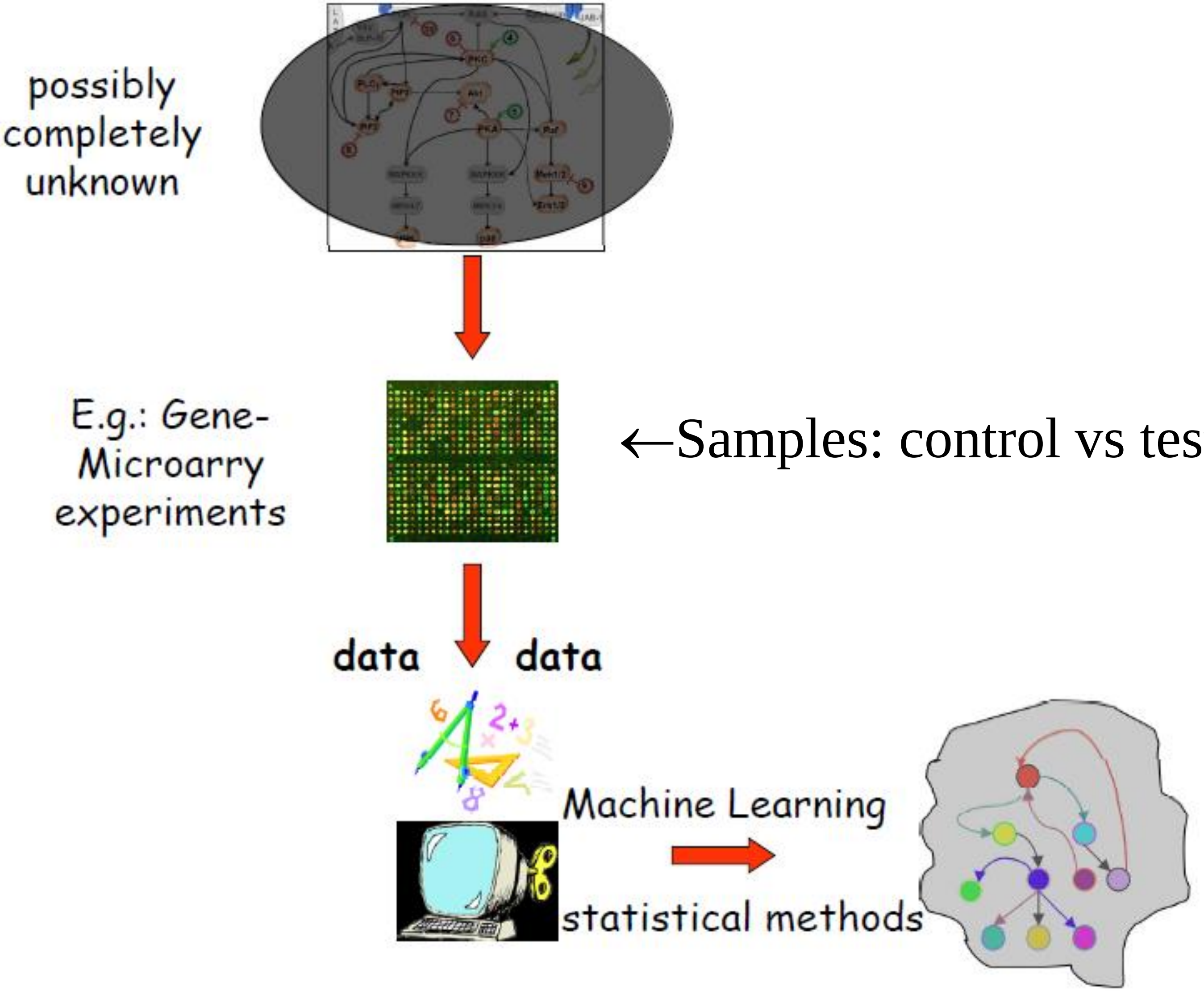
GLOBEX Bioinformatics (Summer 2015)

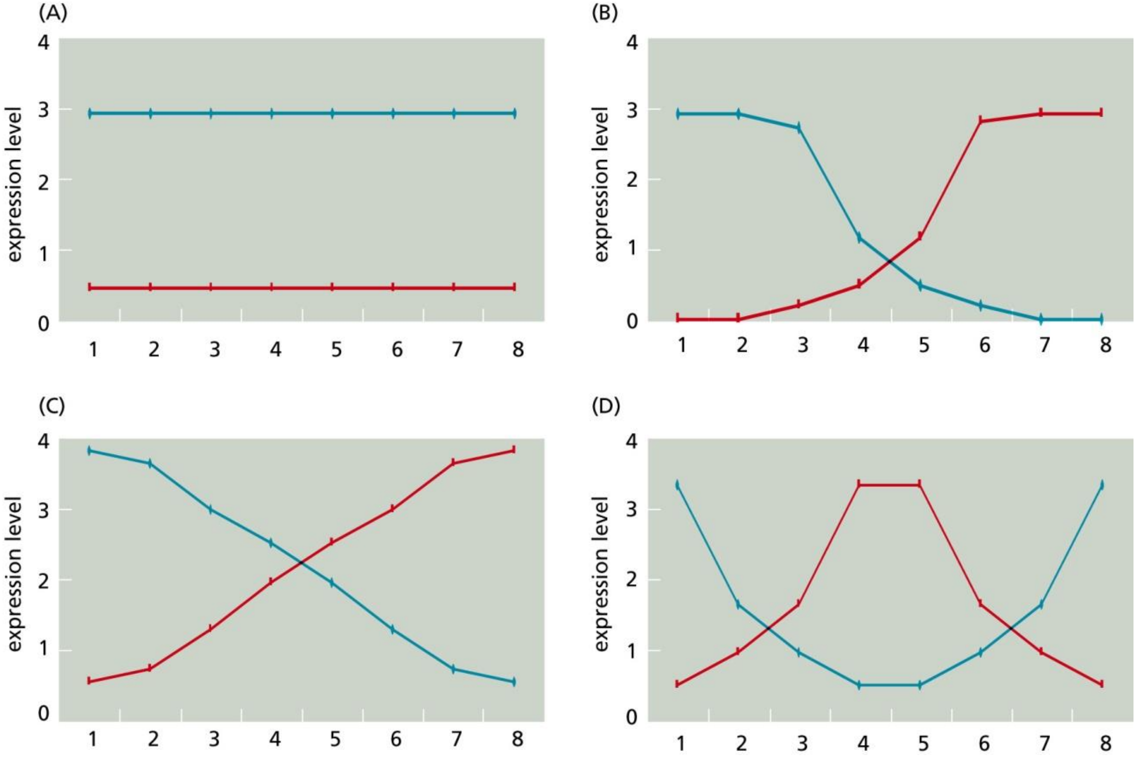
Systems biology: Gene expressions profiling and clustering

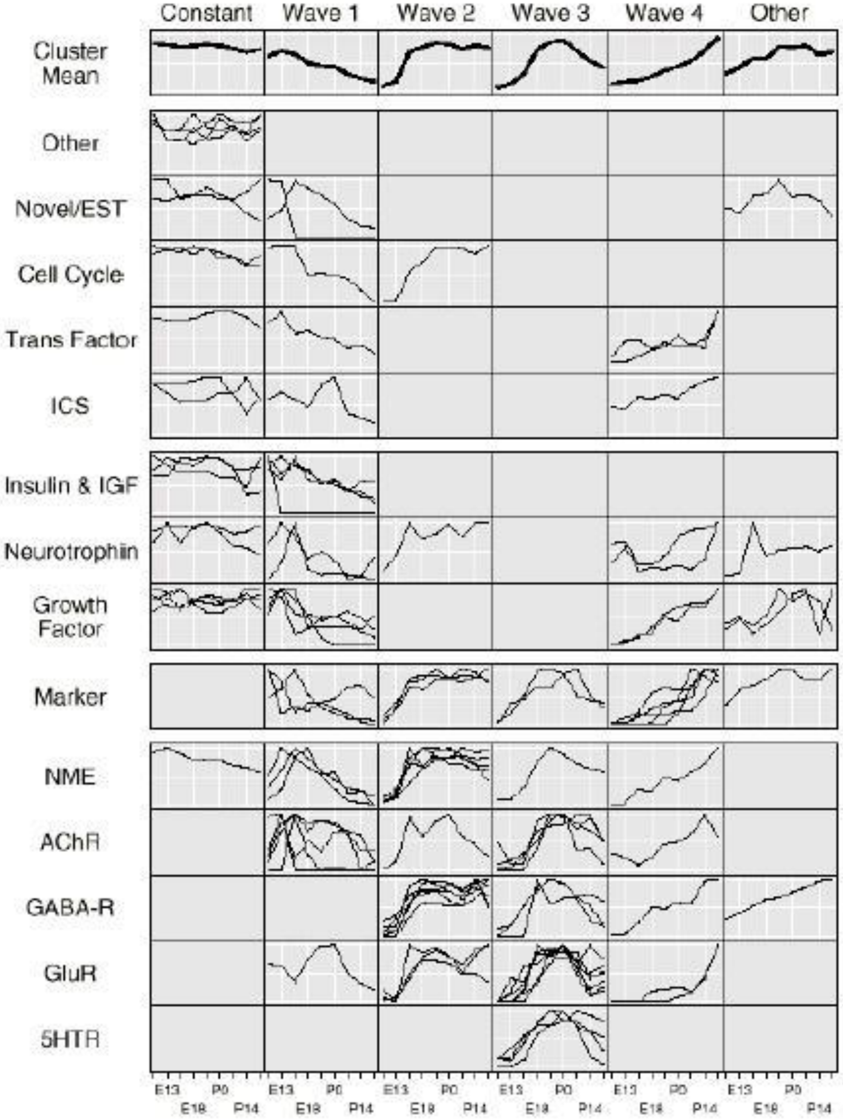
Genotypes ⇔ Phenotypes



←Samples: control vs test

Typical expression profiles




Constant
Wave 1
Wave 2
Wave 3
Wave 4
Other
Cluster Mean
Other
Novel/EST
Cell Cycle
Trans Factor
ICS
Insulin & IGF
Neurotrophin
Growth Factor
Marker
NME
AChR
GABA-R
GluR
5HTR
E13
P0
E18
P14

Why clustering?

