BIO306: Bioinformatics

Lecture 13

Gene Ontology (GO) and Enrichment Analysis

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Review last lecture

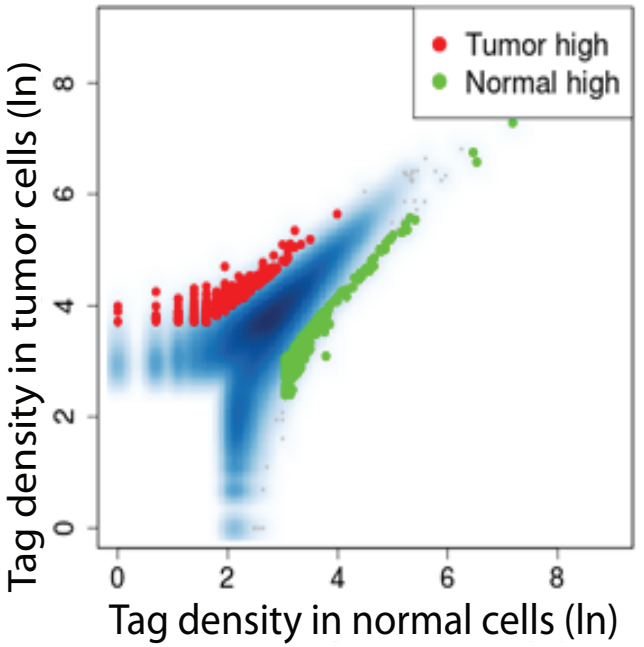
- Definition of epigenetics?
- How to detect genome-wide DNA methylation?
- How to detect genome-wide nucleosome positioning and chromatin accessibility?
- How to identify genome-wide TF binging sites ?
- What is Hi-C? How to identify the significant interaction?

How to interprete data?

- Driven by experimental questions, but with a long list of siginficant genes
  - which genes are of interest?
  - what's special about the differentially expressed genes?

Common phenomenon

- Genes expression or epigenetic profile changes when cell status change. What is the features of these genes?



Gene Ontology:
provides a controlled vocabulary to describe gene and gene product attributes in any organism

Purpose: unified, unambiguous, structured nomaclatures for the use of computer program

Ontology (Philosophical)

- Philosophical study of the nature of being, existence or reality as such, as well as the basic categories of being and their relations.
- Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning:
  - what entities exist or can be said to exist
  - how such entities can be grouped, related within a hierarchy
  - subdivided according to similarities and differences.

Ontology (Information scientist)

Terminologies with which axioms and definitions are associated, formulated in ways which make them suitable for supporting software applications (for example in some Description Logic framework)

Object orientated programming languages

Not only define annotation, but also define sophisticated relationships so that reasoning can be formulated

Beginning

- DNA sequencing
- Protein database (PDB)

- SwissPort: protein sequence
- Flybase: drosophila genetics and molecular biology
- AceDB: c elegan
- SGD: yeast

...

Databases use different terminologies, more and more data and entities...

Yeast, Mouse and Fly representives founded the Gene Ontology Consortium

“Gene Ontology: tool for the unification of biology” 2000

“structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism”

Gene Ontology: database

- By June 2003:
  1297 cellular components
  5396 molecular functions
  7290 biological processes
  in total: 11020

- Under the bigger umbrella of OBO
  Open Biological Ontology
  http://obo.sourceforge.net

Gene Ontology: construction

- User-driven: like wiki, an annotation with its source
  - literature
  - database
  - computational analysis
- Curation team
- Conflicts -> case by case discussion
- Automatic quality controls

Gene Ontology: usage

- integrating gene/protein information from different organisms
- assigning functions to protein domains
- finding functional similarities in genes
- predicting the likelihood that a particular gene is involved in diseases that haven‘t yet been mapped to specific genes
- analyzing groups of genes that are co-expressed during development;
- developing automated ways of deriving information about gene function from the literature

Gene Ontology: usage (cont'ed)

- developing automated ways of deriving information about gene function from the literature;
- verifying models of genetic, metabolic and product interaction networks.

Building Ontology

- The ontologies in OBO serve as controlled vocabularies for expressing the results of biological science.
- 'A relation B' (where 'A' and 'B' are terms in a biological ontology and 'relation' stands in for 'part_of' or some similar expression) general statements about biological classes or types.

Building Ontology

- Assertions about corresponding instances (I.e. the mass of this particular specimen in this particular Petri dish) do not belong to the general statements of biological science (outside the scope of OBO).
- Yet such assertions are still relevant to ontologies: *only through instances definitions and rules for coding, relations between classes can be formulated in an intuitive and unambiguous way*.

Gene Ontology: Basics

GO terms: Three fundamental sets:

(1)cellular component (where is it)
(2)molecular function (what it does there)
(3) biological process (how it does it)

Biological Process

- a recognized series of events (more than one step)
  - cell cycle, development, metabolism, signal transduction

