Correlation Networks and Inference of Gene Regulatory Networks
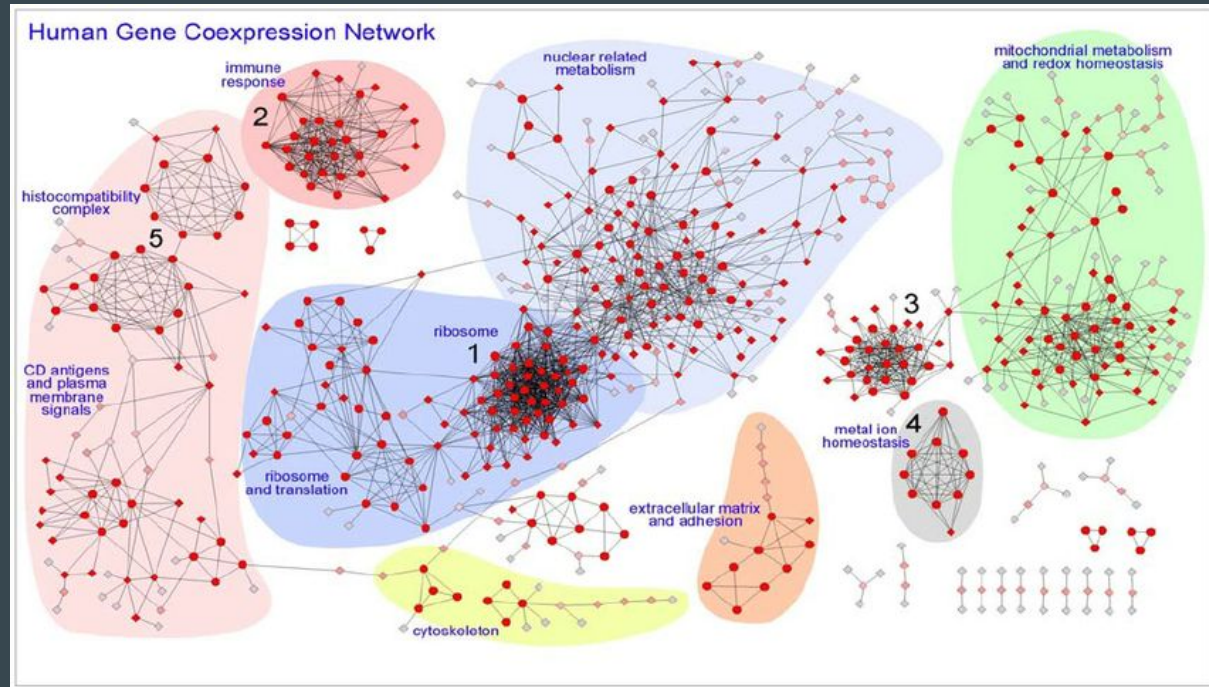
• • •

Costas Bouyioukos, Université Paris Diderot, EDC-UMR7216
Anaïs Baudot, CNRS Marseille

Outline

1.

