List of Figures

1 Preprocessing

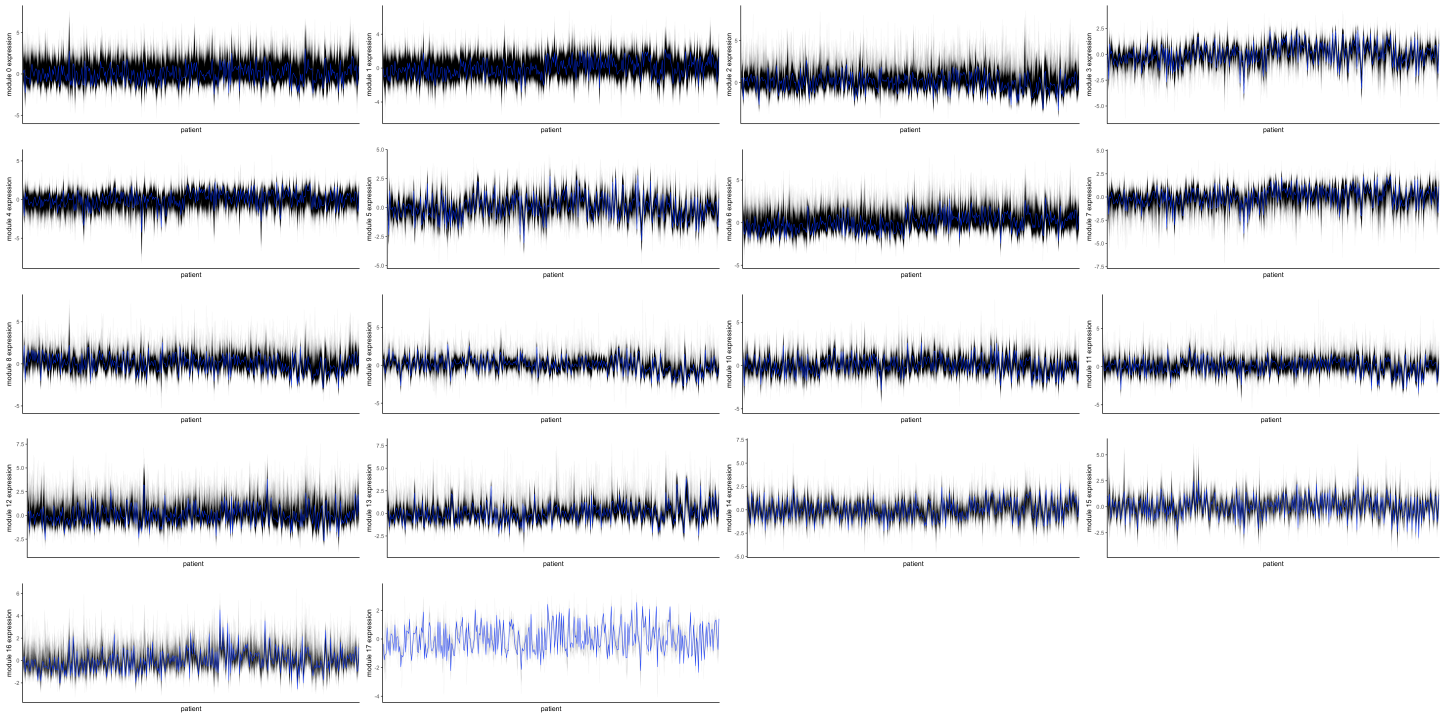
1.1 Gene expression array data

Gene expression array data for 344 AML patients was downloaded from GEO (GSE 14468). Normalization was performed in Gentles et al 2015 [1]. Annotation of the probes to gene with nearest transcription start site were available from GEO. We query for the Gene Ontology terms of the corresponding genes and filtered the probes based on their GO biological process annotation to obtain genes related to methylation and transcriptional regulation. After filter for genes that have keyword "methylation" or "transcrip", 3417 out of 17788 are left for the downstream analysis.

