
Abstract

1. Introduction

2. Methods

2.1. Figueroa et al. (2010) DNA methylation and gene expression data

DNA methylation and gene expression microarray data of 344 AML cases from Figueroa et al. (2010) were downloaded from the GEO repository (GEO Accession number: GSE18700, GSE14468). Briefly, methylation data were obtained from custom design human promoter microarray for HELP assay, and gene expression data was obtained from Affymetrix Human Genome U133 Plus 2.0 Array.

Preprocessed gene expression data were obtained from Gentles et al. (2015), which normalized the raw CEL files with Affymetrix MAS5 algorithm and \log_2 transformed. We queried the GO terms on molecular function and biological process based on the HUGO gene symbol of the probes and filtered for probes that contain keywords related to transcription and methylation, which serves as predictors of our model.

For methylation data analysis, we followed Figueroa et al. (2010) and performed unsupervised hierarchical clustering of patients from methylation microarray using subset of the probes with standard deviation > 1 across all AML cases ($n = 3745$). Lingoes transformed 1 - Pearson correlation distance and Ward's minimum variance criterion were used. The patient clusters from the original study was reproduced. Next, we clustered the probes with similar methylation pattern using consensus clustering (Monti et al., 2003), which produces visual and quantitative stability evidence to a given number of clusters k and cluster assignments. R package *ConsensusClusterPlus* (Wilkerson & Hayes, 2010) provides an implementation of the method and were used

to determine the optimal number of clusters. We used the following options: 80 % probe subsampling, Ward's criterion with Lingoes transformation on 1 - Pearson correlation distance, 50 replicates for each k and maximum $k = 10$. We decides the number of optimal number of clusters based on the consensus matrix and area under CDF. The enrichment of biological processes represented by genes in each methylation cluster were examined using PANTHER overrepresentation test using Gene Ontology Biological process annotation data (Mi et al., 2013) with $FDR < 0.05$.

2.2. Selection of significant gene expression probes

Variational Bayesian spike regression (vBsr) (Logsdon et al., 2012) is a penalized Bayesian regression model that imposes sparsity constraint on the regression coefficients using a spike-and-slab prior and utilized mean-field approximation to achieve fast computation. In addition, the algorithm was ran multiple times with random initialization to identify multiple local maxima of lower bound and provides the option of Bayesian Model Averaging (BMA) over the identified models to produce an estimate to reduce the model uncertainty. vBsr also defines a test statistic z_{vb} associated with each penalized coefficients that allow control over the family-wise error rate by tuning the penalty parameter l_0 such that z_{vb} statistics are approximately $\mathcal{N}(0, 1)$ under the null hypothesis. We used vBsr to select the subset of gene expression probes that significantly associate with the observed pattern of variation of methylation level across patients. We ran 100 random starts for each model and used BMA option. We tuned the penalty parameter l_0 such that a feature will have a posterior probability of 0.95 if it passes a Bonferroni correction in the multivariate model to control the Type I error rate. Gene expression probes that were significant for z_{vb} at $\widehat{FDR} = 0.1$ were selected.

2.3. TCGA LAML RNA-seq and DNA methylation data

TCGA Level 3 mRNASeq and DNA methylation data was downloaded from Broad TCGA GDAC site for LAML. mRNASeq probes with missing values were imputed by KNN using R packages *impute* (Hastie et al., 2001). We queried the GO terms on molecular function and biological process based on the HUGO symbol of the probes and filtered for probes that contain keywords related to transcription and

¹Biomedical Informatics Training Program, Stanford University, CA, USA ²Department of Biomedical Data Sciences, Stanford University, CA, USA. Correspondence to: Andrew J. Gentles <andrewg@stanford.edu>.

