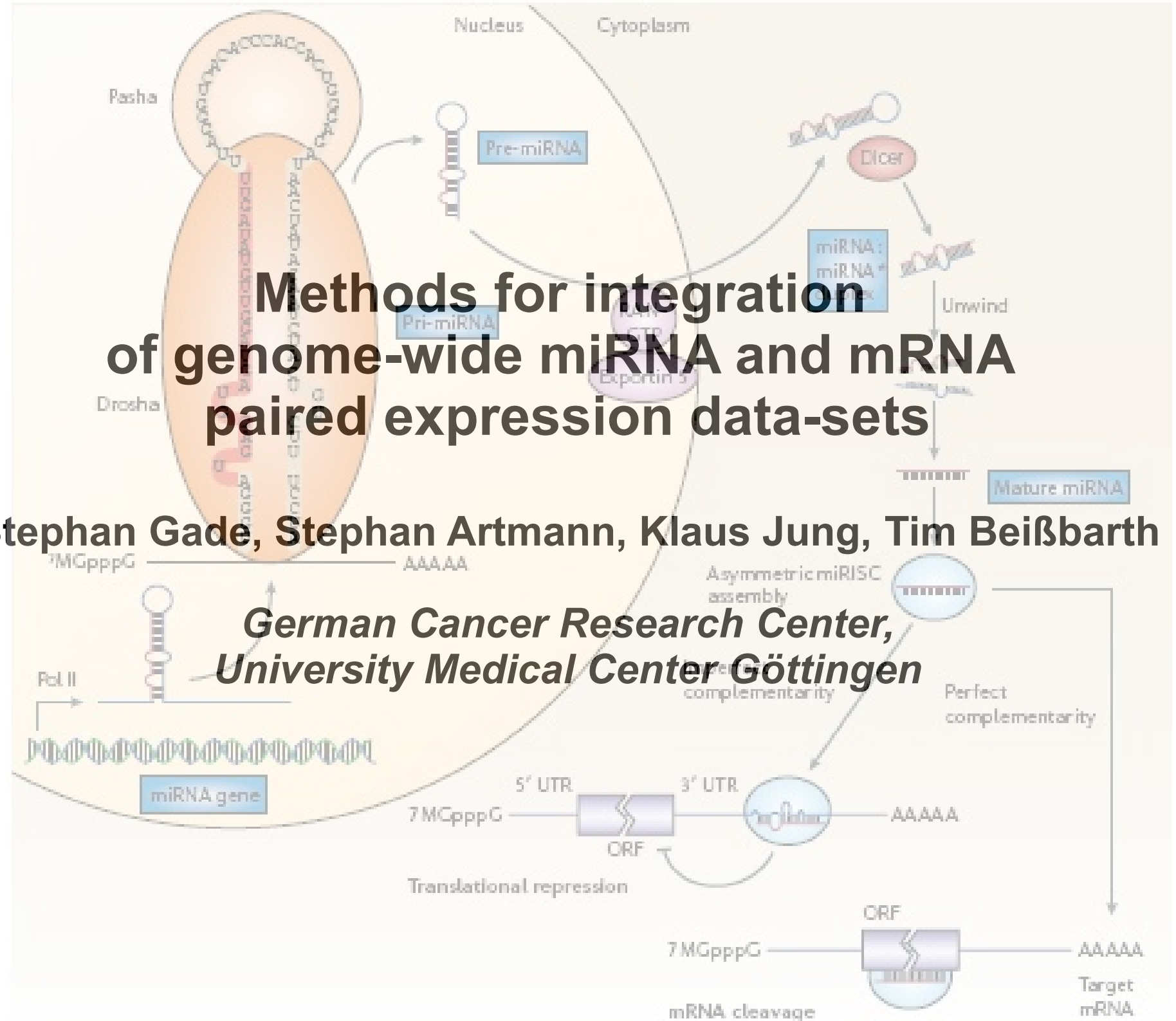


Methods for integration of genome-wide miRNA and mRNA paired expression data-sets

