Downstream analysis of transcriptomic data

Shamith Samarajiwa

CRUK Bioinformatics Summer School

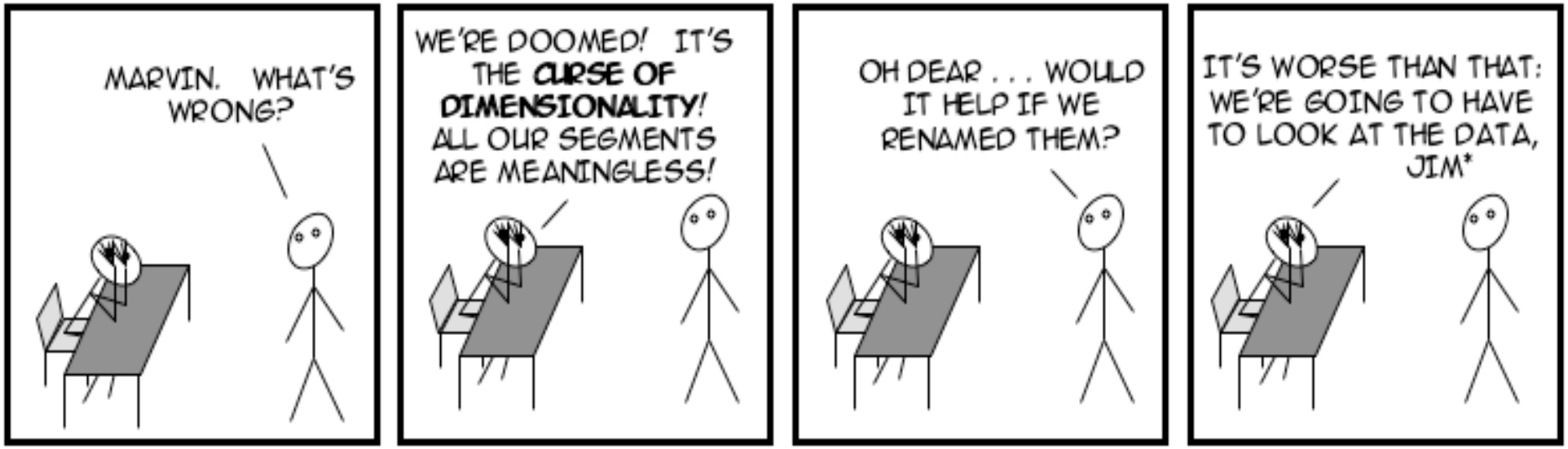
July 2015

General Methods

- Dimensionality reduction methods (clustering, PCA, MDS)
- Visualizing Patterns (heatmaps, dendrograms)
- Over representation Analysis (ORA)
  - Ontology enrichment
  - Gene Set Enrichment analysis (GSEA/GSA)
  - Pathway enrichment analysis
- Network biology methods
- Promoter analysis of co-regulated genes
- Gene/Protein interaction analysis
- Biomedical Bibliomics

The curse of dimensionality

- **Curse of dimensionality** (Bellman 1961) phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings.
- Caused by number of features > number of samples
- high dimensional (d): hundreds or thousands of dimensions.
- gene expression: d ~ 10^4 – 10^5
- SNP data: d ~ 10^6



A common theme of these situations is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance.

