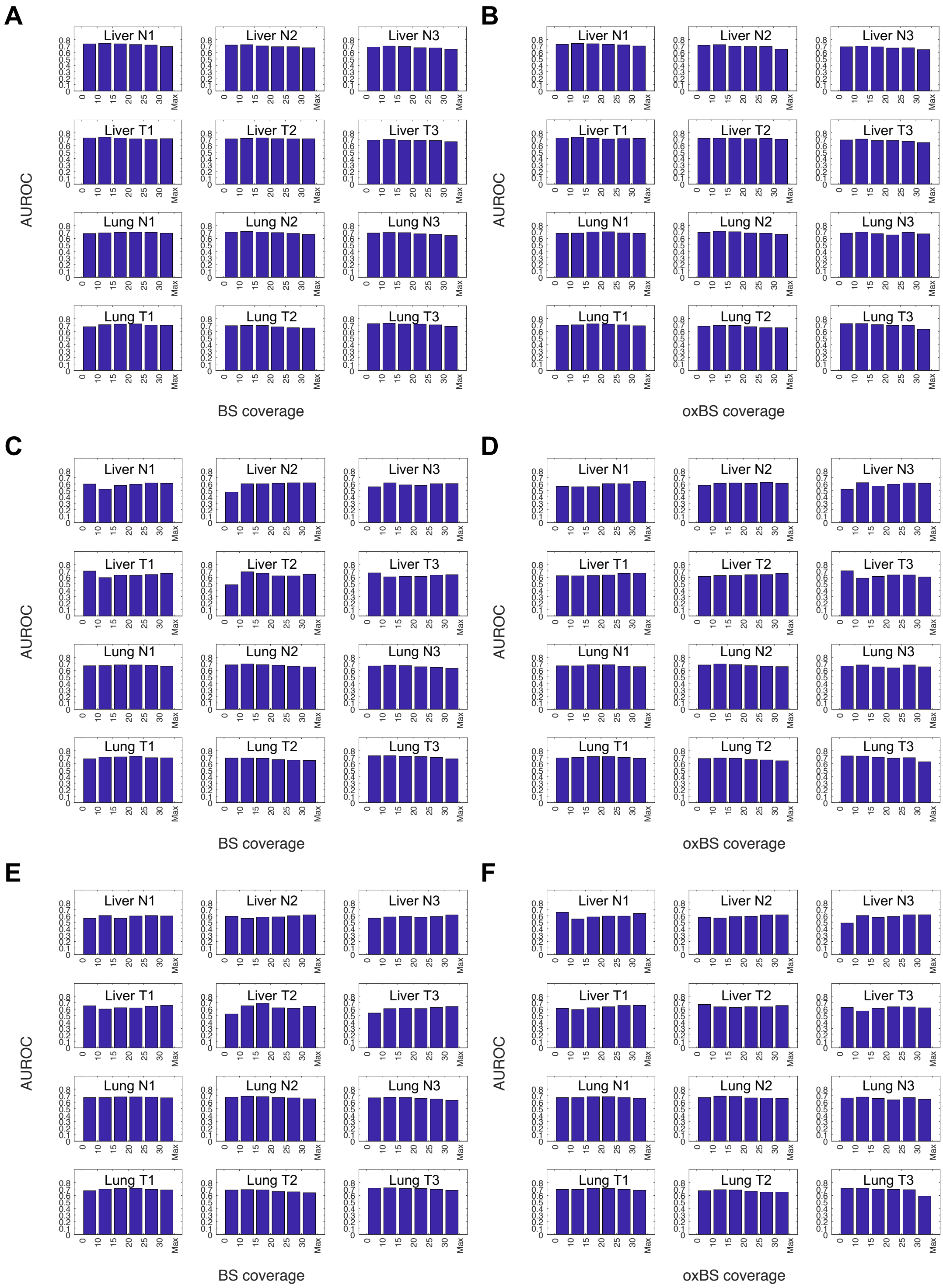

Figure S3: Modeling accuracy for transcripts with different read coverage. The model involving methylation features from all 16 regions were applied to infer the expression class of transcripts in the cross-validation setting. The features included were either both WGBS and oxWGBS (**A** and **B**), WGBS only (**C** and **D**), or oxWGBS only (**E** and **F**). Transcripts with different BS (**A**, **C** and **E**) and oxBS (**B**, **D** and **F**) read coverage were then separated into different bins, and the average AUROC values of the transcripts in each bin were computed separately.

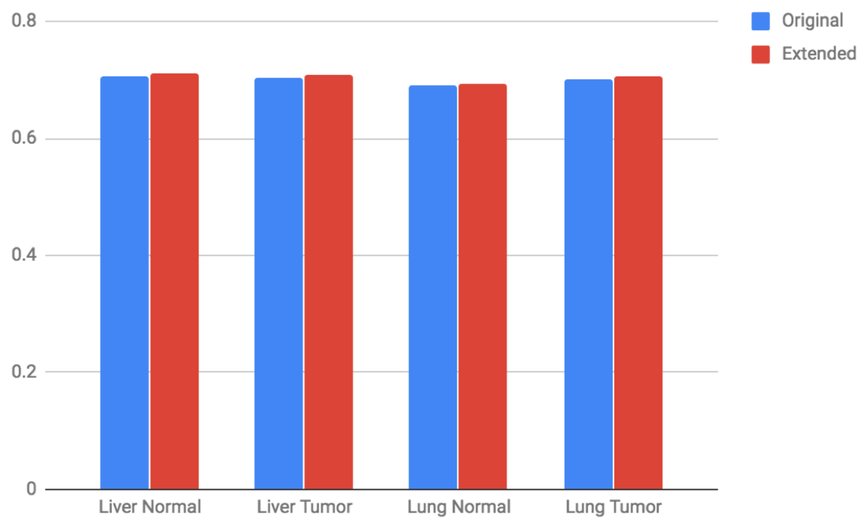

Figure S4: Comparison of models involving only features from the 16 regions or also features from extended flanking regions. The original models involved all types of methylation features from the 16 regions associated with each transcript. The extended models involved three additional 5000bp bins upstream of Up5 and three additional 5000bp bins downstream of Down5.

