Simulation-based Evaluation for Discovering Significant Pathways Associated with Survival Time

Jinheum Kim[1], Seungyeoun Lee[2], Sunho Lee[2]

[1]Department of Applied Statistics, University of Suwon
[2]Department of Mathematics & Statistics, Sejong University

May 27, 2011

Outline

- Gene set analysis: review
- Simulation experiments for comparison of the tests
- An application to ovarian cancer data
- Concluding remarks

Gene set analysis

- What is gene-set analysis?
  - A microarray data analysis which uses **existing knowledge of biological pathways** or sets of individual genes that are linked via related biological functions
- Objective: To discover gene sets the expression of which is associated with a phenotype of interest
- Advantage over single-gene analysis
  - Be more useful in interpreting the results to gain insights into biological mechanisms
  - Reduce the multiple testing problems because there are typically much fewer pathways than genes

Gene Set Enrichment Analysis

- Proposed by Mootha et al. (2003), Subramanian et al. (2005)
- S : a priori defined set of genes in a total of M genes on a microarray dataset
  - n_s : the number of genes in the set S
- Want to test that the expression pattern of S is associated with a phenotype of interest

Procedure

- Compute a correlation or an association measure between each of the M genes with a phenotype, say $r_j, j = 1, \ldots, M$
- Order the M genes by the values of r_j's from the maximum to minimum
- Compute the Enrichment Score (ES) as follows:
  - Start with ES=0
  - Sum up from the top rank ($j = 1$) to the last rank ($j = M$), **increasing by**

$$\frac{|r_j|^p}{\sum_{k \in S} |r_k|^p},$$

    if the jth gene belongs to the gene set S, and **decreasing by**

$$\frac{1}{M - n_S},$$

    otherwise
- Take the absolute value of **the maximum deviation from zero of the ES values** among the M genes as the test statistic for the gene set S

Procedure (continued)

- Randomly **assign the original phenotype variable to samples** and follow the previous steps. Repeat these procedures for several times of permutations
- Obtain the significance level by comparing the observed value of the test statistic and its permutation distribution obtained
- Remarks
  - GSEA with $p = 0$ is **a normalized Kolmogorov-Smirnov statistic** (Mootha et al., 2003)
  - Weak point: GSEA with $p = 0$ may have high scores of ES for set clustered near the middle of the ranked list (Subramanian et al., 2005)

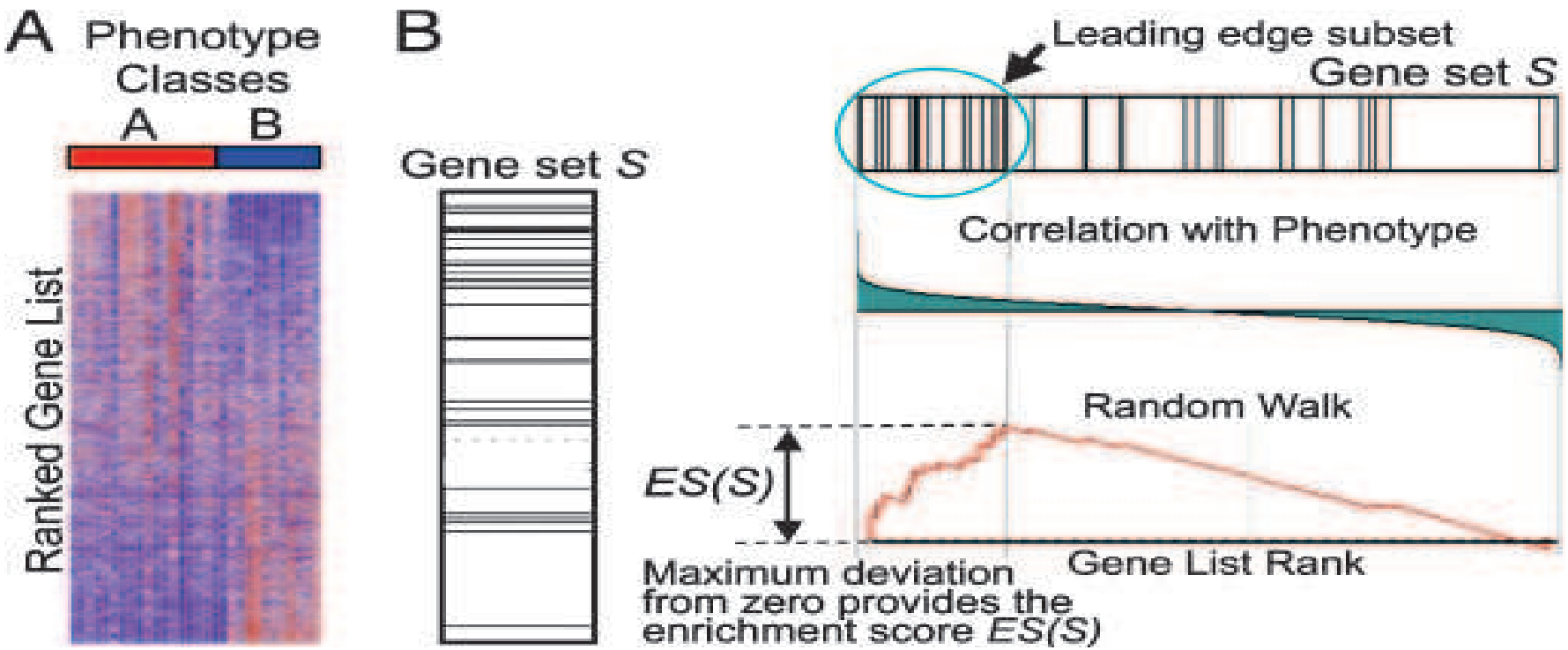


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the ''gene tags,'' i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

