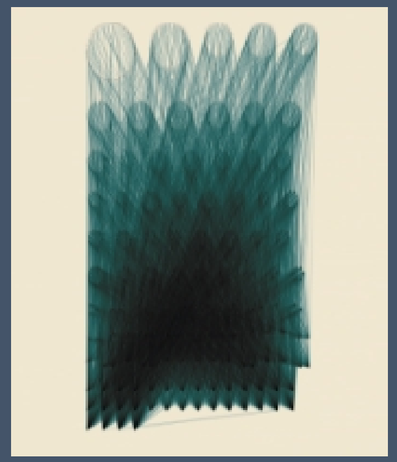
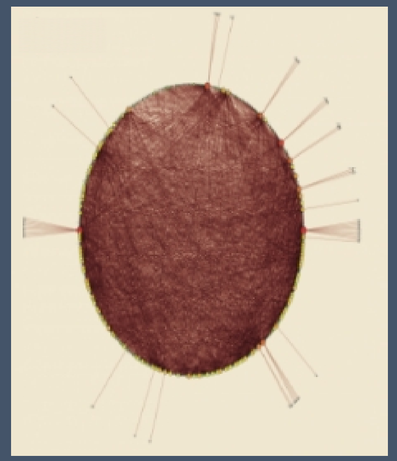
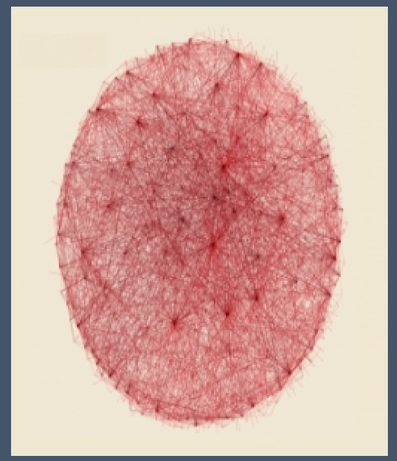
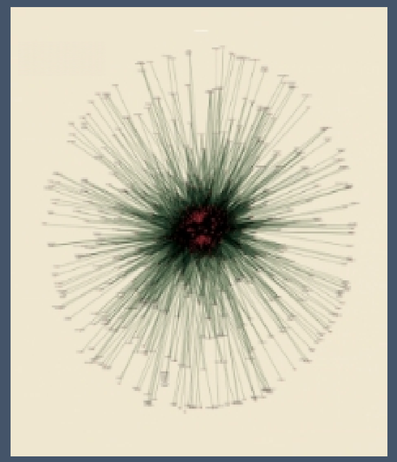
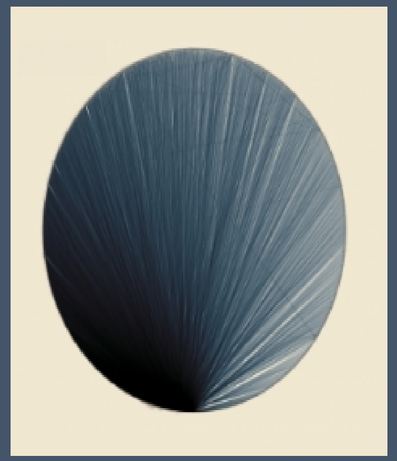
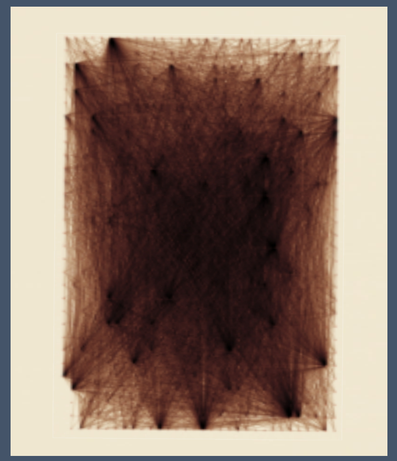


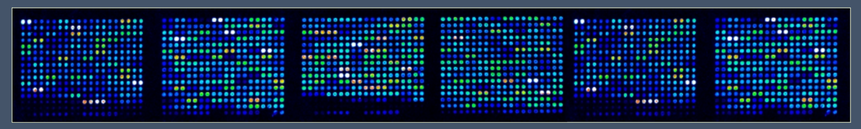












TUTORIAL

# FROM MICROARRAYS TO NETWORKS

Anastasiadou's LAB

# HELLO THERE!

**Degree in MathematicsDepartment of Mathematics and Applied Mathematics, University of Crete. (UoC)**

**M.Sc in Bioinformatics**

**PhD candidate at Medical School of Athens**

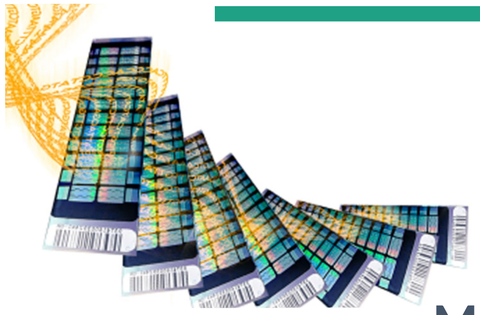
**Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens. (UoA)**

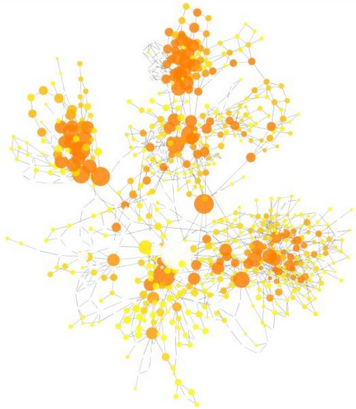
I am Vicky Filippa

You can find me here:
vfilippa@di.uoa.gr
vickyrougefilippa@gmail.com

# FROM MICROARRAYS TO NETWORKS

AN OVERVIEW





## MICROARRAYS

- Obtain Data
- Manipulation of Data
- Differential Expression Analysis (DE)
- Suggested Biomarkers and Visualization of Results

## NETWORKS

- Co-Expression
- Edge-lists-Network Construction
- Types of Networks and Network Visualization
- Network Annotation and Metrics / Obtaining Information

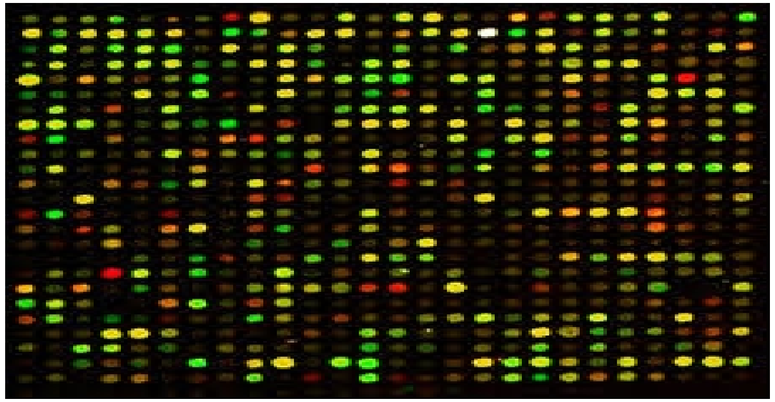

# [1] MICROARRAYS

# [1] MICROARRAYS

**DNA microarrays:**
(*otherwise known as gene or genomic chip, DNA chip or gene array*)
Are collections of microscopic unique DNA spots(probes) attached to a solid surface (glass,silicone). The probes can be long (500-1500bp) cDNA sequences.

**The cDNA technology is a complex electrical-optical-chemical process:**

- cDNA slide fabrication
- mRNA preparation
- fluorescence dye labeling
- gene hybridization
- robotic spotting
- green and red fluorophores excitation by lasers
- imaging using optics
- slide scanning
- analog to digital conversion using either charge-coupled devices (CCD) or photomultiplier tubes (PMT)
- image storage and archiving



Only the pathological sample

Only the "control" sample

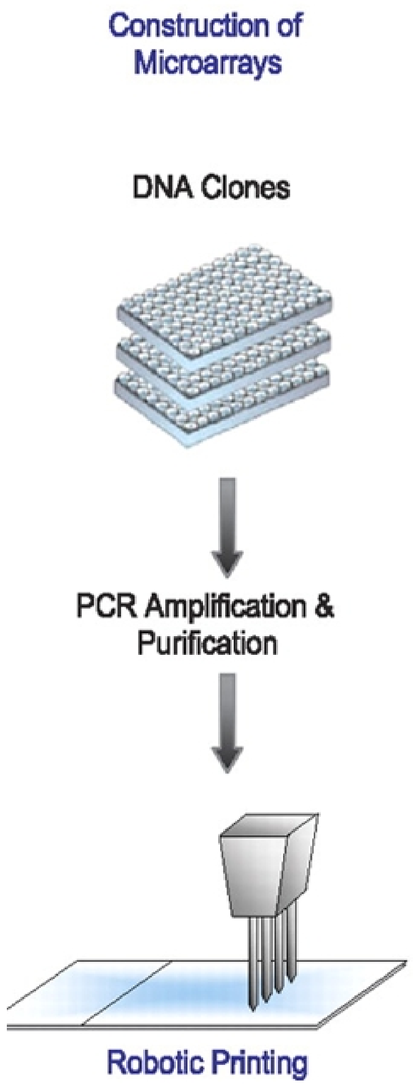
Equal amounts of the gene in pathological and "control" cells

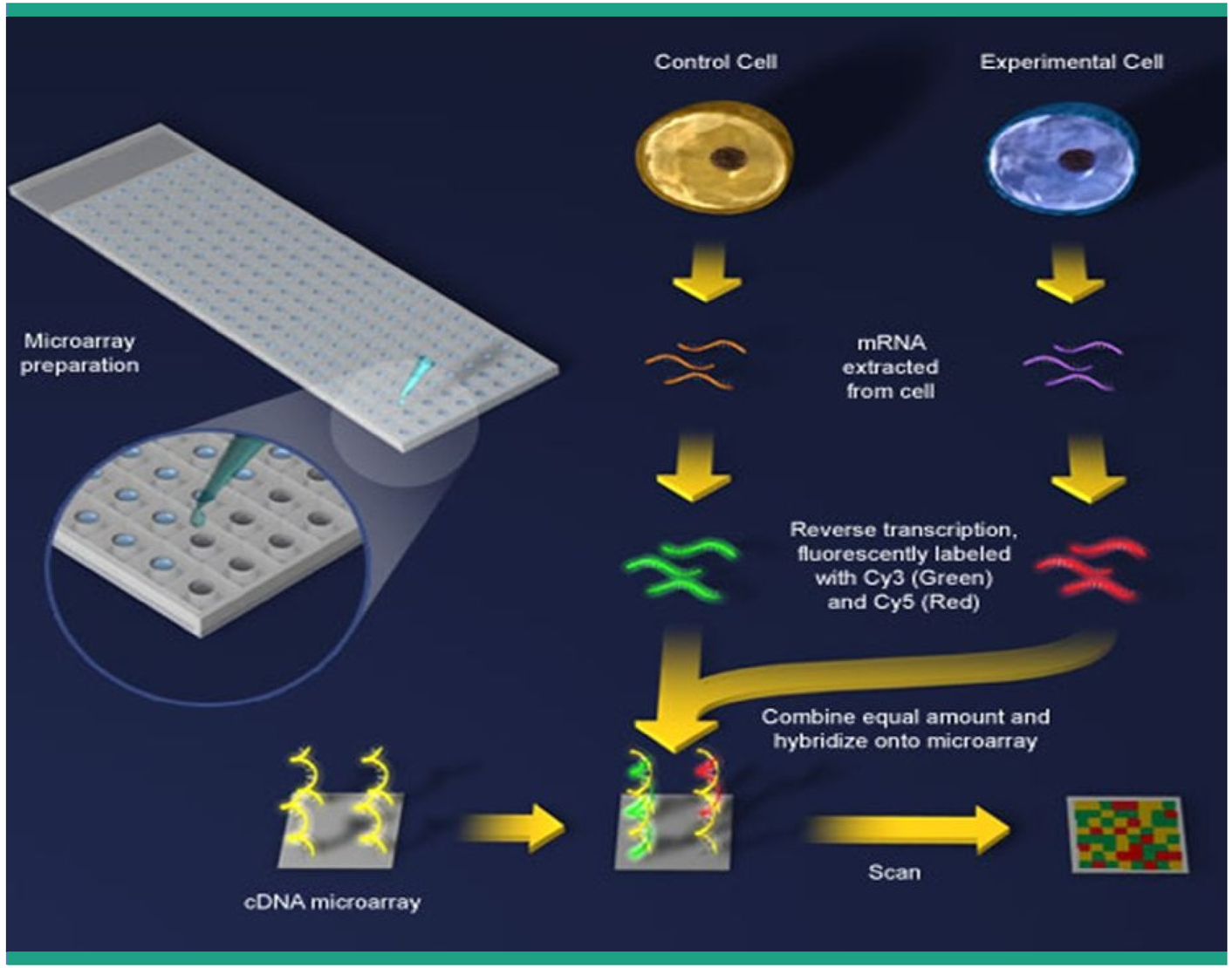
More of the gene's amount(signal) in pathological cells than in "control" cells

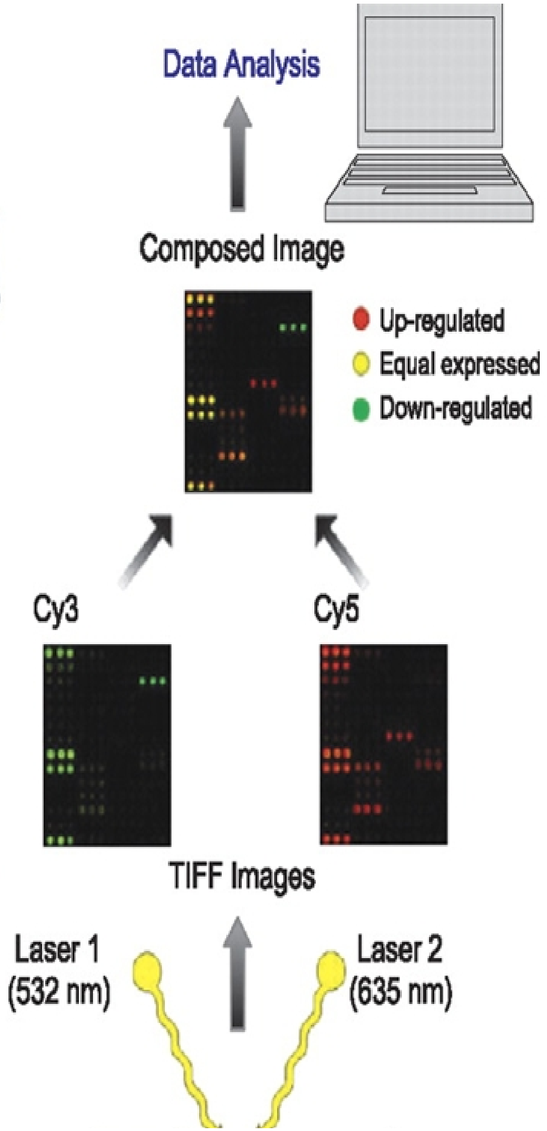
No gene in either pathological or "control" cells

# Microarrays Experiment

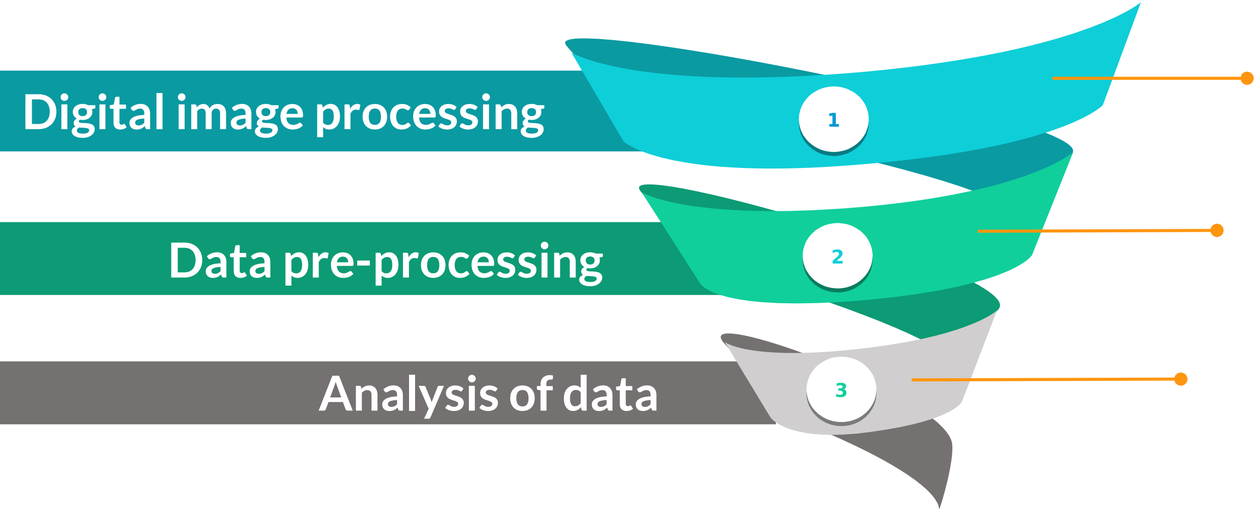
https://www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/



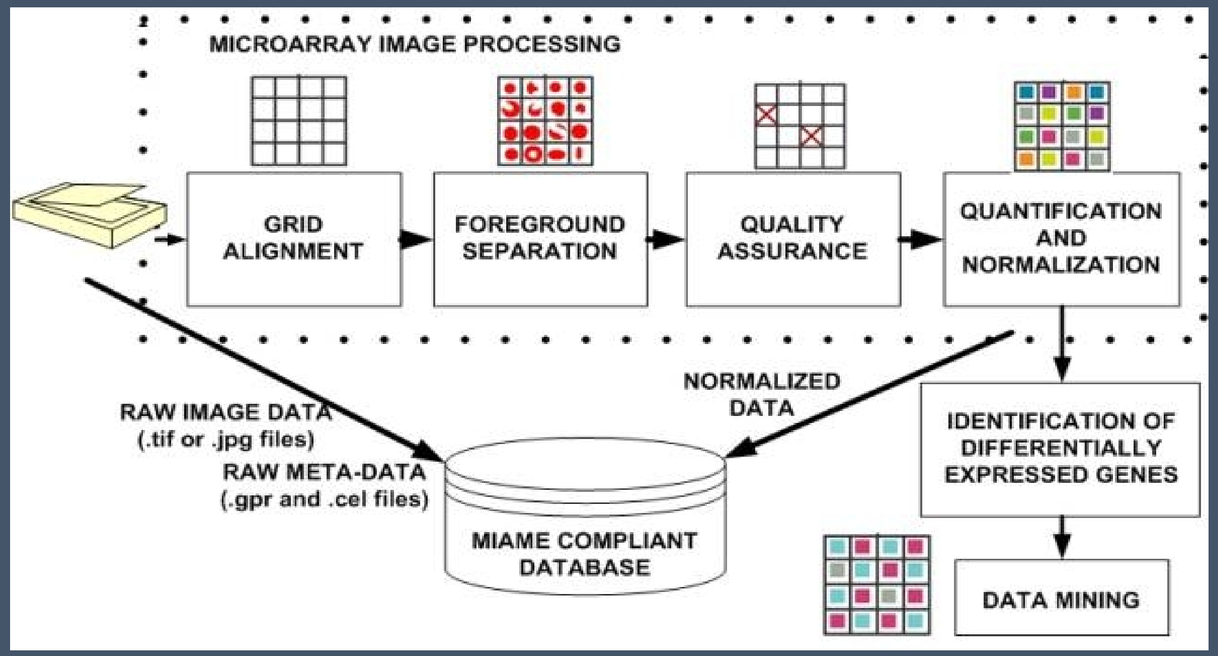
# Steps taken on the data processing part




Segmentation, grid application, export intensity through quantification of the signal

Background correction, normalization, filtering

Statistical methods for the detection of significant differentially expressed genes




- Raw data for each assay (e.g., CEL or FASTQ files)
- Final processed (normalized) data for the set of assays in the study (e.g., the gene expression data count matrix used to draw the conclusions in the study)
- Essential sample annotation (e.g., tissue, sex and age) and the experimental factors and their values (e.g., compound and dose in a dose response study)
- Experimental design including sample data relationships (e.g., which raw data file relates to which sample, which assays are technical, which are biological replicates)
- Sufficient annotation of the array or sequence features examines (e.g., gene identifiers, genomic coordinates)
- Essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

# [2] OBTAINING DATA

# [2] OBTAINING DATA

Search GEO Database for proper datasets :
The Gene Expression Omnibus Genomic Database (GEO), is a public repository of the National Center for Biotechnology Information (NCBI) of high performance experiments. https://www.ncbi.nlm.nih.gov/geo/

NCBI Resources How To Sign in to NCBI

GEO DataSets | GEO DataSets | (nafld) AND "Homo sapiens"[porgn:__txid9606] | Search

Create alert Advanced Help

**Entry type**
DataSets (1)
Series (21)
Samples (0)
Platforms (0)

**Organism**
Customize ...

**Study type** clear
✓ Expression profiling by array
Methylation profiling by array
Customize ...

**Author**
Customize ...

**Attribute name**
tissue (13)
strain (0)
Customize ...

**Publication dates**
30 days
1 year
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

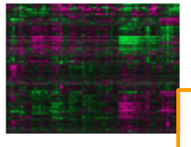
## Search results

**Items: 1 to 20 of 22**

<< First < Prev Page 1 of 2 Next > Last >>

Filters activated: Expression profiling by array. Clear all to show 446 items.