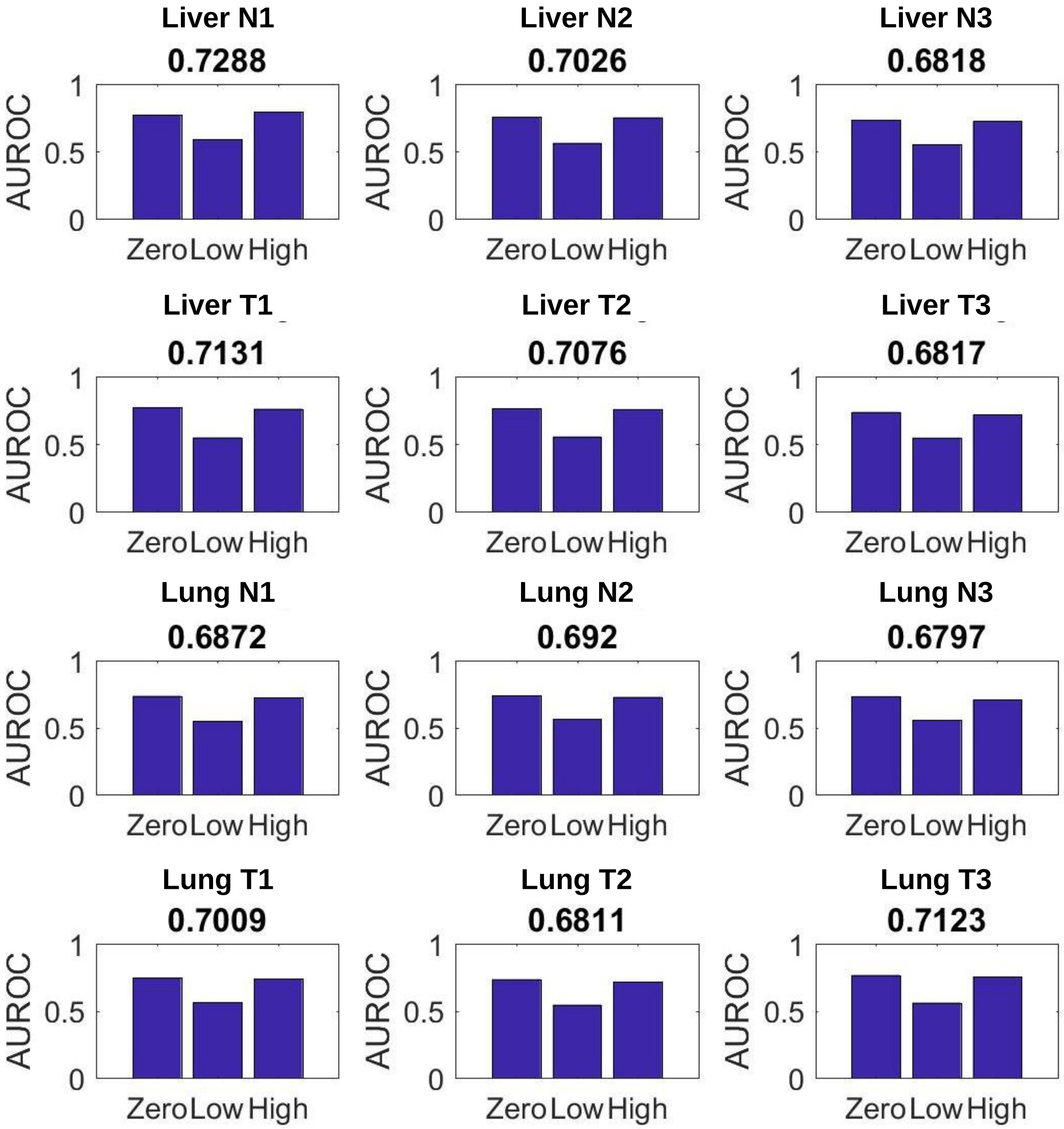

Figure S5: Modeling accuracy of the three expression classes involving all methylation features at all 16 regions associated with each transcript based on the large data set. Each bar shows the AUROC value of the cross-validation result when that expression class was considered the positive class. The number above each panel is the average AUROC of the three classes weighted by their transcript counts.

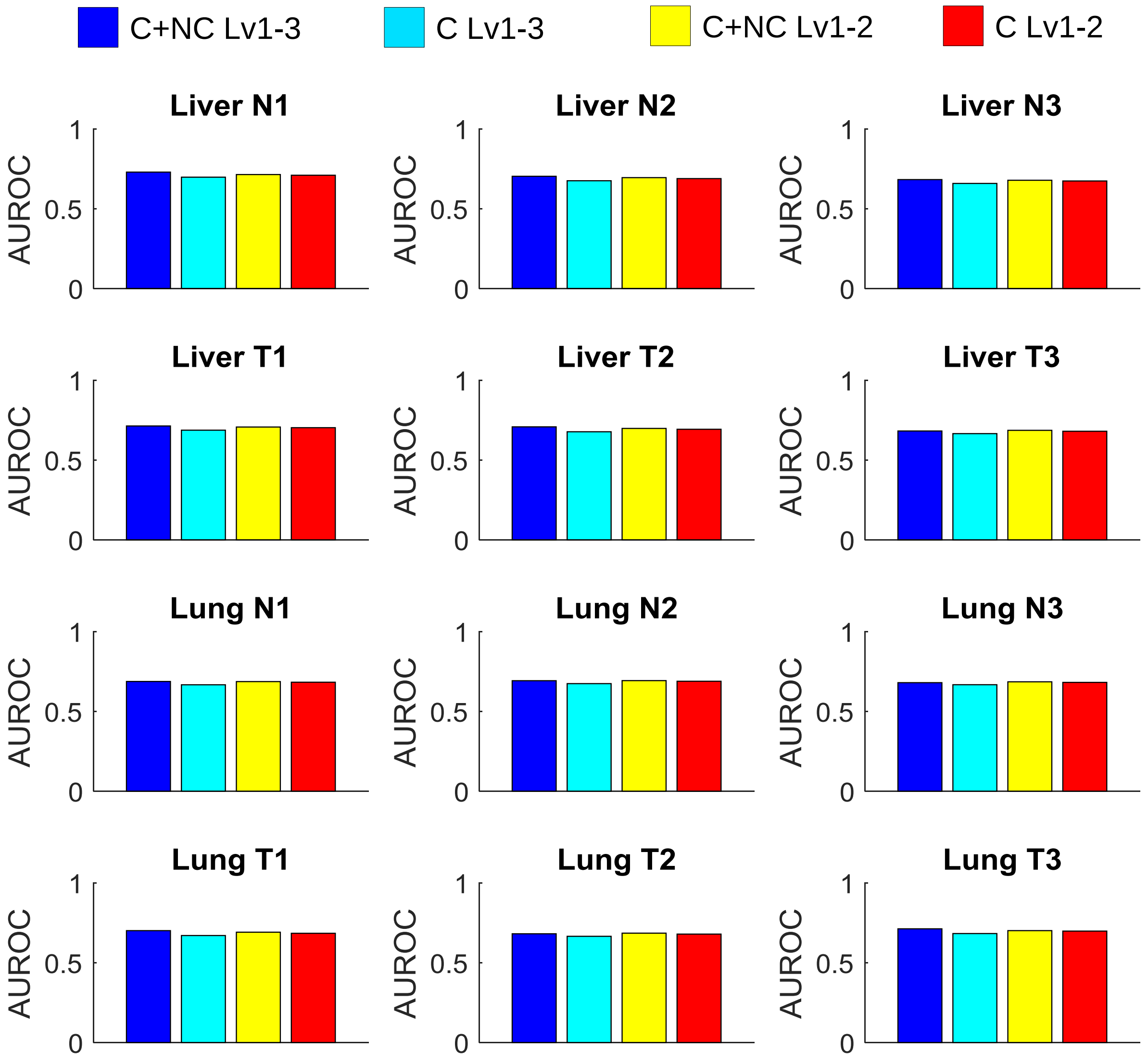

Figure S6: Average modeling accuracy of the three expression classes when all methylation features at all 16 regions associated with each transcript was considered, involving i) either both protein-coding and non-coding transcripts (“C+NC”) or protein-coding transcripts only (“C”), and ii) either GENCODE levels 1-3 transcripts (“Lv1-3”) or only GENCODE levels 1-2 transcripts (“Lv1-2”), based on the large data set. Each bar shows the average AUROC value of the three expression classes, weighted by their sizes, where the AUROC value of each expression class is the cross-validation result when this class was considered the positive class.

