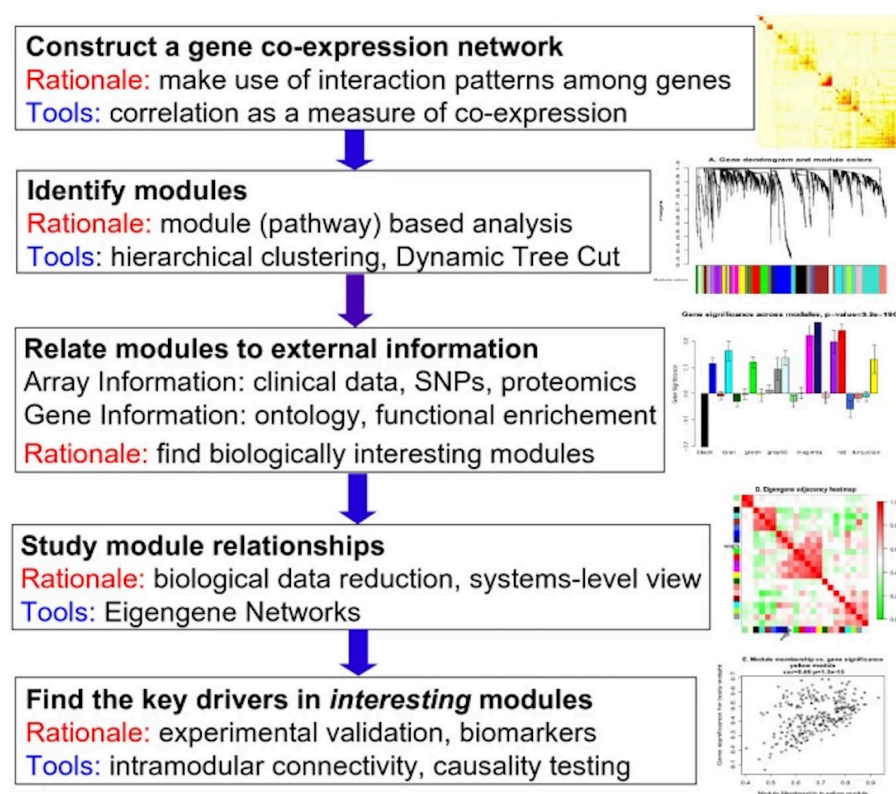


## Weighted Gene Co-Expression Network Analysis (WGCNA)

WGCNA stands for Weighted Gene Co-expression Network Analysis. It's a widely used bioinformatics method for studying biological networks, particularly in the context of gene expression data. WGCNA allows researchers to identify groups of genes that are co-expressed across different samples or conditions. By identifying these gene modules, researchers can gain insights into the underlying biological processes and regulatory mechanisms.

In WGCNA, genes are grouped into modules based on their patterns of co-expression across samples. The method constructs a network where genes are represented as nodes, and the strength of their co-expression is represented as edges. Modules are then identified as branches or clusters within this network, often through hierarchical clustering or other methods.