Stephan Gade, Stephan Artmann, Klaus Jung, Tim Beißbarth

German Cancer Research Center, University Medical Center Göttingen

Challenges in personalized medicine

Patients

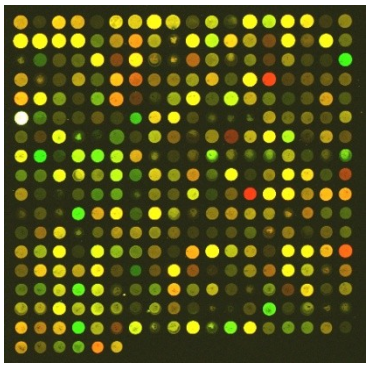
Therapy

Responder
e.g. Reduction of Tumor

Vs.

Non-Responder
e.g. No Reduction of Tumor

Measure



Gene Expression Signatures

and/or

miRNA Signature

Predict?

Different Endpoints

Therapy response, e.g.

- Reduction of

  Tumor Size / Stage
- Tumor Regression Grade

Patient prognosis, e.g.

- Overall Survival
- Disease Free Survival

Different Aims

Finding differential genes, e.g.

- limma
- Cox-Proportional Hazards Regression

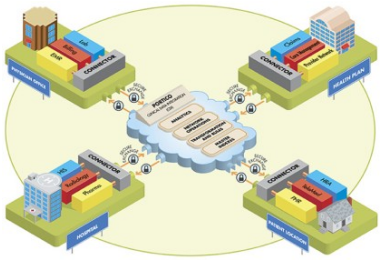
Training a classification model, e.g.

- Linear Discriminant Analysis
- Support Vector Machines
- Boosting

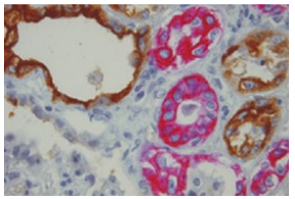
Different Types of Data

Patient data:

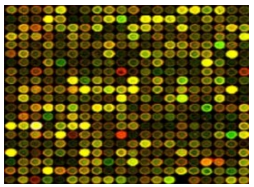
Clinical data

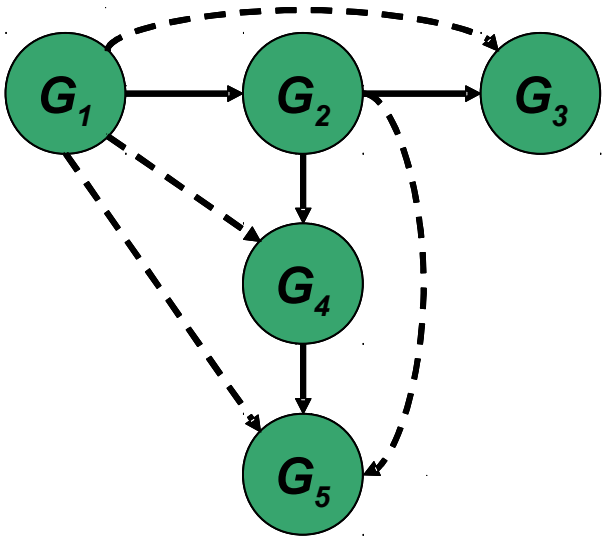


Histological biomarkers

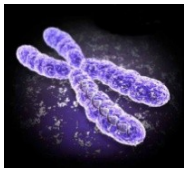


Gene expression profiling (mRNA, miRNA)

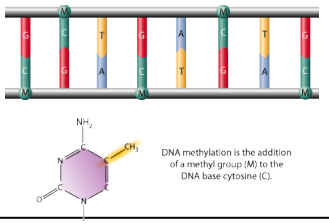




SNPs / Genome aberrations

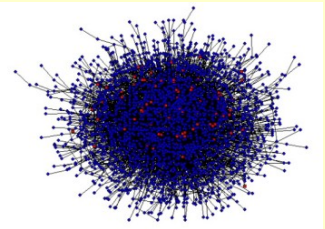


Methylation



External knowledge:

Biological knowledge (Pathway, PPI,TFBS)

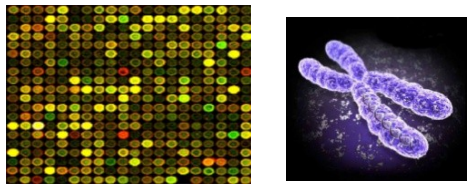


Data fusion of diverse data-types

Methods for data fusion are not widely applied or easily available.

Different concepts for data fusion

Analyze each data-set individually

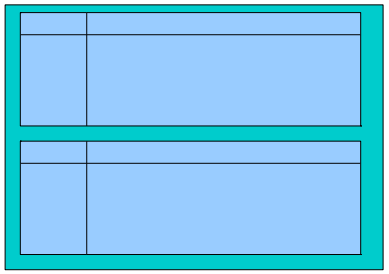


gene list 1, classifier 1

gene list 2, classifier 2

- most common
- usually try to interpret different results manually

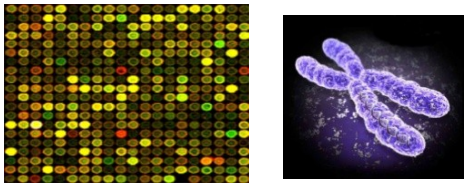
Paste Matrices, analyze features individually



common feature list, classifier

- not usually advisable
- different scales/properties of data
- different weighting

Meta-Analysis



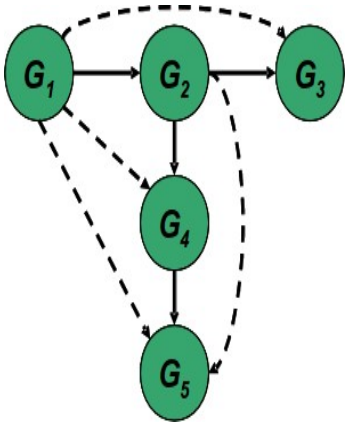
pvalue list 1, classifier 1

pvalue list 2, classifier 2

combined pvalues, meta-classifier

- very flexible
- does not model relations between features of the different data-types

Integrative-Model



- have to understand properties of each of the data-types.

Contents of this talk

- Meta-Analysis approach to find differential miRNAs:

  Detection of simultaneous Group Effects in microRNA Expression and related Target Gene Sets

  Stephan Artmann, Klaus Jung

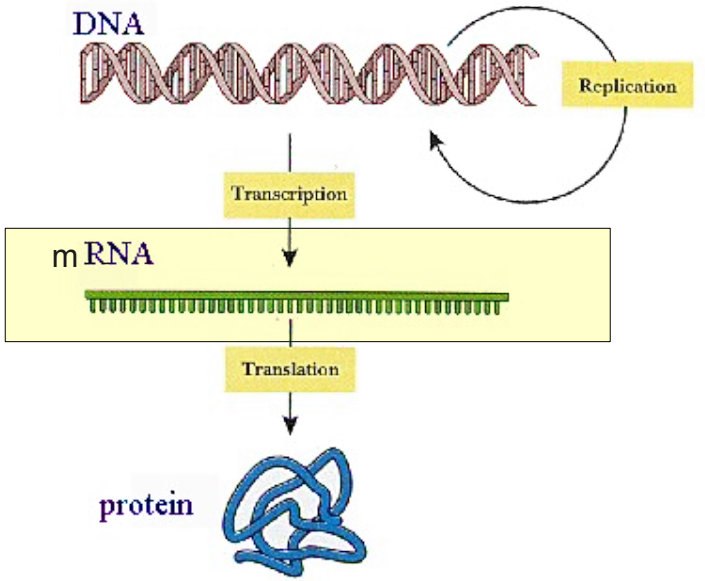
- Classification approach that combines mRNA and microRNA data:

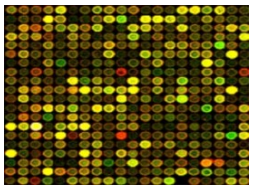
  Graph based fusion of miRNA and mRNA expression data improves prediction of relapse time in prostate cancer

  Stephan Gade

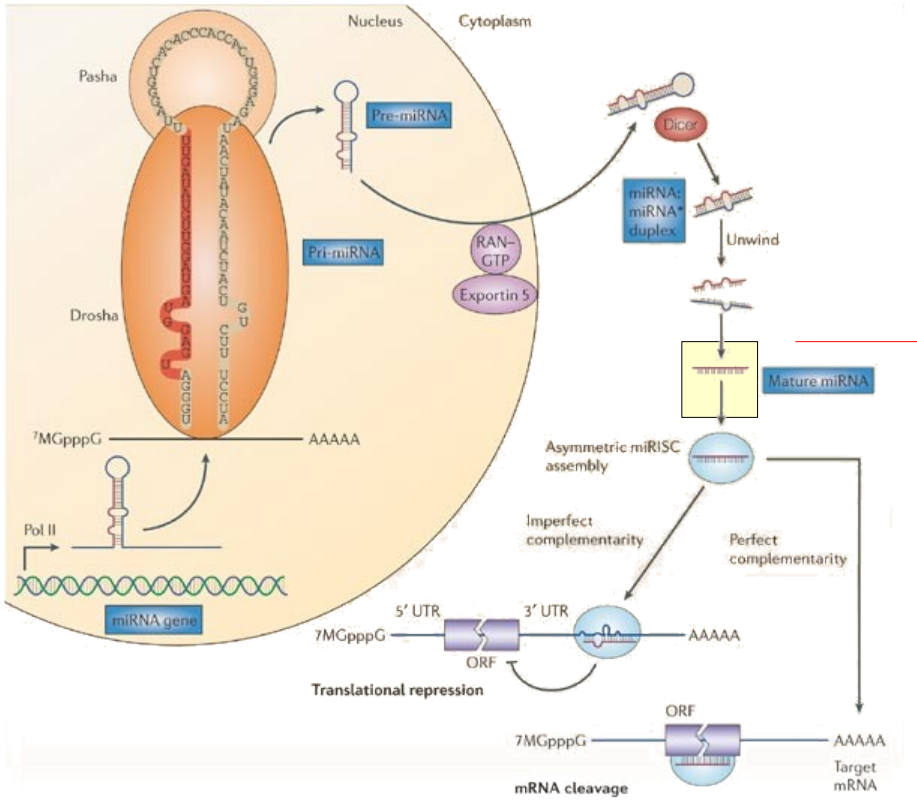
Two different kinds of microarrays

