

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:

$y_g = X \times \beta_g$

PTSD 5d-1d 10d-1d Expression values

$y_{g,1}$ 1 1 0 0 0 β_g^{ptsd}
 $y_{g,2}$ 1 1 0 0 0 β_g^{5d-1d}
 $y_{g,3}$ 0 1 0 0 0 β_g^{5d-7d}
 $y_{g,4}$ 0 1 0 0 0 β_g^{10d-1d}
 $y_{g,5}$ 1 0 1 0 0 $\beta_g^{10d-42d}$
 $y_{g,6}$ 1 0 1 0 0
 $y_{g,7}$ 0 0 1 0 0
 $y_{g,8}$ 0 0 1 0 0
 $y_{g,9}$ 1 0 0 1 0
 $y_{g,10}$ 1 0 0 1 0
 $y_{g,11}$ 0 0 0 1 0
 $y_{g,12}$ 0 0 0 1 0
 $y_{g,13}$ 1 0 0 0 1
 $y_{g,14}$ 1 0 0 0 1
 $y_{g,15}$ 0 0 0 0 1
 $y_{g,16}$ 0 0 0 0 1

- 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

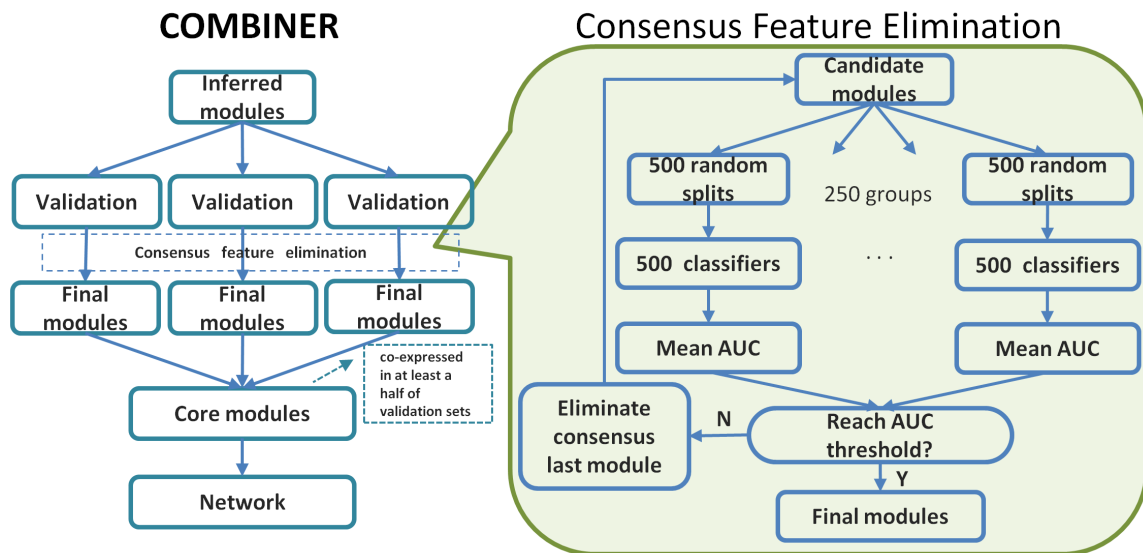
7.2. Multiple tissues:

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]
```

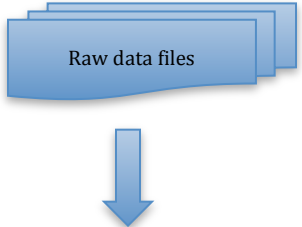
```
T-test: [p,t]=mattest(Data(:,Class==1), Data(:,Class==0));
Permutation: [p,t]=mattest(Data(:,Class==1), Data(:,Class==0),
'permute', 20000);
LDA:
Class_est=Classify(Data_test,Data_training,Class_training);
SVM:
Svmstruct=svmtrain(Data_training,Class,'kernel_function','linear',
'method','SMO');
Class_est=svmclassify(svmstruct,Data_test);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

