Statistical Methods for Gene Set Analysis of Gene Expression Data

吳漢銘 助理教授
淡江大學 數學系

hmwu@mail.tku.edu.tw
http://www.hmwu.idv.tw

2011/12/13
Content

- Recent Progress

- Finding Differential Expressed Genes
  - t-test and Significance Analysis of Microarrays (SAM)

- Gene Set Analysis (GSA)
  - Gene Set Enrichment Analysis (GSEA)
  - SAM-GS
  - Maxmean

- Null Hypothesis, Statistical Significance Of Gene Set Scores (P-values)

- Gene Set Analysis Tools, Methods Comparison
Incorporating **biological knowledge** into analysis.

- Meta-analysis: **pooling**
- Well-curated publicly **data set**.
- Development of **standardized testing platforms** (e.g., AffyComp).
- **Quality-control** assessment.
- **Gene set** analysis (GSA)
Recent Progress

Ref: Avak Kahvejian, John Quackenbush & John F Thompson, 2008, What would you do if you could sequence everything? Nature Biotechnology 26, 1125 - 1133
Finding Differentially Expressed Genes (DEGs)

More than two samples

Two-sample (independent)

Paired-sample (dependent)

p-values or Statistics

(1) fixed number

(2) above some level
Individual Gene Analysis (IGA)

- IGA evaluates the significance of individual genes between two groups of samples compared and yields a list of altered genes.
- The list is investigated with biologically defined gene sets derived from Gene Ontology or some pathway databases to assess the enrichment of specific biological themes in the list.
**Hypothesis Testing and P-Values**

*Biological Question* → *Statistical Formulation*

**Null Hypothesis:**

- $H_0$: no differential expressed.
- $H_0$: no difference in the mean gene expression in the group tested.
- $H_0$: the gene will have equal means across every group.

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \ldots = \mu_n \]

**P-Values:**

Probability of observing your data under the assumption that the null hypothesis is true.

**Decision Rule:**

Reject $H_0$ if *p-value* is less than alpha. $(p < 0.05$ commonly used$)$. The lower the *p-value*, the more significant.
**t-test Statistics**

**One sample t-test**

\[ H_0 : \mu = \mu_0 \]
\[ H_1 : \mu \neq \mu_0 \text{ (two-tailed).} \]
\[ \mu: \text{ population mean.} \]
\[ \alpha: \text{ significant level (e.g., 0.05).} \]

Test Statistic:

\[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \]

\[ \bar{X}: \text{ sample mean.} \]
\[ S: \text{ sample standard deviation.} \]
\[ n: \text{ number of observations in the sample.} \]

- Reject \( H_0 \) if \(|t_0| > t_{\alpha/2,n-1}\).
- Power = 1 − \( \beta \).
- \((1 − \alpha)100\% \) Confidence Interval for \( \mu \):
  \[ \bar{X} - t_{\alpha/2}S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2}S/\sqrt{n} \]
- \( p\)-value = \( P_{H_0}(|T| > t_0), \quad T \sim t_{n-1} \).

**Two Sample t-test (Unpaired)**

\[ H_0 : \mu_x - \mu_y = \mu_0 \]
\[ H_0 : \mu_x - \mu_y \neq \mu_0 \]
\[ \alpha: \text{ significant level (e.g., 0.05).} \]

Test Statistic:

\[ t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \]

for homogeneous variances:
\[ df = n + m - 2 \]
for heterogeneous variances:
adjusted \( df \)

Reject \( H_0 \) if \(|t_0| > t_{\alpha/2,df}\)
Other t-Statistics for Microarray Data

B-statistic
Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic

\[ t = \frac{\bar{M}}{(a+s)/\sqrt{n}} \]

Penalized t-statistic
Tusher et al (2001, PNAS, SAM)
Efron et al (2001, JASA)

\[ t = \frac{\bar{M}}{(a+s)/\sqrt{n}} \]

General Penalized t-statistic
(Lonnstedt et al 2001)

\[ t = \frac{b}{s^* \times SE} \]

Penalized two-sample t-statistic
\[ t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \text{ where } s^* = \sqrt{a + s^2} \]

Robust General Penalized t-statistic
The Main Problems of IGA

- The final result of IGA is significantly affected by the selected **threshold** (which is normally chosen arbitrarily).
  - **Example:** Pan et al. showed that different choices of the threshold value severely alter the biological conclusions (enrichment of specific function categories in the gene list).

- Many genes with **moderate but meaningful** expression changes are discarded by the strict cutoff value, which leads to a reduction in **statistical power**.

- All the statistical methods applied are based on the **wrong assumption of independent gene** (or gene group) sampling, which increases false positive predictions.
Significance Analysis of Microarrays (SAM)
SAM: Significance Analysis of Microarrays

SAM: Supervised learning software for genomic expression data mining
http://www-stat.stanford.edu/~tibs/SAM/


- SAM has facilities for Gene Set Analysis (GSA).
- GSA uses the "maxmean" statistic.
**SAM:** Significance Analysis of Microarrays

Two class, unpaired data

\[ y_j = 1 \text{ or } 2 \]

\[ r_i = \overline{x}_{i2} - \overline{x}_{i1} \]

Make variation in \( d(i) \) similar across genes of all intensity levels

- **Calculation**
  
  \[ d_i = \frac{r_i}{s_i + s_0} \]

- **Data**
  
  \( x_{ij} \)

- **Response**
  
  \( y_j \)

- **Samples**
  
  \( j = 1, 2, \ldots, n \)

- **Genes**
  
  \( i = 1, 2, \ldots, p \)

- **Standard deviation**
  
  \( s_i \)

- **Exchangeability factor**
  
  \( s_0 \)

- **Order statistics**
  
  \( d_i \):

  \[ d_1, d_2, \ldots, d_p \]

- **Sort**
  
  \( d_{(1)} \), \( d_{(2)} \), \ldots, \( d_{(p)} \)

- **Large positive difference**

- **Large negative difference**
**SAM:**

**Expected Test Statistics**

- **Response:** \( y_j = 1, 1, \ldots, 2, \ldots, 2 \)
- **Permutation:** 1, 2, 1, 2, 1, \ldots, 1
- **\( r_i^* = \bar{x}_{i2} - \bar{x}_{i1} \)**
- **\( d_i^* = \frac{r_i^*}{s_i^* + s_0^*} \)**

- \( \bar{d}_{(1)} = (1/B) \sum_b d_{(i)}^{*b} \)
- Expected order statistics

\[
\begin{align*}
    \bar{d}_{(p)} \\
    \vdots \\
    d_{(2)}^{*b} \\
    d_{(1)}^{*b} \\
    b = 1, 2, \ldots B
\end{align*}
\]
Points for genes with evidence of induction:

\[ d(i) - \bar{d}(i) > \Delta \]

significant positive

\[ d(i) = \bar{d}(i) \]

upper cut-point \( \text{cut}_{up}(\Delta) \)

Points for genes with evidence of repression:

\[ \bar{d}(i) - d(i) > \Delta \]

significant negative

lower cut-point \( \text{cut}_{low}(\Delta) \)
Gene Set Analysis
(GSA)
Gene Sets

- A gene set
  - a group of genes with related functions.
  - sets of genes or pathways, for their association with a phenotype.
  - Examples: metabolic pathway, protein complex, or GO (gene ontology) category.
- Identified from a prior biological knowledge.
- May better reflect the true underlying biology.
- May be more appropriate units for analysis.

M gene sets

<table>
<thead>
<tr>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41bbPathway</td>
<td>TRAF2</td>
<td>MAP3K1</td>
</tr>
<tr>
<td>2</td>
<td>ace2Pathway</td>
<td>COL4A1</td>
<td>COL4A5</td>
</tr>
<tr>
<td>3</td>
<td>acetaminophenPathway</td>
<td>CYP3A</td>
<td>PTGS2</td>
</tr>
<tr>
<td>4</td>
<td>achPathway</td>
<td>TERT</td>
<td>MUSK</td>
</tr>
<tr>
<td>5</td>
<td>actinYPPathway</td>
<td>ABI-2</td>
<td>WASL</td>
</tr>
<tr>
<td>6</td>
<td>agpPathway</td>
<td>GNGT1</td>
<td>PRKACB</td>
</tr>
<tr>
<td>7</td>
<td>ahspPathway</td>
<td>HMBS</td>
<td>ALAS1</td>
</tr>
<tr>
<td>8</td>
<td>allPathway</td>
<td>ADPR</td>
<td>PDCD8</td>
</tr>
<tr>
<td>9</td>
<td>akap13Pathway</td>
<td>PRKACC</td>
<td>PRKAR2A</td>
</tr>
</tbody>
</table>

m1 m2 m3 ...