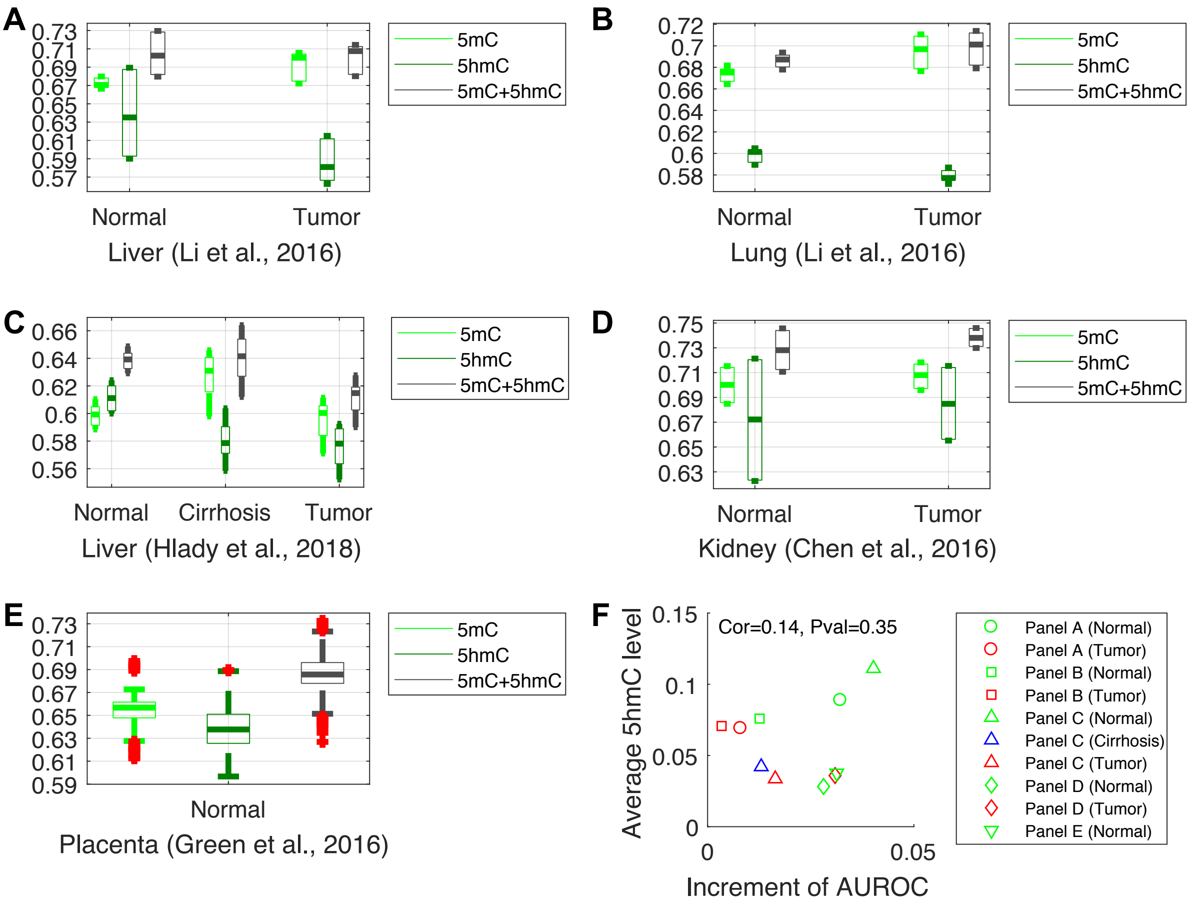

Figure S7: Average modeling accuracy of the three expression classes involving methylation features at all 16 regions associated with each transcript based on the large data set and data from additional tissue types. **A**-**E** Each bar represents the distribution of AUROC values across the three expression classes of the samples in each sample group from the five data set. **F** Relationship between the genome-wide average 5hmC level and the increment of AUROC value when comparing models involving both 5mC and 5hmC features with models involving 5mC features alone.

