Web-based supplementary materials for "Identification of differentially expressed gene sets using the Generalized Berk-Jones statistic"

Sheila M. Gaynor*, Ryan Sun*, Xihong Lin, John Quackenbush

February 14, 2019

Supplementary Methods

Because the cumulative logit model used to simultaneously model grades 1, 2, and 3 in the sensitivity analysis is a non-canonical-link model with atypical interpretation, we briefly review the model here. Let

$$\gamma_k(\mathbf{G}_i) = \Pr\left(Y_i \leq k | \mathbf{G}_i\right).$$

We define the cumulative logits as

$$\begin{aligned}
\text{logit}\left(\gamma_k(\mathbf{G}_i)\right) &= \log\left(\frac{\gamma_k(\mathbf{G}_i)}{1-\gamma_k(\mathbf{G}_i)}\right), \\
&= \alpha_{k0} + \mathbf{G}_i^T \boldsymbol{\beta}_k \\
\boldsymbol{\beta}_k &= (\beta_{1k}, \beta_{2k}, ..., \beta_{dk}) \\
\mathbf{G}_i &= (G_{i1}, .., .G_{id})
\end{aligned}$$

for $k = 1, 2$. Here each element of $\boldsymbol{\beta}_k$ can be interpreted as a log-cumulative odds ratio. We further make the proportional odds assumption that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{\mathbf{G}}$ to reduce the large number of parameters in our testing problem and create a more parsimonious model.

The global null hypothesis of no association between tumor grade and expression can again be expressed as $H_0 : \boldsymbol{\beta}_{\mathbf{G}} = \mathbf{0}_{d\times 1}$. To estimate the correlation structure of these regression coefficients we use a nonparametric bootstrap. For each set, each gene expression in the set is first placed one by one into the base cumulative logit model (with only an intercept) to produce a test statistic summarizing the marginal evidence of association. These test statistics take the place of Z_j in Equation (2) of the main manuscript.

We then permute the outcomes $Y_i, i = 1, 2, ..., n$ 100 different times, each time breaking any existing associations between gene expression and outcome. After each permutation, we again place each gene expression into the model one by one and record the test statistic; therefore, after 100 permutations, we have a $100 \times d$ matrix of test statistics calculated under the null. We find the correlation structures of these bootstrapped results, and this correlation is used in performing the Generalized Berk-Jones test.

This additional analysis demonstrates the exceptional flexibility of the GBJ framework. Any number of models may be used to determine the marginal associations between an element in the set of interest and the outcome. While we provided the standard GLM formulation in the main text for simplicity, it is also possible to build a more specialized cumulative logit model, mixed model, or generalized estimating equation, for example. Thus the individual-level data can be modeled in the manner most appropriate for each specific application, with GBJ employed afterwards for set-based inference.

Supplementary Figures

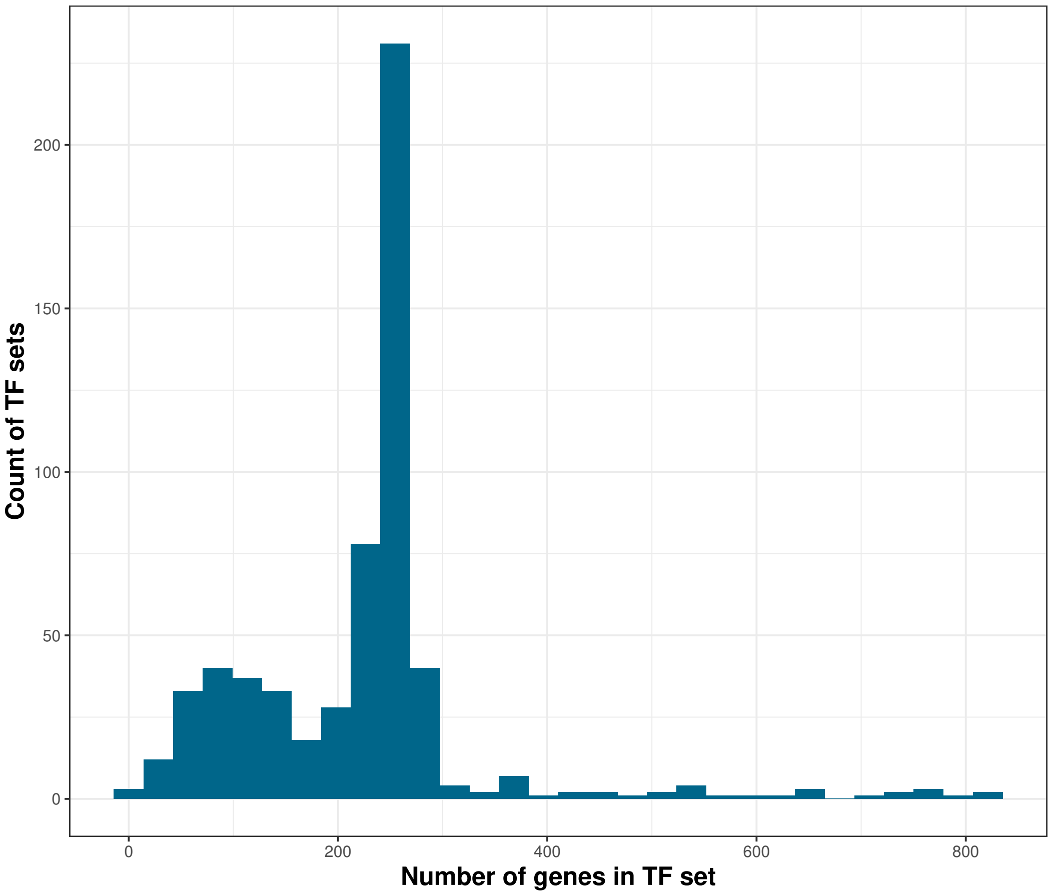

Figure S1: Histogram illustrating the number of genes included in each of the 593 sets tested. The median number of genes in a set is 243.

