Supplementary materials
Gene Set Analysis: Limitations in popular existing methods and proposed improvements

Pashupati Mishra, Petri Törönen, Yrjö Leino, Liisa Holm

Contents

1 Extreme value distributions .......................... 2
   1.1 Extreme value type I distribution .............. 2
   1.2 Extreme value type II distribution .............. 3
   1.3 Extreme value type III distribution .............. 3
   1.4 General extreme value distribution .............. 3

2 Series expansion for p-value calculation from EVD and GEVD 4

3 Schematic representation of mGSZ 5

4 Evaluation of the p-value calculation methods with p53 data 6
   4.1 Asymptotic and empirical p-values for mGSZ scores .......... 7
   4.2 Asymptotic and empirical p-values for mGSA scores .......... 8
   4.3 Asymptotic and empirical p-values for mAllez scores .......... 9
   4.4 Asymptotic and empirical p-values for SUM scores .......... 10

5 Evaluation of the p-value calculation methods with gender data 10
   5.1 Asymptotic and empirical p-values for mGSZ scores .......... 10
   5.2 Asymptotic and empirical p-values for mGSA scores .......... 11
   5.3 Asymptotic and empirical p-values for mAllez scores .......... 12
   5.4 Asymptotic and empirical p-values for SUM scores .......... 12
   5.5 Asymptotic approximation of empirical null distribution with gender data . 12

6 Comparison of stability of gene set scores with gender data 14

7 Calculation of prior variance in Gene Set Z-score function 15
1 Extreme value distributions

Extreme value distributions are limiting distributions for extreme values (maximums or minimums) of a sample of independent and identically distributed (i.i.d) random variables as the sample size increases (Kotz and Nadarajah, 2000). The extreme value distributions can be based either on the smallest extreme or the largest extreme. We used extreme value distributions to model mGSZ and mGSA scores. Both mGSZ and mGSA scores are maximum values from a group of data. For that reason, we will focus mainly on the extreme value distribution based on the largest extreme. The class of Extreme Value Distributions involves three types of extreme value distributions, type I, type II and type III.

1.1 Extreme value type I distribution

The extreme value type I distribution is also referred to as the Gumbel distribution. The probability density function of the extreme value type I distribution is,

\[ f(x) = \frac{1}{\beta} e^{-(x-\mu)/\beta} e^{-e^{-(x-\mu)/\beta}} \]  

The cumulative distribution function of the extreme value type I distribution is,

\[ F(x) = e^{-e^{-(x-\mu)/\beta}} \]
where, $\mu$ is the location parameter and $\beta > 0$ is the scale parameter. Standard extreme value type I distribution has $\mu = 0$ and $\beta = 1$.

### 1.2 Extreme value type II distribution

Extreme value type II distribution is also referred as to Fréchet distribution. The probability density function of the extreme value type II distribution is,

$$f(x) = \frac{a}{\beta} \left( \frac{x - \mu}{\beta} \right)^{-1-a} e^{-\frac{x-\mu}{\beta}}$$

The cumulative distribution function of the extreme value type II distribution is,

$$F(x) = e^{-\frac{x-\mu}{\beta}}$$

where, $\mu$ is the location parameter, $\beta$ is the scale parameter and $a$ is the shape parameter.

### 1.3 Extreme value type III distribution

Extreme value type III distribution is also called as Weibull distribution. The probability density function of the extreme value type III distribution is,

$$f(x) = \begin{cases} 
\frac{a}{\beta} \left( \frac{x}{\beta} \right)^{a-1} e^{-(x/\beta)^a} & x \geq 0 \\
0 & x < 0 
\end{cases}$$

The cumulative distribution function of the extreme value type III distribution is,

$$F(x) = \begin{cases} 
1 - e^{-(x/\beta)^a} & x \geq 0 \\
0 & x < 0 
\end{cases}$$

where, $\mu > 0$ is the location parameter, $\beta > 0$ is the scale parameter and $a$ is the shape parameter.