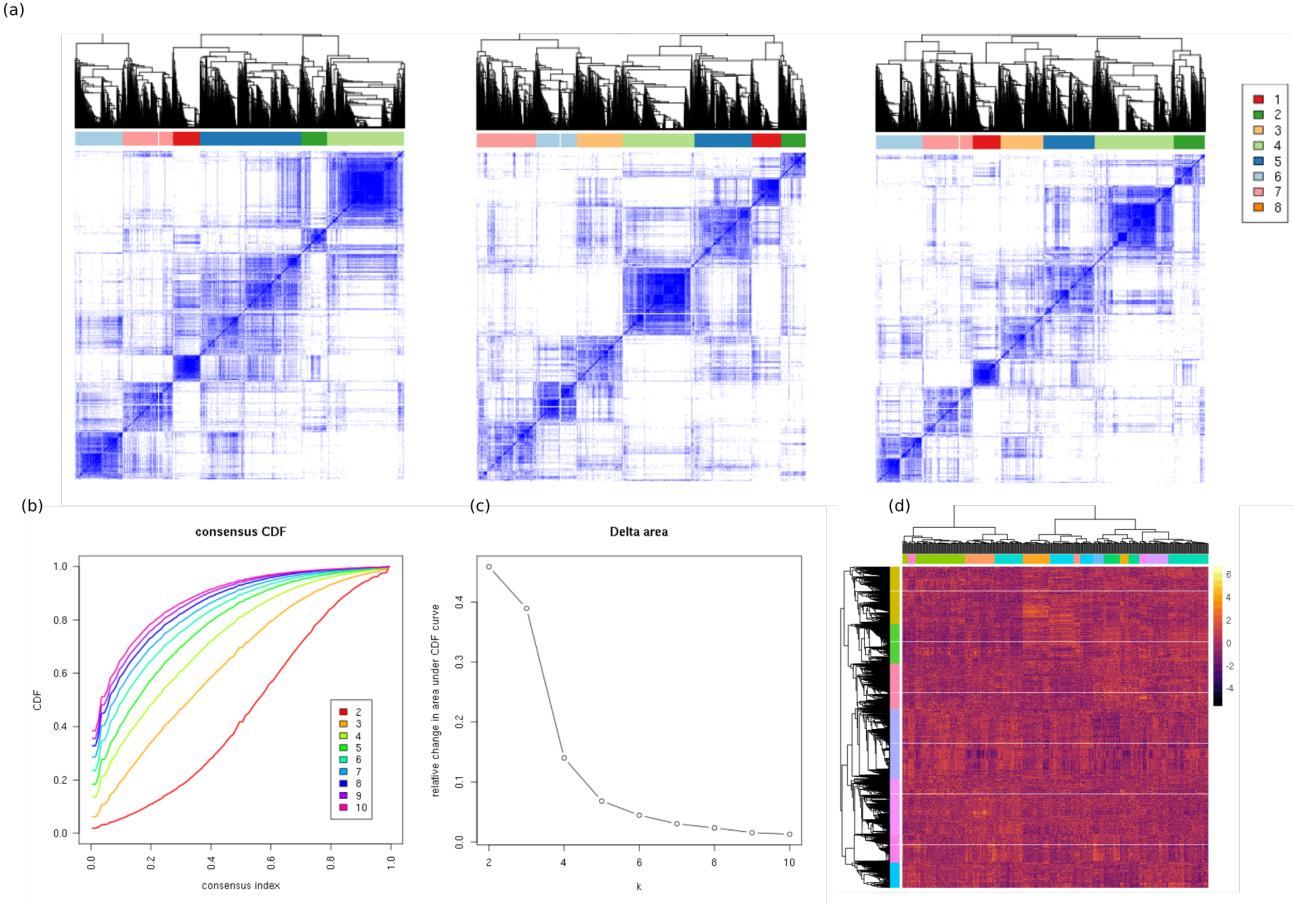


Figure 1. Consensus clustering of methylation probes for Figueroa et al. (2010) (a) consensus matrix for $k = 6, 7, 8$ (b) cumulative distribution function of consensus matrix at each $k = 1, \dots, 10$ (c) Area under CDF of consensus matrix for $k = 1, \dots, 10$ (d) Hierarchical clustering of cases and probes. Each row represents a probe and each column represents a patients. Methylation intensity level were row and column normalized. The 16 AML cases from Figueroa et al. (2010) were reproduced. Probes were clustered using Ward's method with lingo transformation of 1 - Pearson correlation distance transformed to Euclidean space and $k = 7$ were chosen as the cutoff from CC.

methylation as predictors for our model.

Methylation array data was obtained from Illumina Infinium HumanMethylation450 BeadChip. Probes that are on chromosome X and Y were removed. Probes with UCSC RefGene group annotation as TSS 1500 and located within UCSC CpG island annotation were selected. The correlation between the Beta value of filtered methylation probes and RSEM level of the corresponding genes in mRNASeq data were computed and significant methylation probes was filtered with FDR adjusted p-value (FDR < 0.1) of 0.05. The resulting probes were clustered using hierarchical clustering with euclidean distance using Ward's method. 20 clusters were set as the cutoff for assigning cluster membership. The average methylation level within the cluster of each patient was computed as the response vector.

