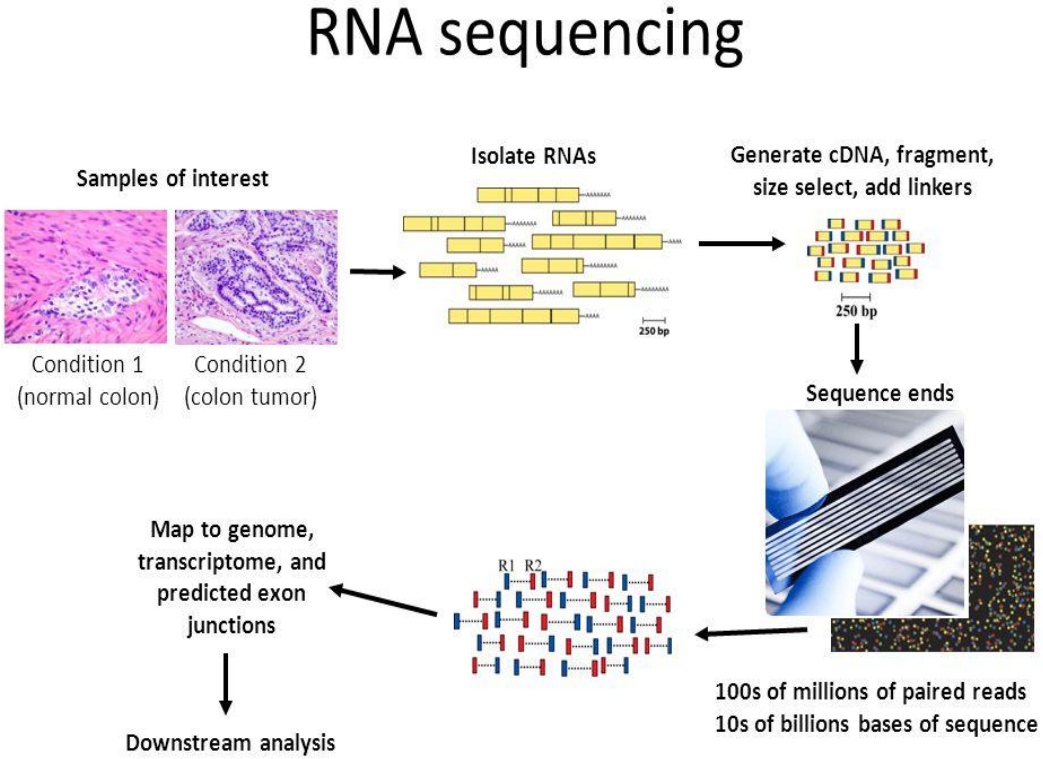
Gene expression clustering using gene ontology and biological networks