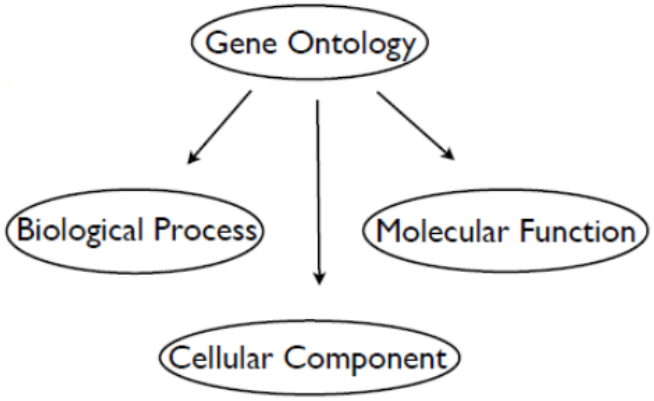
Molecular Function

- what does a gene's product do?
  - binding, enzyme, ligand

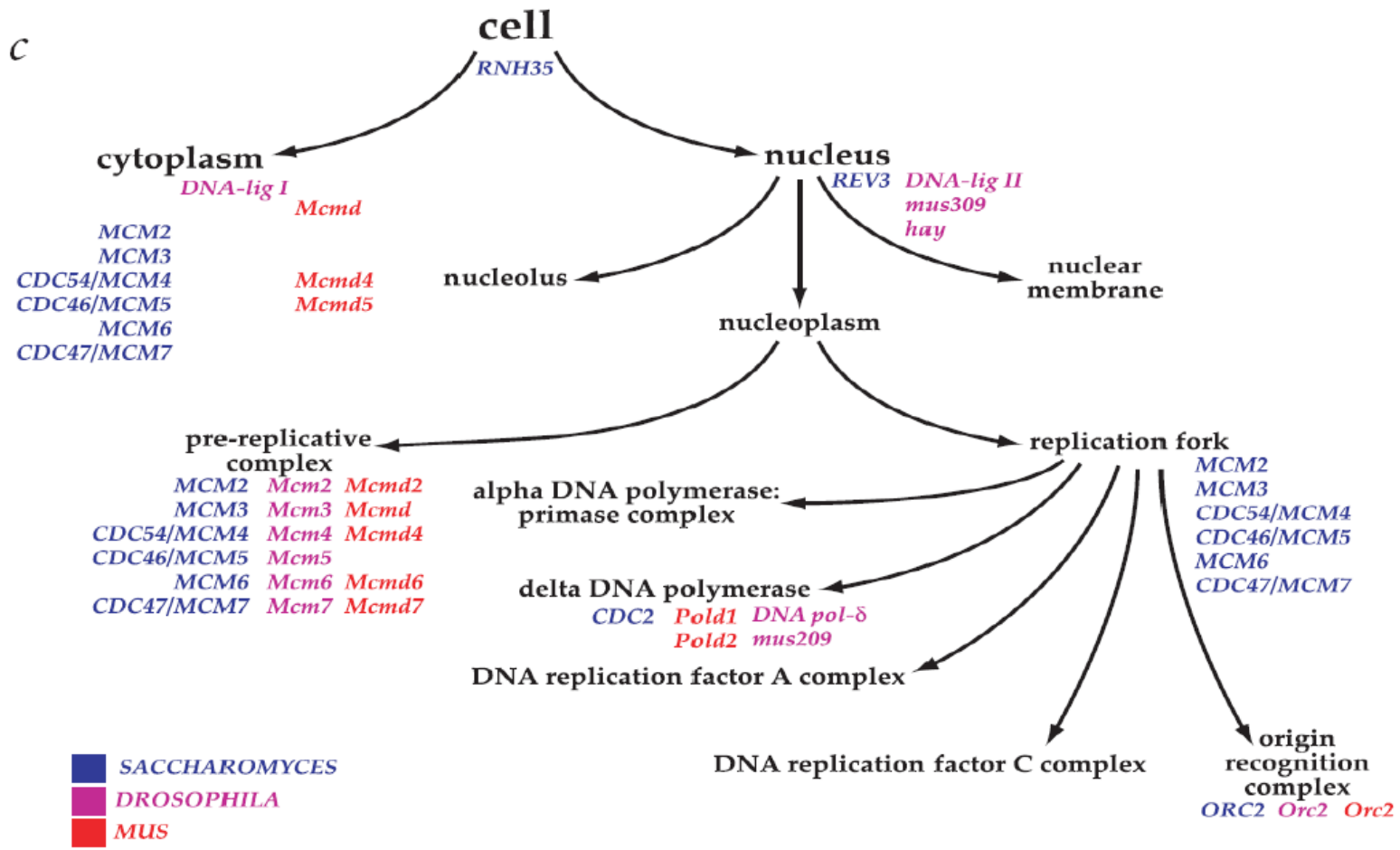
Cellular Component

- localization within a cell
  - membrane, nucleus, complex

It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

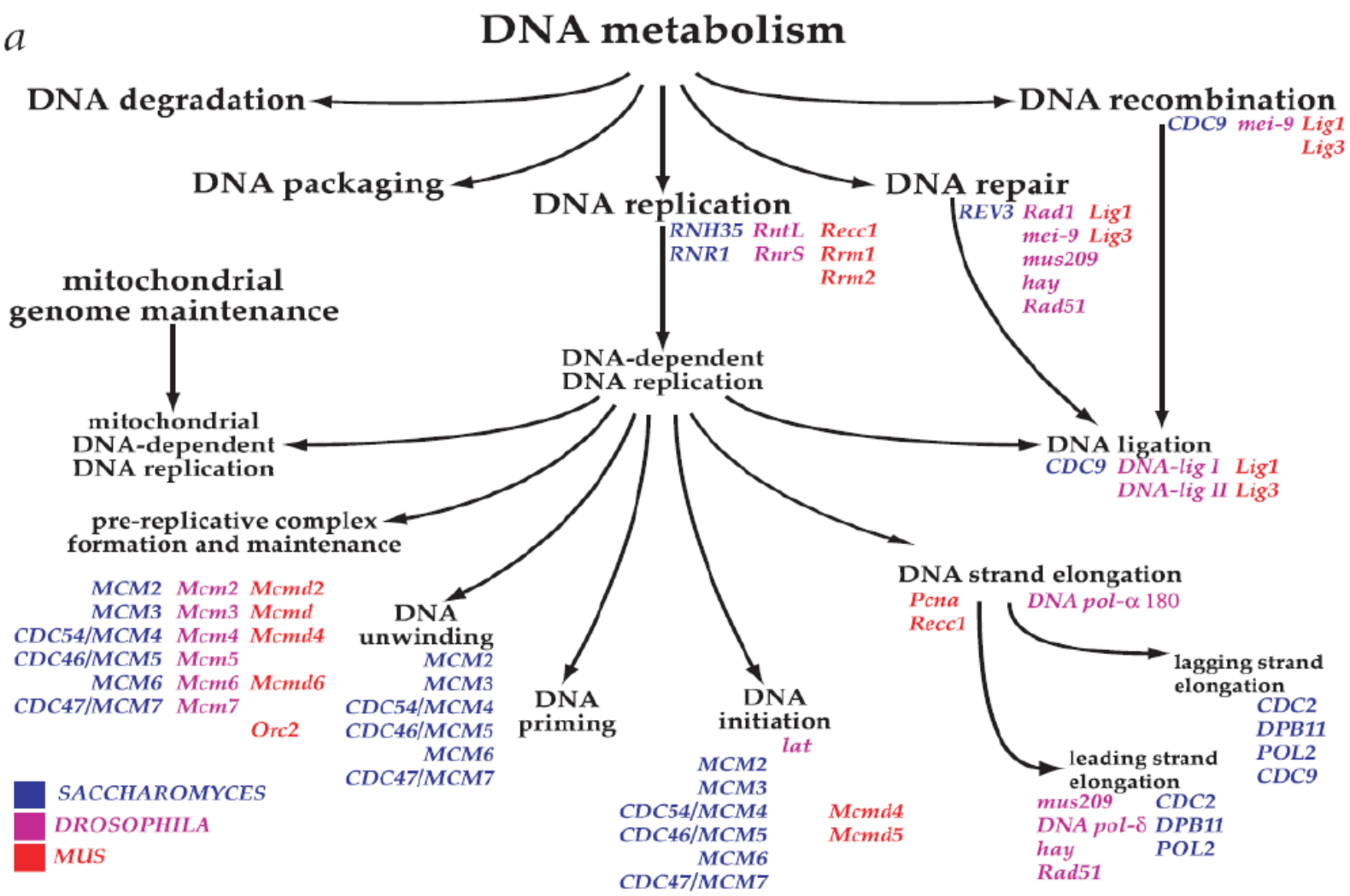


Gene Ontology -Cellular Component