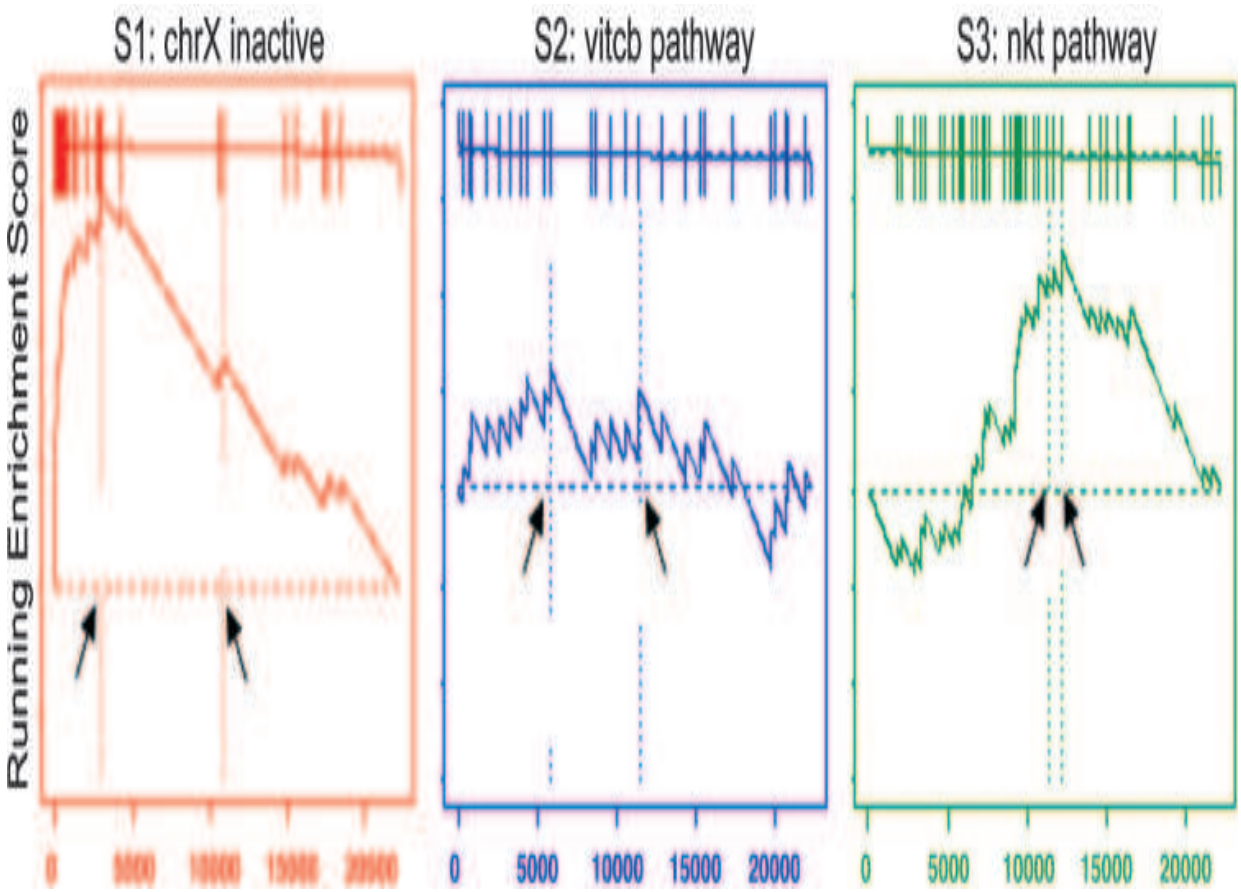


Fig. 2. Original (4) enrichment score behavior. The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male/female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. Table 1 compares the nominal *P* values for S1, S2, and S3 by using the original and new method. The new method reduces the significance of sets like S3.