Student:
Peter Leontev, SPbAU

Scientific advisor:
Son Pham, UCSD

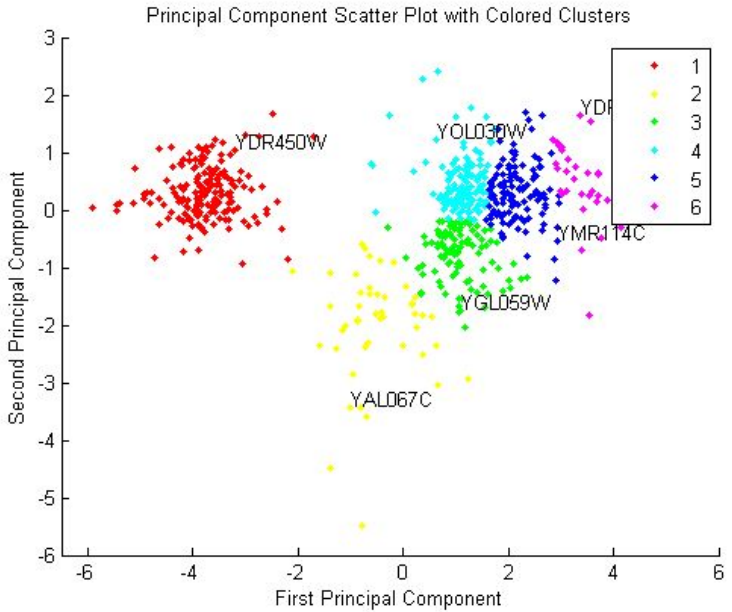
Gene expression data source

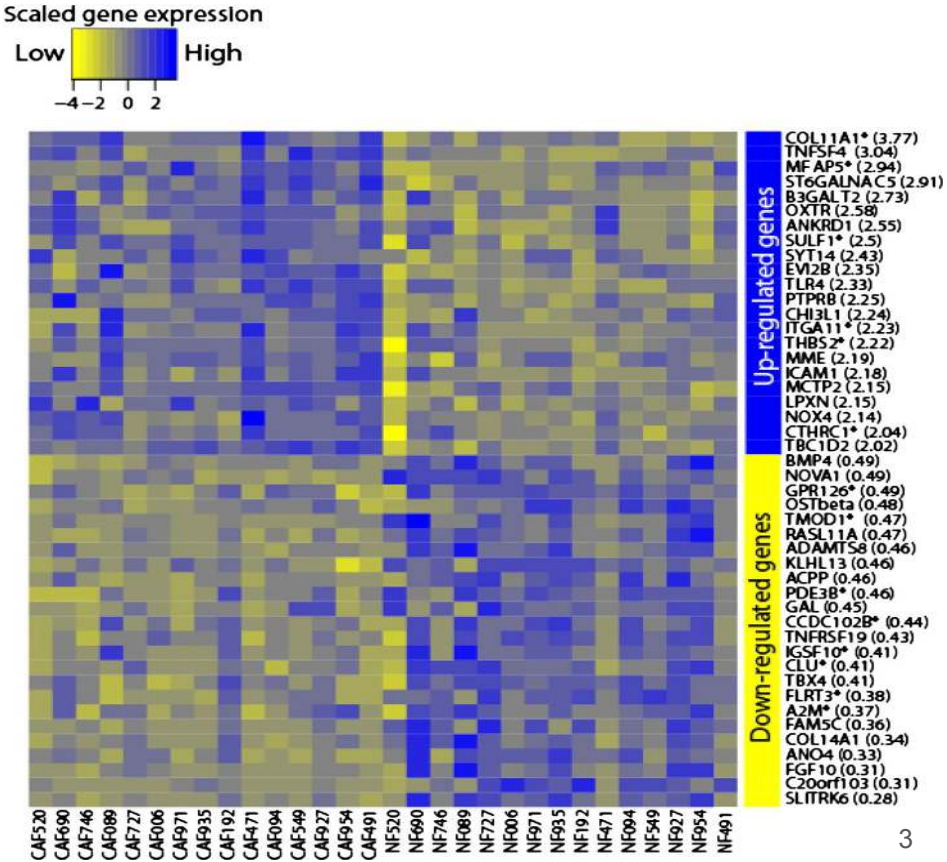
- Microarrays
- RNA-seq
- Single-cell RNA-seq



Gene expression data representation

- CPM, FPKM, RPKM, TPM
- Heatmap/PCA/tSNE/...





Clustering in a nutshell

- *Similar* items should fall into the same clusters whereas *dissimilar* items should fall in different clusters.

- There is no single best criterion for obtaining a partition because *no single and precise definition of cluster exist*.

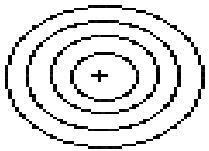
- As a result, there are tremendous amount of methods.

