Challenges of integrating and analyzing different genomic data types

Jim Cavalcoli
University of Michigan
ISMB 2012 Bioinformatics Core Workshop
Outline

• Technological Challenges – Different platforms, different software and methods yields variation in results *(for Discussion)*

• Comparing and Integrating data from different methods and within and between samples

• Tools for adding biological relevance and integrating data
Method Comparisons

• Different platforms – same(?) biological measurement
  – RNA-Seq v. Microarray
  – Exome Variants v SNP arrays

• Same Samples – Different Biological Measurement
  – RNA expression & TF Binding
  – Genome Variants & Metabolite changes
Challenges in Comparing Across Experiments

• Same Samples, but looking across different platforms
  • “From the same tissues/treatments we want to examine RNA-Seq and ChIP for TF SP1”
  • Or RNA-Seq and Methylation or Histone Binding profile

• **What RNA-seq comparison is best to integrate with the ChIP-seq data?** For the transcription factor studied, we had RNA-seq data for WT versus KO of the TF. We also had untreated WT versus WT with a treatment that stimulated the TF activity.
  – Although it might be expected that the WT versus KO for the TF would have served as the best comparison to integrated with the ChIP-seq data, it turned out that the WT treated versus untreated correlated better with the ChIP-Seq results.
Challenges Cont’d

• **What peak regions regulate which gene(s)?** A lot of ChIP-Seq bindings do not apparently regulate any genes. They may serve a different purpose, no purpose in the context under study, or they might not even be functional.

• Some genes share promoters, and some peaks are far from any gene.

• Given all these complexities, it’s difficult to know which gene(s), if any, a peak should be paired with.

• Typically, the further you get from a gene’s TSS, the less likely the TF is to regulate the corresponding gene. However, the fall off appears to be different for different TFs.
TF binding sites near TSSs correspond to differential expression

Gene rank by distance between peak and TSS

Expression change compared to knockout:
- No change
- Reduction
- Induction

% genes significantly expressed

Gene rank by distance between peak and TSS

0          20        40         60         80        100

1k 2k 3k 5k 10k 20k
Integrating Biological Relevance

- Geneset enrichment – consensus by crowd
  - What sets are meaningful?
- Categorization and clustering – guilt by association
  - What level of binning will be helpful?
- Linking into Literature reference information
  - Accuracy of what is published? And NLP derived?
- Workflows to maintain consistency and provenance *(outside the scope of today’s talk)*
Gene Set Enrichment

- GSEA (Broad Institute)
  - http://www.broadinstitute.org/gsea/index.jsp
- LRPath (NCIBI)
  - http://lrpath.ncibi.org/
- DAVID
  - http://david.abcc.ncifcrf.gov/home.jsp
- ConceptGen
  - http://Conceptgen.ncibi.org
Cytoscape Plug-in: Metscape

- Provide the context for experimental data
- Utilize prior knowledge of metabolic networks
- Display multiple measurements across observations, time points, experimental conditions etc.
- Integrate multidimensional data
  - Can be gene expression and metabolomics data
- Provide broader view of metabolic networks
- Link to diseases


http://metscape.ncibi.org/
### Metab2MeSH – Compound Annotation with MeSH Terms

**http://metab2mesh.ncbi.nlm.org**


#### Compound search examples:
- methylmalonic acid
- glucose-6-phosphate

**MeSH term search examples:**
- diabetes mellitus
- metabolism, inborn errors

Trouble finding the right MeSH term? Check the MeSH browser.

