GS01 0163
Analysis of Microarray Data

Keith Baggerly and Bradley Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

4 November 2010

Lecture 19: Gene Set Enrichment Analysis

- Gene Set Enrichment Analysis
- TCGA Analysis Project

Night Sky

Go to a remote location (preferably in the Southern Hemisphere) late at night when the weather is clear and look up.

What do you see?

Gene Set Enrichment Analysis (GSEA)

Last week, we saw that we can use known information about gene functions and gene relationships to help understand the biology behind a list of differentially expressed genes:

- Derive a list of significantly differentially expressed genes, while controlling for false discovery,
- Determine pathways containing (many of) the genes concerned,
- Gain biological insight . . .

Will this algorithm find all significantly affected pathways?

Overview of GSEA

Detecting modest changes in gene expression datasets is hard, due to:

- the large number of variables,
- the high variability between samples, and
- the limited number of samples.

The goal of GSEA is to detect modest but coordinated changes in prespecified sets of related genes.

Such a set might include all the genes in a specific pathway, on a specific chromosome, or in a specific cytoband, for instance.

GSEA Publications

Mootha et al., Nature Genetics 34, 267–273 (2003)

Subramanian et al., PNAS 102(43), 15545–15550 (2005) (http://www.pnas.org/content/102/43/15545.abstract).

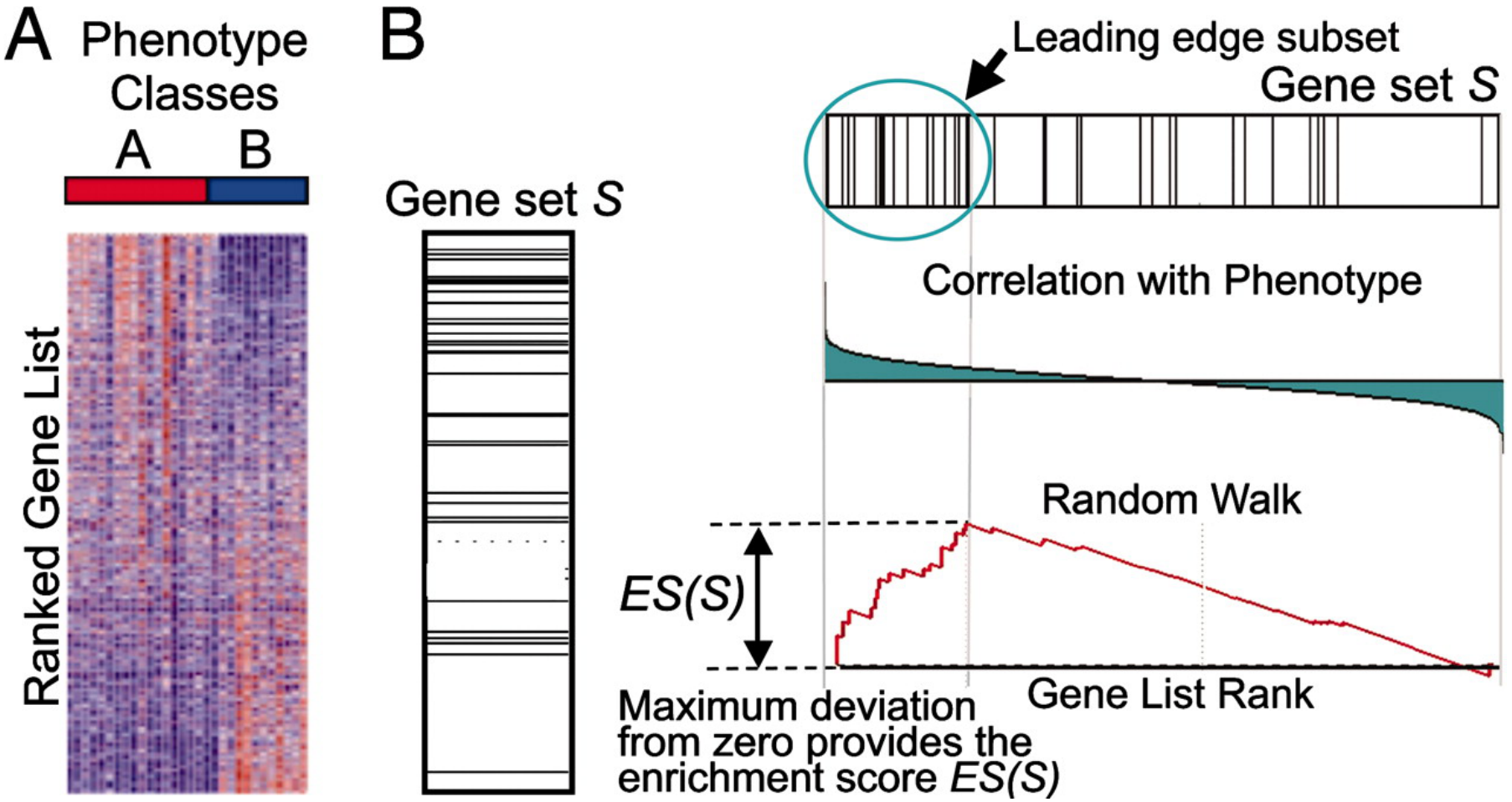
The description of GSEA changed between the two papers. We follow the second formulation.