Dimensionality Reduction

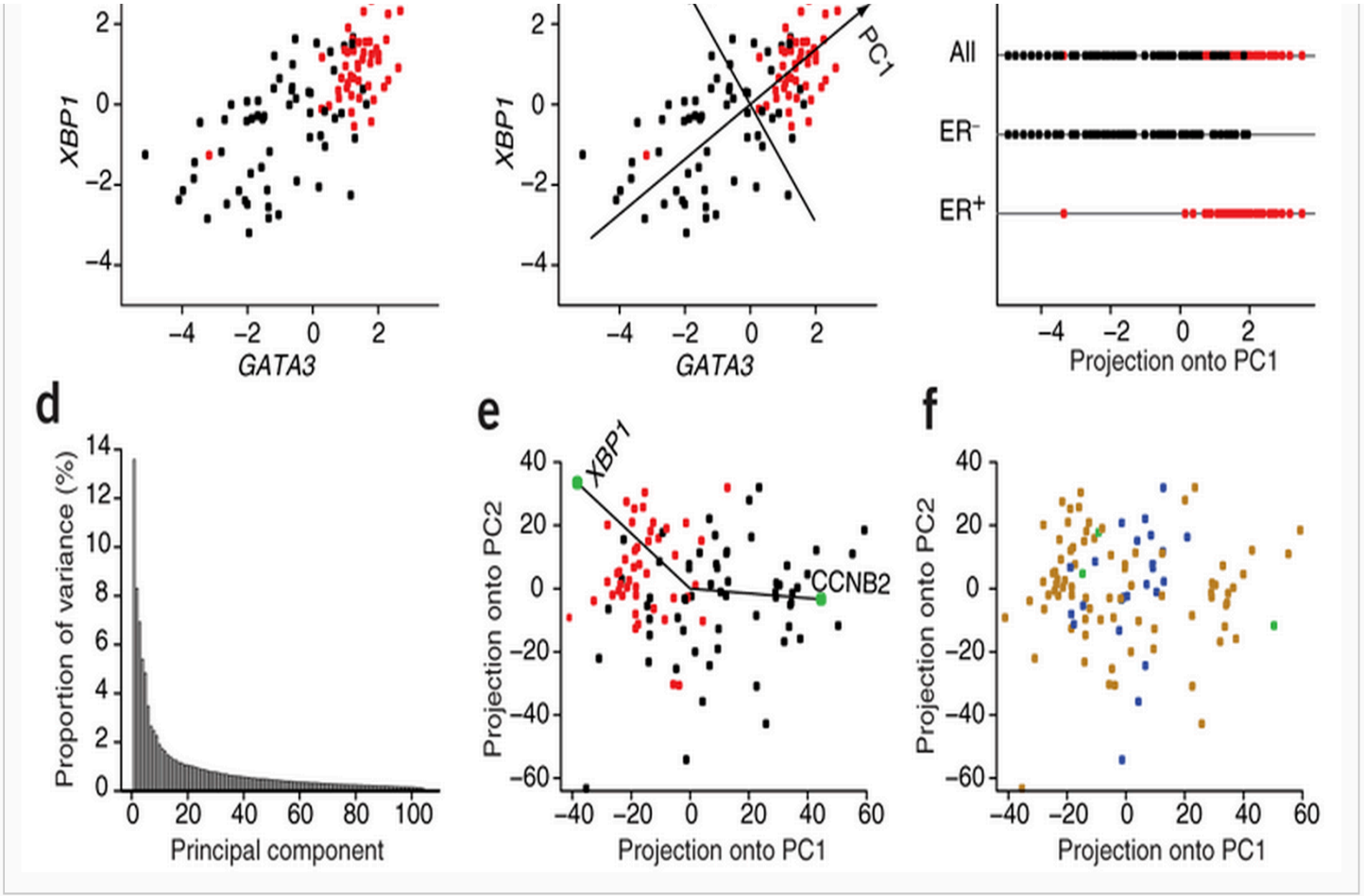
- Dimensionality reduction techniques are a common approach for dealing with noisy, high-dimensional data.

- These **unsupervised** methods can help uncover interesting structure in complex datasets.

- However they give very little insight into the biological and technical aspects that might explain the uncovered structure.

Principal Component Analysis (PCA)

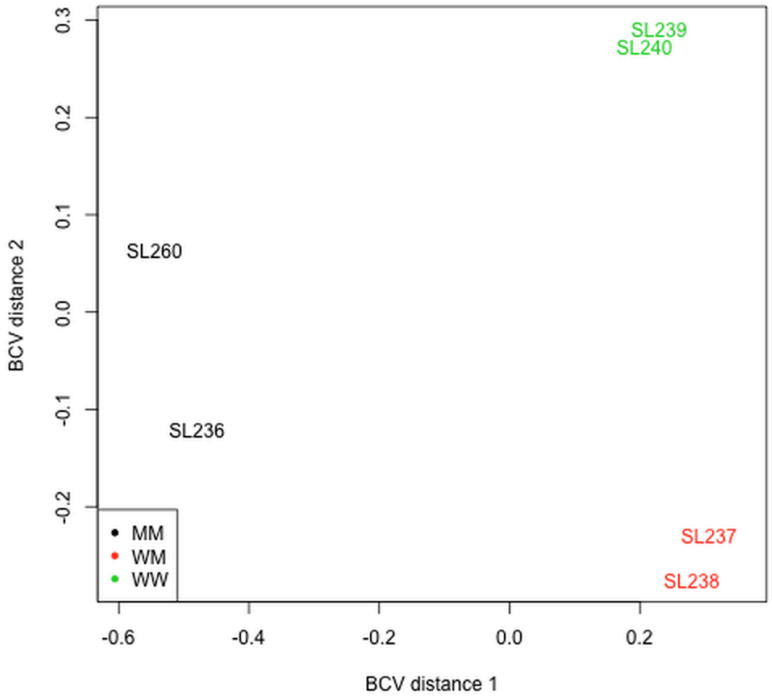
- Principal component analysis (PCA) reduces the dimensionality of the data while retaining most of the variance in the data set.
- It accomplishes this reduction by identifying directions (Eigenvectors + Eigenvalues), called principal components, along which the variance in the data is maximal. By using a few components, each sample can be represented by a relatively few numbers instead of by values for thousands of variables. Data can then be visualized, making it possible to assess similarities and differences between samples and determine whether samples are grouped together.



