Data Integration: hands-on session

July 21, 2017
Karan Uppal, PhD
Assistant Professor of Medicine
Director of Computational Metabolomics & Integrative Omics,
Clinical Biomarkers Laboratory
Emory University
1. MetaboAnalyst: Integrated Pathway Analysis

URL: http://www.metaboanalyst.ca/
Click on “Click here to start”
Select the “Integrated Pathway Analysis” module
Data upload – select the “Use our example data” option

<table>
<thead>
<tr>
<th>Gene List with optional fold changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Entrez</td>
</tr>
<tr>
<td>1737</td>
</tr>
<tr>
<td>83440</td>
</tr>
<tr>
<td>3939</td>
</tr>
<tr>
<td>10911</td>
</tr>
<tr>
<td>10690</td>
</tr>
<tr>
<td>10010</td>
</tr>
<tr>
<td>11224</td>
</tr>
<tr>
<td>63826</td>
</tr>
<tr>
<td>11031</td>
</tr>
<tr>
<td>4190</td>
</tr>
<tr>
<td>10782</td>
</tr>
<tr>
<td>10993</td>
</tr>
<tr>
<td>10455</td>
</tr>
<tr>
<td>10963</td>
</tr>
<tr>
<td>10282</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metabolite List with optional fold changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>#KEGG</td>
</tr>
<tr>
<td>C00116</td>
</tr>
<tr>
<td>C00586</td>
</tr>
<tr>
<td>C00033</td>
</tr>
<tr>
<td>C00583</td>
</tr>
<tr>
<td>C00222</td>
</tr>
<tr>
<td>C00719</td>
</tr>
<tr>
<td>C05984</td>
</tr>
<tr>
<td>C00207</td>
</tr>
<tr>
<td>C00065</td>
</tr>
<tr>
<td>C00031</td>
</tr>
<tr>
<td>C00079</td>
</tr>
<tr>
<td>C02632</td>
</tr>
<tr>
<td>C00664</td>
</tr>
<tr>
<td>C00114</td>
</tr>
<tr>
<td>C00073</td>
</tr>
</tbody>
</table>

Specify organism: Homo sapiens (human)

Use our example data

Submit
Enrichment Analysis

Enrichment analysis aims to evaluate whether the observed genes and metabolites in a particular pathway are significantly enriched (appear more than expected by random chance) within the dataset. You can choose over-representation analysis (ORA) based on either hypergenomics analysis or Fisher's exact method.

- Hypergeometric Test
- Fisher's Exact Test

Topology Analysis

The topology analysis aims to evaluate whether a given gene or metabolite plays an important role in a biological response based on its position within a pathway. **Degree Centrality** measures the number of links that connect to a node (representing either a gene or metabolite) within a pathway; **Closeness Centrality** measures the overall distance from a given node to all other nodes in a pathway; **Betweenness Centrality** measures the number of shortest paths from all nodes to all the others that pass through a given node within a pathway.

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality

Pathway Databases

Users can choose one of three different modes of pathways: - the gene-metabolite mode (default) allows joint-analysis and visualization of both significant genes and metabolites; while the gene-centric or metabolite-centric mode allows users to identify enriched pathways driven by significant genes or metabolites, respectively.

- Gene-metabolite pathways
- Gene-centric pathways
- Metabolite-centric pathways
Results
2. 3Omics - homepage
Features

• Correlation analysis and network visualization
  – Pairwise Pearson correlation analysis

• Database-derived relationships in correlation analysis
  – Uses an internal database based on NCBI Entrez gene, Uniprot proteins, and KEGG metabolites to determine gene-protein-metabolite relationship

• Coexpression analysis
  – Two-way hierarchical clustering analysis
  – Rows: variables (Genes + proteins + metabolites, genes+metabolites, etc.)
  – Columns: samples

• Phenotype analysis
  – Uses OMIM databases to link genes with phenotypes

• Pathway and Gene Ontology Enrichment analysis
  – Using KEGG, HumanCyc, and DAVID
Data upload

Example: 3omics.cmdm.tw/example/T_example.csv
Example: 3omics.cmdm.tw/example/P_example.csv
Example: 3omics.cmdm.tw/example/M_example_pubchem.csv
# Data format

(http://3omics.cmdm.tw/help.php#examples)

