Google Summer of Code 2014: Connect PathVisio and WikiPathways to Expression Atlas

Jonathan Melius
Introduction

The core idea:

• Build a first PathVisio plugin that connects to Expression Atlas.
• Next focus is a plugin on human tissue data available from Expression Atlas.
• As a final step in the project, the 'tissue' plugin is used to on WikiPathways.
Contents

I. AtlasPlugin

II. TissueAnalyzer

III. Connect from WikiPathways
Context - PathVisio

• Open-source pathway analysis and drawing software

www.pathvisio.org
Context - WikiPathways

• Open collaborative platform dedicated to the curation of biological pathways

www.wikipathways.org/
Context - Expression Atlas

• Microarray and RNA-sequencing experiments from ArrayExpress manually curated, annotated and standardized

www.ebi.ac.uk/gxa
I. AtlasPlugin

Conception

• The user give the experiment(s) id
• Connect to Expression Atlas via RDF
• Download the dataset
• Parsing all the informations necessary (gene identifier, p-value, t-statistic)
I. AtlasPlugin
PathVisio Module

• Pass all the information through the importation module
• Create automatic visualisation
• Comprehensible data expression table
I. AtlasPlugin

RDF Schema

Gene Expression Atlas RDF schema

Namespaces:
- atlas: <http://rdf.ebi.ac.uk/terms/atlas>
- efo: <http://www.obo-owl.org/efo/>
- vocab: <http://purl.obolibrary.org/obo/vocab/>
- xsd: <http://www.w3.org/2001/XMLSchema#>
- miriam: <http://identifiers.org/>
- foaf: <http://xmlns.com/foaf/0.1/>
- dcterms: <http://purl.org/dc/terms/>
- swo: <http://www.ebi.ac.uk/swo/>
I. AtlasPlugin Wizard

![Atlas Expression data import wizard](image)

Choose an experiments and the file locations

- **Experiment ID**: E-GEOID-18842 E-GEOID-6731 E-GEOID-8977
- **Output file**: abc
- **Gene database**: idmapper-pgdb:/home/mael/Pathy/Hs_Derby_20130701.bridge

[Back] [Next] [Cancel]
## I. AtlasPlugin Preview

### Choose column types

**Select primary identifier column:** Gene_ID

**Use the same system code for all rows:** Ensembl

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gene_ID</td>
<td>E-GEOD-18842-probe</td>
<td>E-GEOD-18842-propertyValue</td>
<td>E-GEOD-18842-pValue</td>
<td>E-GEOD-18842-...</td>
</tr>
<tr>
<td>2</td>
<td>ENSG00000144909</td>
<td>218304_s_at</td>
<td>non-small cell lung cancer</td>
<td>3.3279266E-6</td>
<td>-5.3317847</td>
</tr>
<tr>
<td>3</td>
<td>ENSG00000169946</td>
<td>219778_at</td>
<td>normal</td>
<td>6.7637376E-7</td>
<td>5.732229</td>
</tr>
<tr>
<td>4</td>
<td>ENSG00000164330</td>
<td>229487_at</td>
<td>non-small cell lung cancer</td>
<td>5.3668773E-6</td>
<td>-5.2085423</td>
</tr>
<tr>
<td>5</td>
<td>ENSG00000164330</td>
<td>229487_at</td>
<td>normal</td>
<td>5.3668773E-6</td>
<td>5.2085423</td>
</tr>
</tbody>
</table>
I. AtlasPlugin Visualisation
I. AtlasPlugin