1.2 Methylation array data

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

Standardization was performed on the probes across patients. PhenoGraph [2] were used to cluster probes. PhenoGraph first constructs a nearest-neighbor graph ($k = 30$) that has individual probe as a node. The edge weight between probe pairs are computed as the Jaccard similarity coefficients between their nearest neighbor set. Louvain community detection was performed in this algorithm to find the clustering assignment that maximizes the modularity on the graph. After applying PhenoGraph on the 22759 methylation probes, 20 communities (modules) were obtained. The counts of probes in each module along with the first principal component of each module are 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 3.

1.3 Selection of significant gene expression probes using vBsr

We used Variational Bayesian spike regression (vBsr) [3] to select gene expression probes that are significantly associated with module methylation profile. vBsr is a penalized Bayesian regression model that uses a spike-and-slab prior to impose sparsity constraint on the regression coefficients. Fast computation were achieved by utilizing mean-field approximation. The algorithm was ran 50 times with random initialization to identify multiple local maxima of lower bound and Bayesian Model Averaging (BMA) was used to produce a unique estimate over all identified models. vBsr defines a test statistic z_{vb} and corresponding p-value. We applied Bonferroni correction to the p-values at $\alpha = 0.05$. Figure 2 shows the $-\log p$ values of the gene probes for modules that have at least one significant hits.

