Learning from oncogenomic data

Gunes Gundem

Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

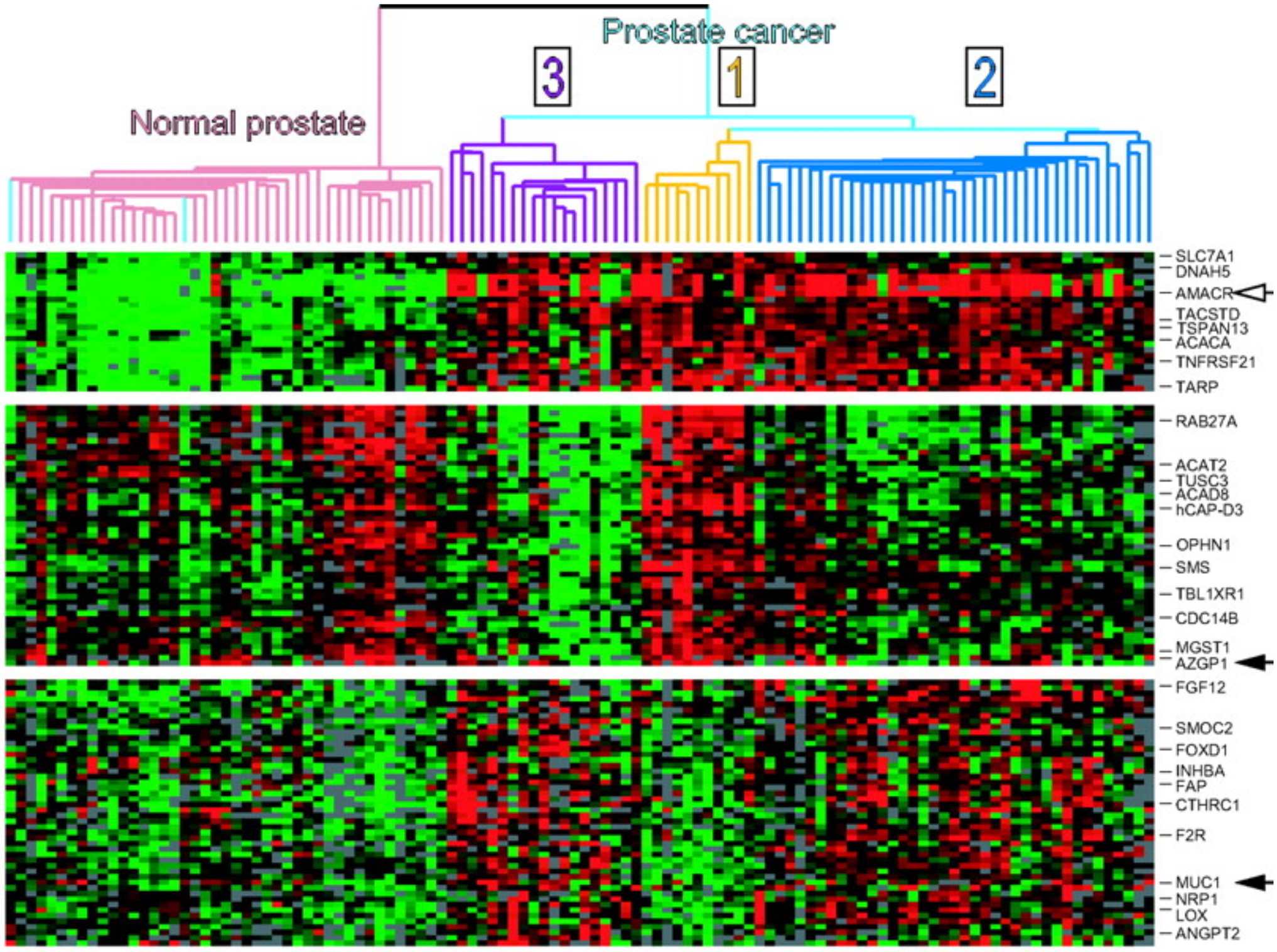
different approaches

- unsupervised approach
  - unbiased, no assumptions
  - hierarchical clustering
  - classification of the data into biologically meaningful subtypes

- supervised approach
  - partition data according to clinical end point
    - drug resistance, prognosis
  - detect the genes that explain the best the difference between groups

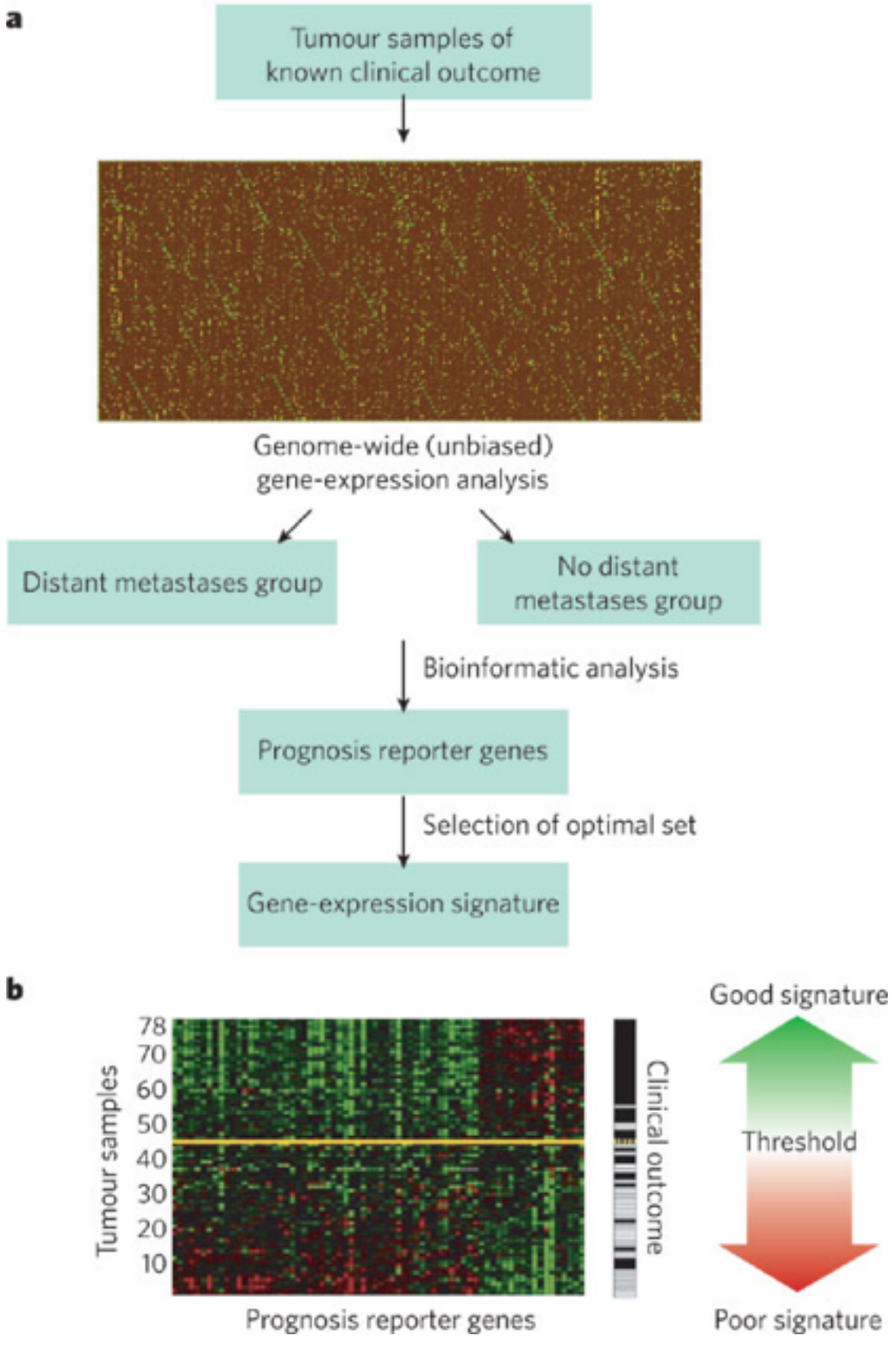
Learning from oncogenomic data

unsupervised approach



Learning from oncogenomic data

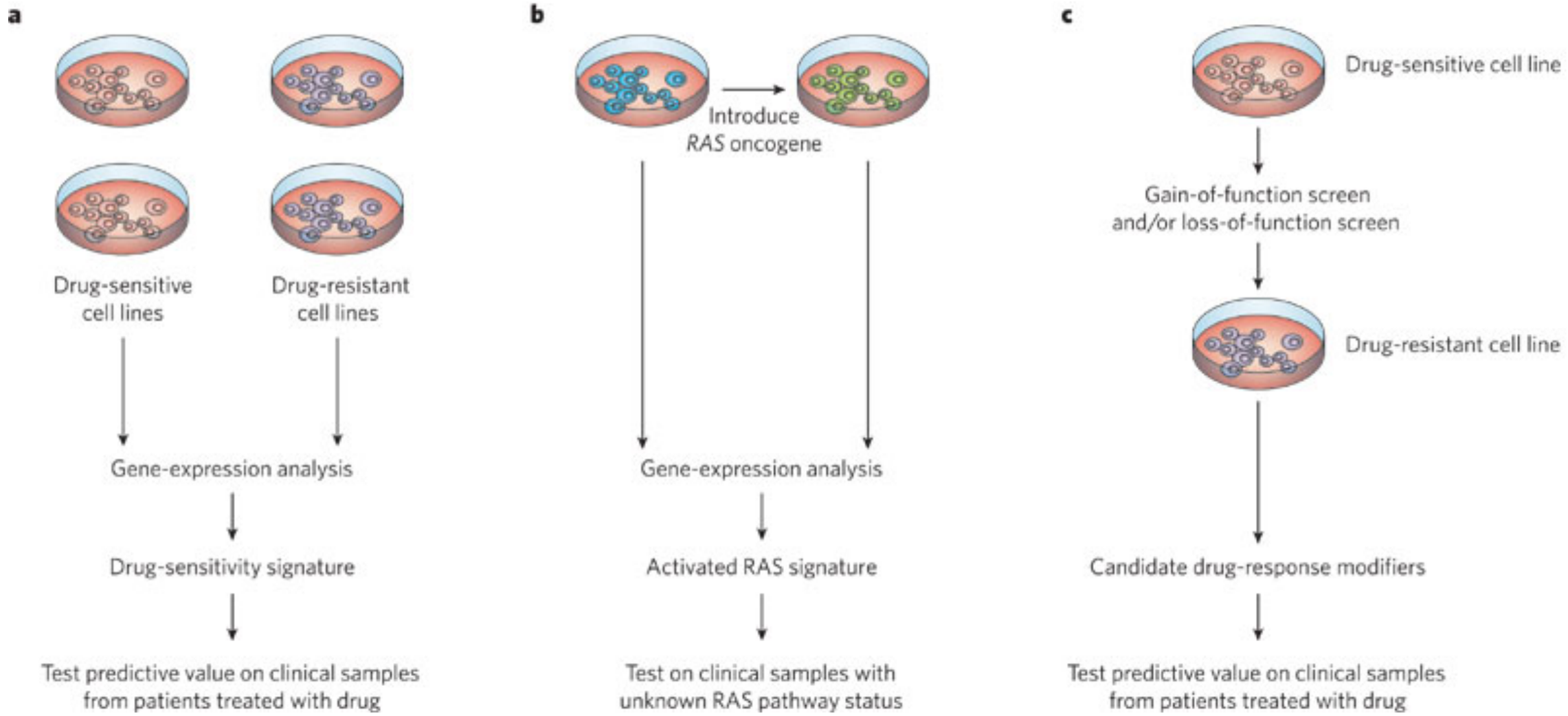
supervised approach



v'ant Veer LJ et al., Nature 2008

Learning from oncogenomic data

supervised approach



- Bild AH et al., Nature, 2006
- Huang E et al., Nature Gen, 2003
- Sorlie T et al., PNAS, 2001

- Berns K et al., Cancer Cell, 2007
- Huang E et al., Nature Gen, 2003
- Sorlie T et al., PNAS, 2001

v'ant Veer LJ et al., Nature 2008

Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

unsupervised approach: clustering

- Cluster: a collection of data objects
- Cluster analysis: grouping a set of data objects into clusters such that the data objects are
  - similar to one another within the same cluster
  - dissimilar to the objects in other clusters
- The quality of a clustering result depends on
  - similarity measure (distance metric)
  - clustering method
- unsupervised (no training set, no predefined classes)

Learning from oncogenomic data

clustering: requirements

- scalability
- ability to deal with different types of attributes
- discovery of classed with arbitrary shape
- minimal requirements for domain knowledge to determine input parameters
- able to deal with noise and outliers
- high-dimensionality
- incorporation of user-specified constraints
- interpretability and usability

Learning from oncogenomic data

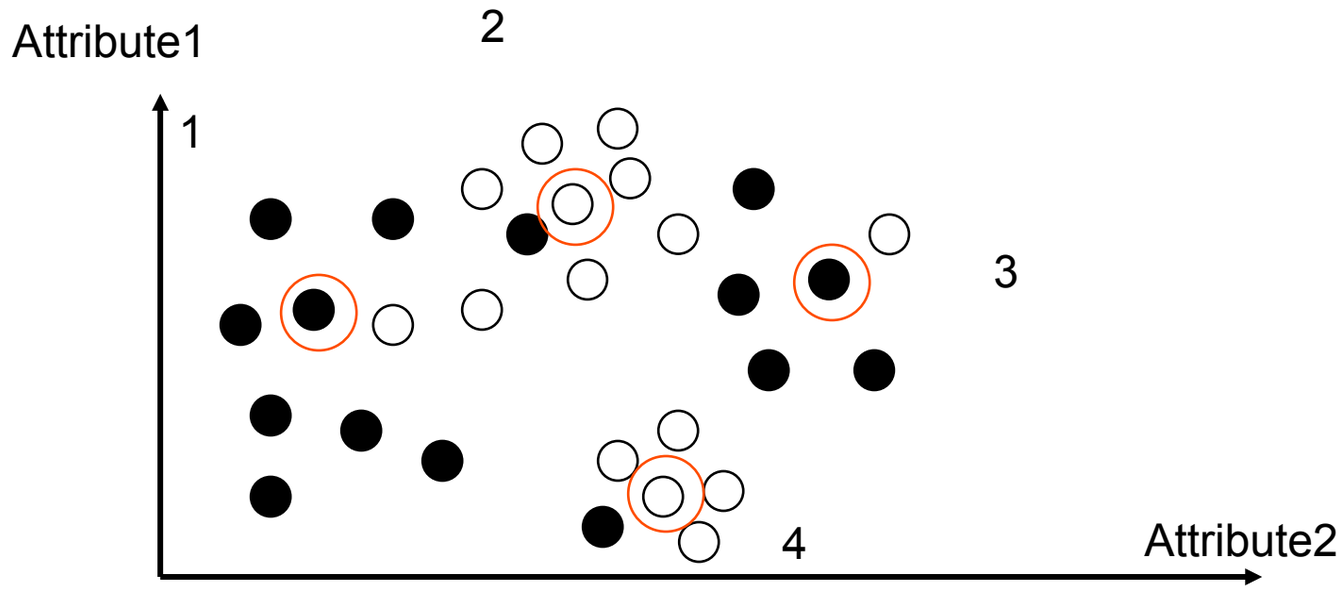
major clustering approaches

- partitioning algorithms/representative-based/prototype-based:
  construct various partitions and then evaluate them by some criterion
- hierarchical clustering
  create a hierarchical representation of the data set using some criterion
- density-based
  based on connectivity and density functions
- grid-based
  based on a multiple-level granularity structure
- model-based
  hypothesize a model for each cluster and find the best fit that model each

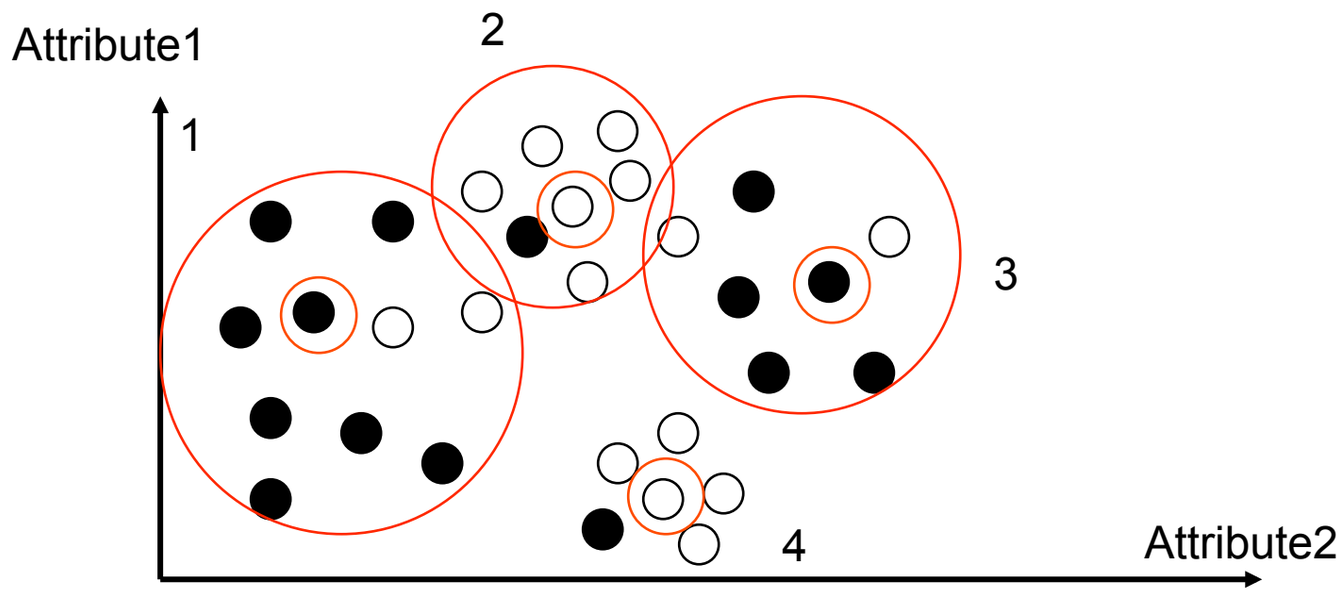
Learning from oncogenomic data

representative-based clustering

- find *representatives*



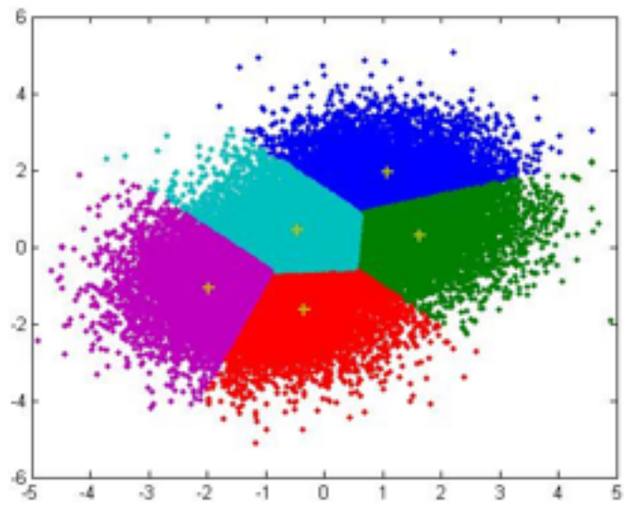
- cluster the remaining around *representatives*