Biological motivation

- Find co-expressed genes (by using the correlation of their gene expression profiles) this is a proxy for possible co-regulation.
- Find co-functional modules (clusters) this is a proxy for common function regarding a given phenotype.
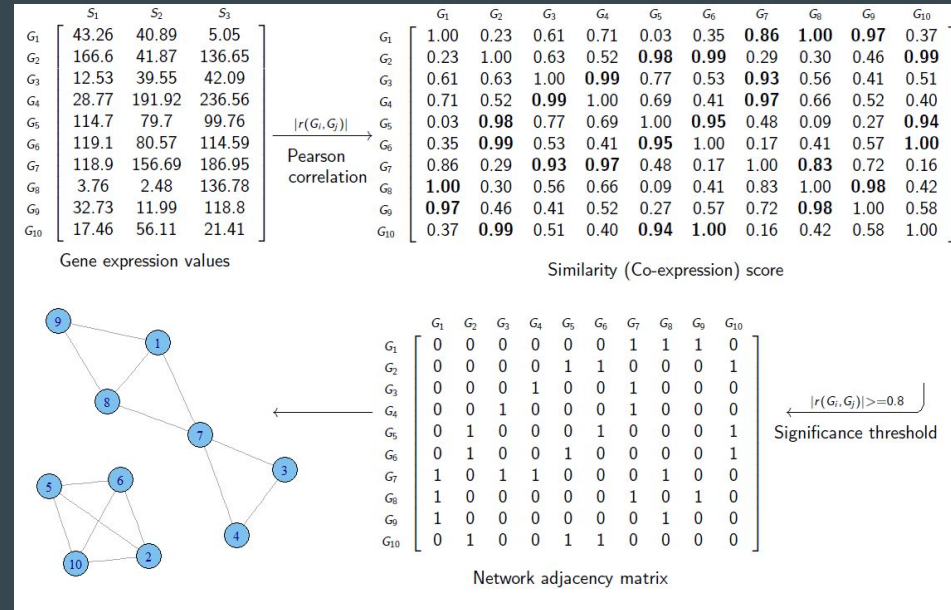
Gene Correlation matrices

- Correlation -> Is a measure of similarity between gene expression profiles.
- For a gene expression matrix, if we compute the correlation coefficient of all vs. all gene expression profiles we obtain the expression correlation matrix.
- Correlation is only ONE measure of similarity, we can use different coefficients (Pearson, Spearman etc.) and different measures (MI etc.).
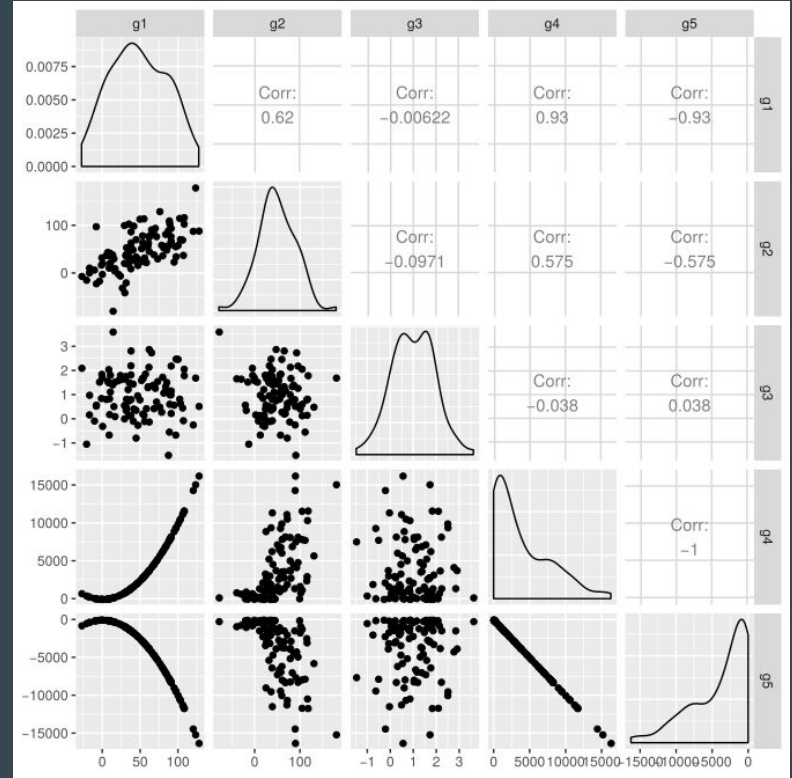
From matrices to correlation Networks

- A correlation matrix represents a (fully connected) weighted network
- Correlation sign (+/-) represents the type of association between biological entities.
- However a fully connected network is never useful. Solution:
  - Apply a threshold
- Then construct the adjacency matrix

  (which is the network)

# Network inference

- There are numerous algorithms to construct co-expression networks from "similarity" matrices.
- The Weighted Correlation Networks Analysis (WGCNA)
  - Identify clusters (modules) of highly correlated genes.
  - Summarise these modules by finding the "eigengene" (1st Principal Component)
  - Relate modules to external biological traits and to each other via the eigen gene.

