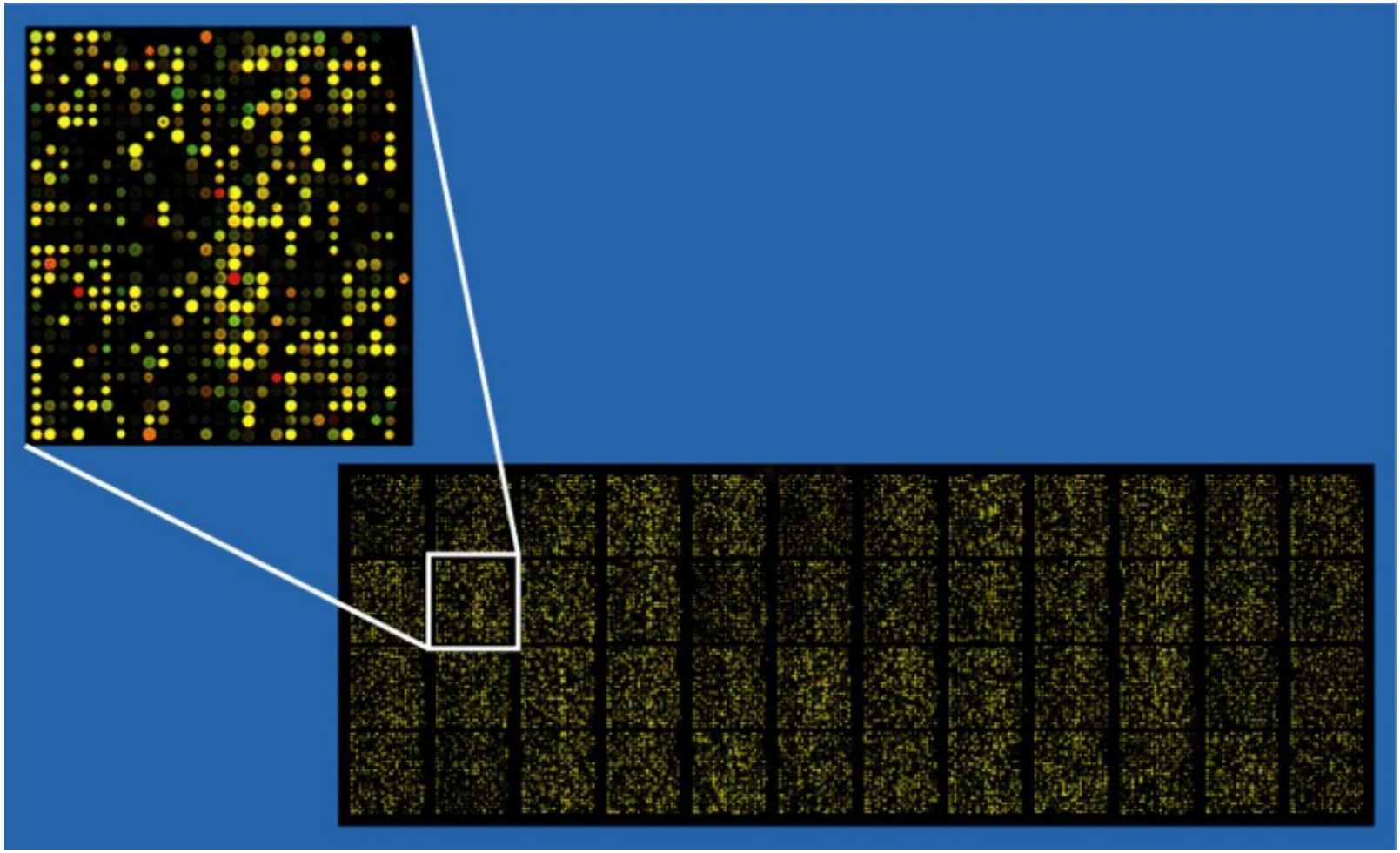
# CMSC423: Bioinformatic Algorithms, Databases and Tools Lecture 19