GSEA Algorithm: Step 1

Calculate an Enrichment Score:

- Rank genes based on the correlation between their expression and the class distinction
- For an *a priori* set of genes S, compute the cumulative sum over ranked genes:
  - Increase sum when gene in S, decrease it otherwise.
  - Magnitude of increment depends on correlation of gene with phenotype.
- Record the maximum deviation from zero as the enrichment score

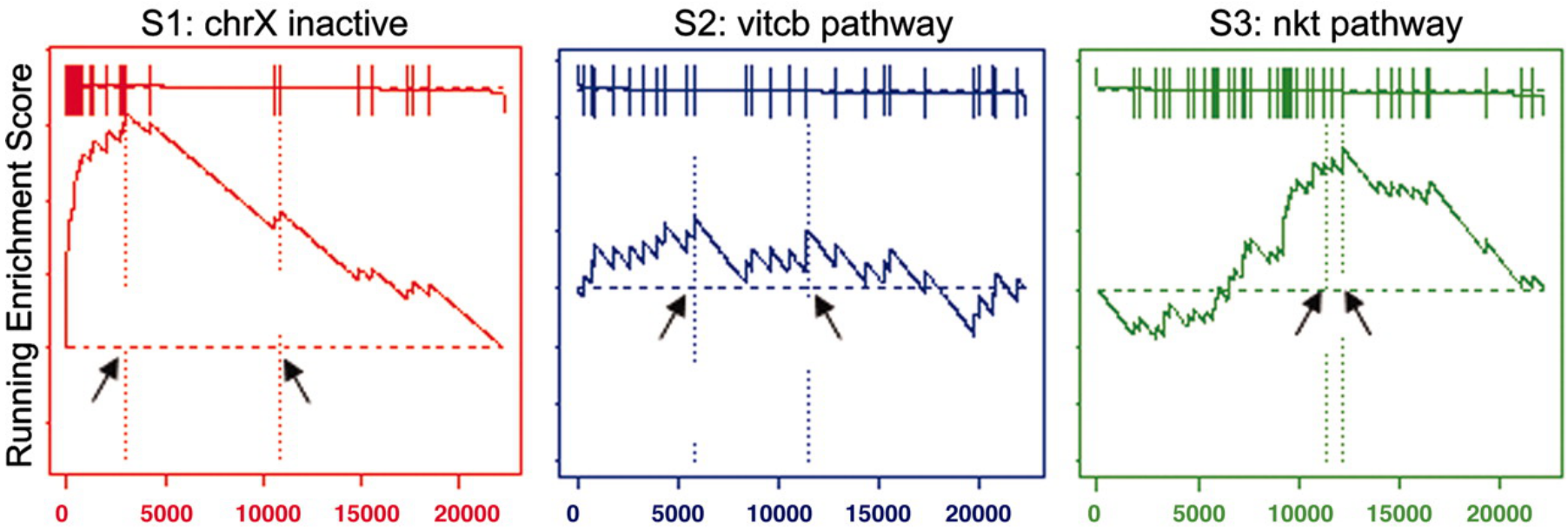
Schematic overview of GSEA



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA Algorithm: Step 1

Three examples:



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

Vertical lines indicate locations of maximum enrichment score and zero correlation between expression and phenotype.