Table 1. *P* value comparison of gene sets by using original and new methods

Gene set	Original method nominal *P* value	New method nominal *P* value
S1: chrX inactive	0.007	<0.001
S2: vitcb pathway	0.51	0.38
S3: nkt pathway	0.023	0.54

Significance analysis of microarray for gene-set

- Want to test that the mean vectors of expressions of genes in a gene set do not differ by the phenotype (binary)
- BUT, the Hotelling's T^2 for a two-sample mean test cannot be applied when $n_s > N - 2$, where N : total number of samples
- Based on individual t-like statistics from SAM (Tusher et al., 2001)
- SAM-GS is defined by

$$\text{SAM-GS} = \sum_{k=1}^{n_s} \left(\frac{\bar{x}_1(k) - \bar{x}_2(k)}{s(k) + s_0} \right)^2$$

  - $s(j)$: pooled standard deviation over the two groups of the phenotype
  - s_0 : small positive constant that adjusts for the small variability encountered in microarray data

Global test

- The Global test is **a score test based on random-effect modeling of parameters** corresponding to the coefficients of the individual genes in the pathway.
- Goeman et al. (2004) originally proposed the Global test based on the generalized linear model and then extended this test to the survival time in the Cox proportional hazards model (Goeman et al., 2005)

The model

- Data: $\{(t_i, d_i, x_{ik}, z_{il}), k = 1, \ldots, n_s; l = 1, \ldots, p; i = 1, \ldots, N\}$
  - t_i : observed survival time
  - d_i : censoring indicator, i.e., $d_i = 1$ indicates death and $d_i = 0$ indicates censoring
  - x_{ik} : gene expression measurements
  - z_{il} : covariates
  - Assume $p < N$, but no restriction on n_s
- The Cox's PH model: $h_i(t) = h_0(t)e^{c_i + r_i}$
  - $h_0(t)$: unknown baseline hazard function
  - $c_i = \sum_{l=1}^{p} \gamma_l z_{il}$, $r_i = \sum_{k=1}^{n_s} \beta_k x_{ik}$: linear predictors
- Hypothesis of interest: $H_0 : \beta_1 = \cdots = \beta_{n_s} = 0$
  - When n_s : small, use classical tests
  - How can we test for general n_s? Assume $\beta_k, k = 1, \ldots, n_s$, are random and independent with mean 0 and common variance τ^2
  - Simply reduced to $H_0 : \tau^2 = 0$

Derivation for the test: 1st stage

- Assume $h_0(t)$ and γ_l's are known, i.e., $H_0(t) = \int_0^t h_0(s)ds$ and c_i are known
- Use the log marginal likelihood because r_i are not observed, i.e.,

$$L(\tau^2) = \log\{E_r(\mathcal{L})\},$$

where $\mathcal{L} = e^l$ and

$$l = \log\{\prod_{i=1}^{N} h_i(t_i)^{d_i} e^{-H_i(t_i)}\} = \sum_{i=1}^{N} d_i\{\log h_0(t_i) + c_i + r_i\} - H_0(t_i)e^{c_i+r_i}$$

- Make a score test using

$$\frac{\partial L(0)}{\partial \tau^2} = \frac{1}{2}\{(d-u)'R(d-u) - \text{trace}(RU)\}$$

  - $d = (d_1, \ldots, d_N)'$; $u = (u_1, \ldots, u_N)'$, $u_i = H_0(t_i)e^{c_i}$; $U = \text{diag}(u_i)$
  - $R = XX'$, $X = (x_{ik}) : N \times n_s$

Derivation for the test: 2nd stage

- We shall plug in $\hat{H}_0(t_i)$, but still assume that c_i's are known
- Use Breslow estimator

$$\hat{H}_0(t_i) = \sum_{t_j \le t_i} \frac{d_j}{\sum_{t_k \ge t_j} e^{c_k}}$$

- Propose a test statistic

$$T = (d - \hat{u})R(d - \hat{u}) - \text{trace}(R\hat{U})$$

  - $\hat{U} = \text{diag}(\hat{u}_i)$, $\hat{u}_i = \hat{H}_0(t_i)e^{c_i}$
  - $\mathrm{E}(T) =?$ $\mathrm{Var}(T) =?$ Quite technical!
  - Instead use a simpler form $T_0 = (d - \hat{u})R(d - \hat{u})$ from the relation

$$Q = \frac{T - \hat{\mathrm{E}}(T)}{\hat{\mathrm{Var}}(T)} = \frac{T_0 - \hat{\mathrm{E}}(T_0)}{\hat{\mathrm{Var}}(T_0)}$$

  - Why is Q the global test?

Derivation for the test: 3rd stage

- Replace γ_l by their MLEs, but still a valid score test is possible!
- Approximate $T_0, \hat{\mathrm{E}}(T_0), \hat{\mathrm{Var}}(T_0)$ and plug in them into Q, correspondingly
- Two ways to calculate p-value of Q
  - Use a normal approximation from martingale CLT
  - Use the Permutation test for small samples
  - Permute martingale residuals, i.e., redistributed the vectors of gene expression measurements over the individuals while keeping the relationship between the fixed covariates and survival the same

Wald-type test

- Proposed by Adewale et al. (2008) as a sort of SAM-GS
- For the kth gene ($k = 1, \ldots, n_s$), assume the Cox model with covariates like

$$h_i(t) = h_0(t) \exp(\beta_k x_{ik} + \sum_{l=1}^{p} \gamma_l z_{il})$$

- Combine component-wise test statistic for testing the significance of a pathway, i.e.,

$$W = \sum_{k=1}^{n_s} r_k^2$$

  - $r_k = b_k / s_k, k = 1, \ldots, n_s$: t-test statistic
  - b_k : parameter estimate of the log-hazard ratio, β_k, associated with the expression of the kth gene on the survival time
  - s_k : standard error of b_k

