

Course: BNFO 644-852

Student: Maria Somoza

UCID: mes23@njit.edu

Professor: Jason Wang

Subject: Final Project

Unsupervised GRN Inference Tool

Table of Contents

1.Introduction……………………………………………………………3

2. Purpose…………………………………………………………………..3

3. Requirements………………………………………………………….3-4

4. Downloading Data from Data Websites…………………4

4.1 NCBI Website…………………………………………………………………4-5

4.1.1 Use Search Bar for Organisms/Species………………………5

4.1.2 Download the TAR File…………………………………………………5-6

4.1.3 Extract files…………………………………………………………………7

5. Installing the Packages in R………………………………8-9

6. Starting the R Algorithm………………………………………9-12

7. Download the ARACNe Software………………………..12

7.1 Download geWorkbench………………………………………………………13

7.2 Sign up and Register………….…………………………………………………..13

7.3 Click on the geWorkbench icon to open……………………………….17

7.4 Getting the Algorithm……………………………………………………………18-19

8. Run the Analysis…………………………………………………….20

8.1 Go to file……………………………………………………………………………..20

8.2 Load the Genes for Analysis…………………………………………………22

9. Results…………………………………………………………………24

9.1 To see the Results of Graph…………………………………………………25

9.1.2 Correlation Threshold…………………………………………………….26

9.2 Histogram*……………………………………………………………………………27*

*9.3 “Universe Style” Image…………………………………………….………….30-32*

10. References………………………………………………………34

1. Introduction

What is an unsupervised gene regulatory network (GRN)? Unsupervised methods rely on expression data and tend to achieve lower prediction accuracies compared to supervised methods (2).

There are a variety of unsupervised GRN tools available such as: ARACNe, context likelihood of relatedness (CLR), relevance networks (RN), etc. (1)

2. Purpose

The main purpose of this project is to analyze the top 100 variant genes of the data GSE19587 obtained from GEO database. I will infer regulatory networks from gene expression data; extracting the affymetrix formats.

3. Requirements

Software

The system design requires the following software:

|  |  |
| --- | --- |
| Component | Requirement |
| Programming Language | R x64 3.0.1 |
| Unsupervised GRN Inference Tool | ARACNe |

Hardware

The system design requires the following hardware:

|  |  |
| --- | --- |
| Component | Requirement |
| Platform | Windows 7  Hp (64bits) |

Downloading Required Programming Language

For programming, download the following:

 R x64 3.0.1

<http://www.r-project.org/>

Downloading Required Software

For Unsupervised GRN Inference download the following tool:

 ARACNe

<https://wiki.c2b2.columbia.edu/workbench/index.php/ARACNe>

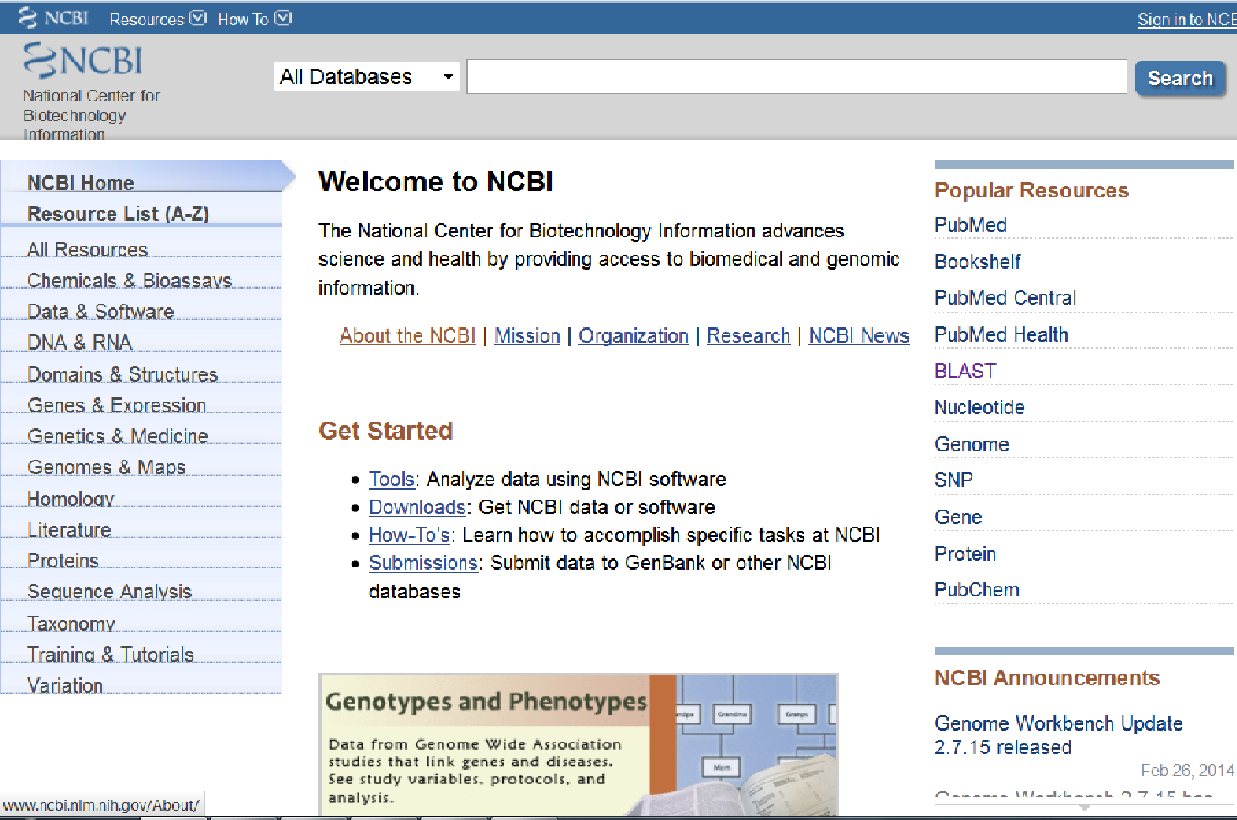
4. Downloading Data from Data Websites

4.1 Go to the NCBI website and type the accession number on the “GEO accession”

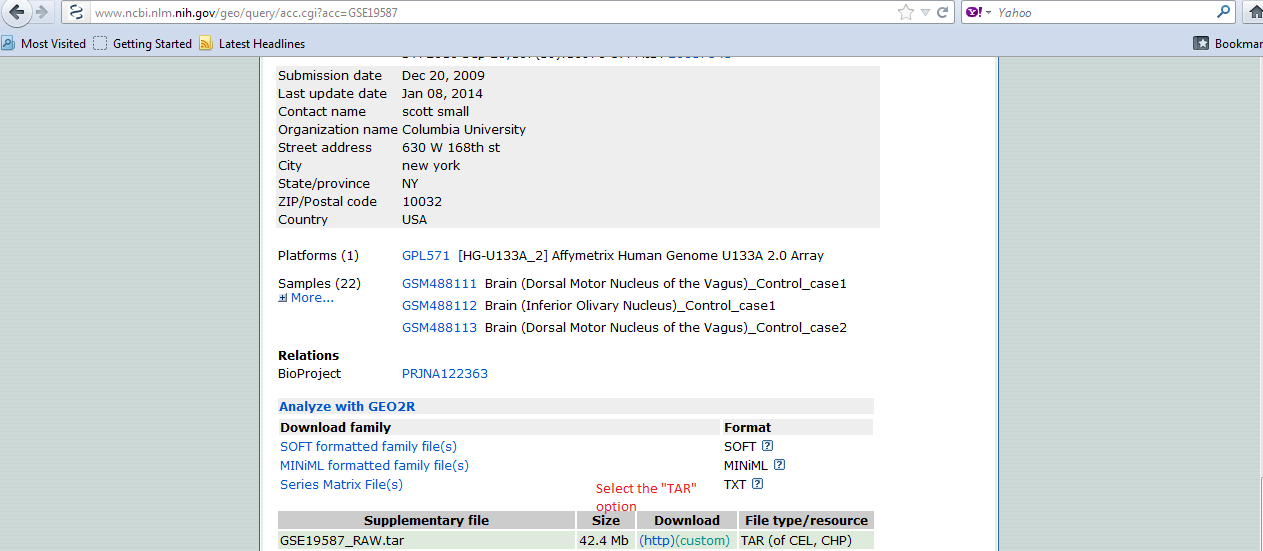
<http://www.ncbi.nlm.nih.gov/>



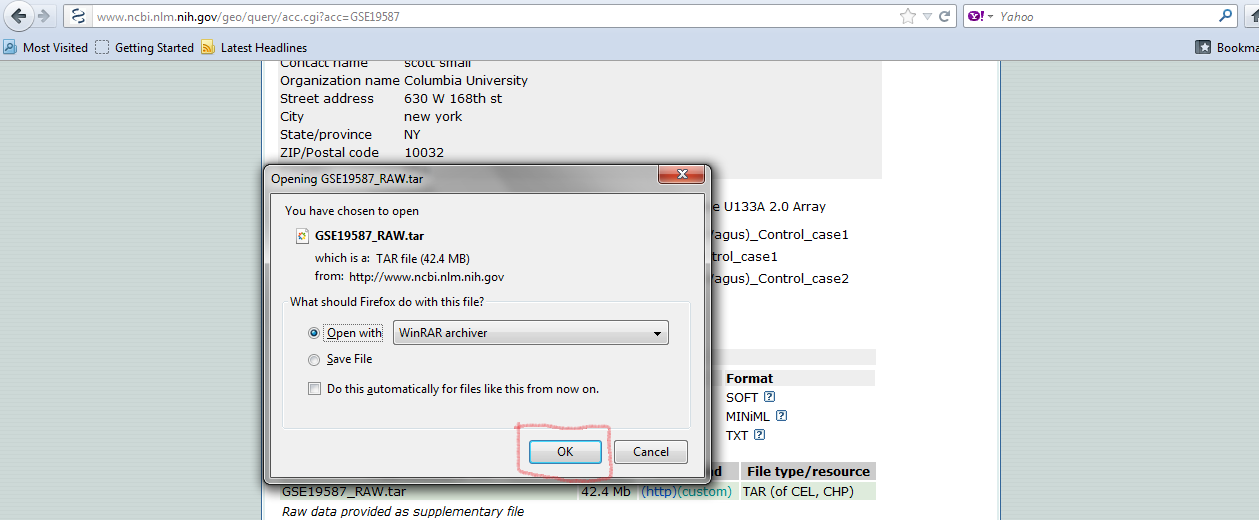
4.1.1 Or type in the name of the organisms/species



4.1.2 Download the TAR file

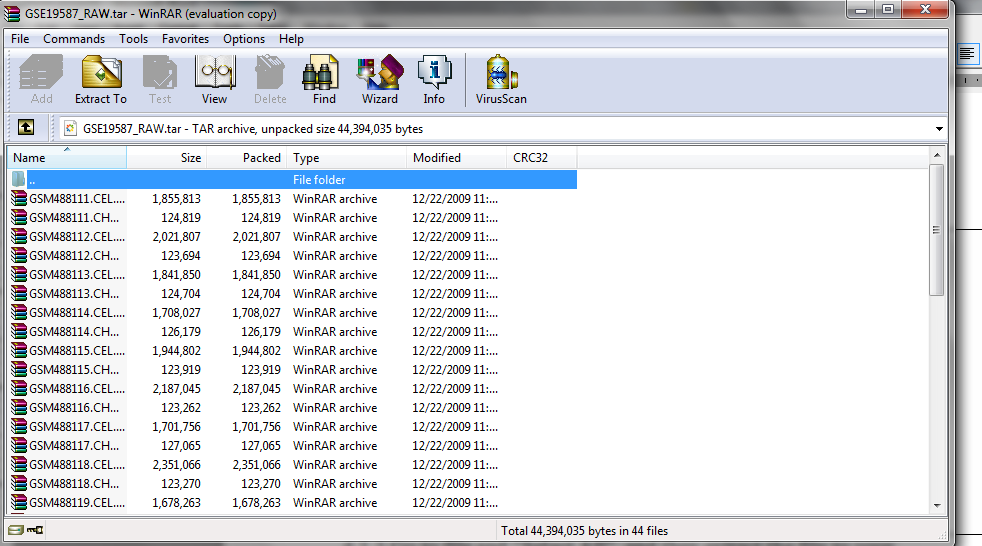


Click OK



Note: Go to <http://www.win-rar.com/download.html?&L=0> to download WinRAR to unzip the files.

