GO HERE!

- [http://hpc.ilri.cgiar.org/beca/training/AdvancedBFX2013/index2.html](http://hpc.ilri.cgiar.org/beca/training/AdvancedBFX2013/index2.html)
Data Mining: Clustering and Statistical Analysis with MeV

Marcus Jones
Infectious Disease/Genomic Medicine
J. Craig Venter Institute
28 August 2013
Data Analysis Concepts
The last stage of analysis involves the analysis of multiple hybridizations.

- Time Courses
- Distinct Experimental Groups
- Replicates of one direct measurement of interest
**Expression Ratios**

**Dyes**
- Cy5
- Cy3

**Measured Intensities**
- Cy5: 216108
- Cy3: 77103

**Ratio**
\[
\frac{216108}{77103} = 2.798
\]

**Gene Expression**
- *In vivo*
- *Gene Expression*
  - RNA
  - cDNA synthesis
  - 384-well qRT-PCR

**Data Analysis**
Log\textsubscript{2} Expression Ratios: Benefits

Log ratios are easier to work with than regular ratios.

A five-fold change in expression level could be represented by one of two regular expression ratios:

\[
\frac{500}{100} = \text{Ratio } 5.0 \quad \text{OR} \quad \frac{100}{500} = \text{Ratio } 0.2
\]

Note the asymmetrical nature of the ratio values.

However, if the regular expression ratios are converted to log expression ratios:

\[
\frac{500}{100} = \text{Ratio } 5.0 \quad \rightarrow \quad \log_2(5.0) = 2.32
\]

OR

\[
\frac{100}{500} = \text{Ratio } 0.2 \quad \rightarrow \quad \log_2(0.2) = -2.32
\]

The log ratios equal in magnitude and the sign indicates which channel had the higher intensity.
## Expression Vectors: Shape

### Gene A

<table>
<thead>
<tr>
<th>Expt</th>
<th>Expt 1</th>
<th>Expt 2</th>
<th>Expt 3</th>
<th>Expt 4</th>
<th>Expt 5</th>
<th>Expt 6</th>
<th>Expt 7</th>
<th>Expt 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>-0.8</td>
<td>1.5</td>
<td>1.8</td>
<td>0.5</td>
<td>-0.4</td>
<td>-1.3</td>
<td>0.8</td>
<td>1.5</td>
</tr>
</tbody>
</table>

### Log₂ Ratios

Log₂(cy5/cy3)

![Graph of Log₂(cy5/cy3) with values ranging from -0.8 to 1.5 across experiments]
We can study the expression levels of multiple genes across a series of experimental conditions by stacking the corresponding expression vectors.

<table>
<thead>
<tr>
<th></th>
<th>Expt 1</th>
<th>Expt 2</th>
<th>Expt 3</th>
<th>Expt 4</th>
<th>Expt 5</th>
<th>Expt 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>-2.32</td>
<td>-1.69</td>
<td>-0.87</td>
<td>-0.12</td>
<td>0.73</td>
<td>1.42</td>
</tr>
<tr>
<td>Gene B</td>
<td>2.71</td>
<td>2.09</td>
<td>1.24</td>
<td>0.70</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td>Gene C</td>
<td>-1.55</td>
<td>-0.49</td>
<td>0.97</td>
<td>1.32</td>
<td>0.59</td>
<td>-0.38</td>
</tr>
</tbody>
</table>

This is called an *Expression Matrix*. 
Each element is a log ratio:
\[ \log_2 \left( \frac{\text{Cy5}}{\text{Cy3}} \right) \]

- Red indicates a positive log ratio (i.e. \( \text{Cy5} > \text{Cy3} \))
- Green indicates a negative log ratio (i.e. \( \text{Cy5} < \text{Cy3} \))
- Black indicates a log ratio of ~0 (i.e. \( \text{Cy5} = \text{Cy3} \) or \( \text{Cy5} \approx \text{Cy3} \))
- Gray indicates missing data

Expression Matrices are commonly represented as a grid of red and green cells:
Basic Analysis Approaches

- **General Clustering**
  - For finding gene sets with coherent patterns of expression, e.g. hierarchical clustering, k-means, SOTA, SOM

- **Hypothesis Driven Analysis**
  - Statistical tests based on defined experimental design, e.g. t-test, Significance Analysis of Microarrays (SAM), ANOVA, 2-Factor ANOVA

- **Biological Role Identification**
  - Finding biological meaning from gene lists, EASE
What is RNASeq?

