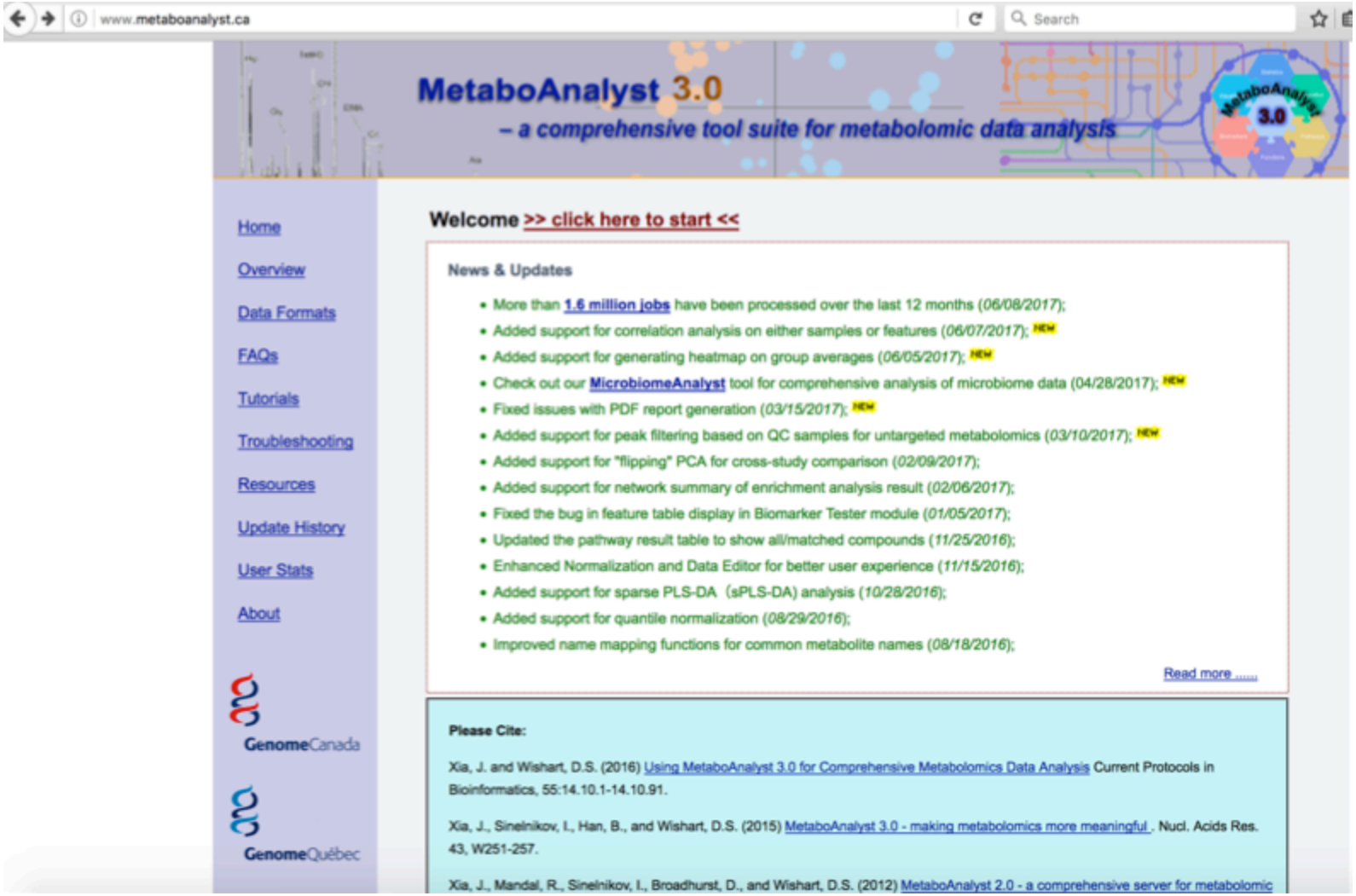
Data Integration: hands-on session

July 21, 2017

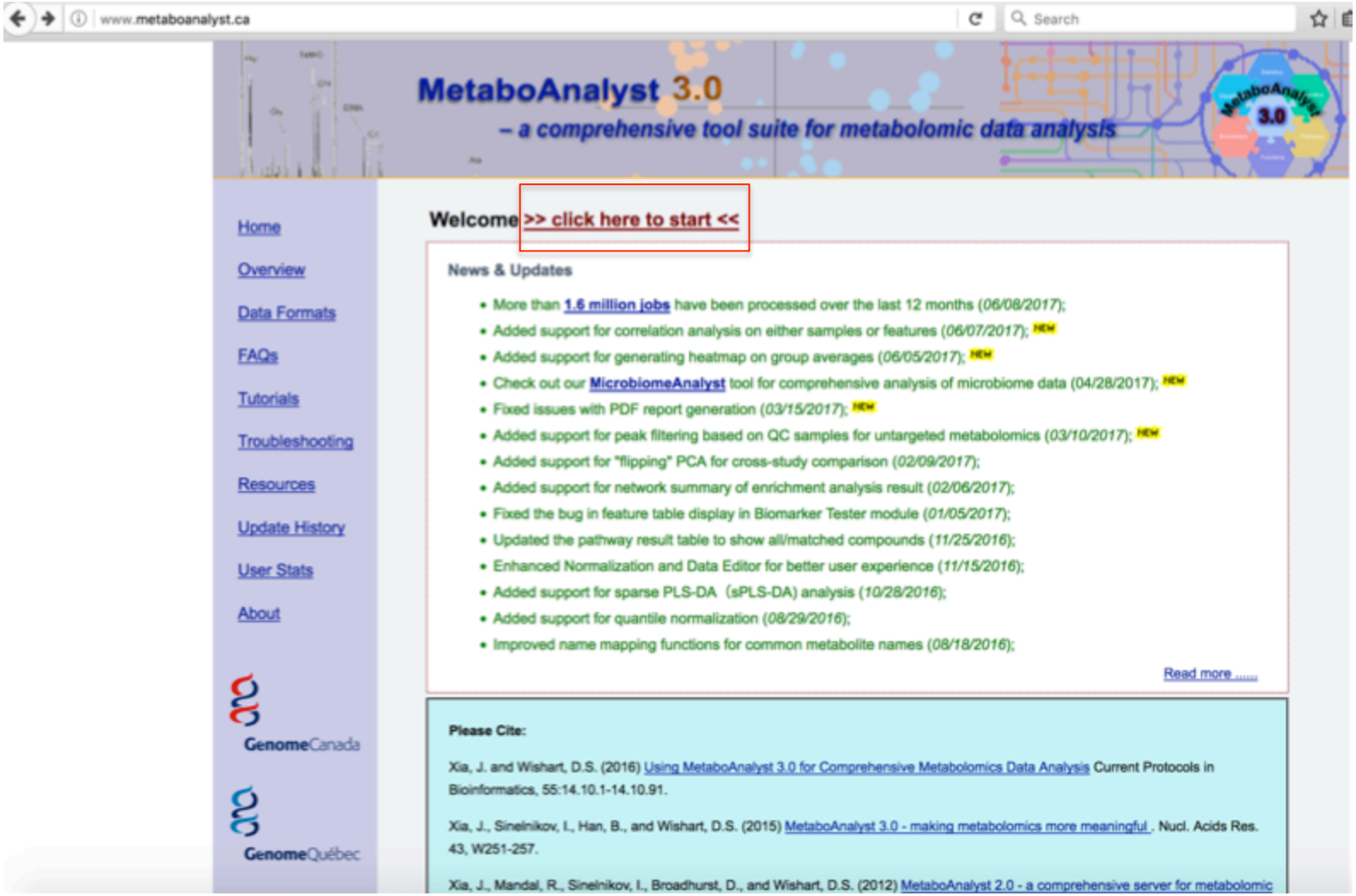
Karan Uppal, PhD
Assistant Professor of Medicine
Director of Computational Metabolomics & Integrative Omics,
Clinical Biomarkers Laboratory
Emory University