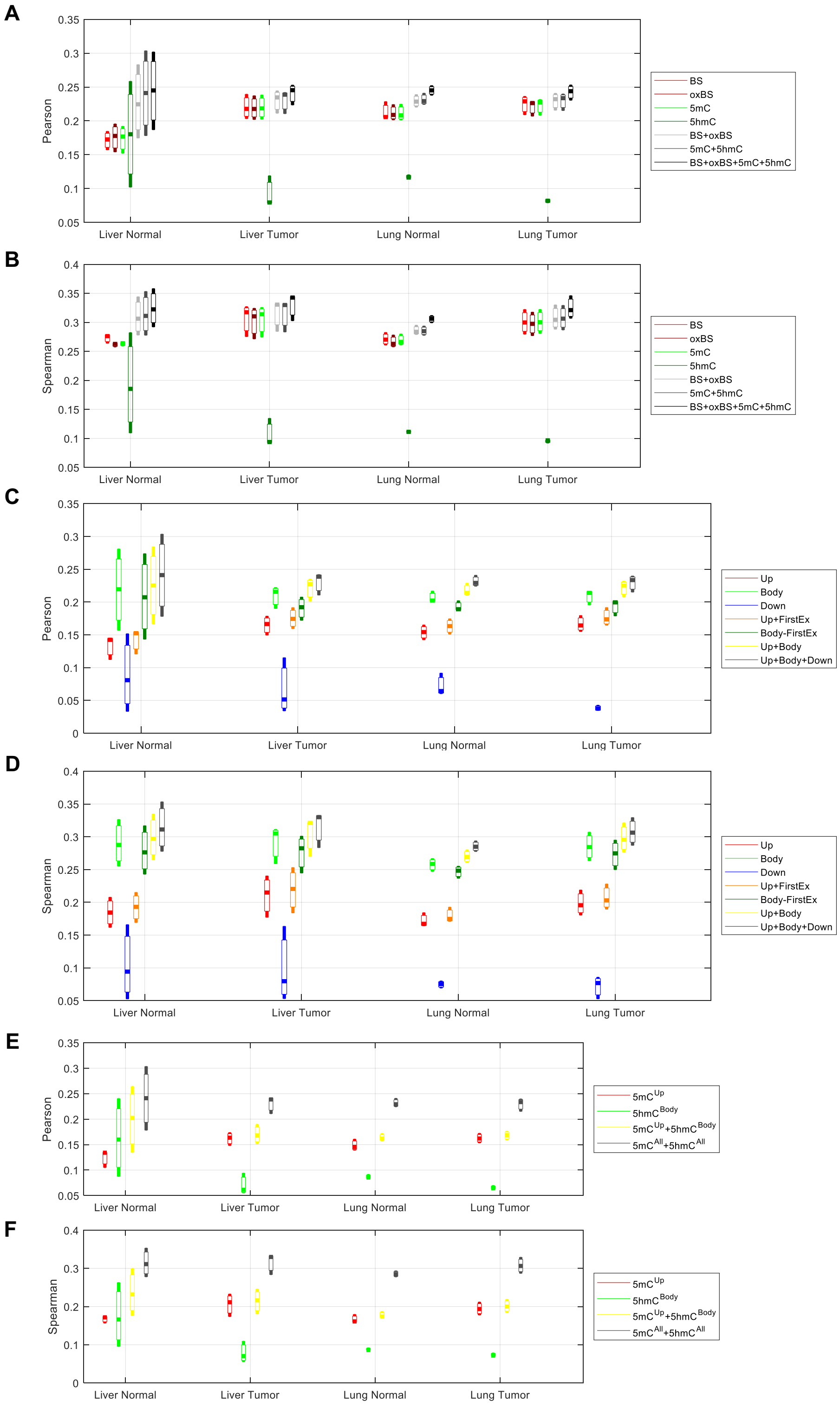

Figure S8: Accuracy of the models for inferring log expression levels based on the large data set. Each bar represents the distribution of correlation values across the different cross-validation folds of three samples in each sample group. **A**,**B** Comparison of models involving different combinations of methylation features from all genomic regions associated with each transcript. **C**,**D** Comparison of models involving all types of methylation features from different combinations of genomic regions associated with each transcript. **E**,**F** Comparison of several knowledge-driven models. In these six panels, the models were evaluated by Pearson's correlation (**A**,**C**,**E**) or Spearman's correlation (**B**,**D**,**F**) of the cross-validation results.

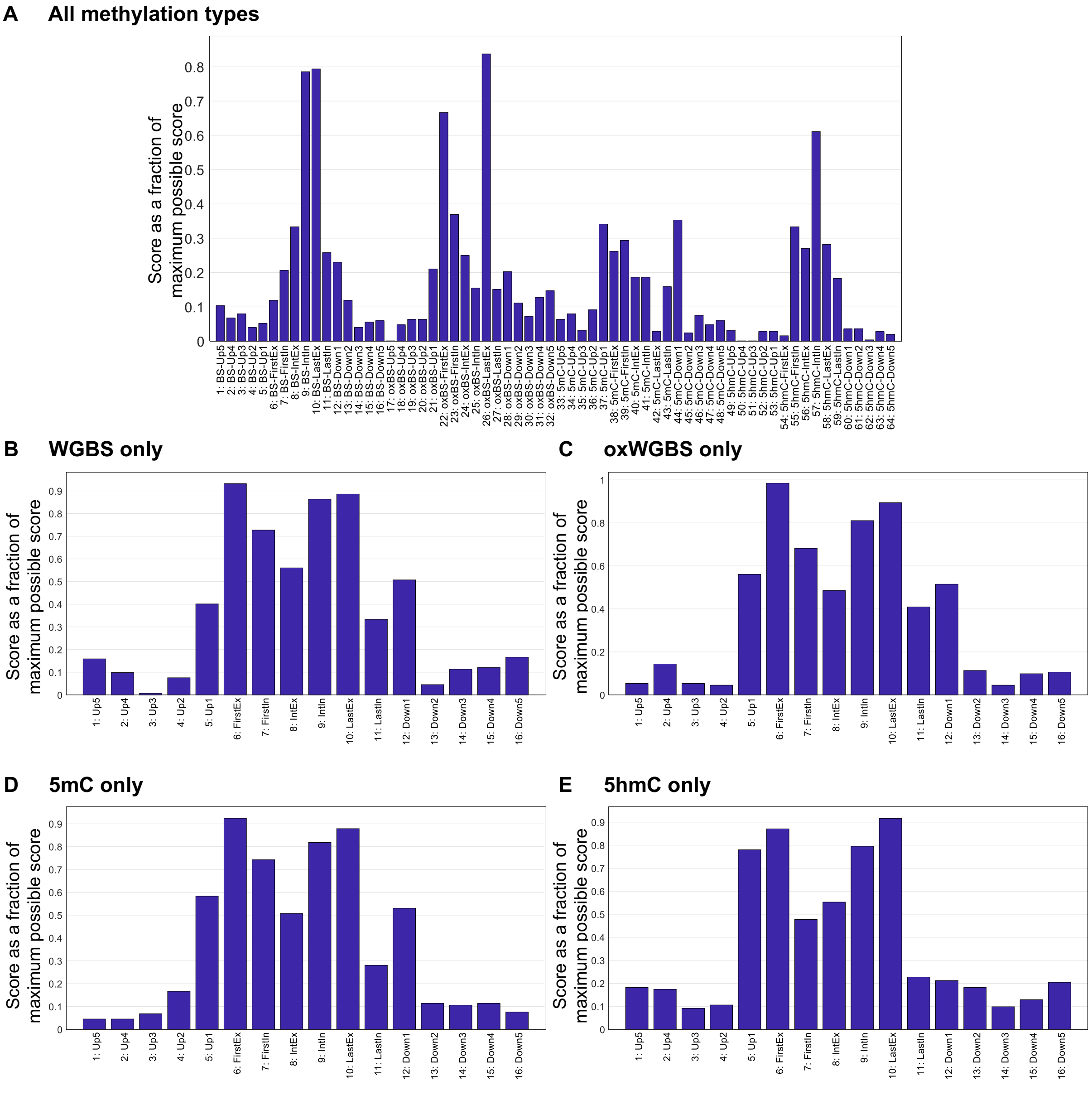

Figure S9: The most useful methylation features for inferring expression classes according to the forward-search procedure of feature selection, considering all methylation features types (**A**), WGBS only (**B**), oxWGBS only (**C**), 5mC only (**D**) or 5hmC only (**E**), based on the large data set. For each sample, the top feature was given a score of x, the second top feature was given a score of x-1, and so on, for the top x features, where x=32 in Panel **A** and x=11 for Panels **B-E**. The total score of each feature across all the samples is shown as a percentage of the maximum possible score.