- Sample type
- Sample/Library preparation
- Instrument for processing
- Mapping vs *de novo* assembly
- RNASeq measurement values
  - RPKM
  - FPKM
Characterization of Host and Pathogen Expression

- Total RNA
- Enriched Bacterial mRNA
- HiSeq 2000

J. Craig Venter
INSTITUTE
• Facultative anaerobe, Gram-positive

• Clinical presentation:
  o skin infections, abscess
  o pneumonia
  o meningitis
  o persistent infections
  o artificial joints, bone
The Concern Over MRSA

Hospital MRSA infections in the USA

S. aureus Impact on Health Care

Number of deaths

http://www.destiny-pharma.demon.co.uk/images/MRSA%20Deaths%202004.gif
Virulence Factors

S. aureus USA300 replacing USA400
Experimental Goal

- Understand the host-pathogen interaction during *S. aureus* infection.
- Identify host pathways differentially expressed during infection.
- Determine host tissue specific expression.
- Determine pathogen differential expression.
Steps in analysis

• Clustering to find expression trends
• Statistical analysis of significant genes
• Functional analysis of differentially expressed genes
• Visual representation of differentially expressed genes
One goal may be to identify genes which have “similar” patterns of expression (i.e. similar expression vectors).

“Clustering Algorithms” are a popular method for doing this.

Some Clustering Algorithm Types:
* Agglomerative: Hierarchical Trees
* Divisive: $k$-means, Self-Organizing Maps
* Nonclustering: Principal Component Analysis

Now we just need to decide what it means to be “similar”…
The ability to calculate a \textit{distance} (or \textit{similarity} - its inverse) between two expression vectors is fundamental to clustering algorithms.

Distance and Similarity

Distance between vectors is the basis upon which decisions are made when grouping similar patterns of expression.

Selection of a \textit{distance metric} defines the concept of distance for a particular experiment.
k-Means Clustering (KMC)
K-Means Clustering (KMC)

1. Specify number of clusters, e.g., 5.

2. Randomly assign genes to clusters.
3. Calculate mean / median expression profile of each cluster.

4. Select a gene and move it to the cluster having the closest mean profile.

5. If the gene is shifted to a new cluster, recalculate means for the winning and losing clusters.

6. Repeat steps 4 and 5 until genes cannot be shuffled around any more, OR a user-specified number of iterations has been reached.
Hands On

• Cluster sample data by KMC analysis
Statistical Methods
Statistical tests can often be applied to data sets where replicate samples fall into distinct experimental groups.

- Stat tests can be used to find genes that are differentially expressed in accordance with the various conditions under study.
- Unlike general clustering, these tests can provide measures of confidence when reporting genes that are differentially expressed across experimental conditions.
Finding Significant Genes

- Average Fold Change Difference for each gene suffers from being arbitrary and not taking into account systematic variation in the data.
Finding Significant Genes

- Assume we will compare two conditions with multiple replicate hybs for each condition.

- Our goal is to find genes that have significantly different mean expression between these conditions.

- These are the genes that we will use for later data mining such as biological role analysis.
MeV’s Analysis Modules

Hands On

and

Demonstration
Matching Methods to Designs

Experimental designs help to dictate which methods are appropriate to apply.

• **One Sample T-test** – Finds genes within a single experimental group that are over or under expressed relative to the reference sample.