... ... ...

c2.symbols.gmt
Gene Set Analysis

- Whether some functionally predefined classes of genes are differentially expressed?

- A statistical test to determine significance of a gene class is referred to as gene class testing (GCT) or gene set analysis (GSA).

IGA and GSA

- GSA aims to identify gene sets with ‘subtle but coordinated’ expression changes that cannot be detected by IGA methods.
  - even weak expression changes in individual genes gathered to a large gene set can show a significant pattern.
- Results of GSA are not affected by arbitrarily chosen cutoffs.
- GSA does not provide information as detailed as IGA.

Source: Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, Briefings In Bioinformatics. 2008, 9(3). 189-197.
Several benefits of GSA

**From a statistical point of view:**
GSA typically increases power and reduces the dimensionality of the underlying statistical problem.

**From the biological perspective:**
GSA help to understand the functional mechanism in a cell.
(1) is a certain pathway activated in a given tissue under some treatment x?
(2) is the pathway more active than other pathway?
Literature Review

- Global Test (global model with random effects): Goeman et al., 2004
- ANCOVA Global Test: Mansmann and Meister, 2005
- GSEA: Subramanian et al., 2005
- Principal component analysis (PCA): Kong et al., 2006
- Significance analysis of microarray for gene sets (SAM-GS): Dinu et al., 2007
- Gene list analysis with prediction accuracy (GLAPA): Maglietta et al., 2007
- Maxmean: Efron and Tibshirani, 2007
- exSAM-GS: Adewale et al. 2008
- Multivariate analysis of variance test (MANOVA, modified Hotelling’s T2): Tsai and Chen, 2009
- Linear combination Test (LCT): Wang, Dinu, Liu and Yasui, 2011

Gene Set Enrichment Analysis (GSEA)

GSEA (Subramanian et al., PNAS, 2005)
GSEA: Gene Set Enrichment Analysis

- GSEA was introduced by Mootha et al. 2003, and was used to identify pre-defined gene sets which exhibited significant differences in expression between samples from normal and diabetic patients.
- The methodology was subsequently refined by Subramanian et al. 2005.


Step 1: Enrichment Score (ES)

Expression data set  Ranked gene list  \( N_H \) genes

\[
SNR = \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}
\]

Evaluate the fraction of genes in \( S \) ("hits") weighted by their correlation and the fraction of genes not in \( S \) ("misses") present up to a given position \( i \) in \( L \).

\[
P_{\text{hit}}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R} \quad N_R = \sum \frac{|r_j|^p}{g_j \in S}
\]

\[
P_{\text{miss}}(S, i) = \sum_{g_j \notin S, j \leq i} \frac{1}{N - N_H}
\]

\[
ES(S) = \max_i \left\{ P_{\text{hit}}(S, i) - P_{\text{miss}}(S, i) \right\}
\]

\( ES(S) > 0 \): gene set enrichment at the top of the ranked list.

\( ES(S) < 0 \): gene set enrichment at the bottom of the ranked list.
Enrichment Plot

\[ ES(S) = \max_i \{ P_{\text{hit}}(S, i) - P_{\text{miss}}(S, i) \} \]

- If \( p=0 \)
  \[ ES(S) = \text{Kolmogorov-Smirnov statistic} \]
- Set \( p=1 \).

- For a randomly distributed \( S \), \( ES(S) \) will be relatively small.
- It is concentrated at the top or bottom of the list, or nonrandomly distributed, then \( ES(S) \) will be corresponding high.

Subramanian et al., PNAS 102(43), 15545–15550 (2005).
Step 2: Estimating Significance

Assess the significance of an observed ES by comparing it with the set of score $ES(null)$ computed with randomly assigned phenotype.

$$SNR = \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

- For positive ES
- For negative ES

$$ES^{(b)}(S), b = 1, \cdots, 1000$$

$$p-value \approx \frac{\#\{ES^{(b)} > ES_{obs}\}}{\#\text{permutations}}$$
**Step 3: Multiple Hypothesis Testing**

**X: false positive gene**

\[
P(X \geq 1) = 1 - P(X = 0) = 1 - 0.95^n
\]

<table>
<thead>
<tr>
<th>Number of genes tested (N)</th>
<th>False positives incidence</th>
<th>Probability of calling 1 or more false positives by chance (100(1-0.95^N))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/20</td>
<td>5%</td>
</tr>
<tr>
<td>2</td>
<td>1/10</td>
<td>10%</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>64%</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>99.4%</td>
</tr>
</tbody>
</table>
Step 3: Multiple Hypothesis Testing

- When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing.
  - Normalize ES for each gene set to account for the size of the set (NES).
  - Control the proportion of false positives by calculating the false discovery rate (FDR) corresponding to each NES.

- FDR
  - It is the estimated probability that a set with a given NES represents a false positive finding.
  - It is computed by comparing the tails of the observed and null distributions for the NES.
GSEA Software

Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- Download the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- Explore the Molecular Signatures Database (MSigDB), a collection of annotated gene sets for use with GSEA software.
- View documentation describing GSEA.

What's New

29-Mar-2011: Version 3.7 of this web site was released. We updated the text for improved clarity, and ARKv3.7 was released with new features and bug fixes.

09-Sep-2010: We are pleased to announce the release of a new, more expanded version of the MSigDB database. This updated version includes additional gene sets in different species and biological processes. New gene sets have been added for species other than human.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 23 no. 23 2007, pages 3251–3253
doi:10.1093/bioinformatics/btm369

Gene expression

GSEA-P: a desktop application for Gene Set Enrichment Analysis

Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo and Jill P. Mesirov*

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Received on June 18, 2007; revised on June 18, 2007; accepted on July 9, 2007

Advance Access publication July 20, 2007

Associate Editor: Olga Troyanskaya

*Correspondence to: Jill P. Mesirov

Published by National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.
**Downloads (register first!)**


**Quick Tour:** [http://www.broadinstitute.org/gsea/doc/desktop_tutorial.jsp](http://www.broadinstitute.org/gsea/doc/desktop_tutorial.jsp)

---

**Downloads**

The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the GSEA/MSigDB license for more details.

**Software**

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 1.6 or higher. If your computer has Java 1.5 and cannot upgrade to Java 1.6, please see the FAQ.

<table>
<thead>
<tr>
<th>Software</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JavaGSEA Desktop Application</strong></td>
<td>Easy-to-use graphical user interface</td>
</tr>
<tr>
<td></td>
<td>Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 1.6+</td>
</tr>
<tr>
<td></td>
<td>Produces richly annotated reports of enrichment results</td>
</tr>
<tr>
<td></td>
<td>Integrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms</td>
</tr>
<tr>
<td></td>
<td>The GSEA team suggests always starting GSEA by using these Launch buttons, or by clicking the icon that the application installs on your desktop, in order to ensure optimal memory allocation</td>
</tr>
<tr>
<td><strong>javaGSEA Java Jar file</strong></td>
<td>Command line usage</td>
</tr>
<tr>
<td></td>
<td>Runs on any platform that supports Java 1.6+</td>
</tr>
<tr>
<td></td>
<td>We recommend using the 'Launch' buttons above instead of this mode for most users</td>
</tr>
<tr>
<td><strong>GSEA Java Source Code</strong></td>
<td>100% Java implementation of GSEA</td>
</tr>
<tr>
<td><strong>Java source files</strong></td>
<td>Incorporate GSEA into your own data analysis pipeline</td>
</tr>
<tr>
<td></td>
<td>Programmatically call the open source GSEA java API</td>
</tr>
<tr>
<td><strong>R-GSEA R Script</strong></td>
<td>Usage from within the R programming environment</td>
</tr>
<tr>
<td></td>
<td>Easily inspect, learn and tweak the algorithm</td>
</tr>
<tr>
<td></td>
<td>Incorporate GSEA into your own data analysis pipeline</td>
</tr>
<tr>
<td></td>
<td>Programmatically call the open source GSEA R API</td>
</tr>
<tr>
<td></td>
<td>Click here to learn more about the R-GSEA script</td>
</tr>
<tr>
<td><strong>GenePattern GSEA Module</strong></td>
<td>Use GSEA from within GenePattern</td>
</tr>
<tr>
<td></td>
<td>Use GSEA in concert with a large suite of other analytics found in GenePattern (a powerful and flexible analysis platform developed at the Broad Institute)</td>
</tr>
</tbody>
</table>

Download

- [GSEA-P-R-1.0.zip](http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html)
- GenePattern site
Molecular Signatures database (MsigDB)

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this website, you can:

- Search for gene sets by keyword.
- Browse gene sets by name or collection.
- Examine a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- Download gene sets.
- Investigate gene sets:
  - Compute overlaps between your gene set and gene sets in MSigDB.
  - Categorize members of a gene set by gene families.
  - View the expression profile of a gene set in any of the three provided public expression compendia.

