Sample size considerations for the efficiency of extracting regulatory connections from a combined miRNA and gene expression data set

Smriti Shridhar

Chair of Bioinformatics
Boku University Vienna, Austria.

BioToP Biomolecular Technology of Proteins

FWF Der Wissenschaftsfonds.

BOKU VIBT Vienna Institute of BioTechnology

miRNAs

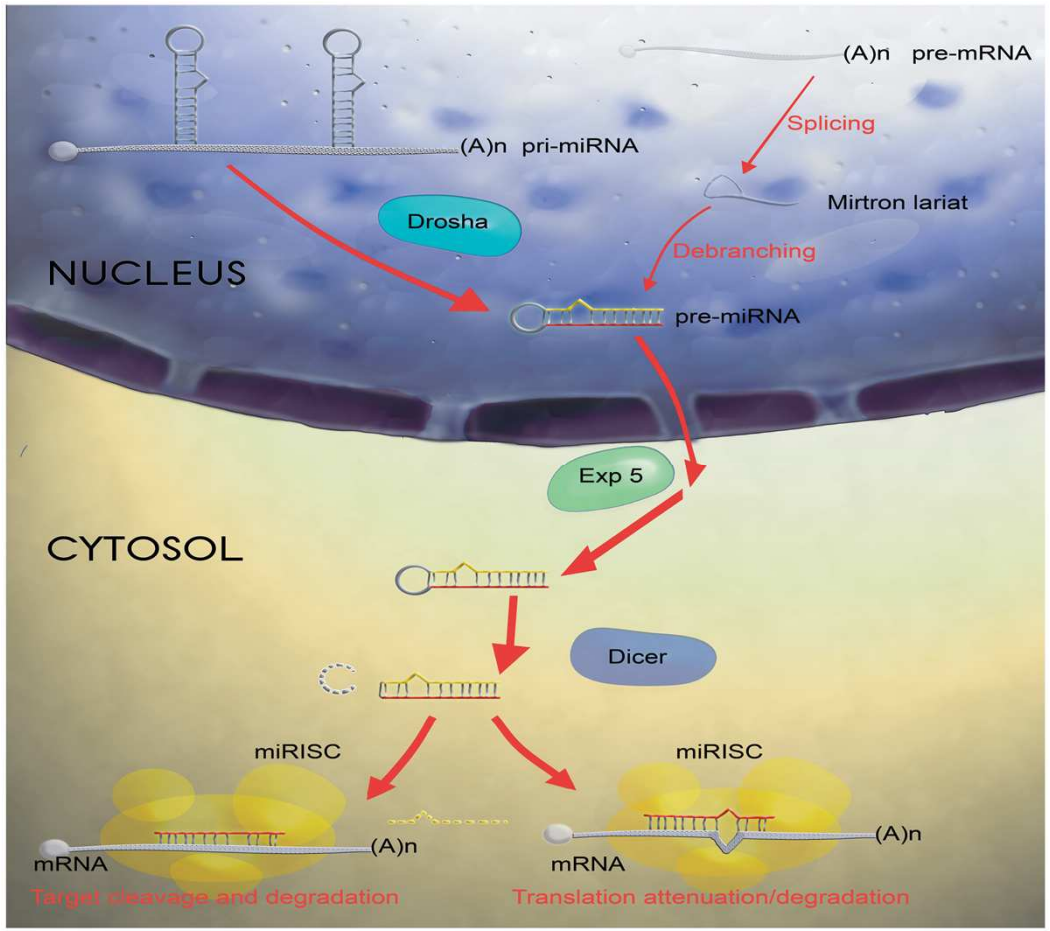
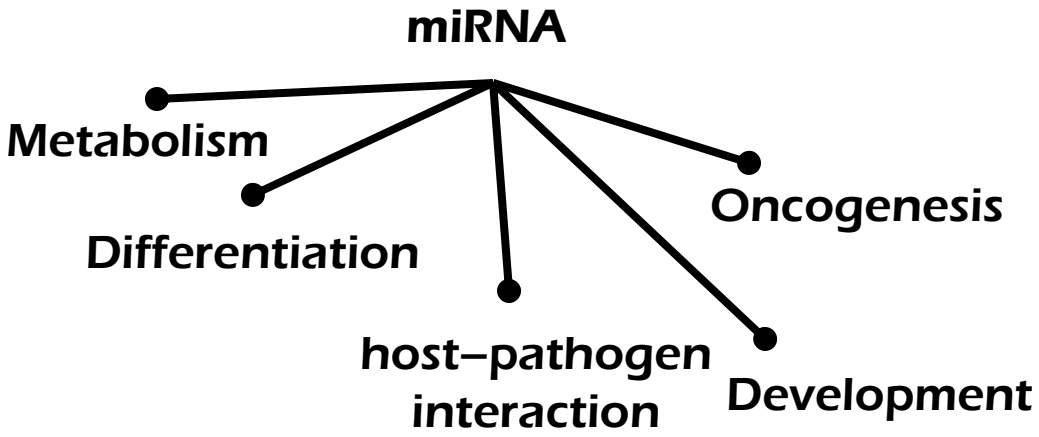
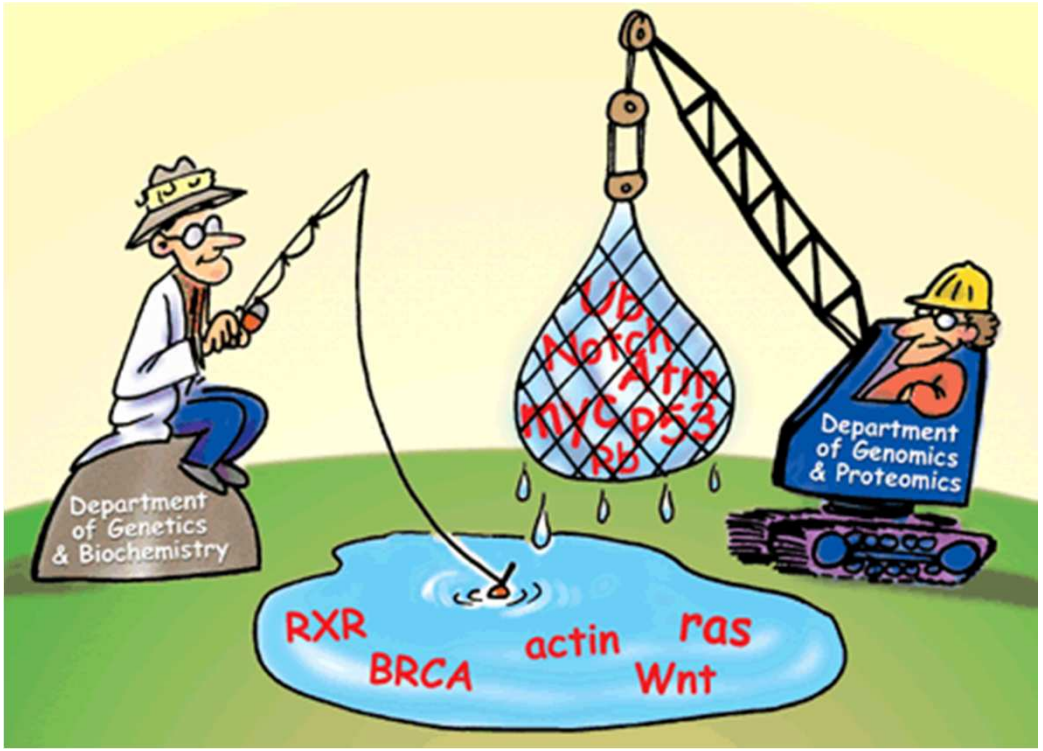