## Samples

<table>
<thead>
<tr>
<th>Variables</th>
<th>timepoint1</th>
<th>timepoint2</th>
<th>timepoint3</th>
<th>timepoint4</th>
<th>timepoint5</th>
</tr>
</thead>
<tbody>
<tr>
<td>akap9</td>
<td>-0.24</td>
<td>-0.6</td>
<td>-0.47</td>
<td>-0.38</td>
<td>-0.31</td>
</tr>
<tr>
<td>macf1</td>
<td>-0.3</td>
<td>-0.3</td>
<td>0.48</td>
<td>0.07</td>
<td>-0.36</td>
</tr>
<tr>
<td>RNPEP</td>
<td>0.24</td>
<td>0.85</td>
<td>0.15</td>
<td>0.79</td>
<td>0.69</td>
</tr>
<tr>
<td>SDHA</td>
<td>0.1</td>
<td>0.37</td>
<td>0.18</td>
<td>0.23</td>
<td>0.33</td>
</tr>
<tr>
<td>EEF1B2</td>
<td>-0.04</td>
<td>-0.31</td>
<td>0.06</td>
<td>-0.39</td>
<td>-0.46</td>
</tr>
<tr>
<td>EEF1D</td>
<td>0.07</td>
<td>0.29</td>
<td>0.22</td>
<td>0.75</td>
<td>0.47</td>
</tr>
<tr>
<td>EIF4A1</td>
<td>0.42</td>
<td>0.65</td>
<td>0.66</td>
<td>0.97</td>
<td>0.78</td>
</tr>
<tr>
<td>WARS</td>
<td>1.47</td>
<td>1.72</td>
<td>0.58</td>
<td>1.79</td>
<td>1.69</td>
</tr>
<tr>
<td>G3BP2</td>
<td>0.15</td>
<td>0.09</td>
<td>0.1</td>
<td>0.2</td>
<td>-0.22</td>
</tr>
<tr>
<td>PAK2</td>
<td>-0.21</td>
<td>-0.14</td>
<td>-0.15</td>
<td>-0.31</td>
<td>-0.4</td>
</tr>
<tr>
<td>PPP4C</td>
<td>-0.13</td>
<td>0.05</td>
<td>-0.09</td>
<td>0.21</td>
<td>-0.12</td>
</tr>
<tr>
<td>ZNF224</td>
<td>-0.06</td>
<td>0.31</td>
<td>0.17</td>
<td>0.27</td>
<td>0.61</td>
</tr>
<tr>
<td>ZNF268</td>
<td>-0.23</td>
<td>0.08</td>
<td>0.01</td>
<td>0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>TRRAP</td>
<td>0.07</td>
<td>-0.12</td>
<td>0.41</td>
<td>0.45</td>
<td>-0.09</td>
</tr>
<tr>
<td>RAD23B</td>
<td>-0.07</td>
<td>-0.32</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.44</td>
</tr>
<tr>
<td>TARDBP</td>
<td>0.23</td>
<td>0.18</td>
<td>0.39</td>
<td>0.63</td>
<td>0.23</td>
</tr>
<tr>
<td>CSTF2</td>
<td>0.51</td>
<td>0.65</td>
<td>0.71</td>
<td>1.18</td>
<td>0.89</td>
</tr>
<tr>
<td>PSMC2</td>
<td>0.82</td>
<td>0.57</td>
<td>1.15</td>
<td>1.75</td>
<td>0.58</td>
</tr>
<tr>
<td>F8</td>
<td>-0.19</td>
<td>-0.02</td>
<td>-0.35</td>
<td>-0.82</td>
<td>-0.81</td>
</tr>
<tr>
<td>MYOM1</td>
<td>-0.28</td>
<td>-0.29</td>
<td>-0.54</td>
<td>-1.06</td>
<td>-1.03</td>
</tr>
<tr>
<td>ACTR3</td>
<td>0.57</td>
<td>0.48</td>
<td>0.39</td>
<td>0.32</td>
<td>0.72</td>
</tr>
<tr>
<td>ITPR2</td>
<td>0.62574</td>
<td>1.771</td>
<td>-0.057392</td>
<td>1.2612</td>
<td>1.7769</td>
</tr>
<tr>
<td>NUCB2</td>
<td>-1.1943</td>
<td>-0.96016</td>
<td>-0.71549</td>
<td>-1.1877</td>
<td>-0.70604</td>
</tr>
<tr>
<td>CAMK1</td>
<td>0.33342</td>
<td>0.87499</td>
<td>0.059355</td>
<td>0.062122</td>
<td>0.53605</td>
</tr>
<tr>
<td>BCL2A1</td>
<td>2.2913</td>
<td>3.8479</td>
<td>-0.12343</td>
<td>1.6604</td>
<td>3.3933</td>
</tr>
<tr>
<td>PDCD6IP</td>
<td>0.46362</td>
<td>0.88049</td>
<td>0.20539</td>
<td>0.36177</td>
<td>0.62012</td>
</tr>
</tbody>
</table>
Select checkbox next to “Use example data”
Wait for upload to complete
View Results – Correlation Network
View Results – Co-expression analysis