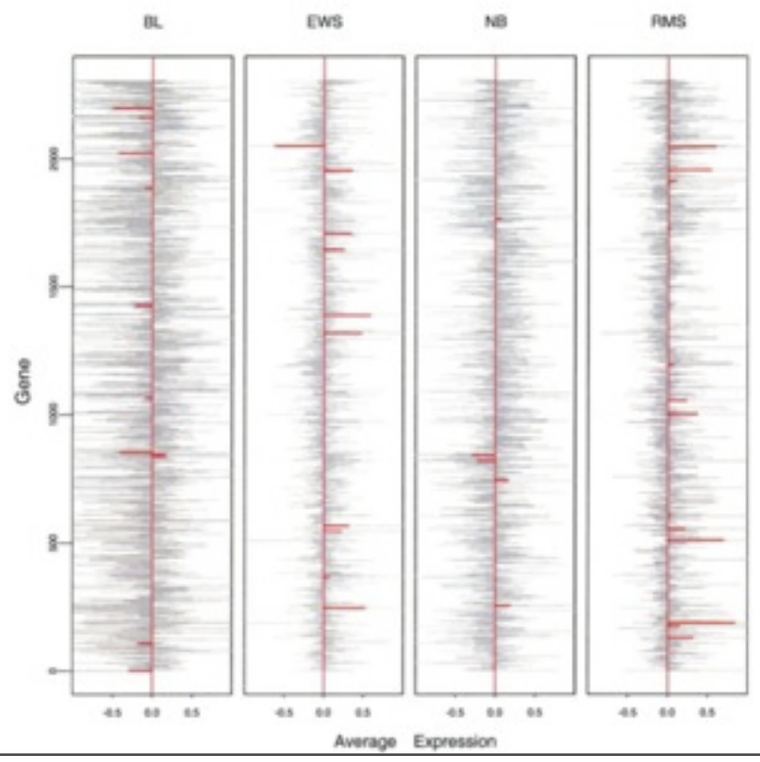
Learning from oncogenomic data

representative-based clustering

- construct a partition of *n* objects into a set of *k* clusters
- given a *k*, find a partition of *k* clusters that optimizes the partition criterion
- examples:
  - K-MEANS: each cluster is represented by the center of the cluster



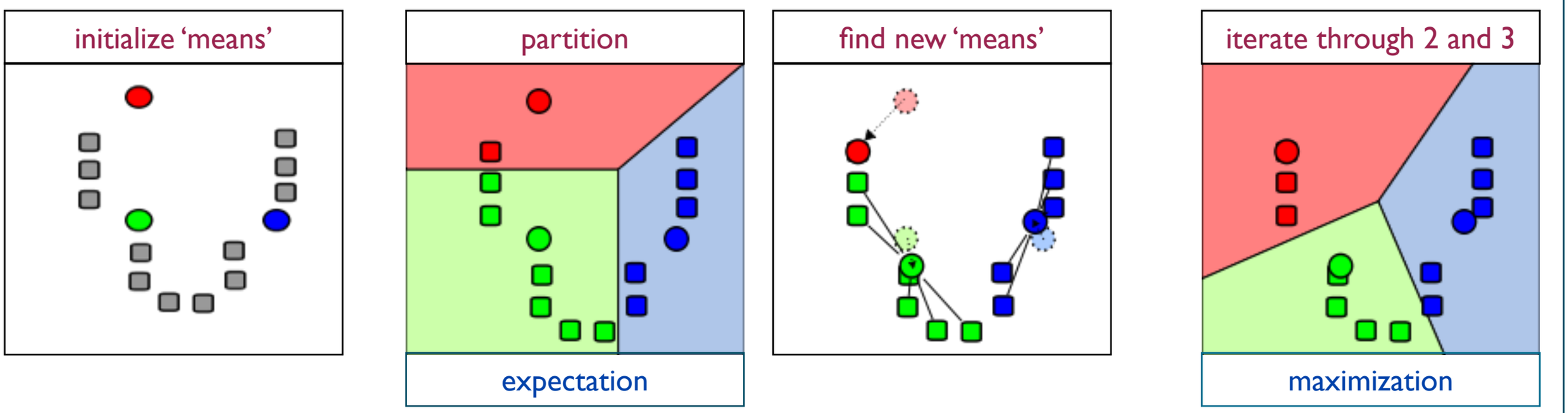
  - PAM: each cluster is represented by one of the objects in the cluster



Learning from oncogenomic data

representative-based clustering: k-means

pros	▸easy to use; well-studied ▸when the number of clusters are known ▸applicable to binary data
cons	▸cluster number is used-defined ▸heuristic (local vs. global optimum) ▸sensitive to noise and outliers ▸sensitive to initialization



Learning from oncogenomic data

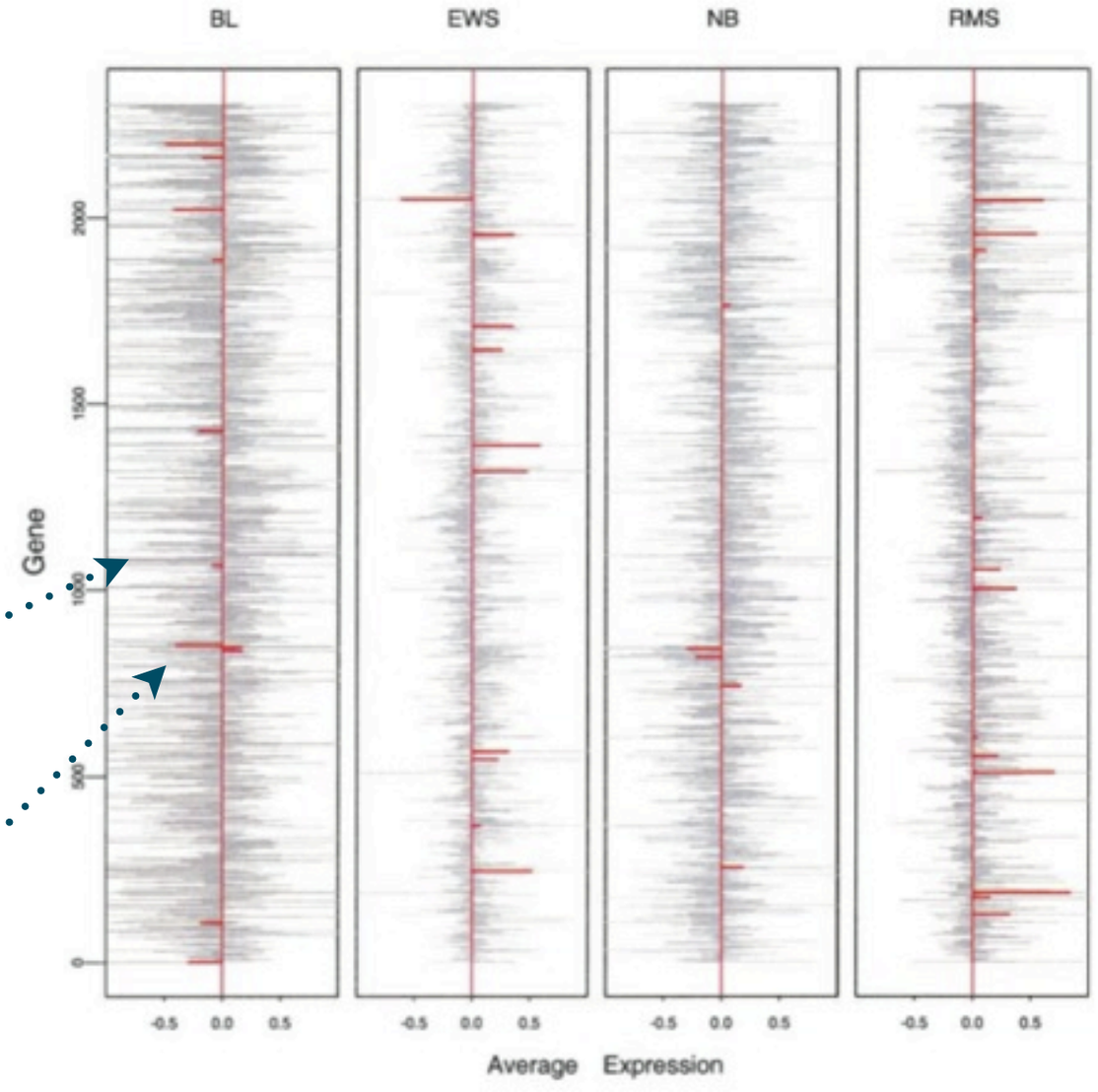
representative-based clustering: PAM

- nearest shrunken centroids and **PAM** (prediction analysis of microarrays - Tibshirani R et al. PNAS 2002)
- shrink the class centroids towards to the overall centroids after standardizing by the within-class stdev

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)},$$

centroids

shrunken centroids

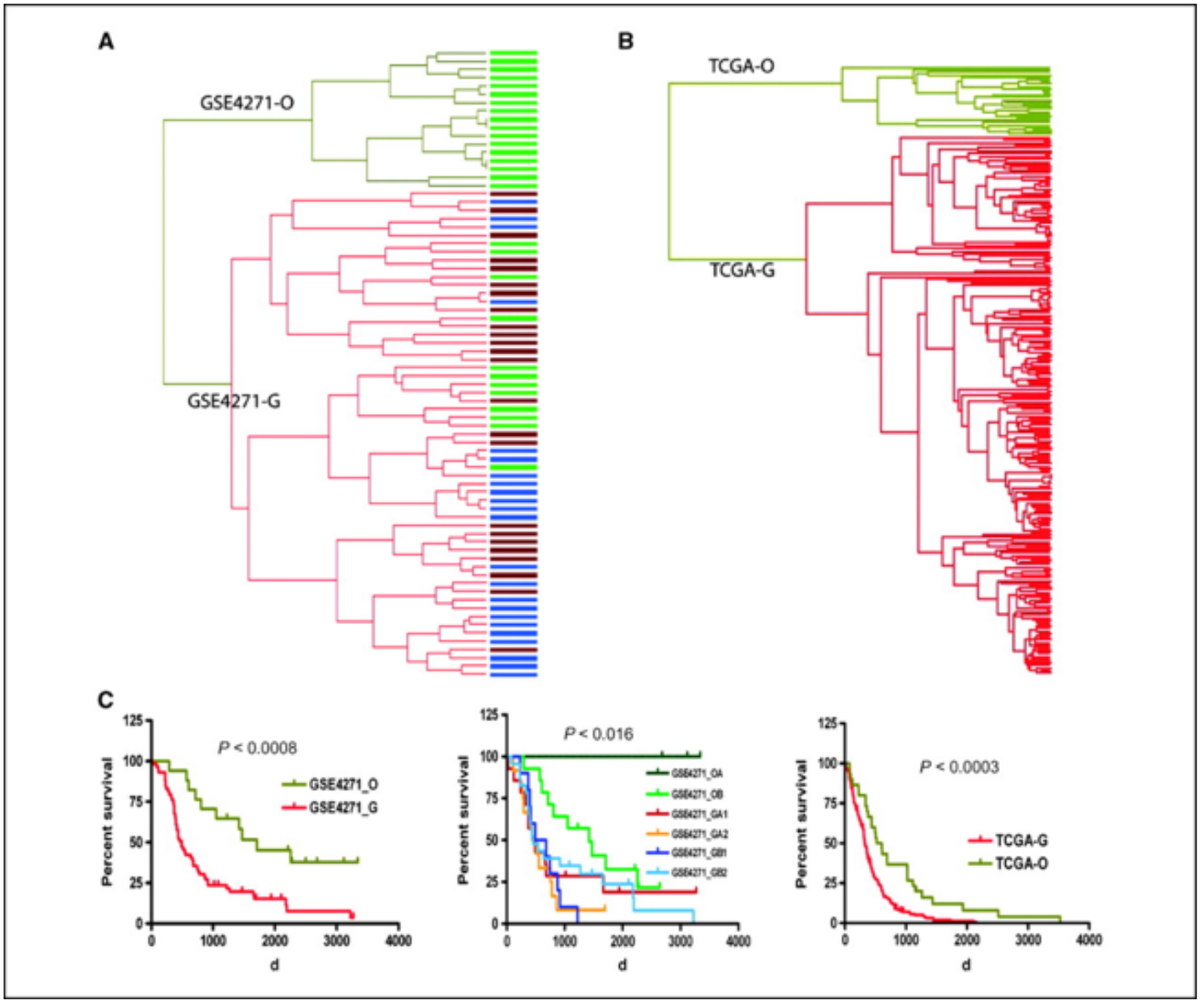


Learning from oncogenomic data

representative-based clustering: PAM

- k-means and NMF identified two main glioma subtypes:

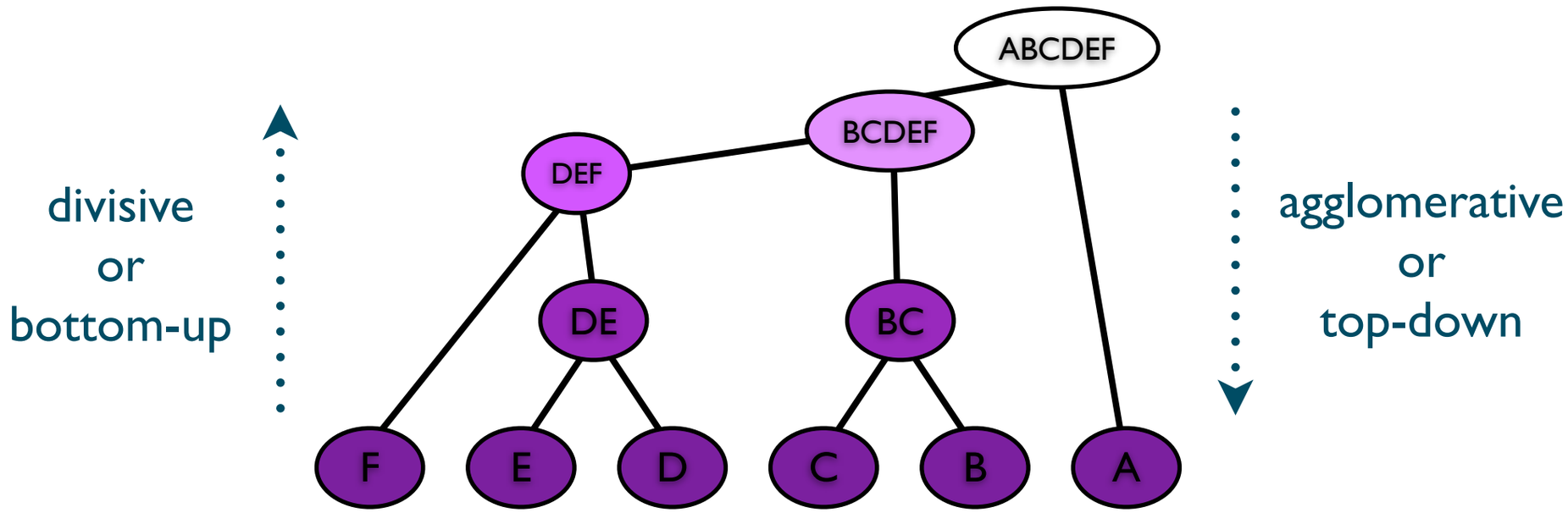
  G (glioblastoma) and *O* (oligodendroglial) among other 6 Glioma Subtypes



Li A and Walling J et al., Cancer Res, 2009

Learning from oncogenomic data

hierarchical clustering

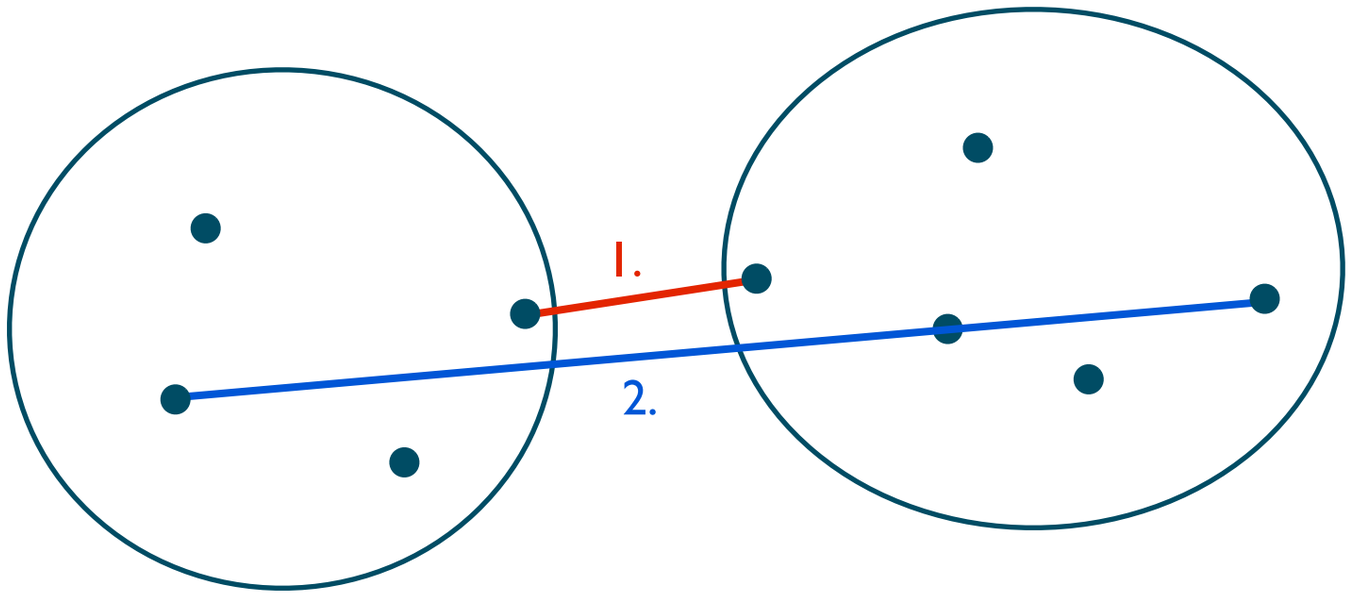


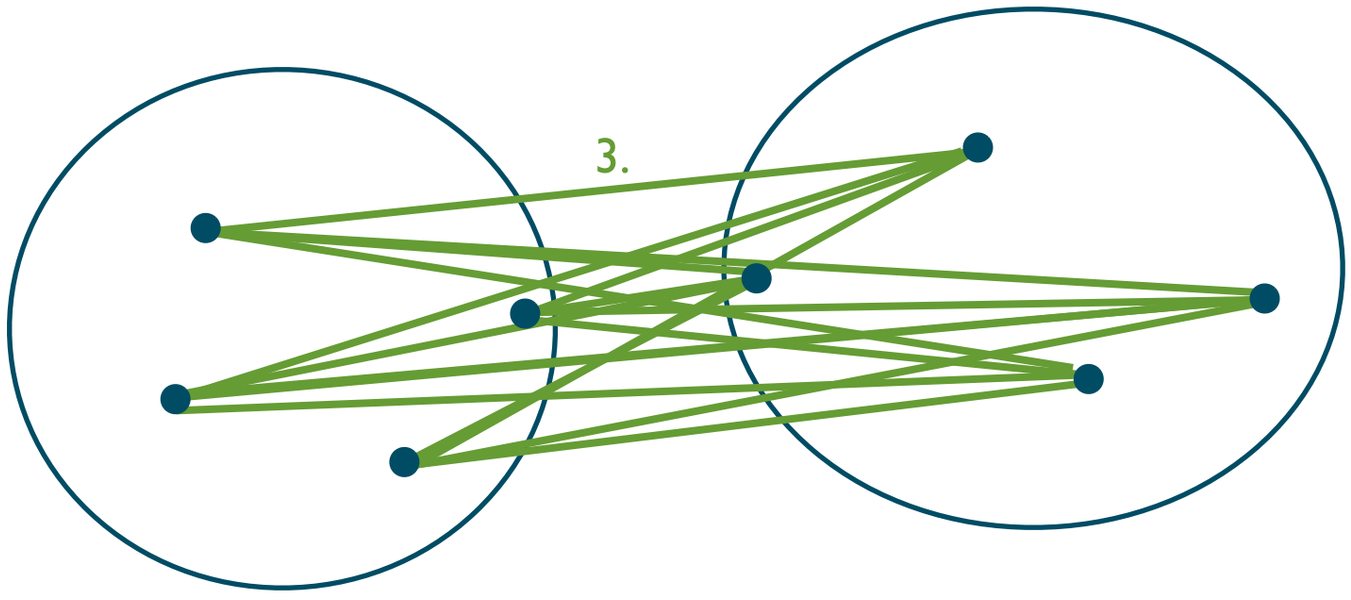
HIERARCHICAL	
USEFUL	▸intuitive visualization ▸any distance measure can be used ▸easy, fast and well-known
DRAWBACK	▸tree structure imposed on data ▸cannot handle partially observed data

dif. approaches
1.linkage algorithms 2.probabilistic models 3.More Bayesian approches

