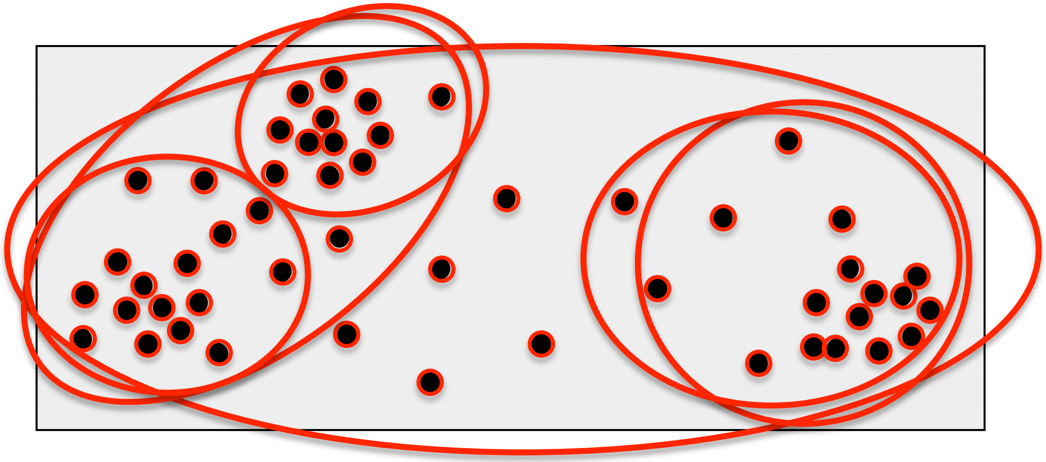
Clustering

Clustering/Class Discovery

- Aim: Partition data (e.g. genes or samples) into sub-groups (clusters), such that points of the same cluster are "more similar".
- Example:
  How many clusters?

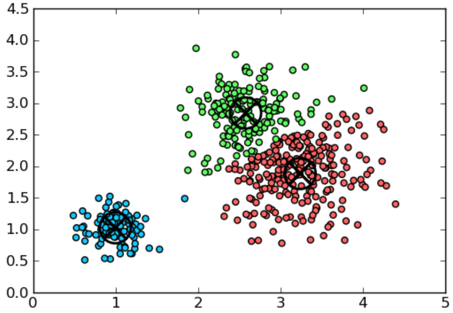


- One has to choose:
  - Clustering method
  - Similarity/distance measure
  - Evaluate clusters

Clustering in GenePattern

- Representative based:

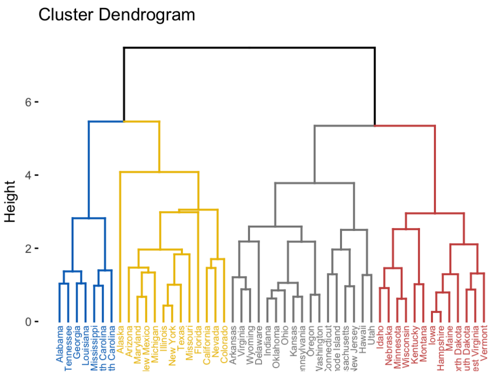
  Find representatives/centroids of the dataset
  - *K*-means
  - Self Organizing Maps (SOM)



K-Means Clustering

- Bottom-up (Agglomerative)

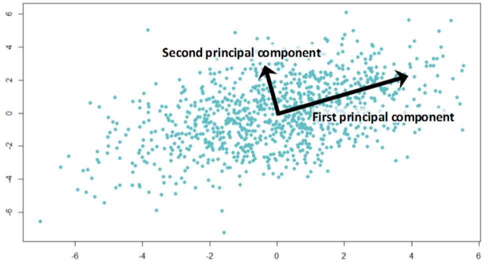
  Create an ordering of the data by closeness
  - Hierarchical clustering



Hierarchical Clustering

- Clustering-like:

  Reduce the data to a smaller number of dimensions containing the majority of the information content
  - NMF (Non-Negative Matrix Factorization)
  - PCA (Principal Components Analysis)