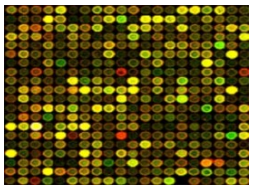
- Gene Expression





- miRNA Expression

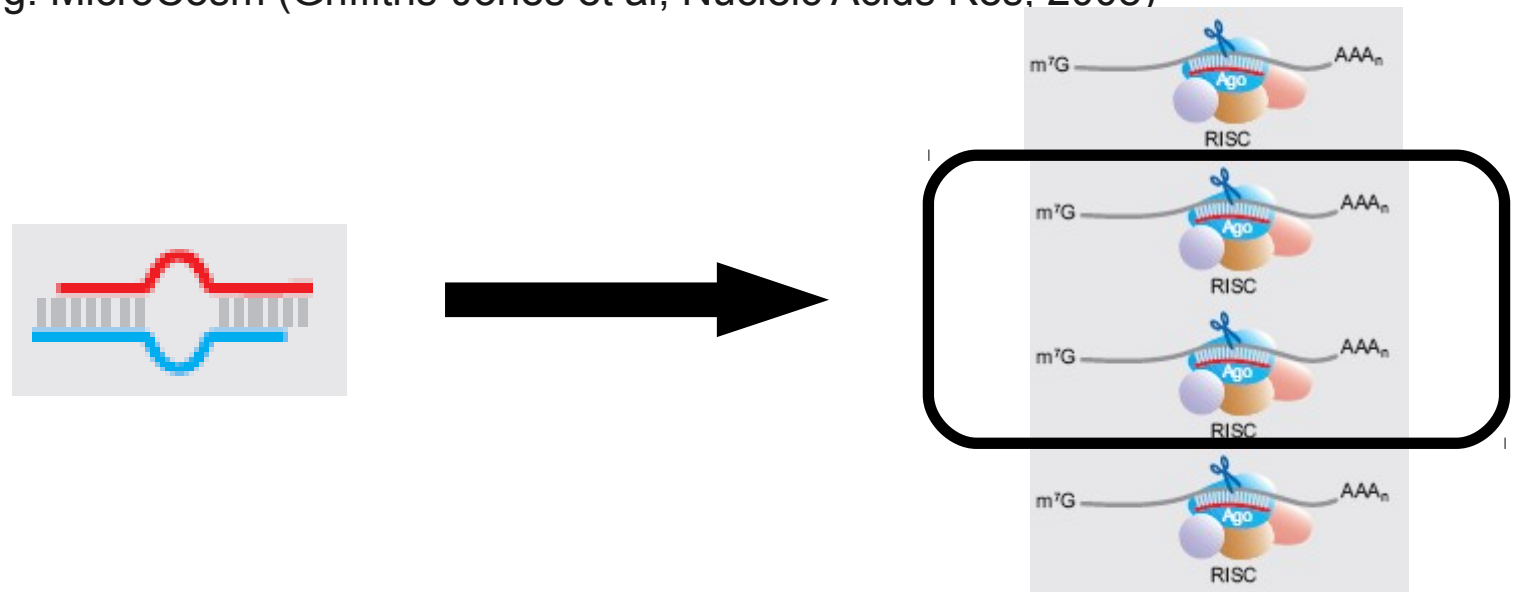




Sources of Information

- Expression of miRNAs
- Expression of mRNAs
- Target Prediction: which miRNA influences which mRNA?

