6.874/... Recitation 2

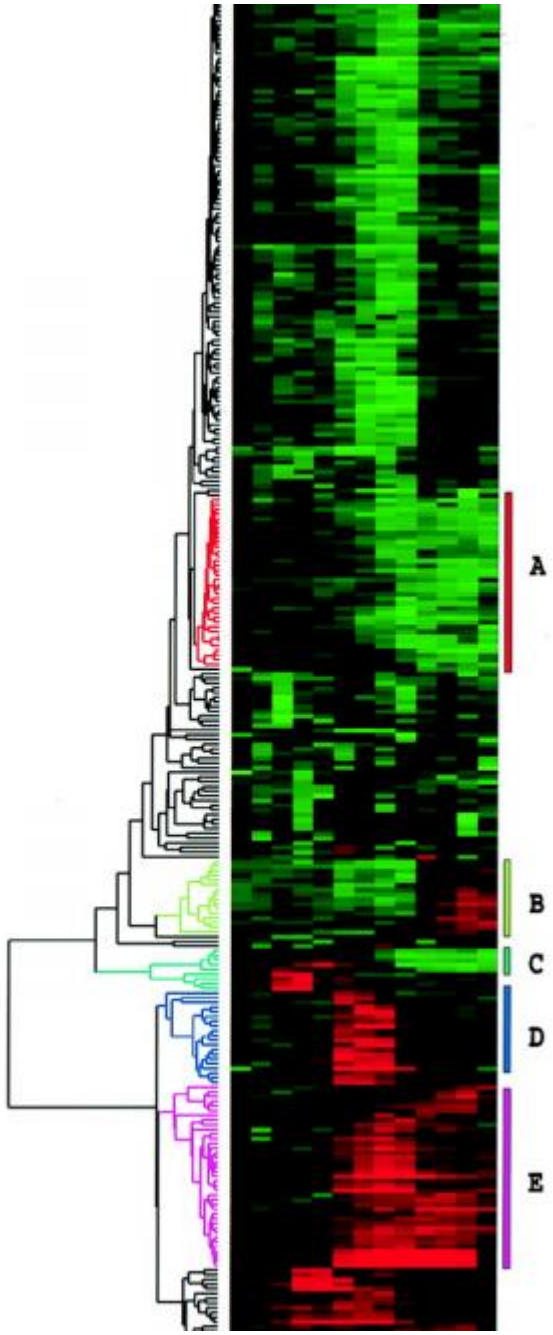
Courtesy of an MIT Teaching Assistant.

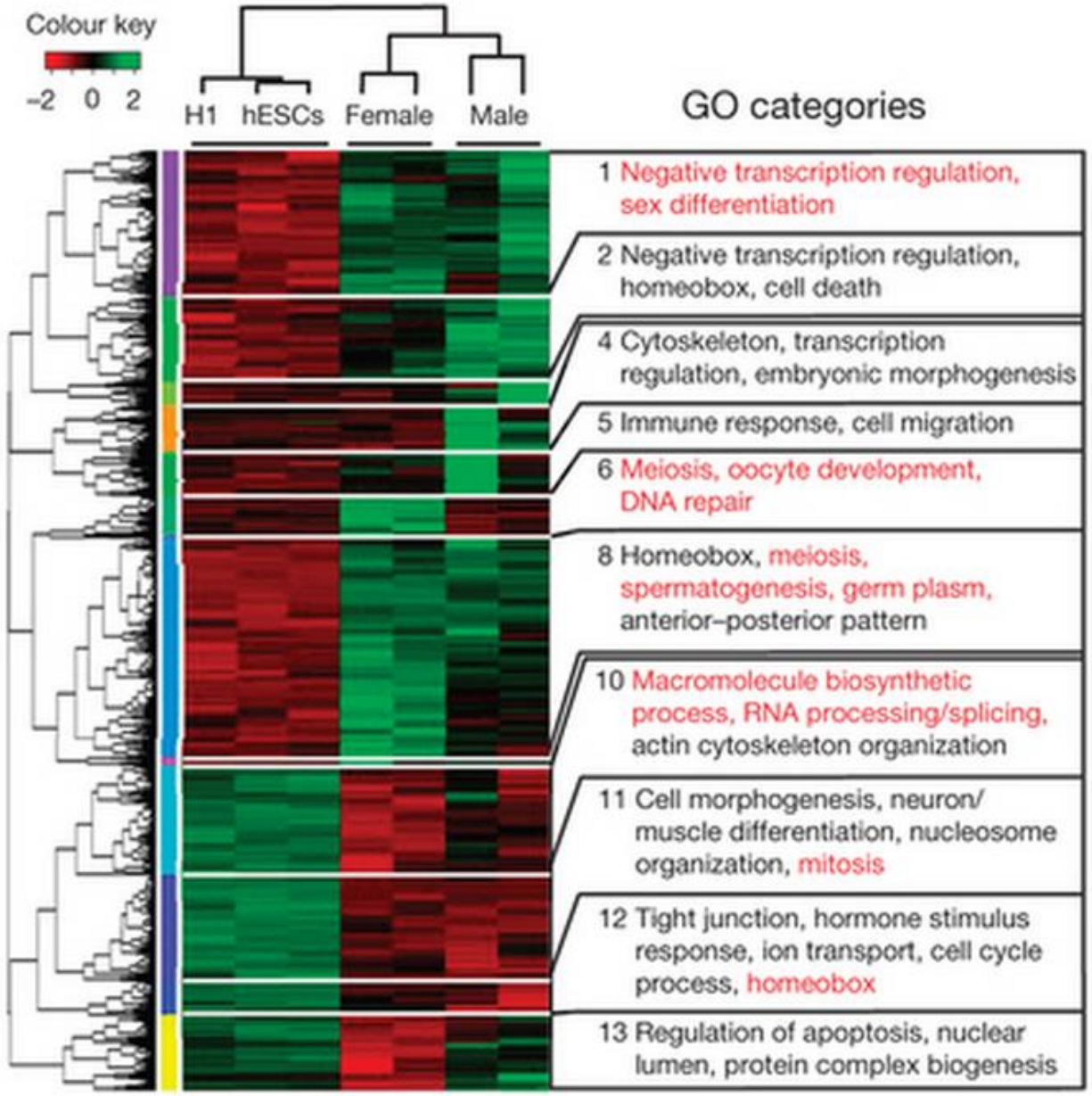
Reminders

- Pset 1 posted – due Feb 20th (no extra problem)
- Pset 2 posted – due Mar 13th
- Project teams due – Feb 25th
  - Interests and background directory has been posted
- Lecture videos will be posted on MITx soon – next week?

Today

- Clustering (6.874 topic)
- Biology review
- Alignment





Source: Gkountela, Sofia, Ziwei Li, et al. "The Ontogeny of cKIT+ Human Primordial Germ Cells Proves to be a Resource for Human Germ Line Reprogramming, Imprint Erasure and in Vitro Differentiation." *Nature Cell Biology* 15, no. 1 (2013): 113-22.

Clustering – K-means

- Group points together based on how ‘close’ they are to each other
- Dataset of unlabelled points: $X = \{x_1, x_2, \dots, x_N\}, x_n \in R^d$
- Assume K clusters – each is defined by a centroid μ_k
- $r_{nk} = 1$ if x_n belongs to cluster k
- Find unknowns μ_k and r_{nk}

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

Algorithm 10.1 *K-Means Clustering*

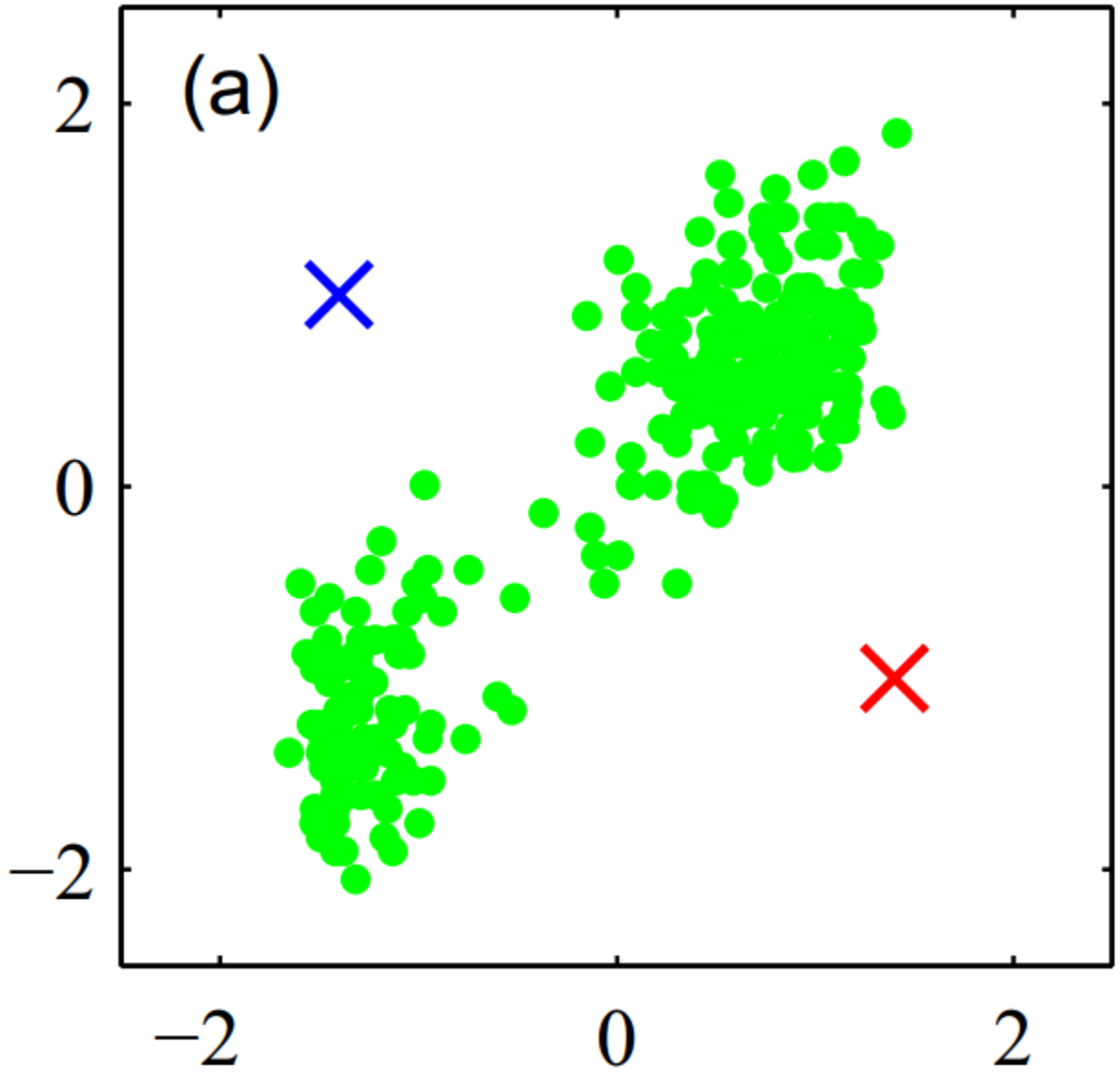
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