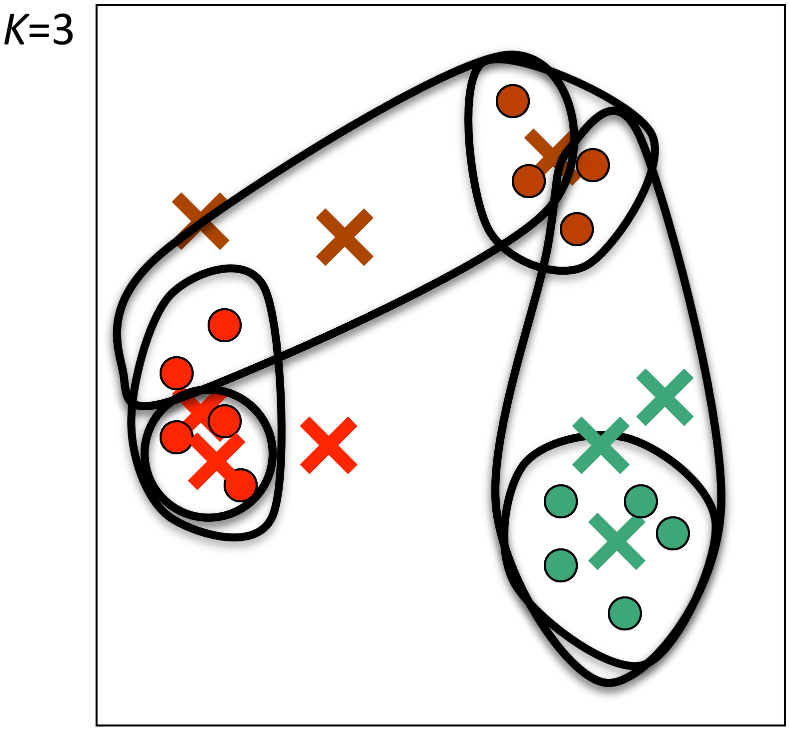


Principal Components Analysis

K-means Clustering

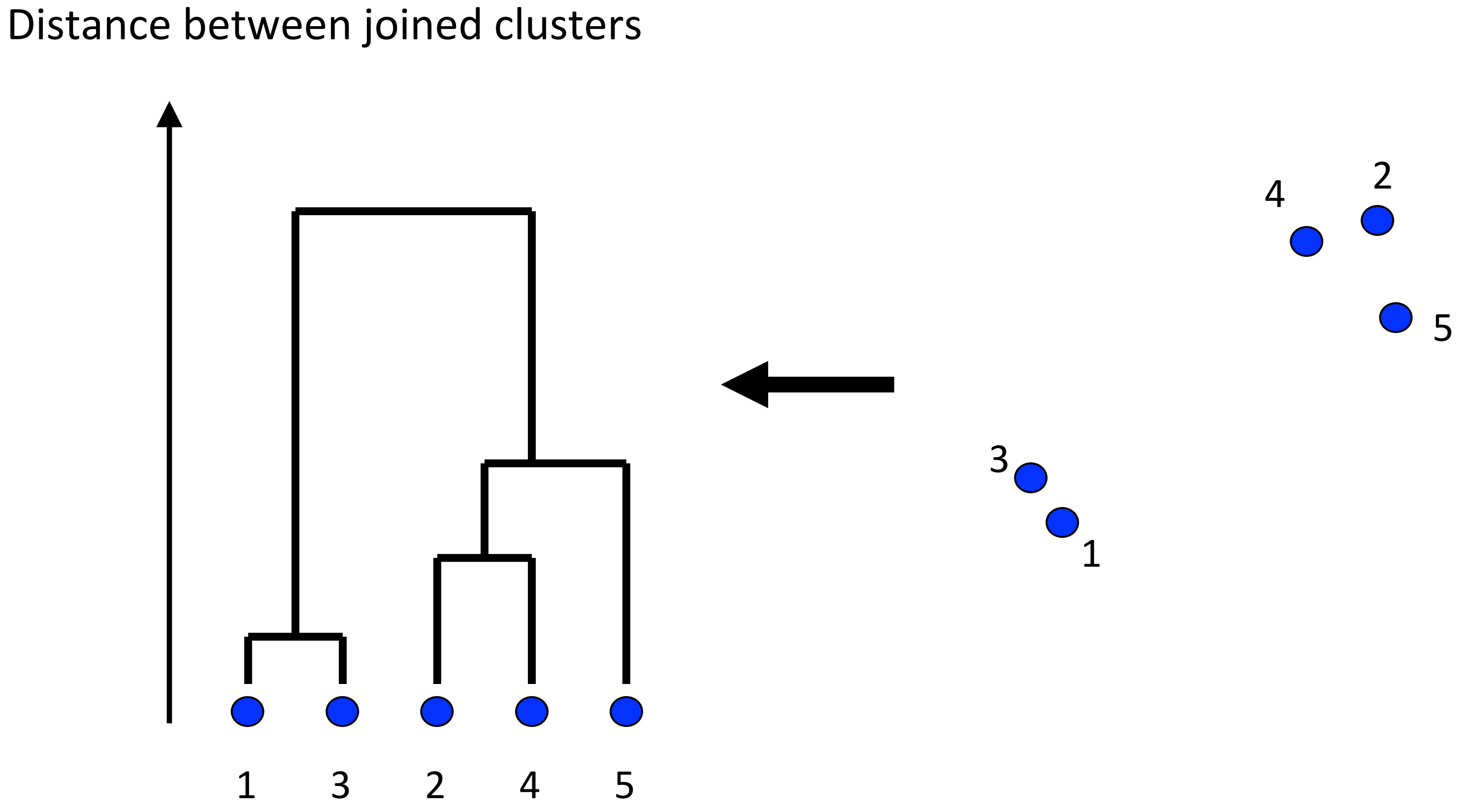
- Initialize centroids at random positions
- Iterate:
  - Assign each data point to its closest centroid
  - Move centroids to center of assigned points
- Stop when converged