- "cardiovascular diseases"

#### Filter by top level MeSH Heading:

---Show All---

<table>
<thead>
<tr>
<th>Compound Name</th>
<th>Compound ID(s)</th>
<th>MeSH Heading(s)</th>
<th>MeSH Descriptor</th>
<th>MeSH Qualifier</th>
<th>PubMed Articles*</th>
<th>P-Value</th>
<th>Q-Value</th>
<th>Fold Change</th>
<th>ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Homocysteine</td>
<td>91552</td>
<td>Diseases</td>
<td>Cardiovascular Diseases</td>
<td>-</td>
<td>1031</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>25.5</td>
<td>24022.0</td>
</tr>
<tr>
<td>Dl-Homocysteine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-ammonio-4-sulfanyl</td>
<td>446925</td>
<td>Chemicals and Drugs</td>
<td>Carotenoids</td>
<td>-</td>
<td>1422</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>645.8</td>
<td>910388.2</td>
</tr>
<tr>
<td>(25)-2-agaraniylo-4-sulfanyl</td>
<td>446922</td>
<td>Chemicals and Drugs</td>
<td>Beta Carotenoids</td>
<td>-</td>
<td>206</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>390.9</td>
<td>195022.3</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Organisms Technology, Industry, Agriculture</td>
<td>Lycopersicon esculentum</td>
<td>-</td>
<td>234</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>541.7</td>
<td>178117</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446923</td>
<td>Chemicals and Drugs</td>
<td>Lutein</td>
<td>-</td>
<td>215</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>728.4</td>
<td>154005</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Chemicals and Drugs</td>
<td>Xanthophylls</td>
<td>-</td>
<td>200</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>482.2</td>
<td>96048.9</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Chemicals and Drugs</td>
<td>Anticarcinogenic Agents</td>
<td>-</td>
<td>231</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>166.3</td>
<td>37464.5</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446923</td>
<td>Chemicals and Drugs</td>
<td>Antioxidants</td>
<td>-</td>
<td>569</td>
<td>0.00e-1</td>
<td>0.00e-1</td>
<td>49.7</td>
<td>26833.8</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Chemicals and Drugs</td>
<td>Vitamin E</td>
<td>-</td>
<td>188</td>
<td>1.00e-21</td>
<td>8.77e-21</td>
<td>36.8</td>
<td>6515.5</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Diseases</td>
<td>Prostatic Neoplasms</td>
<td>-</td>
<td>168</td>
<td>2.95e-10</td>
<td>2.21e-10</td>
<td>32.6</td>
<td>5091.4</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Anatomy Technology, Industry, Agriculture</td>
<td>Fruit</td>
<td>-</td>
<td>137</td>
<td>1.51e-18</td>
<td>1.13e-18</td>
<td>60.2</td>
<td>7805.5</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Chemicals and Drugs</td>
<td>Vitamin A</td>
<td>-</td>
<td>144</td>
<td>5.24e-13</td>
<td>3.47e-13</td>
<td>33.3</td>
<td>4486.1</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446923</td>
<td>Technology, Industry, Agriculture</td>
<td>Dietary Supplements</td>
<td>-</td>
<td>120</td>
<td>2.91e-16</td>
<td>1.64e-16</td>
<td>33.8</td>
<td>3790.7</td>
</tr>
<tr>
<td>Lycopene</td>
<td>446925</td>
<td>Organisms Technology, Industry, Agriculture</td>
<td>Venenables</td>
<td>-</td>
<td>84</td>
<td>1.09e-11</td>
<td>5.26e-11</td>
<td>58.7</td>
<td>4704.7</td>
</tr>
</tbody>
</table>
Web Services for NCIBI Tools

http://ws.ncibi.org/

• Data Services
  – Natural Language Processing Pipeline for PubMed and PMCOA
  – Gene2MeSH
  – Metab2MeSH
  – Michigan Molecular Interactions Database (MiMI)
  – Metabolomics

• Computational Analysis Services
  – Natural Language Processing
    • Sentence Segmentation
    • Phrase Structure Parsing
  – Gene Set Enrichment Analysis
    • LRPath
    • ThinkBACK
In Summary

• There are many remaining challenges in Technical areas yet the potential benefits to science and medicine are huge.

• It’s very important to harness all the knowledge that’s out there and this comes from integrating multiple data and annotation streams.