1. MetaboAnalyst: Integrated Pathway Analysis

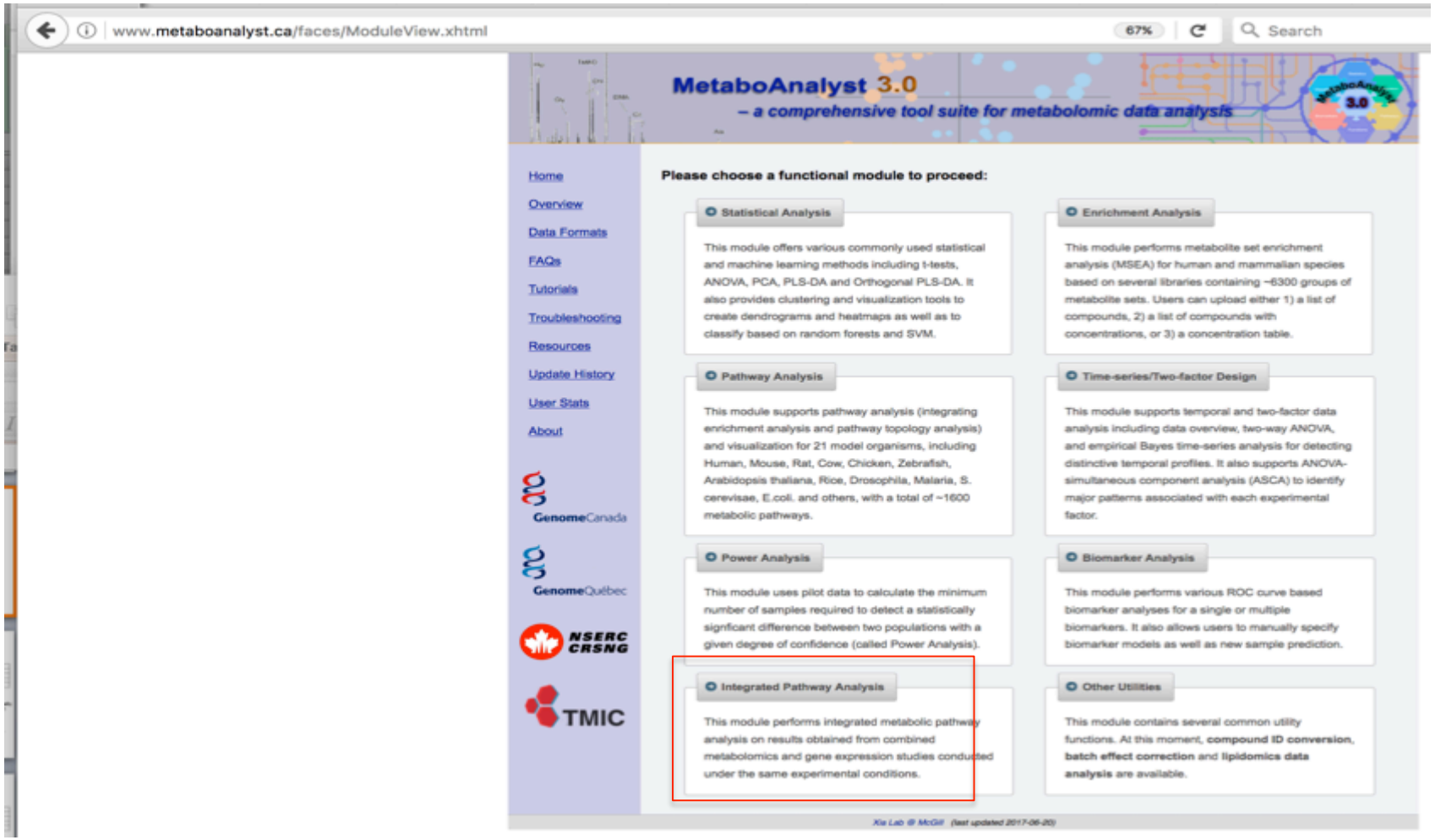
URL: http://www.metaboanalyst.ca/



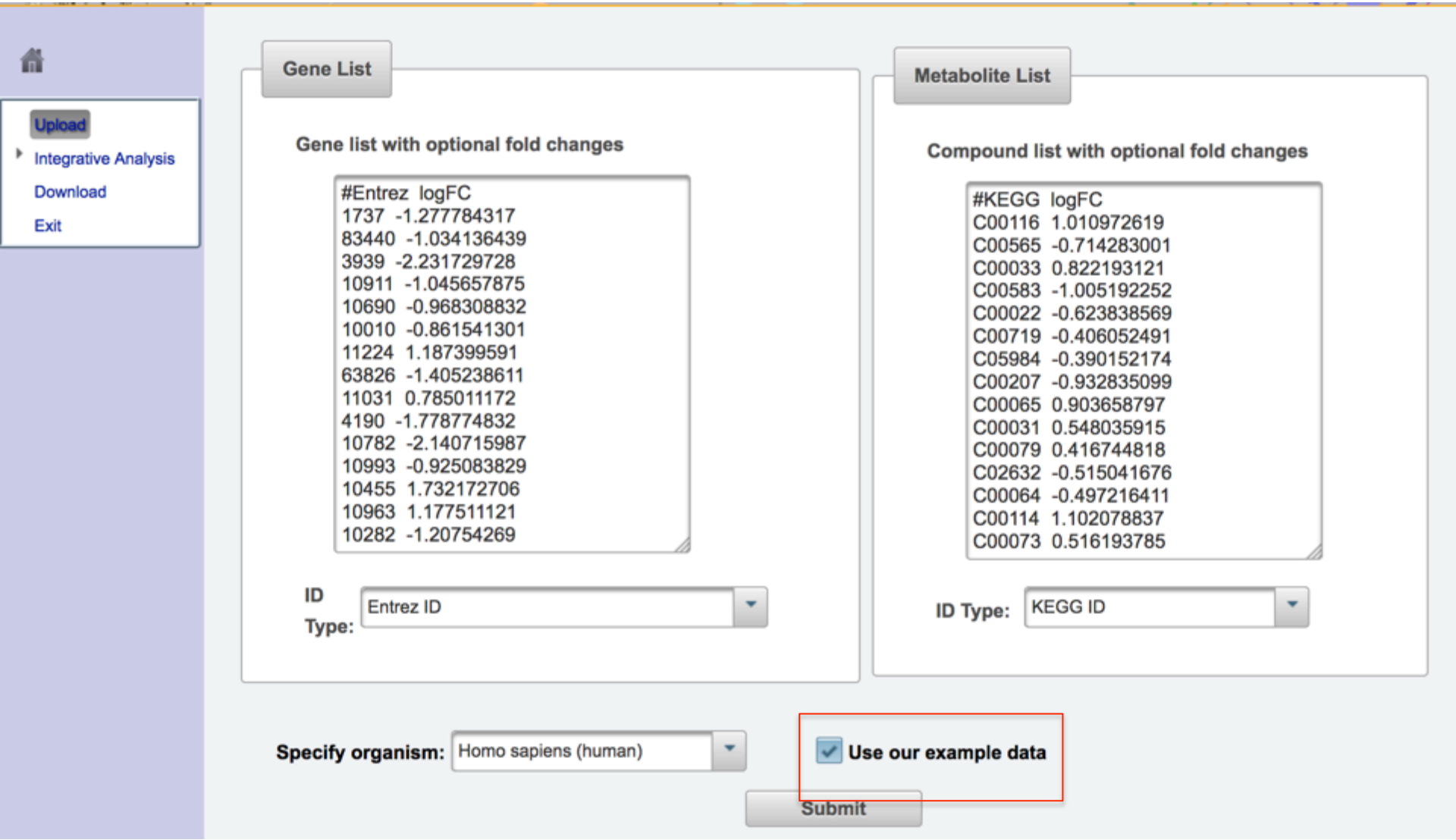
Click on “Click here to start”



Select the “Integrated Pathway Analysis” module



Data upload – select the “Use our example data” option



Enrichment Analysis

Enrichment analysis aims to evaluate whether the observed genes and metabolites in a particular pathway are significantly enriched (appeatr more than expected by random chance) within the dataset. You can choose over-representation analysis (ORA) based on either hypergenometrics 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

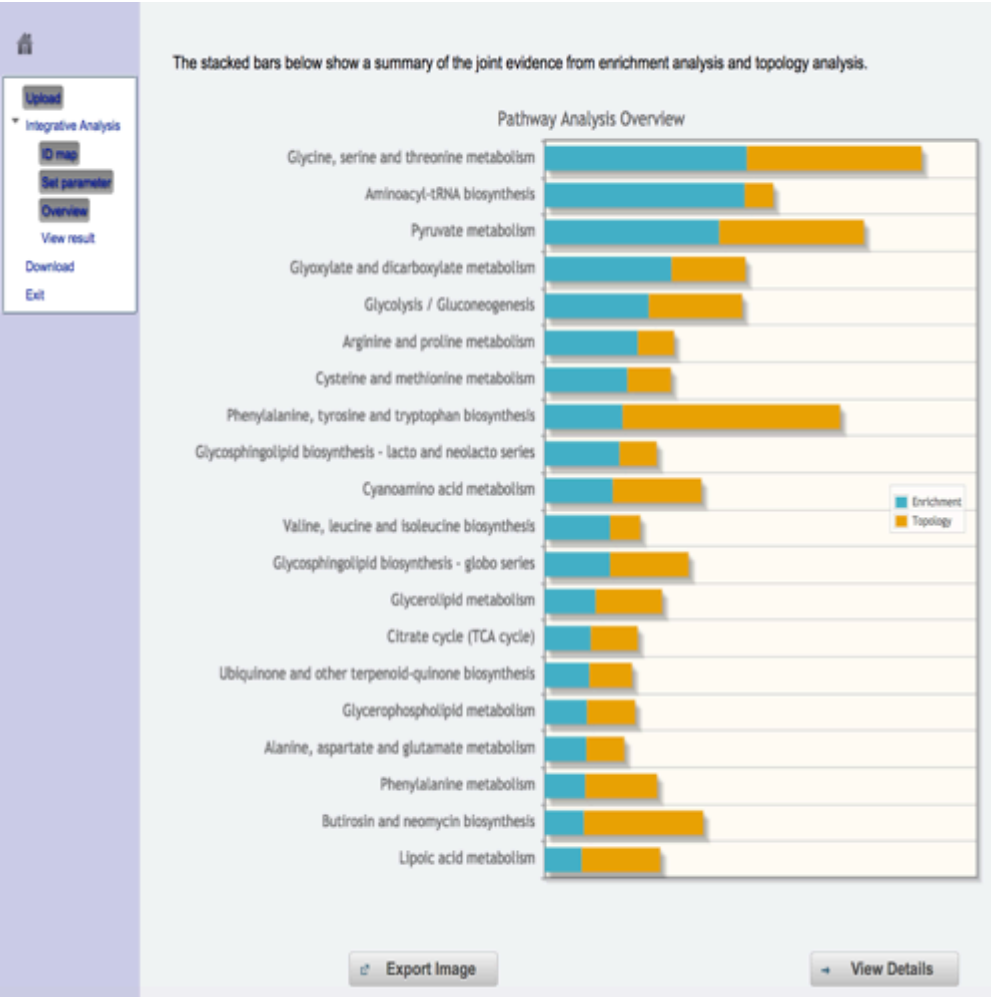
Users can choose one of three different modes of pathways: - the gene-metabolite mode (default) allows joint-analysis and visualization of both significant genes and metabolites; while the gene-centric or metabolite-centric mode allows users to identify enriched pathways driven by significant genes or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

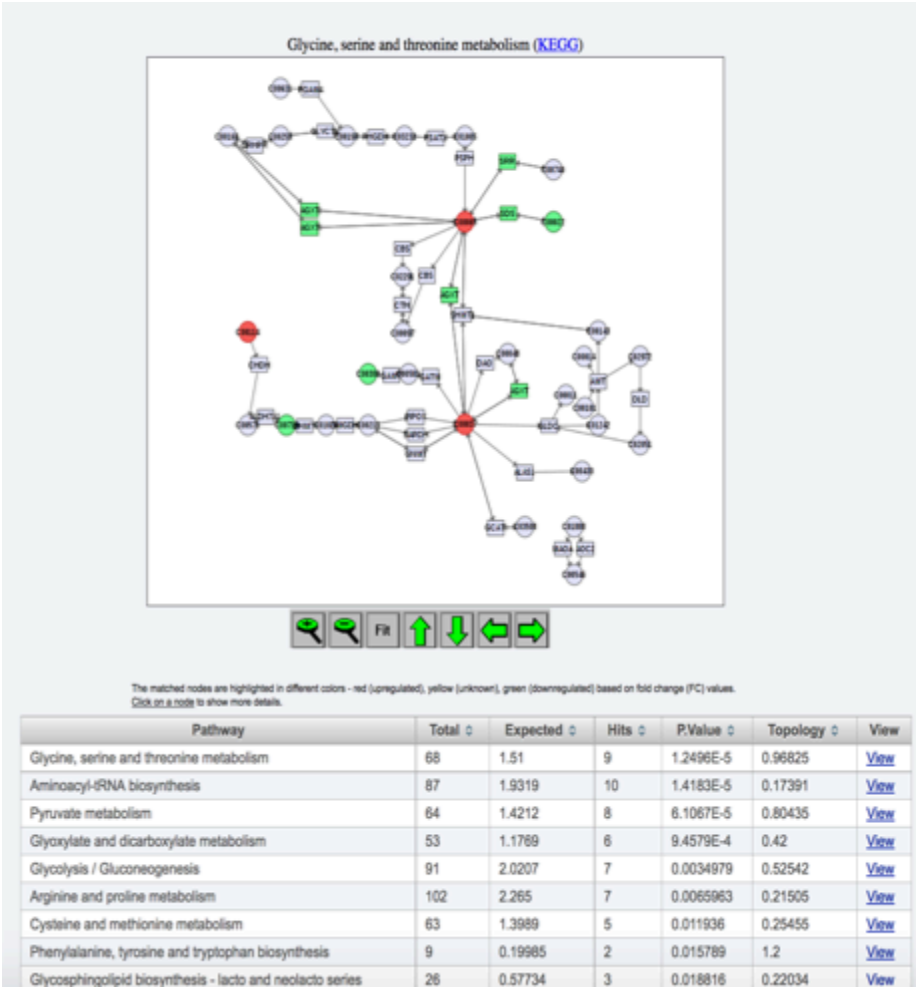
- (•) Gene-metabolite pathways
- ( ) Gene-centric pathways
- ( ) Metabolite-centric pathways