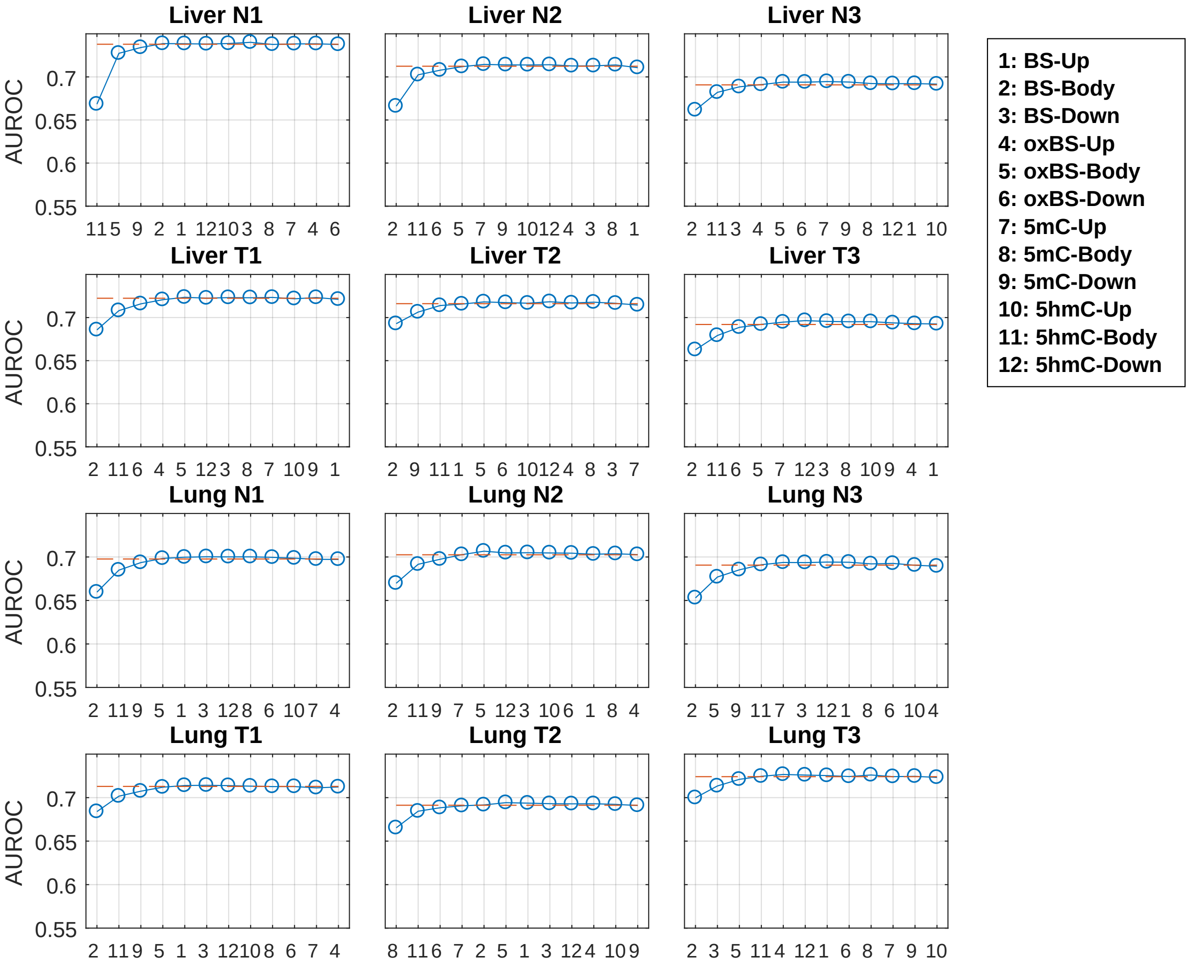

Figure S10: Change of model accuracy with each additional feature block during the forward search procedure, based on the large data set. Numbers along the x-axis show the feature block IDs, with the corresponding feature blocks stated in the box on the right. The red dash line shows the “best case” AUROC of the model involving all feature blocks.

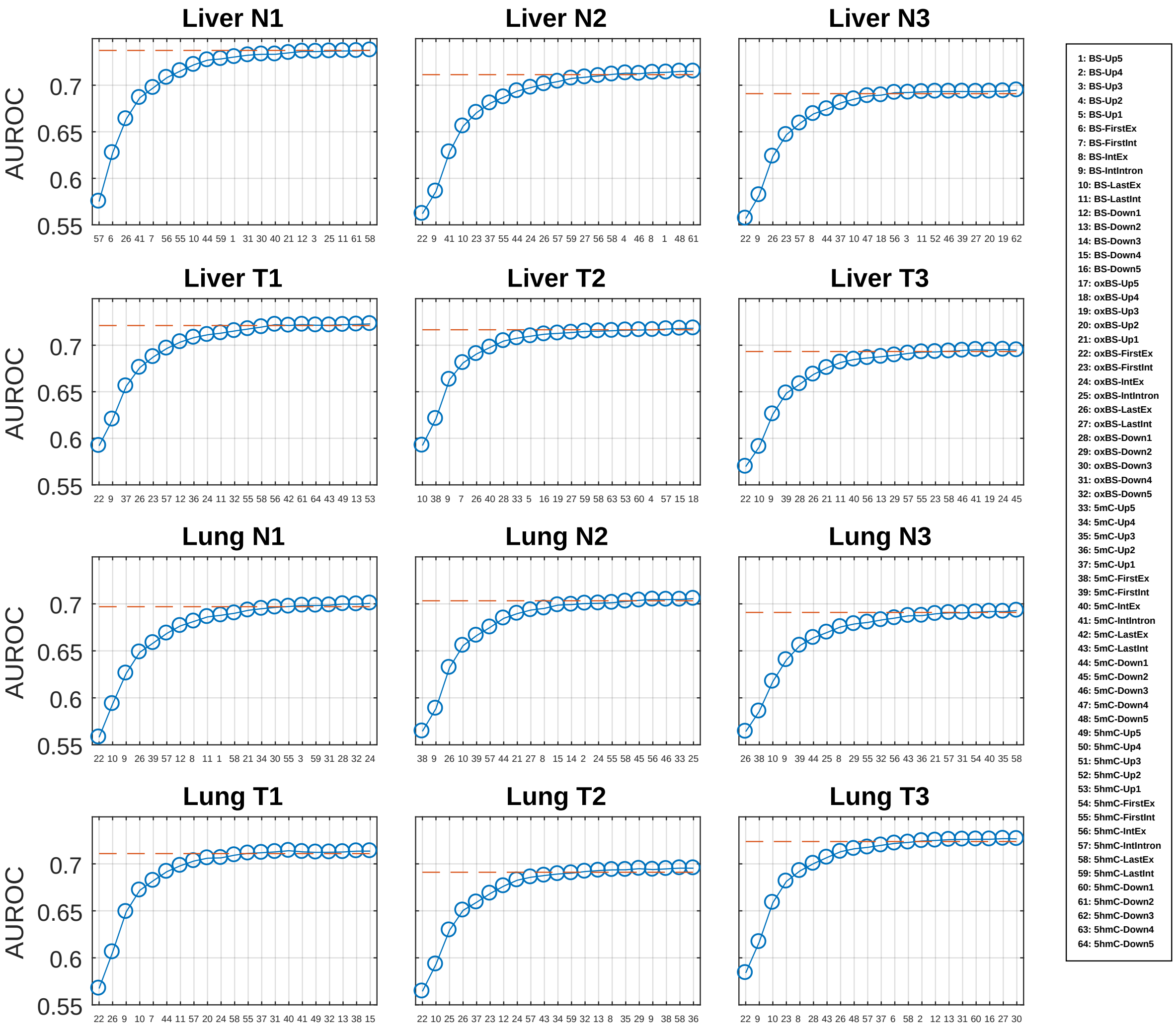

Figure S11: Change of model accuracy with each additional feature during the forward search procedure, based on the large data set. Numbers along the x-axis show the feature IDs, with the corresponding features stated in the box on the right. The red dash line shows the “best case” AUROC of the model involving all features.

