Tools for Microarray Data Analysis

吳漢銘 助理教授
淡江大學 數學系

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

2011/06/16
Content

- Microarray Life Cycle
- Statistical Issues and Recent Progress

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

- Gene Set Analysis (GSA)
  - Gene Set Enrichment Analysis (GSEA)
Microarray Life Cycle

**Biological Questions**
1. differentially expressed genes
2. relationships between gene, tissues or treatments
3. classification of tissues and samples

**Experimental Design**
1. design
2. sample size and power

**Microarray Experiments**
1. target preparation
2. hybridization
3. washing
4. image acquisition

**Preprocessing**
1. image analysis
2. quantify expression
3. quality assessment
4. normalization

**Statistical Analysis**
1. estimation
2. testing
3. clustering
4. classification
5. gene network

**Domain knowledge**

**Biological Verification and Interpretation**
1. RT-PCR
Basic Statistical Issues

- Data Preprocessing: image processing, normalization
- Gene Filtering, Missing Values Imputation
- Finding Differential Expressed Genes
- Visualization (including dimension reduction)
- Clustering
- Classification
- ...
Advance Statistical Issues

- Experimental Design
- Time Course Microarray Experiments
- Gene Regulatory Networks/Pathway
- Annotations/Databases
- Comparisons, Sample Size, Dye Swap, Replicates, ...
- Web Resource, Software Design
- ...

...
Recent Progress

- Incorporating **biological knowledge** into analysis.
- Meta-analysis: pooling
- Well-curated publicly data set.
- Quality-control assessment.
- Development of standardized testing platforms (e.g., AffyComp).
- 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 Differential Expressed Genes (DEGs)
Finding Differentially Expressed Genes

- More than two samples
- Two-sample (independent)
- Paired-sample (dependent)

Microarray Data Matrix

<table>
<thead>
<tr>
<th>Gene</th>
<th>exp01</th>
<th>exp02</th>
<th>exp03</th>
<th>exp04</th>
<th>exp05</th>
<th>exp06</th>
<th>exp07</th>
<th>exp08</th>
<th>exp09</th>
<th>exp10</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene001</td>
<td>-0.48</td>
<td>-0.42</td>
<td>0.87</td>
<td>0.92</td>
<td>0.67</td>
<td>-0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene002</td>
<td>-0.39</td>
<td>-0.58</td>
<td>1.08</td>
<td>1.21</td>
<td>0.52</td>
<td>-0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene003</td>
<td>0.87</td>
<td>0.25</td>
<td>-0.17</td>
<td>0.18</td>
<td>-0.13</td>
<td>-0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene004</td>
<td>1.57</td>
<td>1.03</td>
<td>1.22</td>
<td>0.31</td>
<td>0.16</td>
<td>-1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene005</td>
<td>-1.15</td>
<td>-0.98</td>
<td>1.21</td>
<td>1.62</td>
<td>1.12</td>
<td>-0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene006</td>
<td>0.04</td>
<td>-0.12</td>
<td>0.31</td>
<td>0.16</td>
<td>0.17</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene007</td>
<td>2.95</td>
<td>0.45</td>
<td>-0.40</td>
<td>-0.66</td>
<td>-0.59</td>
<td>-0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene008</td>
<td>-1.22</td>
<td>-0.74</td>
<td>1.34</td>
<td>1.50</td>
<td>0.83</td>
<td>-0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene009</td>
<td>-0.73</td>
<td>-1.06</td>
<td>-0.79</td>
<td>-0.02</td>
<td>0.16</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene010</td>
<td>-0.58</td>
<td>-0.40</td>
<td>0.13</td>
<td>0.59</td>
<td>-0.09</td>
<td>-0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene011</td>
<td>-0.50</td>
<td>-0.42</td>
<td>0.66</td>
<td>1.05</td>
<td>0.68</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene012</td>
<td>-0.06</td>
<td>-0.29</td>
<td>0.42</td>
<td>0.46</td>
<td>0.30</td>
<td>-0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene013</td>
<td>-0.16</td>
<td>0.29</td>
<td>0.17</td>
<td>-0.28</td>
<td>-0.02</td>
<td>-0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene014</td>
<td>-0.36</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.08</td>
<td>-0.23</td>
<td>-0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene015</td>
<td>-0.72</td>
<td>-0.05</td>
<td>0.64</td>
<td>1.04</td>
<td>0.04</td>
<td>-0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene016</td>
<td>-0.78</td>
<td>-0.52</td>
<td>0.26</td>
<td>0.20</td>
<td>0.48</td>
<td>-0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene017</td>
<td>0.60</td>
<td>0.55</td>
<td>0.41</td>
<td>0.46</td>
<td>0.18</td>
<td>-1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene018</td>
<td>-0.20</td>
<td>-0.67</td>
<td>0.13</td>
<td>0.10</td>
<td>0.38</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene019</td>
<td>-2.29</td>
<td>-0.64</td>
<td>0.77</td>
<td>1.60</td>
<td>0.53</td>
<td>-0.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene020</td>
<td>-1.46</td>
<td>-0.76</td>
<td>1.06</td>
<td>1.50</td>
<td>0.74</td>
<td>-0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene021</td>
<td>-0.57</td>
<td>0.42</td>
<td>1.03</td>
<td>1.35</td>
<td>0.64</td>
<td>-0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene022</td>
<td>-0.11</td>
<td>0.13</td>
<td>0.41</td>
<td>0.60</td>
<td>0.23</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