Collections

The MSigDB gene sets are divided into five major collections:

- **c1** positional gene sets for each human chromosome and each cytogenetic band.

- **c2** curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

- **c3** motif gene sets based on conserved co-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

- **c4** computational gene sets defined by expression neighborhoods centered on 380 cancer-associated genes.

- **c5** GO gene sets consist of genes annotated by the same GO terms.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

...
## Example Datasets

<table>
<thead>
<tr>
<th>DATASET</th>
<th>DESCRIPTION</th>
<th>RELEVANT DATA (save link to download)</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Transcriptional profiles from male and female lymphoblastoid cell lines. Results of C1 GSEA analysis of this dataset. Results of C2 GSEA analysis of this dataset.</td>
<td>Gender_hgu133a.gct GenderCollapsed.gct Gender.cls</td>
<td>Unpublished</td>
</tr>
<tr>
<td>p53</td>
<td>Transcriptional profiles from p53+ and p53 mutant cancer cell lines. Results of C2 GSEA analysis of this dataset.</td>
<td>p53_hgu95av2.gct p53Collapsed.gct p53.cls</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Gene sets</td>
<td>Archived gene sets from the GSEA PNAS 2005 publication. <strong>Note</strong>: This collection of gene sets is not the latest version, so when beginning a new analysis you might want to download the current collection of gene sets from the Downloads page.</td>
<td>C1_symbols.gmt (positional) C2_symbols.gmt (curated)</td>
<td>Subramaniam and Tamayo PNAS 2005</td>
</tr>
</tbody>
</table>
P53 Status in Cancer Cell Lines

- **NCI-60 collection of cancer cell lines.**
  - Past usage: to identify targets of the transcription factor p53, which regulates gene expression in response to various signals of cellular stress.
  - The mutational status of the p53 gene has been reported for 50 of the NCI-60 cell lines: 17 normal, and 33 mutations.

**GSEA:** to identify functional gene sets (C2) correlated with p53 status.
- (p53+ > p53-): five gene sets.

**LES:** (p53- > p53+) whether three gene sets reflect a common biological function.
- resulting 16, 11, 13 genes.
- 4 gene in common: MAPK pathway.

<table>
<thead>
<tr>
<th>Gene set</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set: p53 status in NCI-60 cell lines</td>
<td></td>
</tr>
<tr>
<td>Enriched in p53 mutant</td>
<td></td>
</tr>
<tr>
<td>Ras signaling pathway</td>
<td>0.171</td>
</tr>
<tr>
<td>Enriched in p53 wild type</td>
<td></td>
</tr>
<tr>
<td>Hypoxia and p53 in the cardiovascular system</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stress induction of HSP regulation</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p53 signaling pathway</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>p53 up-regulated genes</td>
<td>0.013</td>
</tr>
<tr>
<td>Radiation sensitivity genes</td>
<td>0.078</td>
</tr>
</tbody>
</table>

![Fig. 3. Leading edge overlap for p53 study. This plot shows the ras, ngf, and igf1 gene sets correlated with p53^{-} clustered by their leading-edge subsets indicated in dark blue. A common subgroup of genes, apparent as a dark vertical stripe, consists of MAP2K1, PIK3CA, ELK1, and RAF1 and represents a subsection of the MAPK pathway.](image)
### Demo Dataset: Transcriptional profiles from p53+ and p53 mutant cancer cell lines

<table>
<thead>
<tr>
<th>Data File</th>
<th>Content</th>
<th>Format</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression dataset</td>
<td>Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).</td>
<td>res, gct, pcl, or txt</td>
<td>You create the file.</td>
</tr>
<tr>
<td>Phenotype labels</td>
<td>Contains phenotype labels and associates each sample with a phenotype.</td>
<td>cls</td>
<td>You create the file or have GSEA create it for you.</td>
</tr>
</tbody>
</table>

**P53_hgu95av2.gct**

<table>
<thead>
<tr>
<th>#.1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12625</td>
<td>50</td>
<td>786-0</td>
<td>BT-549</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>NAME</td>
<td>Description</td>
<td>786-0</td>
<td>BT-549</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>100_g_at</td>
<td>na</td>
<td>215.37</td>
<td>132.94</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1000_at</td>
<td>na</td>
<td>328.68</td>
<td>234.31</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1001_at</td>
<td>na</td>
<td>39.64</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1002_f_at</td>
<td>na</td>
<td>18.46</td>
<td>12.14</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1003_s_at</td>
<td>na</td>
<td>60.83</td>
<td>30.19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1004_at</td>
<td>na</td>
<td>68.02</td>
<td>54.41</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1005_at</td>
<td>na</td>
<td>610.35</td>
<td>65.93</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1006_at</td>
<td>na</td>
<td>12.79</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1007_s_at</td>
<td>na</td>
<td>354.92</td>
<td>208.33</td>
<td></td>
</tr>
</tbody>
</table>

**P53Collapsed_symbols.gct**

<table>
<thead>
<tr>
<th>#.1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10100</td>
<td>50</td>
<td>786-0</td>
<td>BT-549</td>
<td>C</td>
<td>CCRF-CEM</td>
</tr>
<tr>
<td>2</td>
<td>NAME</td>
<td>DESCRIPTION</td>
<td>786-0</td>
<td>BT-549</td>
<td>C</td>
<td>CCRF-CEM</td>
</tr>
<tr>
<td>3</td>
<td>TACC2</td>
<td>na</td>
<td>46.05</td>
<td>82.17</td>
<td>16.87</td>
<td>98.6</td>
</tr>
<tr>
<td>4</td>
<td>C14orf132</td>
<td>na</td>
<td>108.34</td>
<td>59.04</td>
<td>25.61</td>
<td>33.11</td>
</tr>
<tr>
<td>5</td>
<td>AGER</td>
<td>na</td>
<td>42.2</td>
<td>25.75</td>
<td>76.01</td>
<td>40.41</td>
</tr>
<tr>
<td>6</td>
<td>32385_at</td>
<td>na</td>
<td>7.43</td>
<td>13.94</td>
<td>8.55</td>
<td>21.13</td>
</tr>
<tr>
<td>7</td>
<td>RBM17</td>
<td>na</td>
<td>11.4</td>
<td>3</td>
<td>3.16</td>
<td>2.34</td>
</tr>
<tr>
<td>8</td>
<td>DYT1</td>
<td>na</td>
<td>148.09</td>
<td>317.17</td>
<td>316.66</td>
<td>147.23</td>
</tr>
<tr>
<td>9</td>
<td>CORO1A</td>
<td>na</td>
<td>8.62</td>
<td>9.12</td>
<td>1672.53</td>
<td>5.91</td>
</tr>
<tr>
<td>10</td>
<td>WT1</td>
<td>na</td>
<td>206.74</td>
<td>136.71</td>
<td>141.34</td>
<td>129.09</td>
</tr>
</tbody>
</table>

**P53.cls**

<table>
<thead>
<tr>
<th>#.1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>INDEX</th>
<th>MUT</th>
<th>WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MUT</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td>MUT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Input for GSEA (2)

