Supplementary Material

“Integrative Gene Set Enrichment Analysis Utilizing Isoform-Specific Expression”

This document provides a list of important notation, additional simulation results and results from our breast cancer example mentioned in the main body of the paper.

S1 Notation

K	Number of RNA-seq studies to be combined
S_k	Number of samples in study k
P	Number of pathways in the pathway database to be tested
G	Total number of genes that appear in at least one component study
Z_{gp}	Indicator variable of gene g's membership in pathway p: $Z_{gp} = 1$ if gene g is in pathway p; $Z_{gp} = 0$, otherwise
I_g	Number of isoforms within gene g
T_{kg}	Indicator variable of gene g's presence in study k: $T_{kg} = 1$ if gene g is present in study k; $T_{kg} = 0$, otherwise
X_{ksgi}	Expression level of isoform i of gene g for sample s in study k
Y_{ks}	Phenotype of sample s in study k
$\mathbf{U}_{kg}$	Score statistic of gene g in study k
$\mathbf{V}_{kg}$	Estimated covariance matrix of $\mathbf{U}_{kg}$
Q_g	Gene-level quadratic statistic of gene g
$p(Q_g)$	P-value measuring significance of gene $g's$ association with Y
$\boldsymbol{\beta}_{kg}$	Vector of isoform effects of gene g in study k
$\boldsymbol{\mu}_g$	Overall isoform effects of gene g across studies
$\boldsymbol{\xi}_{kg}$	Study-specific deviations of gene g from $\boldsymbol{\mu}_g$ in study k
$\boldsymbol{\Sigma}_g$	Covariance matrix of $\boldsymbol{\xi}_{kg}$
ω_p	Enrichment score of pathway p
c_p	Size of pathway p
d_p	Number of genes in the genome but not in pathway p
ω_p^*	Size-adjusted enrichment score of pathway p
$p(\omega_p^*)$	P-value measuring significance of pathway $p's$ enrichment
α	Strength of pathway enrichment signal
λ	Gene sampling rate
v	Mean isoform effect

S2 The approximate limiting distribution of the RE test statistic

As mentioned in Section 2.1, the limiting distribution of the RE test statistic (5) is approximately chi-square with $I_g + 1$ degrees of freedom. Below we show empirical evidence that the approximation is generally adequate.

We consider both discrete and continuous cases, and use the same methods described in Section 4 of the main paper to generate data for null-case genes. We set the number of studies $K = \{5, 10\}$ (each study with 40 samples) and $\lambda = 1$, and consider genes with $I_g = \{2, 4, 6, 8, 10\}$ isoforms. We generate 1000 genes for each combination of the parameters under $H_0 : \mu_g = 0$ and $\Sigma_g = 0$ and calculate the RE statistics Q_g's. Figure S1 compares the empirical cumulative distribution function (CDF) of Q_g (the dotted curve) with $\chi^2_{I_g+1}$ (the solid curve) under each setting. Clearly, the solid and dotted curves are grouped by the number of degrees of freedom. It seems that the approximation is reasonably good. Similar observations can be made for other settings in the paper as well (results omitted for concision).

