LARGE SCALE DATA GENERATION AND ANALYSIS

Advance Cell and Molecular Biology (ICG601)

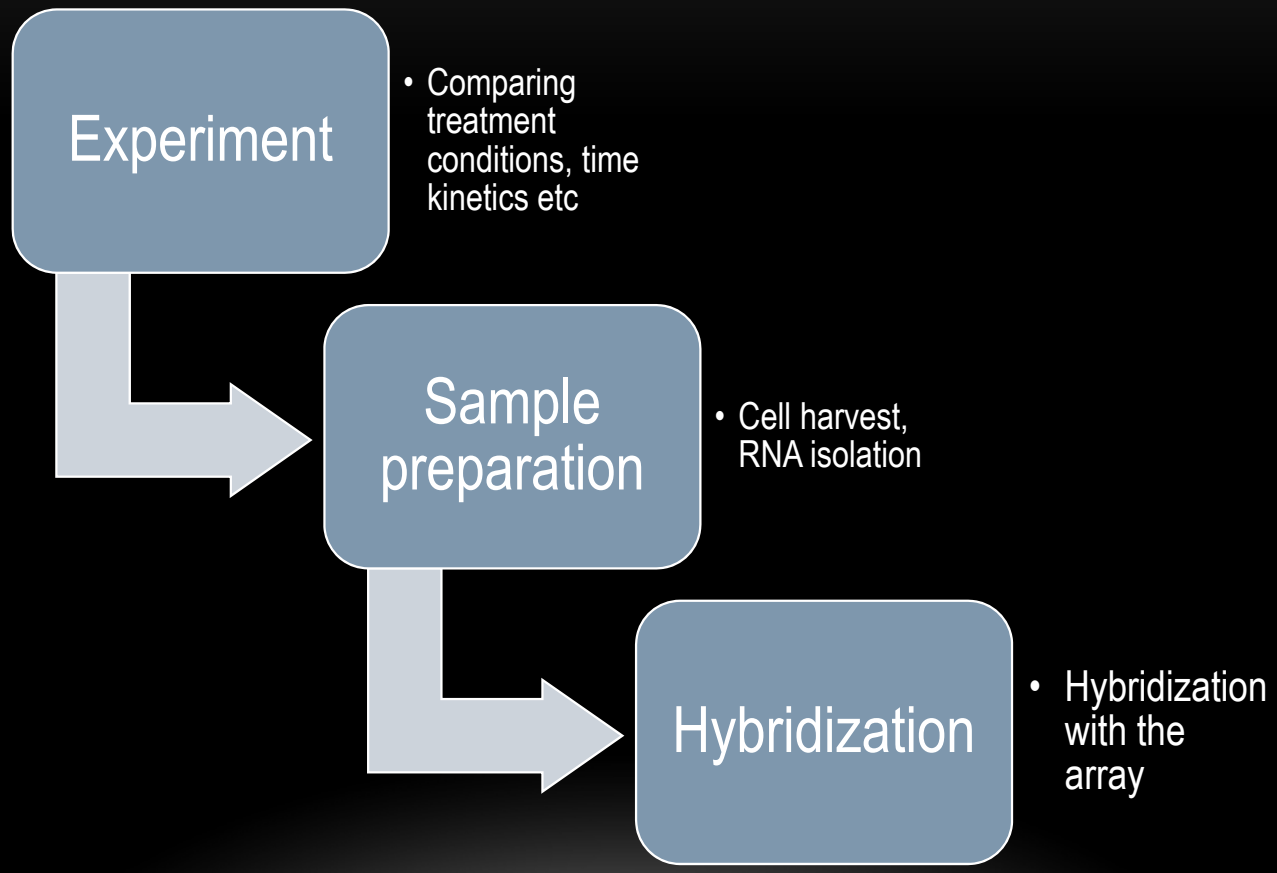
Dr. Dhiraj Kumar

Immunology Group

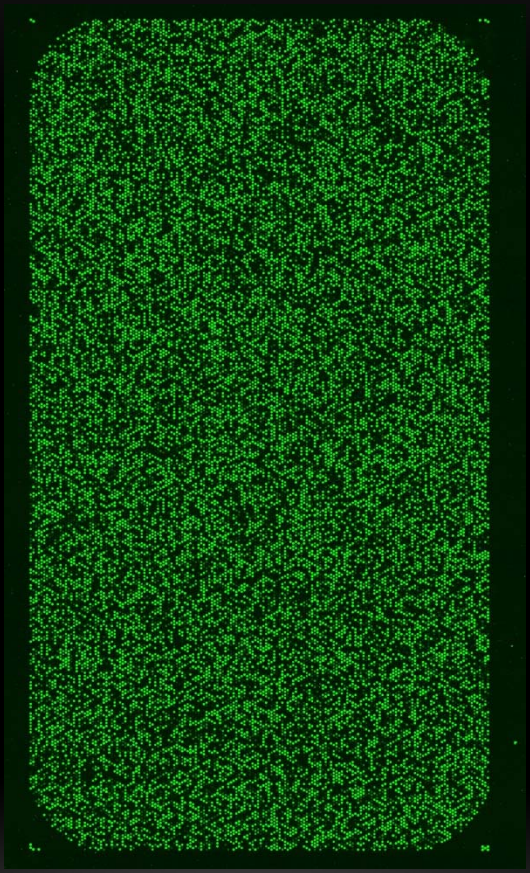
LARGE SCALE DATA ANALYSIS

- Microarray data Analysis
- Pathway mapping and enrichment
- Clustering (clustergram, PCA, k-means cluster)
- Functional annotation of clusters
- Analysis of high-content screening data (Hit identification and validation)