• **Two Sample T-test** – Finds genes that have mean expression values that are different between two experimental groups.
Matching Methods to Designs

- **ANOVA** – similar to t-test but used for the analysis of multiple experimental groups

  \[\text{condition 1, condition 2, \ldots, condition } n\]

- **Two Factor ANOVA** – useful for the analysis of multiple experimental groups where two experimental factors are under study.

  \(2\times3\) Design

  \[
  \begin{array}{ccc}
  \text{strain 1} & \text{condition 1} & \text{condition 2} \\
  \text{strain 2} & \text{condition 3} \\
  \end{array}
  \]
Matching Methods to Designs

- SAM – Significance Analysis of Microarrays provides options to cover many experimental designs.
  - One Class – Single experimental group of replicate hybs
  - Two Class – Two experimental groups
  - Two Class Paired – Two experimental groups where related samples across groups can be paired
  - Multi-class – Multiple experimental groups
SAM

- Developed by Tusher and Tibshirani at Stanford to address the problems of multiple tests on type I error, false positives.

- Uses an estimated FDR as a criterion for significance.

- Interactive means of selecting gene lists while monitoring FDR.

- Provides testing modes that cover many common experimental designs.
SAM Graph

-plots expected vs. observed d-scores.

-slider alters the delta value (obs – exp, dashed lines), from the diagonal line (obs = exp).

-the number of genes selected (outside the bounds) and the estimated number of false calls is reported as you adjust delta.
Hands On SAM
Genes Unique to Infection Models
Genes Shared in Infection Models
Lung Specific Genes
Statistical Significance and Biological Significance

Statistical tests provide lists of genes that show significant changes in expression, however at least two important considerations remain:

What is the magnitude of the expression change? Relatively small changes may be significant due to low variability of the measurements.

Is there a common biological system implicated by multiple genes that show changes in expression?
Expression Analysis Systematic Explorer (EASE)

Exploring Prevalent Biological Roles within Gene Lists
EASE analysis identifies prevalent biological themes within gene clusters.

The significance of each identified theme is determined by its prevalence in the cluster and in the gene population of genes from which the cluster was created.
Consider a population of genes representing a diverse set of biological roles or themes shown below as different colors.
Many algorithms can be applied to expression data to partition genes based on expression profiles over multiple conditions.

Many of these techniques work solely on expression data and disregard biological information.
Consider a particular cluster...

- What are the some of the predominant biological themes represented in the cluster and how should significance be assigned to a discovered biological theme?
Example:

Population Size: 40 genes
Cluster size: 12 genes

10 genes, shown in green, have a common biological theme and 8 occur within the cluster.
EASE Results

• Consider all of the Results
  EASE reports all themes represented in a cluster and although some themes may not meet statistical significance it may still be important to note that particular biological roles or pathways are represented in the cluster.

• Independently Verify Roles
  Once found, biological themes should be independently verified using annotation resources.
Basic EASE Requirements

Annotation keys; identifiers for each gene must be loaded with the data into MeV.

EASE file system; EASE uses a file system to link annotation keys to biological themes.
EASE File System

Gene List (MeV Cluster)

Optional Conversion File

Locus Link | GenBank#
-----------|----------
86         | AA410394
87         | AA669042
88         | AA775521
89         | AA196000
90         | AI261580
90         | AA136910
91         | AA436008
92         | AA682819
92         | H23277
95         | AA402915
97         | W80489
98         | N49204
101        | AA279186
102        | AA872057

Classification File

Locus Link | GO_Biological_Process
-----------|------------------------
10076      | protein amino acid dephosphorylation
10076      | biological_process unknown
10076      | transmembrane receptor protein tyrosine phosphatase
10076      | cell adhesion
10076      | ubiquitin cycle
10076      | cell-matrix adhesion
10077      | biological_process unknown
10077      | cell-cell signaling
10079      | cation transport
10079      | metabolism
1008       | homophilic cell adhesion
10082      | posttranslational membrane targeting
10083      | hearing
10083      | intracellular signaling cascade
10083      | vision
10084      | regulation of transcription, DNA-dependent
10085      | development
10085      | cell adhesion
10087      | immune response
10087      | protein amino acid phosphorylation