4.1.3 Go to file and “Select All”, and then extract the file to your chosen directory.

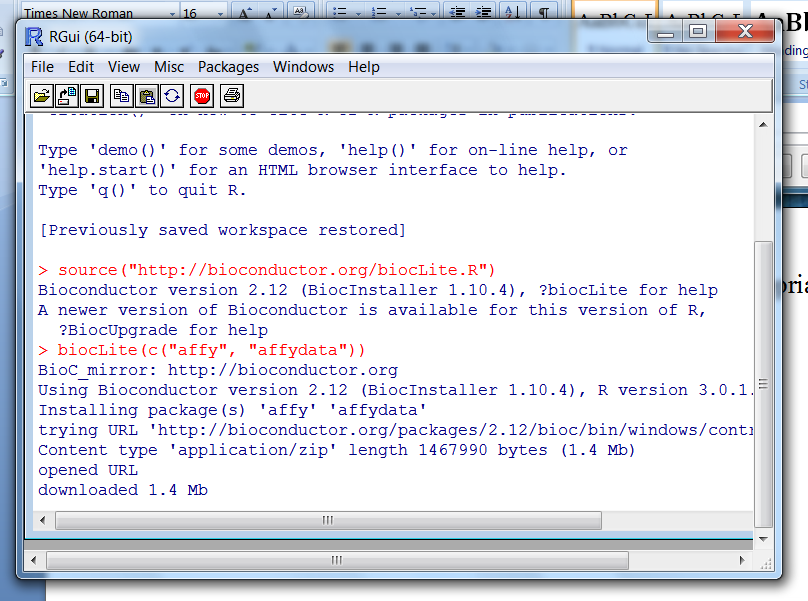


5. Installing the packages in R

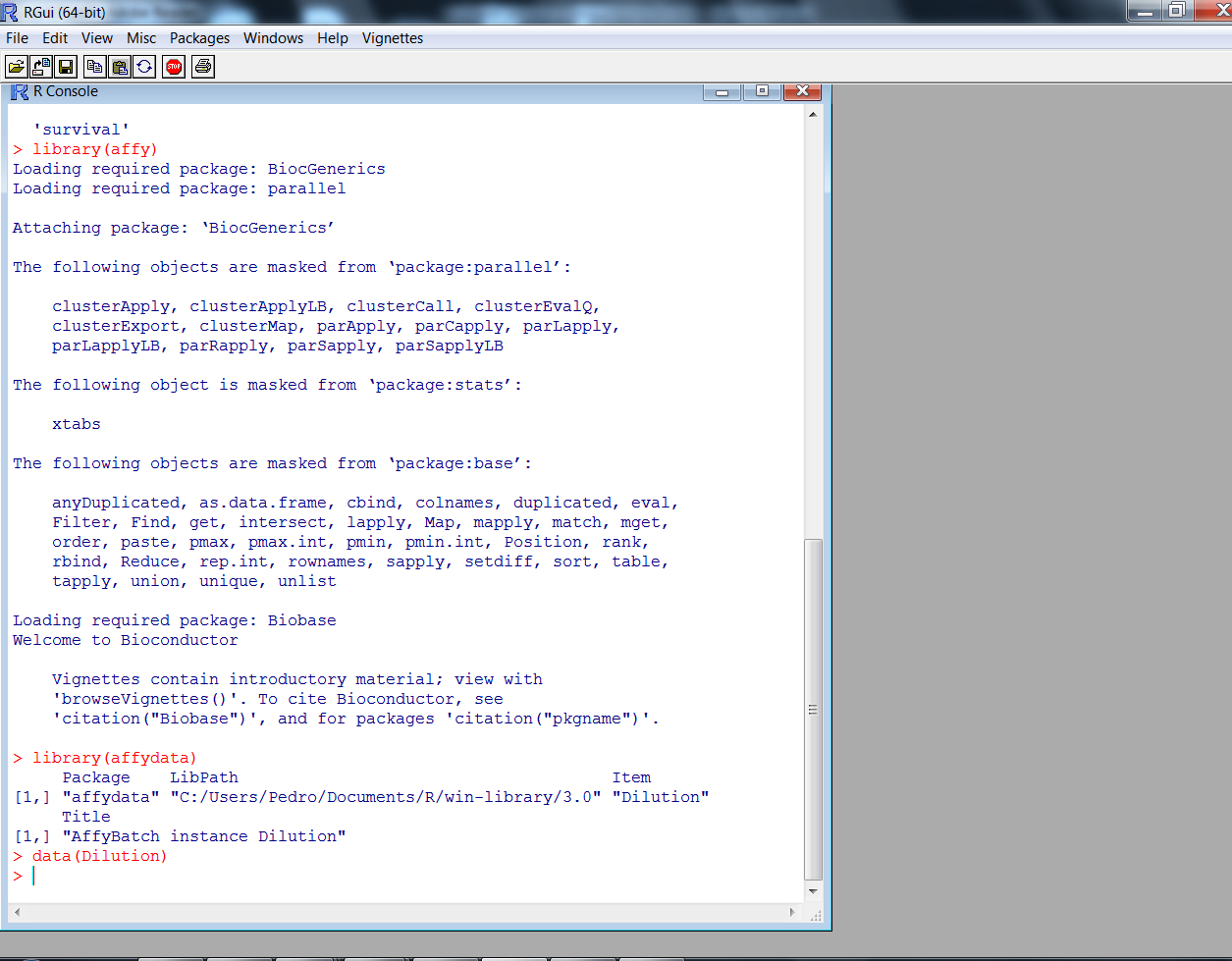
1. Assuming that the R language is downloaded appropriately

2. Install the following packages in R

* **source("http://bioconductor.org/biocLite.R") -🡪 > biocLite(c("affy", "affydata"))**



* **library(affy)🡪 library(affydata) 🡪 data(Dilution)**



6. Begin Writing the R Algorithm

# Normalize using rma and specify the location of the extracted files

>Data<-ReadAffy(celfile.path="where/your/extracted/zipped/files/directory/ ", compress=TRUE)

\*\*Notice: This may take a while to read

>eset<-rma(Data)

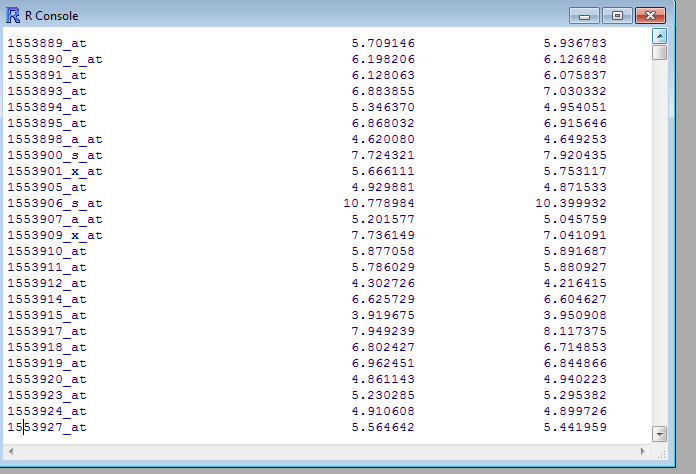
Note: If this function doesn’t work, then use justRMA() function

>eset<-justRMA(celfile.path=" where/your/extracted/zipped/files/directory/ ", compress=TRUE)

##Return the normalized data, each row a gene and each column a

sample

>exprs(eset)

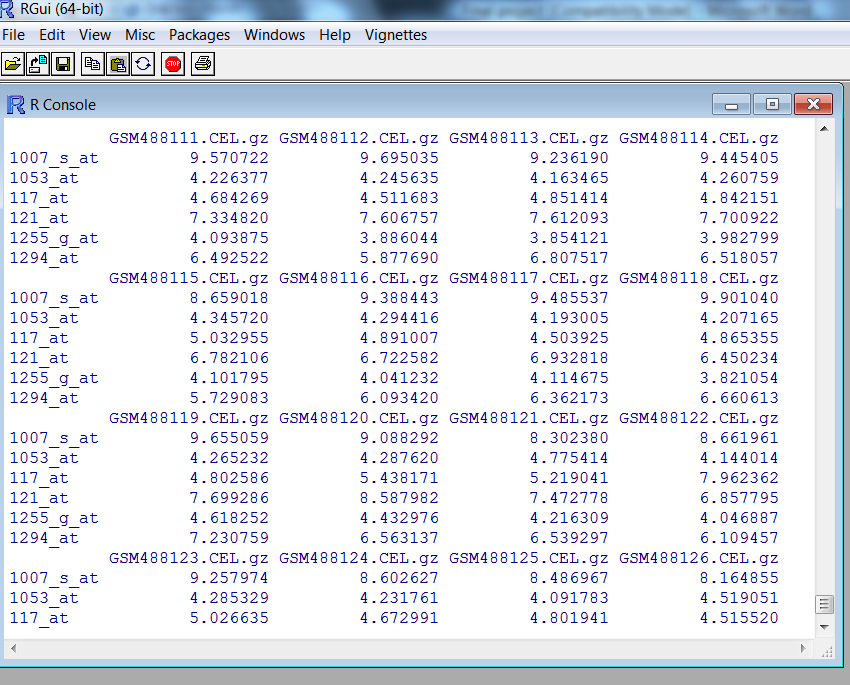


#Extract the normalized data from eset object

>norm.expr<-exprs(eset)

# To see how the data look like

>head(norm.expr)



#Output your normalized data into a file

>write.exprs(eset, file="Parkinsondata.txt")

#Now read your normalized data into R using read.table()

>geneExpr=read.table("Parkinsondata.txt")

#Get Gene number

>nGene=nrow(geneExpr)

# Initialize variance to be 0

>geneVar=rep(0, nGene)

>for (i in 1:nGene) {

geneVar[i]<-var(unlist(geneExpr[i,]))

}

#Now we will get the top 100 variant genes

>cutoff=quantile(geneVar, 1-100/nGene)

>top100=geneExpr[geneVar>cutoff, ]

#Name the samples to be more readable

>colnames(top100)=c(paste("Dorsal\_Motor\_Nucleus of the Vagus\_Control",1:4,sep=""), paste("Inferior\_Olivary\_Nucleus\_Control",1:4,sep=""), paste("Dorsal\_motor\_Nucleus of the Vagus\_Parkinson",1:6, sep=""), paste("Inferior\_Olivary\_Nucleus\_Parkinson",1:6,sep=""))