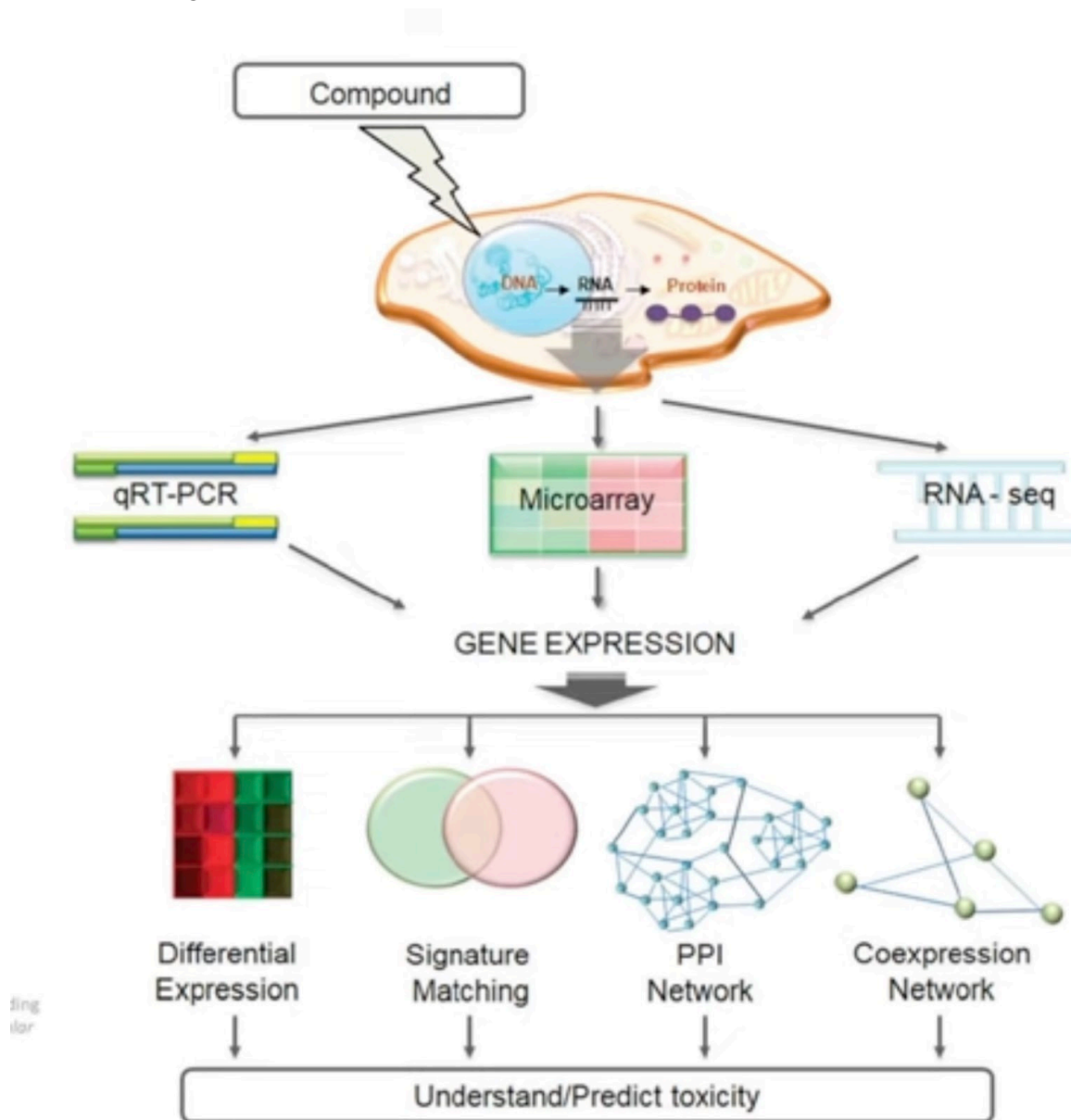
WGCNA has been applied in various fields, including cancer research, neuroscience, and plant biology, among others. It's valuable for identifying candidate biomarkers, understanding gene regulatory networks, and uncovering potential therapeutic targets.



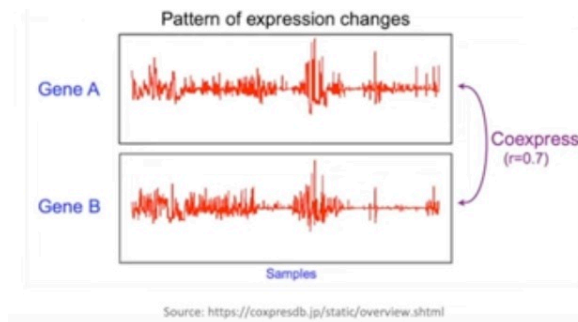
## Overview

Let's say there is an experimental drug being tested for cancer, and there are some patients that aren't responding to the therapy. We are interested in observing the systemic level changes that contribute to resistance. You want to identify the pathways involved in building this resistance, thus leading to typing to identify the genes involved in the pathway.