Learning from oncogenomic data

hierarchical clustering



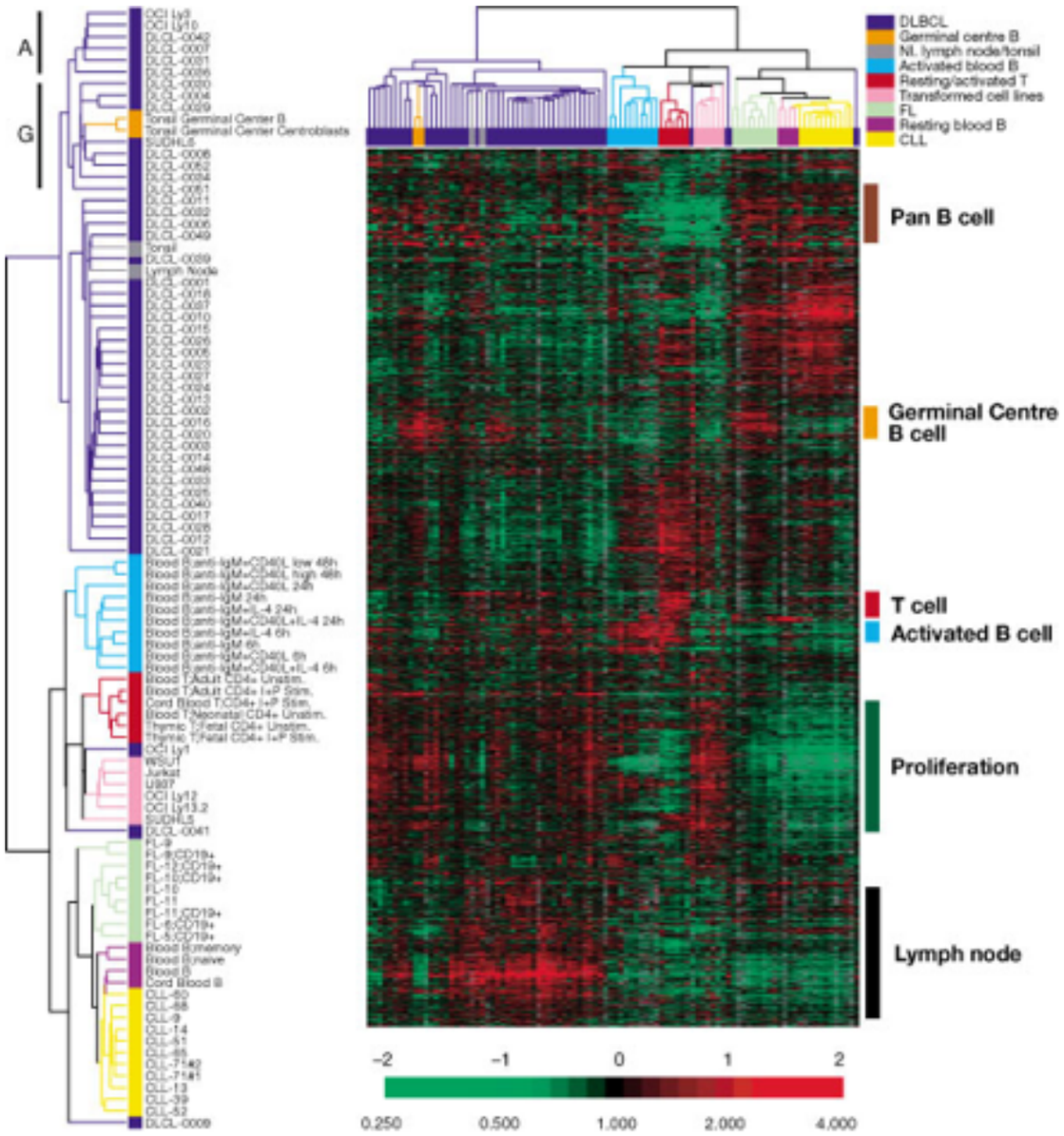


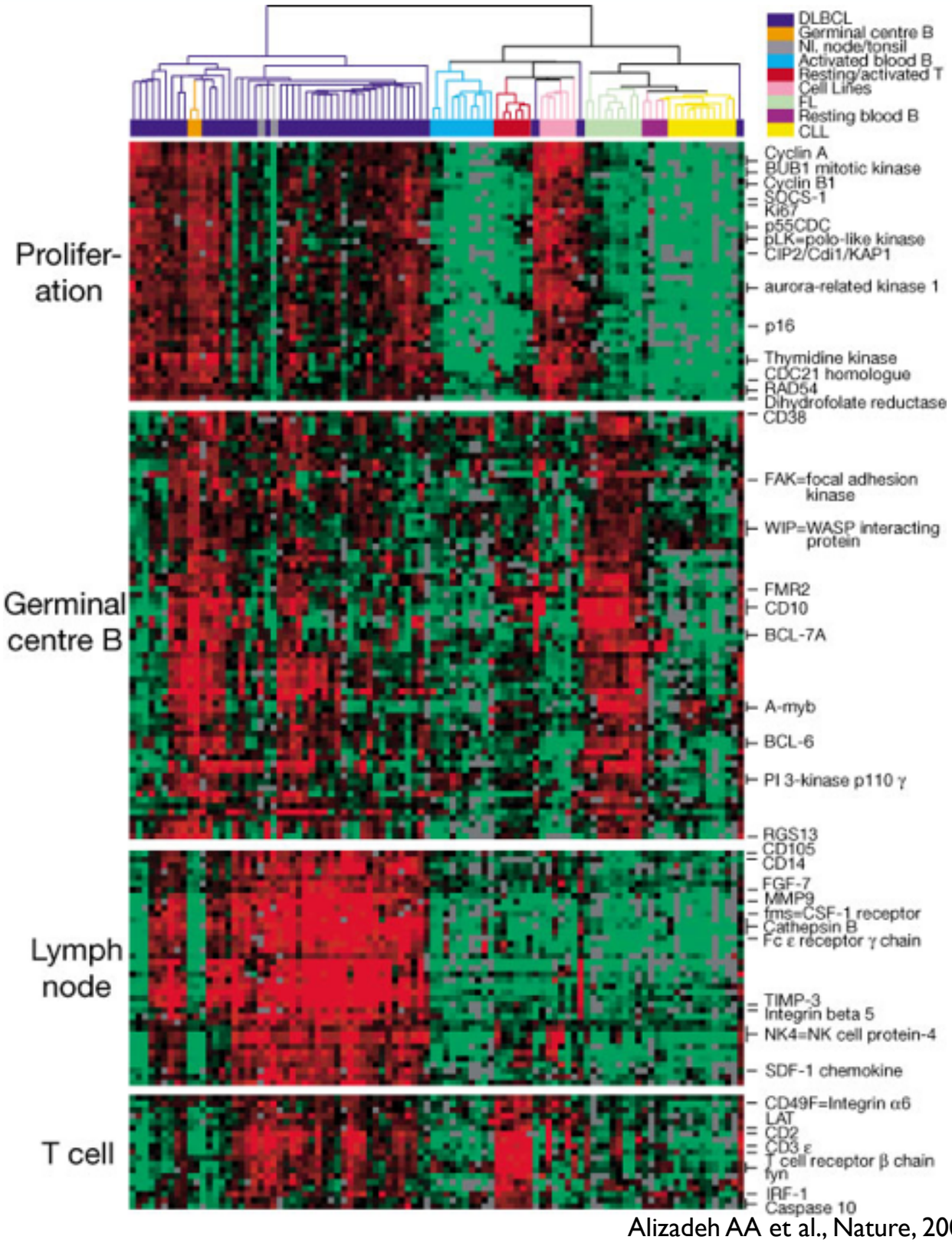
by linkage criterion
1.minimum or single linkage
2.maximum or complete linkage
3.mean or average linkage