#Matlab command:

Pipeline Flow:

Raw Data (Agilent) Fig 1

TYPE	Protocol_Name	Protocol_date	Scan_ScannerName	Scan_NumChannels	Scan_Date	Scan_MicronsPerPixel	Scan_MicronsPerPixel	Scan_OriginalGUID	Scan_NumScanPass	Grid_Nbr
1	DATA	5/29/2009 17:04	Agilent Technologies Scanner	2	5/23/2010 18:36			5	5	5
2	DATA	8.19009	8.19009	1.98950	779007	8.37204			1.00526	0
3	DATA	1	1	1	260	1	0E_BrightCorner	3488.5	270	3488.5
4	DATA	2	1	2	66	1	DarkCorner	3533.5	270	3533.5
5	DATA	3	1	3	66	1	DarkCorner	3339.13	270	3339.13
6	DATA	4	1	4	66	1	DarkCorner	3564.77	269.81	3564.77
7	DATA	5	1	5	66	1	DarkCorner	3389.85	270.017	3389.85
8	DATA	6	1	6	66	1	DarkCorner	3615.42	270.066	3615.42
9	DATA	7	1	7	66	1	DarkCorner	3650.79	269.896	3650.79
10	DATA	8	1	8	66	1	DarkCorner	3666.14	270.118	3666.14
11	DATA	9	1	9	66	1	DarkCorner	3691.59	269.908	3691.59
12	DATA	10	1	10	66	1	DarkCorner	3717.18	269.681	3717.18
13	DATA	11	1	11	66	1	DarkCorner	3742.49	269.667	3742.49
14	DATA	12	1	12	0	0	A_52_P010356	NM_009912	3767.88	269.808
15	DATA	13	1	13	0	0	A_52_P000882	NM_008725	3793.19	269.974
16	DATA	14	1	14	0	0	A_52_P003405	NM_007473	3818.54	269.887
17	DATA	15	1	15	0	0	A_52_P019156	AK06412	3843.93	269.974
18	DATA	16	1	16	0	0	A_51_P031831	NM_028752	3869.15	269.917
19	DATA	18	1	18	0	0	A_51_P030630	NM_008159	3919.61	269.911
20	DATA	19	1	19	0	0	A_52_P002357	AK084122	3945.6	269.795
21	DATA	20	1	20	0	0	A_52_P009864	ENSMUST000034401	3970.91	269.896
22	DATA	21	1	21	0	0	A_51_P056389	AK039774	3996.13	270.12
23	DATA	22	1	22	0	0	A_52_P048402	NM_013782	4021.57	269.987
24	DATA	23	1	23	0	0	A_51_P014208	NM_028622	4047.14	269.725
25	DATA	24	1	24	0	0	A_52_P009018	AK11948	4072.99	269.896
26	DATA	25	1	25	0	0	A_52_P013688	NM_198993	4097.63	269.958
27	DATA	26	1	26	0	0	A_52_P008194	NM_145323	4123.13	270
28	DATA	27	1	27	0	0	A_52_P229271	NM_027060	4148.05	269.934



Filtered Data (GeneSpring Output) (Fig 2)