“Searching for meaningful information patterns and dependencies among genes, in order to provide a basis for hypothesis testing, typically includes the initial step of grouping genes, with similar changes in expression into groups or ‘clusters’”.

Exploratory and unsupervised

Clustering the microarray matrix can be achieved in three ways:

i. Genes can form a group which show similar expression across conditions
ii. Samples can form a group which show similar GE across all genes
iii. Bi-clusters: genes and samples are clustered simultaneously, giving rise to a subset of genes and a subset of samples.

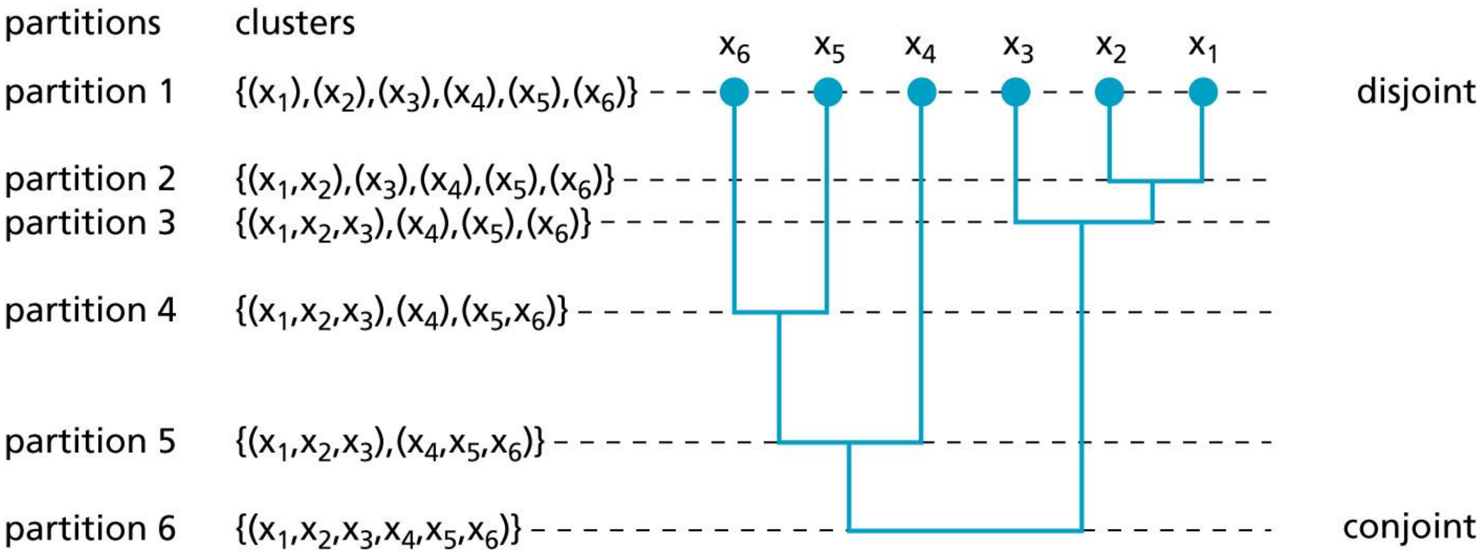
Clustering types:

- One level v.s. Hierarchical
- Exclusive v.s. overlapping
- Boolean v.s. fuzzy

Clustering prerequisites:

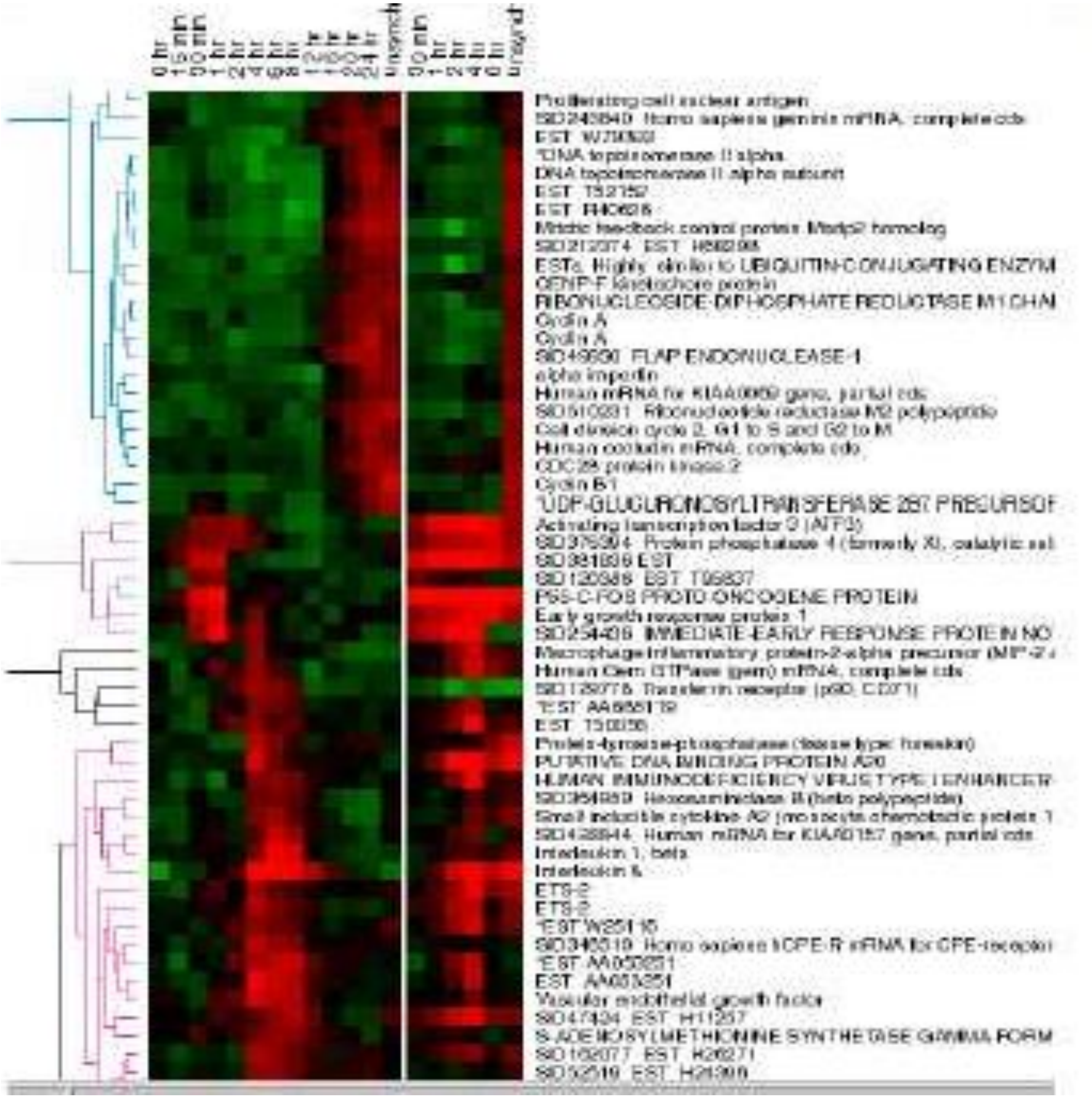
- Pattern representation; feature selection and extraction, e.g. PCA
- Definition of pattern proximity – measure of "distance", e.g., Euclidean distance, Mahalanobis distance, correlation distances
- Clustering
- Data abstraction
- Assessment of clusters: validation – internal, external and relative

Hierarchical clustering

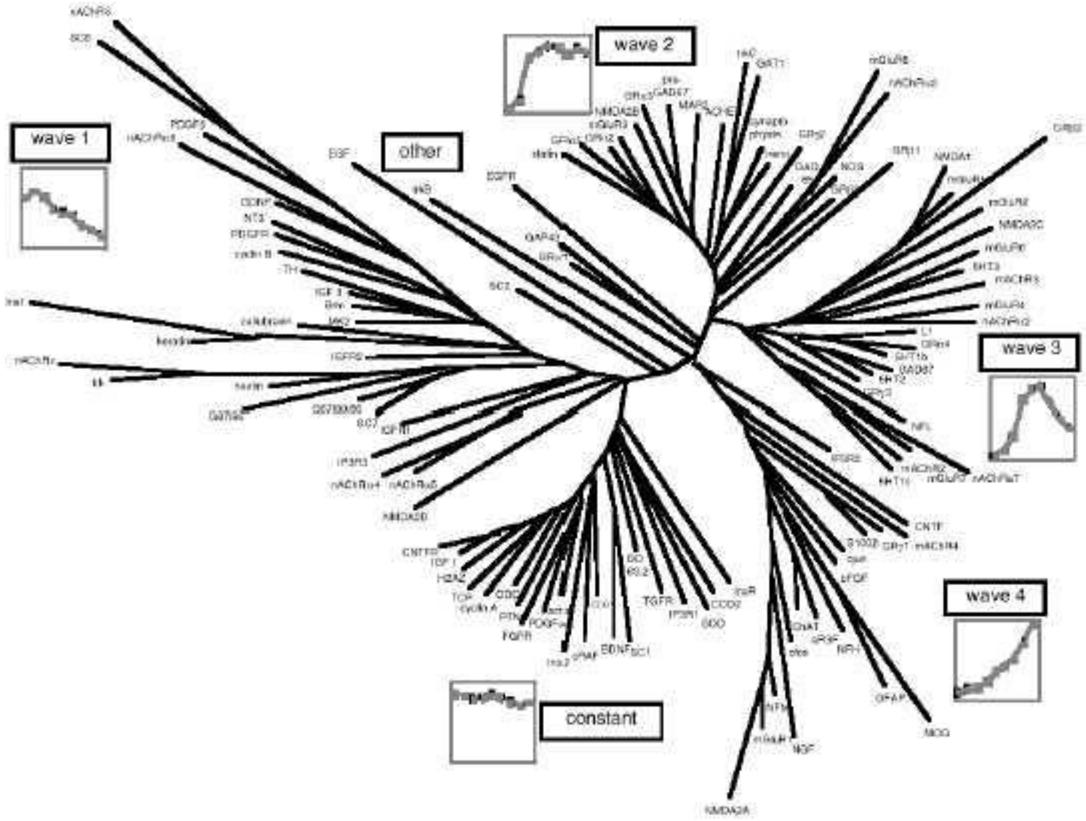


partitions	clusters
partition 1	$\{(x_1),(x_2),(x_3),(x_4),(x_5),(x_6)\}$
partition 2	$\{(x_1,x_2),(x_3),(x_4),(x_5),(x_6)\}$
partition 3	$\{(x_1,x_2,x_3),(x_4),(x_5),(x_6)\}$
partition 4	$\{(x_1,x_2,x_3),(x_4),(x_5,x_6)\}$
partition 5	$\{(x_1,x_2,x_3),(x_4,x_5,x_6)\}$
partition 6	$\{(x_1,x_2,x_3,x_4,x_5,x_6)\}$

