

## Supplemental Methods

### Methylated CpG Pre-processing

Within each sample, the % methylation at each CpG was calculated using the bisseq pipeline (<https://pbtech-vc.med.cornell.edu/git/thk2008/bisseq>). Filtering by coverage was then performed for all samples and controls using a custom python script to select CpGs with 10 – 400x coverage. We evaluated whether stringent coverage cutoffs would affect the proportion of various represented genomic features. While requiring coverage in a higher number of samples decreased the overall number of CpGs considered, the relative proportion of gene parts, CpG islands, or CpG island shores remained similar.

### Hierarchical clustering

In order to determine the optimal set of patient clusters based on AML epigenetic variation, we chose to cluster using the CpGs that were most variable across the patient samples. We evaluated the standard deviation across all samples displayed by each CpG covered in all patients (n=950,953). The resulting distribution showed several modes, and we opted to select a cutoff that included only the most variable mode, corresponding to the 85<sup>th</sup> standard deviation percentile. Using these CpGs, we then performed a hierarchical clustering using a euclidean distance metric and Ward's method ('ward.D' in R). In this clustering, patients with classic translocations such as t(8;21) and t(v;11q23) were in close proximity. We then chose a cutoff that preserved this clustering and did not group them with neighboring cases not harboring these translocations.

### Annotating CpGs to genomic features

In order to determine which genomic features overlapped a given CpG, the ChIPseeqerAnnotate program from the ChIPseeqer package was used (version 2.1). The output was reformatted using a set of custom python scripts such that a one-to-many annotation of each CpG resulted. For example, a CpG within one gene promoter that overlaps an exon on the opposite strand would be annotated twice, once for each feature. For subsequent analyses, CpGs associated with more than one feature were prioritized according to the hierarchy: promoter > exon > intron > neighborhood > intergenic. For analyses specifically focusing on exons or introns, CpGs within the first exon or intron were omitted due to the proximity to the promoter.

To annotate CpG islands, ChIPseeqerNongenicAnnotate was used. To annotate CpG island shores, active enhancers, and poised enhancers, this program was modified from its original format to include these new annotations.

### Annotation of enhancers

In order to activate active and poised enhancers, we utilized data from the mobilized CD34+ cells in the epigenome roadmap. For all samples with H3K4me1, H3K27ac, H3K4me3, and input ChIP-seq data (01480, 01536, 01549), we called peaks using ChIPseeqer (version 2.1) using the default parameters.

For active enhancers, intersecting H3K4me1 and H3K27ac peaks were identified in each sample and combined peaks overlapping H3K4me3 were removed. Peaks overlapping promoters (TSS +/- 2kb) were removed as well. The union of these remaining peaks was then used as the set of CD34+ active enhancers in subsequent analyses. Poised enhancers were identified in a similar fashion, except that regions with H3K27ac peaks were excluded.

### Predictive ability of genomic features

To evaluate the contribution of each genomic feature to the overall clustering, we re-analyzed the samples using only the CpGs contained within various genomic subsets. For each subset of CpGs, we clustered again using CpGs above the 85<sup>th</sup> percentile of variability within the subset. To quantify the relative similarity of the subset to overall clustering, we used the adjusted rand index. To determine the significance of the subset clustering, the same procedure was performed 500 times using the same number of CpGs chosen at random from the set of CpGs covered in all samples. The actual subset adjusted rand index was then compared to this distribution of rand indices using a z test.

### Cluster-specific clustering accuracy by genomic feature

The ability of CpGs within each genomic feature was assessed by inspecting the dendrograms produced using the various genomic subsets. If a cluster was entirely preserved, the clustering was labeled as 'intact'. If the original cluster was split into two major groups, it was labeled 'partial'. If the clustering showed greater perturbation, it was considered to be 'lost'.

### Calculating methylation change

In order to determine the methylation change within each cluster, the group of cases in each was compared to the 22 CD34+ normal controls using MethylSig with a minimum of 3 CpGs per group. CpGs with  $\geq 25\%$  change from controls and  $q \leq 0.01$  were considered to be differentially methylated CpGs (DMCs).

### Methylation changes at genomic features

In order to determine overall methylation changes at various genomic features, CpGs were binned across the feature and flanking sequence. In the enhancer analysis shown in **Figure 2a**, CpGs within every active enhancer and the 10kb flanking sequence was considered with the directionality determined according to that of the closest gene. CpGs within each active enhancer were divided into 10 bins, and the average difference was calculated. CpGs within each flank were also divided into 10 bins and averaged. This was done regardless of the active enhancer size or CpG coverage within a given enhancer or bin. For features such as a TSS which occupies a single point, the two flanks were divided into 15 bins each and each was averaged to preserve the overall number of bins considered.

To calculate this difference for DMCs only, the CpGs were pre-filtered to include only those with  $\geq 25\%$  difference from controls and a q-value  $\leq 0.01$  as determined by MethylSig.

### DMC enrichment calculation

In order to determine the enrichment of DMCs for a given cluster within a given genomic compartment, a two by two table was constructed containing the counts of DMCs within the genomic compartment, non DMCs within the compartment, DMCs outside the genomic compartment, and non DMCs outside the genomic compartment. A one-tailed Fisher's exact test was used to determine whether DMCs were enriched within the specified genomic compartment for the cluster being considered. This was done for all clusters and all genomic compartments.

### IDH2, IDH1/DNMT3A, and DNMT3A methylation analysis

Analysis of *IDH2*, *IDH1/DNMT3A*, and *DNMT3A* cases was done separately from the initial clustering. Cases with an *IDH1* or *IDH2* mutation without *DNMT3A* mutation, *DNMT3A* mutation without *IDH1* or *IDH2* mutation, and *IDH1* or *IDH2* + *DNMT3A* mutation were identified and assigned to separate categories. As with the original clustering, the group with *IDH1* or *IDH2* mutation alone was composed almost entirely of *IDH2* mutated AMLs. The *IDH1* or *IDH2* + *DNMT3A* category was likewise composed almost entirely of *IDH1* / *DNMT3A* cases. These categories were subsequently described as *IDH2* and *IDH1* / *DNMT3A*. DMCs within each of these categories was determined as before by comparing them to the 22 CD34+ controls using MethylSig with a minimum coverage of 3 CpGs per category required.

In order to determine the average methylation change at DMCs in each of these categories, the same approach was taken as before, using a custom script to analyze DMCs within each feature divided into 10 bins. The average methylation change within each of these bins for each category is shown (**Figure 4a**). The average methylation change at promoters and active enhancers for these categories was performed in the same way as it was for the 14 clusters.

### Methylation 'canyon' analysis

Large relatively hypomethylated canyons were previously described within the mouse genome (1). These annotated regions were then mapped to hg19 using the UCSC genome browser liftover tool. Of 1104 canyons originally described in the mouse genome, 227 were conserved in the human genome. DNA methylation changes at these regions and their flanking sequences were analyzed as previously described by dividing the regions and their flanking sequences into 10 bins each and averaging the methylation change in each bin (**Figure 4a-d**).

### Gene expression analysis

Gene expression analysis was performed using affymetrix hgu133 plus 2 arrays as previously described. Probes were assigned to genes according to the brainarray annotation for affymetrix hgu 133 plus 2 arrays. Normalization was performed using RMA, and differential expression relative to CD34+ normal bone marrow controls was assessed using the limma package. In analyses of differential expression, genes with  $q \leq 0.05$  and a fold change  $\geq 1.5$  were considered to be significant. Probe sets mapping to the same gene symbol were averaged prior to plotting.

### Assessment of the ability of specific genetic lesions to predict epigenetic clustering

To determine the ability of specific genetic lesions to predict epigenetic clustering, a multinomial logistic regression model was built using epigenetic categories as the output and specific lesions as the inputs. A base model was built using only the classical cytogenetic lesions t(8;21), inv(16), t(v;11q23), t(15;17). To this model, every single genetic lesion and pair of genetic lesions was added. The goodness of fit of each model was measured using the Akaike Information Criterion (AIC) (2) and compared to the base model using the formula  $\exp((AIC_{\text{translocation}} - AIC_{\text{mutation+translocation}})/2)$ , which was then plotted as a heatmap (**Figure S1a**).

Using this approach, CEBPA double mutation, NPM1 mutation, DNMT3A mutation, CEBPA silencing, and IDH2 mutation were identified as strong contributors to the determination of epigenetic cluster identity. These lesions were then added to the base model, and the same analysis was performed (**Figure S1b**).

### DMC enrichment

To determine which genomic compartments showed specific differential methylation, we constructed a 2 x 2 matrix of each compartment with rows containing counts of differentially and non-differentially methylated CpGs, and columns identifying CpGs inside and outside of the compartment of interest. Significance was assessed using a one-sided Fisher's Exact Test, and then corrected using the Benjamini-Yekutieli approach (3).

To determine enrichment of differential methylation within enhancers in each relevant genomic compartment, an analogous approach was taken. Enhancers were defined as described in the main text using ChIP-seq data from mobilized CD34+ cells from the epigenome roadmap (4). 2 x 2 matrices were constructed with rows containing counts of differentially and non-differentially methylated CpGs and columns containing enhancer CpGs within the region of interest or CpGs outside. Significance testing and correction for multiple testing was performed as before.

