Supplementary Materials for "Outcome-Guided Disease Subtyping for High-Dimensional Omics Data"

Peng Liu, Yusi Fang, Ren Zhao, Lu Tang[*]and George C. Tseng[†]

January 12, 2022

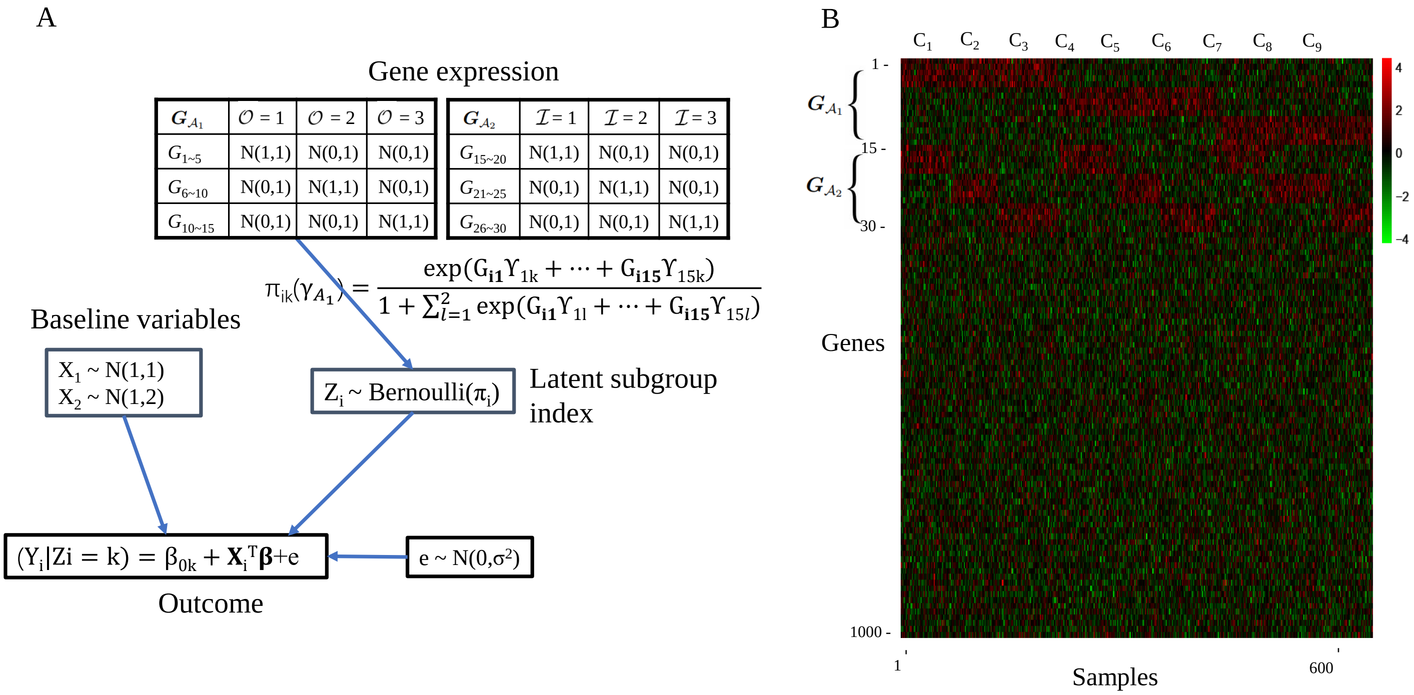
[*] Correspond to: lutang@pitt.edu
[†] Correspond to: ctseng@pitt.edu

Table S1: Comparison of sparse k-means (SKM), penalized model based clustering (PMBC), supervised clustering (SC) and outcome-guided clustering (ogClust) under four simulation model settings with 600 observations and 2 baseline covariates, 1000 genes and $G_{j|j\in\mathcal{A}_2} \sim N(3,1)$ in 100 repetitions.

