生物信息学与系统生物学

张世华
中国科学院数学与系统科学研究院

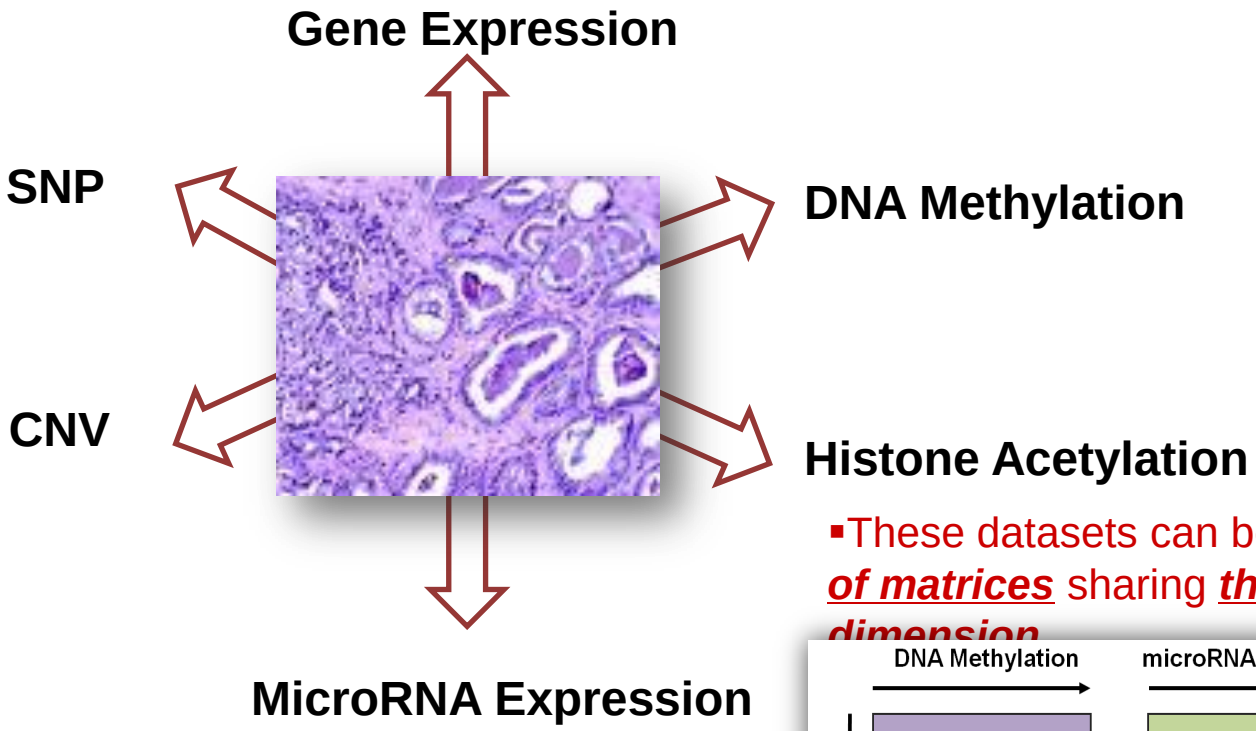
MicroRNA-gene co-module identification via semi-supervised machine learning technique

Shihua Zhang

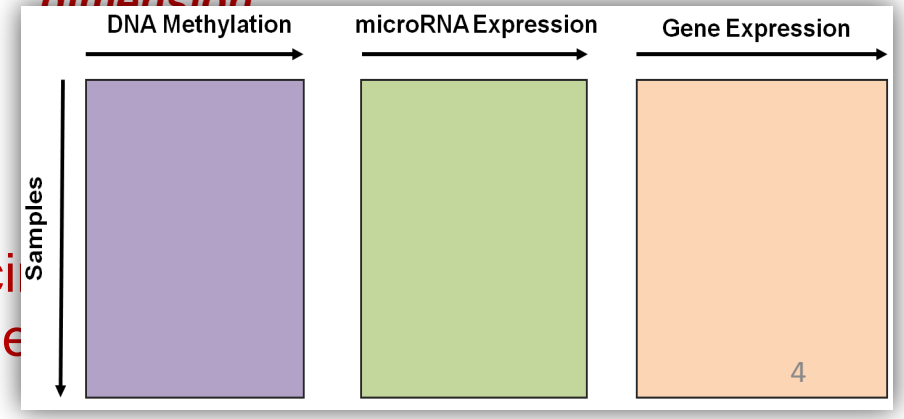
◆**Part I: Integrating multi-dimensional genomic data to identify multi-dimensional modules in an unsupervised manner;**

◆**Part II: Integrating multiple types of data to discover miRNA-gene co-modules in a semi-supervised manner.**

Background: Multi-Dimensional genome-wide profiling of same samples

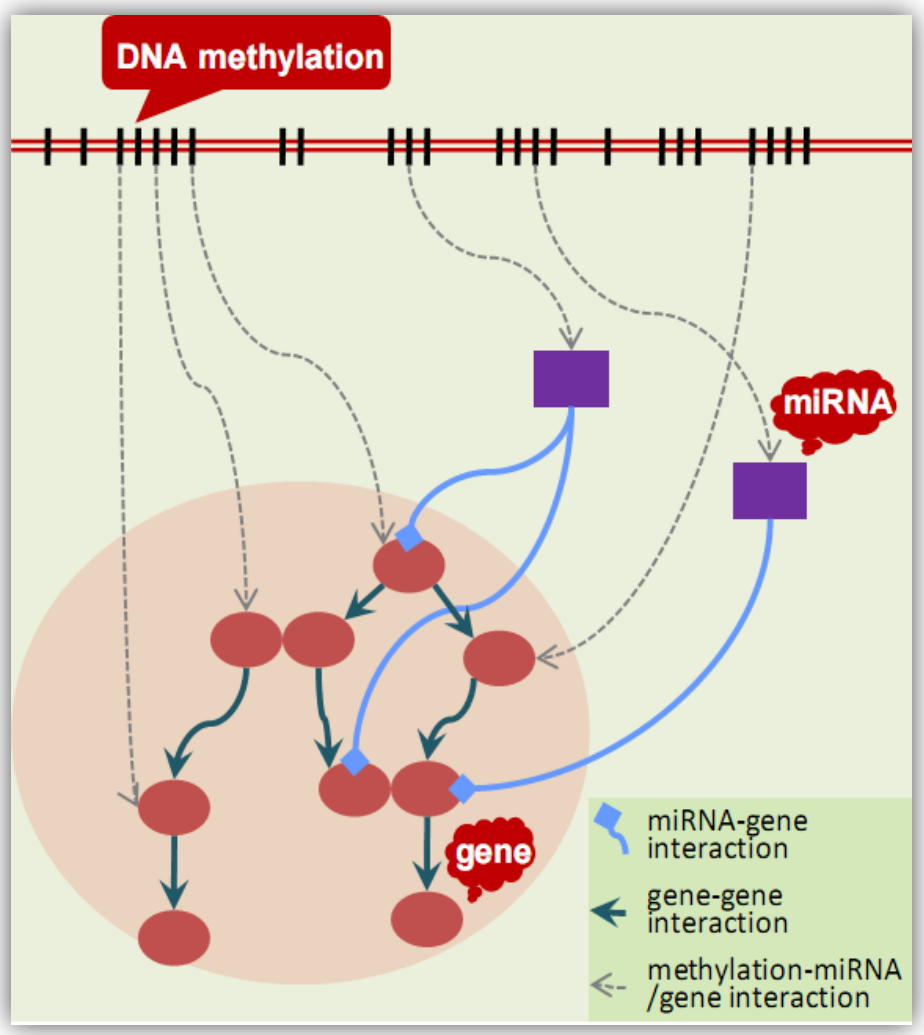


- These datasets can be represented as ***a set of matrices*** sharing ***the same sample dimension***