3. Results

3.1. Methylation probes clustering

Figure 1 shows the results of consensus clustering on probes. Consensus of a pair of probe is defined to be the proportion of clustering runs on the resampled dataset that two probes are clustered together. Consensus matrices for cluster count $k = 1, \dots, 10$ were examined and $k = 7$ shows high intra-cluster consensus and low intercluster consensus among all cluster counts. Figure 1(a) shows heat map of consensus matrix for cluster count $k = 6, 7, 8$. The CDF plot Figure 1(b) shows that $k = 7$ approaches the maximum consensus distribution. The Delta Area Plot (Figure 1(c)) shows the area under the CDF curves and $k = 7$ is observed to have the largest k with a appreciable increase in consensus. These evidences suggest a cluster count of 7. Figure 1(d) shows

Enriched biological process	ref	overlap	expected	fold enrich	raw P	FDR
Cluster 1 (size 326)						
cell communication	5693	97	63.04	1.54	1.8e-6	2.4e-3
signaling	5578	96	61.77	1.55	1.5e-6	2.3e-3
Cluster 3 (size 671)						
cellular process	15478	436	379.56	1.15	6.1e-9	9.5e-5
Cluster 5 (size 816)						
developmental process	5654	310	199.38	1.55	9.2e-18	1.4e-13
anatomical structure development	5299	293	186.86	1.57	5.6e-17	4.4e-13
multicellular organism development	4918	275	173.42	1.59	2.7e-16	1.1e-12
system development	4309	244	151.95	1.61	1.0e-14	2.6e-11
Cluster 6 (size 656)						
developmental process	5654	199	147.52	1.35	2.2e-6	5.7e-3
cellular component organization or biogenesis	5773	196	150.62	1.30	3.1e-5	2.7e-2
cellular component organization	5584	192	145.69	1.32	2.0e-5	2.1e-2
anatomical structure development	5299	190	138.25	1.37	1.3e-6	5.2e-3
multicellular organism development	4918	184	128.31	1.43	1.1e-7	1.6e-3
Cluster 7 (size 461)						
negative regulation of biological process	5129	132	89.21	1.48	8.2e-07	1.3e-2

Table 1. Enriched biological processes in Figueroa et al. (2010) methylation probes cluster PANTHER overrepresentation test were performed using the associated gene with each methylation probe with control of $FDR < 0.05$. The table shows the test gene set (genes in each cluster), reference set size, overlap between the two sets, fold enrichment and corresponding raw p-value and FDR value. The hits with overlap between test and reference gene set at least 25% of the cluster size were shown in the table. Cluster 2 (size 290) and 4 (size 525) result in no significant hits.

the heat map of methylation level with 16 column cluster of patients and 7 row cluster of probes.

We examine each clusters for enrichment of specific biological processes using PANTHER overrepresentation test (Mi et al., 2013) and the annotations associated with each cluster were shown below in Table 1. Cluster 2 and 4 showed no significant enrichment results.

3.2. Selection of gene expression probes associated with methylation pattern

For each cluster, we averaged the methylation level across patients. Variational Bayes spike regression was used to select the gene expression probes that are significantly associated with the average methylation profile. Figure 2 shows the heat map for methylation level, average methylation profile, and selected gene expression level for cluster 1, 3, 5, 6, 7. Table 3.2 showed the significantly associated gene probes with each cluster and corresponding annotation and coefficient from vBsr model.

4. Discussion

References

Figueroa, Maria E, Lugthart, Sanne, Li, Yushan, Erpelinck-Verschueren, Claudia, Deng, Xutao, Christos, Paul J, Schifano, Elizabeth, Booth, James, van Putten, Wim, Skrabanek, Lucy, et al. Dna methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer cell*, 17(1):13–27, 2010.

Gentles, Andrew J, Newman, Aaron M, Liu, Chih Long, Bratman, Scott V, Feng, Weiguo, Kim, Dongkyoon, Nair, Viswam S, Xu, Yue, Khuong, Amanda, Hoang, Chuong D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine*, 21(8):938, 2015.

Hastie, Trevor, Tibshirani, Robert, Narasimhan, Balasubramanian, and Chu, Gilbert. impute: Imputation for microarray data. *Bioinformatics*, 17(6):520–525, 2001.

Logsdon, Benjamin A, Carty, Cara L, Reiner, Alexander P, Dai, James Y, and Kooperberg, Charles. A novel variational bayes multiple locus z-statistic for genome-wide