Expression data

GeneProduct annotation

Name: CFD
Identifier: 1675
Database: Entrez Gene
Synonyms: ADN

Expression data

<table>
<thead>
<tr>
<th>Identifier</th>
<th>ENSG00000197766</th>
<th>ENSG00000197766</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOID-18842-probe</td>
<td>205382_s_at</td>
<td>205382_s_at</td>
</tr>
<tr>
<td>E-GEOID-18842-propertyValue</td>
<td>normal</td>
<td>non-small cell lung cancer</td>
</tr>
<tr>
<td>E-GEOID-18842-pValue</td>
<td>1.0667746E-26</td>
<td>1.0667746E-26</td>
</tr>
<tr>
<td>E-GEOID-18842-stat</td>
<td>16.387613</td>
<td>-16.387613</td>
</tr>
<tr>
<td>E-GEOID-8977-probe</td>
<td>205382_s_at</td>
<td>205382_s_at</td>
</tr>
<tr>
<td>E-GEOID-8977-propertyValue</td>
<td>normal</td>
<td>invasive ductal carcinoma</td>
</tr>
<tr>
<td>E-GEOID-8977-pValue</td>
<td>2.243653E-4</td>
<td>2.243653E-4</td>
</tr>
<tr>
<td>E-GEOID-8977-stat</td>
<td>6.7572465</td>
<td>-6.7572465</td>
</tr>
</tbody>
</table>
I. AtlasPlugin

Discussion

• RDF changes required
• Browser for the experiment / Querying
• Connection to statistics module
II. TissueAnalyzer Datasets

- REST interface
- E-MTAB-1733: RNA-seq of coding RNA representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes
- E-MTAB-513: RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)
II. TissueAnalyzer Datasets analysis

Density plot Inter: Illumina

Density plot Inter: 27 tissues

N = 33683  Bandwidth = 0.5

N = 33687  Bandwidth = 0.5
II. TissueAnalyzer Conception

- User chooses one of the tissue dataset
- Download the whole dataset
- Or filtered by tissue selection
- Provide a tissue side panel
II. TissueAnalyzer Selection
II. TissueAnalyzer Side Panel
II. TissueAnalyzer
Common pathways
II. TissueAnalyzer Discussion

• More advance statistics methods
• Not currently available in RDF
• User Feedback
III. Connect from WikiPathways Conception

• Create *TissueAnalyzer* standalone to calculate statistics for pathways
• Integrate *TissueAnalyzer* in Wikipathways
• Build a user interface component for the WikiPathways web portal
III. Connect from WikiPathways Tissue page

- **TissueAnalyzer** result from E-MTAB-1733 dataset
- Ranking pathways for each tissue

http://test2.wikipathways.org/index.php/Special:TissueAnalyzer
III. Connect from WikiPathways
Tissue page selection

Select tissue: adipose tissue  submit

☐ Show common pathways

human body features
### III. Connect from WikiPathways

**Tissue selected**

Select tissue: [heart]  [submit]

- [ ] Show common pathways

**Gradient color scale**

<table>
<thead>
<tr>
<th>Viewer</th>
<th>Pathway name</th>
<th>Median expression</th>
<th>Mean expression</th>
<th>Ratio active gene measured</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Striated_Muscle_Contraction</td>
<td>7.57</td>
<td>6.83</td>
<td>20/38</td>
<td>52</td>
</tr>
<tr>
<td>1</td>
<td>TCA_Cycle_Copy_Test28</td>
<td>6.88</td>
<td>6.79</td>
<td>11/17</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>Glycolysis_and_Gluconeogenesis</td>
<td>5.93</td>
<td>4.74</td>
<td>18/49</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>Cori_Cycle</td>
<td>5.81</td>
<td>5.38</td>
<td>5/15</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>Arachidonate_Epoxidase_-_Epoxide_Hydrolase</td>
<td>5.67</td>
<td>4.86</td>
<td>3/7</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>Glycogen_Metabolism</td>
<td>4.86</td>
<td>4.64</td>
<td>11/43</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>Synthesis_and_Degradation_of_Ketone_Bodies</td>
<td>4.81</td>
<td>4.77</td>
<td>2/5</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>miR-targeted_genes_in_adipocytes_-_TarBase</td>
<td>4.81</td>
<td>4.62</td>
<td>1/17</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Translation_Factors</td>
<td>4.81</td>
<td>5.15</td>
<td>11/50</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>Dopamine_metabolism</td>
<td>4.75</td>
<td>3.65</td>
<td>2/13</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>Sulindac_Metabolic_Pathway</td>
<td>4.7</td>
<td>3.9</td>
<td>0/5</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>Physiological_and_Pathological_Hypertrophy_of_the_Heart</td>
<td>4.48</td>
<td>4.38</td>
<td>4/26</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>Complement_Activation_Classical_Pathway</td>
<td>4.46</td>
<td>3.39</td>
<td>2/17</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>Diurnally_Regulated_Genes_with_Circadian_Orthologs</td>
<td>4.46</td>
<td>4.21</td>
<td>6/48</td>
<td>12</td>
</tr>
</tbody>
</table>
III. Connect from WikiPathways