WGCNA Module - network Detection
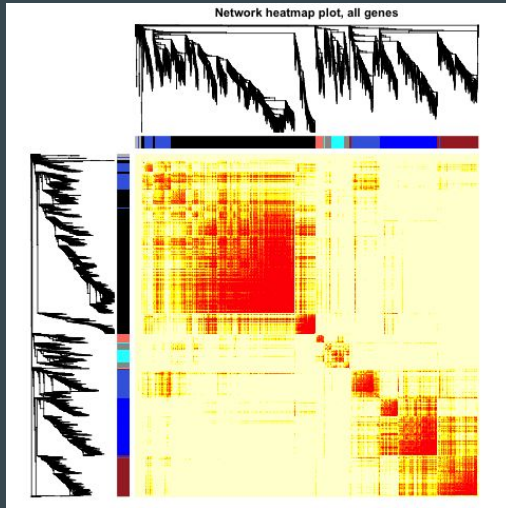

- Start by the similarity/correlation matrix.
- Apply the a "soft thresholding" approach.
- Hierarchical clustering of the "similarity" (correlation) matrix.
- DeepTreeCut to Identify Topological Overlapping Modules to separate the modules
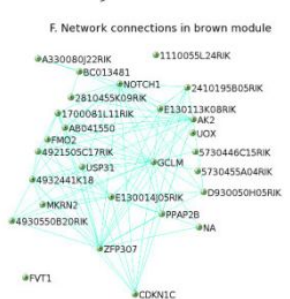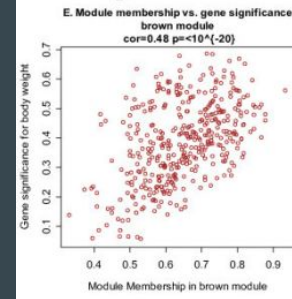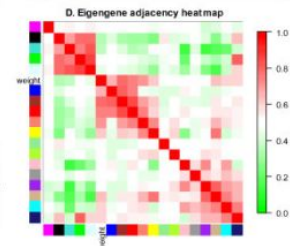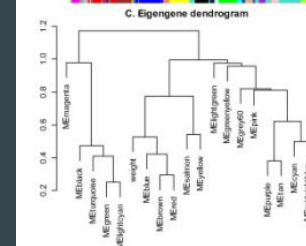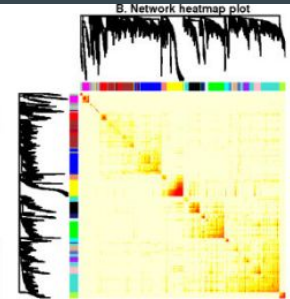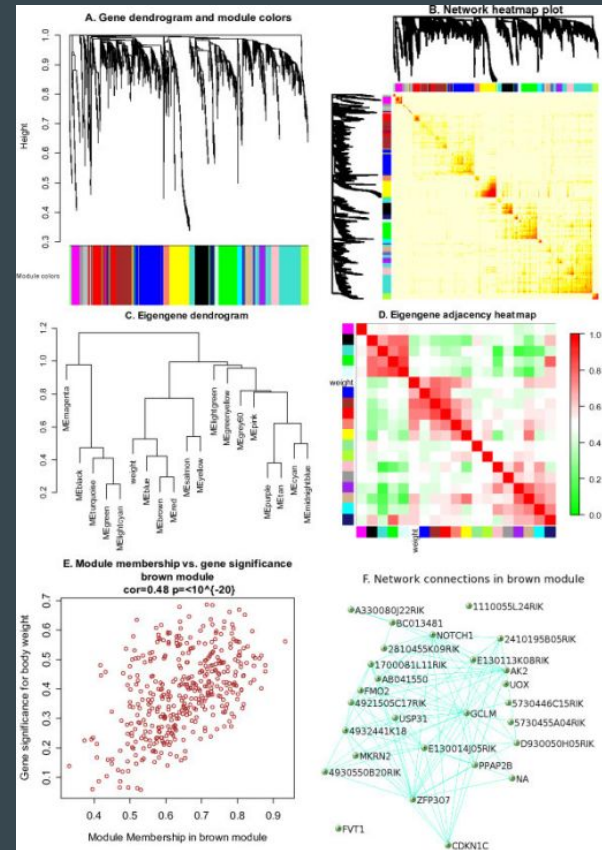
Eigen-genes and phenotypes