Thanks to the next-generation sequenci
dimensional genomics datasets will eme

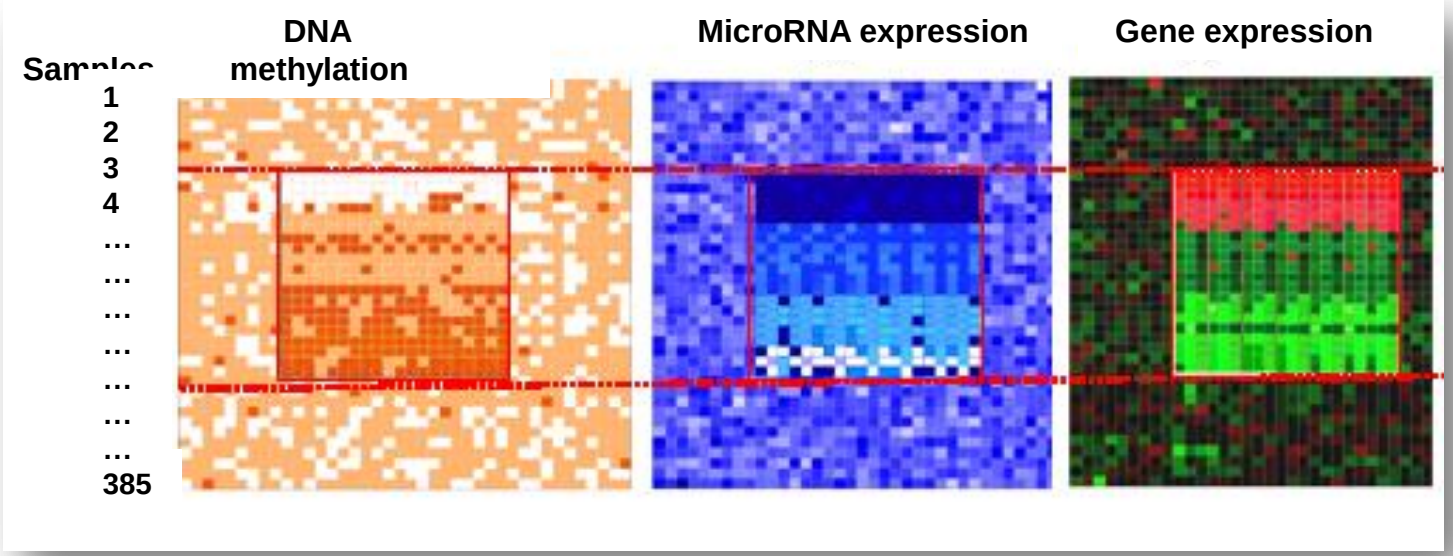
Motivation: Multi-dimensional combinatorial regulatory relationships



The multiple factors may form coordinated ***regulatory programs or modules*** to achieve specific functions

Our goal

- Identify <u>multi-dimensional modules</u> across multiple types of genomics data



A ***multi-dimensional module*** is a set of **DNA methylation markers**, **miRNAs** and **genes that show correlated profiles** across a **subset of samples**.

Our approach

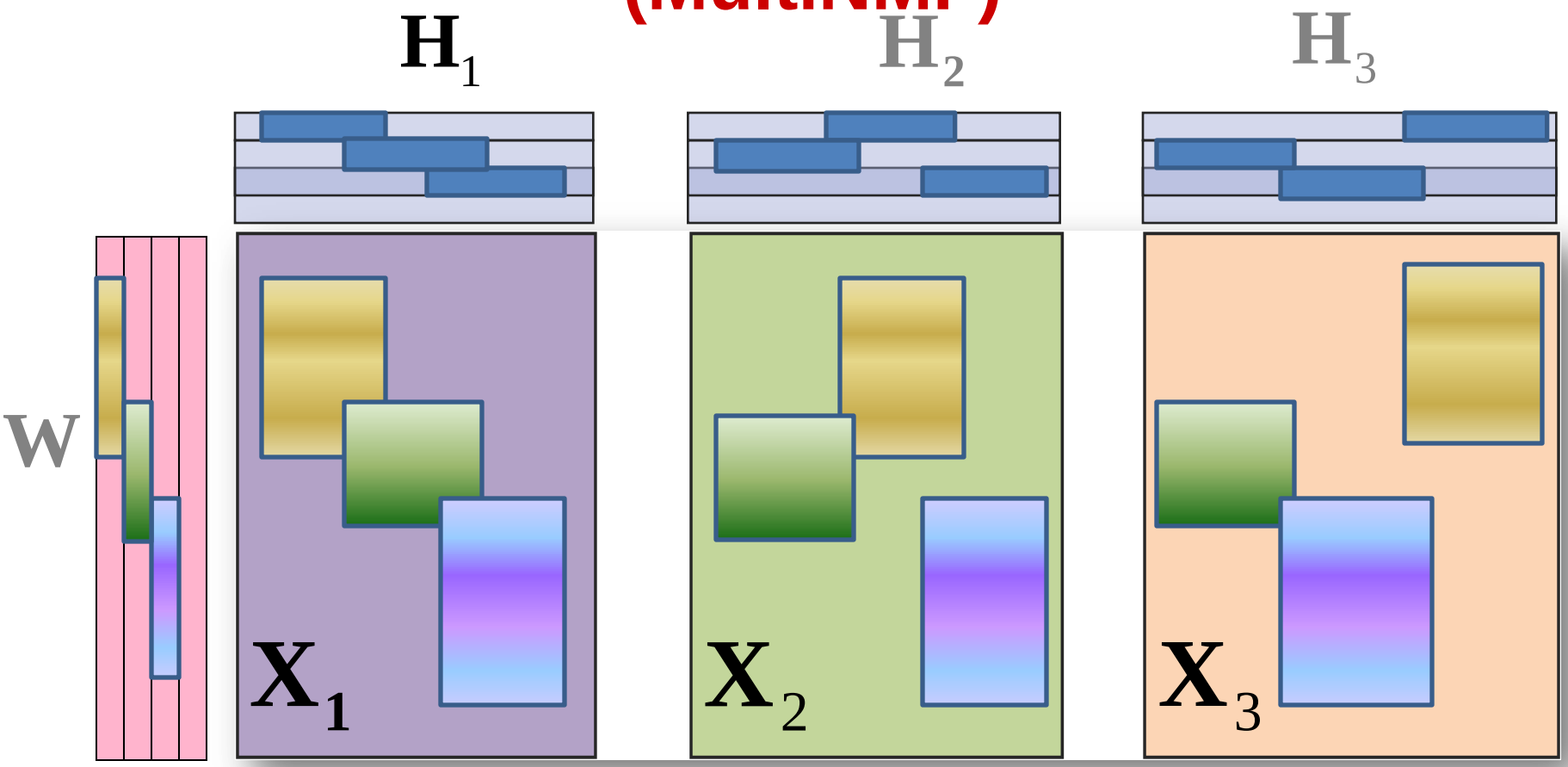
- Non-negative Matrix Factorization (NMF): Given a non-negative matrix X find non-negative matrix factors W and H such that X ≈ WH.

