List of Figures

1 Preprocessing

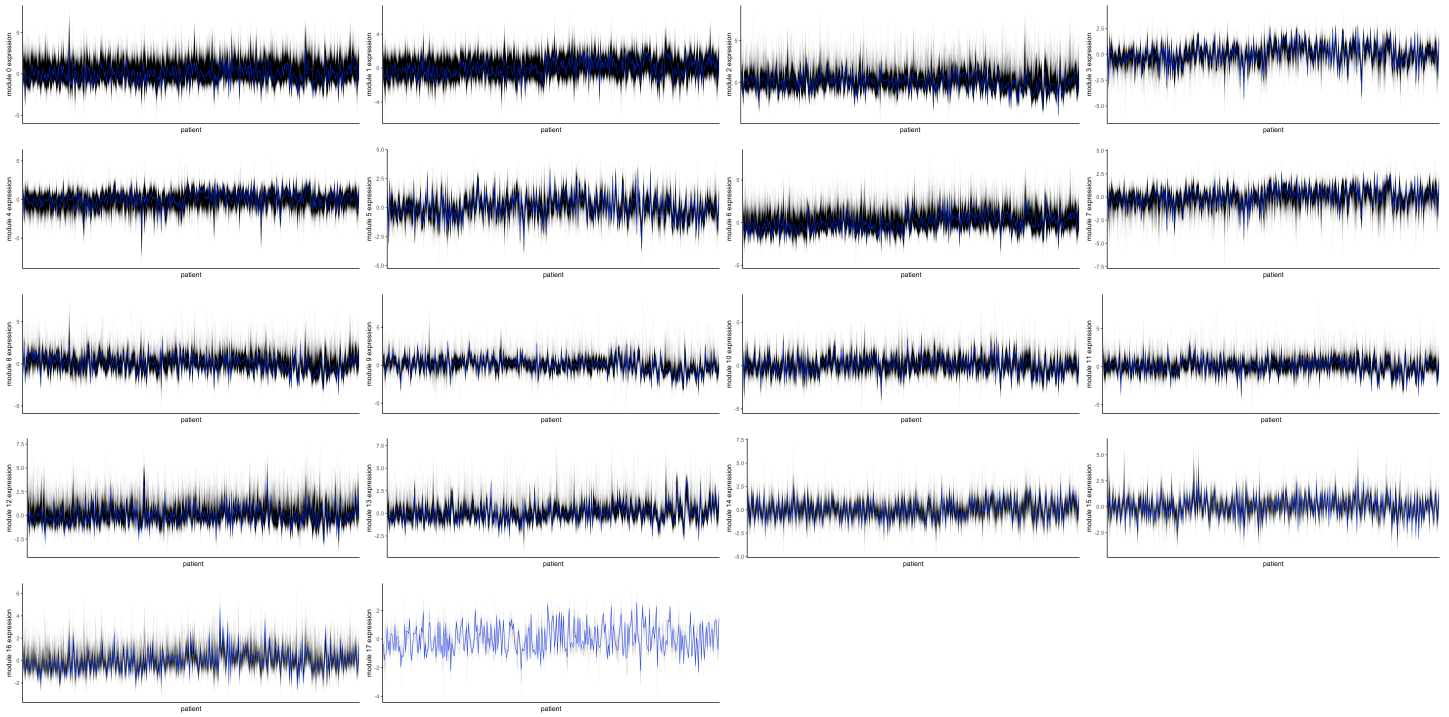
1.1 Gene expression array data

Gene expression array data for 344 AML patients was downloaded from GEO (GSE 14468). Normalization was performed in [**?**]. Each probe was annotated to the gene with nearest transcription start site and the corresponding gene symbol was provided in the file. The gene symbols of the probes are used to query the Gene Ontology [**?**] for biological process annotation. We are interested in genes related to methylation and transcriptional regulation. To remove irrelevant genes, we filter for genes that have keyword "methylation" and "transcrip" in their biological process annotation. In the end, 3417 out of 17788 remained for the downstream analysis.

1.2 Methylation array data

Methylation array data for 344 AML patients was downloaded from GEO (GSE 18700). Each probe was annotated to the nearest transcription start site allowing for a maximum distance of 5 kb from the TSS. Any probes that were mapped to a gene are filtered for downstream analysis and in total 22759 out of 25626 probes are left after the filtering step.

Normalization was performed on the probes across patients. PhenoGraph [**?**] were used to cluster probes. PhenoGraph first constructs a nearest-neighbor graph ($k = 30$) which has individual probe as a node and edge weight between each probe pair as Jaccard similarity coefficients between their nearest neighbor set. Louvain community detection algorithm was performed to find the clustering assignment that maximizes the modularity on the graph. 20 communities (modules) were obtained as a result. The counts of probes in each cluster along with the "eigengene" profile was computed for each module as their first principal component, shown in the Figure 1.



module	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
probes	3312	2810	2140	2013	1853	1842	1583	1325	1085	995	861	723	723	630	326	277	237	24

Figure 1: **First principal component of methylation modules** Each panel represent a module and the black lines represent the normalized methylation level of a probe across 344 patients. The blue lines show the first principal component of each module. The size of each module is shown in the table

In addition, the heatmap of expression level of the clusters are shown in Appendix Figure 2 and annotated by hits on GSEA.

2 Variational Bayes Spike Regression

A Methylation array heatmap

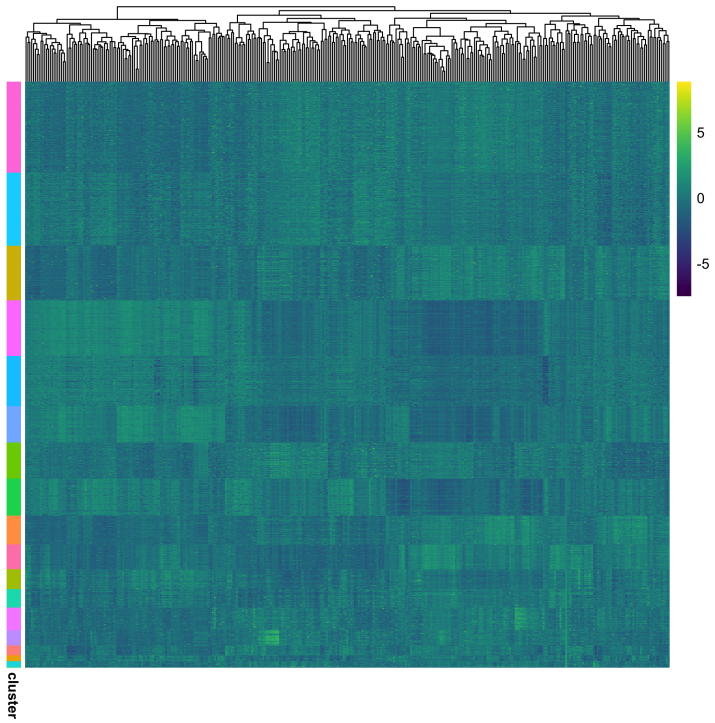


Figure 2: **Heatmap of methylation modules** Heatmap of methylation array expression. The rows are annotated with the PhenoGraph clusters and columns are arrange by hierarchical clustering on patients using Euclidean distance.