Objective

- To compare the statistical performance of the four gene-set analysis tests, namely two different GSEA tests (**GSEA1** and **GSEA2**), Global test (**GT**) and Wald-type test (**WT**) for assessing differential expression associated with survival time phenotype based on the simulation dataset and a real dataset of ovarian cancer patients
- Remarks
  - For GSEA tests, define $r_j, j = 1, \ldots, M$, as the t-statistics in the Cox model, i.e., $r_j = b_j/s_j$
  - GSEA1=GSEA with $p = 0$; GESA2=GSEA with $p = 1$
- *cf* Dinu et al. (2007), Liu et al. (2007): binary phenotype

Data generation procedure

- Generate **the gene expression variables** from a multivariate normal, MVN$(0, \Sigma)$
- Generate **regression coefficients**, β_j's, from either an uniform distribution or a normal distribution, which represent the association between the survival time and gene expressions
- Construct **a survival time** from a Cox model with gene expression variables and a specified baseline hazard function
- Generate **a censoring time** from an exponential distribution with a parameter λ

Design parameters

- M : total number of genes (=200)
- n_s : size of gene set (=20,50)
- $h_0(t)$: baseline hazard function (=0.005)
- p_s : proportion of significant genes in each gene-set (=0.1,0.3,0.5)
- Number of permutation=1000
- For checking the **size** of the tests,
  - $\Sigma = 0.2 I_M$, where I_M : identity matrix of the order of M
  - $\beta_j = 0, j = 1, \ldots, M$
  - N : sample size (=50,80)
  - c_p : fraction of censoring (=0,0.1,0.3,0.5)
  - Number of replication=500

Design parameters (continued)

- For the **power** calculation,
- Three scenarios for Σ
  - Case (I): $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 0.2, i = 1, \ldots, M; \sigma_{ij} = 0, i \neq j, i, j = 1, \ldots, M$
  - Case (II): $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 0.2$ if $i = 1, \ldots, M; \sigma_{ij} = 0.02$ if $i \neq j, i, j = 1, \ldots, [n_s \times p_s]; \sigma_{ij} = 0.02$, if $i \neq j, i, j = [n_s \times p_s] + 1, \ldots, n_s$, $\sigma_{ij} = 0$, otherwise
  - Case (III): $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 0.2$ if $i = 1, \ldots, M; \sigma_{ij} = 0.02$, if $i \neq j, i, j = 1, \ldots, [n_s \times p_s], \sigma_{ij} = 0$, otherwise
- Three different ways for generating $\beta_j, j = 1, \ldots, [n_s \times p_s]$, along with $\beta_j = 0, j = [n_s \times p_s] + 1, \ldots, M$
  - Case (A): $U(0.2, 0.6)$
  - Case (B): $U(-0.6, -0.2)$ and $U(0.2, 0.6)$ equally
  - Case (C): $N(0, 0.5^2)$
- N : sample size(=80)
- c_p : fraction of censoring(=0,0.1,0.3)
- Number of replication=200

Table 1. The estimated size of the four tests based on 500 iterations

N	n_s	c_p	GSEA1	GSEA2	GT	WT
80	20	0	0.058	0.052	0.050	0.050
		10	0.060	0.052	0.058	0.058
		30	0.034	0.054	0.054	0.066
		50	0.062	0.042	0.040	0.032
	50	0	0.050	0.048	0.044	0.042
		10	0.050	0.056	0.058	0.058
		30	0.054	0.040	0.054	0.048
		50	0.068	0.060	0.058	0.058
50	20	0	0.054	0.042	0.036	0.032
		10	0.062	0.056	0.046	0.046
		30	0.056	0.036	0.044	0.038
		50	0.046	0.042	0.038	0.046
	50	0	0.070	0.046	0.052	0.044
		10	0.044	0.048	0.042	0.046
		30	0.050	0.038	0.060	0.058
		50	0.046	0.046	0.058	0.040

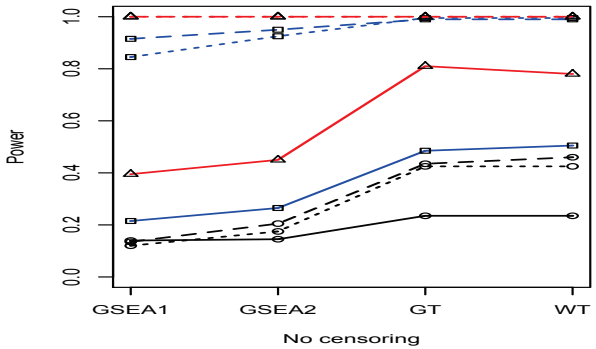
Note: N=Sample size; n_s=the size of gene set; c_p=censoring proportion. GSEA1=GSEA with the equal weight; GSEA2=weighted GSEA depending on the correlation between genes and the phenotype; GT=Global test; WT=Wald-type test