- The concept of "Eigengene"
  - A "virtual" gene with a gene expression profile such that it represents better than anything else all the genes in the module.
  - Can be visualised as the "barycentre" of a module.
  - Mathematically is the first eigenvector (principal component) of a eigen decomposition of the correlation matrix.
- We use this module signature to relate modules between them ad between modules and external biological traits (phenotypes, clinical data, features). By calculating their correlation coefficient.
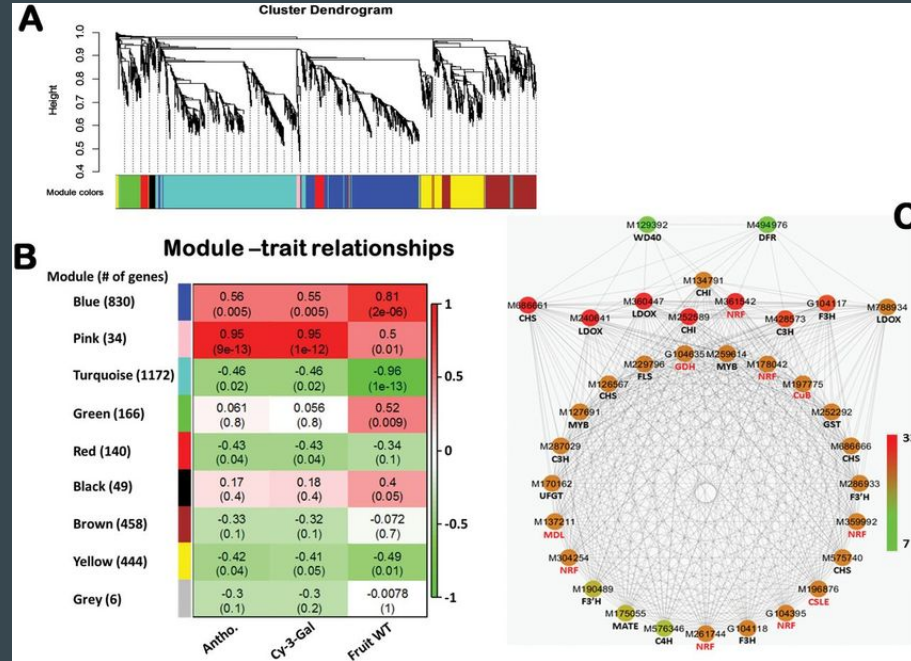
Module detection