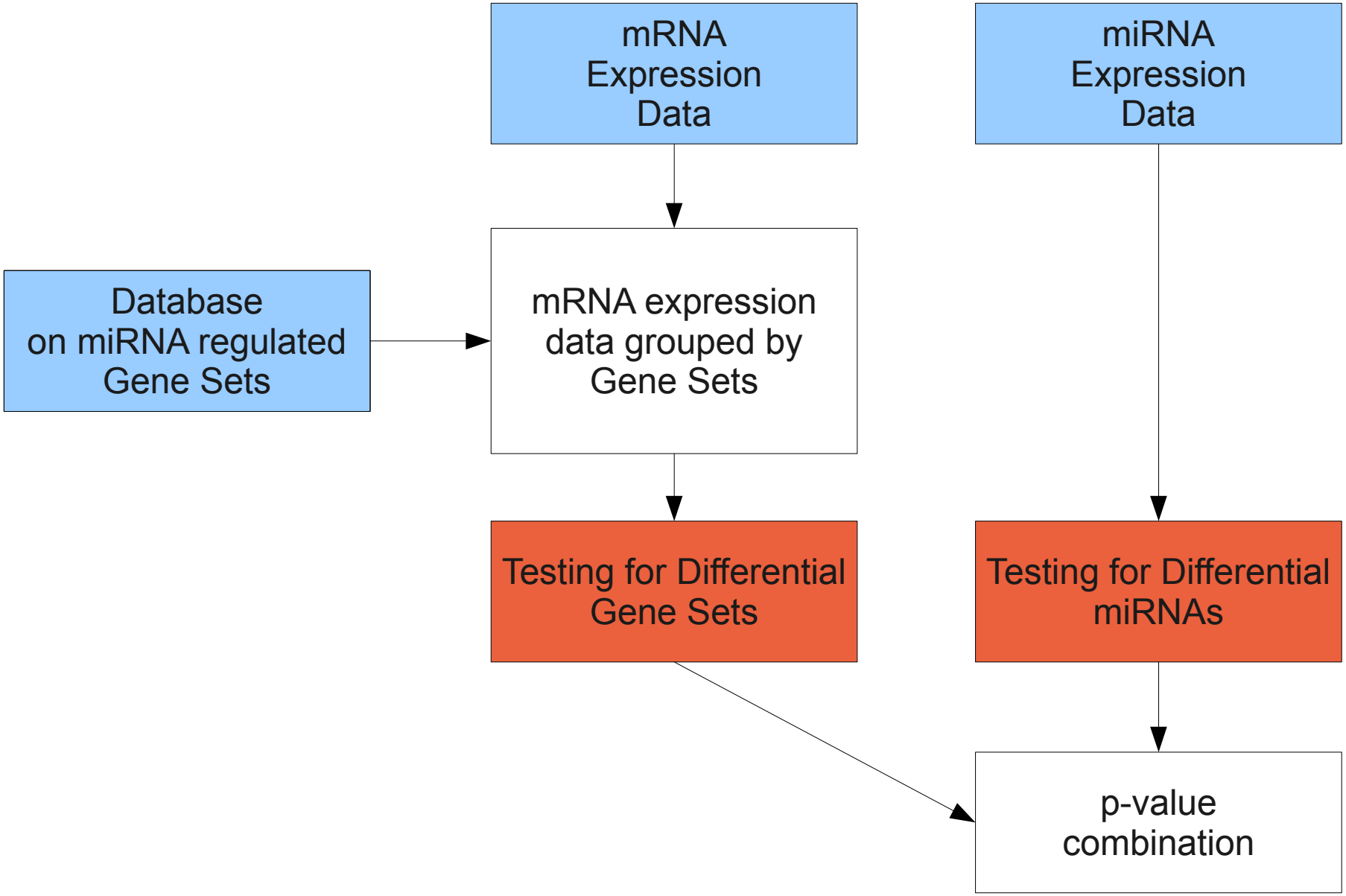
e.g. MicroCosm (Griffiths-Jones et al, Nucleic Acids Res, 2008)