p-values or Statistics

<table>
<thead>
<tr>
<th>Gene</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene001</td>
<td>0.067</td>
</tr>
<tr>
<td>gene002</td>
<td>0.052</td>
</tr>
<tr>
<td>gene003</td>
<td>0.013</td>
</tr>
<tr>
<td>gene004</td>
<td>0.016</td>
</tr>
<tr>
<td>gene005</td>
<td>0.011</td>
</tr>
<tr>
<td>gene006</td>
<td>0.017</td>
</tr>
<tr>
<td>gene007</td>
<td>0.059</td>
</tr>
<tr>
<td>gene008</td>
<td>0.063</td>
</tr>
<tr>
<td>gene009</td>
<td>0.516</td>
</tr>
<tr>
<td>gene010</td>
<td>0.009</td>
</tr>
<tr>
<td>gene011</td>
<td>0.068</td>
</tr>
<tr>
<td>gene012</td>
<td>0.030</td>
</tr>
<tr>
<td>gene013</td>
<td>0.002</td>
</tr>
<tr>
<td>gene014</td>
<td>4.23</td>
</tr>
<tr>
<td>gene015</td>
<td>0.084</td>
</tr>
<tr>
<td>gene016</td>
<td>0.048</td>
</tr>
<tr>
<td>gene017</td>
<td>0.018</td>
</tr>
<tr>
<td>gene018</td>
<td>0.538</td>
</tr>
<tr>
<td>gene019</td>
<td>0.053</td>
</tr>
<tr>
<td>gene020</td>
<td>0.074</td>
</tr>
<tr>
<td>gene021</td>
<td>0.764</td>
</tr>
<tr>
<td>gene022</td>
<td>0.423</td>
</tr>
</tbody>
</table>

Microarray Data Matrix

<table>
<thead>
<tr>
<th>Gene</th>
<th>exp01</th>
<th>exp02</th>
<th>exp03</th>
<th>exp04</th>
<th>exp05</th>
<th>exp06</th>
<th>exp07</th>
<th>exp08</th>
<th>exp09</th>
<th>exp10</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene001</td>
<td>-0.48</td>
<td>-0.42</td>
<td>0.87</td>
<td>0.92</td>
<td>0.67</td>
<td>-0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene002</td>
<td>-0.11</td>
<td>0.13</td>
<td>0.41</td>
<td>0.60</td>
<td>0.23</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cy 5: treatment
Cy 3: control

- p-values or Statistics
- ranking
- fix number
- above some level
Paired Data: Breast Cancer Dataset

cDNA Microarrays Data:

- **#Samples**: 20 breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy
- **#Genes**: 9216 genes.

Paired Data:

- Two measurements from each patient, one before treatment and one after treatment.

Interests:

- the difference between the two measurements (the log ratio).
- whether a gene has been **up-regulated** or **down-regulated** in breast cancer following that treatment.


Unpaired Data: Leukemia Dataset

**Affymetrix Microarray Data**

- **#Samples**: Bone marrow
  - #ALL (acute lymphoblastic leukemia): 27 patients (急性淋巴細胞白血病)
  - #AML (acute myeloid leukemia): 11 patients (急性骨髓性白血病)
- **#Genes**: 7070 genes.