1. Postbariatric, morbidly obese patients with nonalcoholic fatty liver disease: liver biopsies



Analysis of liver from morbidly obese patient representing nonalcoholic fatty liver disease (**NAFLD**) subtypes steatosis and nonalcoholic steatohepatitis (NASH), post-bariatric surgery. Results provide insight into molecular basis of the **NAFLD** liver phenotypes and into postbariatric molecular changes.
Organism: **Homo sapiens**
Type: **Expression profiling by array**, transformed count, 4 disease state, 3 protocol sets
Platform: GPL11532 Series: GSE48452 73 Samples
Download data: CEL
DataSet Accession: GDS4881 ID: 4881
PubMed Similar studies GEO Profiles Analyze DataSet

2. Gene expression profiling from high-fat medium (HFM)-treated and growth medium (GM)-treated Sk-hep1 cells

(Submitter supplied) Non-alcoholic fatty liver disease (**NAFLD**) is a major problem in obese peoples and caused by unbalanced uptake of fatty acid. Novel drug identification is necessary to develop effective therapies. We combine LOPAC® and High-Content system to identify compounds significantly reducing intracellular lipid droplets after high fat medium (HFM) treatment. Among 1280 compounds, 5 show efficacy in

**Filters:** Manage Filters
Top Organisms [Tree]
Homo sapiens *(22)*

**Find related data**
Database: Select
Find items

**Search details**
```
nafld[All Fields] AND "Homo
sapiens"[porgn] AND "Expression
profiling by array"[Filter]
```
Search See more...

**Recent activity**
Turn Off Clear
(nafld) AND "Homo sapiens"[porgn] AND ("Expression profiling by a... (22) GEO DataSets

# [2] OBTAINING DATA

**Search GEO Database for proper datasets :**

The Gene Expression Omnibus Genomic Database (GEO), is a public repository of the National Center for Biotechnology Information (NCBI) of high performance experiments. https://www.ncbi.nlm.nih.gov/geo/

**Series GSE89632** Query DataSets for GSE89632

| | |
|---|---|
| Status | Public on Nov 08, 2016 |
| Title | Genome-wide analysis of hepatic gene expression in patients with non-alcoholic fatty liver disease and in healthy donors in relation to hepatic fatty acid composition and other nutritional factors |

| | |
|---|---|
| Submission date | Nov 07, 2016 |
| Last update date | Dec 22, 2017 |
| Contact name | Johane P. Allard |
| E-mail | johane.allard@uhn.on.ca |
| Phone | 416-340-5159 |
| Organization name | University Health Network |
| Department | Medicine |
| Street address | 200 Elizabeth St, 9-NU-973 |
| City | Toronto |
| State/province | Ontario |
| ZIP/Postal code | M5G 2C4 |
| Country | Canada |

| | |
|---|---|
| Platforms (1) | GPL14951 Illumina HumanHT-12 WG-DASL V4.0 R2 expression beadchip |
| Samples (63) | GSM2385720 liver_SS_CL-86 |
| ⊞ More... | GSM2385721 liver_NASH_CL-87 |
| | GSM2385722 liver_SS_CL-88 |

**Relations**

| | |
|---|---|
| BioProject | PRJNA352744 |

Analyze with GEO2R

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

AFFYMETRIX
AGILENT
ILUMINA

GSM2385720 liver_SS_CL-86
GSM2385721 liver_NASH_CL-87
GSM2385722 liver_SS_CL-88
GSM2385723 liver_SS_CL-90
GSM2385724 liver_SS_CL-91
GSM2385725 liver_SS_CL-92
GSM2385726 liver_SS_CL-95
GSM2385727 liver_SS_CL-96
GSM2385728 liver_NASH_CL-97
GSM2385729 liver_NASH_CL-98
GSM2385730 liver_SS_CL-100
GSM2385731 liver_NASH_CL-103
GSM2385732 liver_NASH_CL-106
GSM2385733 liver_SS_CL-108
GSM2385734 liver_SS_CL-110
GSM2385735 liver_NASH_CL-111
GSM2385736 liver_NASH_CL-112
GSM2385737 liver_NASH_CL-113
GSM2385738 liver_SS_CL-114
GSM2385739 liver_NASH_CL-116
GSM2385740 liver_SS_CL-117
GSM2385741 liver_NASH_CL-118
GSM2385742 liver_NASH_CL-128
GSM2385743 liver_NASH_CL-132
GSM2385744 liver_SS_CL-134
GSM2385745 liver_SS_CL-136
GSM2385746 liver_SS_CL-140
GSM2385747 liver_SS_CL-142
GSM2385748 liver_NASH_CL-144
GSM2385749 liver_SS_CL-145
GSM2385750 liver_NASH_CL-147
GSM2385751 liver_NASH_CL-152
GSM2385752 liver_NASH_CL-155
GSM2385753 liver_NASH_CL-157
GSM2385754 liver_NASH_CL-160
GSM2385755 liver_SS_CL-161
GSM2385756 liver_NASH_CL-167

**"DISEASE"**
**2 CONDITIONS**

GSM2385757 liver_HC_HLD-1
GSM2385758 liver_HC_HLD-2
GSM2385759 liver_HC_HLD-3
GSM2385760 liver_HC_HLD-4
GSM2385761 liver_HC_HLD-5
GSM2385762 liver_HC_HLD-7
GSM2385763 liver_HC_HLD-8
GSM2385764 liver_HC_HLD-10
GSM2385765 liver_HC_HLD-11
GSM2385766 liver_HC_HLD-13
GSM2385767 liver_HC_HLD-14
GSM2385768 liver_HC_HLD-21
GSM2385769 liver_HC_HLD-23

**"CONTROLS"**

# SERIES MATRIX

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| !Sample_ | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM521 |
| !Sample_ | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or | Public or |
| !Sample_ | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Feb 11 20 | Mar 11 20 |
| !Sample_ | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C | Oct 13 2C |
| !Sample_ | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA | RNA |
| !Sample_ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| !Sample_ | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort | Postmort |
| !Sample_ | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa | Homo sa |
| !Sample_ | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s | disease s |
| !Sample_ | age: 57 | age: 73 | age: 80 | age: 84 | age: 70 | age: 82 | age: 70 | age: 80 | age: 94 | age: 70 | age: 79 | age: 67 | age: 54 | age: 73 | age: 82 | age: 72 | age: 73 | age: 75 | age: 74 | age: 72 | age: 79 | age: 67 | age: 75 | age: 81 | age: 55 | age: 59 |
| !Sample_ | gender: r | gender: r | gender: f | gender: f | gender: f | gender: f | gender: r | gender: r | gender: f | gender: r | gender: r | gender: r | gender: r | gender: f | gender: r | gender: f | gender: f | gender: r | gender: r | gender: f | gender: f | gender: r | gender: r | gender: r | gender: r | gender: r |
| !Sample_ | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi | brain regi |
| !Sample_ | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA | total RNA |
| !Sample_ | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( | RNeasy ( |
| !Sample_ | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin | biotin |
| !Sample_ | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc |
| !Sample_ | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 | 9606 |
| !Sample_ | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc |
| !Sample_ | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc | Standarc |
| !Sample_ | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain | To obtain |
| !Sample_ | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 | GPL96 |
| !Sample_ | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I | Frank,A,I |
| !Sample_ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ | middletf@ |
| !Sample_ | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- | 315-464- |
| !Sample_ | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci | Neurosci |
| !Sample_ | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up | SUNY Up |
| !Sample_ | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East | 750 East |
| !Sample_ | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse | Syracuse |
| !Sample_ | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY | NY |
| !Sample_ | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 | 13210 |
| !Sample_ | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| !Sample_ | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n |
| !Sample_ | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n | ftp://ftp.n |
| !Sample_ | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 | 22283 |
| !series_matrix_table_begin | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ID_REF | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM508 | GSM521 |
| 1007_s_a | 11.286 | 11.985 | 12.117 | 12.554 | 12.22 | 11.833 | 12.096 | 12.39 | 11.637 | 12.053 | 11.823 | 11.827 | 11.743 | 11.324 | 11.525 | 11.693 | 11.608 | 12.047 | 12.209 | 11.075 | 11.801 | 11.898 | 11.877 | 11.866 | 12.018 | 11.64 |
| 1053_at | 5.8439 | 5.8409 | 5.7939 | 5.9169 | 5.8513 | 5.8303 | 5.8266 | 5.896 | 5.8289 | 5.9001 | 5.8436 | 5.8333 | 5.8274 | 5.8743 | 5.8516 | 5.9841 | 6.0419 | 5.8597 | 5.958 | 5.8624 | 5.9845 | 5.9387 | 5.9516 | 5.851 | 5.9288 | 5.6548 |
| 117_at | 7.3687 | 7.3774 | 7.3197 | 7.6517 | 7.1858 | 7.6661 | 7.1125 | 7.3284 | 7.4564 | 7.2357 | 7.3389 | 7.8559 | 7.3171 | 7.3556 | 7.0625 | 7.173 | 7.1668 | 11.089 | 7.3357 | 7.3594 | 7.4564 | 7.44 | 7.3902 | 7.2228 | 7.0503 | 7.1193 |
| 121_at | 9.4967 | 9.939 | 9.594 | 9.6975 | 9.3741 | 9.3762 | 9.7555 | 9.858 | 9.2954 | 9.4759 | 9.8099 | 9.663 | 9.5933 | 9.6647 | 9.3443 | 9.4641 | 9.3279 | 9.163 | 9.1242 | 8.901 | 9.3748 | 9.6676 | 9.2057 | 9.6187 | 9.2363 | 9.4453 |
| 1255_g_. | 6.0813 | 5.5575 | 5.4817 | 5.4541 | 5.3776 | 5.4103 | 5.5126 | 5.5368 | 5.4891 | 5.5909 | 5.6156 | 5.4347 | 5.8033 | 5.4955 | 5.5338 | 5.6893 | 5.5616 | 5.3583 | 5.4656 | 5.4112 | 5.3274 | 5.6793 | 5.7865 | 5.4825 | 5.6432 | 5.6236 |
| 1294_at | 8.0129 | 8.4014 | 8.2456 | 8.4021 | 8.4227 | 8.477 | 8.2562 | 8.7169 | 8.2674 | 8.5618 | 8.0539 | 8.2826 | 8.1606 | 8.2388 | 8.2395 | 8.2247 | 8.3783 | 7.9749 | 8.2636 | 8.2876 | 8.2959 | 8.4626 | 8.0784 | 8.3181 | 8.0265 | 7.9776 |
| 1316_at | 6.4968 | 6.8414 | 6.6959 | 6.3892 | 6.4517 | 6.538 | 6.4141 | 6.8714 | 6.3958 | 6.5436 | 6.4827 | 6.7134 | 6.3492 | 7.106 | 6.8789 | 6.7811 | 7.0821 | 6.3847 | 6.3128 | 6.2544 | 6.2163 | 6.9094 | 6.32 | 6.546 | 6.2326 | 6.336 |
| 1320_at | 6.0841 | 6.3006 | 6.1628 | 6.1983 | 6.1515 | 6.1784 | 6.0482 | 6.1184 | 6.0847 | 6.0949 | 6.1182 | 6.1003 | 6.1547 | 6.1795 | 6.13 | 6.1384 | 6.171 | 6.0087 | 5.9978 | 5.9893 | 6.0243 | 6.515 | 5.8398 | 6.0967 | 5.9701 | 5.9455 |
| 1405_i_a | 5.649 | 5.3781 | 5.6378 | 5.7095 | 5.4237 | 5.6548 | 5.321 | 5.6809 | 5.554 | 5.4542 | 5.4717 | 5.687 | 5.446 | 5.397 | 5.646 | 5.4874 | 5.4506 | 5.3091 | 5.3627 | 5.4967 | 5.2396 | 5.5268 | 5.3208 | 5.8871 | 5.2221 | 5.3696 |
| 1431_at | 5.1778 | 5.2764 | 5.1422 | 5.1335 | 5.1768 | 5.095 | 5.0227 | 5.0939 | 5.1333 | 5.1226 | 5.0742 | 5.0909 | 5.0971 | 5.1266 | 4.9404 | 5.0063 | 5.0274 | 4.9198 | 5.0458 | 5.0255 | 4.8439 | 5.1383 | 5.1746 | 5.1409 | 5.1292 | 5.0429 |

# ANNOTATION TABLE/PLATFORM

#ID = Unique identifier for the probe (across all products and species)
#Transcript = Internal transcript id
#Species =
#Source = Transcript sequence source name
#Search_Key = Internal id useful for custom design array
#ILMN_Gene = Internal gene symbol
#Source_Reference_ID = Id in the source database
#RefSeq_ID = Refseq id
#Entrez_Gene_ID = Entrez gene id
#GI = Genbank id
#Accession = Genbank accession number
#Symbol = Gene symbol from the source database
#Protein_Product = Genbank protein accession number
#Array_Address_Id = Decoder id
#Probe_Type = Information about what this probe is targeting
#Probe_Start = Position of the probe relative to the 5' of the source transcript sequence
#SEQUENCE = Probe sequence
#Chromosome = Chromosome
#Probe_Chr_Orientation = Orientation on the NCBI genome built
#Probe_Coordinates = genomic position of the probe on the NCBI genome build 36 vers
#Cytoband =
#Definition = Gene description from the source
#Ontology_Component = Cellular component annotations from Gene Ontology project
#Ontology_Process = Biological process annotations from Gene Ontology project
#Ontology_Function = Molecular function annotations from Gene Ontology project
#Synonyms = Gene symbol synonyms from Refseq
#Obsolete_Probe_Id = Identifier of probe id before bgx time
#GB_ACC = GenBank accession

| ID | Transcript | Species | Source | Search_Key | ILMN_Gene | Source_Reference_ID | RefSeq_ID | Entrez_Gene_ID | GI | Accession | Symbol |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ILMN_1736555 | ILMN_13581 | Homo sapiens | RefSeq | NM_001002844.1 | ZNF280D | NM_001002844.1 | NM_001002844.1 | 54816 | 50811874 | NM_001002844.1 | ZNF280D |
| ILMN_1664176 | ILMN_29187 | Homo sapiens | RefSeq | NM_006329.2 | FBLN5 | NM_006329.2 | NM_006329.2 | 10516 | 19743802 | NM_006329.2 | FBLN5 |
| ILMN_2223941 | ILMN_29187 | Homo sapiens | RefSeq | NM_006329.2 | FBLN5 | NM_006329.2 | NM_006329.2 | 10516 | 19743802 | NM_006329.2 | FBLN5 |
| ILMN_2399503 | ILMN_172742 | Homo sapiens | RefSeq | NM_001079514.1 | UBN1 | NM_001079514.1 | NM_001079514.1 | 29855 | 118572602 | NM_001079514.1 | UBN1 |
| ILMN_2290089 | ILMN_172742 | Homo sapiens | RefSeq | NM_001079514.1 | UBN1 | NM_001079514.1 | NM_001079514.1 | 29855 | 118572602 | NM_001079514.1 | UBN1 |
| ILMN_1762294 | ILMN_23416 | Homo sapiens | RefSeq | NM_025008.2 | ADAMTSL4 | NM_025008.3 | NM_025008.3 | 54507 | 83281434 | NM_025008.3 | ADAMTSL4 |
| ILMN_1687035 | ILMN_23416 | Homo sapiens | RefSeq | NM_025008.2 | ADAMTSL4 | NM_025008.3 | NM_025008.3 | 54507 | 83281434 | NM_025008.3 | ADAMTSL4 |
| ILMN_2174296 | ILMN_168524 | Homo sapiens | RefSeq | NM_014377.1 | DNAJC2 | NM_014377.1 | NM_014377.1 | 27000 | 94538369 | NM_014377.1 | DNAJC2 |
| ILMN_1697634 | ILMN_183260 | Homo sapiens | RefSeq | NM_173616.1 | FLJ35894 | XM_001131199.1 | XM_001131199.1 | 283847 | 113426471 | XM_001131199.1 | FLJ35894 |
| ILMN_1758315 | ILMN_20716 | Homo sapiens | RefSeq | NM_173653.1 | SLC9A9 | NM_173653.1 | NM_173653.1 | 285195 | 27734934 | NM_173653.1 | SLC9A9 |
| ILMN_2166696 | ILMN_15984 | Homo sapiens | RefSeq | NM_178127.2 | ANGPTL5 | NM_178127.2 | NM_178127.2 | 253935 | 31342398 | NM_178127.2 | ANGPTL5 |
| ILMN_1681234 | ILMN_8091 | Homo sapiens | RefSeq | NM_007185.3 | TNRC4 | NM_007185.3 | NM_007185.3 | 11189 | 71164893 | NM_007185.3 | TNRC4 |
| ILMN_1710329 | ILMN_8872 | Homo sapiens | RefSeq | NM_016132.2 | MYEF2 | NM_016132.3 | NM_016132.3 | 50804 | 154146212 | NM_016132.3 | MYEF2 |
| ILMN_1813671 | ILMN_28181 | Homo sapiens | RefSeq | NM_005984.1 | SLC25A1 | NM_005984.1 | NM_005984.1 | 6576 | 21389314 | NM_005984.1 | SLC25A1 |
| ILMN_1700633 | ILMN_4184 | Homo sapiens | RefSeq | NM_022060.2 | ABHD4 | NM_022060.2 | NM_022060.2 | 63874 | 50658086 | NM_022060.2 | ABHD4 |
| ILMN_1752229 | ILMN_15521 | Homo sapiens | RefSeq | NM_001001563.1 | TIMM50 | NM_001001563.1 | NM_001001563.1 | 92609 | 48526508 | NM_001001563.1 | TIMM50 |
| ILMN_2332691 | ILMN_8100 | Homo sapiens | RefSeq | NM_173087.1 | CAPN3 | NM_173087.1 | NM_173087.1 | 825 | 27765073 | NM_173087.1 | CAPN3 |
| ILMN_1734794 | ILMN_10770 | Homo sapiens | RefSeq | NM_139289.1 | AKAP4 | NM_139289.1 | NM_139289.1 | 8852 | 21493038 | NM_139289.1 | AKAP4 |

# [2] DATA PRE-PROCESSING

# [2] DATA PRE-PROCESSING

1

## Background correction

RMA (& GCRMA)
MAS5 (no log2)
Limma R package neqc(), backgroundCorrect()

Normalization of the intensity values by filtering the data of low intensity (of questionable quality ).
The normalization step is key to reducing volatility so that to adjust data and to remove systematic errors.

2

## Logarithmic transformation of data (Improving Graphic Imaging and Interpretation)

- The variance of the logarithmic intensity values depends less on the absolute values
- Normalization takes place additionally
- Normalizing high asymmetric distributions
- Gives a more real picture of the variance

3

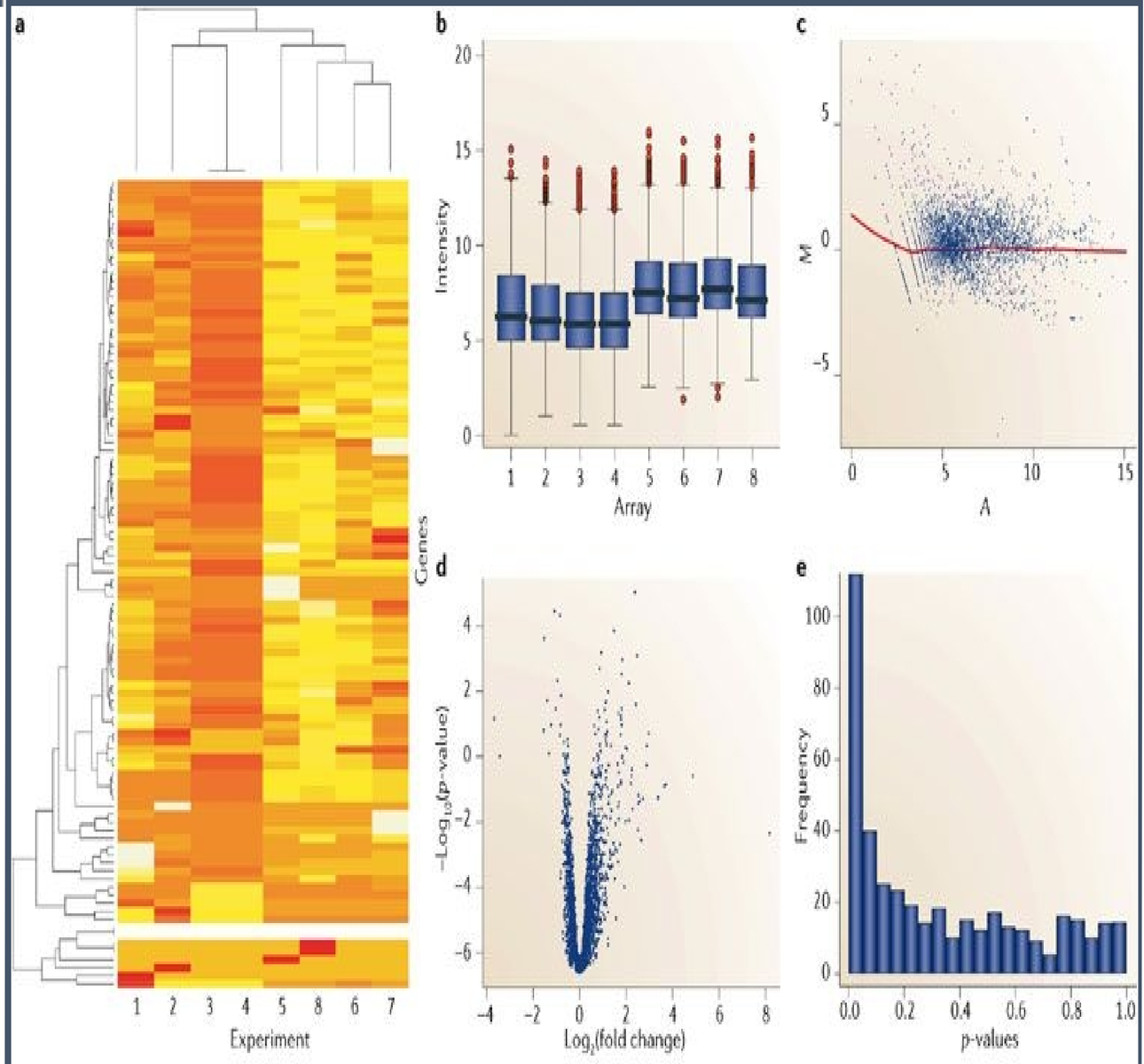
## Normalization

**(Correction of system error of fluorescence intensities)**

The noise must be removed to receive the real signal.

