??

Meng Wang [1] **Andrew J. Gentles** [2]

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

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

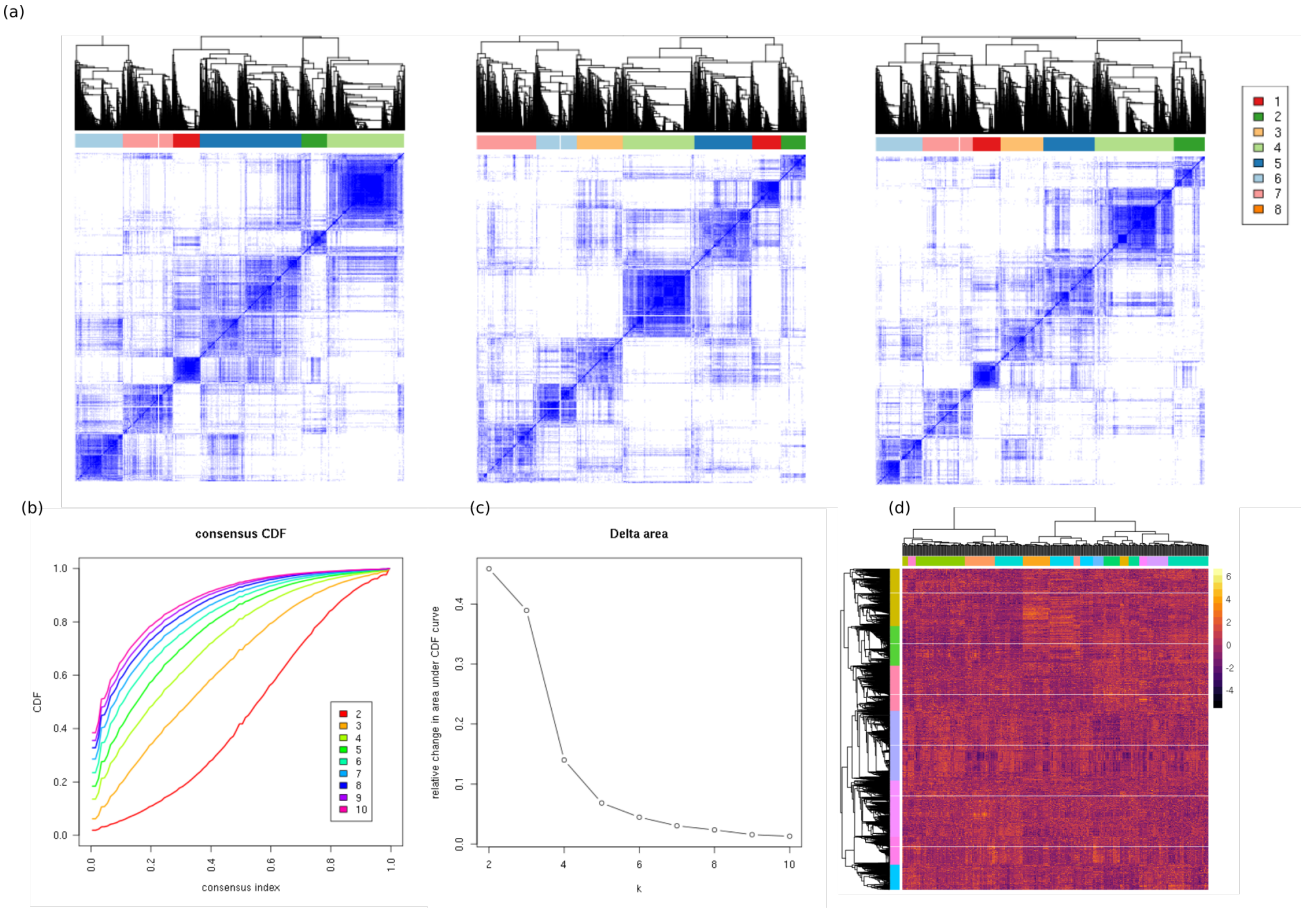

Figure 1. **Consensus clustering of methylation probes for Figueroa et al. (2010)** (a) consensus matrix for k = 6, 7, 8. $k = 7$ shows high intra-cluster consensus and low intercluster consensus. (b) cumulative distribution function of consensus matrix at each $k = 1, \ldots, 10$. $k = 7$ approaches the maximum consensus distribution (c) Area under CDF of consensus matrix for $k = 1, \ldots, 10$. $k = 7$ is the largest k with a appreciable increase in consensus (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. The16 clusters of AML cases from Figueroa et al. (2010) were reproduced. Probes were clustered using Ward's method with 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. Significant associated gene expression probes in Figueroa et al. (2010)

Figure 1 shows the results of consensus clustering on probes and significantly enriched biological processes in each cluster is shown below in Table 1.

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

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**

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 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. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12): 1572–1573, 2010.