Standard Procedure For Analysis

Project Name: Transcriptomics Analysis

Requestor:

Names and contact info of Analyst:

Brief Description of Analysis Types: mRNA microarray data

Procedure or Analysis Revision Date:

Analysis Results/Deliverable Date:

Scientific context or hypothesis for analysis: (Varies)

Systems approach to identify biomarkers

Checklist before starting analysis:

- Scanned mRNA chips
- Experiment design
- Sample annotation
- Upload the raw data to Sysbiocube
- Genespring, R or matlab (Almost R)

Sources of input required:

- Experiment design (Control, Disease, Sampling time point),
- Sample annotation,
- Species and human analogue genes
- Tissue and other supporting information

Input data:

- Agilent Two Color microarray .txt file format (Raw data, Pipeline Flow Fig 1)
- Downloaded from Sysbiocube
- File Name convention (Standard for Raw data)
    - For Example:
        - Std_instrument_tag_species_tissues_group_datatype.txt
          e.g.
          20110825_mus_ heart _Balb_raw_10d1d.txt (if mouse)
          20110825_homo_heart_all_raw.txt (if human)

Analysis plan or steps taken:

1. Quality Control Steps:

Before and after normalization of the data (this step can be accomplished using

- Quality control on microarray chips (check RIN number, 260/230 and 260/280 ratios)
- Generate reports using arrayQCReport or arrayQualityMatrix R bioconductor packages
- Histogram and Intensity Distribution Boxplot
- RNA degradation plot
- MVA plot

2. Preprocess the microarray data

A. Lowess normalization by Genespring FE (feature extraction) 10.x version
B. Import Control type, probe name, signal channels and feature columns
C. Flag the data (detected, not detected, compromised)
    a. If feature is not positive or not significant (not detected)
    b. Not uniform (compromised)
    c. Not above background (not detected)
    d. Saturated or population outlier (compromised)
    e. Otherwise (detected)
D. Ratio computation, log transformation
E. Quantile normalization
F. Missing data imputation (k-nearest neighbor algorithm, often choose 9-11 neighbors)

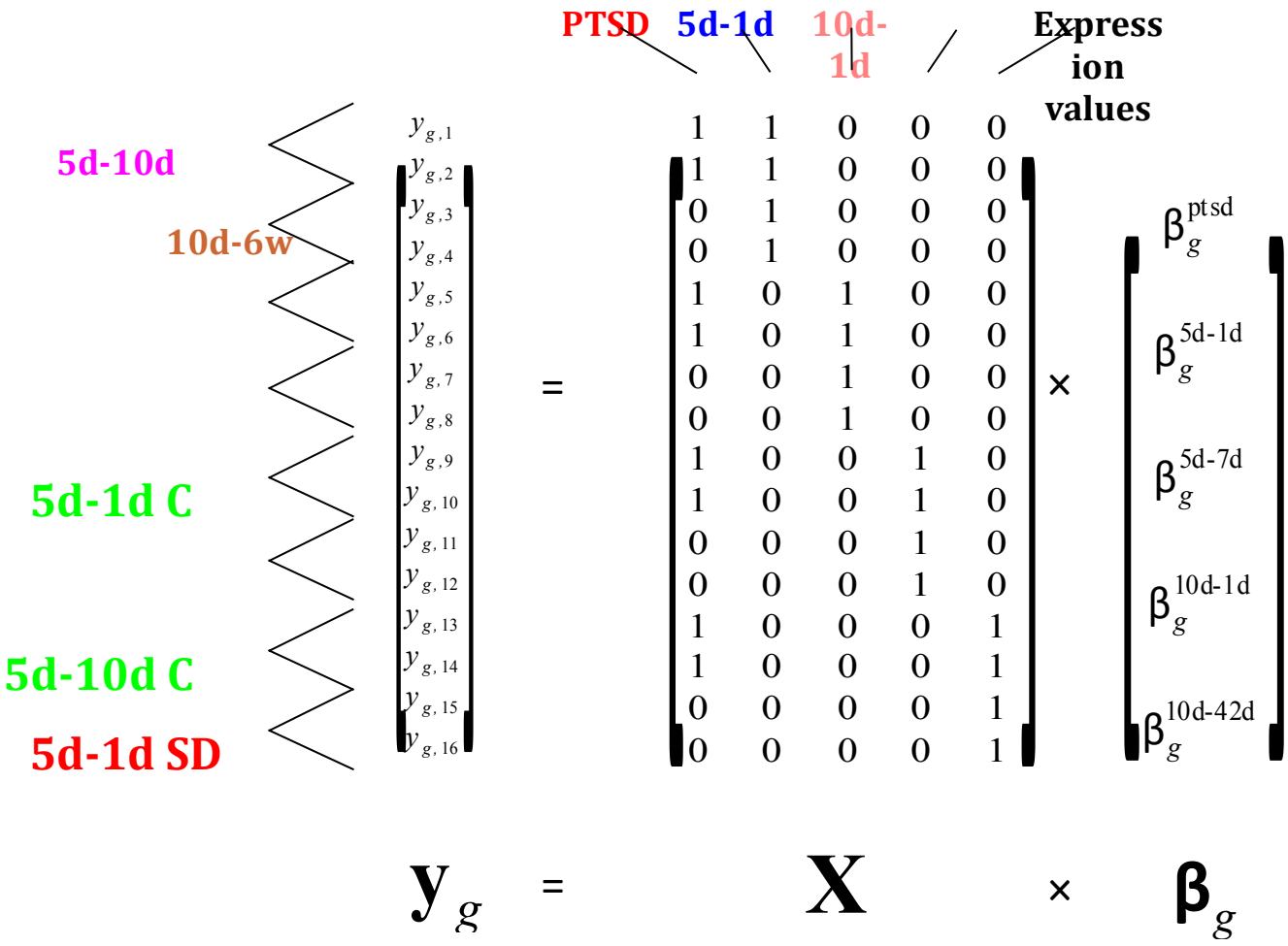
In recent applications including the pipeline, we tended to use LIMMA package for preprocessing. There are three steps in handling the feature extraction files:

1. Background correction (optional) typically normexp.
2. Loess normalization
3. Quantile normalization (optional) if the distribution is non-uniform

Then we check the batch effects using PCA or SWAMP or heatmap, and use COMBAT if the effects are known and not highly correlated to independent variables, otherwise use SVA.

3. DEG expression analysis

A. Unpaired t-test, unequal-variance, significance level 0.05
B. Bioconductor Limma package (The coefficient matrix for multi-segment data is recommended as follows:



A. Permutation test (by default 20,000 random permutes)
B. False discovery rate (Benjamini and Hochberg method, Storey's Q-value

4. **Hierarchical clustering** (by matlab or R, use Euclidian distance, green/red heatmap range: [-3, 3])
   - Time-series clustering:
   - STEM (shape-based clustering): clusters, pathway/GO enrichment in each cluster
   - FACT (feature-based clustering): clusters, pathway/GO enrichment in each cluster, pathway expression/GO term dynamics; Gene response time/dynamics statistics; Within/between pathway analysis;

5. **Pathway/GO term enrichment**
   Over representation analysis (hypergeometric test) by R or Matlab based on MSigDB database v3.1 or DAVID, Cluego in Cytoscape

6. **Classification**
   A. Support vector machine (Linear, nonlinear (not often used)) with Recursive Feature Elimination (RFE)

B. Regulated Linear Discriminant Analysis with RFE
C. Nearest shrunken centroid

7. COMBINER

7. 1. Single tissue:

a. Identify overlaps of DEGs in three aforementioned DEG analysis methods
b. Extract tissue/functional specific DEGs
c. Consensus feature elimination:

i. Start from 100 top DEGs (ranked by average p-values in the three methods)

ii. Run 250 groups of 500 classifiers in parallel using LDA with RFE

iii. Remove the bottom features until the cutoff criteria (Max AUC or a specific AUC value) is reached