Heatmaps of both these analyses are shown in **Figure 4**, with a color scale that starts at a significance level of  $q > 0.05$ .

### Gene Set Analysis of IDH2, DNMT3A, and IDH1/DNMT3A categories

IDH2, DNMT3A, and IDH1/DNMT3A categories were defined as above. Differential expression was also determined as above. When multiple probe sets mapped to a single gene symbol, their differential expression was averaged. To determine the gene sets perturbed in each of these categories, average differential expression values were analyzed using Gene Set Enrichment Analysis using a pre-ranked gene list approach and default parameters.

## References:

1. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature genetics*. 2014;46:17-23.
2. Burnham KP, Anderson DR, Burnham KP. *Model selection and multimodel inference : a practical information-theoretic approach*. 2nd ed. New York: Springer; 2002.
3. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125:279-84.
4. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317-30.

**SUPPLEMENTARY FIGURE LEGENDS**

**Figure S1: Ability of specific genomic lesions to predict epigenetic clustering. A:** Specific lesions adjusted for translocation status. A multinomial logistic regression was built to assess the ability of specific genetic lesions to predict epigenetic clustering. A baseline model was built using t(8;21), inv(16), t(15;17), and t(v;11q23) status. All possible individual and pairs of assayed lesions were then added to this model, and the relative likelihood of improvement in the model's goodness of fit as assessed by the Akaike Information Criterion (AIC) was assessed and plotted as a 2-dimensional heatmap. Individual mutations are shown on the diagonal, and pairs are shown on the off-diagonal tiles. The scale ranges from worsened -> improved predictive ability, with red showing worsened predictive ability, black with minimal change, and green showing improved predictive ability. As shown here, only a minority of lesions improve the predictive ability of the translocation data. **B:** Comparable analysis to that shown in (A), now with *CEBPA-dm*, *NPM1*, *DNMT3A*, *CEBPA-silenced*, and *IDH2* added to the base model. As shown here, no additional mutation or pair of mutations improved the predictive ability of the model.

**Figure S2: ERRBS coverage. A:** Gene parts. Pie chart representation of ERRBS coverage of CpGs within and around genes. **B:** CG dense regions. Pie chart representation of ERRBS coverage of CpGs within and around CpG islands. **C:** Bar plot representation of percentages captured for each of the specified genomic features interrogated by ERRBS in this cohort. **D:** Enhancers. Bar plot representing the genomic distribution of ERRBS coverage within active and poised enhancers.

**Figure S3: Differential methylation at promoters and active enhancers. A:** Overall methylation change including both DMCs and non-DMCs shows relative enrichment at active enhancers. **B:** Differential methylation (DMCs only) is enriched at active enhancers. **C:** Overall methylation change including both DMCs and non-DMCs shows relative restriction at promoters. **D:** Differential methylation (DMCs only) shows relative restriction at promoters.

**Figure S4: DMC distribution. A:** Average raw methylation at active enhancers. **B:** Average raw methylation at promoters. **C:** Differentially methylated CpGs as a percentage of all CpGs covered by ERRBS is compared across clusters. **D:** Total differentially methylated CpG counts are compared to total ERRBS coverage.

**Figure S5: Transcription factor binding site analysis at differentially methylated active enhancers in IDH2, DNMT3A, and IDH1/DNMT3A AMLs.** **A:** De-novo CG containing motifs enriched in differentially methylated active enhancers in the IDH2 AML cohort. **B:** De-novo CG containing motifs enriched within differentially methylated active enhancers in the DNMT3A AML cohort. **C:** De-novo CG containing motifs enriched within differentially methylated active enhancers in the IDH1/DNMT3A AML cohort. **D:** Raw log<sub>2</sub> expression levels of transcription factors corresponding to de-novo discovered motifs within active enhancers. **E:** Analysis of transcription factor binding sites within differentially methylated active enhancers using ChIP-seq based transcription factor binding sites rather than motif-based discovery. Enrichments are plotted using  $-\log_{10}(pval)$  assessed using Fisher's exact test. Only results with  $p < 0.01$  in at least one of the cohorts are shown.

## **SUPPLEMENTARY TABLE LEGENDS**

Supplementary Table 1: Detailed information including ERRBS cluster identity, HELP cluster identity, cytogenetics, molecular profile, and limited profile of mutations and genetic perturbations.

Supplementary Table 2: Extended high throughput mutation profiling.

Supplementary Table 3: DMC identity and annotation for all clusters.

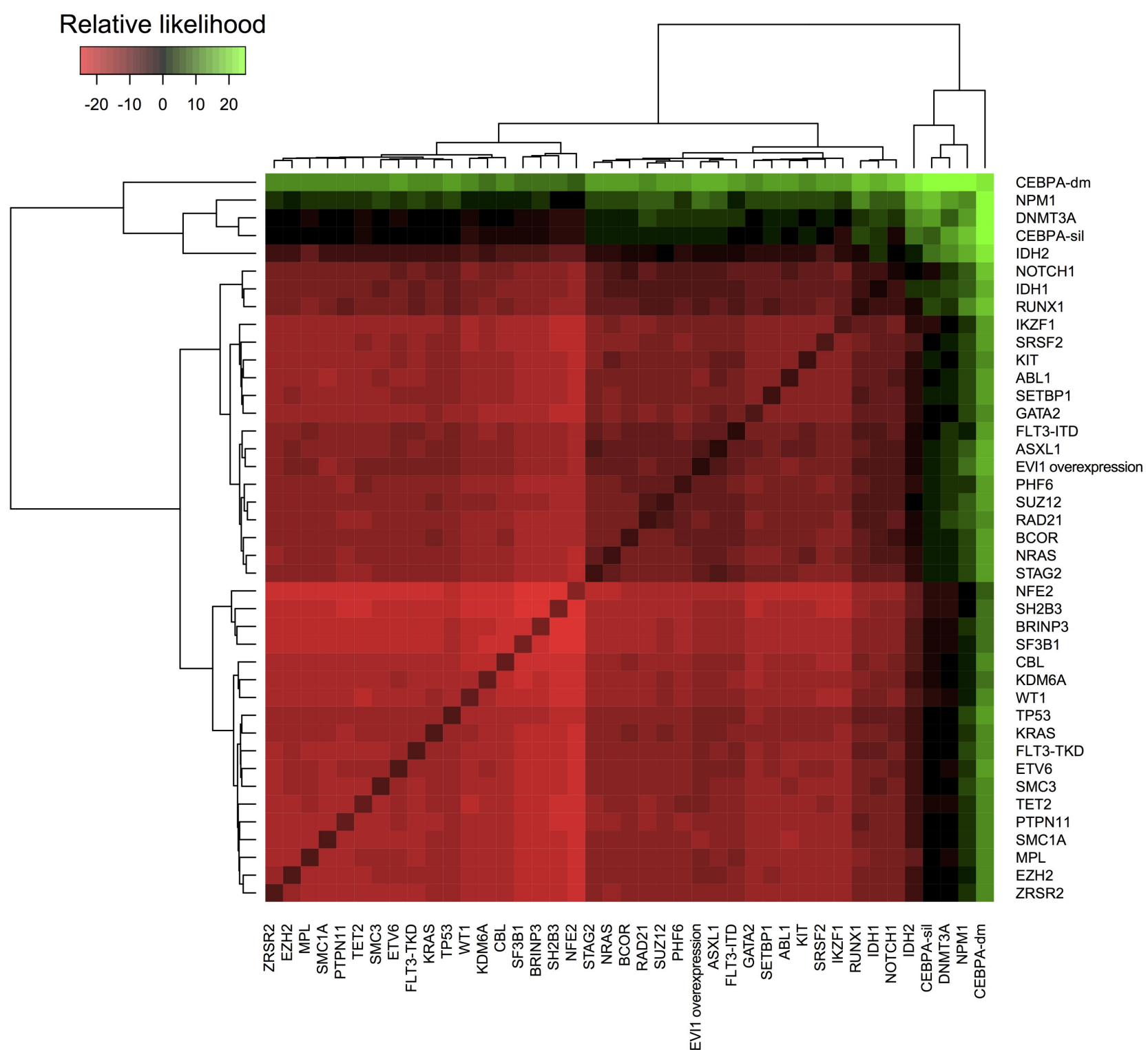
Supplementary Table 4: DMC identity and annotation for IDH2, DNMT3A, and IDH1/DNMT3A cohorts.

Supplementary Table 5: DMC identity and annotation for Idh2, Dnmt3a, and Idh2/Dnmt3a murine models.