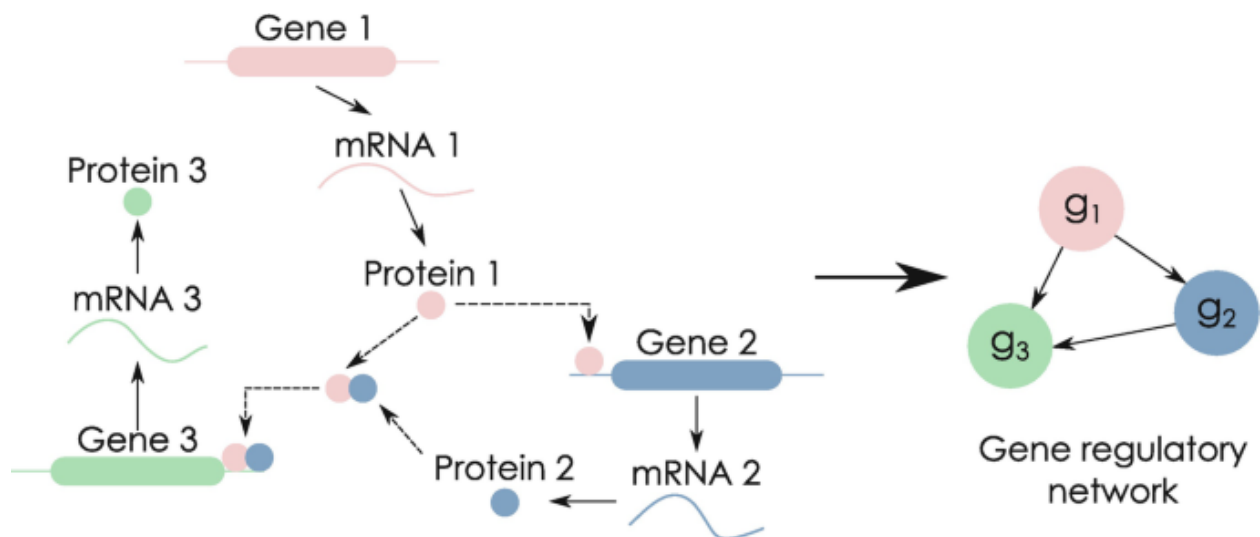
You can get a gene expression matrix through various transcriptome analysis. After the matrix, we can perform computational analysis. But these will only give you individual gene expression, not the actual correlation between genes.



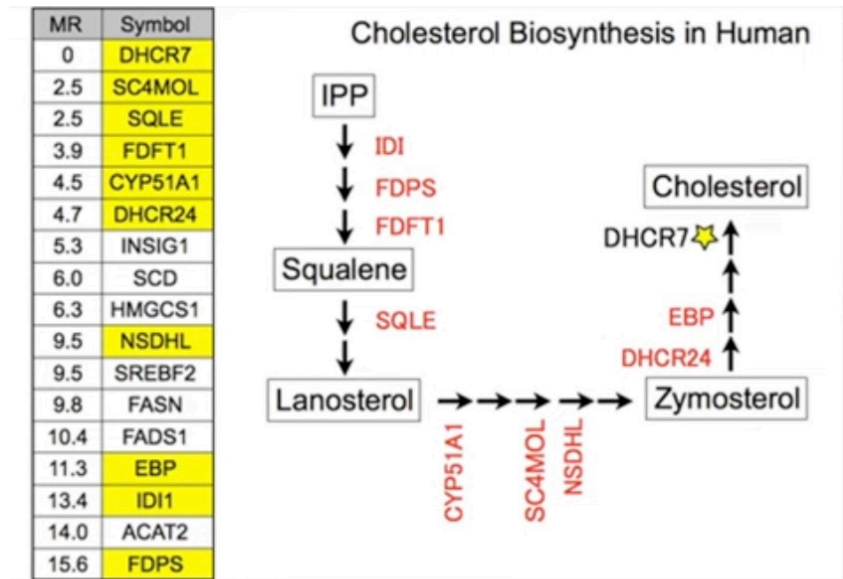
**Genes with similar expression patterns are functionally associated suggesting :  
This means that genes are correlated through similar gene expression patterns**



- Part of the same complexes
- Involved in the same pathway
- May influence each other by a gene expression regulation mechanism i.e, they may be part of the same gene regulatory network



Example of a gene regulatory network between 3 genes indicating function association