### 1.4 General extreme value distribution

General extreme value distribution (GEVD) combines the extreme value type I, II and III distributions. The probability density function of the GEVD is,
The cumulative distribution function is,

\[
F(x) = \begin{cases} 
\frac{1}{\beta} e \left( (1 + az)^{-1/a} \right) & a \neq 0 \\
\frac{1}{\beta} e \left( -z - e(-z) \right) & a = 0
\end{cases}
\]

where, \( z = (x - \mu) / \beta \) and \( a, \beta, \mu \) are shape, scale and location parameters respectively. The scale must be positive, the shape and location can take on any real value.

2 Series expansion for p-value calculation from \textit{EVD} and \textit{GEVD}

Log p-values from EVD and GEVD are calculated as,

\[
F(x) = -\log(p\text{-value}_{EVD}) = -\ln(1 - e^{-e^x})
\]

where,

\[
x = -(z - \mu) / \beta
\]

\( z \) is the absolute mGSZ or mGSA score value for the analyzed gene set and \( \mu \) and \( \beta \) are location and scale parameters respectively for \textit{EVD}.

and

\[
G(x) = -\log(p\text{-value}_{GEVD}) = -\ln \left( 1 - e^{-(1+ax)^{-1/a}} \right), \ a > 0
\]

where,

\[
x = (z - \mu) / \beta
\]
$z$ is the absolute mGSZ or mGSA score value for the analyzed gene set and $\mu$, $\beta$ and $a$ are the location, scale and shape parameters respectively for GEVD.

We used two well known power series expansions for calculation of extremely small $p$-values from $EVD$ and $GEVD$, namely

$$e^{-t} = 1 - t + t^2/2 - t^3/6 + t^4/24 + \ldots$$

and

$$\ln(1 + t) = t - t^2/2 + t^3/3 - t^4/4 + \ldots$$

Both expansions converge rapidly when $|t|$ is reasonably small.

Now, $F(x)$ and $G(x)$ are essentially of the similar form $p(t) = \ln(1 - e^{-t})$, so we shall start by finding the power series expansion for $p(t)$ near $t = 0$. Applying the first expansion above for the exponential function, the rule for the logarithm of a product $\ln(ab) = \ln(a)\ln(b)$, and finally the second expansion with $t$ equalling the whole power series that remains inside the logarithmic function in addition to the constant term 1, we get for our auxiliary function $p(t)$ the result

$$p(t) = \ln(t) - t/2 + t^2/24 - t^4/2880 + t^6/181440 + \ldots$$

To get the series expansion for function $F(x)$, we replace $t$ in $p(t)$ by $e^x$. We shall apply the resulting series for large negative values of $x$, which means that the series converges rapidly. Taking into account the minus sign in the definition of $F(x)$ we end up with the expansion:

$$F_s(x) = -x + e^x/2 - e^{2x}/24 + e^{4x}/2880 - e^{6x}/181440 + \ldots$$

This result shows us that when $x \to -\infty$, the value of $F_s(x)$ approaches $-x$ monotonously from above, the difference being of order $e^x/2$.

In a completely analogous fashion it can be shown that for the function $G(x)$ the corresponding expansion takes the form

$$G_s(x) = \frac{\ln(1 + ax)}{a} + \frac{(1 + ax)^{-1/a}}{2} - \frac{(1 + ax)^{-2/a}}{24} + \frac{(1 + ax)^{-4/a}}{2880} - \frac{(1 + ax)^{-6/a}}{181440} - \ldots$$

With $x \to \infty$, function $G_s(x)$ approaches $-(\ln(1 + ax))/a$ from below, and for large values of $x$, the difference is of order $((1 + ax)^{-1/a})/2$.