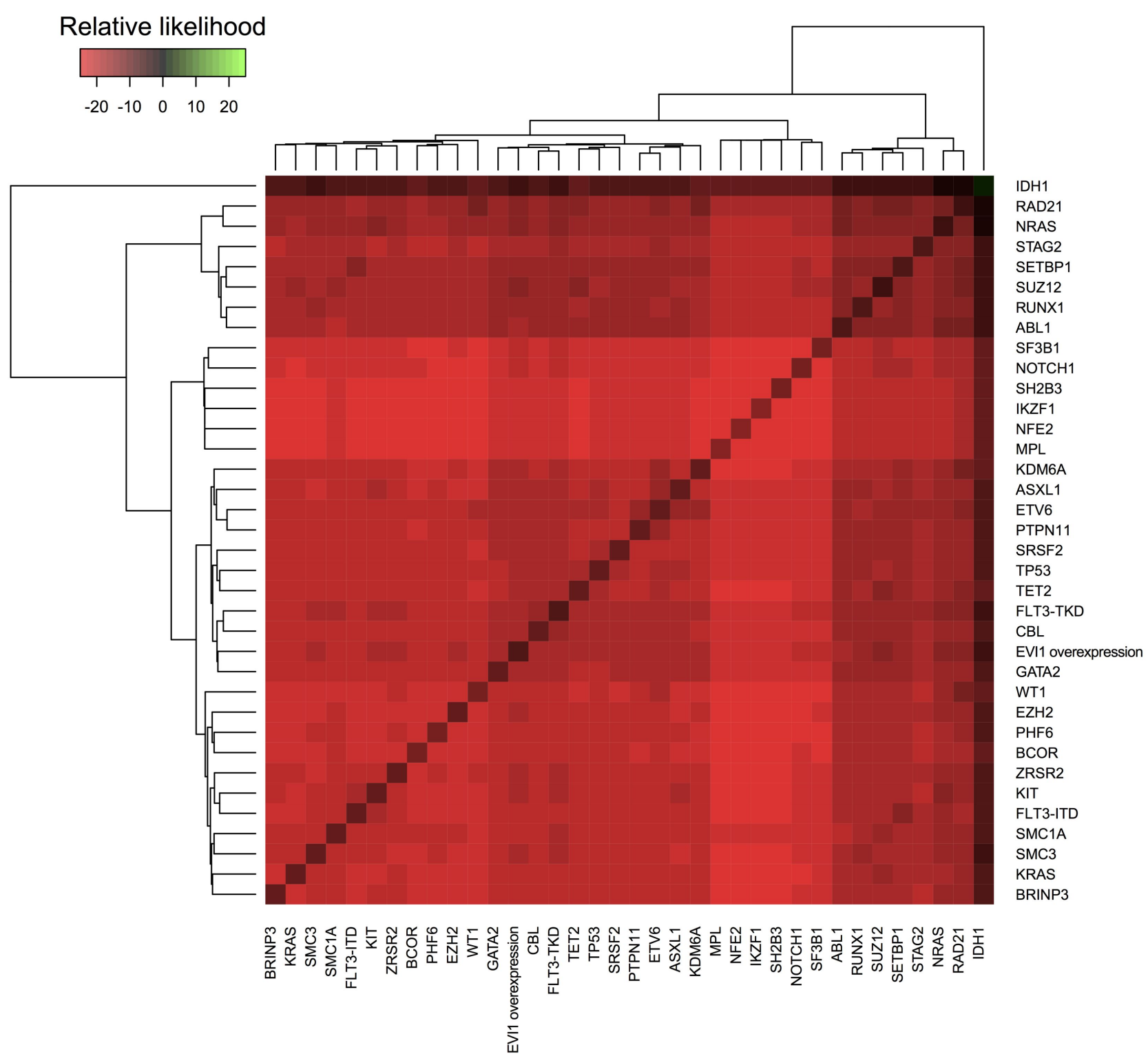


# Supplemental Figure 1.

a)

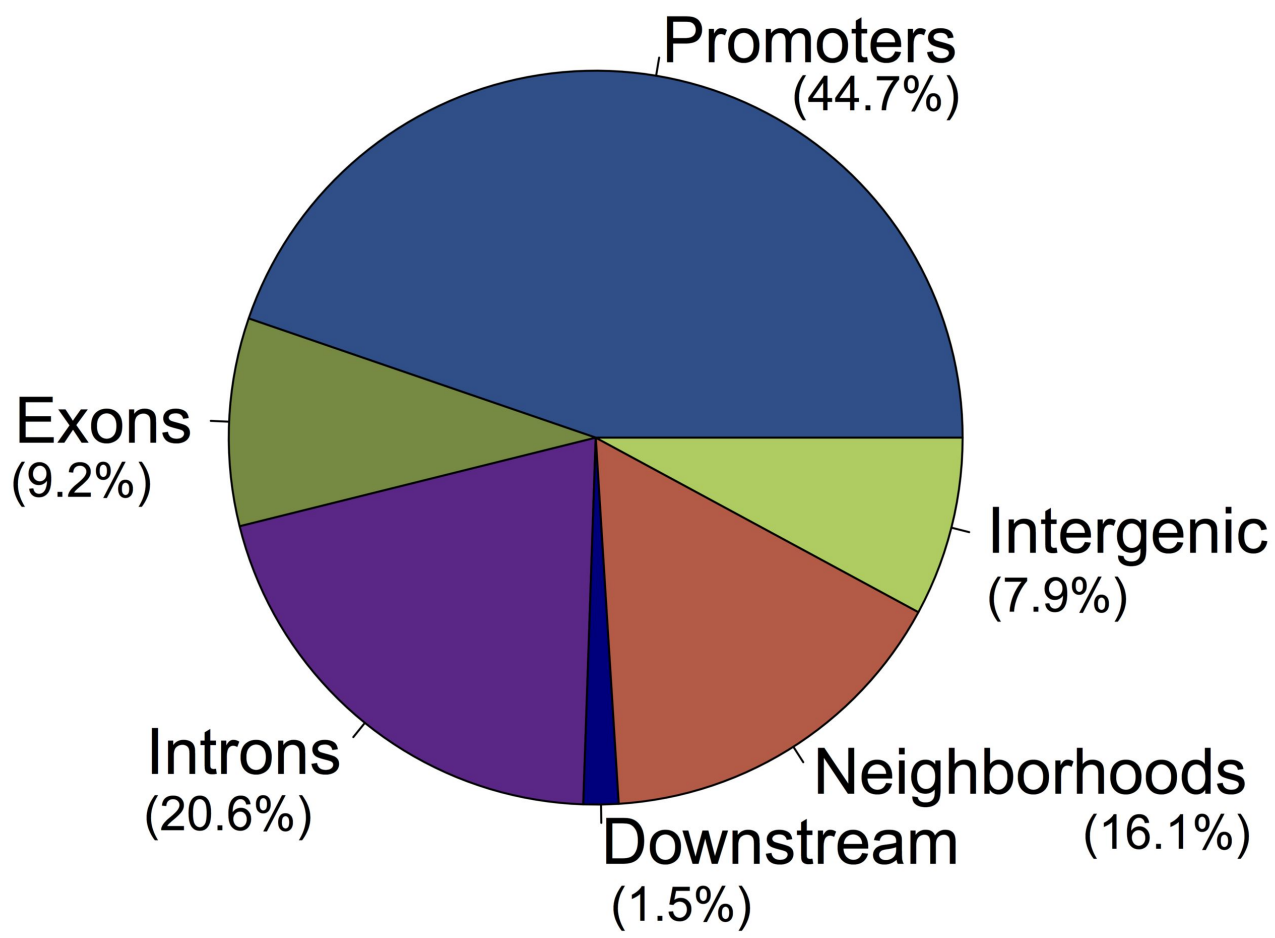


b)