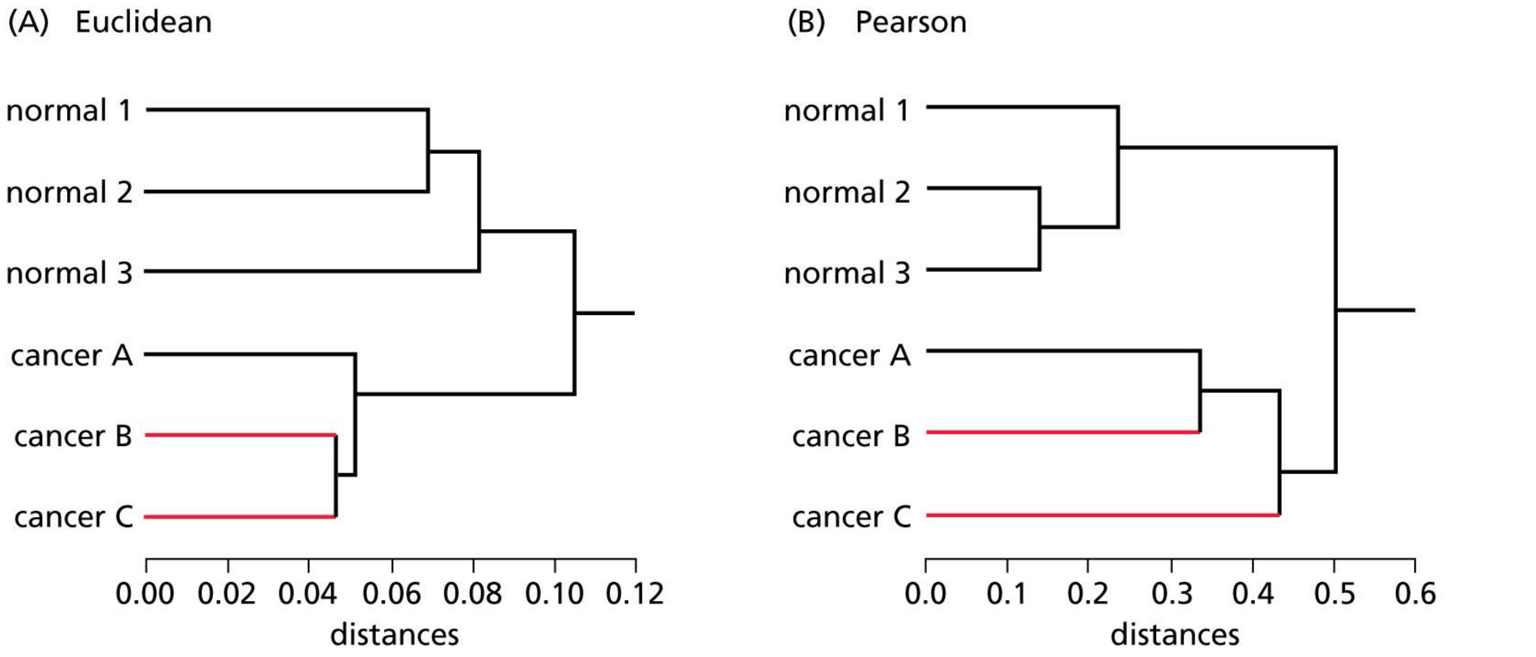
Hierarchical clustering



Russ Altman


wave 1
wave 2
other
wave 3
wave 4
constant
PDGFb
EGF
EGFR
GAP43
SC2
CNTF
NMDA2A
NGF
GFAP
MOG
bFGF
ChAT
TGFR
IP3R1
IP3R3
InsR
BDNF
SC1
NMDA2B
NMDA2C
NMDA2D
NMDA1
mGluR2
mGluR3
5HT2
5HT3
L1
NT3
PDGFR
TH
MK2
IGF I
CNTFR
TCP
cyclin A
cyclin B
H2AZ
FGFR
PDGFa
GAD67
GAT1
MAP2
ACHE
nestin
NFL
NFH
S100b
EGF

Effects of various metrics for measuring distance



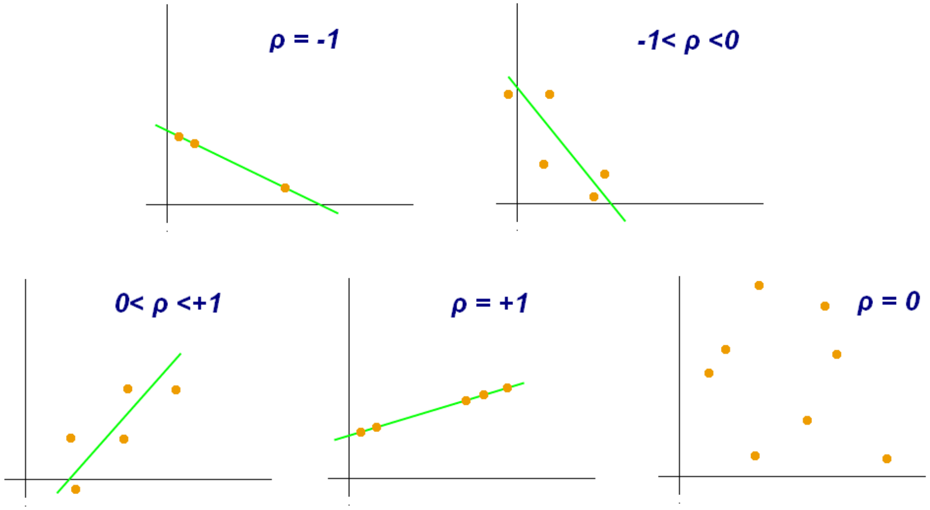
$d_{x,y} = \sqrt{} \sum(x_i - y_i)^2$

$d_{X,Y} = 1 - \rho_{X,Y}.$

Pearson correlation coefficient

For a population

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$



For a sample

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Pearson distance: $d_{X,Y} = 1 - \rho_{X,Y}.$