(**a**) Each dot represents a breast cancer sample plotted against its expression levels for two genes. (In **a–c**, **e**, samples are colored according to estrogen receptor (ER) status: ER^+, red; ER^- , black). (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread. (**c**) Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER^+, ER^- and all samples separately. (**d**) The variance of the principal components when PCA is applied to all 8,534 genes with expression levels for all samples.

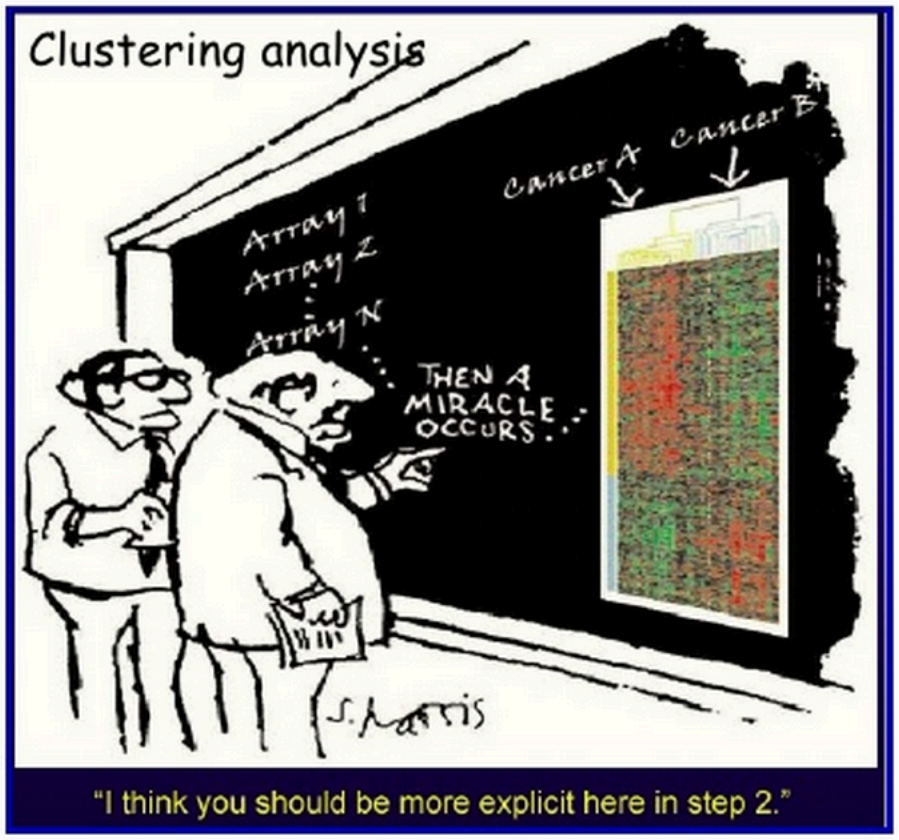
Multi-dimensional Scaling (MDS)

- **Multidimensional scaling** (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix.