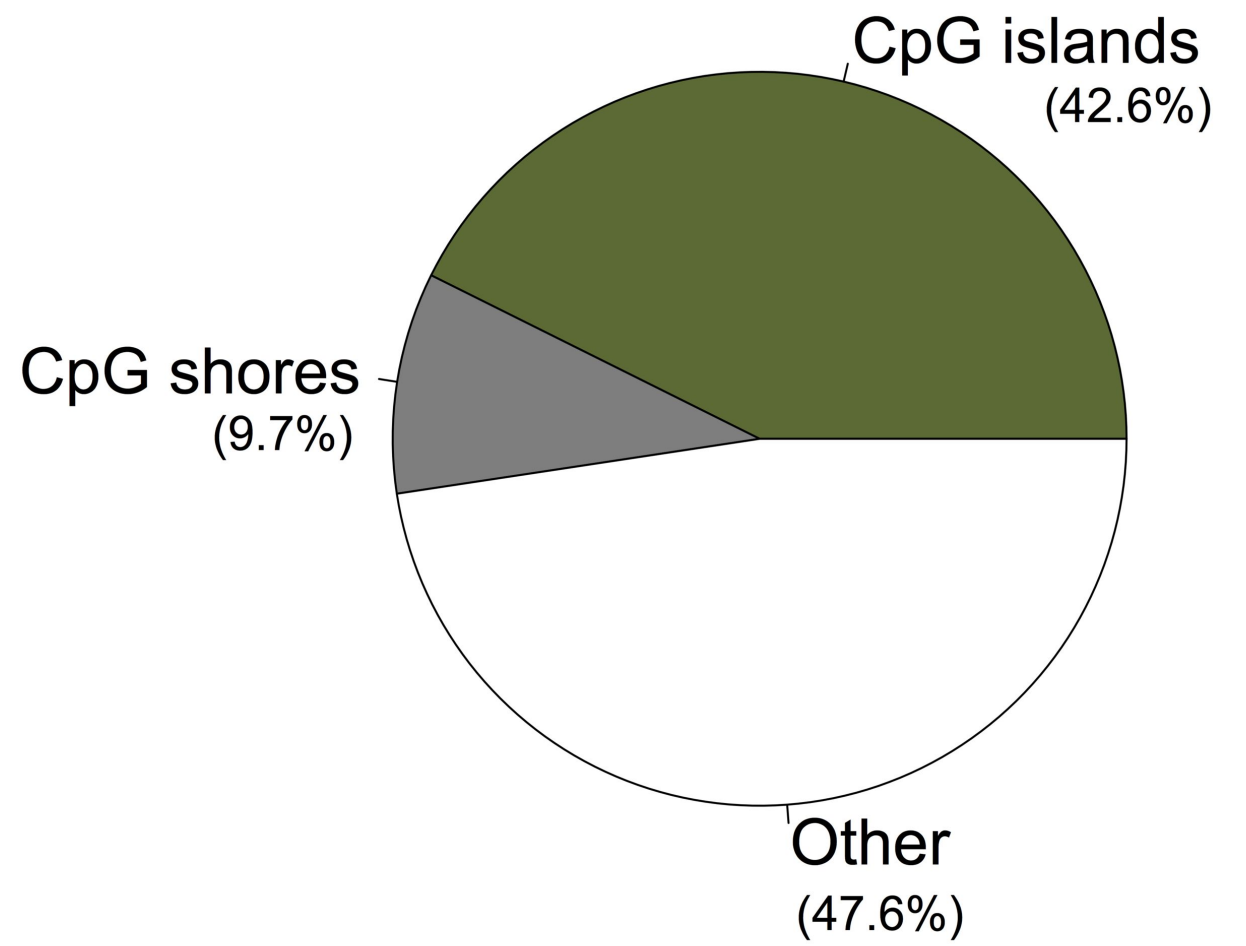


# Supplemental Figure 2.

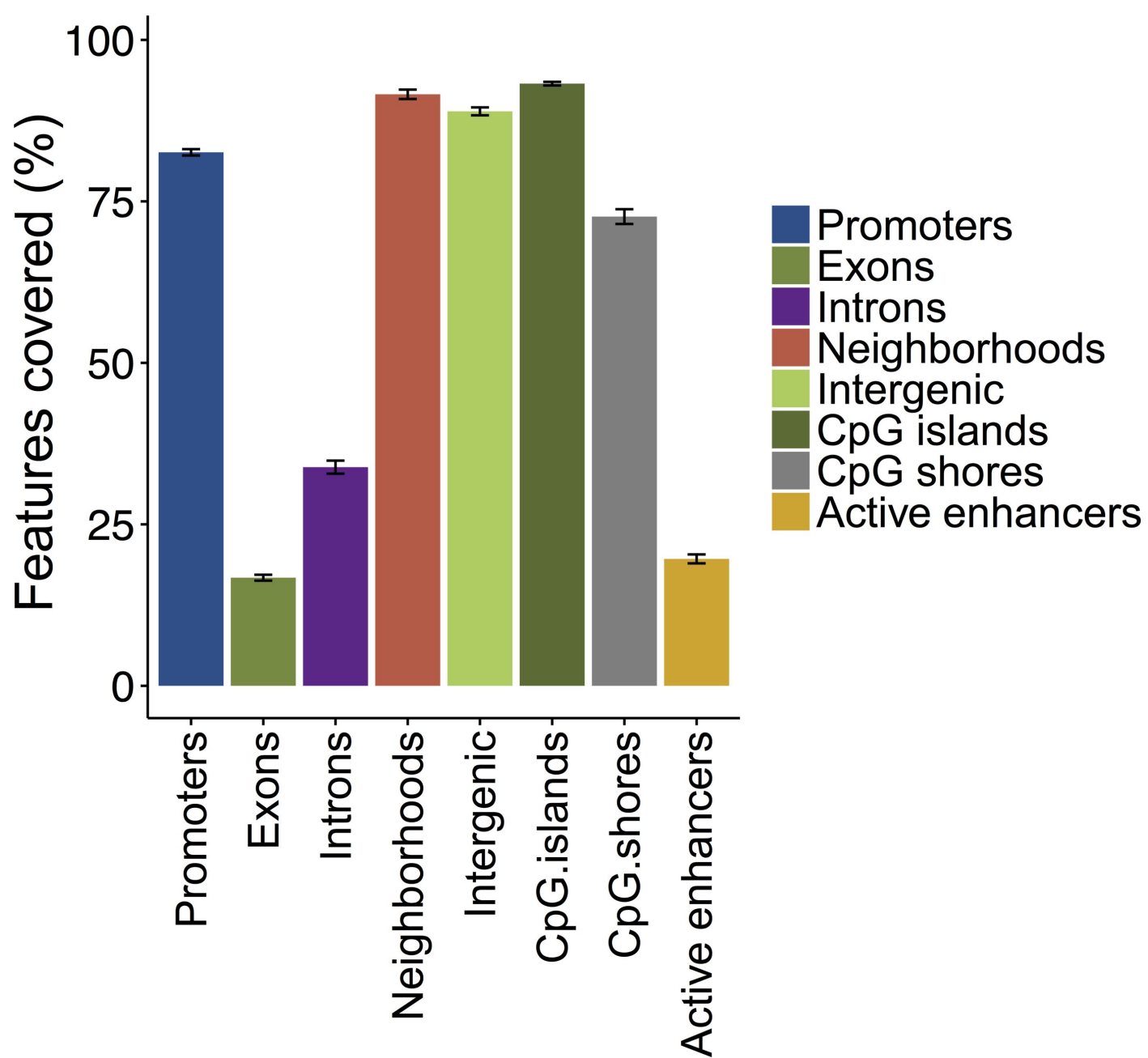
a)



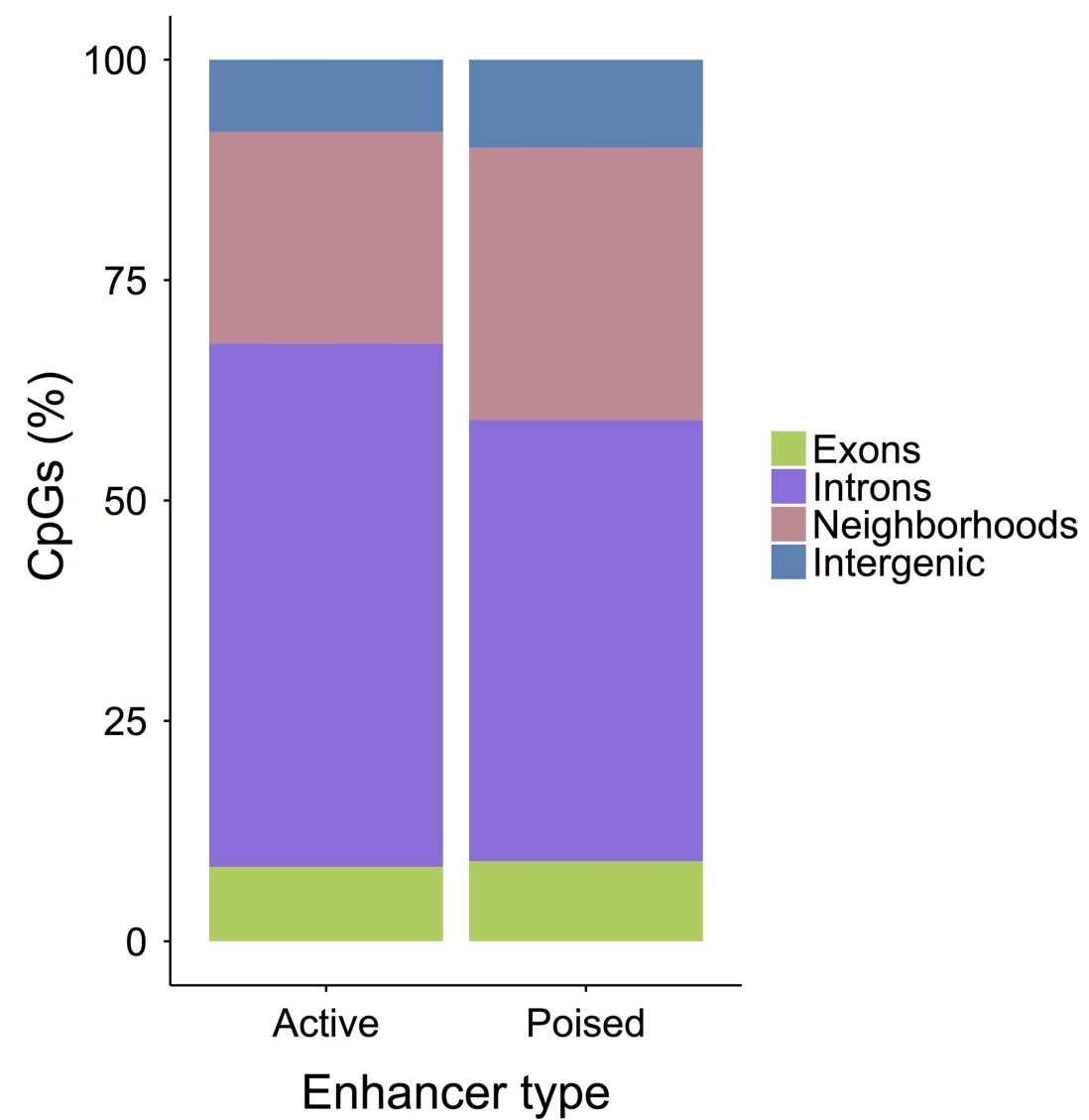
b)



c)