Tissue page viewer

Highlighting legend
- Active gene (expression > 6)
- Non-active gene (expression < 6)

View at WikiPathways
III. Connect from WikiPathways

Pathway page

TCA Cycle (Homo sapiens)
Alexander Pico, Martijn van Iersel, Thomas Keider, Kdahiquist, et al.

Curation Tags

Description
The citric acid cycle, also known as the tricarboxylic acid cycle (TCA cycle) or the Krebs cycle, is a series of enzyme-catalysed chemical reactions of central importance in all living cells that use oxygen as part of cellular respiration. In eukaryotes, the citric acid cycle occurs in the matrix of the mitochondria. The components and reactions of the citric acid cycle were established by seminal work from both Hans Krebs and Hans Otto. Source: Wikipedia [wikipedia:citric_acid_cycle]

Ontology Tags
Pathway Ontology: TCA cycle, citrate cycle pathway
III. Connect from WikiPathways Pathway page tissue table

### Tissue Table

<table>
<thead>
<tr>
<th>Tissue name</th>
<th>Median expression</th>
<th>Mean expression</th>
<th>Active gene measured (%)</th>
<th>Ensembl active gene list (data base link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart</td>
<td>6.88</td>
<td>6.79</td>
<td>64.71</td>
<td>ACO2 OGDH IDH2 IDH3A SUCLG1 MDH2 SDHA CS DLD SDHB SDHD</td>
</tr>
<tr>
<td>kidney</td>
<td>6.51</td>
<td>6.43</td>
<td>41.18</td>
<td>SUCLG2 ACO2 IDH2 SUCLG1 SDHC SDHB SDHD</td>
</tr>
<tr>
<td>duodenum</td>
<td>6.29</td>
<td>6.31</td>
<td>35.29</td>
<td>SUCLG2 ACO2 SUCLG1 SDHA SDHB SDHD</td>
</tr>
<tr>
<td>colon</td>
<td>6.17</td>
<td>6.08</td>
<td>23.53</td>
<td>SUCLG2 SUCLG1 MDH2 SDHD</td>
</tr>
<tr>
<td>small intestine</td>
<td>6.07</td>
<td>6.14</td>
<td>23.53</td>
<td>SUCLG2 ACO2 SUCLG1 SDHD</td>
</tr>
</tbody>
</table>
III. Connect from WikiPathways

Discussion

• Integration new statistical methods when available
• Add body map to the pathway page
• Better way to define common pathways
Summary
Conclusion

• Two new PathVisio plugins
• Atlas plugin
  - Differential expression datasets
  - RDF
  - Automatic visualisation
  - TODO: include not significant genes & logFC
    → enable pathway statistics
Conclusion

• TissueAnalyzer
  - Pathway statistics (median, mean, percentage) for tissue activity
  - Automatic visualisation
  - Standalone version for integration into WikiPathways
  - TODO: more advanced statistical methods data retrieval from RDF
Conclusion

• WikiPathways
  – Tissue page: find active pathways for tissues
  – Pathway page: show tissue expression for pathways
  – TODO: body map on the pathway page
Merci pour votre attention