- Guaranteed to reach a local minimum



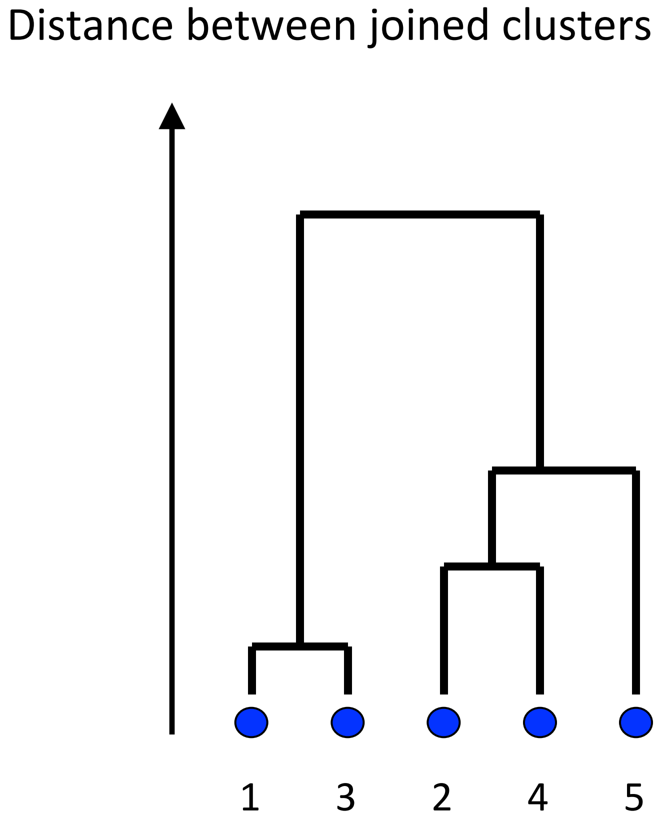
Iteration = 2

Hierarchical Clustering

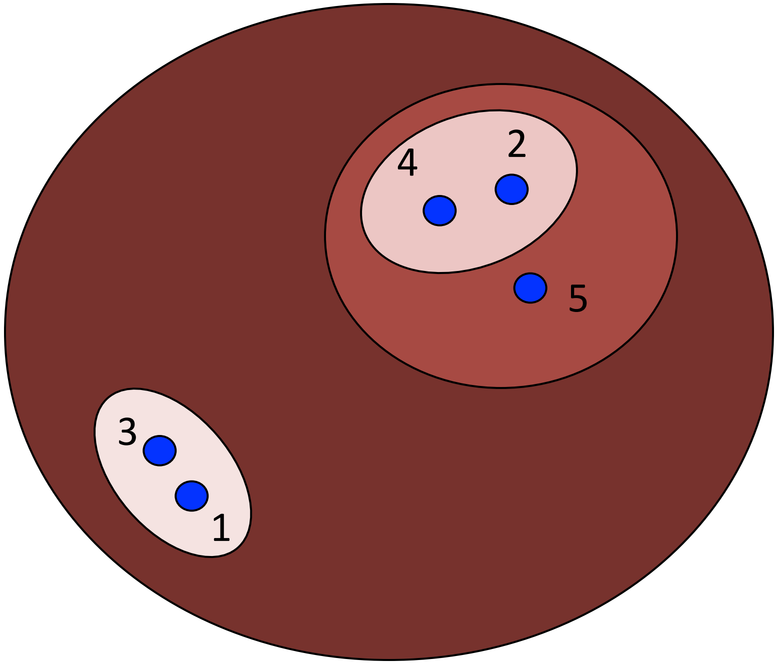


Dendrogram

Hierarchical Clustering



Dendrogram



Linkage is the method for linking clusters based on the distance between them.

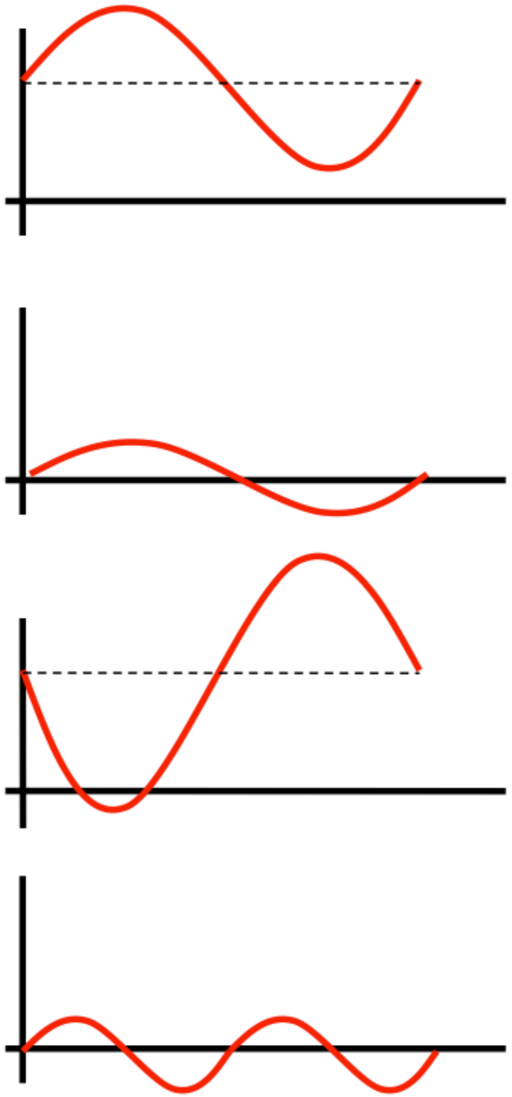
Average Linkage: average distance between all pairs

Complete Linkage: farthest distance between all pairs

Single Linkage: closest distance between all pairs

Distance metrics: Pearson and Euclidean

- Pearson correlation
  - Measures linear dependence between genes
  - “General purpose” distance metric
  - Invariant to scaling
  - Invariant to addition by a constant
- Euclidean distance
  - Measures standard distance between two points
  - Sensitive to scaling
  - Appropriate for row-normalized data

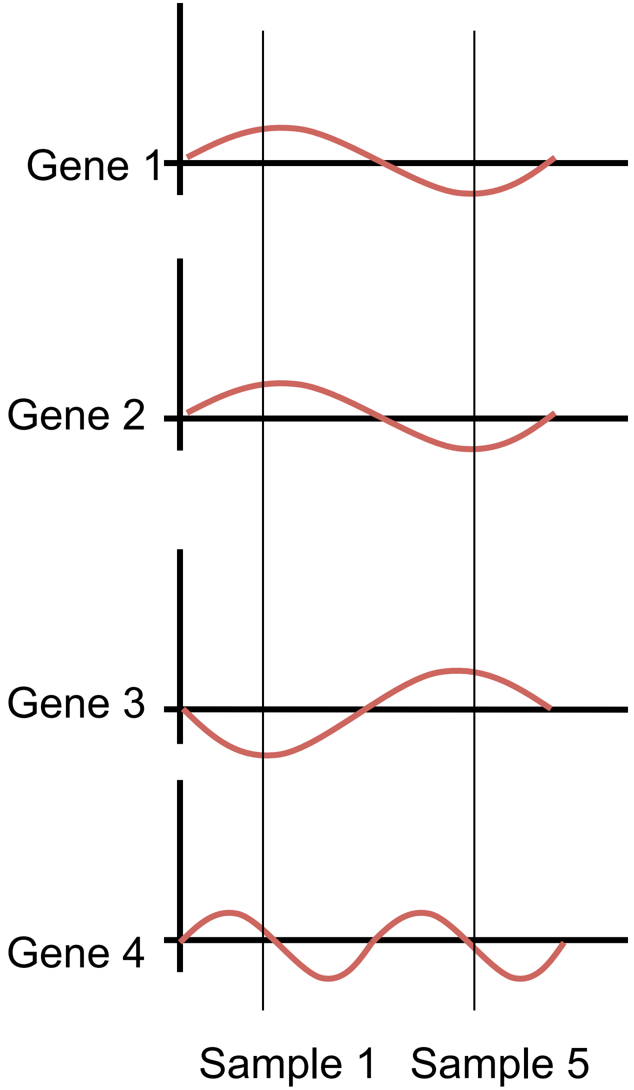


Reasonable Distance Measure

Euclidean distance on samples and genes on row-centered and normalized data.