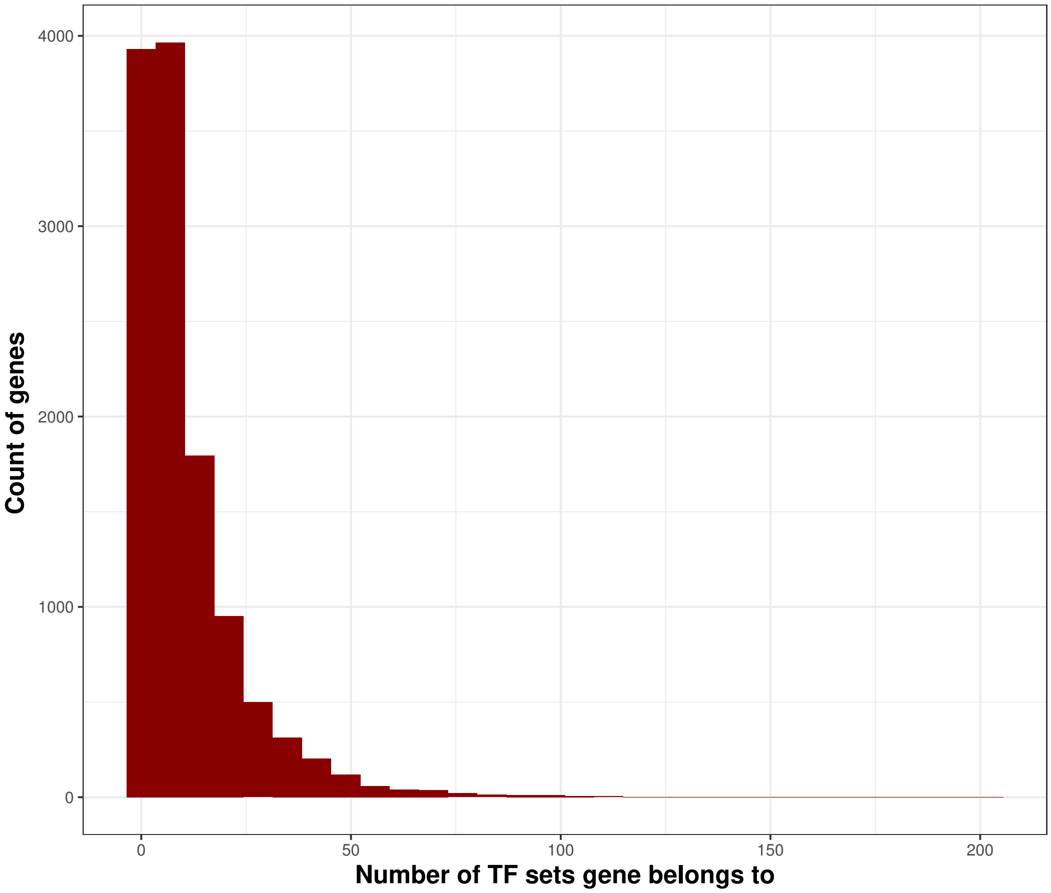

Figure S2: Histogram illustrating the number of different sets to which each gene belongs. If a certain gene appears in every set, then that gene would have value 593 in this graphic. The median number of sets that gene belongs to is 6.

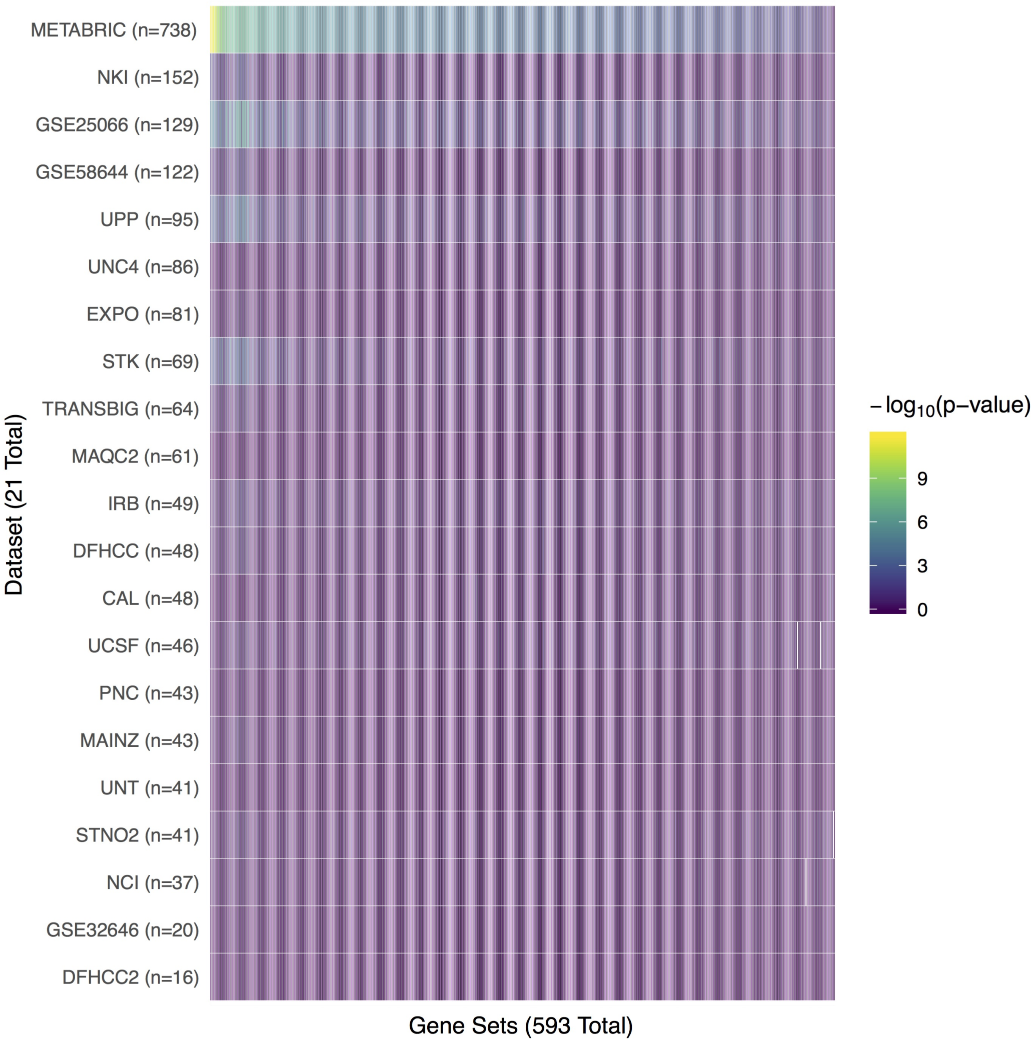

Figure S3: Heatmap illustrating the GBJ $-\log_{10}(p - value)$ for each gene set in each study. There are 21 total datasets and 593 total gene sets. In contrast to Figure 3 of the main manuscript, this figure uses the actual p-values of each set instead of their p-value rank. METABRIC possesses by far the largest sample size out of all the datasets, and so the METABRIC p-values demonstrates far more significance than the p-values of the other studies.