**Unpaired Data**:

- Two groups of patients (ALL, AML).

**Interests**:

- To identify the genes that are up- or down-regulated in ALL relative to AML.
- (i.e., differentially expressed between the two groups.)

Cancer Genomics Program at Whitehead Institute for Genome Research http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Fold-Change Method (1)

1) Calculate fold-change.
2) Rank the genes.
3) Select genes.
Method 1: Select genes based on Numbers

- average differential expression > FC.

Problems:

- FC is an arbitrary threshold.
- FC does not take into account individuals and sample size.

Example:

- s2 (200) close to BG (100), the difference could represent noise.
- credible: a gene is regulated 2-fold with 10000, 5000 units.
**Method 2: Select genes based on %**

- Choose 5% of genes that have the largest expression ratios.

**Problems:**

- Possible that no genes have statistically significantly different gene expression.
null hypothesis:

H₀: No differential expressed.
H₀: no difference in the mean gene expression in the group tested.
H₀: The gene will have equal means across every group.
H₀: \( \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\ldots = \mu_n) \)
The *p*-values

*p*-values
- Probability of **false positives** (Reject $H_0 \mid H_0$ true).
- Probability of observing your data under the assumption that the null hypothesis is true.
- *p*-value = 0.03: only a 3% chance of drawing the sample if the null hypothesis was true.

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

**Use p-value to select genes**
- Select differentially expressed genes based on their *p*-value (not FC).
- The smaller the *p*-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.
One Sample t-test

One sample t-test

\[ H_0 : \mu = \mu_0 \]
\[ H_1 : \mu \neq \mu_0 \text{ (two-tailed).} \]
\( \mu \): population mean.
\( \alpha \): significant level (e.g., 0.05).
Test Statistic:
\[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \]
\( \bar{X} \): sample mean.
\( S \): sample standard deviation.
\( n \): 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 = \( P_{H_0}(|T| > t_0), \quad T \sim t_{n-1} \).

Question

- whether a gene is differentially expressed for a condition with respect to baseline expression?
- \( H_0 : \mu=0 \) (log ratio)
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 \( |t_0| > t_{\alpha/2, df} \)

Applied to a Gene From Leukemia Dataset

\textit{metallothionein IB}

- The gene \textit{metallothionein IB} is on the Affymetrix array used for the leukemia data.

Two-sample t-test

- \( t = -3.4177, p = 0.0016. \)

Conclusion

- the expression of \textit{metallothionein IB} is significantly higher in AML than in ALL at the 1% level.
Two Sample t-test (Paired)

Paired Sample t-test

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

Test Statistic:
\[ T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}} \]

\( \bar{d} \): average of sample differences.
\( S_d \): standard deviation of sample difference
\( n \): number of pairs.

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

ACAT2

- The gene acetyl-Coenzyme A acetyltransferase 2 (ACAT2) is on the microarray used for the breast cancer data.

Paired t-test

- \( t=3.22 \). (two-tailed)
- \( p\)-value = 0.0045, which is significant at a 1% confidence level.

Conclusion

- ACAT2 has been significantly down-regulated following chemotherapy at the 1% level.
Assumptions of t-test

**Be Normal**
- paired t-test,
  the distribution of the subtracted data that must be normal.
- unpaired t-test,
  the distribution of both data sets must be normal.

**How to Detect Normality**
- **Plots**: Histogram, Density Plot, QQplot,…
- **Test for Normality**: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test.

**Homogeneous**
- the variances of the two population are equal.
- Test for equality of the two variances: Variance ratio F-test.
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
  \[ t = \frac{\bar{M}}{(a+s^2)/\sqrt{n}} \], where \( a \) is estimated from the mean and standard deviation of the sample variances \( s^2 \).