- Develop a **joint** Non-negative Matrix Factorization (NMF) approach

$$\min_{W,H_1,H_2,H_3\geq 0} \sum_{\mathbf{i=1,2,3}} \left\| \mathbf{X_i} - \mathbf{WH_i} \right\|_{\mathbf{F}}^{\mathbf{2}}$$

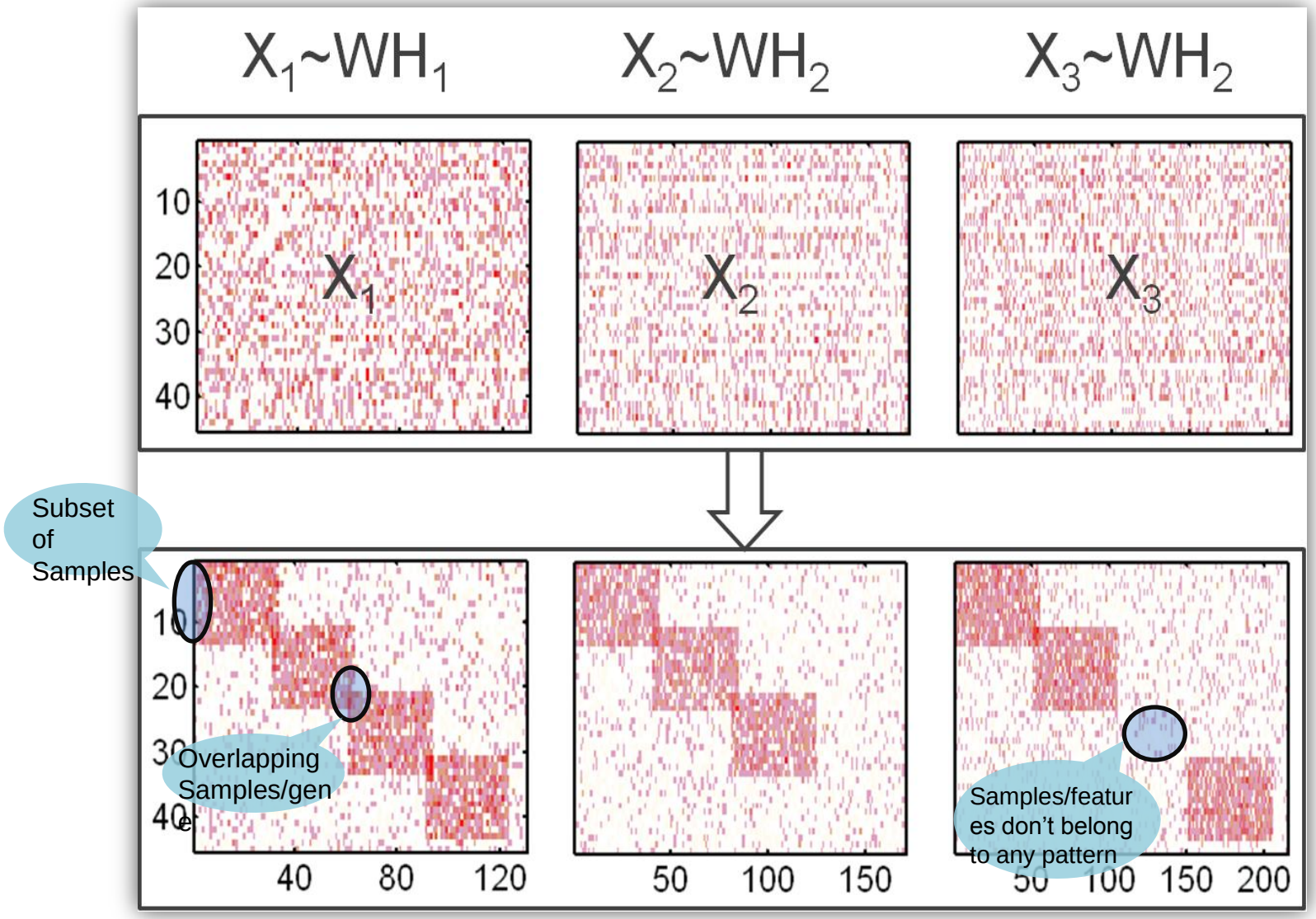
X_i: the data matrix of the *i*-th type of genomics data
W: the component matrix
H_i: the association or loading matrix of the *i*-th type of genomics d

Coherent patterns in matrices (MultiNMF)



$$\min_{W,H_1,H_2,H_3\geq 0} \sum_{i=1,2,3} \left\| X_i - WH_i \right\|_F^2$$

Test on a set of simulated examples



The Cancer Genome Atlas (TCGA) data

- The Cancer Genome Atlas (TCGA)
- We compiled DNA methylation (15418 markers), MicroRNA (799 miRNAs) and gene exrpession data (17811 genes) for **385** ovarian tumor samples.
- Preprocessing and normalization: 1) Transforming the expression data into positive values. 2) Scaling the sum of squares of each of the three matrices to be equal.
- We obtained 200 mRNA-Methylation-microRNA programs (or multi-dimensional modules) covering 2008 DNA methylation loci, 270 MicroRNAs and 2985 genes.
- Most of them are statistically significant (Permutation test: P-value<0.01).

