Computational methods for data integration

March 21, 2018
Karan Uppal, PhD
Assistant Professor of Medicine
Director of Computational Metabolomics & Integrative Omics,
Clinical Biomarkers Laboratory
Emory University
Learning Objectives

• Understanding of different data integration approaches
• Familiarity with tools for data integration and network visualization
Introduction: A Systems Biology Framework

• The goal of **Systems Biology**:
  – Systems-level understanding of biological systems
  – Analyze not only individual components, but their interactions as well and emergent behavior

“Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery”

C. Henry and C. Washington
Dissecting the Biological system via -omics

Dissecting the Biological system via -omics

"Information Overload": >10,000 variables per -omics experiment

Why data integration?

• Systems level analysis provides:
  – more detailed overview of underlying mechanisms;
  – exploration of interactions between different biomedical entities (genes, proteins, metabolites, etc.)

• Combining multiple types of data compensates for noise or unreliable information in a single data type

• More confidence in results if multiple sources of evidence pointing to the same gene or pathway
Paired integrative –omics analysis

• Discover networks of associations or correlated variables (genes, proteins, metabolites, microbiome, epigenetic alterations, clinical variables, etc.) from paired –omics data measured across same samples
  – Univariate or multivariate regression
  – Example: explaining protein abundance with respect to gene expression
• Determine if different –omics data point to same disease mechanism
• Generate novel hypotheses for further investigation
Main approaches for data integration

• Pathway or knowledge-based integration
  – Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  – Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore

• Data-driven integration using meta-dimensional analysis
  – Integration is performed globally such that data from multiple omics layers are combined simultaneously
  – Examples: 3Oomics, mixOmics, xMWAS

• Using literature-derived associations for integration
  – Using co-occurrence criteria for establishing relationship
  – Examples: CoPub, ArrowSmith, SEACOIN2.0
Main approaches for data integration

- **Pathway or knowledge-based integration**
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore

- **Data-driven integration using meta-dimensional analysis**
  - Integration is performed globally such that data from multiple omics layers are combined simultaneously
  - Examples: 3Oomics, mixOmics, xMWAS

- **Using literature-derived associations for integration**
  - Using co-occurrence criteria for establishing relationship
  - Examples: CoPub, ArrowSmith, SEACOIN2.0
Pathway or knowledge-based integration

Metabolomics data (n subjects X p metabolites)

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>-</th>
<th>Mp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>199</td>
<td>19</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>50</td>
<td>30</td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

Differentially expressed metabolites

Pathway analysis for metabolites

Common pathways or pathway rank aggregation

Transcriptomics data (n subjects X q genes)

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>-</th>
<th>Gq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>19</td>
<td>19</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>10</td>
<td>40</td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>

Differentially expressed genes

Pathway analysis for genes
Pathway or knowledge-based integration

**Metabolomics data**
(n subjects X p metabolites)

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th></th>
<th>Mp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>199</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>50</td>
<td>30</td>
<td>-</td>
<td>20</td>
</tr>
</tbody>
</table>

**Transcriptomics data**
(n subjects X q genes)

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th></th>
<th>Gq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>19</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td>-</td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>10</td>
<td>40</td>
<td>-</td>
<td>50</td>
</tr>
</tbody>
</table>

Differentially expressed metabolites

Differentially expressed genes

Pathway analysis using genes and metabolites (joint)
MetaboAnalyst
(http://www.metaboanalyst.ca/faces/ModuleView.xhtml)
Upload data and submit
<table>
<thead>
<tr>
<th>Query</th>
<th>Hit</th>
<th>HMDB</th>
<th>HMDB</th>
<th>KEGG</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>C00116</td>
<td>Glycerol</td>
<td>HMDB0000131</td>
<td>C00116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00565</td>
<td>Trimethylamine</td>
<td>HMDB0000006</td>
<td>C00565</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00033</td>
<td>Acetic acid</td>
<td>HMDB0000042</td>
<td>C00033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00583</td>
<td>Propylene glycol</td>
<td>HMDB0001881</td>
<td>C00583</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00222</td>
<td>Pyruvic acid</td>
<td>HMDB0000243</td>
<td>C00022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00719</td>
<td>Betaine</td>
<td>HMDB0000043</td>
<td>C00719</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C05984</td>
<td>2-Hydroxybutyric acid</td>
<td>HMDB0000008</td>
<td>C05984</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00207</td>
<td>Acetone</td>
<td>HMDB00001659</td>
<td>C00207</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00065</td>
<td>L-Serine</td>
<td>HMDB0000187</td>
<td>C00065</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00031</td>
<td>D-Glucose</td>
<td>HMDB0000122</td>
<td>C00031</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00079</td>
<td>L-Phenylalanine</td>
<td>HMDB0000159</td>
<td>C00079</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Set parameters
Results - Overview

Click on “View Details” for detailed results
Results - Details

MetaboAnalyst -- a comprehensive tool for metabolomics analysis and interpretation

Pyruvate metabolism (KEGG)