- Minimize systematic errors in expressions of the same tile
- Multiple tile comparison

MA PLOT / BOXPLOT / HISTOGRAM/ HEATMAP / VOLCANO PLOT




https://www.nature.com/articles/nrg1749

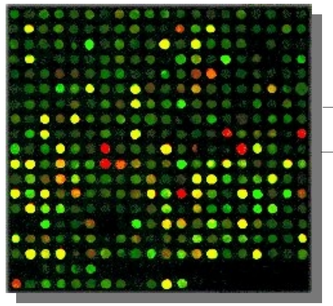

Nature Reviews | Genetics

# Data Manipulation

## In order to proceed to Differential Expression Analysis

**class label**

**Intensities**



| | 0 | 0 | 1 | 1 | 1 | 0 |
|---|---|---|---|---|---|---|
| Probe_ids | GSM85513 | GSM85514 | GSM85515 | GSM85516 | GSM85517 | GSM85518 |
| 1007_s_at | 10.89995638 | 10.74526353 | 10.50083858 | 11.2726252 | 10.60303061 | 10.51979712 |
| 1053_at | 7.468384894 | 7.430974084 | 7.420949239 | 7.436356951 | 7.290826637 | 7.477311326 |
| 117_at | 7.207236391 | 7.26356269 | 7.781946982 | 7.471031924 | 7.485945207 | 7.3282518 |
| 121_at | 8.353033802 | 8.56736164 | 8.332319117 | 8.4445769 | 8.591138364 | 8.408815903 |
| 1255_g_at | 5.574130479 | 5.704594501 | 5.885603827 | 5.885586309 | 5.758336321 | 5.961284775 |
| 1294_at | 8.069341874 | 8.179376232 | 7.927065718 | 8.201891815 | 8.340239995 | 7.861902724 |
| 1316_at | 7.265441773 | 7.108672652 | 7.254406739 | 7.374890809 | 7.328710986 | 7.106973059 |
| 1320_at | 6.790096837 | 6.913872911 | 6.897717413 | 6.868249603 | 7.09995898 | 7.085878237 |
| 1405_i_at | 7.363228044 | 7.906012279 | 6.542664319 | 6.908755827 | 7.912683879 | 6.628585202 |
| 1431_at | 6.21241268 | 6.087831923 | 6.24642086 | 6.299956669 | 6.219591861 | 6.083369456 |
| 1438_at | 8.277196412 | 9.21636985 | 8.274914806 | 8.432758606 | 8.467300931 | 8.236656207 |
| 1487_at | 7.591805822 | 7.999810386 | 7.948621465 | 7.642282983 | 7.829295792 | 7.862653058 |
| 1494_f_at | 6.715724238 | 7.449884809 | 7.127940993 | 9.307620768 | 7.030783092 | 7.223499587 |
| 1552256_a | 8.793423918 | 8.859005322 | 8.759661541 | 8.621644114 | 9.211286117 | 8.736552504 |
| 1552257_a | 8.699622421 | 8.627003057 | 8.618195742 | 8.371665794 | 8.384612261 | 8.145684922 |
| 1552258_a | 7.006313472 | 6.764953779 | 6.66018007 | 7.232681267 | 6.905257119 | 7.141939973 |
| 1552261_a | 6.927855586 | 7.003592668 | 7.048399402 | 6.962205321 | 6.903561396 | 7.26718513 |
| 1552263_a | 6.868319881 | 6.743734981 | 6.247523567 | 6.7457923 | 6.673621769 | 6.760738668 |

**1** From Illumina Platform we keep only the **Gene Symbol** and **ID** columns

**2** We create a class label row based on the series matrix file, filled with **0** for Normal and **1** for Diseased

**3** Normality test (**Boxplot**)

**Data table header descriptions**

**ID_REF**

**VALUE** quantile normalized

**Detection Pval**

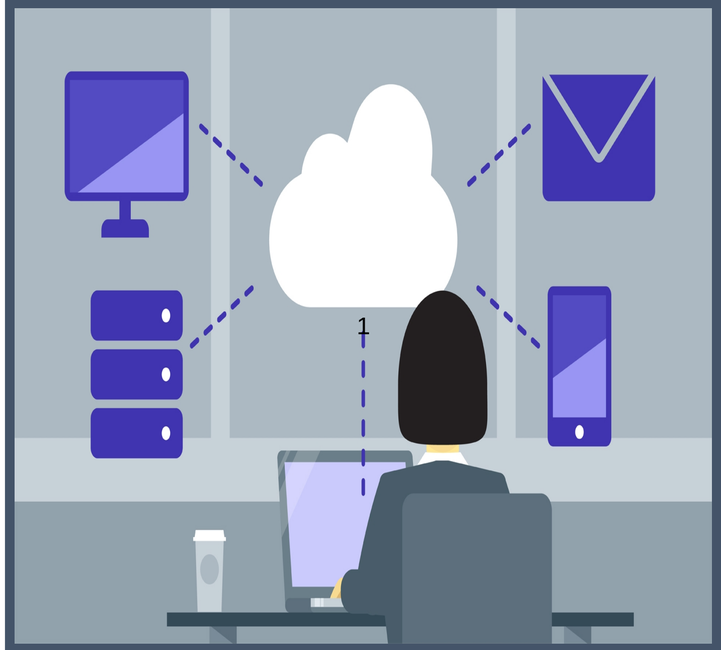
**Data table**

| ID_REF | VALUE |
|---|---|
| ILMN_1802380 | 12.358835 |
| ILMN_1792389 | 10.625872 |
| ILMN_3308818 | 9.074651 |
| ILMN_3242405 | 14.51026 |
| ILMN_2375156 | 10.159377 |
| ILMN_1697642 | 13.586493 |
| ILMN_1788184 | 8.83774 |

# Data Manipulation

In order to proceed to Differential Expression Analysis

| | |
|---|---|
| 200012 x at | RPL21 /// RPL21P28 /// SNORA27 /// SNORD102 |



| | |
|---|---|
| 200593_s_at | HNRNPU |
| 200594_x_at | HNRNPU |
| 200595_s_at | EIF3A |
| 200596_s_at | EIF3A |
| 200597_at | EIF3A |
| 200598_s_at | HSP90B1 |
| 200599_s_at | HSP90B1 |

| | |
|---|---|
| 200012_x_at | RPL21 |
| 200013_at | RPL24 |
| 200014_s_at | HNRNPC |
| 200015_s_at | |
| 200016_x_at | HNRNPA1 |

**4** From the Gene Symbol column of the Illumina Platform, we keep only the **first** name of each symbol

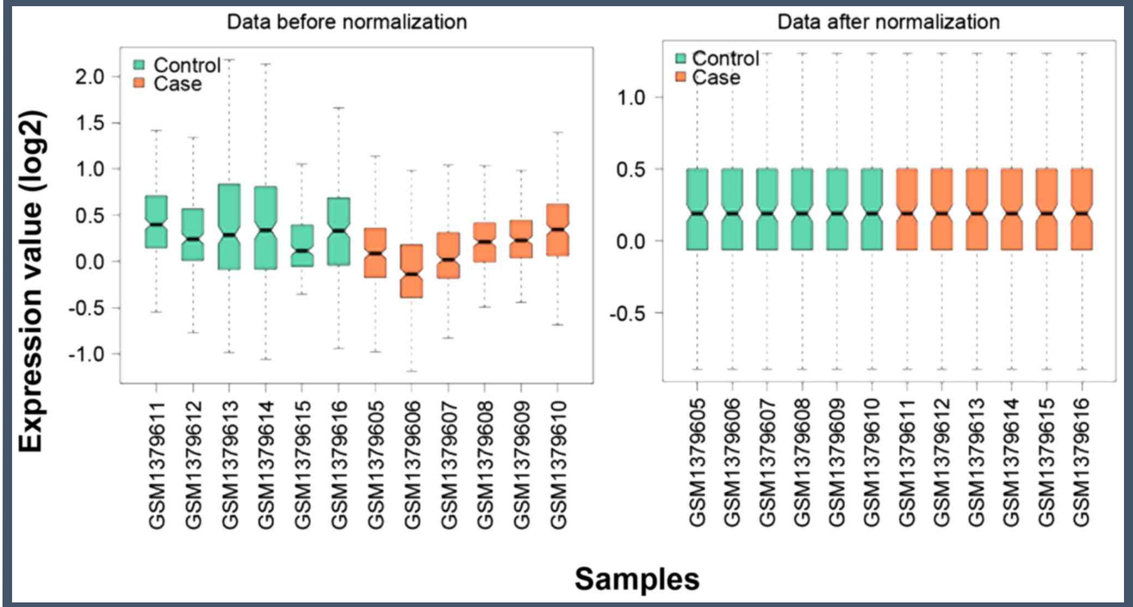
**5** We keep only the **unique** rows

**6** We **vanish** the entries (rows) with empty cells in the Gene Symbol column

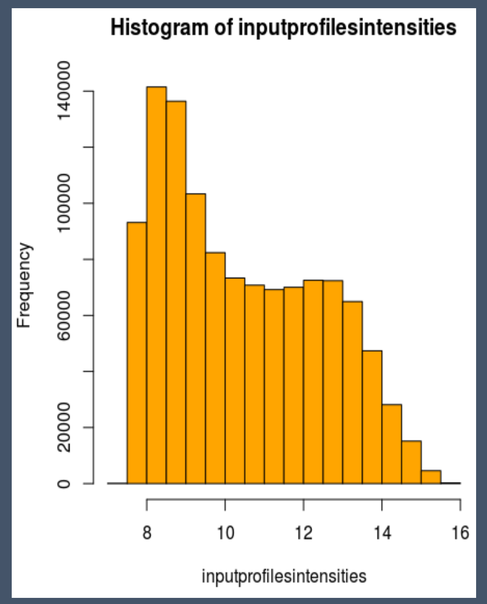
# [3] EXPLORATORY ANALYSIS

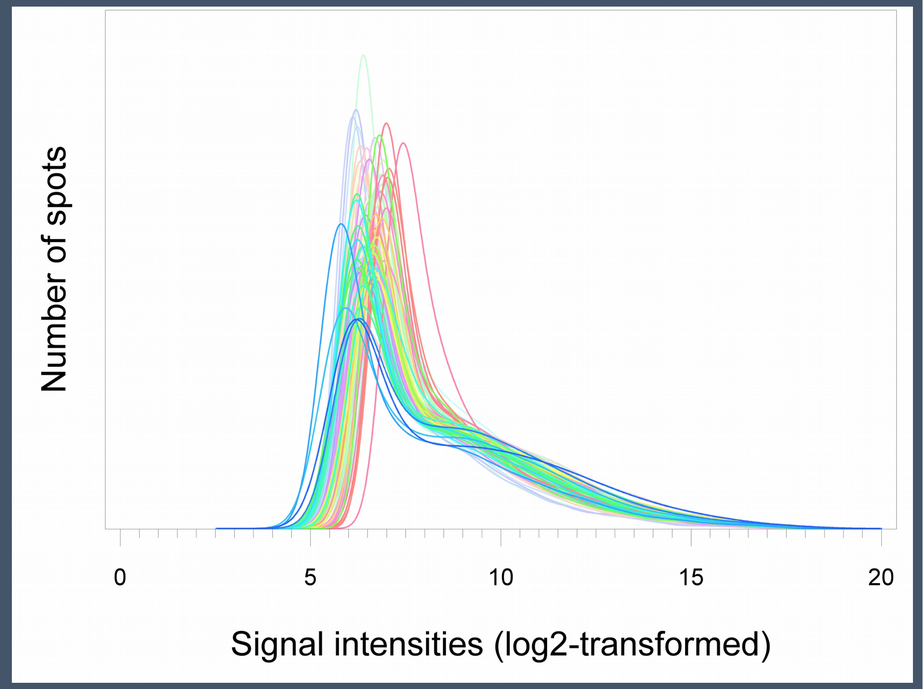
# [3] EXPLORATORY ANALYSIS on our data

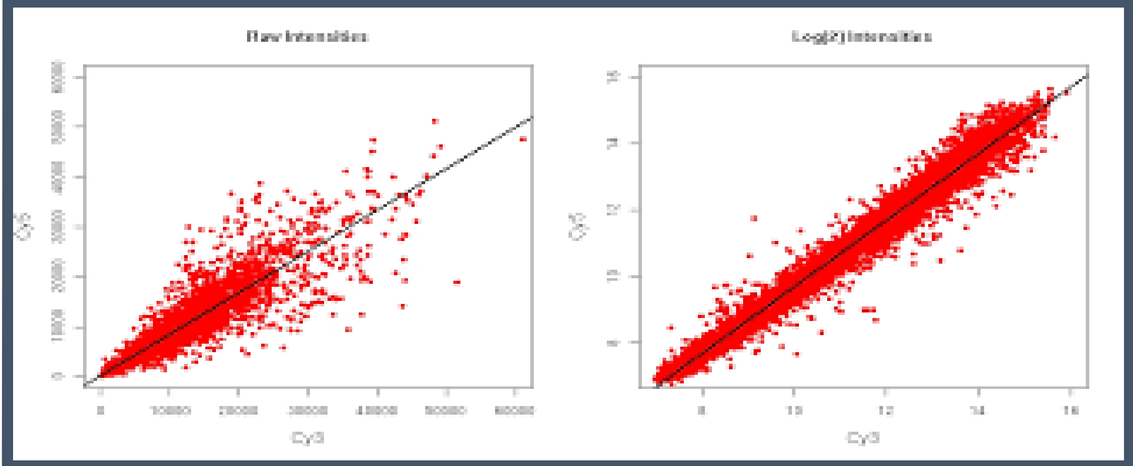
normIntensities<-normalizeQuantiles(Intensities)
Boxplot(Intensities)



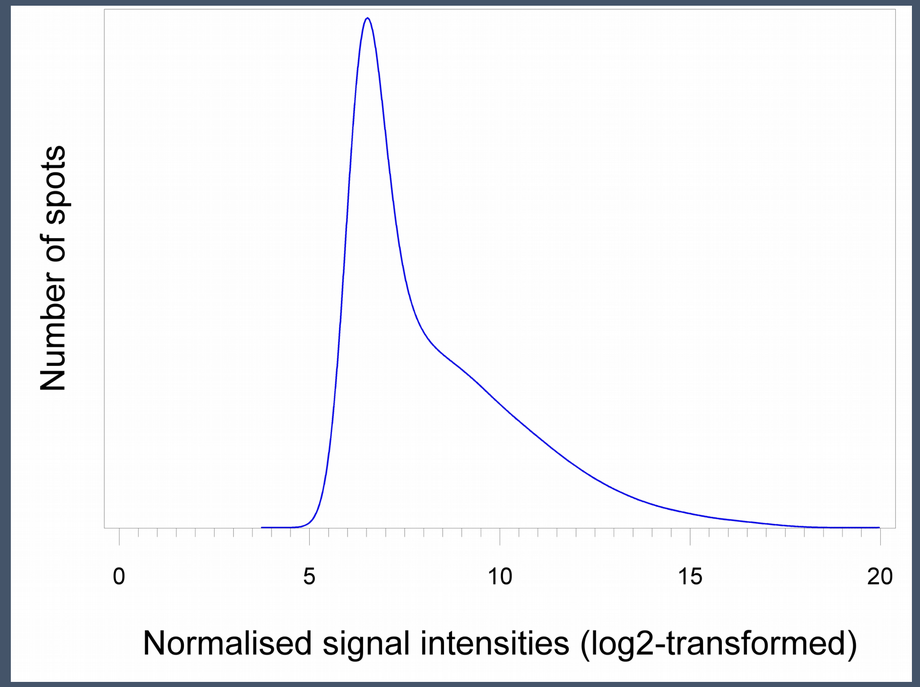

hist(Intensities)



MA plot
Before and after log2 transformation



# [4] DE ANALYSIS-STEPS

**Differential Expression Analysis**

Taking the normalised data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups.To understand the effect of a drug we may ask which genes are **up-regulated** (increased in expression) or **down-regulated** (decreased in expression) between treatment and control groups.

**Statistical Analysis /Control**

Biological phenomenon or random variation in mRNA levels ?

**t-test**

Calculation of statistical t:

**>> t,** the smaller the likelihood that the two average values will be identical

**<< t,** the greater the likelihood that the two average values will be identical

**Statistical Measures**

**P-value**

The lower the p-value, the lower the probability that the two mean of the values will be the same, and therefore the two conditions.

Significant p-value < 0.05 !

**Fold Change**

Measure that describes the amount of change that occurs from an initial to a final state. Is calculated simply as the ratio of the difference between final value and the initial value over the original value.

**Log2 transformation on expression data**

**logFC**

Average(Pathological Expression Values) = A

Average(Normal Expression Values) = B

$$FC = \frac{A}{B} \rightarrow logFC = log\left(\frac{A}{B}\right) = logA - logB$$

The logarithm in the logFC is typically calculated for the base 2. That means one unit of the logFCs translates to a **two-fold** change in expression. The FCs can be calculated from the logFCs as **FC = 2^logFC.**

# DE analysis and output table with statistics

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

LIMMA R package

Functions used :

- model.matrix
- lmFit
- ebayes
- topTable

| ID | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|
| ILMN_1343291 | -7.666201e-02 | 14.571054 | -1.5676111441 | 1.248267e-01 | 0.53911119 | -4.8195564 |
| ILMN_1651209 | -1.410753e-02 | 8.697243 | -0.1562860927 | 8.765913e-01 | 0.97350558 | -5.9481047 |
| ILMN_1651228 | -1.524472e-01 | 13.805876 | -2.1449309794 | 3.806640e-02 | 0.34193062 | -3.8810590 |
| ILMN_1651229 | -2.253579e-03 | 11.986144 | -0.0237858585 | 9.811413e-01 | 0.99567184 | -5.9595301 |
| ILMN_1651235 | 3.458802e-02 | 8.929563 | 0.4302032298 | 6.693501e-01 | 0.91190915 | -5.8713638 |
| ILMN_1651236 | -3.887025e-02 | 8.960563 | -0.2439385567 | 8.085216e-01 | 0.95465784 | -5.9313191 |
| ILMN_1651237 | 2.809551e-01 | 9.066146 | 1.7814996938 | 8.240245e-02 | 0.46004121 | -4.5003741 |
| ILMN_1651238 | 1.102311e-01 | 9.267276 | 0.6689484677 | 5.073598e-01 | 0.84878748 | -5.7467078 |
| ILMN_1651254 | 6.487446e-03 | 13.887543 | 0.1142756747 | 9.095893e-01 | 0.98242598 | -5.9535467 |
| ILMN_1651260 | 1.236105e-01 | 8.365303 | 1.0115687257 | 3.178067e-01 | 0.74041973 | -5.4761819 |
| ILMN_1651262 | 1.844195e-01 | 13.689201 | 1.6612552116 | 1.044587e-01 | 0.50624480 | -4.6841349 |
| ILMN_1651268 | 1.100932e-01 | 9.465302 | 0.7922183627 | 4.328931e-01 | 0.81301728 | -5.6616429 |
| ILMN_1651278 | -9.841123e-02 | 11.183806 | -0.9587149614 | 3.434448e-01 | 0.75837183 | -5.5248097 |
| ILMN_1651282 | -2.758284e-01 | 8.526321 | -1.1105237055 | 2.733853e-01 | 0.70853806 | -5.3785225 |
| ILMN_1651285 | -6.135972e-02 | 10.074949 | -0.3718502307 | 7.119613e-01 | 0.92751915 | -5.8936872 |
| ILMN_1651286 | -1.717950e-01 | 10.361193 | -1.2577241367 | 2.157635e-01 | 0.65366638 | -5.2175608 |

# Filtering and Sorting

platform_two_col

| ID | Gene Symbol |
|---|---|
| 1007_s_at | DDR1 |
| 1053_at | RFC2 |
| 117_at | HSPA6 |
| 121_at | PAX8 |
| 1255_g_at | GUCA1A |
| 1294_at | MIR5193 |
| 1316_at | THRA |
| 1320_at | PTPN21 |
| 1405_i_at | CCL5 |
| 1431_at | CYP2E1 |
| 1438_at | EPHB3 |
| 1487_at | ESRRA |
| 1494_f_at | CYP2A6 |
| 1598_g_at | GAS6 |
| 160020_at | MMP14 |
| 1729_at | TRADD |
| 1773_at | CHURC1-FNTB |
| 177_at | PLD1 |
| 179_at | DTX2P1-UPK3BP1-PMS2P11 |
| 1861_at | BAD |
| 200000_s_at | PRPF8 |
| 200001_at | CAPNS1 |
| 200002_at | RPL35 |
| 200003_s_at | MIR6805 |
| 200004_at | EIF4G2 |
| 200005_at | EIF3D |
| 200012_x_at | RPL21 |
| 200013_at | RPL24 |
| 200014_s_at | HNRNPC |
| 200015_s_at | |
| 200016_x_at | HNRNPA1 |

merge(platform_two_col, toptable, by= “ID”)