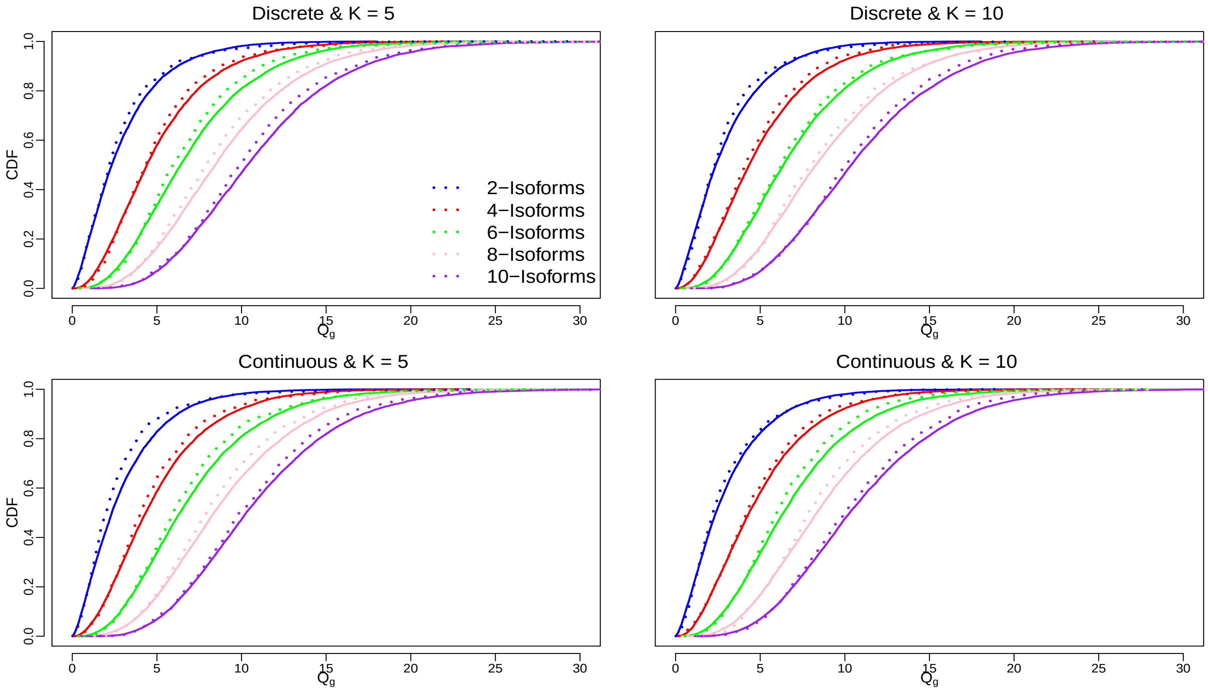

Figure S1: Comparison between the empirical CDF of Q_g (dotted curve) and the limiting chi-square distribution (solid curve) under various settings

S3 Comparison of Algorithms 1 and 2 in estimating stage I P-values

We compare the performance of Algorithms 1 and 2 in estimating P-values of the first-stage analysis in our proposed procedures (i.e., testing $H_0 : \mu_g = 0$ under the FE model and testing $H_0 : \mu_g = 0$ and $\Sigma_g = 0$ under the RE model) for both discrete and continuous cases, by setting $K = 6$, $\alpha = 0.3$, $\lambda = 0.7$ and $v = 0.75$. Figure S2 shows histograms of estimated P-values based on Algorithms 1 and 2 for data generated from the FE model in the discrete case, where the left four histograms are for the FE algorithms, the right four for the RE algorithms, the top four for isoform-active genes, and the bottom four for non isoform-active (i.e. silent) genes. In the top panel, each black bar stands for genes of which P-values are smaller than the cutoff 0.05; and the proportion of significant genes among all isoform-active genes is also reported for each algorithm. From these proportion values for isoform-active genes, we find that FE Algorithm 2 is the best. As to silent genes, the histograms of FE Algorithm1, FE Algorithm 2 and RE Algorithm 2 are pretty close to the uniform distribution $\mathrm{unif}\,[0, 1]$. For data generated from the RE model in the discrete case, Figure S3 reports that RE Algorithm 2 is the best, in terms of the proportion of significant genes and the distribution of P-values for silent genes. For the continuous case, Figure S4 shows FE Algorithm 2 is the best for data generated from the FE model, and Figure S5 shows RE Algorithm 2 is the best for data generated from the RE model. Thus, in all the cases, Algorithm 2 is better than Algorithm 1 when estimating P-values of the first-stage analysis.

