Web tools for biological data analysis

Hedi Peterson
University of Tartu / Quretec Ltd
Three main approaches

- Computational prediction
  - Discrete Function Prediction: Predicting individual or pairwise interactions
  - Network Reconstruction: Integrating and extending network representations
  - Systems Level Modeling: Modeling networks

- Experimental validation
You had hypothesis and did an (microarray) experiment – what can you ask from your data?
Selection of possible questions

• What functional groups are overrepresented?
• What genes are co-expressed?
• What pathways are affected?
• How can I have a compendious view of what is in my data?
• How does my results relate to previous experiments?
• How is my data connected, what modules are there?
1. Functional annotations
g:Profiler
Toolset for functional profiling of gene lists from large-scale experiments
http://biit.cs.ut.ee/gprofiler
Functional analysis of gene lists

What is previously known?

DNA motif detection

protein-DNA interactions

protein-protein interactions

gene expression

<table>
<thead>
<tr>
<th>Gene List</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLRT2</td>
</tr>
<tr>
<td>RRBP1</td>
</tr>
<tr>
<td>D530005L17RIK</td>
</tr>
<tr>
<td>GPX1</td>
</tr>
<tr>
<td>GREN2</td>
</tr>
<tr>
<td>ATF7</td>
</tr>
<tr>
<td>120000SF10RIK</td>
</tr>
<tr>
<td>MYL4</td>
</tr>
<tr>
<td>RCD1</td>
</tr>
<tr>
<td>PCDH17</td>
</tr>
<tr>
<td>IGF1</td>
</tr>
<tr>
<td>NEBL</td>
</tr>
<tr>
<td>NPL_05145.2</td>
</tr>
<tr>
<td>LOB2</td>
</tr>
<tr>
<td>EVII</td>
</tr>
<tr>
<td>PTPRM</td>
</tr>
<tr>
<td>ARH8AP6</td>
</tr>
<tr>
<td>ESRP6</td>
</tr>
<tr>
<td>BMP2K</td>
</tr>
<tr>
<td>3330417A13RIK</td>
</tr>
<tr>
<td>RPRM</td>
</tr>
<tr>
<td>NRK</td>
</tr>
<tr>
<td>PCDH17</td>
</tr>
<tr>
<td>TNNT1</td>
</tr>
<tr>
<td>ELTD1</td>
</tr>
<tr>
<td>TGFβ2</td>
</tr>
<tr>
<td>CD93</td>
</tr>
<tr>
<td>N/A</td>
</tr>
<tr>
<td>ZFP9</td>
</tr>
<tr>
<td>FBNL</td>
</tr>
<tr>
<td>TGFβ2</td>
</tr>
<tr>
<td>AT597468</td>
</tr>
<tr>
<td>MAV1A</td>
</tr>
</tbody>
</table>
Functional analysis of gene lists

- Gene expression
- Protein-protein interactions
- DNA motif detection
- Protein-DNA interactions
- Long lists of candidate genes

What is previously known?
- Biological processes
- Cell locations
- Molecular functions

Gene Ontology
- mIRBase
- Transfac
- microRNA target sites
- Transcription factor binding sites
- Pathways
- Conservation

KEGG & Reactome
- Ensembl
g:Profiler — What is previously known about my list of genes?
- statistical enrichment tests
- ordered and unordered gene lists
- Gene Ontology
- Pathways (KEGG, Reactome)
- DNA motifs (Transfac, miRBase)

g:Cocoa — analyse and compare multiple gene lists

g:Convert — manage database IDs of genes, proteins and probesets

g:Sorter — find genes with similar expression profiles in public data

g:Orth — fetch conserved genes in related species

g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments
Structured vocabulary of biological terms:
- biological processes
- molecular functions
- cell components
Pathways:
Reactome, KEGG