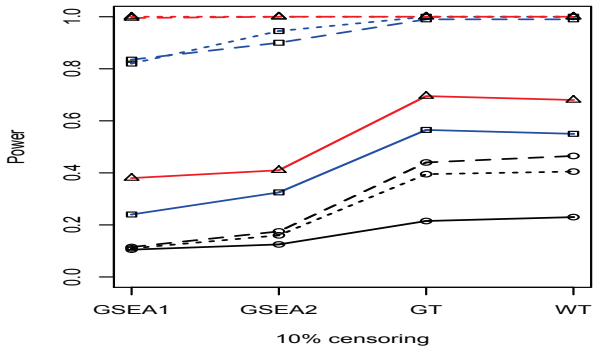
Table 2. The estimated power of the four tests for n_s=50 based on 200 replications

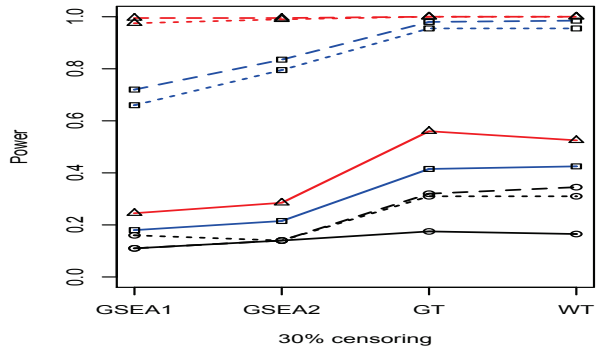
c_p	p_s	Case (I)				Case (II)				Case (III)			
		GSEA1	GSEA2	GT	WT	GSEA1	GSEA2	GT	WT	GSEA1	GSEA2	GT	WT
Case (A): $\beta \sim$U(0.2,0.6)													
0	0.1	0.140	0.145	0.235	0.235	0.135	0.205	0.435	0.460	0.120	0.175	0.425	0.425
	0.3	0.215	0.265	0.485	0.505	0.915	0.950	0.990	0.990	0.845	0.925	0.995	0.995
	0.5	0.395	0.450	0.810	0.780	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0.1	0.105	0.125	0.215	0.230	0.115	0.175	0.440	0.465	0.110	0.160	0.395	0.405
	0.3	0.240	0.325	0.565	0.550	0.835	0.900	0.990	0.990	0.820	0.945	1.000	1.000
	0.5	0.380	0.410	0.695	0.680	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
30	0.1	0.110	0.140	0.175	0.165	0.110	0.140	0.320	0.345	0.160	0.140	0.310	0.310
	0.3	0.180	0.215	0.415	0.425	0.720	0.835	0.980	0.985	0.660	0.795	0.955	0.955
	0.5	0.245	0.285	0.560	0.525	0.995	0.995	1.000	1.000	0.975	0.990	1.000	1.000
Case (B): $\beta \sim$U(0.2,0.6) or $\beta \sim$U(-0.6,-0.2)													
0	0.1	0.115	0.160	0.215	0.235	0.145	0.150	0.230	0.215	0.075	0.120	0.220	0.190
	0.3	0.305	0.370	0.615	0.615	0.210	0.220	0.570	0.560	0.240	0.250	0.535	0.495
	0.5	0.445	0.490	0.805	0.795	0.335	0.380	0.705	0.675	0.275	0.385	0.655	0.595
10	0.1	0.125	0.195	0.255	0.265	0.105	0.150	0.230	0.215	0.070	0.125	0.155	0.180
	0.3	0.215	0.285	0.515	0.490	0.180	0.230	0.470	0.460	0.225	0.280	0.495	0.470
	0.5	0.400	0.425	0.750	0.730	0.350	0.430	0.655	0.565	0.340	0.370	0.675	0.615
30	0.1	0.085	0.065	0.190	0.180	0.075	0.095	0.155	0.165	0.095	0.110	0.190	0.180
	0.3	0.200	0.240	0.440	0.455	0.145	0.115	0.370	0.345	0.200	0.230	0.390	0.370
	0.5	0.235	0.265	0.550	0.520	0.235	0.255	0.475	0.470	0.230	0.285	0.450	0.430
Case (C): $\beta \sim N(0,0.5^2)$													
0	0.1	0.085	0.135	0.345	0.380	0.145	0.140	0.315	0.345	0.110	0.120	0.340	0.340
	0.3	0.275	0.385	0.745	0.785	0.225	0.365	0.730	0.725	0.235	0.375	0.780	0.770
	0.5	0.400	0.515	0.875	0.860	0.455	0.590	0.885	0.865	0.430	0.550	0.870	0.890
10	0.1	0.105	0.095	0.220	0.255	0.115	0.160	0.315	0.350	0.105	0.110	0.320	0.350
	0.3	0.230	0.300	0.700	0.680	0.235	0.345	0.625	0.620	0.250	0.360	0.685	0.680
	0.5	0.345	0.465	0.865	0.860	0.415	0.510	0.845	0.835	0.335	0.410	0.785	0.760
30	0.1	0.125	0.090	0.235	0.215	0.075	0.130	0.260	0.245	0.090	0.110	0.190	0.200
	0.3	0.170	0.215	0.495	0.515	0.220	0.285	0.555	0.600	0.230	0.280	0.570	0.570
	0.5	0.255	0.365	0.720	0.690	0.315	0.410	0.645	0.660	0.375	0.450	0.725	0.715