Effect of different clustering schemes

level

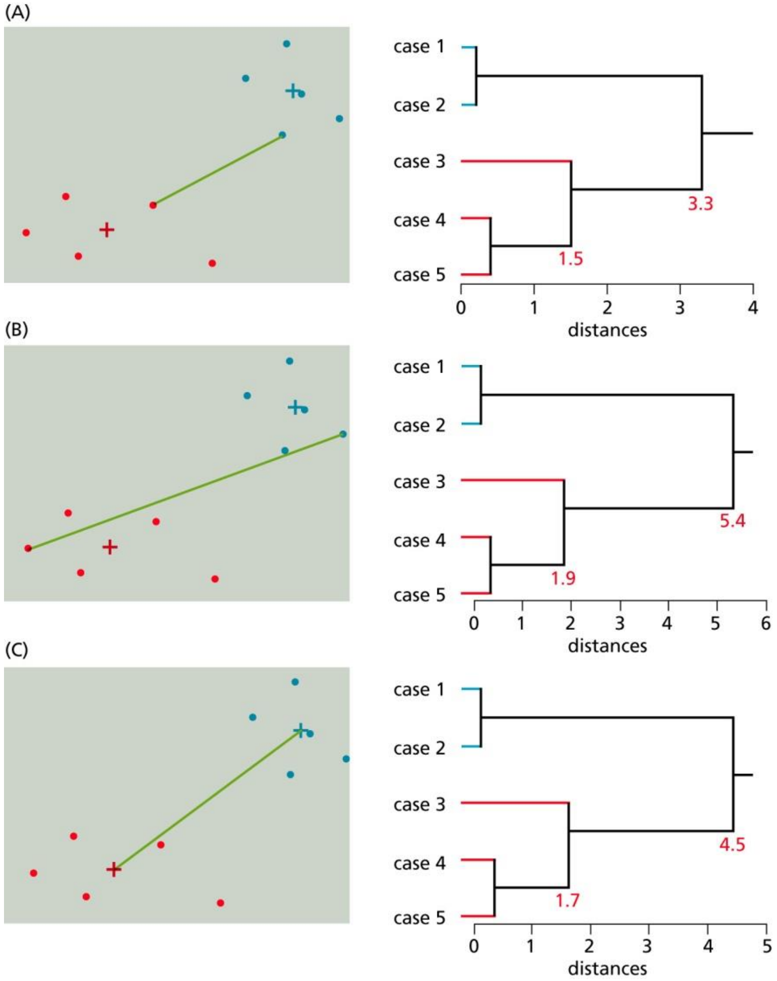
case1: 2.0

case 2: 2.2

case 3: 5.5

case 4: 7.0

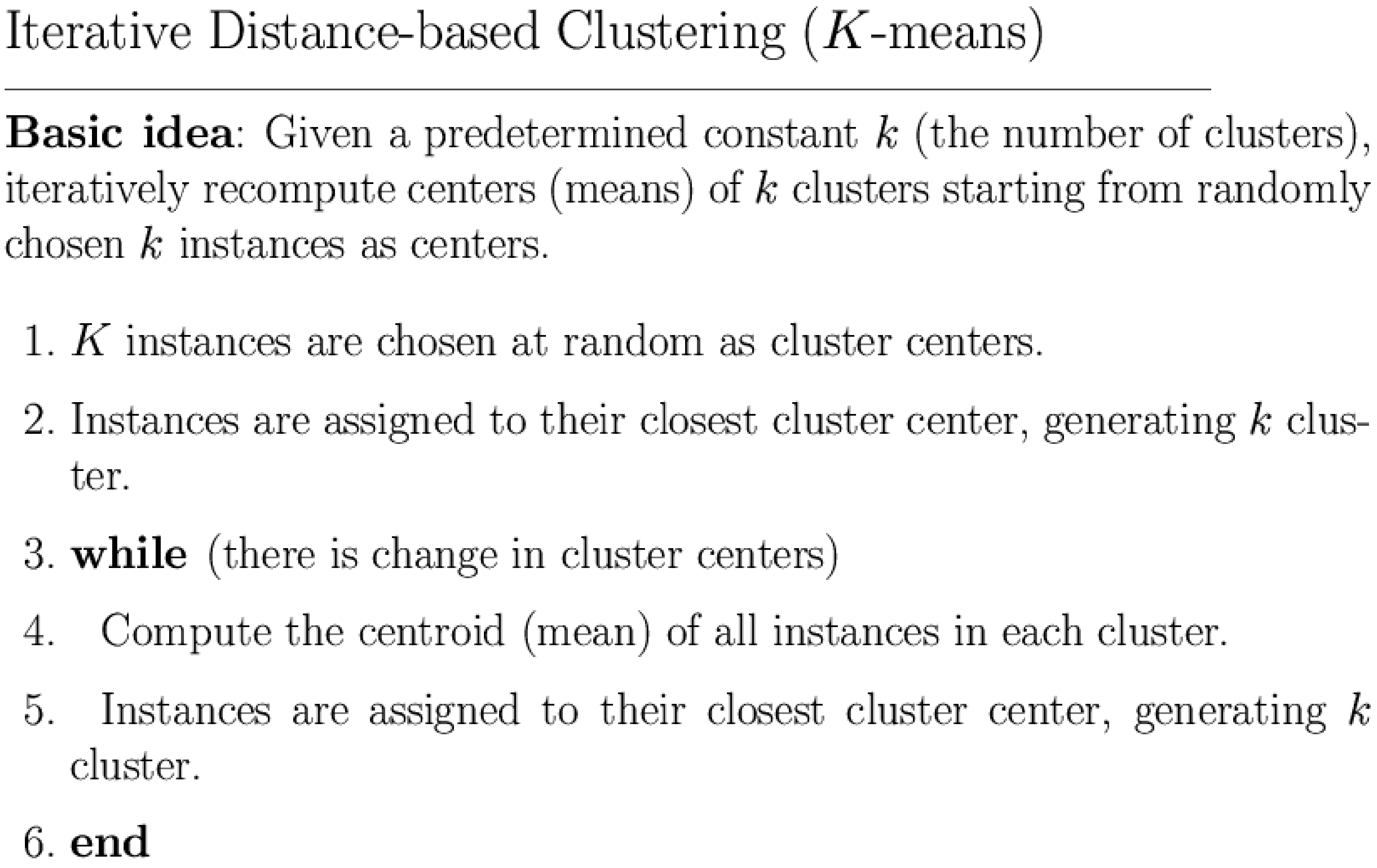
case 5: 7.4



Single linkage

Complete linkage

centroid



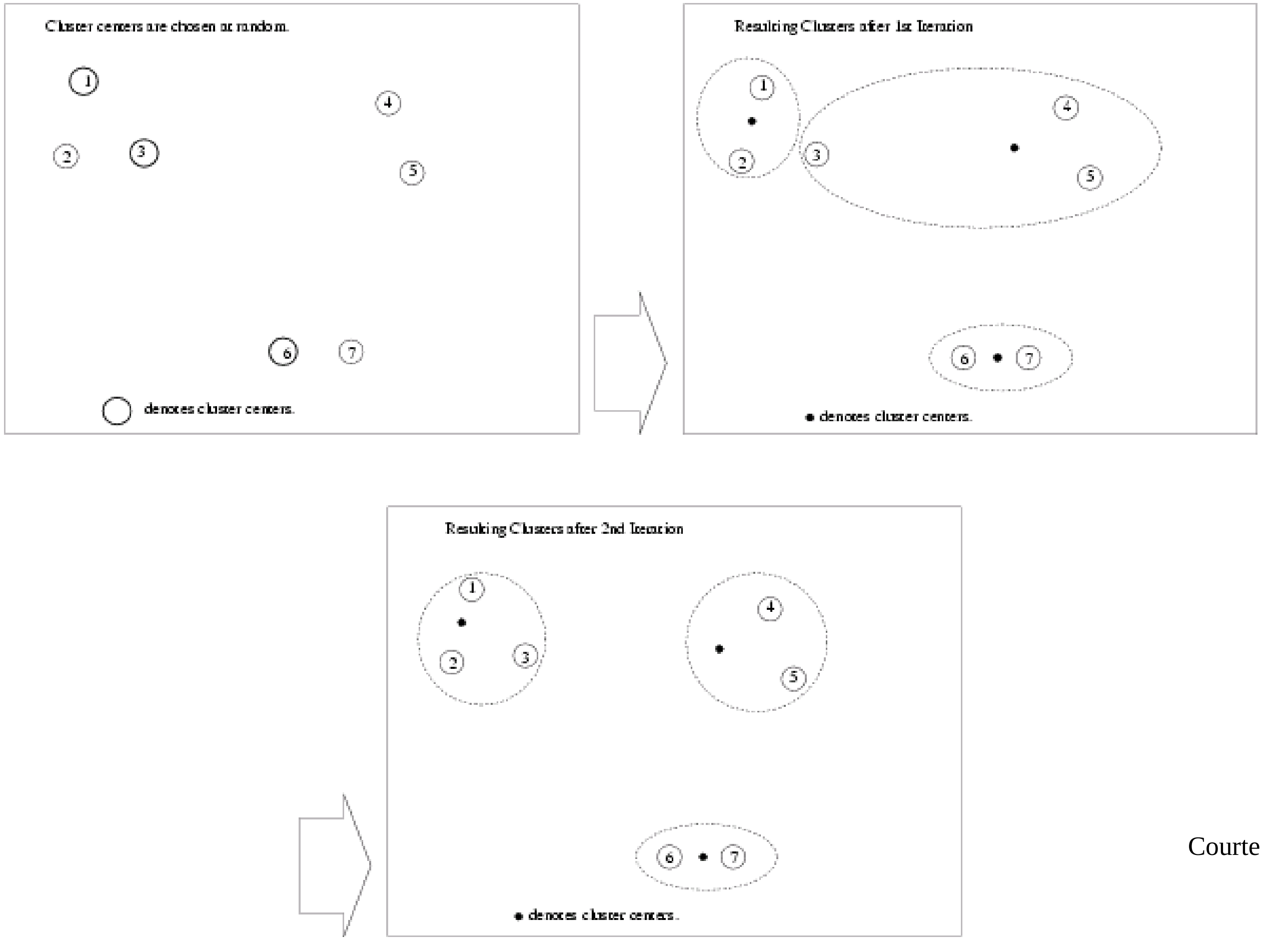
Iterative Distance-based Clustering (K-means)

Basic idea: Given a predetermined constant k (the number of clusters), iteratively recompute centers (means) of k clusters starting from randomly chosen k instances as centers.

```
1. $K$ instances are chosen at random as cluster centers.
2. Instances are assigned to their closest cluster center, generating $k$ cluster.
3. while (there is change in cluster centers)
4.    Compute the centroid (mean) of all instances in each cluster.
5.    Instances are assigned to their closest cluster center, generating $k$ cluster.
6. end
```

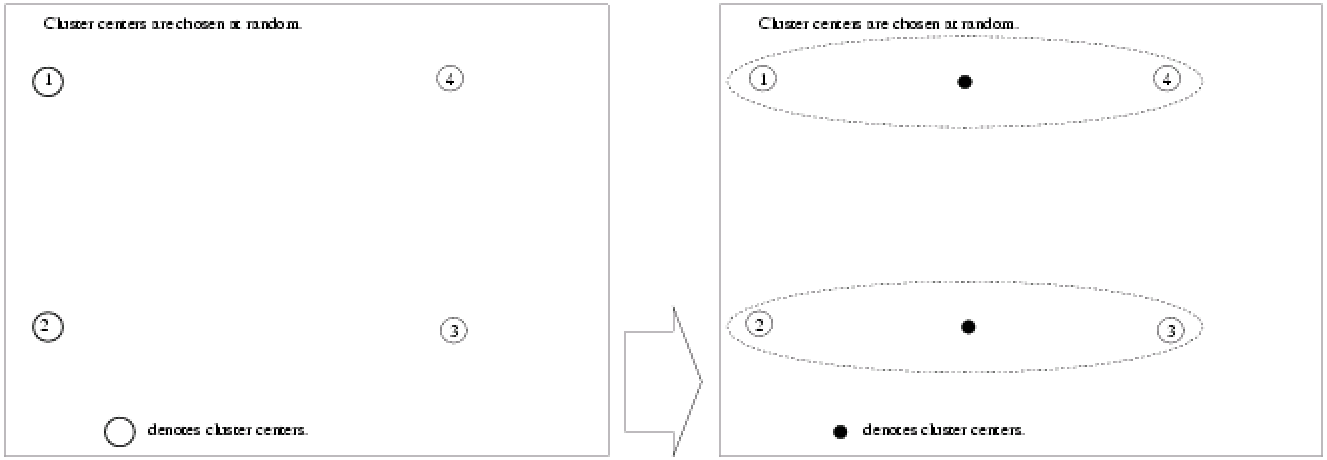
Courtesy of Sun Kim

A Correct Clustering Example


Cluster centers are chosen at random.
1
2
3
4
5
6
7
denotes cluster centers.
Resulting Clusters after 1st Iteration
1
2
3
4
5
6
7
denotes cluster centers.
Resulting Clusters after 2nd Iteration
1
2
3
4
5
6
7
denotes cluster centers.