The matched nodes are highlighted in different colors: red (up-regulated), yellow (unknown), green (down-regulated) based on fold change (FC) values. Click on a node to view more details.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Total</th>
<th>Expected</th>
<th>Hits</th>
<th>PValue</th>
<th>Topology</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyruvate metabolism</td>
<td>41</td>
<td>1.1233</td>
<td>4</td>
<td>0.023442</td>
<td>0.3913</td>
<td>View</td>
</tr>
<tr>
<td>Glycosphingolipid biosynthesis - laco and neolaacto series</td>
<td>26</td>
<td>0.71233</td>
<td>3</td>
<td>0.000614</td>
<td>0.0006</td>
<td>View</td>
</tr>
<tr>
<td>Glycosphingolipid biosynthesis - globo series</td>
<td>13</td>
<td>0.35616</td>
<td>2</td>
<td>0.047147</td>
<td>0.44444</td>
<td>View</td>
</tr>
<tr>
<td>Glutathione, cysteine and glutamic metabolism</td>
<td>13</td>
<td>0.35616</td>
<td>2</td>
<td>0.047147</td>
<td>0.44444</td>
<td>View</td>
</tr>
</tbody>
</table>
Main approaches for data integration

• Pathway or knowledge-based integration
  – Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  – Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore

• Data-driven integration using meta-dimensional analysis
  – Integration is performed globally such that data from multiple omics layers are combined simultaneously
  – Examples: 3Omics, mixOmics, xMWAS

• Using literature-derived associations for integration
  – Using co-occurrence criteria for establishing relationship
  – Examples: CoPub, ArrowSmith, SEACOIN2.0
Metabolomics data (n subjects X p metabolites)

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>-</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>199</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>50</td>
<td>30</td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

Transcriptomics data (n subjects X q genes)

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>-</th>
<th>Gn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>19</td>
<td>19</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>10</td>
<td>40</td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>

### Workflow

**Pathway enrichment**
- Cholesterol biosynthesis
- Methionine and cysteine metabolism

**Relevance networks**

**Clustering**

**Targeted investigation** (e.g.: Arginine x Transcriptome)

**Association matrix**

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>-</th>
<th>Gn</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.4</td>
<td>0.9</td>
<td>-</td>
<td>0.3</td>
</tr>
<tr>
<td>M2</td>
<td>0.7</td>
<td>0.1</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>M3</td>
<td>0.1</td>
<td>0.6</td>
<td></td>
<td>0.8</td>
</tr>
</tbody>
</table>

**Univariate**
- Pearson, Spearman, Partial Correlation
- Tools: 3Omics, MetabNet, etc.

**Multivariate**
- PLS, CCA, sparse PLS
- Tools: mixOmics (Cao 2009), etc.
Relevance networks

• What is a network (or graph)?
  – A set of nodes (vertices) and edges (links)
  – Edges describe a relationship (e.g. correlation) between the nodes

• What is a relevance network?
  – Networks of highly-correlated biomedical/clinical entities (Butte 2000; PNAS)
  – Metabolomics x Proteomics, Transcriptomics x Proteomics, Metabolomics x Microbiome, Metabolomics x Clinical variables/phenotypes, etc.
  – Generate a bipartite graph network using a association threshold (e.g. 0.5) to visualize positive or negative associations

Circles: microbial species
Rectangles: metabolome features
Methods for generating relevance networks

• Univariate
  – Pairwise Pearson or Spearman correlation between data from different biomedical/clinical technologies (Butte et al. 2000, Uppal et al. 2015)
  – 3Omics (Kuo 2013; a web-based tool for analysis, integration and visualization of human transcriptome, proteome and metabolome data)
  – MetabNet (Uppal 2015; R package for performing pairwise correlation analysis and generating relevance networks)

• Multivariate
  – Multivariate regression techniques such as partial least squares (PLS), sparse partial least squares regression (sPLS), multilevel sparse partial least squares (msPLS) regression, etc.
  – xMWAS (Uppal 2018): R package for data-driven integration and differential network analysis
Univariate methods
3Omics (Kuo et al. BMC Systems Biology 2013)

• A web-based tool for analyzing, integrating and visualizing transcriptomic, proteomic and metabolomic data

• http://3omics.cmdm.tw/
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

Please select the desired analysis:

a. Transcriptomics-Proteomics-Metabolomics
b. Transcriptomics-Proteomics
c. Proteomics-Metabolomics
d. Transcriptomics-Metabolomics
e. Transcriptomics only
f. Proteomics only
g. Metabolomics only

Please refer to the help page for more details about each integrating method.

User may upload three kinds of -omic expression data. All analyses will be performed.

- Transcriptomics
  - Browse...
  - No file selected.
  - GenBank ID: e.g. NAT1, ABL1

- Proteomics
  - Browse...
  - No file selected.
  - Uniprot Accession: e.g. P31946, P62258

- Metabolomics
  - Browse...
  - No file selected.
# Data format

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

## Variables

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th>sample4</th>
<th>sample5</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>
Correlation analysis

3Omics generates inter-omic correlation network to display the relationship or common patterns in data over time or experimental conditions for all transcripts, proteins and metabolites. Where users may only have two of the three -omics data-sets, 3Omics supplements the missing transcript, protein or metabolite information by searching HIP database.