Genes: Close -> Correlated

Samples: Similar profile giving Gene 1 and 2 a similar contribution to the distance between sample 1 and 5



Different Distance Measures

Different distance measures can reveal different structures

Normal/Tumor

Center and normalize rows

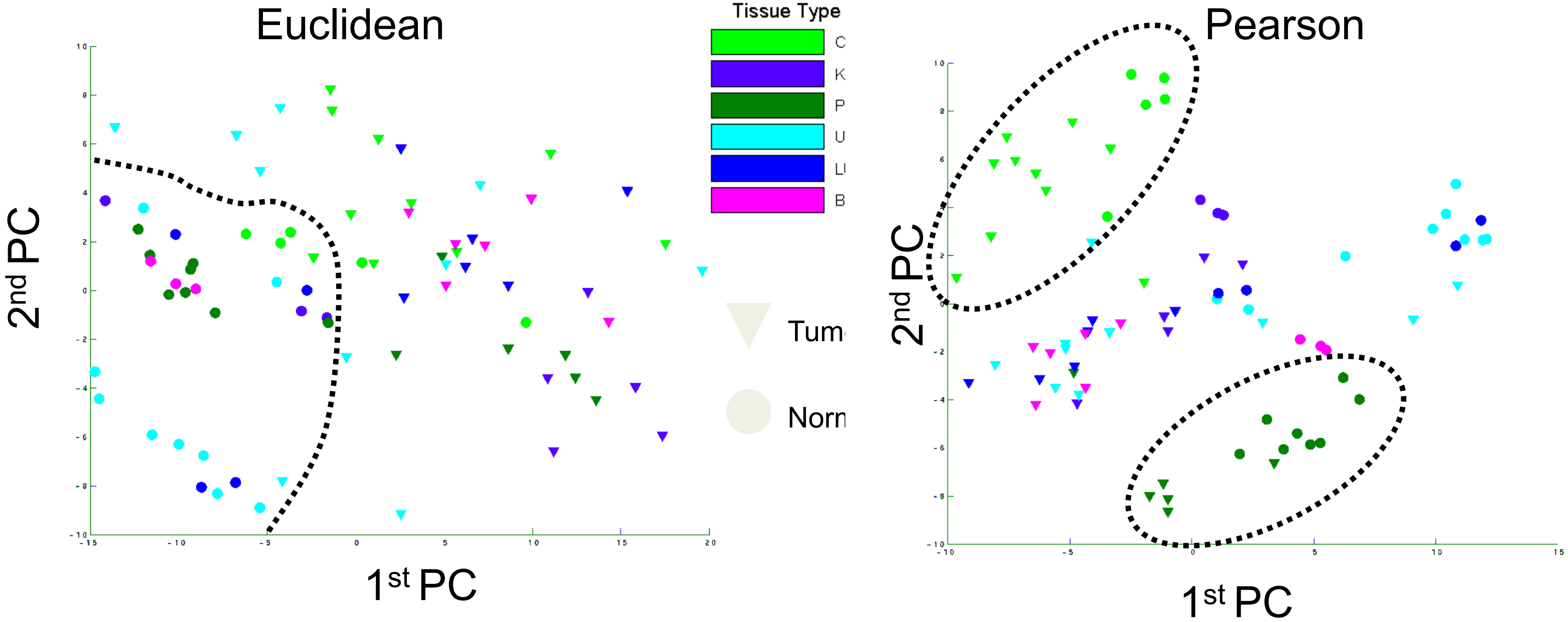
Close samples have similar profiles

Tissue type

Center and normalize rows

Center and normalize columns

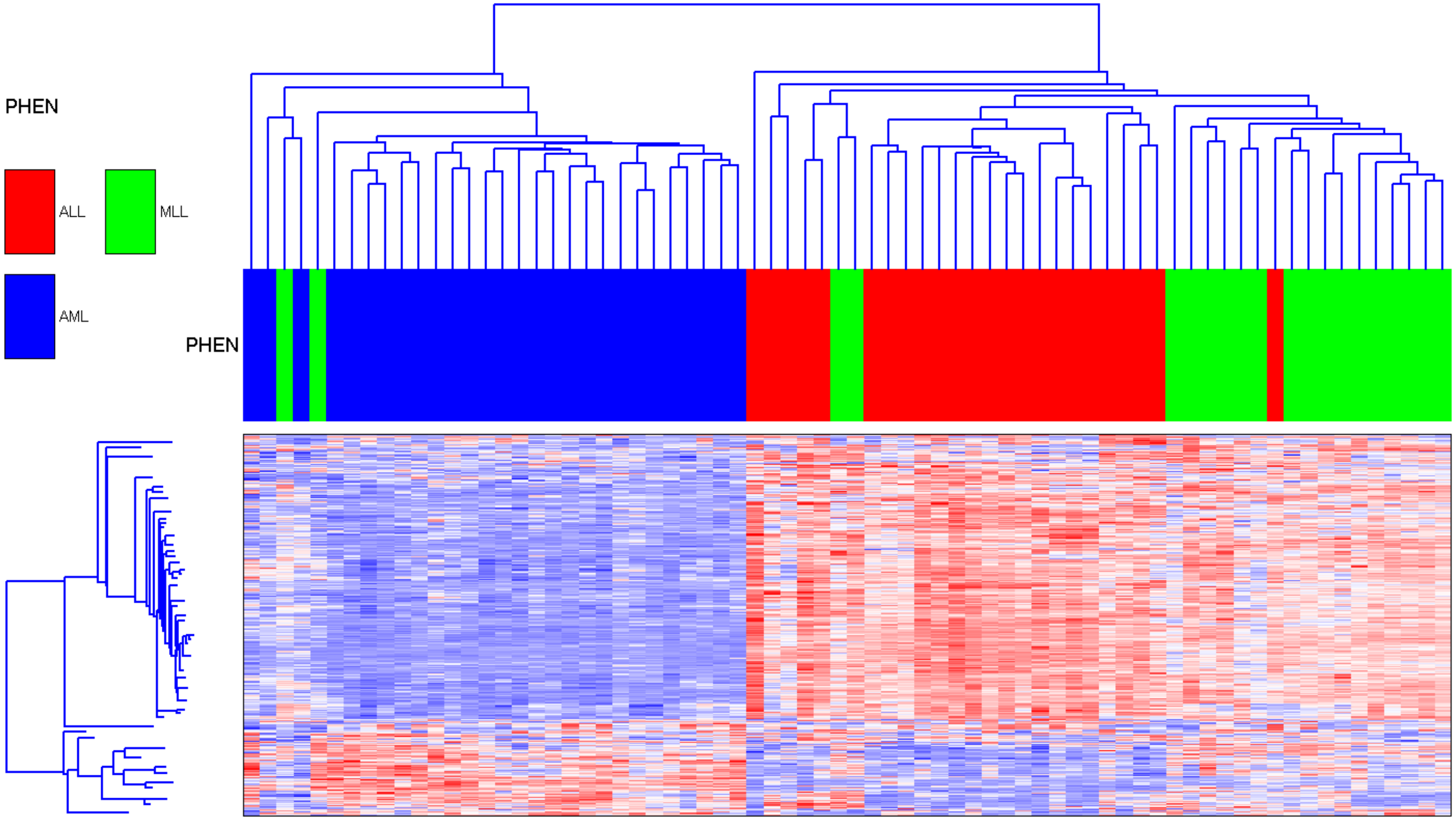