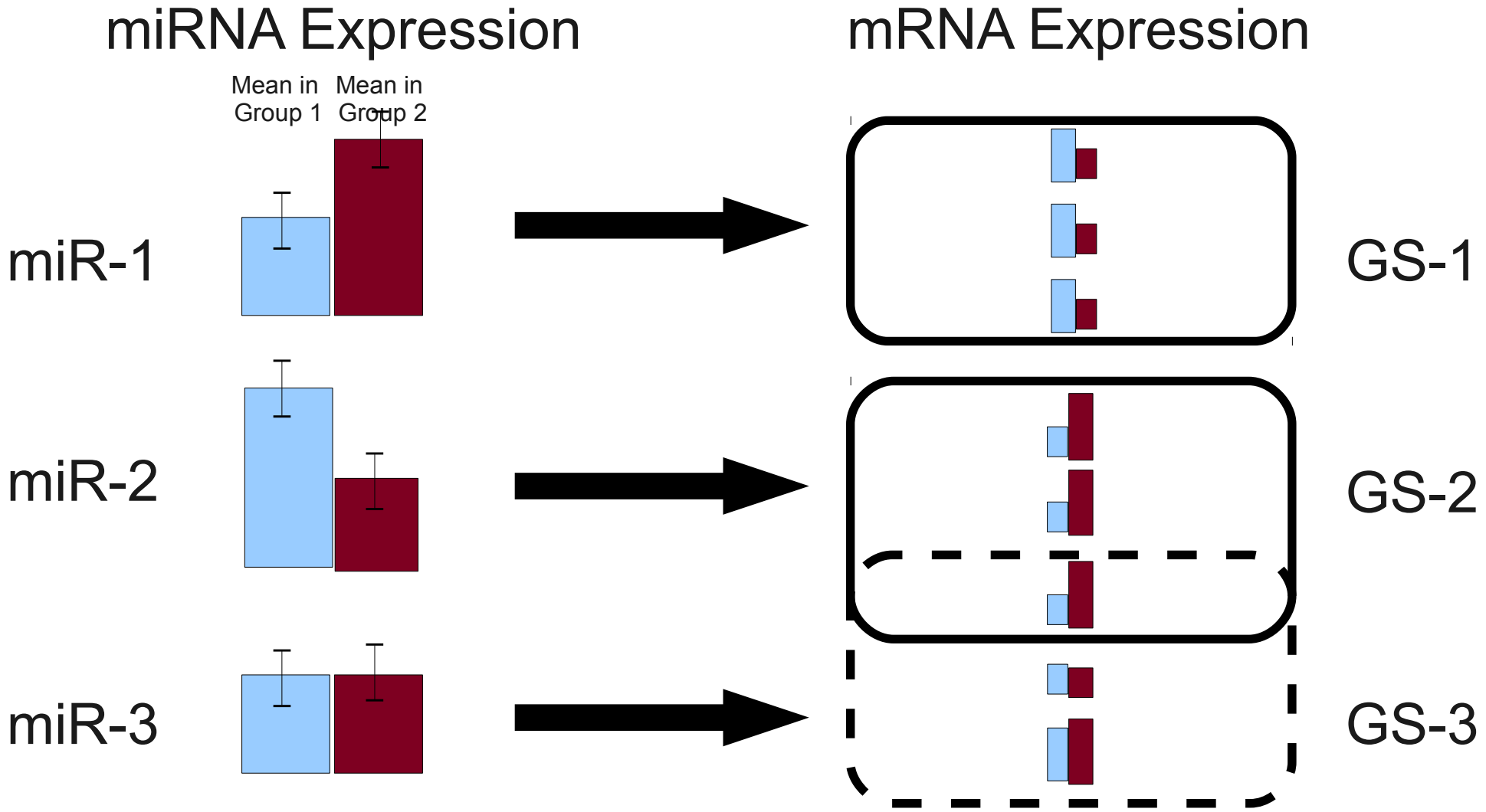
miRNA Target predictions

- MicroCosm target predictions (former miRBase)
  (Griffiths-Jones et al, Nucleic Acids Res, 2008)
- based on miranda algorithm
- energy score for predicted mRNA-miRNA pair
- p-value for energy score based on an extreme value distribution

Approach 1: Combination of Test Results in order to find differential miRNAs.



Tests



LIMMA
(Smyth et al. 2004)