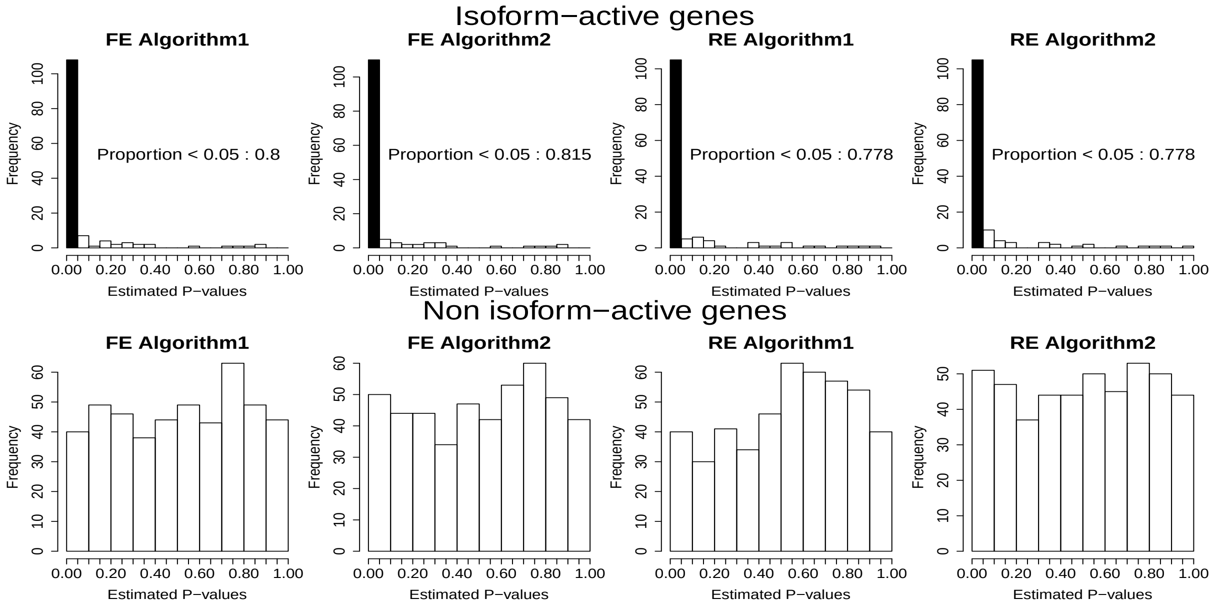

Figure S2: Discrete case: estimated P-values based on Algorithms 1 and 2 for data generated from the FE model

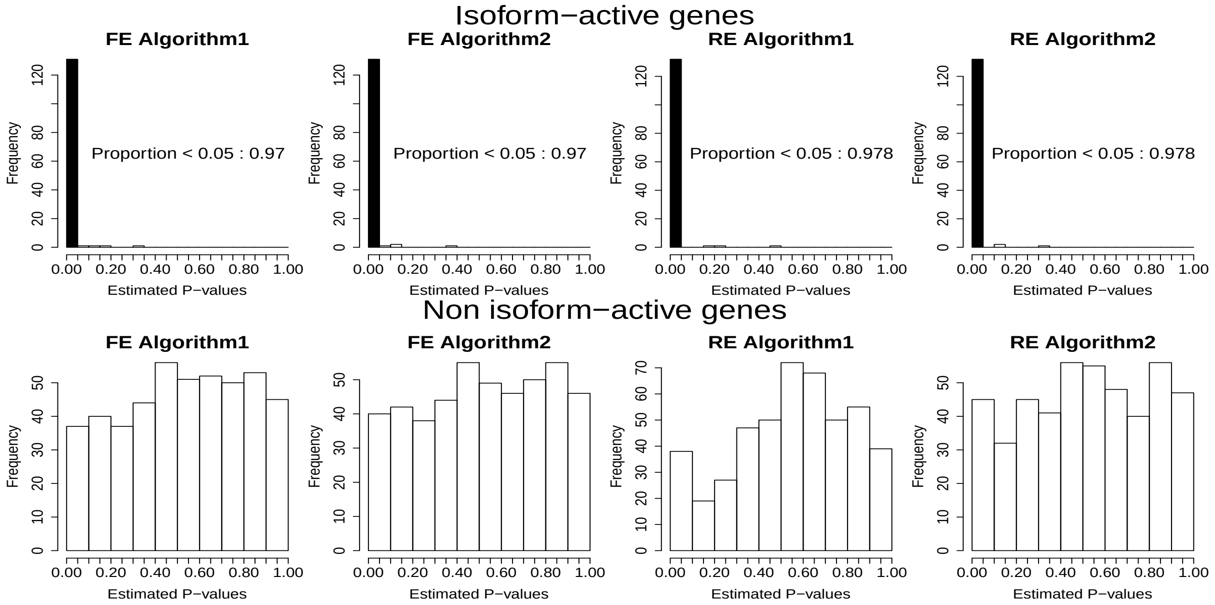

Figure S3: Discrete case: estimated P-values based on Algorithms 1 and 2 for data generated from the RE model