Rows: Variables
Columns: Samples

Co-expression analysis of Transcriptomics, Proteomics & Metabolomics

Summary of Input molecules

Cluster:

Molecules
- IL4 (P05112)
- L-Leucine (HMO:044309, PubchemCID:6106)
- ACTR3
- Doxorubicin (HMO:043492, PubchemCID:31703)
- Sodium ion (HMO:026538, PubchemCID:271)
- Threonine (HMO:0266151, PubchemCID:2375)
- EF4A1
- Dihydrotestosterone
View Results - GO Enrichment Analysis

Gene Ontology functional Profiling

The Gene Ontology (GO) provides defined terms for representing the properties of gene products. GO covers three levels of properties: (i) cellular component, (ii) biological process, (iii) molecular function. Help users to understand information of gene products from the defined three domains.

- biological process
- cellular component
- molecular function

GO Terms with P-value < 0.05

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Coverage</th>
<th>P-value</th>
<th>FDR</th>
<th>Mapped Gene ID</th>
<th>Mapped Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>translation</td>
<td>4</td>
<td>0.0156</td>
<td>EEF1D, EEF1B2, EIF4A1, WARS</td>
<td>1936, 1933, 1973, 7453</td>
<td></td>
</tr>
<tr>
<td>cell death</td>
<td>4</td>
<td>0.1082</td>
<td>PDCD6IP, BCL2A1, TARDBP, PAK2</td>
<td>10015, 597, 23435, 5062</td>
<td></td>
</tr>
<tr>
<td>death</td>
<td>4</td>
<td>0.1104</td>
<td>PDCD6IP, BCL2A1, TARDBP, PAK2</td>
<td>10015, 597, 23435, 5062</td>
<td></td>
</tr>
<tr>
<td>apoptosis</td>
<td>3</td>
<td>0.2565</td>
<td>PDCD6IP, BCL2A1, PAK2</td>
<td>10015, 597, 5062</td>
<td></td>
</tr>
</tbody>
</table>

Biological Process

A biological process is a process of a living organism. Biological processes are made up of any number of chemical reactions or events which result in a transformation. Regulation of biological processes occurs where any process is modulated in its frequency, rate or extent. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.
3. HiPub
Go to: http://hipub.korea.ac.kr/

2. Installation and Usage

i. Install the newest version of Chrome Browser (skip this step if you have Chrome Browser)
ii. Install HiPub through Google Chrome Web Store
iii. Go to PubMed or PubMed Central and HiPub will automatically work.
iv. Usage Example.
   i. PubMed Article (Abstract): PMID24009732
   ii. PubMed Article (Abstract): PMID21262914
   iii. PubMed Article (Abstract): PMID21262914
   iv. PubMed Central Article (Full Text): PMC2733554
   v. PubMed Central Article (Full Text): PMC3641357

3. Documentation


4. Citation

Kyubum Lee, Wonho Shin, Byounggun Kim, Sunwon Lee, Yonghwa Choi, Sunkyu Kim, Minji Jeon, Aik Choon Tan*, and Jaewoo Kang*
HiPub: translating PubMed and PMC texts to networks for knowledge discovery
Installation

• Install Chrome:
  https://www.google.com/chrome/browser/desktop/

• Install the HiPub plugin:
  https://chrome.google.com/webstore/detail/hipub/jlbmiklekmigmbmcodhjgdpooldjcjam
Open the following URL in Google Chrome:
4. xMWAS: R package

URL: https://sourceforge.net/projects/xmwas/files/?source=navbar

A) Install R: https://mirrors.nics.utk.edu/cran/

B) R commands for installing dependencies:
install.packages("mixOmics",repos="http://cran.r-project.org")
install.packages("WGCNA",repos="http://cran.r-project.org",dependencies=TRUE)
install.packages("snow",repos="http://cran.r-project.org")
install.packages("igraph",repos="http://cran.r-project.org")
source("https://bioconductor.org/biocLite.R")
biocLite("graph")
biocLite("RBGL")
install.packages("plsgenomics",repos="http://cran.r-project.org")
install.packages("plyr")