NIAID provided the foundation Java classes upon which the MeV version was built.
<table>
<thead>
<tr>
<th>Term</th>
<th>Pop. Hits</th>
<th>Pop. Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>signal transduction</td>
<td>1940</td>
<td>14845</td>
</tr>
<tr>
<td>transcription, DNA-dependent</td>
<td>1830</td>
<td>14845</td>
</tr>
<tr>
<td>regulation of transcription</td>
<td>1728</td>
<td>14845</td>
</tr>
<tr>
<td>biological_process</td>
<td>1695</td>
<td>14845</td>
</tr>
<tr>
<td>transport</td>
<td>1683</td>
<td>14845</td>
</tr>
<tr>
<td>G-protein coupled receptor protein signaling pathway</td>
<td>1596</td>
<td>14845</td>
</tr>
<tr>
<td>sensory perception of smell</td>
<td>1079</td>
<td>14845</td>
</tr>
<tr>
<td>multicellular organismal development</td>
<td>895</td>
<td>14845</td>
</tr>
<tr>
<td>regulation of transcription, DNA-dependent</td>
<td>785</td>
<td>14845</td>
</tr>
<tr>
<td>metabolic process</td>
<td>645</td>
<td>14845</td>
</tr>
<tr>
<td>oxidation-reduction process</td>
<td>604</td>
<td>14845</td>
</tr>
<tr>
<td>protein phosphorylation</td>
<td>575</td>
<td>14845</td>
</tr>
<tr>
<td>ion transport</td>
<td>573</td>
<td>14845</td>
</tr>
<tr>
<td>cell differentiation</td>
<td>559</td>
<td>14845</td>
</tr>
<tr>
<td>protein transport</td>
<td>508</td>
<td>14845</td>
</tr>
<tr>
<td>modification-dependent protein catabolic process</td>
<td>490</td>
<td>14845</td>
</tr>
<tr>
<td>cell cycle</td>
<td>479</td>
<td>14845</td>
</tr>
<tr>
<td>transmembrane transport</td>
<td>469</td>
<td>14845</td>
</tr>
<tr>
<td>cell adhesion</td>
<td>468</td>
<td>14845</td>
</tr>
<tr>
<td>apoptosis</td>
<td>426</td>
<td>14845</td>
</tr>
<tr>
<td>proteolysis</td>
<td>408</td>
<td>14845</td>
</tr>
<tr>
<td>positive regulation of transcription from RNA polymerase II promoter</td>
<td>341</td>
<td>14845</td>
</tr>
<tr>
<td>immune response</td>
<td>305</td>
<td>14845</td>
</tr>
<tr>
<td>translation</td>
<td>277</td>
<td>14845</td>
</tr>
<tr>
<td>mRNA processing</td>
<td>266</td>
<td>14845</td>
</tr>
<tr>
<td>cell division</td>
<td>265</td>
<td>14845</td>
</tr>
<tr>
<td>intracellular signal transduction</td>
<td>244</td>
<td>14845</td>
</tr>
<tr>
<td>lipid metabolic process</td>
<td>234</td>
<td>14845</td>
</tr>
<tr>
<td>response to DNA damage stimulus</td>
<td>230</td>
<td>14845</td>
</tr>
<tr>
<td>spermatogenesis</td>
<td>216</td>
<td>14845</td>
</tr>
<tr>
<td>negative regulation of transcription from RNA polymerase II promoter</td>
<td>215</td>
<td>14845</td>
</tr>
<tr>
<td>response to stimulus</td>
<td>213</td>
<td>14845</td>
</tr>
<tr>
<td>nervous system development</td>
<td>210</td>
<td>14845</td>
</tr>
<tr>
<td>DNA repair</td>
<td>208</td>
<td>14845</td>
</tr>
<tr>
<td>RNA splicing</td>
<td>203</td>
<td>14845</td>
</tr>
<tr>
<td>small GTPase mediated signal transduction</td>
<td>202</td>
<td>14845</td>
</tr>
<tr>
<td>chromatin modification</td>
<td>195</td>
<td>14845</td>
</tr>
<tr>
<td>in utero embryonic development</td>
<td>192</td>
<td>14845</td>
</tr>
<tr>
<td>mitosis</td>
<td>190</td>
<td>14845</td>
</tr>
<tr>
<td>intracellular protein transport</td>
<td>171</td>
<td>14845</td>
</tr>
<tr>
<td>carbohydrate metabolic process</td>
<td>167</td>
<td>14845</td>
</tr>
<tr>
<td>positive regulation of cell proliferation</td>
<td>163</td>
<td>14845</td>
</tr>
<tr>
<td>potassium ion transport</td>
<td>161</td>
<td>14845</td>
</tr>
<tr>
<td>vesicle-mediated transport</td>
<td>158</td>
<td>14845</td>
</tr>
<tr>
<td>heart development</td>
<td>148</td>
<td>14845</td>
</tr>
<tr>
<td>inflammatory response</td>
<td>142</td>
<td>14845</td>
</tr>
</tbody>
</table>
Heat Map
**PHAGOSOME**