Methods	Estimated K			ARI	Selected Genes		Outcome	
	2	3	>3		FPs	FNs	RMSE	R^2
Model 1: $\gamma = 1; \delta = 2$								
SKM	37	63	0	0.00	286.2	9.9	1.94	0.24
PMBC	0	97	3	0.00	78.7	13.4	1.93	0.24
SC	100	0	0	0.35	45.3	4.8	1.59	0.46
ogClust	41	59	0	0.45	5.8	3.0	1.55	0.51
Model 2: $\gamma = 1; \delta = 3$								
SKM	37	63	0	0.00	286.2	9.9	2.68	0.14
PMBC	0	95	5	0.00	85.3	13.4	2.68	0.13
SC	100	0	0	0.36	28.5	5	2.13	0.45
ogClust	2	98	0	0.86	14.6	0	1.90	0.55
Model 3: $\gamma = 1; \delta = 5$								
SKM	37	63	0	0.00	286.2	9.9	4.25	0.05
PMBC	0	98	2	0.00	92.0	13.2	4.27	0.04
SC	100	0	0	0.36	32.4	4.9	3.25	0.42
ogClust	1	99	0	0.91	14.4	0	2.72	0.61
Model 4: $\gamma = 3; \delta = 3$								
SKM	38	62	0	0.00	406.1	8.5	2.67	0.14
PMBC	0	100	0	0.00	82.0	13.8	2.67	0.13
SC	100	0	0	0.41	17.5	5	2.20	0.41
ogClust	2	98	0	0.88	12.1	0	1.74	0.63

Table S2: Comparison of sparse k-means (SKM), penalized model based clustering (PMBC), supervised clustering (SC) and outcome-guided clustering (ogClust) under four simulation model settings with 600 observations and 2 baseline covariates, 1000 genes and $G_{j|j\in\mathcal{A}_2} \sim N(0.5, 1)$ in 100 repetitions.

Methods	Estimated K			ARI	Selected Genes		Outcome	
	2	3	> 3		FPs	FNs	RMSE	R^2
Model 1: $\gamma = 1; \delta = 2$								
SKM	100	0	0	0.05	794.0	1.6	1.94	0.24
PMBC	73	1	26	0.33	182.7	4.4	1.93	0.24
SC	100	0	0	0.35	47.9	4.8	1.59	0.48
ogClust	66	31	3	0.45	14.0	3.3	1.55	0.51
Model 2: $\gamma = 1; \delta = 3$								
SKM	100	0	0	0.05	794.0	1.6	2.66	0.13
PMBC	74	0	26	0.30	172.0	5.9	2.68	0.13
SC	100	0	0	0.36	51.0	4.8	2.09	0.47
ogClust	2	97	1	0.86	21.3	0.1	1.90	0.56
Model 3: $\gamma = 1; \delta = 5$								
SKM	100	0	0	0.05	794.0	1.6	4.22	0.05
PMBC	69	1	30	0.28	171.8	6.5	4.24	0.04
SC	100	0	0	0.36	47.5	4.8	3.21	0.46
ogClust	0	100	0	0.91	5.1	0.1	2.70	0.61
Model 4: $\gamma = 3; \delta = 3$								
SKM	100	0	0	0.06	769.0	2.0	2.62	0.15
PMBC	77	1	22	0.33	187.8	6.4	2.64	0.15
SC	100	0	0	0.41	17.3	5.0	2.02	0.51
ogClust	0	100	0	0.88	2.5	0.1	1.75	0.63


A
Gene expression
$\pi_{ik}(\gamma_{\mathcal{A}_1}) = \frac{\exp(G_{\mathbf{i1}}\Upsilon_{1k}+\cdots+G_{\mathbf{i15}}\Upsilon_{15k})}{1+\sum_{l=1}^{2}\exp(G_{\mathbf{i1}}\Upsilon_{1l}+\cdots+G_{\mathbf{i15}}\Upsilon_{15l})}$
Baseline variables
$X_1 \sim N(1,1)$
$X_2 \sim N(1,2)$
$Z_i \sim \text{Bernoulli}(\pi_i)$
Latent subgroup index
$(Y_i|Zi = k) = \beta_{0k} + \mathbf{X}_i^T\boldsymbol{\beta}+e$
$e \sim N(0,\sigma^2)$
Outcome
B
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9
$G_{\mathcal{A}_1}$
$G_{\mathcal{A}_2}$
1 -
15 -
30 -
1000 -
Genes
1
600
Samples
4
2
0
−2
−4