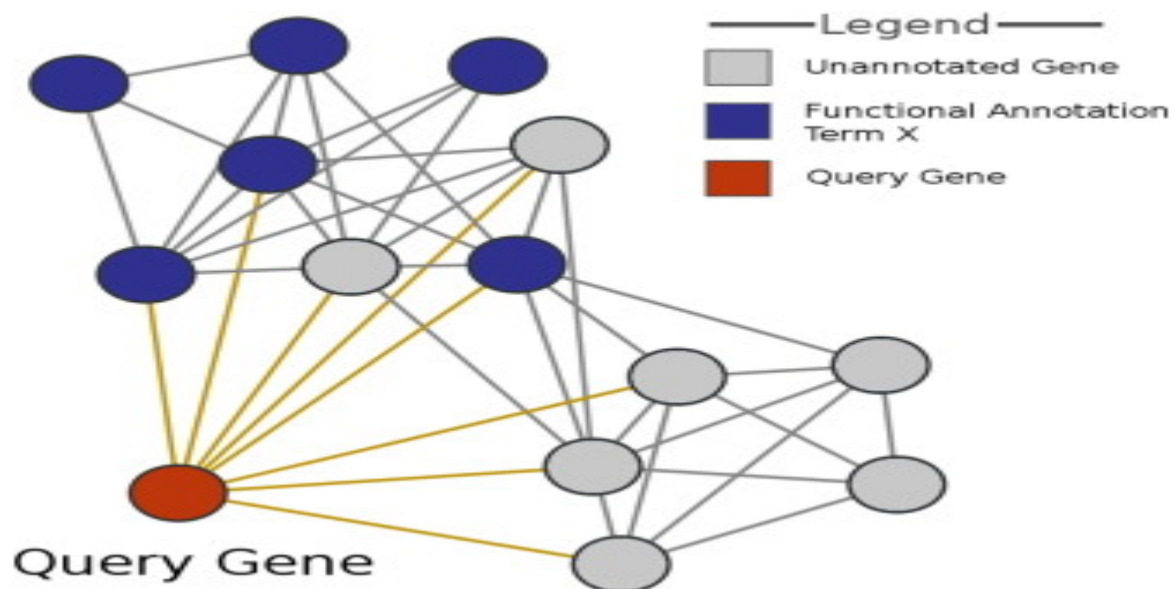


Source: <https://coexpresdb.jp/static/overview.shtml>

Example of genes correlated through the cholesterol biosynthesis pathway

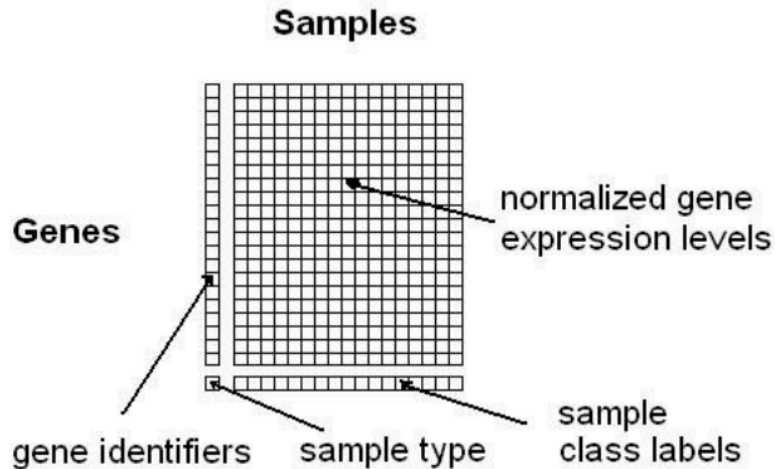
### Terminologies of gene network analysis

- Node
- Edge
- Weighted edge
- Modules : clusters
- Hub : highly related genes
- Module eigengene : representative gene profile

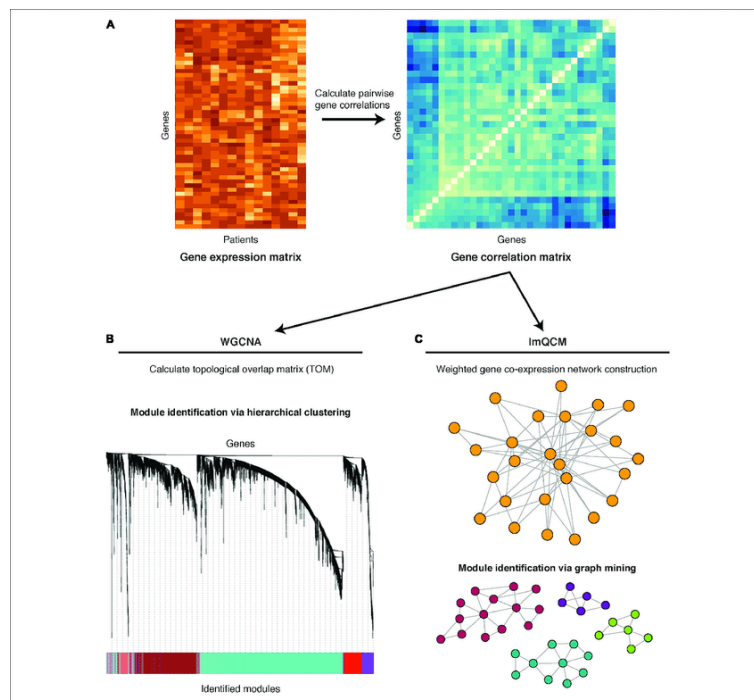


### Gene expression matrix :

- Rows : genes
- Columns : expression values
- All values must be normalised



The WGCNA uses a clustering approach to find correlated genes. It uses the gene expression matrix to attach weighted values to gene correlation and then finds clusters of related genes. These clusters can be identified and then analysed further.