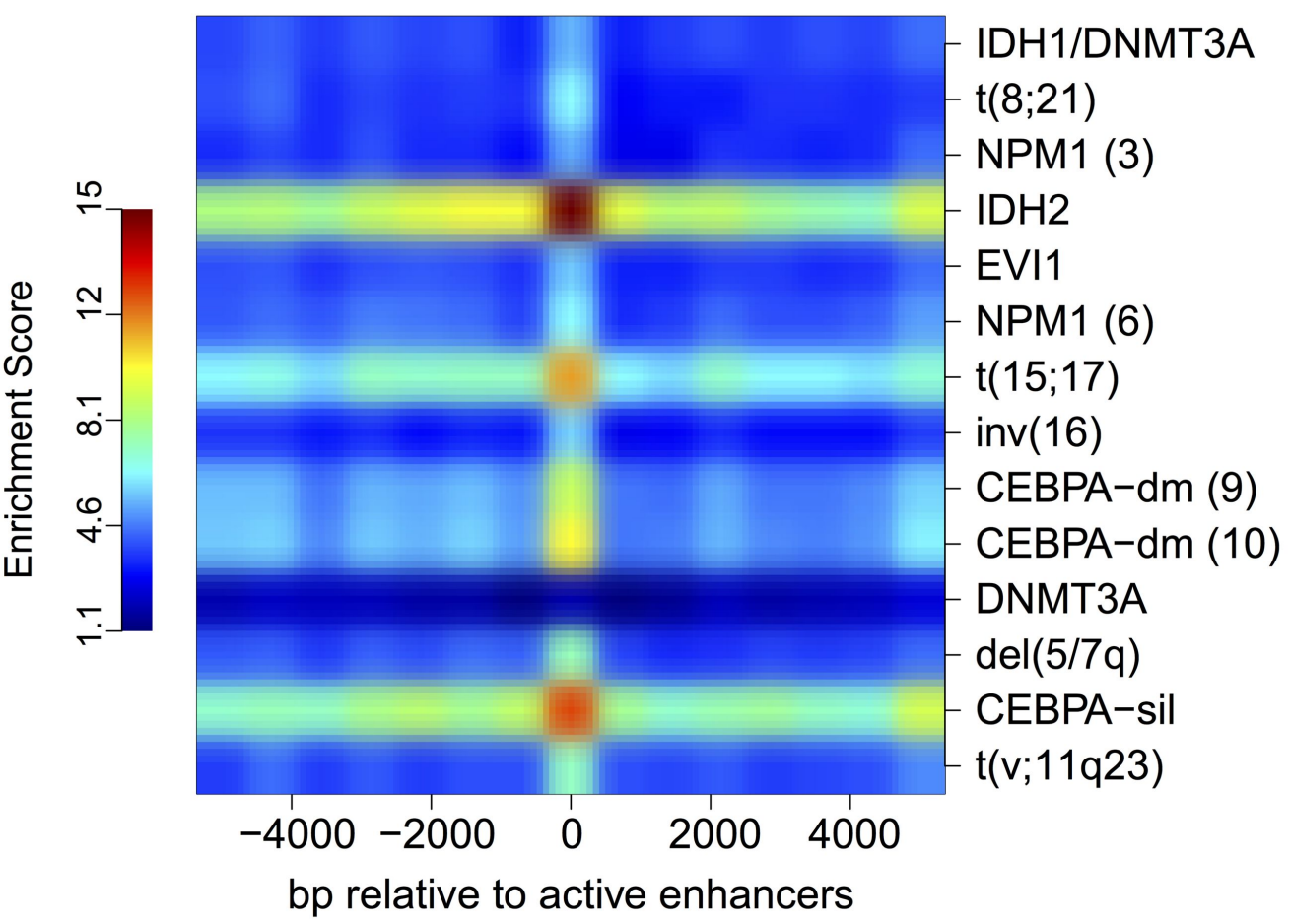


d)