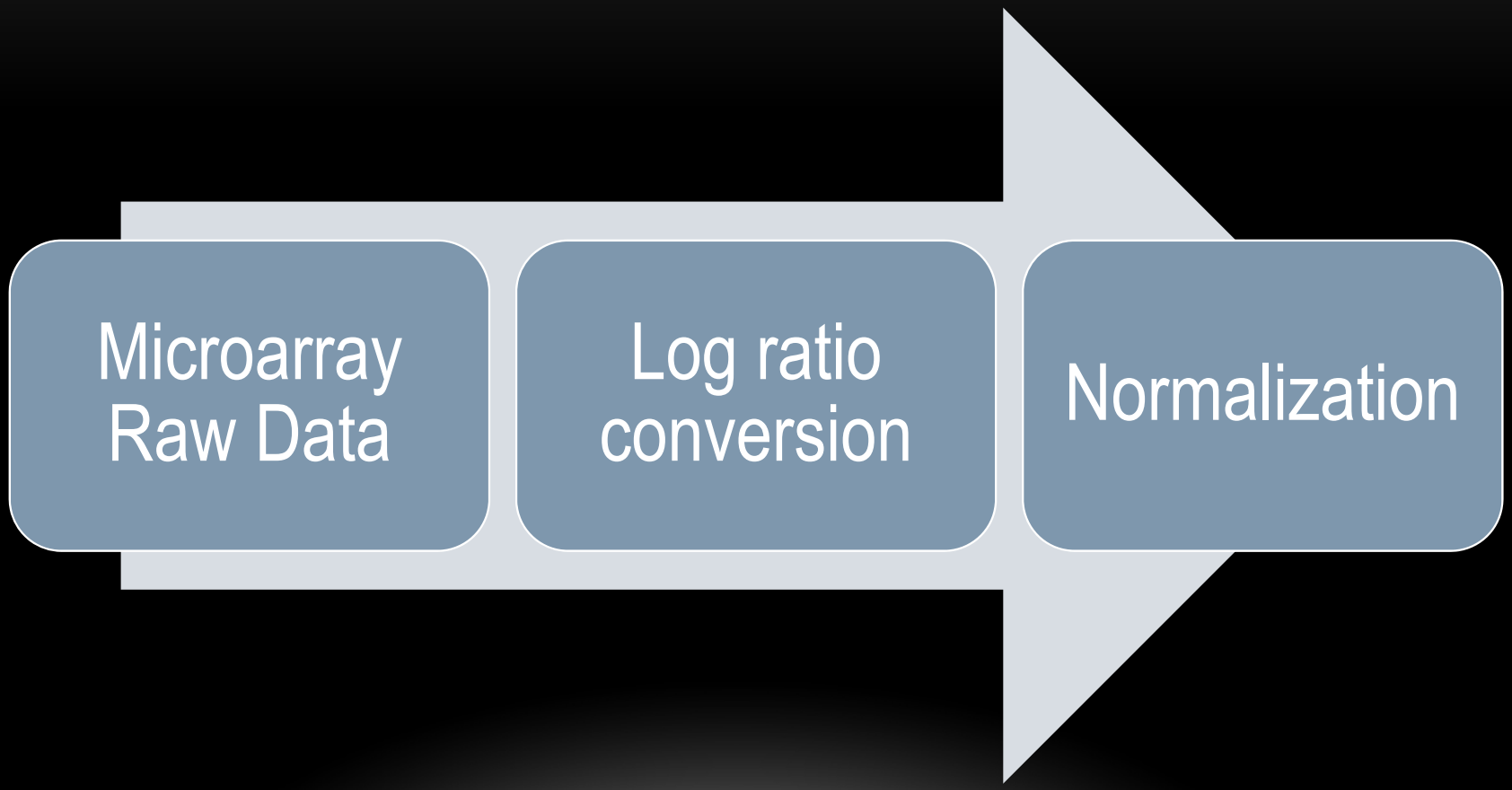
MICROARRAY EXPERIMENT



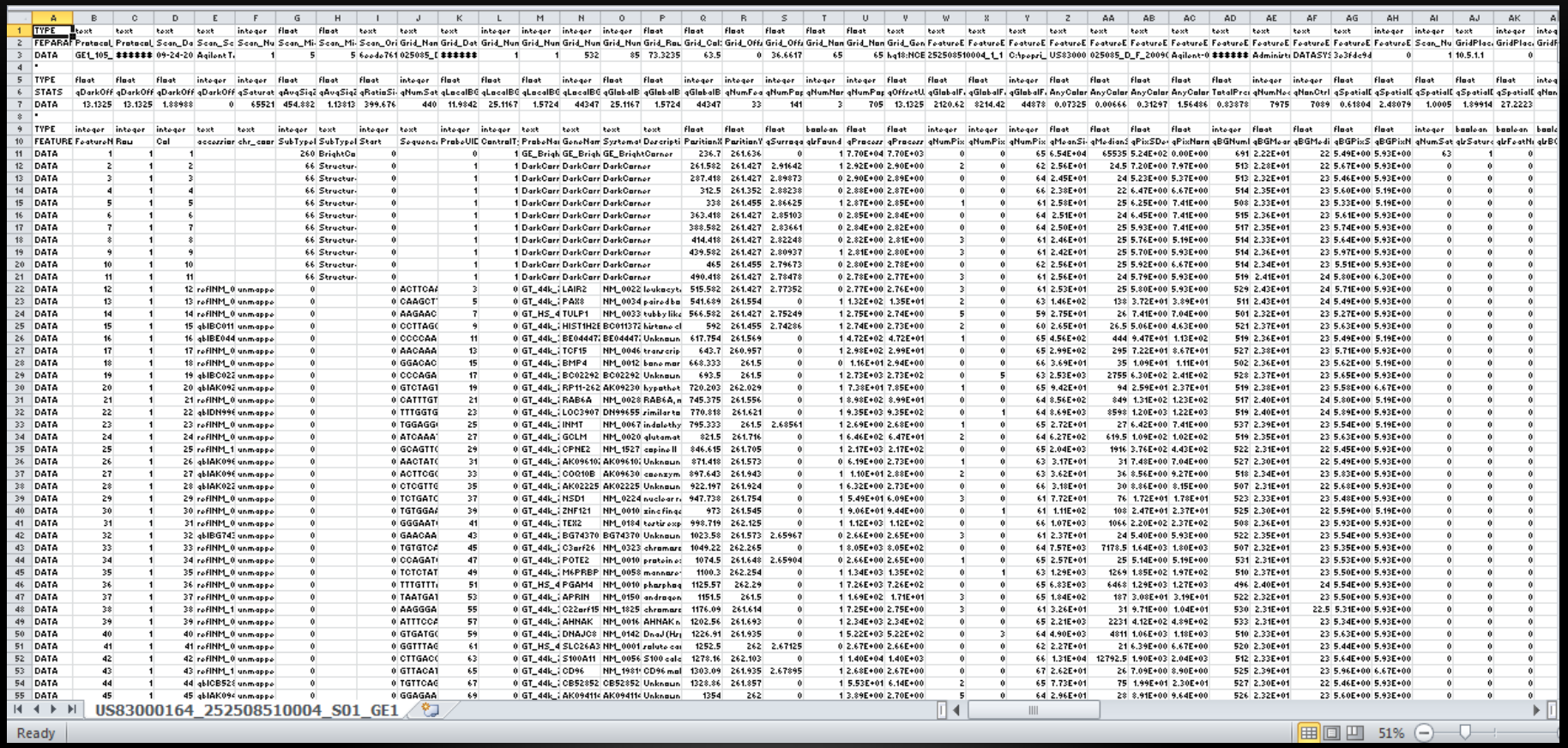
MICROARRAY EXPERIMENT



MICROARRAY FLOW CHART



RAW MICROARRAY DATA



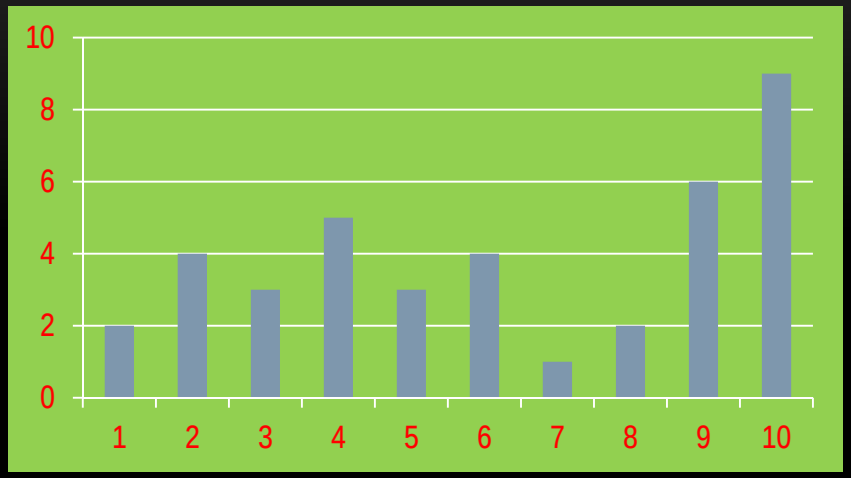
US83000164_252508510004_S01_GE1

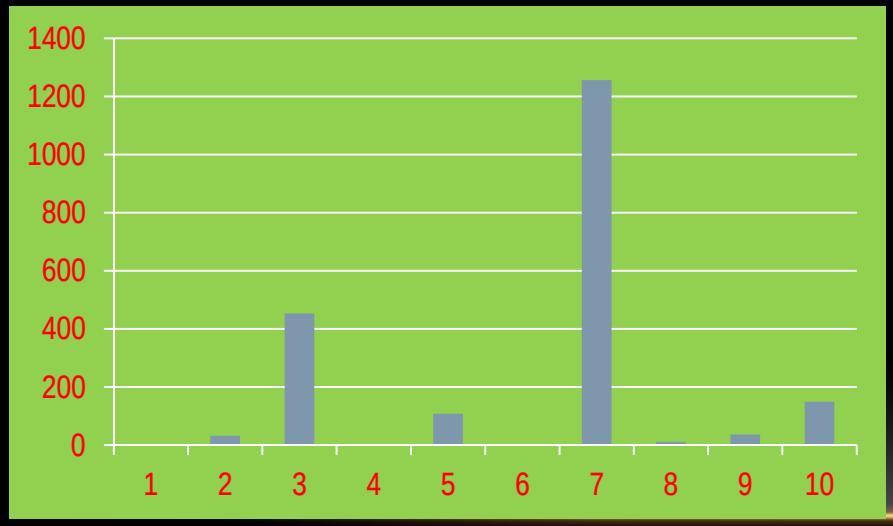
Ready

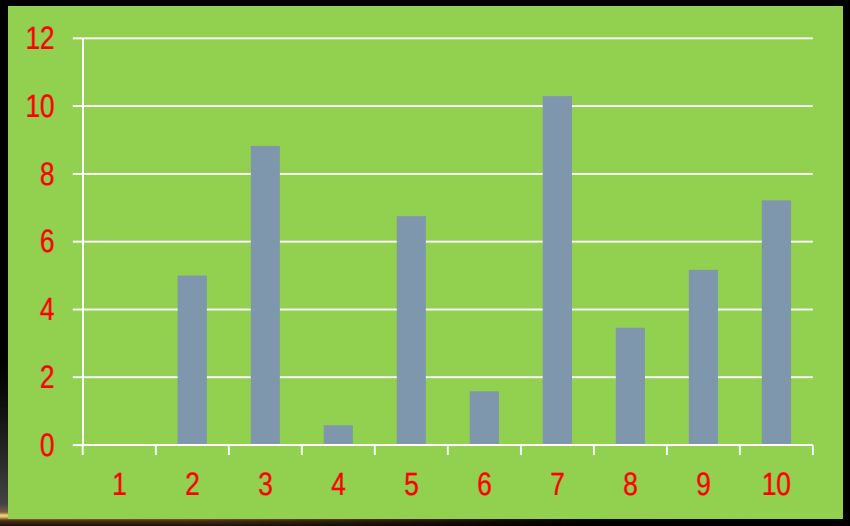
51%

LOG RATIO CONVERSION

Why log ratio conversion?







For analyzing data sets with large deviations

LOG RATIO CONVERSION

1
32
453
1.5
108
3
1256
11
36
149

$$\mathrm{Log}_2(x)=\log_{10}(x)/\log_{10}(2)$$

0
5
8.823367
0.584963
6.754888
1.584963
10.29462
3.459432
5.169925
7.219169

NORMALIZATION

What is normalization???

Comparisons have to be made on similar objects

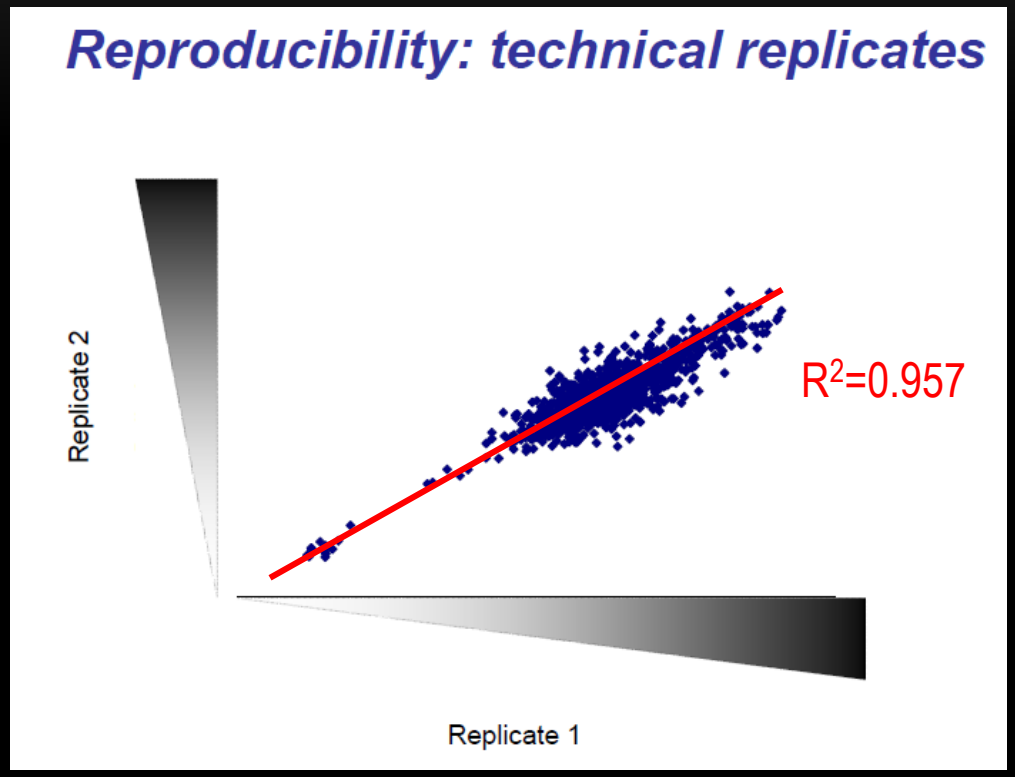
Two approaches:

- Median of the data set
- Corresponding value from control experiment

MEAN, MEDIAN AND MODE

ADVANCE CELL AND MOLECULAR BIOLOGY
(ICG601)