all human pathways in Reactome

KEGG: apoptosis pathway

KEGG: mTOR pathway
Gene Annotations

Gene Ontology

[GO:0032502] developmental process

[GO:0043856] anatomical structure development

[GO:0048869] cellular developmental process

[GO:0030154] cell differentiation

[GO:0048731] system development

[GO:0048468] cell development

[GO:0010001] glial cell differentiation

[GO:0007399] nervous system development

[GO:0048627] myoblast development

[GO:0022008] neurogenesis

Transcription factor binding sites

NRK
PCDH17
TNNI1
ELTD1
TGFBI
CD93

Pathways

ESPRG
BMP2K
3830417A13Rik
RPRM
NRK
PCDH17
TNNI1
ELTD1
TGFBI
CD93
N/A
ZFP9
FBNI
TGFBI
A1597468
MNN1A
Functional enrichment and related statistics

Does my list include more genes with function \( x \) than expected by random chance?

\[
p = \sum_{k}^{\min(n,K)} \frac{(K)(N-K)}{(n-k)(N)}
\]

Hypergeometric p-value
Composing gene lists in g:Profiler

List of candidate genes

Ordered list
Composing gene lists in g:Sorter

List of microarray datasets from Gene Expression Omnibus (GEO)

IGF1

Single gene + most similar genes in a public microarray dataset

top-50 genes with best Pearson correlation

Ordered list for g:Profiler
g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments
1. Functional annotations
2. Clustering the data
Gene selection problem

- 45 000 probes on a chip -> which one to select for profiling?
- Where to put clustering threshold(s)?
- How to compare different gene groups?
- How to do it in a reasonable time?
- How to fit results to one screen?
VisHiC
Visualization of Hierarchical Clustering
http://biit.cs.ut.ee/vishic
Measuring gene activity

- microarray experiments
Hierarchical clustering

- group together genes that behave similarly
- guilt by association
Functional annotations

- Gene Ontology
  - molecular function
  - cellular component
  - biological process

- Pathway enrichments
  - KEGG
  - Reactome

- Motif enrichment
  - Transfac motifs
  - miRBase
VisHiC main idea

Ease the work of a biologist working with gene expression arrays and trying to find biologically important clusters
What VisHiC does?

• Hierarchical clustering
  - Different distance measures (Pearson correlation, Abs. Pearson correlation, Eucledian distance, Minkowski distance, Chord distance)

• Functional annotations
  - Based on g:Profiler (GO, KEGG, Reactome, miRBase, Transfac)

• Dendogram pruning
  - Keep only important clusters
Dendogram pruning

+ 8000px

8000px
VisHiC - input

SELECT DATASET:

Select dataset  GDS558 Myocardial remodeling in response to LVAD
Select distance measure*  Pearson correlation distance
Select correction type*  Analytical threshold

SELECT TREE PARAMETERS

Select strategy*  Best annotation
Display sparse clusters
Insert cluster minimum size*  min: 5, max: 1000
Insert cluster maximum size*  min: 5, max: 1000
Insert additional threshold (double) number, e.g. 1.0 or 1.0E-8
Select gradient*  Exponential

ok
VisHiC - output
<table>
<thead>
<tr>
<th>Id</th>
<th>Size</th>
<th>Score</th>
<th>p-value</th>
<th>Name</th>
<th>Description</th>
<th>Domain Best Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>29887</td>
<td>7</td>
<td>33.63</td>
<td>8.06E-63</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7.7E-34</td>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ml:hsa-miR-363</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mi:hsa-miR-363</td>
<td></td>
</tr>
<tr>
<td>31170</td>
<td>3</td>
<td>32.95</td>
<td>6.36E-61</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.06E-31</td>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ml:hsa-miR-363</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mi:hsa-miR-363</td>
<td></td>
</tr>
<tr>
<td>31024</td>
<td>6</td>
<td>32.52</td>
<td>8.28E-63</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7.7E-34</td>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ml:hsa-miR-363</td>
<td></td>
</tr>
<tr>
<td>28189</td>
<td>4</td>
<td>32.25</td>
<td>6.79E-61</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.45E-32</td>
<td>KEGG:05012</td>
<td>Parkinson's disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ml:hsa-miR-551b*</td>
<td></td>
</tr>
<tr>
<td>32018</td>
<td>9</td>
<td>31.81</td>
<td>3.53E-58</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.14E-32</td>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ml:hsa-miR-768-3p</td>
<td></td>
</tr>
<tr>
<td>31462</td>
<td>5</td>
<td>31.64</td>
<td>1.15E-62</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.69E-32</td>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
<tr>
<td>33458</td>
<td>17</td>
<td>30.52</td>
<td>1.59E-56</td>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>Domain Best Annotations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>generation of precursor metabolites and energy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oxidoreductase activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.24E-29</td>
<td>KEGG:05012</td>
<td>Parkinson's disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REAC:163200</td>
<td></td>
</tr>
</tbody>
</table>
VisHiC - output