**Conventional phagocytosis**
- Binding to receptor
- Endocytosis
- Phagocytosis-promoting receptors
- Recyling endosome
- Early endosome
- Late endosome
- Lysosome

**ER-mediated phagocytosis**
- Phagocytic cup
- F-actin
- NADPH oxidase
- ER (endoplasmic reticulum)
- CALR
- TAP
- MHCI
- MHCI

**Phagosome maturation by interaction with the endocytic pathway**
- VAMP3
- PI(3)F
- TLR
- His
- Fab5
- Fab7
- TRIF
- MAFB
- TUBB
- NOS
- NADPH oxidase
- NADPH oxidase
- MHCII

**Cross presentation**
- Antigen processing and presentation
- TAP
- ER (endoplasmic reticulum)

**The activation mechanism of NADPH oxidase**
- Inactive form
- Active form
- Cytochrome b558
- Membrane
- Rac
- NADPH oxidase complex
- p47phox
- p67phox
- gp91phox
- gp21phox
- p40phox
- p22phox
- gp91phox
- gp21phox
- p40phox
- gp91phox
- gp21phox
- p40phox
- gp91phox
- gp21phox
- p40phox

(c) Kanazawa Laboratories
## Neutrophil Evasion

<table>
<thead>
<tr>
<th>Locus</th>
<th>Symbol</th>
<th>USA300_LUNG</th>
<th>USA300_SKIN</th>
<th>USA300_KIDNEY</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAUSA300_0108</td>
<td>sok</td>
<td>-6.30</td>
<td>N/A</td>
<td>-0.23</td>
</tr>
<tr>
<td>SAUSA300_0113</td>
<td>SpA</td>
<td>N/A</td>
<td>N/A</td>
<td>-11.43</td>
</tr>
<tr>
<td>SAUSA300_0379</td>
<td>ahpF</td>
<td>9.964</td>
<td>N/A</td>
<td>-6.430</td>
</tr>
<tr>
<td>SAUSA300_0380</td>
<td>ahpC</td>
<td>2.150</td>
<td>N/A</td>
<td>-2.384</td>
</tr>
<tr>
<td>SAUSA300_0398</td>
<td>ssl-3/4</td>
<td>0.67</td>
<td>0.98</td>
<td>-0.96</td>
</tr>
<tr>
<td>SAUSA300_0404</td>
<td>ssl-10</td>
<td>-1.79</td>
<td>-6.31</td>
<td>-5.06</td>
</tr>
<tr>
<td>SAUSA300_0407</td>
<td>ssl-11</td>
<td>-0.05</td>
<td>-5.08</td>
<td>-7.65</td>
</tr>
<tr>
<td>SAUSA300_0883</td>
<td>eap</td>
<td>6.64</td>
<td>-1.51</td>
<td>-2.77</td>
</tr>
<tr>
<td>SAUSA300_1052</td>
<td>ecb</td>
<td>-1.65</td>
<td>-4.85</td>
<td>-4.83</td>
</tr>
<tr>
<td>SAUSA300_1055</td>
<td>efb</td>