→ Submit

Set Parameters

Results





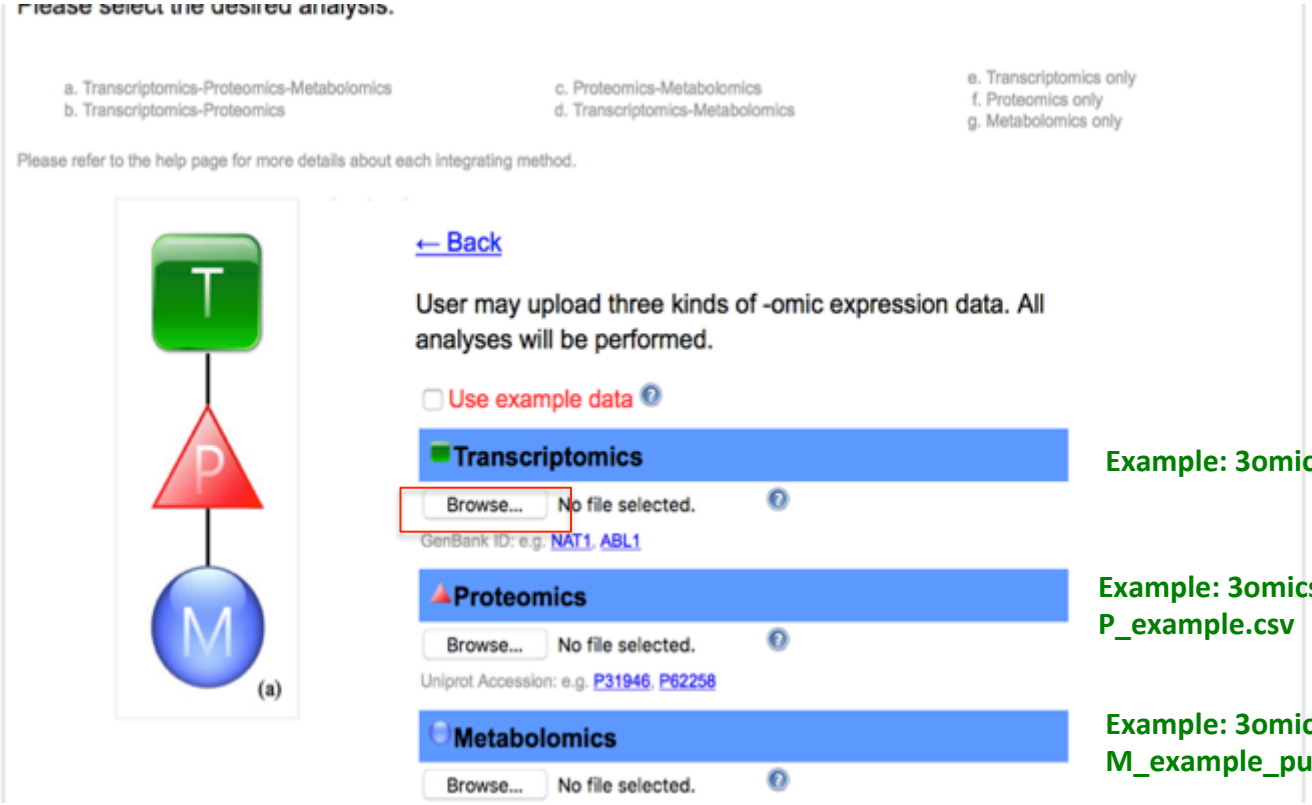
Pathway	Total	Expected	Hits	P.Value	Topology	View
Glycine, serine and threonine metabolism	68	1.51	9	1.2496E-5	0.96825	View
Aminoacyl-tRNA biosynthesis	87	1.9319	10	1.4183E-5	0.17391	View
Pyruvate metabolism	64	1.4212	8	6.1067E-5	0.80435	View
Glyoxylate and dicarboxylate metabolism	53	1.1769	6	9.4579E-4	0.42	View
Glycolysis / Gluconeogenesis	91	2.0207	7	0.0034979	0.52542	View
Arginine and proline metabolism	102	2.265	7	0.0065963	0.21505	View
Cysteine and methionine metabolism	63	1.3989	5	0.011936	0.25455	View
Phenylalanine, tyrosine and tryptophan biosynthesis	9	0.19985	2	0.015789	1.2	View
Glycosphingolipid biosynthesis - lacto and neolacto series	26	0.57734	3	0.018816	0.22034	View

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

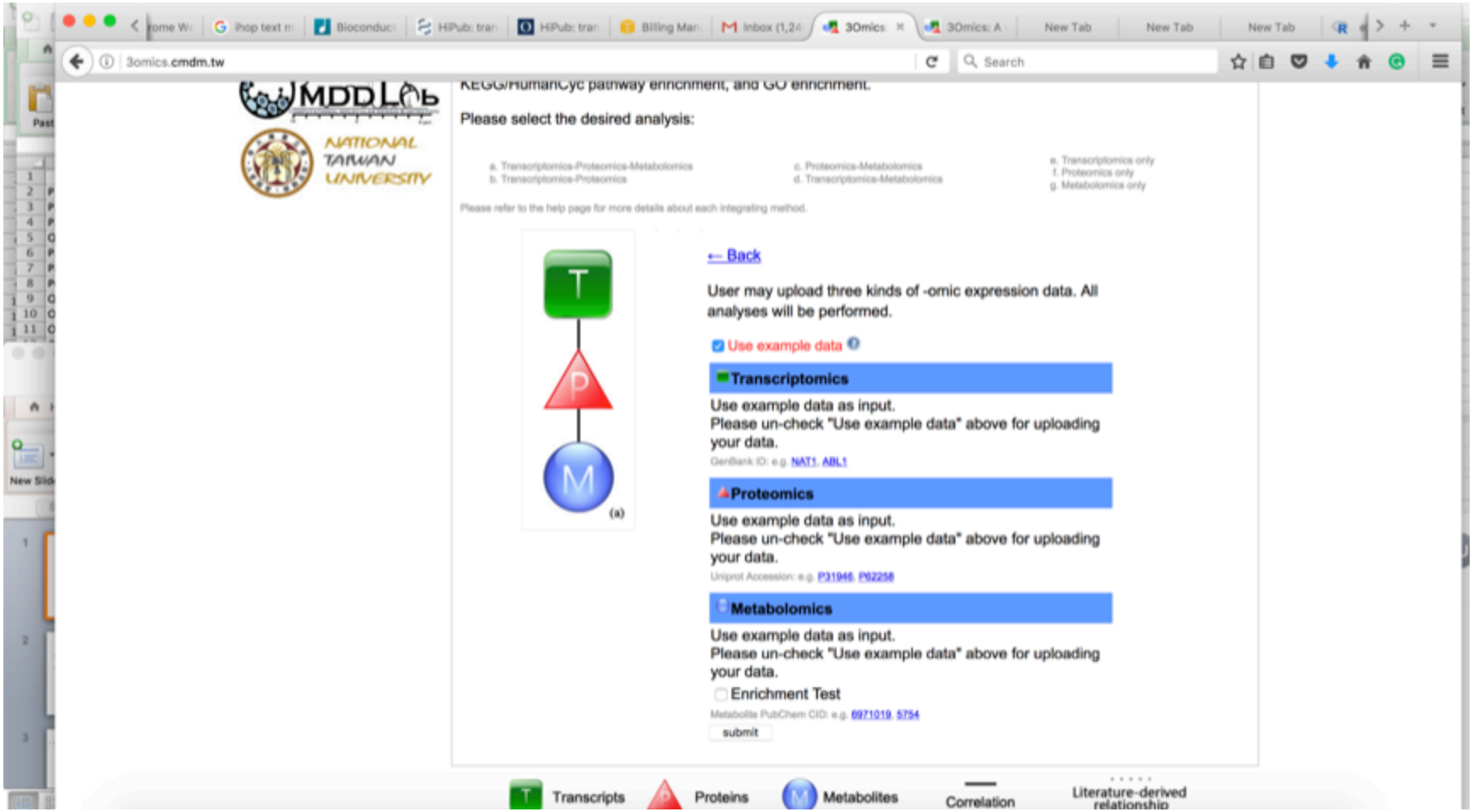
(http://3omics.cmdm.tw/help.php#examples)