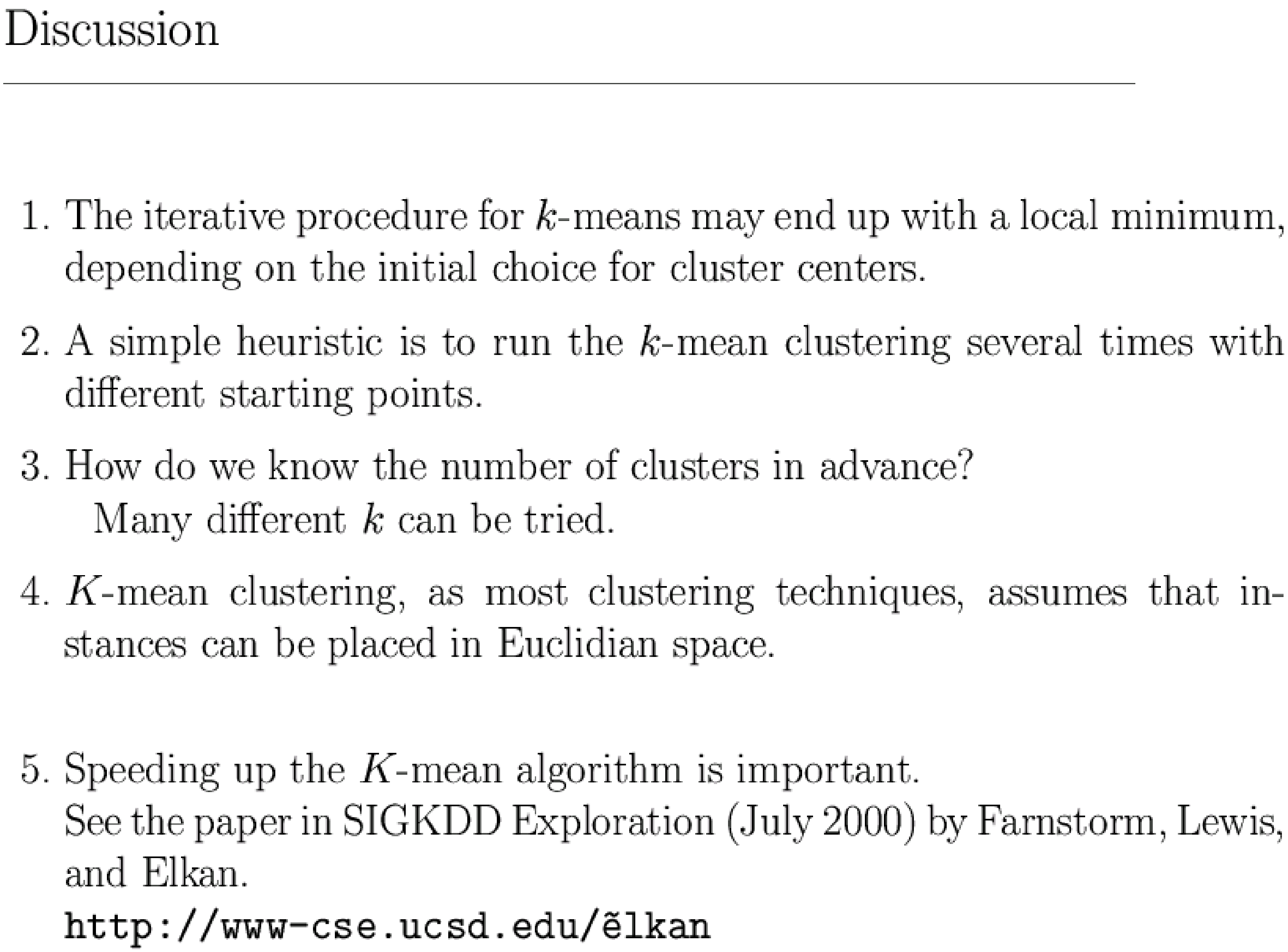
Courtesy of Sun Kim

An Incorrect Clustering Example


Cluster centers are chosen at random.
1
4
2
3
denotes cluster centers.
Cluster centers are chosen at random.
1
4
2
3
denotes cluster centers.

The initial choice of cluster centers, node 1 and node2, leads to an incorrect clustering. Obviously. a different choice of cluster centers, node 1 and node 3, result in a correct clustering.

Courtesy of Sun Kim



Discussion

1. The iterative procedure for k-means may end up with a local minimum, depending on the initial choice for cluster centers.
2. A simple heuristic is to run the k-mean clustering several times with different starting points.
3. How do we know the number of clusters in advance?
   Many different k can be tried.
4. K-mean clustering, as most clustering techniques, assumes that instances can be placed in Euclidian space.
5. Speeding up the K-mean algorithm is important.
   See the paper in SIGKDD Exploration (July 2000) by Farnstorm, Lewis, and Elkan.
   `http://www-cse.ucsd.edu/ẽlkan`

Courtesy of Sun Kim

Fuzzy k-means clustering

Fuzzy membership: Each data point $\mathbf{x}$ has some probability to belong to a cluster w (centered at $\mathbf{u}$).

$$P(w|\mathbf{x})$$

The probabilities of cluster membership for each point are normalized

$$\sum_{i = 1 \text{ to } k} P(w_i|\mathbf{x}_j) = 1 \text{ for } j = 1, \ldots, n \qquad (1)$$

Cluster cost:

$$J = \sum_{i = 1 \text{ to } k} \sum_{j = 1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b \, ||\mathbf{x}_j - \mathbf{u}_i||^2. \qquad (2)$$

Condition for minimum cost:

$$\partial J / \partial \mathbf{u}_i = 0$$

$$\mathbf{u}_i = (\sum_{j = 1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b \, \mathbf{x}_j)/(\sum_{j = 1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b \,) \qquad (3)$$

Update posterior probability as

$$P(w_i|\mathbf{x}_j) = (1/d_{ij})^{1/(b-1)} / \sum_{r=1 \text{ to } k} (1/d_{rj})^{1/(b-1)} \qquad (4)$$

where $d_{ij} = \|\mathbf{x}_j - \mathbf{u}_i\|^2$.

Fuzzy k-means clustering algorithm

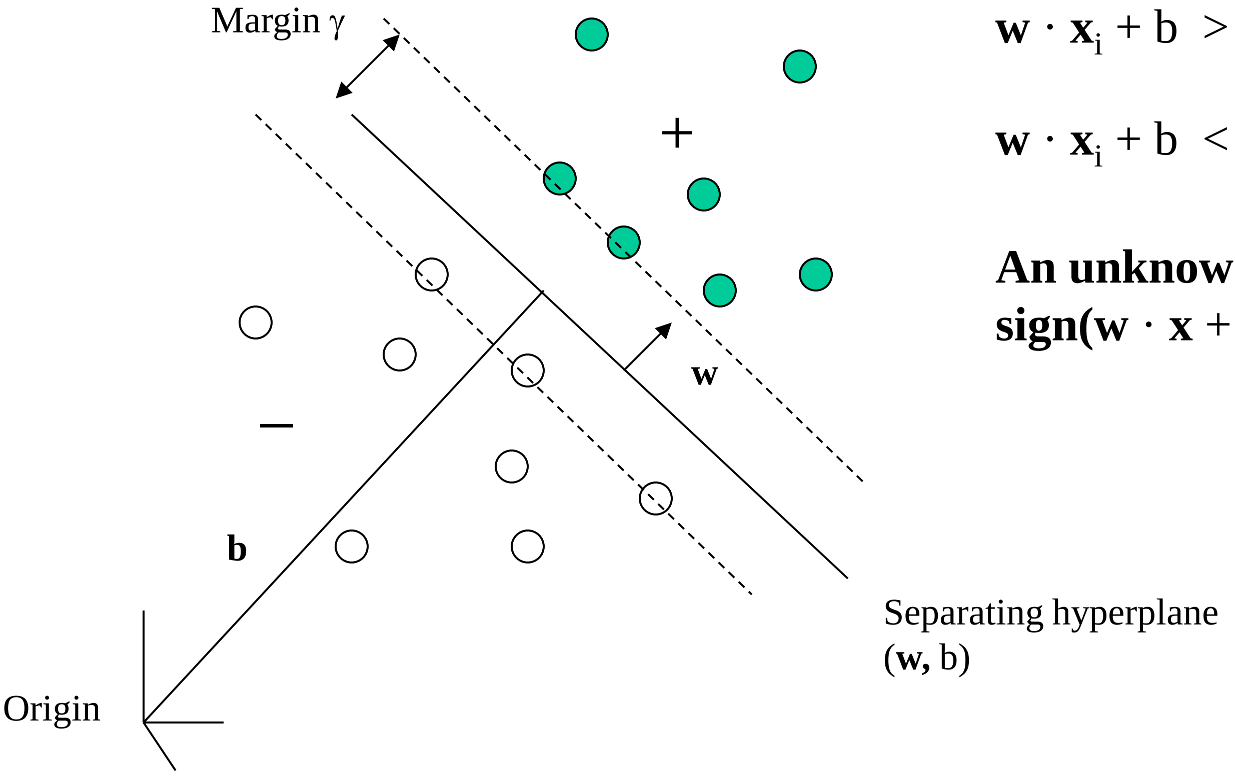
```
initialize $\mathbf{u}_1,\ldots, \mathbf{u}_k$
    normalize $P(w_i|\mathbf{x}_j)$ by eq(1)
    do recompute $\mathbf{u}_i$ for i = 1 to k by eq(3)
           recompute $P(w_i|\mathbf{x}_j)$ by eq(4)
    until small change in $\mathbf{u}_i$ and $P(w_i|\mathbf{x}_j)$
return $\mathbf{u}_1,\ldots, \mathbf{u}_k$.
```