DATASET GDS558 MYOCARDIAL REMODELING IN RESPONSE TO LVAD. CLUSTER ID: 29887

Cluster size: 7
Cluster score: 33.63
Gene group functional profiling: Click here
To convert gene ids: Click here
Orthology search: Click here
### GO: Cellular Component annotation sorted by p-value: (2 of 33)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0:0005739</td>
<td>mitochondrion</td>
<td>8.06E-63</td>
</tr>
<tr>
<td>G0:0044429</td>
<td>mitochondrial part</td>
<td>1.6E-53</td>
</tr>
</tbody>
</table>

### GO: Biological Process annotation sorted by p-value: (2 of 27)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0:0005091</td>
<td>generation of precursor metabolites and energy</td>
<td>3.3E-43</td>
</tr>
<tr>
<td>G0:0022900</td>
<td>electron transport chain</td>
<td>4.5E-27</td>
</tr>
</tbody>
</table>

### GO: Molecular Function annotation sorted by p-value: (2 of 16)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0:0016491</td>
<td>oxidoreductase activity</td>
<td>1.44E-21</td>
</tr>
<tr>
<td>G0:0008137</td>
<td>NADH dehydrogenase (ubiquinone) activity</td>
<td>1.01E-20</td>
</tr>
</tbody>
</table>

### KEGG annotation sorted by p-value: (2 of 13)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG:00190</td>
<td>Oxidative phosphorylation</td>
<td>7.7E-34</td>
</tr>
<tr>
<td>KEGG:05012</td>
<td>Parkinson's disease</td>
<td>1.36E-32</td>
</tr>
</tbody>
</table>

### REACTOME annotation sorted by p-value: (2 of 14)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>REAC:153200</td>
<td>Electron Transport Chain</td>
<td>2.21E-31</td>
</tr>
<tr>
<td>REAC:153217</td>
<td>NADH entries the respiratory chain at Complex I</td>
<td>3.33E-16</td>
</tr>
</tbody>
</table>

### miRNA annotation sorted by p-value: (2 of 9)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M:hsa-miR-363</td>
<td>M:hsa-miR-363</td>
<td>1.66E-6</td>
</tr>
<tr>
<td>M:hsa-miR-551b²</td>
<td>M:hsa-miR-551b²</td>
<td>5.61E-6</td>
</tr>
</tbody>
</table>

### TRANSFAC annotation sorted by p-value: (1 of 1)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Annotation name</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF:M00133_4</td>
<td>NNKGAATTAVTDN:4</td>
<td>2.73E-5</td>
</tr>
</tbody>
</table>
1. Functional annotations
2. Clustering the data
3. Affected pathways
Affected pathways

• Which part of the pathway was affected?
• How much genes change their behaviour?
• Are the affected genes connected?
KEGGanim
Pathway animations for high-throughput data
http://biit.cs.ut.ee/kegganim
Expression matrix - genes - proteins

Biological pathways from KEGG

BIOINFORMATICS

Vol. 00 no. 00 2007

Pages 1-2

KEGGanim: pathway animations for high-throughput data
Prit Adler, Jüri Reimand, Jürgen Jänes, Raivo Kolde, Hedi Peterson, Jaak Vilo