RStudio

toptable

| Gene Symbol | ID | logFC | P.Value |
|---|---|---|---|
| | 215812_s_at | 0.497791327 | 8.07E-07 |
| | 210854_x_at | 0.502950145 | 1.36E-06 |
| | 201658_at | -0.945885085 | 2.18E-06 |
| | 219718_at | -0.448654715 | 5.76E-06 |
| | 213843_x_at | 0.471477818 | 6.17E-06 |
| | 221806_s_at | 0.990861236 | 6.70E-06 |
| | 200601_at | 0.792609727 | 6.79E-06 |
| | 204275_at | 0.318036576 | 7.01E-06 |
| | [illegible]6_at | 1.065401691 | 7.70E-06 |
| | 3[illegible]at | 0.625406867 | 8.85E-06 |
| | 40[illegible] | 0.507797564 | 1.01E-05 |
| | 34206_at | 0.358654612 | 1.44E-05 |
| | 214096_s_at | 0.501498297 | 1.85E-05 |
| | 212778_at | 1.131469812 | 2.09E-05 |
| | 212359_s_at | 0.496572194 | 2.20E-05 |
| | 203206_at | 0.354464527 | 2.34E-05 |
| | 213752_at | 0.579734012 | 4.13E-05 |
| | 204328_at | 0.569882655 | 4.20E-05 |
| | 219114_at | -0.441640382 | 4.53E-05 |
| | 202332_at | 0.699775406 | 4.58E-05 |
| | 218425_at | 0.532921121 | 4.97E-05 |
| | 205546_s_at | 0.485561515 | 5.30E-05 |
| | 214797_s_at | 1.008443012 | 5.46E-05 |
| | 221640_s_at | 0.415141115 | 5.85E-05 |
| | 220142_at | 0.668142073 | 5.89E-05 |
| | 206017_at | -0.56262423 | 6.12E-05 |
| | 218714_at | 0.617553339 | 6.51E-05 |
| | 41160_at | 0.534257733 | 6.89E-05 |
| | 564_at | 0.682972103 | 7.63E-05 |

# Filtering and Sorting

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| We keep only the entries with P-value < 0.05 | We sort the entries based on p-value | We keep only the entries with logFC>1 \| logFC<-1 | We sort the entries based on the abs(logFC) bigger --> smaller |

## Top significant DE genes

# TOP 1000 GENES

| Gene Symbols | Probe ids | logFC | abs(logFC) | P.Value |
|---|---|---|---|---|
| FIGF | 206742_at | -5.32513 | **5.32513** | 5.45E-30 |
| COL17A1 | 204636_at | -3.83636 | **3.83636** | 6.64E-26 |
| KCNJ16 | 219564_at | -2.72097 | **2.72097** | 2.01E-25 |
| FXYD1 | 205384_at | -4.98779 | **4.98779** | 2.75E-25 |
| OXTR | 206825_at | -5.04518 | **5.04518** | 3.94E-23 |
| SCARA5 | 235849_at | -6.1398 | **6.1398** | 4.59E-23 |
| SAMD5 | 228653_at | -4.76737 | **4.76737** | 9.00E-23 |
| TNXA | 216333_x_at | -3.11632 | **3.11632** | 1.21E-22 |
| ... | ... | ... | **...** | ... |
| CASP6 | 211464_x_at | 0.2008 | **0.2008** | 0.049858206 |
| ZNF451 | 215012_at | 0.4392 | **0.4392** | 0.049867915 |

# [5] VISUALIZATION

/home/vicky/Desktop/THESIS_FINAL/ss_vs_nash/heatmap_ss_vs_nash.pdf#master-page3

mirnas_pheatmap

## Heatmap

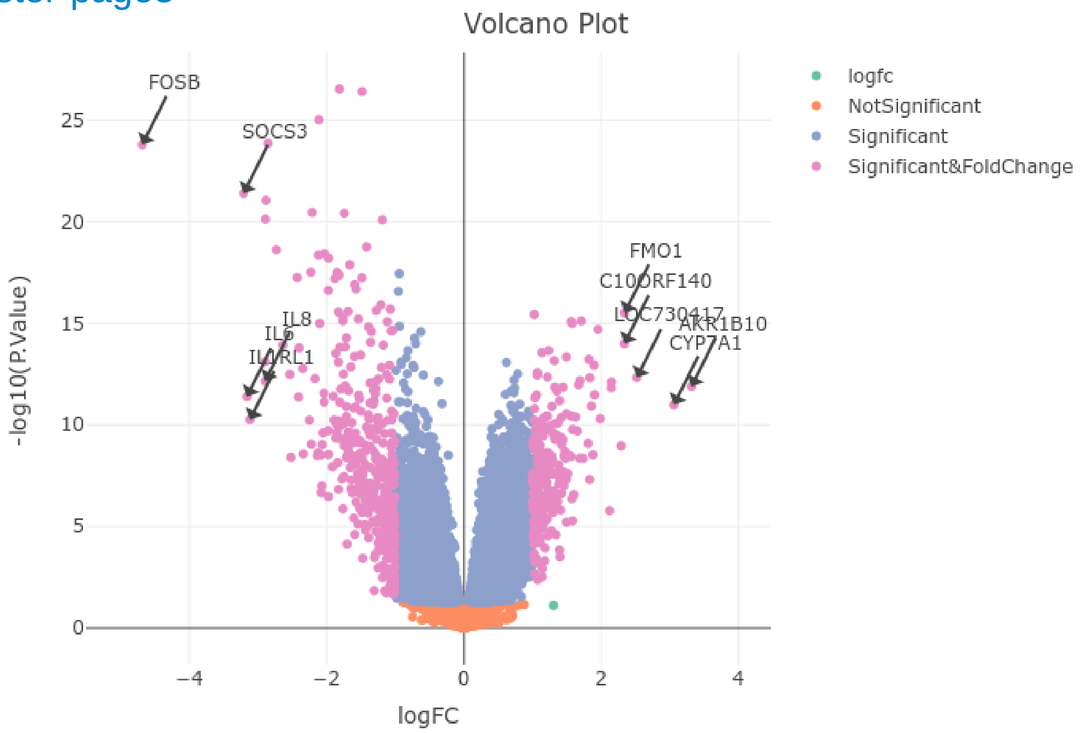
Used in molecular biology to represent the level of expression of many genes across a number of comparable samples.

Heatmap(), pheatmpap()

## Volcano Plots

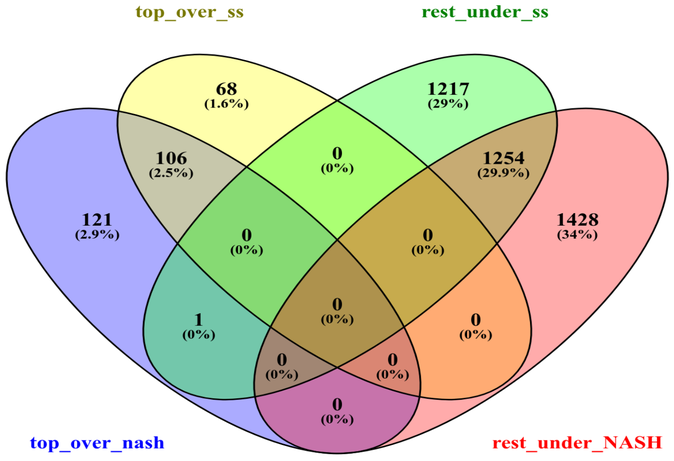
Is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data. It plots significance versus fold-change on the y and x axes, respectively.

ggplot2,plot.ly https://plot.ly/online-chart-maker/




## VENN diagram VENNY, venndiagram()

Shows all possible logical relations between a finite collection of different sets. These diagrams depict elements as points in the plane, and sets as regions inside closed curves.




http://bioinfogp.cnb.csic.es/tools/venny/

# Top DE genes are suggested biomarkers

## Additional steps for the in situ validation of the accuracy of the suggested biomarkers:

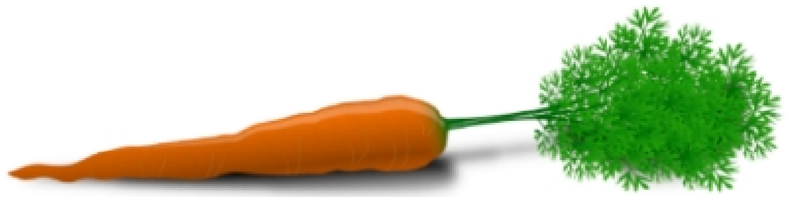
WEKA
The University of Waikato

*Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization*

the caret package

### Caret functionality

- Some preprocessing (cleaning)
- preProcess
- Data splitting
- createDataPartition
- createResample
- createTimeSlices
- Training/testing functions
- train
- predict
- Model comparison
- confusionMatrix



The **caret** package (short for *Classification And REgression Training*) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

http://caret.r-forge.r-project.org/

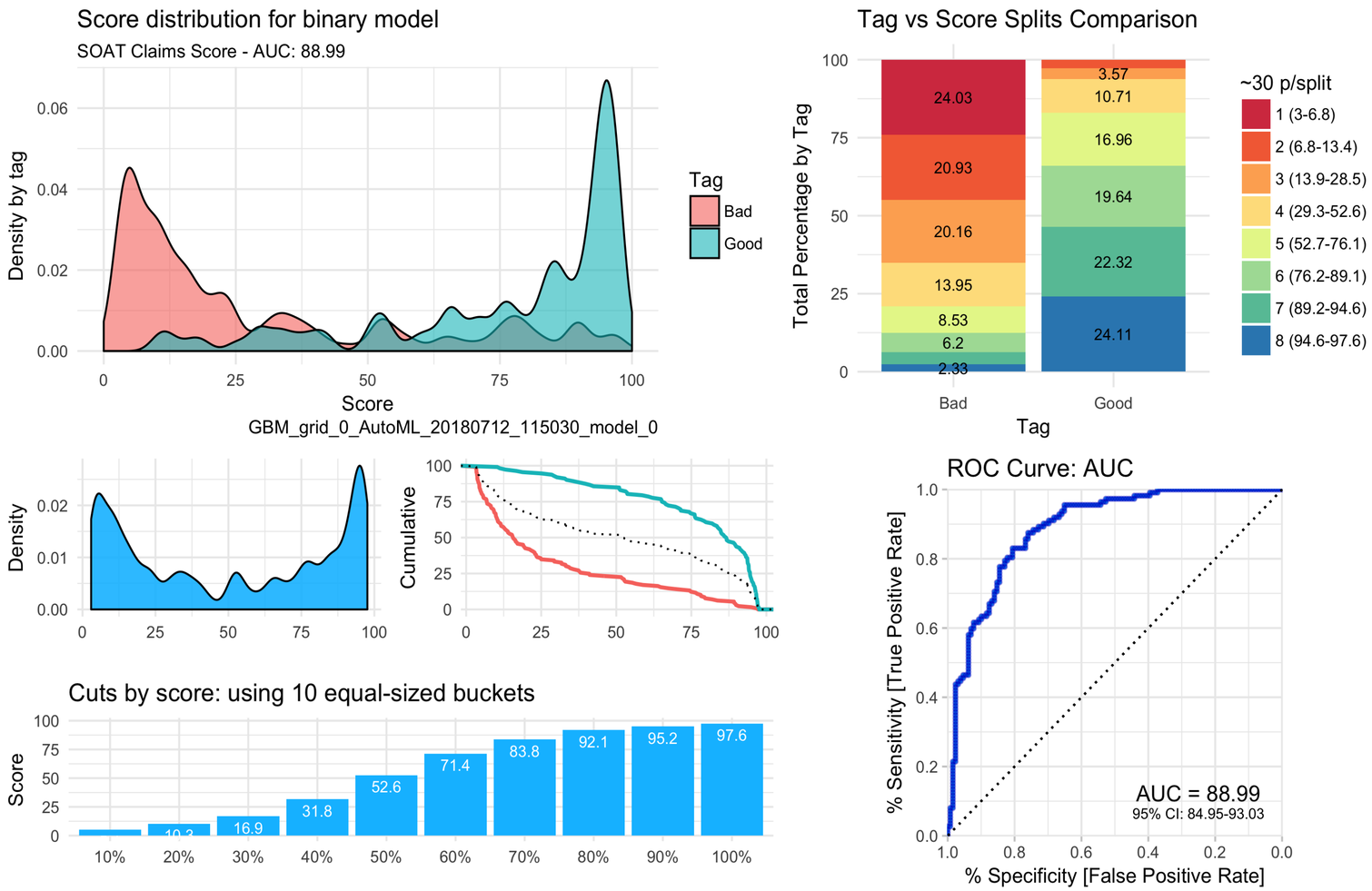
### Machine learning algorithms in R

- Linear discriminant analysis
- Regression
- Naive Bayes
- Support vector machines
- Classification and regression trees
- Random forests
- Boosting
- etc.

# Top DE genes are suggested biomarkers

## Additional steps for the in situ validation of the accuracy of the suggested biomarkers:

### Machine Learning Results in R: one plot to rule them all!

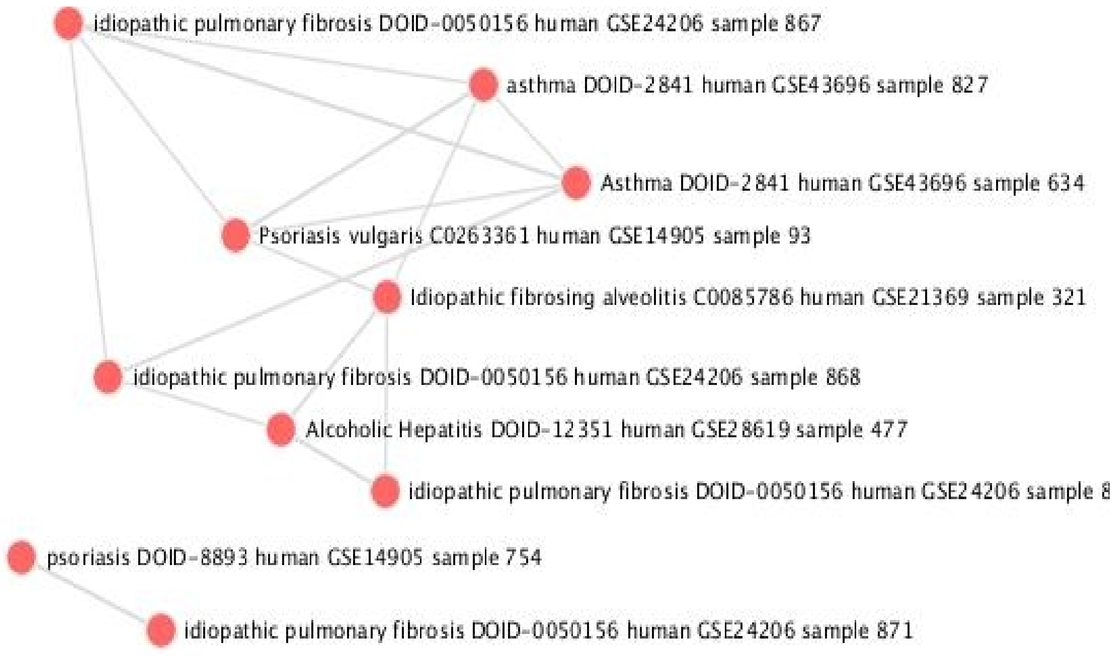



https://datascienceplus.com/machine-learning-results-one-plot-to-rule-them-all/

# [5] Enrichement Analysis

## Enrichment Analysis

Gene set enrichment analysis (GSEA) (also functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes.




## Tools for performing GSEA

1. **Enrichr**
   http://amp.pharm.mssm.edu/Enrichr/
2. **GeneSCF**
   http://genescf.kandurilab.org/
3. **DAVID**
   https://david.ncifcrf.gov/summary.jsp
4. **QuSAGE (R/Bioconductor)**
   http://bioconductor.org/packages/release/bioc/html/qusage.html

DAVID_ Database for Annotation, Visualization, and Integrated Discovery (Laboratory of Human Retrovirology and Immunoinformatics (LHRI); National Institute of Allergies and Infectious Diseases (NIAID); Leidos Biomedical Research, Inc. (LBR).pdf

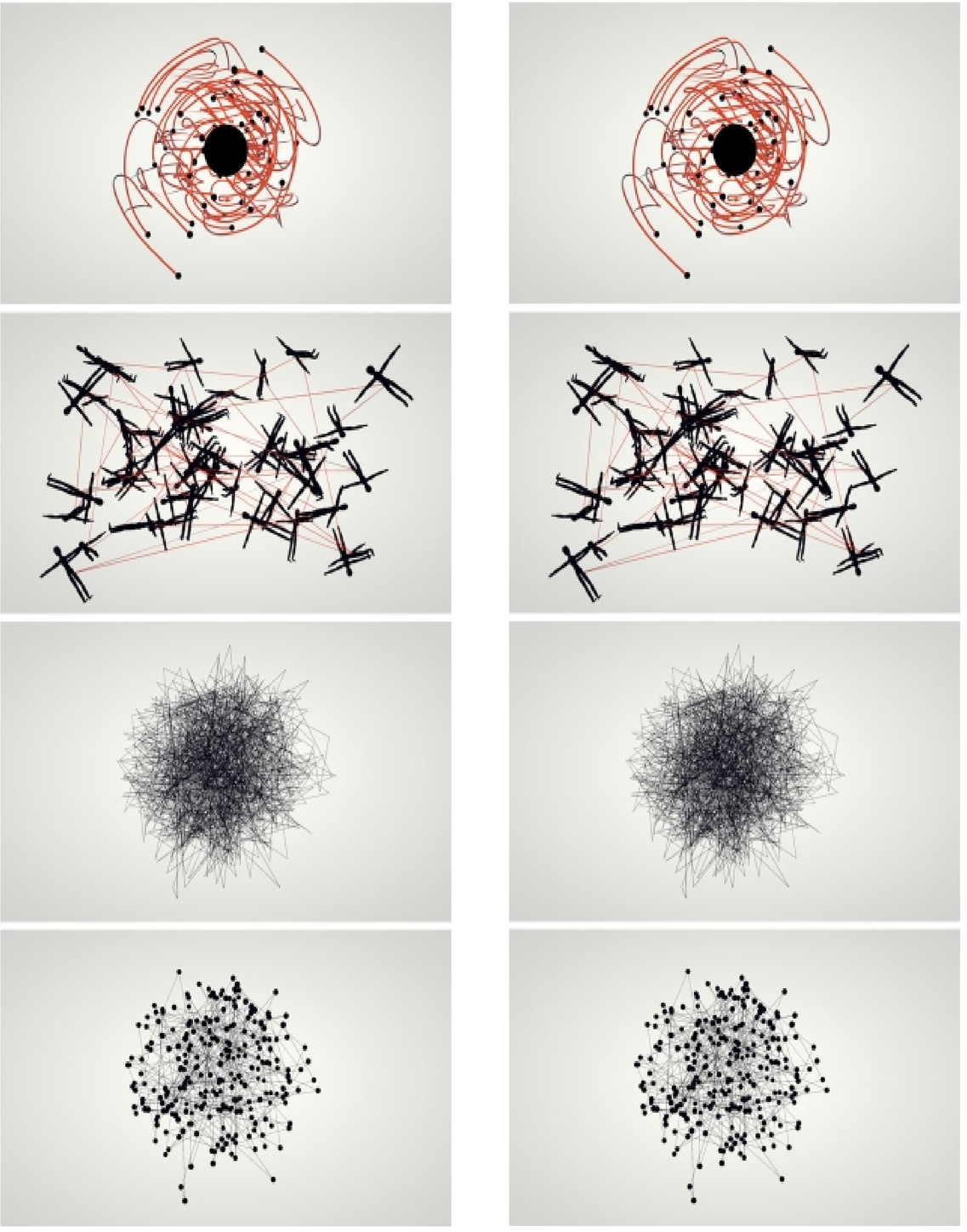
/home/vicky/Desktop/THESIS_FINAL/overview.pdf

# [6] Networks

# [6] Networks

## What is a network ?

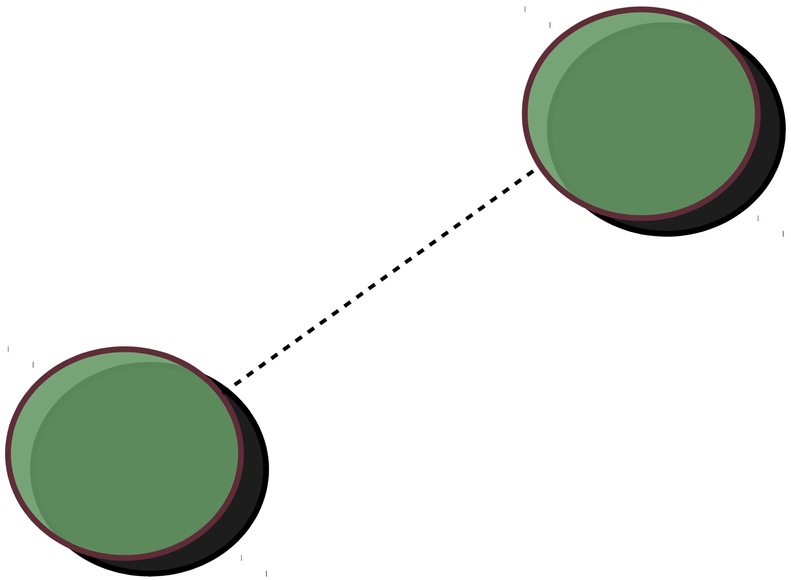
A theoretical structure that describes the relationships between elements that represent it in its form.