Fig. miRNA biogenesis pathway

(Mendes et al., NAR, 2009)

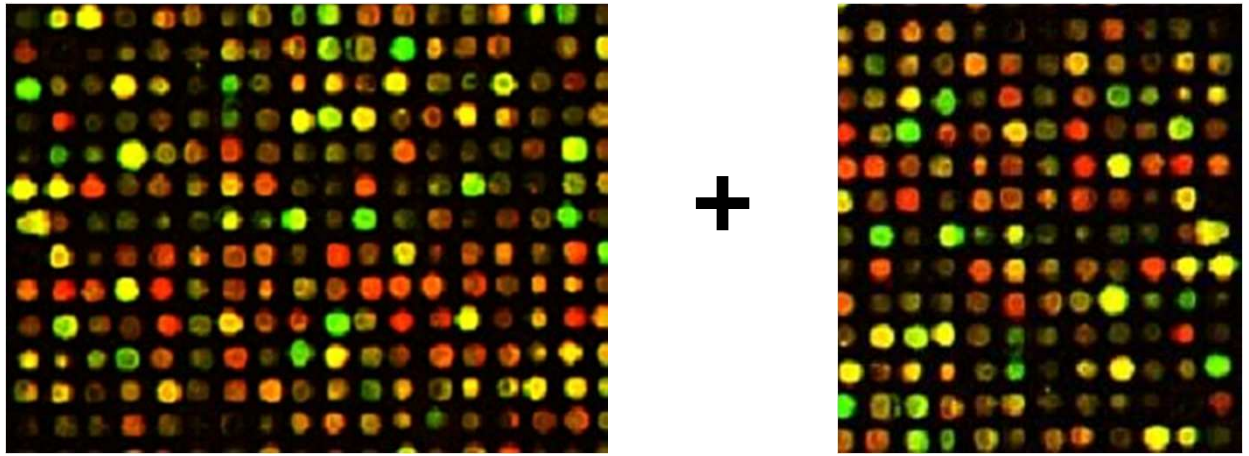
Observed roles of miRNAs





Fields, S. (Science, 2001)

Integrating gene and miRNA expression data



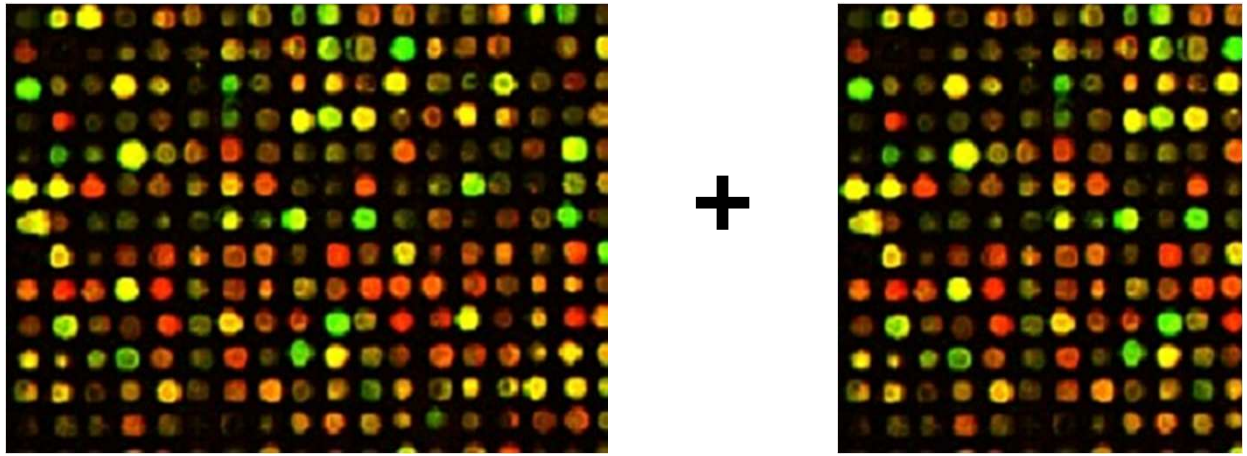
identification of biologically relevant signals!

Challenges to integrating high-dimensional data

- Curse of dimensionality
  → variable selection

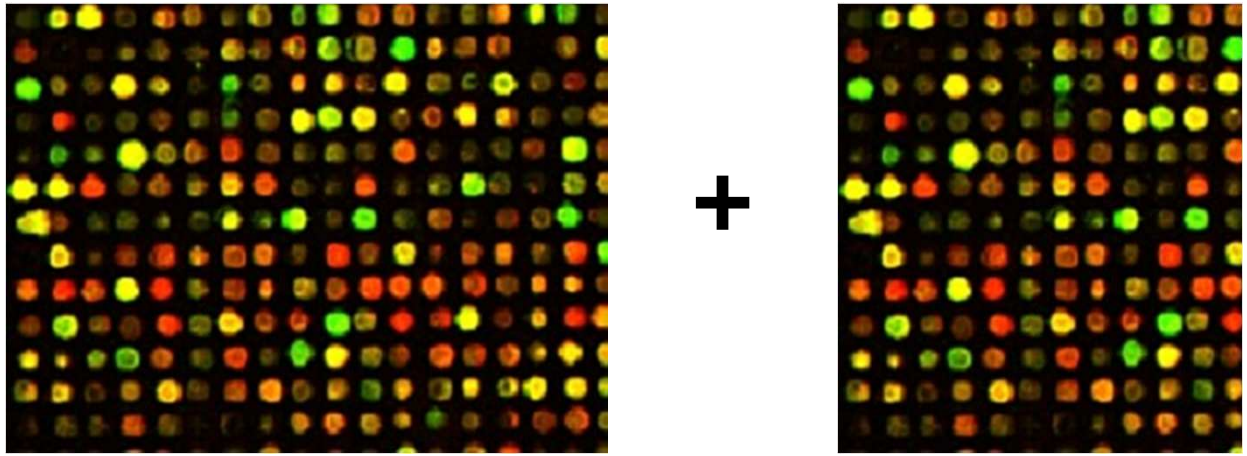
- Combinatorial explosion
  → filter using sequence analysis

