Wide-scale Comparison of Transcriptome Data

Raymond Lim, Pavlidis Lab
Contextualize Studies

Huntington's Disease Study

Other Huntington's Disease Studies

Unrelated Studies
Background: Microarray Study

- Profile expression of thousands of genes
- Differential expression between groups
  - e.g. control vs. treatment
  - Potentially complicated design
- Different platforms
  - Versions
  - Manufacturers, e.g. Affymetrix, Illumina
Overview

<table>
<thead>
<tr>
<th>Gene</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pzp</td>
<td>11586</td>
</tr>
<tr>
<td>Aanat</td>
<td>12671</td>
</tr>
<tr>
<td>Aatk</td>
<td>1022</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>Dataset A Rank</th>
<th>Dataset B Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pzp</td>
<td>334</td>
<td>6721</td>
</tr>
<tr>
<td>Aanat</td>
<td>752</td>
<td>384</td>
</tr>
<tr>
<td>Aatk</td>
<td>76</td>
<td>103</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>0.4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Challenges

• Measuring similarity between gene lists
  – Noisy data

• Validation of results
  – Gold standard

• Confounds
  – Platform effects
  – Difference in statistical power
Data Overview

• Gemma
  – Framework for meta-analysis of gene expression data
  – Automated differential expression analyses
    • Not published results

• Annotations

<table>
<thead>
<tr>
<th>Taxon</th>
<th>NumSamples</th>
<th>Platforms</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>human:224</td>
<td>Min. : 4.00</td>
<td>GPL1261:137</td>
<td>Treatment : 44</td>
</tr>
<tr>
<td></td>
<td>1st Qu.: 12.00</td>
<td>GPL570 : 79</td>
<td>time : 30</td>
</tr>
<tr>
<td></td>
<td>Median : 18.00</td>
<td>GPL81 : 70</td>
<td>DiseaseState : 29</td>
</tr>
<tr>
<td></td>
<td>Mean : 26.46</td>
<td>GPL96 : 46</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td>3rd Qu.: 29.00</td>
<td>GPL339 : 32</td>
<td>Genotype : 22</td>
</tr>
<tr>
<td></td>
<td>Max. :280.00</td>
<td>GPL260 : 29</td>
<td>Genotype</td>
</tr>
<tr>
<td>mouse:349</td>
<td></td>
<td>(Other):180</td>
<td>(Other) :401</td>
</tr>
</tbody>
</table>
Dataset Annotations

- Automated and manual
- Manually clustered subset
- Measure similarity
- Semi-gold standard

<table>
<thead>
<tr>
<th></th>
<th>Brain</th>
<th>Gene X</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset A</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dataset B</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dataset C</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Measuring Gene Signature Similarity

- **Rank based**
  - Spearman, Kendall’s tau
  - Bottom of list noisy
- **Threshold based**
  - Gene overlap
    - Fisher’s exact test
- **Threshold/rank based**
  - ROC method
  - Top-k Kendall
- **Ontology based**
  - Low Coverage
  - Redundancy in GO
    - Filtering
  - Low resolution
  - Performs best

Red: Below threshold

Two colon cancer datasets: GSE12762 (ref) vs. GSE11578

\[
\begin{align*}
\text{Gene A} & \quad \text{Gene D} \\
\text{Gene B} & \quad \text{Gene A} \\
\text{Gene C} & \quad \text{Gene B} \\
\text{Gene D} & \quad \text{Gene E} \\
\text{Gene E} & \quad \text{Gene F} \\
\text{Gene F} & \quad \text{Gene C}
\end{align*}
\]
Measuring Gene Signature Similarity

- Rank based
  - Spearman, Kendall’s tau
  - Bottom of list noisy
- Threshold based
  - Gene overlap
    - Fisher’s exact test
- Threshold/rank based
  - ROC method
  - Top-k Kendall
- Ontology based
  - Low coverage
  - Redundancy in GO
    - Filtering
  - Low resolution

Two colon cancer datasets: GSE12762 (ref) vs. GSE11578
AUC = 0.62

Gene A
Gene B
Gene C
Gene D
Gene E
Gene F

Red: Below threshold
Validation

Gene Expression Signatures

Annotations e.g. brain, kidney, cancer

Pair-wise Comparison

Similarity Scores

Comparison

Annotations e.g. brain, kidney, cancer

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Example

Huntington’s Disease Study

Other Huntington’s Disease Studies

Unrelated Studies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE3634</td>
<td>Molecular pathways involved in neuronal degeneration of polyglutamine mouse models</td>
</tr>
<tr>
<td>GSE3621</td>
<td>R6/1 brain hemisphere time series gene expression</td>
</tr>
<tr>
<td>GSE2161</td>
<td>Identification of genes that are dysregulated in the telencephalon of Dlx1/2 mutants.</td>
</tr>
<tr>
<td>GSE6678</td>
<td>Palmitoyl protein thioesterase-1 knockout mice</td>
</tr>
</tbody>
</table>
Platforms Cluster

- Intersection of genes (~7000)
- GPL1261: Affymetrix Mouse 430 2.0
- GPL81: Affymetrix Mouse U74 2.0
Studies Differ in Statistical Power

- Widely varying numbers of differentially expressed genes
- Datasets with similar power tend to cluster together
Concluding Remarks and Future Direction

• Provide context for gene lists
• Develop better gold standard
• Manual quality control
• Investigate platform effect
• Gene dynamics
• Identify modules of differential expression
Acknowledgements

• Paul Pavlidis
• Leon French
• Artemis Lai
• Members of Pavlidis Lab