<table>
<thead>
<tr>
<th>Data File</th>
<th>Content</th>
<th>Format</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene sets</td>
<td>Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.</td>
<td>gmx or gmt</td>
<td>You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your own gene sets file.</td>
</tr>
<tr>
<td>Chip annotations</td>
<td>Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.</td>
<td>Chip</td>
<td>You use the files on the Broad ftp site, download the files from the GSEA website, or create your own chip file.</td>
</tr>
</tbody>
</table>

### c1.symbols.gmt

<table>
<thead>
<tr>
<th>chr10q24</th>
<th>Cytogenetic band</th>
<th>PITX3</th>
<th>SPFH1</th>
<th>NEURL</th>
<th>C10orf12</th>
<th>NDUFA1</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5q23</td>
<td>Cytogenetic band</td>
<td>ALDH7A1</td>
<td>IL13</td>
<td>8-Sep</td>
<td>IRF1</td>
<td>ACSL6</td>
</tr>
<tr>
<td>chr8q24</td>
<td>Cytogenetic band</td>
<td>HAS2</td>
<td>LRRC14</td>
<td>TSTA3</td>
<td>DGAT1</td>
<td></td>
</tr>
<tr>
<td>chr16q24</td>
<td>Cytogenetic band</td>
<td>RPL13</td>
<td>GALNS</td>
<td>FANCA</td>
<td>CPNE7</td>
<td></td>
</tr>
<tr>
<td>chr13q14</td>
<td>Cytogenetic band</td>
<td>AKAP11</td>
<td>ARL11</td>
<td>ATP7B</td>
<td>C13orf1</td>
<td></td>
</tr>
<tr>
<td>chr7p21</td>
<td>Cytogenetic band</td>
<td>ARL4A</td>
<td>SCIN</td>
<td>GLCCI1</td>
<td>SP8</td>
<td>SOST1</td>
</tr>
<tr>
<td>chr10q23</td>
<td>Cytogenetic band</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### c2.symbols.gmt

<table>
<thead>
<tr>
<th>41bbPathway</th>
<th>TNF-type receptor 4-1BB iso IL2</th>
<th>TRAF2</th>
<th>MAP3K1</th>
<th>IFNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace2Pathway</td>
<td>Angiotensin-converting enz COL4A3</td>
<td>COL4A1</td>
<td>COL4A5</td>
<td>AGT</td>
</tr>
<tr>
<td>acetaminophenPathway</td>
<td>Acetaminophen selectively</td>
<td>CYP3A7</td>
<td>PTGS2</td>
<td>CYP1A2</td>
</tr>
<tr>
<td>achPathway</td>
<td>Nicotinic acetylcholine rece RAPSN</td>
<td>TERT</td>
<td>MUSK</td>
<td>PTK2</td>
</tr>
<tr>
<td>actinYPPathway</td>
<td>The Arp 2/3 complex localiz ACTR3</td>
<td>ABI-2</td>
<td>WASL</td>
<td>ARPC4</td>
</tr>
<tr>
<td>agpcrPathway</td>
<td>G-protein coupled receptor PRKAR2A</td>
<td>GNGT1</td>
<td>PRKACB</td>
<td>PRKCB1</td>
</tr>
<tr>
<td>ahsPPathway</td>
<td>Alpha-hemoglobin stabilizin CPO</td>
<td>HMBS</td>
<td>ALAS1</td>
<td>ERAF</td>
</tr>
<tr>
<td>alfPathway</td>
<td>BLACK</td>
<td>ADPRT</td>
<td>PDCD8</td>
<td>BCL2L1</td>
</tr>
<tr>
<td>a kinase anchor protein 13 EDG4</td>
<td>PRKACC</td>
<td>PRKAR2A</td>
<td>PRKACB</td>
<td></td>
</tr>
</tbody>
</table>
Launch GSEA

Steps in GSEA

1. What you need for GSEA:
   - Expression dataset
   - Phenotype file
   - Gene sets (from MSigDB or your own gene sets)

2. Run GSEA
   - Start with default parameters
   - If you want to collapse probes to genes, specify chip platform

3. View results & leading edge
   - Leading edge finds genes driving enrichment results

Gene Sets Browser

- Browse gene sets in MSigDB
- Search the database of ~2500 gene sets
- Chip2Chip converts gene sets between platforms
- Export gene sets for analysis with GSEA or with other programs

Getting Help

- GSEA website: www.broad.mit.edu/gsea
- GSEA Wiki: www.broad.mit.edu/gsea/wiki
- Email the GSEA team at gsea@broad.mit.edu

GSEA v2.07 (Gene set enrichment analysis -- Broad Institute)
Load Data

1. Load data
2. Method 1: Import data into the application
   - Browse for files...
3. Recent folders (double-click to list content)
Run GSEA

1. Load data
2. Run GSEA
3. Set parameters and launch analysis tools
4. Result status
GSEA Report for Dataset P53_hgu95av2

Enrichment in phenotype: MUT (33 samples)
- 71 / 176 gene sets are upregulated in phenotype MUT
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal p-value < 1%
- 4 gene sets are significantly enriched at nominal p-value < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: WT (17 samples)
- 105 / 176 gene sets are upregulated in phenotype WT
- 15 gene sets are significantly enriched at FDR < 25%
- 15 gene sets are significantly enriched at nominal p-value < 1%
- 15 gene sets are significantly enriched at nominal p-value < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details
- The dataset has 12625 native features
- After collapsing features into gene symbols, there are: 9096 genes

Gene set details
- Gene set size filters (min=15, max=500) resulted in filtering out 143 / 319 gene sets
- The remaining 176 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the MUT versus WT comparison
- The dataset has 9096 features (genes)
- # of markers for phenotype MUT: 4076 (44.8%) with correlation area 42.2%
- # of markers for phenotype WT: 5020 (55.2%) with correlation area 57.8%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Butterfly plot of significant genes

Global statistics and plots
- Plot of p-values vs. NES
- Global ES histogram

Other
- Parameters used for this analysis
Interpretation

Enrichment in phenotype: MUT (33 samples)

- 71 / 176 gene sets are upregulated in phenotype MUT
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 5%

Table: Snapshot of enrichment results

- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results
### Enrichment plot

**Table: GSEA Results Summary**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>P53_hgu95av2_collapsed_to_symbols.P53.cls#MUT-versus_WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>P53.cls#MUT-versus_WT</td>
</tr>
<tr>
<td>Upregulated in class</td>
<td>MUT</td>
</tr>
<tr>
<td>GeneSet</td>
<td>CHR11Q13</td>
</tr>
<tr>
<td>Enrichment Score (ES)</td>
<td>0.45963296</td>
</tr>
<tr>
<td>Normalized Enrichment Score (NES)</td>
<td>1.6873256</td>
</tr>
<tr>
<td>Nominal p-value</td>
<td>0.0</td>
</tr>
<tr>
<td>FDR q-value</td>
<td>1.0</td>
</tr>
<tr>
<td>FWER p-Value</td>
<td>0.6666667</td>
</tr>
</tbody>
</table>