Samples

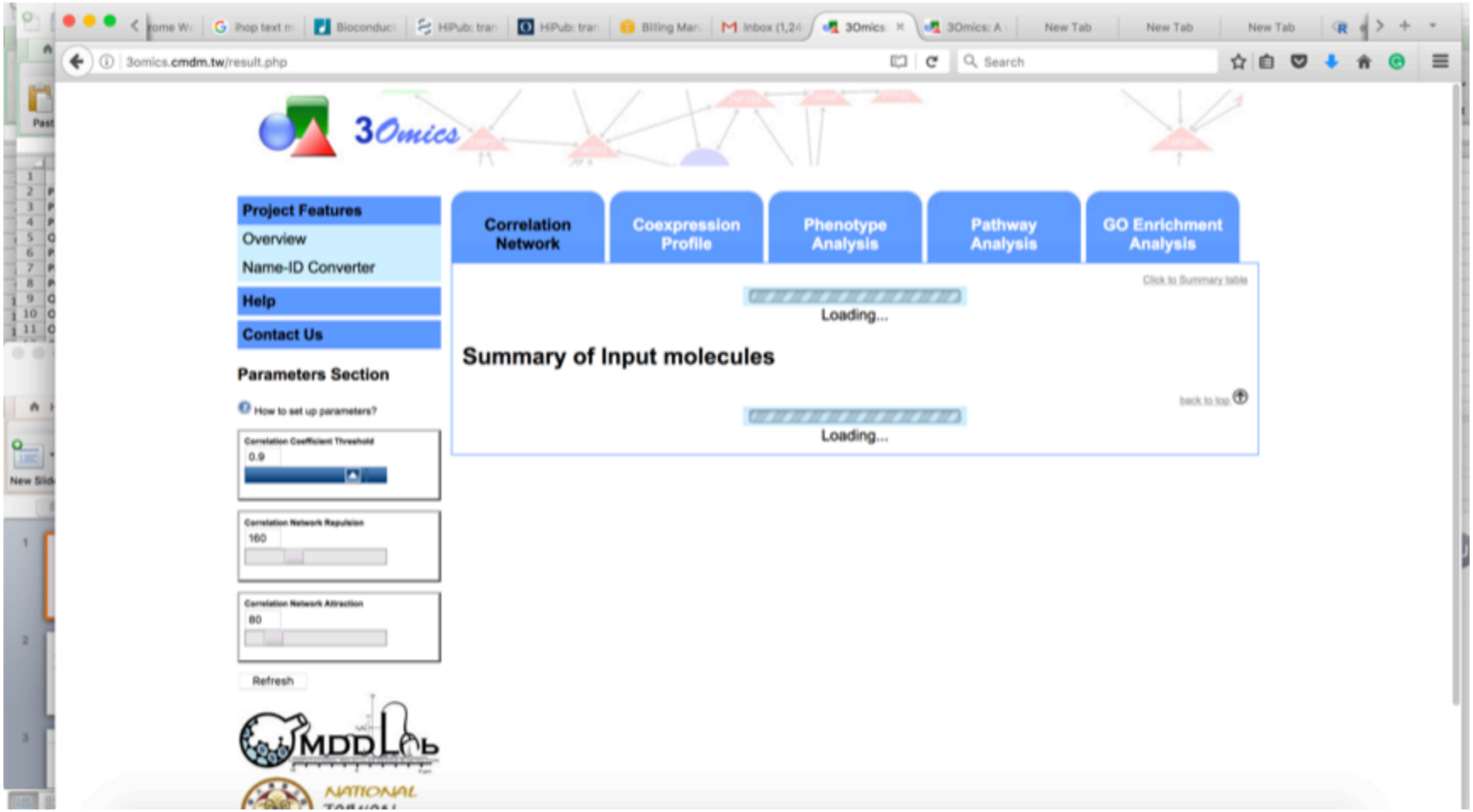
Variables

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

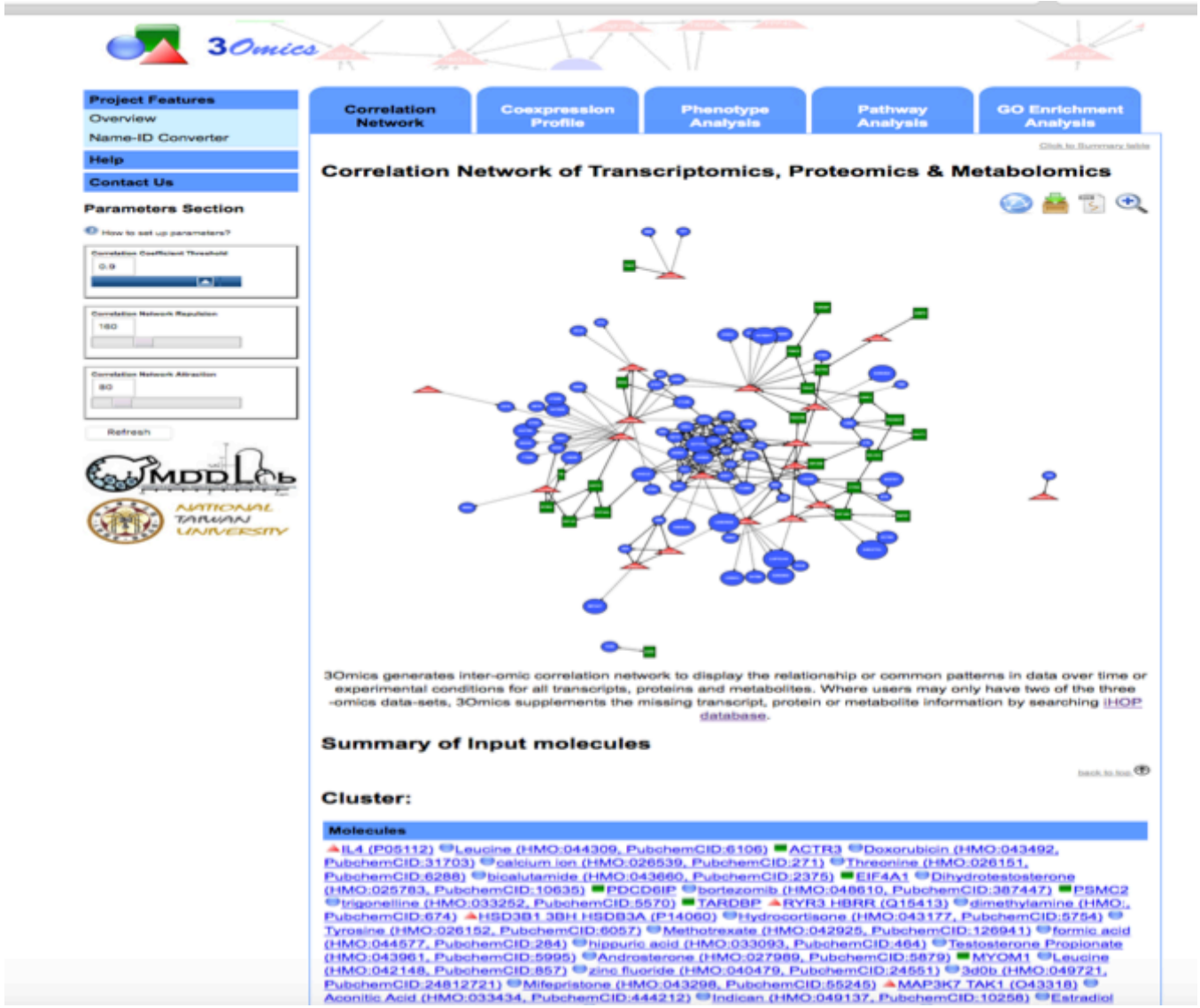
Select checkbox next to “Use example data”