Microarray data analysis

# Microarray data analysis

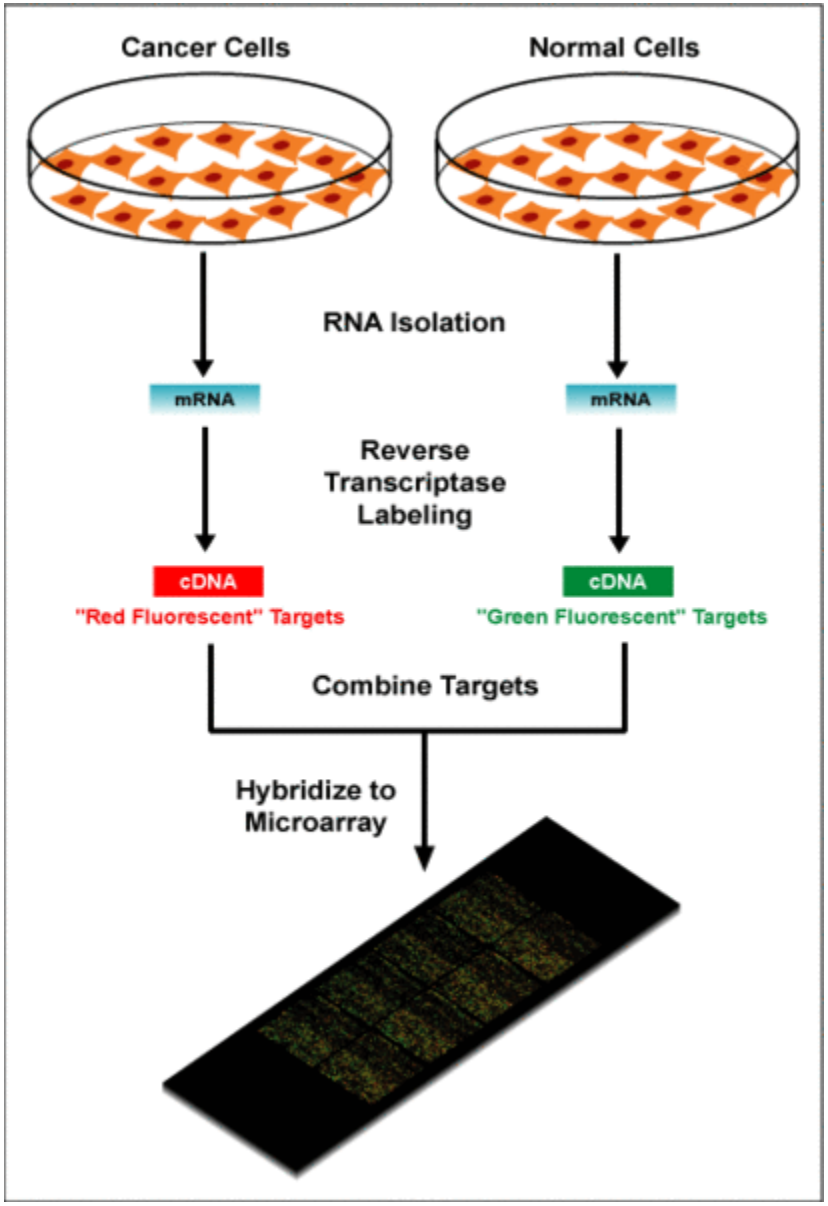


# Types of microarrays

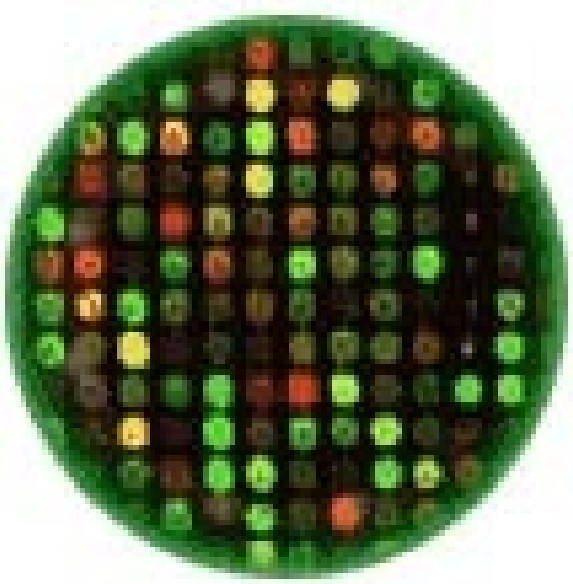
- By technology
  - Spotted
  - Affymetrix
  - Nimblegen
  - Illumina

- By information
  - cDNA (genes or parts of genes)
  - DNA (e.g. sequencing by hybridization)
  - Tiling arrays (whole genome)
  - Protein

# Typical microarray experiment




- Difference in color intensity indicate differences in gene expression levels
- Red – expressed in sample
- Green – expressed in control
- Yellow – expressed in both
- Black – expressed in neither



# Affymetrix arrays

- Instead of full-length RNA – set of 25-bp tags from gene of interest
- Also array mismatch probes – differ by 1 bp from "normal" tags.
- Gene inferred to be present if intensity of "normal" tags is > intensity of mismatch tags

- Building an Affy array – just like a micro-chip
- Individual tags grow by 1bp only if exposed to light