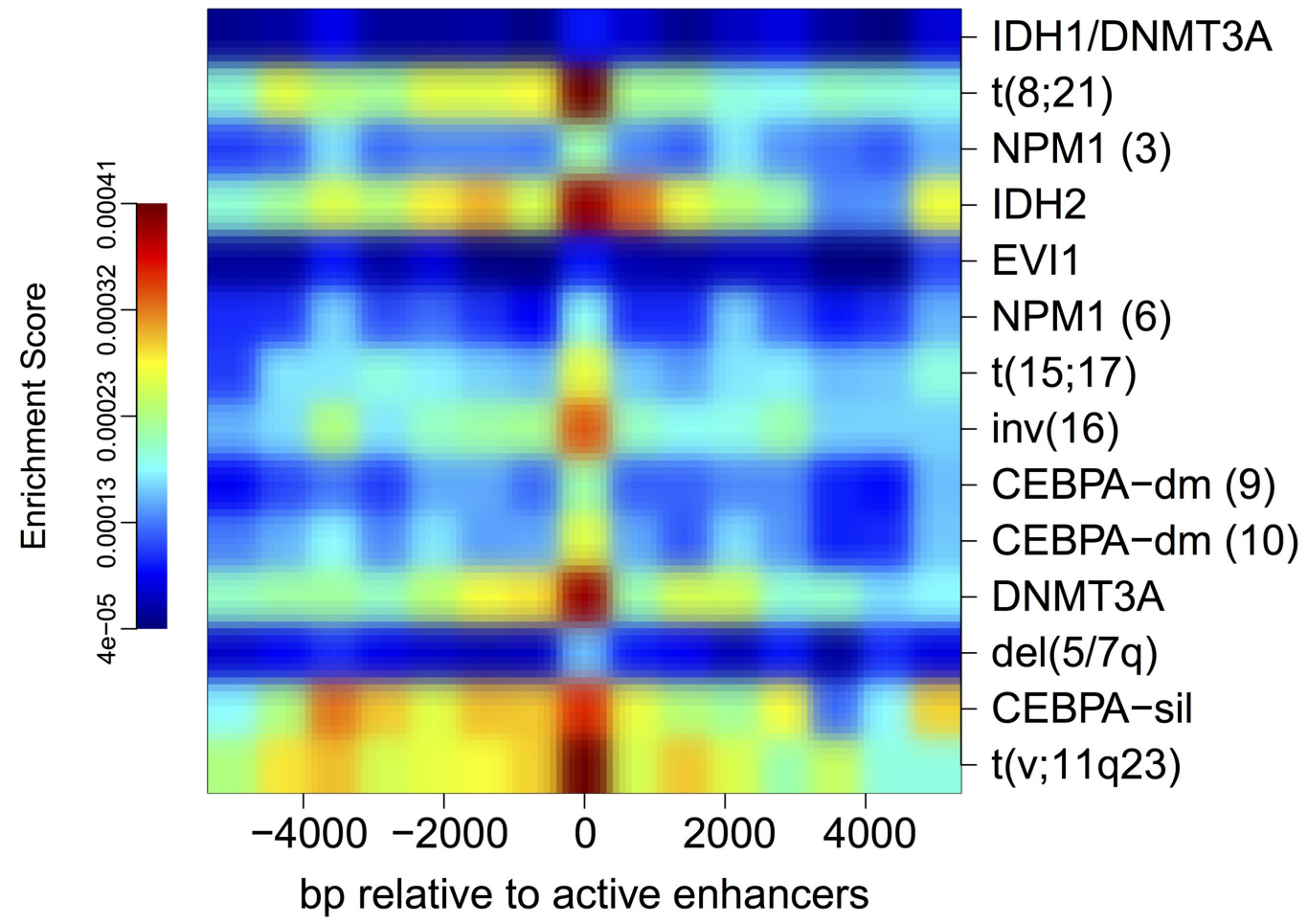


# Supplemental Figure 3.

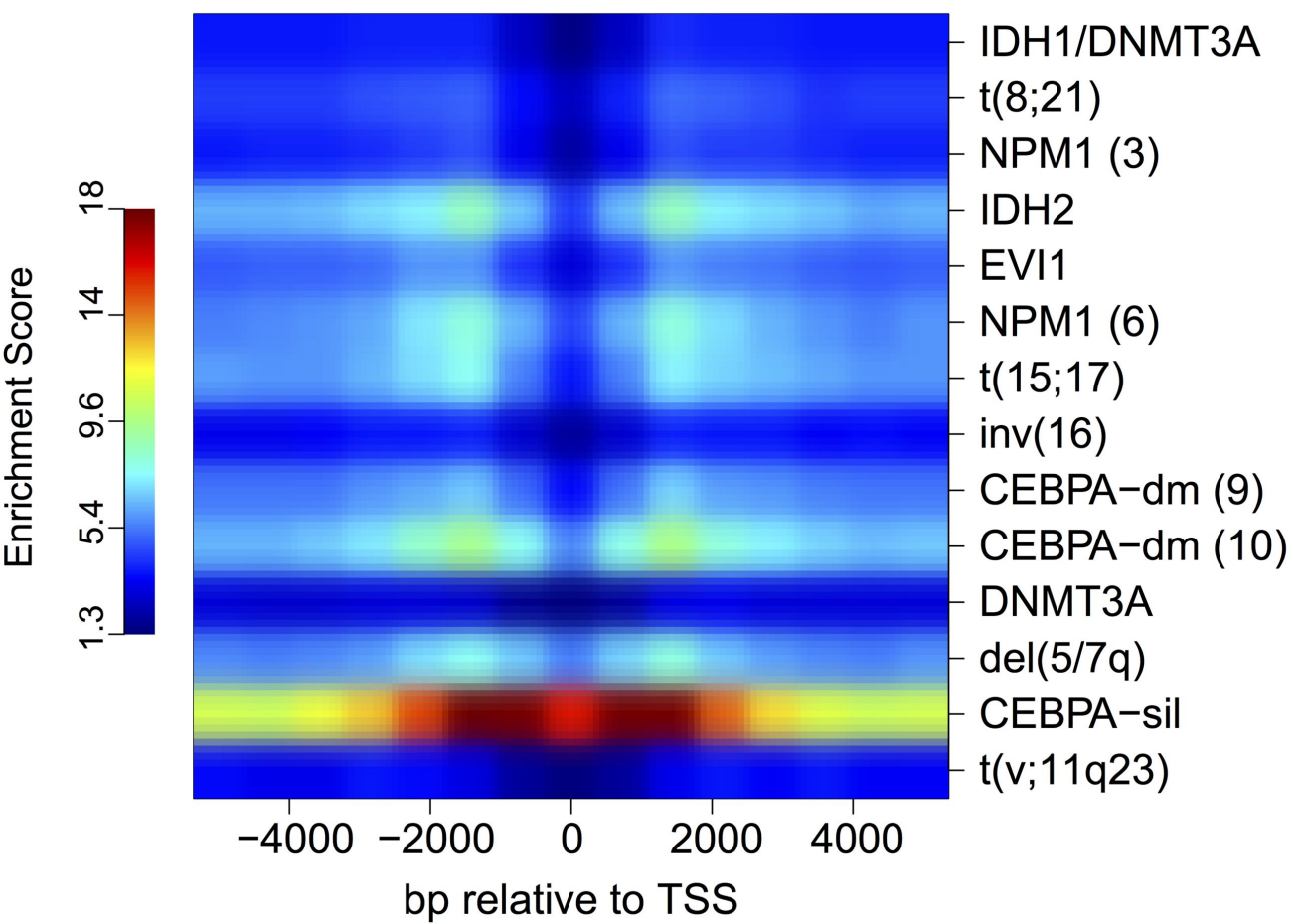
**a)**



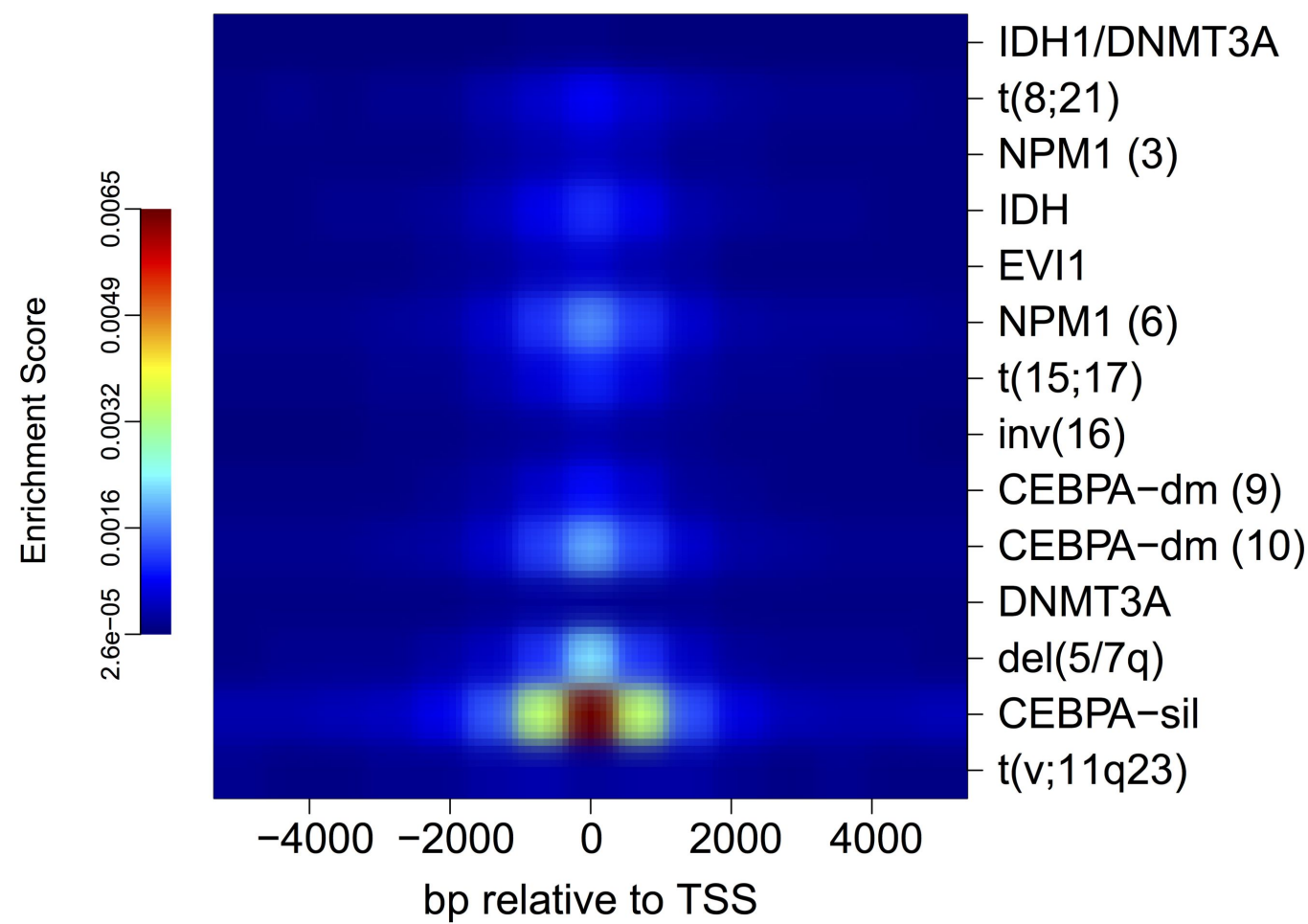