*Estonian Biocentre, Riia 23b, Tartu, Estonia  
University of Tartu, Institute of Computer Science,  
Livii 2, Tartu, Estonia  
QuroTec Inc. Ülikooli 6a, Tartu, Estonia

Pathway analysis: 00200 - 3.439

Pathway enrichment: 00200 - 1.439
Motivation

- **Expression data**
  - DNA -> RNA -> protein (enzyme)

- **KEGG knowledge-base**
  - Describing biological knowledge using graphical representation!
**Public and private datasets**

- Some examples available from AE and GEO
- Support for user accounts – personal data management!
Supported datasets

Upload an experimental dataset:

- **File type:**
  - Soft-file (example)
  - Custom delimited file (example)
- **#\^** Comment symbols
- **TAB** Field delimiter
- **Column configuration:**
  - Number of columns from left not regarded as data values
  - Column # used for gene/protein/probeset reference

Upload a file:

```
/home/adler/data/ex
```

OR specify a dataset:

**example_dataset_01**

Dataset name

Data values

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Homo sapiens</th>
<th>Dataset species</th>
<th>Single Channel (Affymetrix™)</th>
<th>Dual Channel (2-color array)</th>
<th>Apply logarithm normalisation</th>
<th>Apply data centering</th>
<th>Replace missing values (no value or NA)</th>
<th>K Nearest Neighbours (KNN)</th>
<th>Fixed value</th>
<th>Overwrite previous data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1007_at</td>
<td>NA</td>
<td>140.7</td>
<td>339.1</td>
<td>259.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1053_at</td>
<td>67.9</td>
<td>NA</td>
<td>23.3</td>
<td>50.7</td>
<td>7.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>117_at</td>
<td>62.7</td>
<td>68.0</td>
<td>127.4</td>
<td>217.8</td>
<td>67.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>121_at</td>
<td>42.1</td>
<td>395.9</td>
<td>901.3</td>
<td>751.3</td>
<td>437.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1258_at</td>
<td>12</td>
<td>27.1</td>
<td>101.5</td>
<td>119.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1254_at</td>
<td>295.9</td>
<td>329.5</td>
<td>301.2</td>
<td>248.2</td>
<td>206.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1336_at</td>
<td>34.6</td>
<td>26.5</td>
<td>129.3</td>
<td>121.3</td>
<td>6.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1320_at</td>
<td>7.0</td>
<td>30.5</td>
<td>19.0</td>
<td>10</td>
<td>12.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1405_i_at</td>
<td>3662.7</td>
<td>9184.9</td>
<td>559.9</td>
<td>216.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KEGG pathways

- Pathways can be ordered
  - Alphabetically
  - Pathway ID
  - Number of matching probesets

Current dataset:
- **organism**: *hsapiens*
- **dataset**: Human grade II astrocytic tumor vs grade IV tumors Margareto
How to read the picture?

- **# a** - This area shows dataset name, condition number, condition name.
- **# b** - Black rectangles on pathway maps represent enzymes. The rectangle is subdivided horizontally if only one gene corresponds to a given enzyme and the gene is present via **different probesets** on the array. Each vertically subdivided rectangle displays expression level of one probeset in our dataset.
- **# c** - If a given enzyme maps to **several genes** mapped in the KEGG database, the rectangle is subdivided vertically. Each horizontally subdivided rectangle displays expression of different of the enzyme. If gene is not present in our dataset, the rectangle is colored **gray**.
- **# d** - Both of the above conditions may apply simultaneously.
- **# e** - KEGG maps are pathway specific, i.e. enzymes of different species are displayed on a single map. Not all enzymes are present in every organism. Rectangles having E.C. numbers denote enzyme classes that are not described in the current organism.
- **# f** - This area shows additional information about the selected pathway and the animation.
  * The **linegraph** displays deviation of genes throughout the animation (dataset).
  * The **colorbar** displays overall scale of the possible colors on pathway map.
  * The metainfo area shows probeset deviation, number of probes in the pathway and number of probes in the dataset.
Select preferred conditions
Creating static movie

**CineFilm** – single frames attached side by side
1. Functional annotations
2. Clustering the data
3. Affected pathways
4. Other microarray experiments
Given a gene of interest

- study expression properties of the query gene
- what other genes behave similarly to the query gene (“guilt by association”)
- in which conditions?
- find global “neighbors” of the query gene
Multi Experiment Matrix - A web based tool for performing co-expression queries over large collection of gene expression datasets.