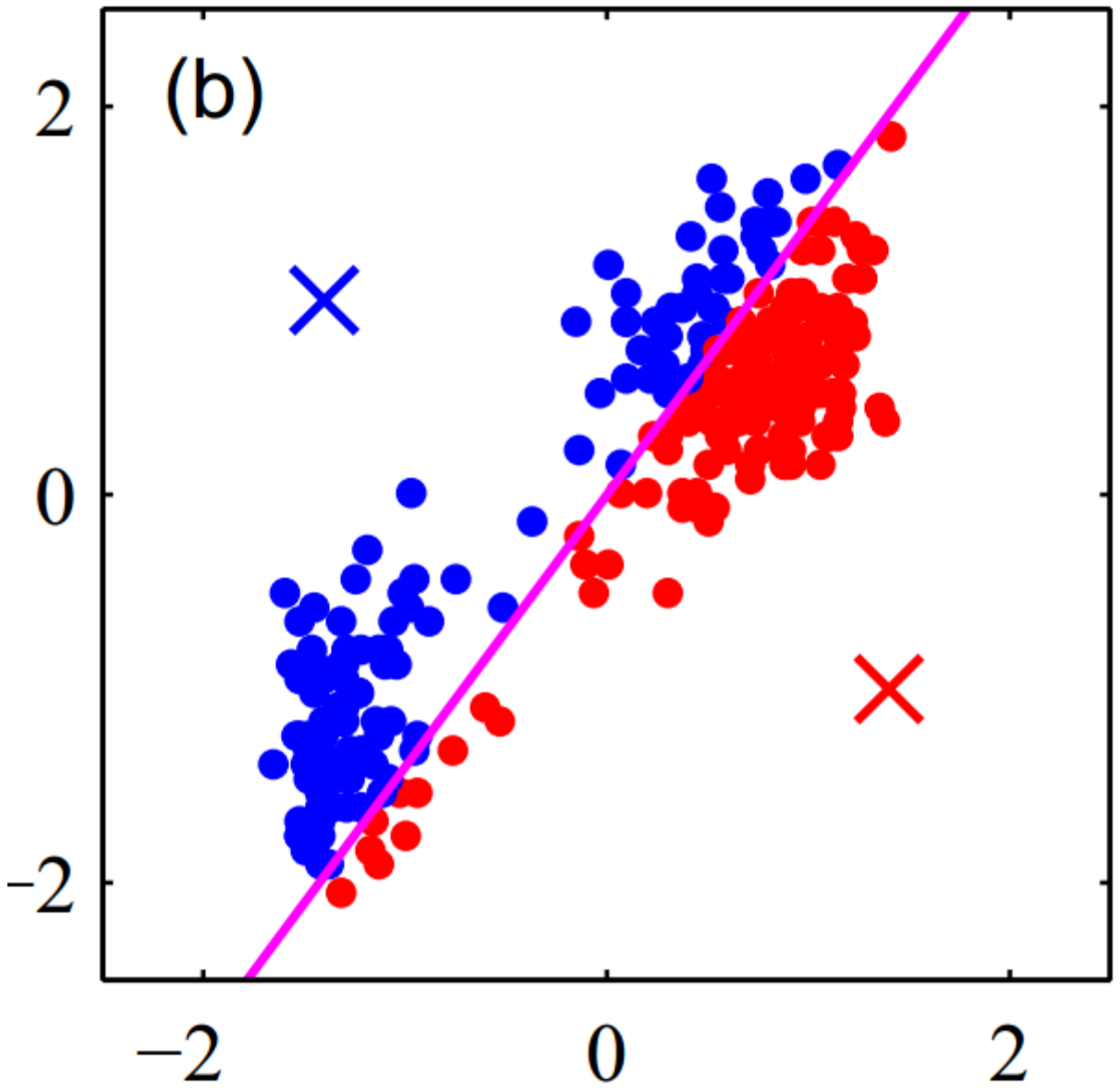
2. Iterate until the cluster assignments stop changing:

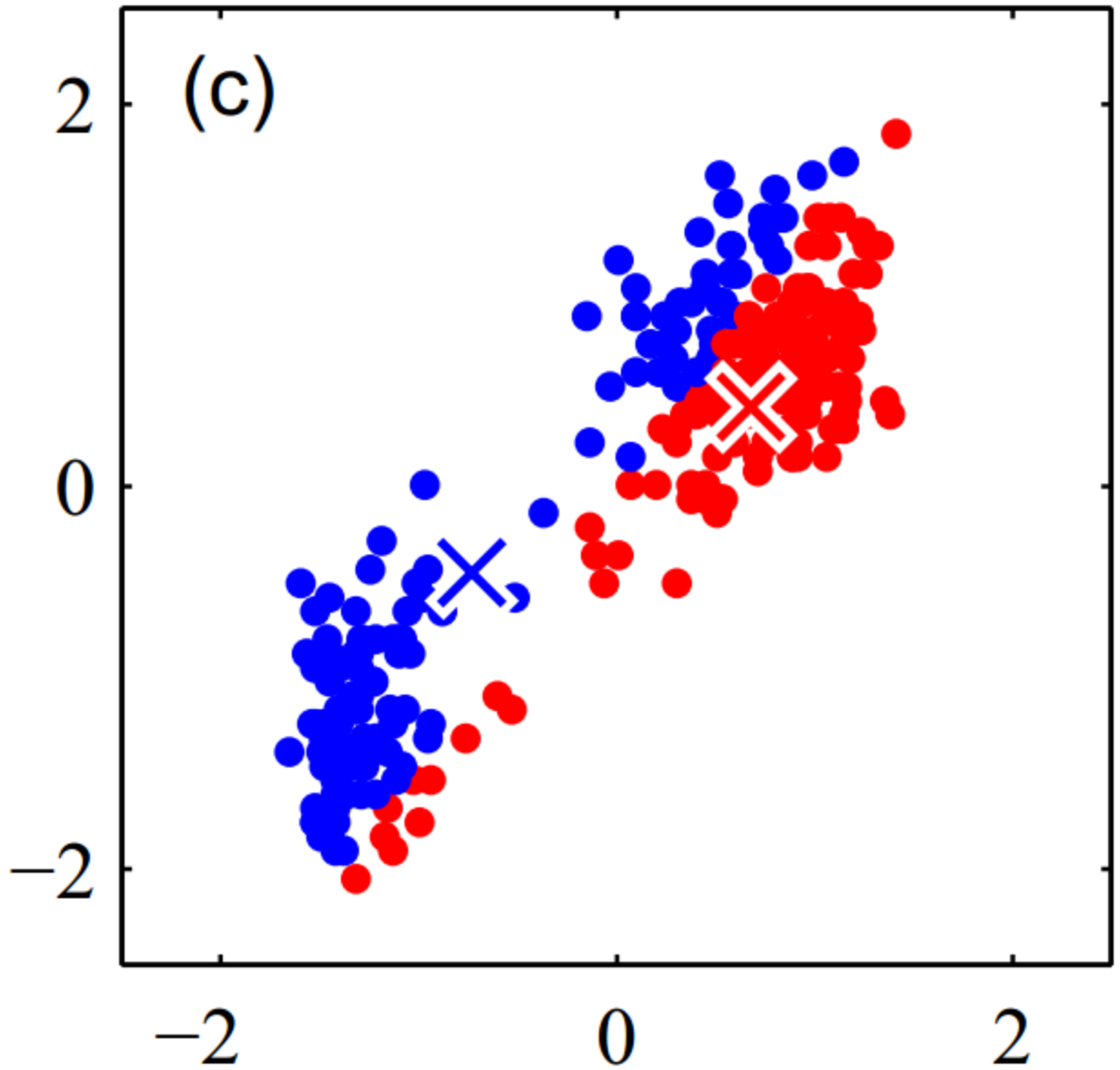
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

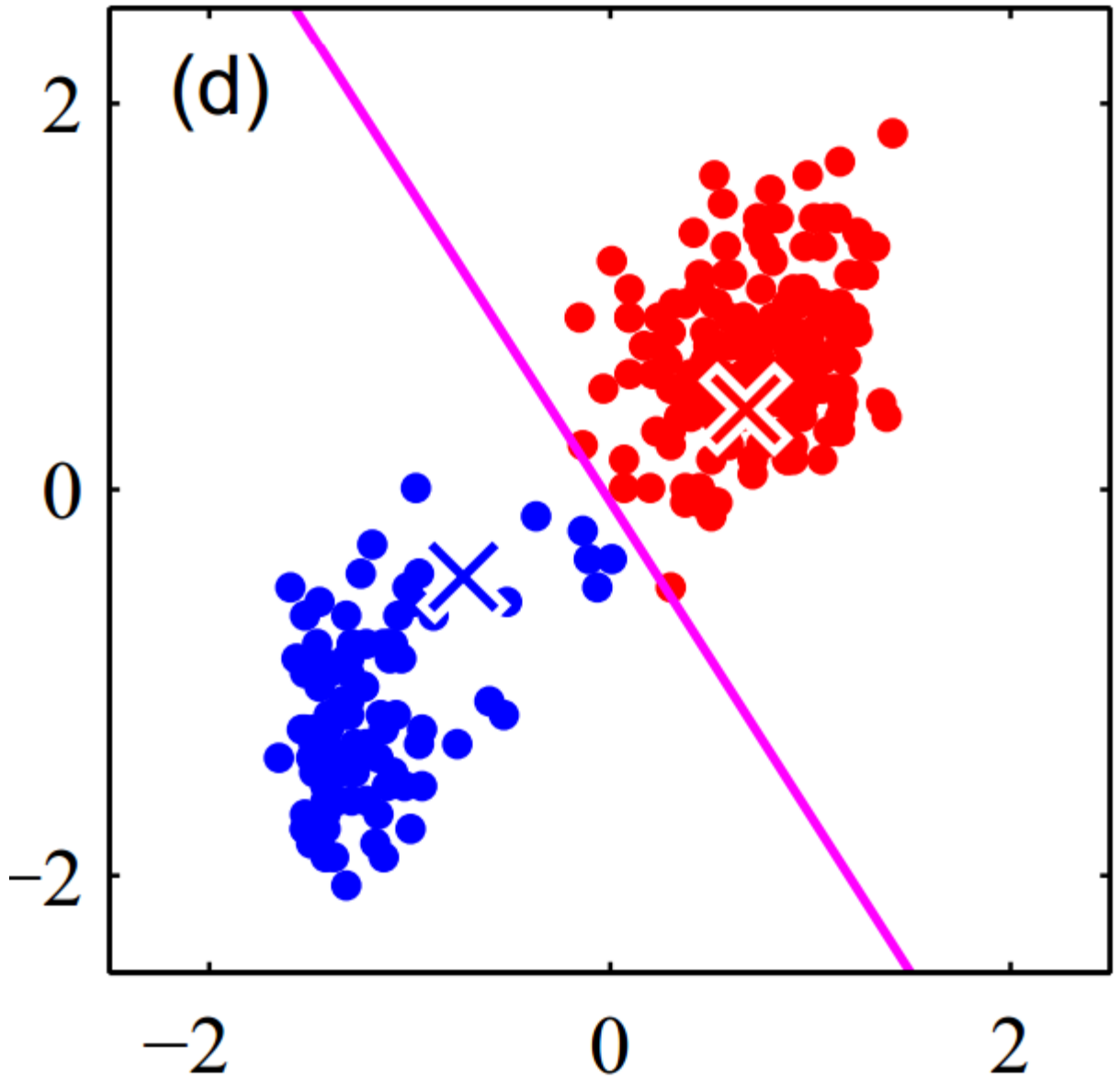
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