**b)**



**c)**

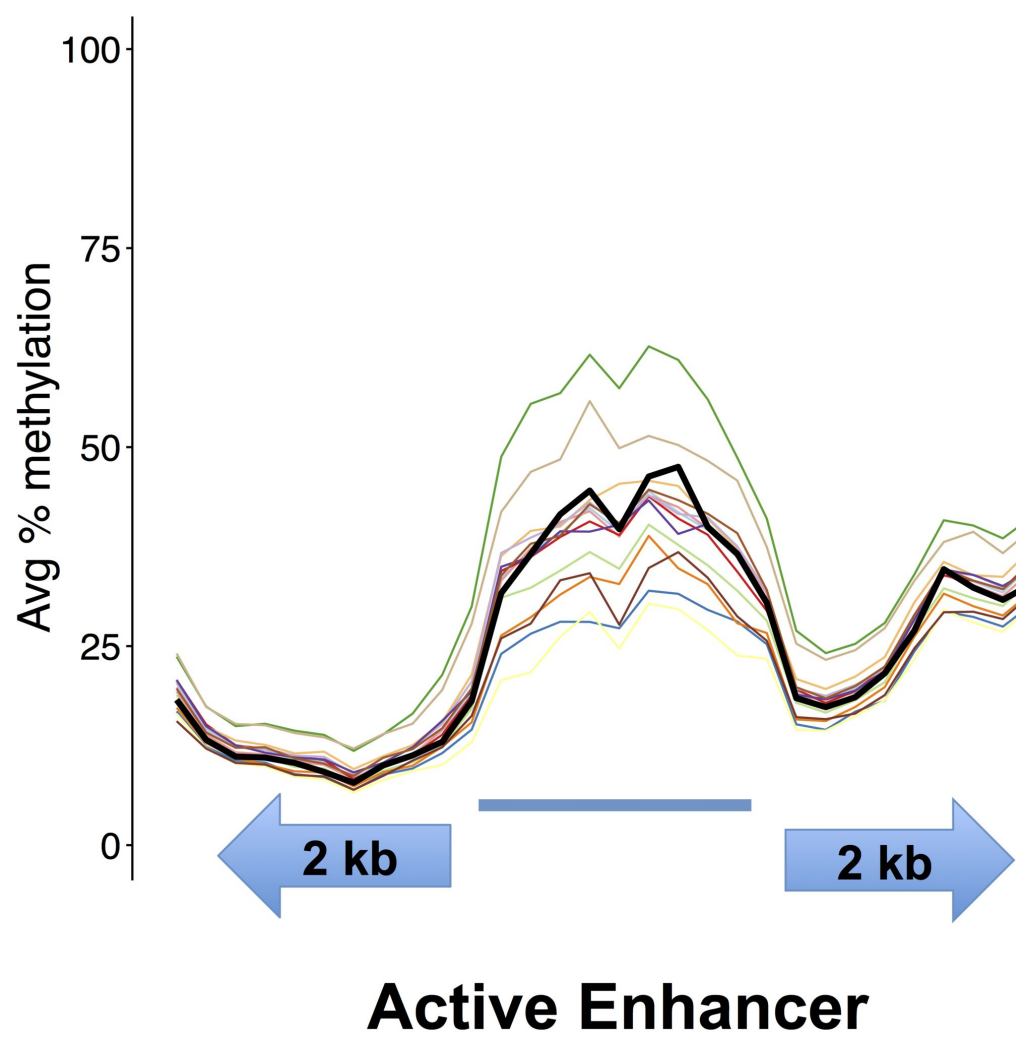


**d)**

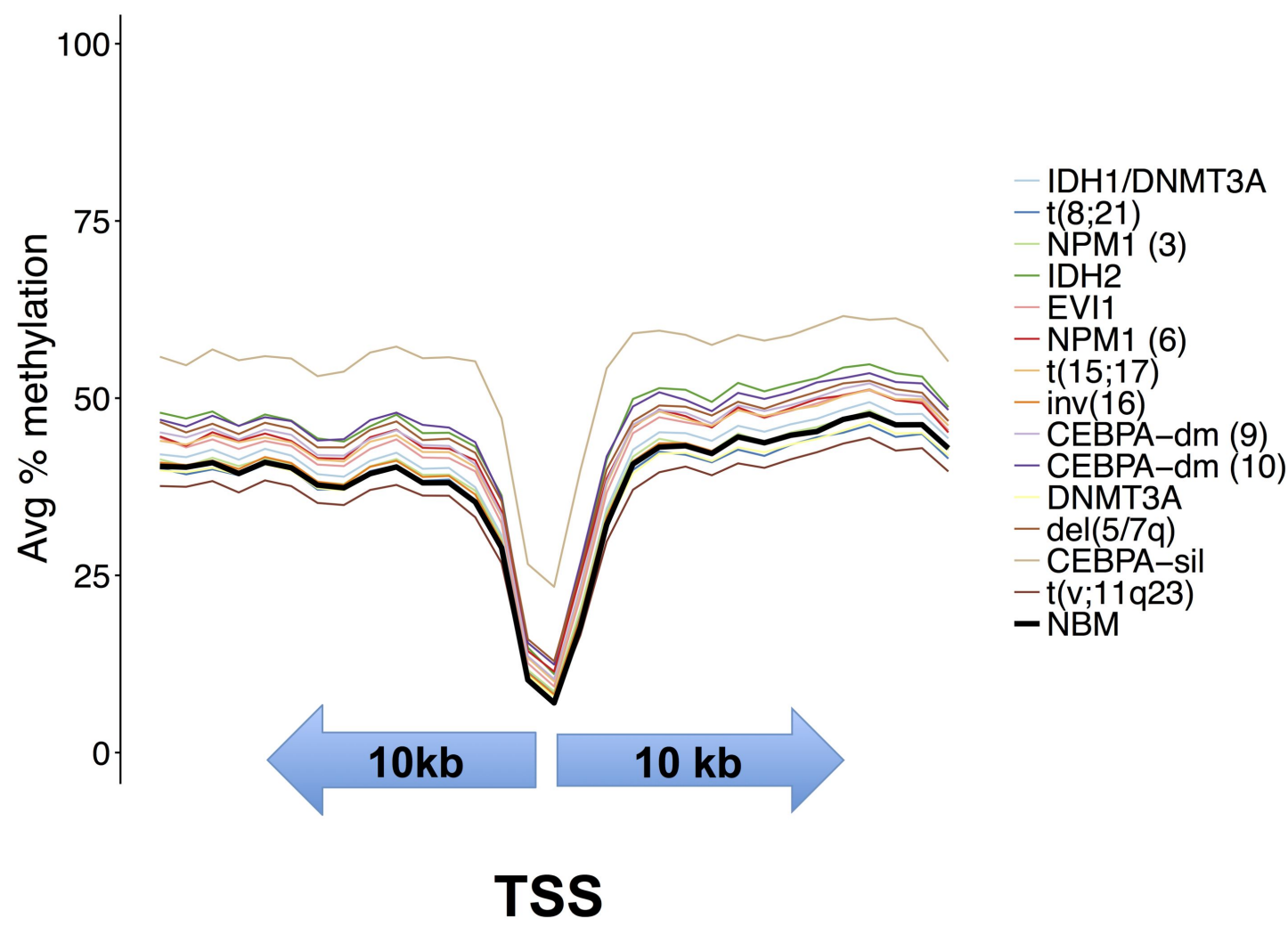


# Supplemental Figure 4.

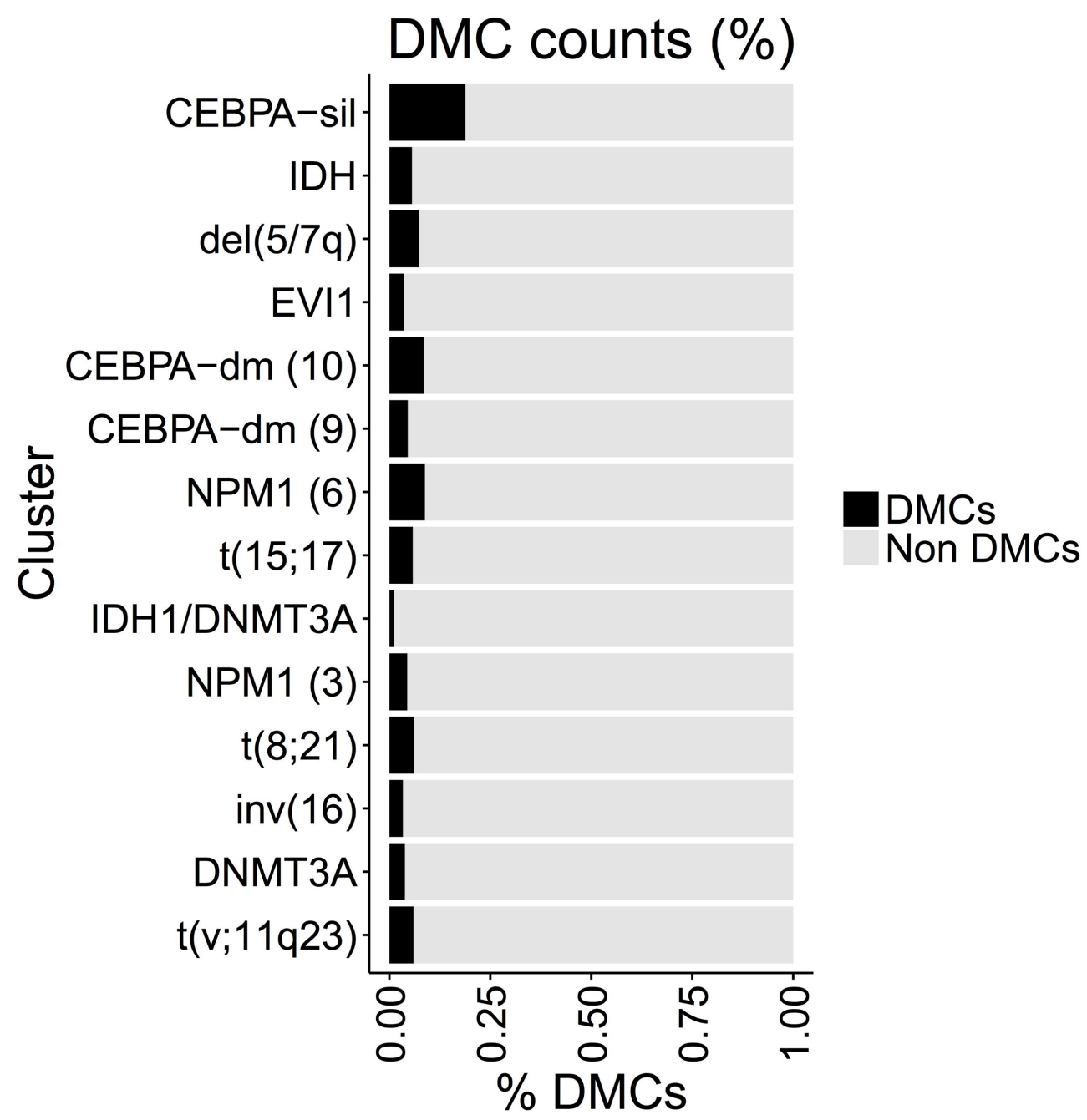
a)