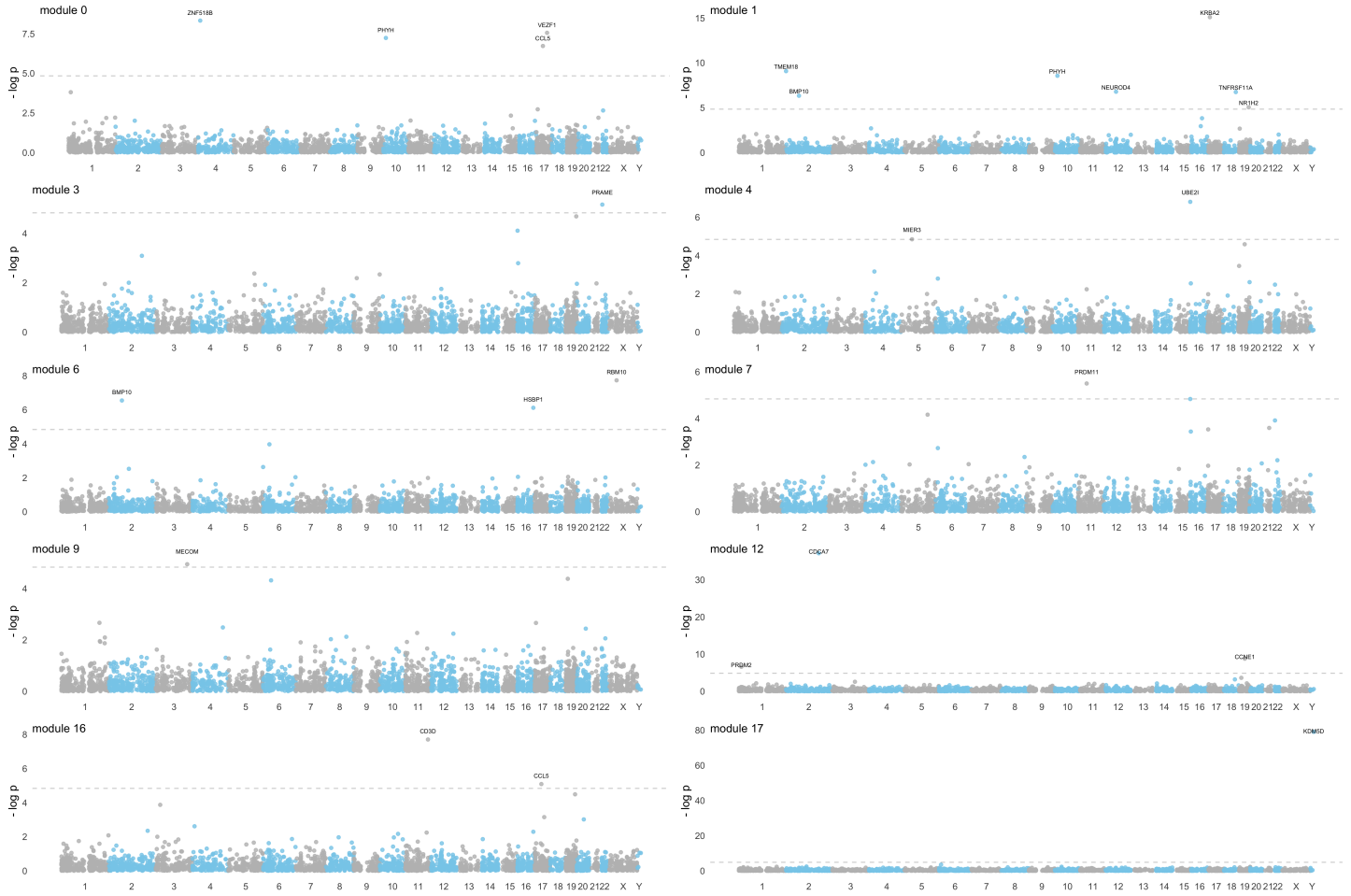


Figure 2: **Gene expression probes significantly associated with methylation module profile selected by vBsr** The Bonferroni corrected significant level is shown as the dashed lines and the significant hits are annotated with corresponding gene symbol.

A Methylation array heatmap

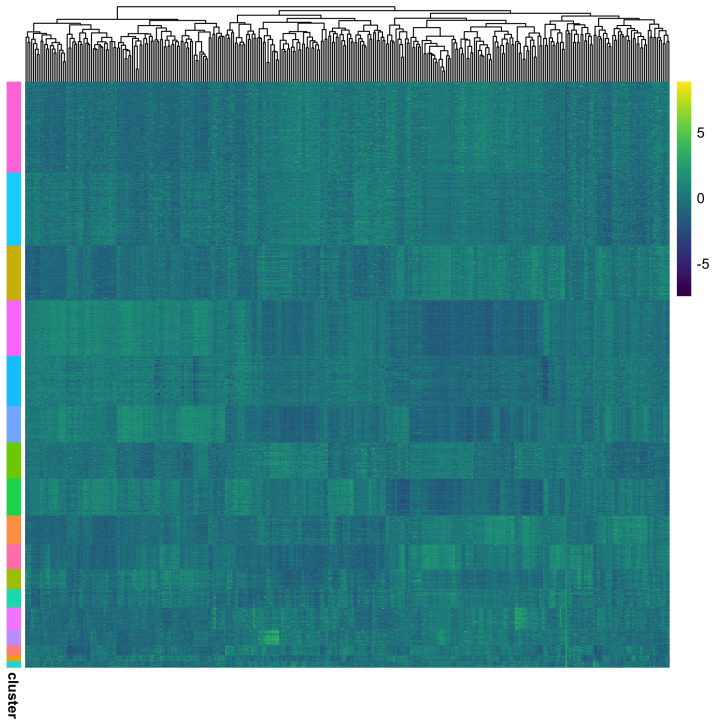


Figure 3: **Heatmap of methylation modules** Heatmap of methylation array expression. The rows are annotated with the PhenoGraph clusters and columns are arrange by hierarchical clustering on patients using Euclidean distance.

References

[1] A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim, V. S. Nair, Y. Xu, A. Khuong, C. D. Hoang, *et al.*, "The prognostic landscape of genes and infiltrating immune cells across human cancers," *Nature medicine*, vol. 21, no. 8, p. 938, 2015.

[2] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, D. A. El-ad, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, *et al.*, "Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis," *Cell*, vol. 162, no. 1, pp. 184–197, 2015.

[3] B. A. Logsdon, C. L. Carty, A. P. Reiner, J. Y. Dai, and C. Kooperberg, "A novel variational bayes multiple locus z-statistic for genome-wide association studies with bayesian model averaging," *Bioinformatics*, vol. 28, no. 13, pp. 1738–1744, 2012.