Close samples have correlated profiles



Data from Lu et al. *Nature*, 2005

Average Linkage

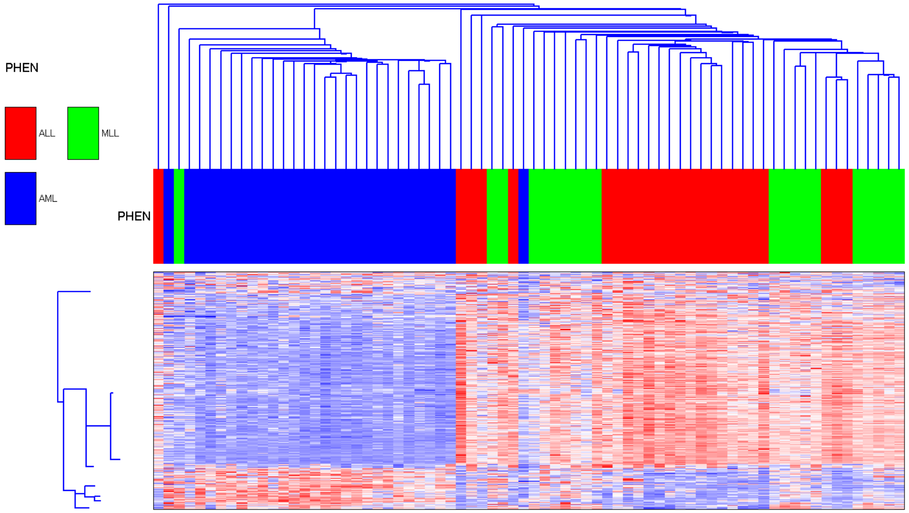
Leukemia samples and genes



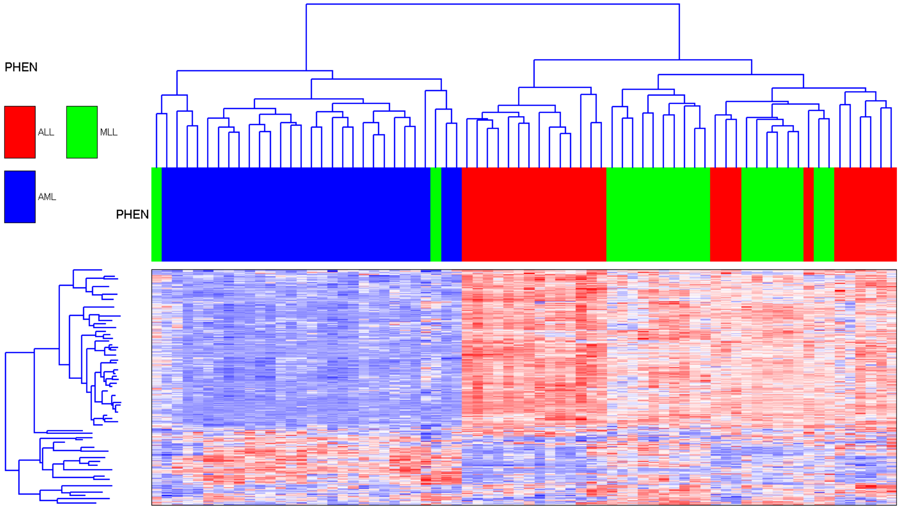
Single and Complete Linkage

Leukemia samples and genes

Single-linkage



Complete-linkage



Similarity/Distance Measures

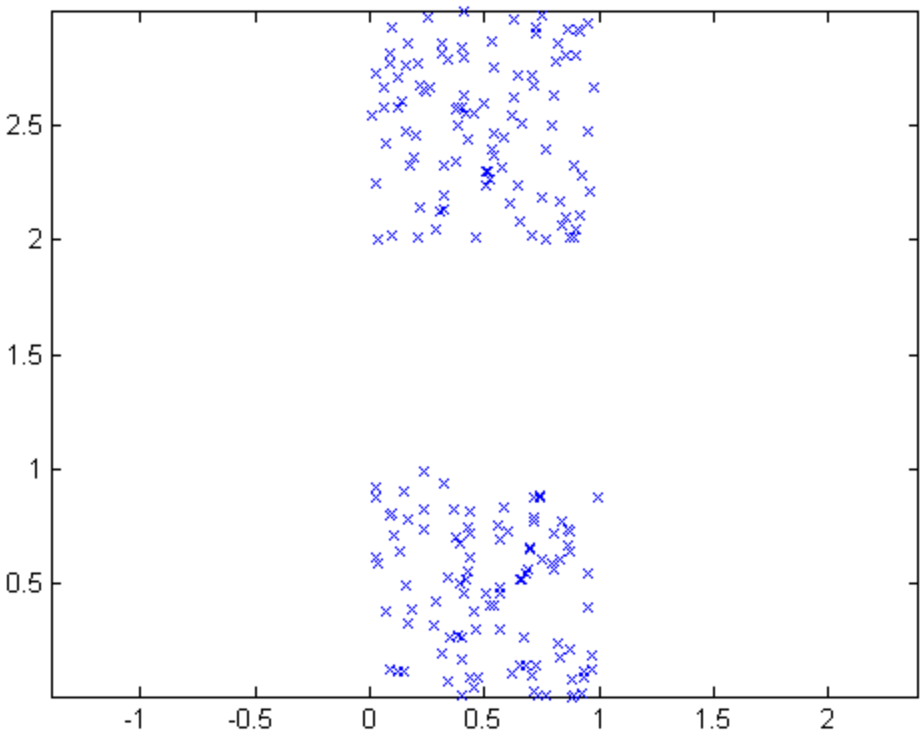
Decide: which samples/genes should be clustered together