Integrating gene and miRNA expression data



- if signal is strong -> smaller no. of samples

Integrating gene and miRNA expression data



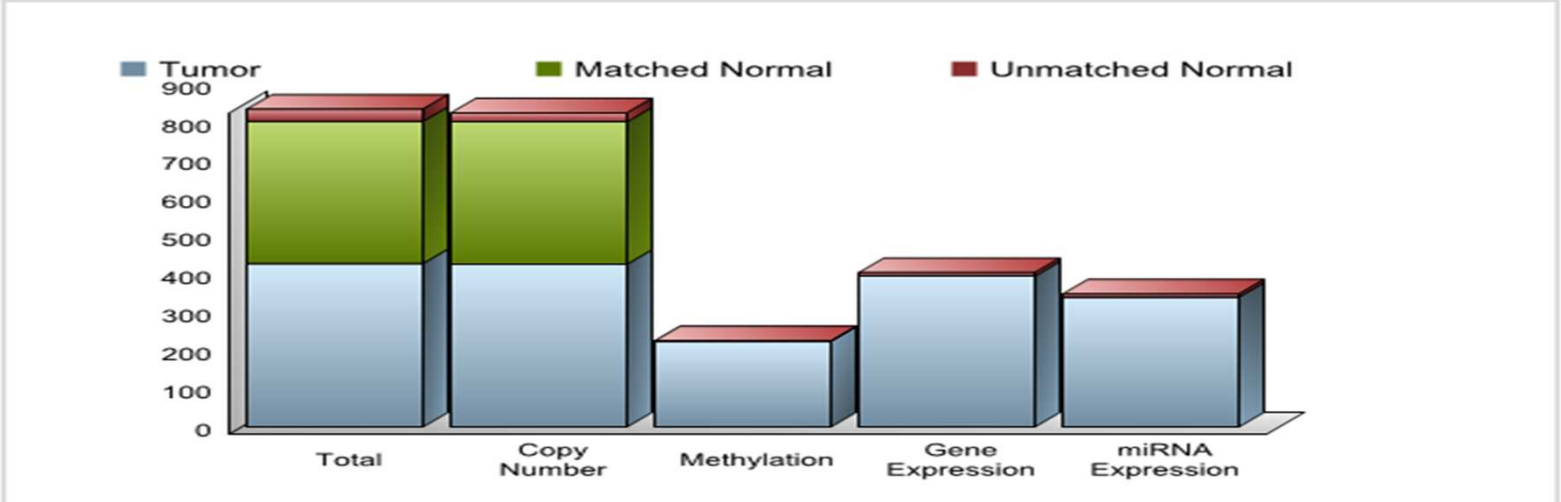
Number of samples required?

Datasets used for comparing effect of sample sizes

- Glioblastoma Multiforme (Dataset 1)

  - malignant brain tumor
  - accounts for about 15 percent of all brain tumors
  - poor prognosis

Glioblastoma multiforme [GBM]	Number of Samples				
	Total	Copy Number	Methylation	Gene Expression	miRNA Expression
Tumor	536	534	281	495	426
Matched Normal	469	469	0	0	0
Unmatched Normal	40	30	1	11	10

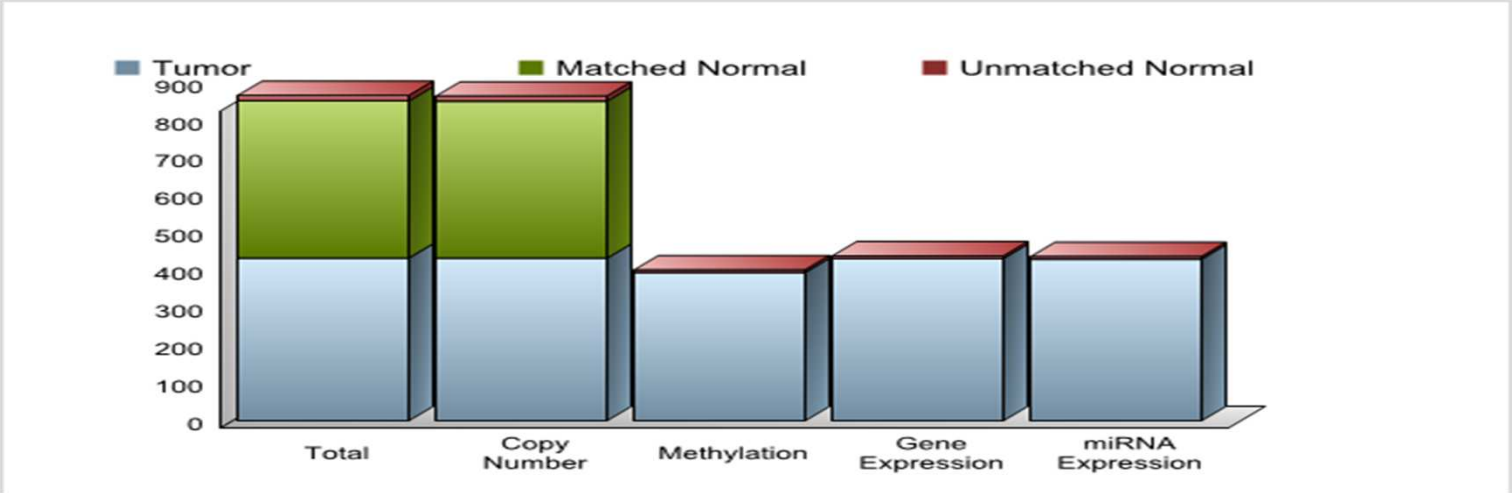


Datasets used for comparing effect of sample sizes

- Ovarian serous cystadenocarcinoma (Dataset2)

  - a type of epithelial ovarian cancer
  - accounts for about 90 percent of all ovarian cancers
  - need for effective screening tests

Ovarian serous cystadenocarcinoma [OV]	Number of Samples				
	Total	Copy Number	Methylation	Gene Expression	miRNA Expression
Tumor	586	585	534	584	582
Matched Normal	569	569	0	0	0
Unmatched Normal	19	18	9	9	8



Comparision of datasets

- **Glioblastomas**

- **Serous cystadenocarcinoma**

- differ in
  - location within the central nervous system
  - sex distribution

Which ultimately affects-

a) tendency for progression
b) response to treatments

Common approaches to integrating high-dimensional data

- Factor analysis, component no. selected by
  - cross validation
  - information criteria (eg. BIC, AIC)