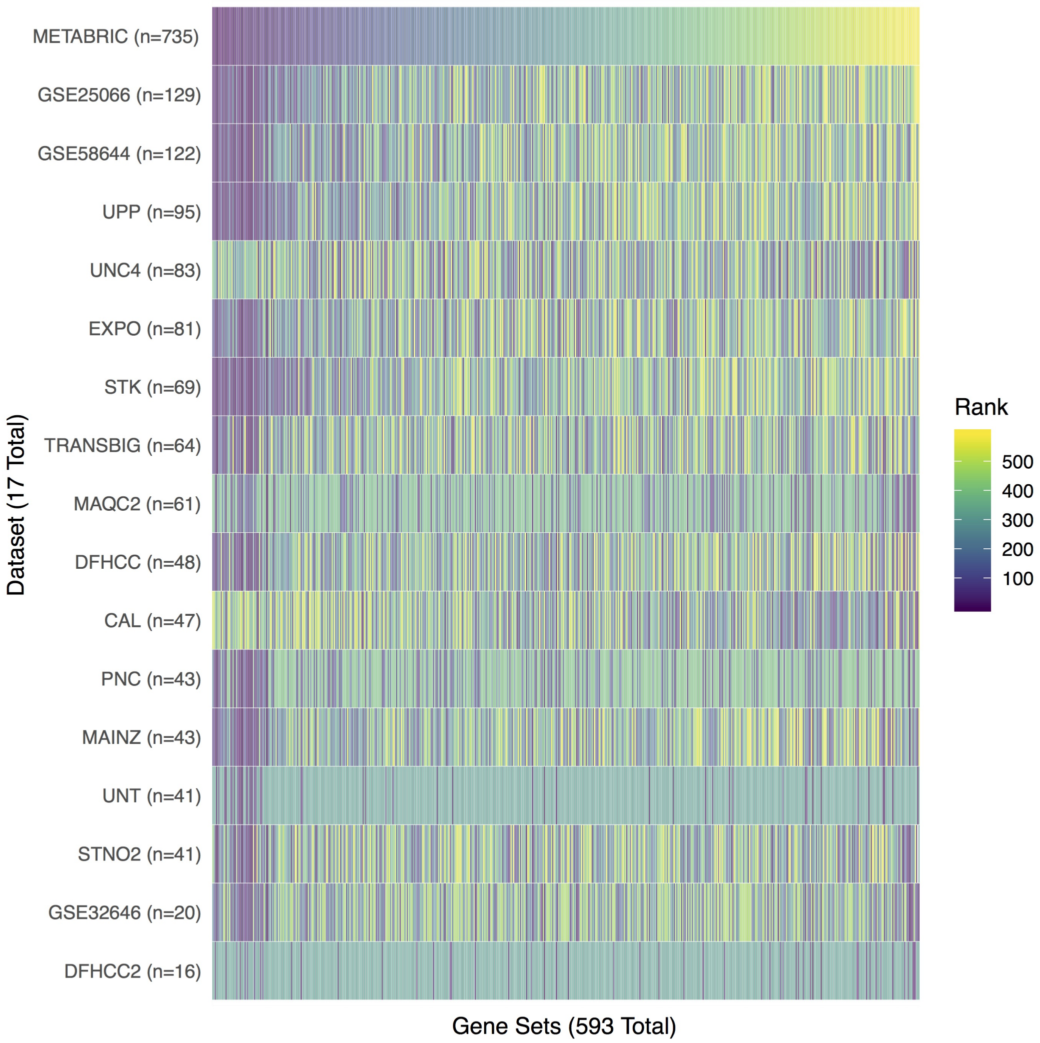

Figure S4: Sensitivity analysis heatmap illustrating the significance rank of gene sets when age-adjusted models are used in place of Equation (1) in the main manuscript. Rank 1 corresponds to the lowest p-value achieved in the study, and rank 593 corresponds to the largest p-value for that study. The studies are ordered by size, and the gene sets are ordered according to their rank in the largest study (METABRIC). We only include 17 studies because only 17 studies contain age information. The results appear very similar to Figure 3 in the main analysis, demonstrating the robustness of GBJ results to different models of marginal association.