- **Euclidean**: the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula
- **Pearson correlation** - a parametric measure of the strength of linear dependence between two variables.
- **Absolute Pearson correlation** - the absolute value of the Pearson correlation
- **Spearman rank correlation** - a non-parametric measure of independence between two variables
- **Uncentered correlation** - same as Pearson but assumes the mean is 0
- **Absolute uncentered correlation** - the absolute value of the uncentered correlation
- **Kendall’s tau** - a non-parametric similarity measure used to measure the degree of correspondence between two rankings
- **City-block/Manhattan** - the distance that would be traveled to get from one point to the other if a grid-like path is followed

Pitfalls in Clustering

- Elongated clusters
- Filament
- Clusters of different sizes

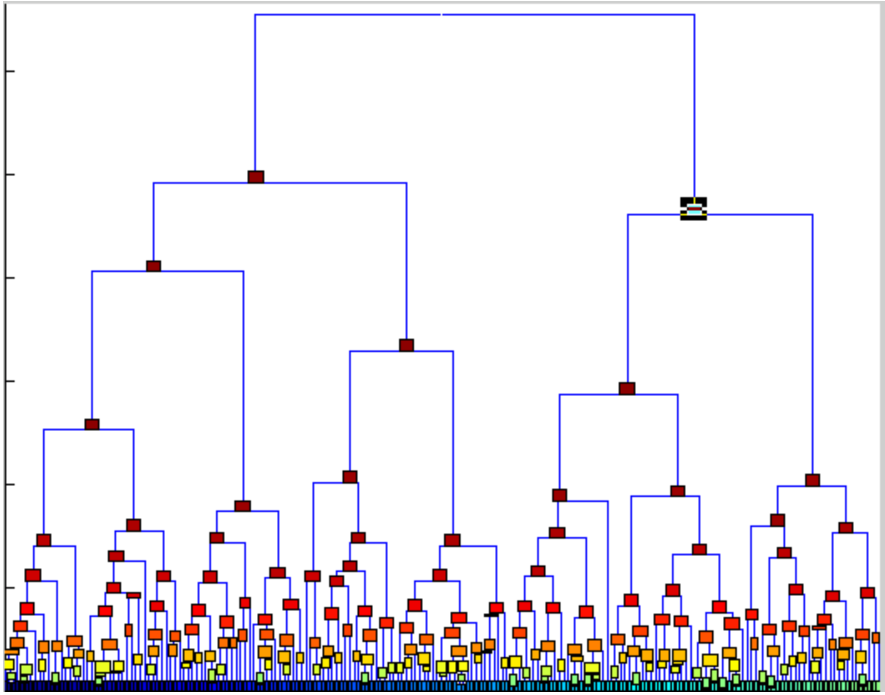
Compact Separated Clusters

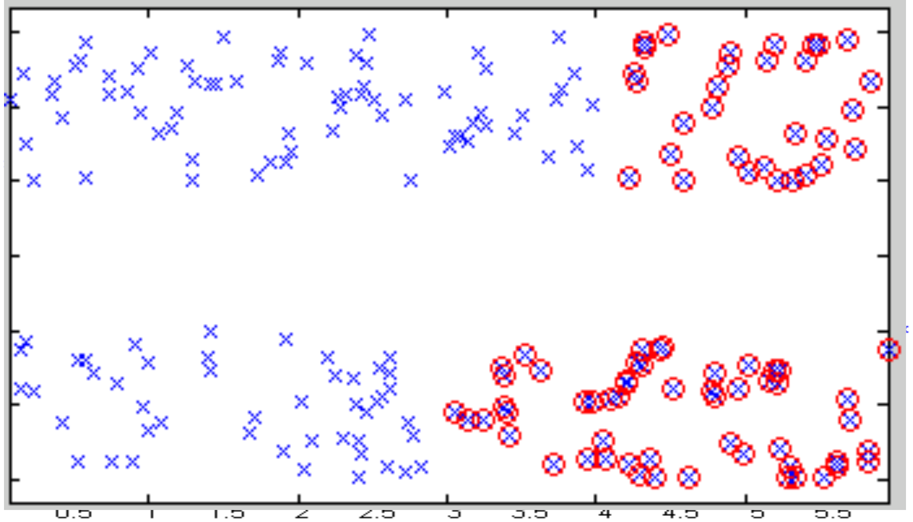


- All methods work

Adapted from E. Domany

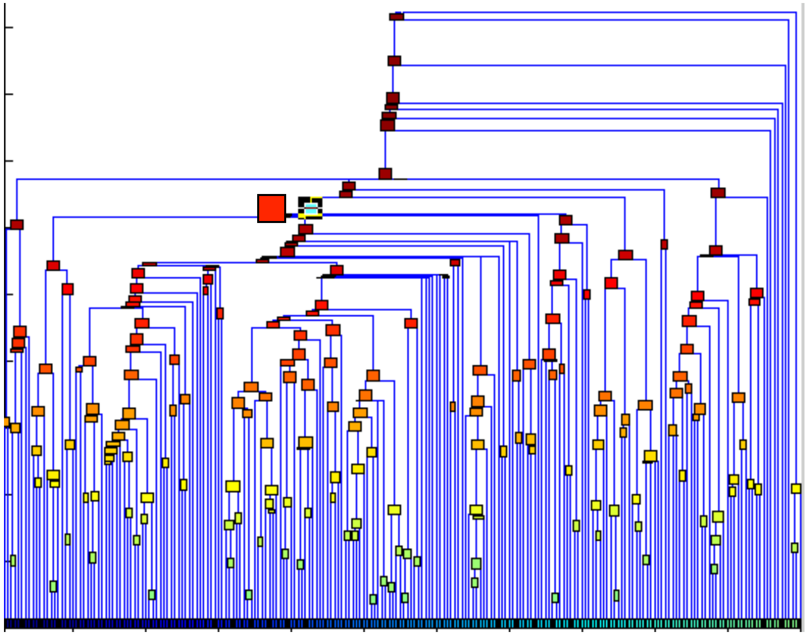
Elongated Clusters

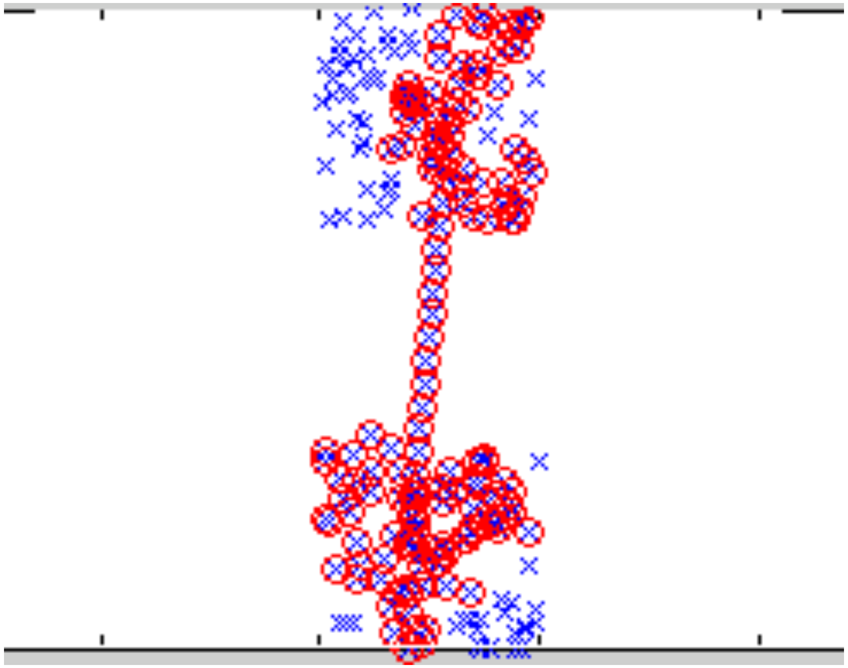




- Single linkage succeeds to partition
- Average linkage fails

Filament

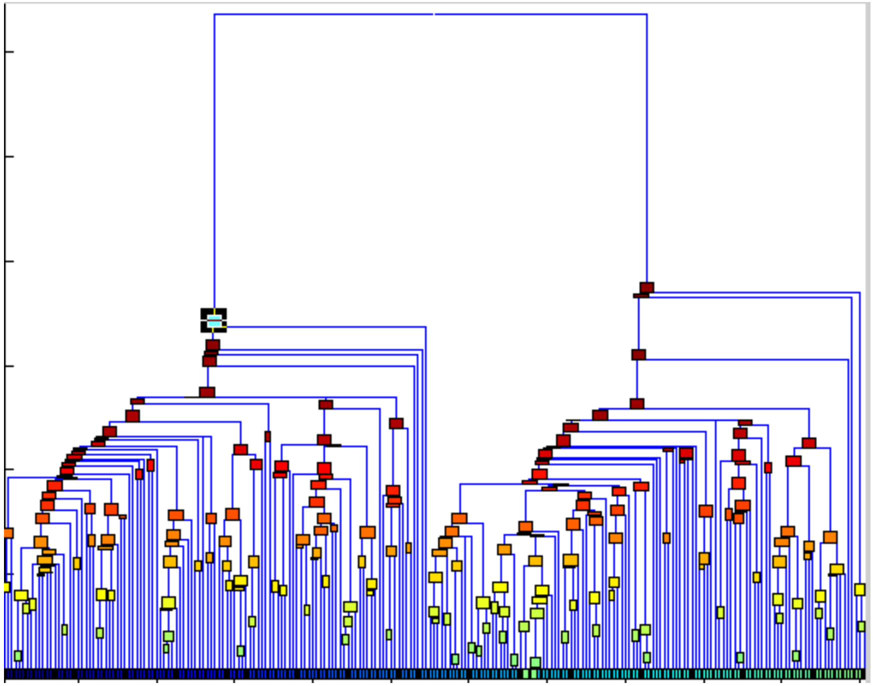


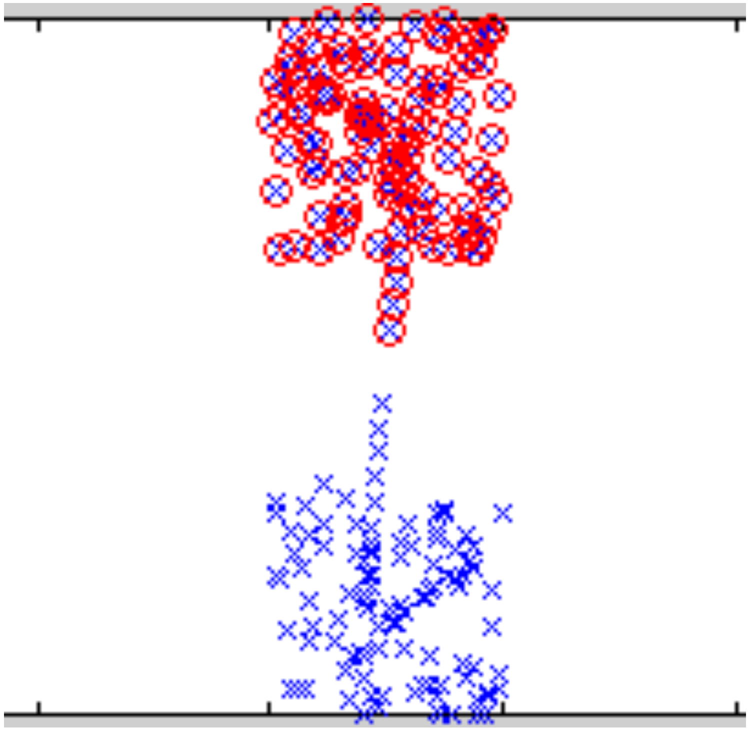


- Single linkage not robust

Adapted from E. Domany

Filament with Point Removed