Proximity measure

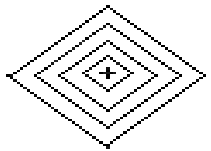
- Gene expresion values can be formalized as numerical vectors:

$$O_i = \{o_{i,j} | 1 <= j <= M\}, 1 <= i <= N, where\ N\ and\ M\ are\ number\ of\ genes\ and\ samples,\ respectively.$$

- There are many proximity metrics such as L1 and L2 norms, Mahalanobis distance, correlation, etc.



Euclidean

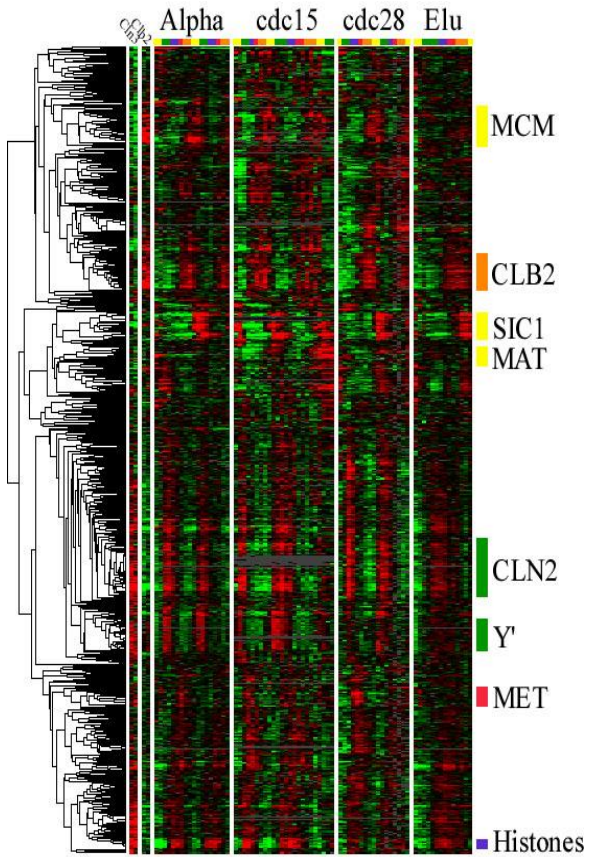


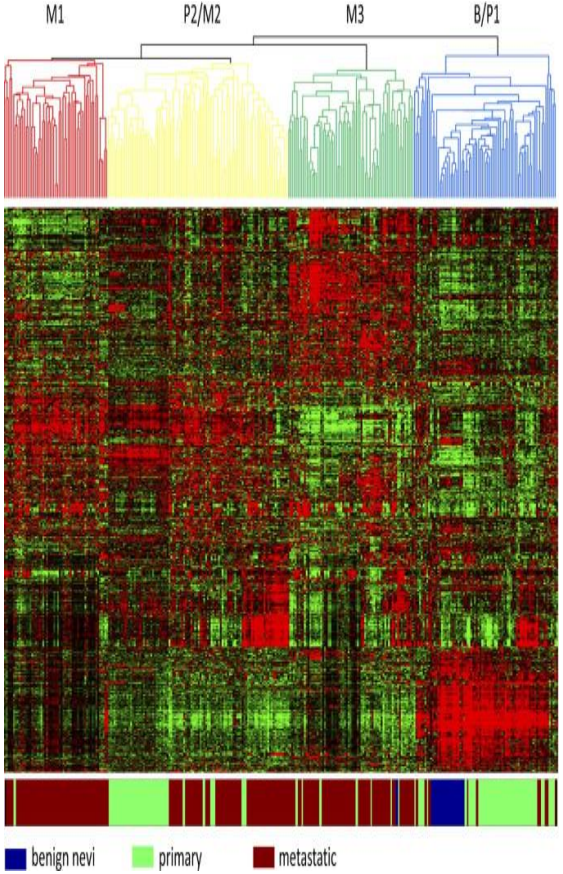
Manhattan

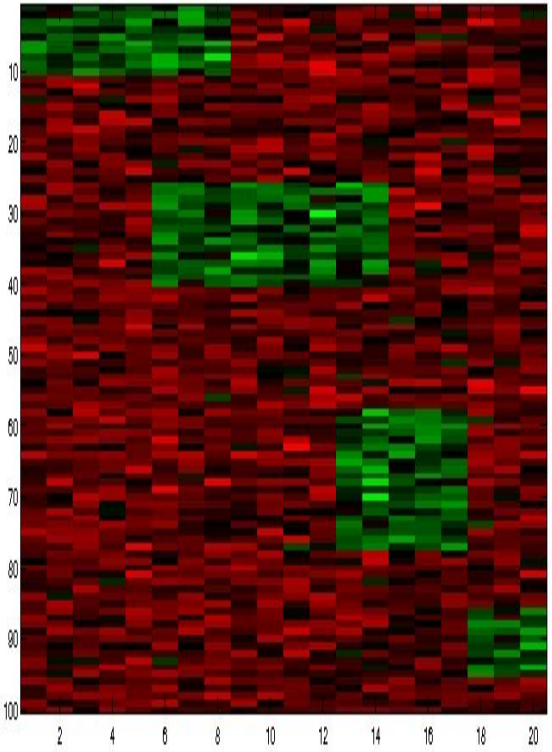
Mahalanobis

$$r = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})(y_i - \bar{y}) \right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Clustering by genes, samples and biclustering