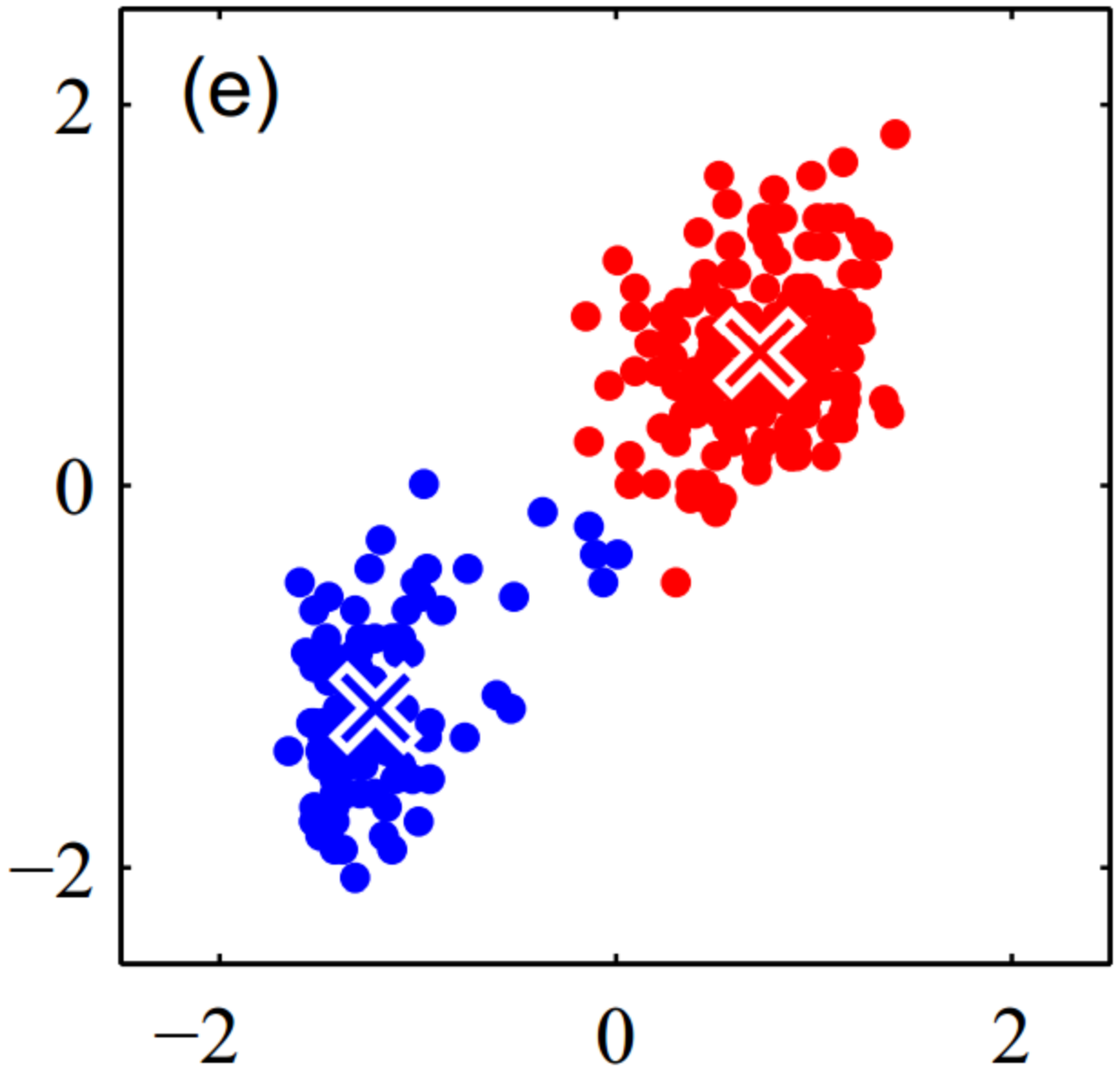
Source: Hastie, Trevor, Robert Tibshirani, et al. The Elements of Statistical Learning. *Springer-Verlag* 2, no. 1, 2009.