Summary of Input molecules

Cluster:

<table>
<thead>
<tr>
<th>Molecules</th>
</tr>
</thead>
</table>
Co-expression analysis

Rows: Variables
Columns: Samples
Phenotype analysis

Phenotype Analysis

A phenotype is defined as any observable characteristic or trait of an organism arising from gene expression, the influence of environmental factors, and the interactions between them. With phenotype-gene association from OMIM, genes and genetic disorders containing information to relate genes in the human genome with specific phenotypes can be identified.

The Transcriptomics data you've input have been used to search through the OMIM database, and the related phenotype and genes can be listed as below:

Please click the link for description and molecular genetic information on OMIM website.

Human-related Phenotype | Related-Gene
--- | ---
OMIM: 611820 | Long QT Syndrome 11
OMIM: 299600 | Leigh Syndrome
OMIM: 123068 | Amyotrophic lateral Sclerosis 10, With or Without Frontotemporal Dementia with TDP43 Inclusions
OMIM: 308700 | Hemophilia A Coagulation Factor VIII, Included

Summary of Input molecules

Cluster:

Molecules
- IL4 (P06712)
- Leucine (OMIM:044309, PubchemCID:6196)
- ACTB3
- Decapentaplegic 5 (OMIM:043266, PubchemCID:31703)
- Calcium ion (OMIM:026539, PubchemCID:271)
- Threonine (OMIM:026151, PubchemCID:6288)
- Bicalutamide (OMIM:043660, PubchemCID:2375)
- EIF4A1
- 3-Hydroxy-3-Methylglutaryl-CoA Reductase (OMIM:025783, PubchemCID:10635)
- PDCD6IP
- Bortezomib (OMIM:048610, PubchemCID:387447)
- PSMC2
- Triglycerol (OMIM:033252, PubchemCID:5570)
- TARDBP
- RYR3
- HBB (Q15413)
- Dimethylamine (OMIM:
- PubchemCID:574)
- HSDB1
- HSDB3 (P14560)
- Hydrocortisone (OMIM:043177, PubchemCID:574)
- Tyrosine (OMIM:026152, PubchemCID:6057)
- Methotrexate (OMIM:042925, PubchemCID:126941)
- Formic acid (OMIM:044577)
- Hipuric acid (OMIM:033093, PubchemCID:464)
- Testosterone Propionate (OMIM:043891, PubchemCID:5985)
- Androsterone (OMIM:027989, PubchemCID:5879)
- MYOM1
- Leucine (OMIM:042148, PubchemCID:857)
- Zinc fluoride (OMIM:040479, PubchemCID:24551)
- 3dihydrotestosterone (OMIM:049721, PubchemCID:24812721)
- Mifepristone (OMIM:043286, PubchemCID:55245)
- MAP2K7
- TAK1 (Q43318)
- Acenic Acid (OMIM:033434, PubchemCID:444212)
- Indican (OMIM:049137, PubchemCID:10258)
- Estradiol (OMIM:029865, PubchemCID:5797)
- NTH (OMIM:049464, PubchemCID:5289054)
- Inositol (OMIM:036496)
Pathway analysis

### Metabolic Pathways

<table>
<thead>
<tr>
<th>Pathway Description</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolic pathways - Homo sapiens (human)</td>
<td>19</td>
</tr>
<tr>
<td>Aminoacyl-tRNA biosynthesis - Homo sapiens (human)</td>
<td>8</td>
</tr>
<tr>
<td>Alanine, aspartate and glutamate metabolism - Homo sapiens (human)</td>
<td>7</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine degradation - Homo sapiens (human)</td>
<td>6</td>
</tr>
<tr>
<td>Cysteine and methionine metabolism - Homo sapiens (human)</td>
<td>6</td>
</tr>
<tr>
<td>Arginine and proline metabolism - Homo sapiens (human)</td>
<td>6</td>
</tr>
<tr>
<td>Phenylalanine metabolism - Homo sapiens (human)</td>
<td>5</td>
</tr>
<tr>
<td>Histidine metabolism - Homo sapiens (human)</td>
<td>5</td>
</tr>
<tr>
<td>Glycine, serine and threonine metabolism - Homo sapiens (human)</td>
<td>5</td>
</tr>
<tr>
<td>Amino sugar and nucleotide sugar metabolism - Homo sapiens (human)</td>
<td>4</td>
</tr>
</tbody>
</table>
GO Enrichment Analysis

Gene Ontology functional Profiling
The Gene Ontology (GO) provides defined terms for representing the properties of gene product. 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>Term Description</th>
<th>P-value</th>
<th>Mapped Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>translation</td>
<td>0.0156</td>
<td>EEF1D, EEF1B2, EIF4A1, WARS</td>
</tr>
<tr>
<td>cell death</td>
<td>0.1082</td>
<td>PDCD6IP, BCL2A1, TARDBP, PAK2</td>
</tr>
<tr>
<td>death</td>
<td>0.1104</td>
<td>PDCD6IP, BCL2A1, TARDBP, PAK2</td>
</tr>
<tr>
<td>apoptosis</td>
<td>0.2565</td>
<td>PDCD6IP, BCL2A1, PAK2</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 other events that results 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.
Multivariate methods
Generating relevance network using sPLS or msPLS techniques (Cao 2009, Liquet 2012)