- Variable selection
  - sPLS with lasso penalization *(Cao et al., Bioinformatics, 2009)*

- Clustering based on Gibbs sampling *(Bonnet et al.,Bioinformatics, 2010)*
  - tight clusters

Learning Module Network Algorithm

Algorithm: Learning Module Network *(LeMoNe)*

Stage1: - two-way clustering of genes and samples by Gibbs sampling

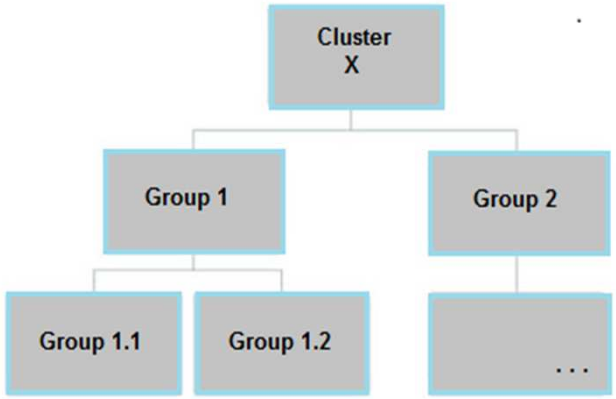
- tight clusters

Stage2: - integration of several heterogeneous regulators

- inferring a prioritized list of potential regulators for each cluster

Learning Network Modules Algorithm

Integration of regulators



Learning Network Modules Algorithm

Inferring a prioritized list of regulators

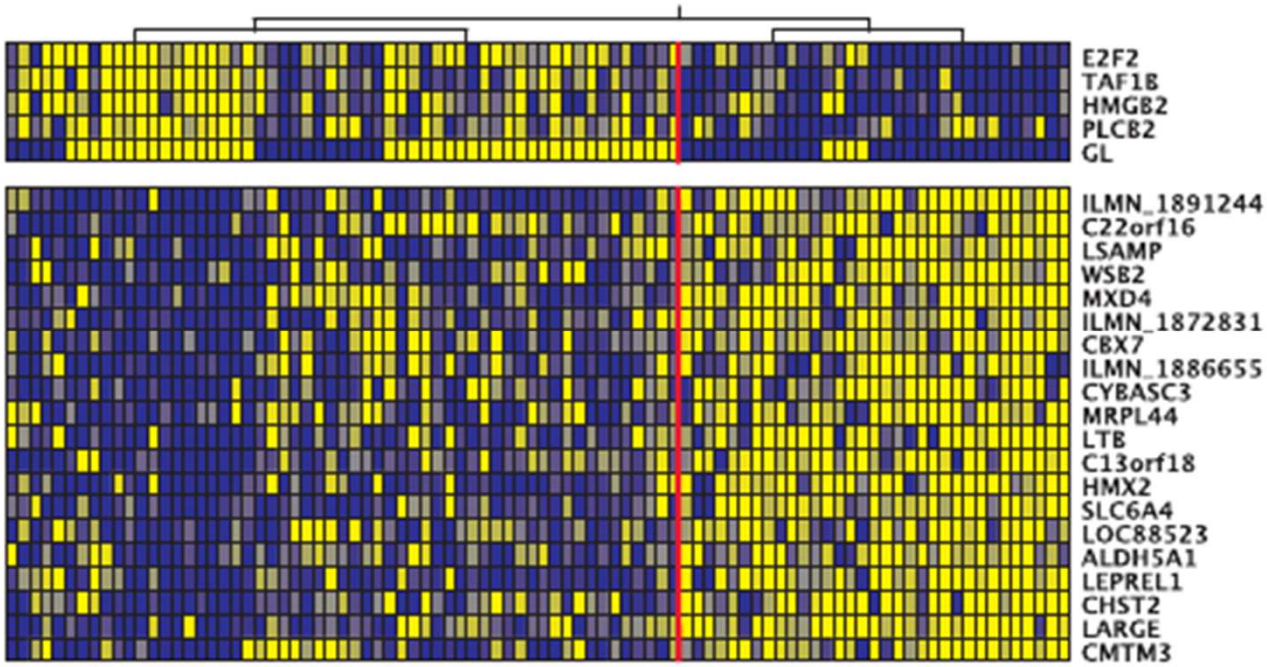


Fig. An example Module. The upper panel represent the high-scoring regulators. The lower panel represent the module genes. Each column represent a different sample. The hierarchical tree on top of the figure is one of the trees used to assign the regulators.