MEM - Multi Experiment Matrix

Enter gene ID(s) (for example: Jun, 203325_s_at, ENSG00000204531, ...)

Select collection

Options: Similarity Output Gene filters Dataset filters

Introduction | Quick-start | Examples | Tips&Tricks | Contact

MEM is a web-based multi experiment gene expression query and visualization tool. It gathers several hundreds of publicly available gene expression data sets from ArrayExpress database. Different data sets feature different tissues, diseases and conditions. For better compatibility and comparability data sets are arranged by the platform type.
Measure of similarity – co-expression

- Co-expression analysis in each dataset separately
- Aggregation of the results
Results

Handpicked datasets: ✓ = * ✓ = * ✓ = * ✓ = * reset all | 419 datasets excluded by filters
Results

Handpicked datasets: ✓: ✓ • ✓: ✓ • ✓: ✓ reset all | 419 datasets excluded by filters

E-GEOD-7069 – Zfx controls the self-renewal of embryonic and hematopoietic stem cells
Tags
Mus musculus Zfx null Zfx-flox Zfx-cko embryonic stem cell hematopoietic stem cell
StDev: 3.298 (over 8 arrays)

E-GEOD-10806 – Mouse iPS cells from neural stem cells by 2 factors (Oct4, KLF4) (Kim, Zaehres et al.)
Tags
Mus musculus
StDev: 3.094 (over 11 arrays)

E-GEOD-9629 – Comparison of normal and beta catenin deficient kidney gene expression profiles at E12.5
Tags
Mus musculus
StDev: 0.396 (over 6 arrays)
51 genes total. Used datasets:
1. microRNA-10a binds the 5' prime UTR of ribosomal protein mRNAs and enhances their translation
2. Translation state array analysis of Mouse embryonic stem cells and embryoid bodies
3. Gene expression analysis of embryonic stem cells expressing VE-cadherin (CD144) during endothelial differentiation
1. Functional annotations
2. Clustering the data
3. Affected pathways
4. Other microarray experiments
5. Network modules
Data as networks

- Genes as nodes
- `-omics` define edges
  - gene expression profiles (e.g. transcription factor knockouts)
  - transcription factor binding (ChIP-chip and ChIP-seq)
  - predicted cis-regulatory motifs
  - protein-protein interactions
  - genetic interactions
  - literature co-occurrence
Data as networks

- Genes as nodes
- *-omics define edges
  - gene expression profiles (e.g. transcription factor knockouts)
  - transcription factor binding (ChIP-chip and ChIP-seq)
  - predicted *cis*-regulatory motifs
  - protein-protein interactions
  - genetic interactions
  - literature co-occurrence
Weighing the evidence I

- Edges are not born equal
  - e.g. gene expression measurements: stronger vs weaker correlation

- Assign local weights to rank edges within a network

- Look for heavy subnetworks

SAGA and SWI/SNF chromatin remodelling complexes
yeast: DNA replication, stress response, transcription
human: SWI/SNF subunits act as tumour suppressors
Data as networks

.. everything appears interconnected

Public datasets for *S. cerevisiae*

- Gene expression data for 269 transcription factor (TF) knockouts (Hu et al. 2007)
- Binding profiles for 135 TFs (Harbison 2004)
- Predicted binding sites for 106 TFs (Maclsaac 2006)
- Protein-protein interactions (PPI) (Collins 2007, Yu 2008)
Finding the modules