$G_{\mathcal{A}_1}$	$\mathcal{O}=1$	$\mathcal{O}=2$	$\mathcal{O}=3$
$G_{1\sim5}$	N(1,1)	N(0,1)	N(0,1)
$G_{6\sim10}$	N(0,1)	N(1,1)	N(0,1)
$G_{10\sim15}$	N(0,1)	N(0,1)	N(1,1)

$G_{\mathcal{A}_2}$	$\mathcal{I}=1$	$\mathcal{I}=2$	$\mathcal{I}=3$
$G_{15\sim20}$	N(1,1)	N(0,1)	N(0,1)
$G_{21\sim25}$	N(0,1)	N(1,1)	N(0,1)
$G_{26\sim30}$	N(0,1)	N(0,1)	N(1,1)

Figure S1: (A) Data generation scheme. $\mathcal{O} = \{1, 2, 3\}$ denotes three clusters defined by genes set $G_{\mathcal{A}_1}$, $\mathcal{A}_1 = \{1, \ldots, 15\}$, and $\mathcal{I} = \{1, 2, 3\}$ denotes another three independent clusters defined by $G_{\mathcal{A}_2}$, $\mathcal{A}_2 = \{16, \ldots, 30\}$. Expression of genes in $G_{\mathcal{A}_1}$ and $G_{\mathcal{A}_2}$ are generated from the distributions listed on the above table. For subject i, only $G_{\mathcal{A}_1}$ have real signals effecting Z_i, which is drawn from a Multinomial distribution with probability $\boldsymbol{\pi}_i = \{\pi_{i1}, \pi_{i2}, 1 - \pi_{i1} - \pi_{i2}\}$. Baseline variables X_1 and X_2 are generated from $N(1, 1)$ and $N(1, 2)$ respectively. Given X_i, G_i and Z_i, the outcome Y_i is generated finally. (B) Heatmap of the expression of 1000 genes across samples. A total of nine subgroups C_1, ..., C_9 are jointly defined by genes sets $G_{\mathcal{A}_1}$ and $G_{\mathcal{A}_2}$.