REPRODUCIBILITY OF RESULTS IN LARGE DATA SETS



For duplicate data sets, calculate R^2 value

FOR MOST BIOLOGICAL STUDIES TRIPLICATE DATA SET IS DESIRABLE

Single Observation

- Unreliable data
- No inferences

Duplicate observations

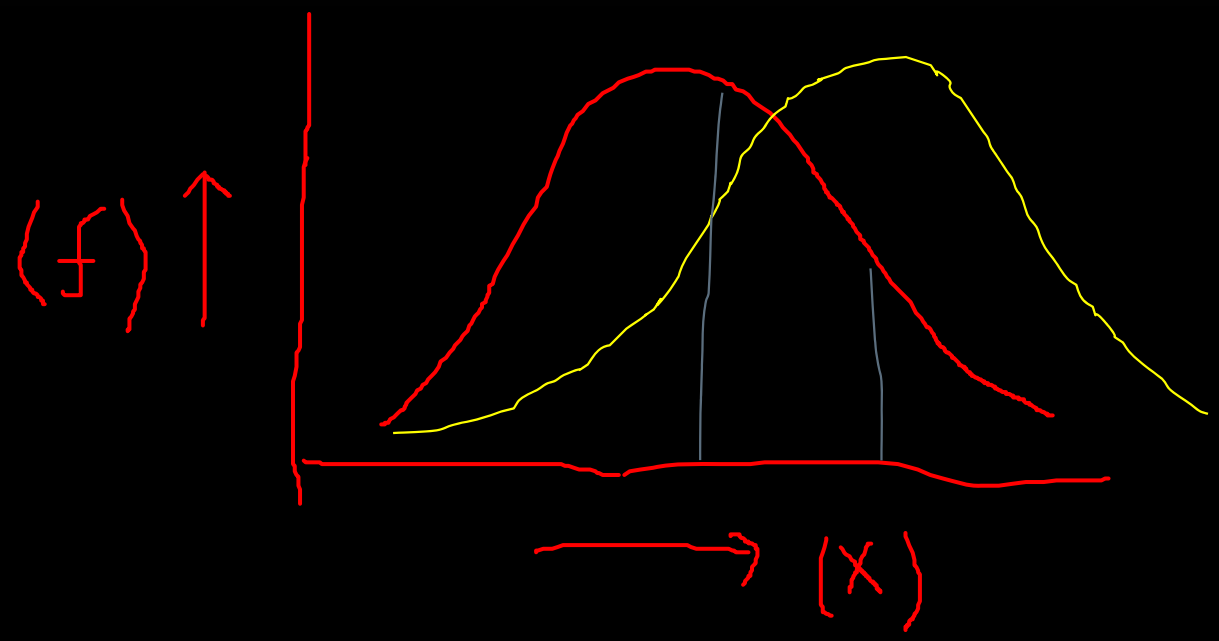
- Get average
- Not very reliable, extreme values??

Triplicate observations

- Get p-values
- Reliable information

P-VALUE

Normal distribution



P-value tells about whether differences in the observation are real or due to sampling bias!!!

USING P-VALUE IN BIOLOGICAL EXPERIMENTS

- Comparing two or more observation/sampling sets

- Identifying whether individuals in a sampling represent same population?

Z-SCORE NORMALIZATION

Z-SCORE

Observations (X)	Z-score
23	-0.60047
26	-0.15012
19	-1.20094
18	-1.35106
34	1.050821
29	0.300235
41	2.101643
12	-2.25176
27	0
32	0.750587
32	0.750587
30	0.450352
28	0.150117
29	0.300235
27	0
28	0.150117
24	-0.45035

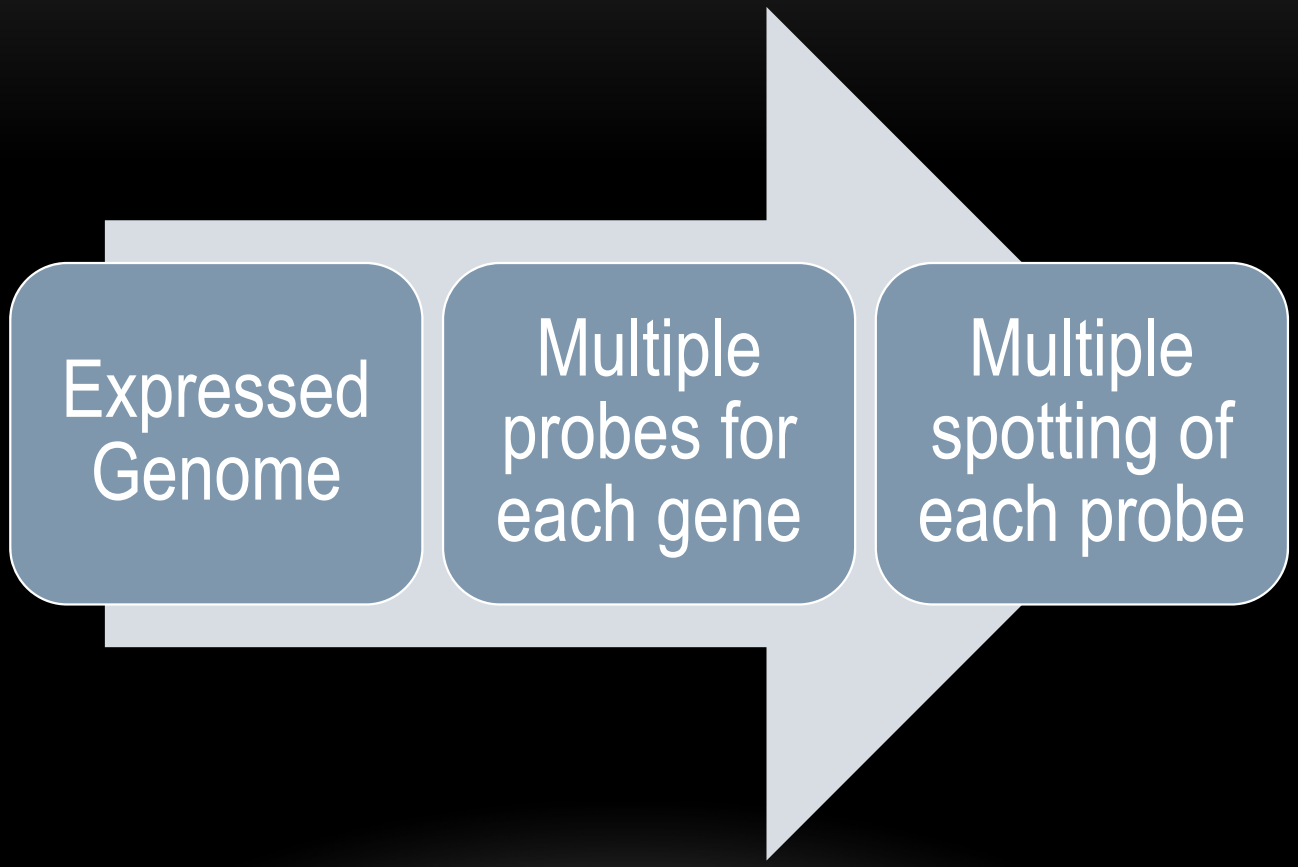
Mean ($\overline{X}$)= 27

Standard Deviation (σ)=6.661456

Z-score= $(X-\overline{X})/\sigma$

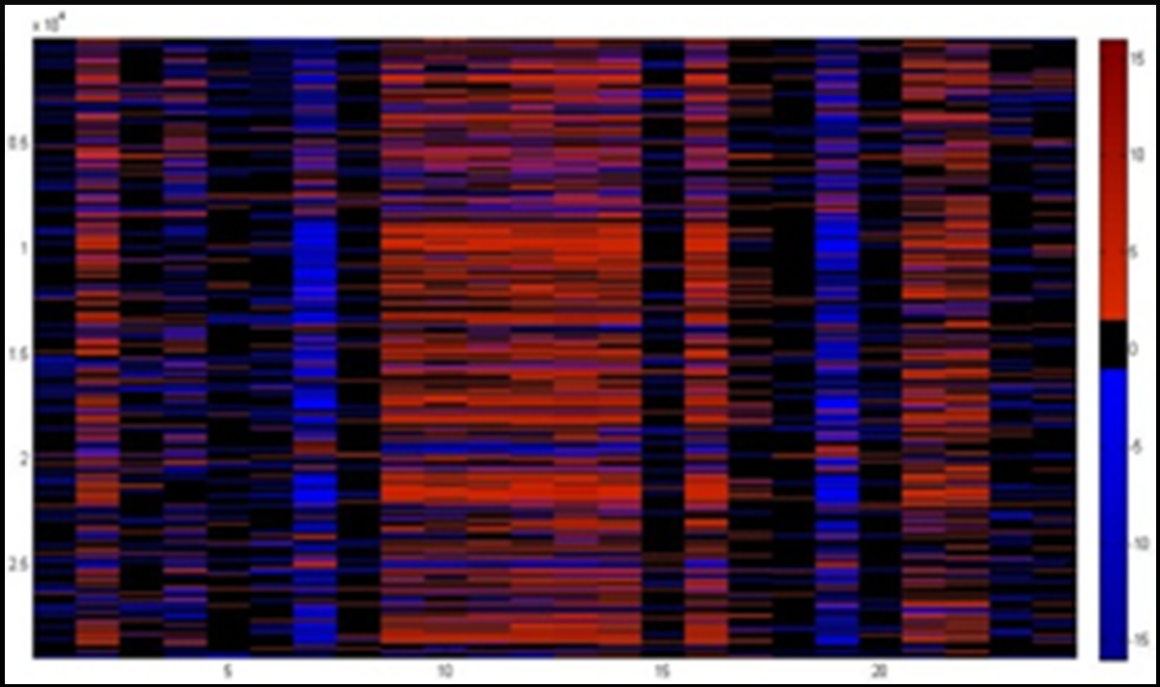
In a different experiment, where mean is different
One can still compare results at the z-score level

REDUNDANCY IN MICROARRAY DATA



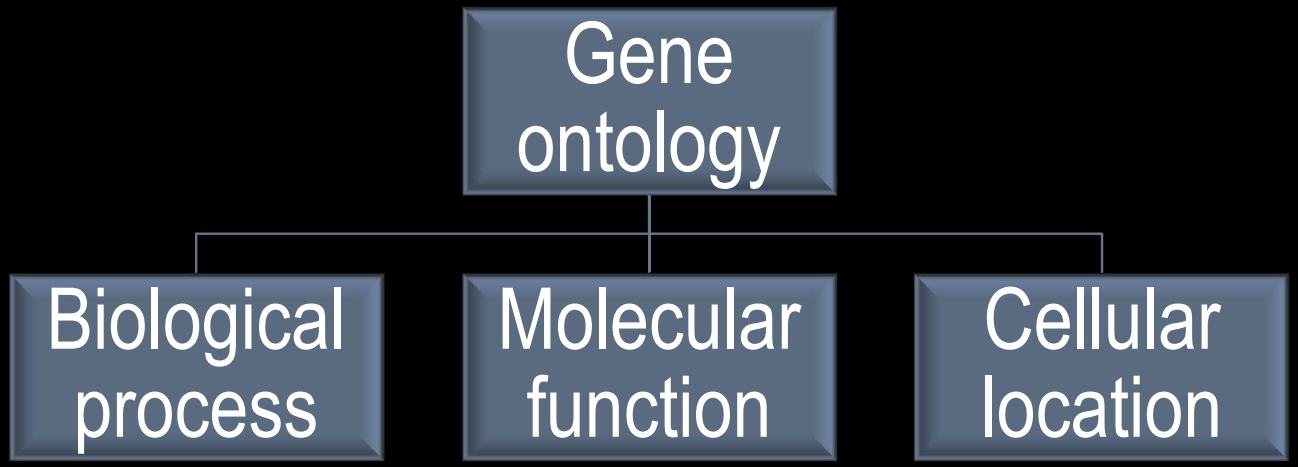
Select non-redundant data set guided by p-value!!!

GENOME-WIDE EXPRESSION DATA



What genome-wide expression data offers?

GENE ONTOLOGY ENRICHMENT (FUNCTIONAL ANNOTATION TO THE PERTURBED GENE SET)

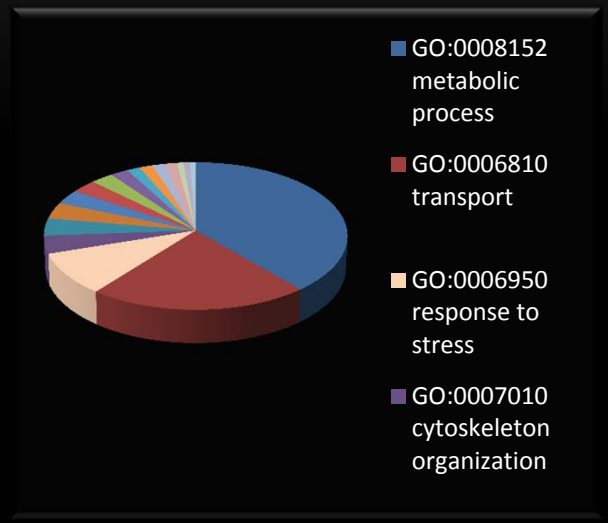


GENE ONTOLOGY ENRICHMENT (FUNCTIONAL ANNOTATION TO THE PERTURBED GENE SET)

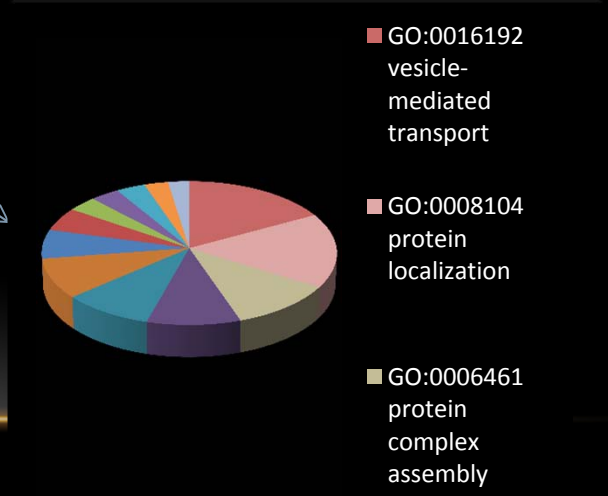
25k genes

3700 up-regulated, 1200 down-regulated

Up



Down



Functional distinction among up- and down-regulated genes

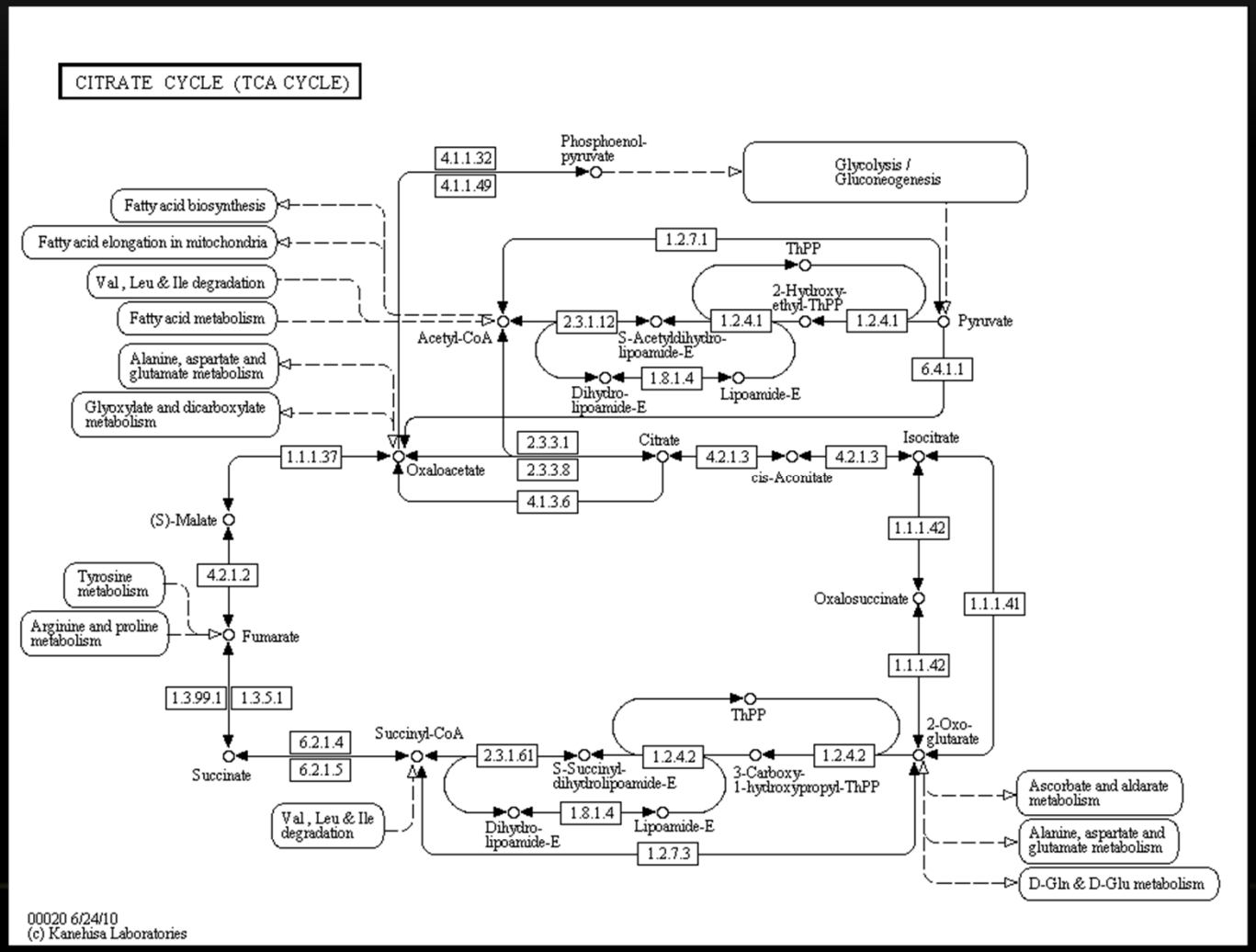
PATHWAY ANALYSIS

How different pathways are effected under given condition?

Slightly more specific than just ontology analysis
e.g. (metabolic function versus glycolysis or TCA, signaling versus cytokine and mitogen signaling)

DATABASE OF ALL KNOWN BIOLOGICAL PATHWAYS, INCLUDING BIOCHEMICAL AND SIGNALING PATHWAYS

TCA CYCLE IN KEGG DATABASE



PATHWAY ENRICHMENT USING ‘R’ PACKAGE

Identify genes significantly regulated at any given time point/tissue/treatment

↓

Sub-pathway-Miner in ‘R’

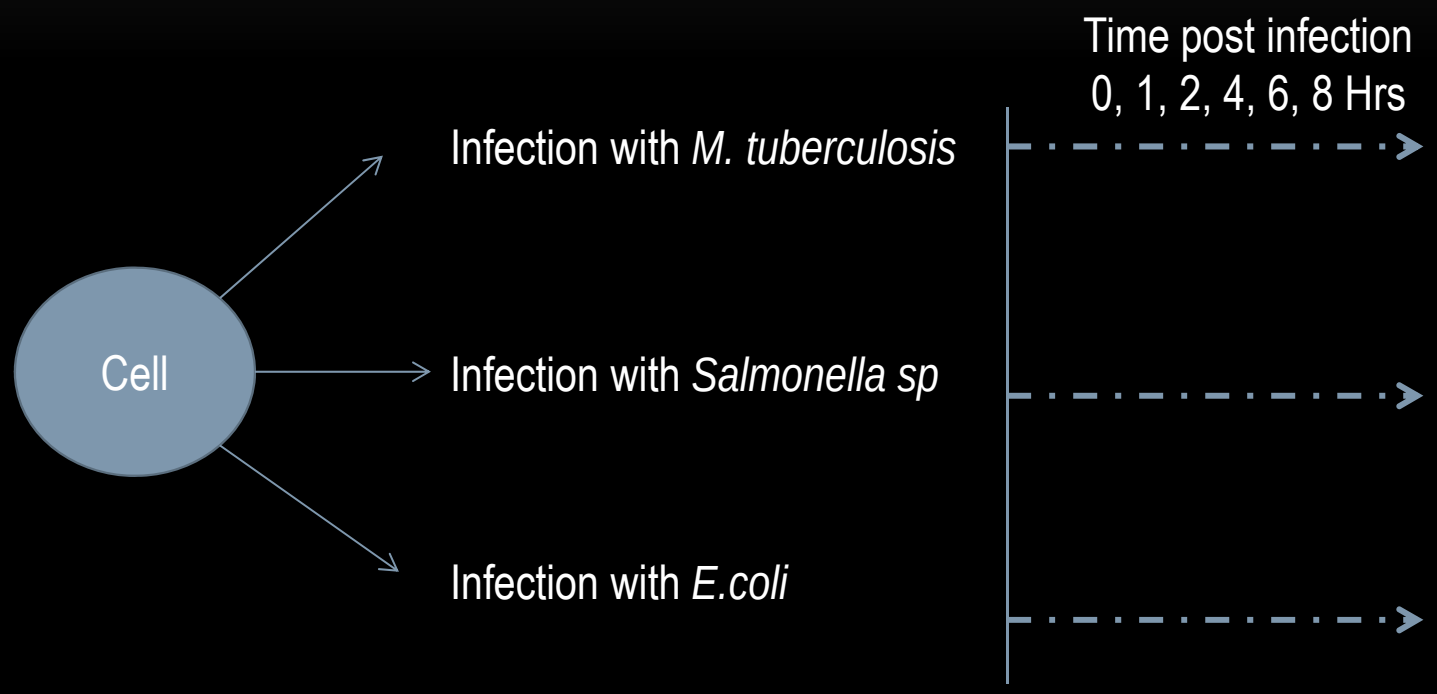
↓

Score gene enrichment of KEGG Pathways
(confidence score >90%; p-value<0.1)

PATHWAY ENRICHMENT

KAL20_W6			
Ribosome	47/6236	117/23798	0.000665
Hypertrophic cardiomyopathy (HCM)	34/6236	87/23798	0.005763
Dilated cardiomyopathy	34/6236	89/23798	0.008609
Malaria	22/6236	54/23798	0.013929
Cardiac muscle contraction	30/6236	81/23798	0.020682
Osteoclast differentiation	41/6236	118/23798	0.02452
Carbohydrate digestion and absorption	16/6236	39/23798	0.031037
Leishmaniasis	24/6236	65/23798	0.037147
Cytosolic DNA-sensing pathway	21/6236	56/23798	0.041678
Nitrogen metabolism	10	23/23798	0.054629
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	26/6236	74/23798	0.055957
Toll-like receptor signaling pathway	34/6236	101/23798	0.058012
Graft-versus-host disease	21/6236	58/23798	0.059828
Type II diabetes mellitus	18/6236	49/23798	0.068245
Pancreatic secretion	35/6236	106/23798	0.070727
Viral myocarditis	31/6236	93/23798	0.076191
Salivary secretion	26/6236	77/23798	0.08593
Aldosterone-regulated sodium reabsorption	16/6236	44/23798	0.089444
PPAR signaling pathway	27/6236	81/23798	0.093069
Biosynthesis of unsaturated fatty acids	10	25/23798	0.093244
Bile secretion	24/6236	71/23798	0.095105
Chagas disease	33/6236	102/23798	0.098182

KINETIC EXPRESSION DATA OR MULTIPLE TREATMENT CONDITIONS

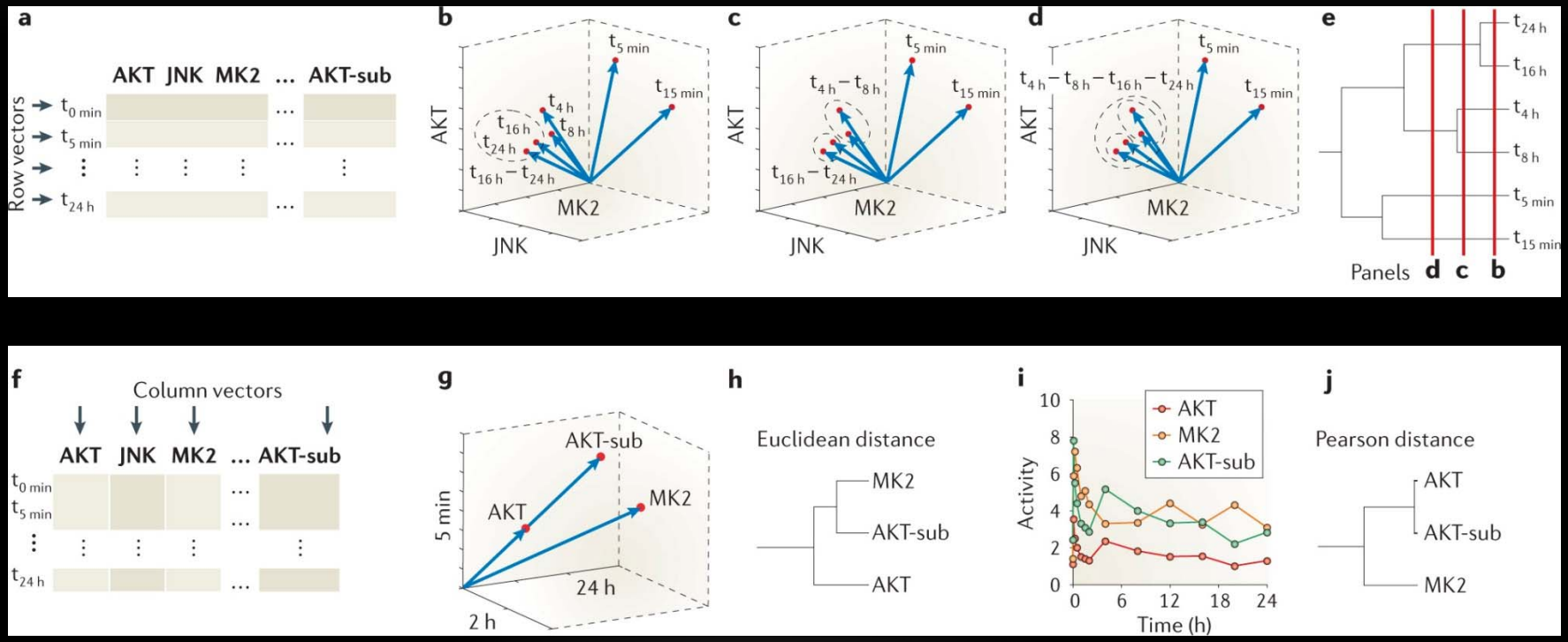


Multivariate analysis

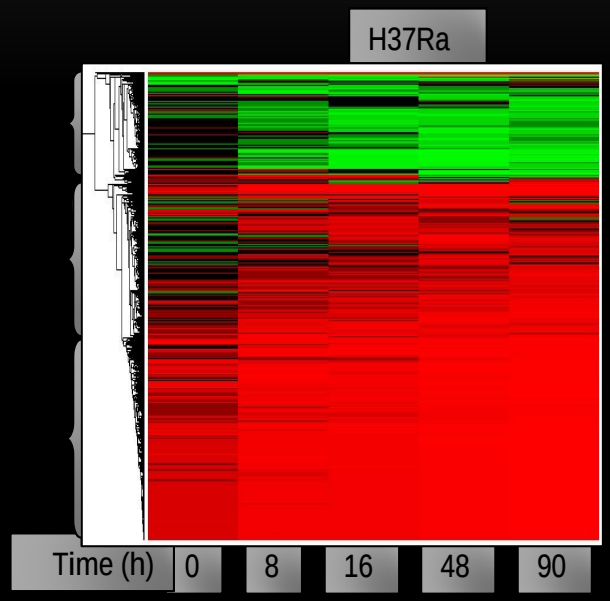
MULTIVARIATE ANALYSIS OF COMPLEX EXPRESSION EXPERIMENTS (IDENTIFYING PATTERNS)

- Clustergram
- K-means clusters
- Principal components

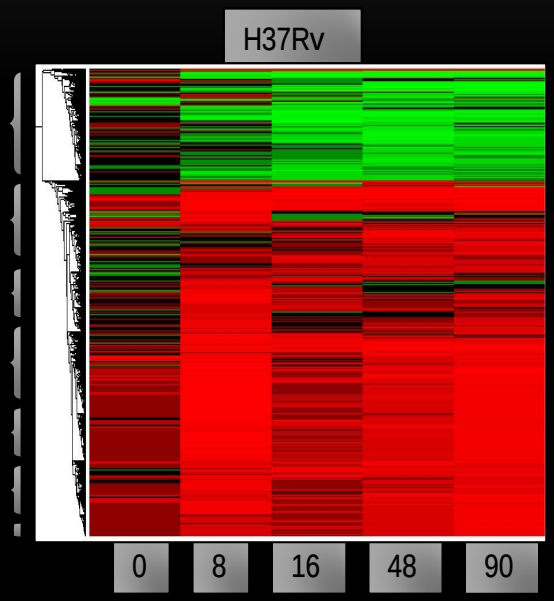
CLUSTERING MULTIVARIATE DATA



IDENTIFY GENES SHOWING SIMILAR EFFECT ON EXPRESSION UPON TREATMENT



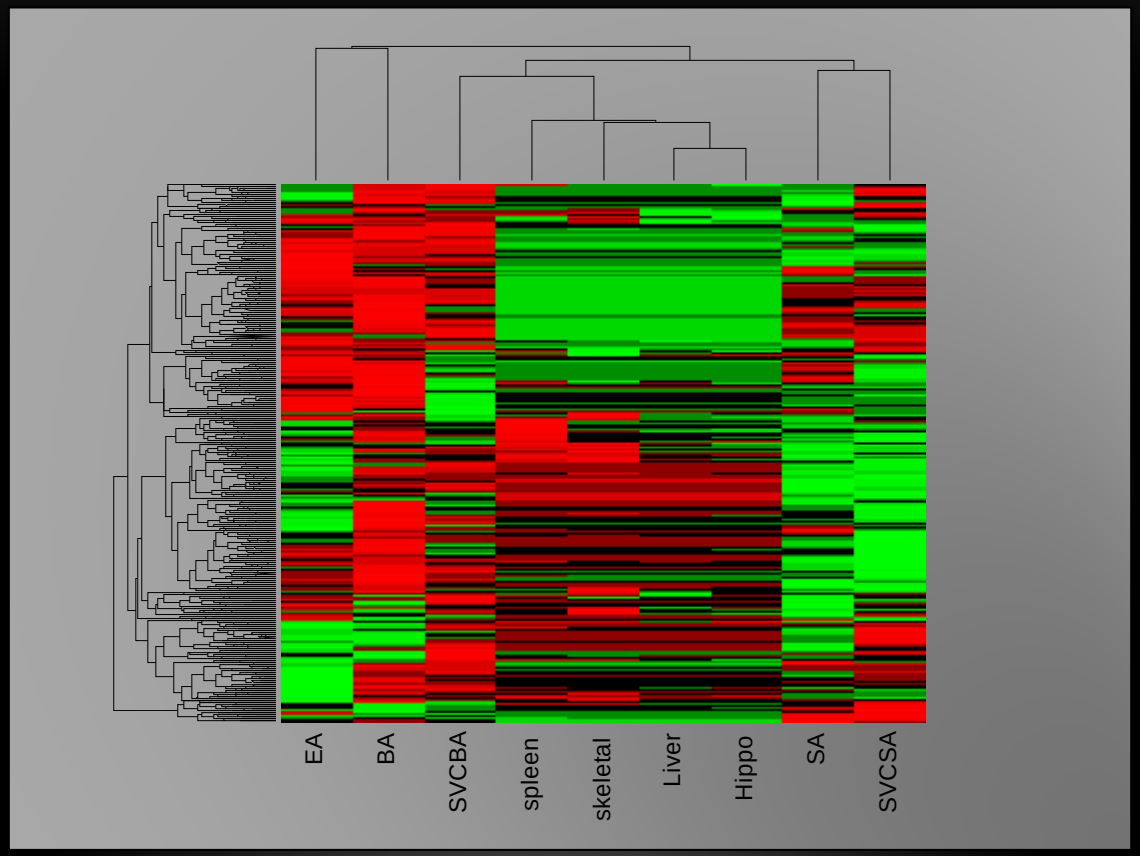
3 clusters



7 clusters

One-way clustergram: based on similarity among genes

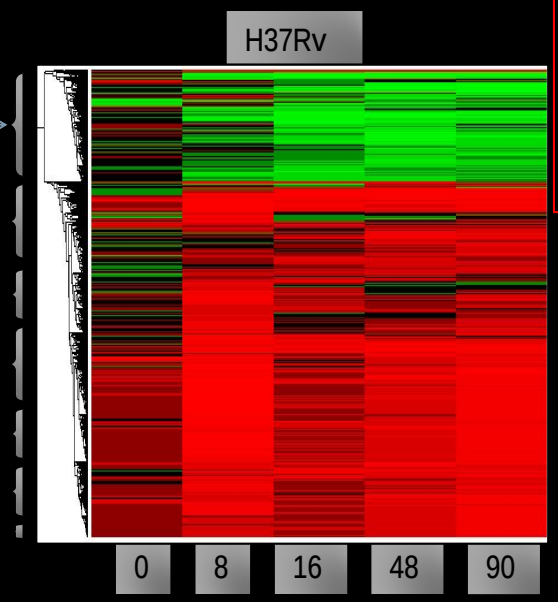
TWO-WAY CLUSTERGRAM: SIMILARITY AMONG GENES AS WELL AS TREATMENTS



Identifies cross relationship among treatment/groups

FUNCTIONAL ENRICHMENT OF CLUSTERS

GO (BP), p-value

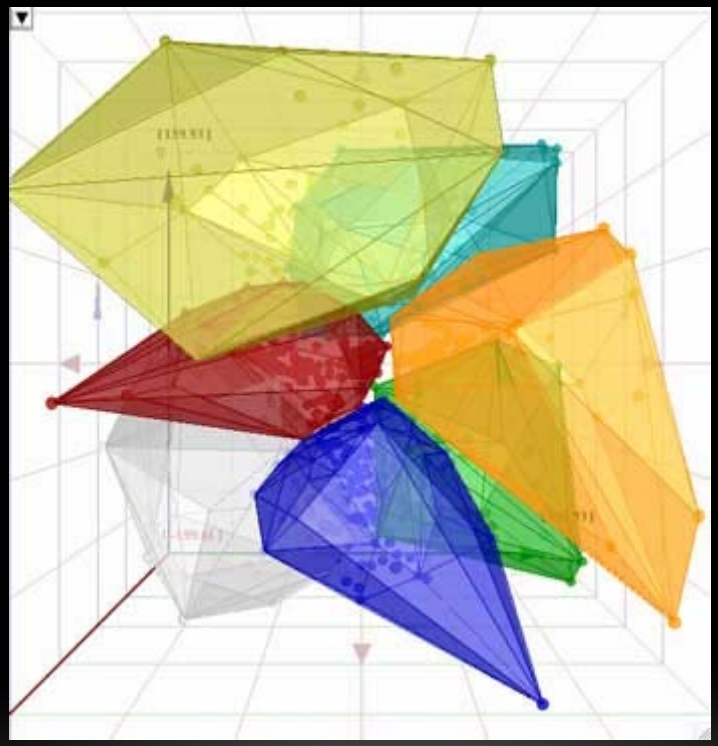


Pro-inflammatory ($p<0.005$)
Immune function ($p<0.01$)
Metabolic ($p<0.05$)
Anti-microbial pathways ($p<0.001$)
Anti-apoptosis ($p<0.01$)

How to get p-values here?

K-MEANS CLUSTERS

Pre-defined number of clusters (bins)



Centroid of the data set

PRINCIPAL COMPONENT ANALYSIS (PCA)

- We have too many observations and dimensions
  - To reason about or obtain insights from
  - To visualize
  - Too much noise in the data
  - Need to “reduce” them to a smaller set of factors
  - Better representation of data without losing much information
  - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition

- Combinations of observed variables may be more effective bases for insights, even if physical meaning is obscure

UNDERLYING CONCEPT

If two items or dimensions are highly correlated or dependent

- They are likely to represent highly related phenomena

UNDERLYING CONCEPT

- What if the dependences and correlations are not so strong or direct?
- Suppose you have 3 variables, or 4, or 5, or 10000?
- Look for the phenomena underlying the observed covariance/co-dependence in a set of variables
  - Once again, phenomena that are uncorrelated or independent, and especially those along which the data show high variance
- These phenomena are called “factors” or “principal components” or “independent components,” depending on the methods used
  - Factor analysis: based on variance/covariance/correlation
  - Independent Component Analysis: based on independence

Principle

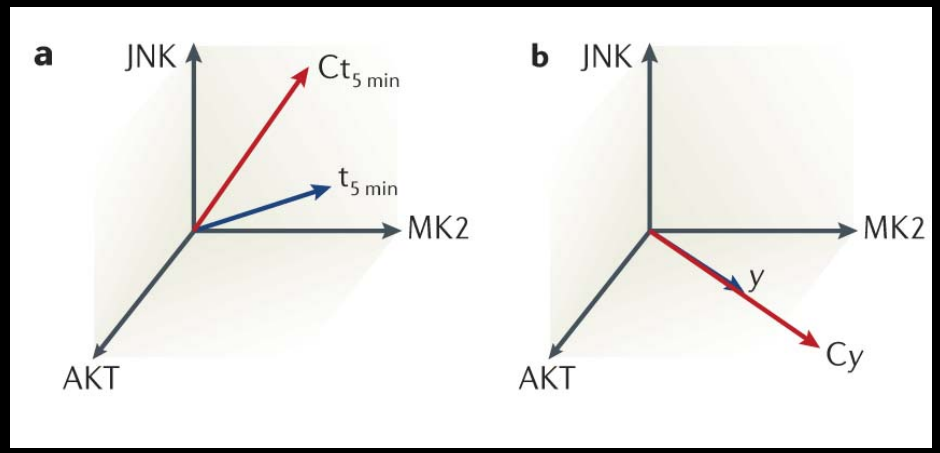
- Linear projection method to reduce the number of parameters
- Transfer a set of correlated variables into a new set of uncorrelated variables
- Map the data into a space of lower dimensionality
- Form of unsupervised learning

PRINCIPAL COMPONENT

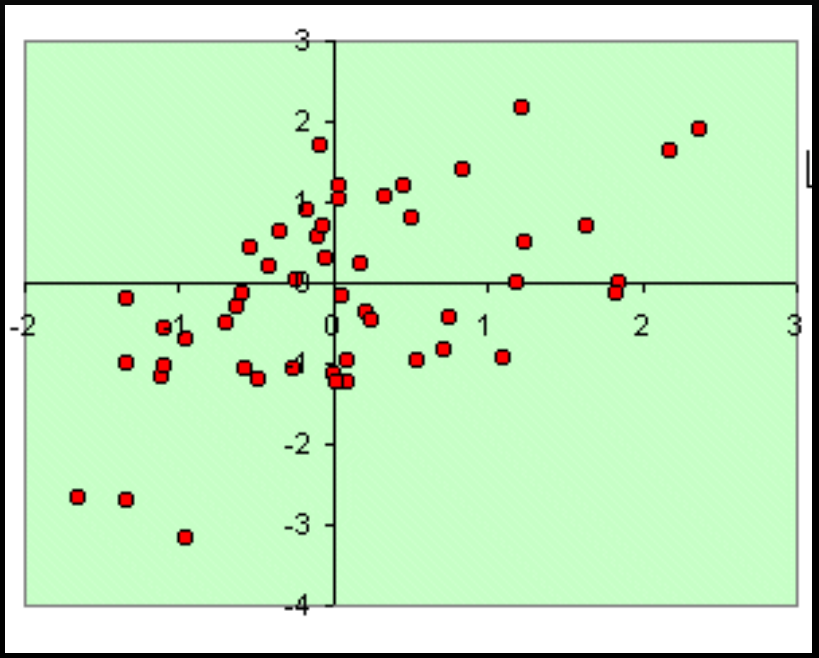
- direction of maximum variance in the input space
- principal eigenvector of the covariance matrix

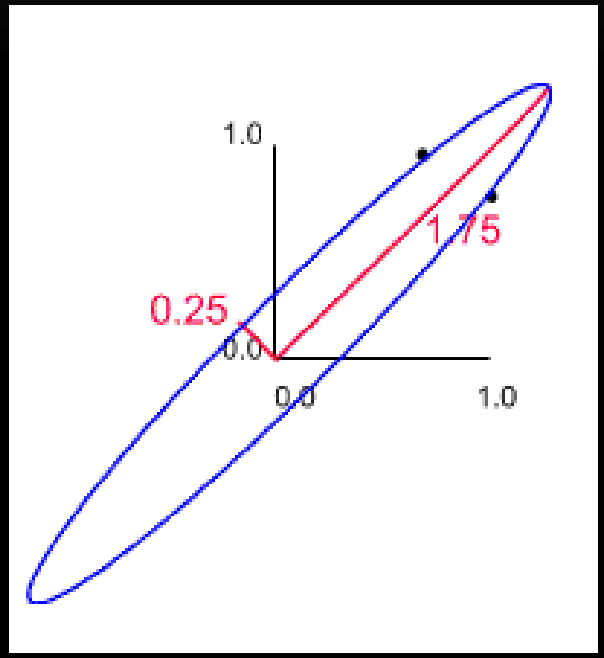
EIGENVALUES AND EIGENVECTORS

$$C = \begin{bmatrix} \text{cov(AKT, AKT)} & \text{cov(AKT, JNK)} & \text{cov(AKT, MK2)} \\ \text{cov(JNK, AKT)} & \text{cov(JNK, JNK)} & \text{cov(JNK, MK2)} \\ \text{cov(MK2, AKT)} & \text{cov(MK2, JNK)} & \text{cov(MK2, MK2)} \end{bmatrix}$$

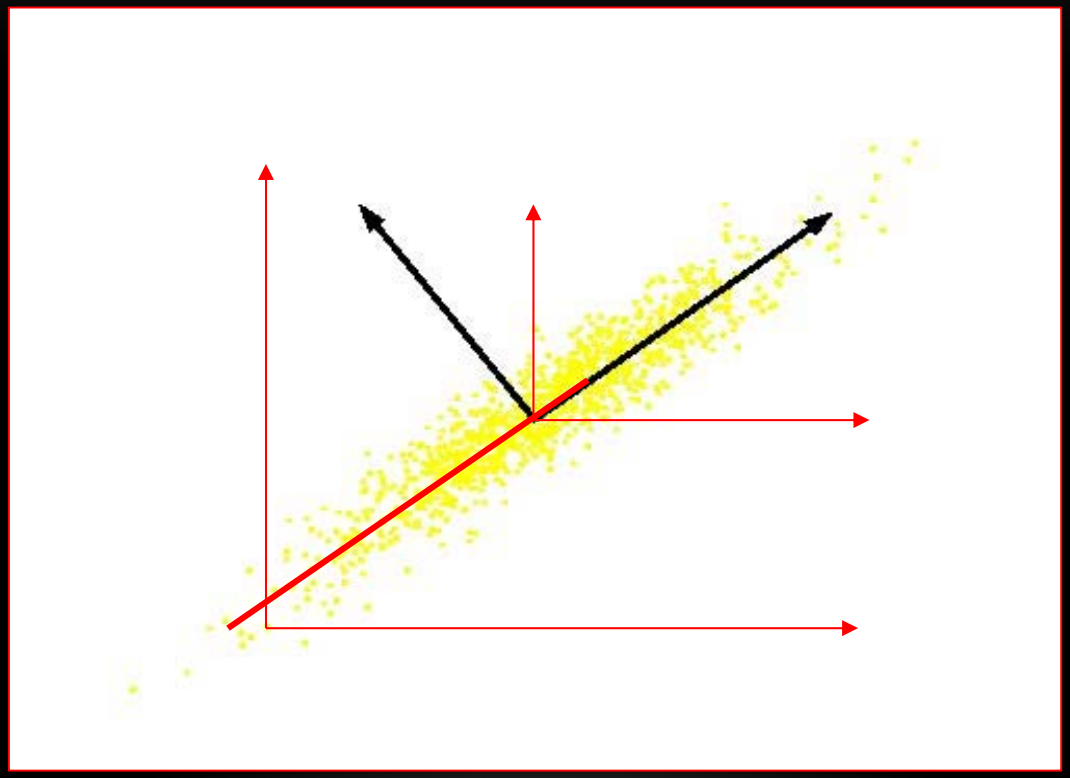


PRINCIPAL COMPONENTS





PCA



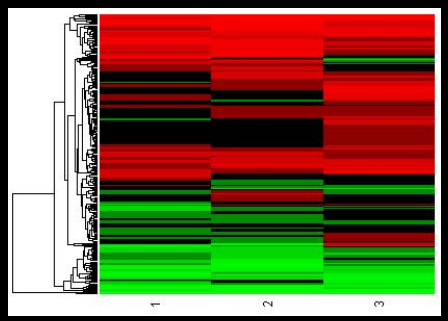
THE OUTCOME OF UNSUPERVISED ANALYSES

Identify salient features of the data and hence of the process under study like

- How many genes are effected in a given process
- How global the effect/influence of a process or treatment is
- Vastness of cellular readjustments to direct the process/responses

- Pathway analysis and GO analysis helps validate the result in view of earlier findings
- Combining clustering with functional analysis may reveal novel associations
- Helps postulate newer hypothesis

MECHANISTIC INSIGHTS FROM GENE EXPRESSION DATA



Clustergram

Regulatory Motifs in Clusters

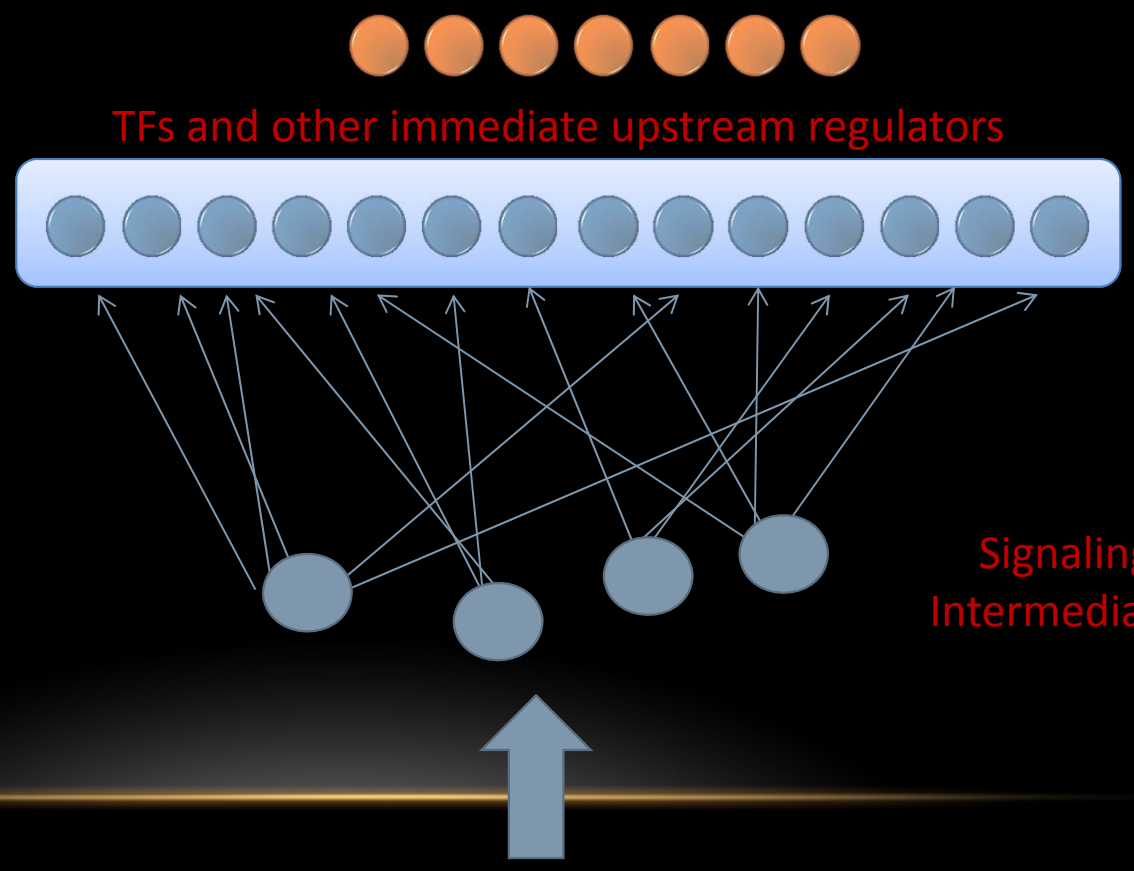
Can we Identify TFs that could be crucial for this set of genes?

UPSTREAM ANALYSIS FROM TF LIST (TRANSFAC)

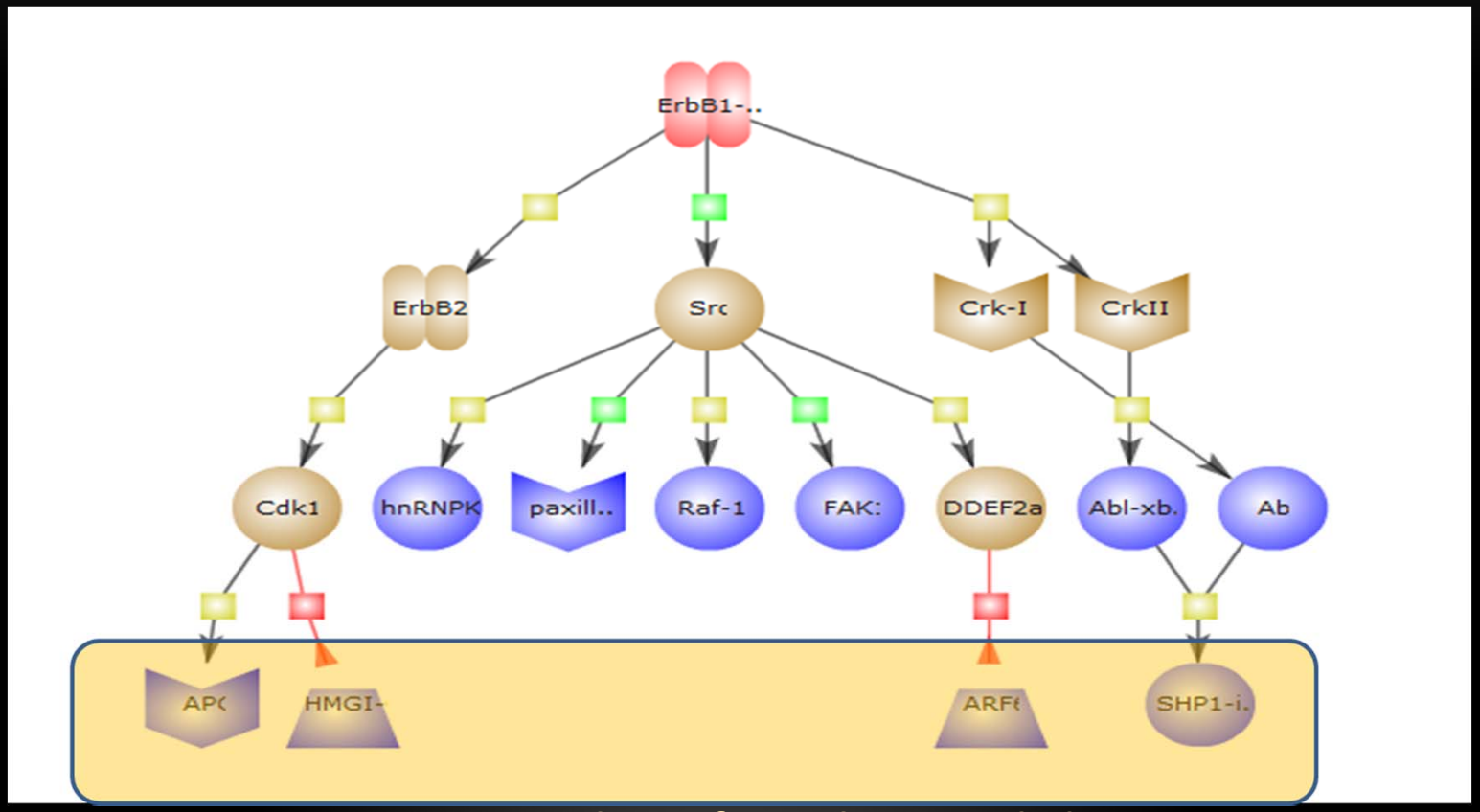
Gene Set

Transfac (Bobase)

Explain



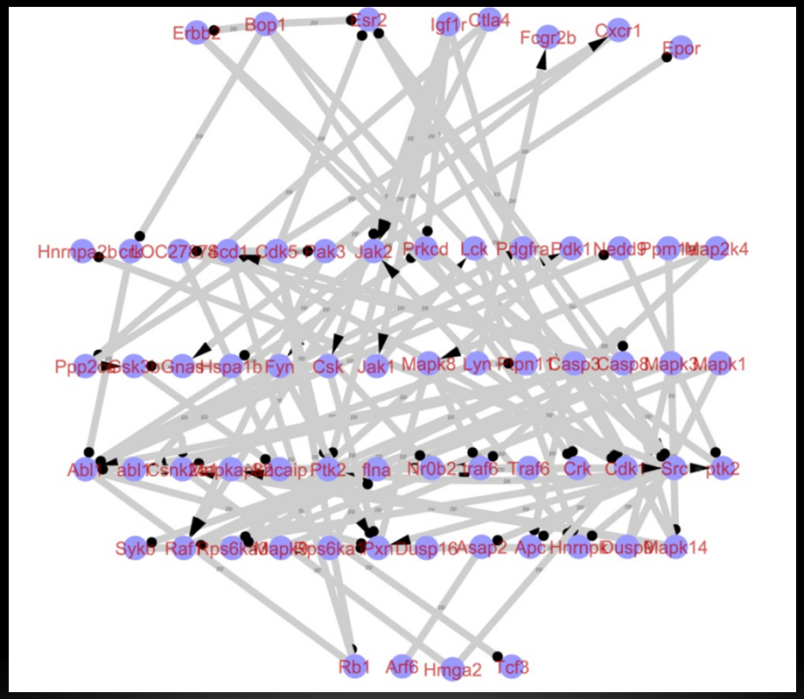
ERBB1 AS KEY NODE FOR A GIVEN GENE SET



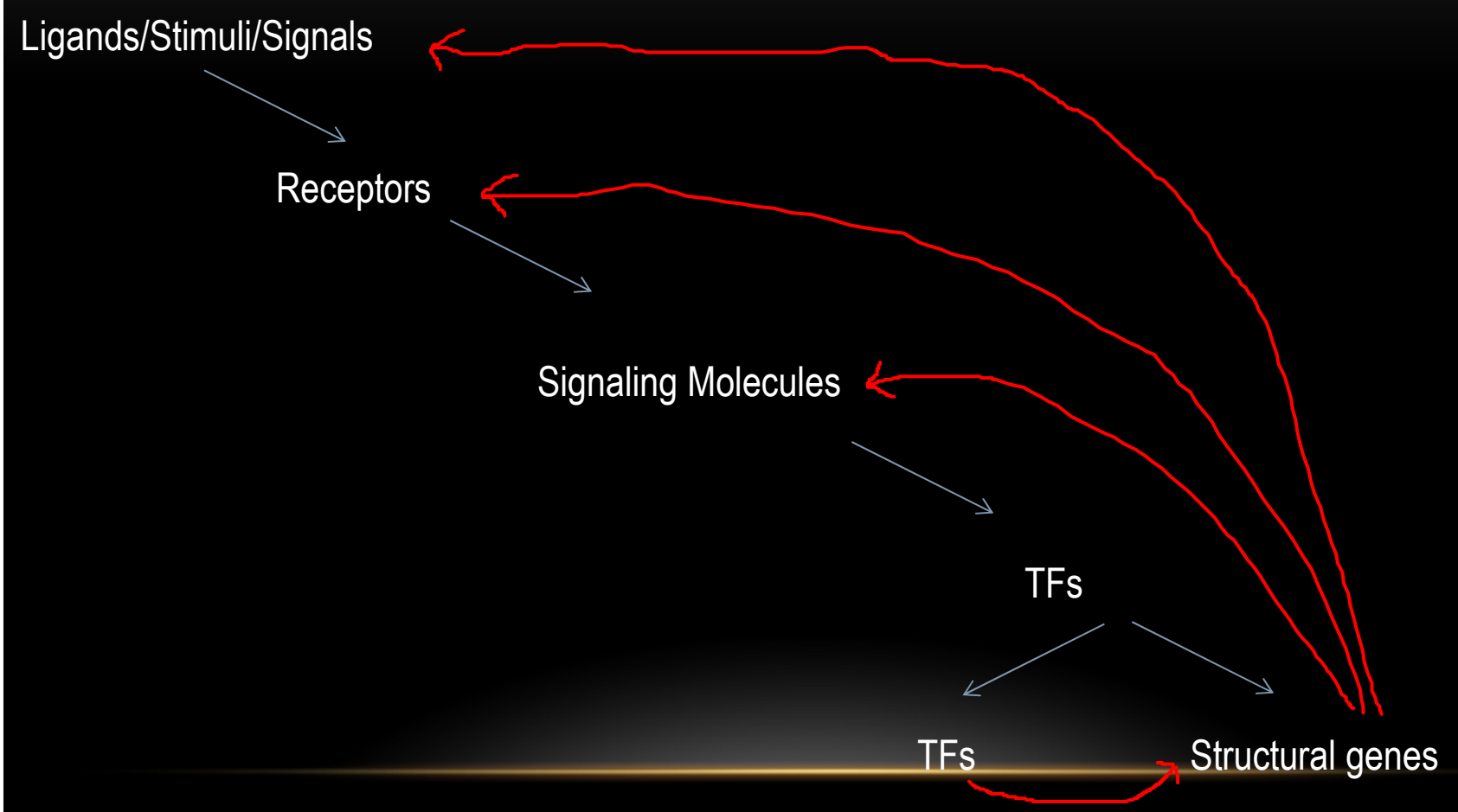
Upstream to subset of gene list provided

IF WE HAVE MANY GENE SETS (IDENTIFIED AS CLUSTERS)

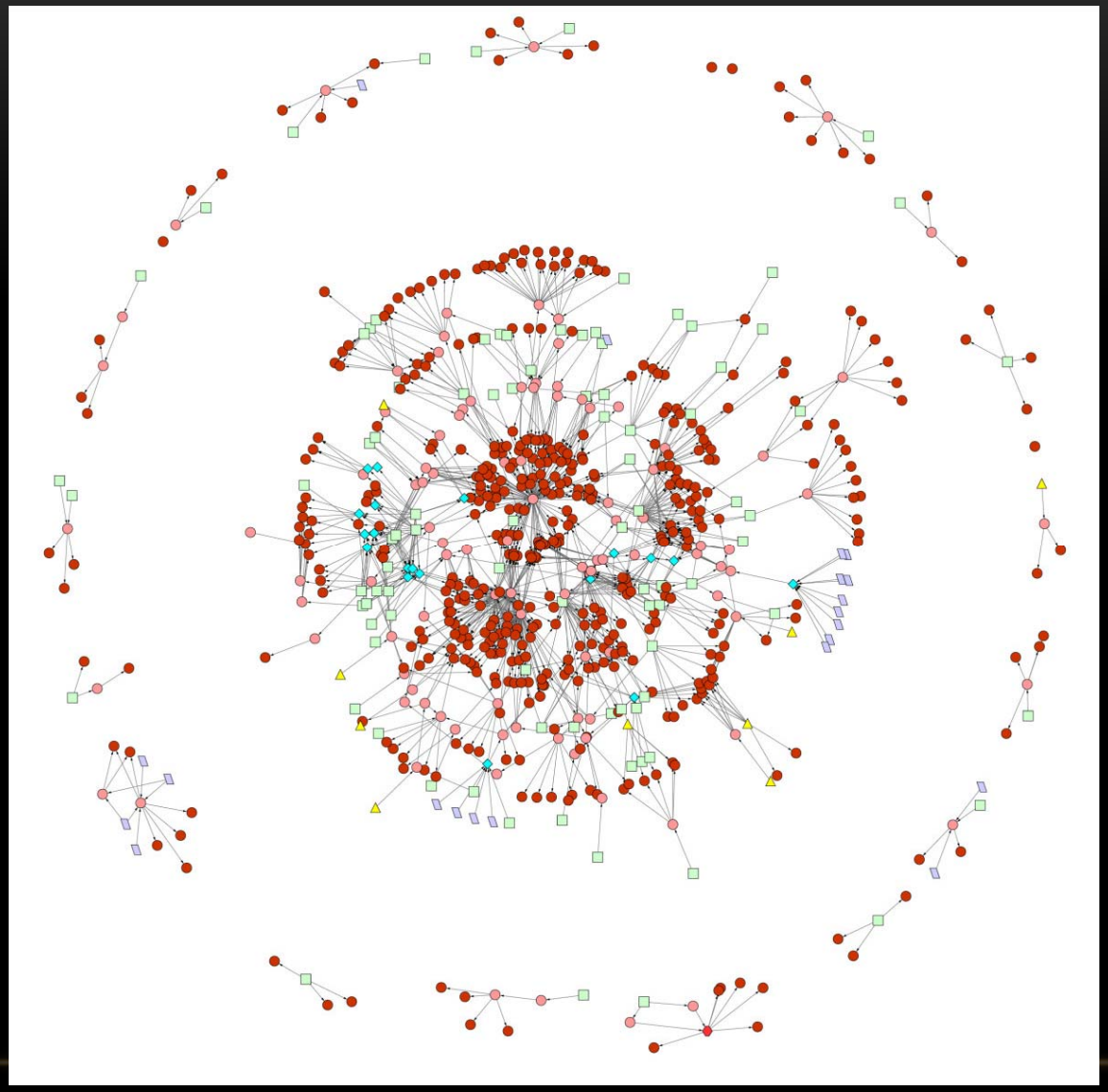
- Get many such key node networks, can be merged together to get entire description of interactions and controls for a given biological process



GENE REGULATORY NETWORK



E. COLI GENE REGULATORY NETWORKS



Map of the transcriptional regulatory network controlling metabolism in E. coli. In this figure, there are genes coding for the TFs (pink circles), genes coding for enzymes (brown circles), external metabolites (green squares), certain internal fluxes (purple parallelograms), stimuli (yellow triangles) and other conditions (blue diamonds).

Samal et al, BMC Sys Bio, 2008

Gene Regulatory Network Analysis

Provide mechanistic insights into the dynamics of gene regulation

Identify novel cross-talk between different signaling/metabolic pathway in gene regulation

Can be further extended for mechanistic insights using approaches such as

Boolean framework analysis (more suitable for prokaryotic system)
Bayesian framework analysis (wide applicability)

MOST OF THE ANALYSIS MAY BE EXTENDED TO PROTEOMIC STUDIES

Transcriptomics versus proteomics

HIGH CONTENT AND GENOME-WIDE SCREENING

Reasons for screening:

Identify novel regulators of complex biological processes (physiological processes like Cell cycle regulators, apoptosis, autophagy or more focused like NFkB activation etc. Also In case of pathogen infections like HIV infection, Mtb infection, Dengue infection etc

If we have large collection of molecules (inhibitors), screen them for some function
Take leads for further validation and into drug development pipeline!!!!

ADVANCE CELL AND MOLECULAR BIOLOGY
(ICG601)