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)

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

multiple regression model

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
Significance Analysis of Microarrays (SAM)
SAM does not do any normalization!

http://www-stat.stanford.edu/~tibs/SAM/

# SAM: Response Type

<table>
<thead>
<tr>
<th>Response type</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td>Real number eg 27.4 or -45.34</td>
</tr>
<tr>
<td>Two class (unpaired)</td>
<td>Integer 1, 2</td>
</tr>
<tr>
<td>Multiclass</td>
<td>Integer 1, 2, 3, ...</td>
</tr>
<tr>
<td>Paired</td>
<td>Integer -1, 1, -2, 2, etc.</td>
</tr>
<tr>
<td></td>
<td>eg - means Before treatment, + means after treatment</td>
</tr>
<tr>
<td></td>
<td>-1 is paired with 1, -2 is paired with 2, etc.</td>
</tr>
<tr>
<td>Survival data</td>
<td>(Time, status) pair like (50,1) or (120,0)</td>
</tr>
<tr>
<td></td>
<td>First number is survival time, second is status (1=died, 0=censored)</td>
</tr>
<tr>
<td>One class</td>
<td>Integer, every entry equal to 1</td>
</tr>
<tr>
<td><strong>Time course, two class (unpaired)</strong></td>
<td>(1 or 2)Time(t)[Start or End]</td>
</tr>
<tr>
<td><strong>Time course, two class (paired)</strong></td>
<td>(-1 or 1 or -2 or 2 etc)Time(t)[Start or End]</td>
</tr>
<tr>
<td><strong>Time course, one class</strong></td>
<td>1Time(t)[Start or End]</td>
</tr>
<tr>
<td>Pattern discovery</td>
<td>eigengenek, where k is one of 1,2,... number of arrays</td>
</tr>
</tbody>
</table>

*SAM Users guide and technical document*
**SAM: Significance Analysis of Microarrays**

- **Two class, unpaired data**
  \[ y_j = 1 \text{ or } 2 \]
  \[ r_i = \bar{x}_{i2} - \bar{x}_{i1} \]

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

- **Data**
  \( y_j \) response

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

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

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

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

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

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

- **Order statistics**
  \[ d_1 \leq d_2 \leq \ldots \leq d_p \]

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

- **Large positive difference**
- **Large negative difference**
SAM: Expected Test Statistics

response
\[ y_j \]
1, 1, …, 2, …, 2

Permutation
1, 2, 1, 2, 1, …, 1

\[ r_i^* = \bar{x}_{i2}^* - \bar{x}_{i1}^* \]
\[ d_i^* = \frac{r_i^*}{s_i^* + s_0^*} \]

\[ \tilde{d}^{(b)} = \frac{1}{B} \sum_b d_{(i)}^{*b} \]

expected order statistics
\[ \tilde{d}(p) \]
\[ \tilde{d}(2) \]
\[ \tilde{d}(1) \]
SAM Plot

Points for genes with evidence of induction

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

significant positive

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

lower cut-point \( \text{cut}_{low}(\Delta) \)

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

Points for genes with evidence of repression

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

significant negative

\[ \bar{d}_{(p)}, \bar{d}_{(2)}, \bar{d}_{(1)} \]

\[ d_{(p)}, d_{(2)}, d_{(1)} \]
Gene Set Enrichment Analysis (GSEA)
Gene Sets

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

- **A gene set (a gene class)**
  - a group of genes with related **functions**.
  - sets of genes or **pathways**, for their association with a phenotype.
  - identified from a **prior** biological knowledge.
  - may better reflect the true underlying biology.
  - may be more appropriate **units** for analysis.

- **Examples**: metabolic pathway, protein complex, or GO (gene ontology) category.

- **Various database**: BioCarta, KEGG, Gene Ontology
A statistical test to determine significance of a gene class is referred to as gene class testing \((\text{GCT})\) or gene set analysis \((\text{GSA})\).

The common approach to the GSA is first to identify a list of genes that express differently among two groups of samples.

The list of differentially expressed genes is then examined with biologically pre-defined gene sets to determine whether any set is overrepresented in the list compared with the whole list.

GSA is becoming a powerful alternative to individual-gene analysis.
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)
Step 1: Enrichment Score (ES)

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 $L$.

$$P_{\text{hit}}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_{j}|^p}{N_R}$$

$$N_R = \sum_{g_j \in S} |r_{j}|^p$$

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

$$E(S) = \max_i \{P_{\text{hit}}(S, i) - P_{\text{miss}}(S, i)\}$$

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

$E(S) < 0$: 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 correspondingly 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 Esnull computed with randomly assigned phenotype.

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

- For positive ES
- For negative ES

\[ p-value \approx \frac{\# \{ ES^{(b)} > ES_{obs} \}}{\# \text{permutations}} \]
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 and MSigDB.

What’s New

29-Mar-2011: Version 3.7 of this web site has been released. We reviewed and updated the text for improved clarity, and fixed some minor bugs. No changes have been made to the MSigDB data or the GSEA desktop software.