GSEA Algorithm: Step 2

Assess significance (nominal P value):

- Permute phenotype labels 1000 times
- Compute ES score as above for each permutation
- Compare ES score for actual data to distribution of ES scores from permuted data

Permuting the phenotype labels instead of the genes maintains the complex correlation structure of the gene expression data.

GSEA Algorithm: Step 3

Adjustment for multiple hypothesis testing:

- Normalize the ES accounting for size of each gene set, yielding normalized enrichment score (NES)
- Control proportion of false positives by calculating FDR corresponding to each NES, by comparing tails of the observed and null distributions for the NES

GSEA Algorithm: Step 4

The original method used equal weights for each gene.

The revised method weighted genes according to their correlation with phenotype.

This may cause an asymmetric distribution of ES scores if there is a big difference in the number of genes highly correlated to each phenotype.

Consequently, the above algorithm is performed twice: one for the positively scoring gene sets and once for the negatively scoring gene sets.

GSEA Availability

GSEA is available from their website http://www.broadinstitute.org/gsea/ .

GSEA is available as both a Java program and an R script. Because we like scripts, we'll use the R version.

The same site provides several existing gene sets, including

- positional gene sets,
- curated gene sets,
- motif gene sets,
- computational gene sets,
- and gene ontology gene sets.

What you need

Download the R source code (`GSEA.1.0.R`), and the gene set databases (`.gmt` files).

To analyze experimental data, you will need to create two text files:

- an gene expression file (`.gct`), and
- a sample phenotype file (`.cls`).

Gene Set Database File

A geneset database (`.gmt`) file is a tab separated text file containing one geneset per line.

The first column is the gene set name.

The second column is a brief description of the gene set.

The remaining columns contain the names of the genes in the gene set. Notes:

- there are many trailing empty columns,
- there are many genes with names like `8-Sep`,
- suggesting whoever developed the “database” used Excel.

Gene Expression File

A gene expression (`.gct`) file is a tab-separated text file.

The first line is the constant `#1.2`.

The second line contains two numbers: the number of genes and the number of samples.

The third line contains column headers for the table on the following lines. The fourth line and below each contain expression values for one gene:

- the first column is the gene name,
- the second column is the gene description,
- the third and subsequent columns are the expression values for each sample.

Sample Phenotype File

A sample phenotype (`.cls`) file is a text file containing three lines.

The first line contains three numbers separated by spaces. The first number is the number of samples. The second and third numbers are the constants 2 and 1, respectively.

The second line begins with `#` and is followed by a space separated list of “long” phenotype names.

The third line consists of a space separated list of “short” phenotype labels for each of the samples in the gene expression file, in the same order they occur there.

- You’ll get nonsense if these two orders ever get out of sync.

Notes:

- The gene names used in the gene expression file must match those in the geneset database file.
- The order of the samples in the gene expression file must match the order of the phenotypes in the sample phenotype file.
- A gene expression file can be read into R as follows:

```
df <- read.delim("file.gct", header=TRUE,
                 sep="\t", skip=2,
                 blank.lines.skip=TRUE)
```

Running GSEA

Create an output directory to hold the results:

```
% mkdir GSEA_Gender_C1
```

Each run of GSEA processes one geneset database file, so it's sensible to include the geneset database name in the directory name.

Source the GSEA to define the GSEA function:

```
source (file.path (Path.To.GSEA.Source.Code,
                   "GSEA.1.0.R"))
```

Call GSEA:

```
GSEA(
  # Input/Output Files :-------------------------
  input.ds =  "Datasets/Gender.gct",
  input.cls = "Datasets/Gender.cls",
  gs.db =     "GeneSetDatabases/C1.gmt",
  output.directory = "GSEA_Gender_C1/",

  #  Program parameters :-------------------------
  doc.string                = "Gender_C1",
  non.interactive.run       = FALSE,
  reshuffling.type          = "sample.labels",
  nperm                     = 1000,
  weighted.score.type       =  1,
  nom.p.val.threshold       = -1,
```