- sparse partial least squares (sPLS) regression or multilevel partial least squares (msPLS) method
- One-step procedure for variable selection as well as integration
- Comparison of different multivariate integration techniques showed that sPLS generates (Cao 2009)
- msPLS – for repeated measures
- Implemented in the R package mixOmics
- Generates association matrix and allows visualization of associations using bipartite relevance networks (Liquet 2012)
sPLS method

• sPLS is a variable selection and dimensionality reduction method that allows integration of heterogeneous omics data from same set of samples
• Robust approximation of Pearson correlation using regression and latent (principal) variates
• Eg: transcriptome (matrix X) and metabolome (matrix Y) data
  where,
  matrix X is an $n \times p$ matrix that includes $n$ samples and $p$ metabolites
  matrix Y is an $n \times q$ matrix that includes $n$ samples and $q$ genes

Objective function
max $\text{cov}(X_u, Y_v)$
  where
  $u_1, u_2...u_H$ and $v_1, v_2...v_H$ are the loading vectors
  $H$ is the number of PLS-DA dimensions

A Lasso based optimization is used to select most relevant variables
Case Study: Application of sPLS technique for integrative -omics. Microbiome-Metabolome Wide Association Study of Lung BAL: Global integration of 5930 m/z features with 153 microbial species using sparse Partial Least Squares regression (Cribbs et al. Microbiome 2016)
A. Association threshold: 0.3

B. Association threshold: 0.4

C. Association threshold: 0.7

D. Using only subset of metabolic features also associated with HIV status (+ve or –ve)
Integrating more than two datasets

### Proteomics data
(n subjects X s proteins)

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>-</th>
<th>Es</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>199</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>50</td>
<td>30</td>
<td>-</td>
<td>20</td>
</tr>
</tbody>
</table>

### Metabolomics data
(n subjects X p metabolites)

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>-</th>
<th>Mp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>199</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>50</td>
<td>30</td>
<td>-</td>
<td>20</td>
</tr>
</tbody>
</table>

### Transcriptomics data
(n subjects X q genes)

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>-</th>
<th>Gq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject1</td>
<td>19</td>
<td>19</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Subject2</td>
<td>10</td>
<td>40</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>SubjectN</td>
<td>10</td>
<td>40</td>
<td>-</td>
<td>50</td>
</tr>
</tbody>
</table>
A. Data matrices from multiple assays
   → Integrative and association analysis
   - Pairwise integrative and network analysis (e.g. $X\leftrightarrow Y$; $X\leftrightarrow Z$, $Y\leftrightarrow Z$) using (sparse) Partial Least Squares regression
   → Generate an edge list matrix, $L_e$, with the list of edges that meet the significance criteria and association score threshold
   → Generate a k-partite graph, $G$, using the union of edge list matrices
   → Community detection, differential centrality and rewiring analysis, and network visualization

xMWAS: R package for data integration and differential network analysis (Uppal 2018, Bioinformatics)

URL: https://kuppal.shinyapps.io/xmwas/

B. Visual diagram of a k-partite graph with nodes colored by type (Genes, Proteins, microRNAs). Edges are colored by correlation type (Positive or Negative).
xMWAS: https://kuppal.shinyapps.io/xmwas/

xMWAS - a data-driven integration and network analysis tool (v0.54)

xMWAS provides an automated workflow for data integration, network visualization, clustering, and differential network analysis of up to four datasets from biochemical and phenotypic assays, and omics platforms.

For installing xMWAS locally in R run:

```r
library(devtools);install_github("kuppal2/xMWAS")
```

Data matrices from multiple assays

Integrative and association analysis

- Pairwise integrative and network analysis (e.g. $X \leftrightarrow Y$, $X \leftrightarrow Z$, $Y \leftrightarrow Z$) using (sparse) Partial Least Squares regression

- Generate an edge list matrix, $L_e$, with the list of edges that meet the significance criteria and association score threshold

- Generate a graph, $G$, using the union of edge list matrices

- Community detection, differential centrality and rewiring analysis, and network visualization


Maintained by Chunyu Ma (chunyu.ma@emory.edu) and Karan Uppal (kuppal2@emory.edu) at Clinical Biomarkers Laboratory, Emory University, Atlanta, GA, USA
Step 1. Upload data files

https://kuppal.shinyapps.io/xmwas/
(See: Help & Support)
## Input data format

### Metabolomics data
**(p metabolites x n subjects)**

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Subject1</th>
<th>Subject2</th>
<th>Subject N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolite 1</td>
<td>199</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>Metabolite 2</td>
<td>10</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Metabolite p</td>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

### Transcriptomics data
**(q genes x n subjects)**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Subject1</th>
<th>Subject2</th>
<th>Subject N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>19</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>Gene 2</td>
<td>10</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gene q</td>
<td>10</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

### Class labels file

<table>
<thead>
<tr>
<th>Class</th>
<th>Subject1</th>
<th>Subject2</th>
<th>Subject N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Subject1</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>Control</td>
<td>Subject2</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>Tumor</td>
<td>SubjectN</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

**Metabolomics data** (p metabolites x n subjects)

**Transcriptomics data** (q genes x n subjects)

**Class labels file**
**User Manual:**

Click [here](https://kuppal.shinyapps.io/xmwas/) to see the user manual.