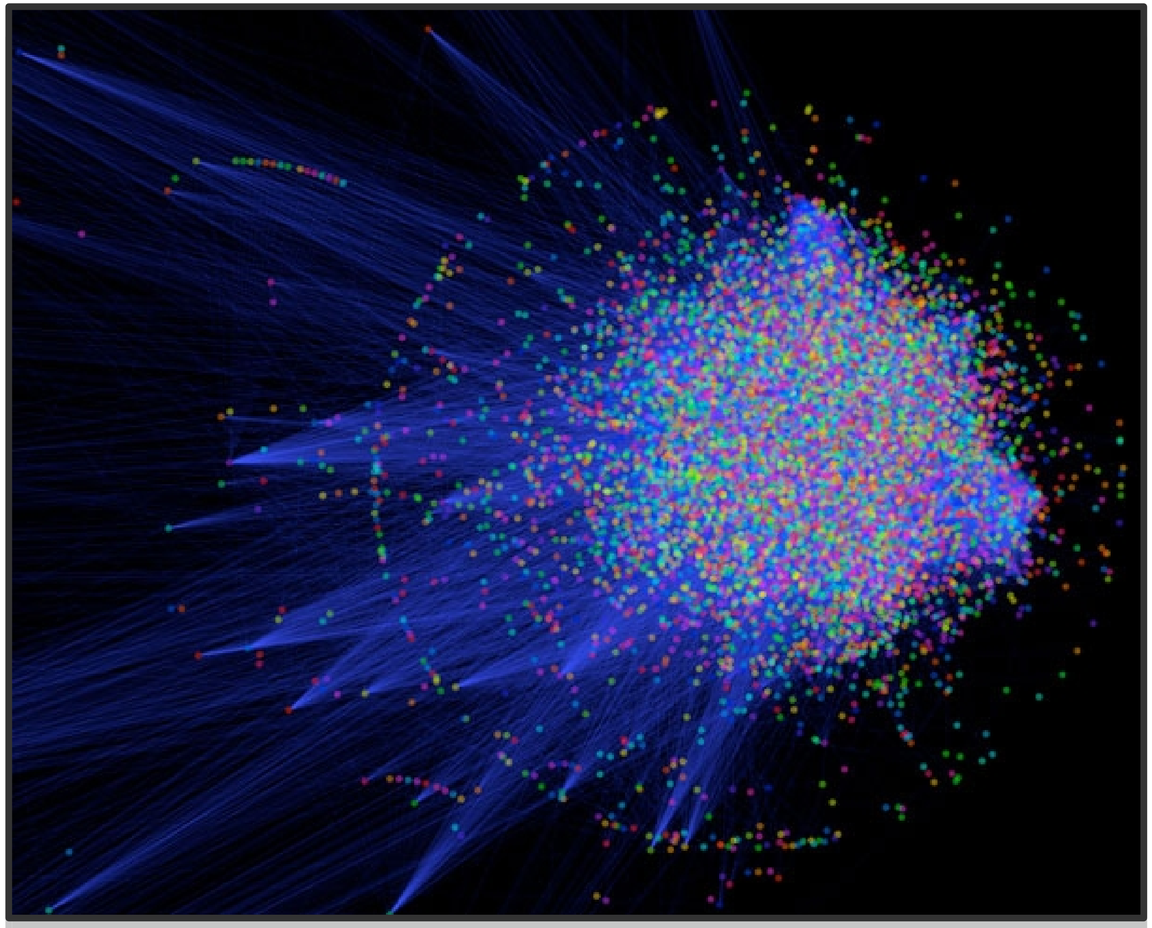
# Networks and Biology

**Biological networks:**
at all levels of study of the life sciences from the most tiny (molecular) to the most macroscopic (ecosystems)

**Genes**
**Proteins**
**Metabolites...**



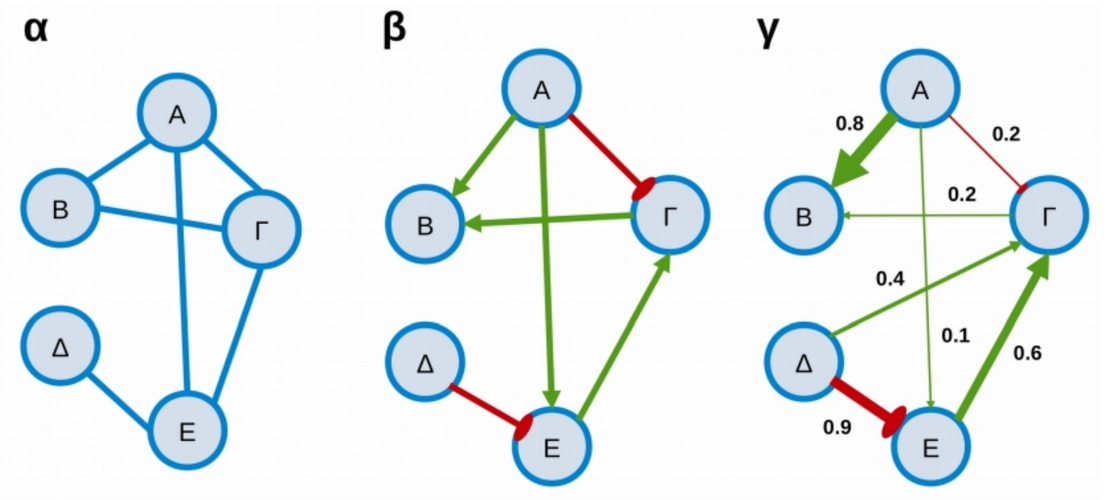
**Physical**
**Biochemical**
**functional**



# Types of biological Networks

## Regulatory Networks :

Regulation of expression between genes




## Metabolic Networks

Nodes --> Enzymes and Metabolites
Edges --> Chemical Reactions
A description of the overall activity of the metabolism

## Signaling/Propagation Networks

Cell signaling processes
Nodes --> proteins
Edges --> Activation reactions that are stages in the transmission of a signal

## Protein Interaction Networks

- All protein-containing biological networks are networks of protein interactions
- Physical Interaction Relationships

Identifying such interactions --> extremely difficult experimental

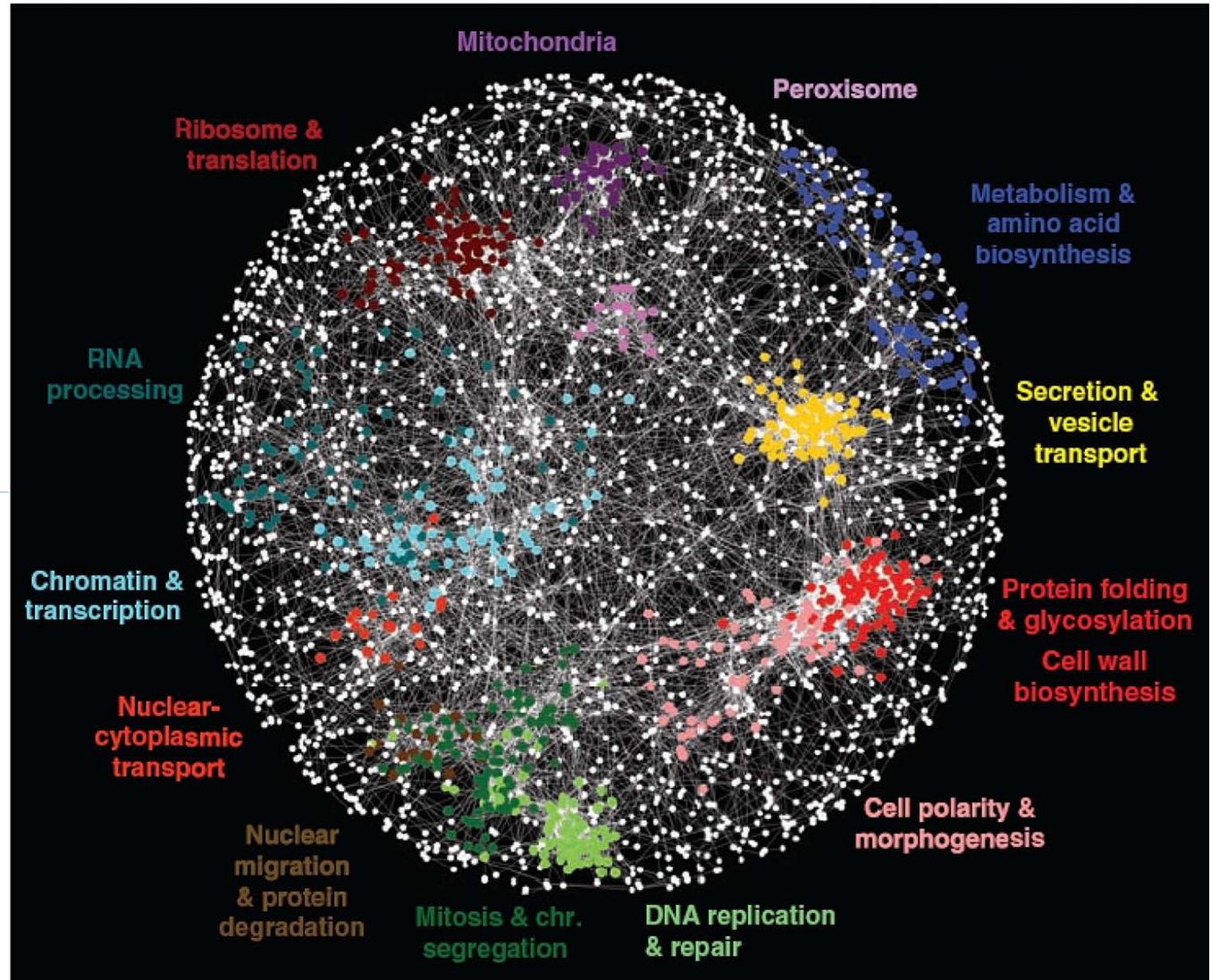
# Molecular Interactions

## Human Interactome > 200.000 interactions

DISEASE - complex interactions disorders
Absence - presence of an interaction

Limited mapping of disrupted molecular interactions

Problem of understanding - investigating diseases

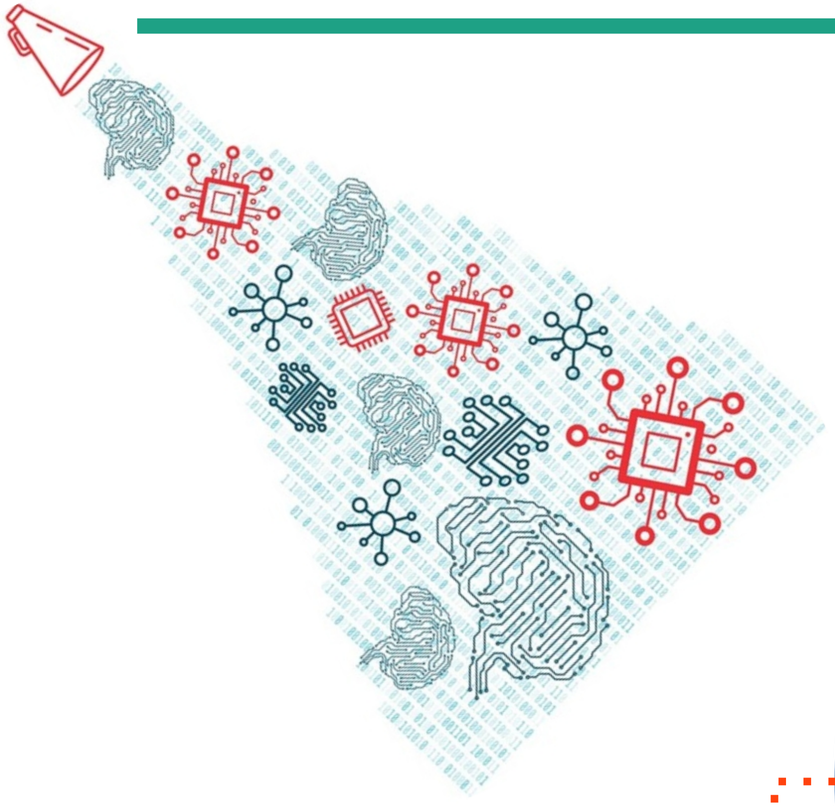


# Databases

- Biomolecular Interaction Network Database(**BIND**)
- Biological General Repository for Interaction Datasets (**BioGRID**)
- Human Protein Reference Database (**HPRD**)
- Molecular Interaction Database (**IntAct**)
- Molecular Interactions Database (**MINT**)

# [7] Co-expression

# Network Inference Methods



need to understand molecular interactions - diseases

**Systems Biology - Development of Statistical Methods of Construction of Co-Expression Networks - Network Relationship Based on Molecular Expression Profiles - Gene-Gene Discovery - Phenotype**

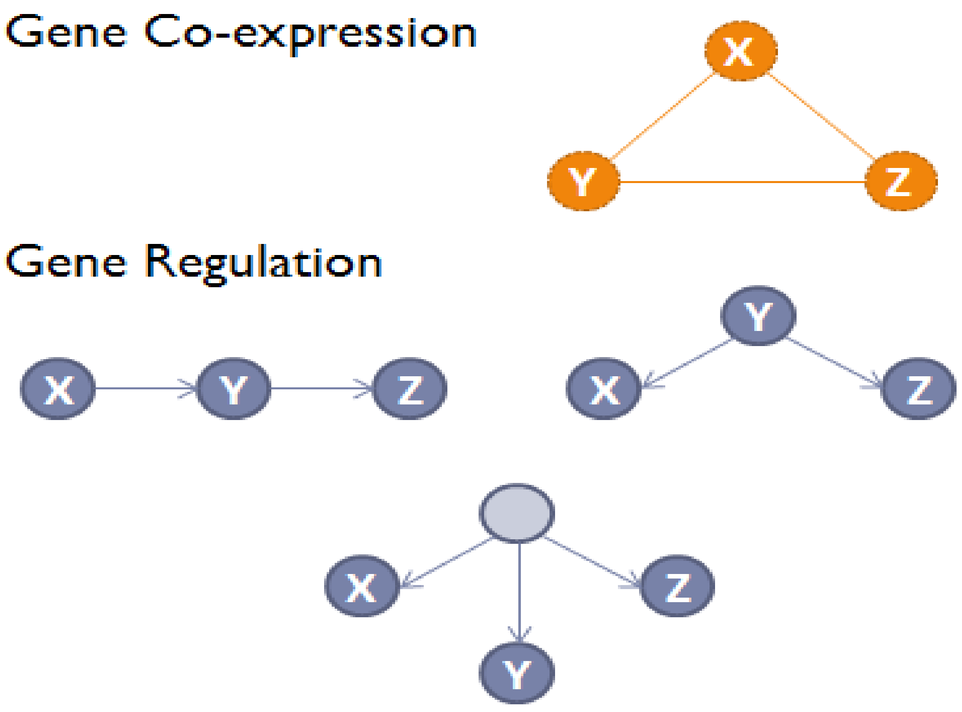
Absence of molecular interactions mapping

Rapid accumulation of molecular expression profiles

# [7] Co-expression

**Use R packages to move from the level of expression to the level of coexpression**

**Gene co-expression network (GCN)**
Is an undirected graph.
**Nodes** --> genes
**Edge** --> a significant co-expression relationship between a pair of genes
**Construction**
looking for pairs of genes which show a similar expression pattern across samples.
**Biological interest**
Co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex.




Constructed using data sets from high-throughput gene expression profiling technologies such as Microarrays or RNA-Seq

The direction and type of relationships are not defined in gene coexpression networks

# Co-expression network construction

**Input**

- Gene expression data [Intensities file]

**nxm matrix where**

**n -->** the number of genes we want to test

**m-->** the number of samples

| GeneSymbols | GSM506037 | GSM506039 | GSM506040 | GSM506041 | GSM506042 | GSM506043 | GSM506044 | GSM506045 | GSM506046 |
|---|---|---|---|---|---|---|---|---|---|
| SYT1 | 10.33908 | 5.343771 | 5.582321 | 5.268273 | 5.225692 | 6.249693 | 5.426984 | 6.364965 | 7.592141 |
| VSNL1 | 11.47426 | 7.370188 | 8.336997 | 6.543741 | 5.69627 | 10.65557 | 6.295036 | 6.405679 | 7.94896 |
| OXR1 | 8.313855 | 5.417594 | 5.497601 | 5.252535 | 4.946619 | 6.36206 | 5.405048 | 6.352633 | 6.768559 |
| ENC1 | 11.74219 | 7.562007 | 8.601264 | 8.365366 | 6.497138 | 9.249153 | 7.266273 | 8.545738 | 9.363238 |
| PRKAR1A | 11.29573 | 8.48222 | 8.733762 | 5.901087 | 7.204596 | 10.23476 | 10.423 | 9.959939 | 10.15759 |
| TCF4 | 9.083672 | 6.547215 | 7.137525 | 8.105058 | 6.327345 | 6.793546 | 8.144326 | 8.259708 | 8.976389 |
| SNAP25 | 11.90147 | 8.774288 | 9.846148 | 8.22824 | 7.821966 | 11.28837 | 8.404737 | 8.243678 | 9.227068 |
| RFC5 | 9.445796 | 7.808323 | 8.229622 | 7.049702 | 6.752701 | 7.960366 | 8.12147 | 8.726924 | 8.842229 |
| TAC1 | 8.264978 | 5.711597 | 6.566486 | 5.552792 | 5.639768 | 6.832999 | 5.955256 | 5.56833 | 6.624945 |
| TTC3 | 10.64062 | 7.132298 | 7.552528 | 7.154346 | 6.25438 | 7.668864 | 9.20298 | 10.06523 | 10.11041 |
| LPPR4 | 10.82668 | 7.562777 | 9.032742 | 7.015502 | 7.257954 | 9.142533 | 7.095432 | 7.889068 | 8.764746 |
| PRKACB | 9.939425 | 8.105348 | 8.48073 | 7.785791 | 5.832049 | 8.740435 | 8.258381 | 8.607728 | 9.566483 |
| PDP1 | 10.18662 | 7.332411 | 7.367985 | 7.18908 | 6.550693 | 7.76814 | 7.482113 | 8.67154 | 9.317524 |
| STMN2 | 11.76233 | 9.537437 | 10.67859 | 8.000459 | 6.469519 | 11.91599 | 7.806339 | 8.730346 | 10.10862 |
| PSD3 | 11.14672 | 7.544622 | 7.410266 | 7.47527 | 6.449869 | 8.721755 | 8.306248 | 9.269829 | 10.00961 |
| PREPL | 10.11161 | 7.416642 | 8.37225 | 8.116906 | 7.392681 | 9.423034 | 7.537574 | 8.042512 | 8.792299 |
| YWHAB | 11.1723 | 8.040752 | 8.98805 | 6.870547 | 6.707229 | 10.23989 | 9.857018 | 9.75049 | 10.00768 |
| SNX10 | 9.686013 | 7.586221 | 8.467611 | 6.598464 | 6.530436 | 8.797027 | 6.638717 | 7.276467 | 8.340045 |

# Co-expression network construction

**Degree of similarity** (coexpression measure)

- It is calculated among the pairs of genes
- Create a new table --> how similar the expression levels of 2 genes are alikes

| | $G_1$ | $G_2$ | $G_3$ | $G_4$ | $G_5$ | $G_6$ | $G_7$ | $G_8$ | $G_9$ | $G_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $G_1$ | 1.00 | 0.23 | 0.61 | 0.71 | 0.03 | 0.35 | **0.86** | **1.00** | **0.97** | 0.37 |
| $G_2$ | 0.23 | 1.00 | 0.63 | 0.52 | **0.98** | **0.99** | 0.29 | 0.30 | 0.46 | **0.99** |
| $G_3$ | 0.61 | 0.63 | 1.00 | **0.99** | 0.77 | 0.53 | **0.93** | 0.56 | 0.41 | 0.51 |
| $G_4$ | 0.71 | 0.52 | **0.99** | 1.00 | 0.69 | 0.41 | **0.97** | 0.66 | 0.52 | 0.40 |
| $G_5$ | 0.03 | **0.98** | 0.77 | 0.69 | 1.00 | **0.95** | 0.48 | 0.09 | 0.27 | **0.94** |
| $G_6$ | 0.35 | **0.99** | 0.53 | 0.41 | **0.95** | 1.00 | 0.17 | 0.41 | 0.57 | **1.00** |
| $G_7$ | 0.86 | 0.29 | **0.93** | **0.97** | 0.48 | 0.17 | 1.00 | **0.83** | 0.72 | 0.16 |
| $G_8$ | **1.00** | 0.30 | 0.56 | 0.66 | 0.09 | 0.41 | 0.83 | 1.00 | **0.98** | 0.42 |
| $G_9$ | **0.97** | 0.46 | 0.41 | 0.52 | 0.27 | 0.57 | 0.72 | **0.98** | 1.00 | 0.58 |
| $G_{10}$ | 0.37 | **0.99** | 0.51 | 0.40 | **0.94** | **1.00** | 0.16 | 0.42 | 0.58 | 1.00 |

# Co-expression network construction

## Co-expression measures

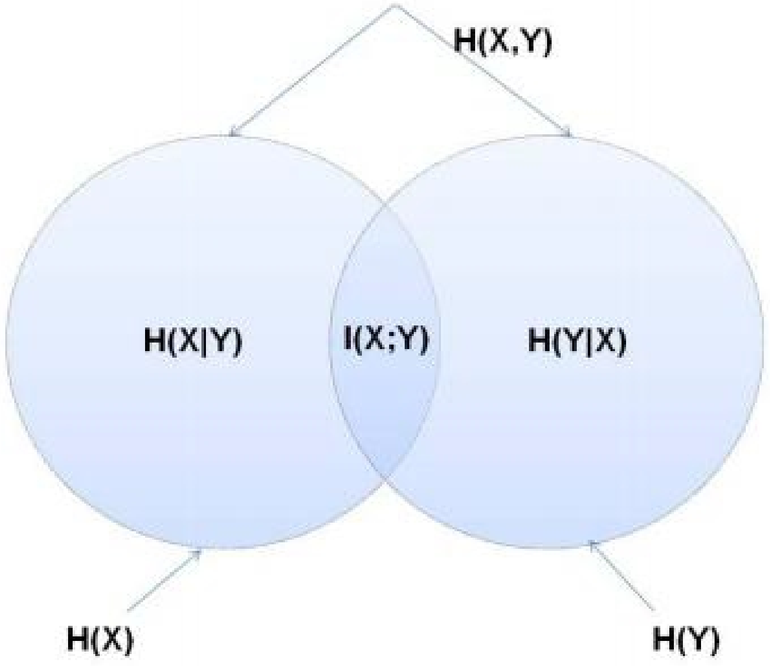
| Correlation | Mutual Information | Tree Based |
| --- | --- | --- |
| Pearson Correlation | ARACNE | Genie 3 |
| Spearman Correlation | CLR | |
| Partial Correlation | MRNET | |
| | C3NET | |

# Mutual information

**Mutual information**

- The information that is shared between two variables
- How much the uncertainty decreases taking into account the expression levels of a gene when we know the expression levels of another gene