- **Cliques**
  - Fully connected graphs -- e.g. protein complexes

- **Hubs**
  - Highly connected nodes -- e.g. transcriptional regulators

- **Sets of neighbours**
  - Specific genes of interest + near neighbours

- **Graph clustering**
  - MCL: Markov clustering (van Dongen, 2000)
  - Betweenness Centrality Clustering
Module evaluation

- Find enriched functions
  - Processes, functions, cell components: Gene Ontology
  - Pathways: KEGG and Reactome
  - cis-regulatory motifs: Transfac and miRBase

[SAGA and SWI/SNF chromatin remodelling and g:Profiler, Reimand et al (2007)]
Module evaluation II

Function $\alpha$ enriched in a module

$$p_\alpha = \sum_k \frac{\min(n, K)}{\binom{n}{k}} \frac{\binom{K}{N-K}}{\binom{N}{n-k}}$$

cumulative hypergeometric test

Functional relevance of a module

$$q = \sum_\alpha -\log (p_\alpha)$$

log sum of all related enrichments

$$q = \frac{1}{n} \sum_\alpha -\log (p_\alpha)$$

log sum of all related enrichments, normalised by module size
GraphWeb 1-2-3
Finding the modules

Network partitioning with MCL: Markov clustering (van Dongen, 2000)

- Gene expression data for 269 transcription factor (TF) knockouts (Hu et al 2007)
- Binding profiles for 135 TFs (Harbison 2004)
- Predicted binding sites for 106 TFs (Maclsaac 2006)
- Protein-protein interactions (Collins 2007, Yu 2008)
Finding the modules

Network partitioning with MCL: Markov clustering  
(van Dongen, 2000)

- Gene expression data for 269 transcription factor (TF) knockouts (Hu et al 2007)
- Binding profiles for 135 TFs (Harbison 2004)
- Predicted binding sites for 106 TFs (Maclsaac 2006)
- Protein-protein interactions (Collins 2007, Yu 2008)
Finding the modules

Network partitioning with MCL: Markov clustering (van Dongen, 2000)

- Gene expression data for 269 transcription factor (TF) knockouts (Hu et al 2007)
- Binding profiles for 135 TFs (Harbison 2004)
- Predicted binding sites for 106 TFs (Maclsaac 2006)
- Protein-protein interactions (Collins 2007, Yu 2008)
Finding the modules

Network partitioning with MCL: Markov clustering (van Dongen, 2000)

- Gene expression data for 269 transcription factor (TF) knockouts (Hu et al. 2007)
- Binding profiles for 135 TFs (Harbison 2004)
- Predicted binding sites for 106 TFs (Maclsaac 2006)
- Protein-protein interactions (Collins 2007, Yu 2008)
Proteasome module

Core complex:
Subunits RPN2/6/10/13;
YHBI -- oxidative and nitrosative stress response;
other proteasome subunits

RPN4 - regulator of proteasome genes;
regulated by stress;
negative feedback loop with proteasome;

Enriched annotations:
  Apoptosis,
cell cycle checkpoints,
DNA synthesis (Reactome);
Proteasome (KEGG);
  proteolysis,
peptidase activity (GO);
RPN4 motifs (Transfac)

Edges:
- Microarray
- TF motifs
- PPI
GraphWeb
http://biit.cs.ut.ee/graphweb

Input: a large network with experimental data
GraphWeb
http://biit.cs.ut.ee/graphweb

Input: a large network with experimental data
GraphWeb
http://biit.cs.ut.ee/graphweb