09-Sep-2010: We are pleased to announce the release of version 3.0 of the Molecular Signatures Database (MSigDB). It contains an extensively revised and expanded version of the C2 collection of canonical pathways and literature gene sets. In addition, we have made several enhancements to the GSEA and MSigDB website, and fixed an error in the Compute Overlaps procedure. For further details, see the release notes.

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.

Contributors

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors. Funded 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)

<table>
<thead>
<tr>
<th>Downloads</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

### 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>JavaGSEA Desktop Application</th>
<th>Easy-to-use graphical user interface</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>JavaGSEA Java Jar file</th>
<th>Command line usage</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GSEA Java Source Code Java source files</th>
<th>100% Java implementation of GSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-GSEA R Script</th>
<th>Usage from within the R programming environment</th>
</tr>
</thead>
<tbody>
<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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GenePattern GSEA Module</th>
<th>Use GSEA from within GenePattern</th>
</tr>
</thead>
<tbody>
<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>

**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 cis-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.
# 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</td>
<td>Gender_hgu133a.gct  GenderCollapsed.gct  Gender.cls</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>Results of C1 GSEA analysis of this dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Results of C2 GSEA analysis of this dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Results of C2 GSEA analysis of this dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>individuals</td>
<td></td>
<td>(3): 267-73</td>
</tr>
<tr>
<td></td>
<td>Results of C2 GSEA analysis of this dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td>Transcriptional profiles from leukemias - ALL and AML</td>
<td>Leukemia_hgu95av2.gct  LeukemiaCollapsed.gct  Leukemia.cls</td>
<td>Armstrong et al. (2002) Nat Genet</td>
</tr>
<tr>
<td></td>
<td>Results of C1 GSEA analysis of this dataset</td>
<td></td>
<td>30(1): 41-7.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>816-24.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bhattacharjee et al. (2001) Proc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Natl Acad Sci U S A 98(24): 13790</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.</td>
</tr>
<tr>
<td>Gene sets</td>
<td>Archived gene sets from the GSEA PNAS 2005 publication.</td>
<td>C1_symbols.gmt (positional)  C2_symbols.gmt (curated)</td>
<td>Subramaniam and Tamayo PNAS</td>
</tr>
<tr>
<td></td>
<td><strong>Note:</strong> This collection of gene sets is not the latest version, so when</td>
<td></td>
<td>2005</td>
</tr>
<tr>
<td></td>
<td>beginning a new analysis you might want to download the current collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>of gene sets from the Downloads page.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**P53 Status in Cancer Cell Lines**

- **NCI-60 collection of cancer cell lines.**
  - The **mutational status** of the p53 gene has been reported for 55 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.
Input for GSEA (1)

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

```
#1.2
12625
50
1
100_g_at
1000_at
1001_at
1002_f_at
1003_s_at
1004_at
1005_at
1006_at
1007_s_at

1
786-0
132.94
215.37
328.68
39.64
18.46
60.83
68.02
610.35
354.92

2
BT-549
234.31
234.31
8.84
12.14
30.19
54.41
65.93
3.57
208.33
```

P53Collapsed_symbols.gct

```
#1.2
1
10100
50
1
TACC2
C14orf132
AGER
32385_at
RBM17
DYT1
CORO1A
WT1

1
786-0
132.94
215.37
328.68
39.64
18.46
60.83
68.02
610.35
354.92

2
BT-549
234.31
234.31
8.84
12.14
30.19
54.41
65.93
3.57
208.33
```