3 Schematic representation of mGSZ
Figure 1: Schematic representation of mGSZ workflow. mGSZ takes differential gene expression test scores and list or matrix of gene sets as input. The gene list with differential gene expression test scores is (a) ordered based on differential gene expression test scores. Subsets of genes are taken at threshold positions placed in between every consecutive pair of genes, and for each subset (b) the difference between the sum of differential gene expression test scores for the member and non-member genes of the analyzed gene set is calculated, the calculated difference is (c) normalized with the estimates of mean and standard deviation, and (d) the largest absolute value is selected as the GSZ score for the analyzed gene set. The steps (a), (b), (c) and (d) is repeated with differential gene expression test scores from permuted (sample permutation) gene expression data. (h) Asymptotic distribution model is fitted to the GSZ scores from permuted gene expression data. (i) Asymptotic p-value is calculated for the analyzed gene set.

4 Evaluation of the p-value calculation methods with p53 data

In this section, we evaluate asymptotic and empirical methods of p-value estimation. Mean and standard deviation of the gene set scores were adjusted to 0 and 1 respectively for all but GAMMA p-value estimation. We considered empirical p-values from 100000 sample permutations as the reference of truth, against which we compared asymptotic p-values and empirical p-values from 500 sample permutations. Our analysis is based on log p-values.
4.1 Asymptotic and empirical p-values for mGSZ scores

The resolution of empirical p-values depends on the number of permutations (Figure 2). As expected, p-values obtained from Gaussian model fitted to mGSZ scores failed miserably. P-values obtained from extreme value (EVD) and general extreme value (GEVD) distributions fitted to mGSZ scores were the best estimates (Figure 2).

![Correlation of empirical and asymptotic p-values](image)

**Figure 2:** Correlation of empirical and asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.
4.2 Asymptotic and empirical p-values for mGSA scores

The best p-values for mGSA scores were the asymptotic p-values estimated from extreme value distribution (EVD), general extreme value distribution (GEVD) and gamma distribution (GAMMA) fitted to mGSA scores (Figure 3). Empirical p-values and asymptotic p-values obtained from Gaussian distribution fitted to mGSA scores had the worst correlation with the reference of truth (Figure 3).

Figure 3: Correlation of empirical and asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.
4.3 Asymptotic and empirical p-values for mAllez scores

Asymptotic p-values for mAllez scores were estimated from Gaussian distribution fitted to mAllez scores. Here too, empirical p-value estimated with 500 sample permutations had the worst correlation with the reference of truth (Figure 4).

Figure 4: Correlation of empirical and asymptotic p-values (NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.
4.4 Asymptotic and empirical p-values for SUM scores

Empirical and asymptotic p-value evaluation for SUM scores showed similar results as that of mAllez (see Figure 5).

![Correlation of empirical and asymptotic p-values for SUM scores with the reference of truth](image)

Figure 5: Correlation of empirical and asymptotic p-values (NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.

5 Evaluation of the p-value calculation methods with gender data

Null distribution of gene set scores from permuted data is mostly dependent on gene set scoring functions. In this section, we show the evaluation of asymptotic and empirical methods of p-value estimation with gender data.

5.1 Asymptotic and empirical p-values for mGSZ scores

P-values estimated with EVD and GEVD still remain the best p-values in gender data as well (Figure 6). However, in contrast to the results from p53 data, empirical p-values are equally good as asymptotic p-values calculated with EVD and GEVD (Figure 6). This is because gender data has relatively lower biological signal as compared to p53 data and thus empirical p-values reach the finest resolution already with 500 sample permutations.
Figure 6: Correlation of empirical and asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.

5.2 Asymptotic and empirical p-values for mGSA scores

The results are similar to that of p53 data except that the empirical p-values show equally good correlation with the reference of truth for the reason described in Section 5.1 (Figure 7).

Figure 7: Correlation of empirical and asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.
5.3 Asymptotic and empirical p-values for mAIllez scores

The results are similar to that of p53 data except that the empirical p-values show equally good correlation with the reference of truth for the reason described in Section 5.1 (Figure 8).

Figure 8: Correlation of empirical and asymptotic p-values (NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.

5.4 Asymptotic and empirical p-values for SUM scores

The results are similar to that of p53 data except that the empirical p-values show equally good correlation with the reference of truth for the reason described in Section 5.1 (Figure 9).