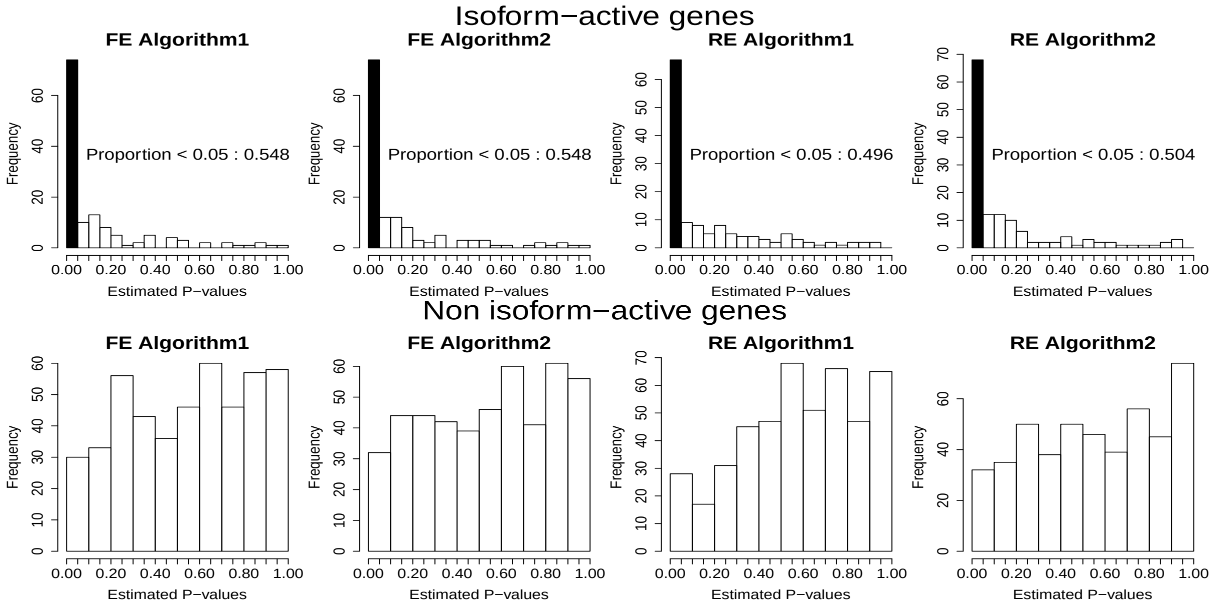

Figure S4: Continuous case: estimated P-values based on Algorithms 1 and 2 for data generated from the FE model