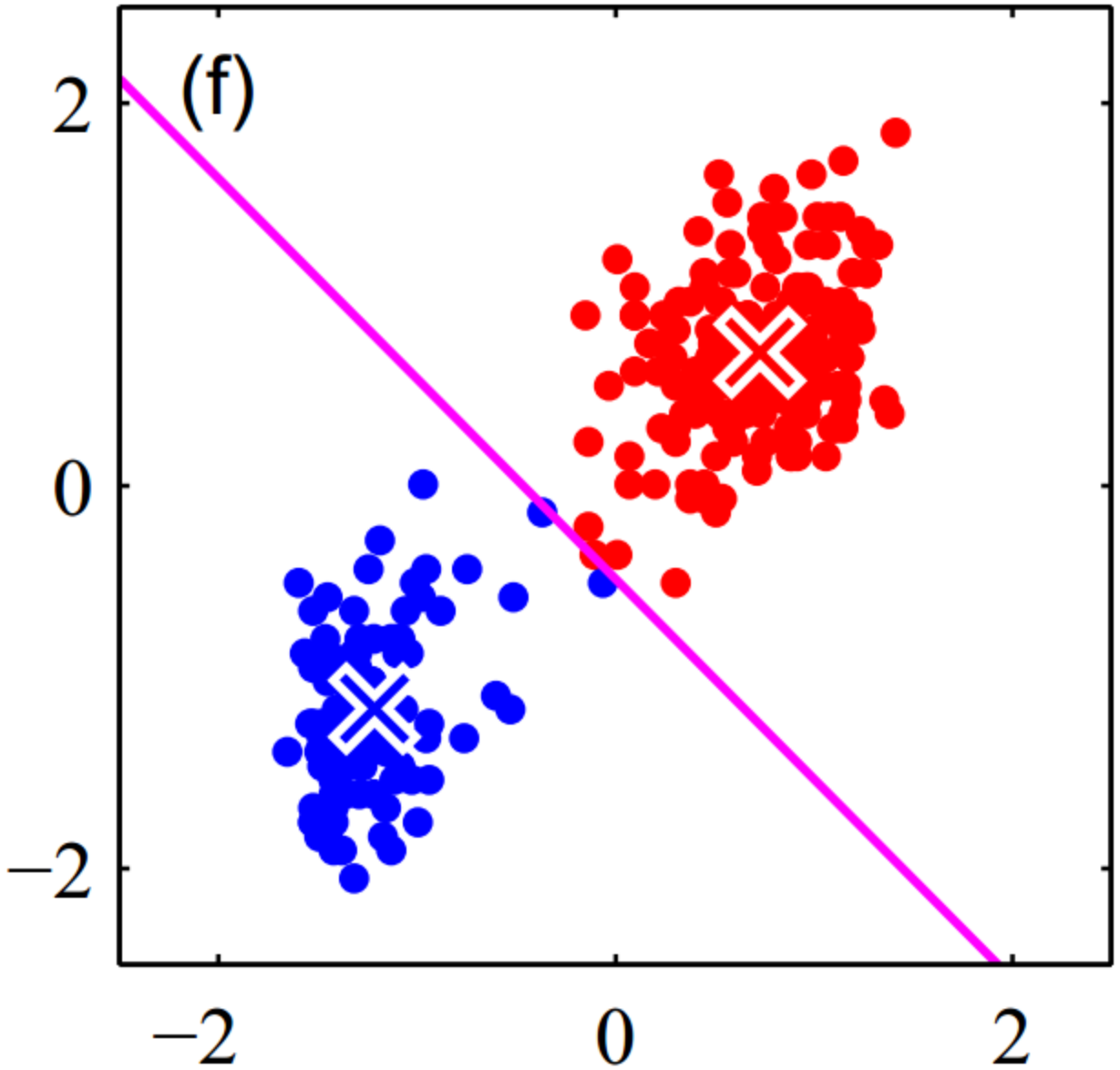


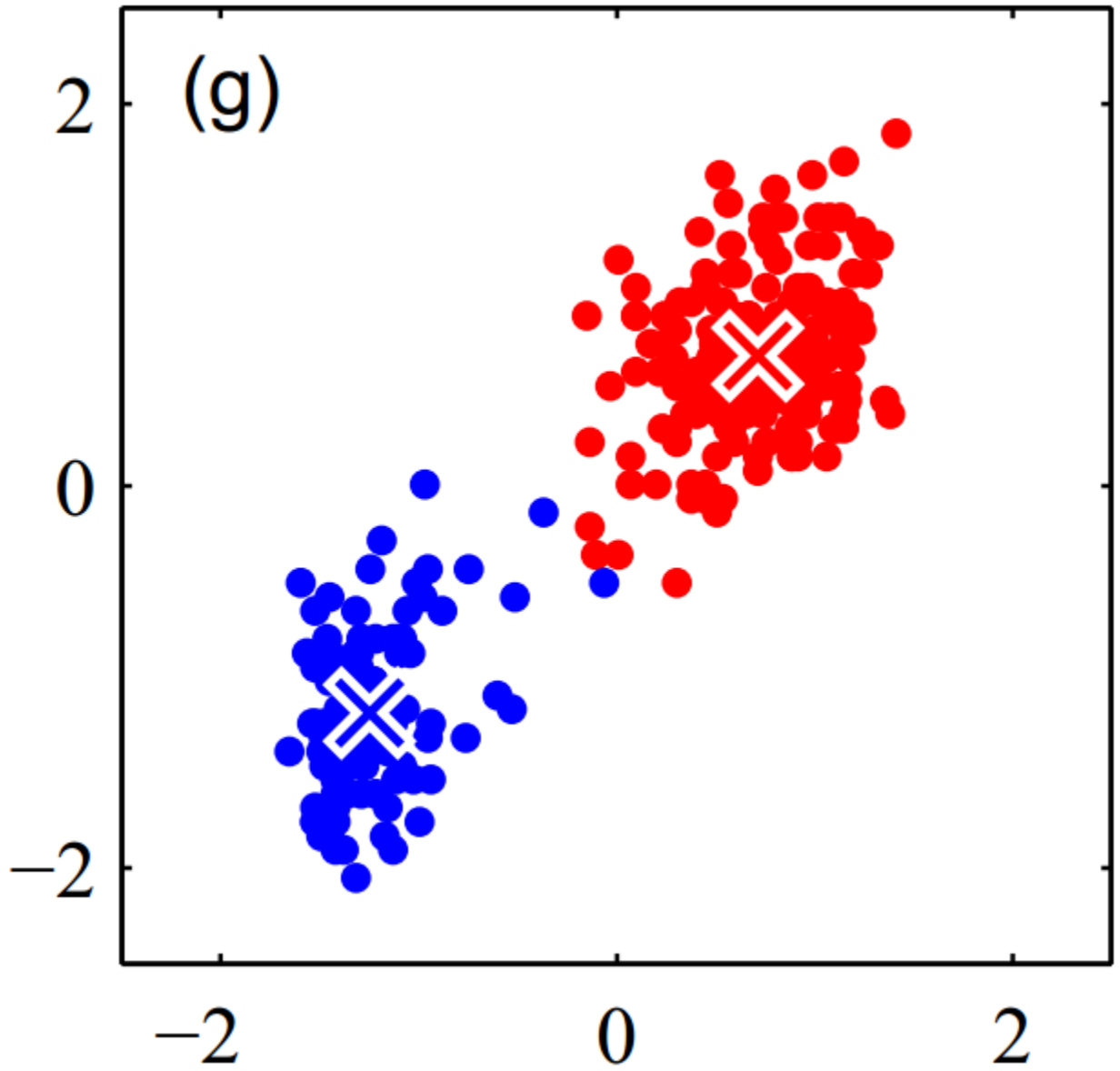


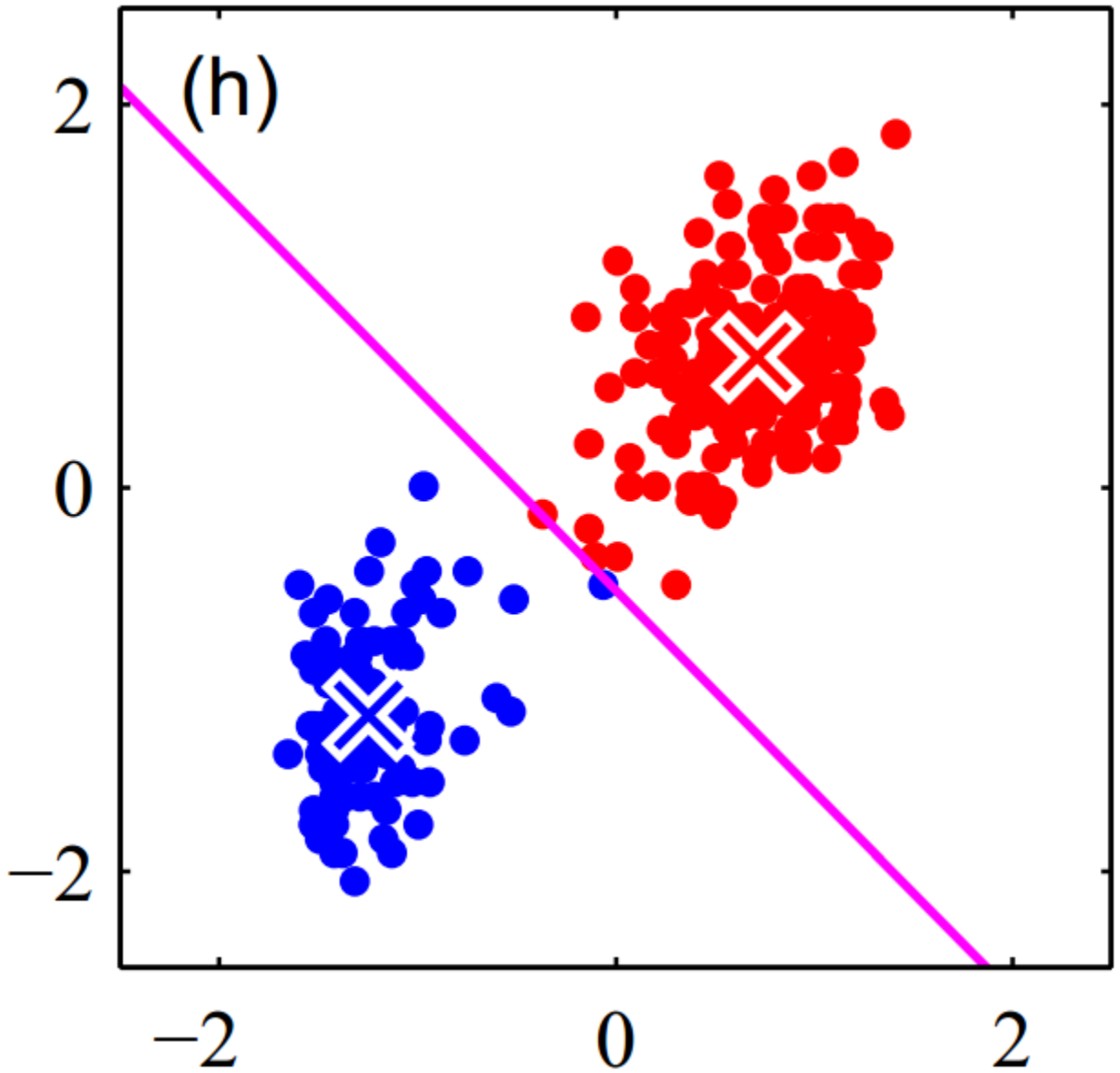


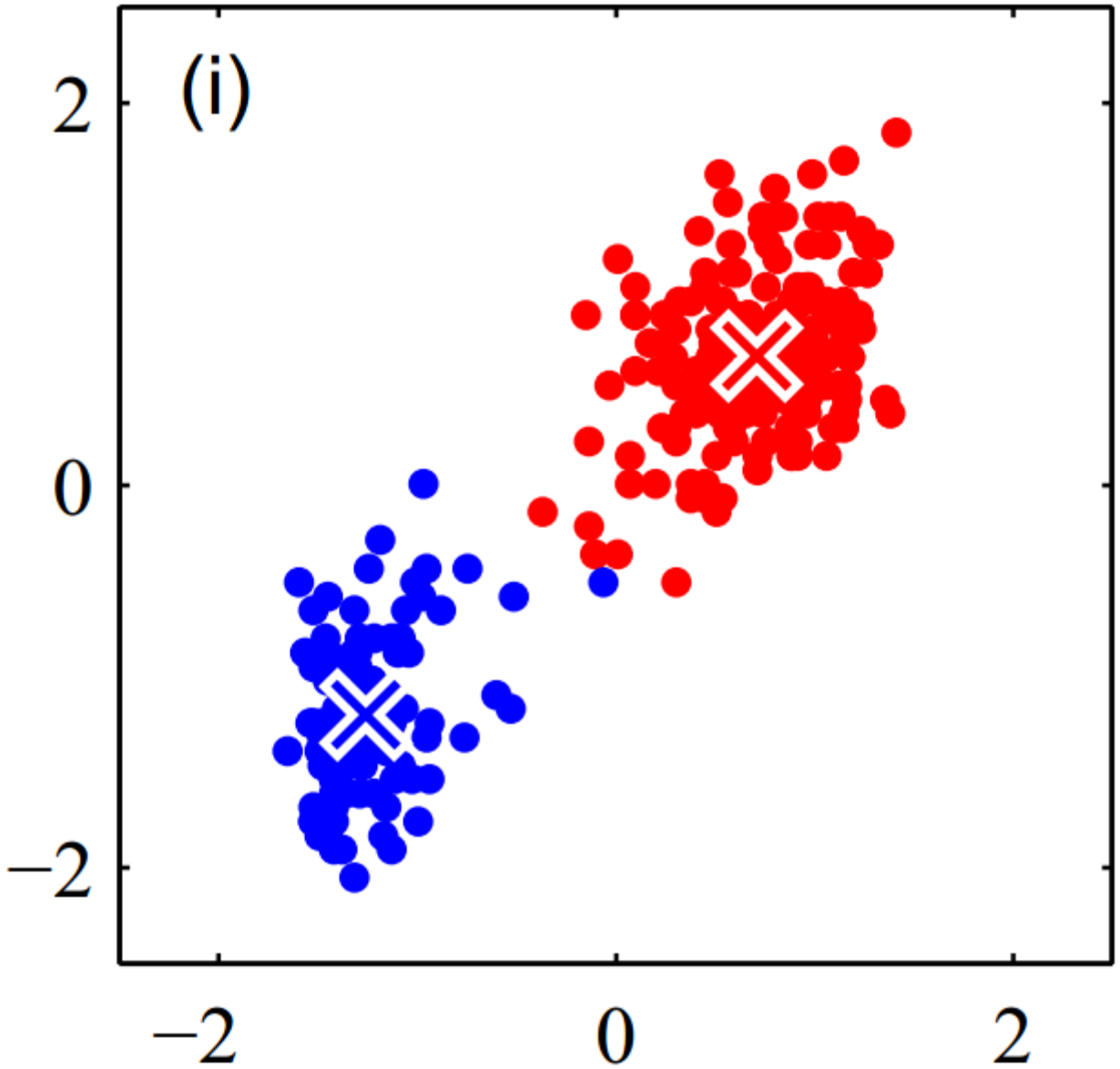












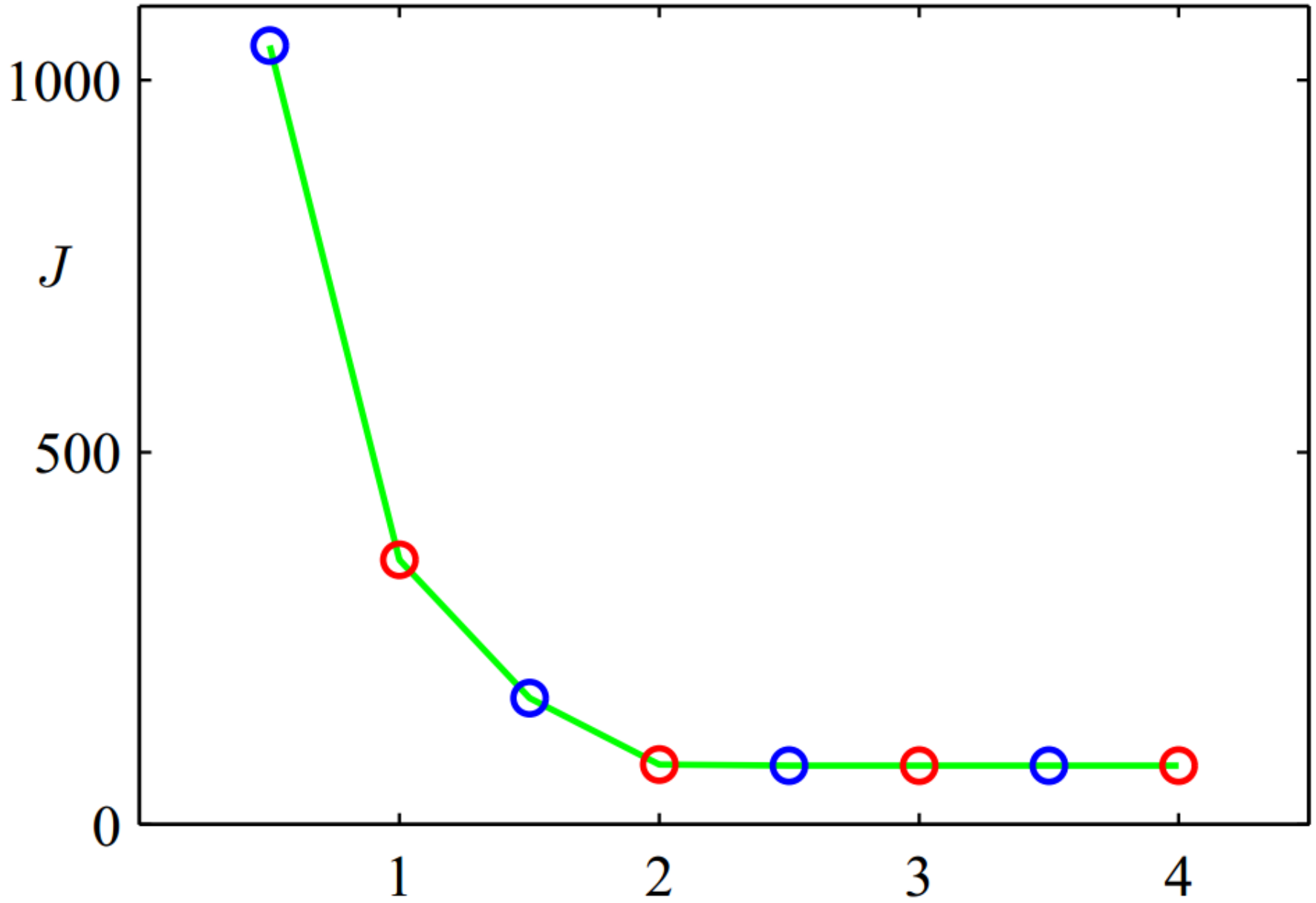


Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (e_f - e_g)' \Sigma^{-1} (e_f - e_g)$, where Σ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace e_{gc} with the rank of e_{gc} within the expression values of gene g across all conditions $c = 1 \ldots C$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