#Output for analysis

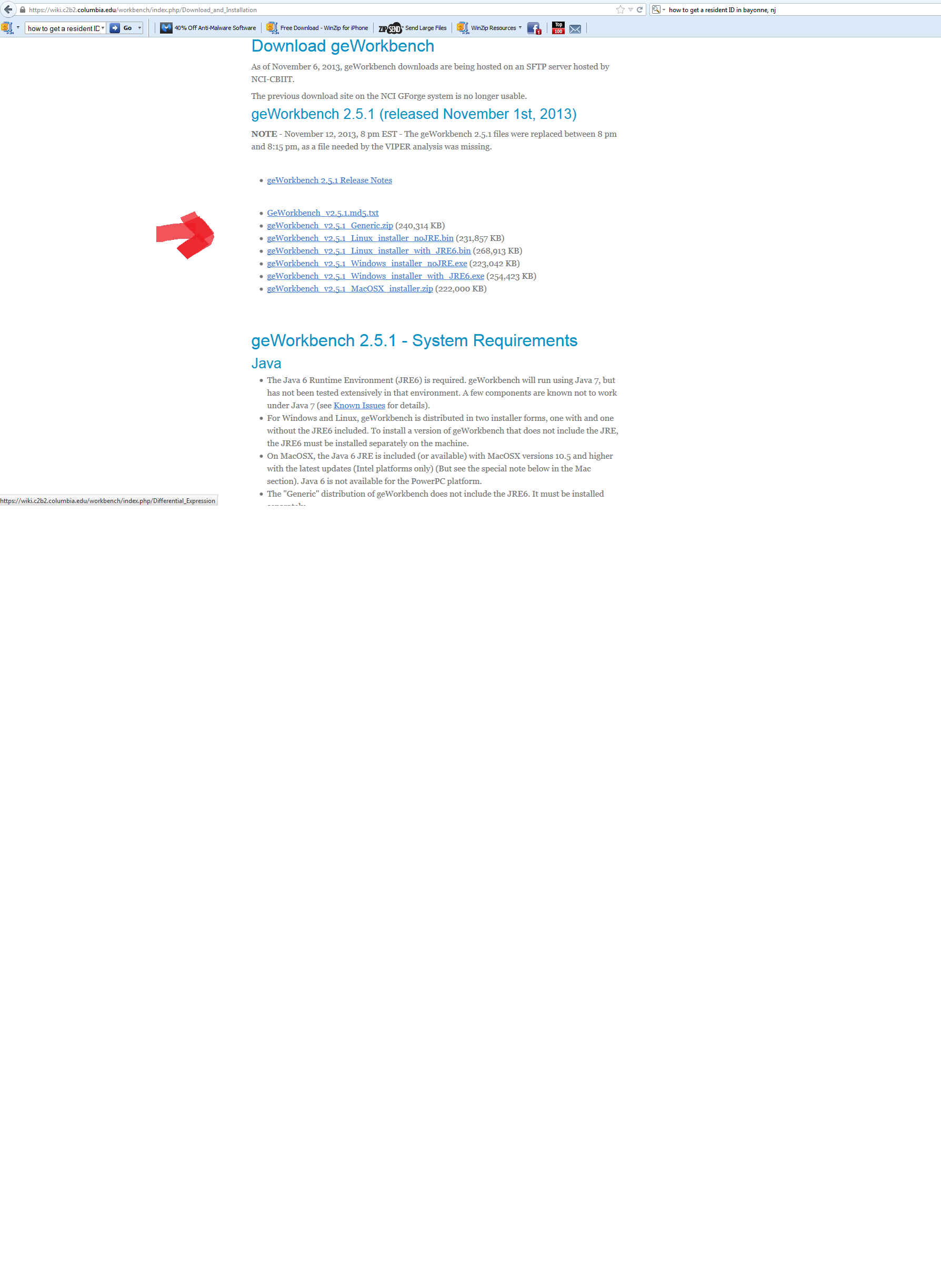
>write.table(data.frame(AffyID=rownames(top100),top100), file="GSE19587.txt",sep="\t", quote=F, row.names=F)

7. Download the ARACNe Software

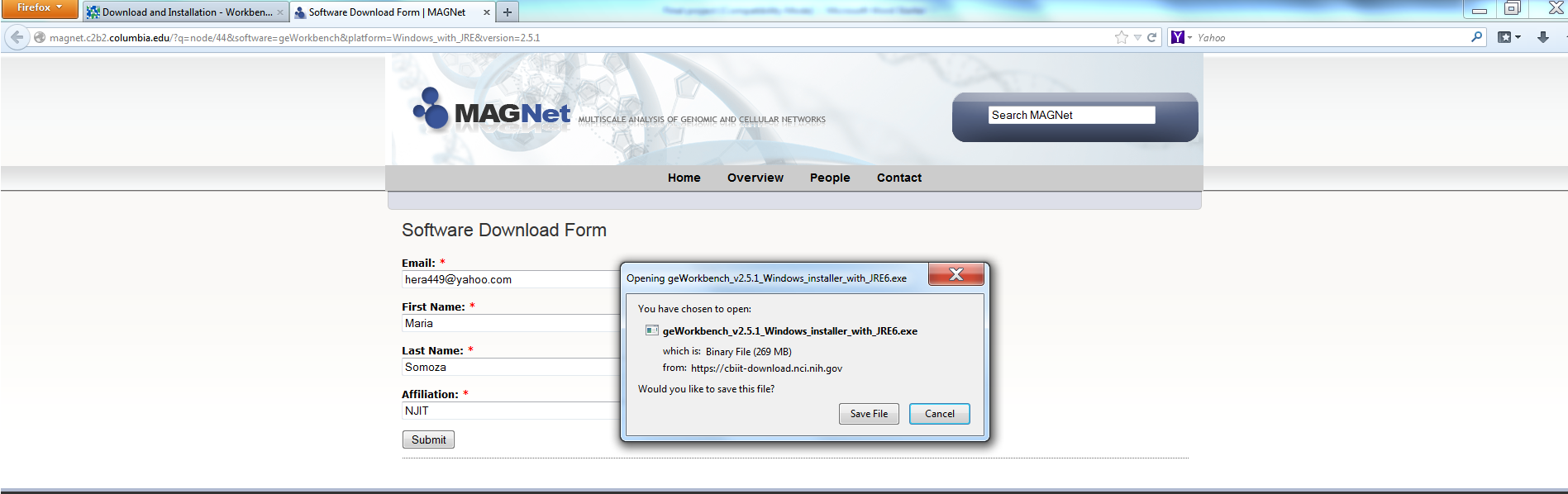
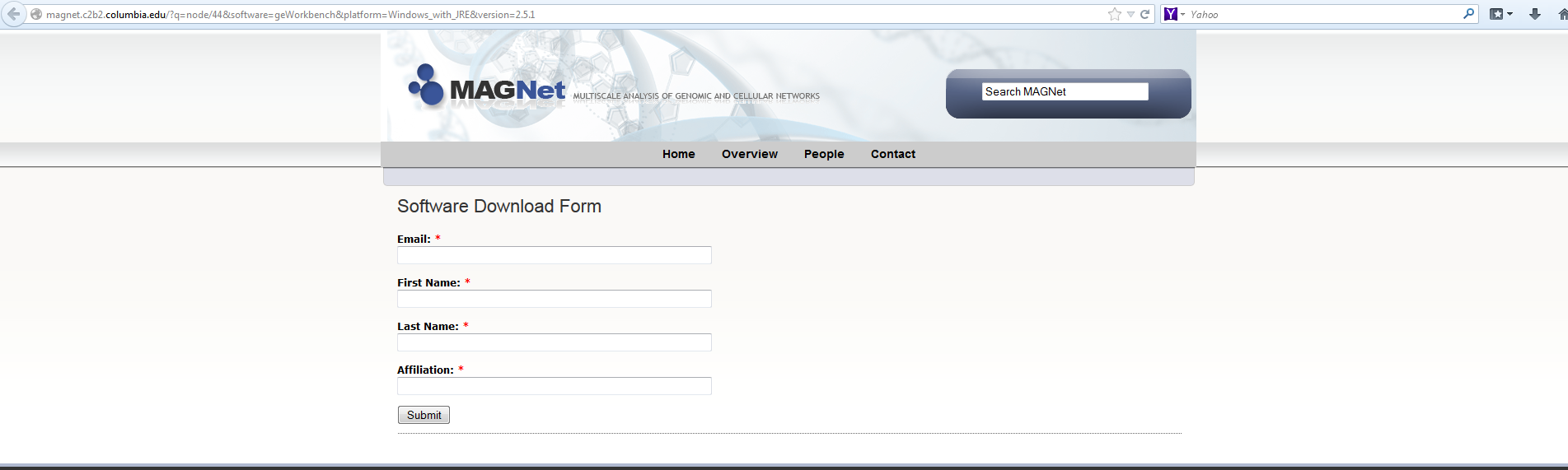
* **General Description of ARACNe**

**ARACNe is an information-theoretic algorithm used to identify transcriptional interactions between gene products using microarray gene expression profile data.**

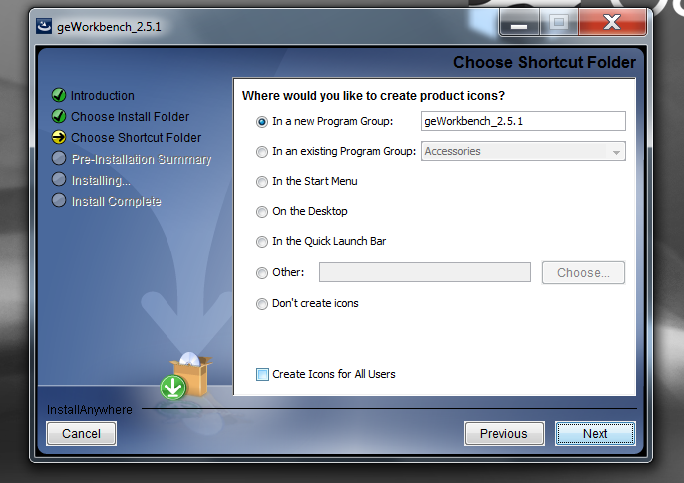
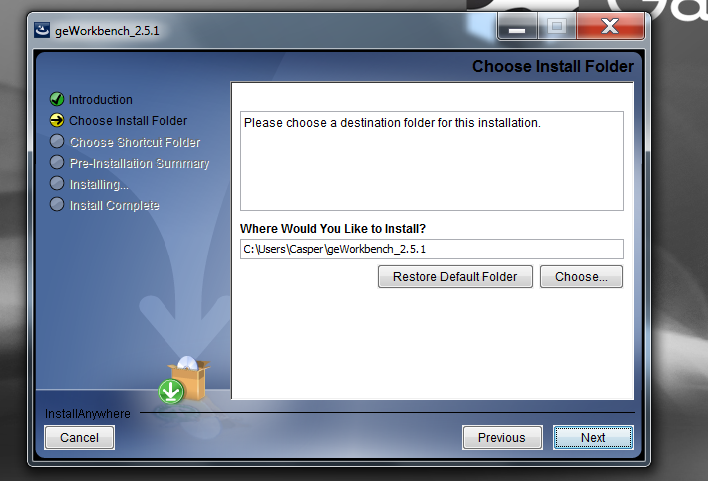
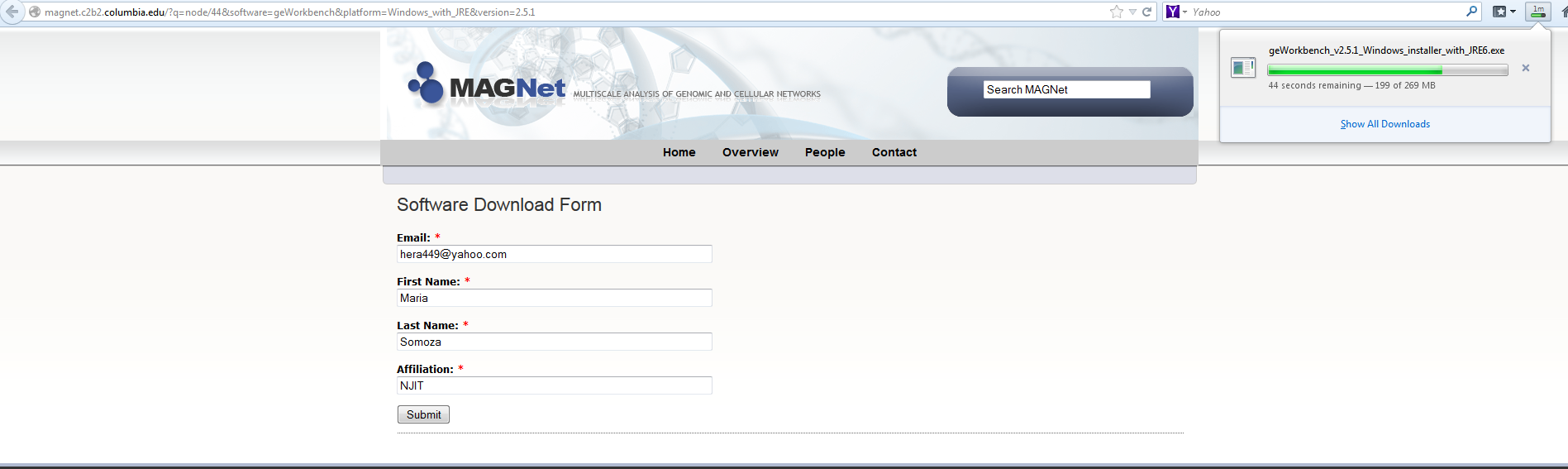
7.1 First, download the appropriate geWorkbench with the appropriate platform