5.5 Asymptotic approximation of empirical null distribution with gender data
Table 1: Asymptotic distribution models fitted on the empirical null distribution generated by the compared methods with gender data and their results. Scores not meeting the criteria for optimal model (Section 2.5.1 from the main article) are highlighted.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Models</th>
<th>m se</th>
<th>m se(subset)</th>
<th>cor</th>
<th>cor(subset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mGSZ</td>
<td>EVD</td>
<td>0.002</td>
<td>0.03</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>GEVD</td>
<td>0.002</td>
<td>0.03</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>GAMMA</td>
<td>0.003</td>
<td>0.05</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>NORM</td>
<td>0.03</td>
<td><strong>0.73</strong></td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>mGSA</td>
<td>EVD</td>
<td>0.002</td>
<td>0.03</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>GEVD</td>
<td>0.001</td>
<td>0.02</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>GAMMA</td>
<td>0.001</td>
<td>0.02</td>
<td>0.99</td>
<td><strong>0.97</strong></td>
</tr>
<tr>
<td></td>
<td>NORM</td>
<td>0.03</td>
<td><strong>0.87</strong></td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>mAllez</td>
<td>NORM</td>
<td>0.0007</td>
<td>0.002</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>WRS</td>
<td>NORM</td>
<td>0.01</td>
<td>0.07</td>
<td><strong>0.96</strong></td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>SS</td>
<td>GAMMA</td>
<td>0.008</td>
<td><strong>0.3</strong></td>
<td>0.97</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>SUM</td>
<td>NORM</td>
<td>0.0005</td>
<td>0.002</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>KS</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>wKS</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
Figure 9: Correlation of empirical and asymptotic p-values (NORM) (X-axis) estimated from 500 sample permutations with the reference of truth (Y-axis). The reference of truth corresponds to empirical p-values estimated from 100000 sample permutations.

6 Comparison of stability of gene set scores with gender data

In this section, we present the comparison of stability of the gene set scores from mGSZ, KS and wKS with gender data. Similar to that of p53 data, here too the stability plot for mGSZ, based permuted data shows that middle part of the seven percentiles of the gene set score profiles of the permuted data stay quite stable across gene list positions (Figure 10a). The maximum percentile shows smaller deviation from zero than the minimum percentile (Figure 10a). This indicates that gene set scores coming from under-representation show stronger signal than those coming from over-representation. The minimum of the gene set score profile from original data comes quite early in the gene list, pointing that gene set members occur near the bottom of the ordered gene list (Figure 10a). Notice that the separation between gene set score profile from original data and permuted data is quite clear across different threshold positions (Figure 10a). The importance of our visualization is more highlighted when we analyze the behavior of wKS and KS with the same gene expression dataset and the same gene set. For wKS the gene set score profiles from permuted data show strange behavior with S shaped profile (Figure 10b) and for KS the middle area shows strong variation (Figure 10c). The separation between gene set score profiles from permuted data and original data is very weak in both wKS and KS (Figure 10c). This comparison is based on the strongest gene set reported by wKS. Note that the mGSZ software reports absolute values for gene set scores.
Figure 10: Visualization of scoring function profiles for mGSZ, wKS and KS for gene set "XINACT MERGED" in gender data. Figure also shows 7 percentiles (0, 10, 25, 50, 75, 90, 100) of gene set score profiles obtained from 1000 sample permutation data. Notice the clear separation of the gene set score profiles from original data with those from permuted data and stability of seven percentiles of gene set score profiles from permuted data in case of GSZ-score.

7 Calculation of prior variance in Gene Set Z-score function

For the calculation of prior variance \(k\) in Equation 2 of the main article, we use; 1) the median of the variance estimates obtained with the analyzed gene set across the whole gene list and multiply it with a weight, \(w2\) \((0 \leq w2 \leq 1)\), and 2) the median of the variance estimates obtained with the gene set of size 10 across the whole gene list and multiply it with a weight, \(w1\) \((0 \leq w1 \leq 1)\). Prior variance is then obtained by summing the weighted variance medians. The selection of the reference class size and the values for \(w1\) and \(w2\) were based on the original paper by Törönen et al., 2009.
8 Similarity of the GSZ and GSA scoring functions