C) Install xMWAS: https://sourceforge.net/projects/xmwas/files/?source=navbar

Installation on Windows:
1) Download the "xMWAS_*\.zip" file (https://sourceforge.net/projects/xmwas/files/?source=navbar)
2) Open R
3) Click on "Packages" under the menu bar
4) Click on "Install packages from local zip files"
5) Browse to the download location of the "xMWAS_*\.zip" file
6) Double click on the file and installation should begin
7) Run "library(xMWAS)" command from within R to make sure the package is successfully installed

xMWAS installation on Mac:
1) Go to "Applications"
2) Open R
3) Go to "Packages & Data"
4) Select "Local Source Package" option from the drop down menu
5) Click on "Install"
6) Browse to the download location
7) Click "Open"
library(xMWAS)

# example dataset that includes mRNA, protein, and miRNA data
data(exnci60)
data(classlabels_casecontrol) # example classlabels file for case vs control design
data(classlabels_repeatmeasures) # example classlabels file for repeat measures design

xMat<-exnci60$mrna
yMat<-exnci60$miRNA
zMat<-exnci60$prot
classlabels<-exnci60$classlabels

output<"~/Users/karanuppal/xMWASoutput/" # change for your computer

xmwas_res<-run_xmwas(Xome_data=xMat,Yome_data=yMat,Zome_data=zMat,Wome_data=NA,outloc=output,
classlabels=classlabels,class_fname=NA,xmwasmethod="spls",plsmode="canonical",max_xvar=1000,max_yvar=1000,
max_zvar=1000,max_wvar=1000,rsd.filt.thresh=1,corthresh=0.8,keepX=100,keepY=100,keepZ=100,keepW=100,
pairedanalysis=FALSE,optselect=TRUE,rawPthresh=0.05,numcomps=10,net_edge_colors=c("blue","red"),
net_node_colors=c("orange","green","cyan","gold"),Xname="X",Yname="Y",Zname="Z",Wname="W",
net_node_shape=c("rectangle","circle","triangle","star"),all.missing.thresh=0.3,maxnodesperclass=100,
seednum=100,label.cex=0.2,vertex.size=6,graphclustering=TRUE,interactive=FALSE,max_connections=2000,
centrality_method="eigenvector",use.X.reference=FALSE,removeRda=TRUE)

suppressWarnings(try(sink(file=NULL),silent=TRUE))
Open R
Copy and paste the R code on slide 18 (after changing the output location)
xMWAS – processing complete
xMWAS output
Input Parameters

```plaintext
# xMwas v0.3 Parameters

xmwasmethod: spls
plsmode: canonical
max_xvar: 1000
max_yvar: 1000
max_zvar: 1000
max_wvar: 1000
rsd_filt.thresh: 1
comthresh: 0.8
keepX: 100
keepY: 100
keepZ: 100
keepW: 100
pairedanalysis: FALSE
optselect: TRUE
rawPthresh: 0.05
numcomps: 10
maxnodesperclass: 100
seednum: 100
graphclustering: TRUE
max_connections: 2000
centrality_method: eigenvector
use.X.reference: FALSE

Loaded packages in the current session:

R version 3.2.2 (2015-08-14)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.12.5 (unknown)
```
Description of output: Readme.txt

"Description of files"

1: X labels correspond to xome fname data, Y labels correspond to yome fname data, Z labels correspond to zome fname data, W labels correspond to wome fname data

2: Pairwise integrative analysis results are under pairwise_results. The files corresponding to each pairwise comparison (X<->Y, X<->Z, Y<->Z...) are: XYassociation_matrix_corthresh0.9.txt (correlation matrix with mapping between node labels and original variable names), XYassociation_networkthresholdX.pdf that includes the pairwise network plots, XYBoolean association matrix corthreshX.txt (same as correlation matrix b correlations meeting the threshold are represented 1, and 0 otherwise)

3: Multiome Network corthreshx.pdf; includes multiome network plot using all significantly associated variables.

4: Multiome Network corthreshx communities.pdf; includes multiome network plot with the communities identified using the multilevel community detection algorithm. Members of each community are assigned colors based on community/module/cluster membership (1: orange; 2: light blue; 3: dark green, and so on).

5: Multiome Network corthreshxyctoscape.omm GML file for all significantly associated variables that can be uploaded to Cytoscape

6: The cluster_membership_centrality_mapped.txt file includes community detection results using the multilevel community detection algorithm and the centrality measures.

7: The matrix_centrality.txt file includes the centrality measures across different conditions for nodes that meet the association criteria and include in the association networks.

8: If the classlabels are provided, network analysis is performed for samples from each class. The results are written in individual subfolders."
Network graphs

- Genes
- Proteins
- microRNAs

Red: +ve correlation; Blue: –ve correlation
Positive correlation
Negative correlation
Community detection and centrality analysis
Pairwise results – X<->Y, X<->Z, Y<->Z
Questions/Comments?