Introduction to Systems Biology
Lecture 7 Part B-1

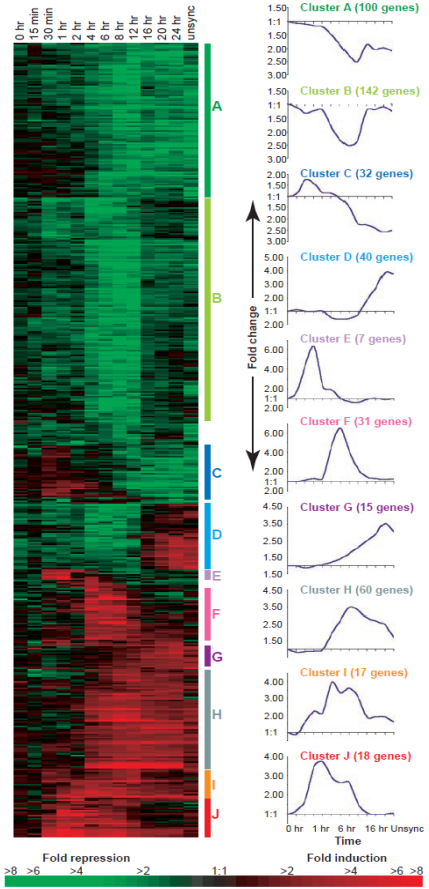
Iyengar

Analyzing Large Data Sets

Identifying differentially expressed genes, proteins , metabolites

Figure shows changes in level of 517 mRNAs at varying time upon stimulation of human fibroblasts with serum
Such representations are called Heatmaps

From HHMI : A –free 26 slide tutorial on how to analyze DNA microarray data
http://www.hhmi.org/biointeractive/how-analyze-dna-microarray-data



Iyer et al (1999)
Science 283:83

Statistical tests to identify differentially expressed genes

T-tests can be used to test if two sets of data are significantly different from each other. Generally used if the test statistic follows a normal distribution

ANOVA analysis of variance – commonly used to test the null hypothesis and determine if there is difference between any two groups when there are more than two groups in an experiment. Significance at a user defined value, p value of 0.05 or 0.01

Mann–Whitney non–parametric test of the null hypothesis. Non-parametric means there is no assumption regarding the distribution of the test statistic

Cluster Analysis – putting entities (e.g.) genes into groups such that entities within a group are more closely related to each other than to entities in another group. Often used to identify groups of genes expressed (or repressed) under a specified condition (perturbation, duration of treatment etc)

Gene-Set Enrichment Analysis

A useful method of grouping based on a user-defined biologically relevant criterion

- Common biological function
- Location on the same chromosome
- Regulated by the same signaling pathway

Uses:
To identify pathways of relevance for a phenotype such as cancer
-able to identify the Ras-MAPK pathway associated with a certain (p53-) type of cancer

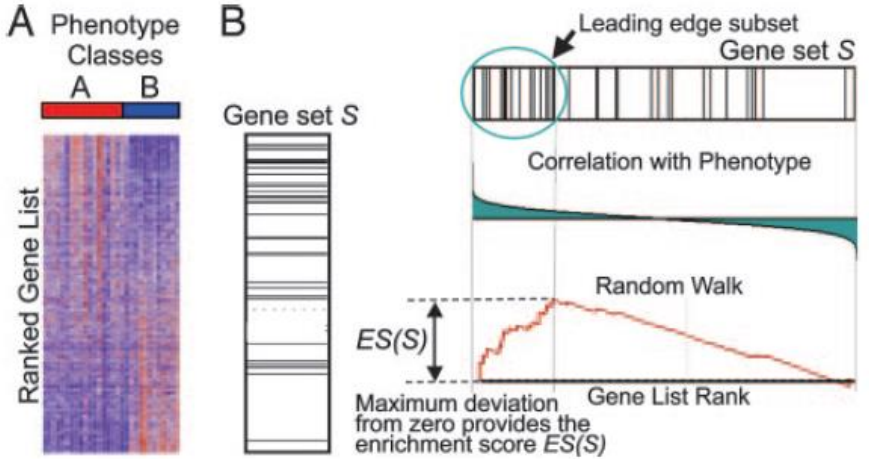


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the ''gene tags,'' i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Subramanian A et.al. (2005) PNAS 102: 15545