For the calculation of max-mean statistics in the GSA method, the differential expression test scores are summed and divided by the size of the whole gene set in both positively and negatively regulated parts of the ordered gene list. The maximum of the absolute values of the two parts is selected as the max-mean statistics (Equation 14).

\[
MaxMean = \max \left( abs \left( \sum_{X_i > 0} X_i \right), abs \left( \sum_{X_i < 0} X_i \right) \right) / N_{GeneSet} \quad (14)
\]

Where, \( X_i \) is the differential expression test score for \( i \)th gene of the analyzed gene set and \( N_{GeneSet} \) is the total number of genes in the analyzed gene set.

GSZ equation uses non-member genes (genes not belonging to the analyzed gene set) in the calculation. However, the result from the list of non-member genes is essentially the result from the list of member genes subtracted from the result from the whole gene list. Thus, the \( Diff_N \) estimate of GSZ (Equation 1 in the main article) is a function of sum in MaxMean score function (Equation 14). Both methods focus only on one (either positively or negatively regulated) part of the gene list. The differences are: 1) the mean and standard deviation estimates for each half in GSA are obtained via gene permutations, whereas in GSZ, they are obtained by estimates from the actual data, 2) Unlike GSA, GSZ scoring function is used to test many threshold positions. However the function can be modified by fixing the threshold position.
9  Log p-values of the top gene sets

The empirical p-values of the top 50 gene sets reported by the compared methods in p53 and gender datasets calculated from 100000 sample permutations were compared. The comparison was considered inconclusive as the ordering of the methods varied between datasets. We tested the sensitivity of the methods to false positive signals in the main article. The p-value analysis should be complemented by the sensitivity test of the methods to false positive signals. Therefore, we present the results of the p-value analysis in the supplementary.

9.1  Log p-values of the top gene sets in p53 dataset

SS and mGSZ are the best performing methods on the upper region of the ordered gene list (OGL) (Figure 11a). However, on the lower region of the list, SS is the best method followed by mGSZ, SUM and mAllez (Figure 11a).

9.2  Log p-values of the top gene sets in gender dataset

Unlike the results from p53 dataset, the best methods on the upper region of the OGL of gender dataset are mGSZ and mAllez (Figure 11b). WRS and mAllez report the best p-values on the lower region of the list (Figure 11b). Surprisingly, SS reports the weakest p-values in the gender dataset (Figure 11b).

Figure 11: Empirical log p-values for the top 50 gene sets estimated by each of the compared methods in p53 and gender dataset.
False positive signal test with p53 data

In addition to the leukemia dataset, the compared methods were also evaluated with p53 dataset and GO gene sets for false positive signal test. Here, we selected GO gene sets instead of curated gene sets from Subramanian et al., 2005 because GO gene sets have wide range of sizes. This allows to test the stability of the methods over wide range of gene sets sizes. Unlike with the leukemia dataset, even though mGSZ, GSA, weighted KS and unweighted KS are more conservative, the difference to the other methods is significantly small (Figure 12).

Figure 12: Empirical log p-values for GO gene sets estimated by the compared methods on randomized p53 dataset.

Comparison of mGSZ, mGSA and mAllez with program packages

False positive signal test shows that mGSZ and most other methods show quite similar behavior (Figure 13). While CAMERA is the most conservative, ROAST shows strong
noise signal (Figure 13). The result points out a major difference between competitive and self-contained gene set analysis methods. Self-contained gene set analysis methods calculate the gene set scores without considering the genes other than member genes. Thus, in a signal rich dataset like leukemia dataset most of the null gene sets are reported as significant by self-contained gene set analysis method like ROAST.

Figure 13: Comparison of mGSZ, mGSA and mAllez with the program packages with randomized gene sets.