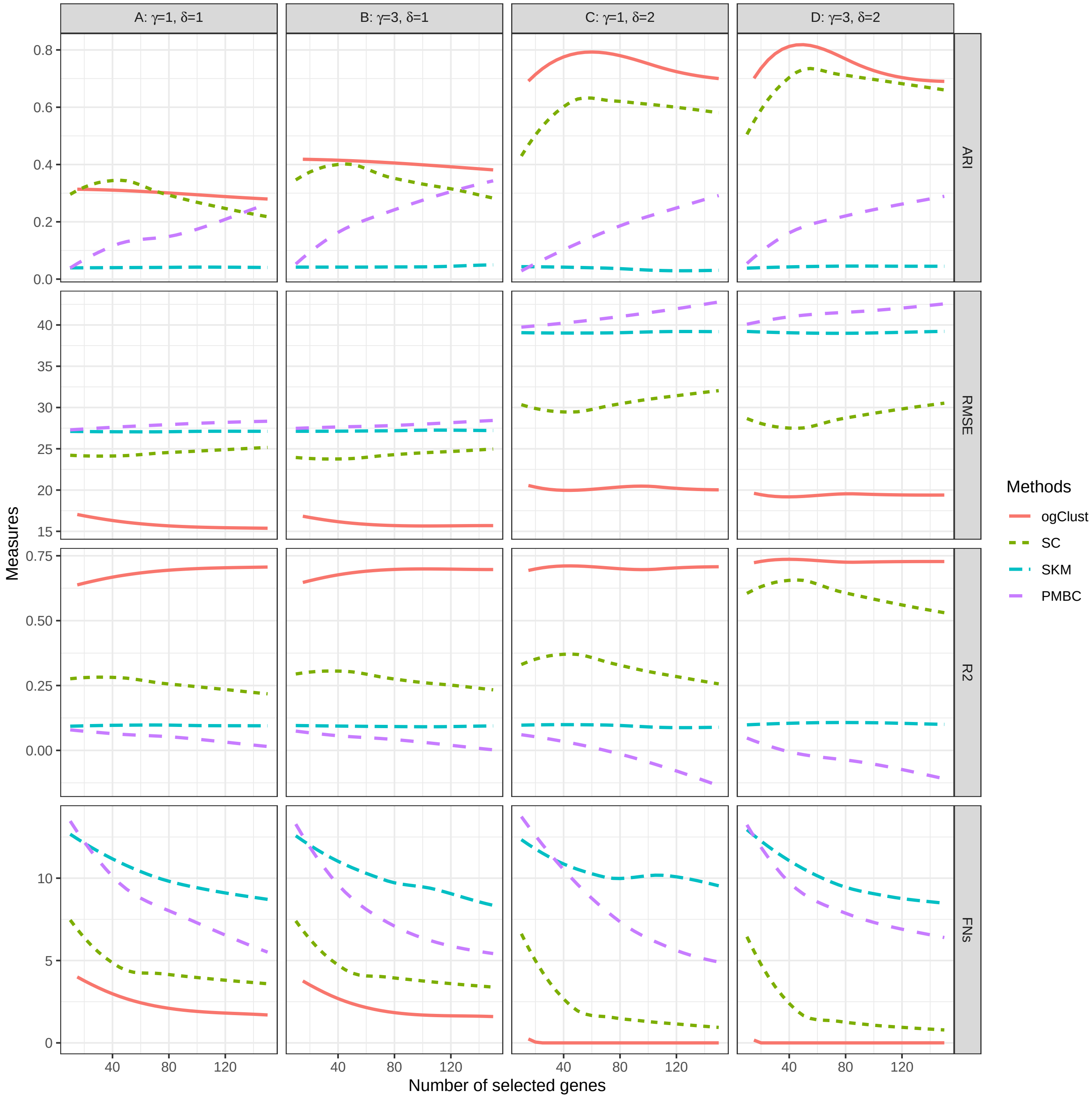


Figure S2: Comparison of ogClust and SC,SKM and PMBC under four simulation settings with survival outcome. We compare RMSE, R^2, ARI and FNs (y-axis) vs number of genes selected in each setting (x-axis).

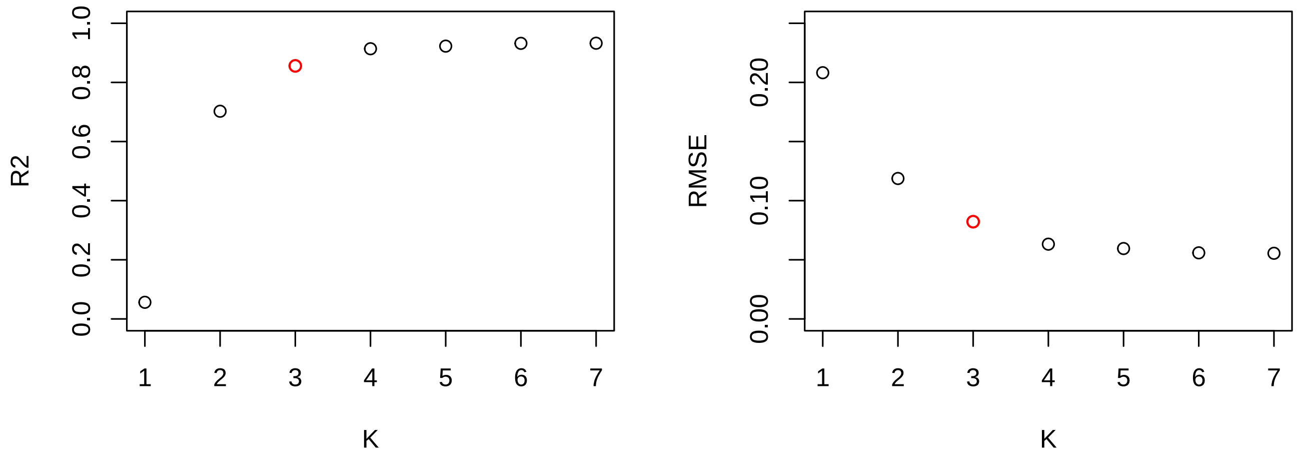


Figure S3: Plot of (A) R^2 and (B) RMSE against the number of clusters.