Note: n_s=the size of gene set; c_p=censoring proportion; p_s=the proportion of significant genes in a gene set. GSEA1=GSEA with the equal weight; GSEA2=weighted GSEA depending on the correlation between genes and the phenotype; GT=Global test; WT=Wald-type test. For Case (I), all genes are independent; for Case (II), there is within-correlation among significant genes and among non-significant genes, separately; and for Case (III), there is within-correlation among significant genes

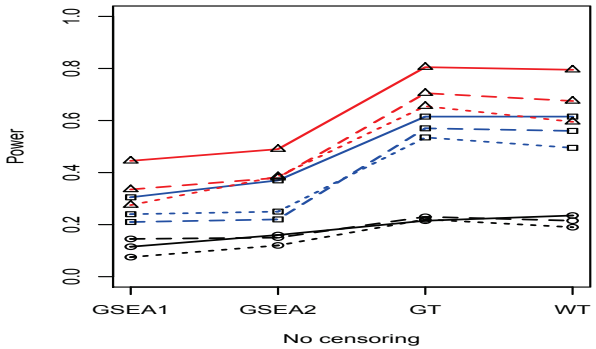
Case (A): β~U(0.2,0.6)

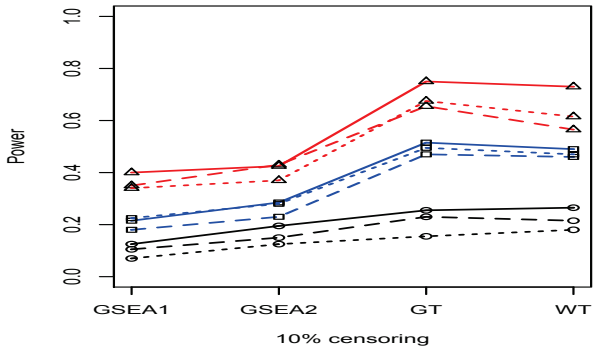


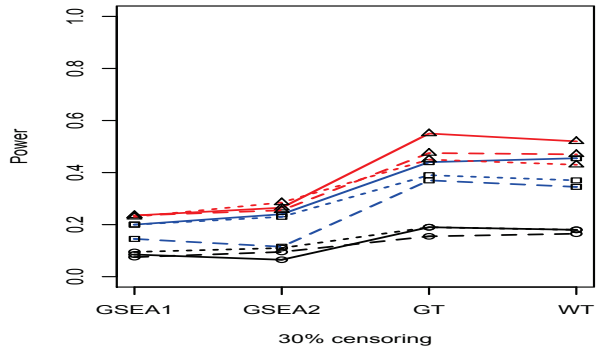




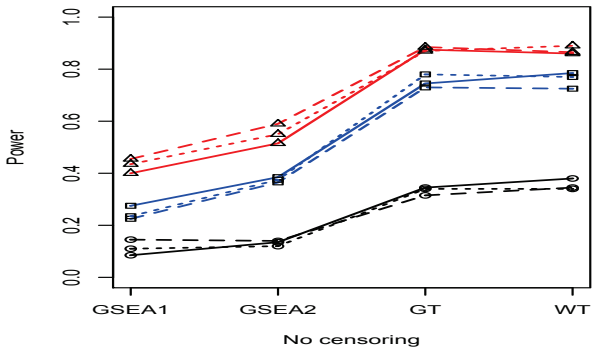
Case (B): β~U(-0.6,-0.2) or U(0.2,0.6)