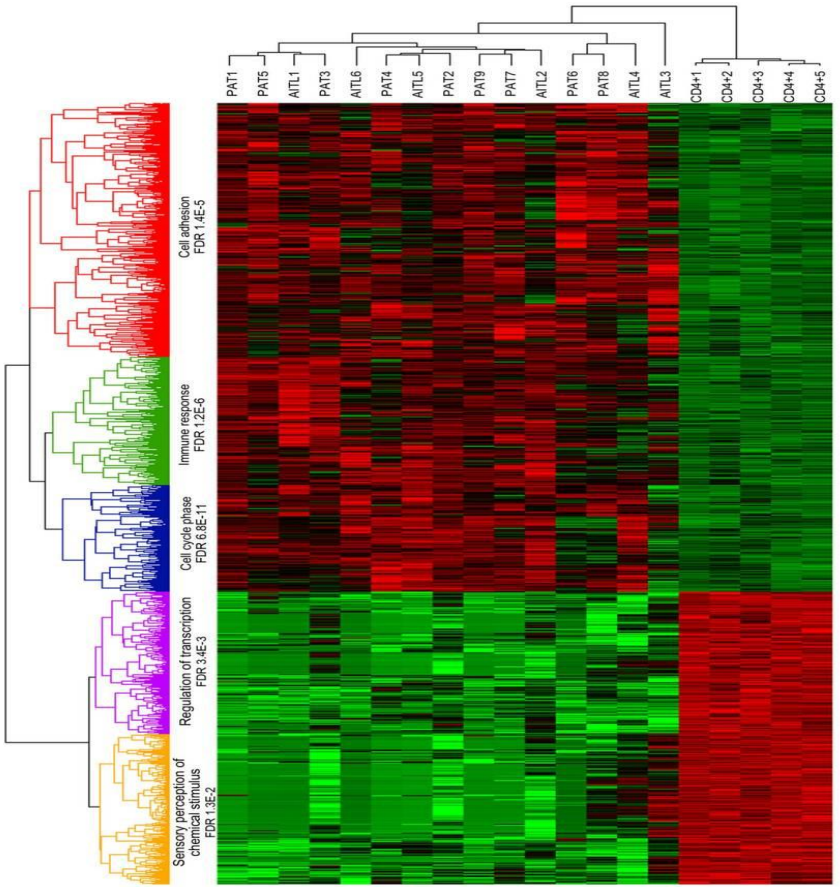






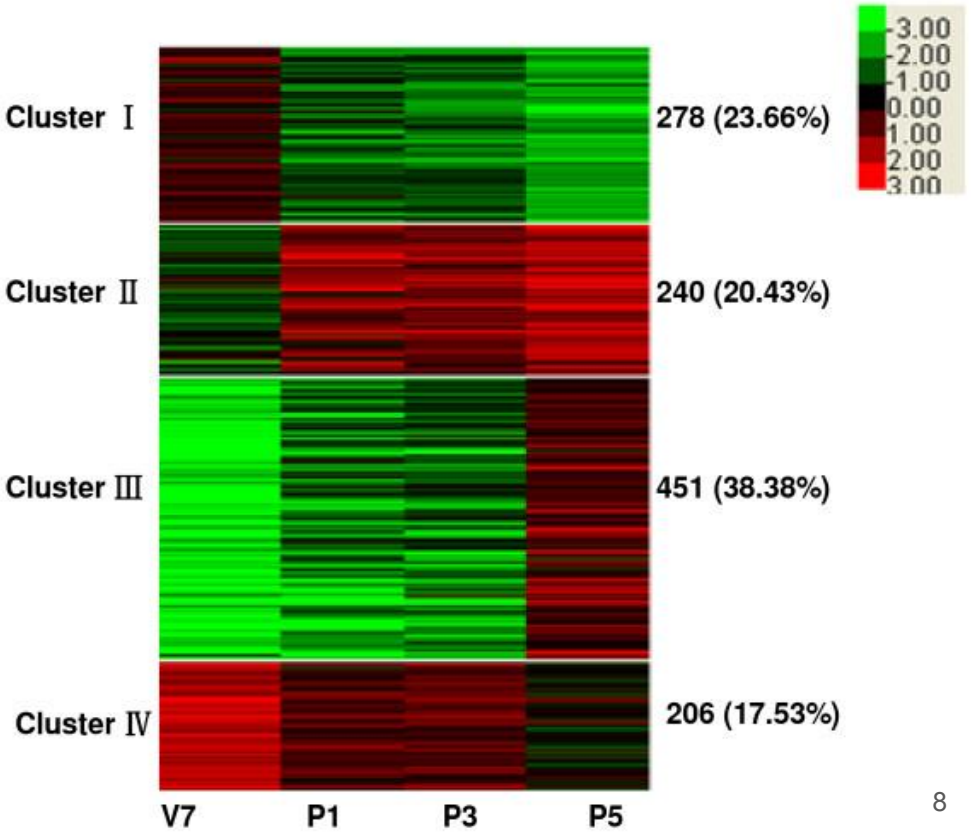
Two popular clustering algorithms

Hierchical clustering (HC) allows us to cluster both genes and samples in one picture and see whole dataset.



Two popular clustering algorithms

K-means is a partitioning method which requires predefined number (K) of clusters.



Disadvantages of proximity-based only algorithms

- Use simple distance metrics for large-scale datasets.
- Consider genes as *independent entities*.
- Ignore known functions of genes.