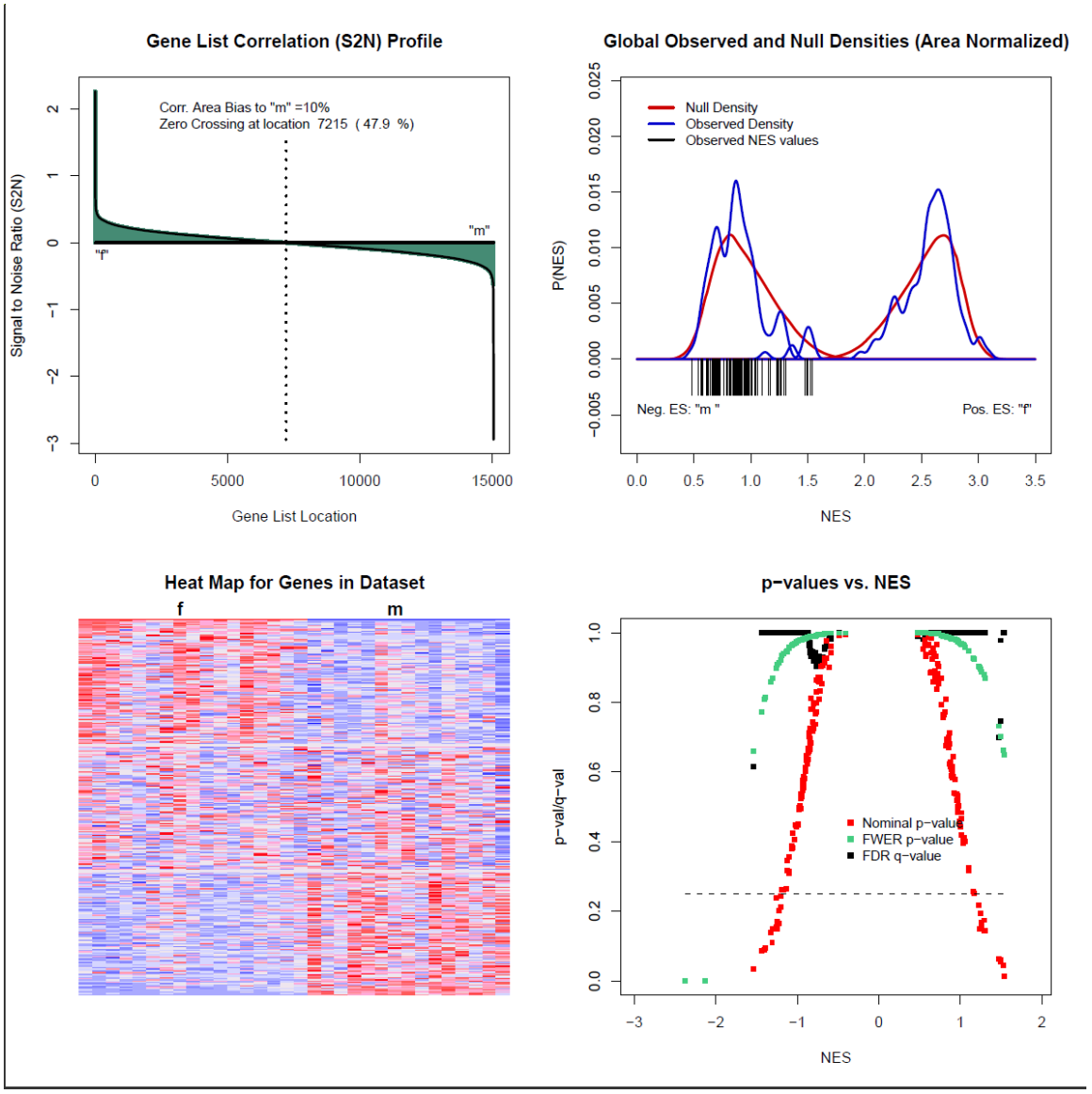
```
  fwer.p.val.threshold  = -1,
  fdr.q.val.threshold   = 0.25,
  topgs                 = 20,
  adjust.FDR.q.val      = FALSE,
  gs.size.threshold.min = 15,
  gs.size.threshold.max = 500,
  reverse.sign          = FALSE,
  preproc.type          = 0,
  random.seed           = 111,

  #  Tweaks for experts only :--------------------
  perm.type             = 0,
  fraction              = 1.0,
  replace               = FALSE,
  save.intermediate.results = FALSE,
```



```
  OLD.GSEA                    = FALSE,
  use.fast.enrichment.routine = TRUE
)
```

```
GSEA.Analyze.Sets(
   directory = "GSEA_Gender_C1/",
   topgs = 20,
   height = 16,
   width = 16
)
```