Classical k-means is a special case when membership is defined as

$$P(w_i|\mathbf{x}_j) = 1 \quad \text{if } \|\mathbf{x}_j - \mathbf{u}_i\| < \|\mathbf{x}_j - \mathbf{u}_{i'}\| \text{ for all } i' \neq i.$$

$$= 0 \quad \text{otherwise.}$$

Support vector machine (SVM)

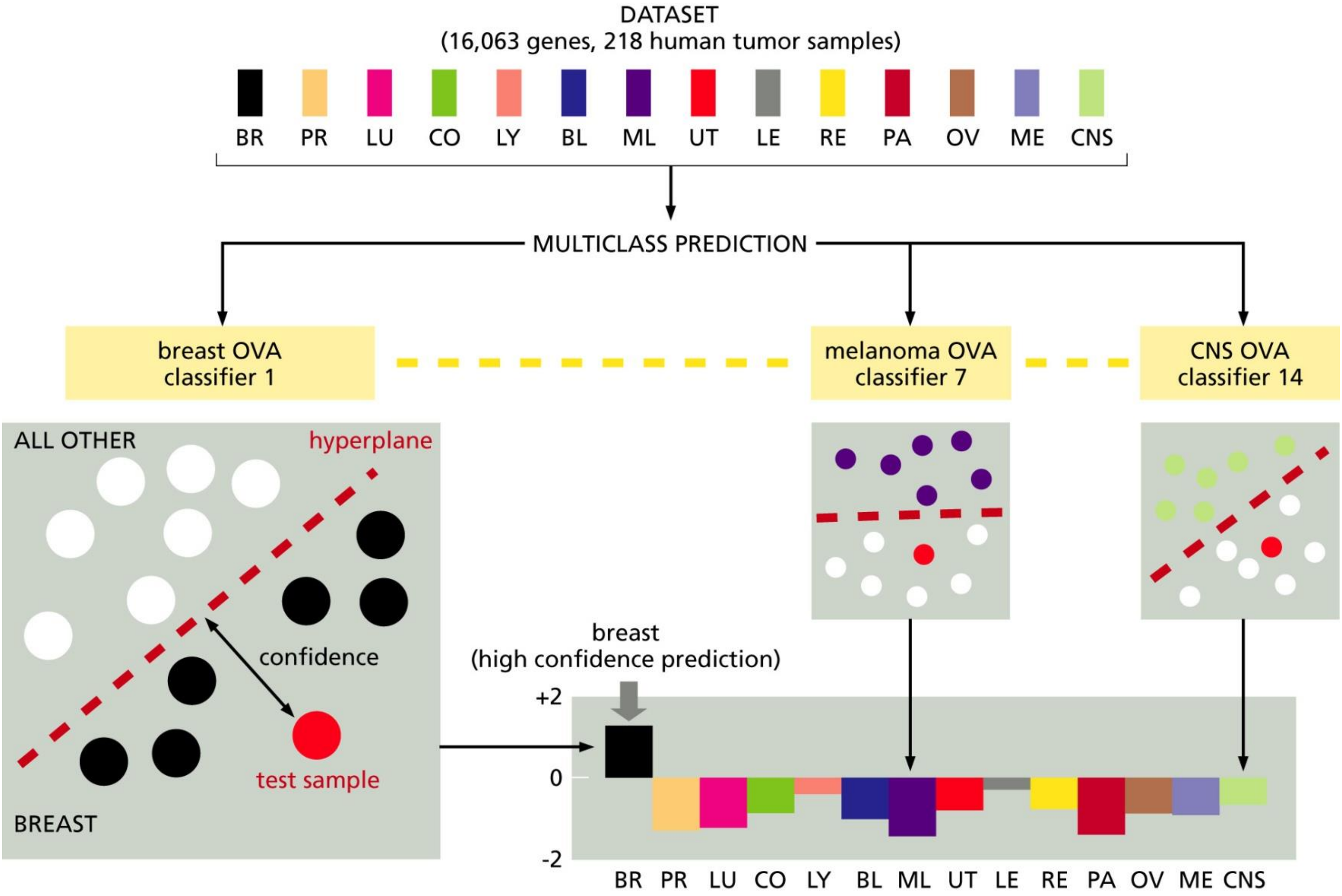


$\mathbf{w} \cdot \mathbf{x}_i + b > 0$ if $y_i = +1$

$\mathbf{w} \cdot \mathbf{x}_i + b < 0$ if $y_i = -1$

An unknown x is classified as sign(w · x + b)

Application of SVM classification



Gene Set Enrichment

Gene set enrichment analysis: Identify sets of related genes ***associated*** with (possibly cause) phenotypic changes.

- Discovery/exploratory mode: discovery of gene sets that were not previously known to be related. (More difficult; see clustering).
- Validation mode: determination and confirmation of which set among a known collection. (Require prior and domain knowledge).

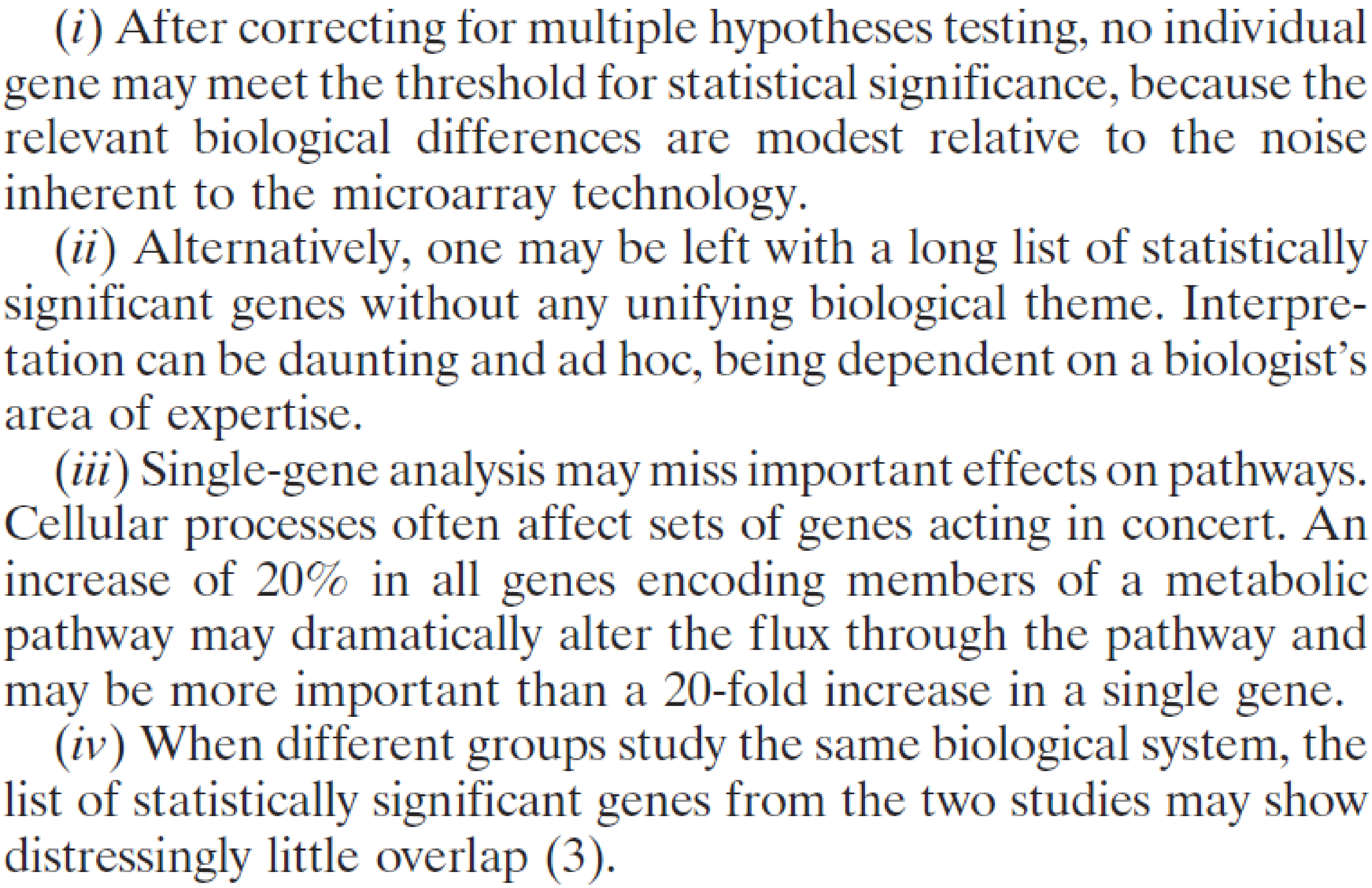
Early approach:

- Examine individual genes by their differential expression between phenotypes, exceeding a preset threshold, say p-value < 0.01. A binary decision for prescreening.
- Use Fisher's exact test to determine if selected genes belong to a pre-specified gene set.

Current approach:

- rank all genes according to differential expression
- determine if a pre-specified gene set is overrepresented toward the top or the bottom of the ranked list.