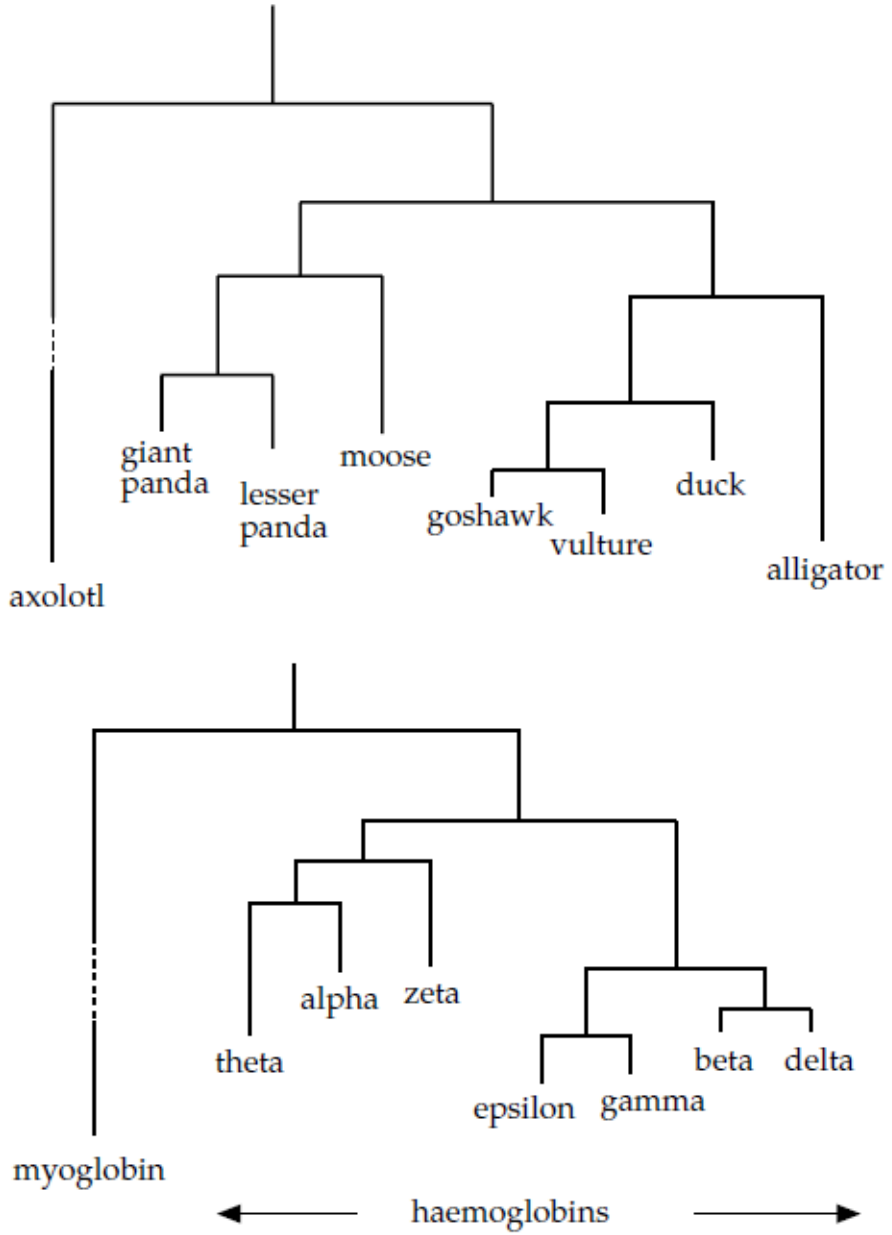
d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

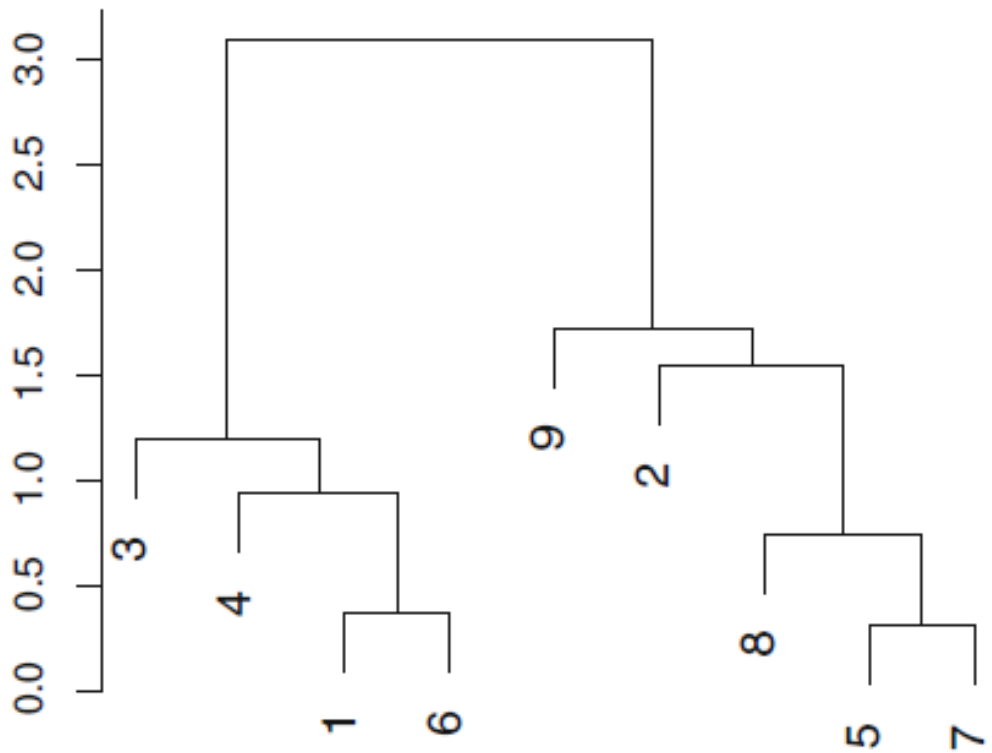
Source: D'haeseleer, Patrik. "How Does Gene Expression Clustering Work?." *Nature Biotechnology* 23, no. 12 (2005): 1499-1502.

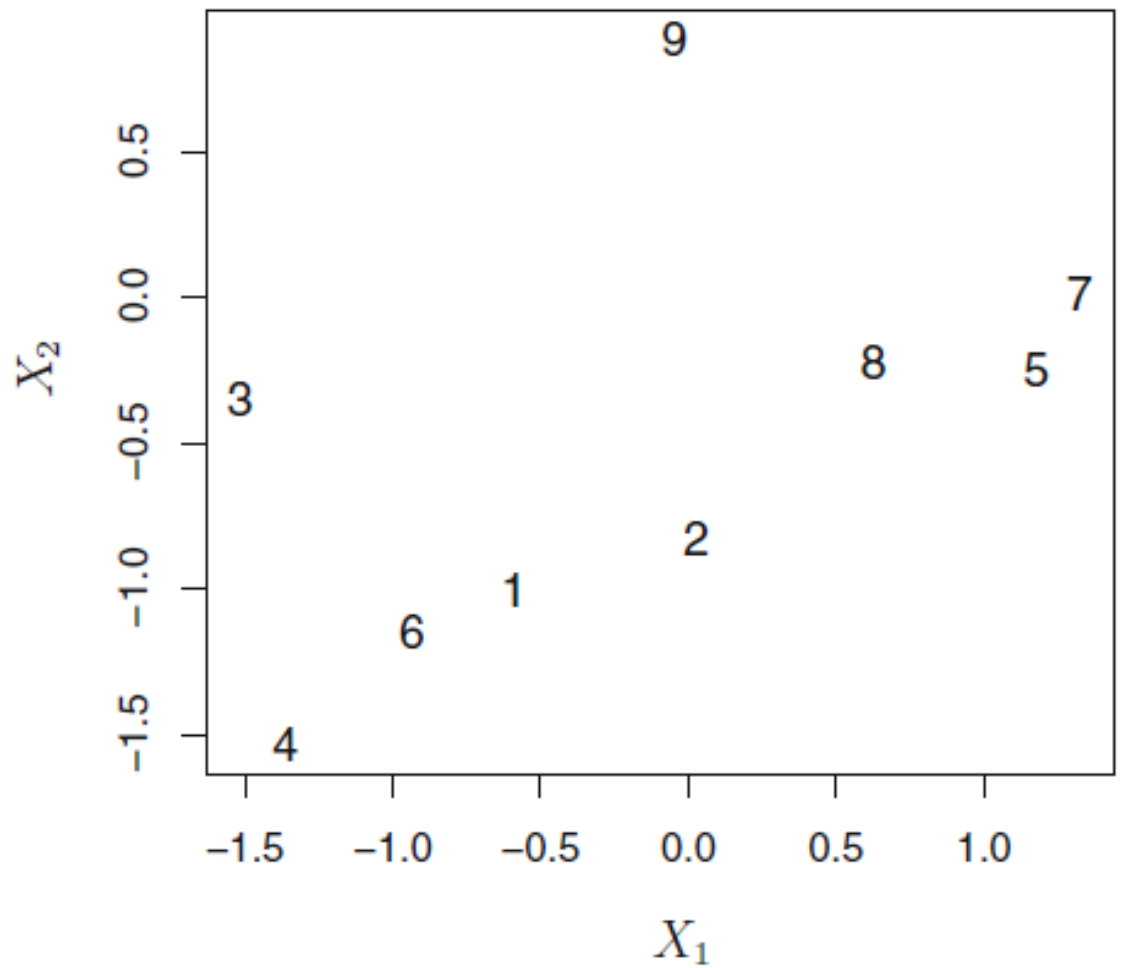
How does gene expression clustering work?
Patrik D'haeseleer
Nature Biotechnology **23**, 1499 - 1501 (2005)
doi:10.1038/nbt1205-1499

Hierarchical clustering

- Organize data in a tree
  - Leaves are individual genes/species
  - Path lengths between leaves are distances
  - Similar points should lie in same lower subtrees
- Used to reveal evolutionary history of sequences




0.0
0.5
1.0
1.5
2.0
2.5
3.0
3
4
1
6
9
2
8
5
7


9
7
8
5
3
2
1
6
4
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_2
X_1

Hierarchical clustering

- What dissimilarity measure should be used?
  - To compare individual points
- What type of linkage should be used?
  - To compare clusters with each other
- Where do we cut dendrogram to obtain clusters?