**Joint Entropy**



# Co-expression matrix

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Parmigene / clr algorithm
Igraph / R package

Functions used :

- Knmi.all
- clr
- Graph.adjacency
- Get.edgelist

Context Likelihood Or Relatedness Network

**CLR** algorithm is an extension of relevance network. Instead of considering the mutual information I(Xi;Xj) between features Xi and Xj, it takes into account the score sqrt(zi2+zj2), where

$$z_i = max\{0, \frac{I(X_i; X_j) - \mu_i}{\sigma_i}\}$$

and mean(Xi) and sd(Xi) are, respectively, the mean and the standard deviation of the empirical distribution of the mutual information values I(Xi,Xk), k=1,...,n

We used iGraph package in order to switch from the co-expression matrix to the final edge list

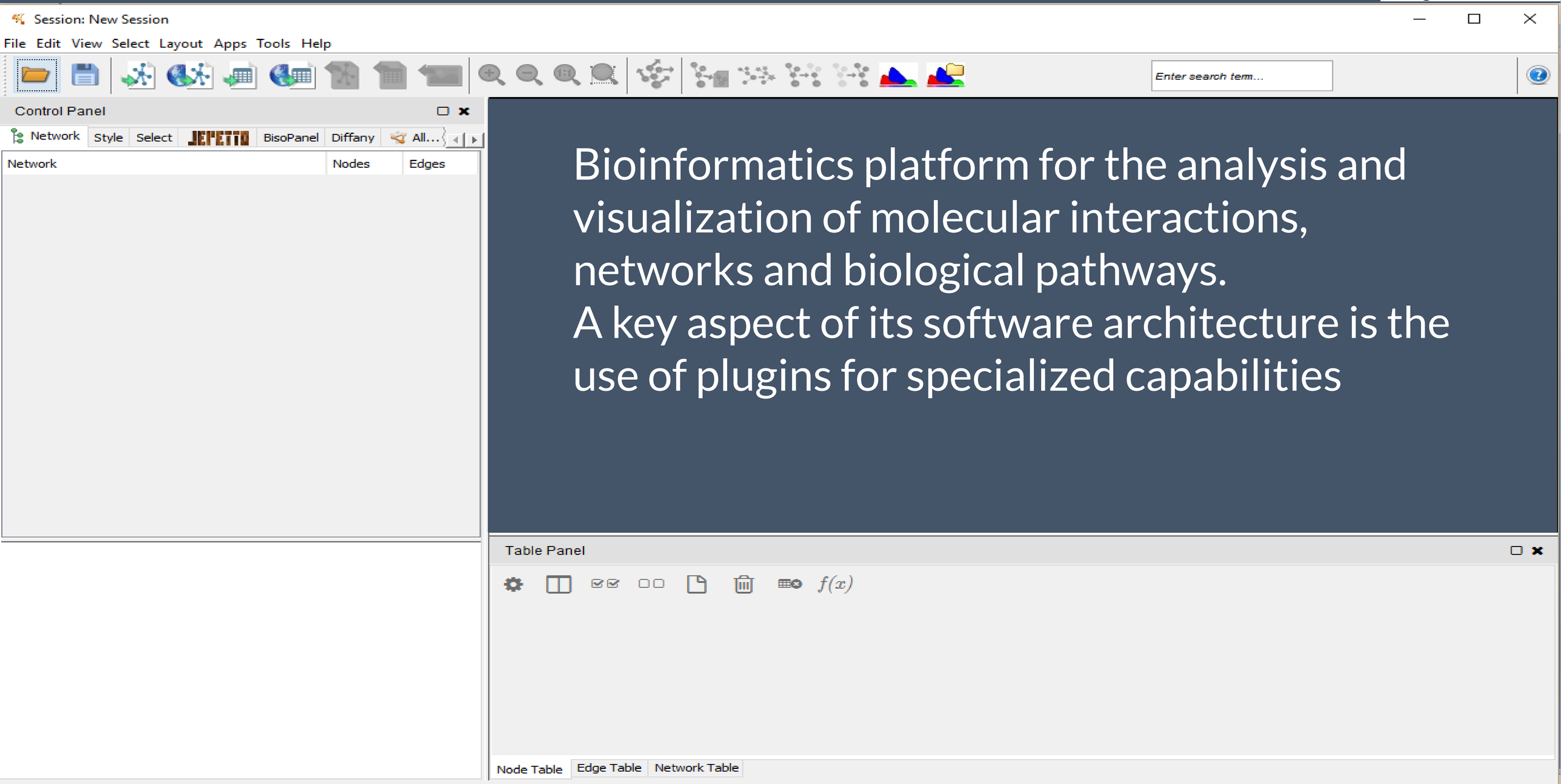
# Adjacency matrix

| | FOSB | IL8 | IL1RL1 | SOCS3 | IL6 | FOS | CYP7A1 |
|---|---|---|---|---|---|---|---|
| FOSB | 0.000000e+00 | 5.127012e+00 | 1.731682e+00 | 2.333716e+00 | 1.357592e+00 | 1.211002e+01 | 0.000000e+00 |
| IL8 | 5.127012e+00 | 0.000000e+00 | 3.493319e-01 | 4.462795e+00 | 8.702143e+00 | 7.379213e+00 | 0.000000e+00 |
| IL1RL1 | 1.731682e+00 | 3.493319e-01 | 0.000000e+00 | 1.383660e+00 | 5.539228e-01 | 2.210812e-01 | 1.498286e+00 |
| SOCS3 | 2.333716e+00 | 4.462795e+00 | 1.383660e+00 | 0.000000e+00 | 3.414252e-01 | 3.446351e+00 | 1.615355e-01 |
| IL6 | 1.357592e+00 | 8.702143e+00 | 5.539228e-01 | 3.414252e-01 | 0.000000e+00 | 1.967844e+00 | 0.000000e+00 |
| FOS | 1.211002e+01 | 7.379213e+00 | 2.210812e-01 | 3.446351e+00 | 1.967844e+00 | 0.000000e+00 | 6.237162e-01 |
| CYP7A1 | 0.000000e+00 | 0.000000e+00 | 1.498286e+00 | 1.615355e-01 | 0.000000e+00 | 6.237162e-01 | 0.000000e+00 |

**edgelist -->** /home/vicky/Desktop/THESIS_FINAL/ss_vs_nash/EDGE_LIST.html

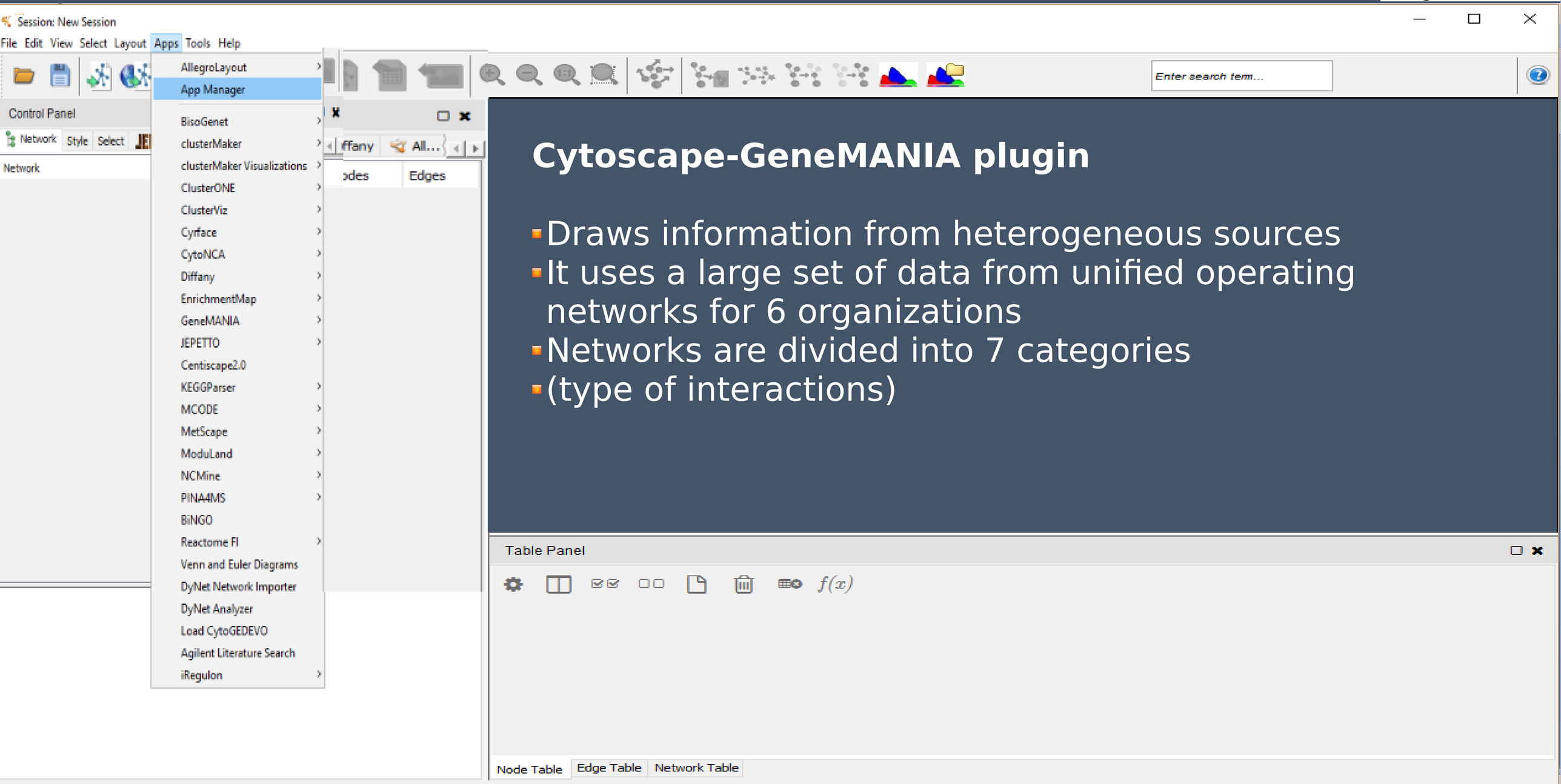
# [8] CYTOSCAPE

# [8] CYTOSCAPE




Bioinformatics platform for the analysis and visualization of molecular interactions, networks and biological pathways.
A key aspect of its software architecture is the use of plugins for specialized capabilities

# [8] CYTOSCAPE




## Cytoscape-GeneMANIA plugin

- Draws information from heterogeneous sources
- It uses a large set of data from unified operating networks for 6 organizations
- Networks are divided into 7 categories
- (type of interactions)

# Types of Interactions BioGRID

**Genetic Interaction:**
Two genes are operably linked if the effects of disruption of a gene are modified by the disruption of another gene (**BioGRID**)

**Co-localization:** Two genes are linked if they are expressed in the same tissue or their products are in the same cellular region.

**Predicted:** Two genes are linked if their products interact with another organism - (**bibliography**)

**Shared protein domains:** Two gene products are linked if they have a similar structure - (**InterPro, SMAR and Pfam**)

**Pathways:** Two genes are linked if they are on the same path.( **Reactome, BioCyc and Pathway Commons**).

**Co-Expression:**
Two genes are linked if their expression levels are similar in a gene expression study. Most of these data - (**Gene Expression Omnibus (GEO) and the corresponding publications**).

**Physical Interaction:**
Two gene products bind if they were found to interact in a protein-protein interaction study. Data - (**BioGRID and Pathway Commons**)

# CYTOSCAPE APP STORE

apps.cytoscape.org

Εφαρμογές | MalaCards - human d | Sci-Hub: removing ba | Program: DNA Microa | Venny | Babelomics 4.3 | MammaPrint Test | Biological Functions c | Remove Duplicate Lin

Cytoscape App Store | Submit an App | Search the App Store | Sign In

All Apps

## Categories

- collections
- data visualization
- network generation
- graph analysis
- online data import
- network analysis
- integrated analysis
- clustering
- utility
- enrichment analysis
- data integration
- systems biology
- layout
- ontology analysis
- visualization
- pathway database
- network comparison
- local data import
- import
- interaction database

more »

## Newest Releases

Get Started with the App Store »

**CytoCopteR** (3.0+)
A Cytoscape plug-in for training logic models

**ANIMO** (3.0+)
ANIMO (Analysis of Networks with Interactive MOdeling) lets you

**GTA** (3.0+)
Module (cluster) detection in PPI network based on gene

**CyNetSVM** (3.0+)
A Cytoscape App for Cancer Biomarker Identification Using

**PTMOracle** (3.0+)
Co-visualisation and co-analysis of PTM and PPI data

**TiCoNE** (3.0+)
Time Course Network Enricher is an interactive clustering method

more newest releases »

## Top Downloaded Apps

**ClueGO** (3.0+)
Creates and visualizes a functionally grouped network of

**BiNGO** (3.0+)
Calculates overrepresented GO terms in the network and display

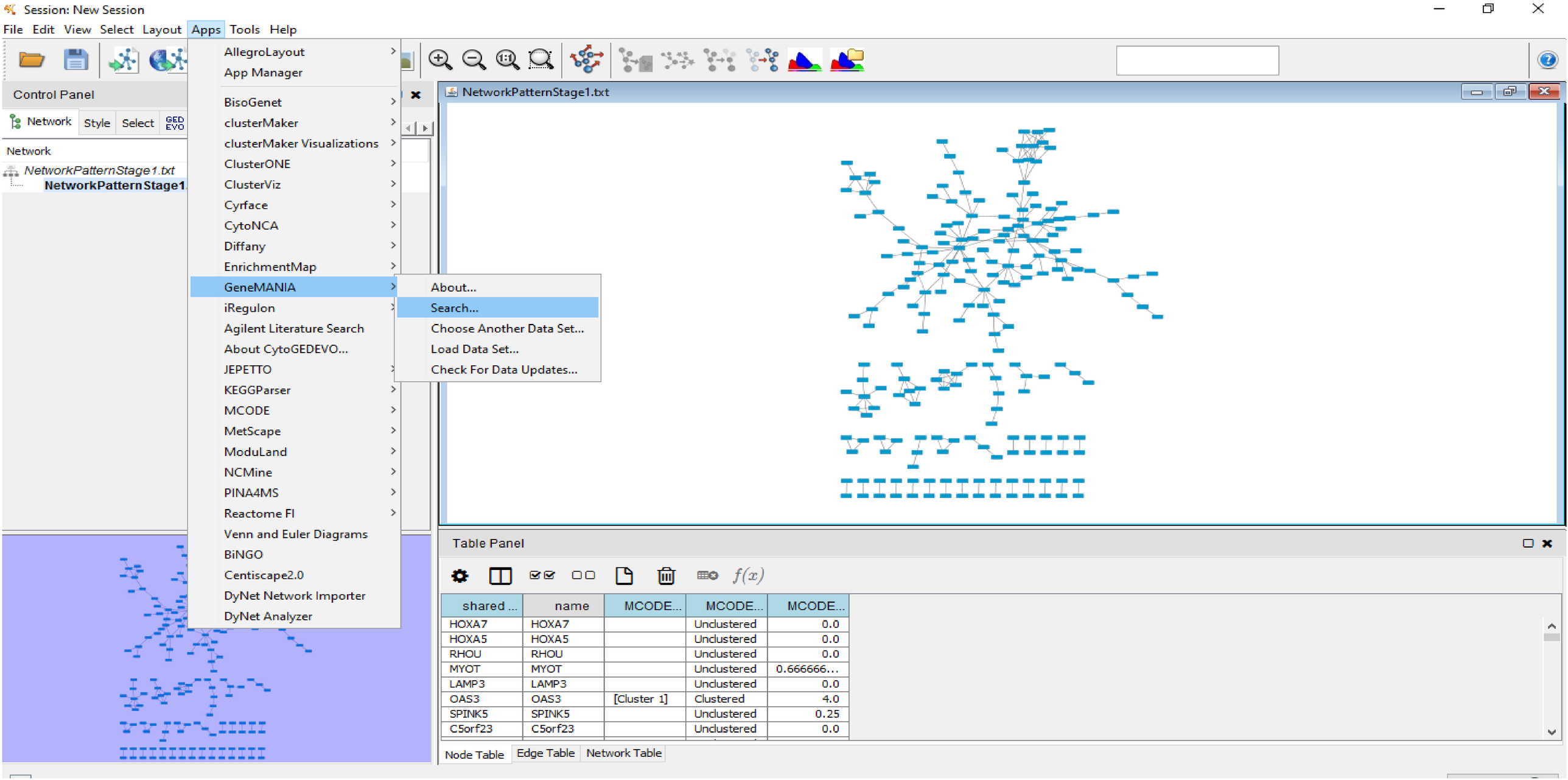
**GeneMANIA** (3.0+)
Imports interaction networks from public databases from a list of

**CluePedia** (3.0+)
CluePedia: A ClueGO plugin for pathway insights using integrated

**AgilentLiteratureSearch** (3.0+)
Mines scientific literature to find publications related to search

**MCODE** (3.0+)
Clusters a given network based on topology to find densely

# CYTOSCAPE APP STORE



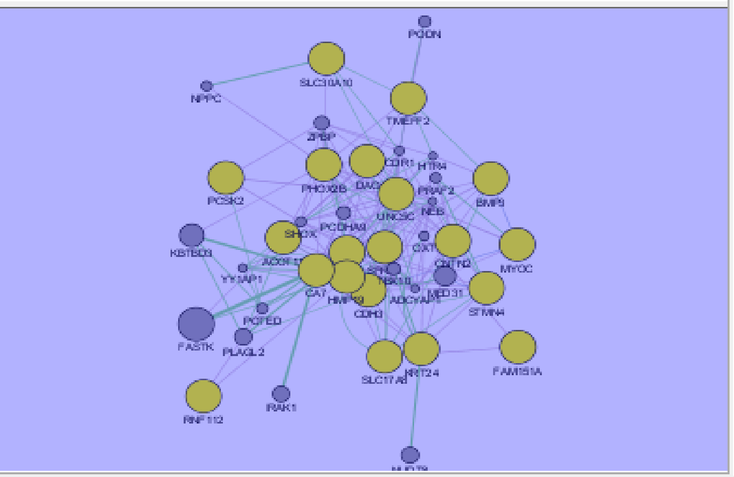
File Edit View Select Layout Apps Tools Help

\10 TMEFF2 USP2 HMP19

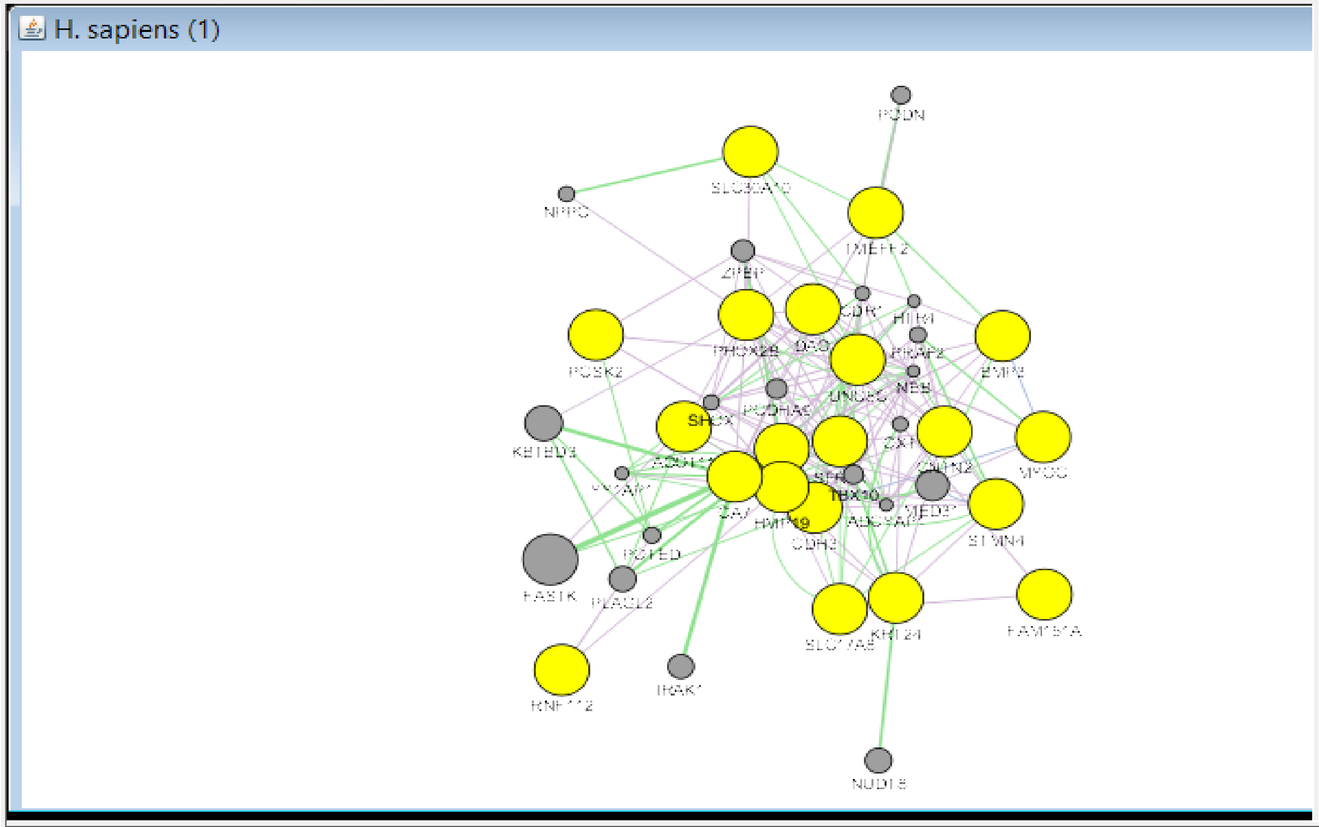
Control Panel

Network | VizMapper | Filters | JEPETTO | Dynamic Network | jA...

| Network | Nodes | Edges |
|---|---|---|
| *H. sapiens (1)* | | |
| **H. sapiens (1)** | 40(20) | 163(0) |



H. sapiens (1)



Results Panel

Node Details | JEPETTO Enrichment | JEPETTO

Organism: H. sapiens

Networks | Genes | Functions

Sort by: name, per cent weight
Expand: all, top-level, none
Enable: all, none

- [x] **Co-expression**
- [x] **Genetic interactions**
- [x] **Co-localization**

Export results...