GSEA Global Plots



GSEA Summary Results

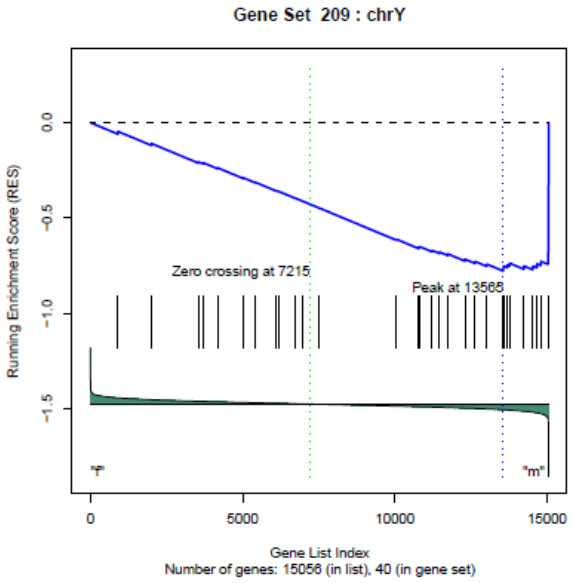
GS	SIZE	SOURCE	ES	NES
chrY	40	Chromosome Y	-0.77758	-2.3729
chrYp11	18	Cytogenetic band	-0.75901	-2.1365
chrYq11	16	Cytogenetic band	-0.88634	-2.1332
chr9q21	35	Cytogenetic band	-0.43408	-1.5349

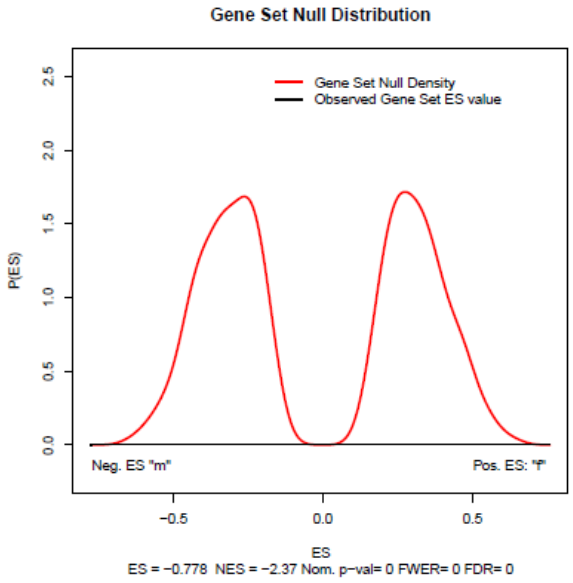
GS	NOM p-val	FDR q-val	FWER p-val	Tag %
chrY	0	0	0	0.475
chrYp11	0	0.00035938	0.001	0.5
chrYq11	0	0.00023958	0.001	0.625
chr9q21	0.03509	0.61654	0.659	0.314

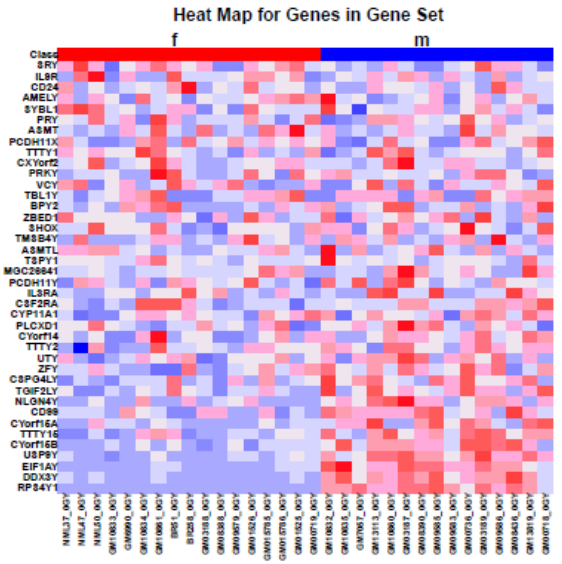
GS	Gene %	Signal	FDR (median)	glob.p.val
chrY	0.0991	0.429	0	0
chrYp11	0.0991	0.451	0	0
chrYq11	0.054	0.592	0	0
chr9q21	0.18	0.258	0.47917	0.224

GSEA Gene Set Plots

GSEA produces plots such as the following for the top (20 in our case) gene sets.