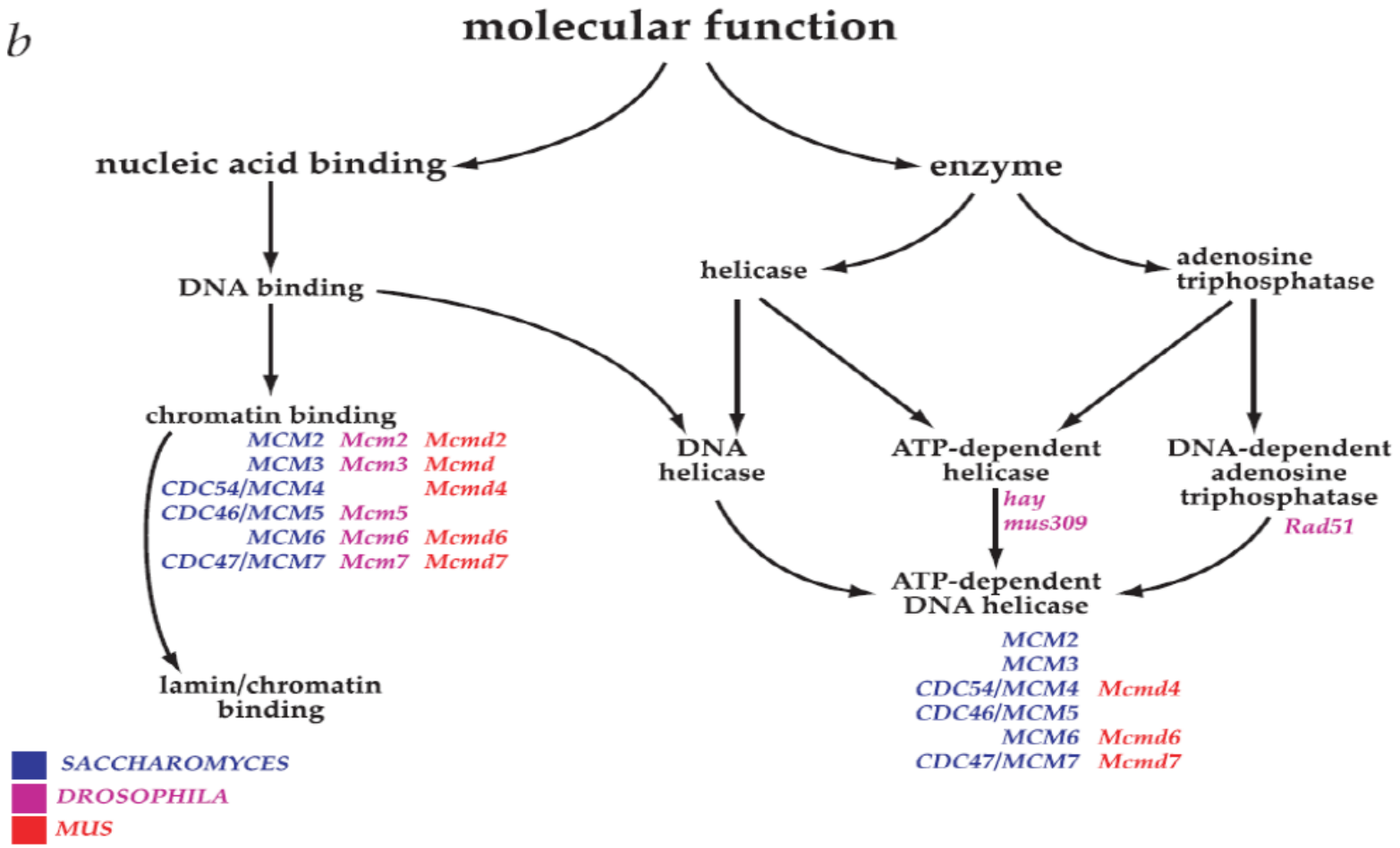
http://www.geneontology.org/GO_nature_genetics_2000.pdf

Gene Ontology –Biological Process



http://www.geneontology.org/GO_nature_genetics_2000.pdf

Gene Ontology – Molecular Function



Gene Ontology: Basics

Structure: Directed Acyclic Graph

- Single direction
- Zero or more Children
- One or more parents
- Linked as synonyms