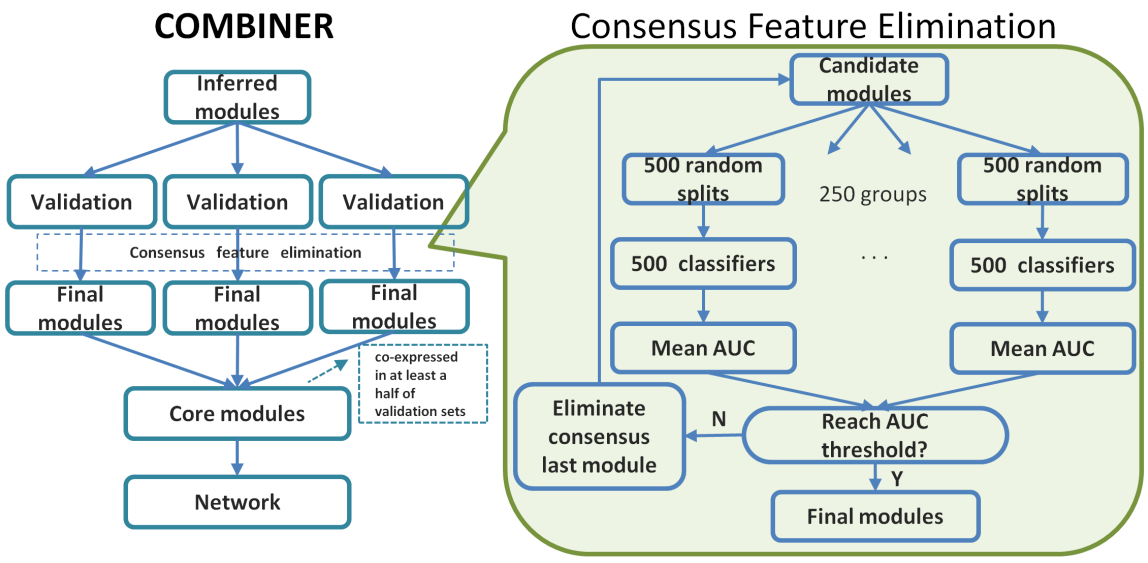
7. 2. Multiple tissues:

i. CORG pathway inference from inference dataset, so that each pathway becomes a vector called "pathway activity" (PA). The pathway database is taken from MsigDB v3.1.

ii. Regenerate PA in validation datasets, and run consensus feature elimination

iii. Conserve core modules: modules co-expressed in at least a half of validation sets

iv. Connect components of core modules based on protein-protein interaction from STRING v9.0



7. Upload normalized data, analysis results to sysbiocube

Statistical Analysis Plan or procedure:

Parameters Used:

Generation of research questions:

- What are the candidate gene/pathway/GO term biomarkers?
- What are the common/different patterns in multiple tissues?
- Does any subtypes exist in the subjects?

Software Used:

- Gene spring,
- Bioconductor R packages,
- Matlab

Software/program/script developed:

- Matlab toolbox (COMBINER, FACT)

Databases and public data sources:

- Such as Kegg pathways, DAVID, biocarta, HMDB or etc..

Data Disposal:

- This will include where analysis files, results are saved at common location at Sysbiocube (upload)
- File names including intermediate files
- File Name convention (Standard for Analyzed data, Pipeline Flow, Fig 2)
    - For Example:
    - Study_species_tissues_group_datatype.txt
      e.g.
      ptsd_mus_heart_all_analy_10d24h_moderated_ttest_p_0.05 _2783_probes.txt (if mouse)
      ptsd_homo_ heart _all_analyzed_parmeteres.txt (if human)

Short Description of results or finding:

- Lists of biomarkers, network figures, AUC figures

Publications and references:

Analysis Tasks performed:

- Analysis steps performed by analyst (Name and Task)

Appendix:

Script/Code:

#Random Forest

```
library(randomForest)

dataFrame<-read.table("fileName.csv", sep=",", header=TRUE, row.names=1)
str(dataFrame)
xm<- dataFrame[,1:n]
dim(xm)
ym<-as.factor(dataFrame[,n+1])
group<-c(rep('N',number of negatives),  rep('P', number of positives))
set.seed(number)
mtry=number
print(date())
rf<-randomForest(xm, ym=as.factor(group), ntree=10000 or any reasonable
number)
imp.temp <- abs(rf$importance[,])
t <- order(imp.temp,decreasing=TRUE)
plot(c(1:ncol(xm)),imp.temp[t],log='x',cex.main=1.5, xlab='Gene
rank',ylab='title',cex.lab=1.5, pch=16,main='number of probes')
gn.imp <- names(imp.temp)[t]
gn.25 <- gn.imp[1:25] # vector of top 25 genes, in order
t <- is.element(colnames(xm),gn.25)
sig.gn <- xm[,t]

write.table(sig.gn, "address/fileNameOutPut.txt", sep="\t")

varImpPlot(rf, n.var=25, main='Top 25 probes')
```