Summary

- We have developed an effective method for ***simultaneously*** analyzing ***multi-dimensional*** genomic data.
- The joint NMF method can identify ***sample-specific multi-dimensional modules***.
- We have applied the proposed method to simulated data and the TCGA ovarian cancer data. The identified multi-dimensional modules showed strong ***biological relevance***.
  - GO enrichment analysis
  - Network and pathway analysis
  - Clinical analysis
  - ……

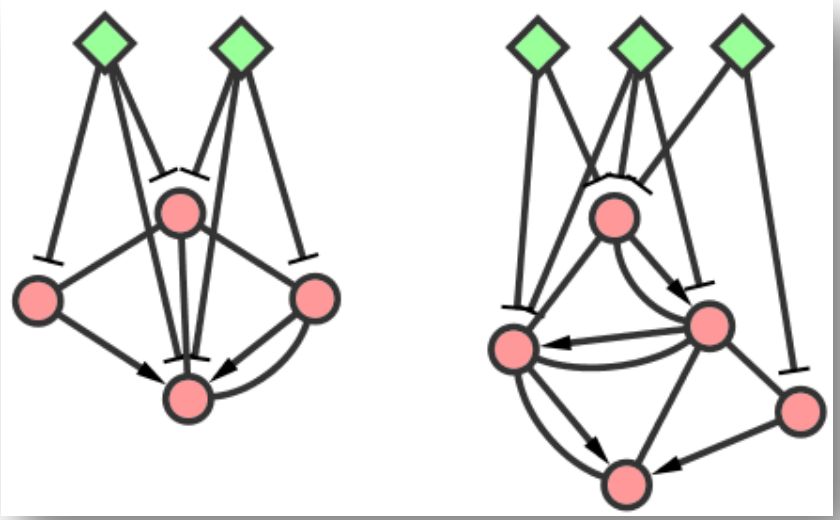
Part II: Simultaneous Integration of multiple types of data to discover miRNA-gene co-modules

Background

- MicroRNAs play crucial regulatory roles in **repressing mRNA translation** or **mediating mRNA degradation** by targeting mRNAs in a sequence-specific manner (Bartel, 2004).
- Great experimental and computational progress has been made on **the problems** of
  - identifying which genes encode miRNAs;
  - predicting the target genes of miRNAs within multiple genomes;
  - characterizing miRNA expression patterns based on microarray data.
- More and more labs are producing **simultaneous expression profiles of miRNA and mRNA on the same set of samples**

Background (cont.)

- How miRNAs, genes and proteins interact **on a systems level**, e.g. global miRNA regulation in cellular networks?
- **Little is known about the modular patterns** in miRNA-gene regulation systems.



How to identify miRNA-gene co-modules?

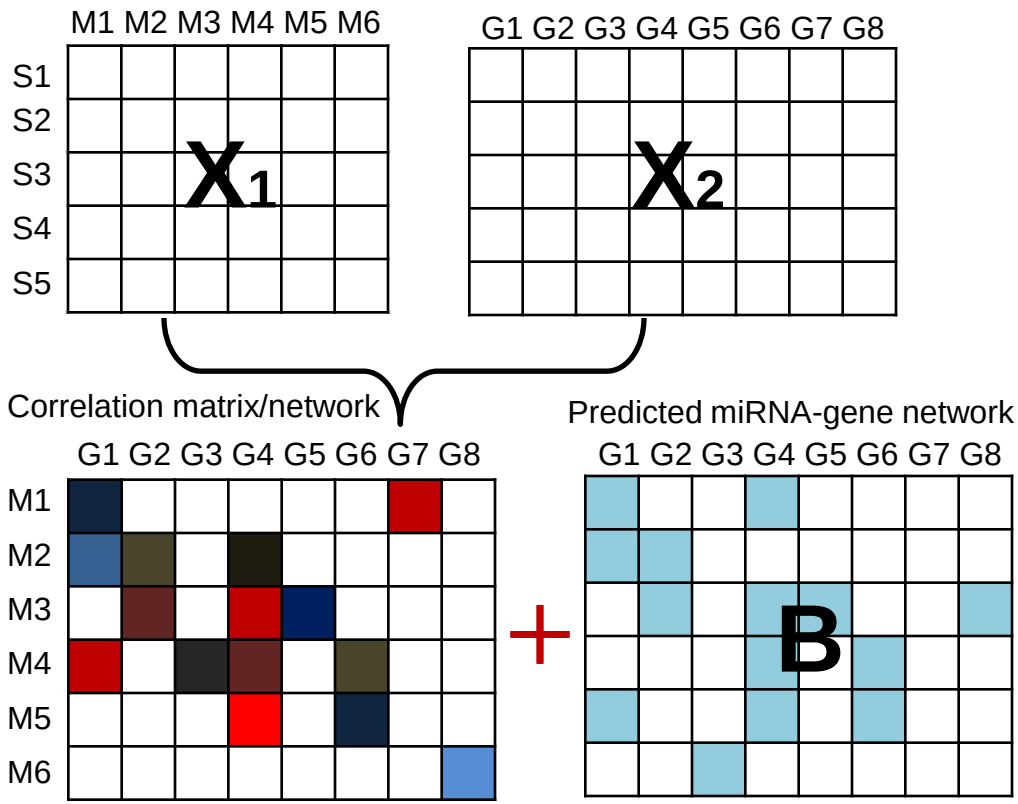
Challenges for miRNA-gene co-modules identification

- **<u>Multiplicity</u>:** One gene can be cooperatively regulated by multiple miRNAs and one miRNA can regulate a large number of genes.
- **<u>Specificity</u>:** The miRNA-mRNA target relationships differ among tissues and conditions.
- **<u>Anti-correlation or not</u>:** Although miRNAs physically interact with mRNAs, ultimately miRNA regulation affects the quantities of proteins in cells rather than the quantities of mRNAs. Thus, the expression levels of miRNAs are not always anti-correlated with those of their target genes.
- **<u>Noisy</u>:** The data are quite noisy and incomplete (e.g., predicated miRNA-gene interactions).

Related studies

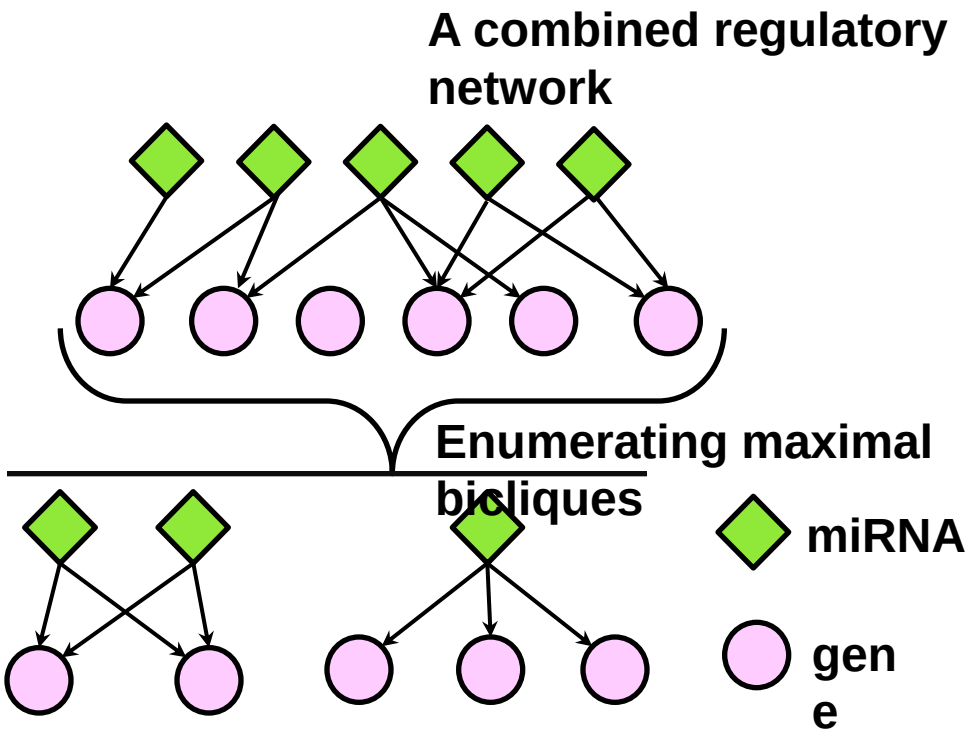
- **Only apply to** miRNA-gene targeting network (ISMB 2005)
- **combine with** miRNA, gene expression profiles (Bioinformatics, 2007).
- **No one considered the gene network**.
- **Enumerating bi-cliques** is sensitive to noise (ISMB 2005; BMC Genomics, 2009)
- Sequential integration of miRNA and gene expression profiles with miRNA-gene network.
- ……