Attributes...

Table Panel

H. sapiens (1)

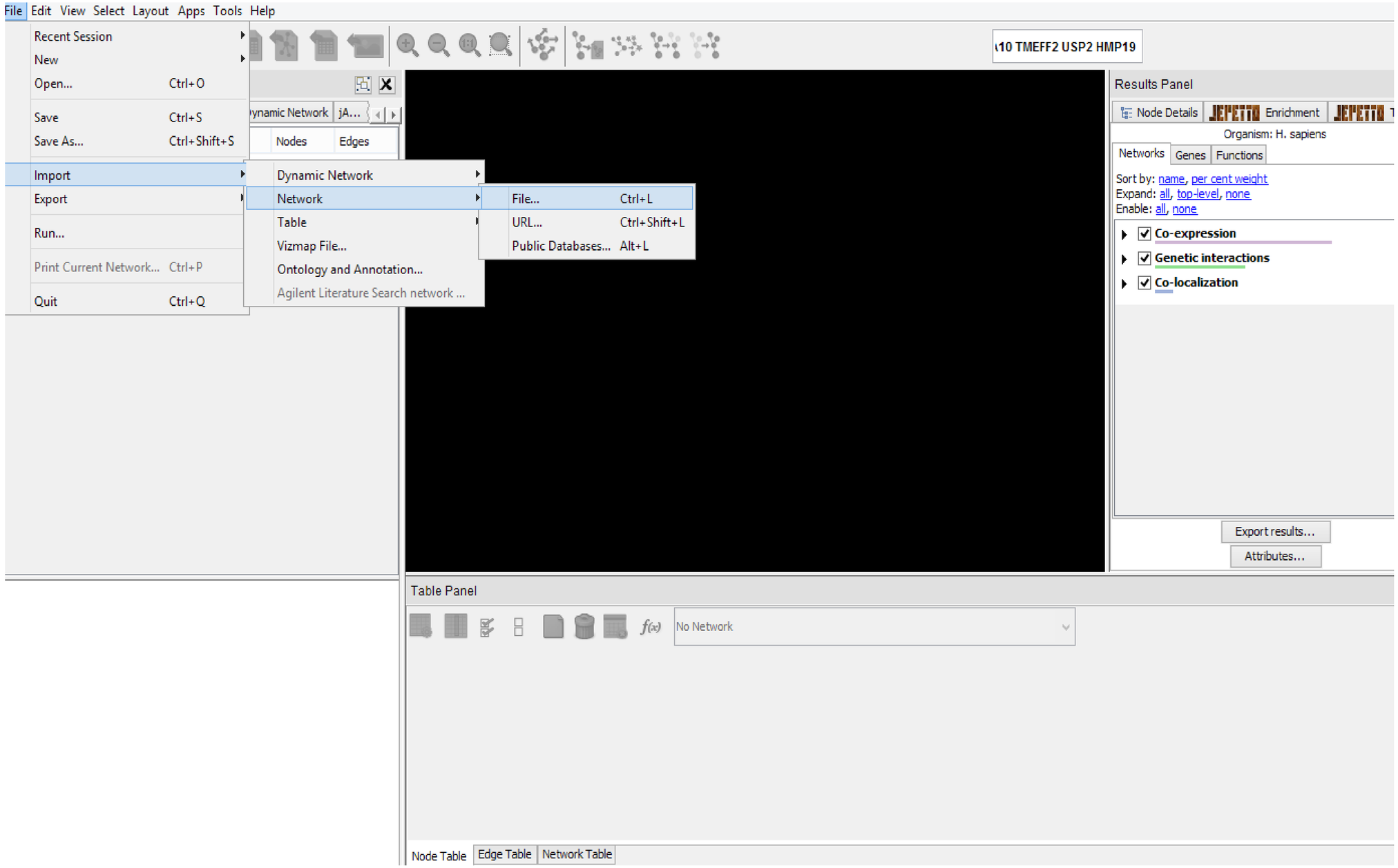
| Synonym | Ensem... | shared ... | node type | RefSeq ... | name | score | log score | Entrez ... | Ensem... | gene n... | RefSeq ... | Uniprot ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZNF179 | | H__sapie... | query | NM_0071... | H__sapie... | 0.873334... | -0.13543... | 7732 | | RNF112 | NP_0090... | Q7Z5V9 |
| | ENSP000... | H__sapie... | query | NM_0159... | H__sapie... | 0.624566... | -0.47069... | 51617 | ENSG000... | HMP19 | NP_0570... | Q9Y328 |
| TPEF | ENSP000... | H__sapie... | query | NM_0161... | H__sapie... | 0.739583... | -0.30166... | 23671 | ENSG000... | TMEFF2 | NP_0572... | TEFF2_H... |
| PMX2B | ENSP000... | H__sapie... | query | NM_0039... | H__sapie... | 0.492147... | -0.70897... | 8929 | ENSG000... | PHOX2B | NP_0039... | Q99453 |
| UNC5H3 | ENSP000... | H__sapie... | query | NM_0037... | H__sapie... | 0.540246... | -0.61572... | 8633 | ENSG000... | UNC5C | NP_0037... | UNC5C_... |
| VGLUT3 | ENSP000... | H__sapie... | query | NM_1393... | H__sapie... | 0.682550... | -0.38191... | 246213 | ENSG000... | SLC17A8 | NP_6474... | VGLU3_H... |
| PCAD | ENSP000... | H__sapie... | query | NM_0017... | H__sapie... | 0.529273... | -0.63624... | 1001 | ENSG000... | CDH3 | NP_0017... | P22223 |
| USP9 | ENSP000... | H__sapie... | query | NM_1719... | H__sapie... | 0.709608... | -0.34304... | 9099 | ENSG000... | USP2 | NP_7419... | UBP2_H... |
| BMP-3A | ENSP000... | H__sapie... | query | NM_0012... | H__sapie... | 0.542663... | -0.61126... | 651 | ENSG000... | BMP3 | NP_0011... | P12645 |
| RB3 | ENSP000 | H__sapie | query | NM_0307 | H__sapie | 0.658641 | -0.41757 | 81551 | ENSG000 | STMN4 | NP_1104 | STMN4 |

Node Table | Edge Table | Network Table

# GENEMANIA EXPORT RESULTS

| Gene 1 | Gene 2 | Weight | Type |
|---|---|---|---|
| CDR1 | PODN | 0.11246 | Co-expression |
| CDR1 | ZPBP | 0.054254 | Co-expression |
| NPPC | PHOX2B | 0.064768 | Co-expression |
| ... | ... | ... | ... |
| YY1AP1 | POTED | 0.04958 | Genetic interactions |
| YY1AP1 | SHOX | 0.027611 | Genetic interactions |
| YY1AP1 | USP2 | 0.305926 | Genetic interactions |
| ZPBP | USP2 | 0.807896 | Genetic interactions |

# NETWORK VISUALIZATION



Session: New Session

File Edit View Select Layout Apps Tools Help

Control Panel

Network Style Select BisoPanel JEPETTO

default

Properties

Def. Map. Byp.

| | |
|---|---|
| | Border Paint |
| 4.0 | Border Width |
| | Fill Color |
| 30.0 | Height |
| | Label |
| | Label Color |
| 12 | Label Font Size |
| | Shape |
| | Size |
| 255 | Transparency |
| 70.0 | Width |

Lock node width and height

Node Edge Network

NetworkPatternStage1.txt

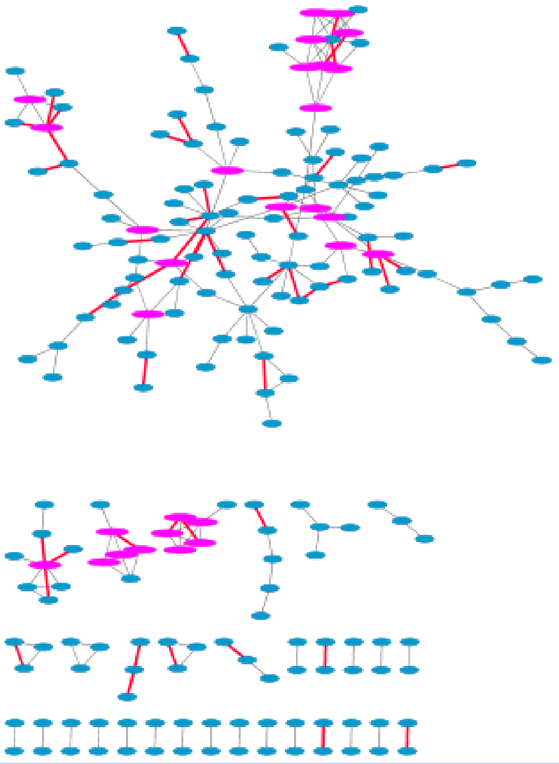


Table Panel

| shared ... | name | degree.... |
|---|---|---|
| HOXA7 | HOXA7 | 1 |
| HOXA5 | HOXA5 | 1 |
| RHOU | RHOU | 1 |
| MYOT | MYOT | 2 |
| LAMP3 | LAMP3 | 1 |
| OAS3 | OAS3 | 5 |
| SPINK5 | SPINK5 | 7 |
| C5orf23 | C5orf23 | 1 |

Node Table Edge Table Network Table

Wednesday, November 30, 2016

Session: New Session

File Edit View Select Layout Apps Tools Help

hsa-miR-208a-5p hsa-miR

Control Panel

Network | Style | Select | Diffany | BisoPanel | AllegroLayout

| Network | Nodes | Edges |
|---|---|---|
| *EdgeListVaggelis.txt* | | |
| EdgeListVaggelis.txt | 32(0) | 28(0) |
| *EdgeListVaggelismirs.txt* | | |
| EdgeListVaggelismirs.txt | 31(0) | 30(0) |
| *EdgeList.txt* | | |
| EdgeList.txt | 41(0) | 58(30) |
| **EdgeList.txt(1)** | 19(0) | 21(0) |

EdgeList.txt(1) View

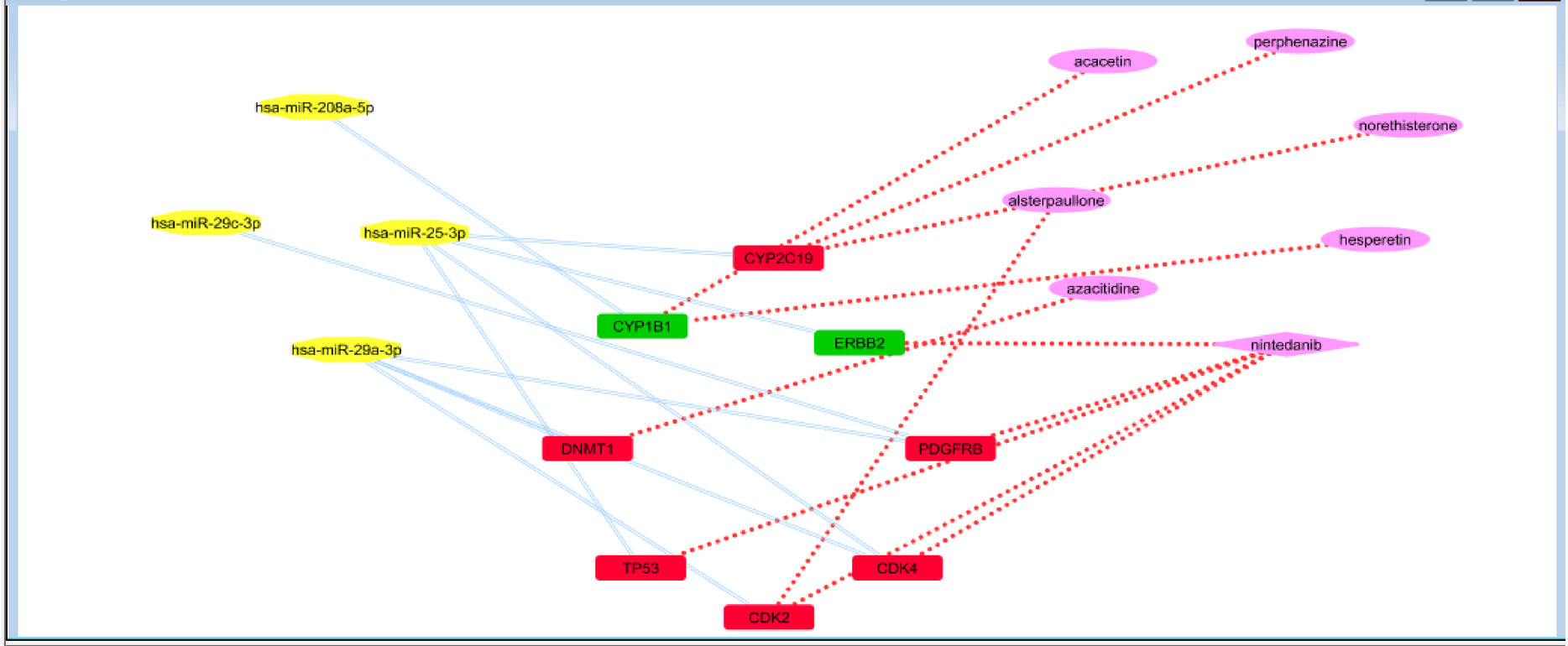



Table Panel

| shared ... | degree.... | name |
|---|---|---|
| acacetin | 1 | acacetin |
| hsa-miR-... | 4 | hsa-miR-... |
| CDK2 | 4 | CDK2 |
| hsa-miR-... | 1 | hsa-miR-... |
| TP53 | 2 | TP53 |
| ERBB2 | 3 | ERBB2 |
| CYP1B1 | 3 | CYP1B1 |
| norethist... | 1 | norethist... |

Node Table | Edge Table | Network Table

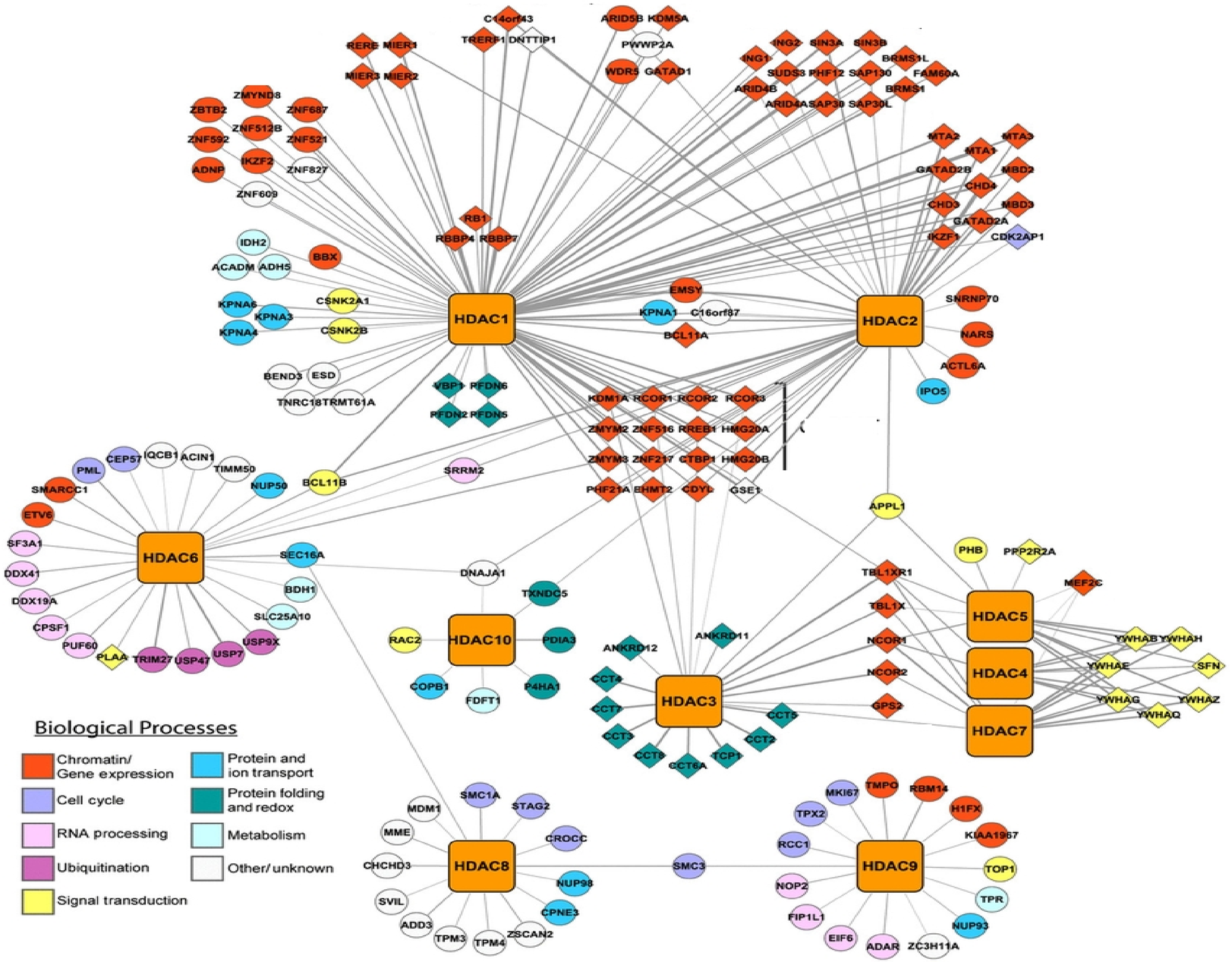
Wednesday, November 23, 2016

# SUPERNETWORK

EdgeList.txt - Notepad

File Edit Format View Help

| Source | Target | Weight |
|---|---|---|
| alsterpaullone | CDK2 | 1 |
| acacetin | CYP1A1 | 1 |
| hesperetin | CYP1A1 | 1 |
| hesperetin | CYP1B1 | 1 |
| perphenazine | CYP2C19 | 1 |
| etoposide | TOP1 | 1 |
| nintedanib | PLK1 | 1 |
| alsterpaullone | CDK5 | 1 |
| azacitidine | DNMT1 | 1 |
| nintedanib | FLT1 | 1 |
| nintedanib | EGFR | 1 |
| nintedanib | CDK2 | 1 |
| nintedanib | CDK4 | 1 |
| alsterpaullone | GSK3B | 1 |
| clobetasol | CYP1A1 | 1 |
| acacetin | CYP1B1 | 1 |
| nintedanib | TP53 | 1 |
| nintedanib | ERBB2 | 1 |
| norethisterone | CYP2C19 | 1 |
| nintedanib | PDGFRB | 1 |
| irinotecan | ABCG2 | 1 |
| etoposide | ABCC3 | 1 |
| irinotecan | TOP1 | 1 |
| nintedanib | IGF1R | 1 |
| hesperetin | SOAT1 | 1 |
| perphenazine | CALM1 | 1 |
| etoposide | TOP2A | 1 |
| nintedanib | SRC | 1 |
| hsa-miR-155-5p | PLK1 | 2 |
| hsa-miR-155-5p | CDK5 | 2 |
| hsa-miR-155-5p | DNMT1 | 2 |
| hsa-miR-155-5p | FLT1 | 2 |
| hsa-miR-155-5p | EGFR | 2 |
| hsa-miR-155-5p | CDK2 | 2 |
| hsa-miR-155-5p | CDK4 | 2 |
| hsa-miR-155-5p | GSK3B | 2 |
| hsa-miR-155-5p | CYP1A1 | 2 |
| hsa-miR-208a-5p | CYP1B1 | 2 |
| hsa-miR-25-3p | TP53 | 2 |
| hsa-miR-25-3p | ERBB2 | 2 |
| hsa-miR-25-3p | CYP2C19 | 2 |
| hsa-miR-29a-3p | PDGFRB | 2 |
| hsa-miR-192-5p | ABCG2 | 2 |



# [8] NETWORK METRICS

# NETWORK METRICS

## DEFINITIONS

**Degree Centrality**
"An important node interacts with a large number of other nodes"
Degree of center corresponds to the number of nodes adjacent to a given node.