Module network inferred by the LeMoNe algorithm

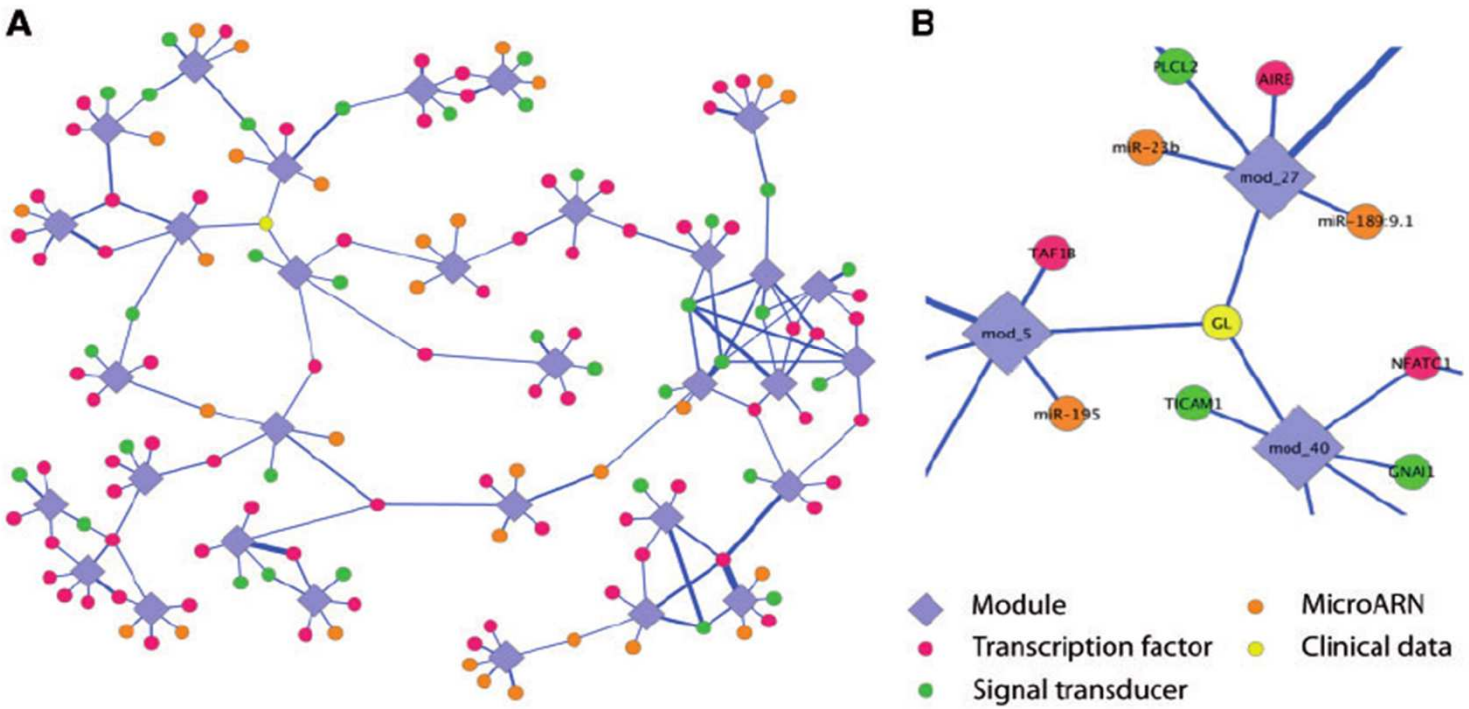


Fig. (A) Clusters of co-expressed genes have diamond shapes, while regulators are symbolized by circles.(B) Zoom on the module network representation.

Assessment of analysis performance

- simulation studies
- ‘story building’:
  validation of results for individual miRNA – target pairs
- relation to functional annotation (e.g., GeneOntology)
- comparison with independent miRNA–target predictions
  (e.g., from sequence based analysis)

Sample size performance

- Evaluation by

  - enrichment of miRNAs known to be associated with Glioblastoma

    (Ruepp et al., Genome Biology, 2010)

  - enrichment of predicted miRNA targets

    (Kozomara et al., NAR, 2011)