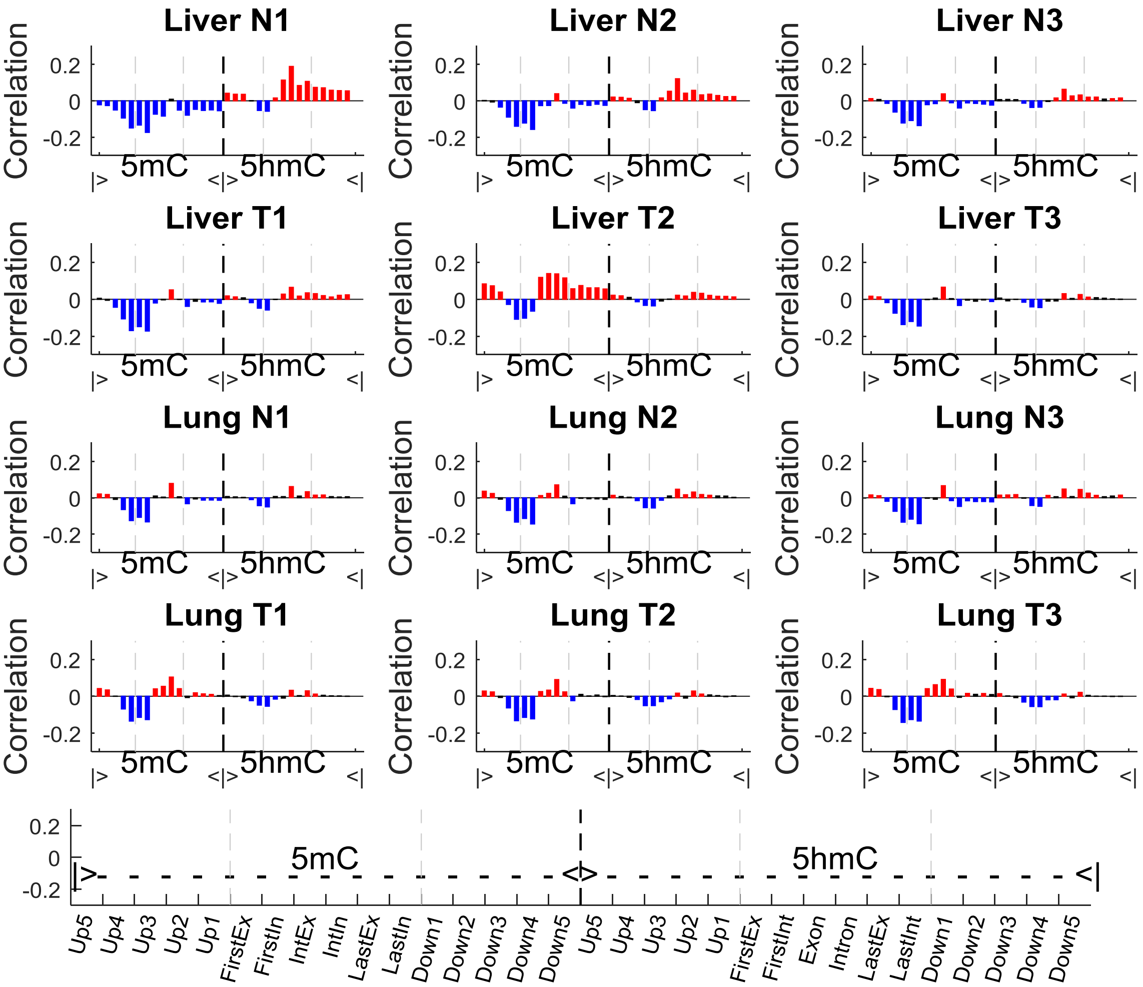

Figure S12: Pearson's correlations between log expression level and 5mC/5hmC levels at individual sub-regions based on the large data set. Statistically significant positive/negative correlations with a Bonferroni corrected p-value of 0.05 are represented by red/blue bars, while insignificant correlations are represented by black bars. These p-values were computed by randomly permuting the methylation levels of the transcripts in the respective region and calculating the resulting correlation with transcript expression levels. P-value was then defined as the fraction of times that the correlation value in the permuted cases was larger than the one in the original unpermuted case, further corrected by the number of tests performed (i.e., number of bars in each panel).