**Enrichment plot: CHR11Q13**

- **Enrichment Score**
- **Leading Edge subset**

**Graphs:**
- **Enrichment plot: CHR12Q13**
- **Enrichment plot: CHR11Q13**
Table: GSEA details

<table>
<thead>
<tr>
<th>PROBE</th>
<th>GENE SYMBOL</th>
<th>GENE_TITLE</th>
<th>RANK IN GENE LIST</th>
<th>RANK METRIC SCORE</th>
<th>RUNNING ES</th>
<th>CORE ENRICHMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CFL1</td>
<td>CFL1 Entrez, Source</td>
<td>cofilin 1 (non-muscle)</td>
<td>22</td>
<td>0.429</td>
<td>0.0258</td>
<td>Yes</td>
</tr>
<tr>
<td>2 SF3B2</td>
<td>SF3B2 Entrez, Source</td>
<td>splicing factor 3b, subunit 2, 145KDa</td>
<td>34</td>
<td>0.408</td>
<td>0.0515</td>
<td>Yes</td>
</tr>
<tr>
<td>3 MRPL49</td>
<td>MRPL49 Entrez, Source</td>
<td>mitochondrial ribosomal protein L49</td>
<td>42</td>
<td>0.390</td>
<td>0.0765</td>
<td>Yes</td>
</tr>
<tr>
<td>4 RELA</td>
<td>RELA Entrez, Source</td>
<td>v-rel reticuloendotheliosis viral oncogene homolog polypeptide gene enhancer in B-cells 3, p65 (avia</td>
<td>48</td>
<td>0.384</td>
<td>0.1012</td>
<td>Yes</td>
</tr>
<tr>
<td>5 PPP2R5B</td>
<td>PPP2R5B Entrez, Source</td>
<td>protein phosphatase 2, regulatory subunit B (B56</td>
<td>65</td>
<td>0.372</td>
<td>0.1239</td>
<td>Yes</td>
</tr>
<tr>
<td>6 HTATIP</td>
<td>HTATIP Entrez, Source</td>
<td>HIV-1 Tat interacting protein, 60KDa</td>
<td>91</td>
<td>0.356</td>
<td>0.1446</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PROBE</th>
<th>GENE SYMBOL</th>
<th>GENE_TITLE</th>
<th>RANK IN GENE LIST</th>
<th>RANK METRIC SCORE</th>
<th>RUNNING ES</th>
<th>CORE ENRICHMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>105 NAALDL1</td>
<td>NAALDL1 Entrez, Source</td>
<td>N-acetylated alpha-linked acidic dipeptidase-like</td>
<td>8011</td>
<td>-0.221</td>
<td>0.0637</td>
<td>No</td>
</tr>
<tr>
<td>106 FLRT1</td>
<td>FLRT1 Entrez, Source</td>
<td>fibronectin leucine rich transmembrane protein 1</td>
<td>8306</td>
<td>-0.246</td>
<td>0.0472</td>
<td>No</td>
</tr>
<tr>
<td>107 PDE2A</td>
<td>PDE2A Entrez, Source</td>
<td>phosphodiesterase 2A, cGMP-stimulated</td>
<td>8419</td>
<td>-0.258</td>
<td>0.0518</td>
<td>No</td>
</tr>
<tr>
<td>108 FOLR3</td>
<td>FOLR3 Entrez, Source</td>
<td>folate receptor 3 (gamma)</td>
<td>8924</td>
<td>-0.354</td>
<td>0.0190</td>
<td>No</td>
</tr>
</tbody>
</table>
Heat Map for Hits

CHR11Q13 : Blue-Pink O' Gram in the Space of the Analyzed GeneSet
Gene Set Null Distribution of ES

CHR11Q13: Random ES distribution.
Gene set null distribution of ES for CHR11Q13
## Detailed Enrichment Results

### Enrichment in phenotype: MUT (33 samples)

- 71 / 176 gene sets are upregulated in phenotype MUT
- 0 gene sets are significant at FDR < 25%
- 4 gene sets are significantly enriched at nominal p-value < 1%
- 4 gene sets are significantly enriched at nominal p-value < 5%
- **Snapshot of enrichment results**
  - Detailed enrichment results in html format
  - Detailed enrichment results in excel format (tab delimited text)
  - Guide to interpret results

### Table: Gene sets enriched in phenotype MUT (33 samples)

<table>
<thead>
<tr>
<th>GS follow link to MSigDB</th>
<th>GS DETAILS</th>
<th>SIZE</th>
<th>ES</th>
<th>NES</th>
<th>NOM p-val</th>
<th>FDR q-val</th>
<th>FWER p-val</th>
<th>RANK AT MAX</th>
<th>LEADING EDGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHR11Q13</td>
<td>Details...</td>
<td>108</td>
<td>0.46</td>
<td>1.69</td>
<td>0.000</td>
<td>1.000</td>
<td>0.667</td>
<td>2479</td>
<td>tags=53%, list=27%, signal=72%</td>
</tr>
<tr>
<td>CHR7P21</td>
<td>Details...</td>
<td>16</td>
<td>0.64</td>
<td>1.66</td>
<td>0.000</td>
<td>0.717</td>
<td>0.800</td>
<td>979</td>
<td>tags=50%, list=11%, signal=56%</td>
</tr>
<tr>
<td>CHRXP11</td>
<td>Details...</td>
<td>66</td>
<td>0.53</td>
<td>1.62</td>
<td>0.182</td>
<td>0.664</td>
<td>0.833</td>
<td>1909</td>
<td>tags=55%, list=21%, signal=69%</td>
</tr>
<tr>
<td>CHR5Q14</td>
<td>Details...</td>
<td>20</td>
<td>0.55</td>
<td>1.62</td>
<td>0.077</td>
<td>0.525</td>
<td>0.833</td>
<td>535</td>
<td>tags=30%, list=6%, signal=32%</td>
</tr>
<tr>
<td>CHR10Q22</td>
<td>Details...</td>
<td>33</td>
<td>0.53</td>
<td>1.57</td>
<td>0.000</td>
<td>0.602</td>
<td>0.933</td>
<td>1649</td>
<td>tags=55%, list=18%, signal=66%</td>
</tr>
<tr>
<td>CHR1P22</td>
<td>Details...</td>
<td>22</td>
<td>0.61</td>
<td>1.46</td>
<td>0.000</td>
<td>0.996</td>
<td>0.967</td>
<td>2510</td>
<td>tags=77%, list=28%, signal=106%</td>
</tr>
<tr>
<td>CHR1P36</td>
<td>Details...</td>
<td>117</td>
<td>0.41</td>
<td>1.42</td>
<td>0.059</td>
<td>1.000</td>
<td>1.000</td>
<td>1852</td>
<td>tags=44%, list=20%, signal=54%</td>
</tr>
<tr>
<td>CHR3P14</td>
<td>Details...</td>
<td>10</td>
<td>0.24</td>
<td>0.82</td>
<td>0.037</td>
<td>1.000</td>
<td>1.000</td>
<td>2590</td>
<td>tags=59%, list=20%, signal=53%</td>
</tr>
<tr>
<td>CHR8P21</td>
<td>Details...</td>
<td>138</td>
<td>0.19</td>
<td>0.56</td>
<td>0.867</td>
<td>0.999</td>
<td>1.000</td>
<td>1718</td>
<td>tags=25%, list=19%, signal=30%</td>
</tr>
<tr>
<td>CHR4Q31</td>
<td>Details...</td>
<td>24</td>
<td>0.18</td>
<td>0.54</td>
<td>1.000</td>
<td>0.994</td>
<td>1.000</td>
<td>2516</td>
<td>tags=38%, list=28%, signal=52%</td>
</tr>
<tr>
<td>CHRXQ22</td>
<td>Details...</td>
<td>21</td>
<td>0.21</td>
<td>0.52</td>
<td>1.000</td>
<td>0.988</td>
<td>1.000</td>
<td>3222</td>
<td>tags=48%, list=35%, signal=74%</td>
</tr>
<tr>
<td>CHR9Q22</td>
<td>Details...</td>
<td>32</td>
<td>0.15</td>
<td>0.48</td>
<td>1.000</td>
<td>0.988</td>
<td>1.000</td>
<td>1821</td>
<td>tags=22%, list=20%, signal=27%</td>
</tr>
</tbody>
</table>
Gene Markers for the MUT versus WT Comparison

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 143 / 319 gene sets
- The remaining 176 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the MUT versus WT comparison

- The dataset has 9096 features (genes)
- # of markers for phenotype MUT: 4076 (44.8%) with correlation area 42.2%
- # of markers for phenotype WT: 5020 (55.2%) with correlation area 57.8%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Butterfly plot of significant genes