Learning from oncogenomic data

hierarchical clustering

- Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

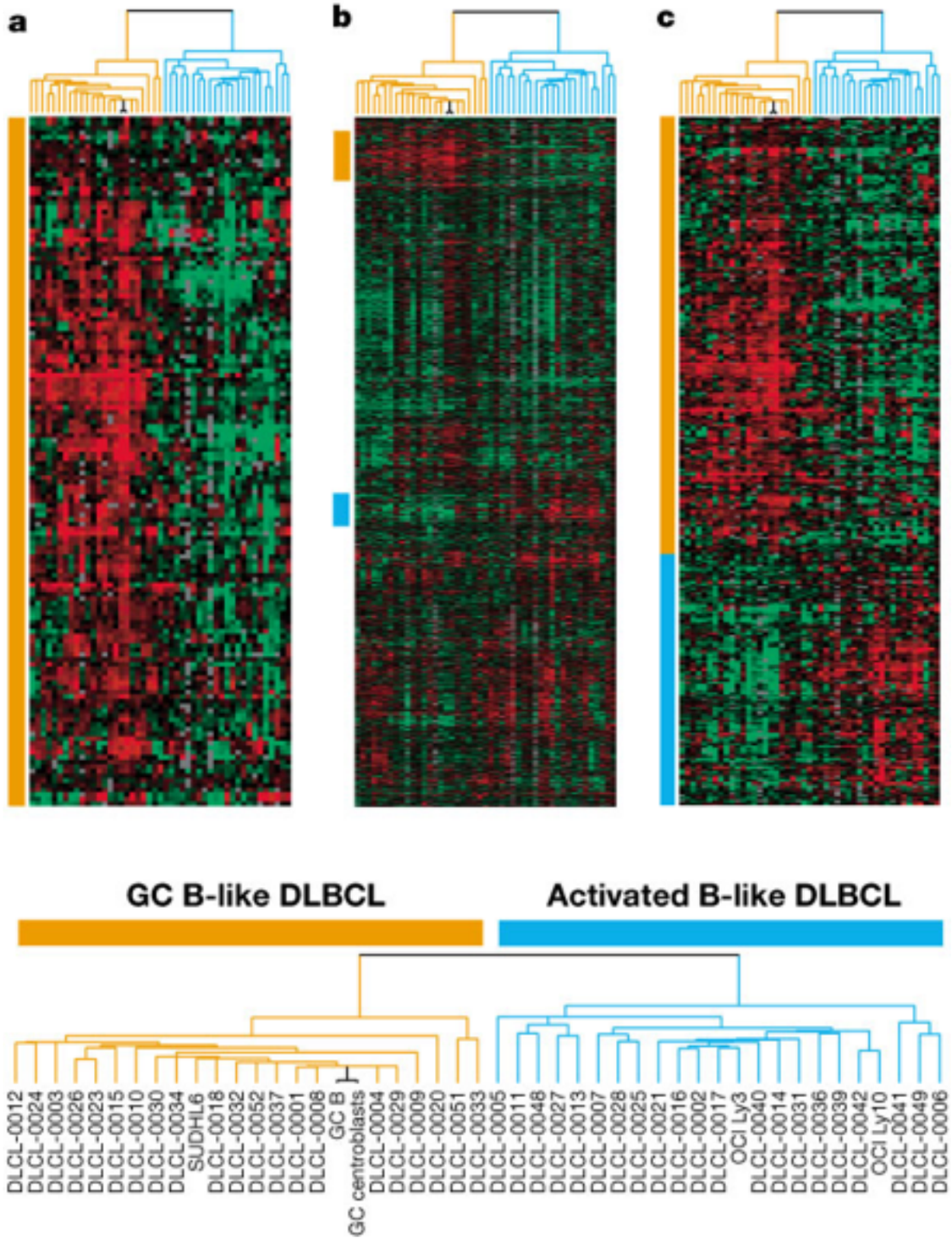




Alizadeh AA et al., Nature, 2000

Learning from oncogenomic data

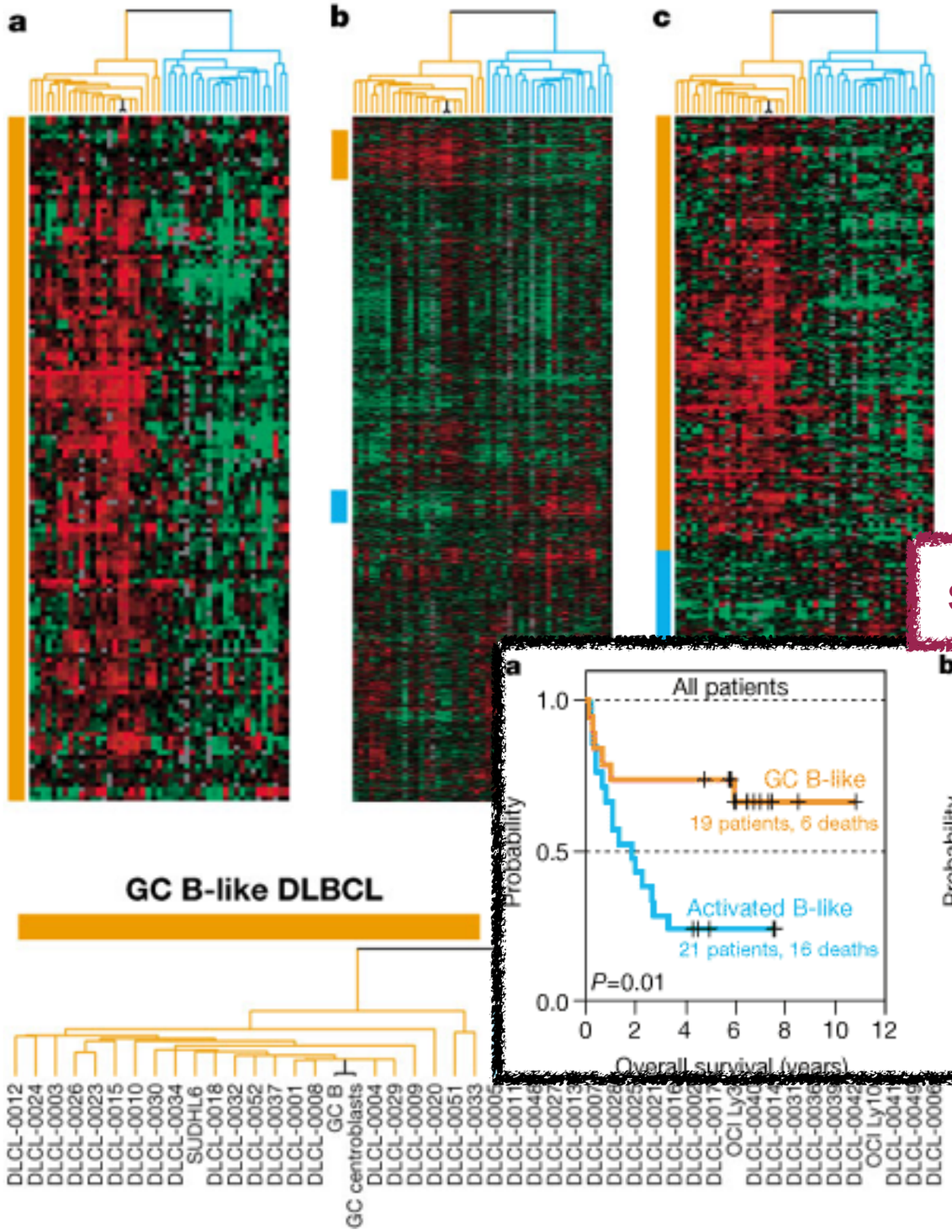
hierarchical clustering



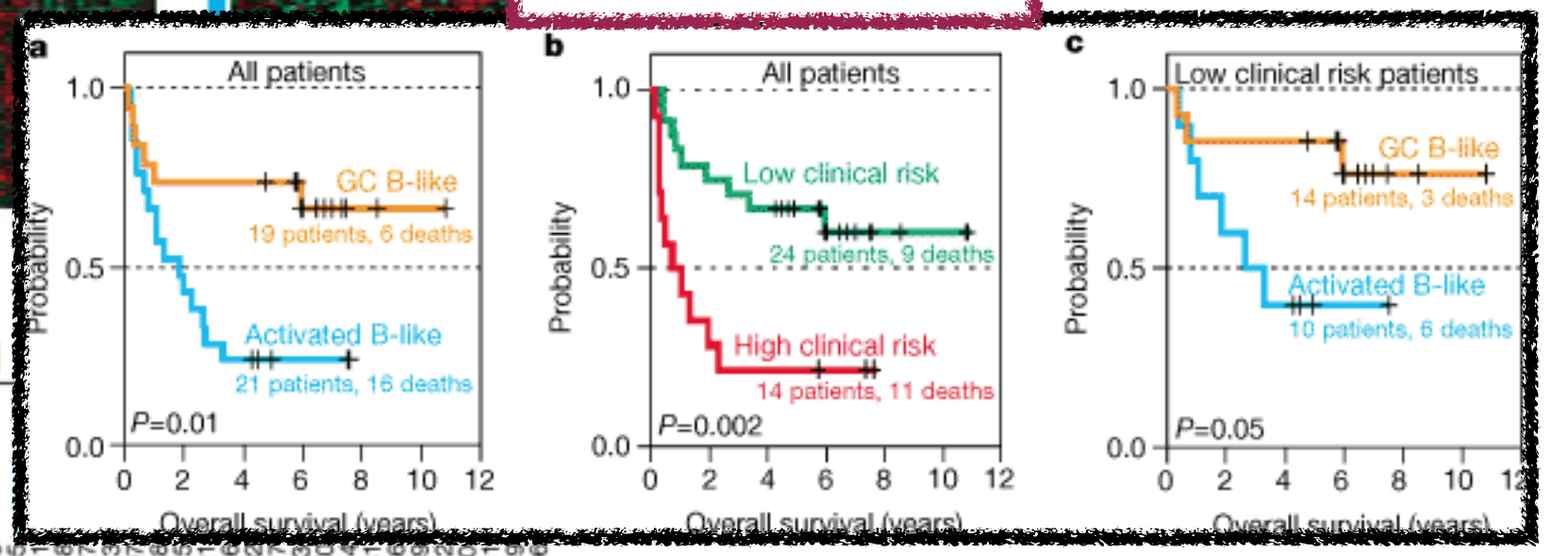
Alizadeh AA et al., Nature, 2000

Learning from oncogenomic data

hierarchical clustering



survival analysis

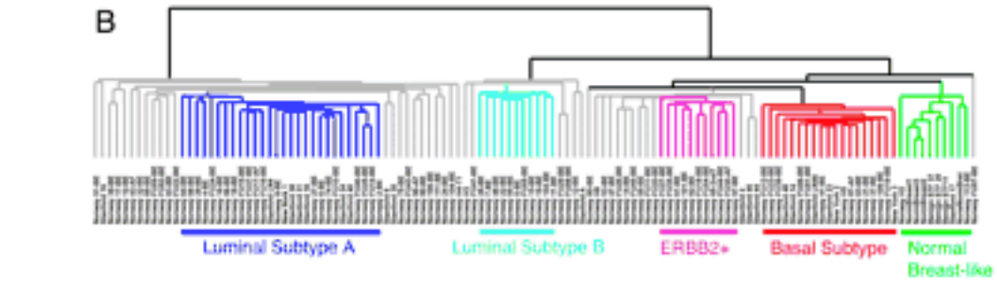


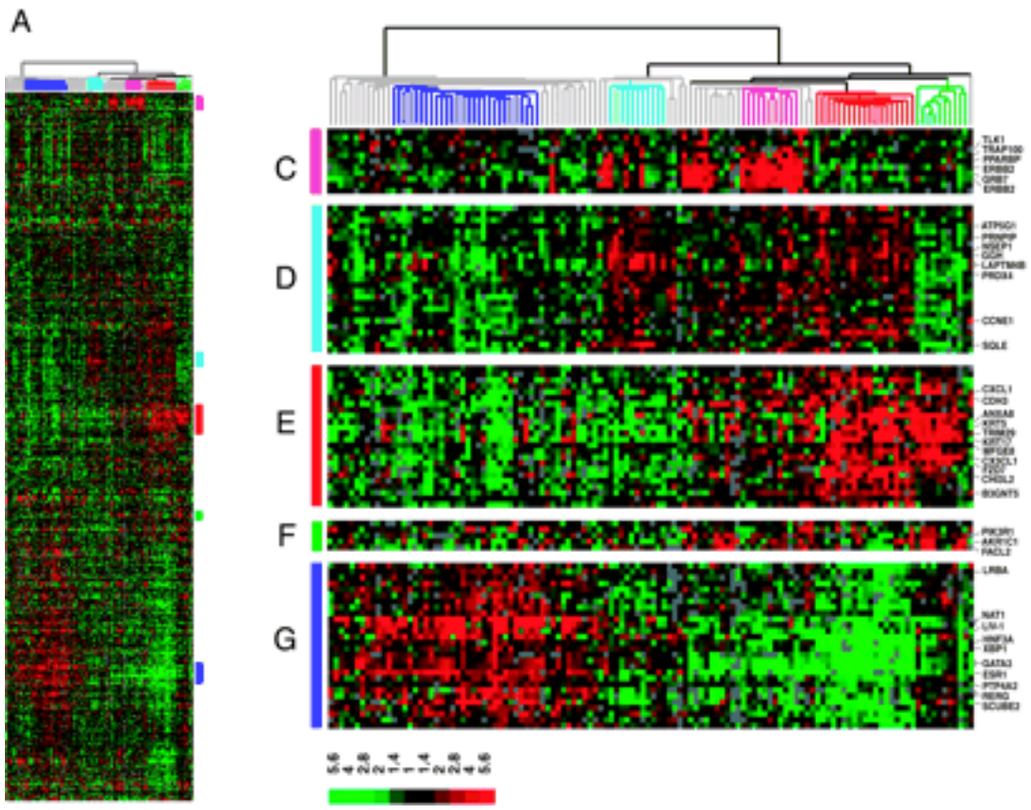
Alizadeh AA et al., Nature, 2000

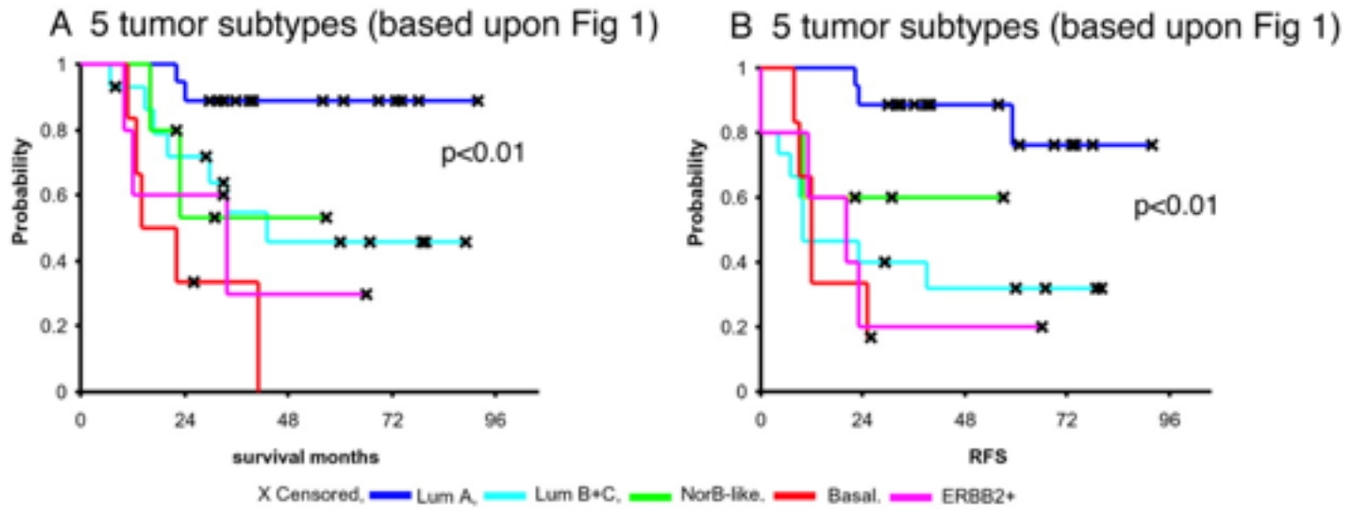
Learning from oncogenomic data

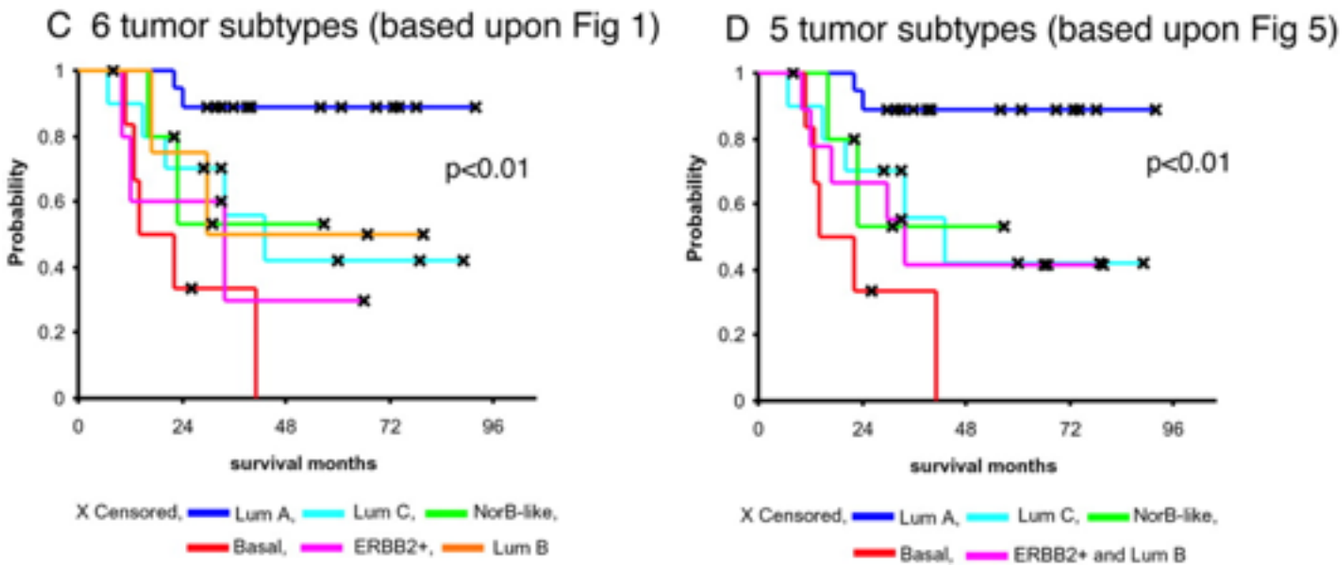
hierarchical clustering

- Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications
- 'intrinsic' gene set detected by SAM (significance analysis of microarrays)
- hierarchical clustering using the intrinsic gene set to cluster breast cancer samples









Sorlie T et al., PNAS, 2001

Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

supervised approach: classification

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attribute is the *class*
- Find a *model* for the class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible
  - A *test set* is used to determine the accuracy of the model.
  - Usually the data set ⋯⋯➤ *test set* to build the model
    ⋮
    ▼
    training set to validate the model

Learning from oncogenomic data

supervised approach: methods

- decision tree based methods
- rule based methods
- memory based reasoning, instance-based learning
- neural networks
- Naive Bayes and Bayesian Belief Networks
- Support Vector Machines
- Ensemble methods

Learning from oncogenomic data

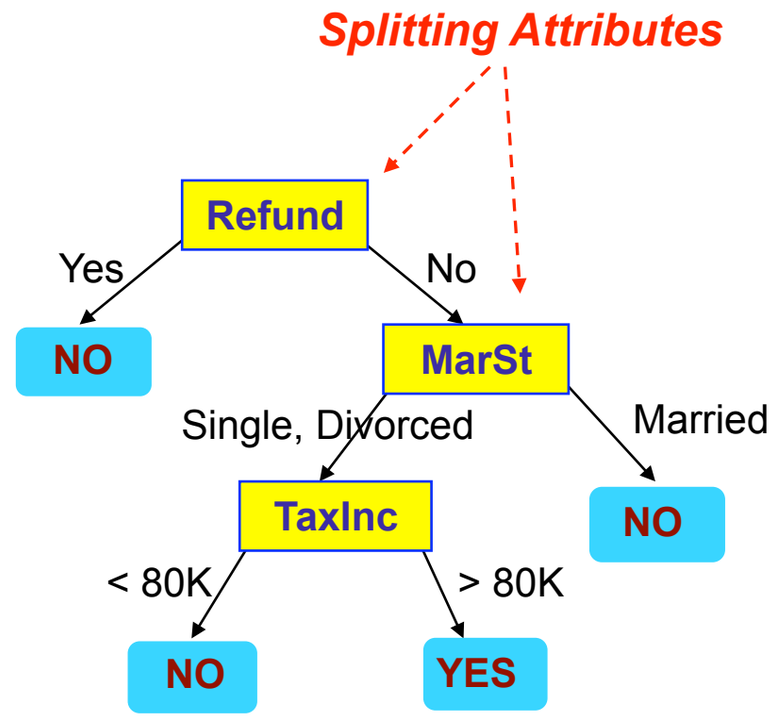
decision trees: a simple example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

categorical, categorical, continuous, class

training data

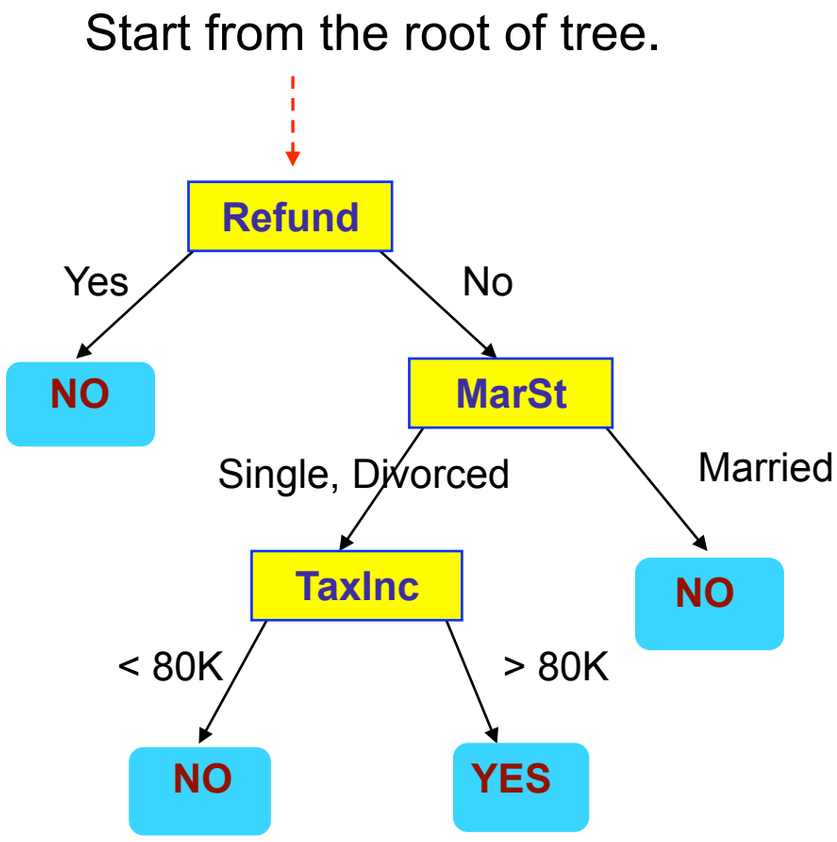
induction



Model: decision tree

Learning from oncogenomic data

decision trees: a simple example

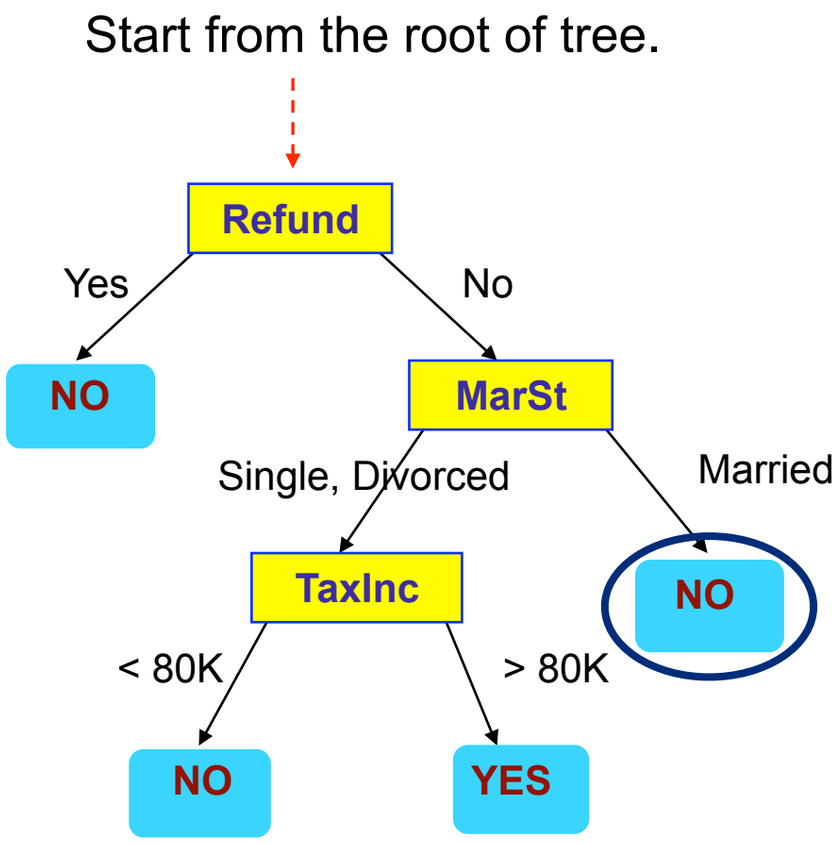


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Learning from oncogenomic data

decision trees: a simple example



Test Data

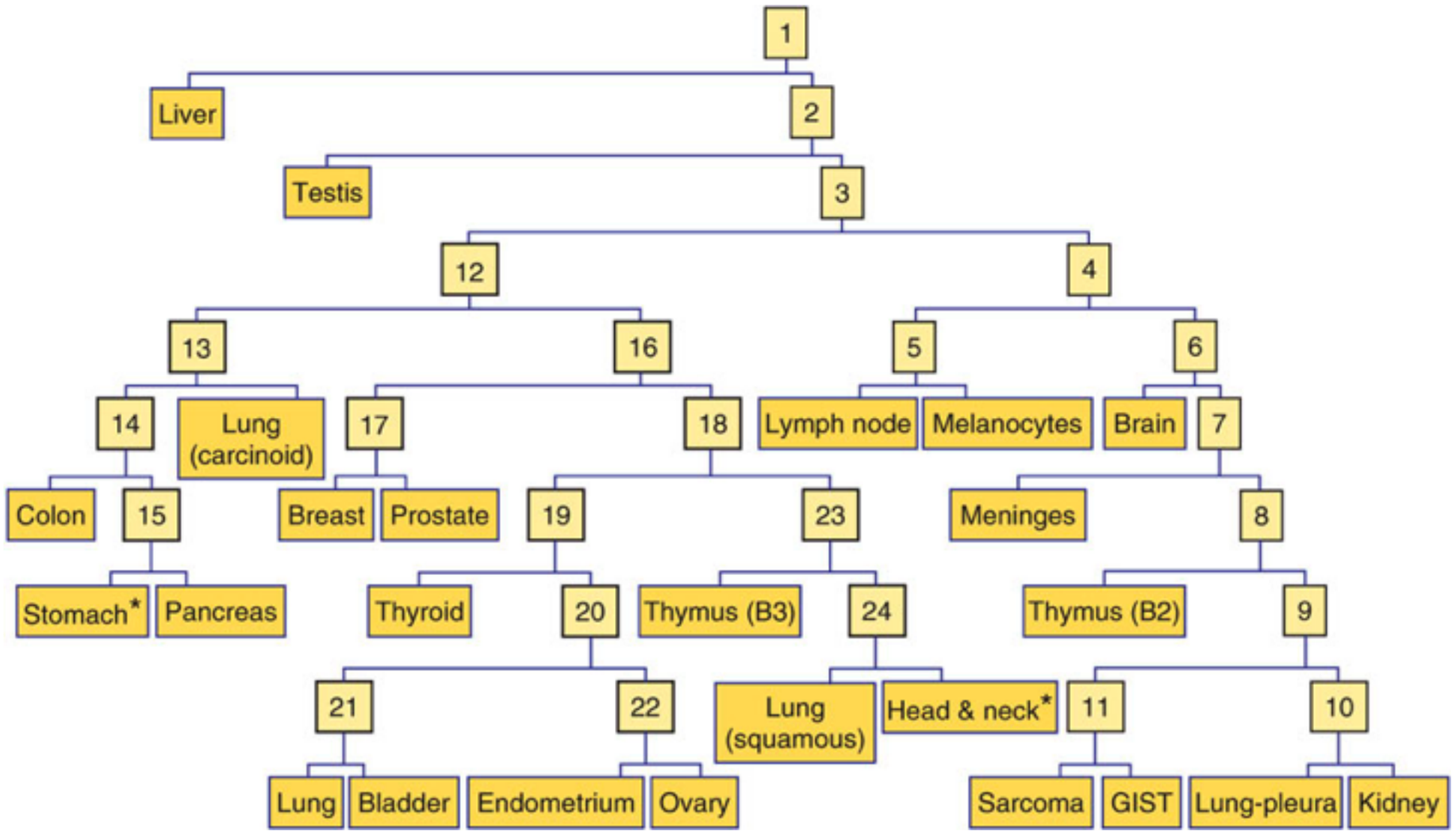
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

deduction

Learning from oncogenomic data

decision trees: real life

- a classifier of only 48 miRNA markers among 22 tissues
- overall accuracy of 90%
- binary decision tree
- uses logistic regression to assign a probability of belonging to one of the two branches of a node