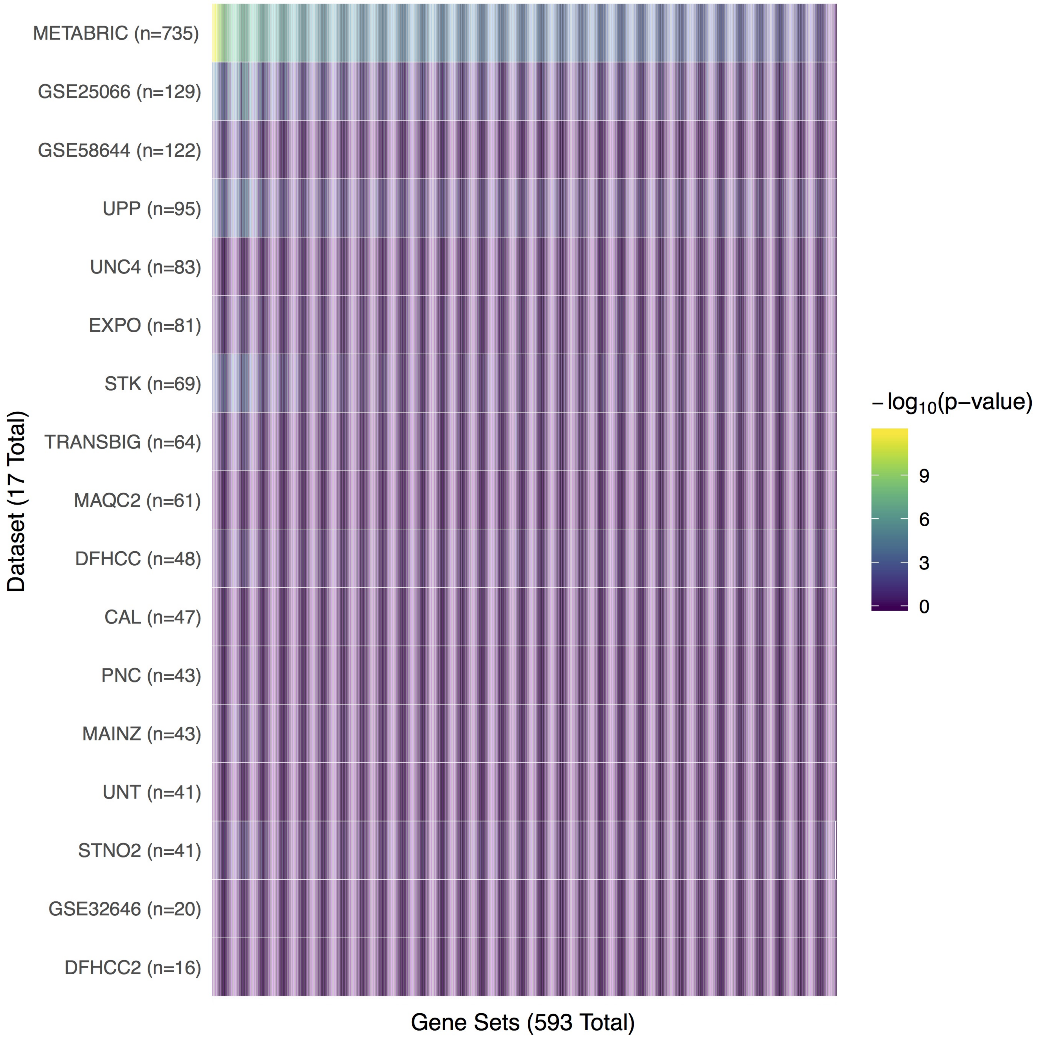

Figure S5: Sensitivity analysis heatmap illustrating the GBJ $-\log_{10}(p - value)$ for each gene set in each study when the analysis is age-adjusted. In contrast to Supplementary Figure 4, this figure uses the actual p-values of each set instead of their p-value rank. METABRIC possesses by far the largest sample size out of all the datasets, and so the METABRIC p-values demonstrates far more significance than the p-values of the other studies. Tthis figure appears very similar to Supplementary Figure S3, demonstrating the robustness of GBJ results to different models of marginal association.

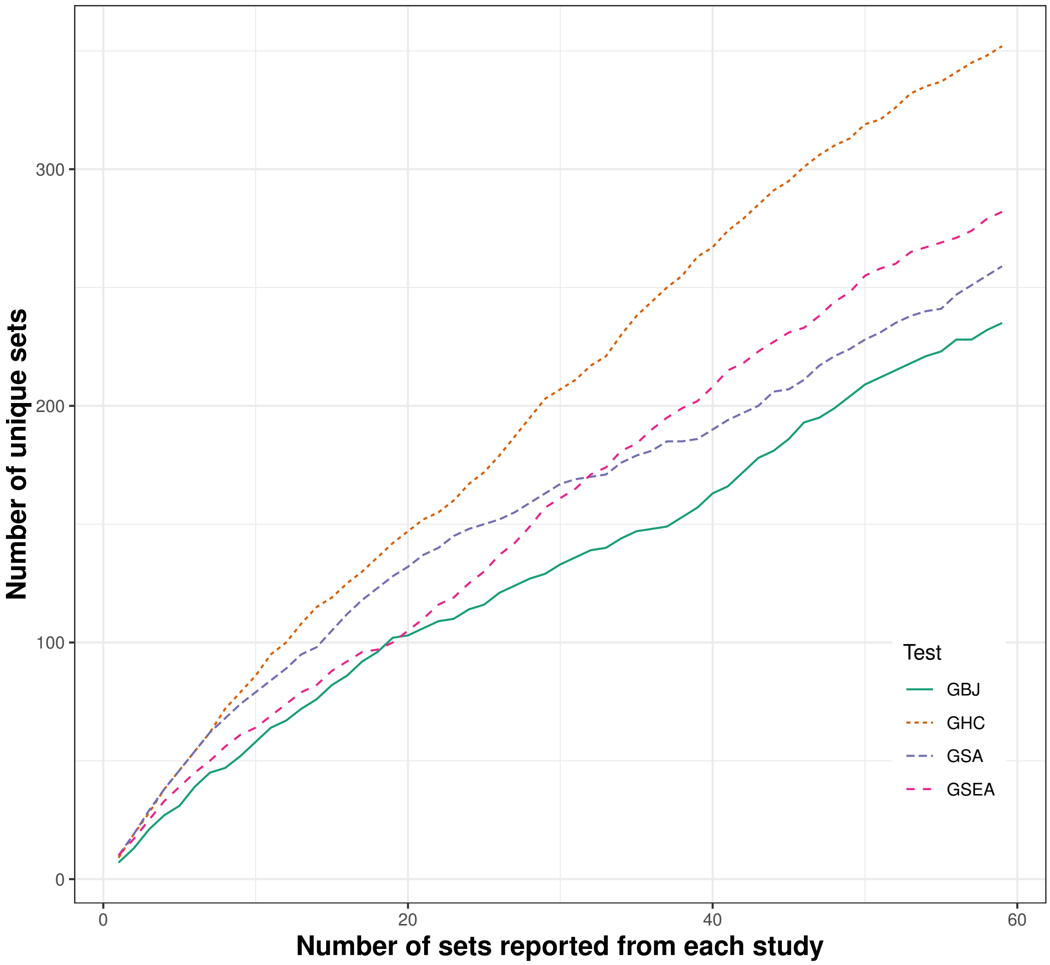

Figure S6: Number of unique gene sets identified across the smallest studies under GBJ, GHC, GSA and GSEA. The x-axis shows the number of sets to report from each study (i.e. at x=1 only the most significant gene set from each study is reported). The y-axis shows the number of unique gene sets reported across all studies. A smaller number of unique gene sets indicates that the method is replicating the same results in multiple data sets. Analysis with GBJ is more likely to report the same top pathways over multiple different studies. Only the ten studies with the smallest sample sizes are considered in this plot.