Motivations and challenges for gene set enrichment analysis

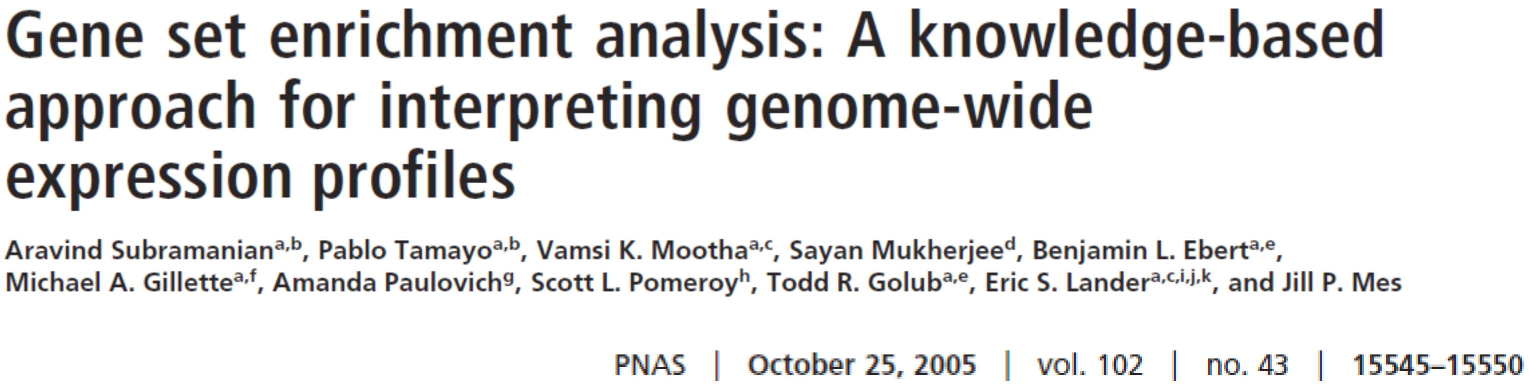


(*i*) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(*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.

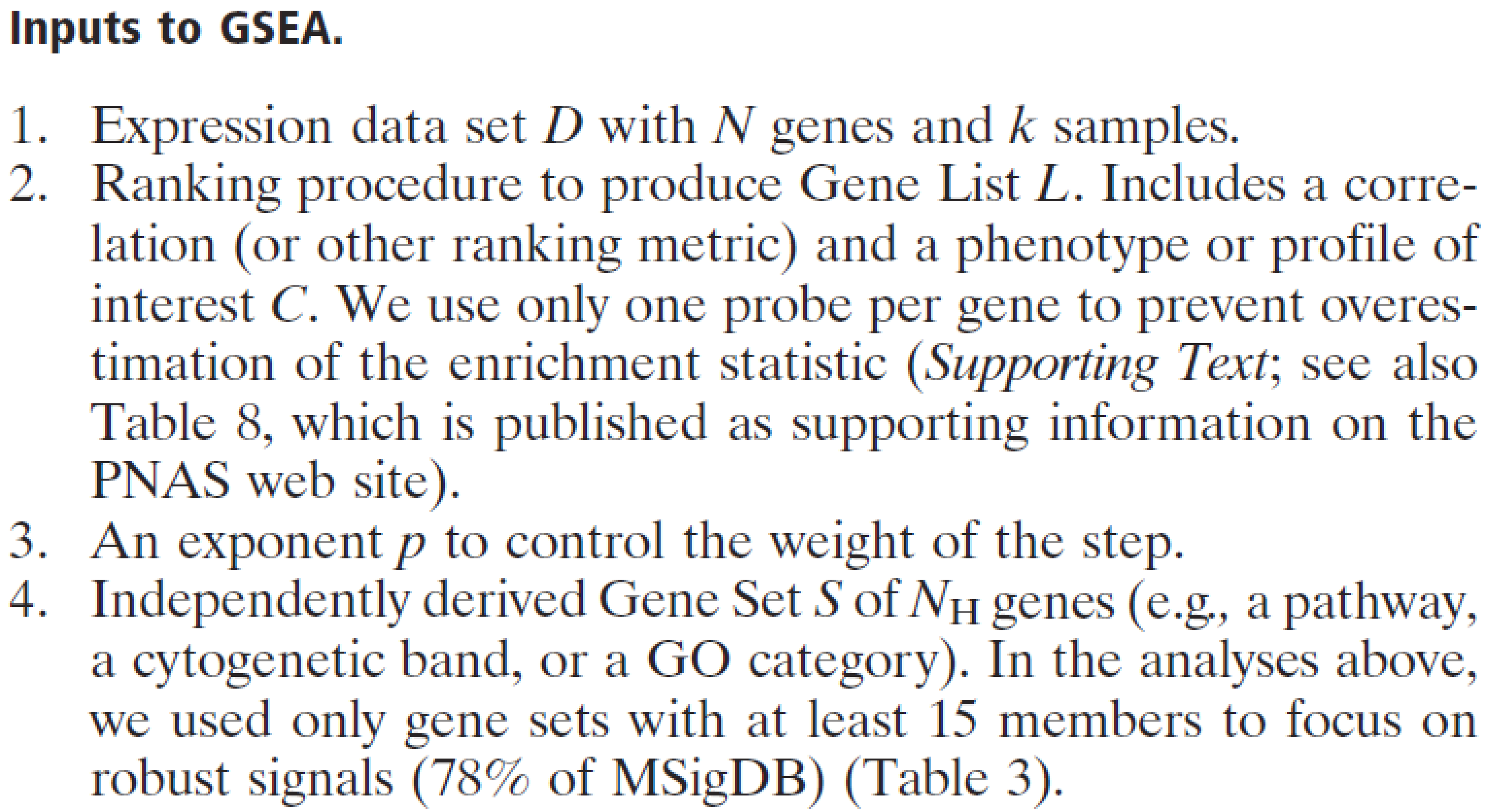
(*iv*) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap (3).



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

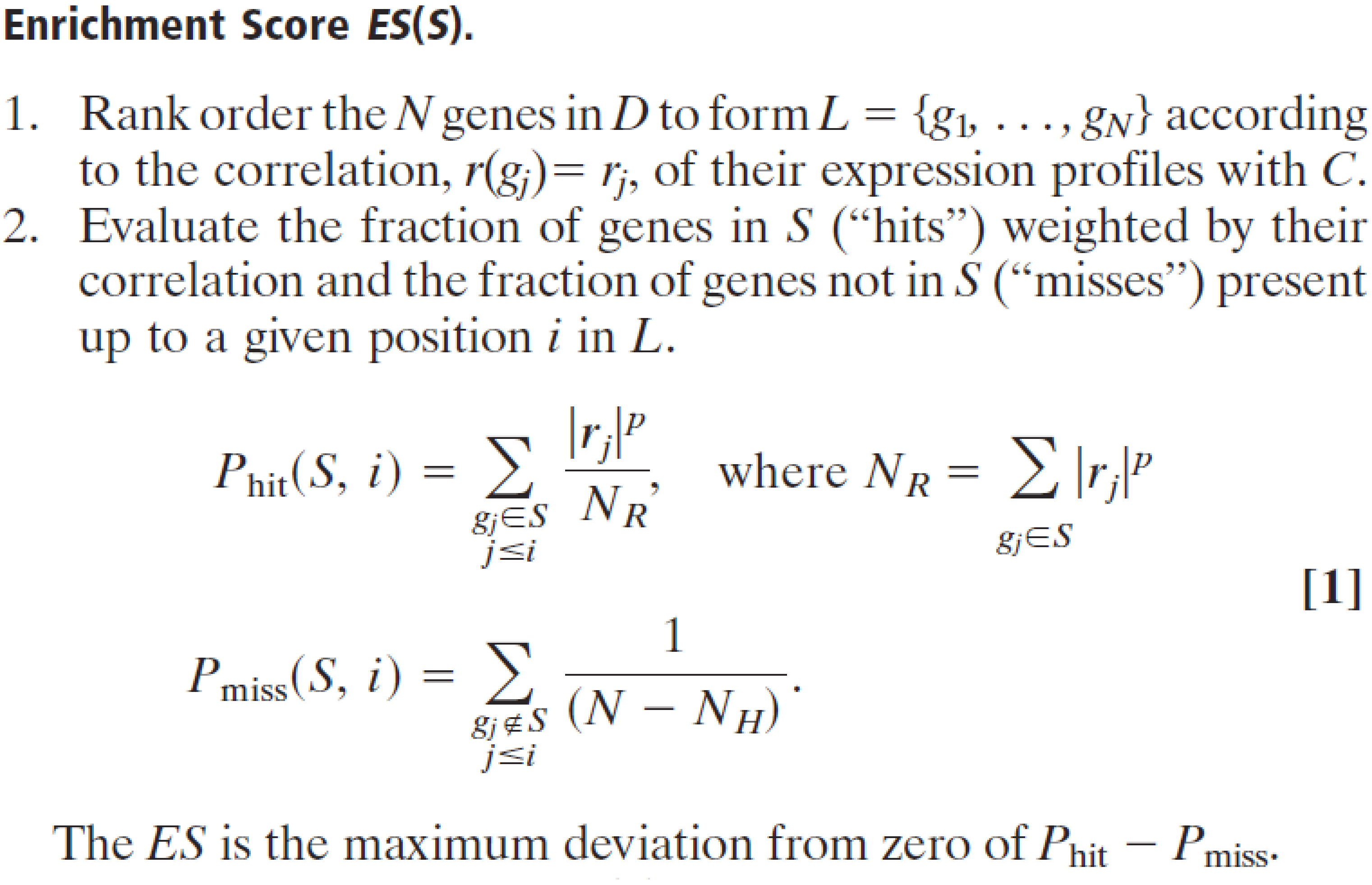
Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mes

PNAS | October 25, 2005 | vol. 102 | no. 43 | 15545–15550



Inputs to GSEA.

1. Expression data set D with N genes and k samples.
2. Ranking procedure to produce Gene List L. Includes a correlation (or other ranking metric) and a phenotype or profile of interest C. We use only one probe per gene to prevent overestimation of the enrichment statistic (*Supporting Text*; see also Table 8, which is published as supporting information on the PNAS web site).
3. An exponent p to control the weight of the step.
4. Independently derived Gene Set S of N_H genes (e.g., a pathway, a cytogenetic band, or a GO category). In the analyses above, we used only gene sets with at least 15 members to focus on robust signals (78% of MSigDB) (Table 3).