- For the calculation of this soft threshold many algorithms have been developed. WGCNA is using DynamicTreeCut (thresholding differently for different modules).
- Clustering of the eigengenes also provides
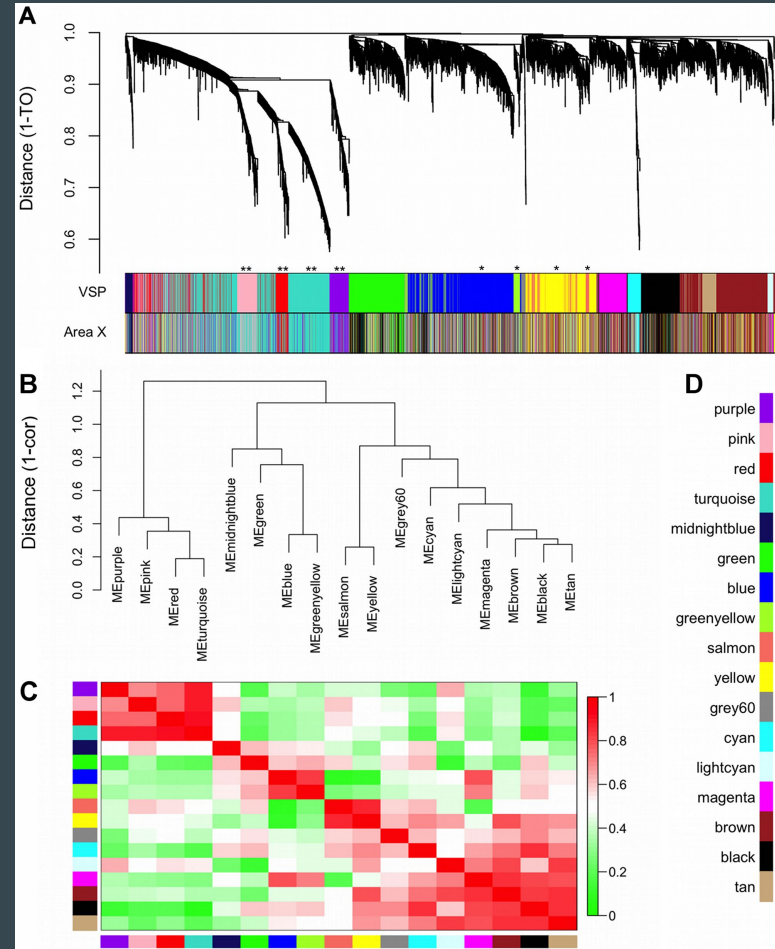
How the network is build

- The central algorithm for network construction is TOM (signed or unsigned)
- TOM stand for e Topological Overlap Measure?Matrix
- As any approach in correlation/coexpression networks is trying to remove spurious connections.

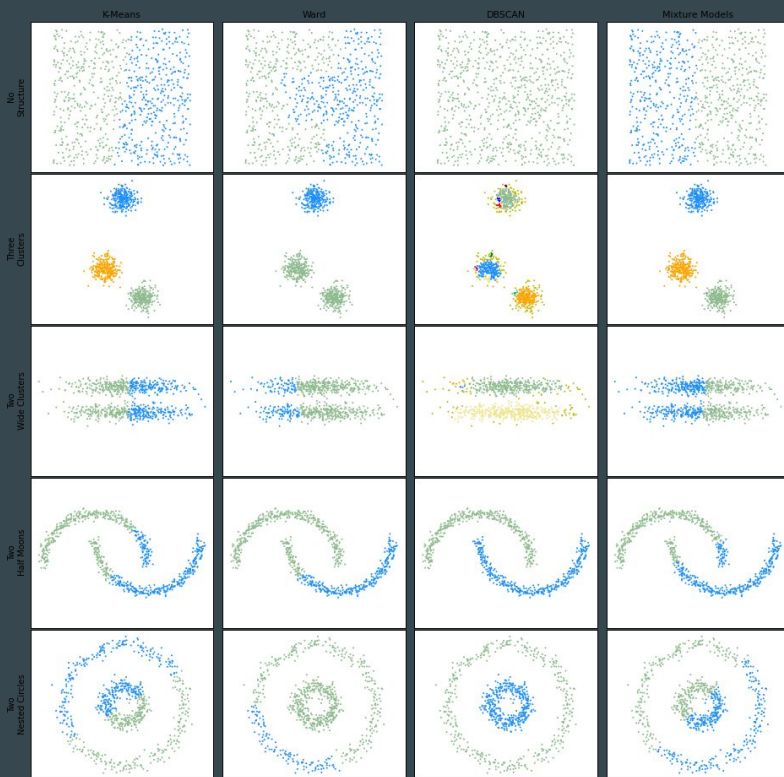Source: E. Ravasz et al., Science 297, 1551 -1555 (2002)
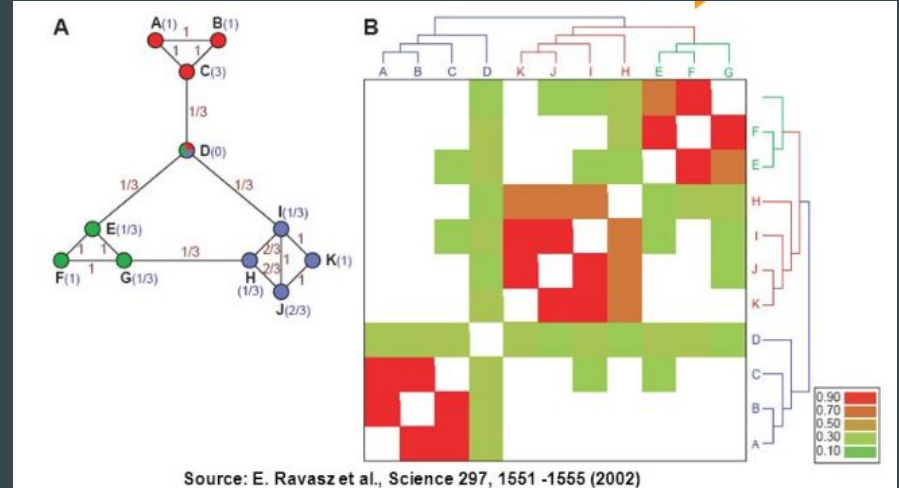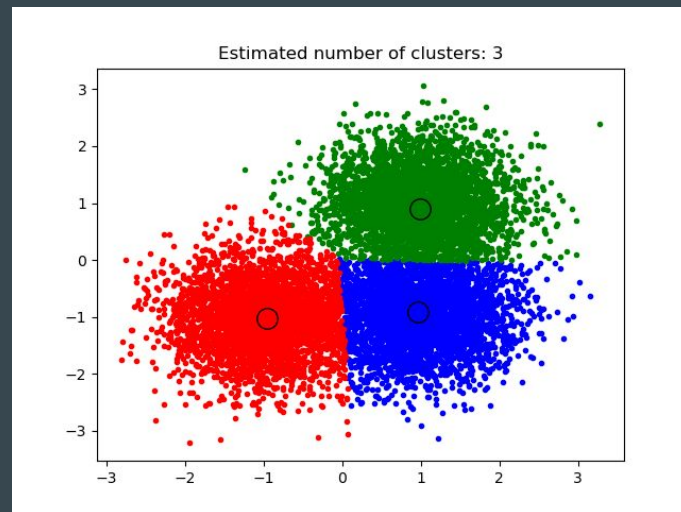
Introduction to clustering

- Clustering, a general family of unsupervised methods to group entities which have similarities.
- Clustering algorithms operate on any kind of similarity measures:
  - Euclidean and other types of distance.
  - Correlation(s)
  - Covariance(s)
-
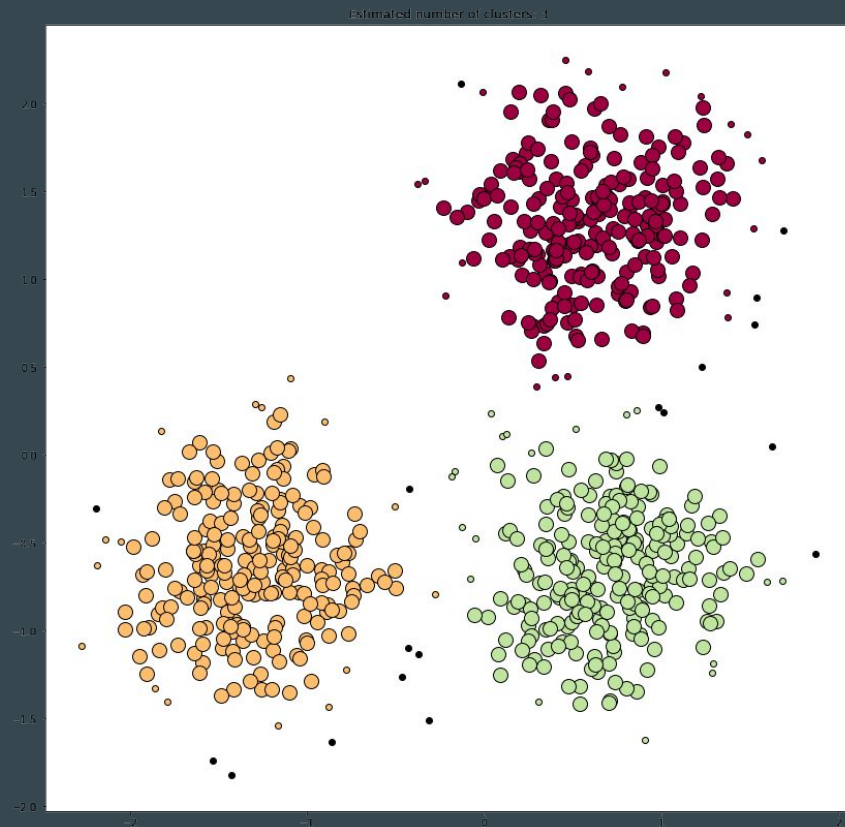
Problems of clustering