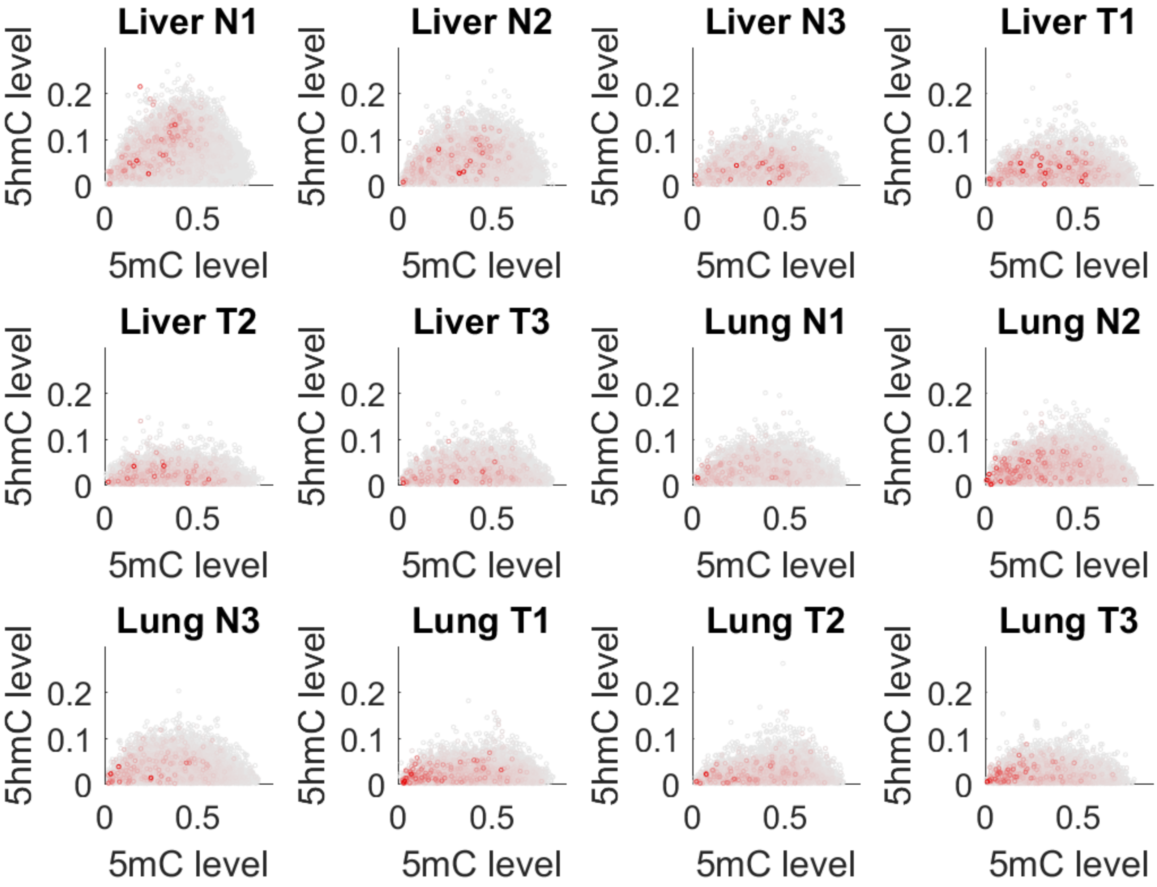

Figure S13: Relationship between transcript expression levels and their 5mC and 5hmC levels at transcript bodies based on the large data set. Each panel corresponds to a sample. In each panel, each circle corresponds to a transcript, with the x-axis and y-axis respectively represent the 5mC and 5hmC levels. The color of a circle indicates the expression level of the transcript, with a darker color indicating a higher expression level. Circles for transcripts with a higher expression level are placed on top of those with a lower expression level.

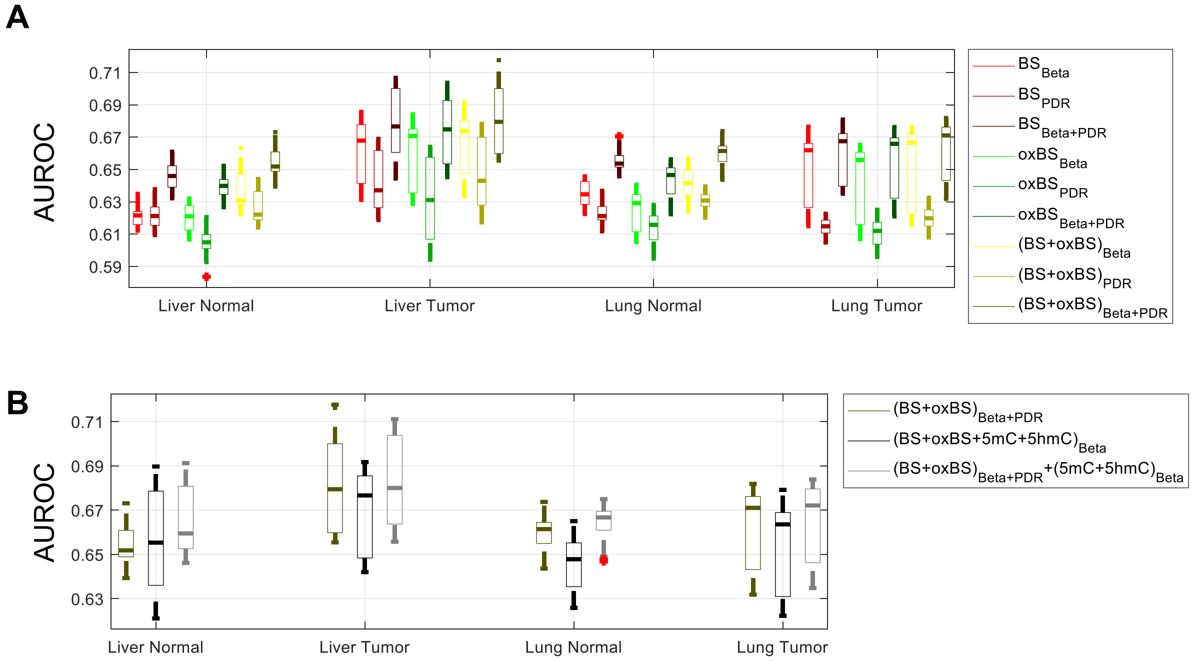

Figure S14: Accuracy of the models for inferring expression classes based on the small data set. Each bar represents the distribution of AUROC values across the three expression classes of the three samples in each sample group. **A** Comparison of models involving different combinations of methylation features from all genomic regions associated with each transcript. **B** Comparison of models that integrate different feature sets. In both panels, red dots indicate outliers.