### Input File Format (no missing values allowed):

#### Dataset File Format:

<table>
<thead>
<tr>
<th>mma_id</th>
<th>CNS.SF_268</th>
<th>CNS.SF_295</th>
<th>CNS.SF_539</th>
<th>CNS.SNBO_19</th>
<th>CNS.SNBO_75</th>
<th>CNS.U251</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-HTOC2_1.mma</td>
<td>0.53</td>
<td>-0.42</td>
<td>0</td>
<td>0.5</td>
<td>-0.27</td>
<td>0.43</td>
</tr>
<tr>
<td>A1BG-AS1_2.mma</td>
<td>0.35</td>
<td>0.54</td>
<td>0.8</td>
<td>-0.24</td>
<td>-0.88</td>
<td>-0.1</td>
</tr>
<tr>
<td>A2LD1_3.mma</td>
<td>-0.05</td>
<td>-1.04</td>
<td>0.85</td>
<td>0.12</td>
<td>-0.36</td>
<td>-0.3</td>
</tr>
<tr>
<td>A2MPL1_4.mma</td>
<td>-1.09</td>
<td>-1.13</td>
<td>0</td>
<td>-0.43</td>
<td>-0.6</td>
<td>0.42</td>
</tr>
<tr>
<td>AGALT_5.mma</td>
<td>-0.86</td>
<td>-0.46</td>
<td>-0.57</td>
<td>0.43</td>
<td>1.38</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Each row is a feature and each column is filename.

#### Class Label File Format (multiclass):

<table>
<thead>
<tr>
<th>File Name</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS.SF_268</td>
<td>CNS</td>
</tr>
<tr>
<td>CNS.SF_295</td>
<td>CNS</td>
</tr>
<tr>
<td>CNS.SF_539</td>
<td>CNS</td>
</tr>
<tr>
<td>CNS.SNBO_19</td>
<td>CNS</td>
</tr>
<tr>
<td>CNS.SNBO_75</td>
<td>CNS</td>
</tr>
<tr>
<td>CNS.U251</td>
<td>CNS</td>
</tr>
<tr>
<td>LE.CCRF_DEM</td>
<td>LE</td>
</tr>
<tr>
<td>LE.HL.00</td>
<td>LE</td>
</tr>
<tr>
<td>LE.K_502</td>
<td>LE</td>
</tr>
<tr>
<td>LE.MOLT_4</td>
<td>LE</td>
</tr>
<tr>
<td>LE.RPMI_8226</td>
<td>LE</td>
</tr>
<tr>
<td>LE.SR</td>
<td>LE</td>
</tr>
</tbody>
</table>

Two columns: the first column is filename; the second column is class. Each row is the information of a file.

#### Class Label File Format (repeated measure with one factor):

<table>
<thead>
<tr>
<th>File Name</th>
<th>Subject</th>
<th>Factor1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Subject001</td>
<td>TP1</td>
</tr>
<tr>
<td>S2</td>
<td>Subject002</td>
<td>TP1</td>
</tr>
<tr>
<td>S3</td>
<td>Subject003</td>
<td>TP1</td>
</tr>
<tr>
<td>S4</td>
<td>Subject004</td>
<td>TP1</td>
</tr>
<tr>
<td>S5</td>
<td>Subject005</td>
<td>TP1</td>
</tr>
<tr>
<td>S6</td>
<td>Subject006</td>
<td>TP1</td>
</tr>
</tbody>
</table>
Step 2. Data preprocessing and filtering

### xMWAS - a data-driven integration and network analysis tool (v0.54)

**Introduction**  
**Analysis**  
**Help and Support**

<table>
<thead>
<tr>
<th><strong>Input Files</strong></th>
<th><strong>Parameter Settings</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choose Files (see help and support)</strong></td>
<td><strong>Relative Standard Deviation (RSD) Threshold (rows):</strong></td>
</tr>
<tr>
<td><strong>Data preparation and filtering</strong></td>
<td><strong>Maximum number of datasetA variables to select based on RSD:</strong></td>
</tr>
<tr>
<td><strong>Integration and association analysis</strong></td>
<td><strong>Maximum number of datasetB variables to select based on RSD:</strong></td>
</tr>
<tr>
<td><strong>Centrality analysis</strong></td>
<td><strong>Minimum ratio of number of samples with a non-missing value to the total number of samples for a variable (rows):</strong></td>
</tr>
<tr>
<td><strong>Graphical options</strong></td>
<td><strong>How are the missing values represented in the data?:</strong></td>
</tr>
</tbody>
</table>

Step 3. Set parameters for integration and association analysis

Step 4. Select method for centrality analysis

- **Eigenvector**: based on the number and quality of connections
- **Betweenness**: based on the extent to which a node lies on the path between other nodes
- **Degree.count**: based on the number of connections
- **Degree.weight**: based on the magnitude of edges (association scores)
- **Closeness**: based on the closeness of a node to all other nodes
Step 5. Set graphical options

[Image of xMWAS interface with options for size of labels, nodes, and graph generation settings.