Algorithm 10.2 *Hierarchical Clustering*

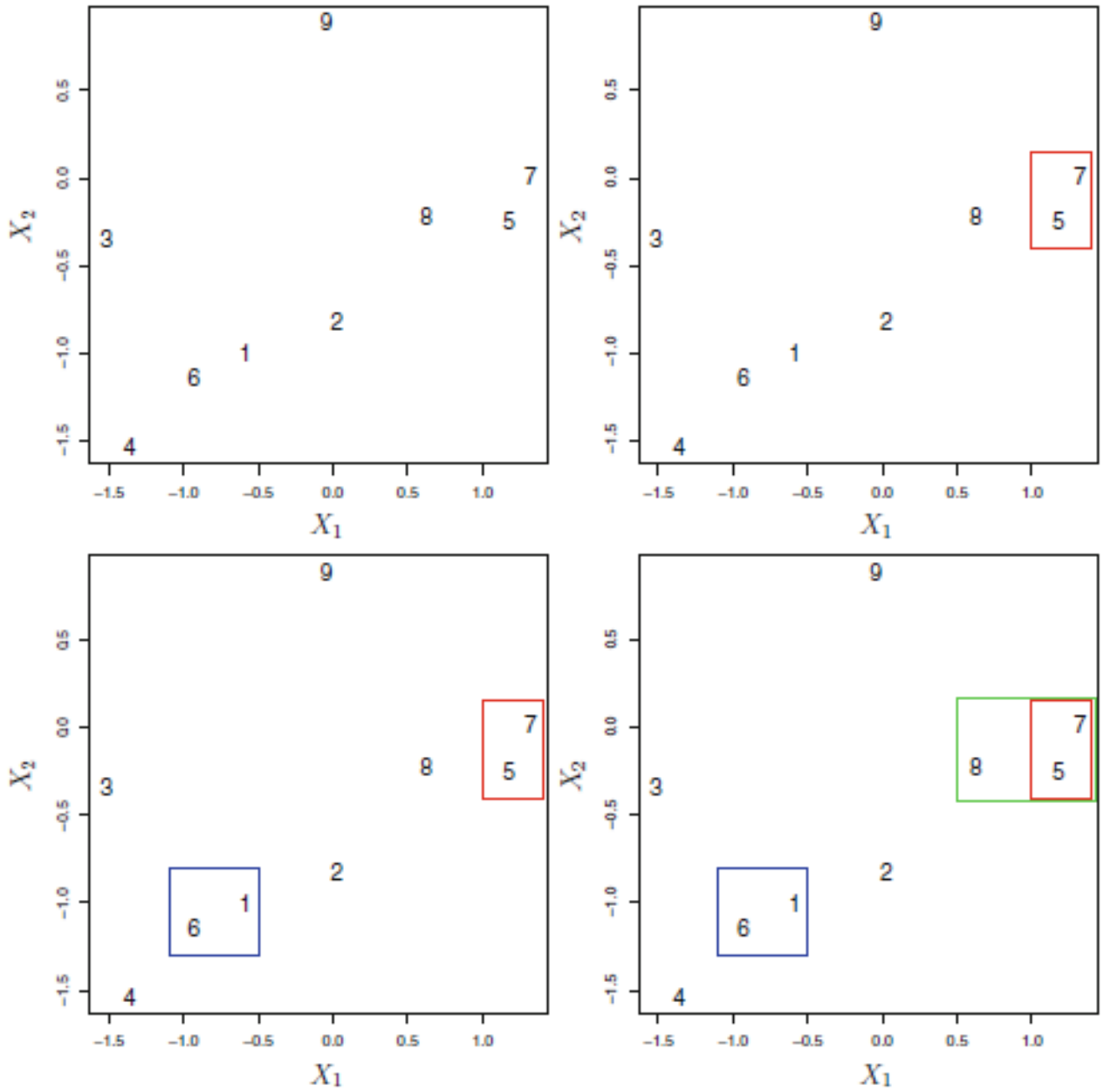
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

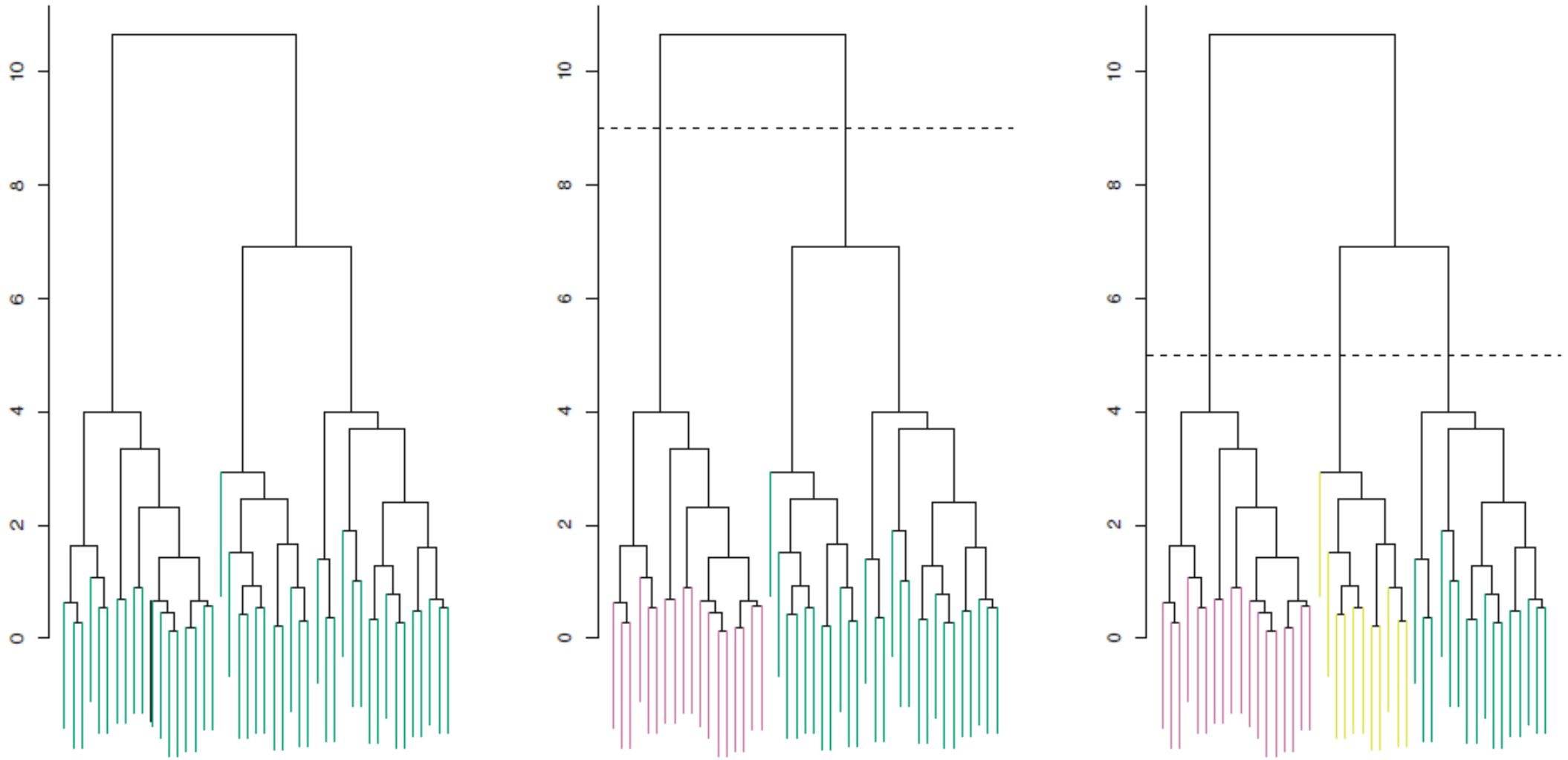
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Source: Hastie, Trevor, Robert Tibshirani, et al. The Elements of Statistical Learning. *Springer-Verlag* 2, no. 1, 2009.



Complete	$d(A,B) = \max_{x \in A, y \in B} d(x,y)$	Maximal intercluster dissimilarity
Single	$d(A,B) = \min_{x \in A, y \in B} d(x,y)$	Minimal intercluster dissimilarity
Average	$d(A,B) = \sum_{x \in A, y \in B} \frac{d(x,y)}{\lvert A \rvert \lvert B \rvert}$	Average intercluster dissimilarity (UPGMA)
Centroid	$d(A,B) = d\left(\frac{\sum_{x \in A} \frac{x}{\lvert A \rvert}, \sum_{y \in B} \frac{y}{\lvert B \rvert}}{}\right)$	Dissimilarity between centroids of each cluster



Model selection

- Hierarchical clustering – cutoff tree at certain point
- K-means – how to choose K?
  - Balance number of clusters (# of parameters – K=n is uninformative) and the variance of the clusters

  - BIC Score – general model selection criterion
  - $\mathrm{BIC} = -2 \times \mathrm{loglikelihood} + \mathrm{d} \times \log(\mathrm{N})$
  - Can use to decide whether to split a cluster
  - Computer BIC score of cluster and two potential child clusters – if BIC score is lower after split, do not accept split

Biclustering

- Simultaneous clustering of rows and columns of a matrix
- Bicluster – subset of rows which exhibit similar behavior across a subset of columns, or vice versa