**Closeness Centrality**
"An important node is relatively close to the other nodes in the network and can communicate quickly with them"
Proximity is defined in the simplest way as the inverse of the total distance of the node v by all other nodes

**Betweenness Centrality**
"An important node will be included in a large number of all the shortest paths among other nodes"
It is calculated as the ratio of the shortest paths running through the node v to the sum of all the shortest paths

## NETWORK METRICS IN BIOLOGY

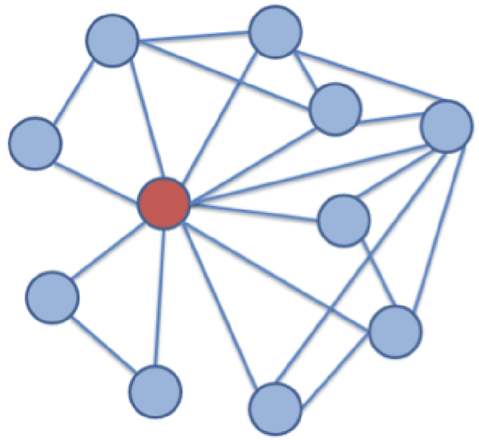
**Degree:** "Hubs" have a central regulatory role

**Closeness:** a "probability" of a protein to be functionally important for several others

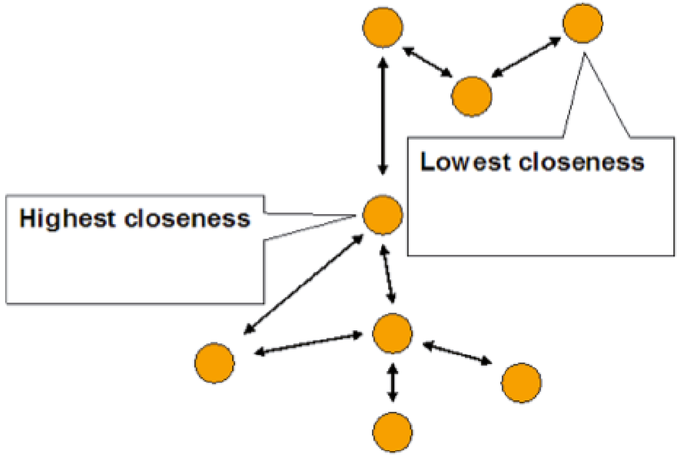
**Betweenness:** ability of a protein to bring distant proteins into communication

# [8] NETWORK METRICS

Degree centrality:
highest number of edges



Closeness centrality:
lowest average shortest distance to all other nodes

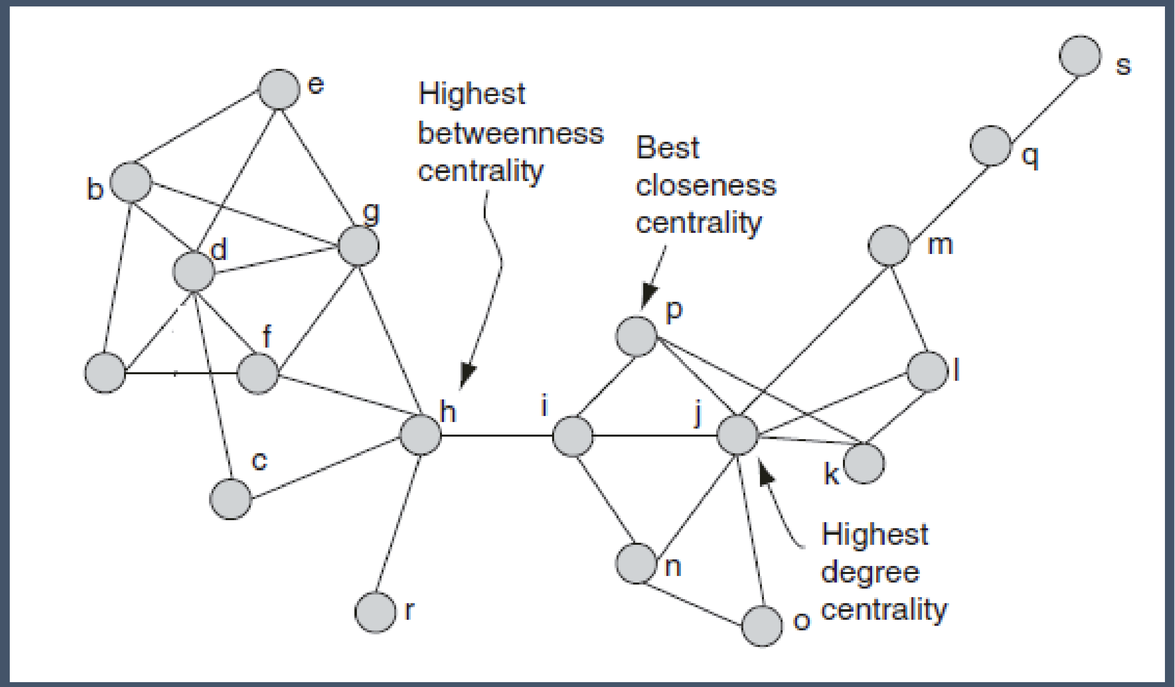



**R**

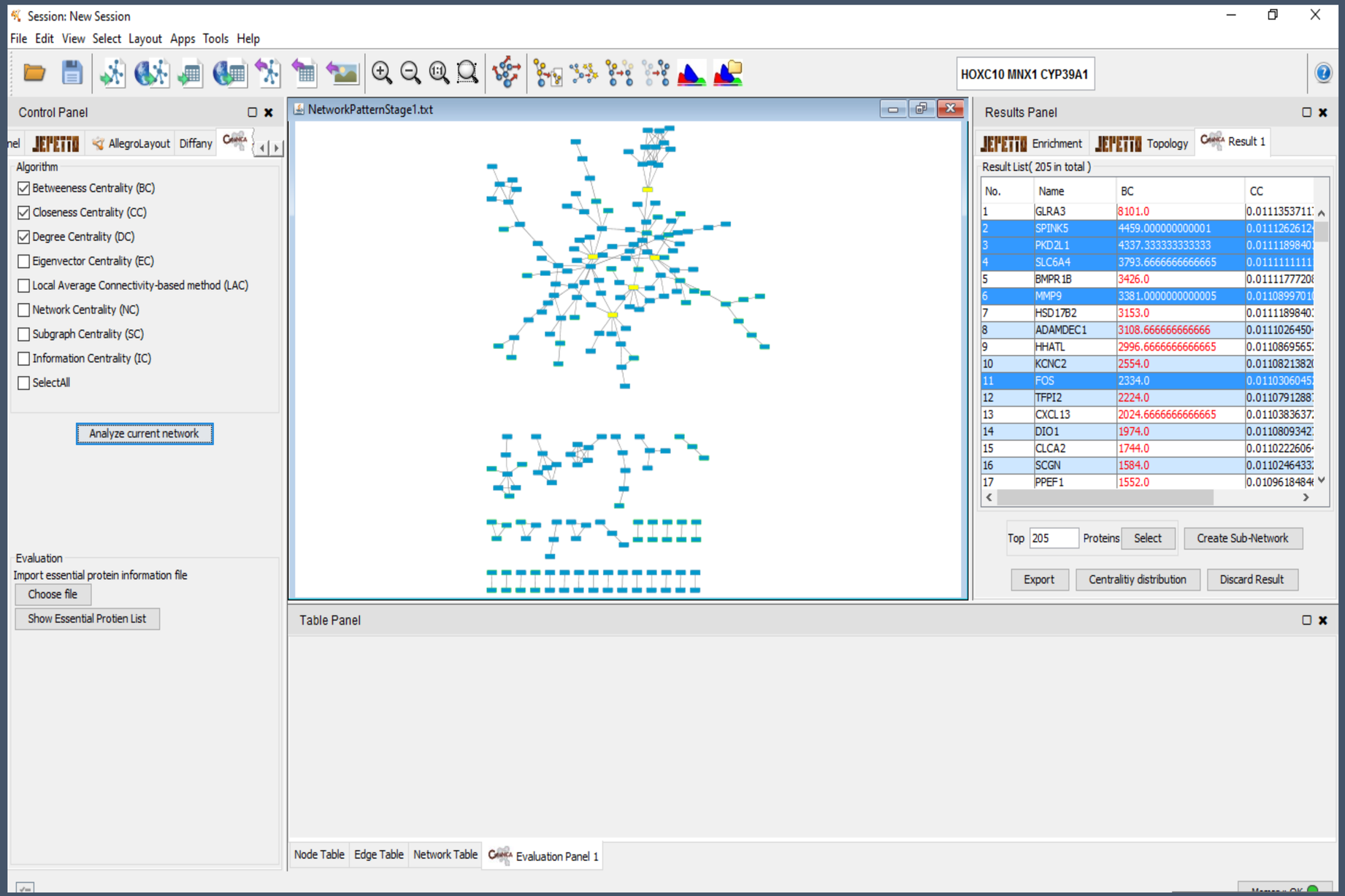
Igraph (! First convert to graph)
(global - local headquarters)

**Cytoscape**

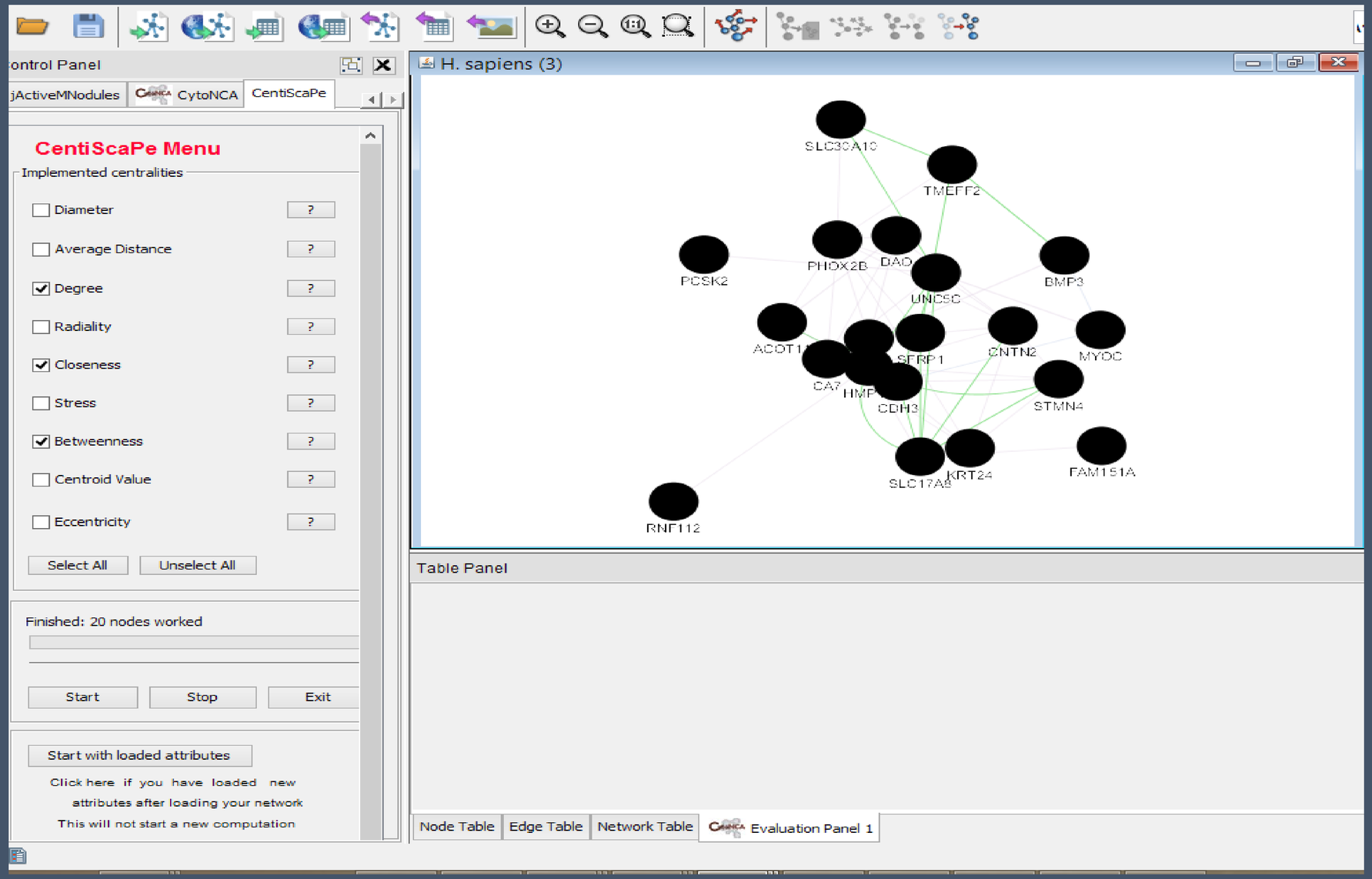
Add CytoNCA
CentiScaPe plugin
Network Analyzer



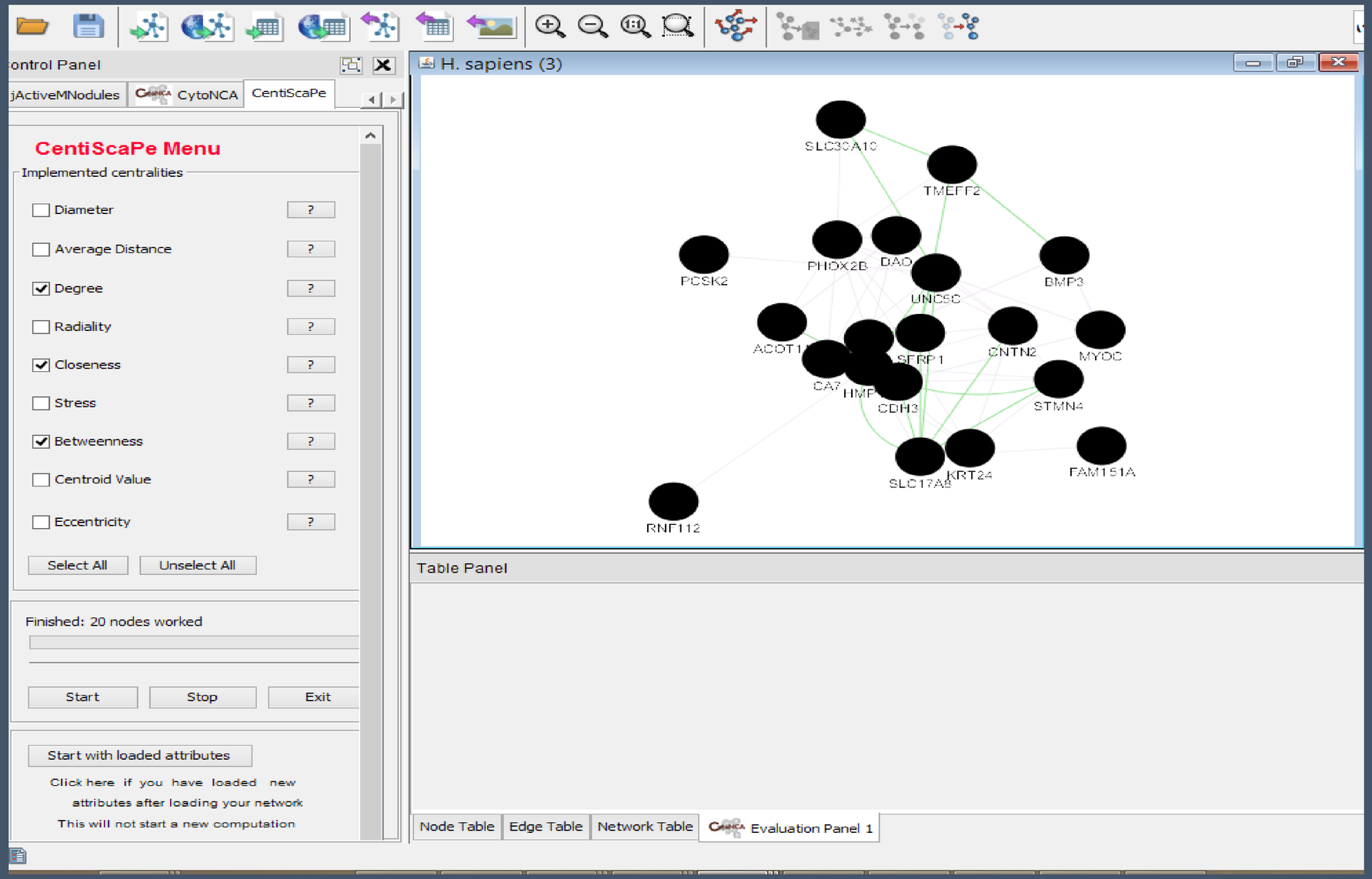
# [8] NETWORK METRICS – CytoNCA



# [8] NETWORK METRICS – CentiScaPe

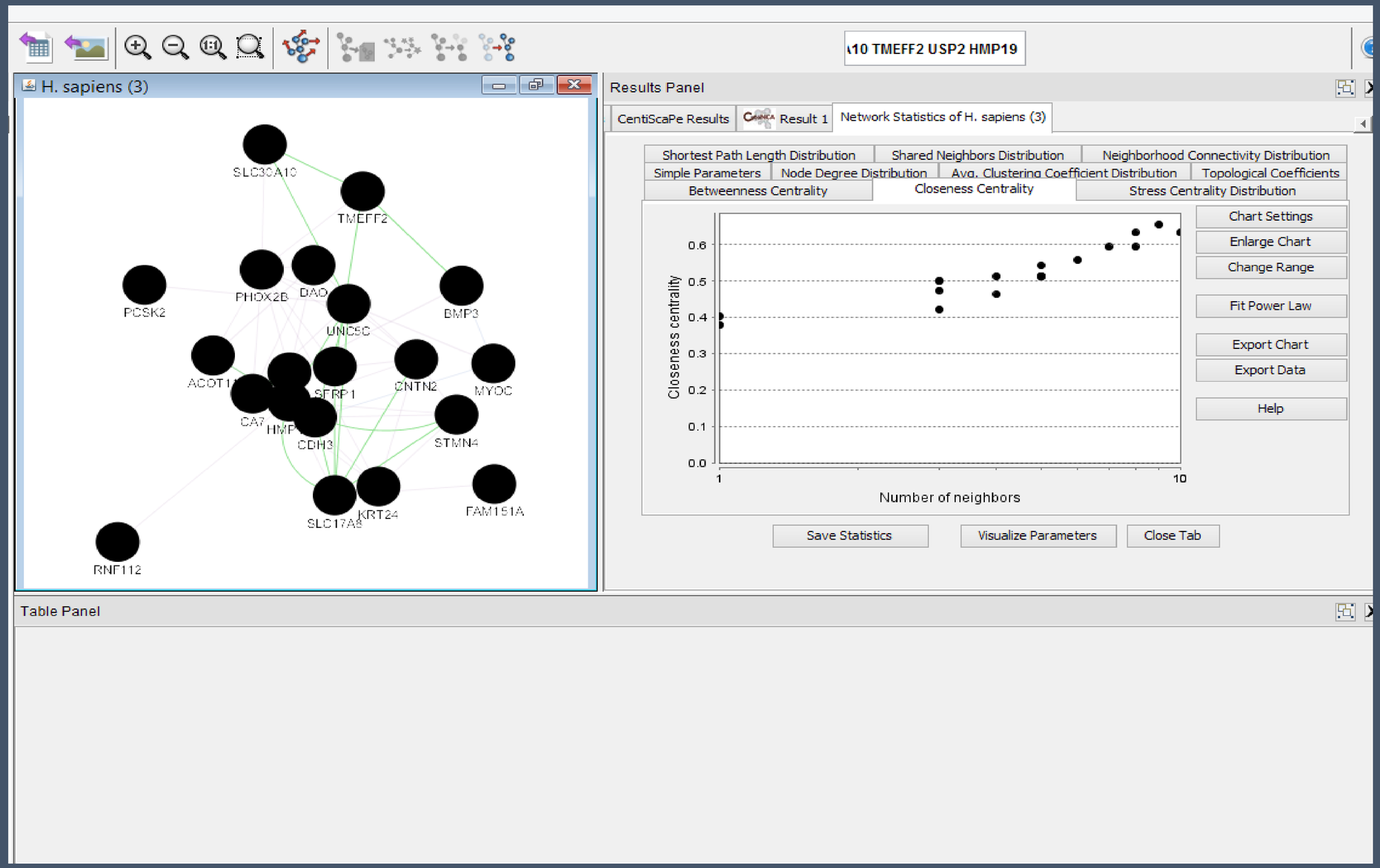


# [8] NETWORK METRICS – CentiScaPe



# [8] NETWORK METRICS

## Network Analyzer



- A = [1,1,0,1,-1]
- B =[1,-1,0,1,-1]

| Γονίδια | Πιθανότητα Εμφάνισης | | | |
|---|---|---|---|---|
| | **P(1)** | **P(0)** | **P(-1)** | **P(1)+P(0)+P(-1)** |
| A | 3/5 | 1/5 | 1/5 | 5/5=1 |
| B | 2/5 | 1/5 | 2/5 | 5/5=1 |

Η Shannon εντροπία των γονιδίων για τις 3 πιθανές καταστάσεις υπολογίζεται ως:

$$H(\gamma o\nu\iota\delta\acute{\iota}o\upsilon) = -\sum_{i=1}^{3} P_i log_2 P_i$$

άρα

$$H(A) = -(\frac{3}{5} log_2 \frac{3}{5} + \frac{1}{5} log_2 \frac{1}{5} + \frac{1}{5} log_2 \frac{1}{5}) = 1.371$$

$$H(B) = -(\frac{2}{5} log_2 \frac{2}{5} + \frac{1}{5} log_2 \frac{1}{5} + \frac{2}{5} log_2 \frac{2}{5} = 1.522$$

Στο επόμενο βήμα εξετάζεται πόσο συχνά τα δύο γονίδια έχουν την ίδια κατάσταση εξετάζοντας όλα τα πιθανά ζεύγη συνδυασμών:

| P(A,B) | Εμφάνιση |
|---|---|
| P(1,1) | 2/5 |
| P(1,0) | 0/5 |
| P(1,-1) | 1/5 |

| P(A,B) | Εμφάνιση |
|---|---|
| P(0,1) | 0/5 |
| P(0,0) | 1/5 |
| P(0,-1) | 0/5 |

| P(A,B) | Εμφάνιση |
|---|---|
| P(-1,1) | 0/5 |
| P(-1,0) | 0/5 |
| P(-1,1) | 1/5 |

Στη συνέχεια υπολογίζεται η από κοινού εντροπία H(A,B):

$$H(A,B) = -\sum_{i,j=1}^{3} P_{ij} log_2 P_{ij}$$

όπου οι τρεις καταστάσεις (1,0 και -1) είναι ανεξάρτητες άρα:

$$H(A,B) = -1\left(\frac{2}{5}log_2\frac{2}{5} + \frac{1}{5}log_2\frac{1}{5} + \frac{1}{5}log_2\frac{1}{5} + \frac{1}{5}log_2\frac{1}{5}\right) = 1.923$$

Για το παραπάνω παράδειγμα η αμοιβαία πληροφορία μεταξύ των δύο προφίλ έκφρασης, η οποία αναπαριστά την συσχέτιση μεταξύ των γονιδίων υπολογίζεται ως:

$$M(A,B) = H(A) + H(B) - H(A,B) = 1.371 + 1{,}522 - 1.923 = 0.970$$

# THE END

https://www.youtube.com/watch?v=stgYW6M5o4k

# QUESTIONS ?