Global statistics and plots

- Plot of p-values vs NES
- Global ES histogram

Other

- Parameters used for this analysis

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME</td>
<td>DESCRIPTION</td>
<td>GENE_SYMBOL</td>
<td>GENE_TITLE</td>
<td>SCORE</td>
</tr>
<tr>
<td>NADK</td>
<td>null</td>
<td>NADK</td>
<td>NAD kinase</td>
<td>0.63814014</td>
</tr>
<tr>
<td>RP2</td>
<td>null</td>
<td>RP2</td>
<td>retinitis pigmentos</td>
<td>0.55928165</td>
</tr>
<tr>
<td>GSPM2</td>
<td>null</td>
<td>GSPM2</td>
<td>G-protein signaling</td>
<td>0.5350833</td>
</tr>
<tr>
<td>TUSC4</td>
<td>null</td>
<td>TUSC4</td>
<td>tumor suppressor</td>
<td>0.5116475</td>
</tr>
<tr>
<td>SLC16A2</td>
<td>null</td>
<td>SLC16A2</td>
<td>solute carrier family</td>
<td>0.48800114</td>
</tr>
<tr>
<td>HS2ST1</td>
<td>null</td>
<td>HS2ST1</td>
<td>heparan sulfate 2-</td>
<td>0.4871485</td>
</tr>
<tr>
<td>FTSJ1</td>
<td>null</td>
<td>FTSJ1</td>
<td>FtsJ homolog 1 (E)</td>
<td>0.47524673</td>
</tr>
<tr>
<td>EXPH5</td>
<td>null</td>
<td>EXPH5</td>
<td>exophilin 5</td>
<td>0.46191633</td>
</tr>
<tr>
<td>LRRC42</td>
<td>null</td>
<td>LRRC42</td>
<td>leucine rich repeat</td>
<td>0.45618612</td>
</tr>
<tr>
<td>PEX14</td>
<td>null</td>
<td>PEX14</td>
<td>peroxisomal biogene</td>
<td>0.4566304</td>
</tr>
<tr>
<td>VAPA</td>
<td>null</td>
<td>VAPA</td>
<td>VAMP (vesicle-as)</td>
<td>0.4549476</td>
</tr>
</tbody>
</table>
Heat Map and Gene Correlation

Heat Map of the top 50 features for each phenotype in P53_hgu95av2_collapsed_to_symbols
Global Statistics and Plots

Global statistics and plots

- Plot of p-values vs. NES
- Global ES histogram

Plot of p-values vs. NES

Global ES histogram
Running the Leading Edge Analysis

c2.symbols.gmt
Results

Heat Map:
the leading edge subsets
(the subset of members that contribute most to the ES)

Set-to-Set:
the overlap between subsets:
the darker the color, the greater the overlap.

Jacquard:
the intersection(LES pair)/union(LES pair)

Gene in Subsets:
the number of subsets in which each gene appears

Number of occurrences:
the number of LES pairs in a particular bin.
Gene Set Analysis:
SAM-GS, Maxmean
SAM-GS

SAM-GS: Significance Analysis of Microarray for Gene Sets (Liu et al, 2007)

\[ \text{SAM-GS} = \sum_{g_j \in S} d_j^2. \]

Significance is assessed by permuting sample labels.

- Qi Liu Irina Dinu Yutaka Yasui, SAM-GS Excel Add-in, version 1.0.2, May 23, 2007
4. Computational issues and software. The developments in the previous two sections lead to our Gene Set Analysis procedure, which we summarize here:

1. Compute a summary statistic \( z_i \) for each gene, for example, the two-sample \( t \)-statistic for two-class data. Let \( z_{\delta} \) be the vector of \( z_i \) values for genes in a gene-set \( \delta \).

2. For each gene-set \( \delta \), choose a summary statistic \( S = s(z) \): choices include the average of \( z_i \) or \( |z_i| \) for genes in \( \delta \), the GSEA statistic or, our recommended choice, the maxmean statistic (3.11).

3. Standardize \( S \) by its randomization mean and standard deviation as in (2.14): \( S' = (S - \text{mean}_S)/\text{stdev}_S \). For summary statistics such as the mean, mean absolute value or maxmean (but not GSEA), this can be computed from the gene-wide means and standard deviations, without having to draw random sets of genes. Note formula (3.13) for the maxmean statistic.

4. Compute permutations of the outcome values (e.g., the class labels in the two-class case) and recompute \( S' \) on each permuted dataset, yielding permutation values \( S'^*1, S'^*2, \ldots, S'^*B \).
Null Hypothesis,
Statistical Significance Of Gene Set Scores (P-values)
**Null Hypothesis of GSA: Q1**

**Q1: competitive null hypothesis**

- Compares the association between the genes in the set and the phenotype is compared with the association of the remainder of the genes and the phenotype.
- **Genes** are the sampling units and the association between the samples and the phenotypes is fixed.

![Diagram showing gene set and phenotype associations](image)
Null Hypothesis of GSA: Q2

**Q2: self-contained null hypothesis**

- Compares the association of the gene set and the phenotype with that of random phenotypes.
- Phenotypes (Samples) are the sampling units while the gene set membership is fixed.

Tian et al (PNAS, 2005) defined Q1, Q2
Geoman and Buhlmann (Bioinformatics, 2007) defined competitive and self-contained.
Null Hypothesis of GSA: Q3

Q3: the "nested null hypothesis"
- Where differential expression of the genes in the gene set is compared to the differential expression of all genes under consideration (both inside and outside the gene set).
- H0: none of the gene sets considered is associated with the phenotype.

The restandardization strategy: mixes the two null hypotheses Q1 and Q2 which may lead to difficulties in the interpretation of the resulting p-values.
One suggests that both gene and sample randomizations should be used because they test two different but complementary null hypotheses.

The other insists that only sample randomization should be used to avoid inherent problems of the gene randomization method.

The interpretation of a P-value greatly depends on the sampling scheme on which the test is based and is related to hypothetical replications of the experiment performed.
GSEA

- Subramanian et al. (2005) included the gene randomization option in their GSEA program and suggested using gene randomization to generate hypothesis when the number of samples is small.

- GSEA utilizes a competitive statistic (Kolmogorov–Smirnov statistic) as a ‘score function’, if not as a test statistic, to represent the relative enrichment of DEGs in each gene set.

- GSEA tests the significance of the entire dataset by applying sample permutation to the scores.

- GSEA is considered a competitive method relative to individual gene sets, but is considered a self-contained method relative to the entire dataset (set of gene sets).
Gene Set Analysis Tools
A General Modular Framework for Gene Set Enrichment Analysis

Ackermann and Strimmer, BMC Bioinformatics 2009
Gene-level statistic: in general, the choice of the gene-level statistic does not seem to have a great impact on the results of the enrichment analysis.

Transformation: The choice of a transformation has quite a substantial effect on the detection rates. The rank squared transformation was most accurate.

Gene set statistics: Overall, the mean or the median work very well. (less sensitive with regard to outliers)
- The median and the rank-based Wilcoxon test may lead to a smaller number of significant findings.
- The GSEA (the enrichment score) is not as reliable as the other gene set test statistics.

Significance assessment: Depending on the data structure, both approaches can yield quite different results.
### Table I: Cutoff-free gene set analysis methods

