Clustering and Pathways

Integrative Genomics module

Michael Inouye
Centre for Systems Genomics
University of Melbourne, Australia

Summer Institute in Statistical Genetics 2016
Seattle, USA

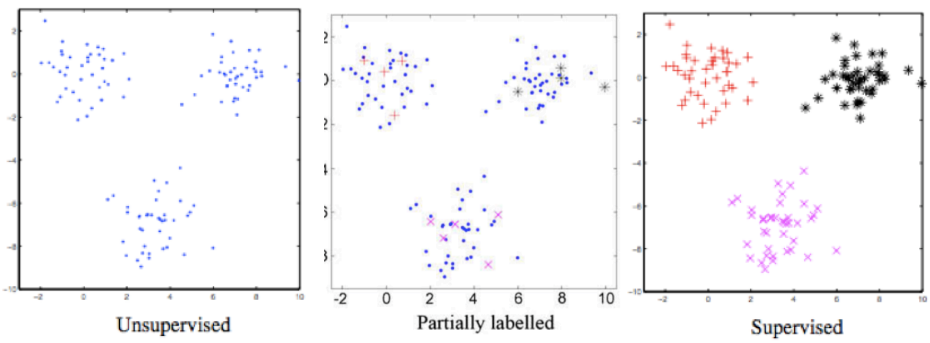
@minouye271
inouyelab.org

This lecture

- **Intro to clustering**
- **Hierarchical clustering**
- **K-means clustering**
- **Comparing clustering algorithms**
- **Intro to pathway analysis**
- **Gene ontology and KEGG pathways**
- **Gene set (GSEA) analysis**

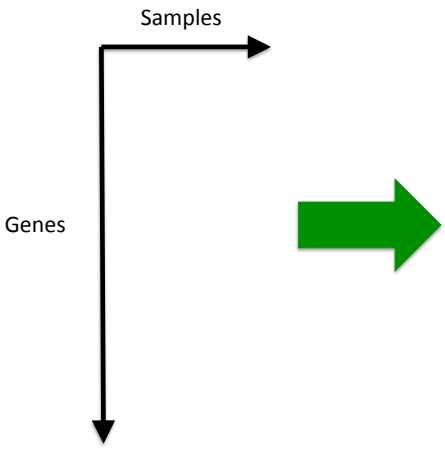
What is clustering?

- **Organization of data into groups**
- **Groups should be relevant to the question at hand**
- **Approaches**
  - Unsupervised
    - *No* knowledge of data labels
  - Semi-supervised
    - *Some* knowledge, e.g. a subset of data labels are known
  - Supervised
    - Technically not clustering (won't go over here, see machine learning)



Lange et al IEEE CVPR 2005

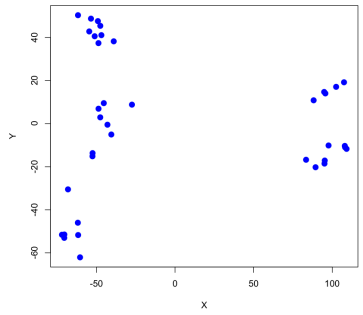
Structure of gene expression data



Are there groups of samples that are more similar to each other than others?

Same for genes

How might this be interpreted?



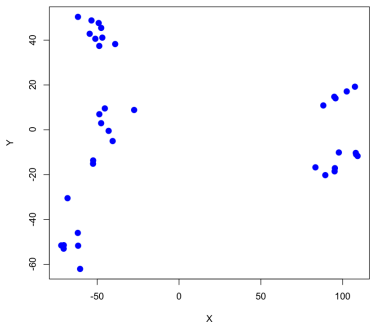
Clustering

- **Wide ranging and fundamental problem in data analysis, exploration, machine learning and other applied fields**
- **There is no 'correct' clustering approach and there is a degree of subjectivity**
  - Depends on what the data is and what is interesting
- **Detection of *clusters* of samples from expression of genes**
- **Detection of *clusters* of gene expression (pathways?) from the samples**
- **What methods can be used to do this?**

Clustering methods

- **There are dozens if not hundreds of methods**
- **Here, we focus on 2 of the most commonly used for gene expression**
  - Hierarchical
  - K-means