Related studies——an example



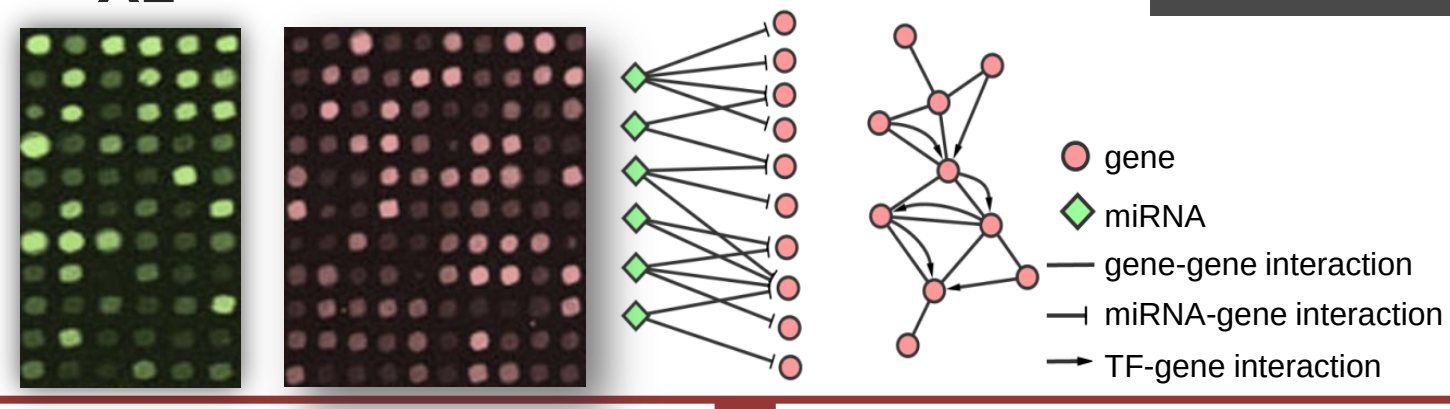
Peng et al. (2009), BMC Genomics

Related studies——an example (cont.)



Our method

Materials



Models

$$\min_{W,H_1,H_2 \geq 0} \sum_{I=1,2} \left\| X_I - WH_I \right\|_F^2 - \lambda_1 Tr(H_1 B H_2^T) - \lambda_2 Tr(H_2 A H_2^T) + \gamma_1 \left\| W \right\|_F^2 + \gamma_2 \left(\sum_j \left\| h_j \right\|_1^2 + \sum_j \left\| h_j \right\|_1^2 \right)$$

Multiplicative update algorithm

Multiplicative update algorithm

Algorithmic Framework for SNMNMF:

- **Step-1**: Initialize W, H_1 and H_2 with non-negative values, and set the iteration index $t = 0$.
- **Step-2:** Fix H_1 and H_2, solve the constrained problem

$$\min_{W \geq 0} \sum_{I=1,2} \|X_I - WH_I\|_F^2 + \gamma_1 \|W\|_F^2$$

That is, update W with

$$w_{ij} \leftarrow w_{ij} \frac{(X_1 H_1^T + X_2 H_2^T)_{ij}}{(W H_1 H_1^T + W H_2 H_2^T + \frac{\gamma_1}{2} W)_{ij}},$$

to find W^{t+1} such that $\mathcal{F}(W^{t+1}, H_1^t, H_2^t) \leq \mathcal{F}(W^t, H_1^t, H_2^t)$.

- **Step-3:** Fix W, solve the constrained problem

$$\min_{H_1, H_2 \geq 0} \sum_{I=1,2} \|X_I - WH_I\|_F^2 - \lambda_1 Tr(H_2 \widehat{A} H_2^T) - \lambda_2 Tr(H_1 \widehat{B} H_2^T + \gamma_2 (\sum_j \|h_j\|_1^2 + \sum_{j'} \|h_{j'}\|_1^2)) \quad (4)$$

That is, update H_1 and H_2 with

$$h_{ij}^1 \leftarrow h_{ij}^1 \frac{(W^T X_1 + \frac{\lambda_2}{2} H_2 B^T)_{ij}}{[(W^T W + \gamma_2 e_{k \times k}) H_1]_{ij}},$$
$$h_{ij}^2 \leftarrow h_{ij}^2 \frac{(W^T X_2 + \lambda_1 H_2 A + \frac{\lambda_2}{2} H_1 B)_{ij}}{[(W^T W + \gamma_2 e_{k \times k} H_2]_{ij}}, \quad (5)$$

to find H_1^{t+1} and H_2^{t+1} such that $\mathcal{F}(W^{t+1}, H_1^{t+1}, H_2^{t+1}) \leq \mathcal{F}(W^{t+1}, H_1^t, H_2^t)$.

- **Step-4:** Let $t \leftarrow t + 1$, repeat **Step-2–3** until convergence criteria are satisfied.

Related techniques in machine learning field

Semi-supervised constraint: The semi-supervised NMF method has been explored recently in machine learning field. The proposed method can considered as a generalization of this type of method.

- Tao Li, Chris Ding, and Michael Jordan. (2007) *Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization. in Proc. IEEE ICDM.*
- Cai, D. *et al. (2008) Non-negative matrix factorization on manifold. in IEEE ICDM, 63-72.*
- Gu, Q., and Zhou, J., (2009) *Local learning regularized nonnegative matrix factorization*, *Proceedings of IJCAI.*

Sparsity constraint: Several different types of sparsity constraints have been proposed for NMF problem. Here we adopted the one suggested by Kim and Park (2007).