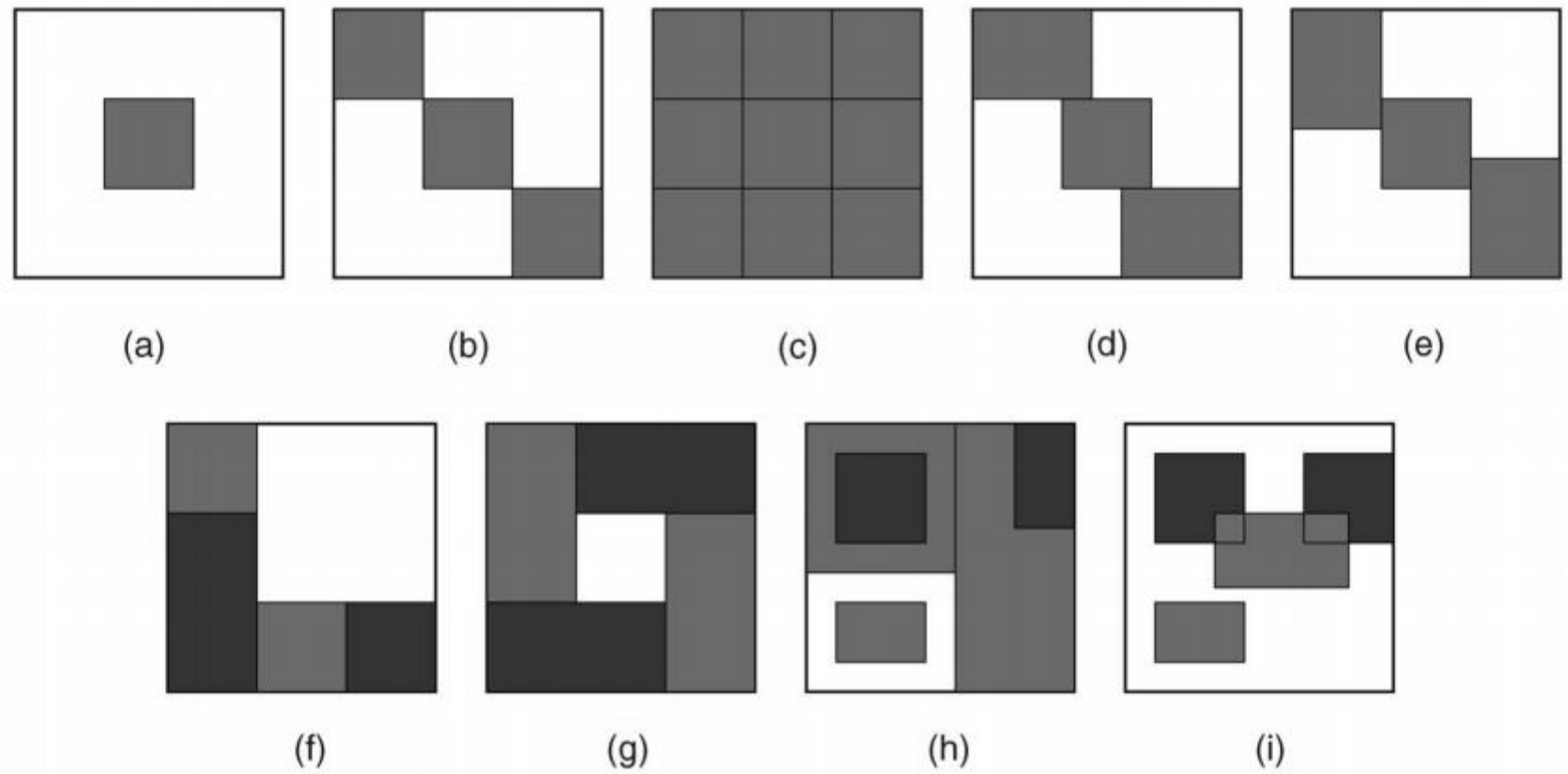
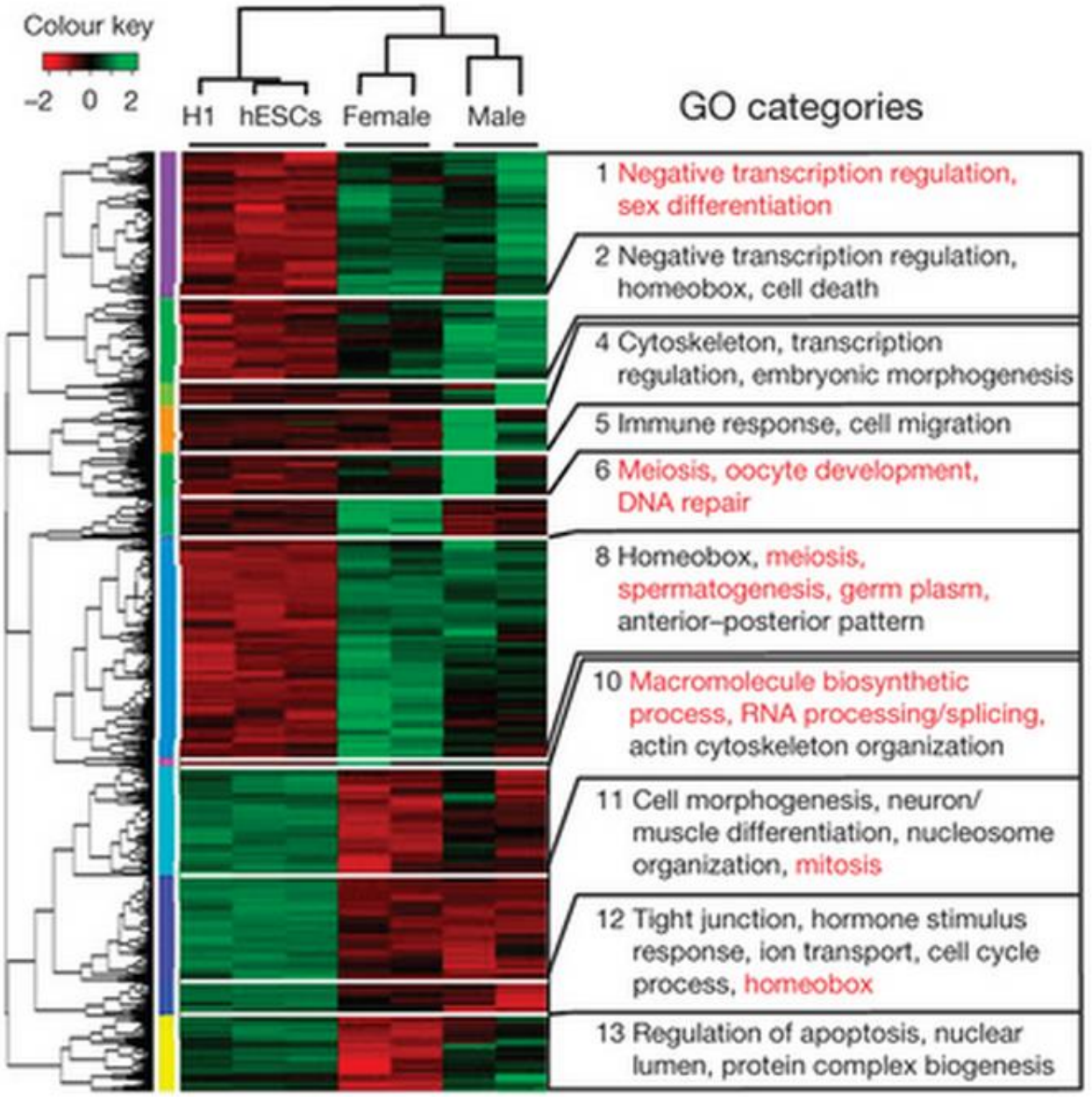


Fig. 4. Bicluster structure. (a) Single bicluster, (b) exclusive row and column biclusters, (c) checkerboard structure, (d) exclusive rows biclusters, (e) exclusive columns biclusters, (f) nonoverlapping biclusters with tree structure, (g) nonoverlapping nonexclusive biclusters, (h) overlapping biclusters with hierarchical structure, and (i) arbitrarily positioned overlapping biclusters.

Source: Madeira, Sara C., and Arlindo L. Oliveira. "Biclustering Algorithms for Biological Data Analysis: A Survey." *Computational Biology and Bioinformatics*, IEEE/ACM Transactions on 1, no. 1 (2004): 24-45.



Source: Gkountela, Sofia, Ziwei Li, et al. "The Ontogeny of cKIT+ Human Primordial Germ Cells Proves to be a Resource for Human Germ Line Reprogramming, Imprint Erasure and in Vitro Differentiation." *Nature Cell Biology* 15, no. 1 (2013): 113-22.

Biology Review

Selection

- Negative selection (purifying/natural selection) – removal of deleterious traits
- Positive selection – increases prevalence of adaptive traits

- Thinking about selection happening at different levels
    - *Protein level: Sequence -> Structure -> Function*
    - RNA level: splicing, degradation/processing (NMD)
    - DNA level: DNA-protein binding sites

Synonymous/Non-synonymous mutations

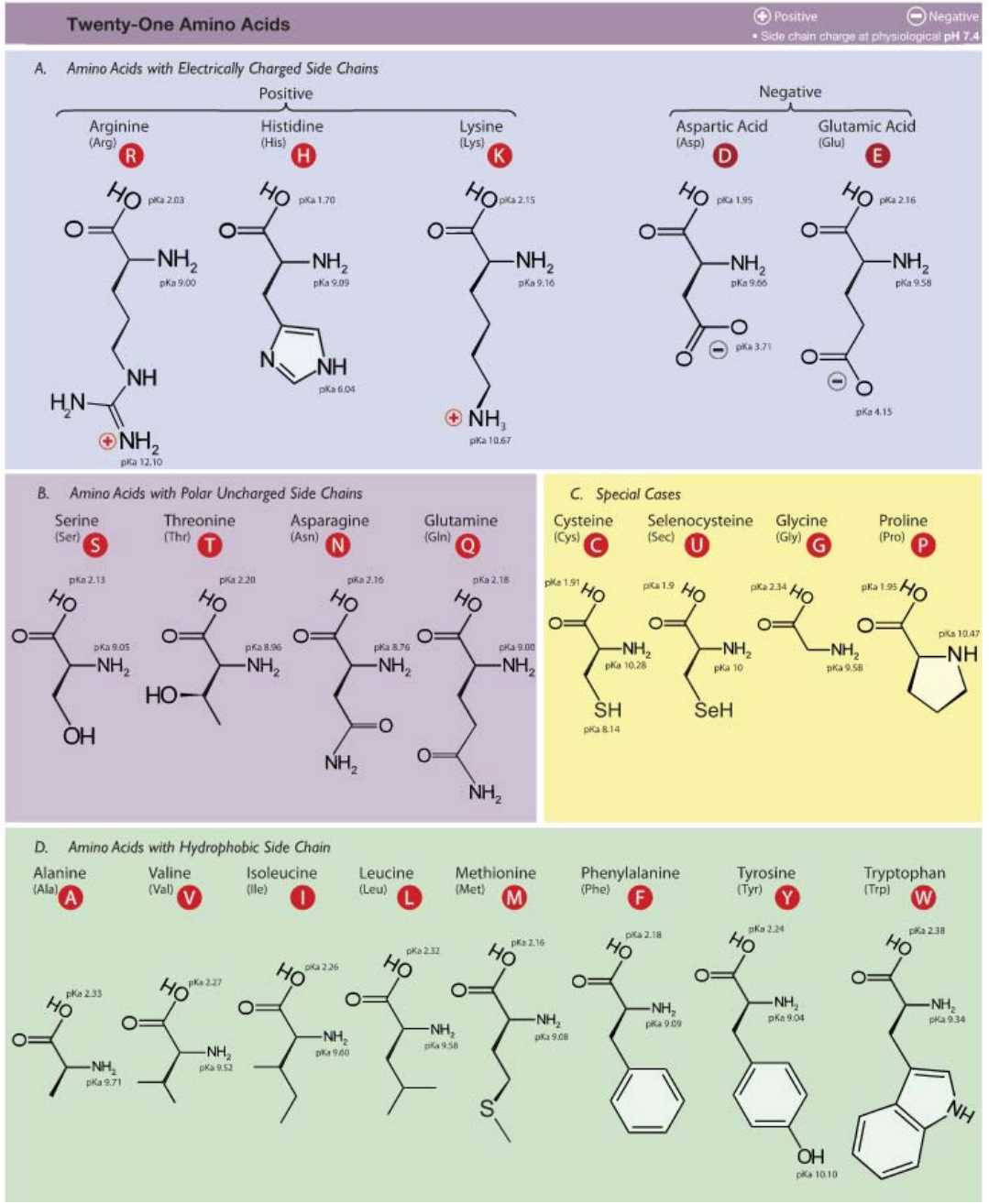
- Redundancy built into the genetic code
- Synonymous – one base changes for another in an exon, but the resulting amino acid sequence is unchanged
- Non-synonymous – new AA
- Can affect splicing, mRNA processing - so may not be silent