Based on the detection of the relevant gene sets from p53 and gender data by mGSZ, mGSA, mAllez and the program packages, mGSZ is clearly the best method (Figure 14a and 14b). Moreover, mGSA and mAllez shows improved performance as compared to GSA and Allez (Figure 14a and 14b).

mGSZ reports the best p-values (based on resolution) with p53 dataset for the top 50 gene sets as compared to the other methods (Figure 15a). However, in case of gender dataset, mGSZ reports the best p-values for the upper region of the gene list and then slightly lags behind mGSA, mAllez and ROAST in the lower region of the gene list (Figure 15b).
12 Evaluation of mGSZ on dataset with small sample size

In this section, we present the results obtained by implementing mGSZpermutaion and mGSZrotation on the simulated dataset described in Section 2.7 of the main article. In
addition, mGSZpermutation and mGSZrotation were also implemented on a simulated dataset with slightly higher signal achieved by adding the constant 1.5, to see if the results vary with varying signals in datasets. The evaluation was based on comparison of log p-values and identification of the positive gene sets by the two methods. In both the comparisons no significant difference was observed (Figures 16 and 17).

![Figure 16: Asymptotic log p-values for the gene sets in the simulated datasets estimated by mGSZrotation and mGSZpermutation.](image)

(a) Simulated data with lower signal  
(b) Simulated data with higher signal

Figure 16: Asymptotic log p-values for the gene sets in the simulated datasets estimated by mGSZrotation and mGSZpermutation.

![Figure 17: Positive gene sets identified by mGSZrotation and mGSZpermutation. Figures represent cumulative count of positive gene sets over the ranked list of the top 50 gene sets reported by each of the compared methods.](image)

(a) Simulated data with lower signal  
(b) Simulated data with higher signal

Figure 17: Positive gene sets identified by mGSZrotation and mGSZpermutation. Figures represent cumulative count of positive gene sets over the ranked list of the top 50 gene sets reported by each of the compared methods.
## Top gene sets identified by standard Allez and mAllez

Table 2: Top 10 gene sets reported by standard Allez and mAllez from p53 dataset. Standard Allez lacks sample permutation and reports z-score. Empirical and asymptotic p-values for mAllez scores were calculated from 1000 sample permutations.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Z-score</th>
<th>Classes</th>
<th>Empirical p-value</th>
<th>Asymptotic p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P53_UP</td>
<td>5.15</td>
<td>p53Pathway</td>
<td>0.001</td>
<td>0.00009</td>
</tr>
<tr>
<td>HTERT_UP</td>
<td>5.03</td>
<td>hsp27Pathway</td>
<td>0.001</td>
<td>0.0002</td>
</tr>
<tr>
<td>p53Pathway</td>
<td>4.53</td>
<td>p53hypoxiaPathway</td>
<td>0.001</td>
<td>0.0003</td>
</tr>
<tr>
<td>GLUT_UP</td>
<td>4.46</td>
<td>P53_UP</td>
<td>0.001</td>
<td>0.0003</td>
</tr>
<tr>
<td>rasPathway</td>
<td>4.19</td>
<td>rasPathway</td>
<td>0.001</td>
<td>0.0003</td>
</tr>
<tr>
<td>GPCRs_Class_A_Rhodopsin-like</td>
<td>4.01</td>
<td>radiation_sensitivity</td>
<td>0.001</td>
<td>0.0004</td>
</tr>
<tr>
<td>HUMAN_CD34_ENRICHED_TF_JP</td>
<td>3.77</td>
<td>SA_DAG1</td>
<td>0.001</td>
<td>0.0009</td>
</tr>
<tr>
<td>hsp27Pathway</td>
<td>3.72</td>
<td>ngfPathway</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>INSULIN_2F_UP</td>
<td>3.67</td>
<td>il4Pathway</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>LEU_DOWN</td>
<td>3.67</td>
<td>fmlpPathway</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>
14 Top gene sets identified by standard GSA and mGSA

Table 3: Gene sets with p-value less than 0.01 reported by GSA and mGSA from p53 dataset with 1000 sample permutations. GSA calculates empirical p-value, whereas mGSA calculates asymptotic p-value.