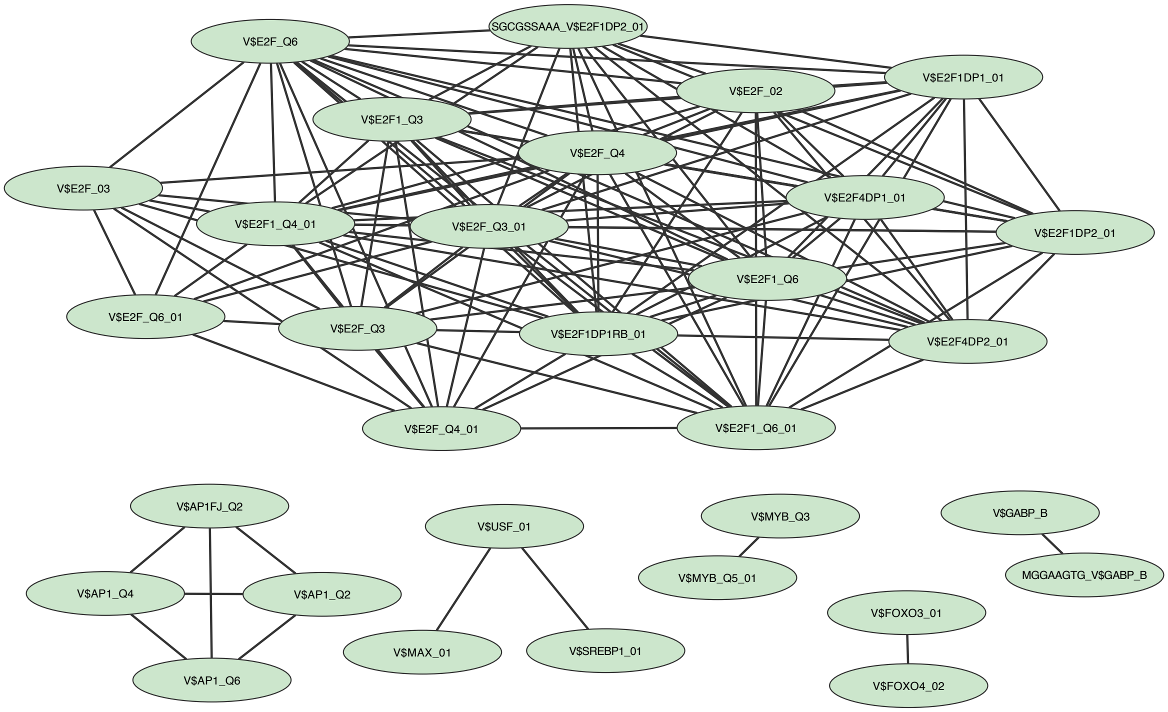

Figure S7: Map of all communities (non-singletons) identified from a network analysis of meta-analyzed GBJ results when using the cumulative logits model and including tumors of all grades. Transcription factor sets are given by nodes, with edges representing genes in common between sets. The largest connected component includes all the same nodes and edges found in the main analysis (Figure 2).

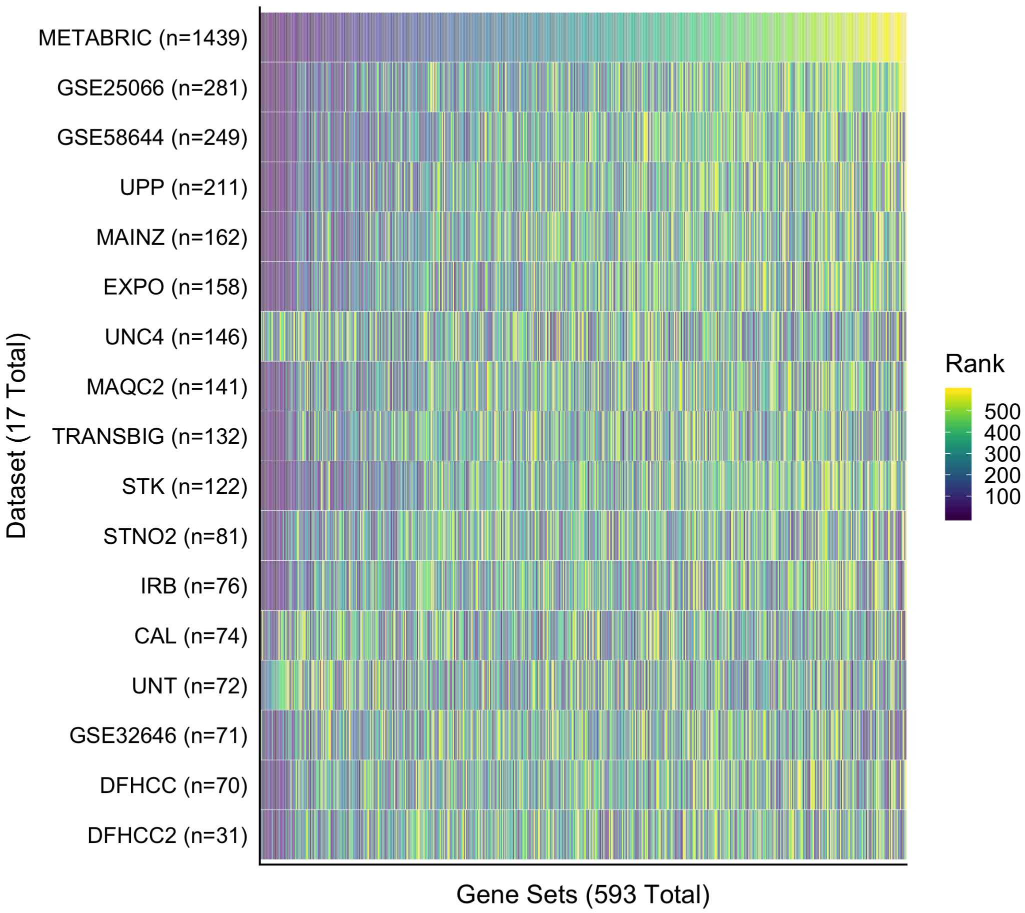

Figure S8: Heatmap illustrating the significance rank of each gene set in each study when using the cumulative logits model with data from tumors of all grades. Note that this figure appears very similar to Figure 3 of the main text, demonstrating the robustness of GBJ results to both modeling frameworks. GBJ is able to replicate the same set of significant gene sets across most studies whether we pursue the logistic regression strategy of the main manuscript or the ordinal cumulative logits model described in the Supplementary Methods.