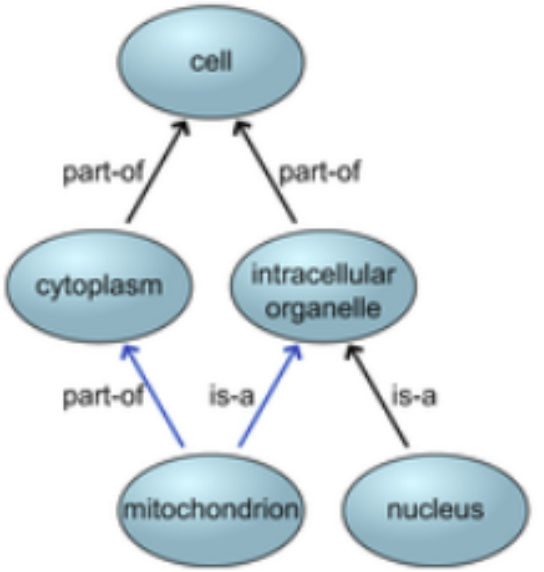


Figure 1. - Directed Acyclic Graph

Gene Ontology: Relationship

- “IS_A”
  one term is a subclass of another class.
  “Golgi apparatus” is a kind of “cell organelle”

- “Part_of”
  one term is a component of another term.
  “Golgi apparatus” is a part of “cell”

Open Biomedical Ontologies OBO

- http://obo.sourceforge.net
- Based on GO success story
- umbrella body for the developers of life-science ontologies
- Contains 60 ontologies

Ontologies Characteristics

- Open: available for use without any constraint or license (applicable to new purposes without restriction).
- receptive to modification (community debate)
- orthogonal (additivity of annotations, modular development)
- syntactically in good order (algorithmic processing)
- employ a common system of identifiers (compatibility with legacy annotations as the ontologies evolve).

Additional Characteristics (OBO Foundry)

- developed in a collaborative effort
- use common relations that are unambiguously defined
- provide procedures for user feedback and for identifying successive versions and
- have a clearly bounded subject-matter (so that an ontology devoted to cell components, for example, should not include terms like ‘database’ or ‘integer’)

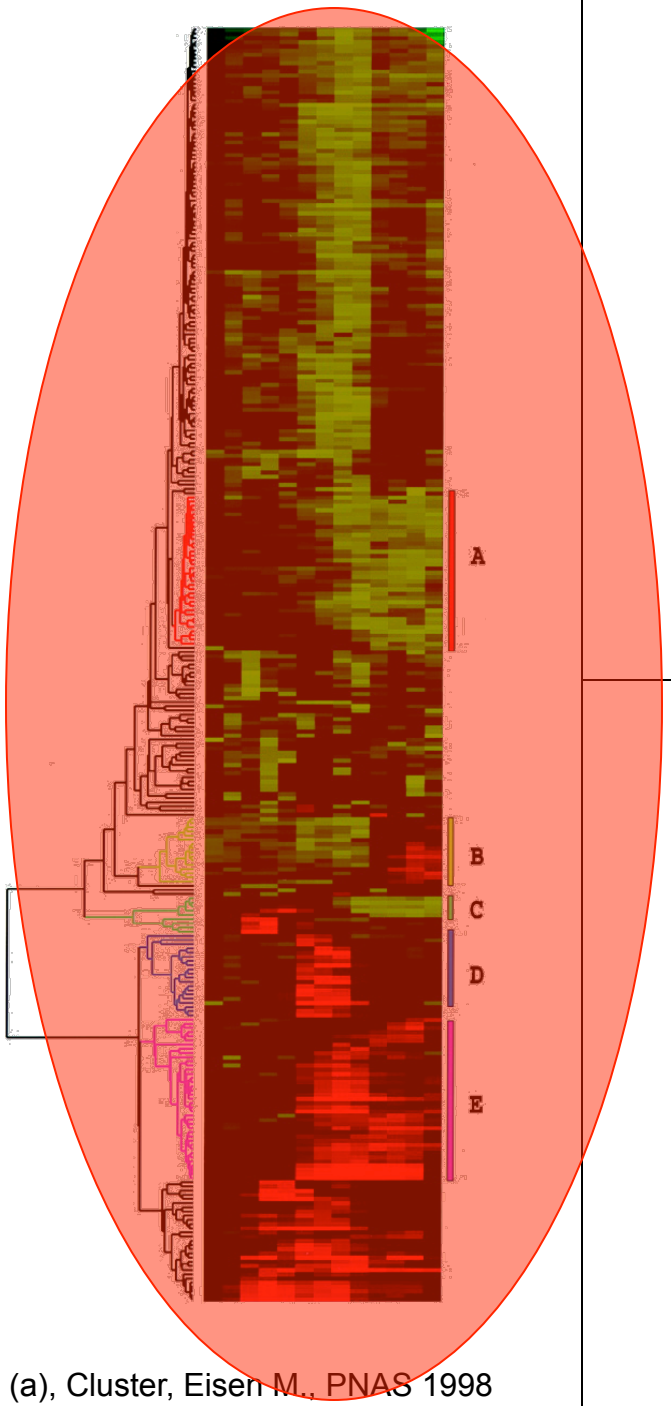
- Serves OBI, ontology for coordinated representation of designs, protocols, instrumentation, materials, processes, data and types of analysis in all areas of biological and biomedical investigation