P53.cls

```
1
50
2
1

2
MUT
WT

3
MUT
MUT
MUT
MUT
MUT
MUT
MUT
MUT
MUT
MUT

...```

...
### 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></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>chr10q24</td>
<td>Cytogenetic band</td>
<td>PITX3</td>
<td>SPFH1</td>
<td>NEURL</td>
<td>C10orf12</td>
</tr>
<tr>
<td>2</td>
<td>chr5q23</td>
<td>Cytogenetic band</td>
<td>ALDH7A1</td>
<td>IL13</td>
<td>8-Sep</td>
<td>IRF1</td>
</tr>
<tr>
<td>3</td>
<td>chr8q24</td>
<td>Cytogenetic band</td>
<td>HAS2</td>
<td>LRRC14</td>
<td>TSTA3</td>
<td>DGAT1</td>
</tr>
<tr>
<td>4</td>
<td>chr16q24</td>
<td>Cytogenetic band</td>
<td>RPL13</td>
<td>GALNS</td>
<td>FANCA</td>
<td>CPNE7</td>
</tr>
<tr>
<td>5</td>
<td>chr13q14</td>
<td>Cytogenetic band</td>
<td>AKAP11</td>
<td>ARL11</td>
<td>ATP7B</td>
<td>C13orf1</td>
</tr>
<tr>
<td>6</td>
<td>chr7p21</td>
<td>Cytogenetic band</td>
<td>ARL4A</td>
<td>SCIN</td>
<td>GLCCI1</td>
<td>SP8</td>
</tr>
<tr>
<td>7</td>
<td>chr10q23</td>
<td>Cytogenetic band</td>
<td>41bbPathway</td>
<td>TNF-type receptor 4-1BB is IL2</td>
<td>TRAF2</td>
<td>MAP3K1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>ace2Pathway</td>
<td>Angiotensin-converting enz</td>
<td>COL4A3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>acetaminophenPathway</td>
<td>Acetaminophen selectively</td>
<td>CYP3A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>achPathway</td>
<td>Nicotinic acetylcholine rece</td>
<td>RAPSN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>actinYPathway</td>
<td>The Arp 2/3 complex localiz</td>
<td>ACTR3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>agpcrPathway</td>
<td>G-protein coupled receptor</td>
<td>PRKAR2A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>ahspPathway</td>
<td>Alpha-hemoglobin stabilizin</td>
<td>CPO</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>alfpPathway</td>
<td>BLACK</td>
<td>ADPRT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>akrk13Pathway</td>
<td>A kinase anchor protein 13</td>
<td>EDG4</td>
</tr>
</tbody>
</table>
Launch 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

Steps in GSEA

BROAD INSTITUTE
Run GSEA

1. Load data
2. Run GSEA
3. Set Parameters and Launch Analysis Tools
4. Results
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 pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 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 pvalue < 1%
- 15 gene sets are significantly enriched at nominal pvalue < 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%

- **Snapshot of enrichment results**
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Table: Snapshot of enrichment results

![Graphs and plots showing enrichment results](image-url)
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 plot: CHR12Q13

Zero cross at 4076

'MUT' (positively correlated)

'WT' (negatively correlated)

Enrichment profile — Hits — Ranking metric scores
## Hits

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 polyepitide 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>
<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 pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 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>
<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>1 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>2 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>3 CHRXP11</td>
<td>Details ...</td>
<td>66</td>
<td>0.53</td>
<td>1.62</td>
<td>0.000</td>
<td>0.864</td>
<td>0.833</td>
<td>1909</td>
<td>tags=55%, list=21%, signal=69%</td>
</tr>
<tr>
<td>4 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>5 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>6 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>7 CHR1P36</td>
<td>Details ...</td>
<td>117</td>
<td>0.41</td>
<td>1.42</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1852</td>
<td>tags=74%, list=20%, signal=54%</td>
</tr>
<tr>
<td>68 CHRSQ14</td>
<td>Details ...</td>
<td>18</td>
<td>0.21</td>
<td>0.52</td>
<td>0.050</td>
<td>1.000</td>
<td>1.000</td>
<td>1850</td>
<td>tags=33%, list=70%, signal=54%</td>
</tr>
<tr>
<td>69 CHRSQ13</td>
<td>Details ...</td>
<td>10</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>70 CHRXQ22</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>71 CHR9Q22</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>72 CRXQ22</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>

Table: Gene sets enriched in phenotype MUT (33 samples) [plain text format]
Gene Markers for the MUT versus WT Comparison

Gene set details
- Gene set size filters \( (\text{min}=15, \text{max}=500) \) 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>GPSM2</td>
<td>null</td>
<td>GPSM2</td>
<td>G-protein signalin</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 fami</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>0.46191833</td>
</tr>
<tr>
<td>LRRC42</td>
<td>null</td>
<td>LRRC42</td>
<td>leucine rich repeat</td>
<td>0.4581682</td>
</tr>
<tr>
<td>PEX14</td>
<td>null</td>
<td>PEX14</td>
<td>peroxisomal bioge</td>
<td>0.4568304</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

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

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

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

Number of occurrences:
the number of LES pairs in a particular bin.