<table>
<thead>
<tr>
<th>Classes</th>
<th>GSA Empirical p-value</th>
<th>Classes</th>
<th>mGSA Empirical p-value</th>
<th>Asymptotic p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsp27Pathway</td>
<td></td>
<td>p53Pathway</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>P53hypoxiaPathway</td>
<td>0</td>
<td>hsp27Pathway</td>
<td>0.001</td>
<td>0.0001</td>
</tr>
<tr>
<td>P53Pathway</td>
<td>0</td>
<td>radiation_sensitivity</td>
<td>0.001</td>
<td>0.0003</td>
</tr>
<tr>
<td>SA_G1_AND_S_PHASES</td>
<td>0</td>
<td>P53hypoxiaPathway</td>
<td>0.001</td>
<td>0.0005</td>
</tr>
<tr>
<td>fmlppathway</td>
<td>0</td>
<td>P53_UP</td>
<td>0.001</td>
<td>0.0006</td>
</tr>
<tr>
<td>rasPathway</td>
<td>0</td>
<td>SA_G1_AND_S1_PHASES</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>radiation_sensitivity</td>
<td>0.001</td>
<td>fmlppathway</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>P53_UP</td>
<td>0.001</td>
<td>SA_DAG1</td>
<td>0.002</td>
<td>0.0007</td>
</tr>
<tr>
<td>SA_DAG1</td>
<td>0.002</td>
<td>rasPathway</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>BadPathway</td>
<td>0.003</td>
<td>BadPathway</td>
<td>0.004</td>
<td>0.009</td>
</tr>
<tr>
<td>ngfPathway</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPK_Cascade</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>il4Pathway</td>
<td>0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP00790_Folate_bios</td>
<td>0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15 List of gene sets that are strongly relevant to p53 activity

Top fifty gene sets reported by each of the compared methods were pooled. Each of the pooled gene sets were then analyzed separately for their relevance to p53 activity. A total of forty gene sets highly relevant to p53 activity were selected as relevant gene sets for p53 data.
Table 4: Gene sets and their relevance to p53 gene activity