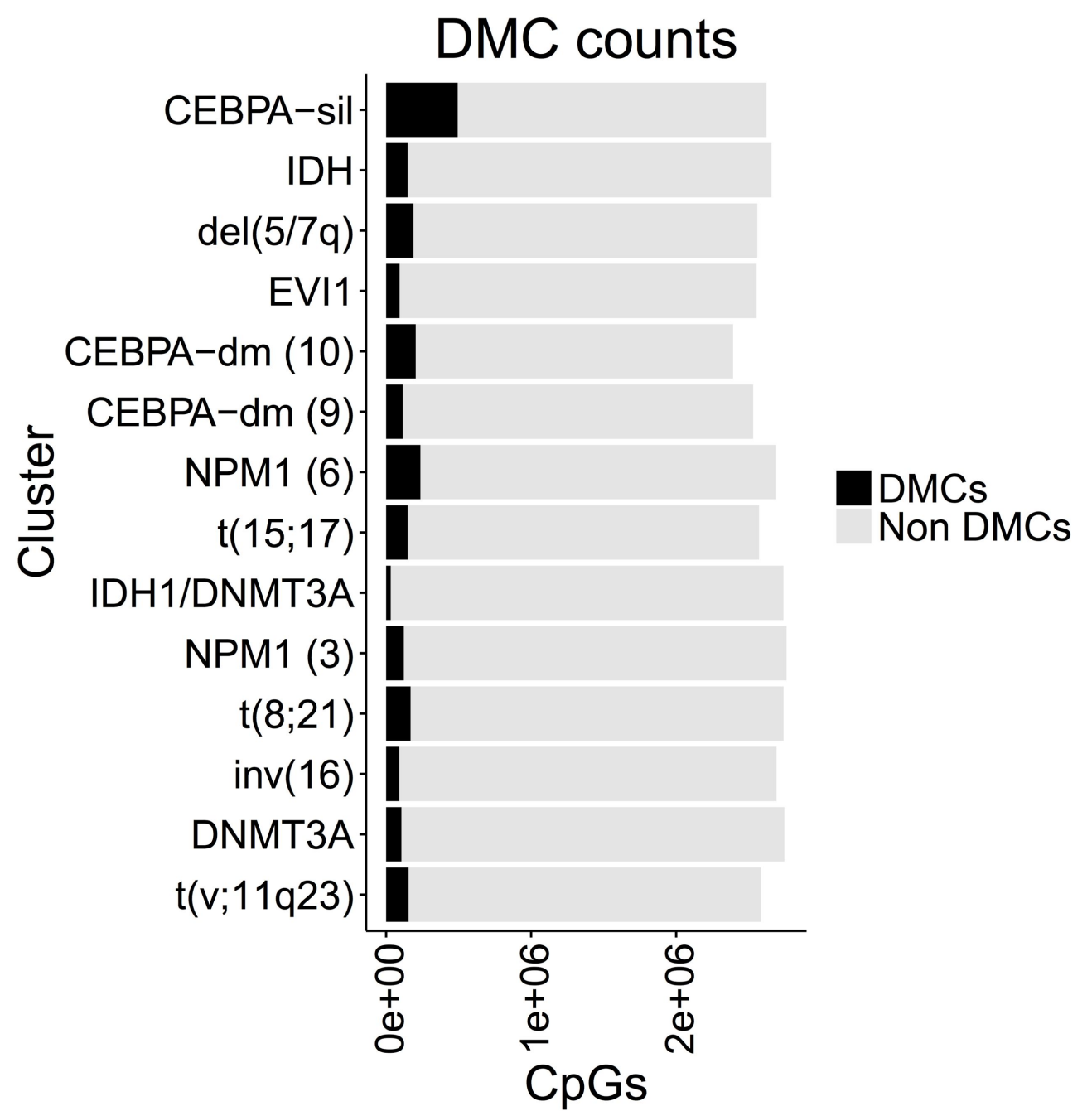
b)



c)






d)



# Supplemental Figure 5.



a)

## IDH2 motifs

Motif	Best match	log(p)	% targets	% background
	SMAD3	-4.7e+01	9.18%	0.45%
	RXRA	-3.7e+01	7.65%	0.53%
	USF2	-2.8e+01	9.18%	1.19%


b)

## DNMT3A motifs

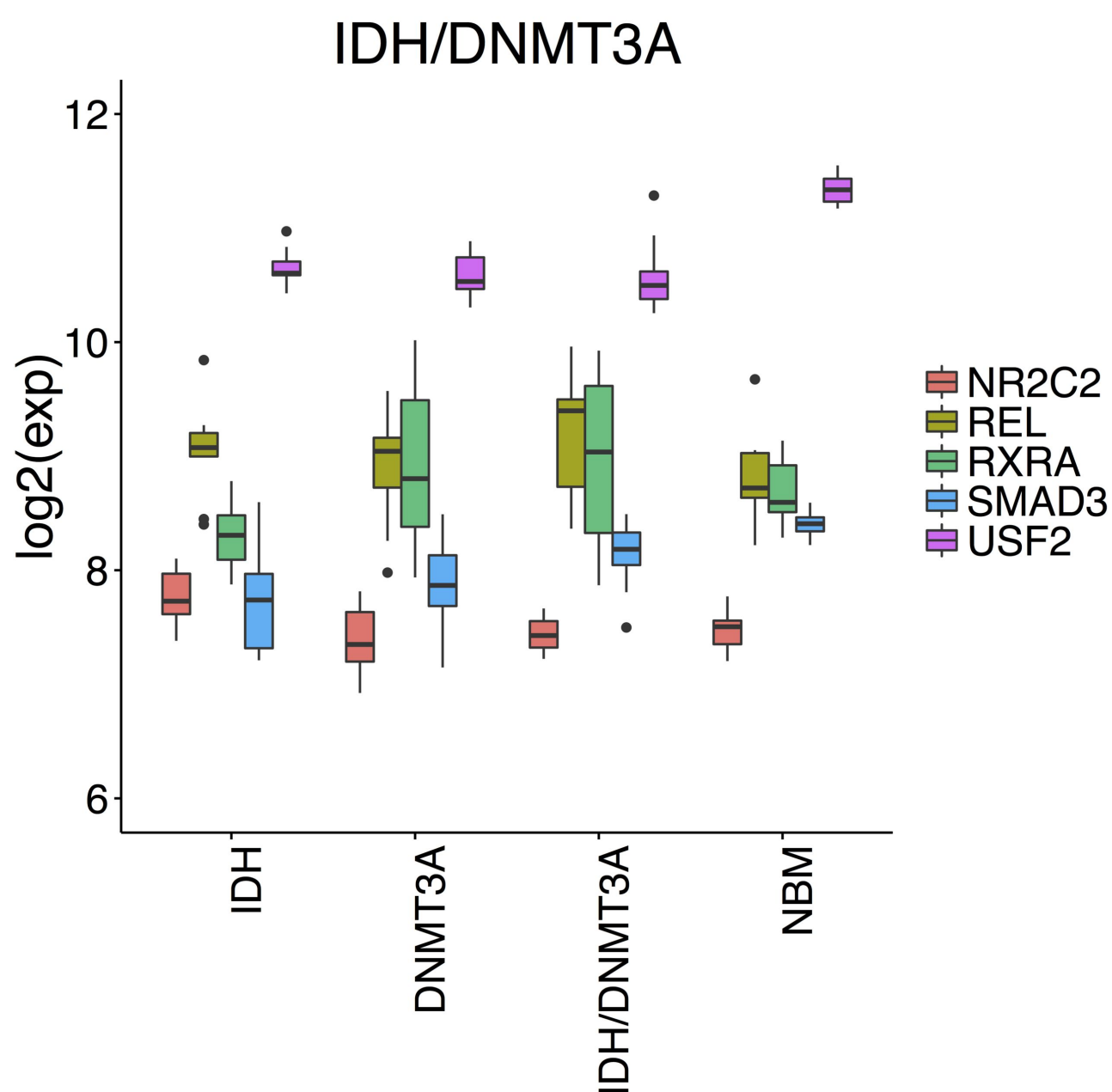
Motif	Best match	log(p)	% targets	% background
	SMAD3	-2.9e+01	20.38%	4.52%
	REL	-2.8e+01	7.01%	0.21%

c)

## IDH1/DNMT3A motif

Motif	Best match	log(p)	% targets	% background
	NR2C2	-2.3e+01	11.39%	0.59%

d)



e)