Enrichments analysis

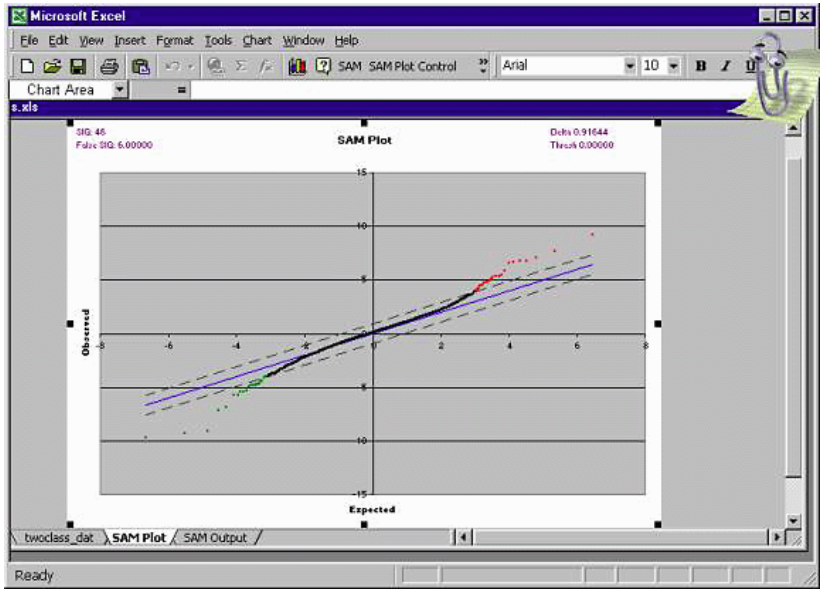
- Enrichment approaches could be classified into 3 categories:
  - Singular Enrichment Analysis (SEA)
  - Gene Set Enrichment Analysis (GSEA)
  - Modular Enrichment Analysis (MEA)
- Simlarly to gene grouping these 3 methods are concerned with unannotated and annotated lists and networks

SEA

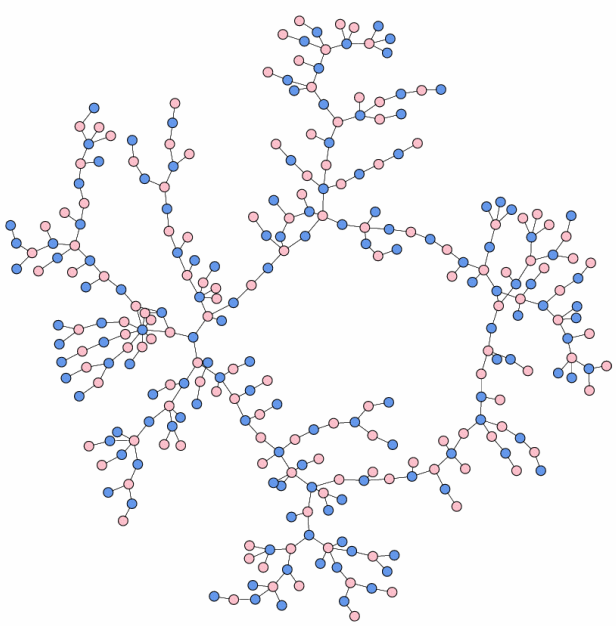
Singular Enrichment Analysis



(a), Cluster, Eisen M., PNAS 1998



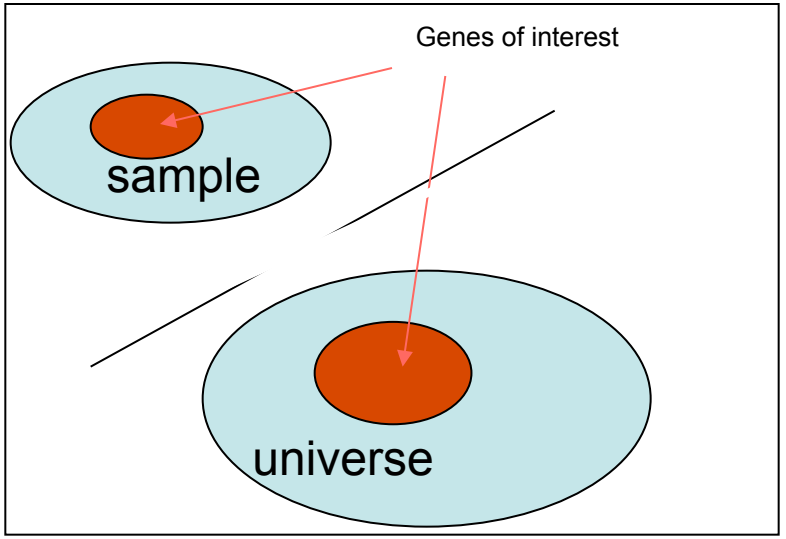
(b) SAM, Tusher V., PNAS, 2001



(c) MNI, di Bernardo D., Nat. Bitechnology, 2005

Enrichment

- Used to give general '*meaning*' to a cluster
- Intuitively: if the proportion of items (genes) in a cluster that belong to a given category (molecular function) is '*big enough*', then we can assess that all the genes in the cluster are related to this same category
- Evaluates the proportion of the frequencies observed in the cluster (sample) of interest with respect to the frequencies expected by random sampling in the global pool of items (genes).
- Expected frequencies are defined by the frequency of the genes in the genome or in the array (universe)



ENRICHMENT

$$\eta_{i,j} = \frac{f_i^{observed}}{f_i^{\exp ected}}$$

Enrichment

- One way (but not exclusive) to evaluate the significance of the enrichment the hypergeometric distribution is used, to calculate the probability α of being wrong when assuming that the enrichment is not due to chance alone.

Hypergeometric distribution

- Allows to evaluate the likelyhood of extracting from a sample of size *n* a number *r* of items of a given type, without resampling, knowing that the sample comes from a pool of *N* items,that contains n_1 items of the given type of interest

p-value

$$p_{i,j}^{\eta} = 1 - \sum_{i=1}^{r-1} \frac{\binom{n_1}{i} \cdot \binom{N - n_1}{n - i}}{\binom{N}{n}}$$

R. A. Fisher and the lady tasting tea

EMBL-EBI

2×2 contigency table

	Patient with disease *D*	Healthy control subject	Total
Elevated level of compound *C*	4	2	6
Normal level of compound *C*	1	3	4
Total	5	5	10

Expected value for top left corner
from null model (no association): $5 \times 6 / 10 = 3$