Rosenfeld N et al., Nature Biotech 2008

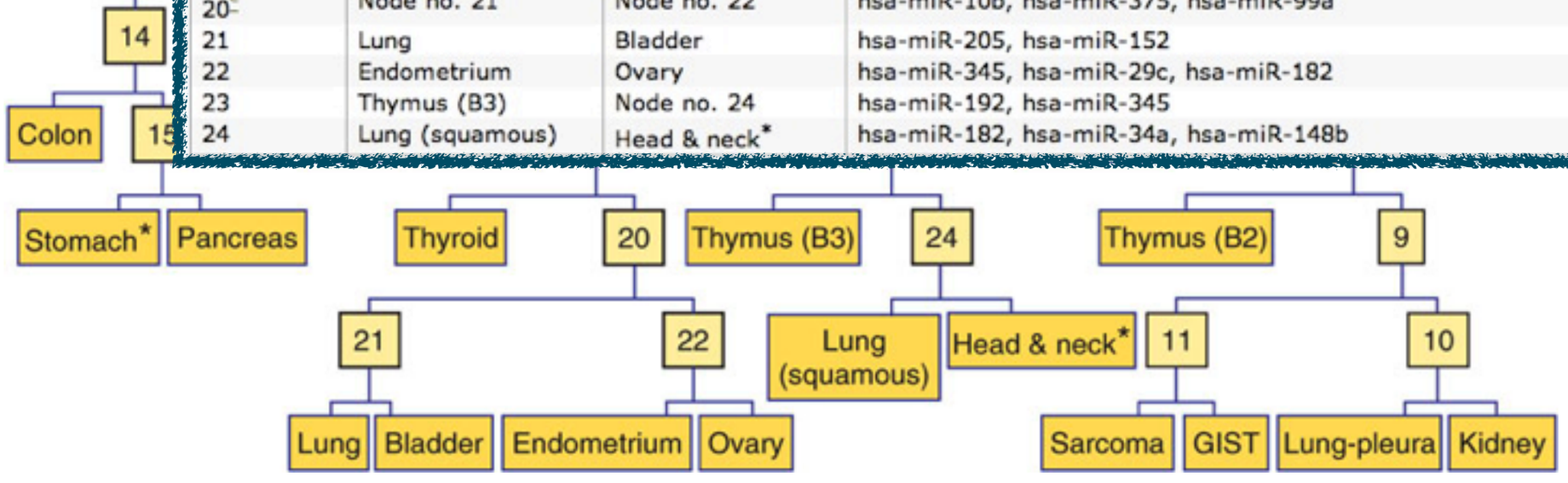
Learning from oncogenomic data

decision trees: real life

- a classifier of only 48 miRNA markers among 22 tissues
- overall accuracy
- binary decision tr
- uses logistic regr

of a node

Node no.	Left branch	Right branch	miRNAs used at the node
1[a]	Liver	Node no. 2	hsa-miR-122a, hsa-miR-200c[b]
2[c]	Testis	Node no. 3	hsa-miR-372
3	Node no. 12	Node no. 4	hsa-miR-200c, hsa-miR-181a, hsa-miR-205
4	Node no. 5	Node no. 6	hsa-miR-146a, hsa-miR-200a, hsa-miR-92a
5	Lymph node	Melanocytes	hsa-miR-142-3p, hsa-miR-509
6	Brain	Node no. 7	hsa-miR-92b, hsa-miR-9*, hsa-miR-124a
7	Meninges	Node no. 8	hsa-miR-152, hsa-miR-130a
8	Thymus (B2)	Node no. 9	hsa-miR-205
9	Node no. 11	Node no. 10	hsa-miR-192, hsa-miR-21, hsa-miR-210, hsa-miR-34b
10	Lung-pleura	Kidney	hsa-miR-194, hsa-miR-382, hsa-miR-210
11	Sarcoma	GIST	hsa-miR-187, hsa-miR-29b
12	Node no. 13	Node no. 16	hsa-miR-145, hsa-miR-194, hsa-miR-205
13	Node no. 14	Lung (carcinoid)	hsa-miR-21, hsa-let-7e
14	Colon	Node no. 15	hsa-let-7i, hsa-miR-29a
15	Stomach*	Pancreas	hsa-miR-214, hsa-miR-19b, hsa-let-7i
16	Node no. 17	Node no. 18	hsa-miR-196a, hsa-miR-363, hsa-miR-31, hsa-miR-193a, hsa-miR-210
17[d]	Breast	Prostate	hsa-miR-27b, hsa-let-7i, hsa-miR-181b
18	Node no. 19	Node no. 23	hsa-miR-205, hsa-miR-141, hsa-miR-193b, hsa-miR-373
19	Thyroid	Node no. 20	hsa-miR-106b, hsa-let-7i, hsa-miR-138
20[e]	Node no. 21	Node no. 22	hsa-miR-10b, hsa-miR-375, hsa-miR-99a
21	Lung	Bladder	hsa-miR-205, hsa-miR-152
22	Endometrium	Ovary	hsa-miR-345, hsa-miR-29c, hsa-miR-182
23	Thymus (B3)	Node no. 24	hsa-miR-192, hsa-miR-345
24	Lung (squamous)	Head & neck*	hsa-miR-182, hsa-miR-34a, hsa-miR-148b



Rosenfeld N et al., Nature Biotech 2008

Learning from oncogenomic data

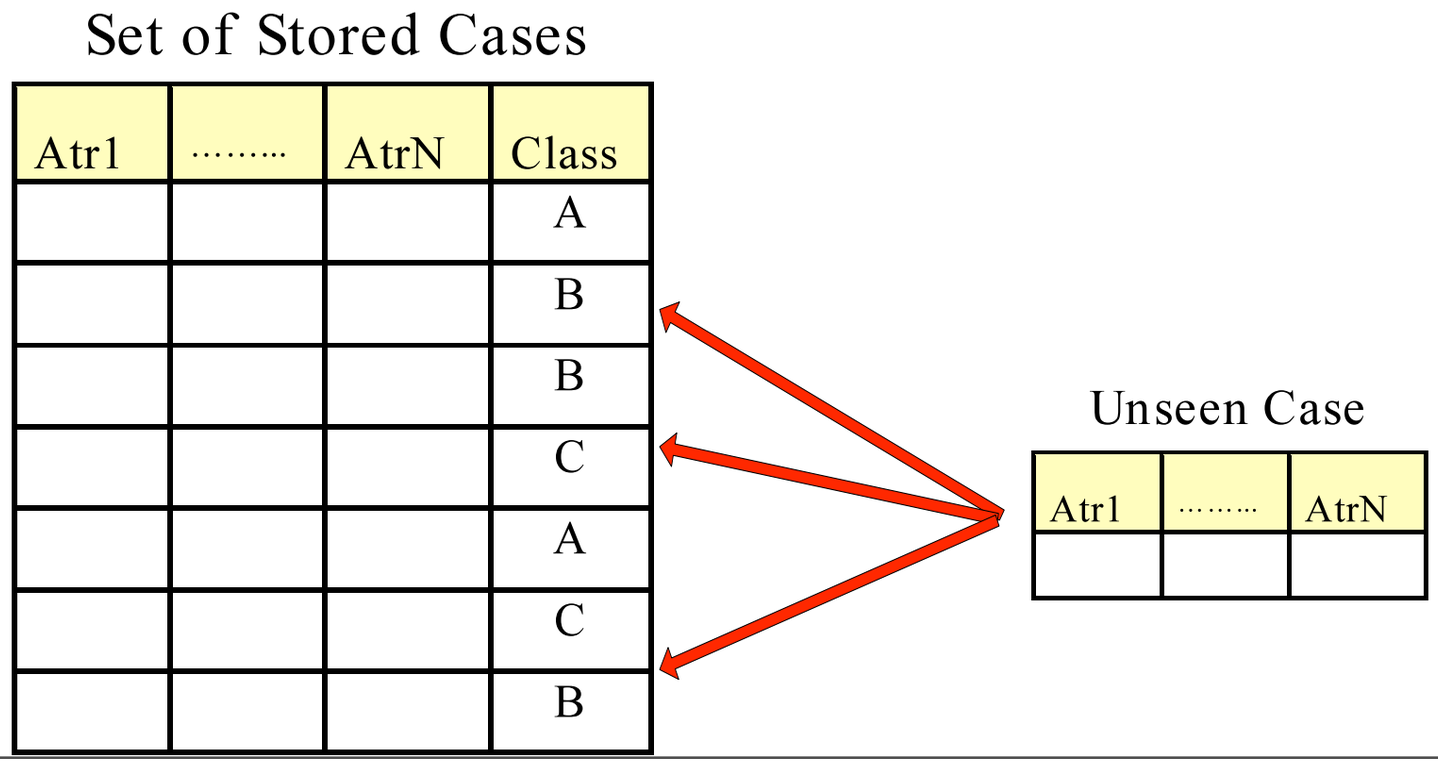
decision trees: pros and cons

- :) inexpensive to construct and extremely fast at classification
- :) can handle both continuous and symbolic attributes
- :) a standard method
- :)no need for distance functions
- :(relies on rectangular appromixation

Learning from oncogenomic data

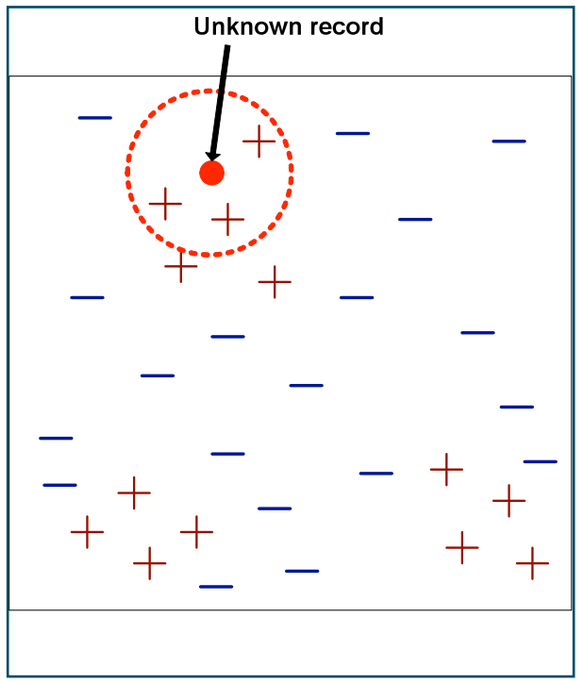
instance-based methods

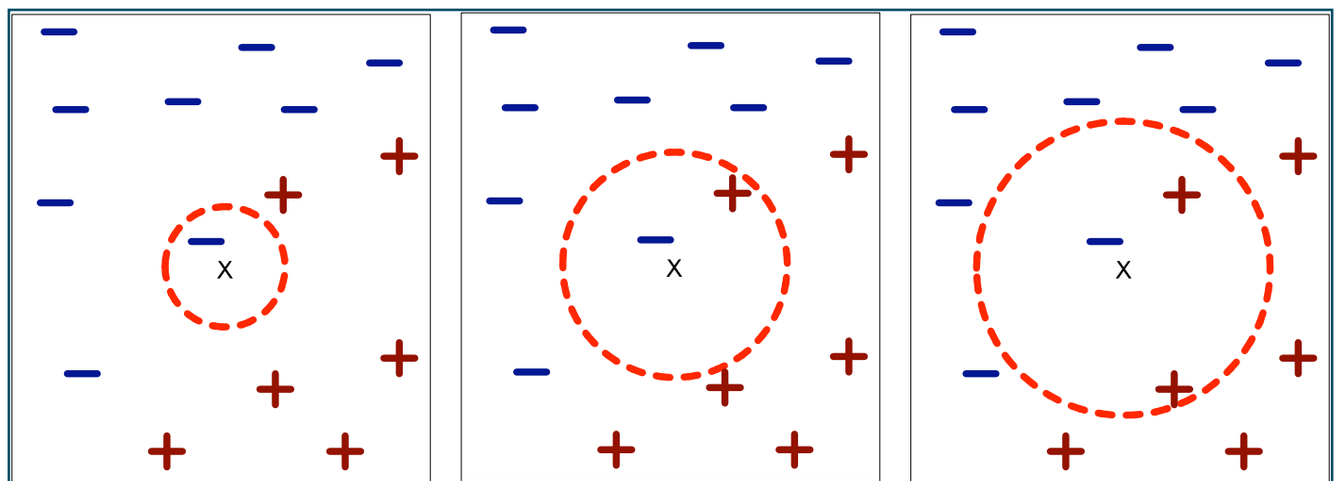
- does not create a model but use training examples directly to classify unseen example (‘*lazy*’ classifiers)
- examples:
  - nearest neighbor
    - chooses k ‘closest’ point (nearest neighbors)



Learning from oncogenomic data

k-NN





Learning from oncogenomic data

k-NN: pros and cons

- choose k
  - if too small, sensitive to noise
  - if too large, indistinct class boundaries
  - use techniques like cross validation
- sensitive to noise, so scaling or selecting features is important
- the quality of the distance function
  - large margin nearest neighbor
- better-represented classes may dominate the prediction
  - weighing
- nearest neighbor search can be expensive
  - linear search, space partitioning, etc.

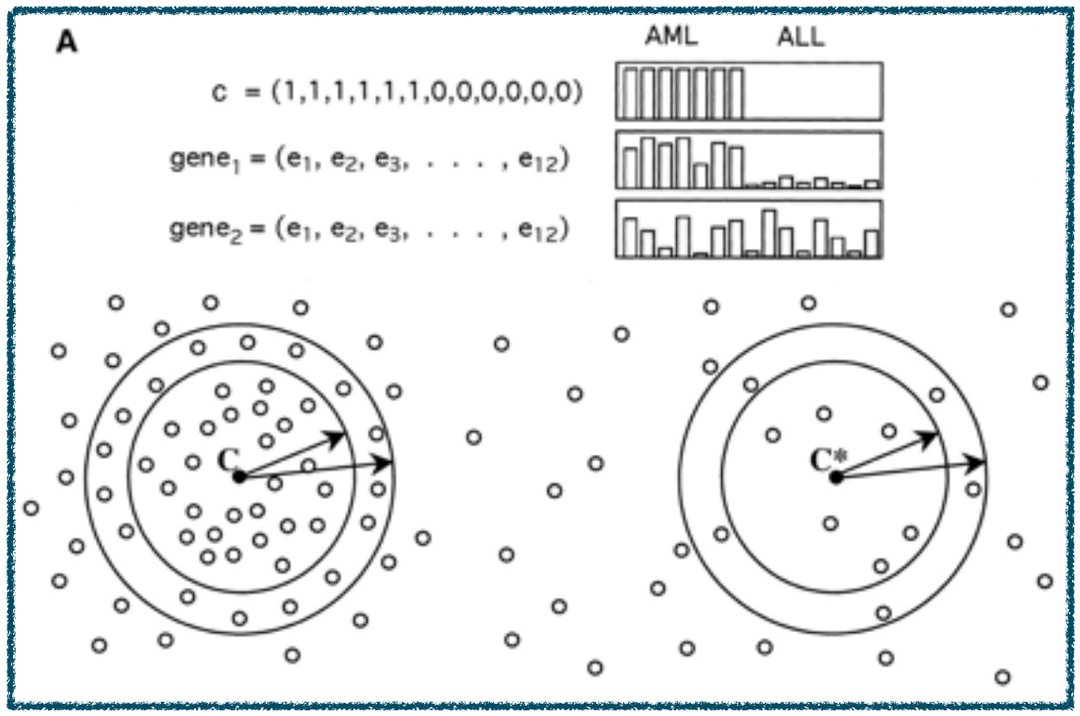
✓ high accuracy

✓ quite popular

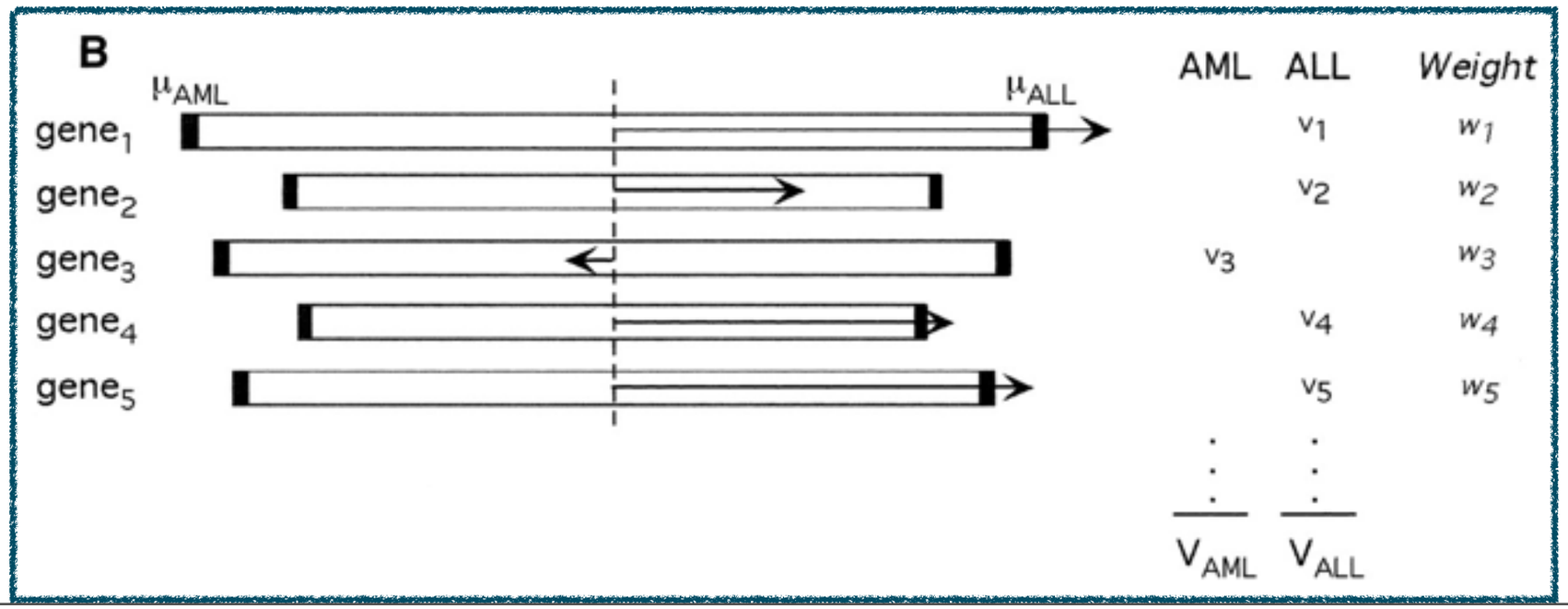
Learning from oncogenomic data

k-NN: pros and cons

- to create a 'class predictor' to classify new, unknown cases of AML and ALL
- '**neighborhood analysis**', the closeness of the gene to idealized expression patterns