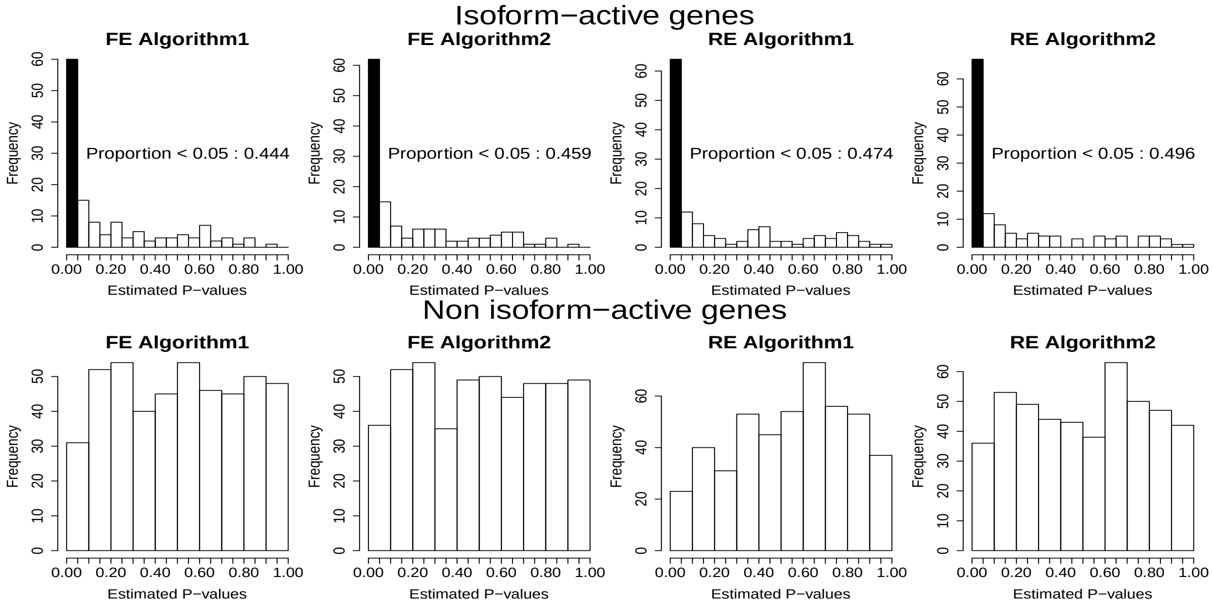

Figure S5: Continuous case: estimated P-values based on Algorithms 1 and 2 for data generated from the RE model

S4 Evaluation of type I error

We examine the type I error of each method at the significance level 0.05 by setting the enrichment signal $\alpha = \alpha_0$ (i.e., the null hypothesis of no enrichment holds for the gene set). We report results in Table S1 for various settings, under each of which 1000 replicate datasets are used. In all the cases, the type I error of the proposed methods, iGSEAi-FE and iGSEAi-RE, although pretty close to 0.05, is consistently smaller than 0.05, which means that the proposed methods are conservative in rejecting the null hypothesis. Among the MAPE methods, the type I error of MAPE_G is close to 0.05; MAPE_P tends to be aggressive for the discrete case, and slightly conservative for the continuous case; and MAPE_I tends to be slightly aggressive for the discrete case but is quite close to 0.05 for the continuous case. Note that the type I error of MAPE_I is somewhere between those of MAPE_G and MAPE_P in all the cases. This might be explained by the fact that MAPE_I is a hybrid of MAPE_G and MAPE_P.

Table SI: Type I error for various methods

Phenotype	Model	λ	v	iGSEAi-FE	iGSEAi-RE	MAPE_G	MAPE_P	MAPE_I
Discrete	FE	0.7	0.5	0.03	0.03	0.05	0.07	0.06
			1	0.03	0.03	0.05	0.08	0.07
		1.0	0.5	0.03	0.04	0.05	0.08	0.05
			1	0.03	0.03	0.04	0.08	0.05
	RE	0.7	0.5	0.03	0.03	0.05	0.07	0.06
			1	0.03	0.03	0.05	0.08	0.07
		1.0	0.5	0.03	0.04	0.05	0.08	0.05
			1	0.03	0.03	0.04	0.08	0.06
Continuous	FE	0.7	0.5	0.04	0.04	0.06	0.03	0.04
			1	0.03	0.04	0.04	0.05	0.05
		1.0	0.5	0.04	0.04	0.05	0.04	0.04
			1	0.03	0.04	0.05	0.04	0.05
	RE	0.7	0.5	0.04	0.03	0.05	0.05	0.05
			1	0.03	0.03	0.05	0.05	0.05
		1.0	0.5	0.04	0.04	0.05	0.05	0.05
			1	0.03	0.03	0.05	0.05	0.05

S5 Control of false discovery rate (FDR)

We evaluate the performance of various methods in FDR control by comparing the empirical FDR versus nominal level (i.e., the Q-value cutoff δ) when testing multiple pathways. Table SII shows results (averaged over 50 replicates) for $\delta \in \{0.01, 0.05, 0.1\}$ using data generated in Simulation I-2 for binary phenotypes and Simulation II-2 for continuous phenotypes of the main paper. We find that our iGSEA methods seem to perform well in FDR control and their empirical values are pretty close to the nominal ones, except for $\delta = 0.1$ in the discrete case, where the values are a bit inflated. Nevertheless, they are consistently better than the MAPEs whose empirical FDR values are inflated in almost all the settings considered, especially for the continuous case.