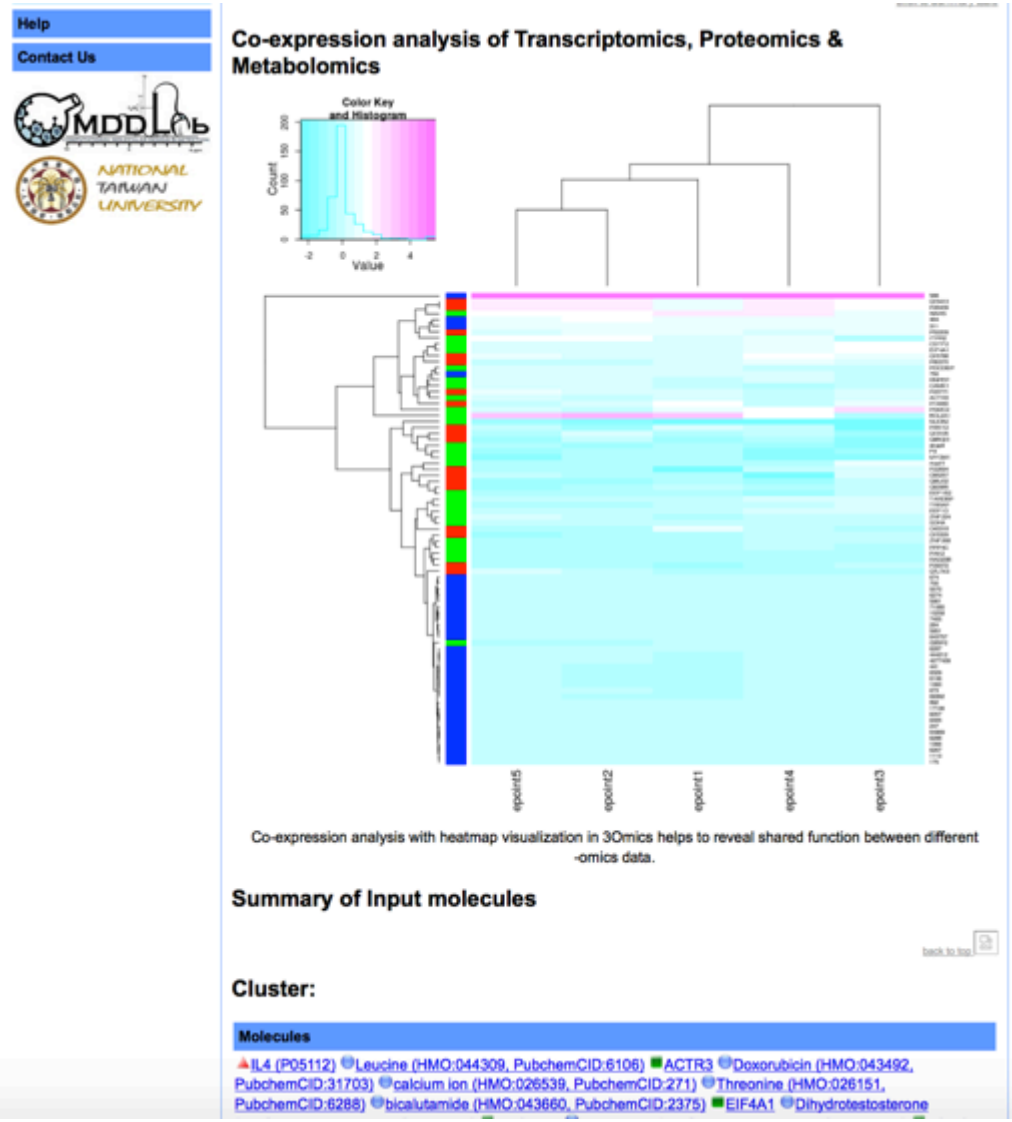
Wait for upload to complete



View Results – Correlation Network



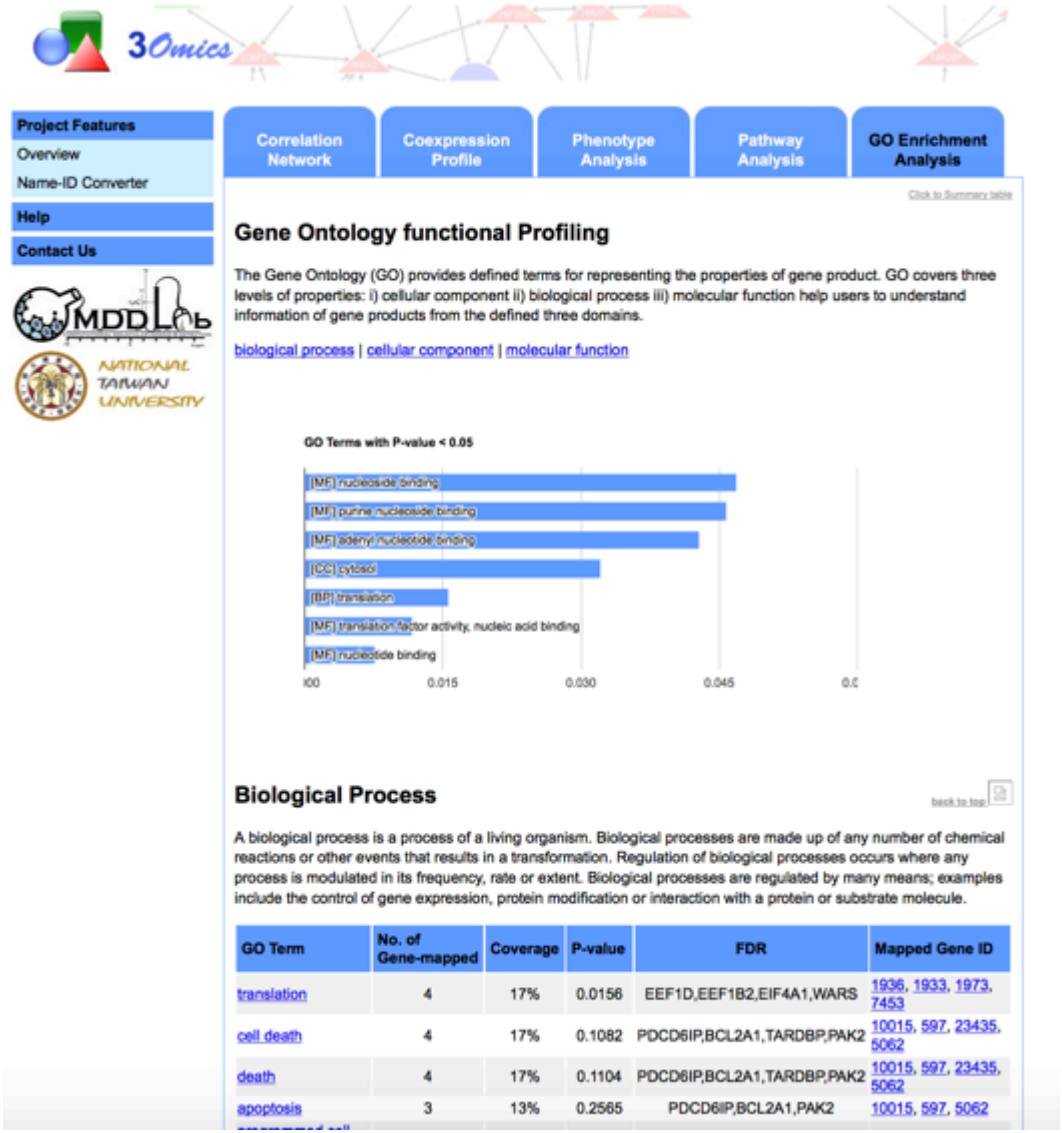
View Results – Co-expression analysis



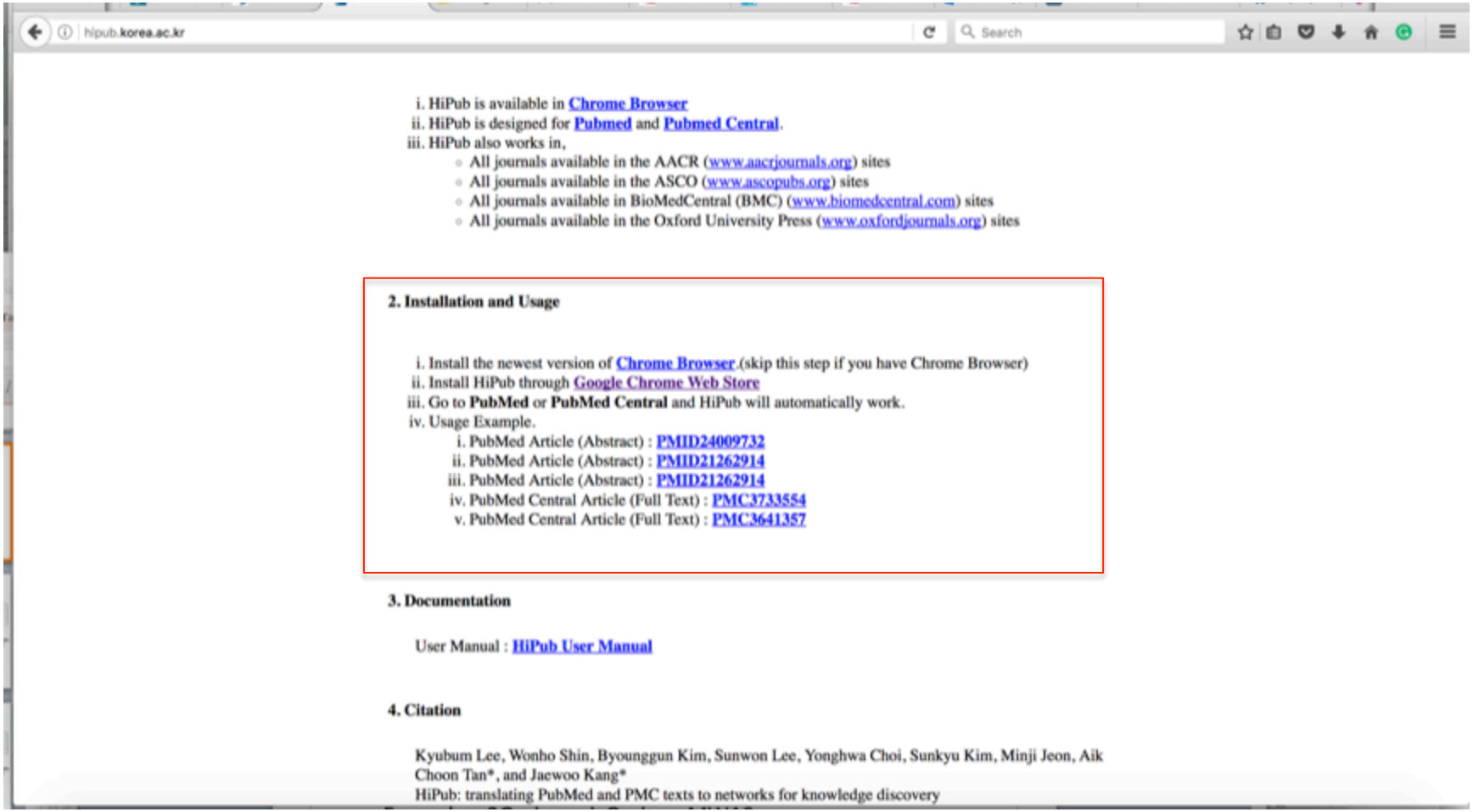
Rows: Variables

Columns: Samples

View Results - GO Enrichment Analysis



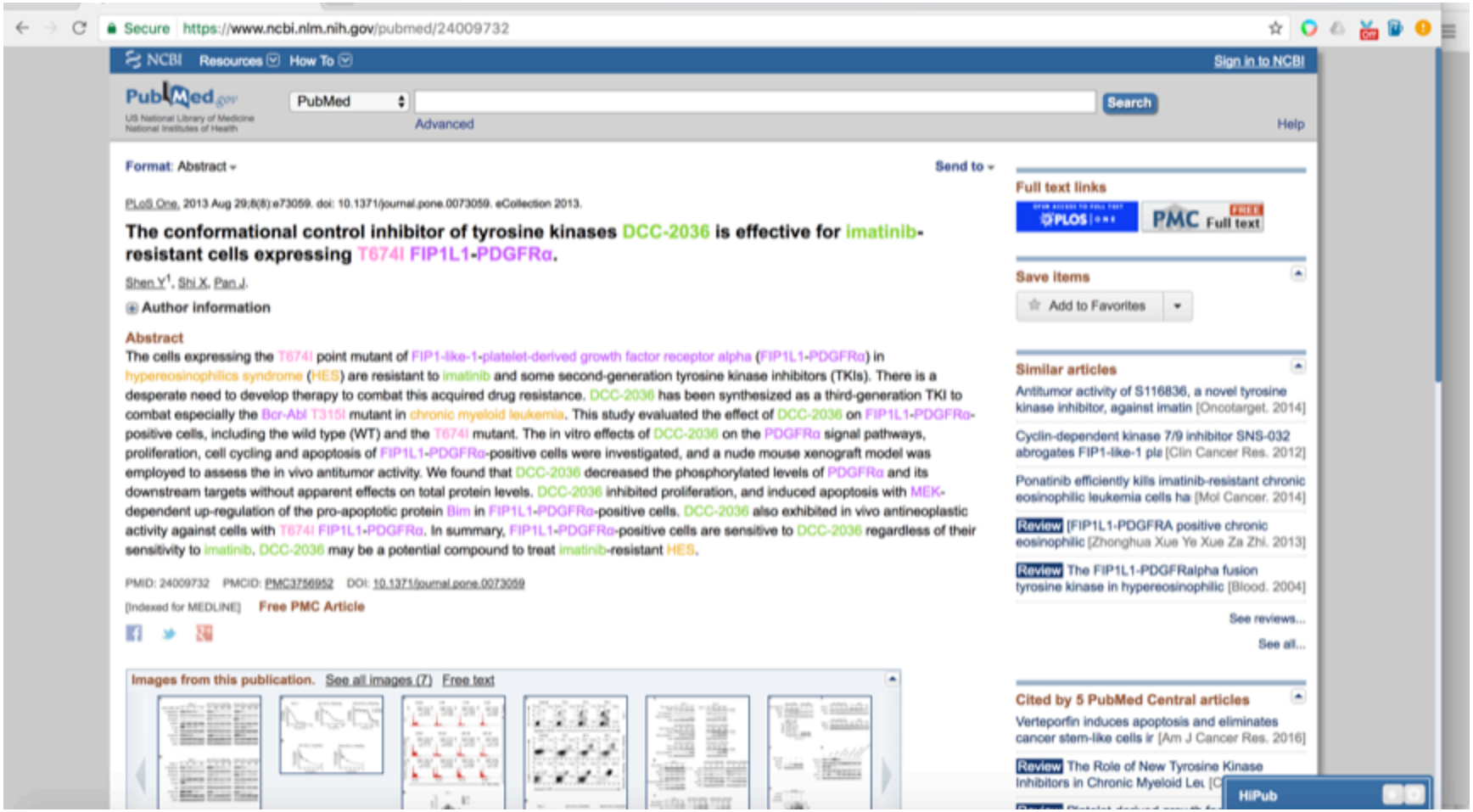
3. HiPub
Go to: http://hipub.korea.ac.kr/



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:
https://www.ncbi.nlm.nih.gov/pubmed/24009732



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"