#NSC:

```
library(pamr)
datalist <- list(x=Data, y=Class, genenames=rownames(Data),
geneid=rownames(Data), samplelabels=colnames(Data), batchlabels=NULL)
train <- pamr.train(datalist) result <- pamr.cv(train, datalist,
folds=as.list(seq(ncol(Data))))
pamr.plotcv(result)
thresh <- max(result$threshold[result$error == min(result$error)]) genes
<- pamr.listgenes(train, datalist, threshold=thresh)
```

#DEG permutation:

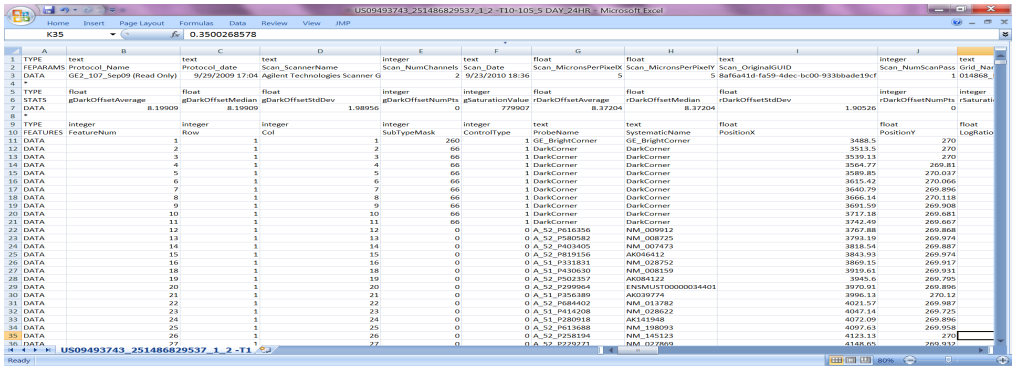
```
library(multtest)
result <- mt.maxT(Data, Class, test="t", side="abs",
fixed.seed.sampling="y", B=1e7) p.values <-
result$rawp[order(result$index)] fwer <-
result$adjp[order(result$index)] fdr <- p.adjust(p.values,
method="fdr")
Library(limma)
result <- eBayes(lmFit(Data, design))
p.values <- result$p.value[,1]
```

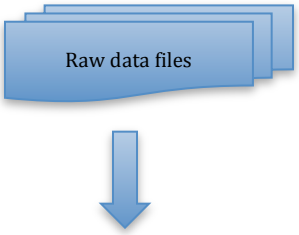
```
T-test:  [p,t]=mattest(Data(:,Class==1), Data(:,Class==0));
Permutation: [p,t]=mattest(Data(:,Class==1), Data(:,Class==0),
'permute', 20000);
LDA:
Class_est=Classify(Data_test,Data_training,Class_training);
SVM:
Svmstruct=svmtrain(Data_training,Class,'kernel_function','linear',
'method','SMO');
Class_est=svmclassify(svmstruct,Data_test);
```

#Matlab command:

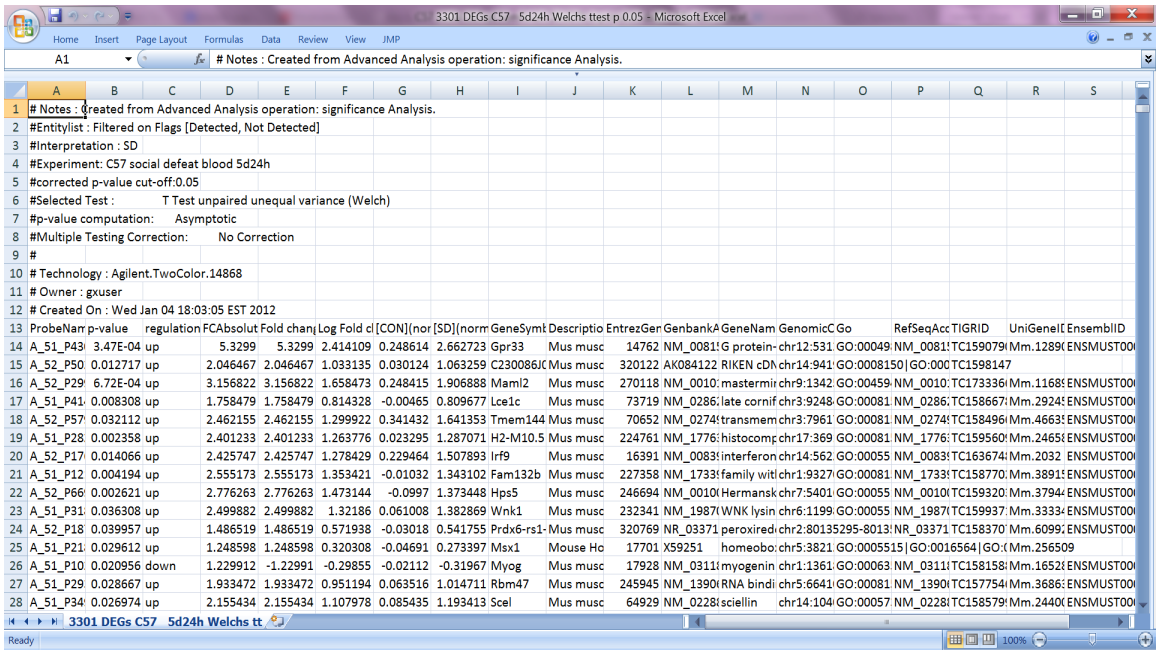
Pipeline Flow:

Raw Data (Agilent) Fig 1

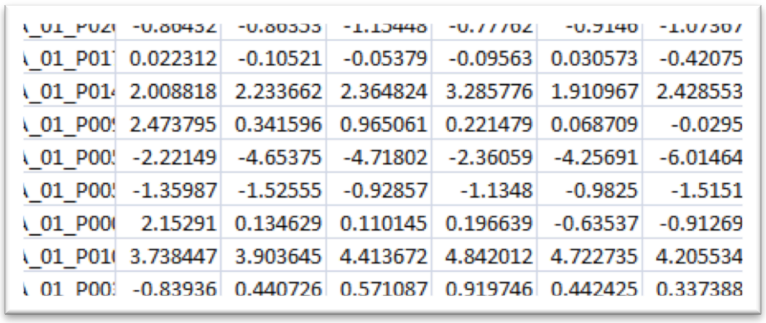




Filtered Data (GeneSpring Output) (Fig 2)



Normalized Data (Quantile Normalization) (Fig 3)



_01_P02	-0.80432	-0.80355	-1.15448	-0.77762	-0.9146	-1.07367
_01_P01	0.022312	-0.10521	-0.05379	-0.09563	0.030573	-0.42075
_01_P01	2.008818	2.233662	2.364824	3.285776	1.910967	2.428553
_01_P00	2.473795	0.341596	0.965061	0.221479	0.068709	-0.0295
_01_P00	-2.22149	-4.65375	-4.71802	-2.36059	-4.25691	-6.01464
_01_P00	-1.35987	-1.52555	-0.92857	-1.1348	-0.9825	-1.5151
_01_P00	2.15291	0.134629	0.110145	0.196639	-0.63537	-0.91269
_01_P01	3.738447	3.903645	4.413672	4.842012	4.722735	4.205534
_01_P00	-0.83936	0.440726	0.571087	0.919746	0.442425	0.337388