Clustering

- The process of grouping together similar entities. Need to define similarity: distance metric
- Hierarchical clustering
- K-means
- SOFM
- PAM
- Biclustering



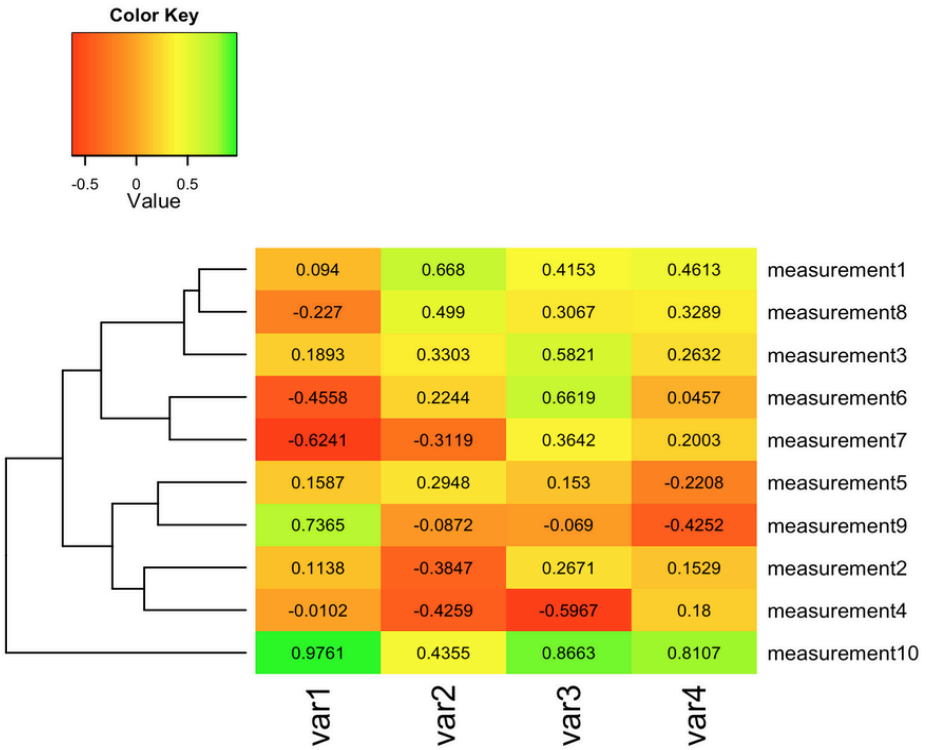
Clustering

- Given enough genes, a set of genes will always cluster.
- Clusters produced by a given algorithm is highly dependent on the distance metric used.
- Same clustering method applied too the same data may produce different results.

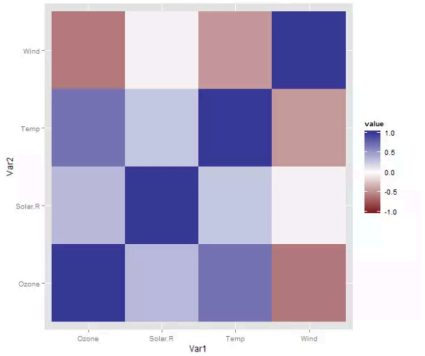
Visualizing complex datasets with heatmaps

- There are multiple R packages and functions that can generate heatmaps.

i. Heatplus
ii. Pheatmap
iii. Heatmap.2{gplots}
iv. Heatmap.plus



Correlation Heatmap

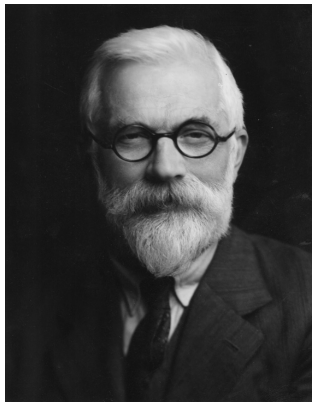


Over Representational Analysis (ORA)

- Given a list of genes/features and one or more lists of annotations, are any of he annotations surprisingly enriched in the gene list?
- How to assess "surprisingly"? -Statistics
- How to correct for repeated testing ?- False Discovery Rate correction
- Test for under and/or over enrichment relative to a background population; selecting the right "background" is important!!

i. Hypergeometric test
ii. Fisher's exact test
iii. Two tailed t-test (sig. difference between means of two distributions)
iv. Kolmogorov–Smirnov test (test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution)

Fisher's Exact Test



- Fisher's exact test is used for ORA of gene lists for a single type of annotation.
- P-value for Fisher's exact test – is "the probability that a random draw of the same size as the gene list from the background population would produce the observed number of annotations in the gene list or more." , – and depends on size of both gene list and background population as well and the number of specific genes in gene list and background.