Enrichment Score *ES*(*S*).

1. Rank order the N genes in D to form $L = \{g_1, \ldots, g_N\}$ according to the correlation, $r(g_j) = r_j$, of their expression profiles with C.
2. Evaluate the fraction of genes in S ("hits") weighted by their correlation and the fraction of genes not in S ("misses") present up to a given position i in L.

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}. \qquad [1]$$

The ES is the maximum deviation from zero of $P_{\text{hit}} - P_{\text{miss}}$.

This is a weighted Kolmogorov-Smirnov like statistic.

Kolmogorov–Smirnov statistic [edit]

The empirical distribution function F_n for n iid observations X_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$$

where $I_{X_i \le x}$ is the indicator function, equal to 1 if $X_i \le x$ and equal to 0 otherwise.

The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

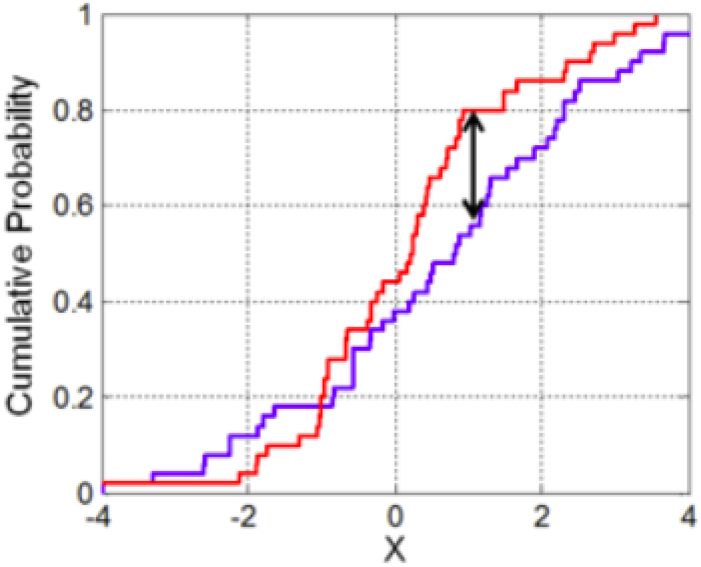
where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively.

The null hypothesis is rejected at level α if

$$D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}}.$$ [7]

The value of $c(\alpha)$ is given in the table below for each level of α[7]

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95



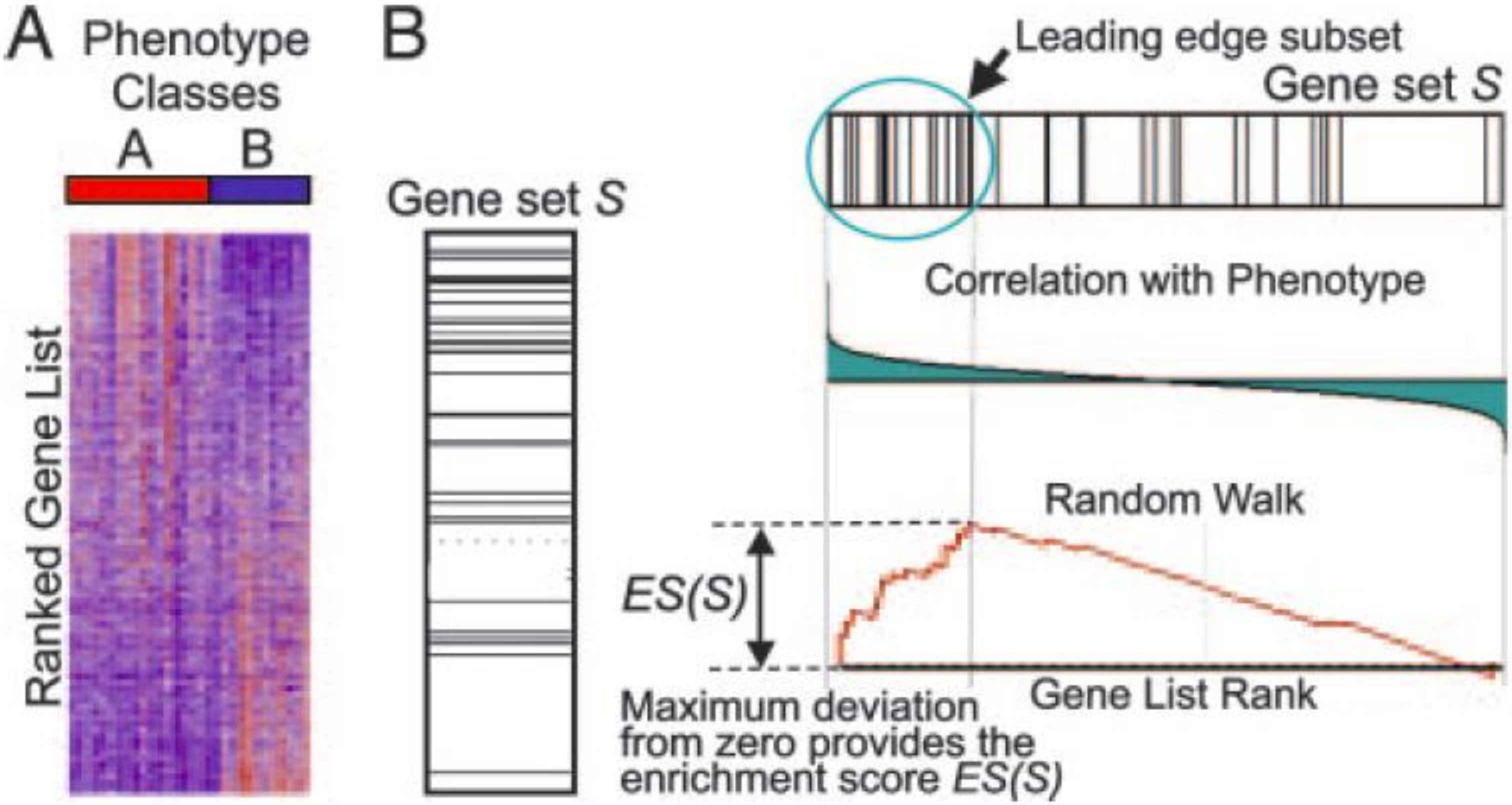
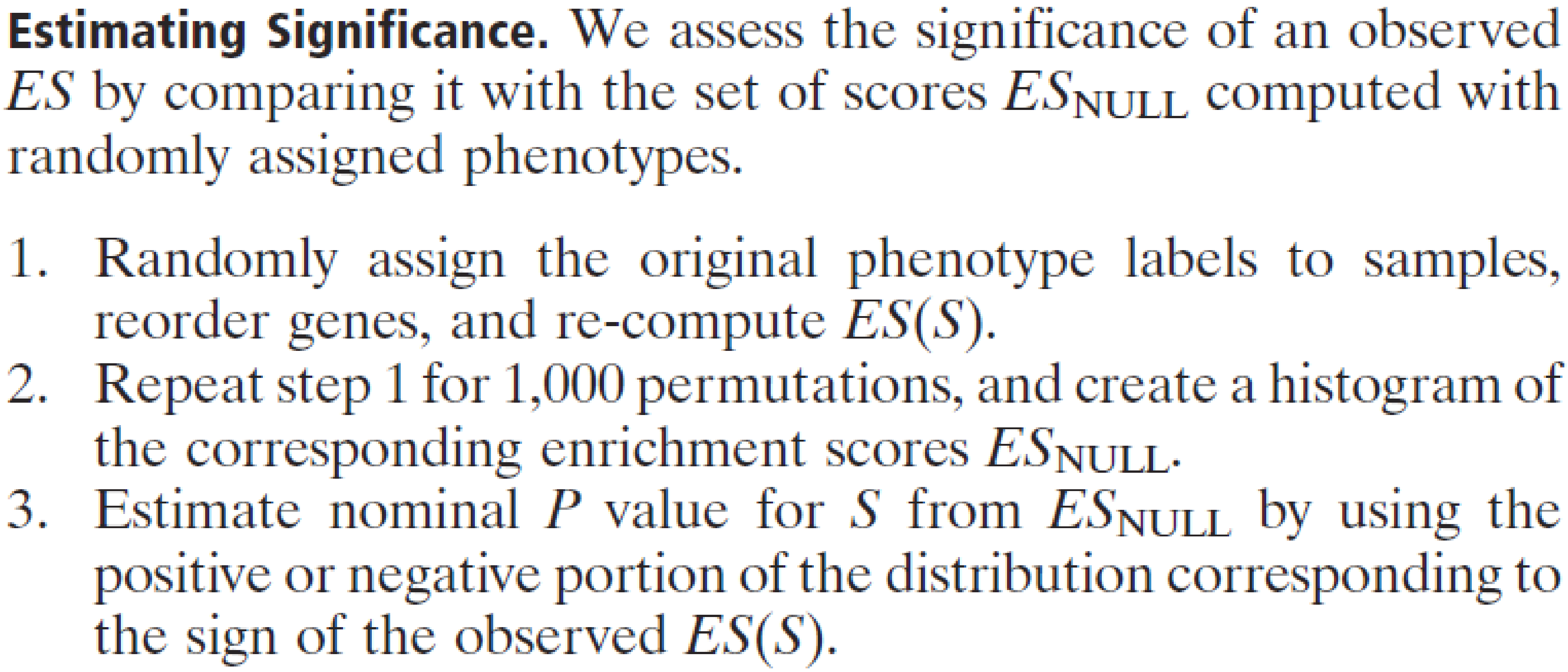


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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



Estimating Significance. We assess the significance of an observed *ES* by comparing it with the set of scores ES_{NULL} computed with randomly assigned phenotypes.

1. Randomly assign the original phenotype labels to samples, reorder genes, and re-compute $ES(S)$.
2. Repeat step 1 for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{NULL}.
3. Estimate nominal *P* value for *S* from ES_{NULL} by using the positive or negative portion of the distribution corresponding to the sign of the observed $ES(S)$.