Ref: When Is “Nearest Neighbor" Meaningful? Beyer et al, 1999

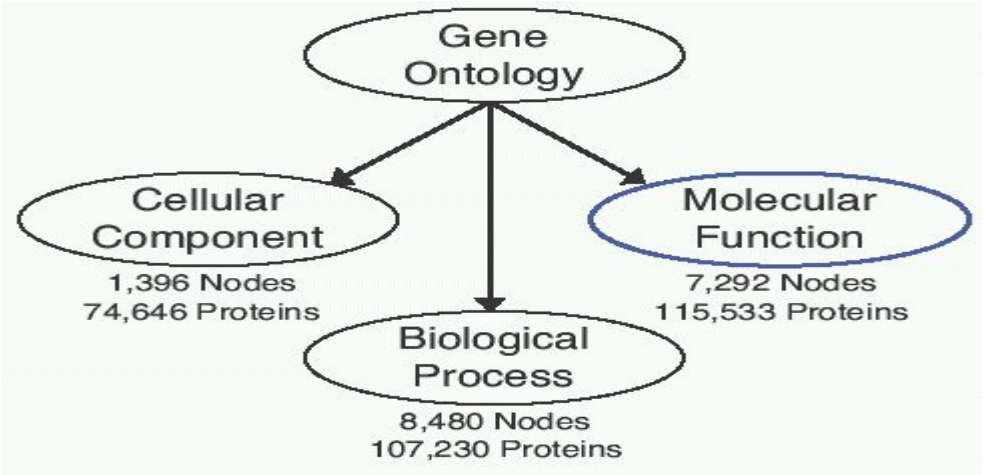
Embedding prior knowledge

- Many clustering algorithms were developed that use knowledge databases in the clustering process.

- However, they follow the principle that integrating one external source of knowledge to guide an algorithm is enough.

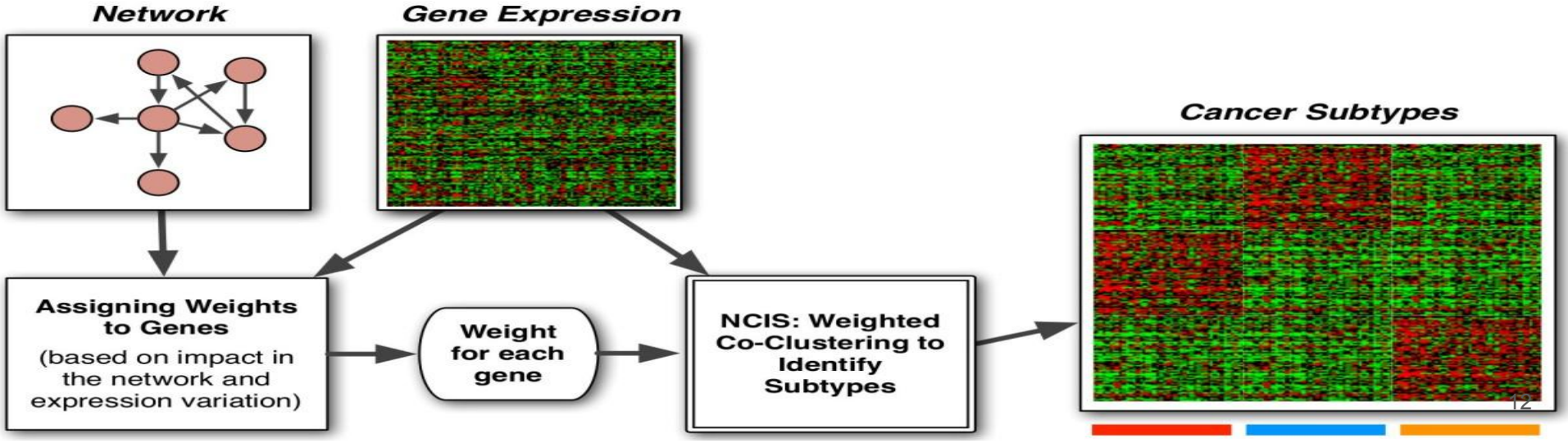
Existing approaches (gene ontology)

- A knowledge-based clustering algorithm driven by Gene Ontology, Cheng et al, 2004.
- Seeing the forest for the trees: using the Gene Ontology to restructure hierarchical clustering, Dotan-Cohen et al, 2009.



Existing approaches (biological networks)