# Typical data analysis process

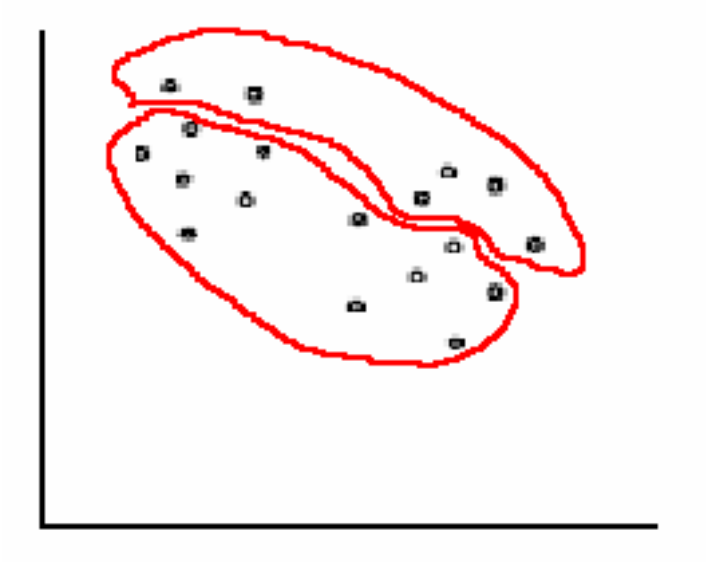
- Image analysis
  - find spots
  - find errors (air bubbles, fingerprints, smears, etc.)
- Normalization
  - make sure total intensity for green and red is the same (otherwise cannot compare intensities)
- Clustering
  - which genes have similar expression?
  - which genes are expressed similarly during a disease?
  - which genes have similar expression patterns over time (time-course experiments)?

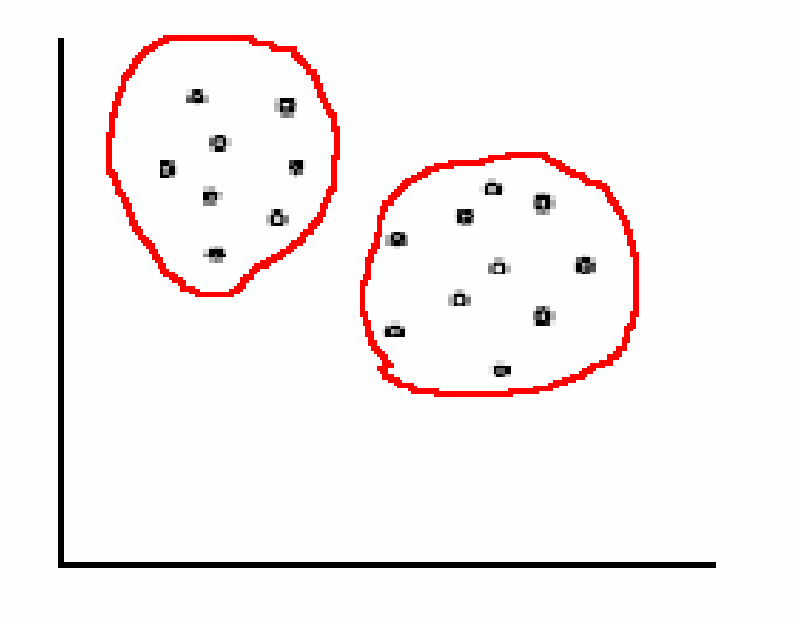
# Data clustering

- Agglomerative
  - Start with single observations
  - Group similar observations into the same cluster
- Divisive
  - All datapoints start in the same cluster
  - Iteratively divide cluster until you find good clustering
- Hierarchical
  - Build a tree – leaves are datapoints, internal nodes represent clusters

# Measures of goodness of clustering

- Homogeneity
  - All points in a cluster must be similar
- Separation
  - Points in different clusters are disimilar





# Microarray clustering

- For each gene can be viewed as an array of numbers
  - expression of gene at different time-points
  - expression of gene in different conditions (normal, variants of a disease, etc.)
- Each time-point or tissue sample can also be viewed as an array of numbers
  - expression levels for all genes

- Basic idea: cluster genes and/or samples to highlight genes involved in disease