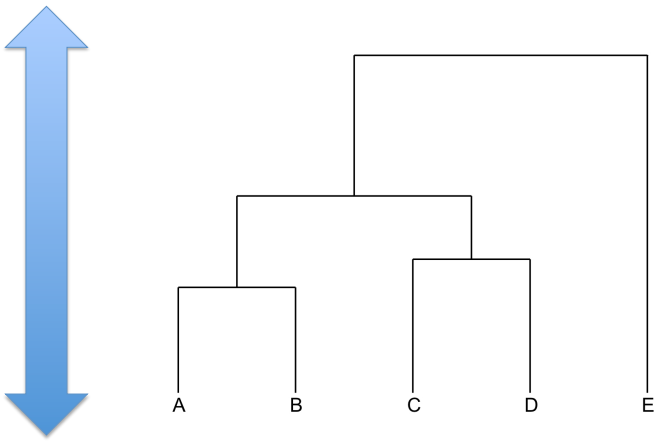
Hierarchical clustering



How do we model distances between individual data points? And groups of data points?

(1) Define a measure of dissimilarity (distance) between data points
(2) Define a 'linkage' criteria – determines distance between groups as function of distances between individual data points
(3) Bottom up vs top down

Dendrogram



Dissimilarity between data points

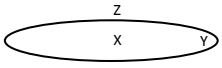
- **Euclidean distance**

$$d_{X,Y} = \sqrt{\Sigma_i (y_i - x_i)^2}$$

- **Manhattan distance**

$$d_{X,Y} = \Sigma_i |y_i - x_i|$$

- **Mahalanobis distance**



- **Correlation**

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho_{X,Y} = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2 \Sigma_i (y_i - \bar{y})^2}}$$

where

x and y are X and Y ordered by rank.

Linkage criteria

- **Given a distance measure…**
  - Complete linkage
    - Distance b/n clusters A and B is max(dist(a,b))
  - Single-linkage
    - Distance is min(dist(a,b))
  - Mean-linkage
    - Distance is mean b/n all pairwise relations in A and B

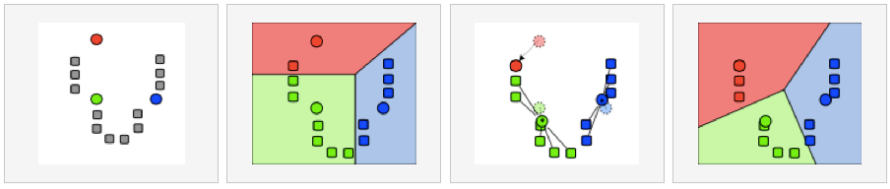
Hierarchical cluster algorithm: Bottom-up

- **Let X = {x1, x2, …, xn} be your data (genes)**
  - Initialize clustering level L(0)=0 & counter=0
  - Find min(dist(clust_i, clust_j)) over all pairs i,j of clusters at current level
  - Increment counter (counter=counter+1)
  - Set L(counter)=dist(clust_i,clust_j)
  - Update distance matrix with new cluster in place of old clusters
  - Loop… until all data points are in 1 cluster
- **Top down is reverse**

K-means clustering

- **Given a set of observations with k clusters, determine cluster centers and data point assignments such that the squared distances between cluster centers and data points are minimized (within-cluster sum of squares)**
- **Optimisation is NP-hard**
  - **Lloyd's algorithm (iterative refinement)**
    - Assign, calculate, update… converge when assignments unchanged
- **Performance bias towards similarly sized clusters**

K-means clustering



1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

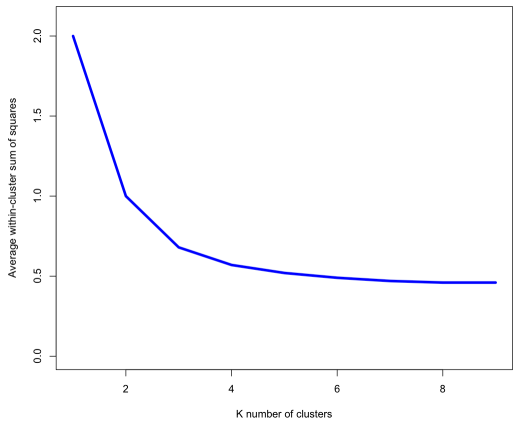
3) The centroid of each of the *k* clusters becomes the new mean.

4) Steps 2 and 3 are repeated until convergence has been reached.

Wikipedia

How many clusters are in the data?