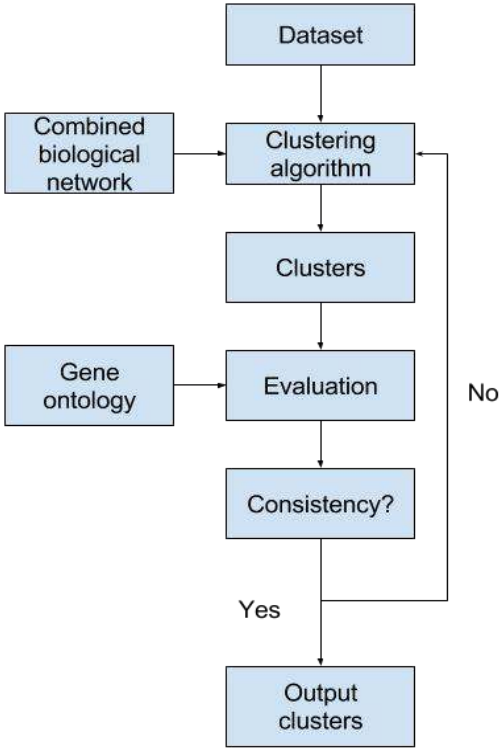
- A network-assisted co-clustering algorithms to discover cancer subtypes based on gene expression, Liu, 2014.
- Improving clustering with metabolic pathway data, Diego et al, 2015.



The goal

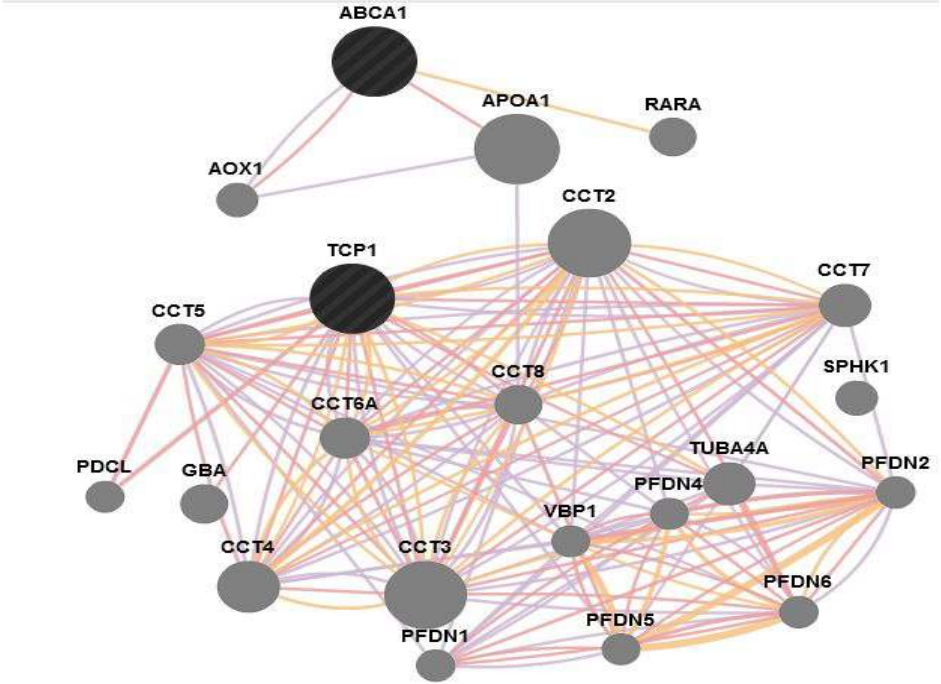
To combine different biological networks and gene ontology to integrate them into gene expression (bi)clustering algorithm.