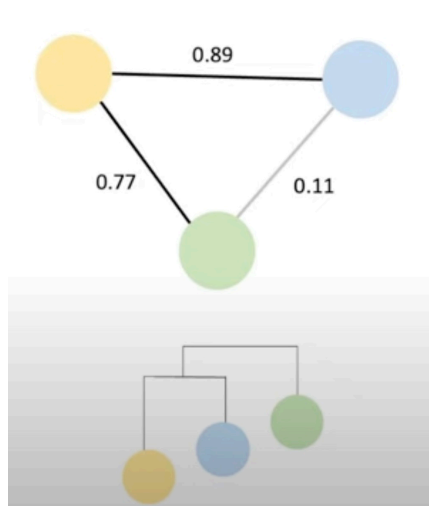


## Methodology

### WGCNA Workflow :

Gene expression data → Construct weighted correlation matrix → Identify Clusters → Correlate clusters with phenotypes/traits → Identify driver genes

### Basic Idea behind constructing Networks :



We use “correlation” to measure similarities. We perform correlation between all pairs of nodes.

1. Have a distance measure that measures similarities between nodes
2. We want to highlight the connections that are important and reduce the noise by **applying a threshold**.
3. This gives us an adjacency matrix whose values give us **proximity measures**.
4. Proximity measures give us **hierarchical clustering**.

### Step 1 : Constructing an adjacency matrix from gene expression data

A similarity matrix measures the correlation between a pair of genes, or rather how similarly the expression of a pair of genes changes together.

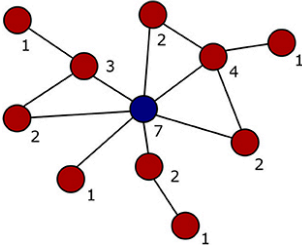
These can be signed or unsigned. Signed networks denote if genes are upregulated or downregulated.

By applying thresholds to these values, we can develop an adjacency matrix.

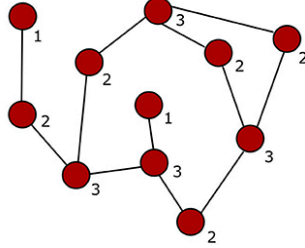
1. **Hard thresholding** : 1 to all values above threshold, and 0 for all others. This results in an unweighted network.
2. **Soft thresholding** : Use correlation values and raise it to a power term. This results in a weighted matrix.

## Step 2 : Understanding types of networks

A Scale-free network



B Random network

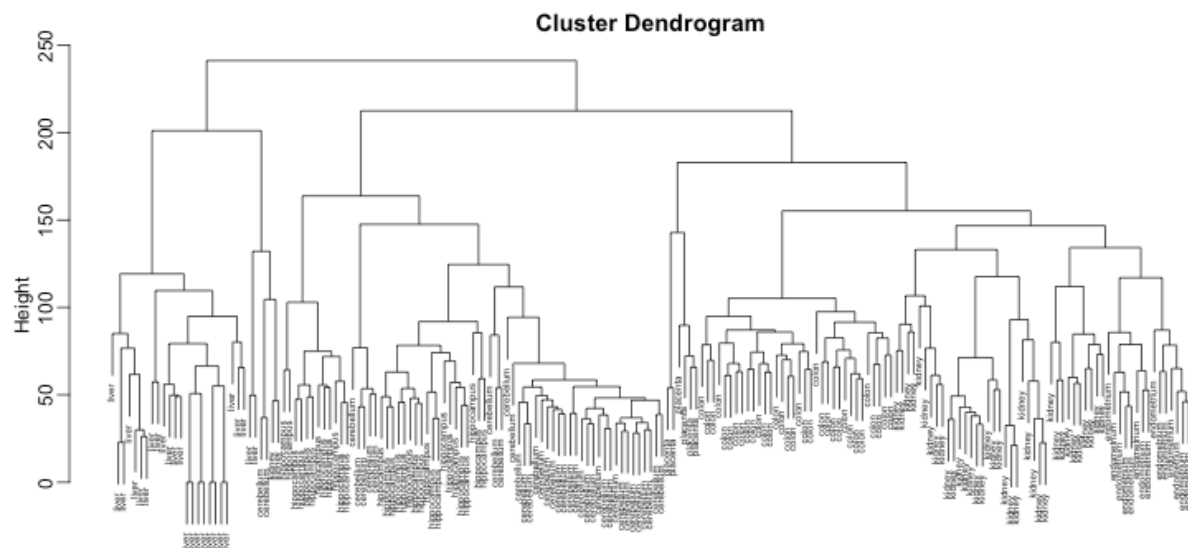


We want to build a scale free network where we can identify the hubs where nodes are highly connected