<td>4.85</td>
<td>-6.54</td>
<td>-13.44</td>
</tr>
<tr>
<td>SAUSA300_1058</td>
<td>hla</td>
<td>8.49</td>
<td>N/A</td>
<td>-4.10</td>
</tr>
<tr>
<td>SAUSA300_1067</td>
<td>PSM beta</td>
<td>N/A</td>
<td>N/A</td>
<td>-6.95</td>
</tr>
<tr>
<td>SAUSA300_1255</td>
<td>mprF</td>
<td>6.68</td>
<td>1.56</td>
<td>-7.97</td>
</tr>
<tr>
<td>SAUSA300_1381</td>
<td>lukF</td>
<td>7.87</td>
<td>4.10</td>
<td>3.84</td>
</tr>
<tr>
<td>SAUSA300_1382</td>
<td>lukS</td>
<td>-4.32</td>
<td>-3.65</td>
<td>-3.40</td>
</tr>
<tr>
<td>SAUSA300_1445</td>
<td>scpA</td>
<td>-6.37</td>
<td>-7.04</td>
<td>-10.33</td>
</tr>
<tr>
<td>SAUSA300_1768</td>
<td>lukD</td>
<td>6.65</td>
<td>4.26</td>
<td>-5.69</td>
</tr>
<tr>
<td>SAUSA300_1769</td>
<td>lukE</td>
<td>1.61</td>
<td>-5.29</td>
<td>-7.09</td>
</tr>
<tr>
<td>SAUSA300_1890</td>
<td>Staphopain</td>
<td>-5.63</td>
<td>-5.60</td>
<td>-6.99</td>
</tr>
<tr>
<td>SAUSA300_1919</td>
<td>SCIN</td>
<td>14.44</td>
<td>5.25</td>
<td>-5.23</td>
</tr>
<tr>
<td>SAUSA300_1920</td>
<td>chips</td>
<td>12.17</td>
<td>9.12</td>
<td>-1.99</td>
</tr>
<tr>
<td>SAUSA300_1922</td>
<td>sak</td>
<td>5.28</td>
<td>4.24</td>
<td>-2.05</td>
</tr>
<tr>
<td>SAUSA300_1974</td>
<td>lukG</td>
<td>6.97</td>
<td>-2.63</td>
<td>-5.62</td>
</tr>
<tr>
<td>SAUSA300_1975</td>
<td>lukH</td>
<td>8.52</td>
<td>1.50</td>
<td>-2.86</td>
</tr>
<tr>
<td>SAUSA300_2364</td>
<td>SBI</td>
<td>1.70</td>
<td>4.46</td>
<td>4.16</td>
</tr>
<tr>
<td>SAUSA300_2365</td>
<td>hlgA</td>
<td>19.85</td>
<td>5.49</td>
<td>5.46</td>
</tr>
<tr>
<td>SAUSA300_2366</td>
<td>hlgB</td>
<td>17.73</td>
<td>0.59</td>
<td>0.89</td>
</tr>
<tr>
<td>SAUSA300_2367</td>
<td>hlgC</td>
<td>6.54</td>
<td>-3.38</td>
<td>-3.62</td>
</tr>
<tr>
<td>SAUSA300_2504</td>
<td>OatA</td>
<td>4.15</td>
<td>-5.30</td>
<td>-1.63</td>
</tr>
<tr>
<td>SAUSA300_2572</td>
<td>aureolysin</td>
<td>-8.64</td>
<td>5.51</td>
<td>-5.39</td>
</tr>
</tbody>
</table>
S. aureus shared expression