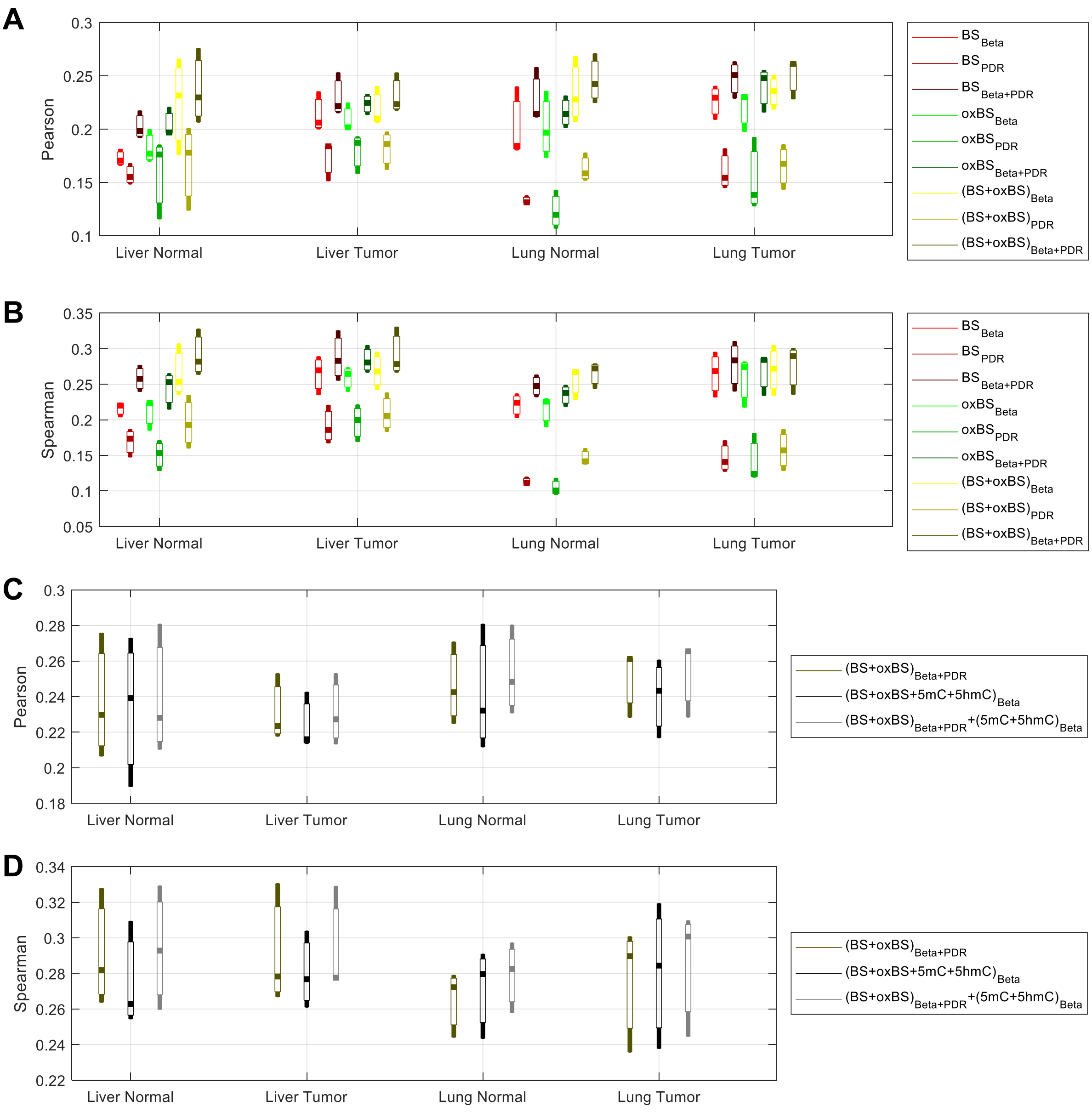

Figure S15: Accuracy of the models for inferring log expression levels based on the small data set. Each bar represents the correlation values across the three samples in each sample group. **A**,**B** Comparison of models involving different combinations of methylation features from all genomic regions associated with each transcript in terms of Pearson's correlation (**A**) or Spearman's correlation (**B**). **C**,**D** Comparison of models that integrate different feature sets in terms of Pearson's correlation (**C**) or Spearman's correlation (**D**).

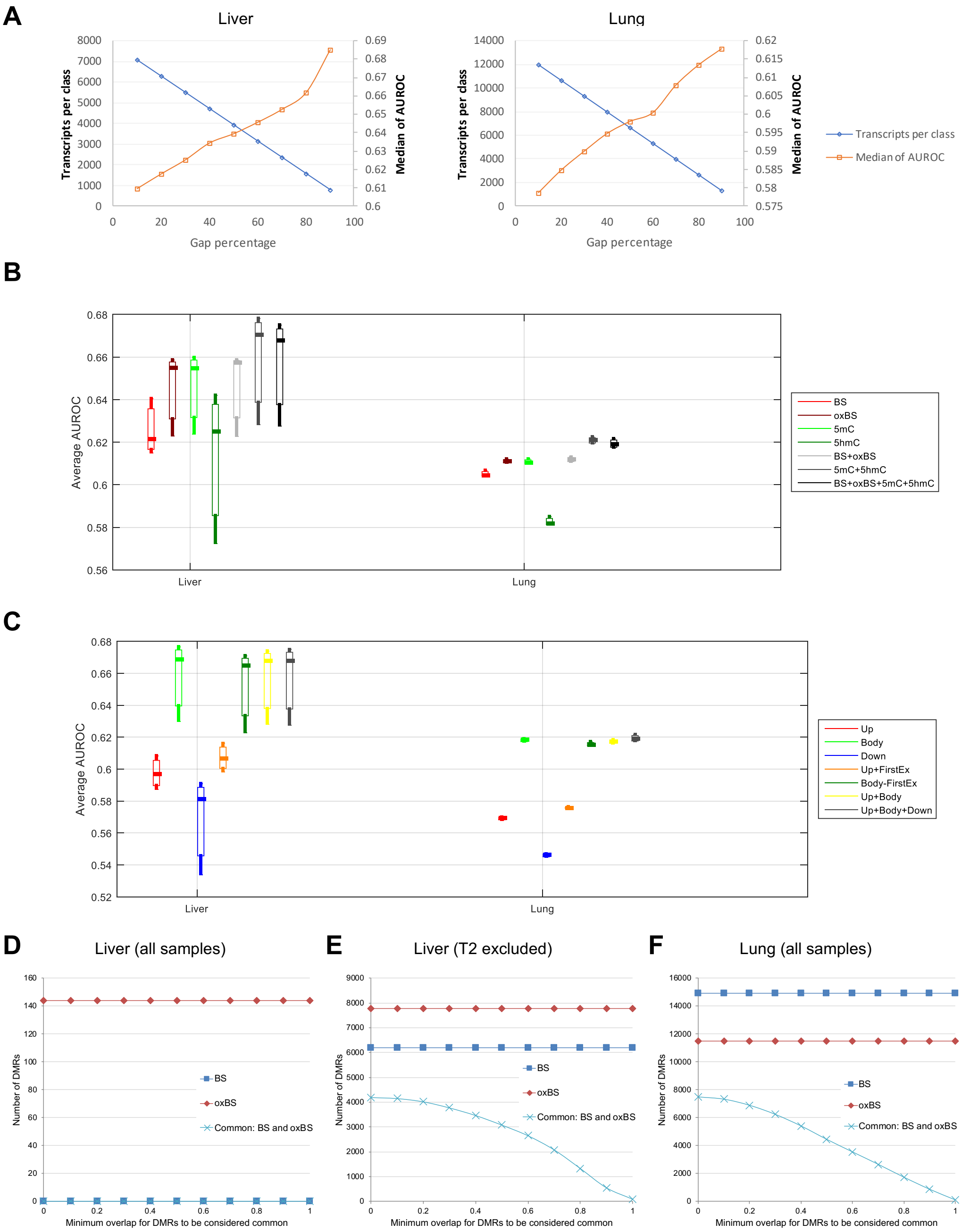

Figure S16: Additional results about differential methylation and differential expression between tumor and matched normal tissue pairs. **A** Number of transcripts and distribution of median AUROC values at various gap percentages between the strong and weak differential expression classes. A larger gap percentage makes the transcripts in the strong differential expression classes having differential expression values much stronger than those in the weak differential expression classes, at the expense of including less transcripts in these classes. **B**,**C** Accuracy of the models for inferring differential expression classes based on the large data set with an inter-class gap percentage of 80%. Each bar represents the distribution of AUROC values across the different cross-validation folds of three pairs of samples in each tissue type. **B** Comparison of models involving different combinations of methylation features from all associated genomic regions of the transcripts. **C** Comparison of models involving all types of methylation features from different combinations of genomic regions. **D**-**F** Overlap of DMRs identified using only WGBS data or oxWGBS data, for all liver samples (**D**), all liver samples except tumor T2 (**E**), and all lung samples (**F**) using dmrseq.