- **Short answer: It depends… translation: we don't know but we can make an educated guess**



Which clustering method is 'best'?

- **Depends of what aspect(s) you care about most**
- **Best guide is to read method comparisons for gene expression data**
  - Thalamuthu et al *Bioinformatics* 22(19):2405 2006
  - Datta & Datta *Bioinformatics* 19(4): 459 2003
  - Yeung et al *Bioinformatics* 17(4):309 2001
  - Song et al *BMC Bioinformatics* 13:328 2012

Pathway analysis

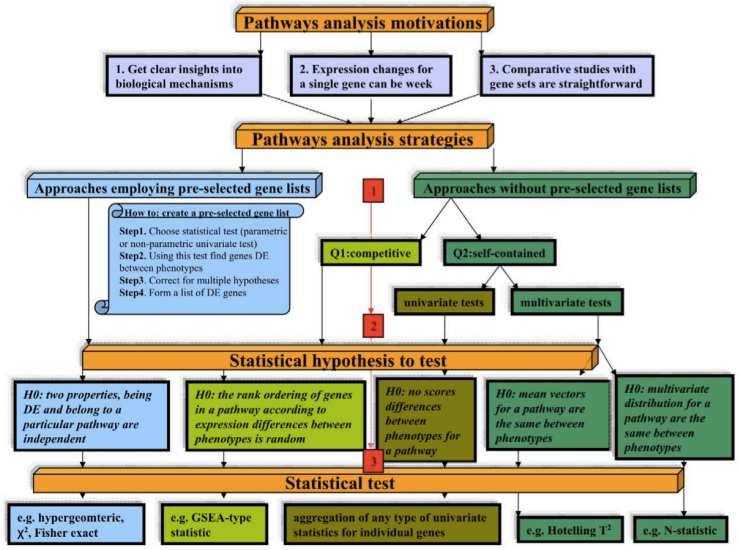
- **Derive a set or list of interesting genes derived from data, make inferences about biological function**
  - Exercise caution due to arbitrary thresholds and power
- **Given sets of genes from external source, assess expression changes of sets in the data**

What tools are there?

Name	Organism[a]	Application Type	URL
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/erminej/
Eu.Gene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html
GOAL	H, M	Web server	http://microarrays.unife.it
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/
MEGO	H	Windows standalone	http://www.dxy.cn/mego/
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html
T-profiler	Y, CA	Web server	http://www.t-profiler.org/

Nam & Kim, *Briefings in Bioinf*

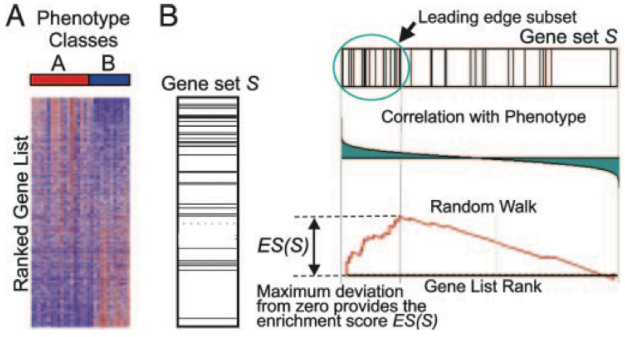
Overview of strategies



Emmert-Streib & Glazko, *PLoS Comp Bio* 2011

GSEA: Gene set enrichment analysis

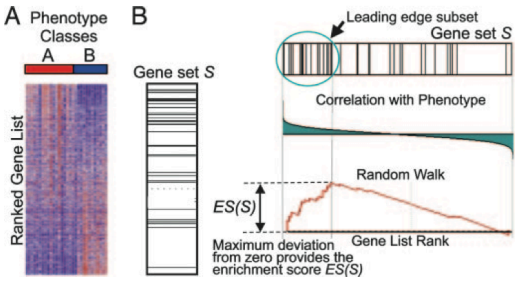
- **Molecular Signatures Database (MSigDB)**
  - Positional, curated, cis-reg motifs, data mining, GO sets, oncogenic, immunological



Subramanian et al *PNAS* 2005

GSEA enrichment score (ES)

- **N genes and k samples**
- **Create a list *L* by ranking the N genes by correlation with phenotype**
- **Walk down list *L* and increase/decrease a running-sum statistic if a gene is/isn't in a predefined set *S***
  - Magnitude of increase/decrease depends on correlation of gene with phenotype
- **ES is the max deviation from zero during random walk**
  - Sort of like a weighted KS statistic
  - ES_null is distribution generated from permutations of phenotypes
  - FDR



Gene set databases and meta-dbs

- **MSigDB**
- **Gene Ontology**
- **KEGG**
- **PLAGE**
- **PAGED**
- **GSA**
- **Gazer**
- **ErmineJ**
- **ASSESS**
- **GeneSetDB**

Reaction Networks

- **Reactions form networks** of interactions in the cell.
- They are presented in databases as **biological pathways**.