- Easy to use -thus- easy to fail methods.
- Difficult to determine a priori the "real" number of clusters.
- Many different algorithms/approaches -difficult to chose.
- Biological data is highly correlated.

Estimated number of clusters: 3

Introduction to network clustering

- Cluster areas of networks that are more connected w.r.t the rest of the network. Network clusters are also called modules.
- Helps to identify related elements and thus relate their function (guilt-by-association)
- Allows the association of modules with biological traits of interest
- Used extensively in PPIs (to identify interactions and complexes) and in GRNs to identify <u>co-expression modules</u>.
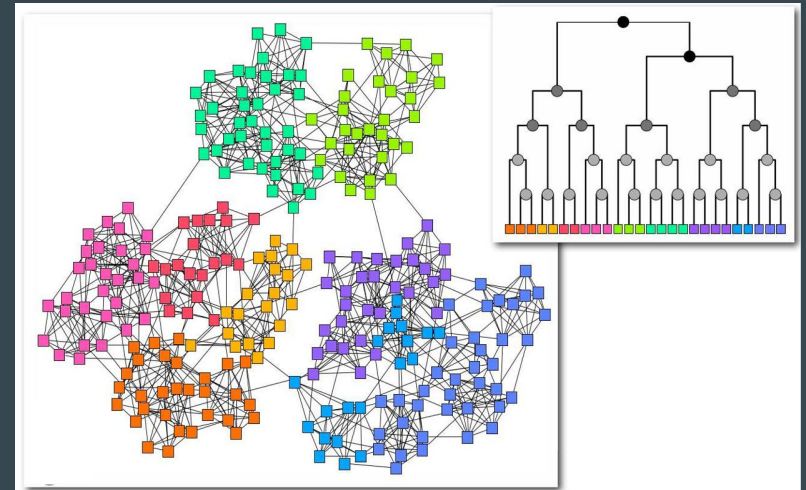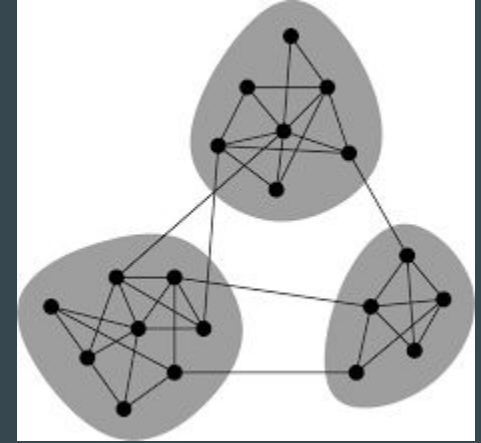
The WGCNA package

- We will work with an "easy to use", "user friendly" R package that implements all the functionalities of the WGCNA method and networks.
- We will export the inferred network to Cytoscape for visualisation.