random idealized expression patterns

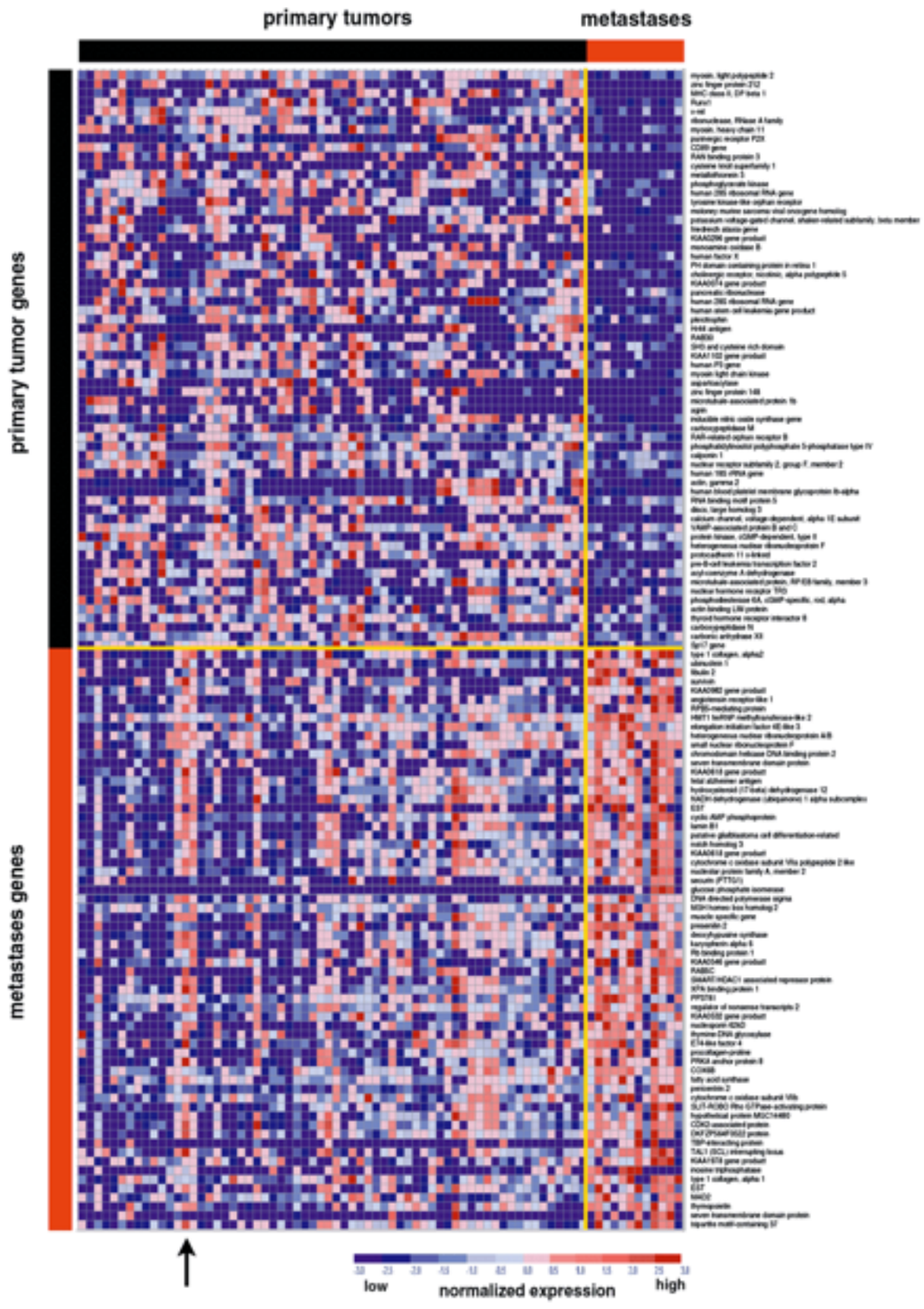


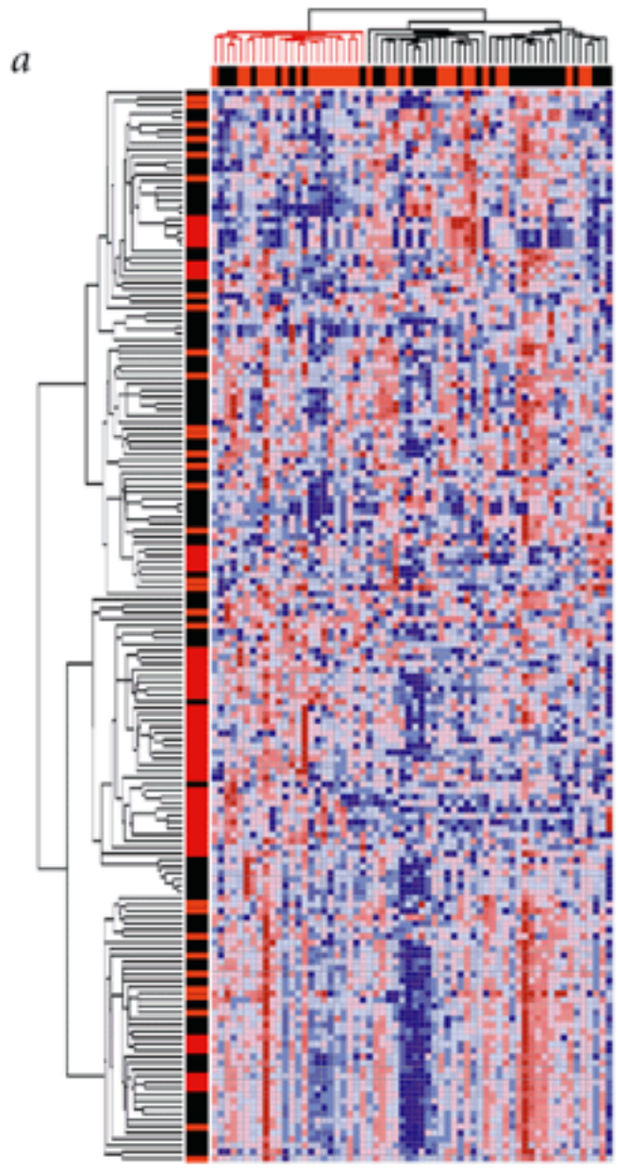
$$v_i = |x_i - (\mu_{AML} + \mu_{ALL})/2|$$

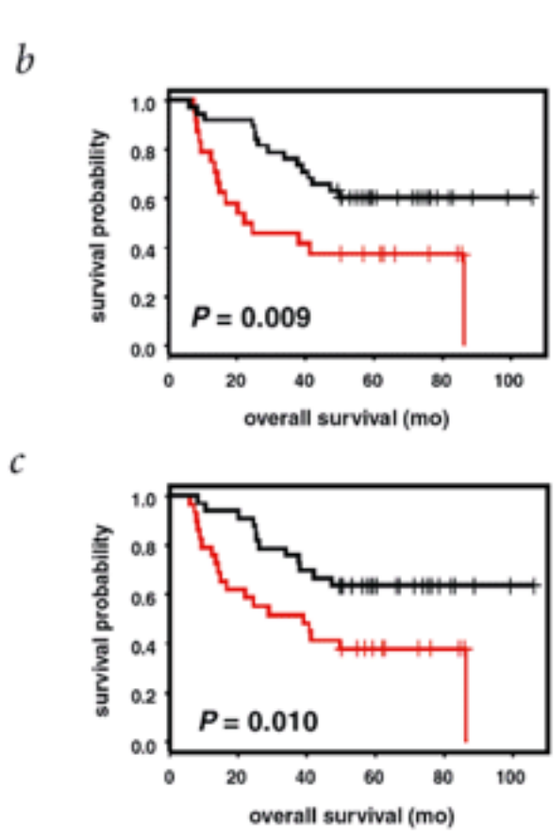
Golub TR et al., Science 1999

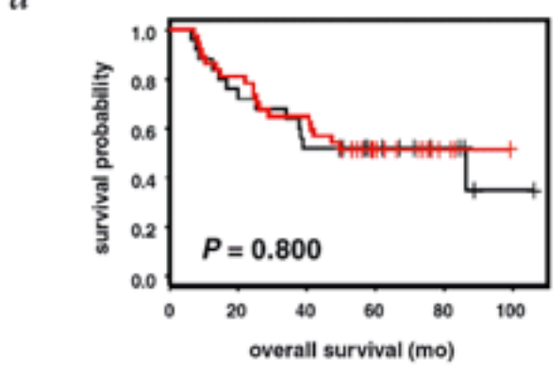
Learning from oncogenomic data

neighborhood analysis







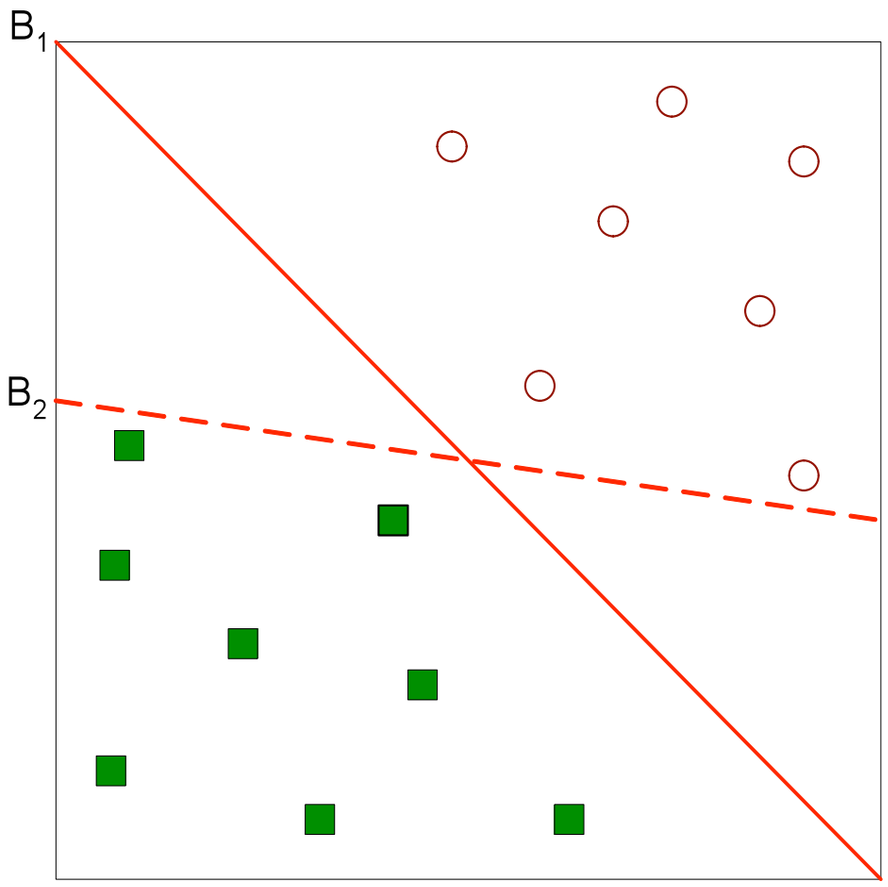


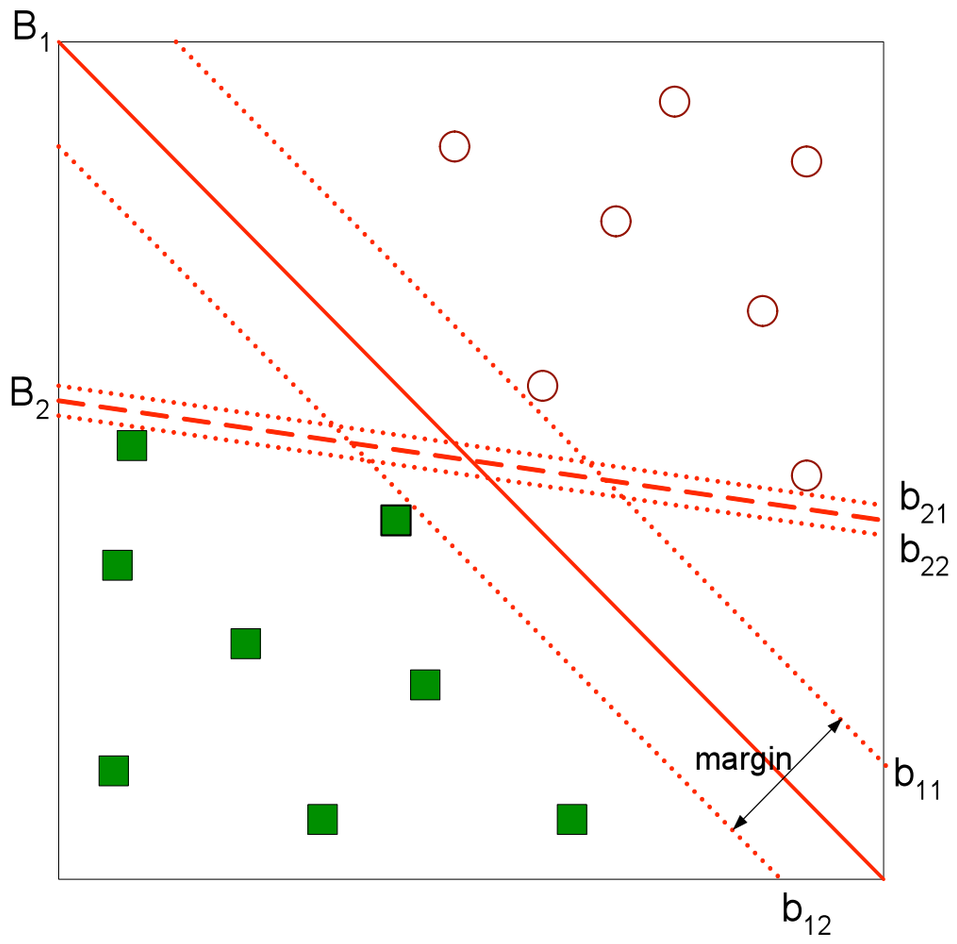
Ramaswamy S et al., Nature Gen 2002

Learning from oncogenomic data

support vector machines

- constructs a hyperplane or a set hyperplanes in a high or infinite dimension space for classification, regression or other tasks





Learning from oncogenomic data

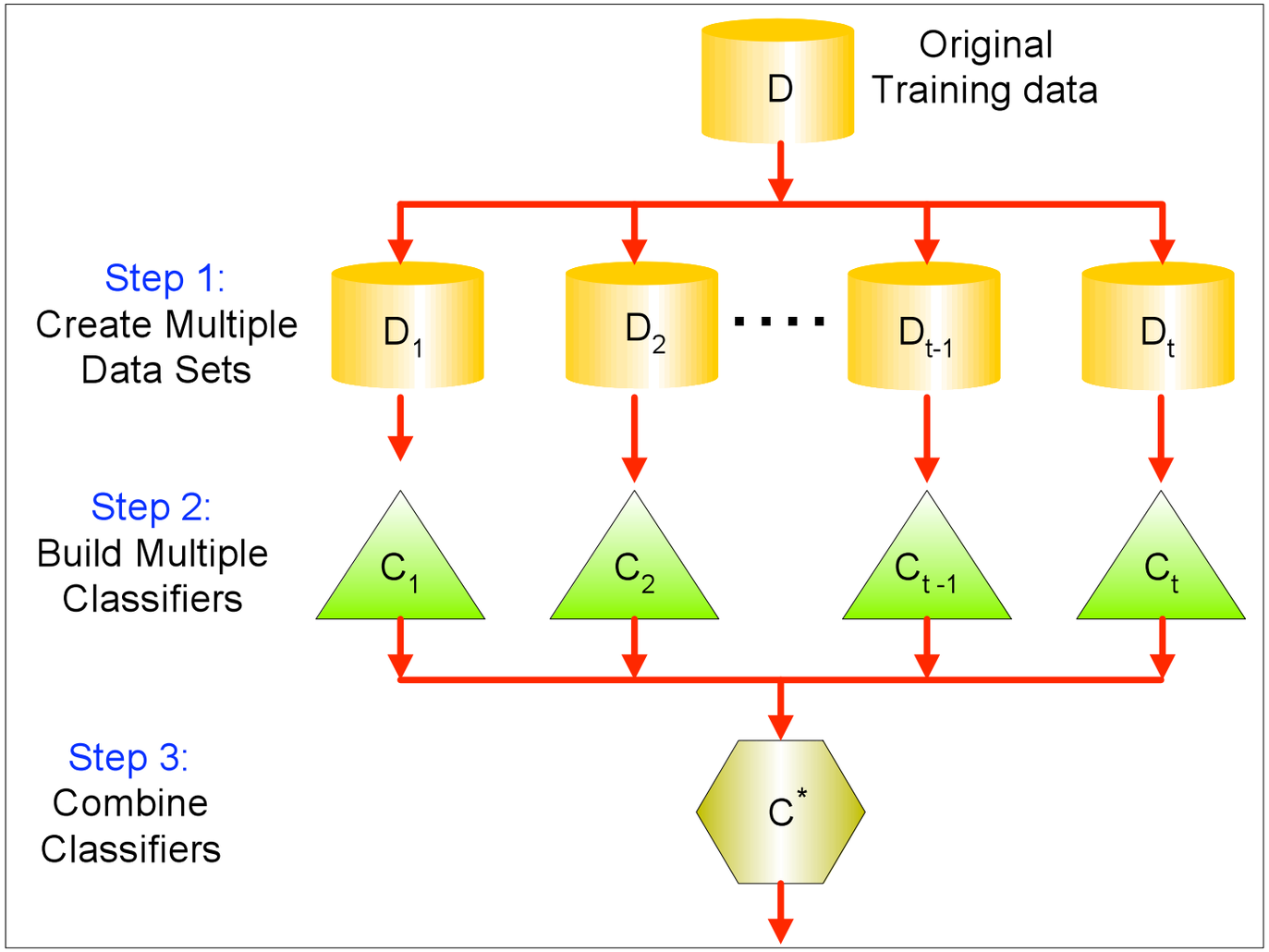
support vector machines

- optimization process is quite slow
- uncalibrated class membership probabilities
- only directly applicable for two-class tasks
  - multi-class SVM

✓ high accuracy

Learning from oncogenomic data

ensemble methods



- if you have 25 independent classifiers with error rate e = 0.35
  - the error rate of the ensemble is $\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$

- examples: bagging (sampling with replacement) and boosting (adaptively change the distribution of training data by focusing on previously misclassified records)