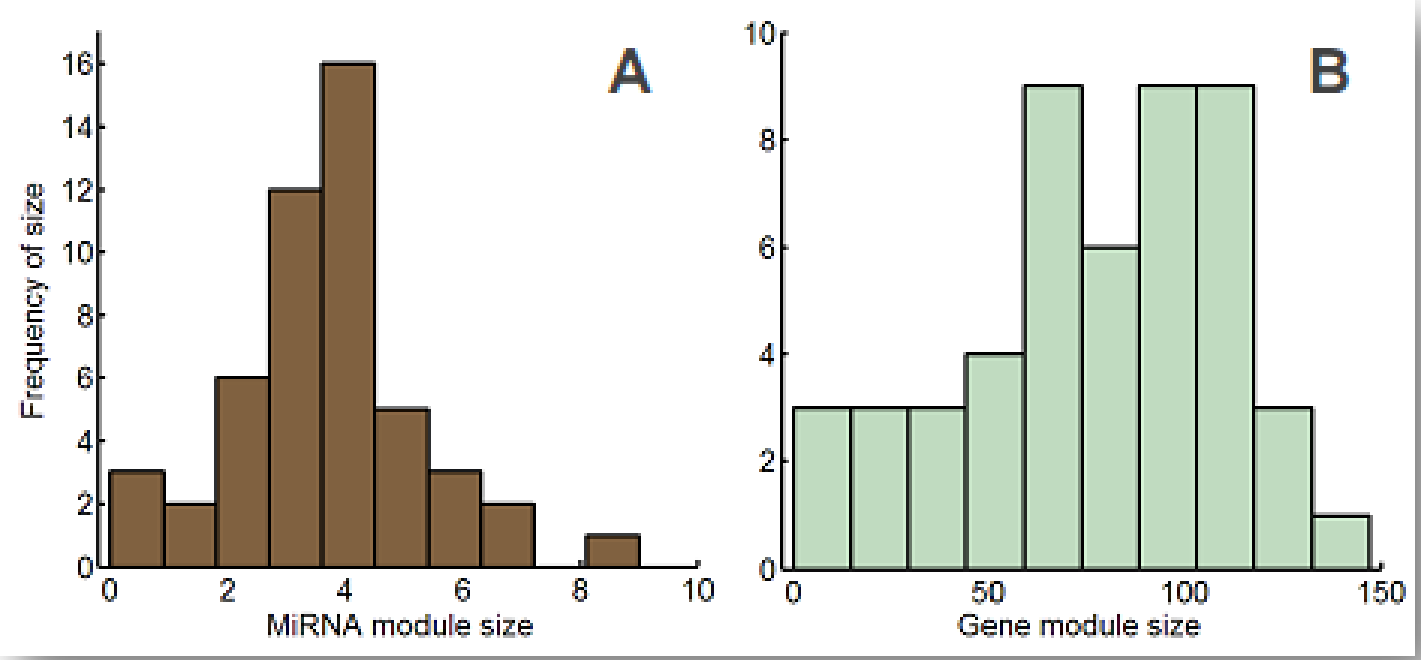
Application

- **miRNA vs. gene expression** profiles of **TCGA** Ovarian cancer data (X1 and X2)
- Predicted **miRNA-gene interaction** network (B) **(MicroCosm website)**
- **Gene network** (A)——protein interaction network (Bossi and Lehner, 2009) and protein-DNA network (TRANSFAC)

- Parameter settings: *k = 50 and ……*

Results

- 49 miRNA-gene co-modules



Size distribution of co-modules with 3.8 miRNAs and 78 gene for each co-module on average.

Validation/Functional analysis

1) Statistical significance test/Permutation tests

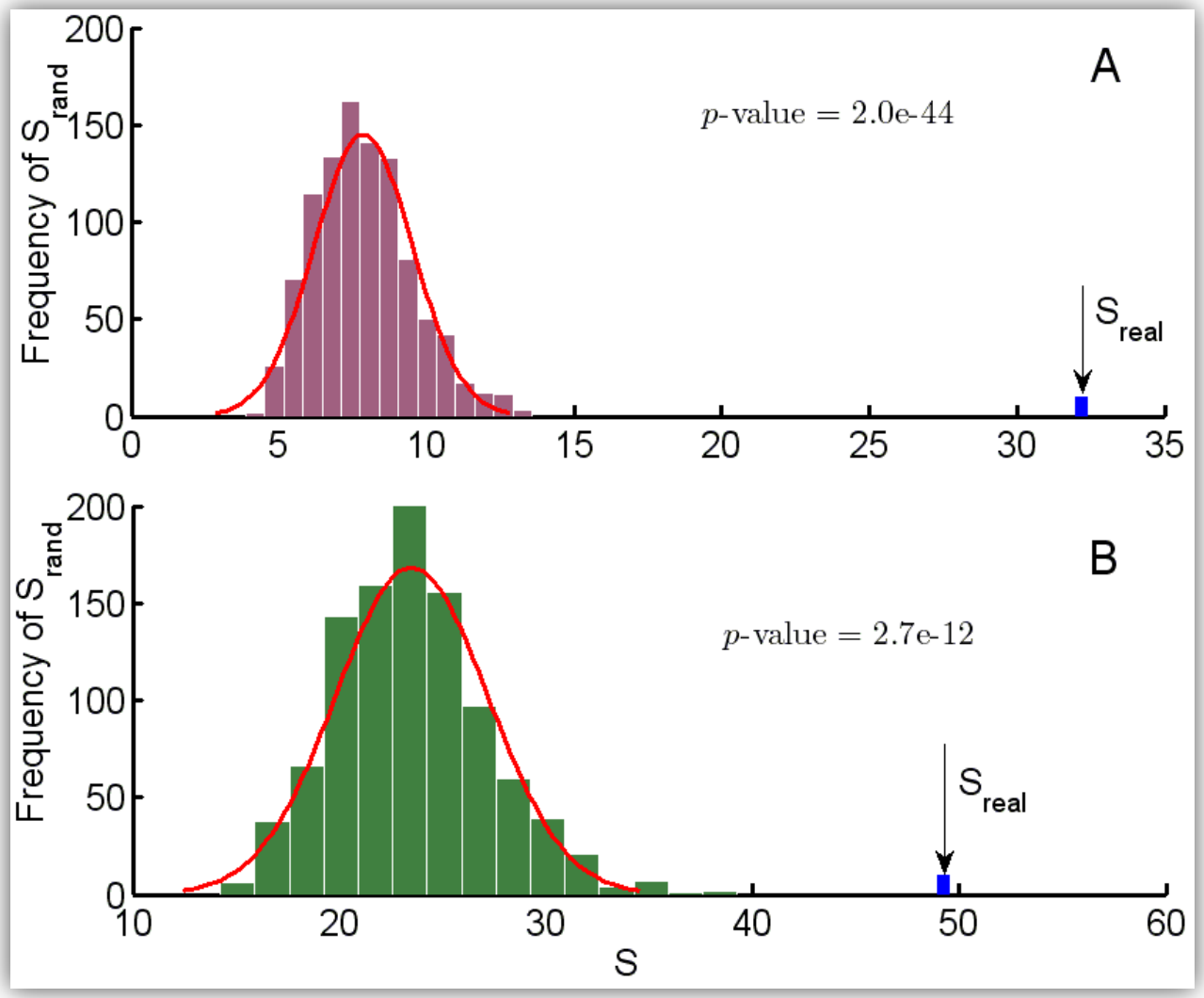
2) miRNA clusters enrichment analysis of miRNA modules