<table>
<thead>
<tr>
<th>Gene sets</th>
<th>Relavance to P53</th>
<th>Gene sets</th>
<th>Relavance to P53</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53hypoxiaPathway</td>
<td>Hypoxia induces p53 mediated apoptosis ^a</td>
<td>Ceramide Pathway</td>
<td>Apoptosis ^b</td>
</tr>
<tr>
<td>p53Pathway</td>
<td>Pathway ^b</td>
<td>Ca_ nf_at_signaling</td>
<td>Apoptosis ^b</td>
</tr>
<tr>
<td>radiation_sensitivity</td>
<td>Mutation in p53 affects radiation sensitivity ^b</td>
<td>Hiv nef pathway</td>
<td>Apoptosis ^b</td>
</tr>
<tr>
<td>SA_PROGRAMMED_CELL_DEATH</td>
<td>Apoptosis ^b</td>
<td>ras pathway</td>
<td>Regulates cell growth, ^b differentiation and death</td>
</tr>
<tr>
<td>P53_UP</td>
<td>Apoptosis ^b</td>
<td>ngf pathway</td>
<td>Deacetylation of p53 ^b</td>
</tr>
<tr>
<td>hq27Pathway</td>
<td>Mediation/regulation of p53 signalling/activity (O'Callaghan-Sunol et al., 2007)</td>
<td>il4 pathway</td>
<td>p53 suppresses IL-4 ^b</td>
</tr>
<tr>
<td>atmPathway</td>
<td>Pathway member ^b</td>
<td>ST-Interleukin-4-pathway</td>
<td>Cytokins ^b</td>
</tr>
<tr>
<td>p53 signalling</td>
<td>Pathway member ^b</td>
<td>ccr3 pathway</td>
<td>Apoptosis ^b</td>
</tr>
<tr>
<td>chemical pathway</td>
<td>Pathway member ^b</td>
<td>cxc4r4 pathway</td>
<td>Apoptosis ^b</td>
</tr>
<tr>
<td>Cr. death</td>
<td>Pathway member ^b</td>
<td>erk pathway</td>
<td>Apoptosis (Cagnol and Chambard, 2010)</td>
</tr>
<tr>
<td>DNA damage signalling</td>
<td>Pathway member ^b</td>
<td>MAPK cascade</td>
<td>Activated by p53 ^b</td>
</tr>
<tr>
<td>G1 pathway</td>
<td>Pathway member ^b</td>
<td>ck1 pathway</td>
<td>Phosphorylates p53 (Alsheich-Bartok et al., 2008)</td>
</tr>
<tr>
<td>G2 pathway</td>
<td>Pathway member ^b</td>
<td>igf1 pathway</td>
<td>Regulated by p53 (Peng, 2010)</td>
</tr>
<tr>
<td>ST FAS signalling Pathway</td>
<td>Pathway member ^b</td>
<td>bcr pathway</td>
<td>Apoptosis (Kroesen et al., 2001)</td>
</tr>
<tr>
<td>Cell cycle pathway</td>
<td>Pathway member ^b</td>
<td>MAP00790_folate biosynthesis</td>
<td>Folate stress induces p53 mediated apoptosis (Hoeferlin et al., 2013)</td>
</tr>
<tr>
<td>Drug resistance and metabolism</td>
<td>Pathway member ^b</td>
<td>ST_Phosphoinositide 3 kinase</td>
<td>p53 mediated apoptosis (Cooper and Nayak, 2012)</td>
</tr>
<tr>
<td>Breast cancer estrogen signalling</td>
<td>Pathway member ^b</td>
<td>mapk pathway</td>
<td>Interacts with p53 ^b</td>
</tr>
<tr>
<td>BAD Pathway</td>
<td>Apoptosis ^b</td>
<td>SA_G2_AND_M_PHASES</td>
<td>Cell cycle ^b</td>
</tr>
<tr>
<td>Mitochondria pathway</td>
<td>Apoptosis ^b</td>
<td>SA_G1_AND_S_PHASES</td>
<td>Cell cycle ^b</td>
</tr>
<tr>
<td>bel2family and reg. network</td>
<td>Apoptosis ^b</td>
<td>SA_DAG1</td>
<td>Regulates cell growth and apoptosis (Wright and McMaster, 2002)</td>
</tr>
</tbody>
</table>

^aManually curated  
^bManually curated
16 List of gene sets that are strongly relevant to gender

Top twenty gene sets reported by each of the compared methods were pooled. Each of the pooled gene sets were then analyzed separately for their relevance to gender. A total of ten gene sets highly relevant to gender were selected as relevant gene sets for gender data.

<table>
<thead>
<tr>
<th>Gene sets</th>
<th>Relevance to gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>XINACT_MERGED</td>
<td>Female specific  &quot;a&quot;</td>
</tr>
<tr>
<td>TESTIS_GENES_FROM_XHX_AND_NETAFFX</td>
<td>Male specific  &quot;a&quot;</td>
</tr>
<tr>
<td>GNF_FEMALE_GENES</td>
<td>Female specific  &quot;a&quot;</td>
</tr>
<tr>
<td>aifPathway</td>
<td>Triggers cell death in males (Du et al., 2004)</td>
</tr>
<tr>
<td>SIG_Regulation_of_the_actin_cytoskeleton_by_Rho_GTPases</td>
<td>Gender specific regulation  &quot;a&quot;</td>
</tr>
<tr>
<td>tercPathway</td>
<td>Telomere length varies with gender (Barrett and Richardson, 2011)</td>
</tr>
<tr>
<td>NFKB_REduced</td>
<td>Variation with gender (Vina et al., 2011)</td>
</tr>
<tr>
<td>caspase_activity</td>
<td>Variation with gender (Liu et al., 2009)</td>
</tr>
<tr>
<td>MAP00252_Alanine_and_aspartate_metabolism</td>
<td>Variation with gender (Mora et al., 2008)</td>
</tr>
<tr>
<td>MAP00910_Nitrogen_metabolism</td>
<td>Variation with gender (Cheney et al., 1987)</td>
</tr>
</tbody>
</table>

"a"Manually curated

References