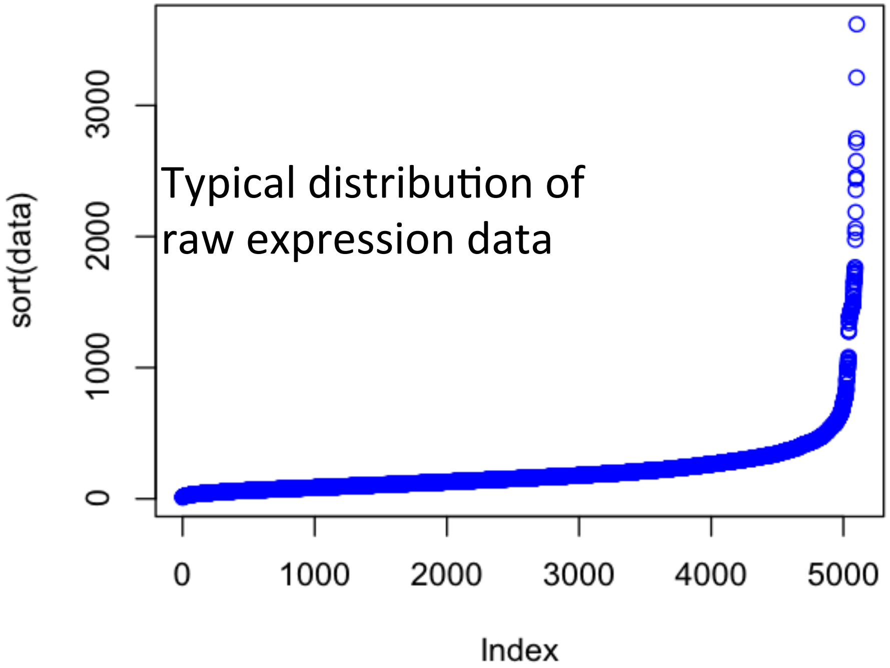


- Single linkage not robust

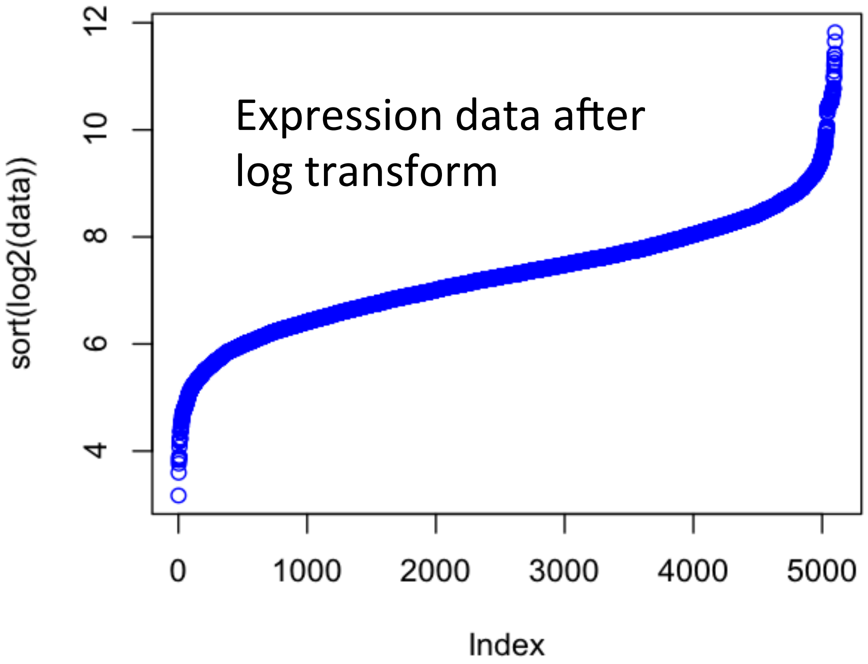
Adapted from E. Domany

Data Preparation

- Row Normalization
  - Makes genes expressed at different levels comparable to each other
- Filtering
  - Removes lowly-expressed (noisy) and invariant genes
- Log transform
  - Removes outliers by scaling distribution



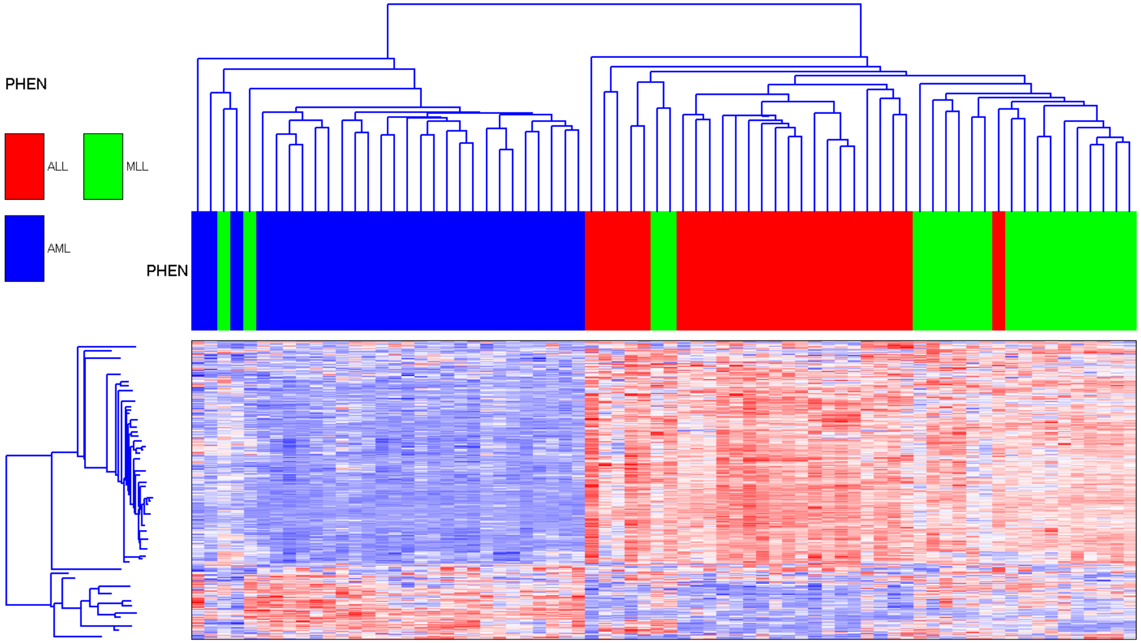
Typical distribution of raw expression data



Expression data after log transform

Two-way Clustering

- Two independent cluster analyses on genes and samples used to reorder the data (two-way clustering):



Clustering Exercise

2018-01-23-15_08_CCMI_Hierarchical Clustering – RNASeq