Step 6. Click on “Start processing”
Results
Case Study: Application of xMWAS for integrative –and differential network analysis of more than 2 dataests. Integrative network analysis of cytokine, metabolome and transcriptome datasets from a study of H1N1 virus infection of mice (Chandler et al. 2016)
Main approaches for data integration

• Pathway or knowledge-based integration
  – Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  – Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore

• Data-driven integration using meta-dimensional analysis
  – Integration is performed globally such that data from multiple omics layers are combined simultaneously
  – Examples: 3Oomics, mixOmics, xMWAS

• Using literature-derived associations for integration
  – Using co-occurrence criteria for establishing relationship
  – Examples: HiPub, CoPub, ArrowSmith
Text mining tools for literature-based relation discovery biomedical text


Association mining based on co-occurrence
HiPub (Lee 2016): http://hipub.korea.ac.kr/
HiPub (Lee 2016):
Summary

• Various tools and techniques are available for integrating and visualization multi –omics data
• Integrative –omics drives systems biology and could play a critical role in personalized medicine
Hands-on exercises
1. HiPub
Go to: http://hipub.korea.ac.kr/

2. Installation and Usage
i. HiPub is available in Chrome Browser
ii. HiPub is designed for Pubmed and Pubmed Central.
iii. HiPub also works in,
   ○ All journals available in the AACR (www.aacrjournals.org) sites
   ○ All journals available in the ASCO (www.ascopepubs.org) sites
   ○ All journals available in BioMedCentral (BMC) (www.biomedcentral.com) sites
   ○ All journals available in the Oxford University Press (www.oxfordjournals.org) sites

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
Steps for installing HiPub

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

• Install the HiPub plugin: https://chrome.google.com/webstore/detail/hipub/jlbmiklekmigmbmcodhjgdpooldjcjam

• Test installation: https://www.ncbi.nlm.nih.gov/pubmed/24009732
  – You should see an annotated title and abstract when you go to the PubMed page above as shown on the next slide
The conformational control inhibitor of tyrosine kinases DCC-2036 is effective for imatinib-resistant cells expressing T674I FIP1L1-PDGFRα.

Abstract

The cells expressing the T674I point mutant of FIP1-like-1-platelet-derived growth factor receptor alpha (FIP1L1-PDGFRα) in hypereosinophilic syndrome (HES) are resistant to imatinib and some second-generation tyrosine kinase inhibitors (TKIs). There is a desperate need to develop therapy to combat this acquired drug resistance. DCC-2036 has been synthesized as a third-generation TKI to combat especially the Bcr-Abl T315I mutant in chronic myeloid leukemia. This study evaluated the effect of DCC-2036 on FIP1L1-PDGFRα-positive cells, including the wild type (WT) and the T674I mutant. The in vitro effects of DCC-2036 on the PDGFRα signal pathways, proliferation, cell cycling, and apoptosis of FIP1L1-PDGFRα-positive cells were investigated, and a nude mouse xenograft model was employed to assess the in vivo antitumor activity. We found that DCC-2036 decreased the phosphorylated levels of PDGFRα and its downstream targets without apparent effects on total protein levels. DCC-2036 inhibited proliferation, and induced apoptosis with MEK-dependent up-regulation of the pro-apoptotic protein Bim in FIP1L1-PDGFRα-positive cells. DCC-2036 also exhibited in vivo antineoplastic activity against cells with T674I FIP1L1-PDGFRα. In summary, FIP1L1-PDGFRα-positive cells are sensitive to DCC-2036 regardless of their sensitivity to imatinib. DCC-2036 may be a potential compound to treat imatinib-resistant HES.

PMID: 24009732  PMCID: PMC3756952  DOI: 10.1371/journal.pone.0073059

Free PMC Article
2. xMWAS: Web version

- URL: https://kuppal.shinyapps.io/xmwas/
- Input files:
  
  https://github.com/kuppal2/xMWAS/upload/master/example_manual_tutorial
  
  – exh1n1_transcriptome.txt
  – exh1n1_metabolome.txt
  – exh1n1_cytokine.txt
  – exh1n1_classlabels.txt
xMWAS - a data-driven integration and network analysis tool (v0.54)

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

Select input file for dataset A ('.csv' or '.txt', 100MB limit)
Browse... exh1n1_metabolome.txt
Name for dataset A:
metabolome

Select input file for dataset B ('.csv' or '.txt', 100MB limit)
Browse... exh1n1_transcriptome.txt
Name for dataset B:
transcriptome

Select input file for dataset C ('.csv' or '.txt', 100MB limit)
Browse... exh1n1_cytokine.txt
Name for dataset C:
cytokine

Add more datasets: + –

Choose a class labels file ('.csv' or '.txt'):
Browse... exh1n1_classlabels.txt

Are there repeated measurements?
- True - Paired (repeated measures)
- False - Unpaired (case-control & multiclass)

Use example data?
- True
- False

Output folder name:
Default: xwarsresults

Compare classes?
- True
- False

Start processing Download results

Output

Slide to go to next figure:
xMWAS - a data-driven integration and network analysis tool (v0.54)