3) Functional enrichment analysis of gene modules

4) Network and pathway analysis based on IPA and literature review

5) Comparison with other methods

Permutation tests



miRNA modules are enriched with miRNA clusters

No.	*q*-value	Overlap miRNAs	Loci[b]	FS
10	0.002	mir-449b, mir-449a	5q11.2	Yes
	0.001	mir-34b*, mir-34c-5p	11q23.1	Yes
14	0.002	mir-143, mir-145	5q32	Yes
16	3.94e-05	mir-182*, mir-96, mir-183	7q32.2	Yes
17	0.001	mir-144, mir-451	17q11.2	Yes
18	0.001	mir-452, mir-224	Xq28	No
19	0.005	mir-30b*, mir-30d*, mir-30d, mir-30b	8q24.22	Yes
20	1.97e-5	mir-96, mir-183, mir-182	7q32.2	Yes
42	0.005	mir-199a-5p, mir-214	1q24.3	Yes
46	0.001	mir-144, mir-451, mir-144*	17q11.2	Yes
48	6.78e-12	mir-513b, mir-513c, mir-508-3p, mir-506, mir-507, mir-509-3-5p, mir-514, mir-509-3p, mir-509-5p	Xq27.3	No
50	0.008	mir-502-3p, mir-500*	Xp11.23	No

miRNA module is enriched with miRNA clusters (cont.)

- For example, in co-module 10, two of the four member miRNAs (mir-449a and 449b) **belong to a miRNA cluster on chromosome 5q11.2**, and the other two (miR-34b* and 34c-5p) belong to a cluster on chromosome 11q23.11.

- In a recent study, miR-449a and 449b have been reported to **have the tumor suppressing function** by regulating Rb/E2F1 activity (Yang et al., 2009). In addition, miR-34b* and 34c-5p were reported to be targeted by p53 and they **cooperatively control cell proliferation in ovarian cancer** (Corney et al., 2007).

Gene module is enriched with known functional sets (GO biological process)

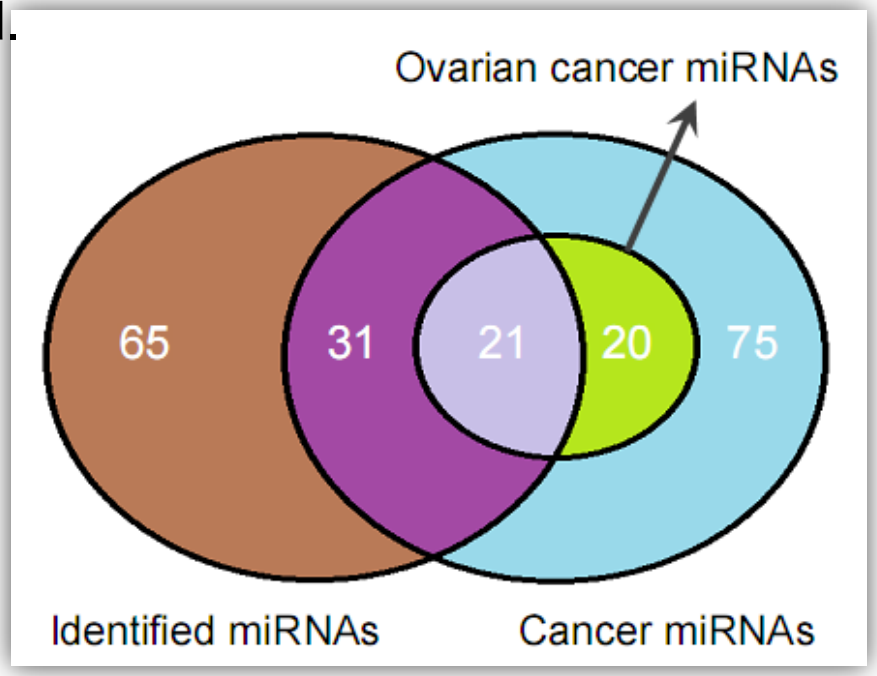
- **Twenty-six (53.1%)** modules have at least one over-represented GO biological process terms with an FDR-corrected q-value < 0.05.

- When we similarly assess a set of random modules, only 3.0%(±2.4%) are enriched in GO biological processes.

miRNA-gene co-module is enriched with cancer miRNAs and genes

Literature survey alysis.

- Overlap test.
- 69.4\% of the modules contain at least two miRNAs that are known to be cancer-related.