Hypergeometric distribution

Probability to get this 2×2 table without an association between *D* and *C*:

$$\frac{\text{Number of ways to choose 4 out of the 5 patients to have elevated } C \times \text{Number of ways to choose 2 out of the 5 controls to have elevated } C}{\text{Number of ways to choose 6 out of the 10 persons to have elevated } C} = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}}$$

in R:

```
> dhyper( 4, 5, 5, 6 )
[1] 0.2380952
```

EMBL-EBI

Hypergeometric distribution

Under the null hypothesis, i.e., the assumption that there is no association between elevated levels of compound *C* and presence of disease *D*, *t*he probability that 4 or even more of the patients have elevated levels of C, is

$$p = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}} + \frac{\binom{5}{5}\binom{5}{1}}{\binom{10}{6}} = 0.26$$

in R:

```
> 1 - phyper( 3, 5, 5, 6 )
[1] 0.2619048
```

EMBL-EBI

Hypergeometric testing of gene sets

Given a list of differentially expressed genes and a collection of gene sets, the following strategy is often employed:

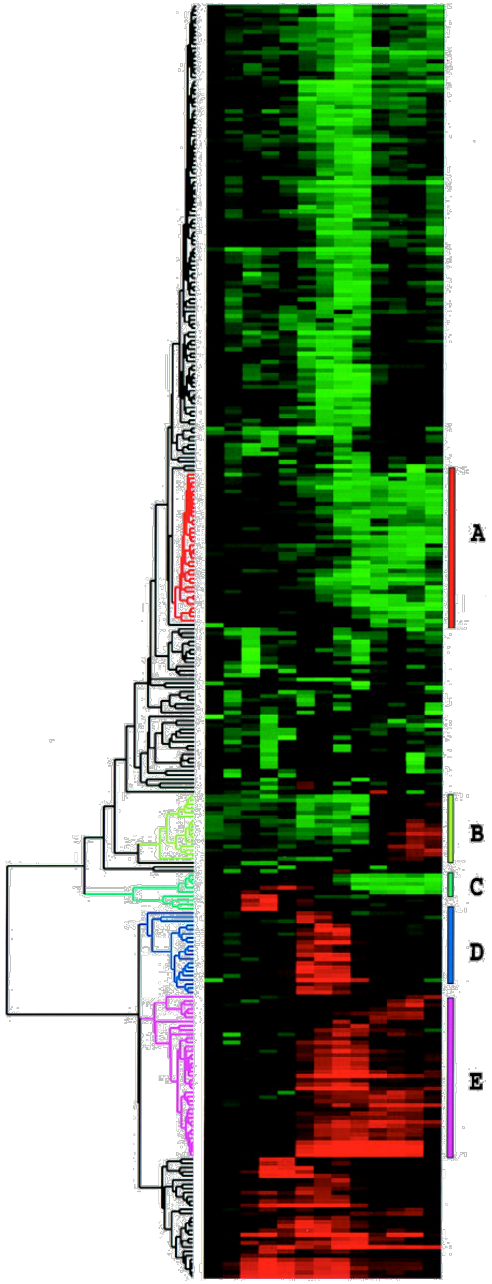
- For each gene set, fill a 2x2 contigency table:

	Differentially expressed	Not differentially expressed	Total
in gene set	.	.	.
not in geneset	.	.	.
Total	.	.	.

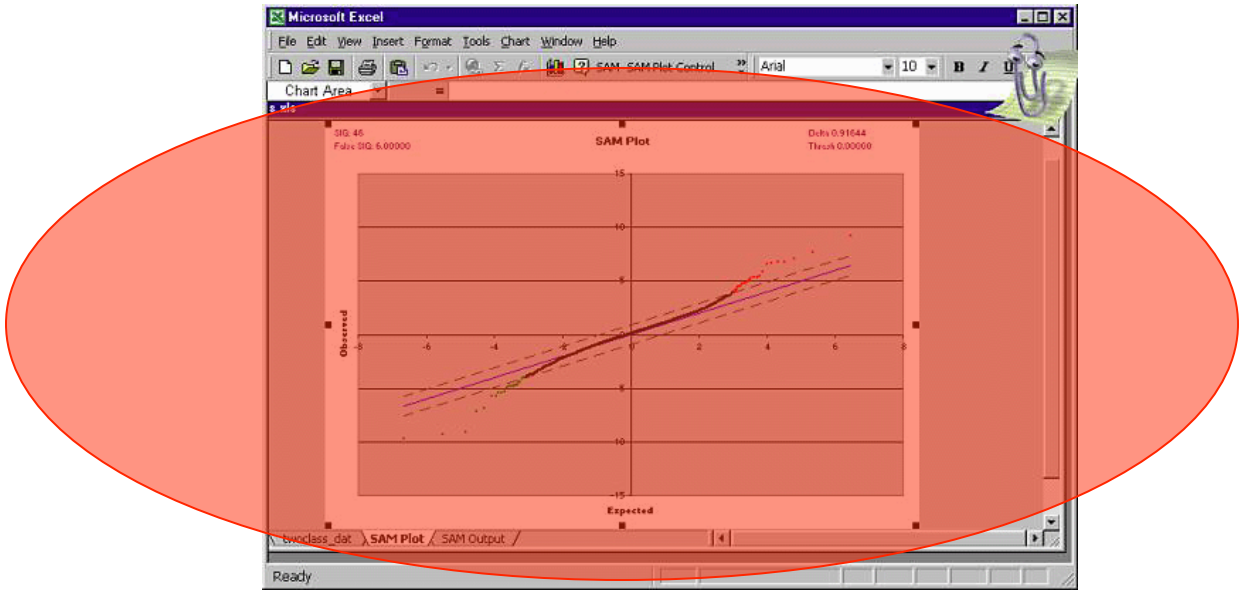
- Calculate *p* value by hypergeometric testing (Fisher's exact test)