Table SII: Comparison of FDR: the empirical FDR versus the Q-value cutoff δ

Phenotype	Model	Q-value cutoff	iGSEAi-FE	iGSEAi-RE	MAPE_G	MAPE_P	MAPE_I
Discrete	FE	0.01	0.01	0.02	0.33	0.03	0.02
		0.05	0.07	0.05	0.09	0.13	0.09
		0.10	0.12	0.14	0.13	0.20	0.16
	RE	0.01	0.01	0.01	0	0.03	0.02
		0.05	0.07	0.08	0	0.10	0.07
		0.10	0.16	0.19	0.16	0.16	0.13
Continuous	FE	0.01	0	0	0.50	0.50	0
		0.05	0.08	0.03	0.60	0.25	0.20
		0.10	0.09	0.08	0.71	0.36	0.35
	RE	0.01	0	0.02	0.50	0.08	0
		0.05	0.05	0.02	0.50	0.21	0.13
		0.10	0.10	0.10	0.53	0.19	0.24

S6 Power comparison based on gene-level expression.

As suggested by one of the reviewers, we simulate the scenario in which each gene only has one isoform and all the methods are compared based on gene-level expression. Figure S6 shows results of power comparison for settings with the number of studies $K = 6$, the enrichment signal $\alpha = 0.3$, the sampling rate $\lambda = 0.7$ and the mean isoform effect size $v \in \{0.25, 0.75, 1\}$ for data generated from both FE and RE models. We find that, based on gene-level expression, for the discrete case, when the data is generated from the FE model, the performance of iGSEAi-FE, iGSEAi-RE and MAPE_P is pretty close, which is slightly better than MAPE_I and substantially better than MAPE_G; when the data is generated from the RE model, iGSEAi-RE is better than the other methods but iGSEAi-FE is worse than MAPE_P and MAPE_I. For the continuous case, our proposed methods are better than MAPEs. Compared with the simulation results reported in the main paper, we find that the improvement of our proposed methods over MAPEs based on isoform-level expression is typically much larger. This seems to suggest that utilizing isoform-level expression can improve the power of iGSEA.