Ontologies and Ontology Enrichment

- Gene Ontology : a controlled vocabulary and machine readable, Directed Acyclic Graph with multiple levels of classification, can be used across species
- 3 types of gene ontology (BP, MF, CC)
- 9 levels (generic to extremely specific)
- “is a” and “part of” relationships
- Evidence codes
- Allows for incomplete knowledge

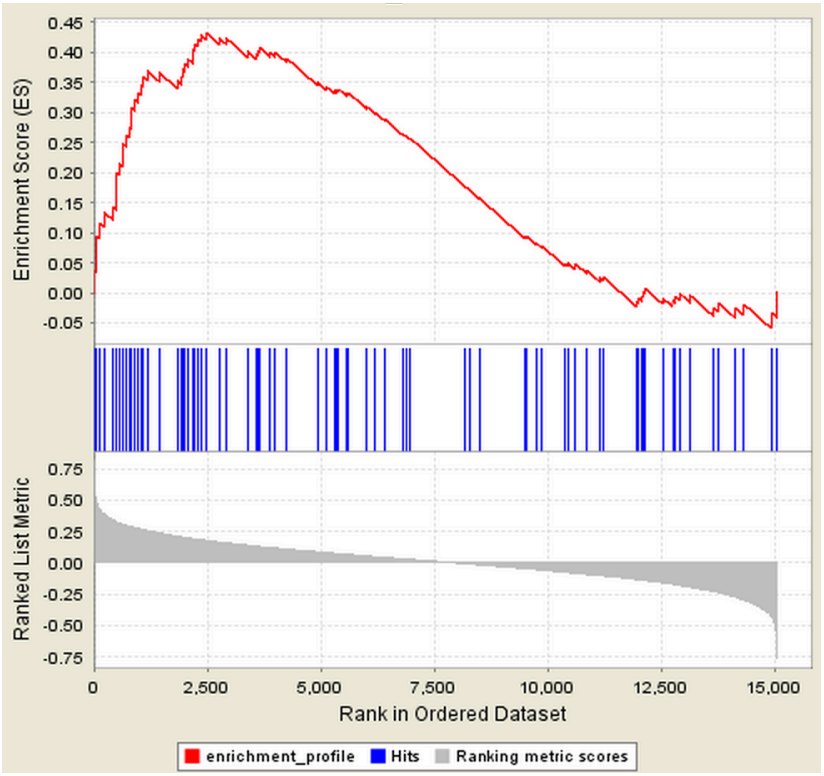
- More than 80 types of biomedical ontologies.
- Enrichment tools: more than 400
- Ex: NIH DAVID, GREAT, GOseq

Pathway Analysis

- Identify pathway that are perturbed or significantly impacted in a given condition or phenotype
- Many pathway databases (~550). Incomplete pathway descriptions.
- Provenance of annotation.
- Signalling, Metabolic, Regulatory pathways

Gene Set Enrichment (GSEA)

- A method for assessing whether a predefined set of genes show statistically significant differences between two conditions.



Msigdb & GMT files

- A collection of gene set
- No centralized curation- user submitted gene sets
- GSEA uses the list rank information without using a threshold.

- The 10348 gene sets in the Molecular Signatures Database (MSigDB) are divided into 8 major collections, and several sub-collections.