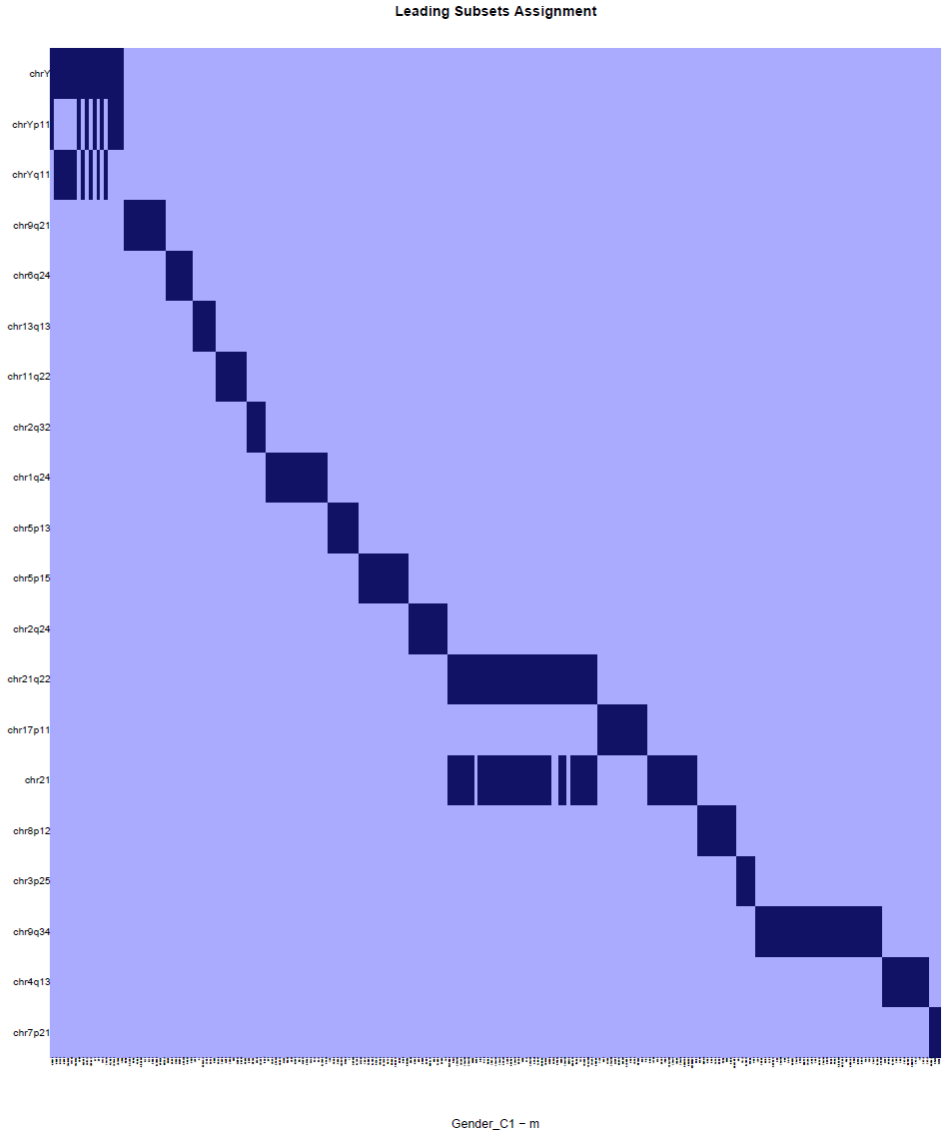




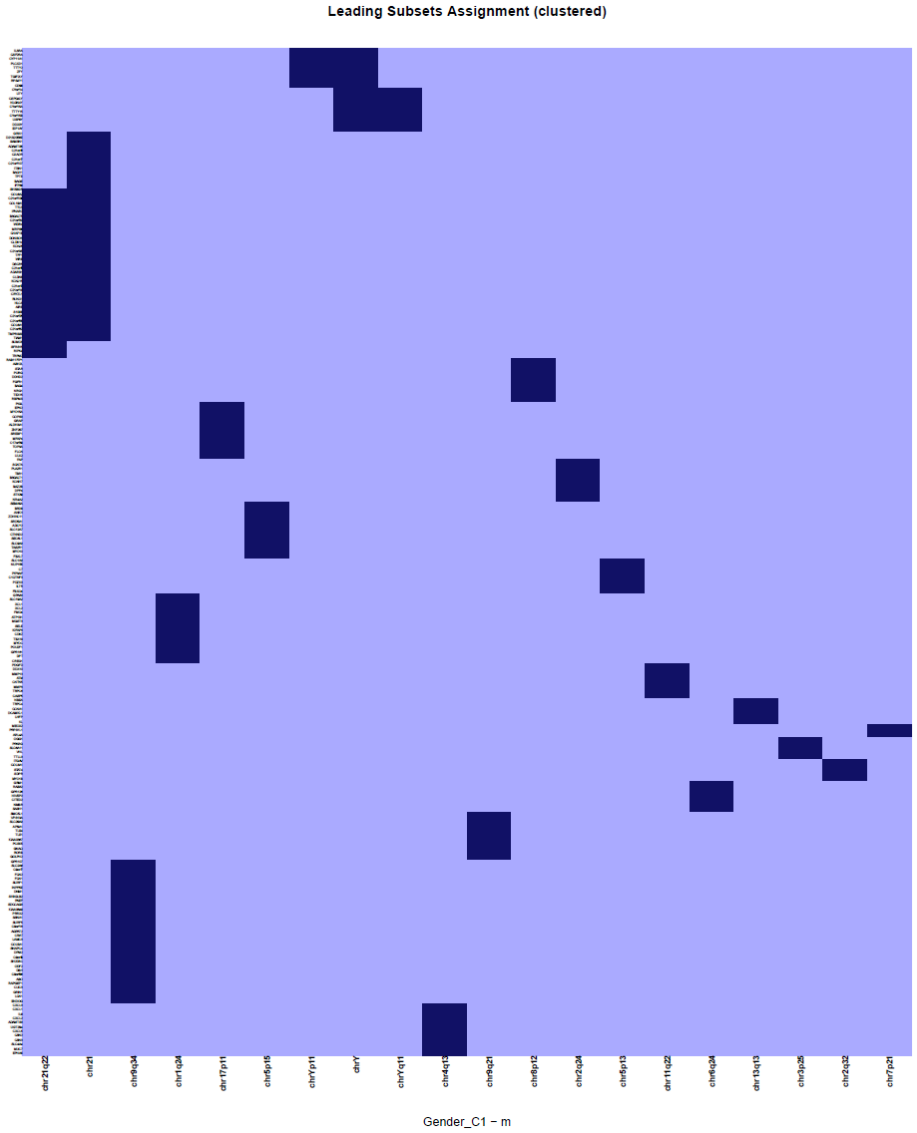
GSEA Gene Set Report

#	GENE	LIST LOC	S2N	RES	CORE_ENRICHMENT
1	RPS4Y1	15056	-2.94	-6.85e-17	YES
2	DDX3Y	15055	-1.84	-0.167	YES
3	EIF1AY	15053	-1.71	-0.272	YES
4	USP9Y	15052	-1.52	-0.369	YES
5	CYorf15B	15051	-0.993	-0.456	YES
6	TTTY15	15050	-0.805	-0.512	YES
7	CYorf15A	15046	-0.705	-0.558	YES
8	CD99	15045	-0.703	-0.598	YES
33	PCDH11X	6105	0.0294	-0.357	NO
34	ASMT	5407	0.05	-0.312	NO
35	PRY	5037	0.0614	-0.29	NO
36	SYBL1	4175	0.089	-0.236	NO
37	AMELY	3720	0.105	-0.211	NO
38	CD24	3543	0.112	-0.205	NO
39	IL9R	2009	0.175	-0.11	NO
40	SRY	904	0.248	-0.046	NO