Sources of variations in experimental data - Genomics

1. Sample preparation - Biological variation - very important especially for tissue samples

2. Number of probes per genes

3. Cross reactivity of the probe

4. Chips from different manufacturers perform differently in CNV and SNP detection (Halper-Stromberg et al (2011) Bioinformatics 27: 1052)

5. Use of appropriate software to "remove" technical artifacts is important (Halper-Stromberg et al (2011) Bioinformatics 27: 1052)

Programs for Analyzing Genomic Data Sets

Most manufacturers of chips now supply programs to analyze data obtained with their chips - generally these work well

For analysis of RNA –Seq data

Cufflinks and Cuffdiff
An open source program that maps RNA-Seq reads to a reference genome to identify transcripts and estimate relative abundance

Cuffdiff can be used to detect change in expression levels of individual transcripts

http://cufflinks.cbcb.umd.edu

Genome-wide Association Studies

Identification of variations in DNA sequence that are associated with increased risk of a disease

Most often focused on SNPs

Define phenotype: categorical or quantitative

Assemble patient population for control and disease group

Sequence whole genome – for better established cases – SNP-Chips

Use of appropriate statistical test to establish association of SNPs with increased risk of disease

Bush W.S and Moore J. H. (2012) PloS Comp Bio 8 : issue 12 e1002822

Sources of Variation in Experimental Technologies - 2

Proteomics Technologies

1. Sample preparation - biological variation
Appropriate preparation of tissue samples

2. Number of peptides per proteins produced by enzymatic degradation

3. Overrepresentation of highly abundant proteins

4. Appropriate data analysis - a 2009 study showed the central importance of appropriate data analysis
Improvement of search engines and data bases

Annotating differentially expressed genes & proteins for sub-cellular functions

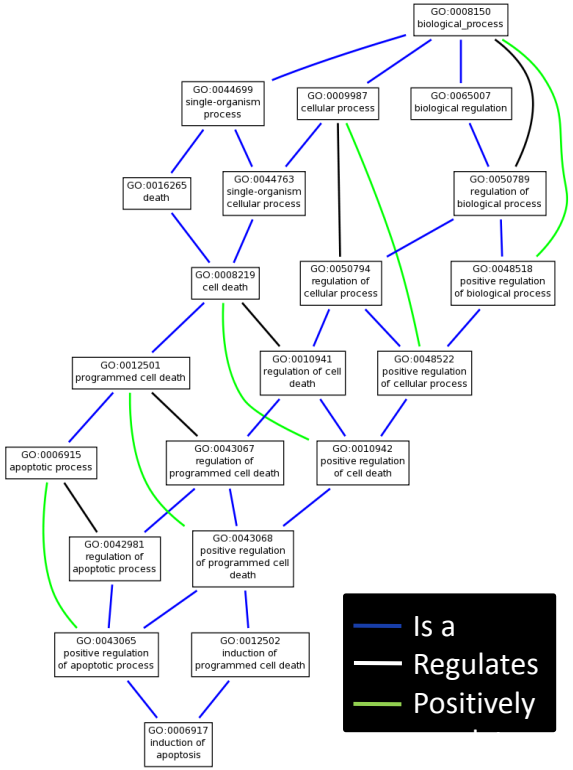
Gene –Ontology – GO

A bioinformatics resource that allows you to categorize genes/gene products (proteins)

It contains three categories: 'Biological Process', 'Cellular Component', 'Molecular Function'

Each of these categories is organized in a hierarchical manner:
- More nonspecific terms are called Parents which have more specific terms are called Children
- The relationship between Parents and Children is further characterized by GO relations (e.g.: 'is a', 'part of', 'has part', 'regulates')

www.geneontology.org



Example: MAPK1

Human MAPK1 is associated with about 100 terms in the three different GO categories

One of these terms is ‘induction of apoptosis’ (‘Biological process’ category)

www.geneontology.org

Lecture 7 – Take Home Points

- High throughput technologies enable us to simultaneously determine many changes in response to perturbations
- Genomic technologies have allowed us obtain a deep understanding of how various changes at the genomic and epigenomic levels can be related to disease origins or progression
- Attention to technical details is critical for obtaining high quality high throughput data
- The use of statistical analyses is critical for extracting knowledge from large sets
- Bioinformatics approaches are useful in connecting molecular changes to cellular processes