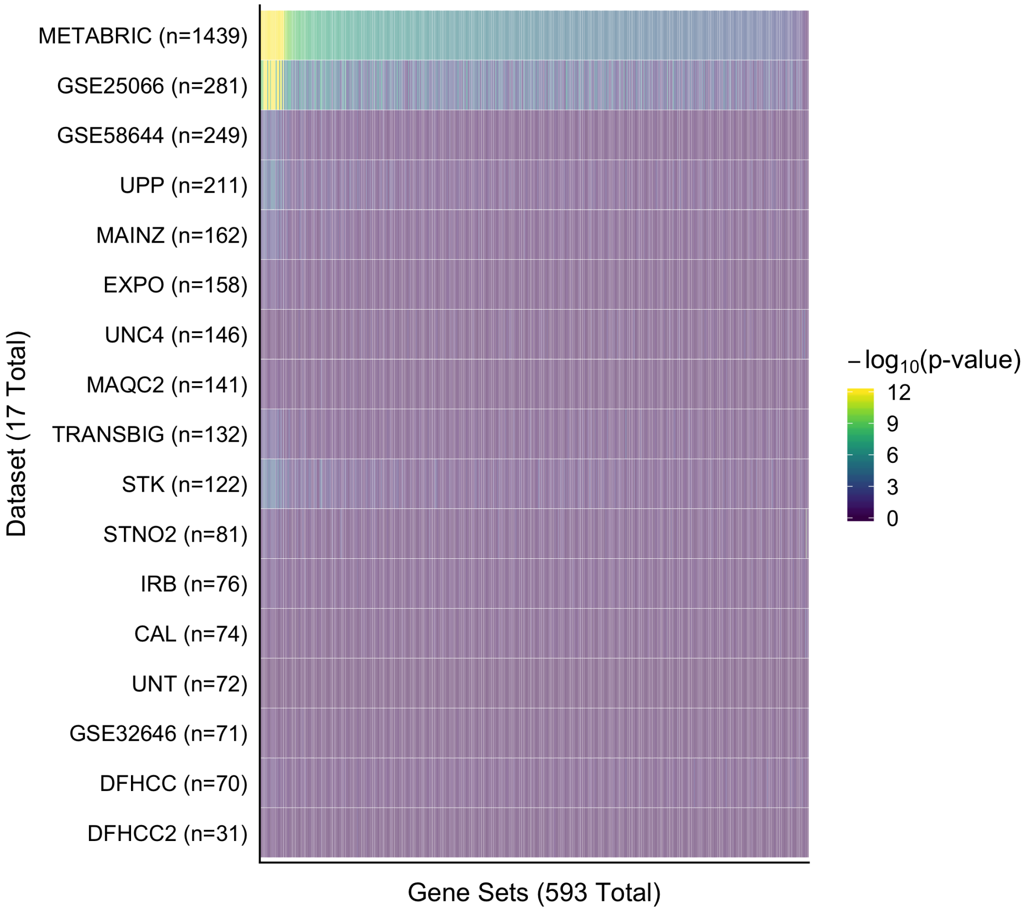

Figure S9: Heatmap illustrating the GBJ $-\log_{10}(p - value)$ for each gene set in each study when using the cumulative logits model with data from tumors of all grades. This figure appears very similar to Supplementary Figure S3.

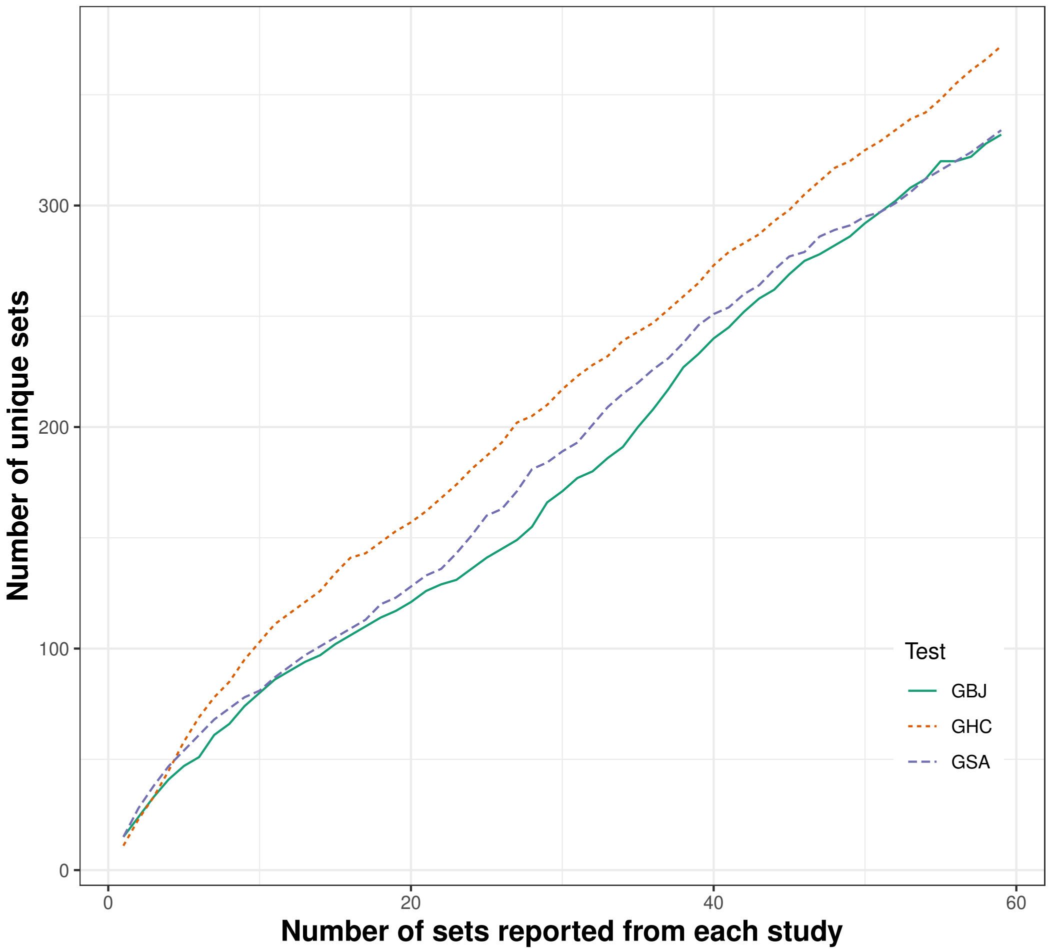

Figure S10: Number of unique gene sets identified across all 17 studies under GBJ, GHC, and GSA when comparing all grades in the cumulative logits analysis. The x-axis specifies how many top-ranked gene sets are reported from each study. The y-axis shows the number of unique gene sets reported across all studies. A smaller number of unique gene sets indicates that the method is replicating the same results in multiple data sets. GBJ is more likely to replicate the same top pathways across studies, as was also observed in the main analysis in Figure 4 and Supplementary Figure S6.