Up regulated

- Lung: 863
- Skin: 260
- Kidney: 323
- Intersections: 95, 42, 37, 73

Down regulated

- Lung: 31
- Skin: 11
- Kidney: 84
- Intersections: 31, 11, 73, 722, 378, 463, 722
Amino Acid Biosynthesis
Biosynthesis of cofactors, prosthetic groups, and carriers
Cell Envelope
Cellular Processes
Central intermediary metabolism
DNA Metabolism
Energy Metabolism
Fatty acid and phospholipid metabolism
Hypothetical
Mobile Genetic Elements
No Data
Protein Fate
Protein Synthesis
Regulatory functions
Signal Transduction
Transcription
Transport and Binding
Unknown

Lung vs in vitro
Skin vs in vitro
Kidney vs in vitro
Metabolic Analysis Lung Model
Linear Expression Maps (LEM)

Expression Organized by Chromosomal Location
• Linear Expression Maps organize expression by locus location on the chromosome or plasmid

• LEMs provide a means to navigate over the genome to find contiguous loci displaying similar patterns of expression.
LEM Requirements

- Locus IDs - each gene of interest should have a gene identifier that can be mapped to the genome
- Chromosomal Location – a separate coordinate file or information in the loaded annotation file should provide a chromosome id and location for each locus.
LEG Analysis

USA300_LUNG -- Chromosome

USA300_SKIN -- Chromosome

USA300_KIDNEY -- Chromosome

ACME

vSAA

ΦSa2usa

vSAβ

ΦSa3usa

Urease cluster
LEM Analysis

chemotaxis–inhibiting protein CHIPS
truncated amidase
staphyloc kinase precursor
aurolysin
phiPVL ORF17–like protein
phi77 ORF044–like protein
phi77 ORF109–like protein
"phi77 ORF002–like protein, phage minor structural protein"
"phi77 ORF004–like protein, putative phage tail component"
"phi77 ORF001–like protein, phage tail tape measure protein"
phi77 ORF100–like protein
hypothetical protein
hypothetical protein
"phi77 ORF020–like protein, phage major tail protein"
phi77 ORF029–like protein
hypothetical protein
phi77 ORF045–like protein
"phi77 ORF006–like protein, putative capsid protein"
"phi77 ORF015–like protein, putative protease"
phage portal protein
Useful Open Source Tools

- Comprehensive Microbial Resource
  - (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi)
- MeV
  - (http://sourceforge.net/projects/mev-tm4/)
- GLAMM
  - (http://www.microbesonline.org/cgi-bin/glamm)
- MicrobesOnline
  - (http://www.microbesonline.org/)
- KEGG Database
  - (http://www.kegg.jp/)
Acknowledgements

- These projects have been funded with federal funds from the National Institute of Allergy and Infectious Diseases at the National Institutes of Health.
- JCVI Faculty and Staff
SAM d-score

The d-score is analogous to a t-value (in t test) and includes a term representing group difference ($r_i$) and variance ($s_i$).

$$d_i = \frac{r_i}{S_i + S_0}$$

$r_i$ is termed the *quantitative response*. It's definition varies depending on the type of test. e.g. for two class unpaired:

$$r_i = \bar{x}_i - \bar{y}_i$$

$S_i$ ; Variance term, def. varies with test type

$S_0$ ; “exchangeability factor” previously called the “fudge factor”
The key point is that a large d-score represents a large response and tight variance. Perhaps significant…

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

How large is “large”? It’s relative….

An “expected” mean d-score is generated for a set of random data permutations. The “observed” d-score is compared to the expected d-score. The larger the difference (observed – expected), the greater the significance of the observed d-score.