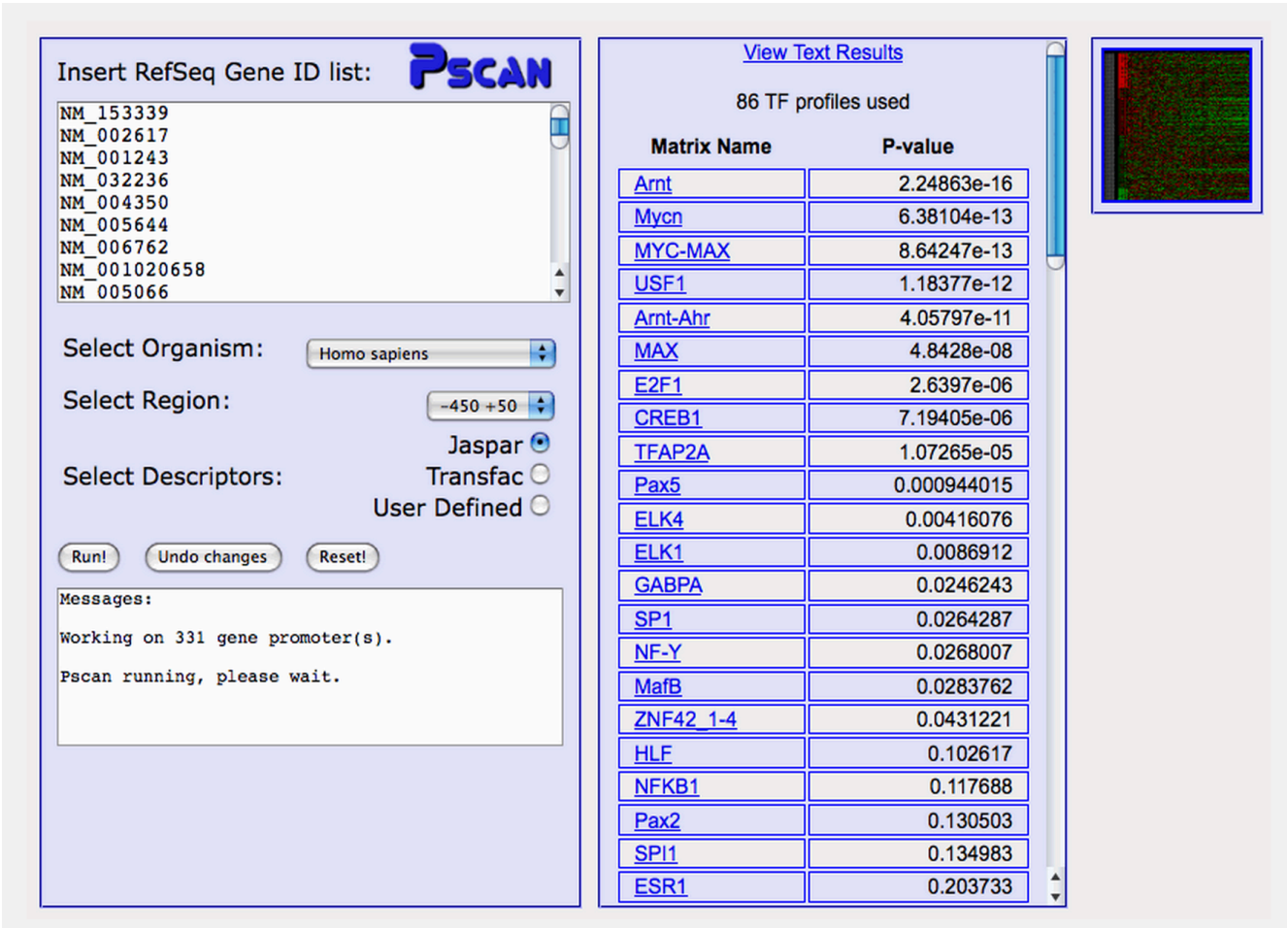
Gene Set Analysis (GSA)

- It differs from a Gene Set Enrichment Analysis (Subramanian et al 2006) in its use of the "maxmean" statistic: this is the mean of the positive or negative part of gene scores in the gene set, whichever is large in absolute values.

Promoter Enrichment Analysis

- Identify co-regulated gene set or signature–Extremely Hard to do!!
- Co regulated genes may have similar biological function or involved in a biological process, show differential expression, involved in the same pathway, belong to the same cluster?
- Identify the upstream regulators of that gene set. How?
- Locate promoters / regulatory elements of the genes set.
- Then search for transcription factor binding sites (TFBS) that are enriched in that promoter set (compared to a random promoter set).

PSCAN


Insert RefSeq Gene ID list:
PSCAN
NM_153339
NM_002617
NM_001243
NM_032236
NM_004350
NM_005644
NM_006762
NM_001020658
NM_005066
Select Organism:
Homo sapiens
Select Region:
-450 +50
Select Descriptors:
Jaspar
Transfac
User Defined
Run!
Undo changes
Reset!
Messages:
Working on 331 gene promoter(s).
Pscan running, please wait.
View Text Results
86 TF profiles used
Matrix Name
P-value
Arnt 2.24863e-16
Mycn 6.38104e-13
MYC-MAX 8.64247e-13
USF1 1.18377e-12
Arnt-Ahr 4.05797e-11
MAX 4.8428e-08
E2F1 2.6397e-06
CREB1 7.19405e-06
TFAP2A 1.07265e-05
Pax5 0.000944015
ELK4 0.00416076
ELK1 0.0086912
GABPA 0.0246243
SP1 0.0264287
NF-Y 0.0268007
MafB 0.0283762
ZNF42_1-4 0.0431221
HLF 0.102617
NFKB1 0.117688
Pax2 0.130503
SPI1 0.134983
ESR1 0.203733