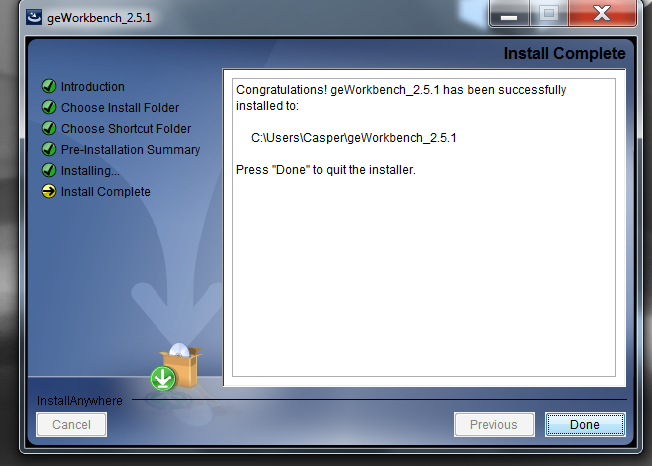
7.2 You have to sign up and register in order to download the software



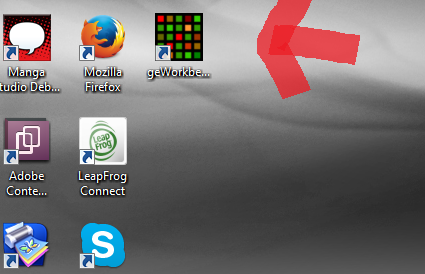
Wait for it to download



Click on ‘Done’ once it finish installing

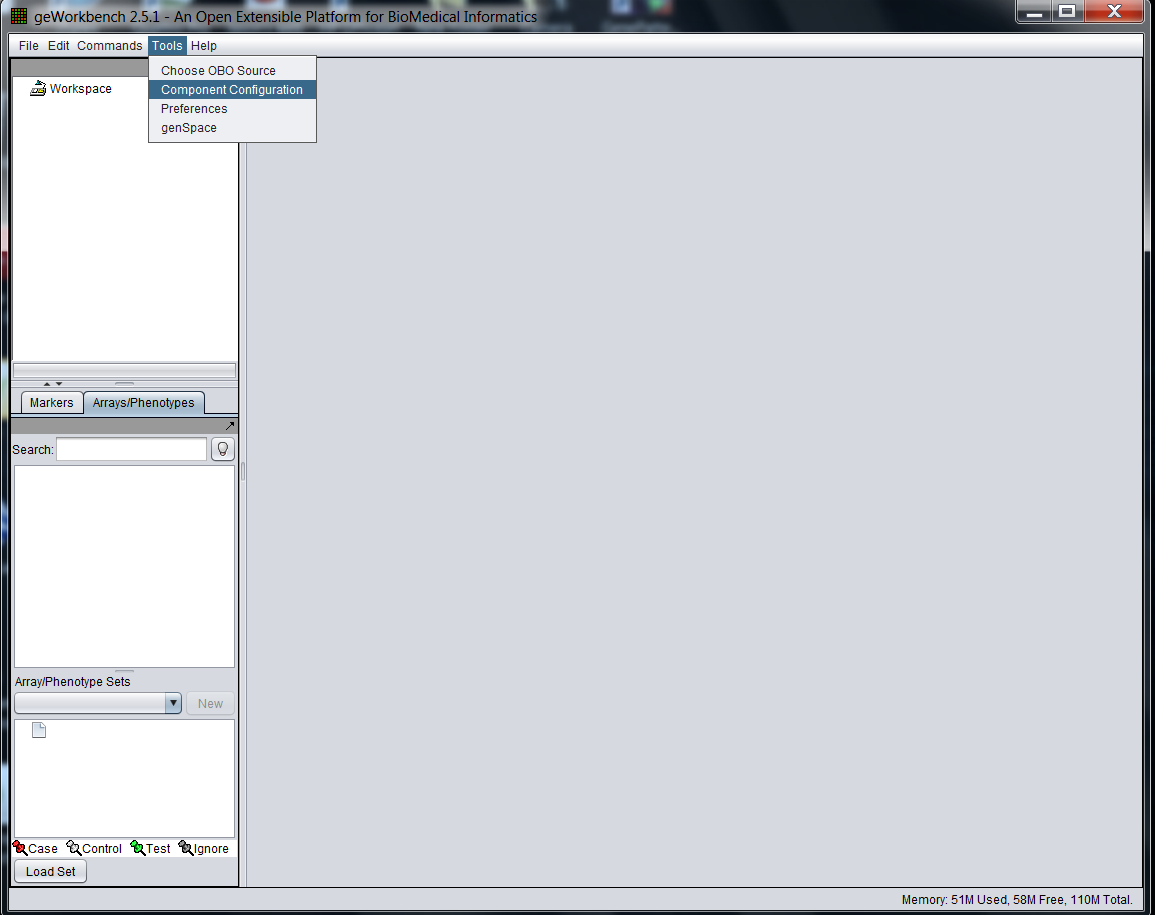


7.3 Click on the geWorkbench icon

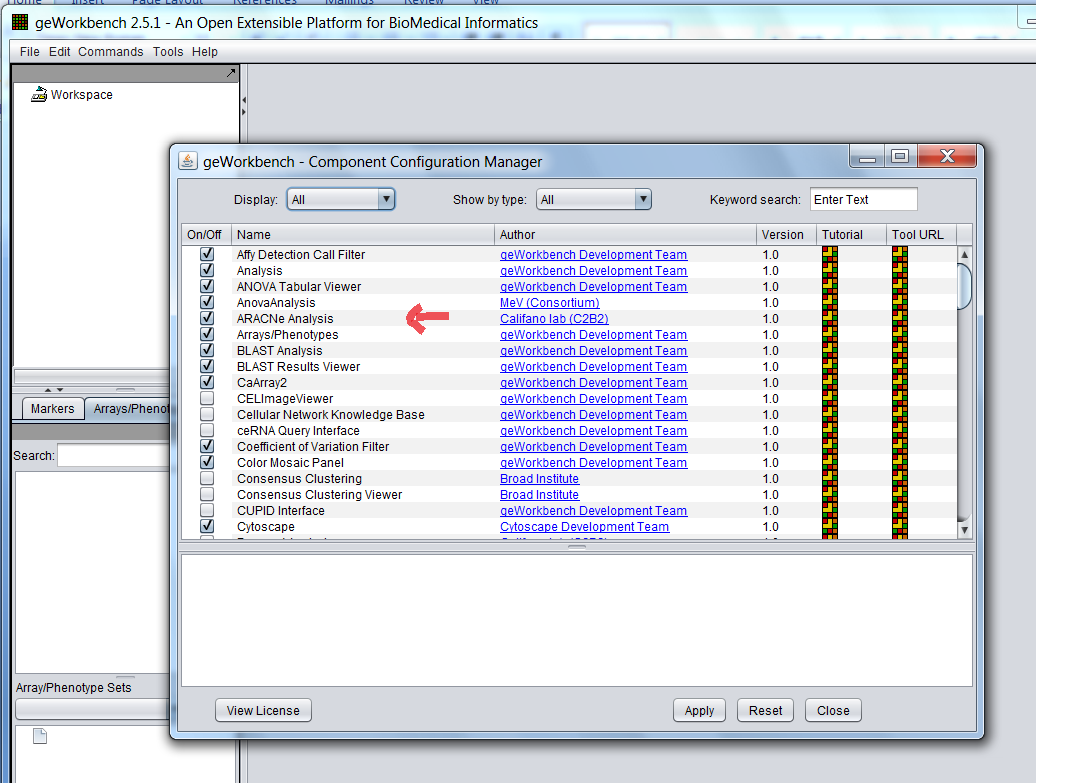


In order to get the algorithm in the software:

Go to Tools 🡪 Component Configuration

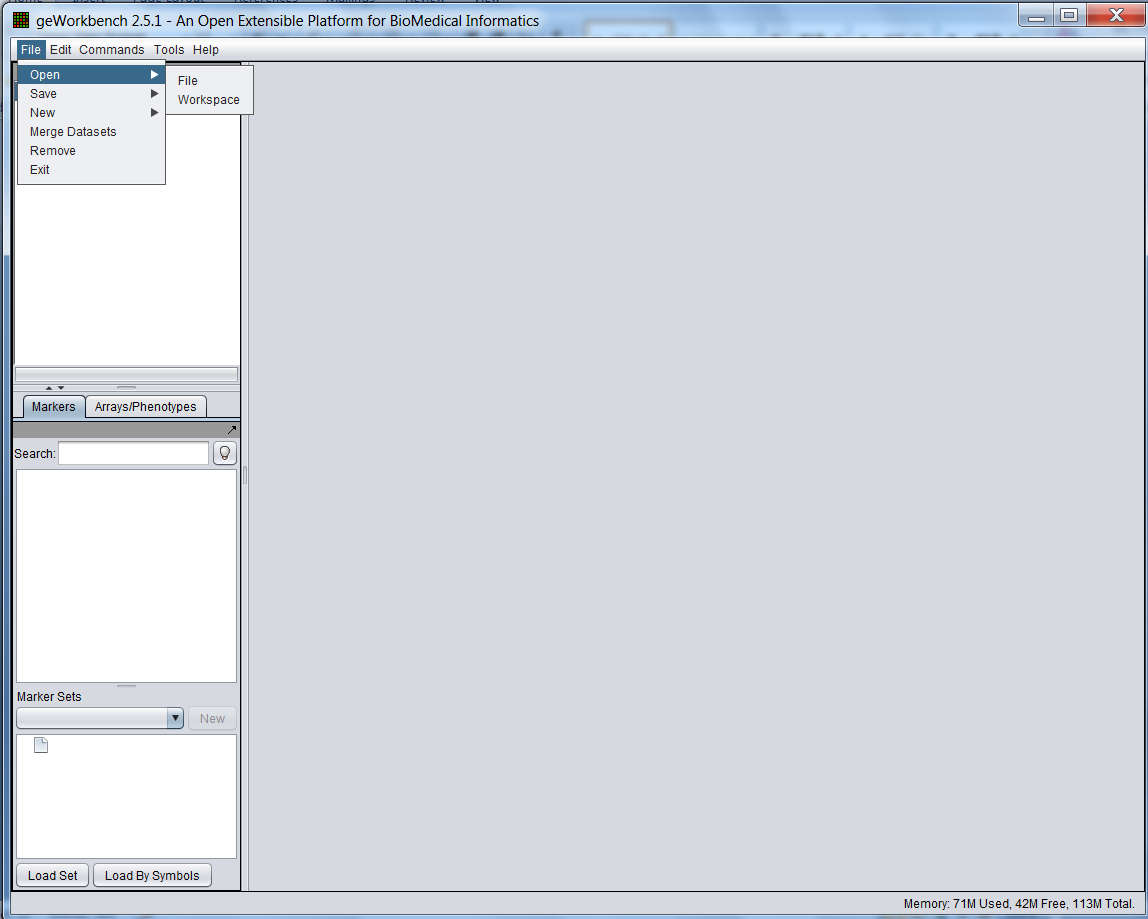


Click on the ‘ARACNe’ Analysis box

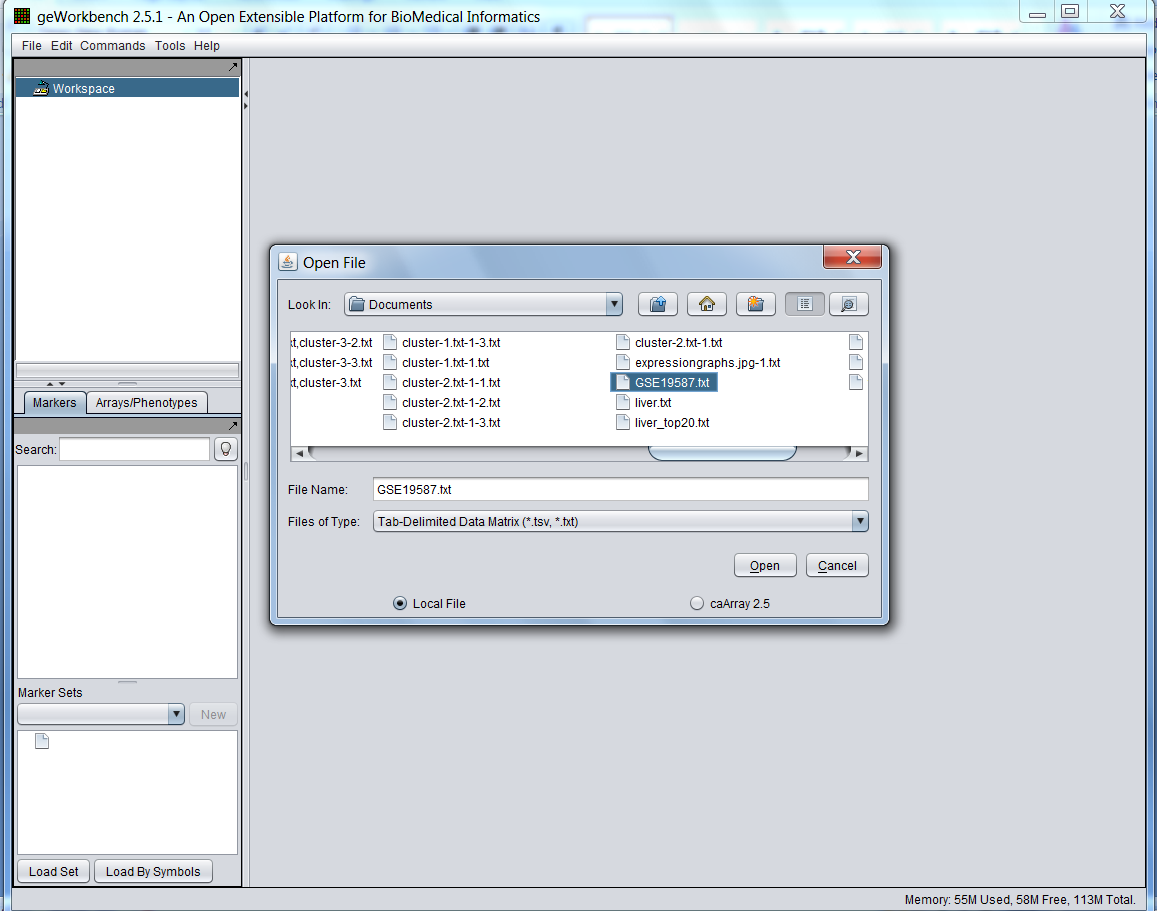


8. Run the Analysis

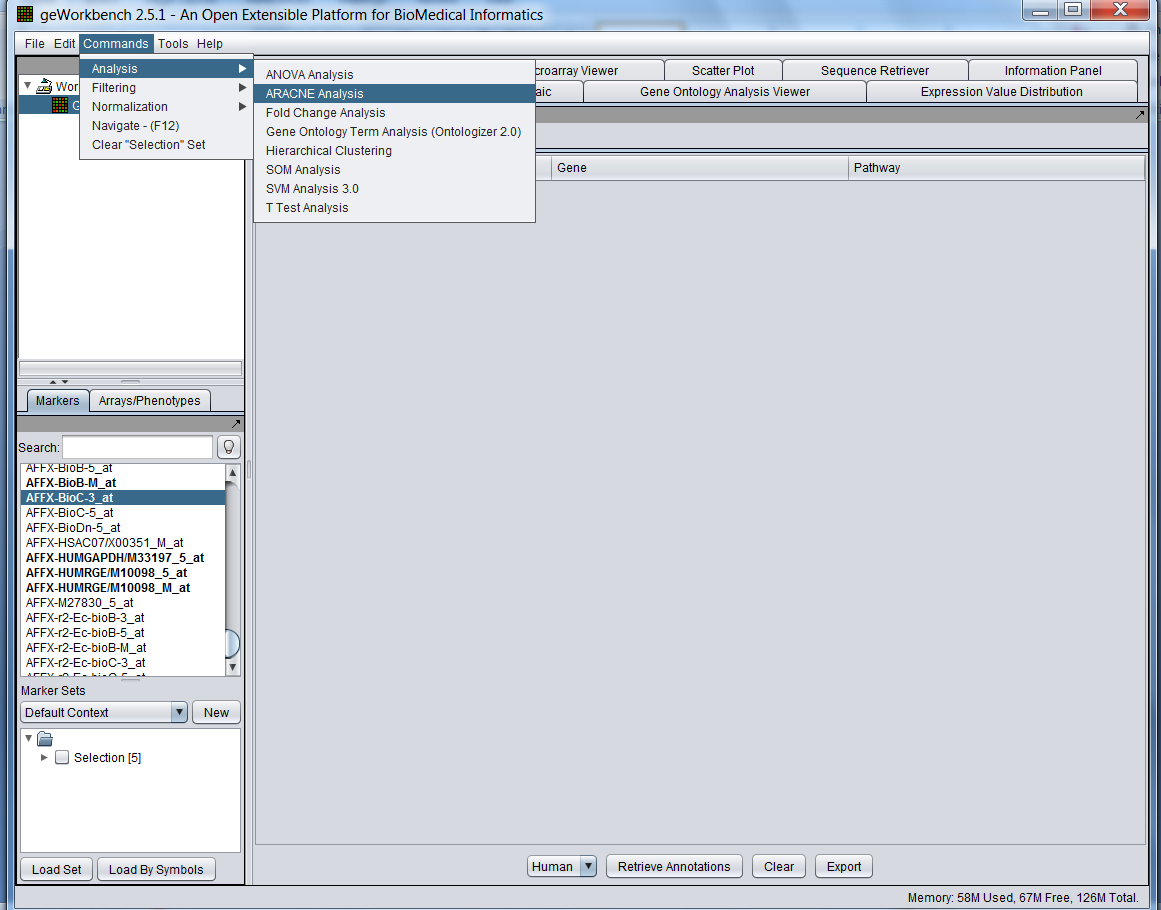
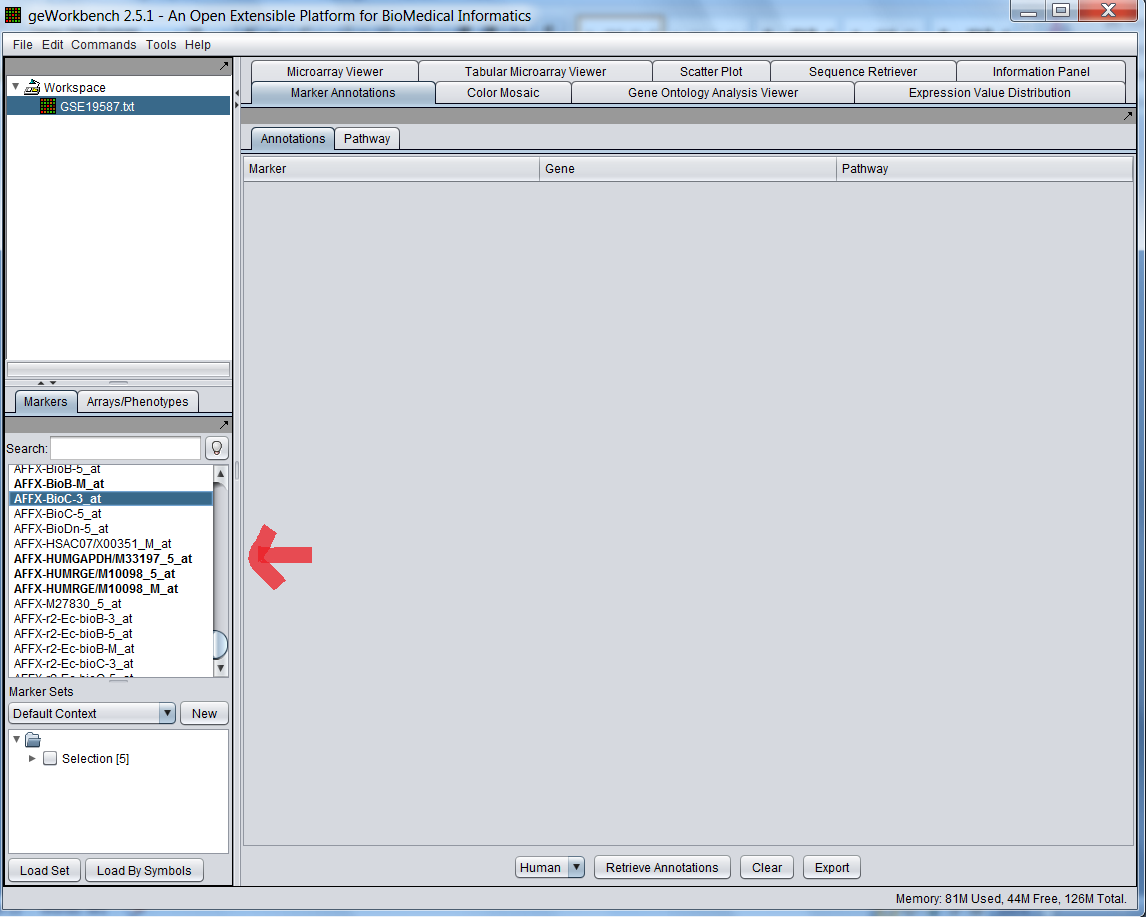
Go to File 🡪 Open 🡪 File



8.1 Locate the saved data and click to open



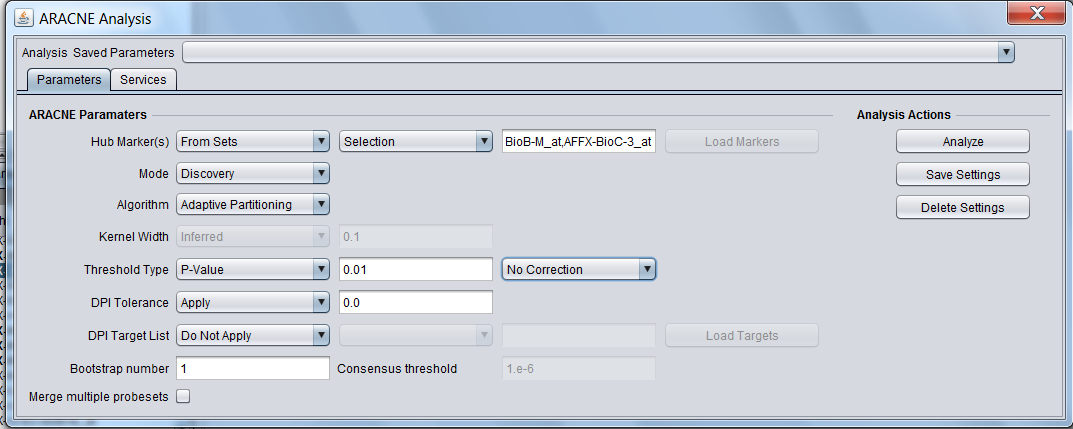
8.2 Select the genes for ARACNe Analysis



On the Parameters section:

1. Select the Hub Markers

2. Select the Mode, Algorithm, Threshold Type, etc.



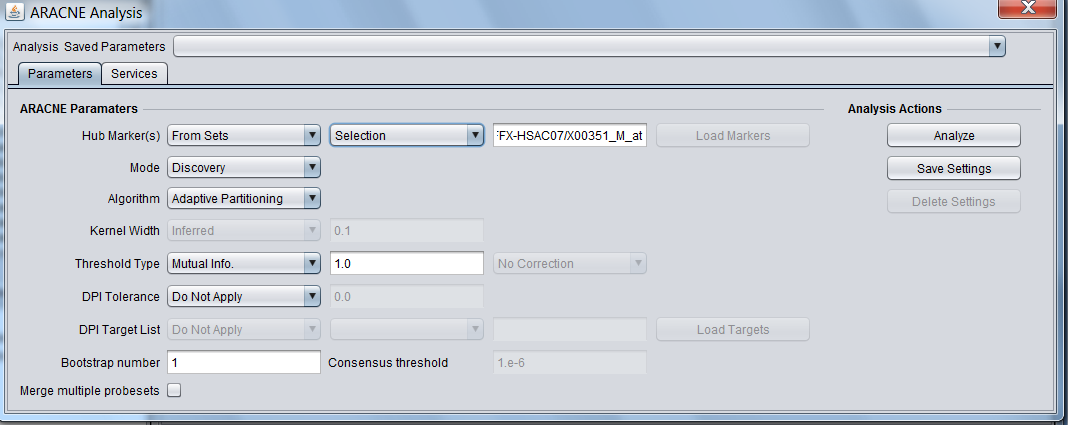
9.Results

The selected genes were analyzed as follows:

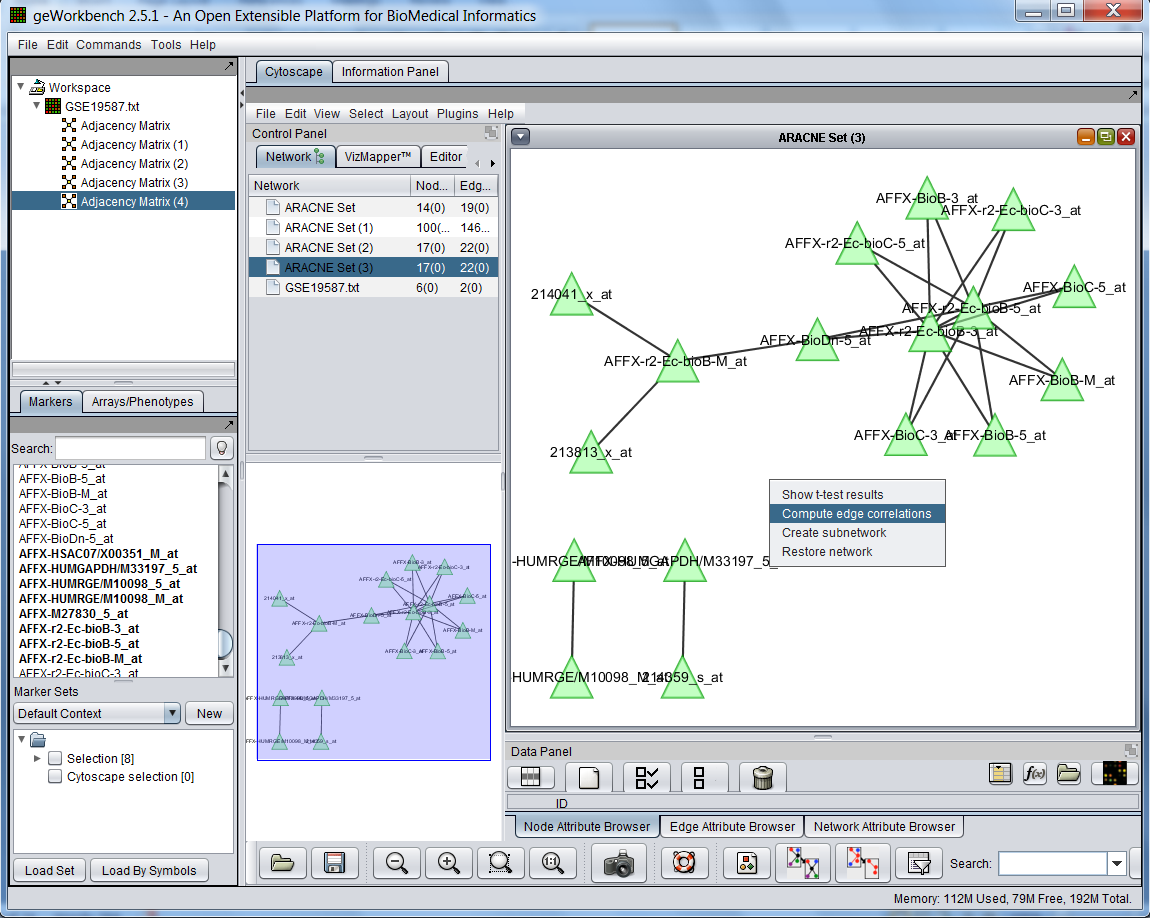
1. The Hub Marker selected was AFFX-HSAC07/X00351\_M\_at

2. The Mode selected was Discovery 🡪 Algorithm: Adaptive Partitioning

3. Threshold Type was MI of 1.0



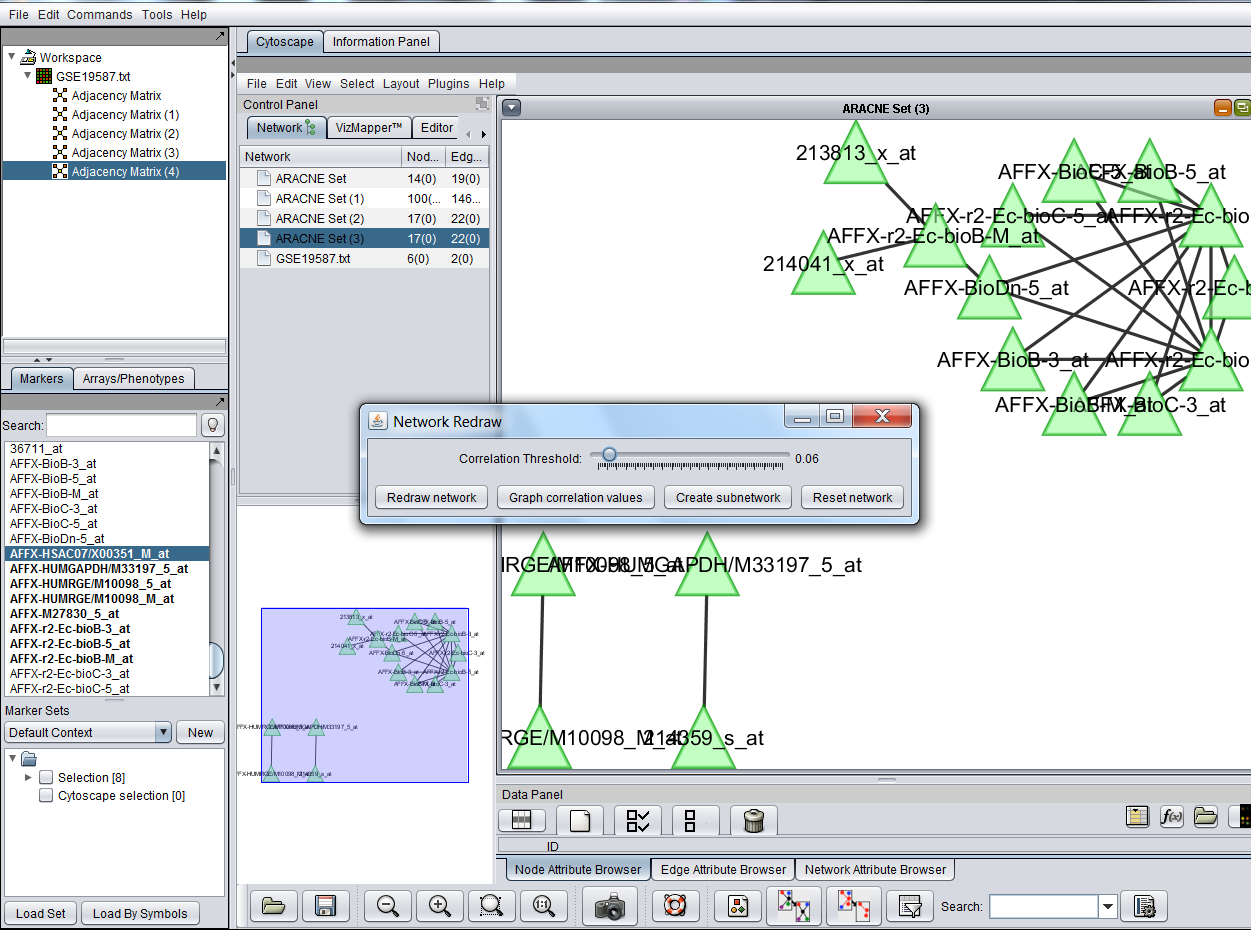
9.1 To see a graph of the correlations, right click and select compute edge correlations



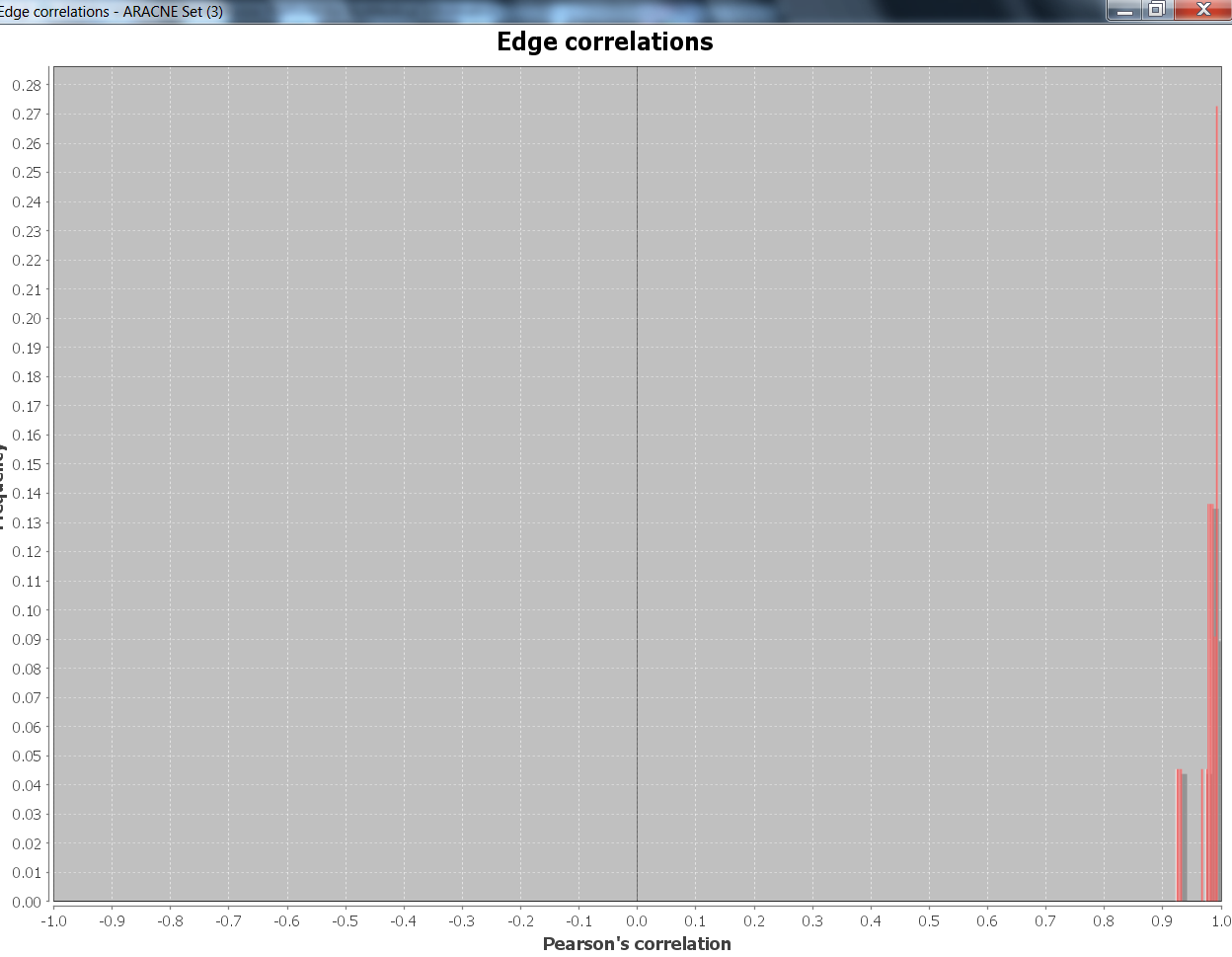
9.1.2 Select the correlation threshold which ranges from 0 to 1.0