**Input Files**
Choose Files (see help and support)

**Parameter Settings**

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

---

**Relative Standard Deviation (RSD) Threshold (rows):**

- 1

**Maximum number of datasetA variables to select based on RSD:**

- 1000

**Maximum number of datasetB variables to select based on RSD:**

- 1000

**Maximum number of datasetC variables to select based on RSD:**

- 1000

**Minimum ratio of number of samples with a non-missing value to the total number of samples for a variable (rows):**

- 0.7

**How are the missing values represented in the data?:**

- 0
xMWAS - a data-driven integration and network analysis tool (v0.54)

Input Files
Choose Files (see help and support)

Parameter Settings
1. Data preparation and filtering
   2. Integration and association analysis
   3. Centrality analysis
   4. Graphical options

Pairwise integrative analysis
Choose a data integration method:
- sPLS

Choose PLS mode:
- canonical

Number of components to use in PLS model:
- 5

Find optimal number of PLS components?
- True
- False

Maximum number of datasetA variables to select in sPLS:
- 100

Maximum number of datasetB variables to select in sPLS:
- 100

Maximum number of datasetC variables to select in sPLS:
- 100

Association analysis
Correlation Threshold:
- 0.7

P-value Threshold For Student’s T-test:
- 0.05

Start processing ▲ Download results

Slide to go to next figure:

Maintained by Chunyu Ma (chunyu.ma@emory.edu) and Karan Uppal (kuppal@emory.edu) at Clinical Biomarkers Laboratory, Emory University, Atlanta, GA, USA
xMWAS - a data-driven integration and network analysis tool (v0.54)

Input Files
Choose Files (see help and support)

Parameter Settings
1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

Size of the Labels:
0.25

Size of the Nodes:
7

Seed for Random Number Generator:
100

Maximum number of associations to include in the network (any numeric value >0 or -1 to use all):

-1

Use dataset A as reference?
True False

Start processing  Download results

Output
Slide to go to next figure:

0 1 2 3 4 5

Starting processing now. Your results will be available for download shortly. The processing time depends on the number of variables. Please use the data filtering options to reduce the run time.
Results
Download results

<table>
<thead>
<tr>
<th>Name</th>
<th>Date Modified</th>
<th>Size</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>class-wise_centrality_matrix.txt</td>
<td>Today at 7:11 AM</td>
<td>12 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>cluster_membership_centrality_table.txt</td>
<td>Today at 7:11 AM</td>
<td>18 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>InputParameters.txt</td>
<td>Today at 7:10 AM</td>
<td>4 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>LogWed_Mar_21_07_10_05_2018.txt</td>
<td>Today at 7:10 AM</td>
<td>789 bytes</td>
<td>Plain Text</td>
</tr>
<tr>
<td>Multidata_Network_threshold0.7_communities.png</td>
<td>Today at 7:11 AM</td>
<td>1.6 MB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Multidata_Network_threshold0.7_linkmatrix.txt</td>
<td>Today at 7:10 AM</td>
<td>160 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>Multidata_Network_threshold0.7.png</td>
<td>Today at 7:10 AM</td>
<td>1.5 MB</td>
<td>PNG image</td>
</tr>
<tr>
<td>Multidata_Network_threshold0.7_cytoscapeall.gml</td>
<td>Today at 7:10 AM</td>
<td>524 KB</td>
<td>Document</td>
</tr>
<tr>
<td>Network_stats.csv</td>
<td>Today at 7:10 AM</td>
<td>126 bytes</td>
<td>comma...values</td>
</tr>
<tr>
<td>NodeID_Name_mapping.txt</td>
<td>Today at 7:10 AM</td>
<td>22 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>pairwise_results</td>
<td>Today at 7:11 AM</td>
<td>--</td>
<td>Folder</td>
</tr>
<tr>
<td>datasetA_x_datasetB_association_matrix_threshold0.7.txt</td>
<td>Today at 7:10 AM</td>
<td>389 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>datasetA_x_datasetB_association_matrix_threshold0.7.png</td>
<td>Today at 7:10 AM</td>
<td>2.4 MB</td>
<td>PNG image</td>
</tr>
<tr>
<td>datasetA_x_datasetB_association_matrix_threshold0.7.png</td>
<td>Today at 7:10 AM</td>
<td>363 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>README.txt</td>
<td>Today at 7:11 AM</td>
<td>2 KB</td>
<td>Plain Text</td>
</tr>
<tr>
<td>xmwasresults20180321111005.zip</td>
<td>Today at 7:11 AM</td>
<td>5.5 MB</td>
<td>ZIP archive</td>
</tr>
</tbody>
</table>
3. xMWAS R package installation instructions for Windows

- Install R: [https://cran.cnr.berkeley.edu/](https://cran.cnr.berkeley.edu/)
- Install R dependencies
  - R command for installation:
    ```r
    source("https://bioconductor.org/biocLite.R");
    biocLite(c("GO.db","graph","RBGL","impute","preprocessCore"),dependencies=TRUE);
    install.packages(c("devtools","WGCNA","mixOmics","snow","igraph","plyr","plsgenomics")
    ,dependencies=TRUE,type="binary", repos="http://cran.r-project.org")
    ```
- Install R package xMWAS
  - R command for installation:
    ```r
    library(devtools); install_github("kuppal2/xMWAS")
    ```
- Test installation:
  - R command for loading the package:
    ```r
    library(xMWAS)
    ```
xMWAS R package installation instructions for Mac OS X