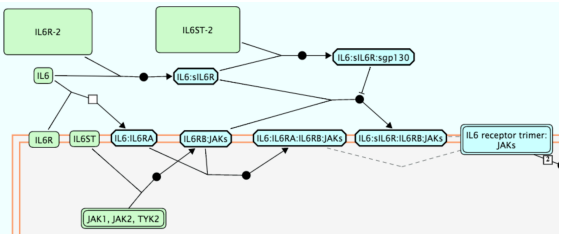
Biological pathway: "A biological pathway is a series of actions among molecules in a cell that leads to a certain product or change in a cell." – US NHGRI

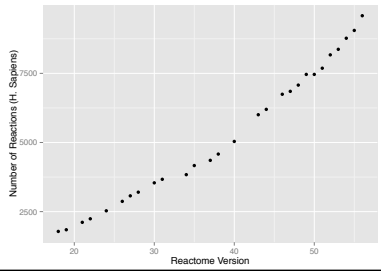
- Reactions and their participants (proteins, complexes) can appear in multiple pathways; this causes **crosstalk**.
- There are **no canonical pathways** – every database, textbook and researcher will differ (sometimes subtly).
- Pathways are **not independent** due to crosstalk, but are often treated as such (e.g. overrepresentation analysis).
- Differ from ontologies/gene sets – they describe both **what** species are involved **and how** they interact.

Reaction Networks

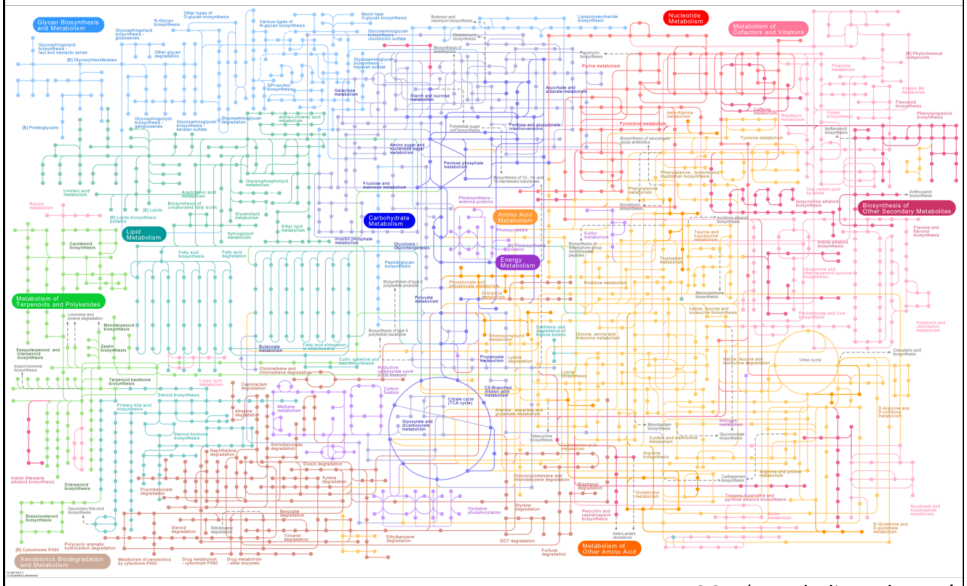
REACTOME
A CURATED PATHWAY DATABASE

- Covers **signal transduction**, **metabolism** and **other cellular processes** (e.g. DNA repair, transcription...).
- 9,238 proteins, 9,422 complexes in 9,584 reactions across 2,007 human pathways (June 2016).
- Annotated with **subcellular locations**; some information about **PTMs** and **disease**.





Reaction Networks - KEGG



KEGG – 'Metabolic Pathways'
http://www.genome.jp/kegg-bin/show_pathway?map01100