# Hierarchical clustering

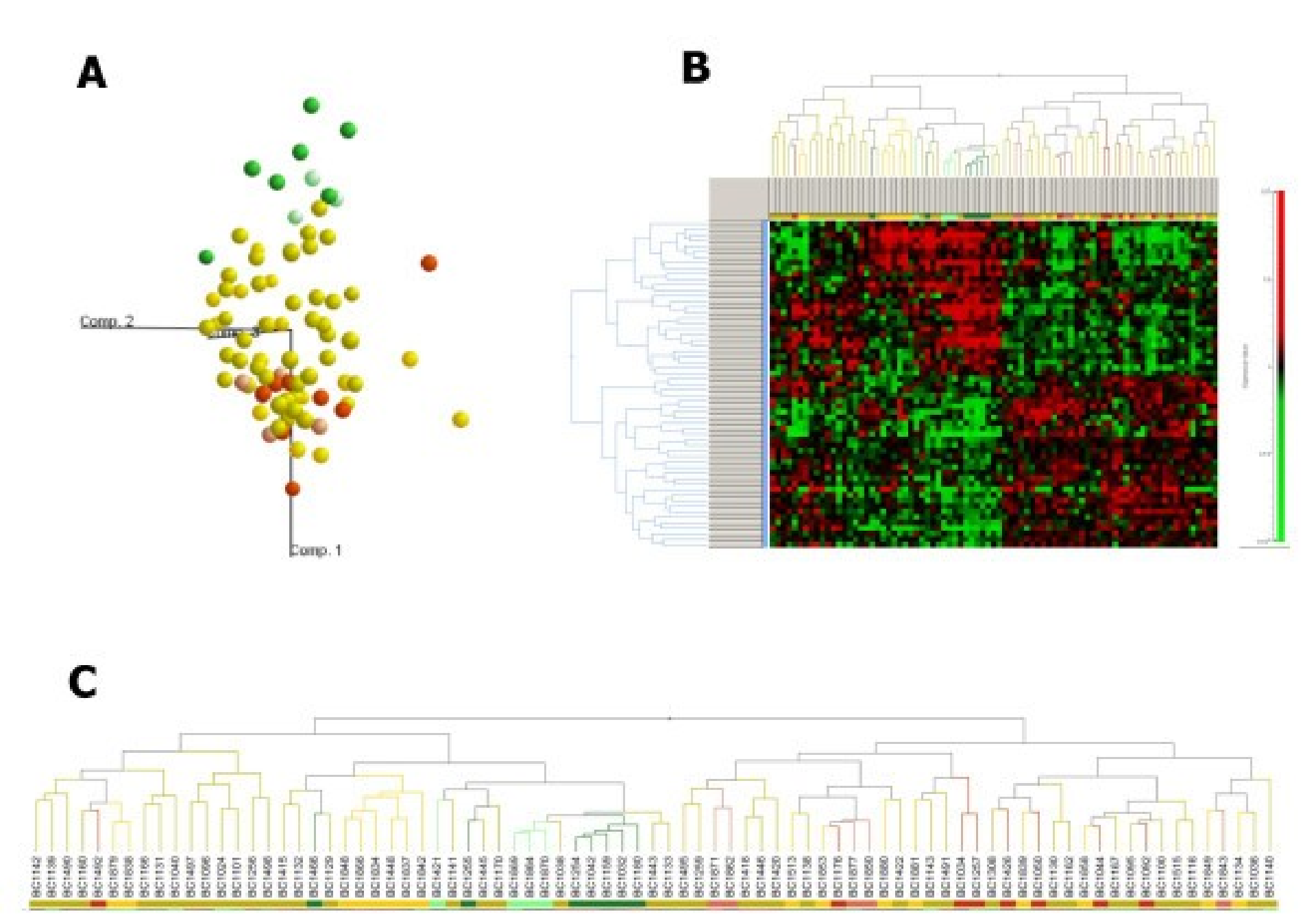
- UPGMA (remember from phylogenetic trees?)
  - compute distance between genes (e.g. euclidean distance of expression vectors)
  - join most similar genes
  - repeat
  - Key element – compute distance between a gene and a cluster, or between two clusters – average distance between all genes in the two clusters – also called “average neighbor”
- Furthest neighbor
  - distance between two clusters = largest distance between all genes
- Nearest neighbor

# Hierarchical clustering...cont

- Irrespective of distance choice, algorithm is the same

  1. compute inter-gene/cluster distances
  2. join together pair of genes/clusters with smallest distance
  3. recompute distances to include the newly created cluster
  4. repeat until all points in one cluster

- Output of program is a tree
- Cluster sets – defined by “cut” nodes – any subset of internal tree nodes defines a set of clusters – the sets of leaves in the corresponding subtrees


A
Comp. 2
Comp. 1
B
C

# k-means clustering

- Split data into exactly k clusters
- Basic algorithm:
  - Create k arbitrary clusters - pick k points as cluster centers and assign each other point to the closest center
  - Re-compute the center of each cluster
  - Re-assign points to clusters
  - Repeat

- Another approach: pick a point at and see if moving it to a different cluster will improve the quality of the overall solution.  Repeat!

# K-means clustering... K=2