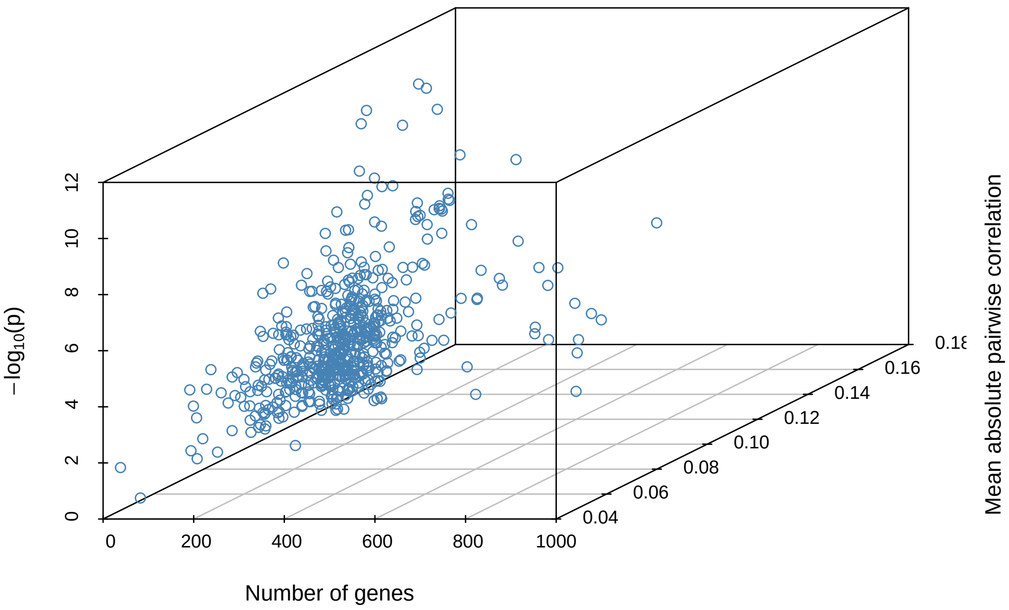

Figure S11: Three-dimensional plot stratifying the results of our real data analysis by correlation structure of the gene set. There is one axis each for the number of genes in a set, the gene set p-value, and the mean absolute pairwise correlation in the correlation structure of the set. We use mean absolute pairwise correlation because it is able to summarize the strengths of ties between different elements into a single number. It is necessary to summarize the correlation structure into a single dimension so that we can plot it along a single axis. We use correlation structures and GBJ p-values from METABRIC because this is the largest study in our dataset, so the estimated correlation and corresponding p-value should be more accurate.

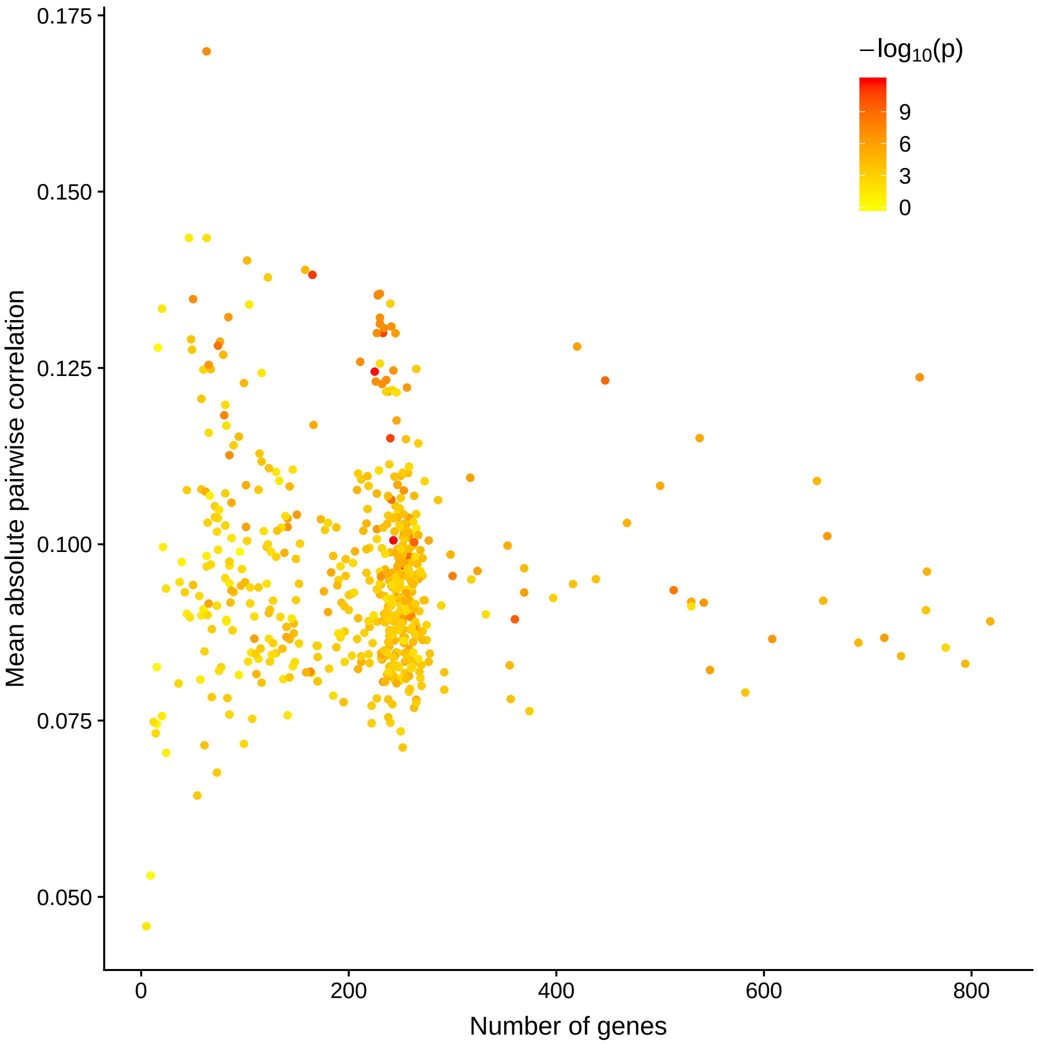

Figure S12: Two-dimensional plot showing the same information as Supplementary Figure S11, but with color in the place of a third axis. Highly significant gene sets are observed at all quantiles of set size and mean absolute bivariate correlation.