1. For this particular one, the threshold is 0.06

2. This set has 17 nodes and 22 edges



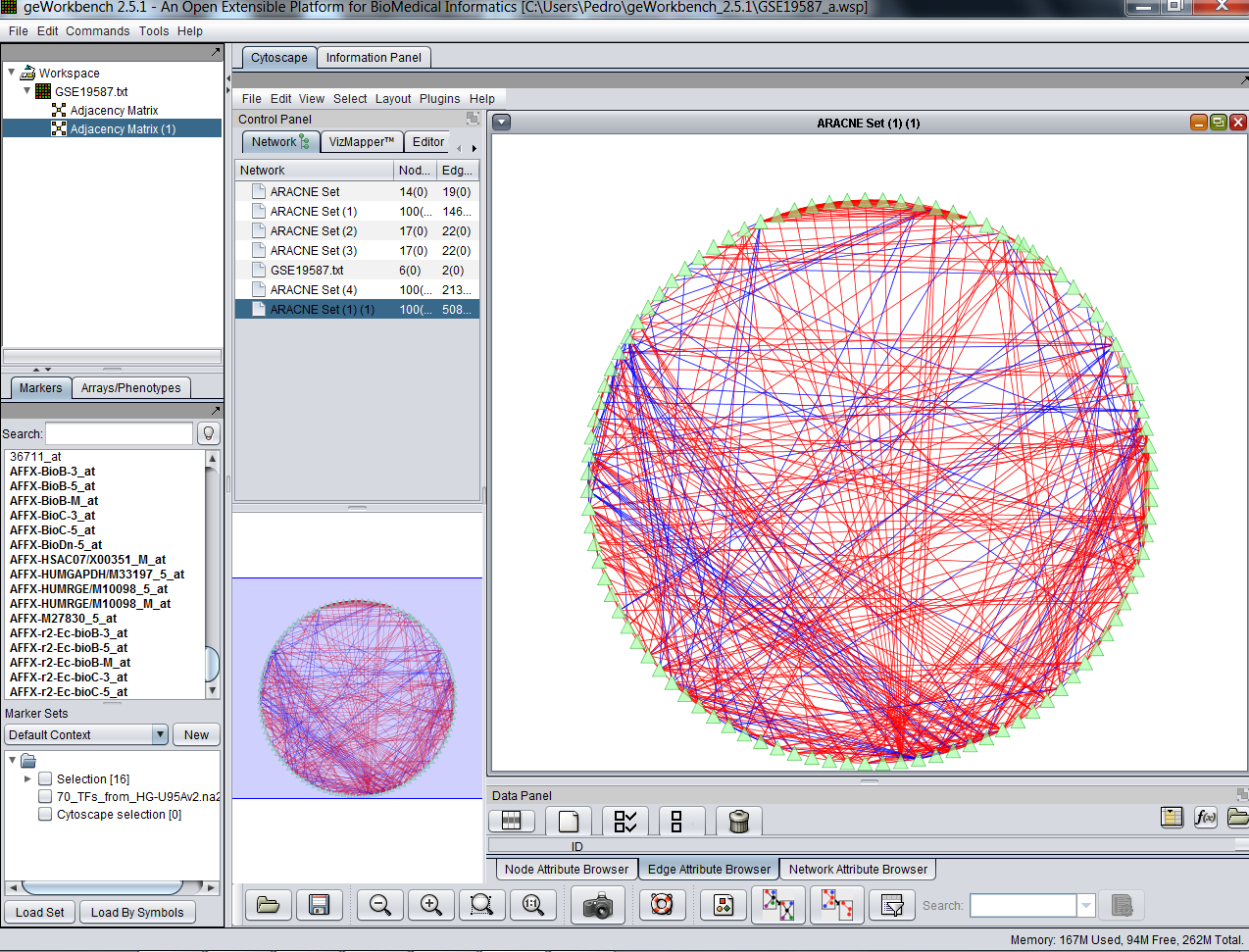
9.2 The following edge correlations histogram is shown: Correlation threshold is 0.05



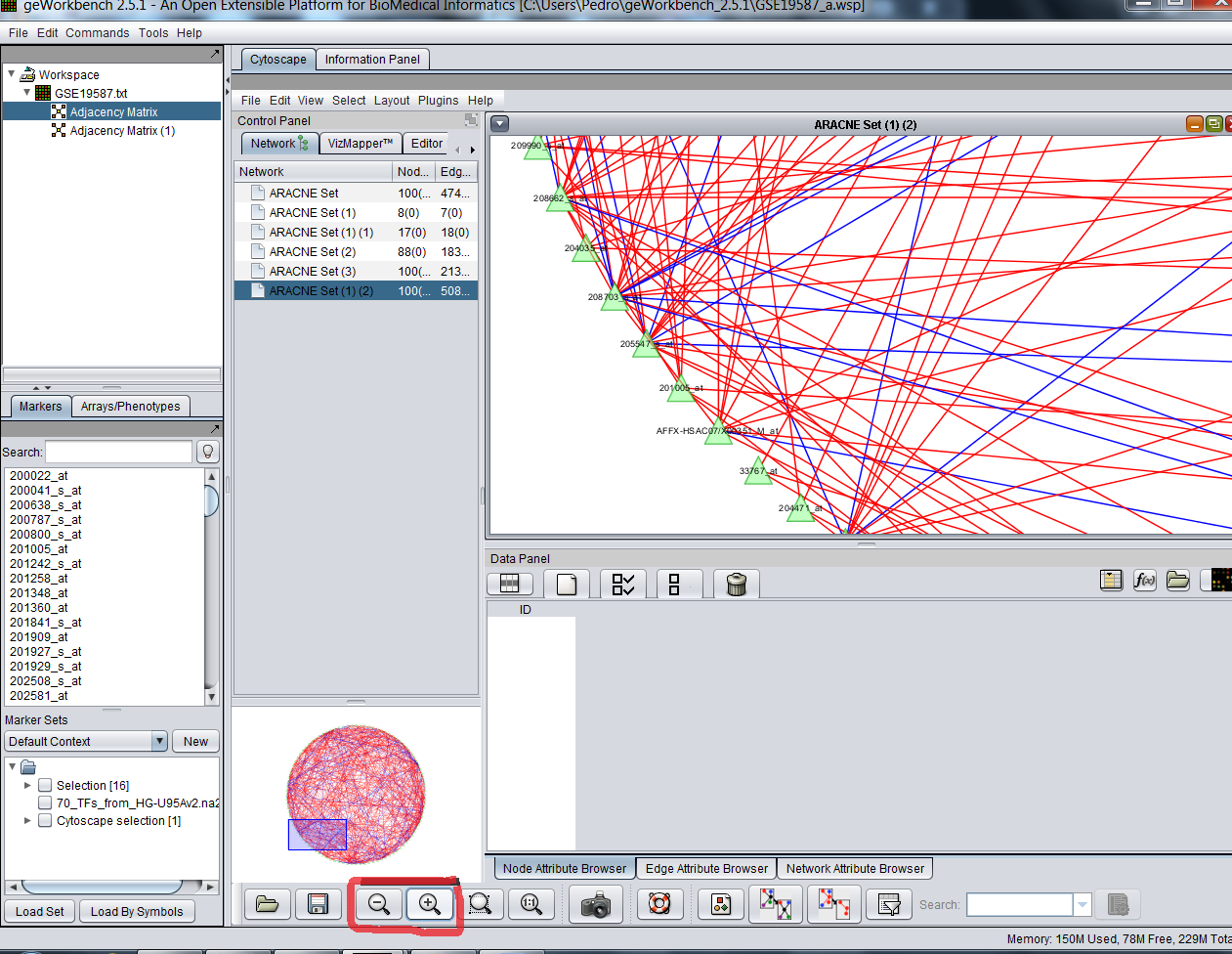
This set has 100 nodes with 508 edges

1. The Mode selected was Discovery 🡪 Algorithm: Adaptive Partitioning

2. The threshold is 0.179 🡪 P value is 0.001

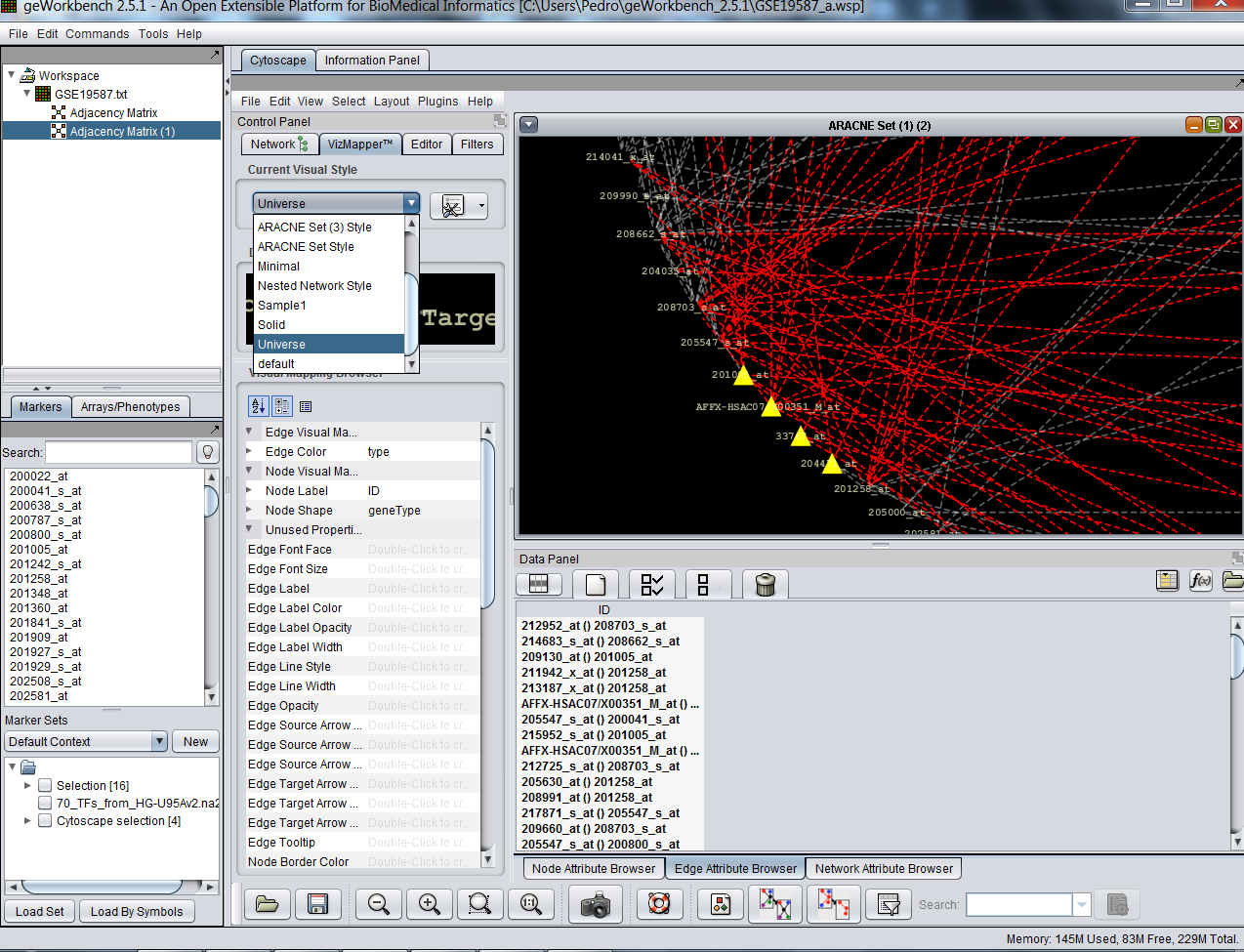


Use the zoom in and out icon to get closer to the image.