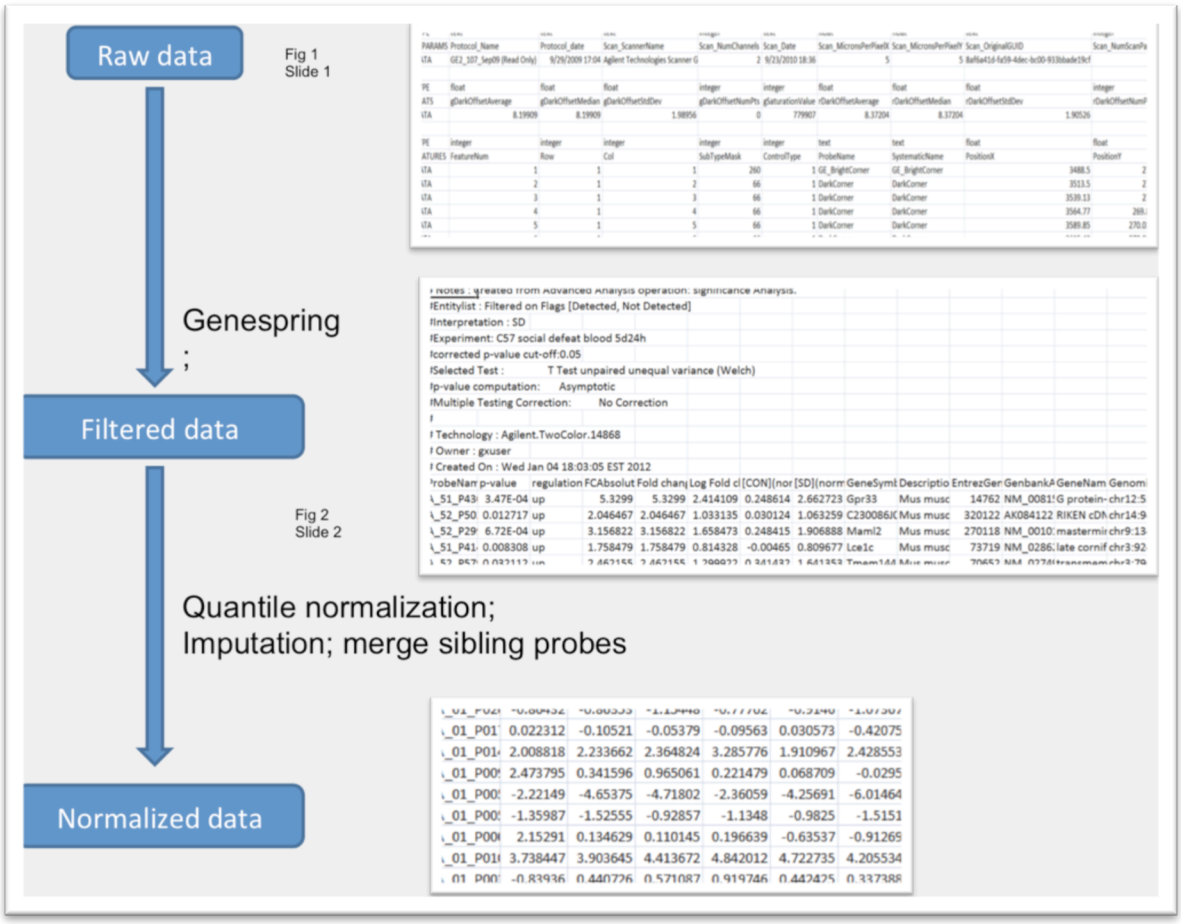
Output after R limma moderated t-test:

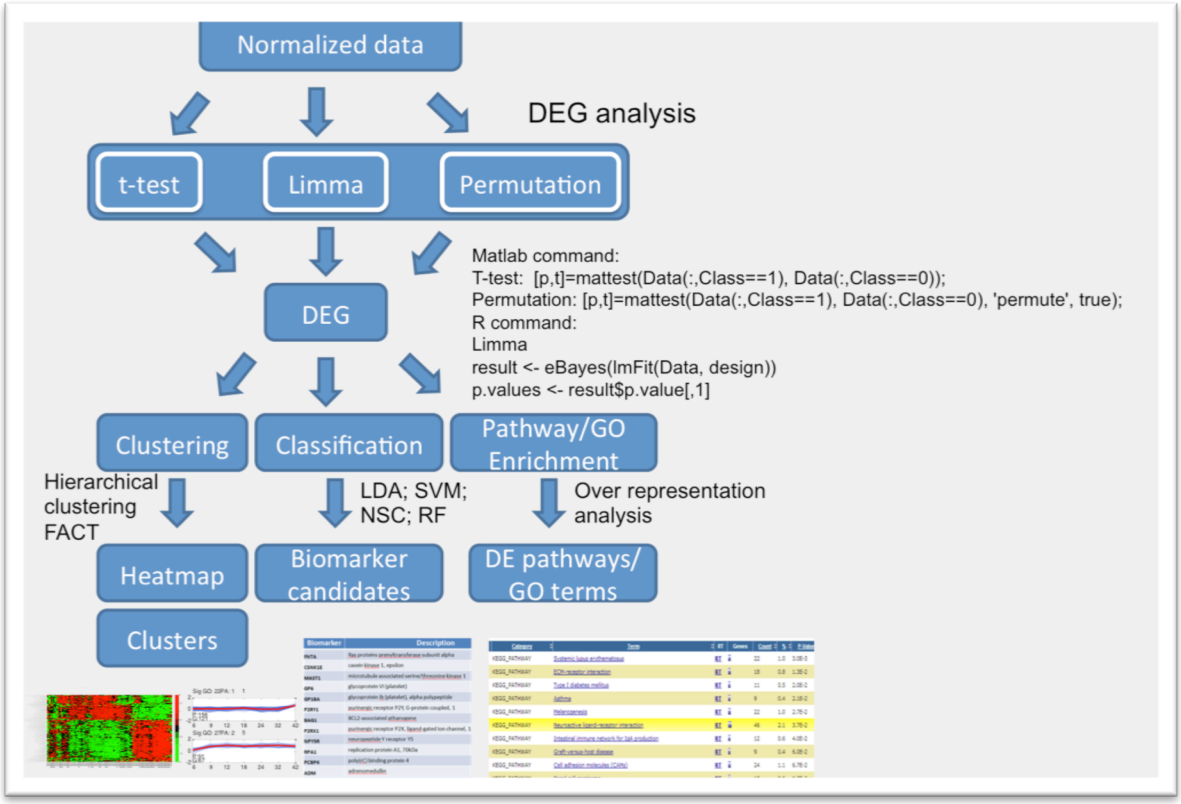
	A	B	C	D	E	F	G	H
1	Entrez	Probe	Symbol	Name	logFC	p.value	fwer	fdr
2	79750	A_23_P1137	ZNF385D	zinc finger protein 385D	-1.462143467	7.00E-07	0.0014927	0.0075848
3	55111	A_23_P2787	PLEKHJ1	pleckstrin homology domain containing, family J member 1	0.216616858	8.00E-07	0.0078026	0.0075848
4	4211	A_24_P3197	MEIS1	Meis homeobox 1	-0.785598753	2.50E-06	0.0096072	0.01441112
5	201625	A_23_P3724	DNAH12	dynein, axonemal, heavy chain 12	-1.144763963	3.50E-06	0.0246376	0.01441112
6	286006	A_24_P3806	C7orf53	chromosome 7 open reading frame 53	-1.274014819	3.80E-06	0.0196622	0.01441112
7	55363	A_23_P4341	HEMGN	hemogen	-1.097886913	7.80E-06	0.0387915	0.019673075
8	2888	A_23_P1545	GRB14	growth factor receptor-bound protein 14	-1.594655812	8.20E-06	0.0303257	0.019673075
9	51471	A_24_P3085	NAT8B	N-acetyltransferase 8B (GCN5-related, putative, gene/pseudogene)	-1.404884907	8.30E-06	0.0410703	0.019673075
10	200879	A_23_P8421	LIPH	lipase, member H	-0.97749106	9.90E-06	0.0419689	0.0208582
11	284207	A_23_P1059	METRNL	meteorin, glial cell differentiation regulator-like	0.370903498	1.56E-05	0.0693556	0.029477291
12	7292	A_23_P1268	TNFSF4	tumor necrosis factor (ligand) superfamily, member 4	-1.643836743	1.71E-05	0.0576478	0.029477291