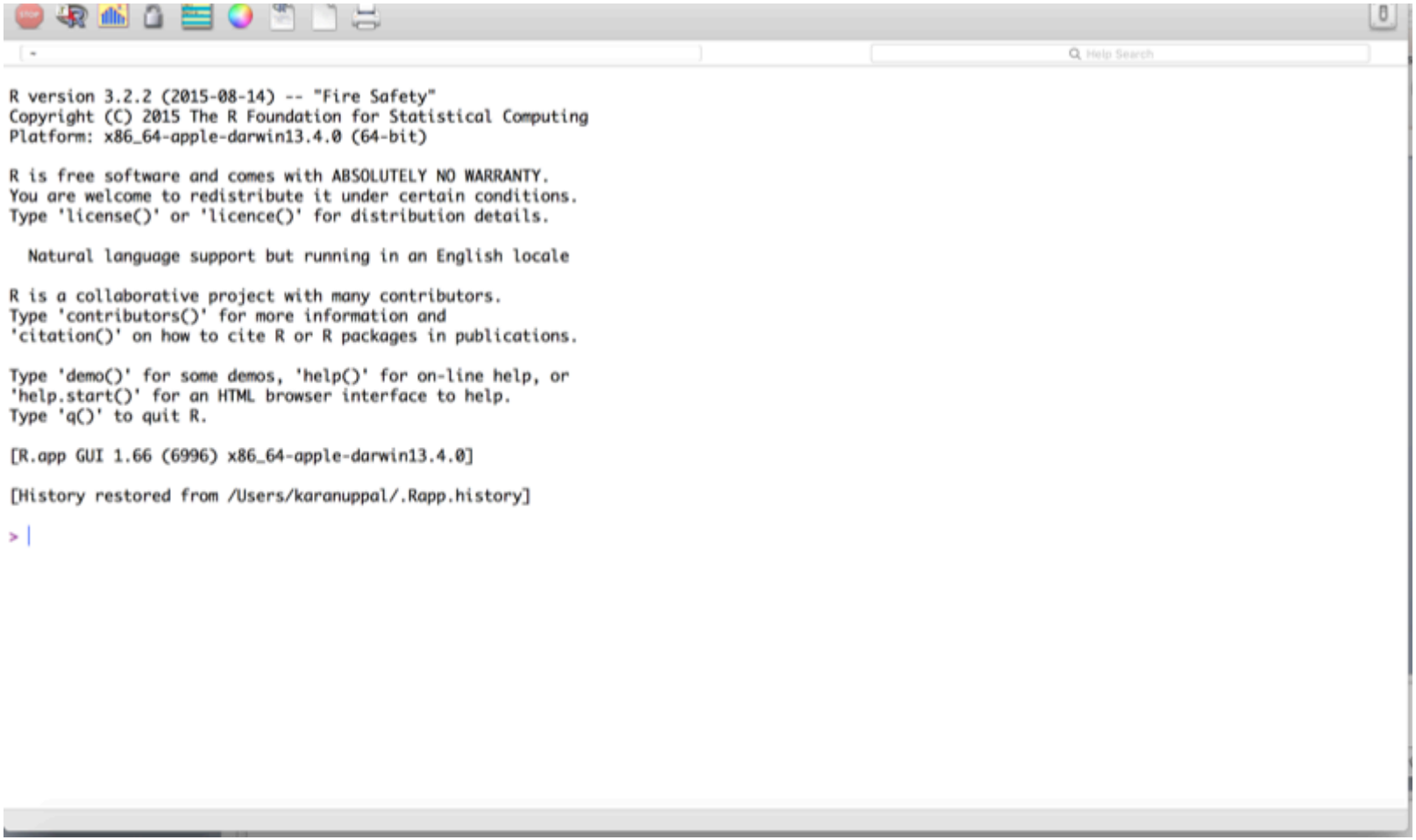
6) Browse to the download location

7) Click "Open"

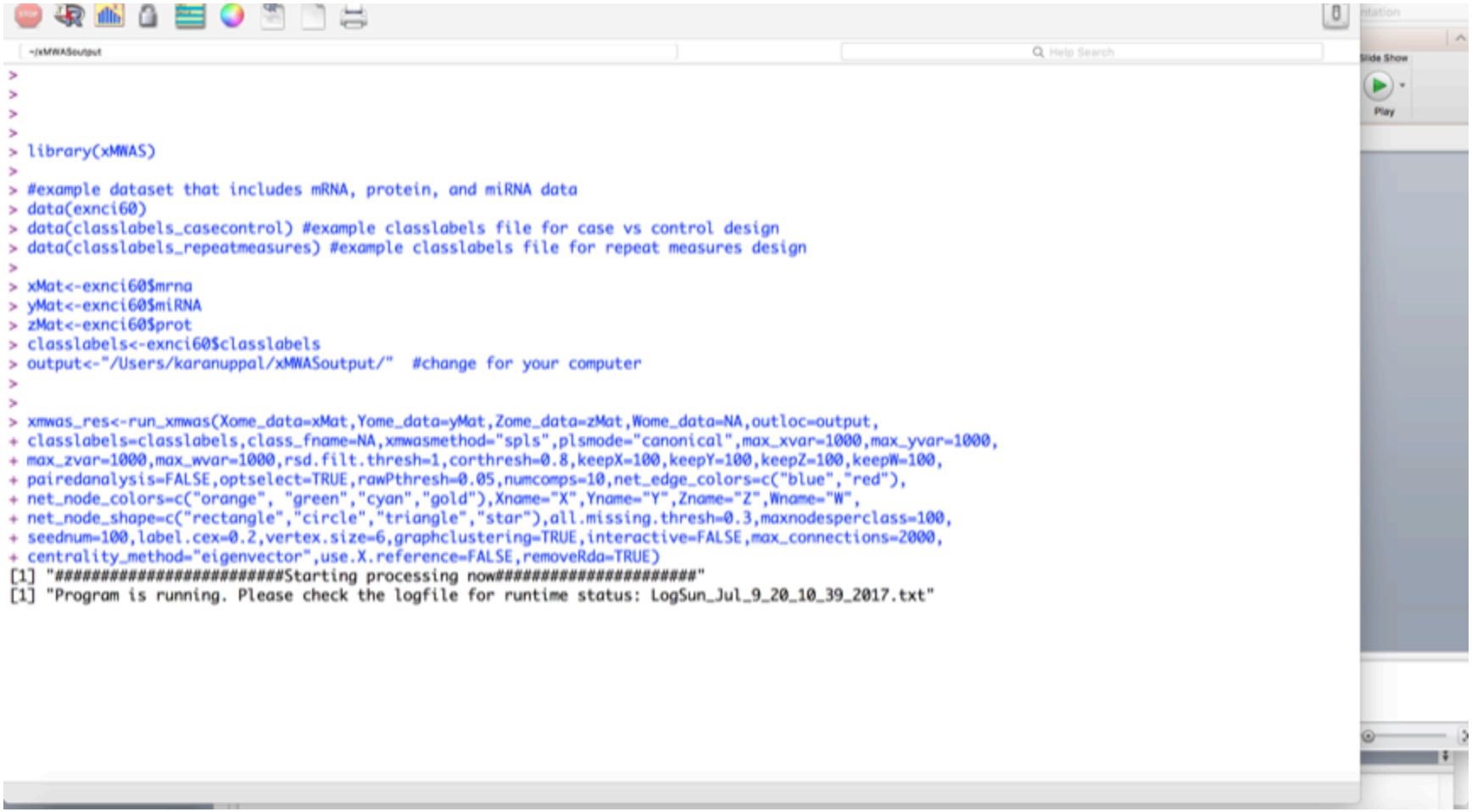
R code

```
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

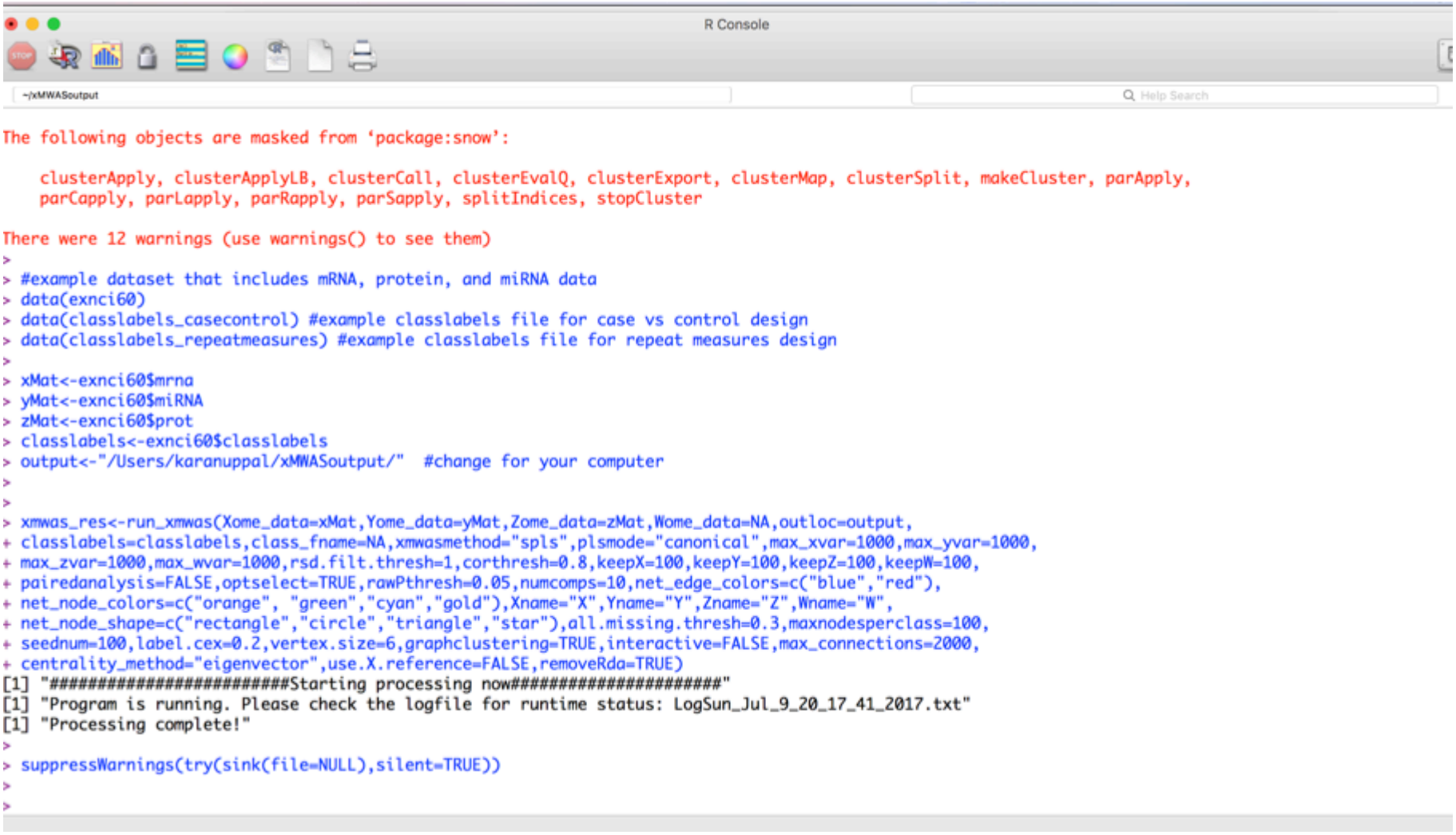

R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[R.app GUI 1.66 (6996) x86_64-apple-darwin13.4.0]
[History restored from /Users/karanuppal/.Rapp.history]
>

Copy and paste the R code on slide 18 (after changing the output location)

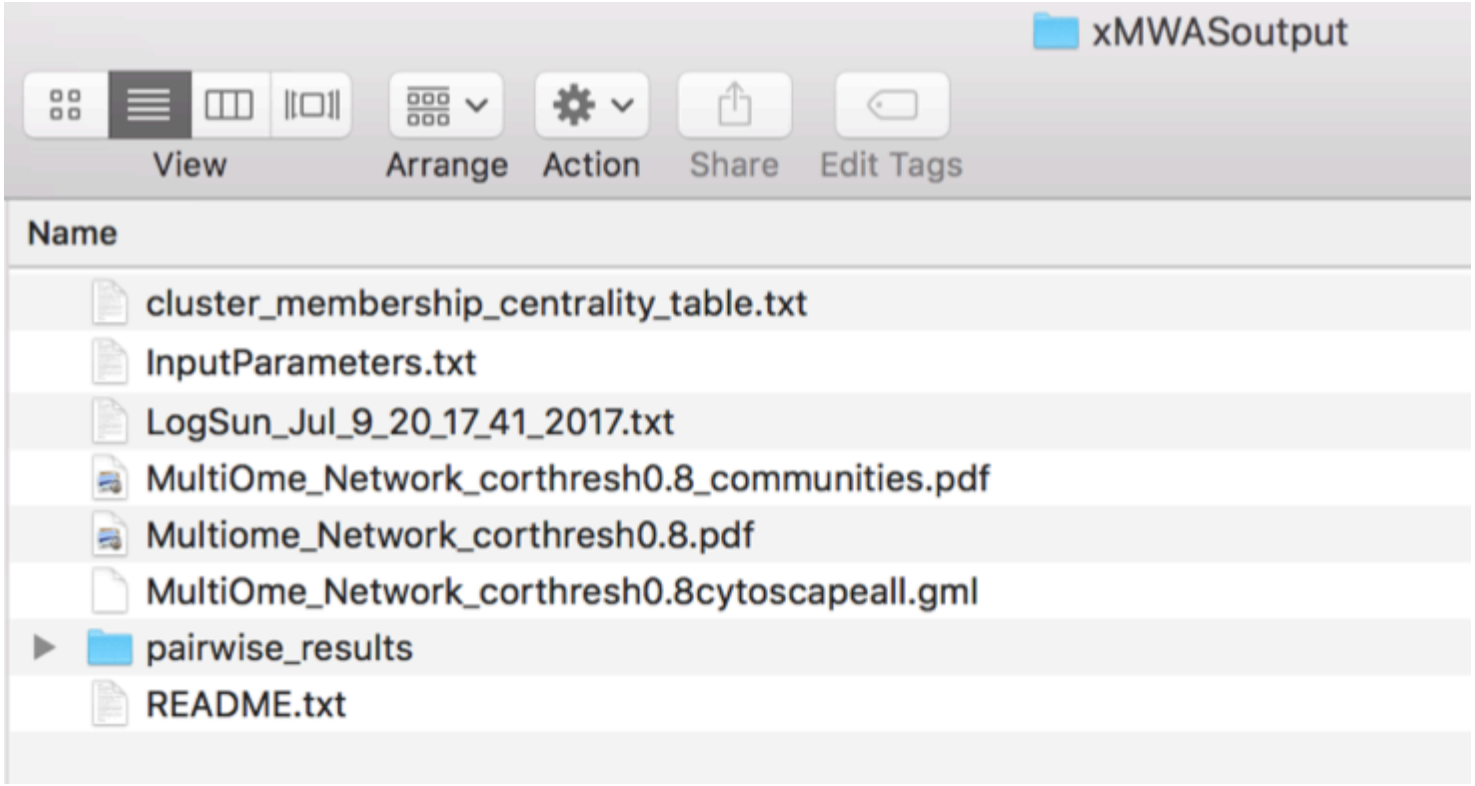