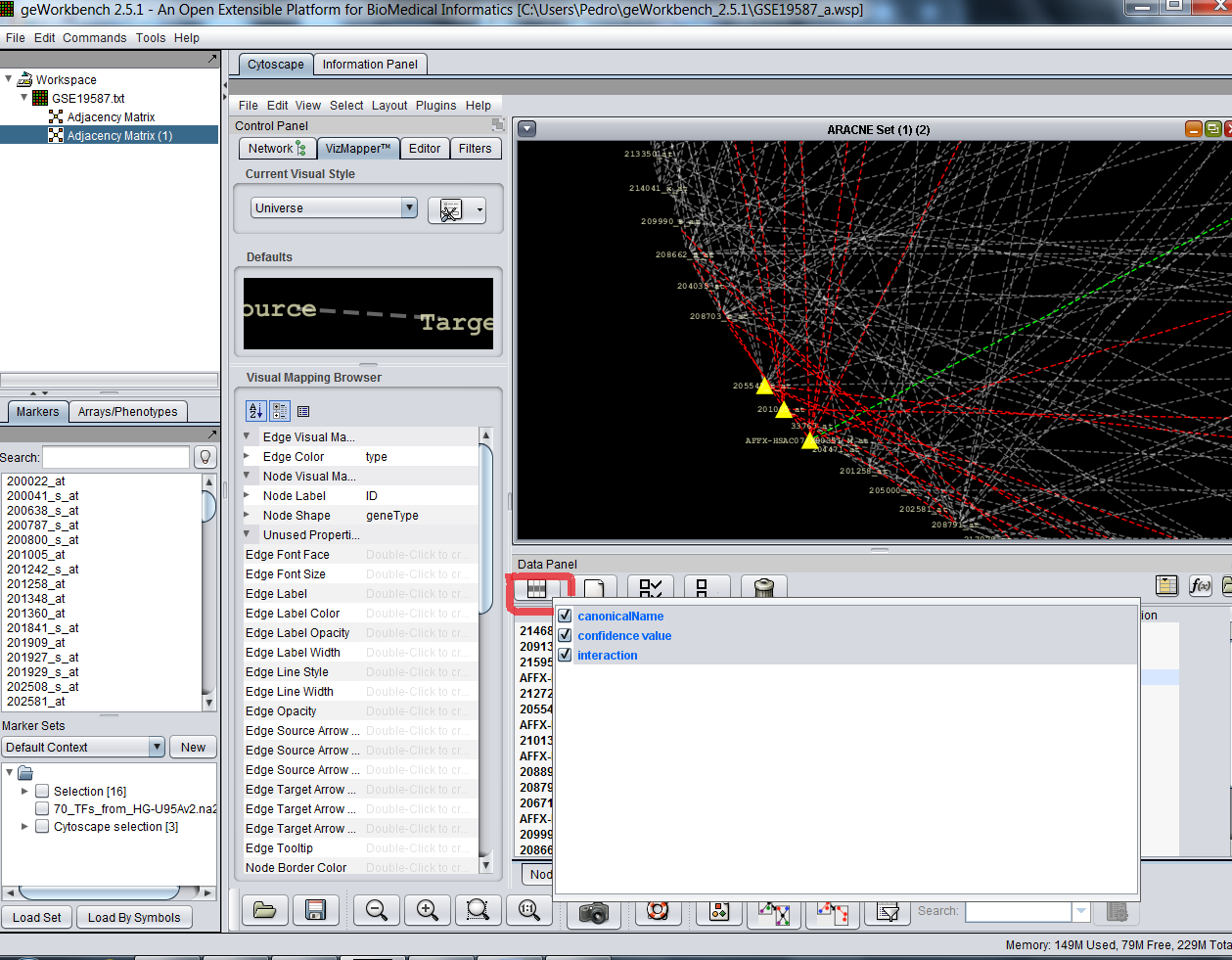


Note: The edges are colored blue or red according to their positive or negative correlations.

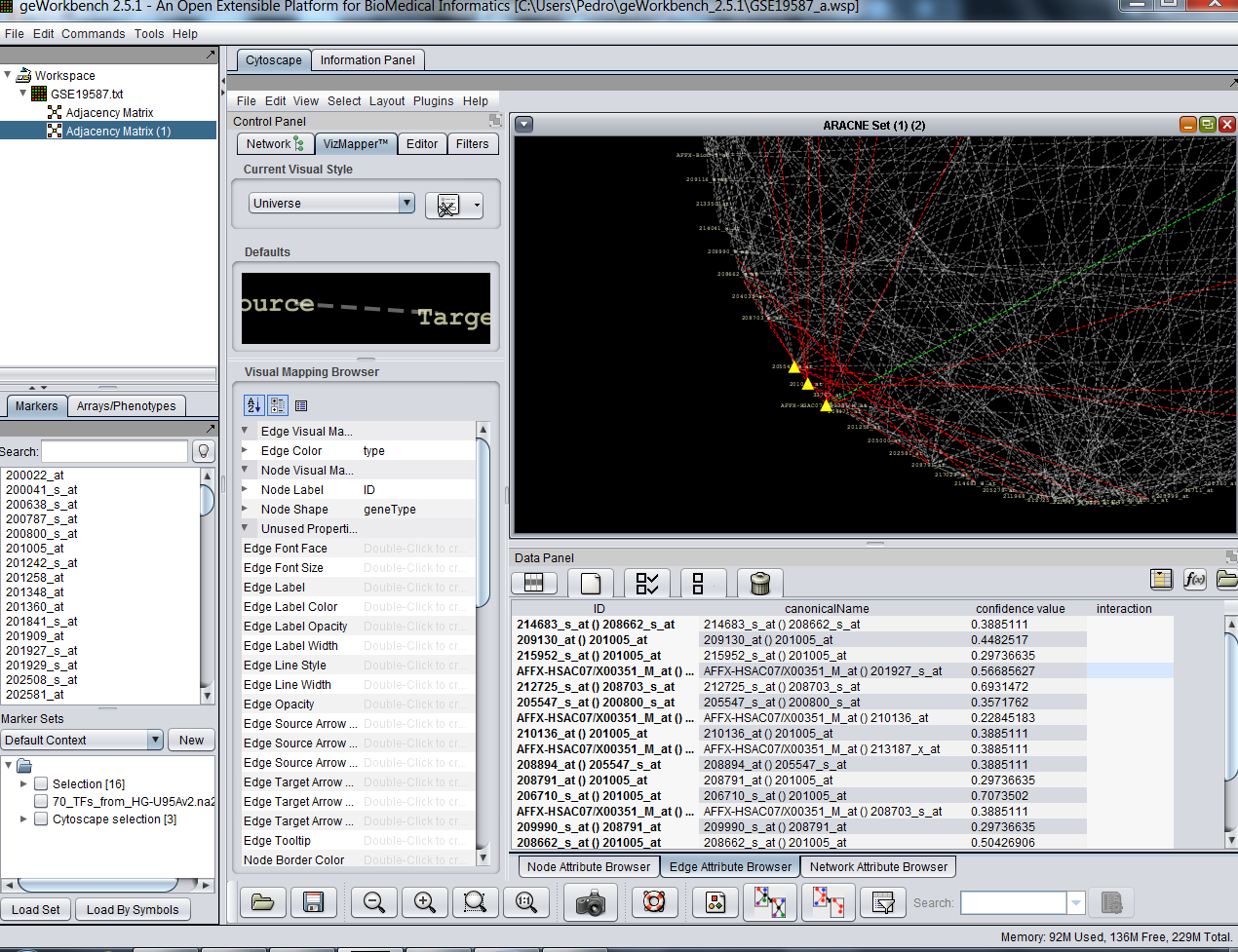
9.3 The figure below shows the result of applying the “Universe” style to a network where four nodes (yellow) are selected.



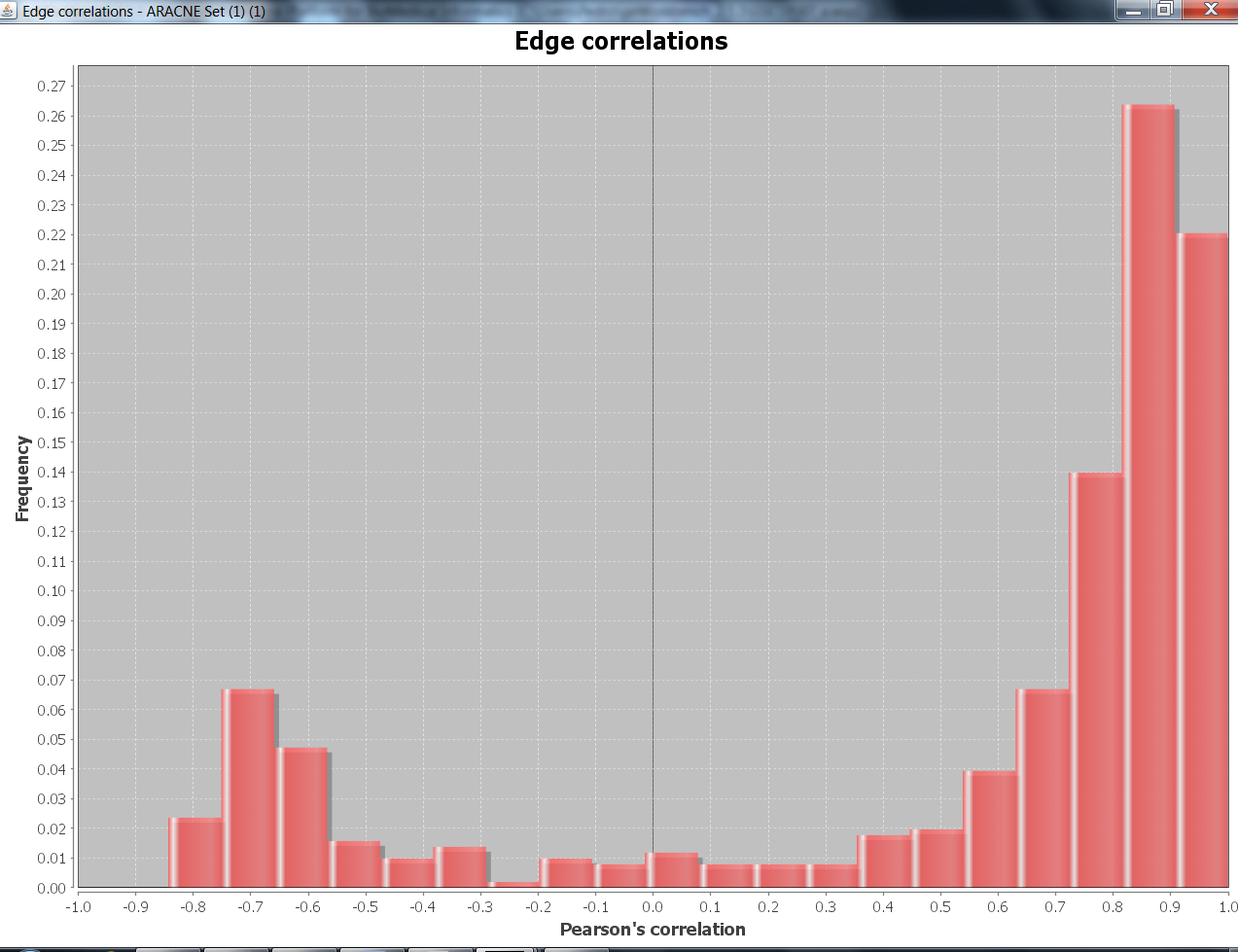
Click on the “select attributes” and mark the square boxes to view the confidence value and interaction between genes



The color green represent interactions between genes



The following edge correlations histogram is shown: Correlation threshold is 1.0



10. References

1. Lingeman, Jesse M. and Shasha, Dennis. (2013) *Network Inference in Molecular Biology: A Hands-on Framework*, California: Springer. Print.

2.Maetschke Stefan R., Madhamshettiwar Piyush B., Davis Melissa J., Ragan Mark A. (2013) *Supervised, semi supervised and*

*unsupervised inference of generegulatory networks,* BRIEFINGS IN BIOINFORMATICS. 1:17, doi:10.1093/bib/bbt034.