GSEA

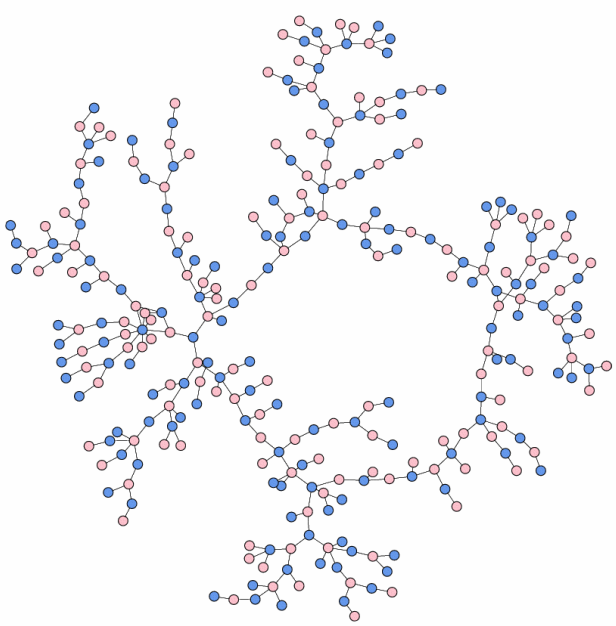
Gene Set Enrichment Analysis



(a), Cluster, Eisen M., PNAS 1998



(b) SAM, Tusher V., PNAS, 2001



(c) MNI, di Bernardo D., Nat. Bitechnology, 2005

GSEA

- ***Focuses on genes sets, that is genes that share common biological function, chromosomal location or regulation***
- ***Does not need a cut-off value to identify interesting, enriched genes***

GSEA

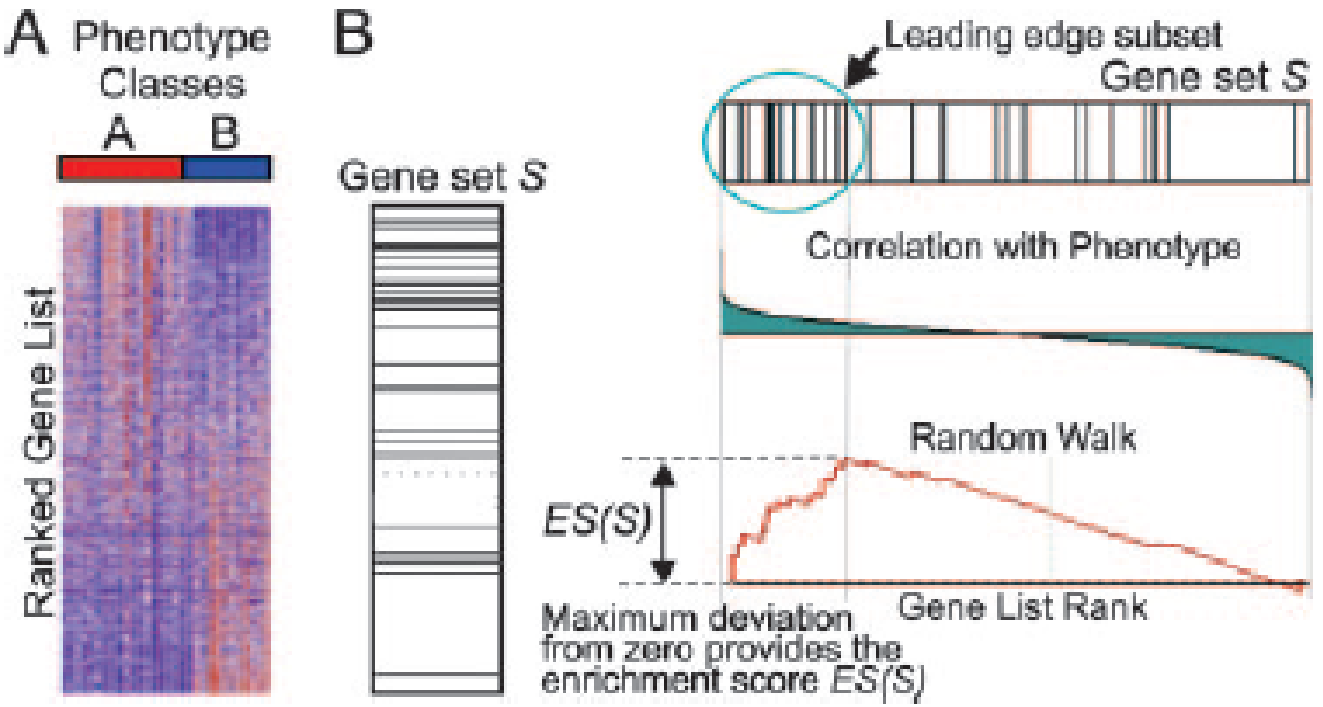
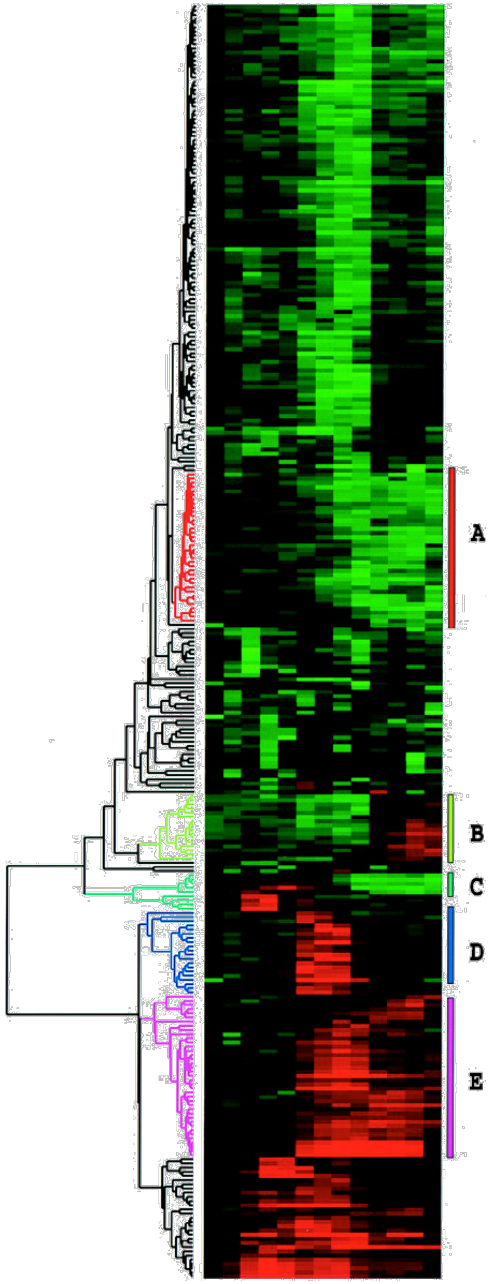