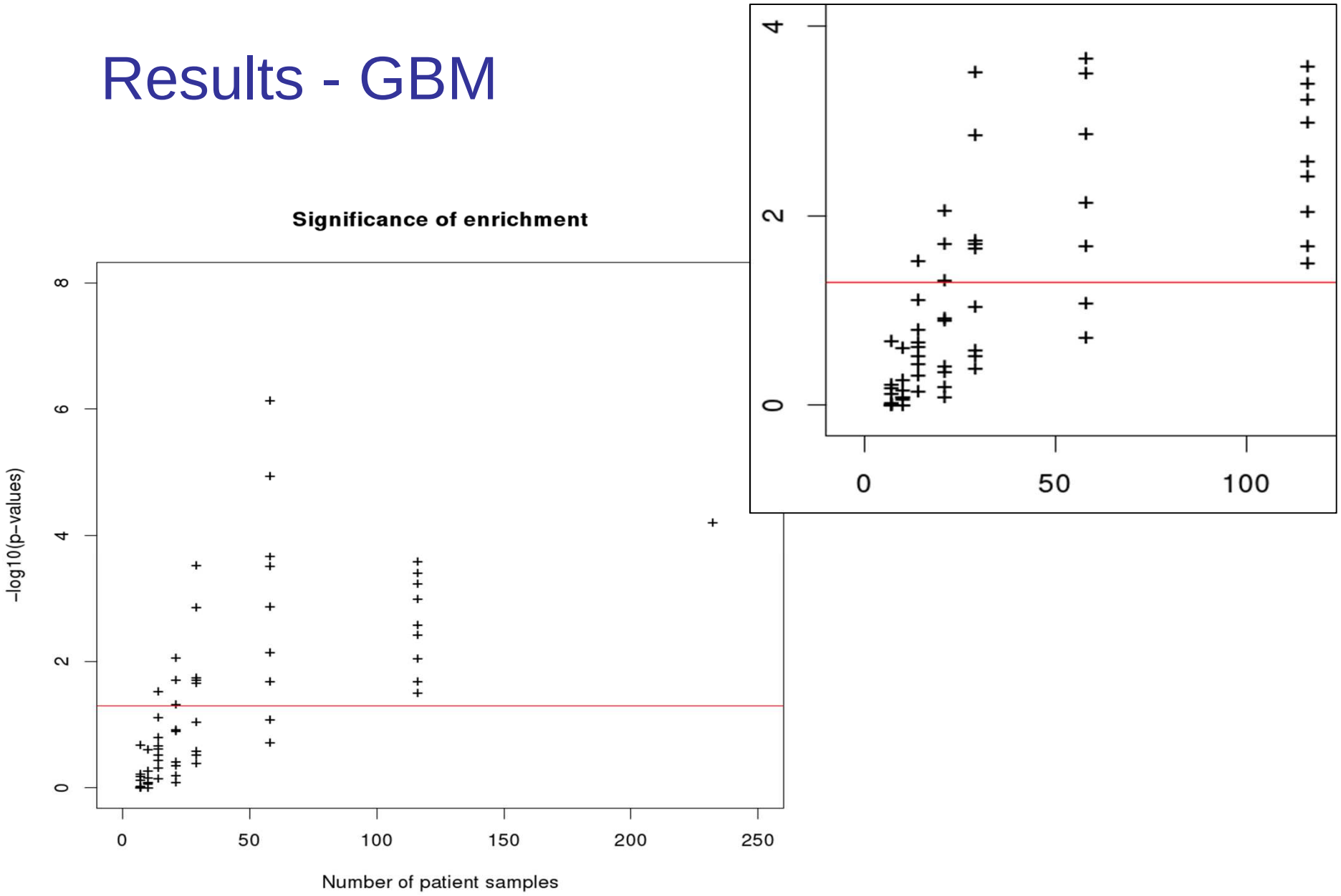
- Enrichment significance by

  - Fisher’s exact test

Data

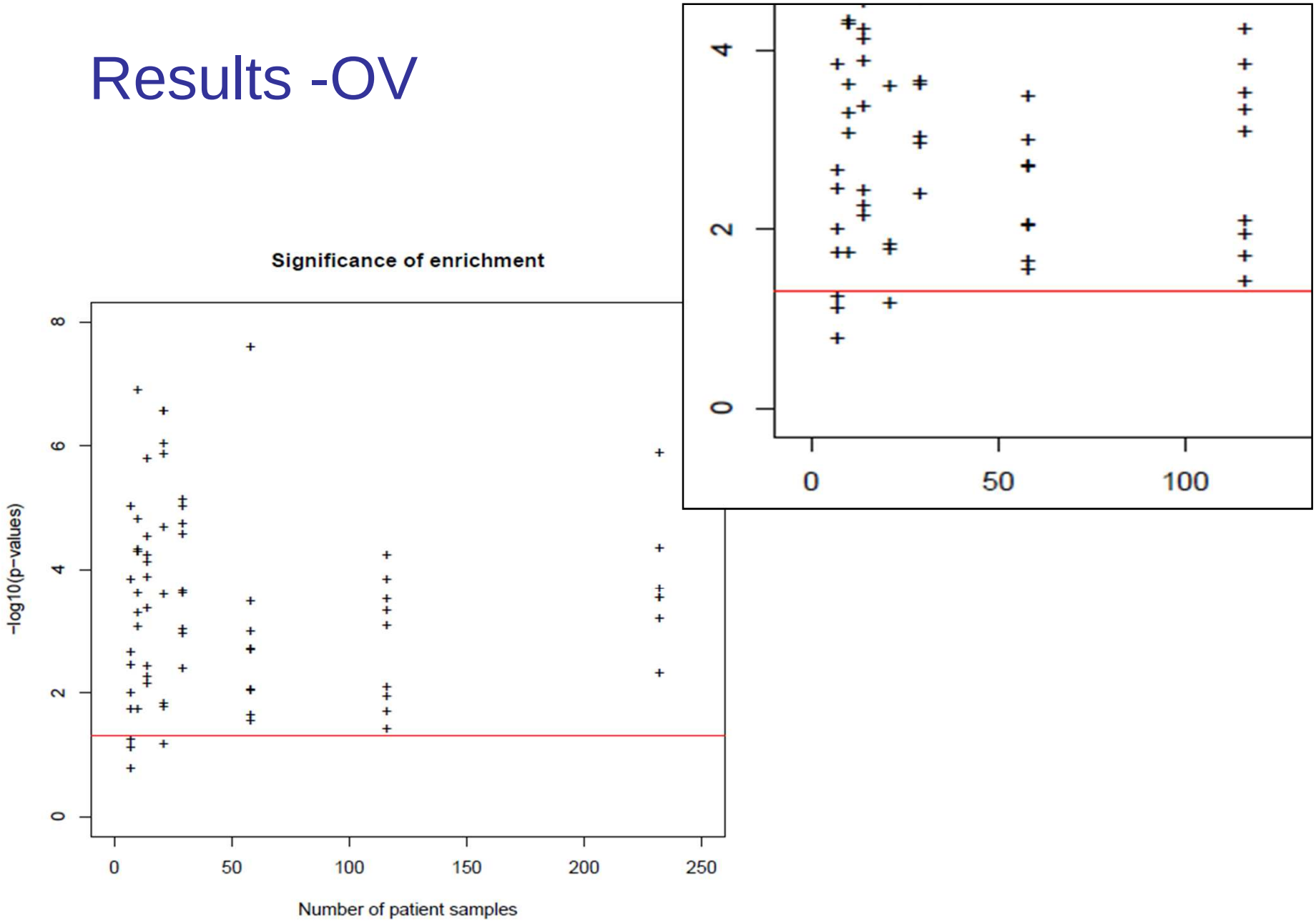
Dataset	No. of expression profiles used		No. of Patient Samples
	Genes	miRNAs	
Glioblastoma	11925	524	232
Ovarian Cancer	17618	799	232

Results - GBM



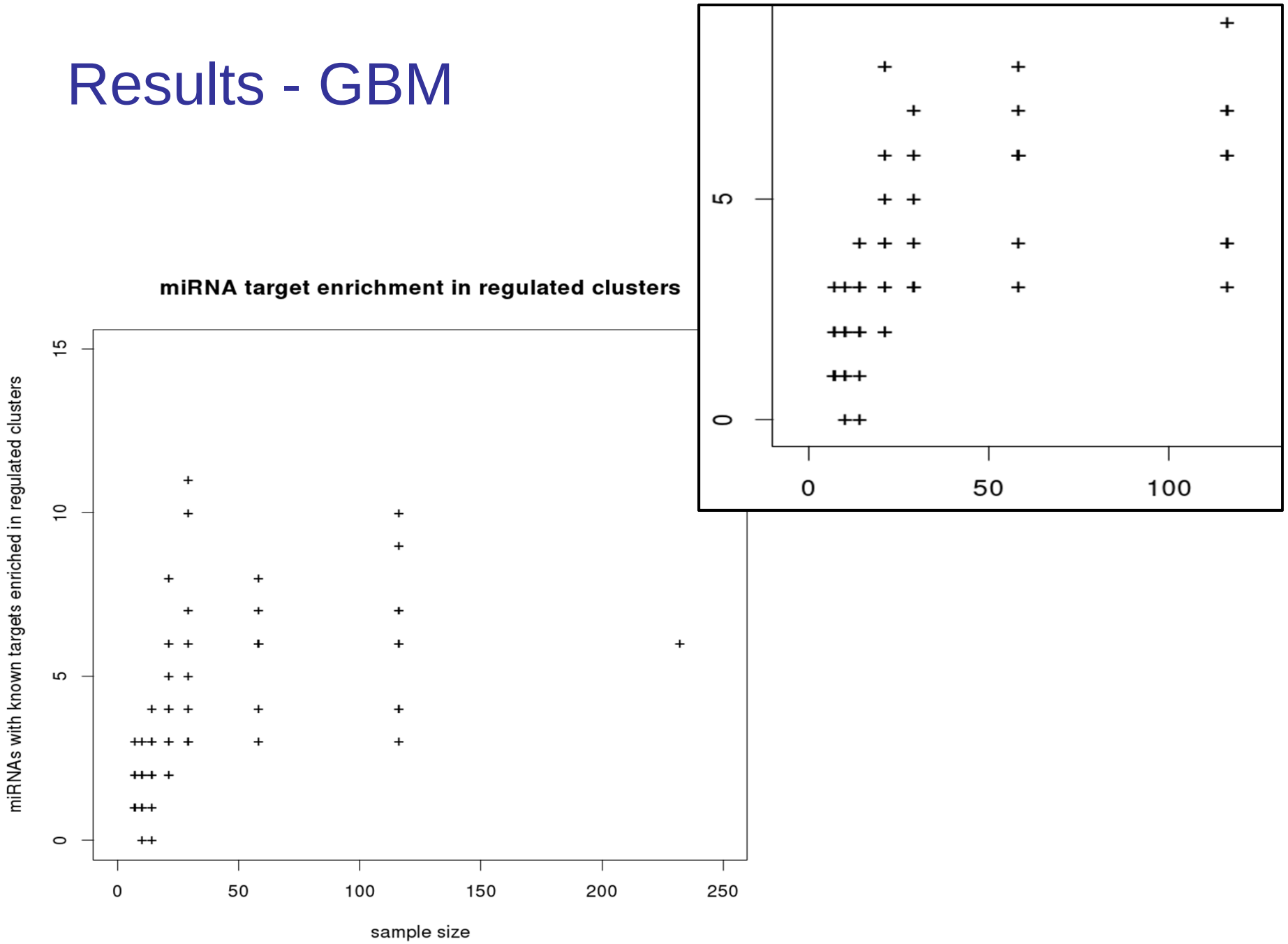
Significance of Enrichment of known Glioblastoma associated miRNAs