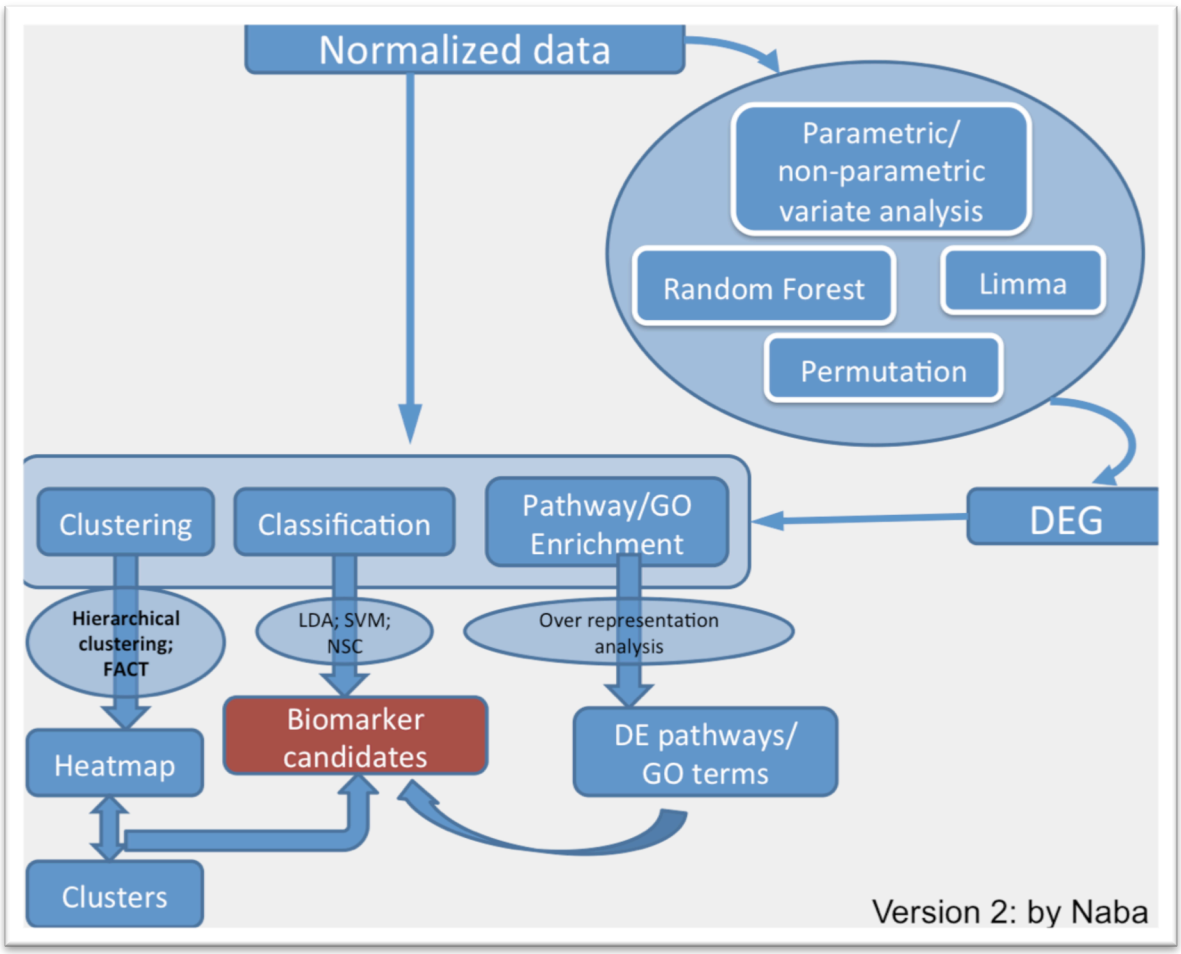
Output after R multtest permutation t-test

	A	B	C	D	E	F	G
1	Entrez	Probe	Symbol	Name	logFC	p.value	fdr
2	79750	A_23_P1137	ZNF385D	zinc finger protein 385D	-1.462143467	1.56E-07	0.0029507
3	4211	A_24_P3197	MEIS1	Meis homeobox 1	-0.785598753	1.43E-06	0.0135335
4	286006	A_24_P3806	C7orf53	chromosome 7 open reading frame 53	-1.274014819	2.70E-06	0.0160258
5	201625	A_23_P3724	DNAH12	dynein, axonemal, heavy chain 12	-1.144763963	3.56E-06	0.0160258
6	2888	A_23_P1545	GRB14	growth factor receptor-bound protein 14	-1.594655812	4.29E-06	0.0160258

Pipeline Flow:




Normalized data
DEG analysis
t-test
Limma
Permutation
Matlab command:
T-test: [p,t]=mattest(Data(:,Class==1), Data(:,Class==0));
Permutation: [p,t]=mattest(Data(:,Class==1), Data(:,Class==0), 'permute', true);
R command:
Limma
result <- eBayes(lmFit(Data, design))
p.values <- result$p.value[,1]
DEG
Clustering
Classification
Pathway/GO Enrichment
Hierarchical clustering FACT
LDA; SVM; NSC; RF
Over representation analysis
Heatmap
Biomarker candidates
DE pathways/ GO terms
Clusters


Normalized data
Parametric/
non-parametric
variate analysis
Random Forest
Limma
Permutation
DEG
Clustering
Classification
Pathway/GO
Enrichment
Hierarchical
clustering;
FACT
LDA; SVM;
NSC
Over representation
analysis
Heatmap
Biomarker
candidates
DE pathways/
GO terms
Clusters
Version 2: by Naba

Normalized data

COMBINER

Pathway inference; Consensus feature elimination

List=combiner(Class1,Class2,expression1,expression2,methodflag)

Protein-protein interaction from STRING

PPI.xls

Core Modules

updown=checkregdirection(List, dataseries);
colist=[~,colist]=xlsread('PPI.xls',1,'A2:B500');
drawmodules(List,updown,colist)

Core Network