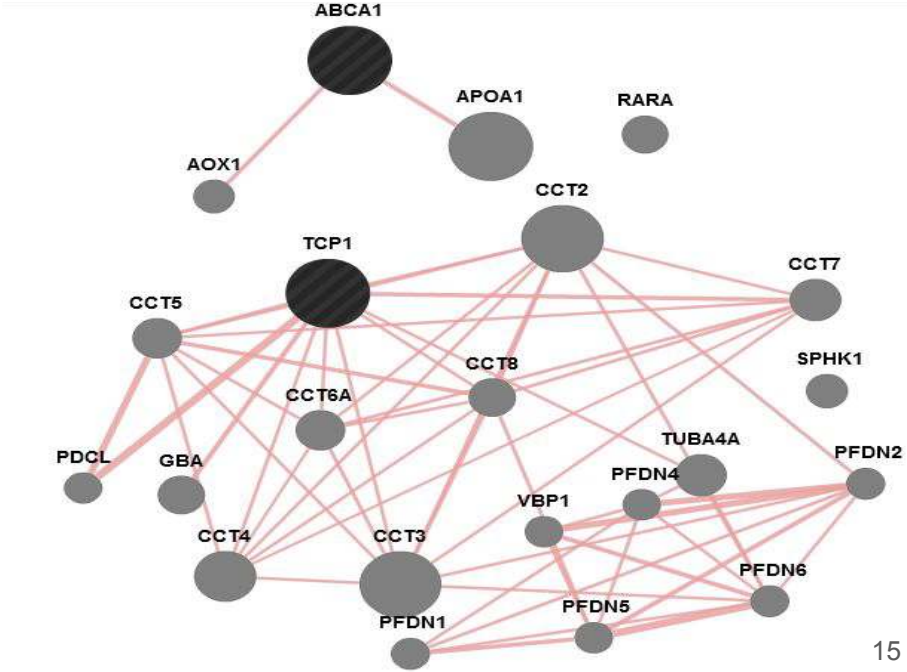
Proposed clustering algorithm



Combining biological networks

- If possible consider any network as directed and weighted.
- Take into account *as much as possible* edge's information.





Evaluation

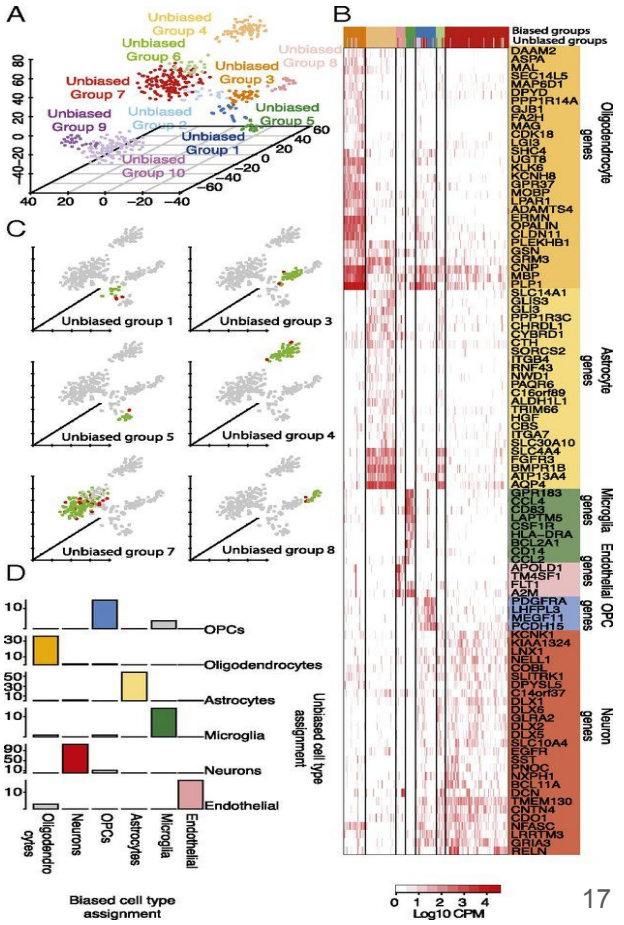
Hypothesis: gene that show similar expression patterns have common function and so may have the similar networks features too.

Evaluation must answer the following questions:

1. How similar gene expression values?
2. Do we have consistency in the gene ontology terms?
3. Can we infer some pathways?(*)

Dataset

- Darmanis S, Sloan SA, Zhang Y, Enge M et al. “A survey of human brain transcriptome diversity at the single cell level” (GSE67835).



Relevance

- Do we have enough knowledge to cluster gene expression data to determine ***cell* types**?

- Do we have enough knowledge to cluster gene expression data to determine ***cell subtypes***?

Thank you!
Questions?