Results -OV

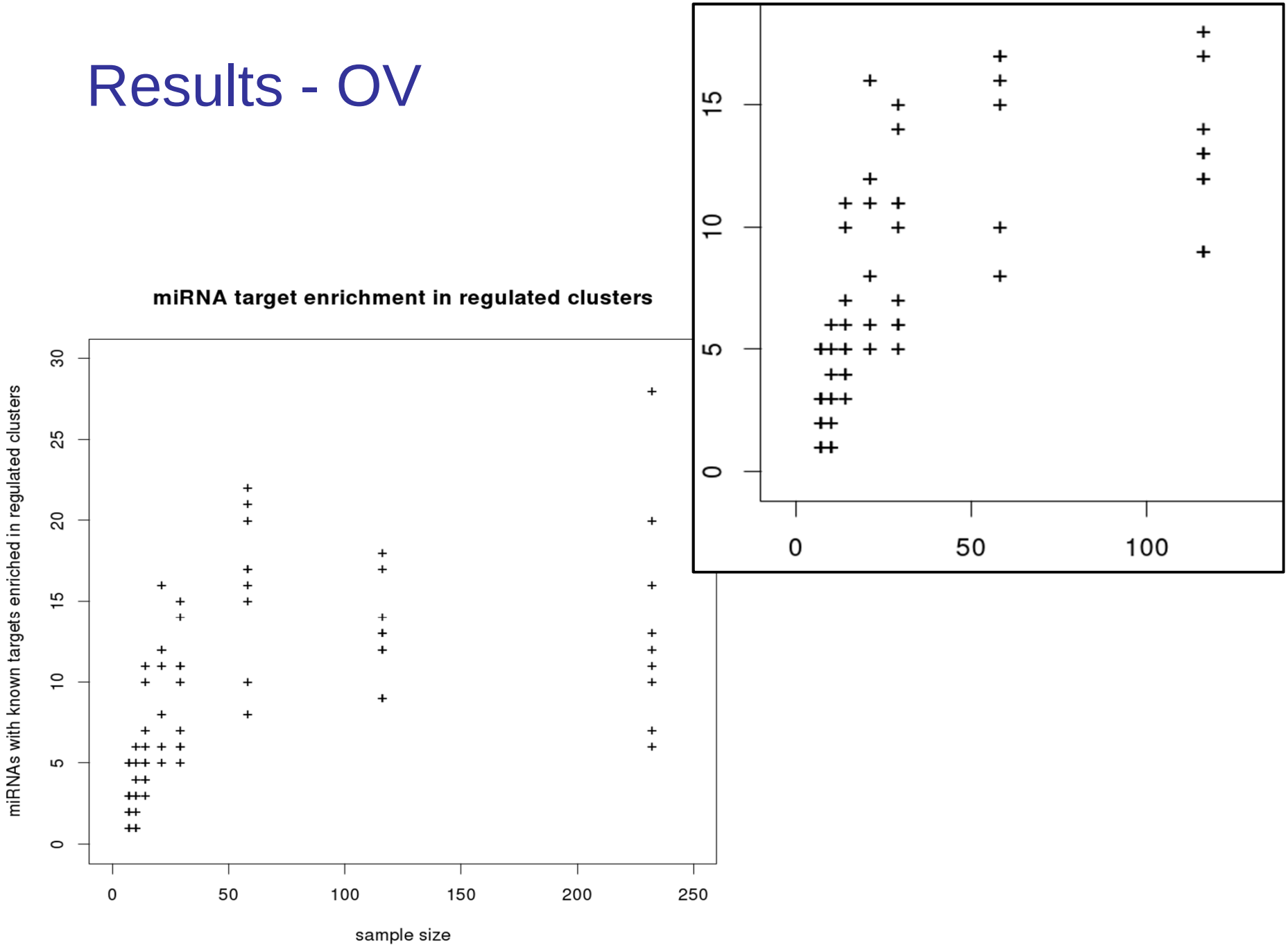


Significance of Enrichment of known Glioblastoma associated miRNAs

Results - GBM


miRNA target enrichment in regulated clusters
miRNAs with known targets enriched in regulated clusters
sample size
0
5
10
15
0
50
100
150
200
250

Results - OV


miRNA target enrichment in regulated clusters
miRNAs with known targets enriched in regulated clusters
sample size
0
5
10
15
20
25
30
0
50
100
150
200
250

Summary

- De novo detection method finds meaningful interactions between miRNAs and mRNAs

- Two approaches were applied for assessment
  - based on known glioblastoma associated miRNAs
  - comparison to sequence based target predictions

- Sub-sampling performed on two different datasets indicate that below a certain sample size the algorithm becomes insufficiently sensitive

Outlook

- Compare power of alternative approaches for different sample sizes
- Test the effect of sample size also on datasets where sample-to-sample variation is smaller
- Investigate alternative functional annotations