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Name</th>
<th>Statistical test</th>
<th>Self-contained versus competitive</th>
<th>Gene versus ample randomization</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtanen et al.</td>
<td>2001</td>
<td>sample randomization</td>
<td>self-contained</td>
<td>sample</td>
<td>[8]</td>
<td></td>
</tr>
<tr>
<td>Pavlidis et al.</td>
<td>2002</td>
<td>gene randomization</td>
<td>competitive</td>
<td>gene</td>
<td>[9]</td>
<td></td>
</tr>
<tr>
<td>Mootha et al.</td>
<td>2003</td>
<td>GSEA</td>
<td>sample randomization</td>
<td>mixed</td>
<td>sample</td>
<td>[7]</td>
</tr>
<tr>
<td>Breslin et al.</td>
<td>2004</td>
<td>Catmap</td>
<td>gene randomization</td>
<td>competitive</td>
<td>gene</td>
<td>[3]</td>
</tr>
<tr>
<td>Smid et al.</td>
<td>2004</td>
<td>GO-Mapper</td>
<td>z-test</td>
<td>competitive</td>
<td>gene</td>
<td>[38]</td>
</tr>
<tr>
<td>Volinia et al.</td>
<td>2004</td>
<td>GOAL</td>
<td>gene randomization</td>
<td>competitive</td>
<td>gene</td>
<td>[39]</td>
</tr>
<tr>
<td>Barry et al.</td>
<td>2005</td>
<td>SAFE</td>
<td>sample randomization</td>
<td>competitive</td>
<td>sample</td>
<td>[9]</td>
</tr>
<tr>
<td>Beh-Shaul et al.</td>
<td>2005</td>
<td>SAFE</td>
<td>sample randomization</td>
<td>competitive</td>
<td>gene</td>
<td>[46]</td>
</tr>
<tr>
<td>Kim et al.</td>
<td>2005</td>
<td>PAGE</td>
<td>z-test</td>
<td>competitive</td>
<td>gene</td>
<td>[14]</td>
</tr>
<tr>
<td>Subramanian et al.</td>
<td>2005</td>
<td>GSEA</td>
<td>sample randomization</td>
<td>mixed</td>
<td>gene</td>
<td>[25]</td>
</tr>
<tr>
<td>Tian et al.</td>
<td>2005</td>
<td>Q1, Q2</td>
<td>gene or sample randomization</td>
<td>competitive or self-contained</td>
<td>gene or sample</td>
<td>[10]</td>
</tr>
<tr>
<td>Tomfohr et al.</td>
<td>2005</td>
<td>PLAGE</td>
<td>sample randomization</td>
<td>self-contained</td>
<td>sample</td>
<td>[20]</td>
</tr>
<tr>
<td>Edelman et al.</td>
<td>2006</td>
<td>ASSESS</td>
<td>sample randomization</td>
<td>competitive</td>
<td>sample</td>
<td>[28]</td>
</tr>
<tr>
<td>Kong et al.</td>
<td>2006</td>
<td>Hotelling’s T squared</td>
<td>self-contained</td>
<td>sample</td>
<td>[21]</td>
<td></td>
</tr>
<tr>
<td>Nam et al.</td>
<td>2006</td>
<td>ADGO</td>
<td>sample randomization</td>
<td>competitive</td>
<td>gene</td>
<td>[29]</td>
</tr>
<tr>
<td>Saxena et al.</td>
<td>2006</td>
<td>AE</td>
<td>sample randomization</td>
<td>competitive</td>
<td>sample</td>
<td>[31]</td>
</tr>
<tr>
<td>Scheer et al.</td>
<td>2006</td>
<td>JProGO</td>
<td>Fisher’s exact test,</td>
<td>competitive</td>
<td>gene</td>
<td>[40]</td>
</tr>
<tr>
<td>Al-Shahrour et al.</td>
<td>2007</td>
<td>Fatiscan</td>
<td>Fisher’s exact test,</td>
<td>competitive</td>
<td>gene</td>
<td>[41]</td>
</tr>
<tr>
<td>Backes et al.</td>
<td>2007</td>
<td>GeneTrail</td>
<td>Fisher’s exact test,</td>
<td>competitive</td>
<td>gene or sample</td>
<td>[42]</td>
</tr>
<tr>
<td>Cavalieri et al.</td>
<td>2007</td>
<td>EuGene Analyzer</td>
<td>Fisher’s exact test,</td>
<td>competitive</td>
<td>gene or sample</td>
<td>[43]</td>
</tr>
<tr>
<td>Dinu et al.</td>
<td>2007</td>
<td>SAM-GS</td>
<td>sample randomization</td>
<td>self-contained</td>
<td>sample</td>
<td>[22]</td>
</tr>
<tr>
<td>Efron et al.</td>
<td>2007</td>
<td>GSA</td>
<td>sample randomization</td>
<td>mixed</td>
<td>sample</td>
<td>[26]</td>
</tr>
<tr>
<td>Newton et al.</td>
<td>2007</td>
<td>Random set</td>
<td>z-test</td>
<td>competitive</td>
<td>gene</td>
<td>[44]</td>
</tr>
</tbody>
</table>
Gene Set Analysis Tools (1)

Table 2: Gene set analysis tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisma</th>
<th>Application Type</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catmap</td>
<td>H</td>
<td>Perl script</td>
<td><a href="http://bioinfo.theplu.se/catmap.html">http://bioinfo.theplu.se/catmap.html</a></td>
<td>[3]</td>
</tr>
<tr>
<td>Analyzer</td>
<td></td>
<td>Web server</td>
<td><a href="http://fatiscan.bioinfo.cipf.es/">http://fatiscan.bioinfo.cipf.es/</a></td>
<td>[41]</td>
</tr>
<tr>
<td>GenetRail</td>
<td>H, M, R, Y, SA, CG, AT</td>
<td>Web server</td>
<td><a href="http://genetrail.bioinf.uni-sb.de/">http://genetrail.bioinf.uni-sb.de/</a></td>
<td>[42]</td>
</tr>
<tr>
<td>Global test</td>
<td>NA</td>
<td>R package</td>
<td><a href="http://bioconductor.org/packages/2.0/bioc/html/globaltest.html">http://bioconductor.org/packages/2.0/bioc/html/globaltest.html</a></td>
<td>[77]</td>
</tr>
<tr>
<td>GOAL</td>
<td>H</td>
<td>Web server</td>
<td><a href="http://microarrays.unife.it">http://microarrays.unife.it</a></td>
<td>[39]</td>
</tr>
<tr>
<td>JProGO</td>
<td>Various prokaryotes</td>
<td>Web server</td>
<td><a href="http://www.jprogo.de/">http://www.jprogo.de/</a></td>
<td>[40]</td>
</tr>
<tr>
<td>MEGO</td>
<td>H</td>
<td>Windows standalone</td>
<td><a href="http://www.dxy.cn/mego/">http://www.dxy.cn/mego/</a></td>
<td>[46]</td>
</tr>
<tr>
<td>PAGE</td>
<td>H, M, R, Y</td>
<td>Python script</td>
<td>From the author (<a href="mailto:kimsy@krib.re.kr">kimsy@krib.re.kr</a>)</td>
<td>[14]</td>
</tr>
<tr>
<td>SAM-GS</td>
<td>NA</td>
<td>Windows Excel Add-In</td>
<td><a href="http://www.ualberta.ca/~yayas/homepage.html">http://www.ualberta.ca/~yayas/homepage.html</a></td>
<td>[22]</td>
</tr>
</tbody>
</table>

*a: Homo sapiens; m: Mus musculus; r: Rattus norvegicus; y: Saccharomyces cerevisiae; b: Bos Taurus; d: Daniel rerio; g: Gallus gallus; c: Caenorhabditis elegans; a: Arabidopsis thaliana; d: Drosophila melanogaster; z: Zebra fish; c: Candida albicans; s: Staphylococcus aureus; g: Corynebacterium glutamicum; a: Arabidopsis thaliana.
## Gene Set Analysis Tools (2)

### Table 2: Tools available for functional profiling by gene-set analysis

<table>
<thead>
<tr>
<th>Tool</th>
<th>Application type or URL for web servers</th>
<th>References</th>
<th>Test</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babelomics (FatiGO + FatiScan)</td>
<td><a href="http://www.babelomics.org">http://www.babelomics.org</a></td>
<td>[26,27,37,40,66]</td>
<td>FE/GS, C</td>
<td>402</td>
</tr>
<tr>
<td>FuncAssociate</td>
<td><a href="http://llama.med.harvard.edu/Software.html">http://llama.med.harvard.edu/Software.html</a></td>
<td>[87]</td>
<td>FE/GS, C</td>
<td>91</td>
</tr>
<tr>
<td>Global test R package</td>
<td></td>
<td>[64]</td>
<td>GS, SC</td>
<td>89</td>
</tr>
<tr>
<td>PAGE</td>
<td>Python script</td>
<td>[65]</td>
<td>GS, C</td>
<td>42</td>
</tr>
<tr>
<td>ErmineJ</td>
<td>Java</td>
<td>[105]</td>
<td>GS, C</td>
<td>35</td>
</tr>
<tr>
<td>FatiScan</td>
<td><a href="http://www.babelomics.org">http://www.babelomics.org</a></td>
<td>[66]</td>
<td>GS, C</td>
<td>34</td>
</tr>
<tr>
<td>GO-mapper</td>
<td>Windows, Perl script</td>
<td>[63]</td>
<td>GS, C</td>
<td>33</td>
</tr>
<tr>
<td>SAFE</td>
<td>R package</td>
<td>[62]</td>
<td>GS, C</td>
<td>27</td>
</tr>
<tr>
<td>GOAL</td>
<td><a href="http://microarrays.unife.it">http://microarrays.unife.it</a></td>
<td>[106]</td>
<td>GS, C</td>
<td>25</td>
</tr>
<tr>
<td>Catmap</td>
<td>Perl script</td>
<td>[107]</td>
<td>GS, C</td>
<td>19</td>
</tr>
<tr>
<td>GODist</td>
<td>Matlab program</td>
<td>[109]</td>
<td>GS, SC</td>
<td>17</td>
</tr>
<tr>
<td>ASSESS</td>
<td>Java</td>
<td>[114]</td>
<td>GS, C</td>
<td>2</td>
</tr>
<tr>
<td>DEA</td>
<td>R package</td>
<td>[115]</td>
<td>GS, C</td>
<td>1</td>
</tr>
<tr>
<td>Global ANCOVA R package</td>
<td></td>
<td>[67]</td>
<td>GS, SC</td>
<td>1</td>
</tr>
<tr>
<td>GAZER</td>
<td><a href="http://integromics.kobic.re.kr/GAzer/index.faces">http://integromics.kobic.re.kr/GAzer/index.faces</a></td>
<td>[116]</td>
<td>GS, C</td>
<td>-</td>
</tr>
<tr>
<td>SAM-GS</td>
<td>Windows excel add-in</td>
<td>[117]</td>
<td>GS, SC</td>
<td>-</td>
</tr>
</tbody>
</table>