## Step 2 : Identify modules for clustering genes

To cluster genes into network modules, we need to use a network proximity measure. WGCNA uses the **topological overlap measure**. If 2 nodes are highly correlated they will have a higher topological overlap measure and lower dissimilarity. This gives us the distance from the gene to every other gene by subtracting the value from 1.

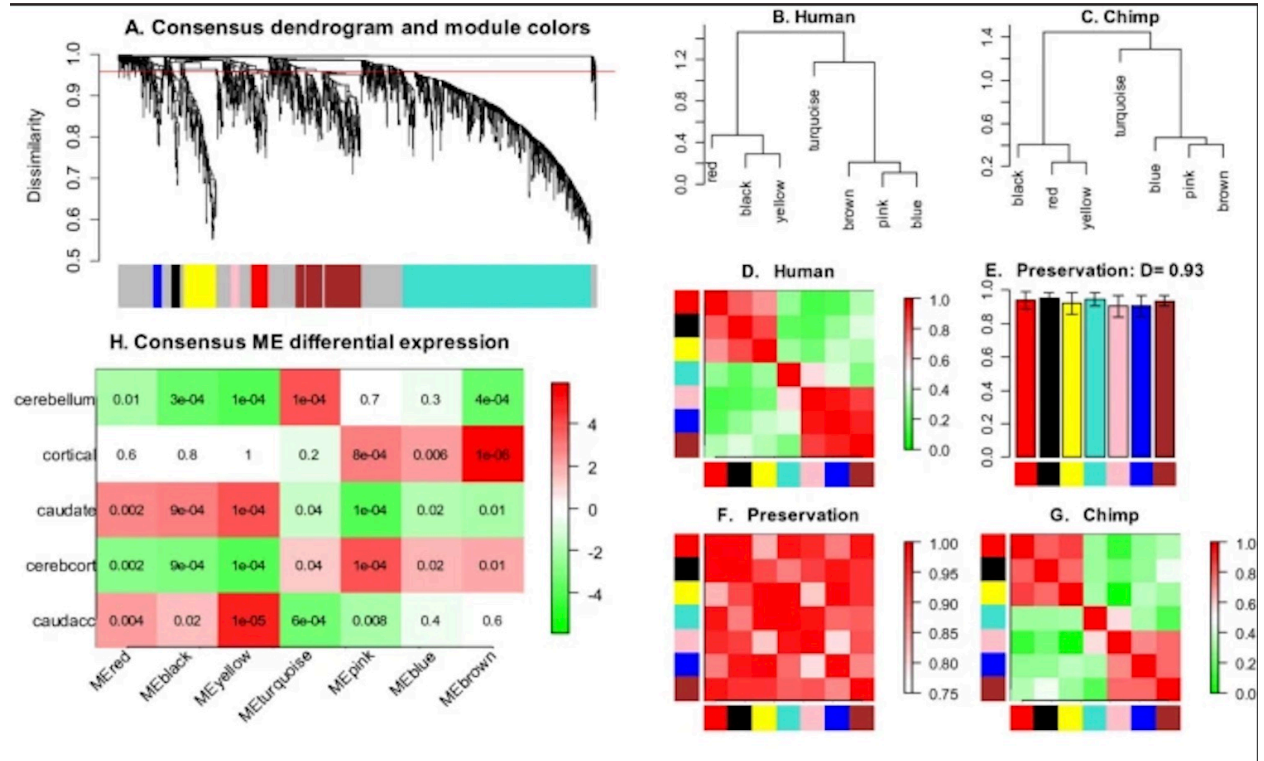
We can divide clusters into modules using **cutree** function. Once we have modules, we can merge them into a larger module of those with similar gene profiles using a dendrogram.





### Step 3 : Mapping modules to traits

We can correlate modules to external phenotypes or traits using module eigengenes. We can map a module to its phenotype by using previously found eigengene values for that trait



We can also identify **driver genes** by **identifying** the genes that are highly connected within the module.