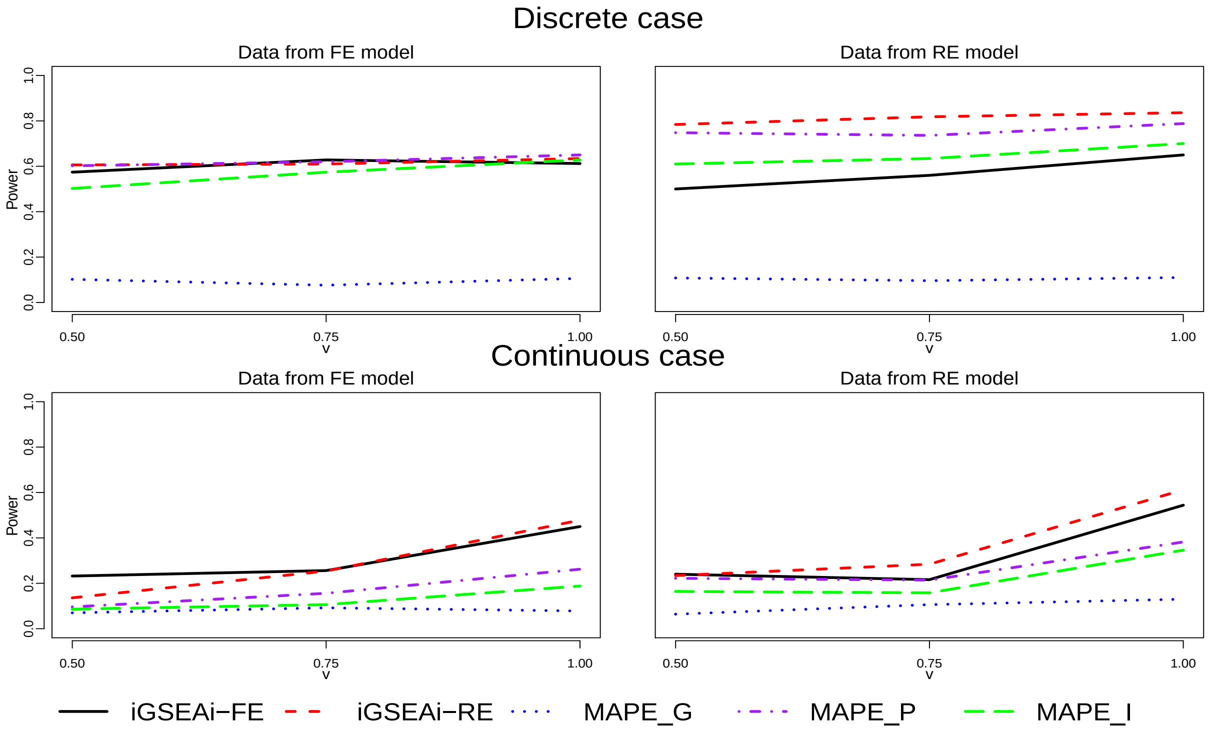

Figure S6: Power comparison based on gene-level expression, where $K = 6$, $\alpha = 0.3$, $\lambda = 0.7$, and $v \in \{0.25, 0.75, 1\}$.

S7 Breast cancer data example: list of KEGG pathways identified

Table SIII summarizes the top pathways identified based on the Q-values determined by iGSEAi-RE, while iGSEAi-FE gives very similar results in this example.

Table SIII: Data example: top KEGG pathways identified by iGSEAi-RE with Q-values reported for all the methods compared.

Pathways	MAPE_P	MAPE_G	MAPE_I	iGSEAi-FE	iGSEAi-RE	
CELL_CYCLE	0.42	0.13	0.20	6.54E-05	4.65E-05	X
PATHWAYS_IN_CANCER	0.60	0.07	0.14	6.54E-05	4.65E-05	X
COLORECTAL_CANCER	0.45	0.12	0.18	6.54E-05	4.65E-05	X
PANCREATIC_CANCER	0.44	0.09	0.15	6.54E-05	4.65E-05	X
GLIOMA	0.57	0.24	0.32	6.54E-05	4.65E-05	X
PROSTATE_CANCER	0.44	0.15	0.21	6.54E-05	4.65E-05	X
CHRONIC_MYELOID_LEUKEMIA	0.40	0.10	0.16	6.54E-05	4.65E-05	X
ACUTE_MYELOID_LEUKEMIA	0.41	0.20	0.27	6.54E-05	4.65E-05	X
SMALL_CELL_LUNG_CANCER	0.42	0.09	0.15	6.54E-05	4.65E-05	X
RIBOSOME	0.00	0.08	0.00	6.54E-05	4.65E-05	+
RNA_DEGRADATION	0.26	0.47	0.23	6.54E-05	4.65E-05	+
SPLICEOSOME	0.30	0.09	0.17	6.54E-05	4.65E-05	+
ERBB_SIGNALING_PATHWAY	0.34	0.12	0.20	6.54E-05	4.65E-05	+
LYSOSOME	0.60	0.44	0.58	6.54E-05	4.65E-05	+
ENDOCYTOSIS	0.64	0.15	0.22	6.54E-05	4.65E-05	+
FOCAL_ADHESION	0.50	0.13	0.16	6.54E-05	4.65E-05	+
ADHERENS_JUNCTION	0.28	0.09	0.15	6.54E-05	4.65E-05	+
NEUROTROPHIN_SIGNALING_PATHWAY	0.37	0.10	0.19	6.54E-05	4.65E-05	+
OOCYTE_MEIOSIS	0.57	0.19	0.24	6.54E-05	4.65E-05	
UBIQUITIN_MEDIATED_PROTEOLYSIS	0.42	0.16	0.23	6.54E-05	4.65E-05	
AXON_GUIDANCE	0.50	0.13	0.19	6.54E-05	4.65E-05	
B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.37	0.09	0.15	6.54E-05	4.65E-05	
FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	0.49	0.11	0.20	6.54E-05	4.65E-05	
REGULATION_OF_ACTIN_CYTOSKELETON	0.64	0.11	0.21	6.54E-05	4.65E-05	
HUNTINGTONS_DISEASE	0.44	0.47	0.48	6.54E-05	4.65E-05	
LYSINE_DEGRADATION	0.61	0.44	0.59	0.000107	4.65E-05	
EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.30	0.21	0.27	0.000107	4.65E-05	
APOPTOSIS	0.58	0.19	0.25	0.000345	4.65E-05	+