First Position (5' end)	Second Position												Third Position (3' end)
	U			C			A			G			
	U	UUU	Phe	UCU	Ser		UAU	Tyr		UGU	Cys		U
		UUC		UCC			UAC			UGC			C
		UUA	Leu	UCA			UAA	*Stop*		UGA	*Stop*		A
		UUG		UCG			UAG	*Stop*		UGG	Trp		G
	C	CUU		CCU	Pro		CAU	His		CGU	Arg		U
		CUC		CCC			CAC			CGC			C
		CUA	Leu	CCA			CAA	Gln		CGA			A
		CUG		CCG			CAG			CGG			G
	A	AUU		ACU	Thr		AAU	Asn		AGU	Ser		U
		AUC	Ile	ACC			AAC			AGC			C
		AUA		ACA			AAA	Lys		AGA	Arg		A
		AUG	Met	ACG			AAG			AGG			G
	G	GUU		GCU	Ala		GAU	Asp		GGU	Gly		U
		GUC	Val	GCC			GAC			GGC			C
		GUA		GCA			GAA	Glu		GGA			A
		GUG		GCG			GAG			GGG			G

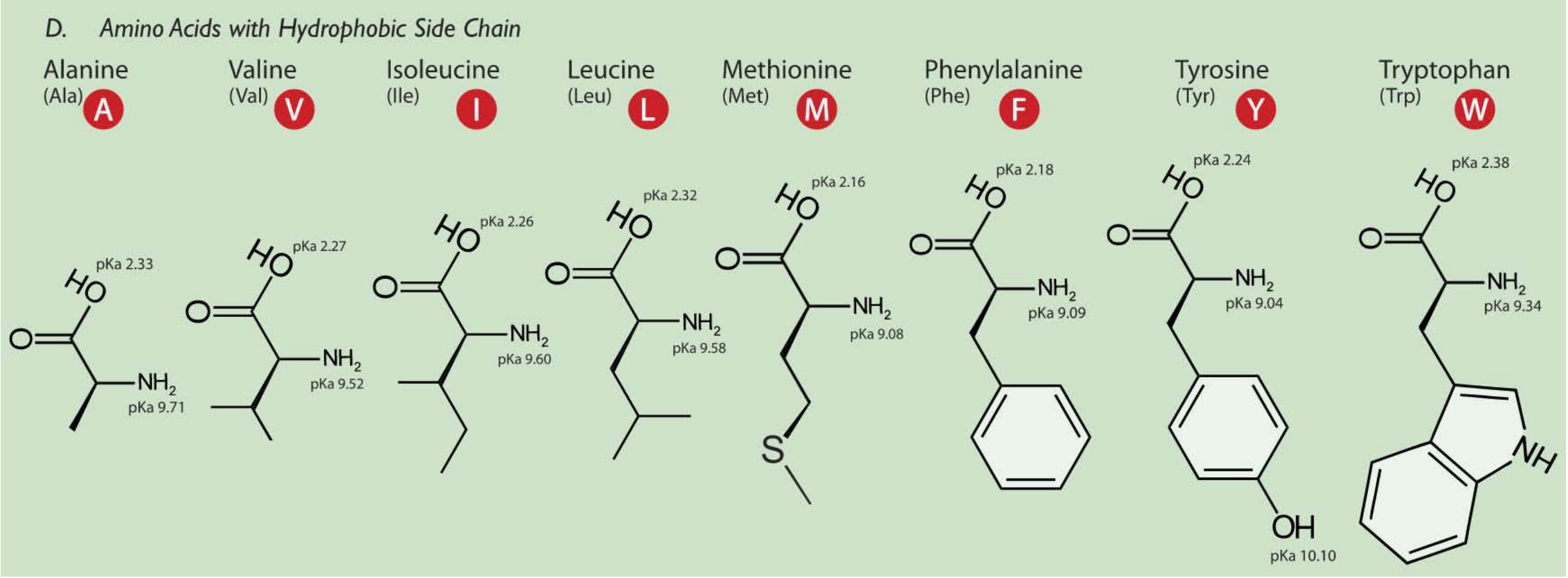
Image by MIT OpenCourseWare.

Side-chain biochemistry

- Amino acids classified by properties of side chains
  - Grouped by general properties
- Substitutions of amino acid with another of similar chemical properties may conserve protein function

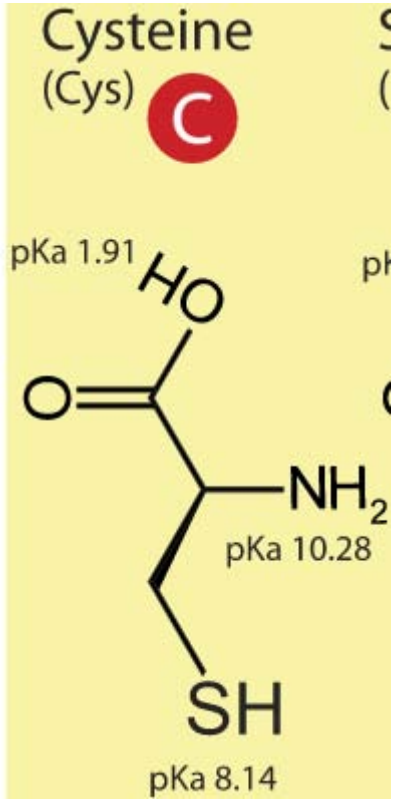


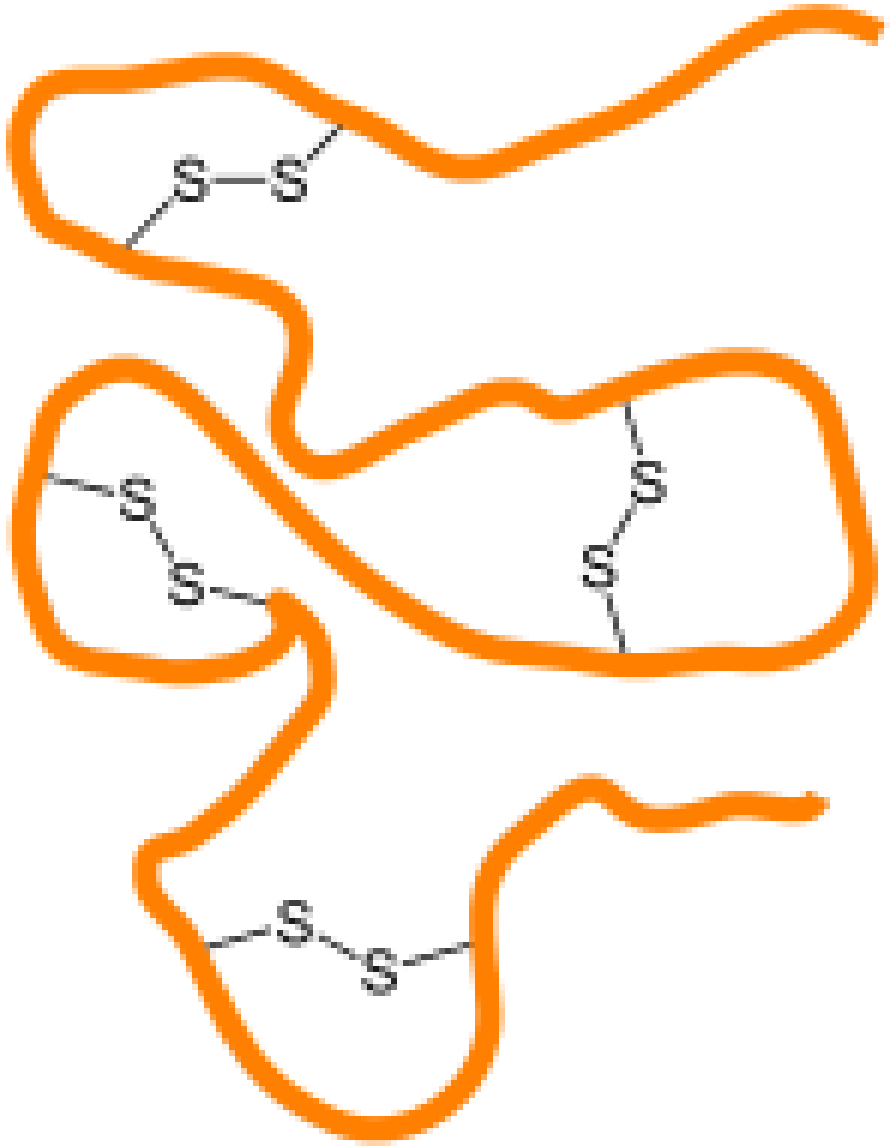
Side chain size (Trp – W)



Disulfide bond

- Important in protein folding – holds two distant portions of protein together
- Occurs between Cys residues





Next-generation sequencing

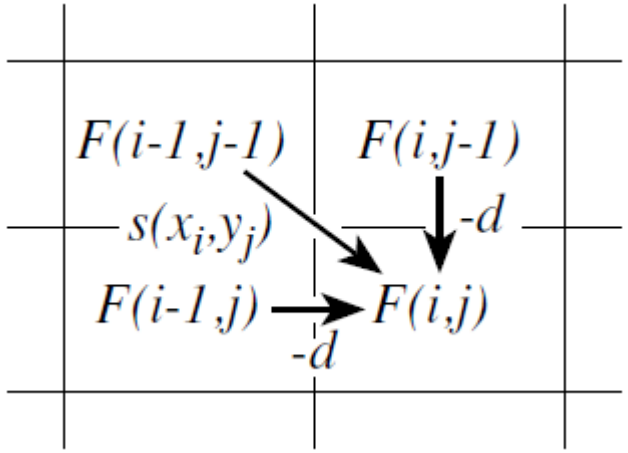
- Sequencing is always of DNA
    - Need to convert RNA to DNA by *reverse transcription* (RT)
- Illumina is current leader in the field
    - 8 lanes on a flow cell
    - Each lane can sequence 200 million 100bp reads – 20 Gbps!
    - Can sequence multiple samples per lane by barcoding
    - Requires (heterogeneous) population of cells to get enough DNA for sample
- Single cell sequencing applications are becoming more common (RNAseq)
- Single molecule technologies are still being developed – PacBio

Alignment

Alignment

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

$$F(i,j) = \max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$



Biological Sequence Analysis - Durbin

Local alignment example

Do a local alignment between these using PAM250 and gap penalty -2:

AWEK

FWEF

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Local alignment solution

	Gap	A	W	E	K
Gap	0	0	0	0	-8
F	0	0	0	0	0
W	0	0	17	15	13
E	0	0	15	21	19
F	0	0	13	19	17

alignment:

W E

W E

Global alignment solution

	Gap	A	W	E	K
Gap	0	-2	-4	-6	-8
F	-2	-4	-2	-4	-6
W	-4	-6	13	11	9
E	-6	-4	-6	17	11
F	-8	-6	-4	15	12

alignment:

A	W	E	K
F	W	E	F

	Global	Semiglobal	Local (gapped)
Penalties at edges?	Yes	No	No
Reset to 0 instead of including negative entries?	No	No	Yes
End of alignment	Bottom right entry	Highest score entry in bottom row or rightmost column	Highest score entry in matrix

Reminders

- Pset 1 posted – due Feb 20th (no extra problem)
- Pset 2 posted – Due Mar 13th
- Project teams due – Feb 25th
  - Interests and background directory has been posted
- Lecture videos will be posted on MITx soon – next week?

7.91J / 20.490J / 20.390J / 7.36J / 6.802 / 6.874 / HST.506 Foundations of Computational and Systems Biology
Spring 2014