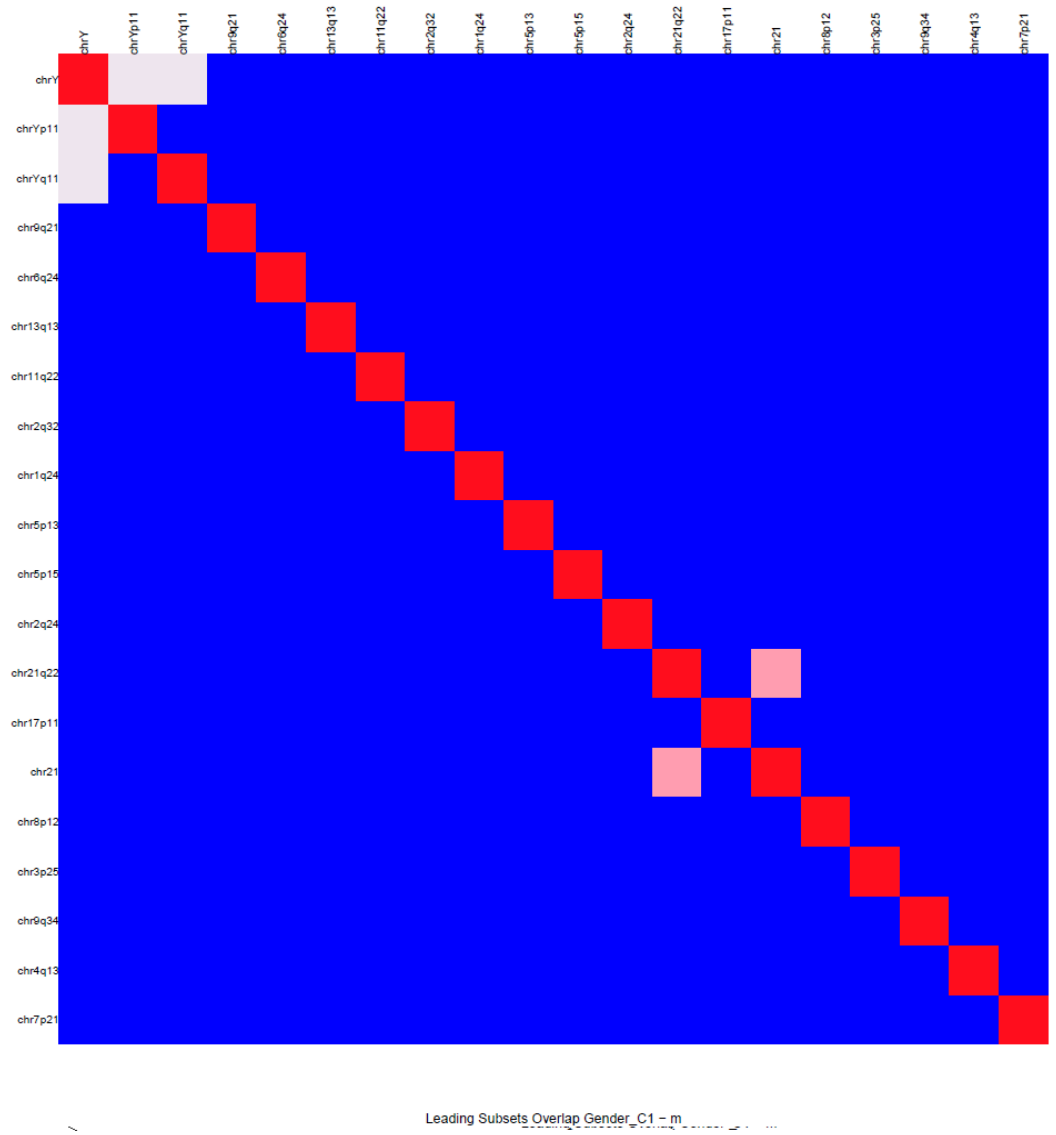
Leading Subsets Assignment



Leading Subsets Assignment - Clustered



Leading Overlap



Defining New Genesets

You aren't limited to using the predefined gene sets.

Indeed, it might be hard to use the existing gene sets together with new data types.

In addition to using sources similar to those used to generate the predefined gene sets, it might be fruitful to use the significant genes found in one analysis as the basis of a geneset.

GSEA facilitates this by create a geneset database file (`.gmt`) file containing a geneset of the leading genes found in the top genesets.

TCGA Analysis Project

As we mentioned last time, there will be a final project in which you will each individually analyze the publicly available TCGA data to find out all you can about a small number of specific genes across multiple assays, and possibly across multiple diseases (Ovarian Cancer and/or Glioblastoma).

TCGA Analysis Project

To enable us to decide which genes each of you will analyze, please submit before the start of next week's lecture (November 11, 3:00 pm):

- a short (one page) list of the genes that you are most interested in analyzing,
- include at least three genes,
- list them in order of descending interest (one is most interesting, two somewhat less so, etc.),
- for each gene include a short list of the specific reasons why you chose it.