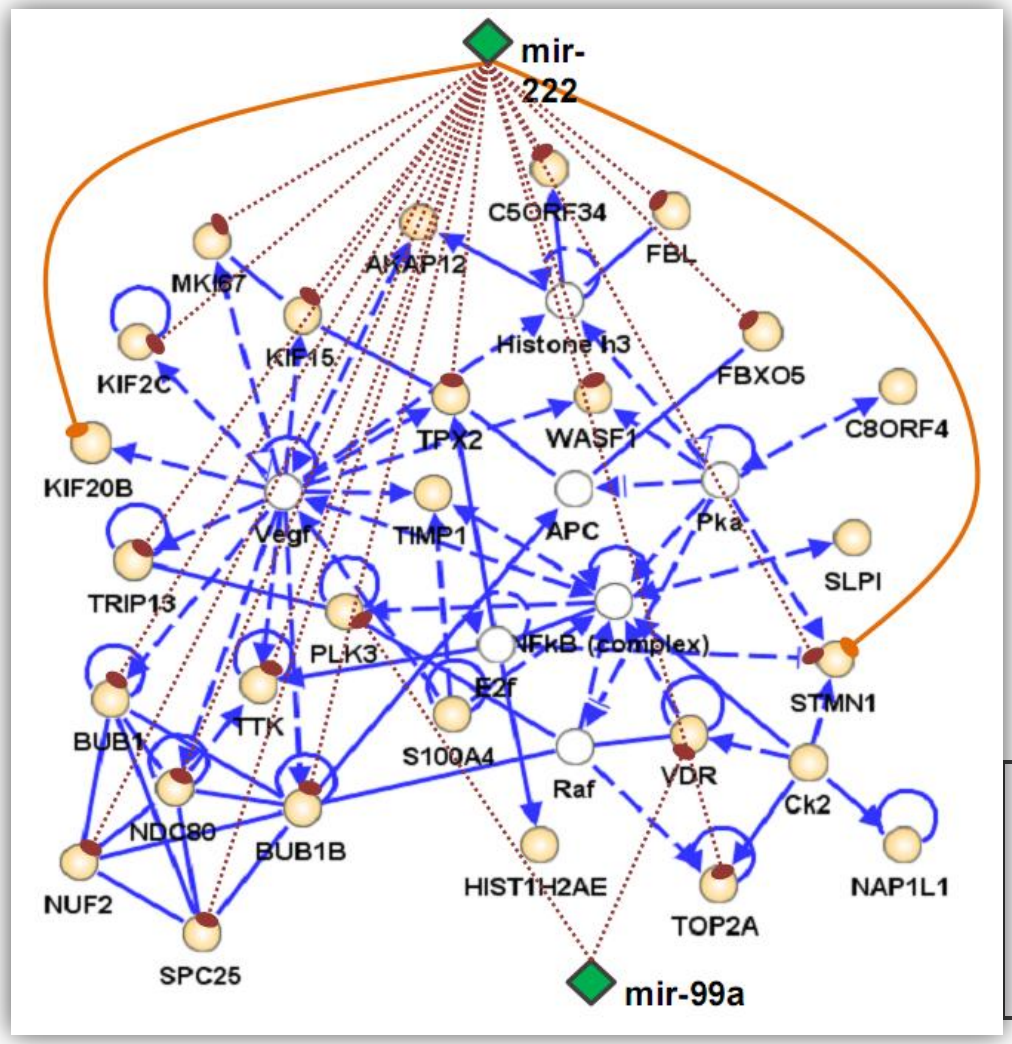
IPA Function and Disease analysis shows

- Most of the modules (63.3\%) are highly enriched in cancer genes (q-value<0.05).
- Moreover, 10 of the modules are significantly enriched in ovarian cancer genes.

Table 2. Functional analysis of selected miRNA-gene co-modules. No.: the serial number of co-modules. CG: cancer genes; PT: permutation test with p-value$\times 50 < 0.05$; Num*: indicate the number of cancer related with this miRNA module as well as the size of this miRNA module. OC miRNAs: ovarian cancer miRNAs are identified in this module.

No.	GO biological process terms	CG	PT	Cancer miRNAs	Num*	OC miRNAs
7	Immune system process; Regulation of cell activation; Regulation of cell proliferation	Yes	4.4e-165	mir-142-5p, mir-142-3p, mir-21*	3/3	mir-21*
15	Immune response; Immune system process; Defense response; Inflammatory response; Response to external stimulus;cell activation	Yes	8.6e-254	mir-142-5p, mir-142-3p, mir-150, mir-146a	4/4	
23	Negative regulation of immune system; Response to external stimulus; Regulation of cell division; Cell adhesion; Regulation of cell migration; Cell Communication;	Yes	1.9e-151	mir-22, mir-199a-5p, mir-145, mir-10b	4/5	mir-22, mir-199a-5p, mir-145, mir-10b
25	Calcium-dependent cell-cell adhesion; Synaptic transmission; Cell adhesion; Extracellular structure organization		4.2e-4	mir-10b*, mir-135b, mir-10b	3/4	mir-10b*, mir-10b
32	Cell cycle process; Organelle organization; Nuclear division; Cell cycle; Cell division;	Yes	2.0-44	mir-133b, mir-145	2/2	mir-145
37	Inflammatory response; Defense response; Immune response; Regulation of apoptosis; Cell chemotaxis; Regulation of DNA binding; Cellular response to stimulus; Regulation of cell death; Anti-apoptosis;	Yes	3.1e-47	mir-223, mir-146a	2/2	mir-223
40	Cell cycle; Cell division; Nuclear division; Mitosis; Organelle fission; Microtubule-based process;	Yes	2.7e-12	mir-99a, mir-135b, mir-222, mir-205	4/4	mir-99a
42	Reproductive developmental process; BMP signaling pathway; Cell differentiation; Regulation of cell development;	Yes	7.5e-136	mir-214,mir-376a, mir-199b-3p, mir-127-3p, mir-199a-5p,	5/7	mir-214, mir-199b-3p, mir-199a-5p, mir-127-3p,

Network analysis of co-module reveals co-regulated signals

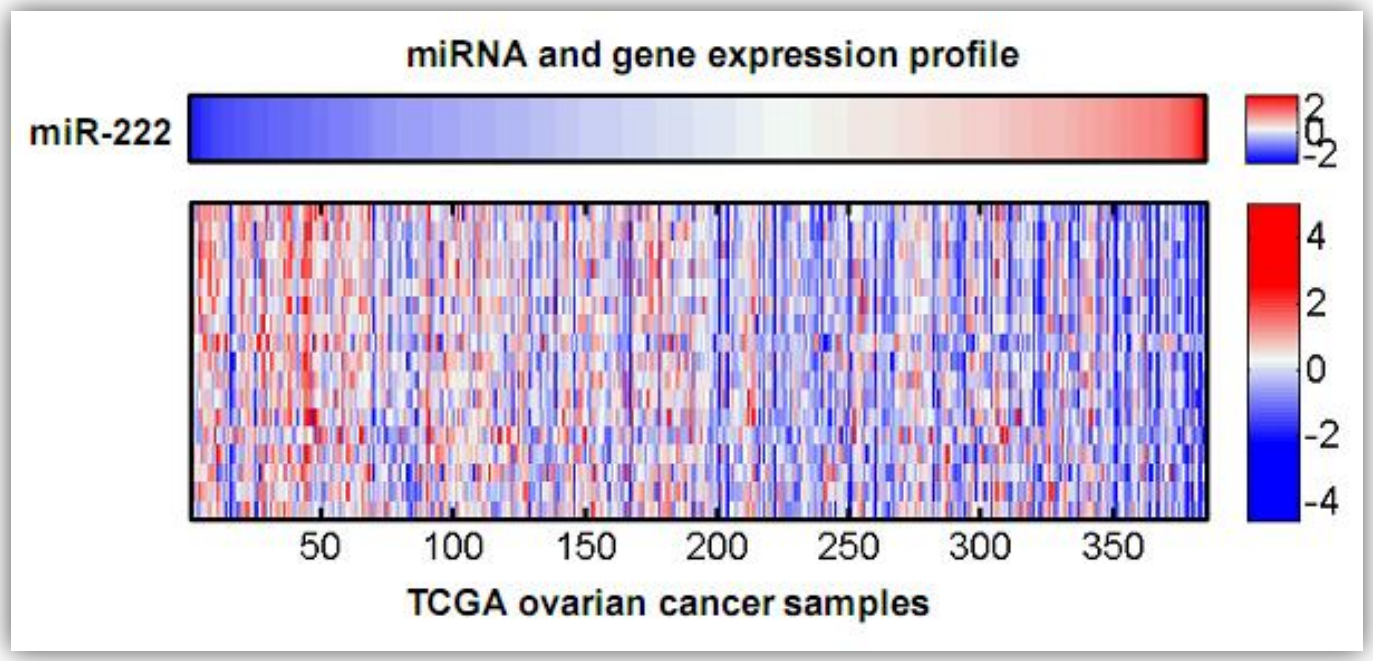


Gene interactions

miR-gene interaction

miR-gene anti-correlation

miRNA-222 expression profile negatively regulates genes' expression in the network



Conclusion

- We propose a method for identification of **miRNA-gene co-modules**
- **Simultaneous** integration
- **Integrating the gene network**
- **Sparsity** constraints
- **Can be apply to** other types biological problems