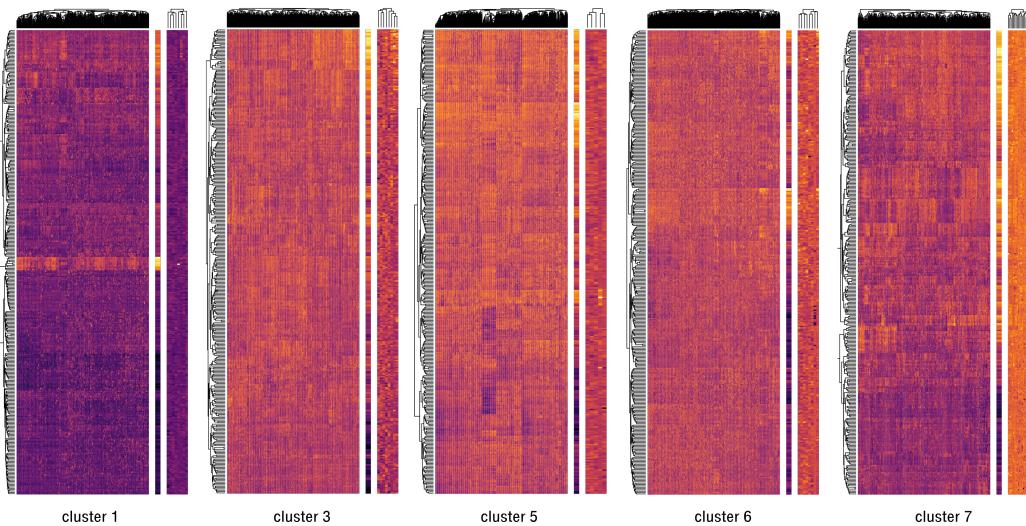


Figure 2. Heatmap of methylation clusters, average methylation level and gene expression probes significantly associate with average methylation level

association studies with bayesian model averaging. *Bioinformatics*, 28(13):1738–1744, 2012.

Mi, Huaiyu, Muruganujan, Anushya, Casagrande, John T, and Thomas, Paul D. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551, 2013.

Monti, Stefano, Tamayo, Pablo, Mesirov, Jill, and Golub, Todd. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

Wilkerson, Matthew D and Hayes, D Neil. Consensus-clusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.

gene	description	coef
Cluster 1		
ERBB2	erb-b2 receptor tyrosine kinase 2	0.002
ZBTB38	zinc finger and BTB domain containing 38	-0.00068
HDAC2	histone deacetylase 2	0.00059
RBPJ	recombination signal binding protein for immunoglobulin kappa J region	-0.00019
POU6F1	POU class 6 homeobox 1	0.0023
TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	0.0002
CDCA7	cell division cycle associated 7	-0.00015
AURKB	aurora kinase B	0.0004
Cluster 3		
MED13L	mediator complex subunit 13 like	-0.00015
INTS6	integrator complex subunit 6	-0.00021
MGMT	O-6-methylguanine-DNA methyltransferase	0.0006
SOX15	SRY-box 15	0.00071
TRAF7	TNF receptor associated factor 7	0.00028
ABHD14B	abhydrolase domain containing 14B	-0.00012
PDLIM1	PDZ and LIM domain 1	1.1e-05
ZMYM4	zinc finger MYM-type containing 4	0.00018
Cluster 5		
TLR6	toll like receptor 6	0.0007
SSU72	SSU72 homolog, RNA polymerase II CTD phosphatase	0.00013
BCL7A	BCL tumor suppressor 7A	0.00021
SARS	seryl-tRNA synthetase	-0.0003
SFPQ	splicing factor proline and glutamine rich	-2.8e-05
Cluster 6		
DEAF1	DEAF1, transcription factor	8.1e-05
ATF7IP	activating transcription factor 7 interacting protein	-5.1e-05
PSEN1	presenilin 1	-0.00022
CREBZF	CREB/ATF bZIP transcription factor	0.00031
RBL2	RB transcriptional corepressor like 2	0.00024
VLDLR	very low density lipoprotein receptor	-6.4e-05
ZC3H8	zinc finger CCCH-type containing 8	-0.00045
ZBTB24	zinc finger and BTB domain containing 24	0.00015
Cluster 7		
MAMLD1	mastermind like domain containing 1	-0.00059
AKT1	AKT serine/threonine kinase 1	9e-05
F2RL1	F2R like trypsin receptor 1	-0.00015
GFI1	growth factor independent 1 transcriptional repressor	8.6e-05
GLO1	glyoxalase I	-3.8e-05
HNMT	histamine N-methyltransferase	0.00092
MED1	mediator complex subunit 1	-0.00018
PAK6	p21 (RAC1) activated kinase 6	0.00033
ZNF286A	zinc finger protein 286A	-0.004
KIAA1549	KIAA1549	0.00047
ZBED5	zinc finger BED-type containing 5	-0.00034
SCML1	Scm polycomb group protein like 1	-0.00043
CRTC3	CREB regulated transcription coactivator 3	-8.9e-05
SLC11A1	solute carrier family 11 member 1	0.00019
TP53BP1	tumor protein p53 binding protein 1	0.00032
CBX2	chromobox 2	-0.00037
TRIM15	tripartite motif containing 15	0.0005

Table 2. Gene expression probes significantly associated with methylation cluster