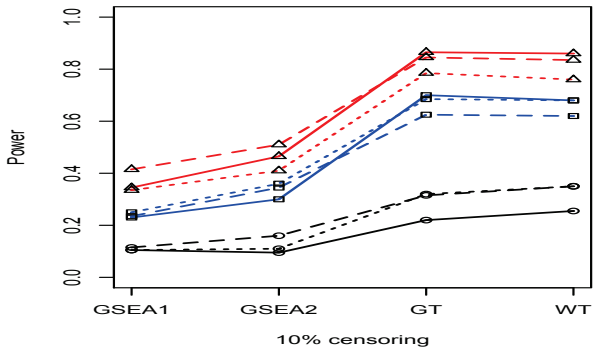


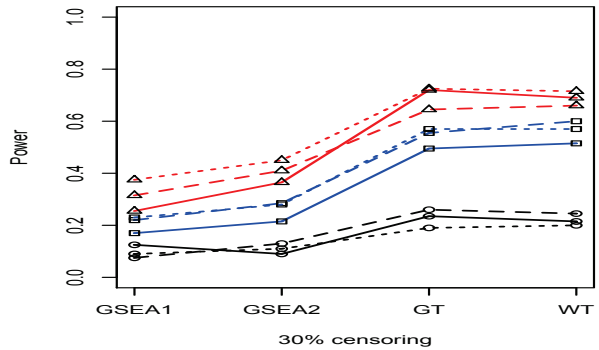


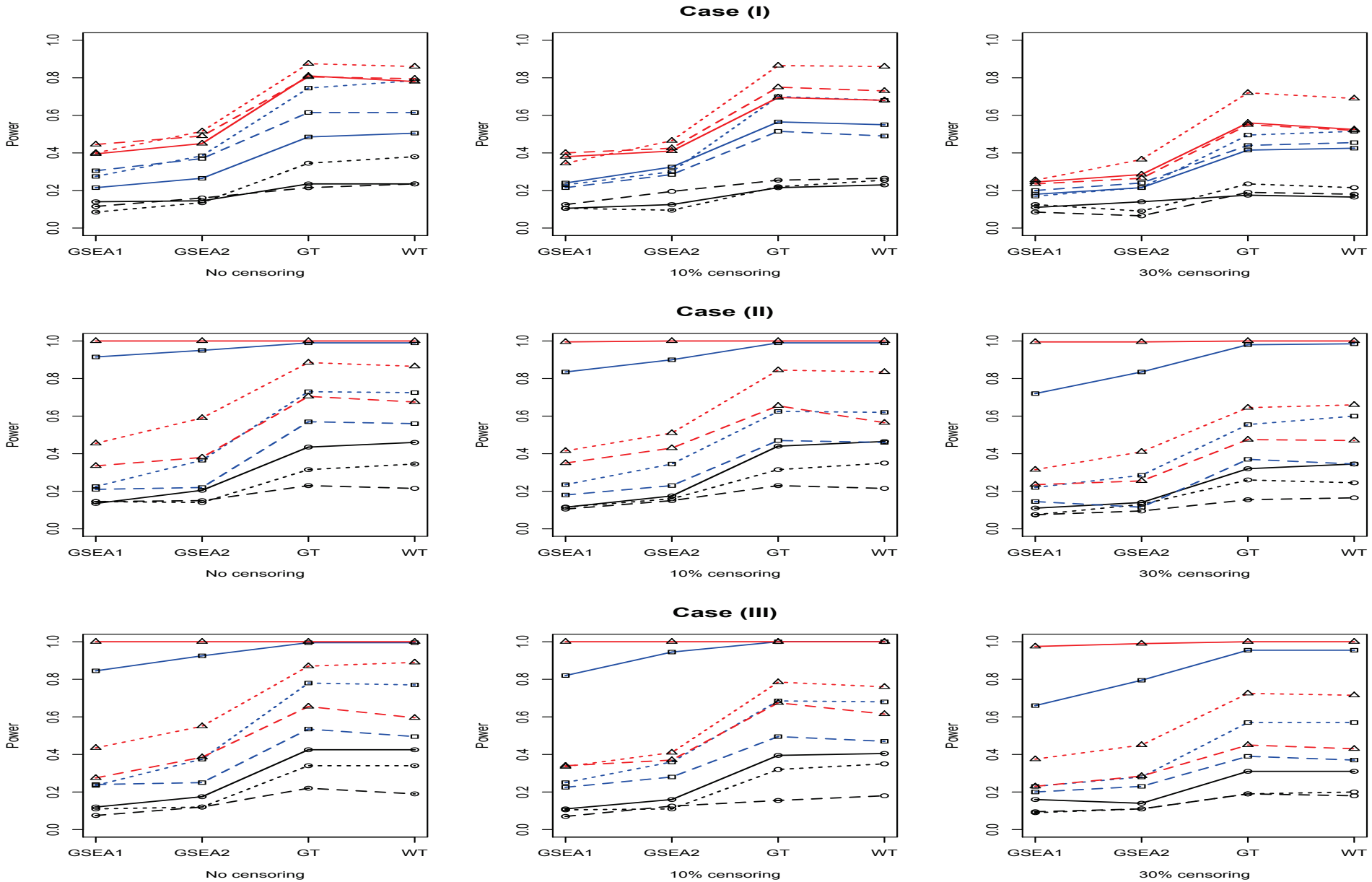


Case (C): β~N(0, 0.5^2)








Case (I)
Case (II)
Case (III)
Power
0.0
0.2
0.4
0.6
0.8
1.0
GSEA1
GSEA2
GT
WT
No censoring
10% censoring
30% censoring

Ovarian cancer data

- Source: 119 ovarian cancer patients who were obtained at the initial cytoreductive surgery from patients treated at Duke University Medical Center and H. Lee Moffitt Cancer Center and Research Institute (Dressman et al., 2007)
- Clinicopathologic variables: age, stage, grade, surgical debulking, chemotherapy and serum CA-125
- 22115 gene expression levels
- 204 pathways were identified by KEGG
  - The smallest pathway consists of 5 genes while the largest one includes 474 genes
  - Average number of genes across 204 pathways is about 80
- Compare the performance of the four tests and also investigate whether the four tests also confirm the profiles and pathways identified by Dressman et al. (2007) and Crijns et al. (2009)

Ovarian cancer data (continued)

- **23 pathways** which are significant at the nominal 0.01 level by the permutation test of at least one of the four tests
- The underlined pathways are those whose profiles were significantly identified by Dressman et al. (2007) and Crijns et al. (2009)
- Both GT and WT tests detect pathways more powerfully than GSEA1 and GSEA2
  - GT and WT detect 15 and 13 significant pathways among 204 pathways, respectively, while only GSEA2 detects only one pathway, with $p < 0.01$
  - Under FDR with $q < 0.1$, **GT** identifies 3 significant pathways, **Pentose phosphate** with 39 genes, **Histidine metabolism** with 54 genes, and **Jak-STAT signaling** with 240 genes while **WT** detects only one pathway, **Histidine metabolism**
  - Among those, **Jak-STAT signaling** pathway was commonly identified by Crijn et al. (2009)

Table 4.