Learning from oncogenomic data

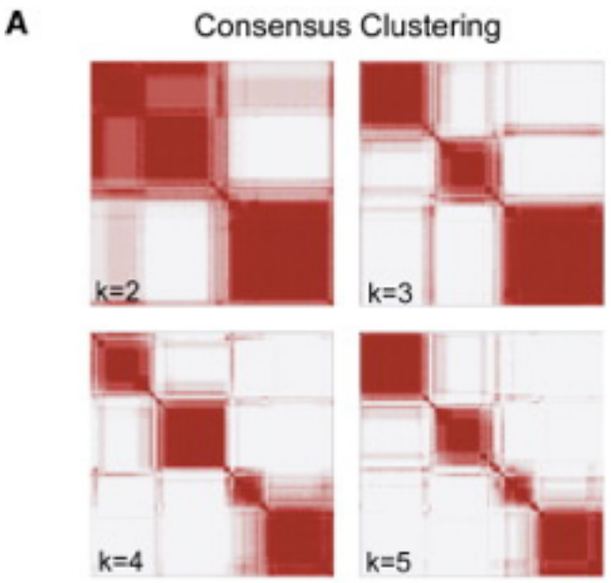
one example to describe it all

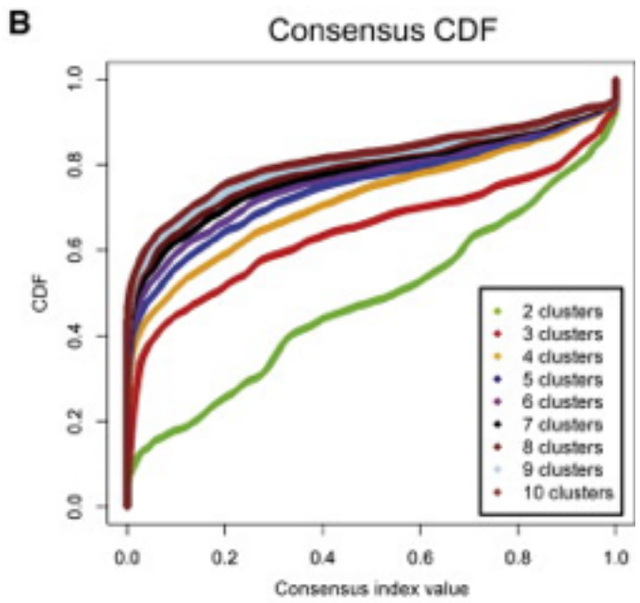
A, B) consensus clustering (Monti et al. 2003)

C) SigClust for evealuation of cluster significance (Liu et al. 2008)

D) silhouette plots to detect core samples

ClaNC to build a 840-gene classifier

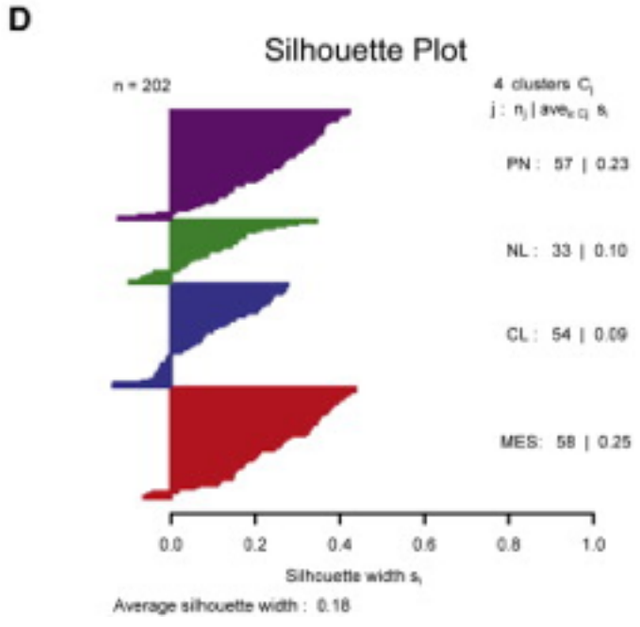




C

SigClust: All Pairwise

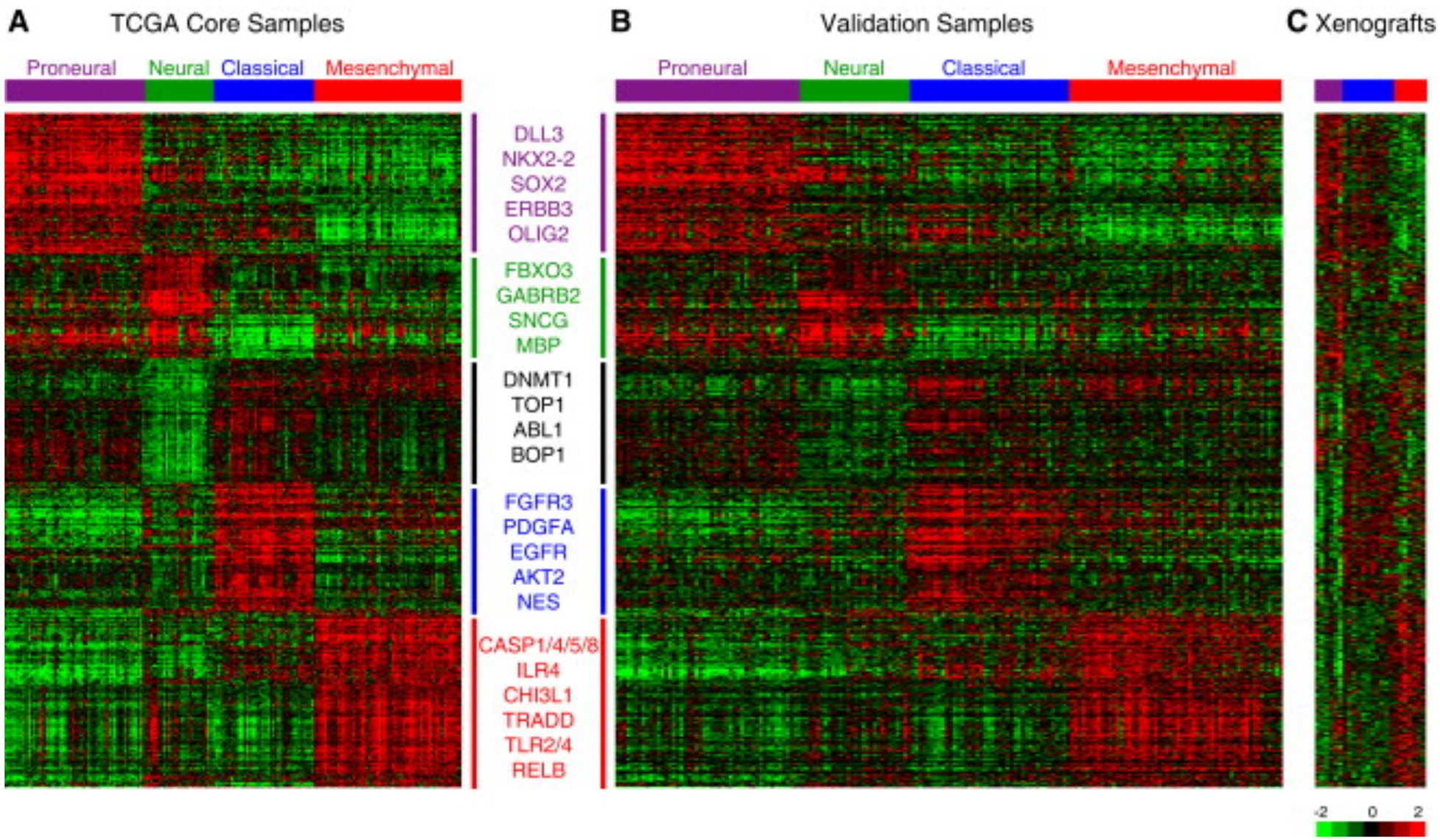
Cluster	PN	NL	CL	MES
PN				
NL	3.90E-03			
CL	1.21E-10	3.70E-04		
MES	5.93E-17	2.69E-07	5.17E-06	



TCGA, Cancer Cell 2010

Learning from oncogenomic data

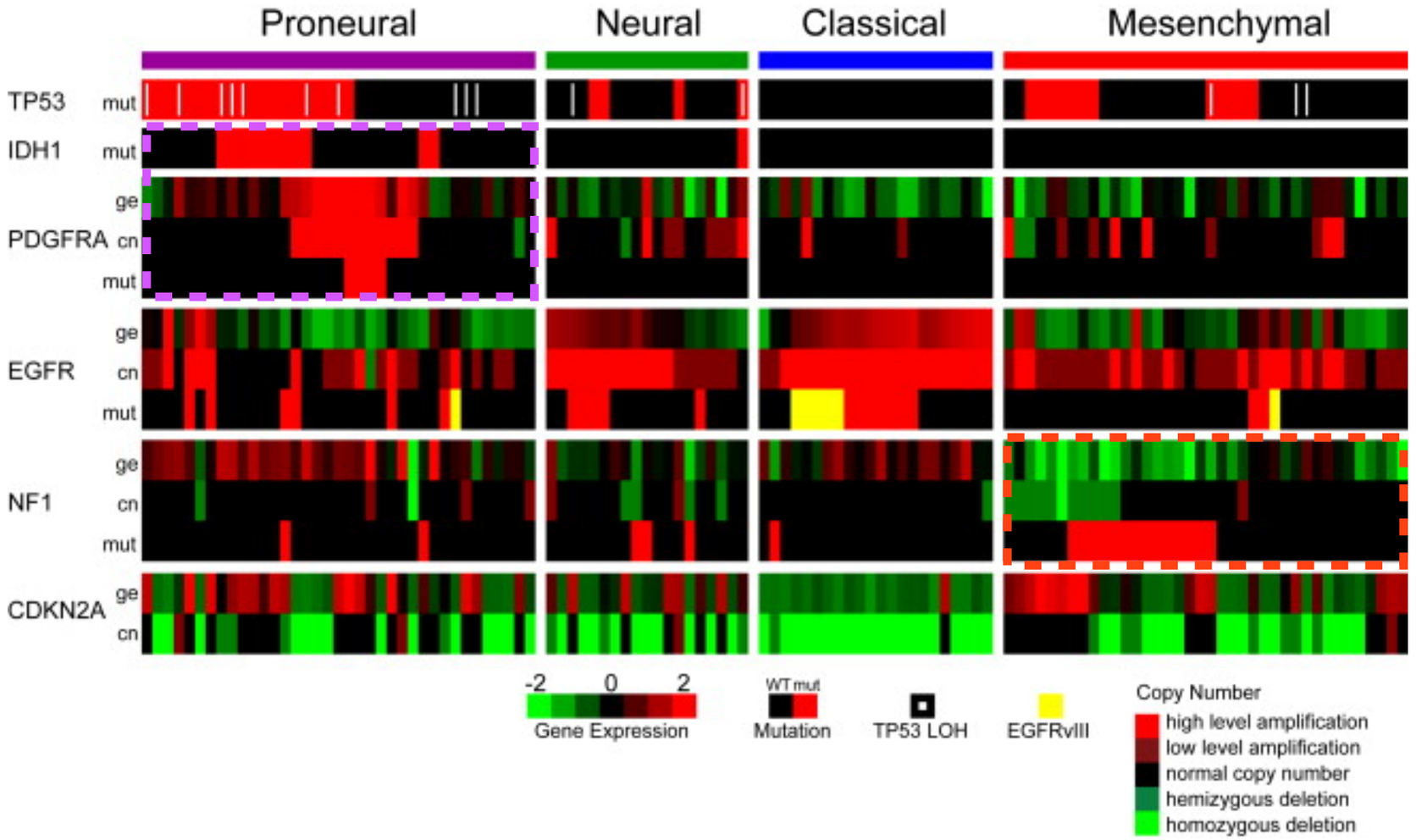
one example to describe it all



TCGA, Cancer Cell 2010

Learning from oncogenomic data

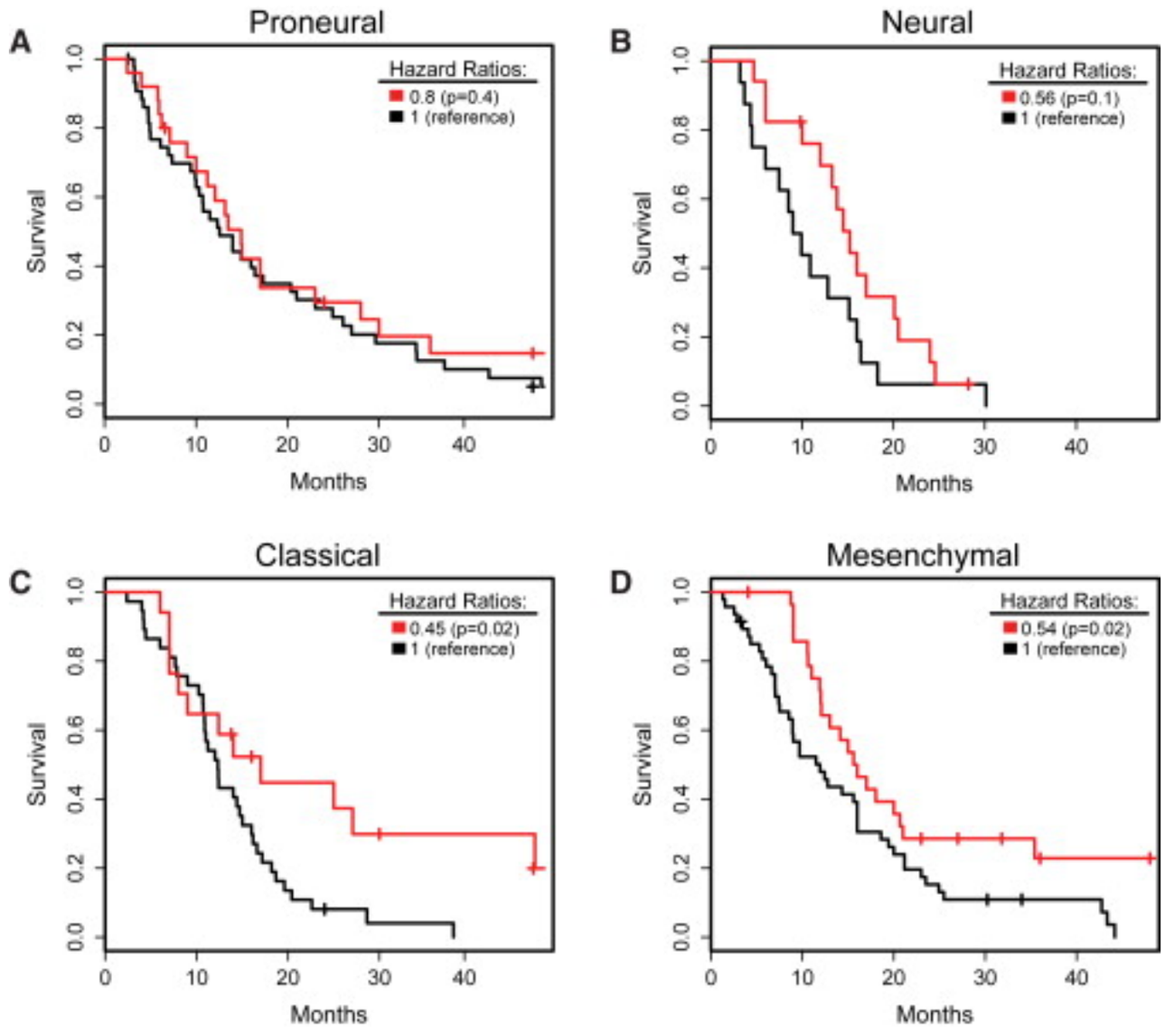
one example to describe it all



TCGA, Cancer Cell 2010

Learning from oncogenomic data

one example to describe it all



TCGA, Cancer Cell 2010

Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

model evaluation

THE IDEA: to focus on the predictive capability of a model

- metrics for performance evaluation
- methods for performance evaluation/estimation
- methods for model comparison

Learning from oncogenomic data

metrics

confusion matrix

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

TP: true positive

FN: false negative

FP: false positive

TN: true negative

the most popular

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Learning from oncogenomic data

metrics

- ACCURACY can be MISLEADING:
- 2-class problem:
  - Class 0 = 9990
  - Class 1 = 10
- if the model everything to be of Class 0, then
  - accuracy = 9990/1000 = 99.9%

Learning from oncogenomic data

metrics

confusion matrix

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

TP: true positive

FN: false negative

FP: false positive

TN: true negative

the most popular

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

cost-sensitive

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

Learning from oncogenomic data

accuracy vs. precision

high accuracy

low precision

low accuracy

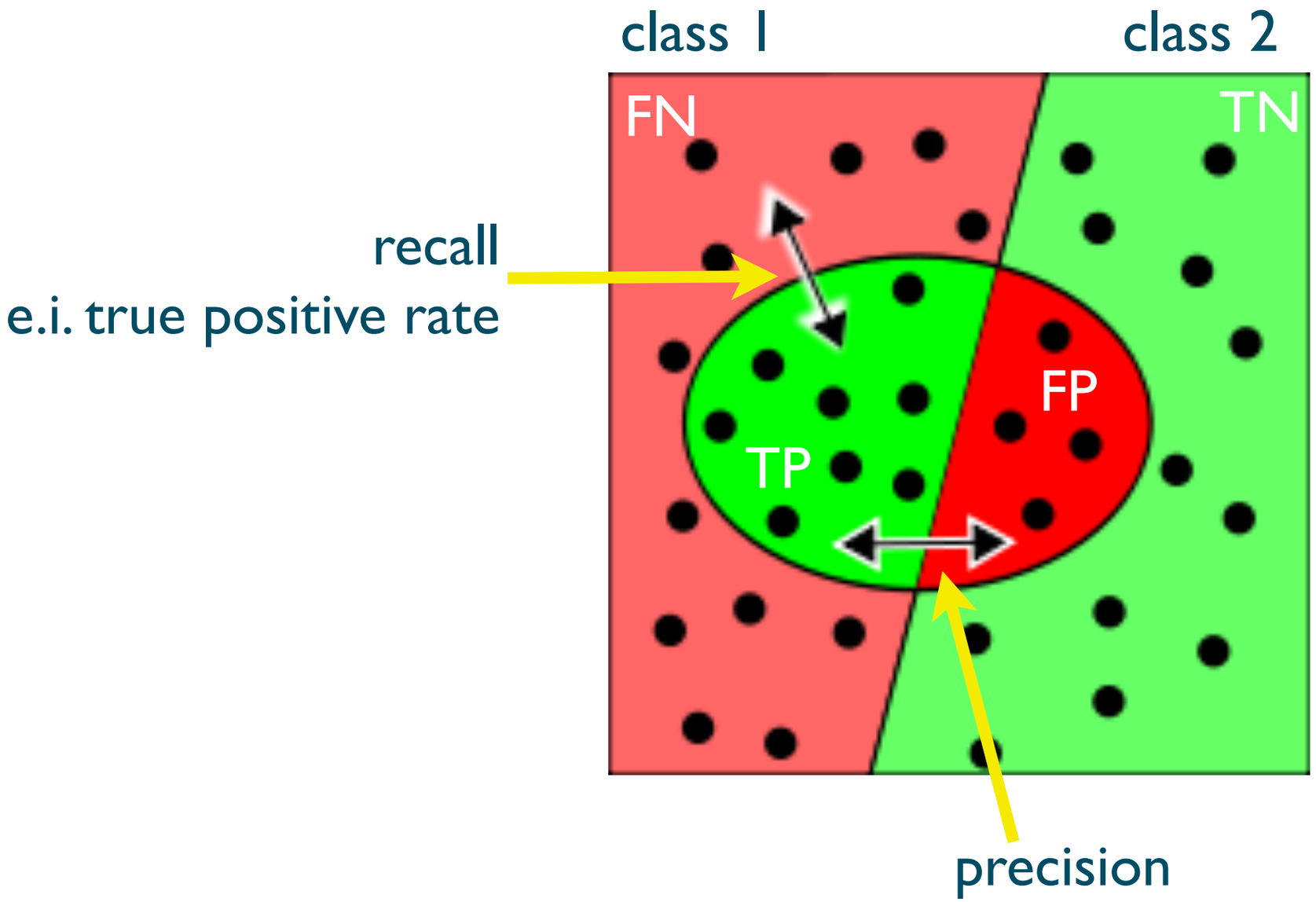
high precision

high accuracy

high precision

Learning from oncogenomic data

precision and recall



Learning from oncogenomic data

performance estimation

- holdout
  - reserve 2/3 for training and 2/3 for testing
- random subsampling
  - repeated holdout
- cross validation
  - partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining
  - leave-one-out: k=n
- stratified sampling
  - over-sampling vs under-sampling
- bootstrap
  - sampling with replacement

Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

survival analysis

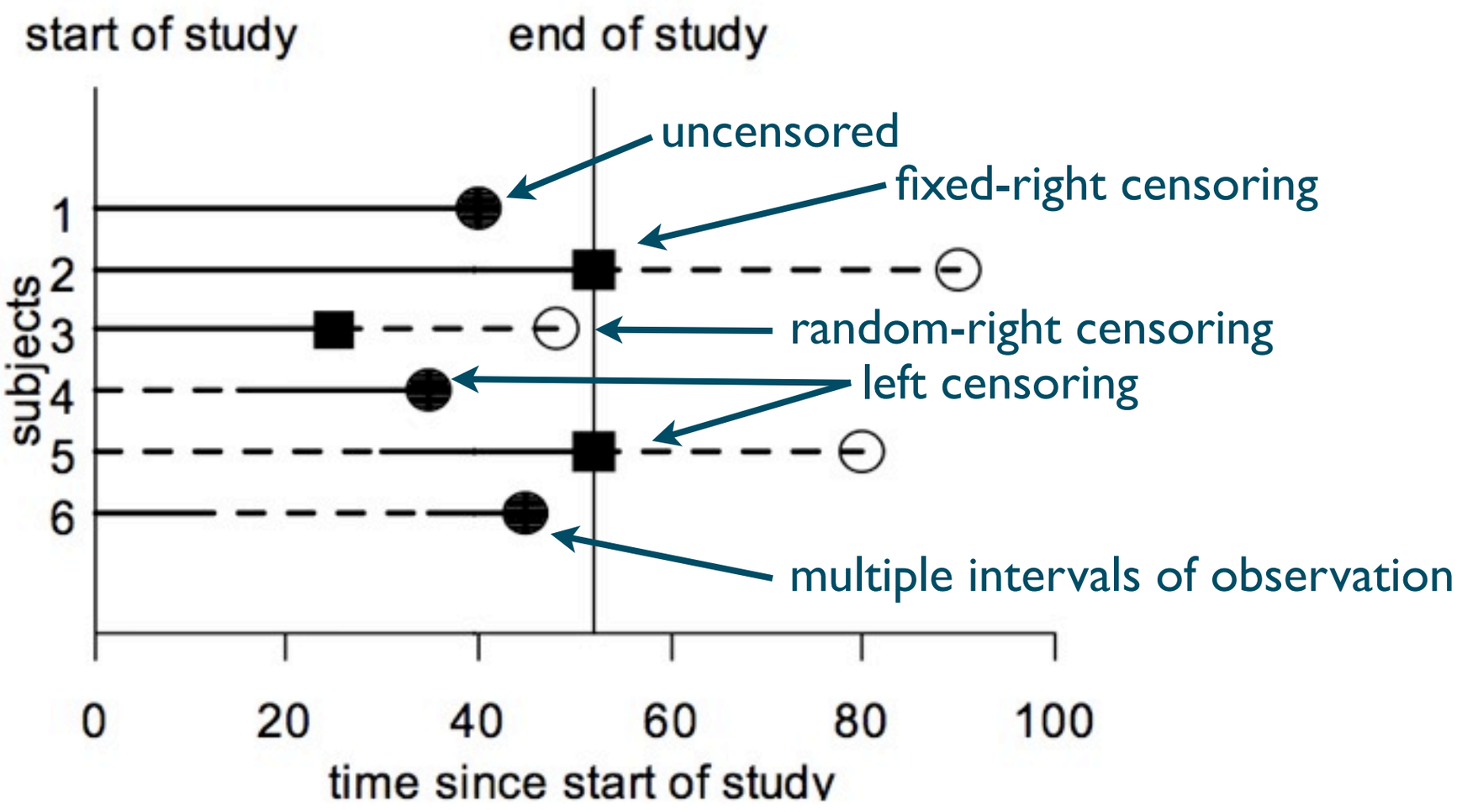
- a variety of methods for analyzing the timing of events (death, failure, recurrence, etc.)
- outline
  - the survival function and the Kaplan-Meier estimator

- ASSUMPTIONS:
- those under study are representative of *all* subjects
- at survival time *t* those subjects who are under observation at that survival time are "*at risk for an event*"
- censoring mechanism is unrelated to survival to survival time

Learning from oncogenomic data

survival analysis

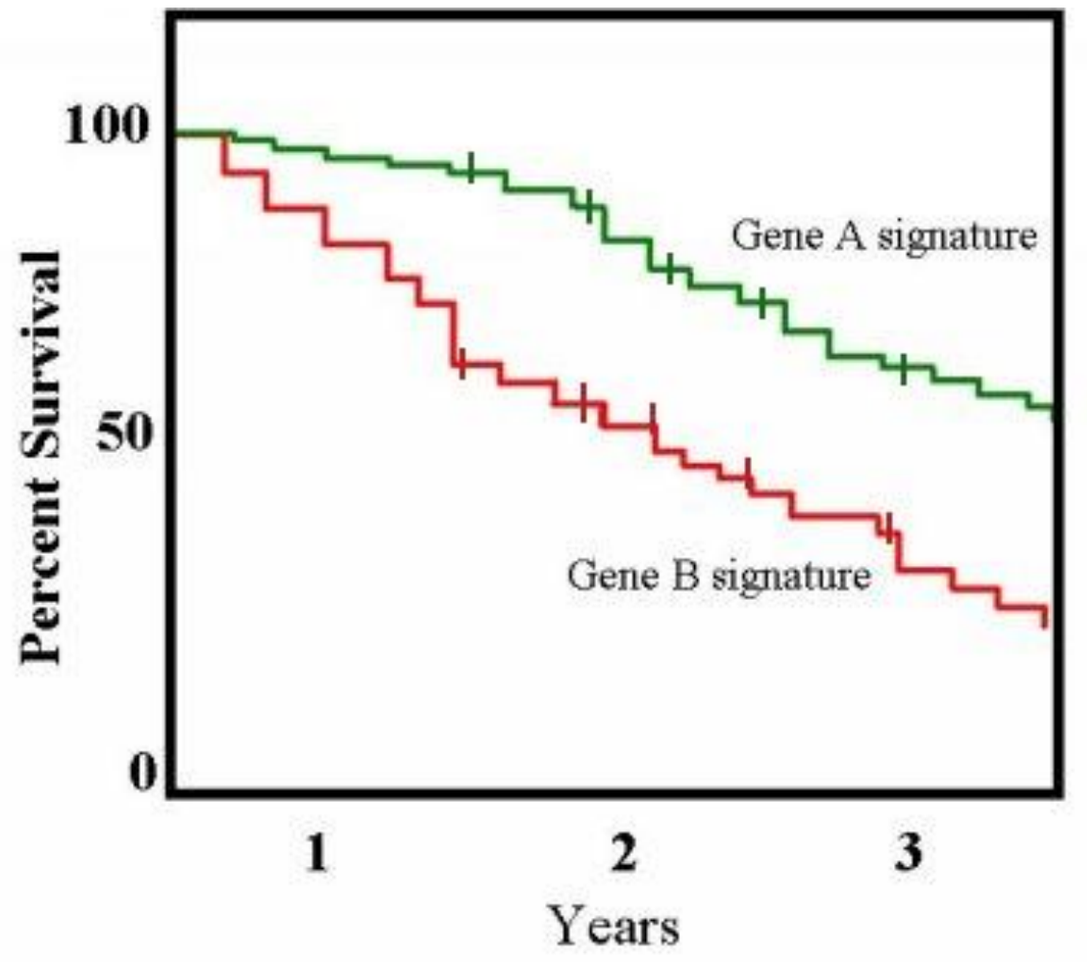
- right-censored time consists of
  - survival time
  - censored/uncensored
  - explanatory variables thought to influence survival time



Learning from oncogenomic data

survival function & Kaplan-Meier Estimator

- SURVIVAL FUNCTION: the probability that a patient will survival beyond a specified time
- KAPLAN-MEIER ESTIMATOR: estimates the survival function for life-time data
  - to measure the fraction of patients that lives or stays free of recurrence beyond a specified time
  - can handle censored data
  - ticks mark right-censoring



Learning from oncogenomic data

outline

- unsupervised approach
  - k-means and PAM
  - hierarchical clustering
- supervised approach
  - decision trees
  - k-NN
  - SVM
- model evaluation
- survival analysis
- BONUS: batch effect

Learning from oncogenomic data

batch effects

"BATCH EFFECTS are sub-groups of measurements that have qualitatively different behaviors across conditions and are unrelated to the biological or scientific variables in a study"

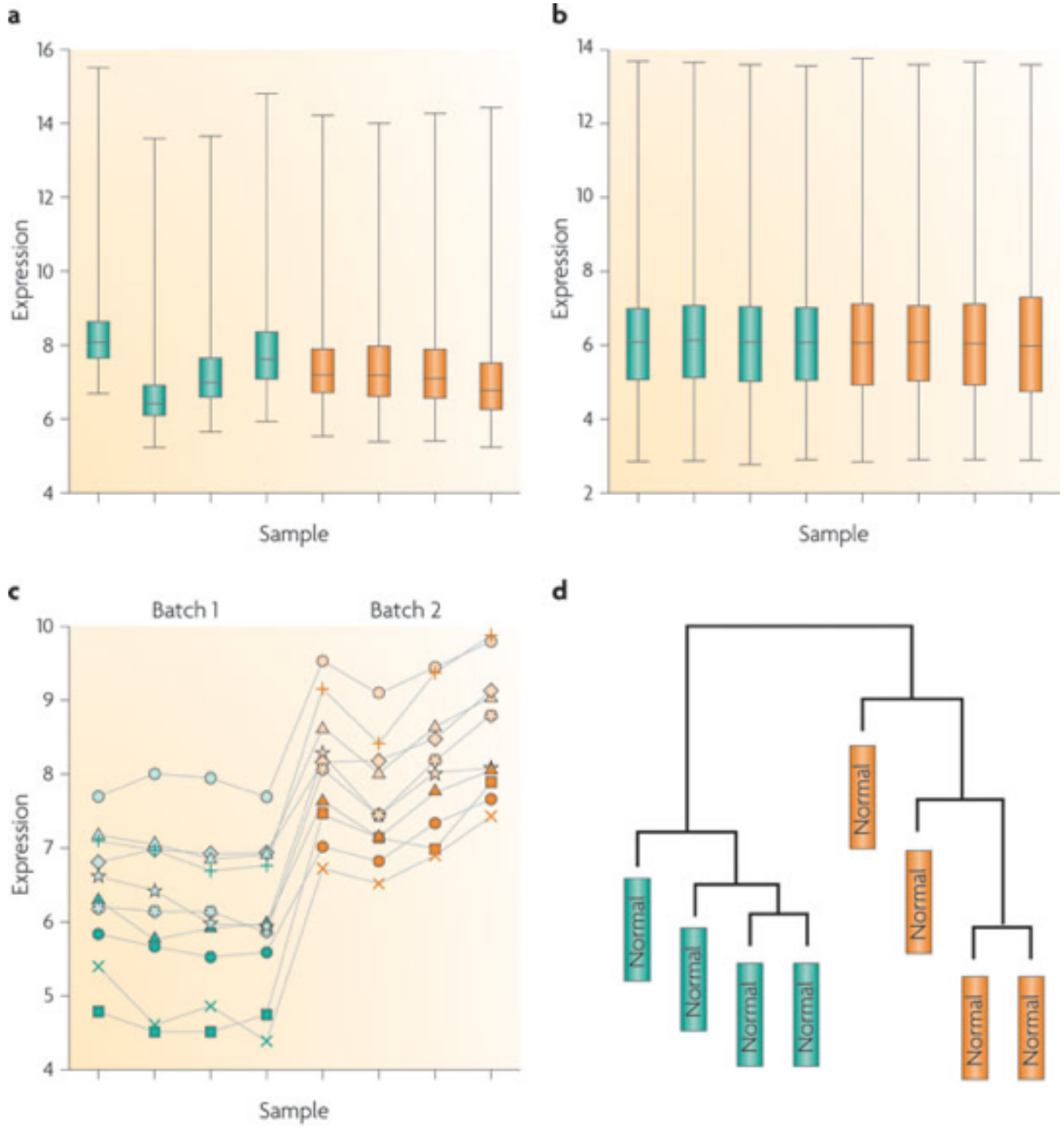
WE DON'T LIKE THEM BECAUSE

- increased variability and decreased power to detect a real biological signal
- correlation between features

Learning from oncogenomic data

batch effects

microarray expression profiling of superficial transitional cell carcinoma of bladder cancer samples with or without surrounding carcinoma in situ



Irizarry R et al., Nature Rev Gen 2010

Learning from oncogenomic data

batch effects

Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome Significant features (%)‡‡	Refs
	Surrogate‡	Confounding (%)§	Susceptible features (%)‖	Principal components rank of surrogate (correlation)¶	Principal components rank of outcome (correlation)#	Susceptible features (%)**		
Data set 1: gene expression microarray, Affymetrix (N_p = 22,283)	Date	29.7	50.5	1 (0.570)	1 (0.649)	91.6	71.9	Dyrskjot L et al., Cancer Res 2004
Data set 2: gene expression, Affymetrix (N_p = 4167)	Date	77.6	73.7	1 (0.922)	1 (0.668)	98.5	62.2	Spielman RS et al., Nature Gen 2007
Data set 3: mass spectrometry (N_p= 15,154)	Processing group	100	51.7	2 (0.344)	2 (0.344)	99.7	51.7	Petricoin EF et al., Lancet 2002
Data set 4: copy number variation, Affymetrix (N_p= 945,806)	Date	29.2	99.5	2 (0.921)	3 (0.485)	99.8	98.8	HapMap, Nature 2003
Data set 5: copy number variation, Affymetrix (N_p= 945,806)	Date	12.2	83.8	1 (0.553)	1 (0.137)	99.8	74.1	Dick DM et al., AJHG 2003
Data set 6: gene expression, Affymetrix (N_p = 22,277)	Processing group	NA	83.8	5 (0.369)	NA	97.1	NA	
Data set 7: gene expression, Agilent (N_p = 17,594)	Date	NA	62.8	2 (0.248)	NA	96.7	NA	TCGA, Nature, 2008
Data set 8: DNA methylation, Agilent (N_p = 27,578)	Processing group	NA	78.6	3 (0.381)	NA	99.8	NA	
Data set 9: DNA sequencing, Solexa (N_p = 2,886)	Date	24.2	32.1	2 (0.846)	2 (0.213)	72.7	16.9	1000 Genomes Project

Irizarry R et al., Nature Rev Gen 2010

Learning from oncogenomic data

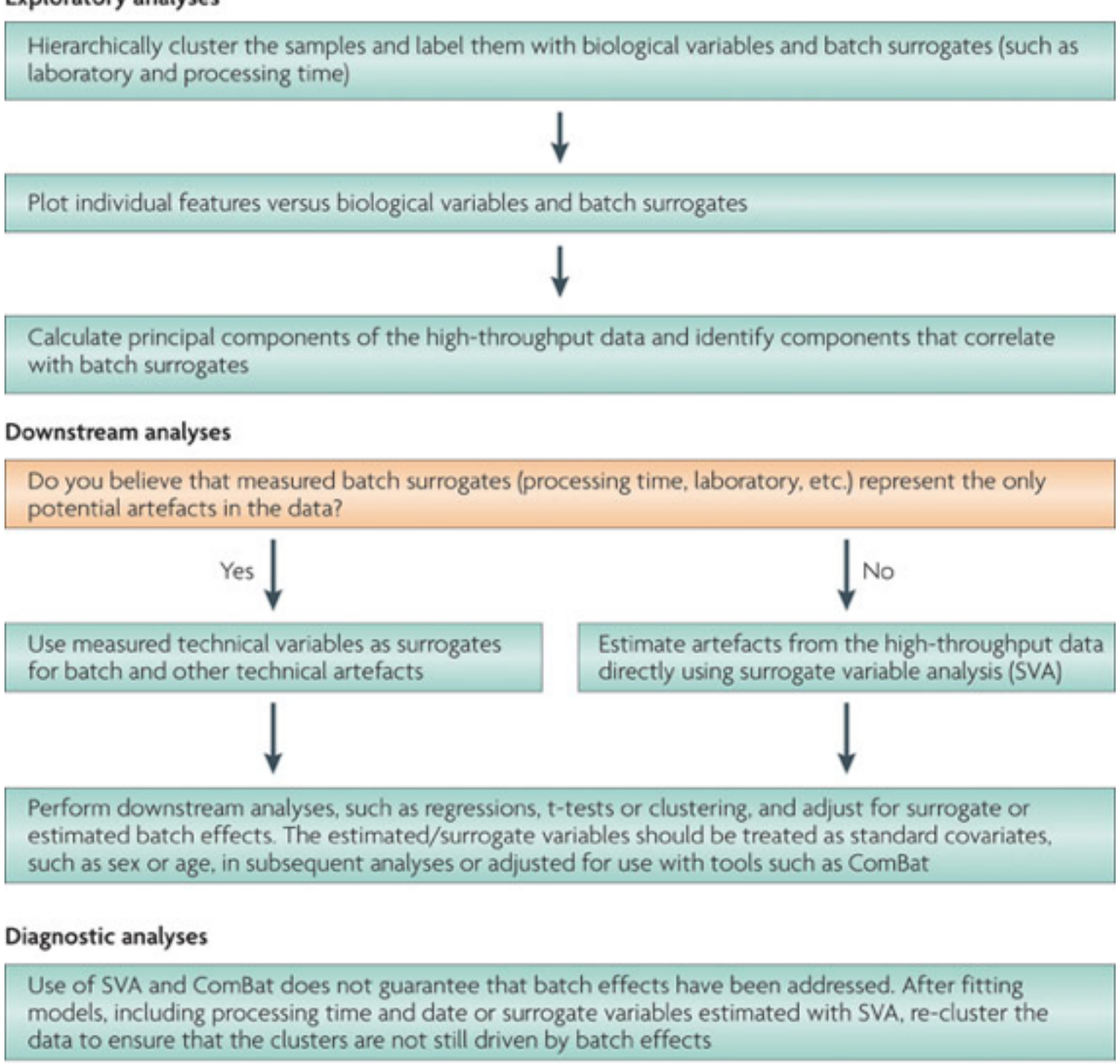
batch effects

WHAT TO DO?

- experiment design solutions
  - distribute batches and other sources of variation across biological groups
  - record information about changes in personnel, reagent, storage, etc.
- statistical solutions

Learning from oncogenomic data

batch effects



Irizarry R et al., Nature Rev Gen 2010