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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

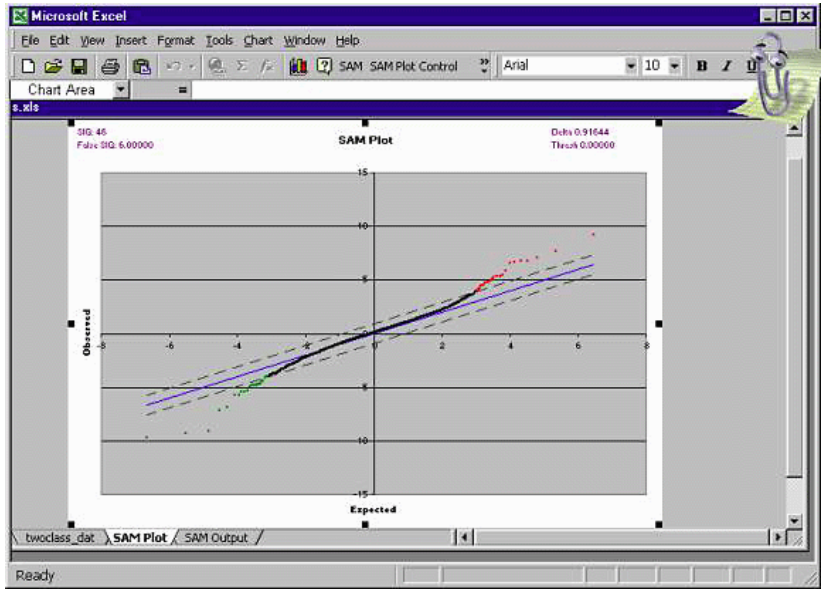
Aravind Subramanian et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, PNAS October 25, 2005 vol. 102 no. 43 15545-15550

MEA

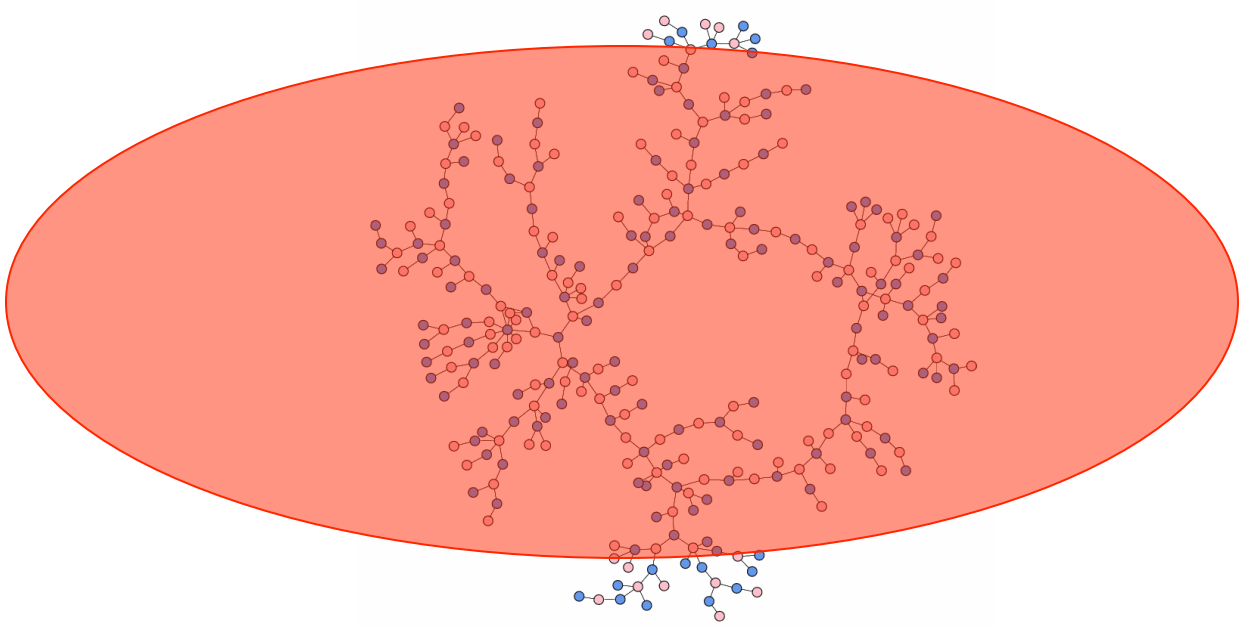
Modular Enrichment Analysis



(a), Cluster, Eisen M., PNAS 1998



(b) SAM, Tusher V., PNAS, 2001



(c) MNI, di Bernardo D., Nat. Bitechnology, 2005

MEA

- Builds on SEA
- Add the possibility to have term to term relationship
- Close to biclustering and networking, allows to construct complex annotations

Enrichment

- k = (Pr(a) - Pr(e))/ (1-Pr(e))
- Pr(a) relative observed agreement among raters
- Pr(e) hypothetical probability of chance agreement
- κ = 1 -> raters in complete agreement.
- κ = 0 -> no agreement among the raters other than what would be expected by chance (as defined by Pr(e)).

Heuristic Multiple Linkage Clustering (DAVID)

- heuristic partitioning procedure allows a gene to participate in more than one cluster.
  1. automatic determination of the optimal numbers of clusters (K)
  2. exclusion of members (genes) that have weak relationships to other members.
- Algorithm:
  - Fuzzy seeding by allowing each gene to serve as a medioid (# neighbor > 4 && cross relevance > 50%)
  - Merge seeding clusters by multiple linkage
  - Repeat 2 until no more merge needed

Enrichment analysis (DVAID)

- The Database for Annotation, Visualization and Integrated Discovery (DAVID)
- https://david.ncifcrf.gov

Thank you for your attention