- Install Xquartz: https://www.xquartz.org/
- Install R: https://cran.cnr.berkeley.edu/
- Install R dependencies
  - R command for installation:
    ```
    source("https://bioconductor.org/biocLite.R");
    biocLite(c("GO.db","graph","RBGL","impute","preprocessCore"),dependencies=TRUE);
    install.packages(c("devtools","WGCNA","mixOmics","snow","igraph","plyr","plsgenomics")
    ,dependencies=TRUE,type="source",repos="http://cran.r-project.org")
    ```
- Install R package xMWAS
  - R command for installation:
    ```
    library(devtools); install_github("kuppal2/xMWAS")
    ```
- Test installation:
  - R command for loading the package: library(xMWAS)
R script for xMWAS using the example dataset
(URL: https://github.com/kuppal2/xMWAS/blob/master/example_manual_tutorial/example_xmwas_runscript_v0.5.R)

#load package
library(xMWAS)

#example dataset that includes metabolome, transcriptome, and cytokine data from the H1N1 mice study (Chandler 2016)
data(exh1n1)
data(classlabels_casecontrol) #example classlabels file for case vs control design
data(classlabels_repeatmeasures) #example classlabels file for repeat measures design
xMat<-exh1n1$metabolome
yMat<-exh1n1$transcriptome
zMat<-exh1n1$cytokine
classlabels<-exh1n1$classlabels

output<-"/home/kuppal2/xMWASv0.54output/"

#call the run_xmwas() function:
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=0.7,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","pink"),Xname="X",Yname="Y",Zname="Z",Wname="W",
net_node_shape=c("rectangle","circle","triangle","star"),all.missing.thresh=0.7,missing.val=0,
seednum=100,label.cex=0.2,vertex.size=6,graphclustering=TRUE,interactive=FALSE,latentthresh=FALSE,
centrality_method="eigenvector",use.X.reference=FALSE,removeRda=TRUE,compare.classes=TRUE,allvar=TRUE,
suppressWarnings(try(sink(file=NULL),silent=TRUE)))
xMWAS output
Input Parameters

```
### xMWS v0.54 Parameters#############
[1] "xmwsmethod: spls"
[1] "plsmode: canonical"
[1] "max_xvar: 1000"
[1] "max_yvar: 1000"
[1] "max_zvar: 1000"
[1] "max_wvar: 5000"
[1] "rsd_fill.thresh: 1"
[1] "all.missing.thresh: 0.7"
[1] "missing.val: 0"
[1] "corthresh: 0.7"
[1] "keepx: 100"
[1] "keepy: 100"
[1] "keepz: 100"
[1] "keepw: 100"
[1] "pairedanalysis: FALSE"
[1] "optselect: TRUE"
[1] "rawPthresh: 0.05"
[1] "numcomps: 10"
[1] "seednum: 100"
[1] "graphclustering: TRUE"
[1] "max_connections: 1e+05"
[1] "centrality_method: eigenvector"
[1] "use.X.reference: FALSE"
[1] "compare.classes: TRUE"
[1] "class.comparison.allvar: TRUE"
[1] "########Loaded packages in the current session##########"
R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.0
LAPACK: /usr/lib/lapack/liblapack.so.3.0

locale:
[1] LC_CTYPE=en_US.UTF-8  LC_NUMERIC=C
[2] LC_TIME=en_US.UTF-8   LC_COLLATE=en_US.UTF-8
[3] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[4] LC_PAPER=en_US.UTF-8   LC_NAME=C
[5] LC_ADDRESS=C          LC_TELEPHONE=C
[6] LC_TEENAGETELEPHONE=C
```

Description of output: **Readme.txt**

```
"Description of files"

"1: X labels correspond to some fname data, Y labels correspond to some fname data, Z labels correspond to some fname data, W labels correspond to some 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 but correlations meeting the threshold are represented as 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 corthreshxcytoscape.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 included 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

A. Colored by data type
- Metabolites
- Genes
- Cytokines

Positive correlation
Negative correlation

B. Colored by community membership

(Edges) Red: +ve correlation; Blue: -ve correlation
(Nodes) Rectangle: X; Circle: Y; Triangle: Z
Using all samples

Each community (C) is represented by a different color:
C1; C2; C3; C4; C5; C6; C7;
Community detection and centrality analysis
Pairwise results – X<->Y, X<->Z, Y<->Z
Clinical Biomarkers Laboratory

Dean Jones, Young-Mi Go, Shuzaho Li, Karan Uppal, Douglas Walker, Josh Chandler, Sophia Banton, Ken Liu, Vilinh Tran, Michael Orr, Bill Liang (not shown)

Lab website: http://clinicalmetabolomics.org/

Funding: ES025632, ES023485, ES019776, OD018006, HL095479, EY022618, HHSN272201200031C

Email: kuppal2@emory.edu
Questions?