<table>
<thead>
<tr>
<th>Module</th>
<th>#Nodes</th>
<th>#Edges</th>
<th>Density</th>
<th>Nodes</th>
<th>Edges</th>
<th>Zoom in</th>
<th>Label distribution</th>
<th>Score</th>
<th>g:Profiler annotations</th>
<th>Visual</th>
<th>SIF export</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1056</td>
<td>1427</td>
<td>0.1%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td>1</td>
<td>1.9e-11 GO:BP cellular process</td>
<td>N/A</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.7e-19 GO:CC intracellular part</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.0e-07 GO:MF protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.9e-06 REACTOME Cooperation of Prefoldin and ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.65e-114 TRANSFAC NNWNRCTCAYTNANMGDANNN:4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>execute g:Profiler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>1052</td>
<td>1847</td>
<td>0.2%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td>0</td>
<td>2.23e-09 GO:BP cytoskeleton organization</td>
<td>N/A</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.29e-10 GO:CC intracellular</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.70e-09 GO:MF protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.10e-55 TRANSFAC NTTACCGG:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>execute g:Profiler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>183</td>
<td>340</td>
<td>1.0%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td>1</td>
<td>1.05e-10 GO:BP microtubule-based process</td>
<td>compact</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.38e-09 GO:CC microtubule cytoskeleton ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.27e-05 KEGG Cell cycle - yeast</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.10e-07 TRANSFAC TTWCYAWNNNGAAANNNGTAAYA:4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>execute g:Profiler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#4</td>
<td>172</td>
<td>571</td>
<td>1.9%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td>1</td>
<td>1.23e-05 GO:BP translation</td>
<td>compact</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03e-07 GO:CC ribosome</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>execute g:Profiler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#5</td>
<td>157</td>
<td>212</td>
<td>0.9%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td></td>
<td></td>
<td>compact</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#6</td>
<td>119</td>
<td>158</td>
<td>1.1%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td></td>
<td></td>
<td>compact</td>
<td>SIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#7</td>
<td>104</td>
<td>146</td>
<td>1.4%</td>
<td></td>
<td></td>
<td>Send to input</td>
<td></td>
<td></td>
<td></td>
<td>compact</td>
<td>SIF</td>
</tr>
</tbody>
</table>

Output: a list of functionally significant gene modules
GraphWeb
http://biit.cs.ut.ee/graphweb

Module: 1

Nodes: 52

Edges: 140

Density: 5.3%

Evaluate edge content and support from datasets

Node/edge statistics & export

Gene expression

Protein interaction

Protein-protein interaction & TF binding site

Visuals & export to Cytoscape

Functional annotations

Score:

1.78e-20 GO:BP modification-dependent ...
4.66e-20 GO:CC proteasome complex...
4.41e-10 GO:MF endopeptidase activity...
1.07e-21 KEGG Proteasome...
1.06e-17 REACTOME G1/S DNA Damage Checkpoints...

execute g:Profiler

compact labeled SIF

export
Everything is connected

- Microarray experiment
- GraphWeb
- g:Profiler
- VisHiC
- Multi Experiment Matrix
- KEGGAnim
g:Profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments

Jüri Reimand¹, Meelis Kull¹,²,³, Hedi Peterson²,³, Jaanus Hansen¹ and Jaak Vilo¹,²,³,*

VisHiC — hierarchical functional enrichment analysis of microarray data

Darya Krushevskaya¹,², Hedi Peterson³,⁴, Jüri Reimand¹, Meelis Kull¹,⁴ and Jaak Vilo¹,²,⁴,*

KEGGanim: pathway animations for high-throughput data

Priot Adler a*, Jüri Reimand b*, Jürgen Jänes b, Raivo Kolde c, Hedi Peterson ac, Jaak Vilo abc†

a Estonian Biocentre, Riia 23b, Tartu, Estonia b University of Tartu, Institute of Computer Science, Liivi 2, Tartu, Estonia c QureTec Inc. Ülikooli 6a, Tartu, Estonia

Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods

Priot Adlera*, Raivo Koldea††, Meelis Kull††, Aleksandr Tkachenko††, Hedi Peterson††, Jüri Reimand† and Jaak Vilo††

GraphWeb: mining heterogeneous biological networks for gene modules with functional significance

Jüri Reimand¹,², Laur Tooming¹, Hedi Peterson³,⁴, Priot Adler³ and Jaak Vilo¹,⁴,*

http://biit.cs.ut.ee/software
Hands on exercises

GO to:
http://biit.cs.ut.ee/software/tutorials

open:
2010_11_23_coh_exercises.pdf

Questions:
peterson@quaretec.com