---

The limitations of the information on current annotation databases for IGA:
(1) incomplete knowledge, (2) time-delayed curation, (3) imprecise or incorrect
electronic annotations, (4) inability to predict new functions and (5) semantic
misclassification of annotations.

The problems are all shared by GSA except for finding more relevant categories
among overlapping gene sets for which GSA is able to assign different scores.

Using overlapping gene sets in GSA substantially improves the analysis.

Table 3: Gene set databases

<table>
<thead>
<tr>
<th>Name</th>
<th>Organism*</th>
<th>Gene sets</th>
<th>Web address</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erminej</td>
<td>H, M, R</td>
<td>GO, composite GO, InterPro, Pathways, TFBS</td>
<td><a href="http://www.bioinformatics.ubc.ca/microannots/">http://www.bioinformatics.ubc.ca/microannots/</a></td>
<td>[16]</td>
</tr>
<tr>
<td>MSigDb</td>
<td>H</td>
<td>Cytobands, curated pathways, motif, computed</td>
<td><a href="http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html">http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html</a></td>
<td>[25]</td>
</tr>
</tbody>
</table>

*H: Homo sapiens; M: Mus musculus; R: Rattus norvegicus; Y: Saccharomyces cerevisiae.
Methods Comparison
Goeman and Buhlmann (2007) has investigated methodological issues in methods that test for differential expression of gene sets.

It has revealed some methodological aspects of popular methods that are inefficient or even incorrect from a statistical point of view.

Compare: the three "self-contained" gene set methods: (1) Global Test, (2) ANCOVA Global Test and (3) SAM-GS.

After the standardization of gene expression, the three methods gave very similar results, with slightly higher statistical significance given by SAM-GS.

Global Test and ANCOVA Global Test were able to analyze continuous and survival phenotypes and to adjust for covariates.

The framework of the competitive hypothesis testing via gene sampling is subject to serious errors in calculating and interpreting statistical significance of gene sets, because of its implicit or explicit untenable assumption of probabilistic independence across genes.
Efron and Tibshirani (2007) introduced five test statistics for a GSEA algorithm: (1) mean, (2) mean.abs, (3) maxmean, (4) GSEA and (5) GSEA.abs.

They concluded that the maxmean statistic is the only method with consistently low P-values in all situations.

Some criticisms of GSEA: the enrichment score can be influenced by the size of a gene set and by the presence or absence of lower-ranking sets.

The appropriate method is chosen depending on either the number of the samples or the property of the DEG sets the user wants to find.

Methods were selected for which software packages were available through the Bioconductor project.

<table>
<thead>
<tr>
<th>Method</th>
<th>Measurement of gene set</th>
<th>Permutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSEA-Category</td>
<td>The sum of the <strong>T-statistics</strong></td>
<td>Sample</td>
</tr>
<tr>
<td>GSEA-limma</td>
<td>The mean of the <strong>T-statistics</strong></td>
<td>Gene</td>
</tr>
<tr>
<td>SAFE</td>
<td>The sum of the ranks of the <strong>T-statistics</strong></td>
<td>Sample</td>
</tr>
<tr>
<td>Globaltest</td>
<td>The mean of the <strong>Q-statistics</strong></td>
<td>Sample</td>
</tr>
<tr>
<td>PCOT2</td>
<td>Hotelling's <strong>$T^2$</strong> (multivariate T-statistic)</td>
<td>Sample</td>
</tr>
<tr>
<td>sigPathway</td>
<td>The mean of the <strong>T-statistics</strong></td>
<td>Sample</td>
</tr>
</tbody>
</table>

**correlation-based methods**

Globaltest $\sim$ PCOT2 $>$ GSEA-Category

Poor: SAFE, sigPathway

- **Compare**: the five self-contained gene set methods and the competitive GSEA method:

- The **self-contained methods** of SAM-GS, Global test and ANCOVA Global outperformed GSEA.

- **General conclusions** regarding the relative performance of the investigated methods **could not be made**, as no simulation studies were completed.
The criterion for choosing a statistical method for GSA:

- If the purpose is to find gene sets relatively enriched with DEGs, a competitive method based on Q1 should be used.

- If the purpose is to find gene sets clearly separated between the two sample groups, a self-contained method based on Q2 should be selected.

- Prefers the mixed approach (Q3) to avoid the clear drawbacks of the other methods, but recommends using all the methods simultaneously, if possible, with biological analyses.

- **Sample randomization** provides statistically sound P-values.
Selecting an optimal tool, which of course depends on the type of experimental data. The first thing to consider is the type of organism.

For a human gene expression dataset with an enough number of samples (more than 10), GSEA is highly recommended because it is a statistically sound method based on sample randomization and provides a user-friendly, standalone program.

- GSA (Efron and Tibshirani, 2007) and SAFE provide potentially better statistical properties than GSEA (R packages).

For mouse, rat or yeast datasets for which the GSEA program is not available, web servers such as Babelomics, GAzer or GeneTrail are recommended.

- When the number of samples is small, gene randomization-based tools such as ErmineJ or GAzer are highly recommended.

Tsai and Chen (2009) proposed using a MANOVA test for gene-set analysis.

They compared it to several methods including:
1. principal component analysis,
2. SAM-GS,
3. GSEA,
4. Maxmean,
5. ANCOVA,

They found the MANOVA approach appeared to perform best, but concluded that most methods, except GSEA and maxmean, were generally comparable in terms of power.

A limitation of the MANOVA method is that it is only applicable to categorical outcomes data.

The self-contained have been reported to be more powerful than competitive methods.

Compare: (1) Kolmogorov–Smirnov test (KS), (2) Fisher’s method (FM), (3) Stouffer’s method (SM), (4) tail strength (TS), (5) a novel modified tail strength statistic (MTS), (6) global model with fixed effects (GMFE), (7) global model with random effects (GMRE), and (8) principal component analysis (PCA).

The simulation scenarios varied according to: (1) number of genes in a gene set, (2) number of genes associated with the phenotype, (3) effect sizes, (4) correlation between expression of genes within a gene set, and (5) the sample size.

Over a variety of scenarios, the FM with empirical p-values or the GMRE were the most powerful analytical approaches for a self-contained gene set analysis.

The analysis based on the first principal component and Kolmogorov-Smirnov test tended to have lowest power.


Further Reading

**METHODOLOGY ARTICLE**

Gene set enrichment meta-learning analysis: next-generation sequencing versus microarrays

Gregor Stiglic*, Mateja Bajgot and Peter Kokol

---

**METHODOLOGY ARTICLE**

Gene set analysis exploiting the topology of a pathway

Maria Sofia Massa¹, Monica Chiogna¹*, Chiara Romualdi²