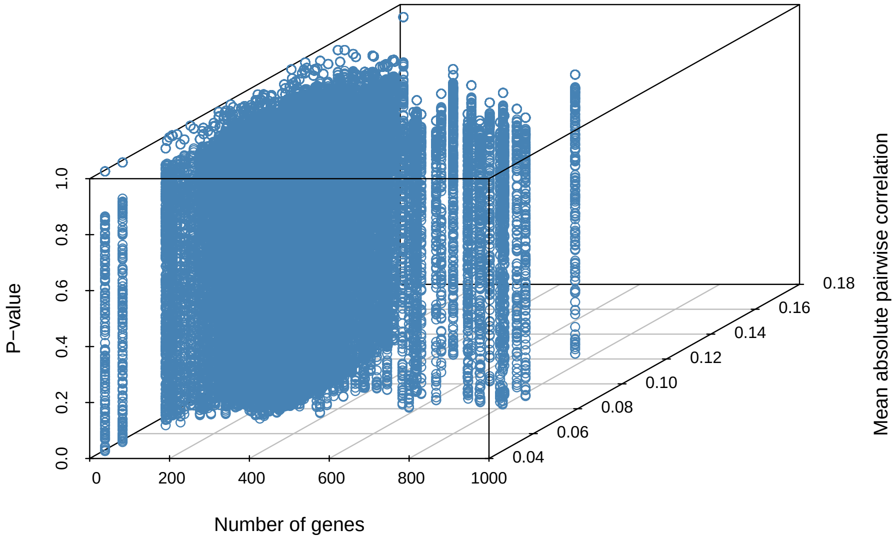

Figure S13: Three-dimensional plot stratifying the results of our Type I error simulation (see Supplementary Table S2) by correlation structure of the gene set. There is one axis each for the number of genes in a set, the gene set p-value, and the mean absolute pairwise correlation in the correlation structure of the set. We use mean absolute pairwise correlation because it is able to summarize the strengths of ties between different elements into a single number to plot along a single axis.

Supplementary Tables

Table S1: GBJ p-value for association of transcription factor gene sets with breast cancer tumor grade in the main logistic regression analysis. P-values are given for all 21 studies. See attached file.

Table S2: GBJ simulated type I error rate using METABRIC expression correlation structures. In each iteration of the simulation, we simulate binary outcomes with a constant success probability of 0.5 for each subject in the METABRIC study. We then draw a gene set at random from those used in the analysis and test it for association with the outcome according to the strategy outlined in the main text. The nominal significance levels are set at 0.05 and 0.005. We perform 50,000 iterations of the simulation and report the Type I error rate below.

α	Attained Size
0.05	0.03491
0.005	0.00567

Table S3: Summary of grade distribution from the 17 MetaGxBreast datasets used in the cumulative logits analysis where grades 1, 2, and 3 are all considered.

Dataset	Grade 1	Grade 2	Grade 3
CAL	10, (8.8%)	42, (37.2%)	61, (54%)
DFHCC	23, (20%)	28, (24.3%)	64, (55.7%)
DFHCC2	10, (11.9%)	16, (19%)	58, (69%)
EXPO	32, (10.8%)	114, (38.4%)	151, (50.8%)
GSE25066	32, (6.8%)	180, (38.2%)	259, (55%)
GSE32646	16, (13.9%)	78, (67.8%)	21, (18.3%)
GSE58644	26, (8.1%)	135, (42.2%)	159, (49.7%)
IRB	27, (20.9%)	32, (24.8%)	70, (54.3%)
MAINZ	29, (14.5%)	136, (68%)	35, (17.5%)
MAQC2	13, (5.7%)	94, (40.9%)	123, (53.5%)
METABRIC	170, (8.9%)	775, (40.7%)	957, (50.3%)
STK	28, (19%)	58, (39.5%)	61, (41.5%)
STNO2	11, (9.7%)	49, (43.4%)	53, (46.9%)
TRANSBIG	30, (15.3%)	83, (42.3%)	83, (42.3%)
UNC4	25, (10.3%)	80, (32.9%)	138, (56.8%)
UNT	32, (28.6%)	51, (45.5%)	29, (25.9%)
UPP	67, (26.9%)	128, (51.4%)	54, (21.7%)

Table S4: Top transcription factor gene sets associated with breast cancer tumor grade after meta-analysis of GBJ results over 17 studies in the cumulative logits analysis where all tumor grades are considered. The 593 sets consist of all genes regulated by a certain transcription factor. All top sets reported in Table 2 from the main logistic regression analysis can also be found below. The p-values in this table are lower than those in Table 2 because the sample sizes are larger when using all three tumor grades.

Set name (MSigDB)	Transcription factor	Meta-analysis p-value
E2F1DP2_01	E2F1	$< 1 \cdot 10^{-16}$
E2F_Q4	E2F4	$< 1 \cdot 10^{-16}$
E2F4DP2_01	E2F4	$< 1 \cdot 10^{-16}$
E2F_Q6	E2F4	$< 1 \cdot 10^{-16}$
E2F1DP2_01	E2F1	$< 1 \cdot 10^{-16}$
E2F_Q6_01	E2F4	$< 1 \cdot 10^{-16}$
E2F1_Q6	E2F1	$< 1 \cdot 10^{-16}$
E2F1DP1_01	E2F1	$< 1 \cdot 10^{-16}$
E2F_02	E2F2	$< 1 \cdot 10^{-16}$
E2F4DP1_01	E2F4	$< 1 \cdot 10^{-16}$
E2F1_Q4_01	E2F1	$< 1 \cdot 10^{-16}$
E2F_Q3_01	E2F4	$< 1 \cdot 10^{-16}$
E2F1_Q3	E2F1	$< 1 \cdot 10^{-16}$
E2F1DP1RB_01	E2F1	$< 1 \cdot 10^{-16}$
E2F_03	E2F1	$< 1 \cdot 10^{-16}$
E2F_Q4_01	E2F4	$< 1 \cdot 10^{-16}$
E2F_Q3	E2F Family	$< 1 \cdot 10^{-16}$
NFY_Q6	NF-Y	$< 1 \cdot 10^{-16}$
E2F1_Q6_01	E2F1	$< 1 \cdot 10^{-16}$
ALPHACP1_01	Alpha-CP1	$< 1 \cdot 10^{-16}$
NFY_C	NF-Y	$< 1 \cdot 10^{-16}$
NFY_01	NF-Y	$< 1 \cdot 10^{-16}$