Pathway name	Gene set size	p-value				q-value			
		GSEA		GT	WT	GSEA		GT	WT
		p=0	p=1			p=0	p=1		
Histidine metabolism	54	0.095	0.012	0.000	0.000	0.925	0.782	0.000	0.000
Pentose phosphate pathway	39	0.012	0.004	0.000	0.001	0.925	0.782	0.000	0.068
One carbon pool by folate	28	0.072	0.029	0.016	0.001	0.925	0.827	0.126	0.068
Tryptophan metabolism	86	0.040	0.021	0.003	0.002	0.925	0.782	0.077	0.082
DNA replication	52	0.018	0.012	0.009	0.002	0.925	0.782	0.114	0.082
Folate biosynthesis	53	0.216	0.048	0.012	0.003	0.925	0.827	0.114	0.102
Purine metabolism	201	0.083	0.023	0.003	0.004	0.925	0.782	0.077	0.102
Nucleotide excision repair	56	0.038	0.062	0.011	0.004	0.925	0.827	0.114	0.102
Type II diabetes mellitus	74	0.043	0.139	0.013	0.006	0.925	0.852	0.114	0.136
Mismatch repair	35	0.208	0.112	0.013	0.008	0.925	0.852	0.114	0.140
N-Glycan biosynthesis	54	0.131	0.070	0.025	0.009	0.925	0.827	0.170	0.140
Aminophosphonate metabolism	21	0.296	0.034	0.001	0.010	0.925	0.827	0.051	0.140
Renal cell carcinoma	139	0.242	0.188	0.035	0.010	0.925	0.852	0.204	0.140
Colorectal cancer	165	0.153	0.148	0.003	0.011	0.925	0.852	0.077	0.140
Lysine degradation	56	0.122	0.242	0.010	0.016	0.925	0.879	0.114	0.165
Parkinson's disease	34	0.203	0.069	0.180	0.016	0.925	0.827	0.317	0.165
Leukocyte transendothelial migration	197	0.350	0.151	0.006	0.017	0.939	0.852	0.102	0.165
Phenylalanine, tyrosine and tryptophan biosynthesis	13	0.502	0.159	0.006	0.018	0.959	0.852	0.102	0.167
Starch and sucrose metabolism	72	0.405	0.150	0.006	0.027	0.959	0.852	0.102	0.197
Glycerophospholipid metabolism	89	0.524	0.235	0.002	0.029	0.960	0.879	0.077	0.197
Androgen and estrogen metabolism	53	0.419	0.178	0.005	0.031	0.959	0.852	0.102	0.198
Styrene degradation	5	0.031	0.055	0.008	0.031	0.925	0.827	0.114	0.192
Jak-STAT signaling pathway	240	0.275	0.644	0.000	0.038	0.925	0.908	0.000	0.199

Concluding remarks

- Since many gene-set analysis methods have been proposed, these methods were compared based on the simulation results and a real example analysis. However, most of studies have been **dealt with a binary phenotype** like the presence or absence of disease, or treatment vs. control
- We focused on the **survival time phenotype** and compare four different gene-set analysis tests
- For the GSEA tests, we replaced the correlation coefficient by the **regression coefficient** from the Cox model. However, the performance of the GSEA tests are not satisfactory except for a few cases since the GSEA tests are nonparametric approach with using the rank-based statistic instead of using the value of the regression coefficient.

Concluding remarks (continued)

- The Global test (GT) assumes **a random-effect model** for the parameters corresponding to the coefficients of the individual genes in the pathway. Therefore, this test does **not depend on the number of genes** in a given set of genes and works under the assumption of a random-effect model with the common variance. The Wald-type test (WT) takes a sum of squares of the Wald statistic for individual genes constituting the pathway from the framework of regression model. Both GT and WT are based on the parametric approach
- From the simulation results, **both GT and WT are more powerful than both GSEA1 and GSEA2**

Concluding remarks (continued)

- Additionally, the power trend of the four tests is substantially affected by the correlation structure of genes and the association between the survival time and genes
  - There might be a **synergistic effect** in the power of detecting significant genes **when the survival is positively associated with genes and the genes are correlated**
  - **When survival is associated with genes in the two opposite directions**, the power is **higher when the genes are independent than when they are correlated**
  - There is **no substantial difference** in power **when survival is randomly associated with genes**
  - **When the genes are independent**, the power of the four tests has **no significant difference irrespective of the association of genes with the survival**
  - **The correlation among genes** has an important advantage **enabling detection** of significant gene-sets **when survival is positively associated with genes**

Thank you!