Acknowledgement

Paweł P. Łabaj

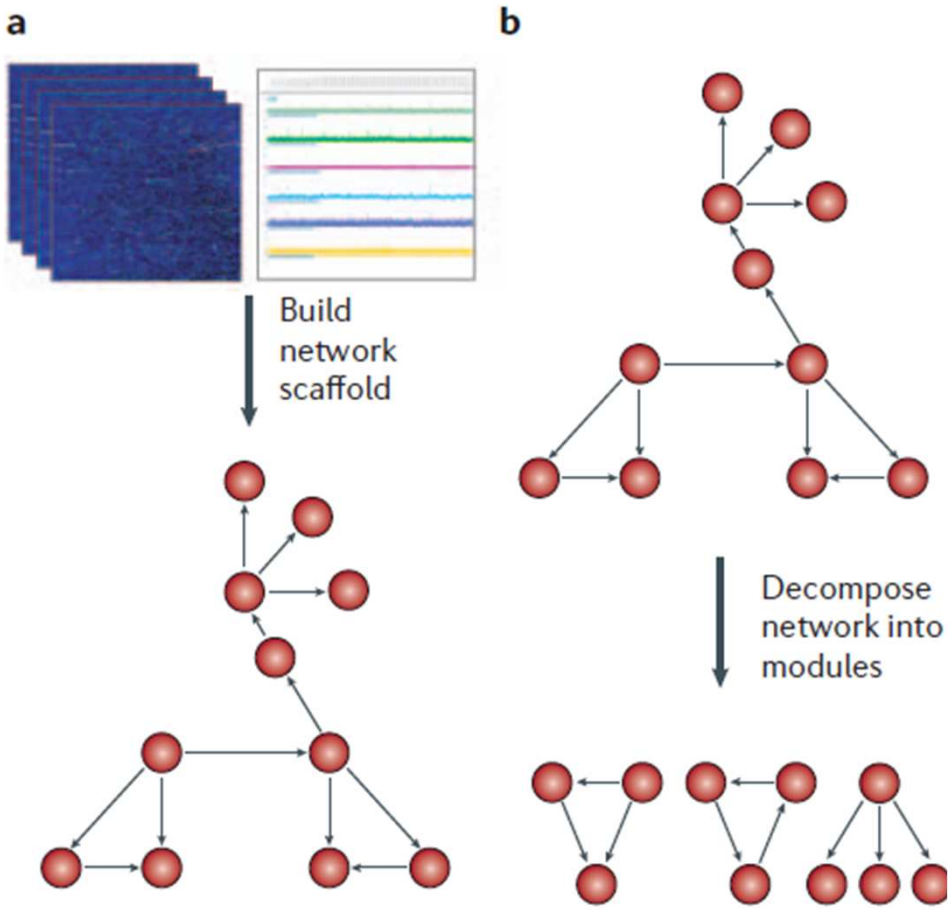
Alexandra Posekany

BioToP Biomolecular Technology of Proteins

FWF Der Wissenschaftsfonds.

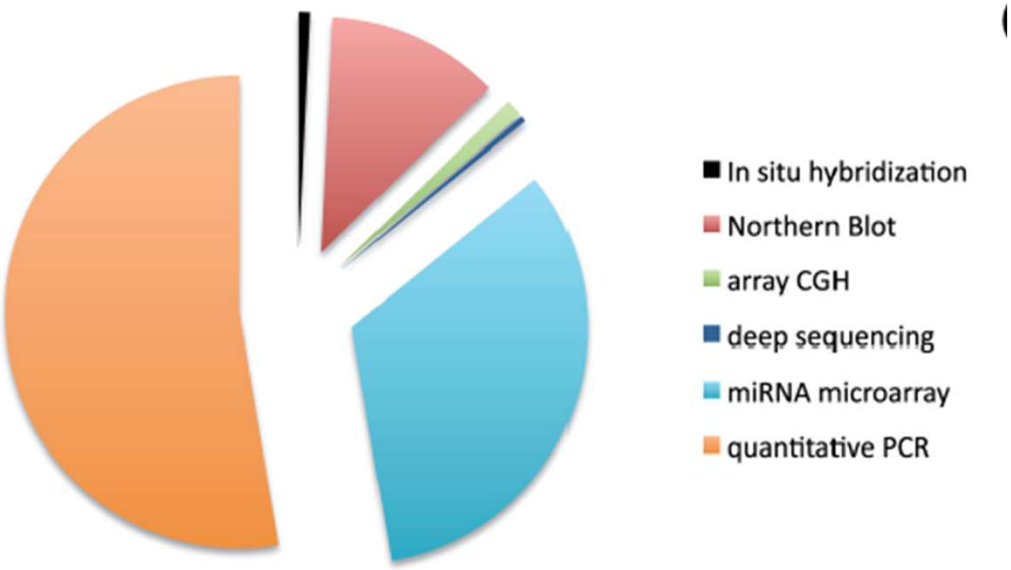
BOKU VIBT Vienna Institute of BioTechnology

The role of modules



Joyce *et al., Nature Reviews Molecular Cell Biology* **7**, 2006

Fraction of annotated miRNA detection methods in PhenomiR



Differences between disease-associated microRNA expression in patients and cell lines!

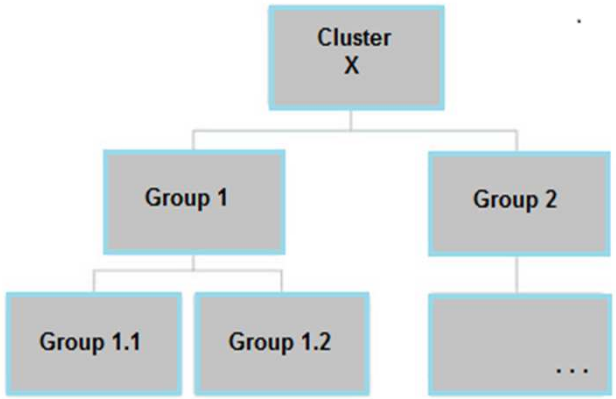
Learning Network Modules Algorithm

Stage 1

- Grouping of genes and conditions based on Gibb's sampling

Stage 2

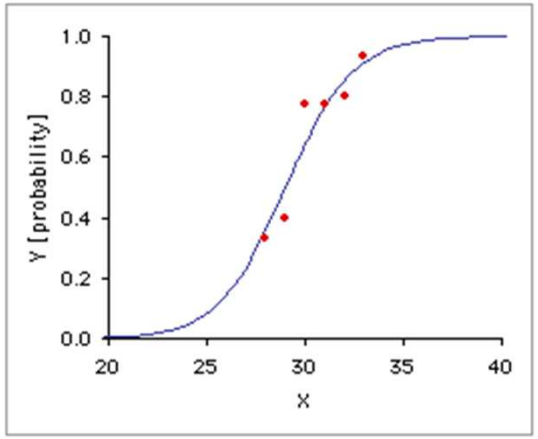
- Regulators assigned to co-expression cluster by logistic regression
- A prioritized list of regulators is obtained for each cluster



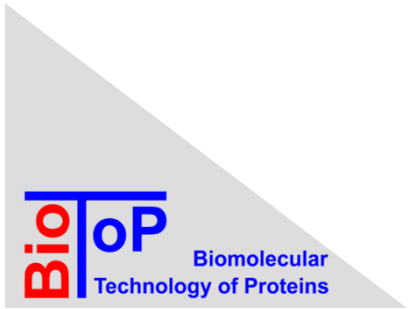
Choice of method

- Sequence information alone
- joint analysis of gene expression profiles and sequence information *(Cheng et al., Plos one, 2008)*
- de novo detection of potential interactions *(Bonnet et al., Bioinformatics, 2010)*

Results



Results


BioToP
Biomolecular
Technology of Proteins


FWF
Der Wissenschaftsfonds.


BOKU
VIBT
Vienna Institute
of BioTechnology