```
> 
> 
> 
> 
> library(xMWAS)
> 
> #example dataset that includes mRNA, protein, and miRNA data
> data(exnci60)
> data(classlabels_casecontrol) #example classlabels file for case vs control design
> data(classlabels_repeatmeasures) #example classlabels file for repeat measures design
> 
> xMat<-exnci60$mrna
> yMat<-exnci60$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)
[1] "########################Starting processing now#####################"
[1] "Program is running. Please check the logfile for runtime status: LogSun_Jul_9_20_10_39_2017.txt"
```

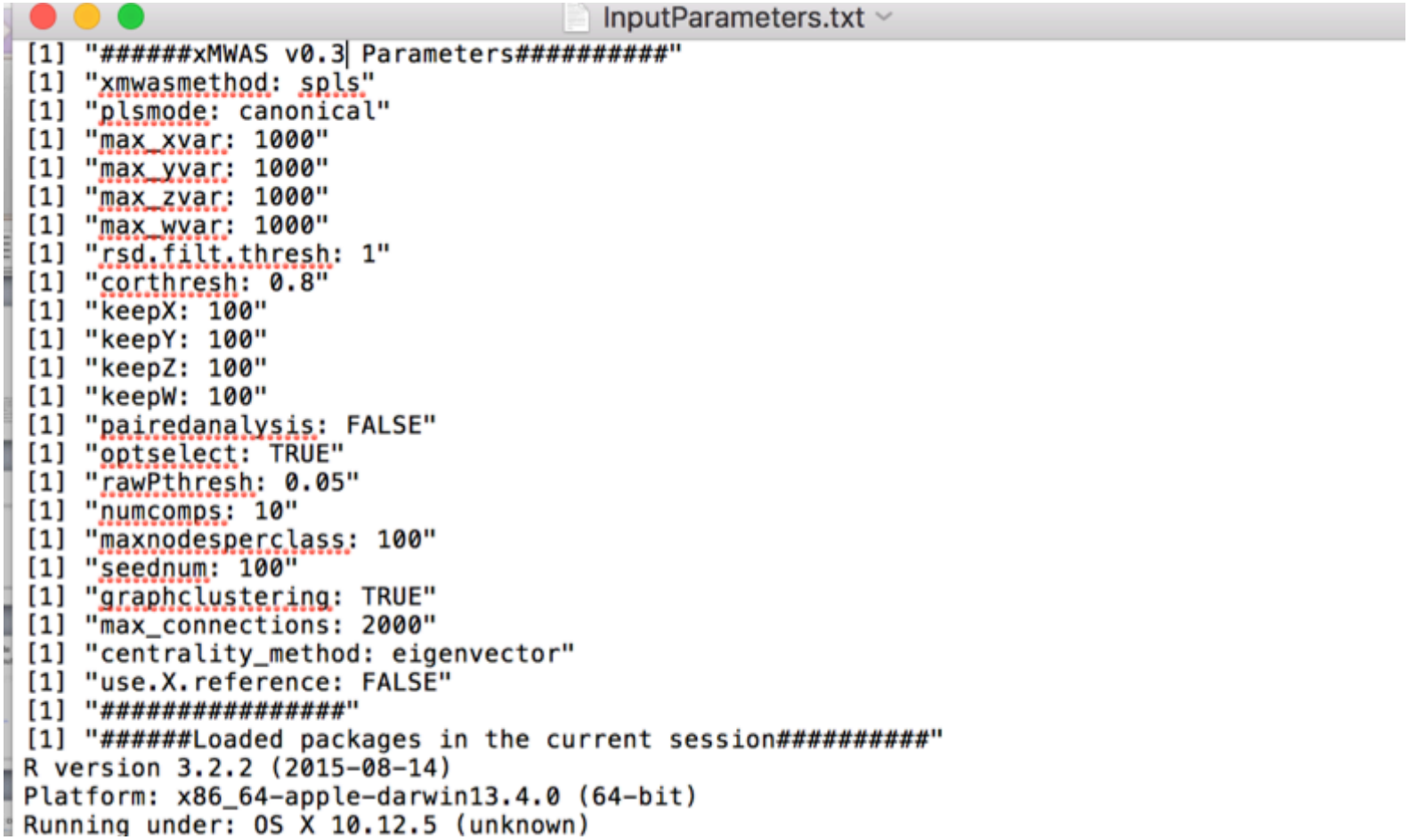
xMWAS – processing complete



xMWAS output



Input Parameters

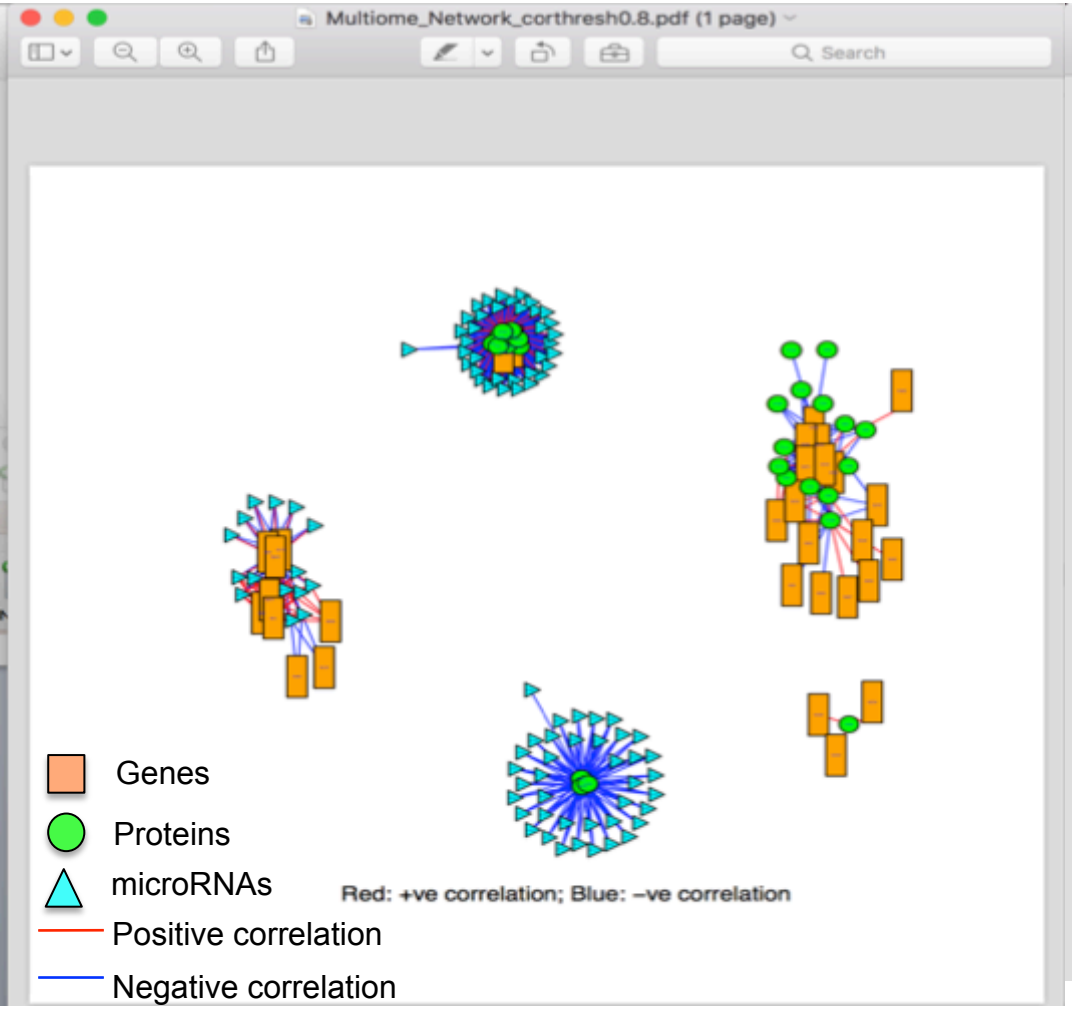


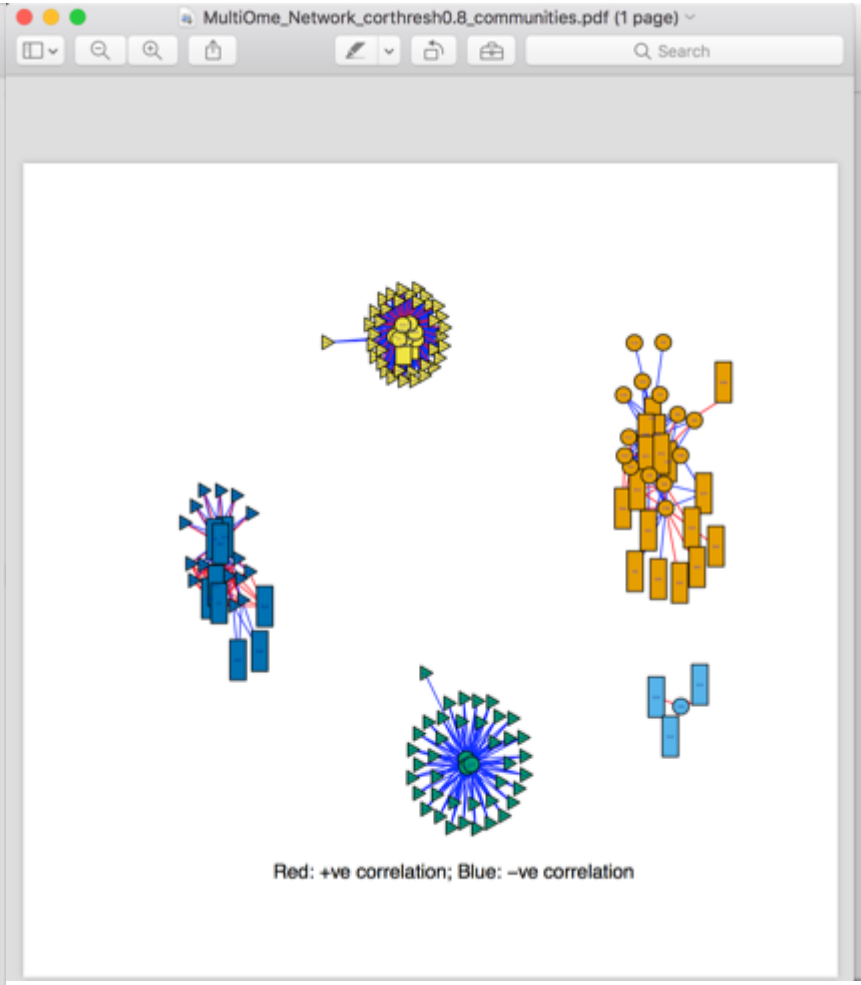
Description of output: Readme.txt

README.txt

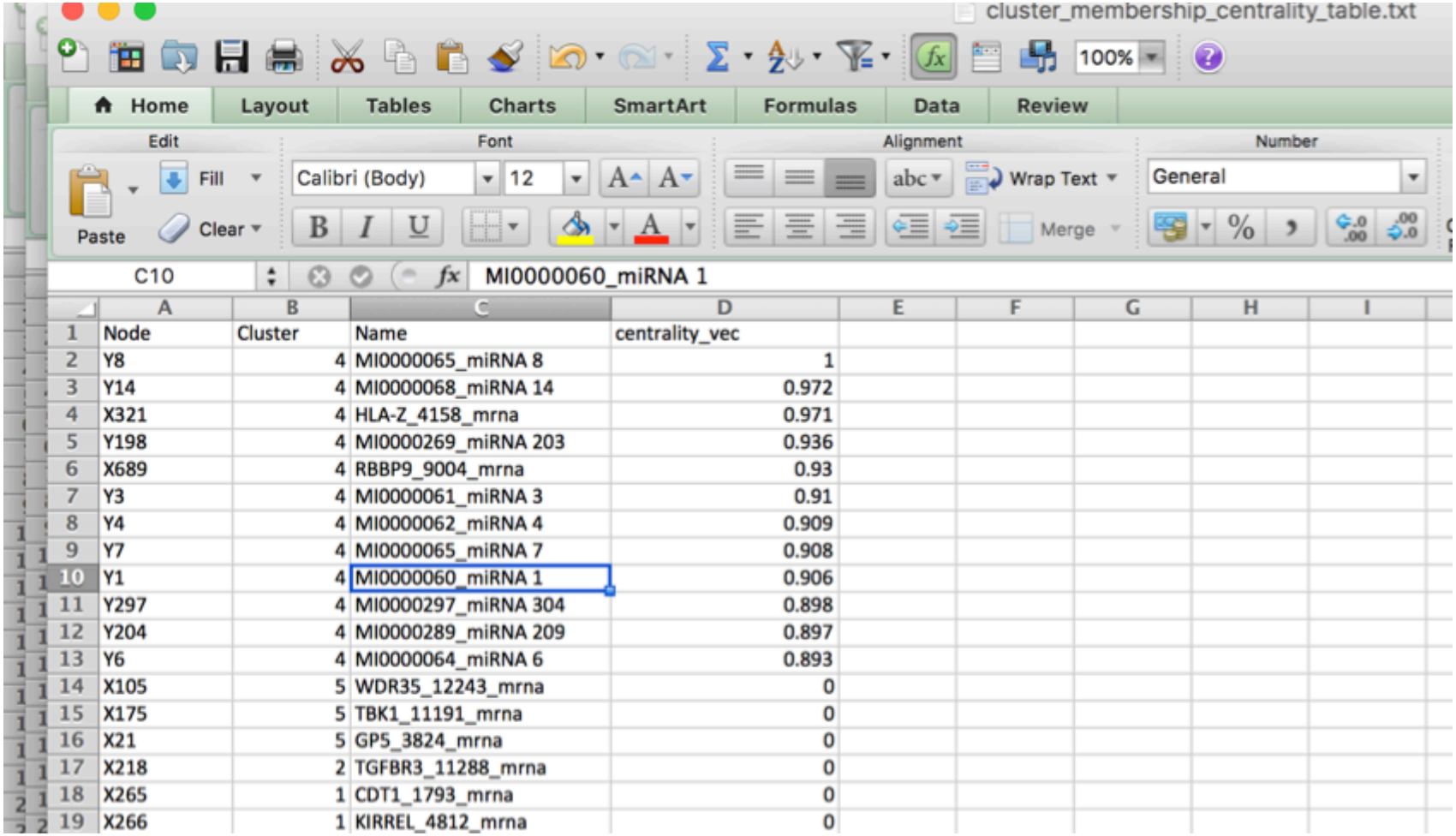
"Description of files"
"1: X labels correspond to xome_fname data, Y labels correspond to yome_fname data, Z labels correspond to zome_fname data, W labels correspond to wome_fname data"