ProbeName	p-value	regulation	FC	Absolut	Fold	chan	Log	Fold	c	[CON]	[nor	[SD]	[norm	GeneSymt	Descriptio	EntrezGer	GenbankA	GeneNam	GenomicCGo	RefSeqAcc	TIGRID	UniGene	EnsemblID
A_51_P43	3.47E-04	up	5.3299	5.3299	2.414109	0.248614	2.662723	Gpr33	Mus musc	14762	NM_00811	G	protein	chr12:531	GO:00049	NM_00811	TC159079	Mm.12891	ENSMUST00				
A_52_P50	0.012717	up	2.046467	2.046467	1.033135	0.030124	1.063259	C230086	Mus musc	320122	AK084122	RIKEN cDN	chr14:941	GO:0008150	GO:000	TC1598147							
A_52_P29	6.72E-04	up	3.156822	3.156822	1.658473	0.248415	1.906888	Maml2	Mus musc	270118	NM_0010	mastermir	chr9:1342	GO:00459	NM_0010	TC173336	Mm.11681	ENSMUST00					
A_51_P41	0.008308	up	1.758479	1.758479	0.814328	-0.00465	0.809677	Lce1c	Mus musc	73719	NM_0286	late cornif	chr3:9248	GO:00081	NM_0286	TC158667	Mm.2924	ENSMUST00					
A_52_P57	0.032112	up	2.462155	2.462155	1.299922	0.341432	1.641353	Tmem144	Mus musc	70652	NM_0274	transmem	chr3:7961	GO:00081	NM_0274	TC158496	Mm.4663	ENSMUST00					
A_51_P28	0.002358	up	2.401233	2.401233	1.263776	0.023295	1.287071	H2-M10.5	Mus musc	224761	NM_1776	histocomp	chr17:369	GO:00081	NM_1776	TC159560	Mm.24651	ENSMUST00					
A_52_P17	0.014066	up	2.425747	2.425747	1.278429	0.229464	1.507893	Irf9	Mus musc	16391	NM_0083	interferon	chr14:562	GO:00055	NM_0083	TC163674	Mm.2032	ENSMUST00					
A_51_P12	0.004194	up	2.555173	2.555173	1.353421	-0.01032	1.343102	Fam132b	Mus musc	227358	NM_1733	family wtl	chr1:9327	GO:00081	NM_1733	TC158770	Mm.3891	ENSMUST00					
A_52_P66	0.002621	up	2.776263	2.776263	1.473144	-0.0997	1.373448	Hps5	Mus musc	246694	NM_0010	Hermansh	chr7:5401	GO:00055	NM_0010	TC159320	Mm.3794	ENSMUST00					
A_51_P31	0.036308	up	2.499882	2.499882	1.32186	0.061008	1.382869	Wnk1	Mus musc	232341	NM_1987	WNK	lysin	chr6:1199	GO:00055	NM_1987	TC159937	Mm.3333	ENSMUST00				
A_52_P18	0.039957	up	1.486519	1.486519	0.571938	-0.03018	0.541755	Prdx6-rs1	Mus musc	320769	NR_03371	peroxidire	chr2:80135	295-8013	NR_03371	TC158370	Mm.6099	ENSMUST00					
A_51_P21	0.029612	up	1.248598	1.248598	0.320308	-0.04691	0.273397	Msx1	Mouse Ho	17701	X59251	homeobc	chr3:3821	GO:0005515	GO:0016564	GO:0016564	Mm.26509						
A_51_P10	0.020956	down	1.229912	-1.22991	-0.29855	-0.02112	-0.31967	Myog	Mus musc	17928	NM_0311	myogenin	chr1:1361	GO:00063	NM_0311	TC158158	Mm.16521	ENSMUST00					
A_51_P29	0.028667	up	1.933472	1.933472	0.951194	0.063516	1.014711	Rbm47	Mus musc	245945	NM_1390	RNA bindi	chr5:6641	GO:00081	NM_1390	TC157754	Mm.3686	ENSMUST00					
A_51_P34	0.026974	up	2.155434	2.155434	1.107978	0.085435	1.193413	Scel	Mus musc	64929	NM_0228	scellin	chr14:104	GO:00057	NM_0228	TC158579	Mm.2440	ENSMUST00					

Normalized Data (Quantile Normalization) (Fig 3)

_01_P02	-0.80452	-0.80555	-1.15448	-0.77702	-0.5140	-1.07507
_01_P01	0.022312	-0.10521	-0.05379	-0.09563	0.030573	-0.42075
_01_P04	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_P010	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:

