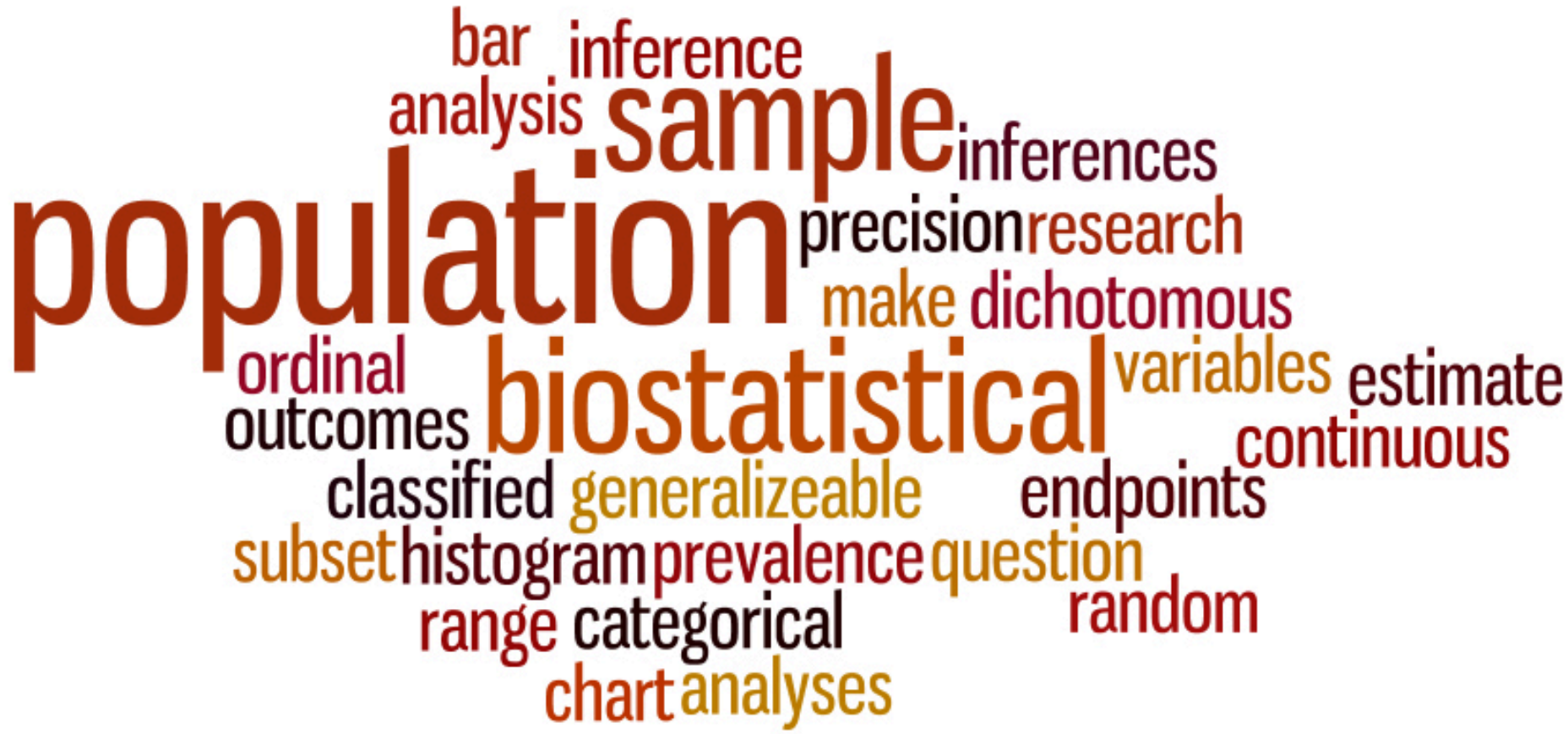


Biostatistics

Xing-Ming Zhao, School of Electronics and Information Engineering, Tongji University

www.comp-sysbio.org

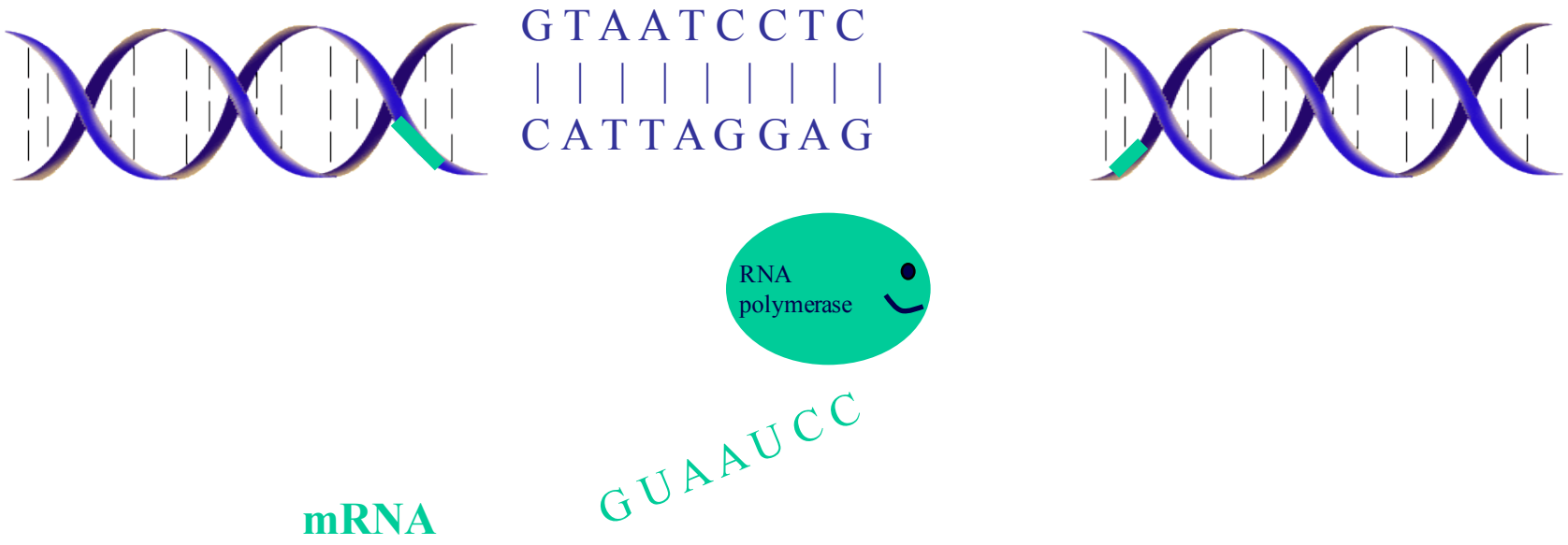
Outline

1. Differential gene expression
2. Gene set analysis
3. Mining gene expression

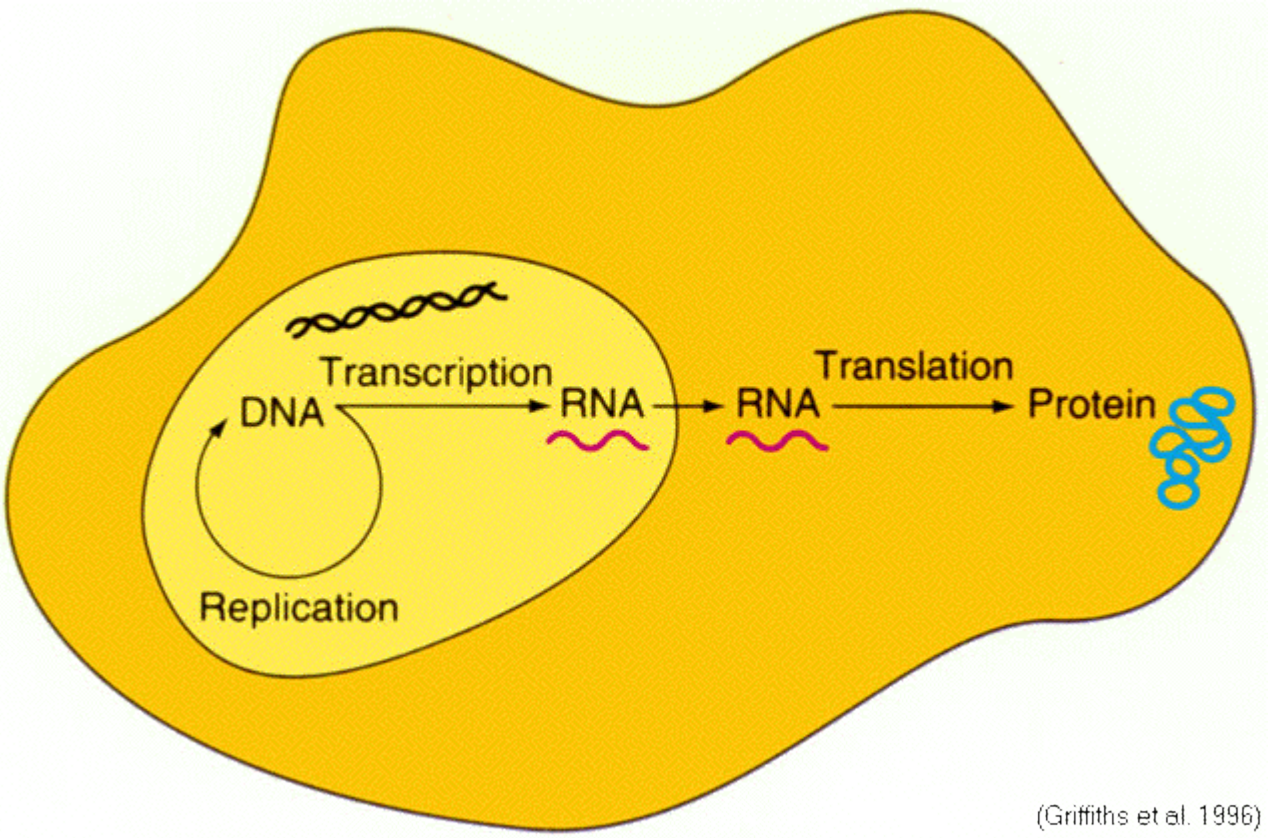
Biological background

Transcription

DNA



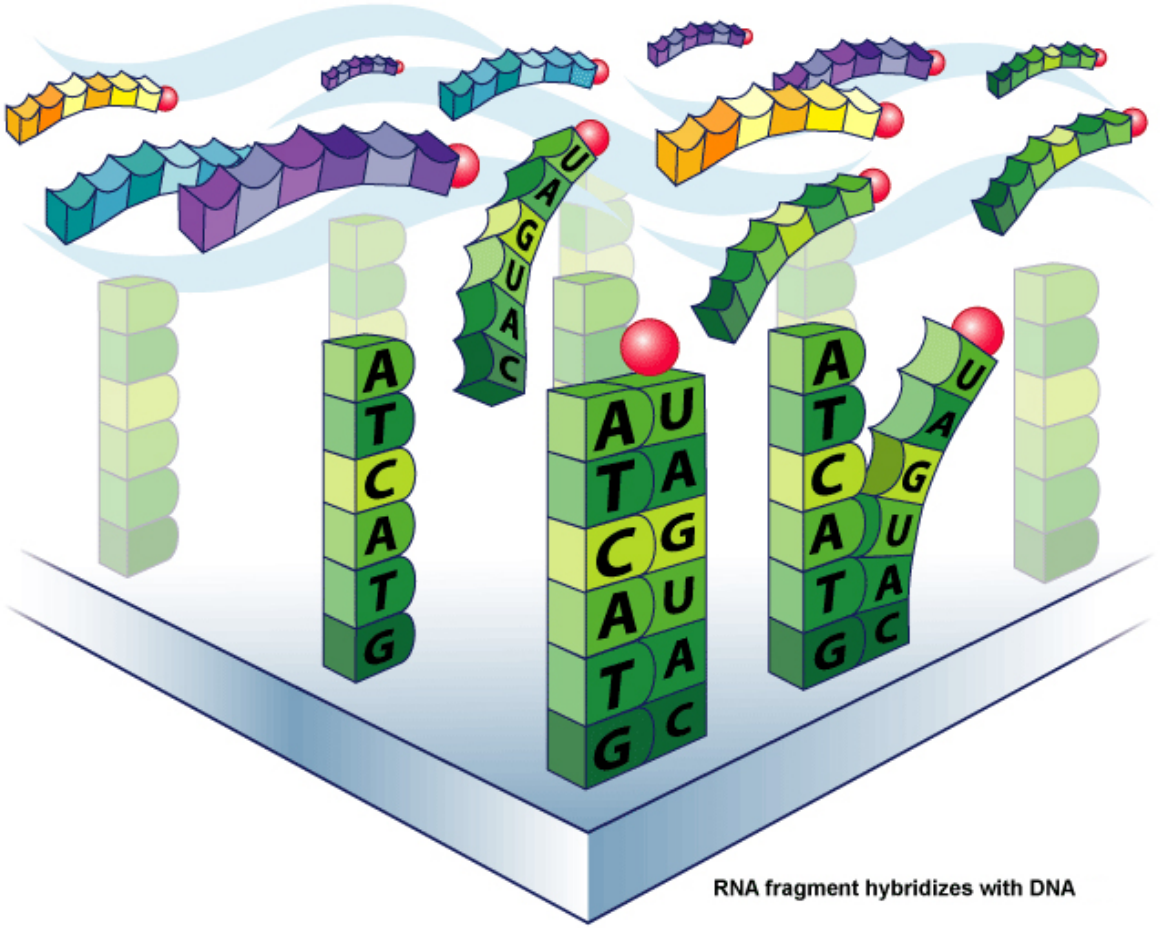
Biological background



the amount of mRNA tell which genes are being expressed in the cell

DNA microarrays

Microarrays are based on the ability of complementary strands of DNA (or DNA and RNA) to hybridize to one another in solution with high specificity.



RNA-seq

RNA-seq is simple:

- isolate all mRNAs
- convert to cDNA using reverse transcriptase
- sequence the cDNA
- map sequences to the genome

The more times a given sequence is detected, the more abundantly transcribed it is. If enough sequences are generated, *a comprehensive and quantitative view* of the entire transcriptome of an organism or tissue can be obtained.

What can we learn from gene expression?

mRNA levels compared in many different contexts

- Different tissues, same organism (brain vs liver)
- Same tissue, same organism (treatment vs control, tumor vs non-tumor)
- Same tissue, different organisms (evolution)
- Time course experiments (effect of treatment, development)

What can we learn from gene expression?

- **Classifications:** for diagnosis, prediction…

  Cell-type, stage-specific, disease-related, treatment-related patterns of gene expression?

- **Gene Networks/Pathways:**

  Functional roles of genes in cellular processes?

  Gene regulation and gene interactions

Public gene expression Repositories

- GeneX at US National Center for Genome Resources

  http://www.ncgr.org/research/genex/

- ArrayExpress at European Bioinformatics Institute

  http://www.ebi.ac.uk/arrayexpress/

Public Repositories

- Stanford University Database http://genome-www4.stanford.edu/MicroArray/SMD/index.html

- Gene Expression Omnibus at US National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/geo/

Differential Gene Expression

Differential Gene Expression

- Cell type specific
  - -e.g. skin cell vs. brain cell
- Developmental stage
  - -e.g. embryonic skin cell vs. adult skin cell
- Disease state
  - -e.g. normal skin cell vs. skin tumor cell
- Environment-specific
  - -e.g. skin cell untreated vs. treated
  - drugs, toxins

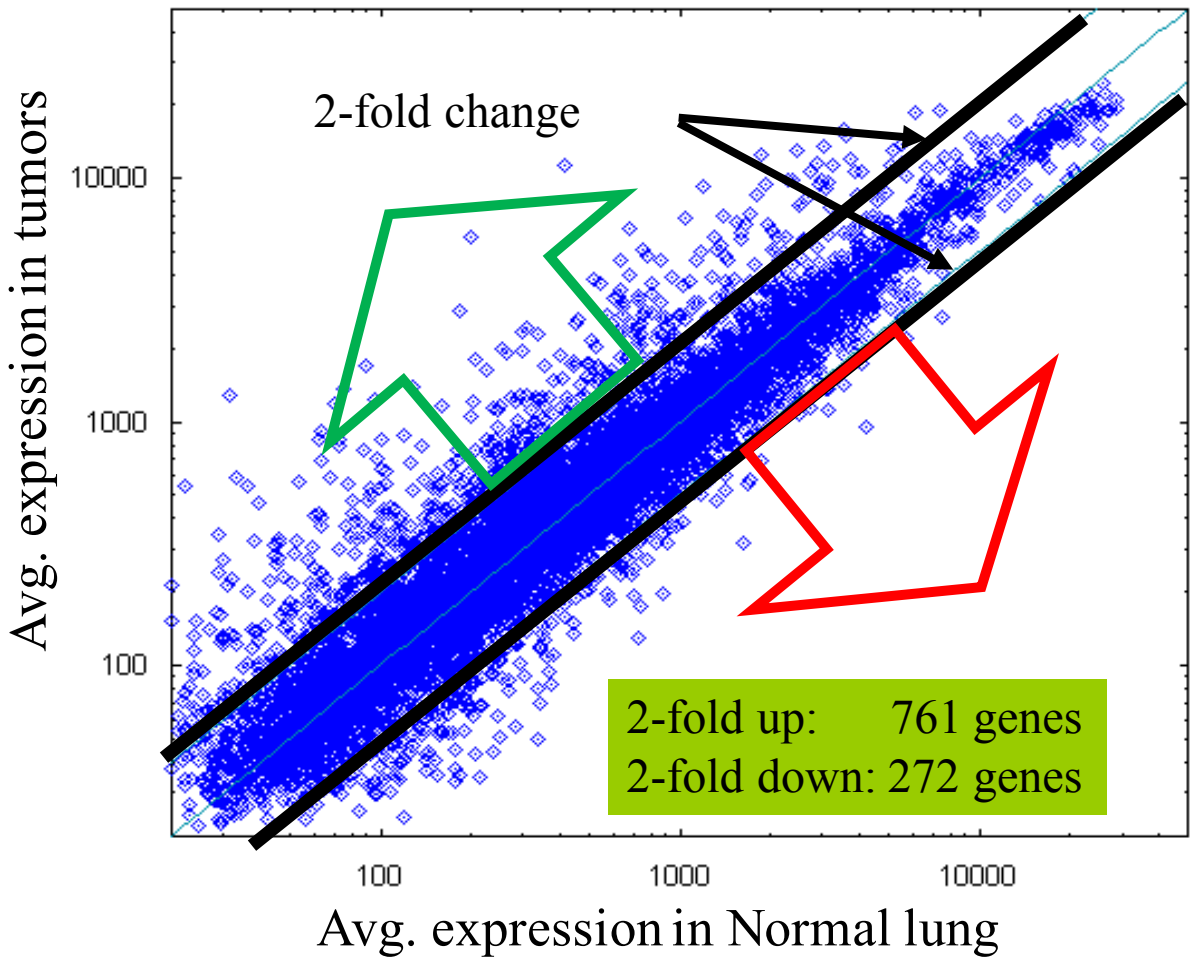
What We Need

- **Score** the genes, hopefully in a meaningful way..

- Attach a measure of **statistical significance** to the score so we can
  - Choose a subset of genes “wisely”
  - Have a measure of how strong our signal is

Fold Change

-Simplest Score

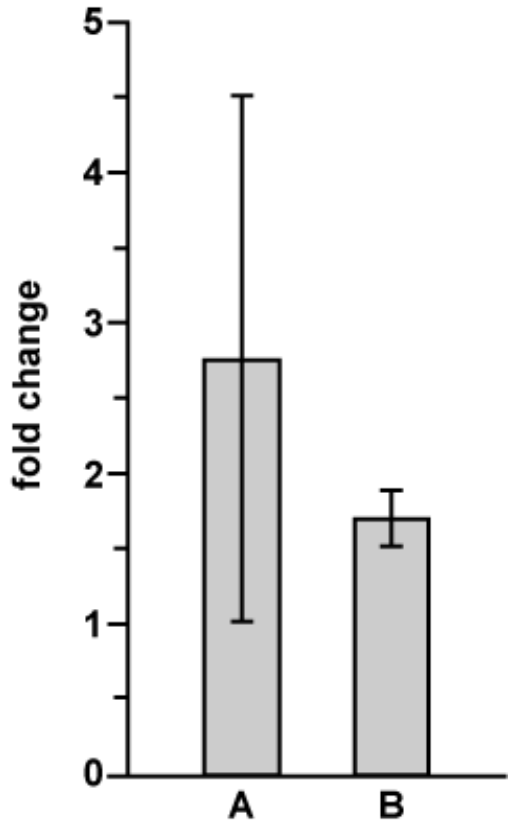


Fold Change

- Arbitrary threshold is used for filtering (e.g. 2- or 3- fold)
  - Fold Change computation: Let E=mean(group1), B=mean(group2)
  - FC= (E-B)/min(E,B)
  - At low expression levels, a 3 fold change can be noise
  - At high expression levels a 1.1 fold change can be biologically important.

Fold Change

- fold change cutoffs (e.g., 2x difference or above)
- not a very satisfying approach:
  - -misses any small changes
  - -doesn't take into account variance



Here, "A" has a fold change >2.5, but varies greatly between replicate experiments. "B" has a fold change of only 1.75, but changes reliably each time the experiment is performed.

Statistical Tests

Using statistical tests is usually a better way of determining which genes are trully changed in your microarray experiment. The ***p-value*** tells you how much confidence to place in your result.

p-value: The chance of rejecting the null hypothesis by coincidence

For gene expression analysis we can say:
the chance that a gene is categorized as differentially expressed by coincidence

Statistical Tests

- Statistical tests seek to make conclusions about parameters (ex. Mean gene expression in groups 1 and 2) on the basis of data in a sample.

- Structure of statistical tests involve testing two hypotheses. The null hypothesis H_0 and the alternative hypothesis H_A.

- The null hypothesis is usually a hypothesis of no difference.

t-test

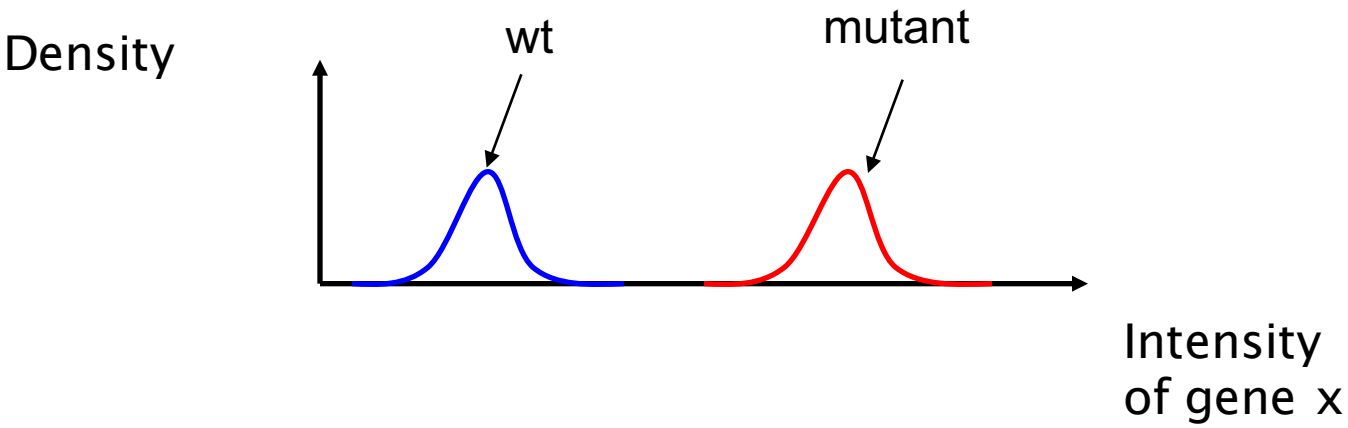
Calculate *T*

$$T = \frac{\overline{X}_2 - \overline{X}_1}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Lookup *T* in a table

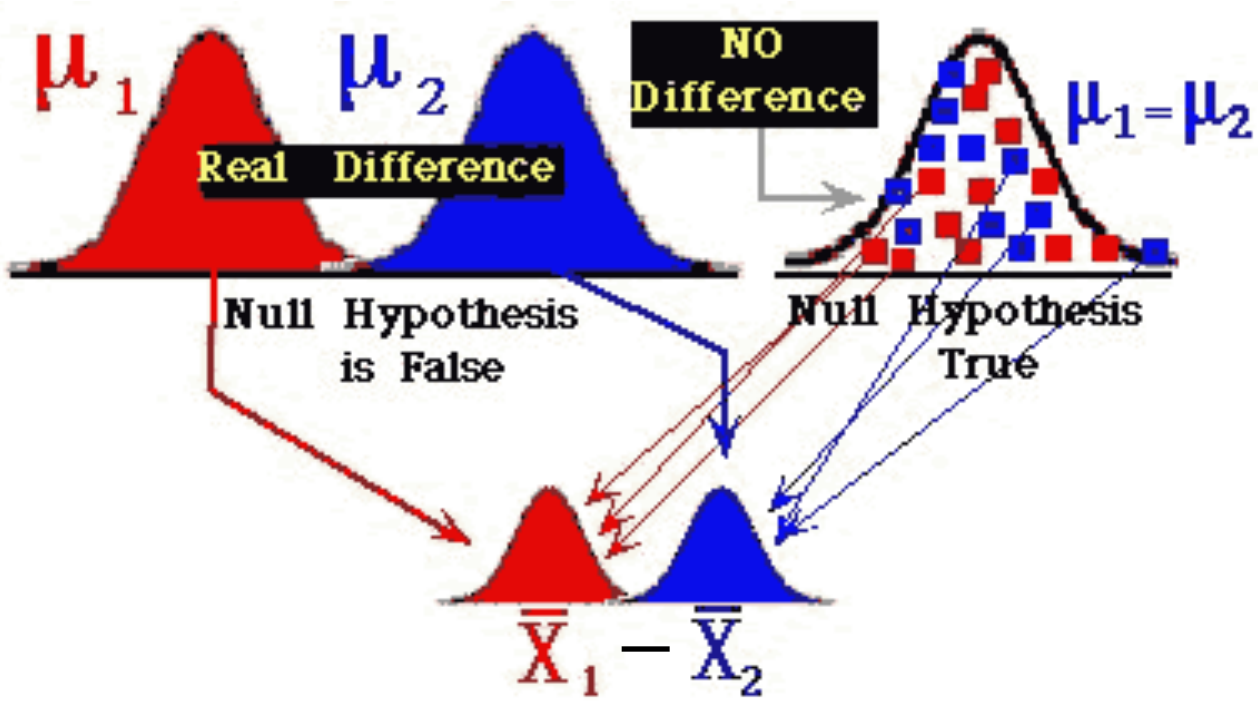
t-test

The *t*-test tests for difference in means



t-test

The *t* statistic is based on the sample mean and variance



$$t = \frac{}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

t-test: example

- Many genes
- 2 conditions
- A few replicates per condition

Gene	Condition 1			Condition 2		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
A	150	160	150	180	190	180
B	50	40	45	50	45	40
C	800	760	680	400	450	425
…						

t-test: example

- Conditions can be control vs treated, different cell types, different time points, etc.

- Typical Question – Which genes' expression levels change due to condition?
  - *t*-test

t-test

- Are we only considering up-regulated or down-regulated genes, or both?
  - If both, perform a 2-tailed test

- Can we assume that the variance of the gene is similar in both samples?
  - Yes => Homoscedastic (the standard t-test)
  - No => Heteroscedastic (Welch's test)

ANOVA

- We want to measure how gene expression changes under different conditions.
  - Only two conditions and an *adequate* number of replicates → t-tests & extensions
  - More than two conditions / more than one factor

Analysis of Variance (ANOVA)

ANOVA

- We want to determine when the variation due to gene expression is significant, but…
- There are multiple sources of variation in measurements *besides just gene expression.*
- We want to know when the variation in measurements is caused by
  - varying levels of gene expression
  - versus other factors.

ANOVA

- Analysis of variance – like a multidimensional *t*-test
- Measure effects of multiple treatments and their interactions
- A thoughtful ANOVA design can help answer several questions with one analysis
- ANOVA can also analyze factors that should be controlled
  - just to confirm absence of confounding effects
- ANOVA generally identifies genes that are influenced by some factor – but then post-hoc tests must be run to identify the specific nature of the influence
  - Ex: *t*-test between all pairs of data

ANOVA

- Some sources of variation in the measurements in microarray experiments are:
  - Array effects
  - Dye effects
  - Variety effects
  - Gene effects
  - Combinations

One-way ANOVA

- Suppose you have a model for each measurement in your experiment:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

  - y_{ij} is j^{th} measurement for i^{th} group.
  - μ : overall mean effect (constant)
  - α_i : i^{th} group effect (constant)
  - ε_{ij} : experimental error term ~N(0,σ^2)

- Therefore, observations from group *i* are distributed with mean $\mu + \alpha_i$ and variance σ^2 .

One-way ANOVA

- Many genes
- Multiple conditions
- A few replicates per condition

One way ANOVA

Gene	Condition 1			Condition 2			Condition 3			…
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	
A	150	160	150	180	190	180	150	155	135	
B	50	40	45	50	45	40	80	90	105	
C	800	760	680	400	450	425	200	220	400	
…										

One-way ANOVA

- Conditions can be treatments or chemicals, cell types, time points, etc.
- Question being asked is whether the expression level for each gene (taken one at a time) changes significantly as a function of dose.
- More specifically, it compares the variability within replicates for a given dose to the variability caused by changing the dose.
- If gene chip contains 1000 genes, then do 1000 ANOVAs.

One-way ANOVA

gene	dose 0		dose 50 mM		dose 75 mM		ANOVA
	rep 1	rep 2	rep 1	rep 2	rep 1	rep 2	p value
AA848268_at	163	164	94	165	178	181	0.35
AA848421_at	-91	-13	-102	-18	-97	-73	0.79
AA848546_at	526	498	424	377	539	410	0.29
AA848563_s_at	283	343	3212	4392	1191	1911	0.02
AA859934_at	65	62	6	1	8	71	0.21
AA875509_at	442	415	259	355	265	280	0.07
AA933181_at	-17	38	4	30	29	27	0.78
AA956437_at	75	95	56	73	38	62	0.20
AA958273_at	42	34	-23	18	25	-24	0.35
AA958274_at	114	42	50	41	0	7	0.18

Multiple Factor ANOVA

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

- The model can be extended by adding
  - More factors (α_i ,β_j, …)
  - Interactions between them ($\alpha\beta_{ij}$, …)
  - Other …
- This is used to model the different sources of variation appearing in microarray experiments

Two-way ANOVA

- Many genes
- 2 Factors, multiple conditions per factor
  - For example, Factor 1 could be dose of a chemical, and Factor 2 could be time point after dosing
  - Can reveal significant effect of time, significant effect of dose, or a significant interaction between the two
- Multiple replicates per condition

Two-way ANOVA

Gene	Dose 1				Dose 2				…
	Time 1		Time 2		Time 1		Time 2		…
	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	
A	150	160	150	180	190	180	150	155	
B	50	40	45	50	45	40	80	90	
C	800	760	680	400	450	425	200	220	
…									

Two-way ANOVA

- Typical Question – Which genes' expression levels change due to time? Due to dose? Due to an interaction between the two?
  - 2-way ANOVA

- Or, eliminate one of the dimensions and ask the same questions as before – At time 1, which doses show similar expression levels among genes?

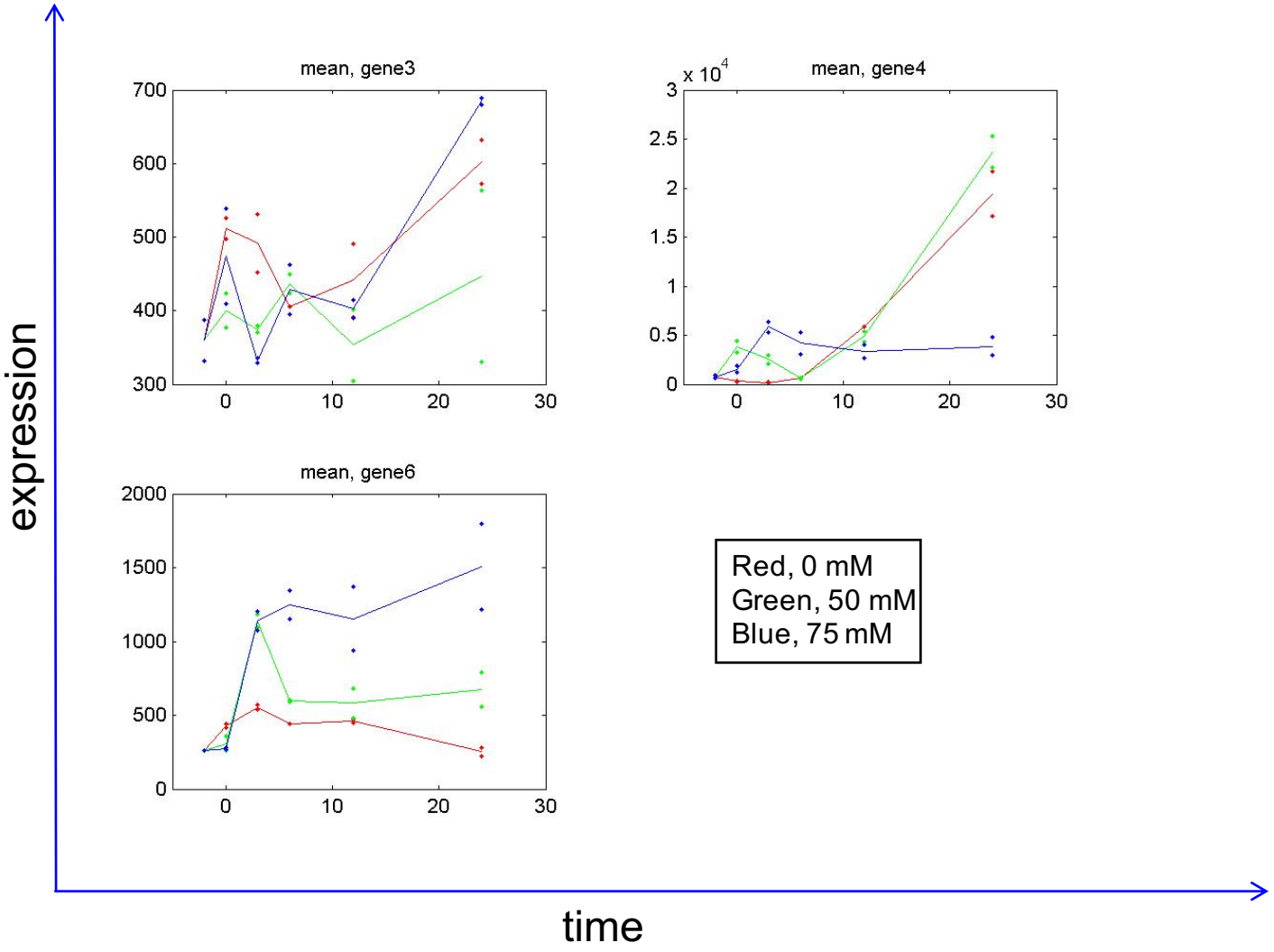
Two-way ANOVA

Source	SS	df	MS	F	P
Time	28724	4	7181	9.20	7.3e-4
Dose	1143	2	572	0.73	0.498
Time*dose	22940	8	2868	3.67	0.016
Error	10930	14	781		
Total	64409	28			

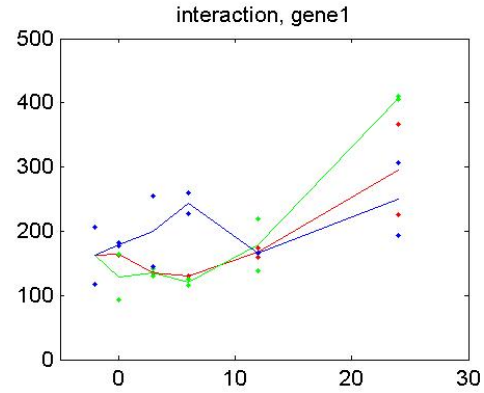
Two-way ANOVA

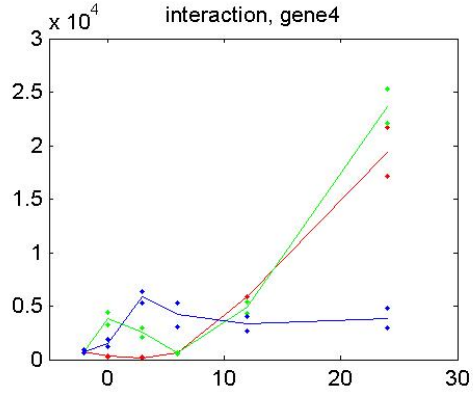
	time	dose	time*dose
AA848268_at	0.000	0.416	0.038
AA848421_at	0.000	0.609	0.913
AA848546_at	0.001	0.024	0.125
AA848563_s_at	0.000	0.000	0.000
AA859934_at	0.206	0.676	0.142
AA875509_at	0.000	0.000	0.001
AA933181_at	0.001	0.498	0.016
AA956437_at	0.434	0.607	0.693
AA958273_at	0.802	0.325	0.283
AA958274_at	0.344	0.242	0.127

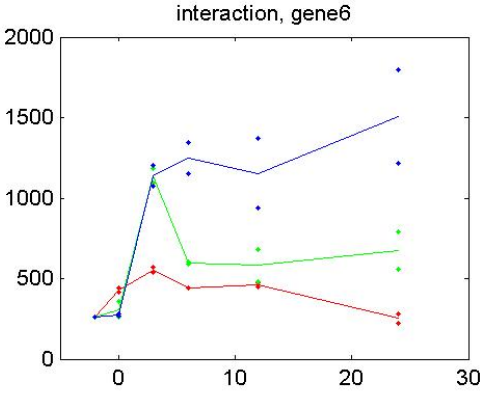
Two-way ANOVA

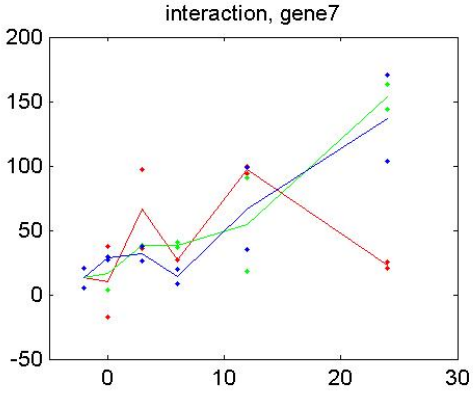


Two-way ANOVA









ANOVA

Let y_{ijkg} be the fluorescent intensity measured from Array i, Dye j, Variety k, and Gene g, on the appropriate scale (such as log). A typical analysis of variance (ANOVA) model is:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (DG)_{jg} + (\boldsymbol{VG})_{kg} + e_{ijkg}$$

- μ, A, D, V are "normalization" terms
- G are the overall gene effects
- AG's are "spot" effects
- DG's are gene-specific dye effects
- VG's are the effects of interest
- *e* is random error

ANOVA

- ANOVA is a robust statistical procedure
- sources of variation, *e.g.* whether variation in gene expression is less in subset of data than in total data set
- Requires moderate levels of replication (4-10 replicates of each treatment)
- Expression judged according to statistical significance instead of by adopting arbitrary thresholds

ANOVA

- Advantage: design does not need reference samples

- Concern: treatments should be randomised and all single differences between treatments should be covered

  E.g., if male kidney and female liver are contrasted on one set, and female kidney and male liver on another, we cannot state whether gender or tissue type is responsible for expression differences observed

Basic Statistical Comparisons between groups

- *t*-Tests & ANOVA
  - Much better than fold change thresholds
  - Need to characterize (or stabilize) variances
  - Assumes normal distribution of expression values. Transformation helps.

- Non-Parametric Tests: Wilcoxon

Wilcoxon Rank Test

- Another gene score:
  - Ignores absolute values
  - Takes into account only **order** of measurements

- Sort the expression values of both groups

  \+ + - - + + + - - + - - + + -

  a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15

- W(g)= sum of ranks of the positive examples:

  $W(g) = 1 + 2 + 5 + 6 + 7 + 10 + 13 + 14 = 58$

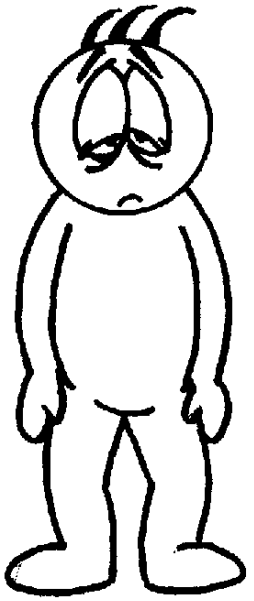
Wilcoxon Rank Test

- A common test in statistics
- Again, we can compute p-values given the null hypothesis H_0
  - $P(W(g) > s|n, k)$ = the probability of getting a score > s given a total of n samples, out of which k are labeled as (+).

P value

We can rank the genes according to the p-value

But, we can't really trust the p-value in a strict statistical way!



P value

Why not!

For two reasons:

1. We are rarely fulfilling all the assumptions of the statistical test

2. We have to take multi-testing into account

The *t*-test Assumptions

1. The observations in the two categories must be independent

2. The observations should be normally distributed

3. The sample size must be ‘large’ (>30 replicates)

Multi-testing?

In a typical microarray analysis we test thousands of genes

If we use a significance level of 0.05 and we test 1000 genes. We expect 50 genes to be significant by chance

$$1000 \times 0.05 = 50$$

Multi-testing correction

- Bonferroni correction
- False discovery rate

Permutation testing

- Permutation testing
  - Scramble (randomly permute) the data to randomize relationship of interest, leaving other structure unchanged
  - Do (any) testing procedure on permuted data
  - Repeat K times.
  - Proportion of times randomly permuted data scores better than actual data is p value
- As K increases, becomes very accurate measure of p value. Makes no assumptions about test, so works with any test.
- Very time consuming (typical K = 1000)

Consequences of multiple testing corrections

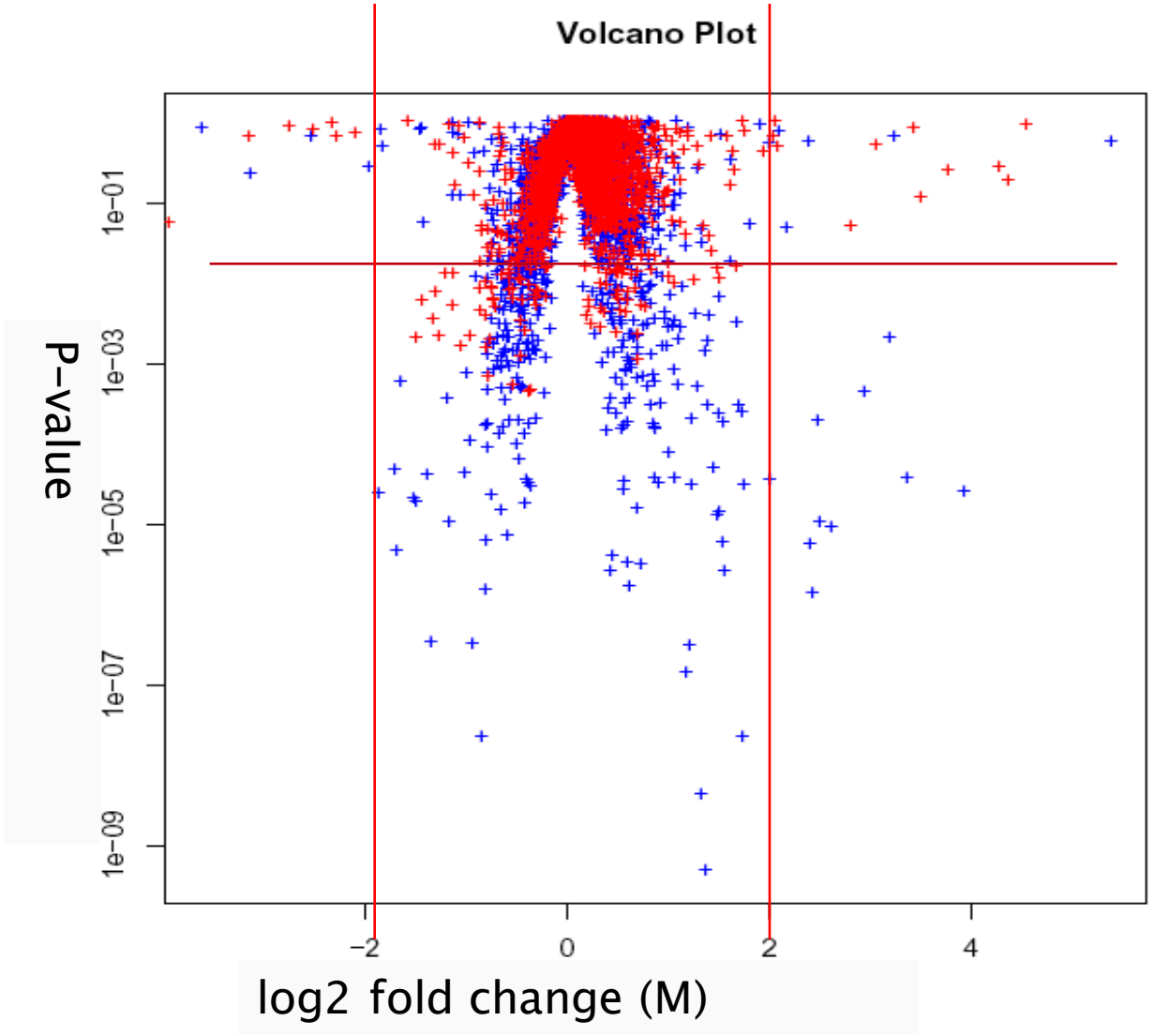
- Multiple testing corrections cost statistical power (probability of correctly rejecting the null when there really is a difference)
  - The more genes tested, the greater the loss

- Therefore, eliminate as many genes as possible from consideration before testing for significant differences.
  - E.g. by eliminating genes that didn't change at all on any chips, regardless of experimental conditions.

Combining *p*-values and fold changes

- What's important biologically?
  - How significant is the difference?
  - How large is the difference?
- Both amounts can be used to identify genes.
- What cutoffs to use?
- How many genes should be selected?
- Where are your positive controls?

Combining *p*-values and fold changes



Signal-to-Noise

Signal-to-Noise (S2N) filter used to select 50 most variable genes out of 19,892 genes on the array.

$$S2N(j) = \left| \frac{\mu_1(j) - \mu_2(j)}{\sigma_1(j) + \sigma_2(j)} \right|$$

where μ and σ indicate means and std. dev. of expression levels of gene *j* for case and control groups.

Gene set analysis

Enrichment analysis

Long lists of differentially expressed genes

What happens next?

- Select some genes for validation?
- Do follow-up experiments on some genes?
- Publish a huge table with the results?
- Try to learn about all the genes on the list (read 100s of papers)?
- ….

Usually, some or all of the above will be done, and more.

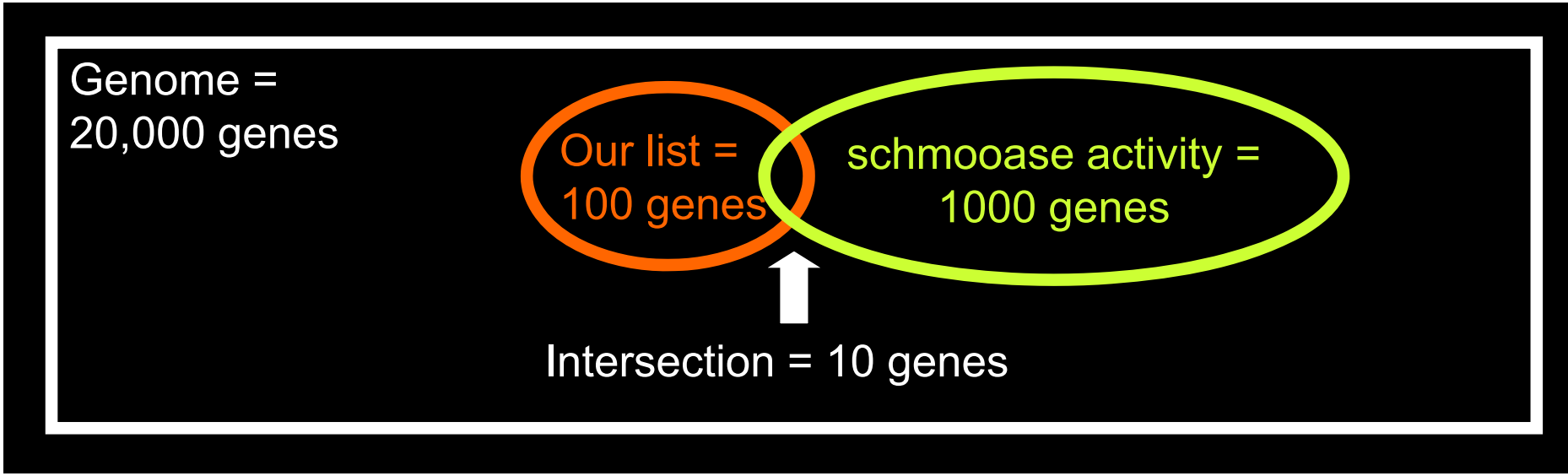
Can we help further at this?

Annotation sources

- Gene Ontology(most popular)
  - biological process, molecular function, cellular component
  - Terms may have >1 “parent” (more general term)
  - GO Slim: includes only general categories
- KEGG; REACTOME pathways
- Genes sharing a motif regulated by the same protein
- Genes found on the same chromosome
- …

Why enrichment analysis?

- Too many genes to examine in detail
- Are we biased?
- How do we know that what we're seeing is by chance?



Main types of enrichment analysis

List-based:

– A subset of all genes chosen by some relevant method

– A list of annotations, each linked to genes

- Rank-based:

  – A set of all genes ranked by some metric (ratio, fold change, etc.)

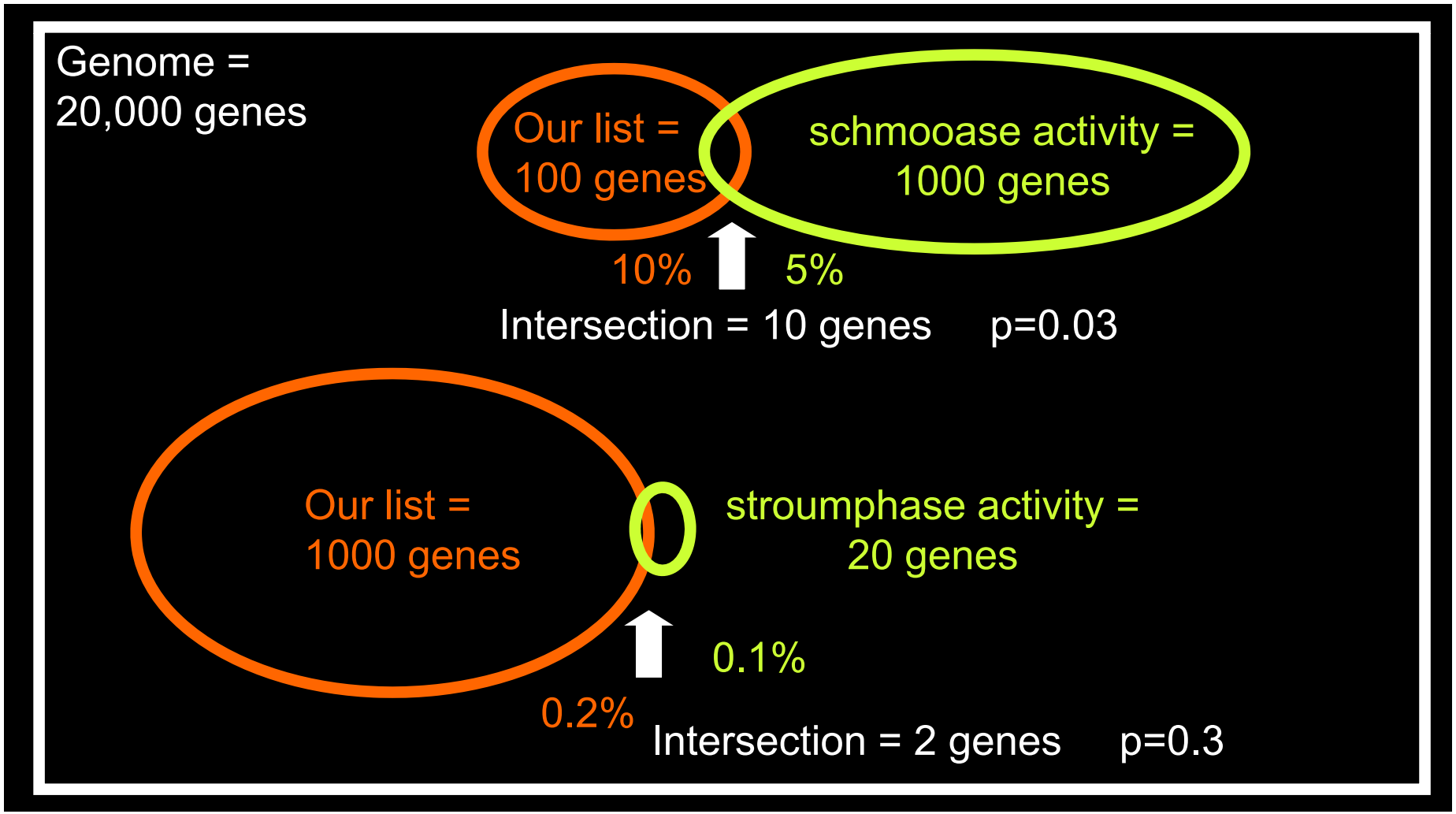
  – A list of annotations, each linked to genes

- List-based with relationships:

  – A subset of all genes

  – A list of annotations, each linked to genes, organized in some relationship (e.g., a hierarchy)

Statistics to test for enrichment



Statistics to test for enrichment

- What is the chance of observing enrichment due to chance?
- Different tests produce very different ranges of p-values
- All look for over-enrichment; some look for under-enrichment
- Recommendation: Use p-values as a tool to rank genes but don't take them literally
- Most methods correct for multiple testing (e.g.,with FDR), which is necessary

Statistics to test for enrichment

- Fisher's exact
- Hypergeometric
- Binomial
- Chi-squared
- Z
- Kolmogorov-Smirnov
- Permutation
- .....

Fisher's test by hand in R

- counts = (matrix(data = c(3, 297, 40, 19960), nrow = 2))
- counts
- fisher.test(counts)
- # is better than
- chisq.test(counts)

	Gene list	Genome
In anno group	3	40
Not in anno group	297	19960

```
        Fisher's Exact Test for Count Data

data:  counts
p-value = 0.02552
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9918169 15.9604612
sample estimates:
odds ratio
  5.039206
```

(3/297) / (40/19960)

Hypergeometric test by hand in R

- min(1 - cumsum(dhyper(0:(3-1), 40, 19960, 300)))
- 0.02193491

	Gene list	Genome
In anno group	3	40
Not in anno group	297	19960

- Equation above tests only for over-enrichment

The Gene Ontology Consortium

http://www.geneontology.org

The goal of the Gene Ontology ™ (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes

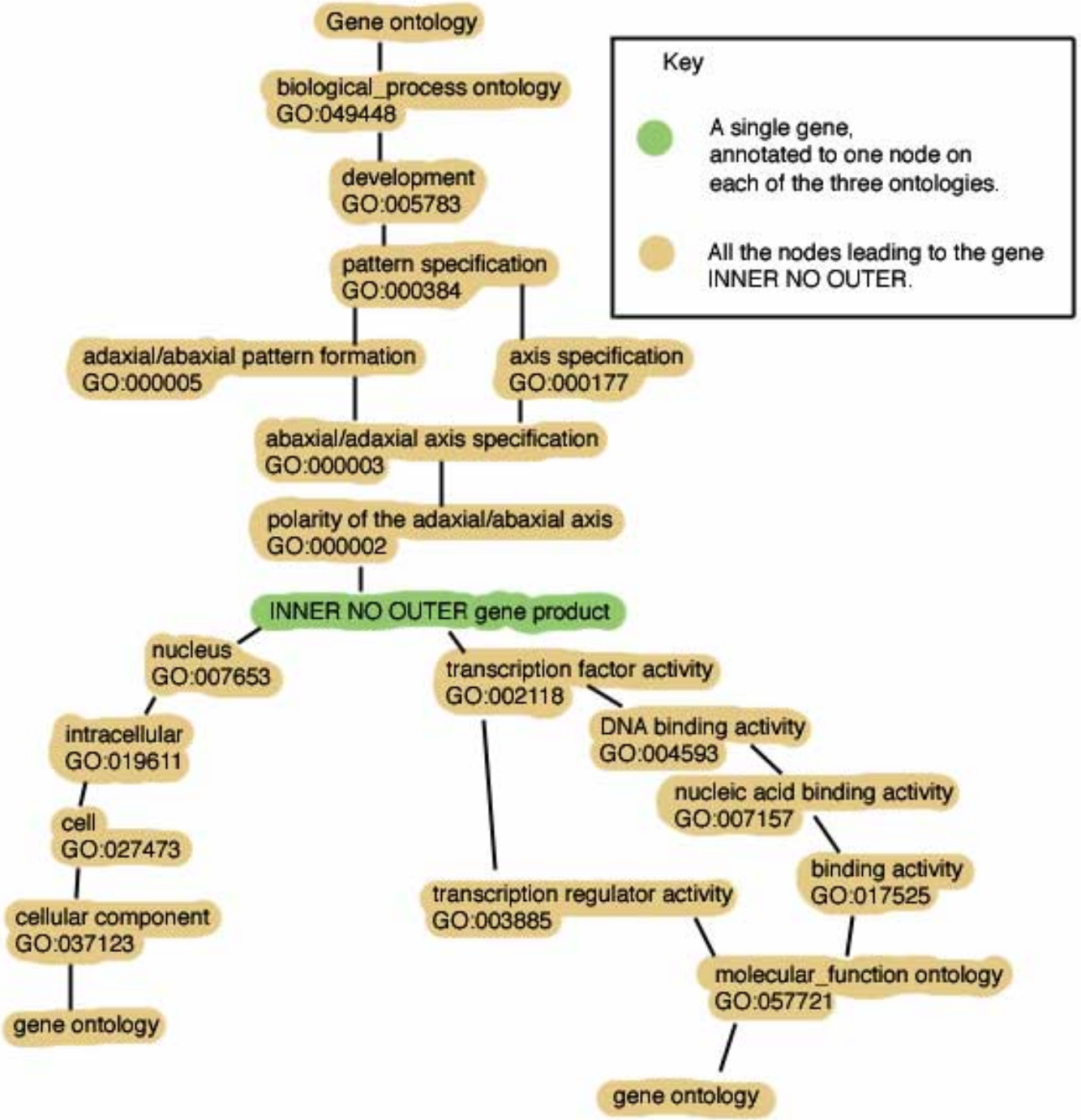
Molecular Function Ontology: the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

Biological Process Ontology: broad biological goals, such as *mitosis* or *purine metabolism,* that are accomplished by ordered assemblies of molecular functions

Cellular Component Ontology: subcellular structures, locations, and macromolecular complexes; examples include *nucleus, telomere,* and *origin recognition complex*

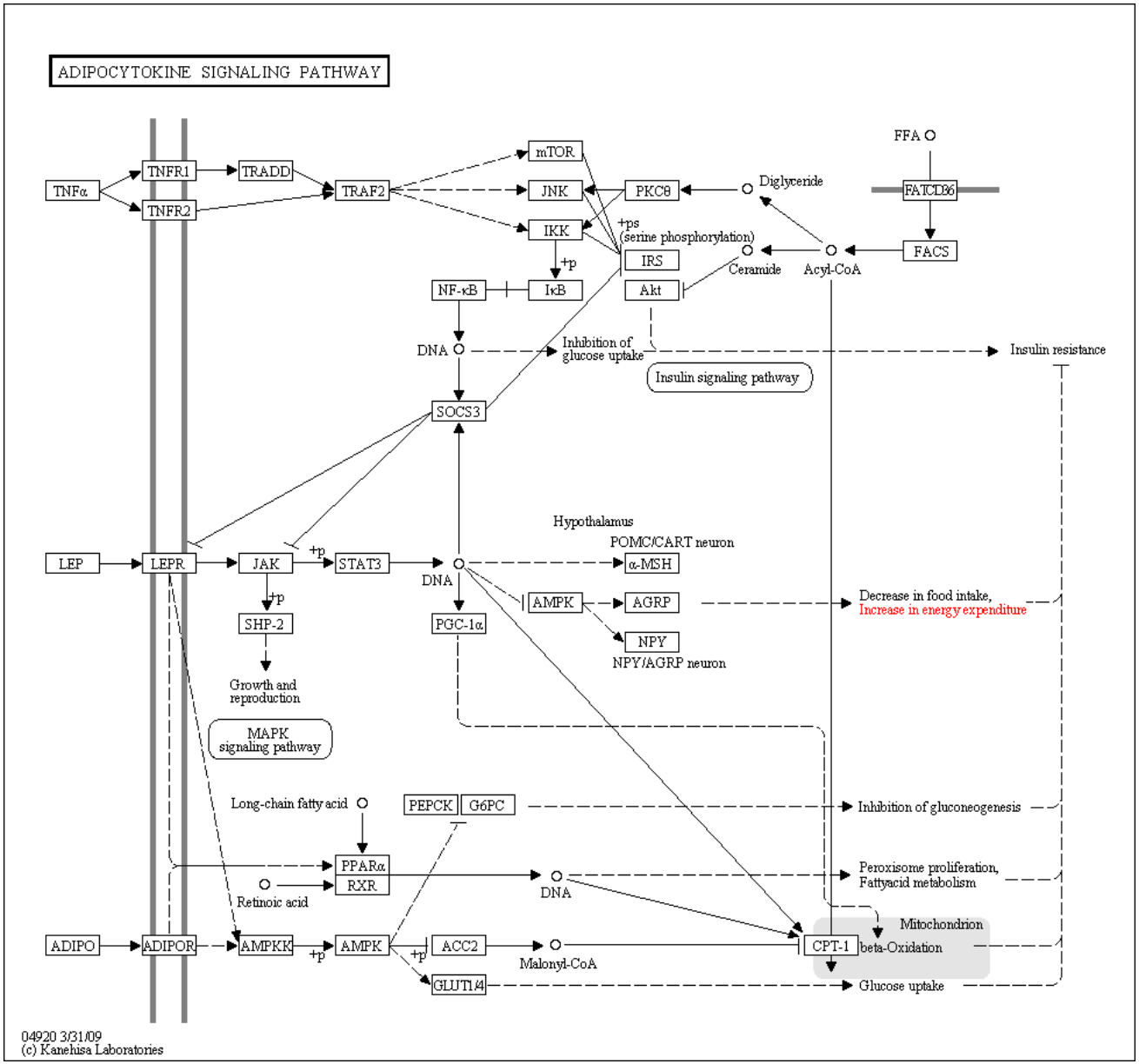
From the GO web site. The *path* back to each ontology from a gene.

We will call each term in a path a *split*.



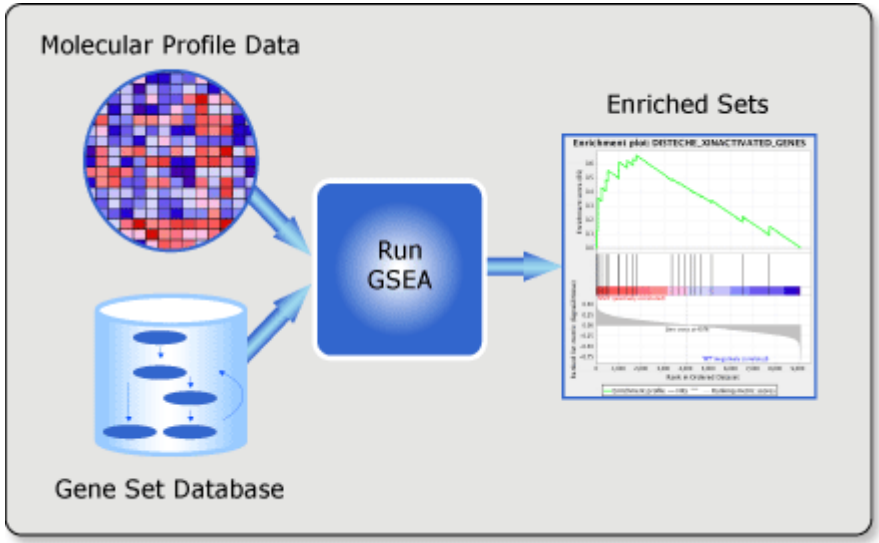
KEGG pathway database

- KEGG = Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
  - The pathway database gives far more detailed information than GO
    - Relationships between genes and gene products
  - But: this detailed information is only available for selected organisms and processes
  - Example: Adipocytokine signaling pathway


ADIPOCYTOKINE SIGNALING PATHWAY
TNFα
TNFR1
TNFR2
TRADD
TRAF2
mTOR
JNK
IKK
PKCθ
Diglyceride
FFA
FATCD36
FACS
Acyl-Coa
Ceramide
+ps
(serine phosphorylation)
IRS
Akt
+p
NF-κB
IκB
DNA
Inhibition of
glucose uptake
Insulin resistance
Insulin signaling pathway
SOCS3
Hypothalamus
POMC/CART neuron
LEP
LEPR
JAK
+p
STAT3
DNA
α-MSH
AMPK
AGRP
NPY
NPY/AGRP neuron
Decrease in food intake,
Increase in energy expenditure
+p
SHP-2
PGC-1α
Growth and
reproduction
MAPK
signaling pathway
Long-chain fatty acid
PEPCK
G6PC
Inhibition of gluconeogenesis
PPARα
RXR
Retinoic acid
DNA
Peroxisome proliferation,
Fattyacid metabolism
Mitochondrion
ADIPO
ADIPOR
AMPKK
+p
AMPK
+p
ACC2
Malonyl-CoA
CPT-1
beta-Oxidation
GLUT1/4
Glucose uptake
04920 3/31/09
(c) Kanehisa Laboratories

Gene Set Enrichment Analysis

- MIT, Broad Institute
- V 2.0 available since Jan 2007



Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [b]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [c]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [d]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [e]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [f]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [g]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [h]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [i]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Eric S. Lander, August 2, 2005

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

microarray

Genomewide expression analysis with DNA microarrays has become a mainstay of genomics research (1, 2). The challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms.

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. The genes can be ordered in a ranked list *L*, according to their differential expression between the classes. The challenge is to extract meaning from this list.

A common approach involves focusing on a handful of genes at the top and bottom of *L* (i.e., those showing the largest difference) to discern telltale biological clues. This approach has a few major limitations.

(*i*) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(*ii*) Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.

(*iii*) Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

(*iv*) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap (3).

To overcome these analytical challenges, we recently developed a method called Gene Set Enrichment Analysis (GSEA) that evaluates microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments. The goal of GSEA is to determine whether members of a gene set *S* tend to occur toward the top (or bottom) of the list *L*, in which case the gene set is correlated with the phenotypic class distinction.

We used a preliminary version of GSEA to analyze data from muscle biopsies from diabetics vs. healthy controls (4). The method revealed that genes involved in oxidative phosphorylation show reduced expression in diabetics, although the average decrease per gene is only 20%. The results from this study have been independently validated by other microarray studies (5) and by *in vivo* functional studies (6).

Given this success, we have developed GSEA into a robust technique for analyzing molecular profiling data. We studied its characteristics and performance and substantially revised and generalized the original method for broader applicability.

In this paper, we provide a full mathematical description of the GSEA methodology and illustrate its utility by applying it to several diverse biological problems. We have also created a software package, called GSEA-P and an initial inventory of gene sets (Molecular Signature Database, MSigDB), both of which are freely available.

Methods

Overview of GSEA. GSEA considers experiments with genomewide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (Fig. 1*A*).

Given an *a priori* defined set of genes *S* (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout *L* or primarily found at the top or bottom. We expect

Freely available online through the PNAS open access option.

Abbreviations: ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; *ES*, enrichment score; FDR, false discovery rate; GSEA, Gene Set Enrichment Analysis; MAPK, mitogen-activated protein kinase; MSigDB, Molecular Signature Database; NES, normalized enrichment score.

See Commentary on page 15278.

[1]A.S. and P.T. contributed equally to this work.

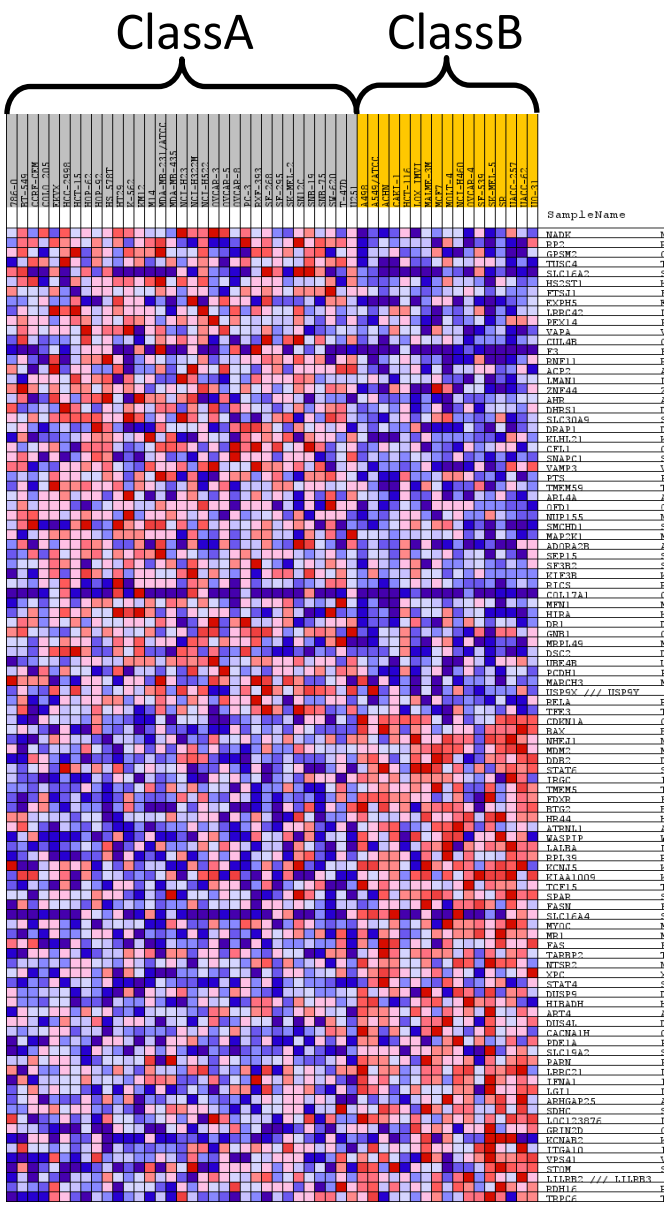
[*]To whom correspondence may be addressed. E-mail: lander@broad.mit.edu or mesirov@broad.mit.edu.

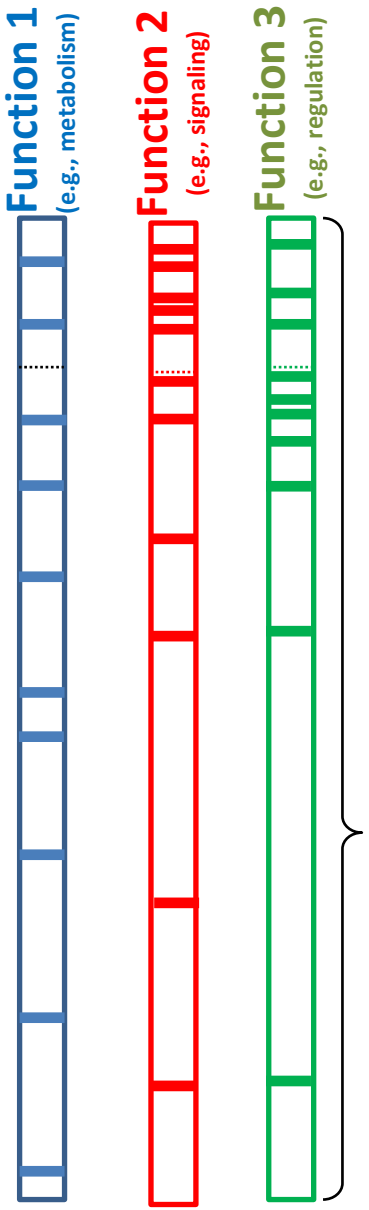
(Subramanian et al. PNAS. 2005.)

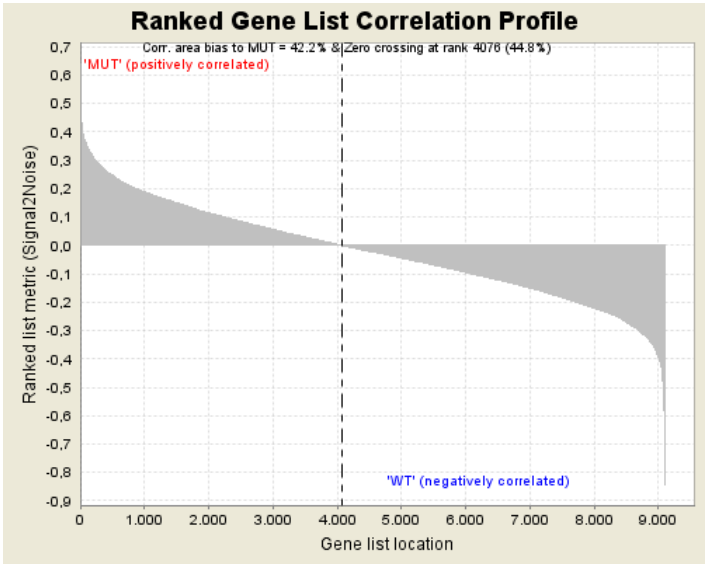
GSEA key features

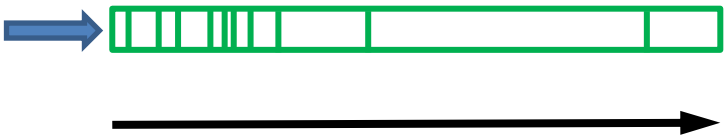
- Calculates a score for the enrichment of a **entire set of genes** rather than single genes!
- Does not require setting a cutoff!
- Identifies the set of relevant genes as part of the analysis!
- Provides a more robust statistical framework!

Gene Set Enrichment Analysis









Running sum:

Increase when gene is in set

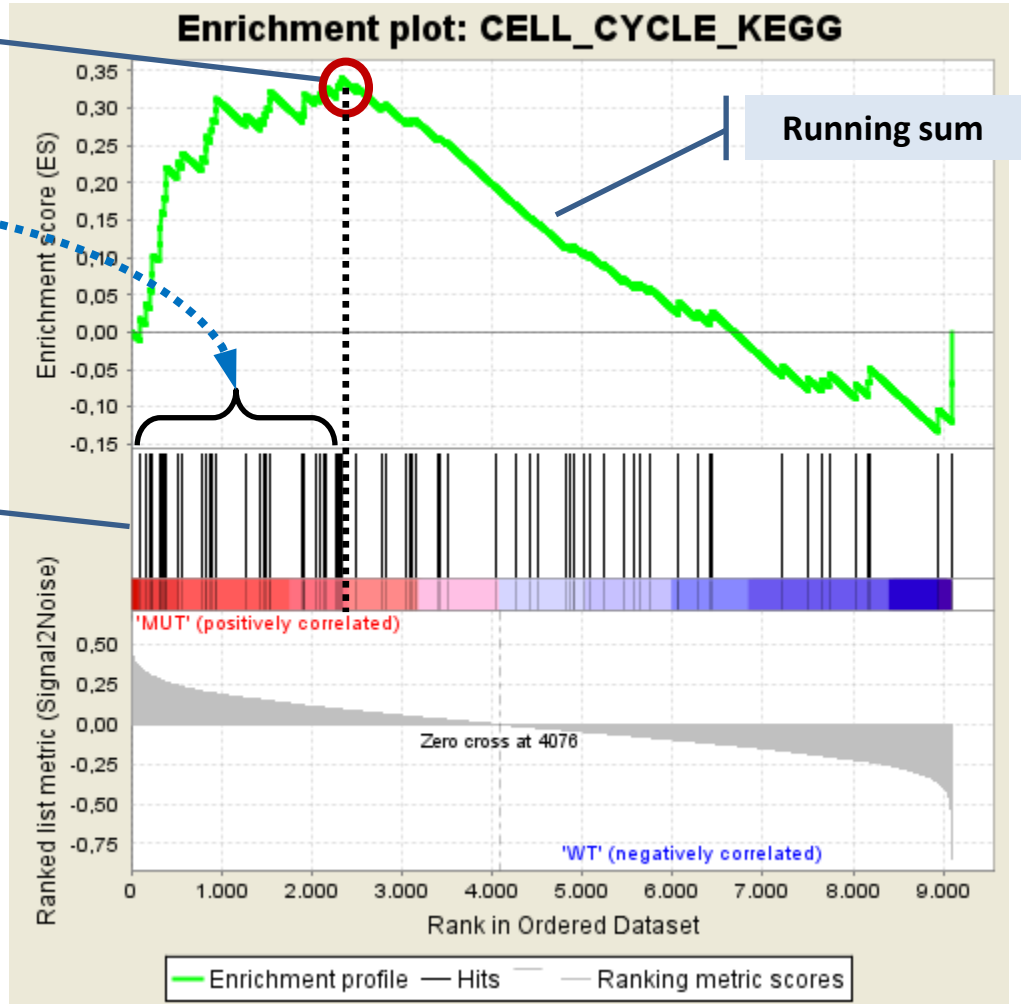
Decrease otherwise

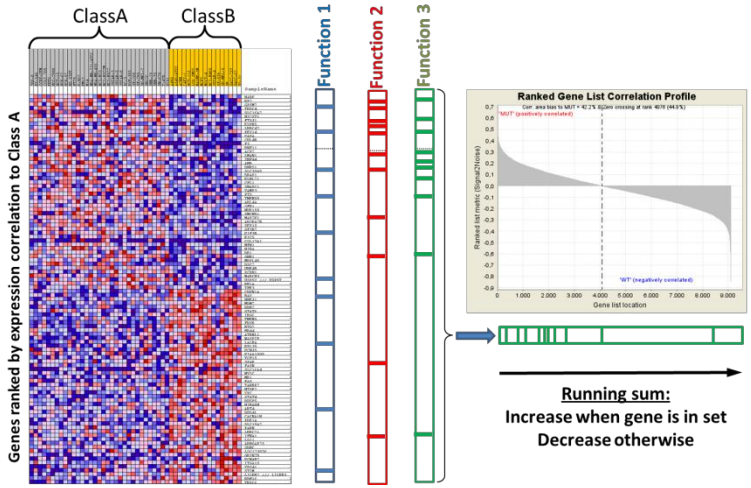
Gene Set Enrichment Analysis

Enrichment score (ES) = max deviation from 0

Leading Edge genes

Genes within functional set (hits)





Other statistical issues

- Goal: Identifying theme(s) of maximal biological significance
  - but this is not perfectly correlated with statistical significance

- What is your background gene set?
  - All genes that could appear in your list

- What about sparse annotation groups?

- Some annotation terms may be subsets of other terms.

Some recommended tools

- DAVID
- GSEA
- BIOBASE (Whitehead has license)
- BiNGO (uses Cytoscape)
- GoMiner: http://discover.nci.nih.gov/gominer
- GOstat: http://gostat.wehi.edu.au

Mining gene expression

Mining gene expression

- After differential analysis, what is next?
  - Genes coregulated
  - Gene network
  - Genes involved in the same biological process

Clustering

- Cluster analysis: dividing samples (genes) into homogeneous groups based on a set of features

- Steps: generate expression summaries
  measure pairwise distances
  cluster

Distance (semi)metrics for pairwise measurements

Euclidean distance

$$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$$

Manhattan distance

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

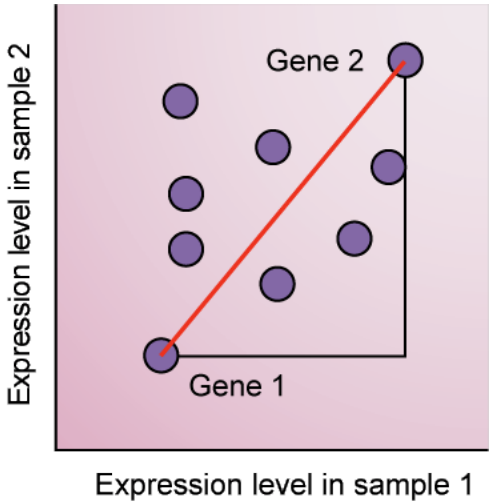
Correlation coefficient

$$r=\sum_{i=1}^{p}\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{p}(y_i-\bar{y})^2}}$$

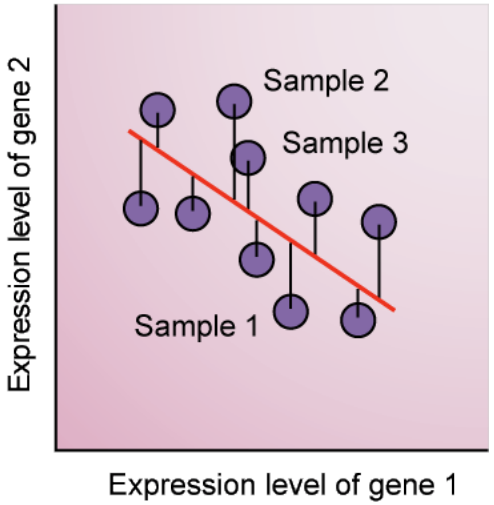
Metrics for gene expression

- Need a method to measure how similar genes are based on expression

- Examples
  - Euclidean distance
  - Pearson correlation coefficient



Euclidean distance



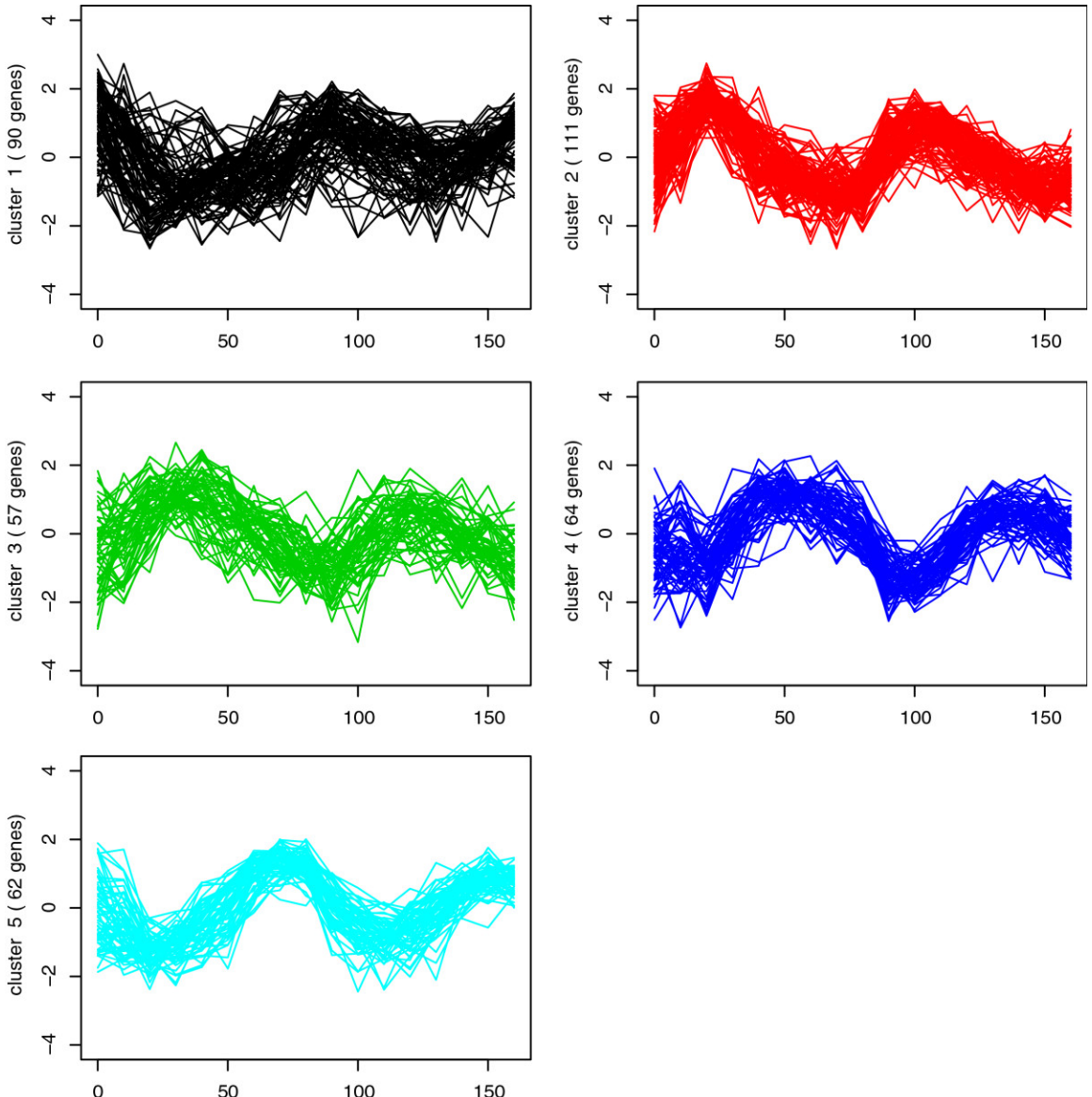
Pearson correlation coefficient

Clustering

- make no assumptions about how the data should behave

- Typical approaches includes
  - hierarchical clustering
  - *k*-means clustering
  - self-organizing maps (SOM)
  - principal component analysis (PCA)

Clustering

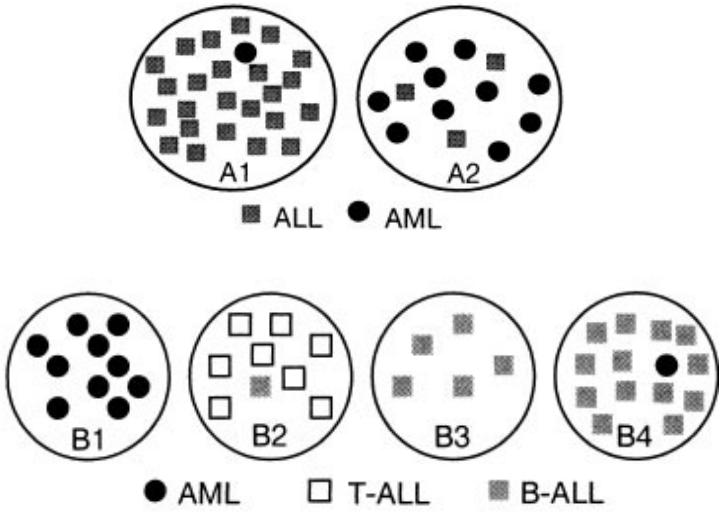

cluster 1 (90 genes)
cluster 2 (111 genes)
cluster 3 (57 genes)
cluster 4 (64 genes)
cluster 5 (62 genes)
4
2
0
−2
−4
0
50
100
150

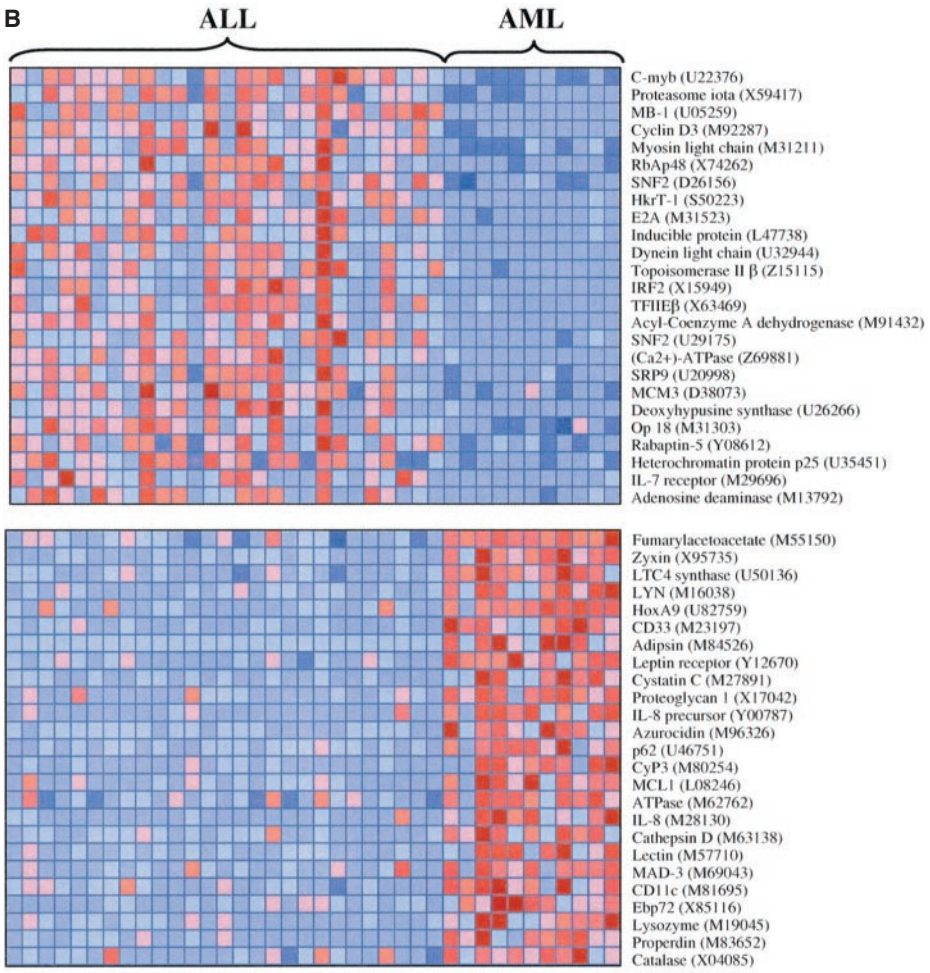
Clustering

REPORTS

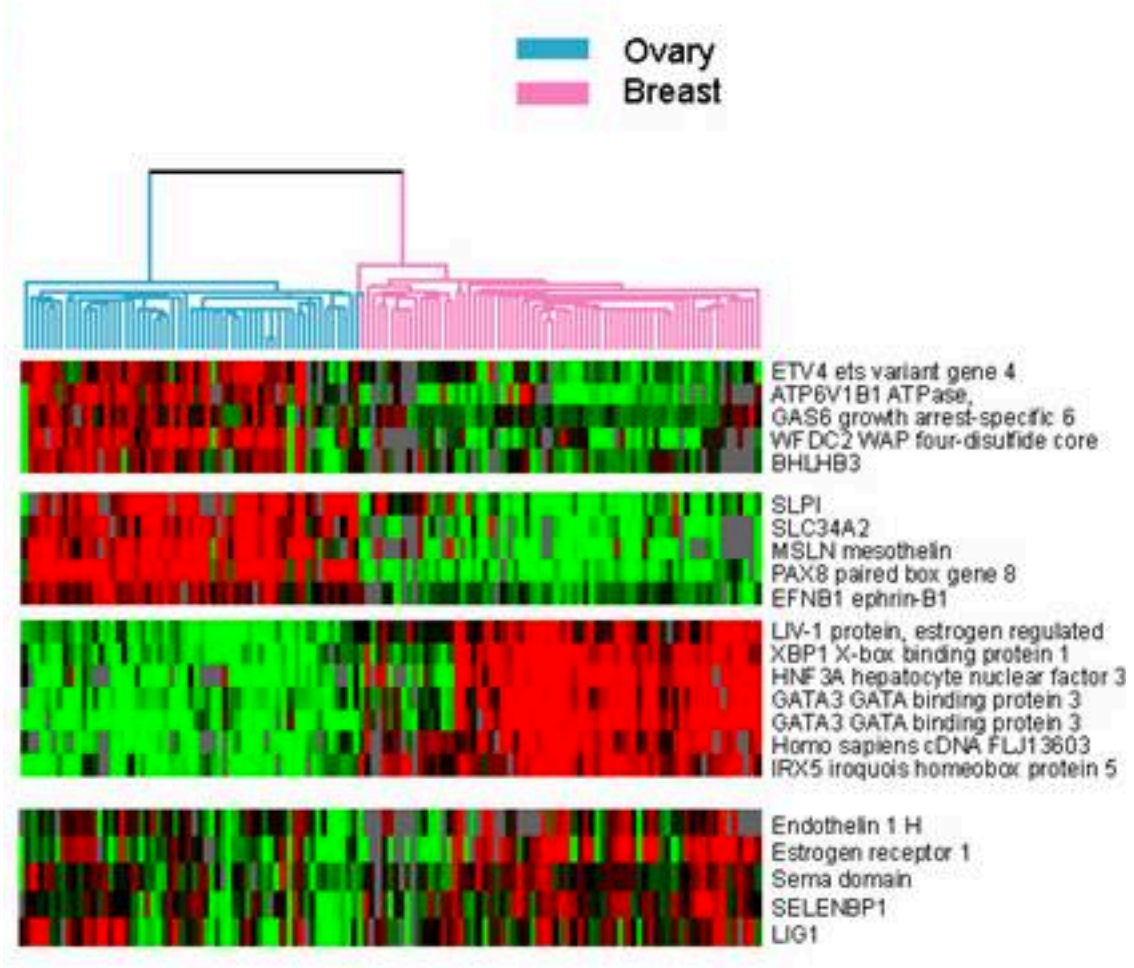
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]





Hierarchical Clustering



Clustering

• Different clustering methods will give you different views of the data

• There is no “correct” clustering method—clustering is just a guide that helps you to see the data in different ways

• On the other hand, there are “incorrect” methods—a certain amount of mathematical validation should be undertaken

• It’s often worth trying a variety of clustering methods to see what is the most useful for your purposes

Principal Component Analysis

- Reduces the dimensionality of the data set
  - Thousands of genes are combined in a few linear combinations to make 2 or 3 Principal Components (PC). Going from thousands of axes, with each axis representing the expression level for a gene, to 2 or 3 axes.

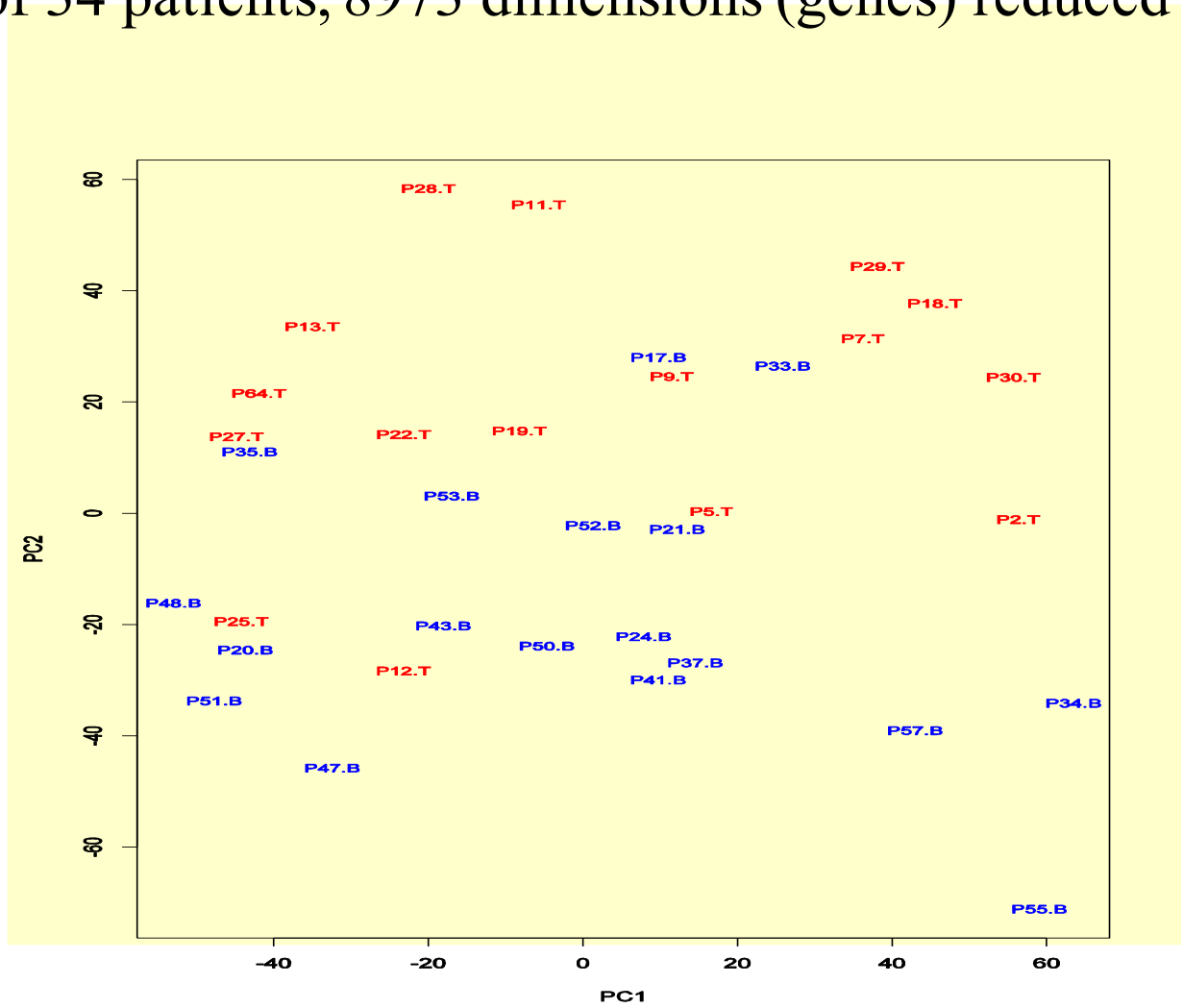
- These few PCs may capture most of the variability of the original data set

- Hope is that the first few PCs extract or expose the cluster structure of the original data set
  - i.e. Another clustering algorithm still needed after PCA

PCA on leukemia data

(precursor B and T)

- Plot of 34 patients, 8973 dimensions (genes) reduced to 2



Student t-test can be used to filter out differential expressed genes

- Compares the means ($\overline{x}_1$& $\overline{x}_2$) of two data sets
  - tells us if they can be assumed to be equal

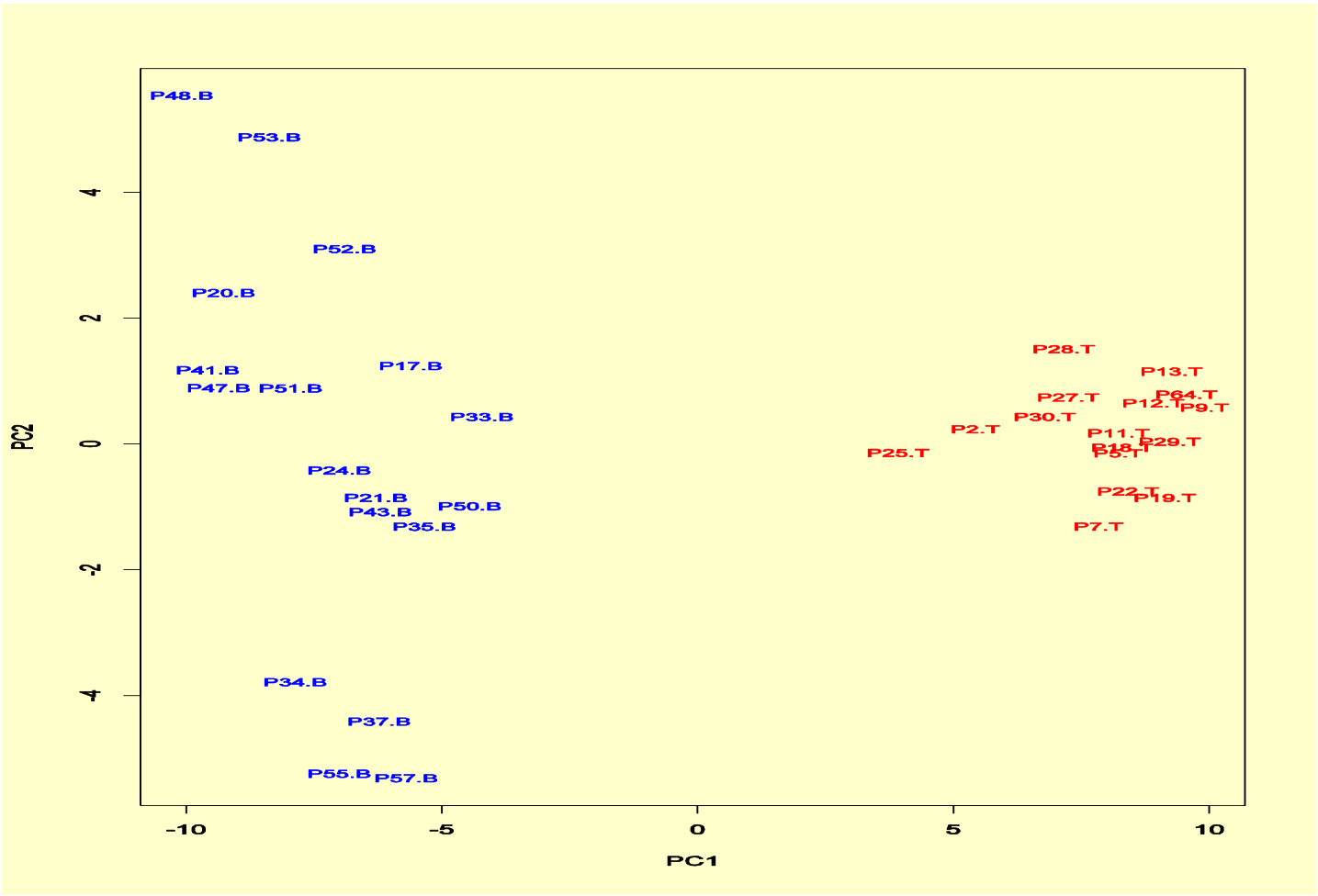
- Can be used to identify significant genes
  - *i.e.* those that change their expression *a lot*!

$$t = \frac{\overline{x}_1 - \overline{x}_2}{SE}, \qquad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

PCA on leukemia data

(precursor B and T)

- Plot of 34 patients, 100 dimensions (genes) reduced to 2



Clustering

Limitation of cluster analysis:

similarity in expression pattern suggests co-regulation but doesn't reveal cause-effect relationships