"2: Pairwise integrative analysis results are under pairwise_results. The files corresponding to each pairwise comparison (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 bu 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 detectic 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.gml: 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 ar written in individual subfolders."

Network graphs



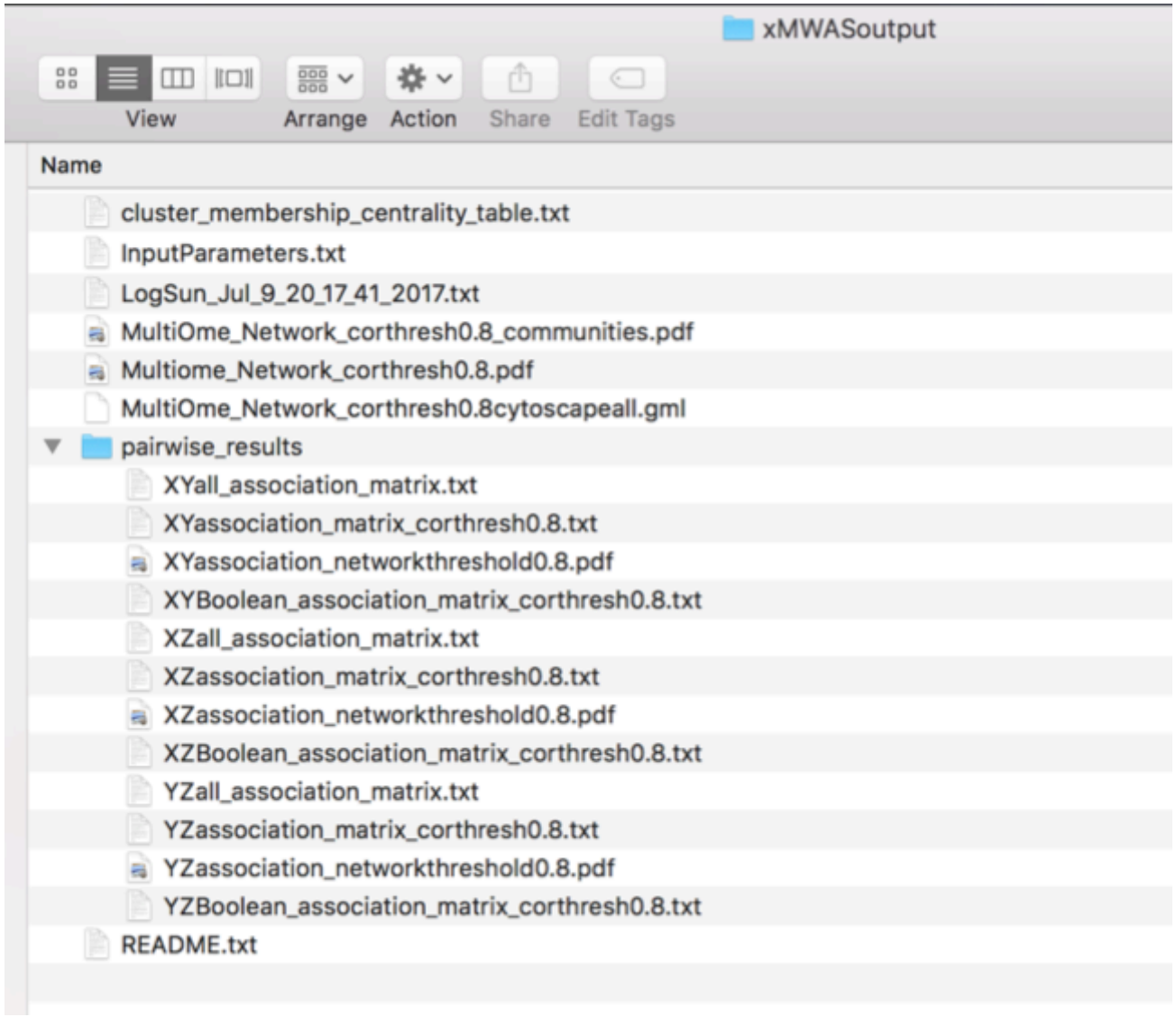


Community detection and centrality analysis



	A	B	C	D	E	F	G	H	I
1	Node	Cluster	Name	centrality_vec					
2	Y8	4	MI0000065_miRNA 8	1					
3	Y14	4	MI0000068_miRNA 14	0.972					
4	X321	4	HLA-Z_4158_mrna	0.971					
5	Y198	4	MI0000269_miRNA 203	0.936					
6	X689	4	RBBP9_9004_mrna	0.93					
7	Y3	4	MI0000061_miRNA 3	0.91					
8	Y4	4	MI0000062_miRNA 4	0.909					
9	Y7	4	MI0000065_miRNA 7	0.908					
10	Y1	4	MI0000060_miRNA 1	0.906					
11	Y297	4	MI0000297_miRNA 304	0.898					
12	Y204	4	MI0000289_miRNA 209	0.897					
13	Y6	4	MI0000064_miRNA 6	0.893					
14	X105	5	WDR35_12243_mrna	0					
15	X175	5	TBK1_11191_mrna	0					
16	X21	5	GP5_3824_mrna	0					
17	X218	2	TGFBR3_11288_mrna	0					
18	X265	1	CDT1_1793_mrna	0					
19	X266	1	KIRREL_4812_mrna	0					

Pairwise results – X<->Y, X<->Z, Y<->Z



Questions/Comments?