Gene Set Enrichment / Globaltest

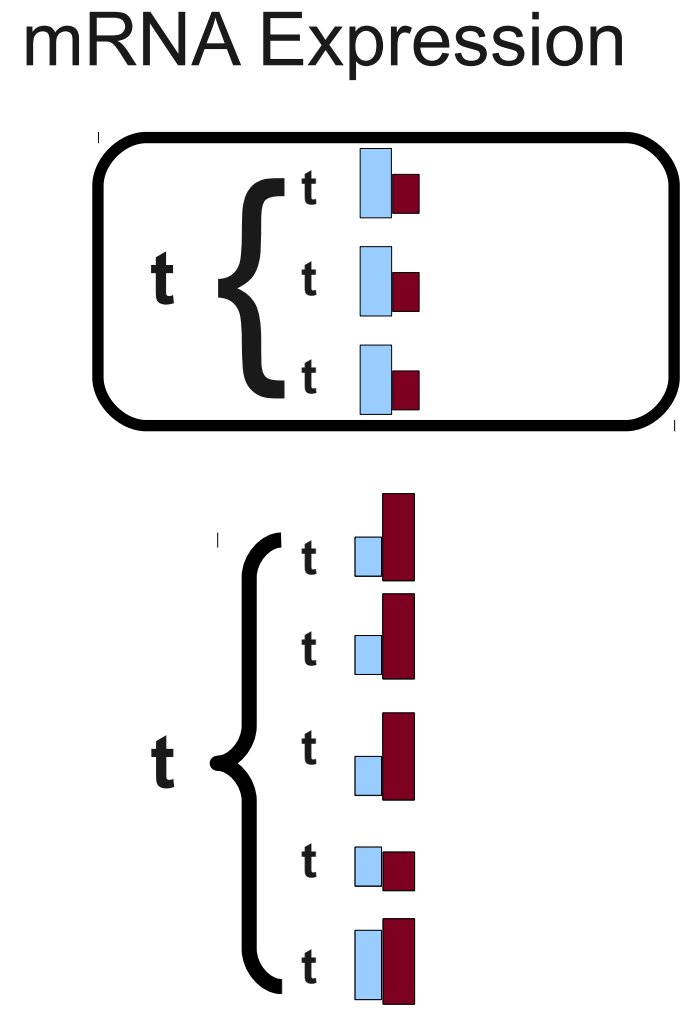
Global vs. Enrichment tests

- Globaltests

  self contained
  Null-Hypothesis

- Enrichment Tests

  competitive
  Null-Hypothesis



Globaltests

- Globaltest

  H_0: $P(Y \mid X) = \mathrm{P}(Y)$

  (Goemann, 2004)

- GlobalAncova

  H_0: $P(X \mid Y = 0) = \mathrm{P}(X \mid Y = 1)$

  (Mansmann & Meister, 2005)

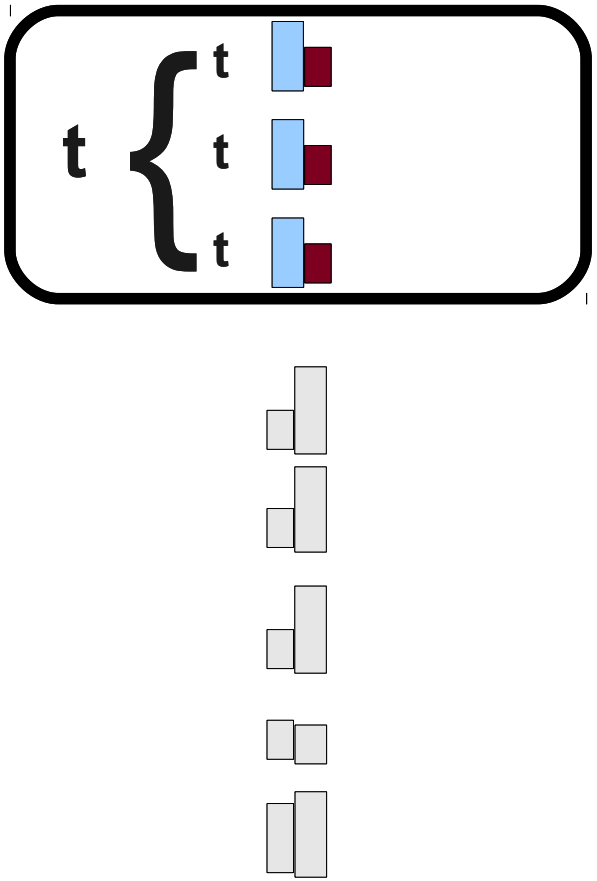
- RepeatedHighDim

  (Jung, 2011)

- ROAST
  - Limma
  - Mean-Statistik
  - Repeated as Random Rotations

  (Wu, 2010)

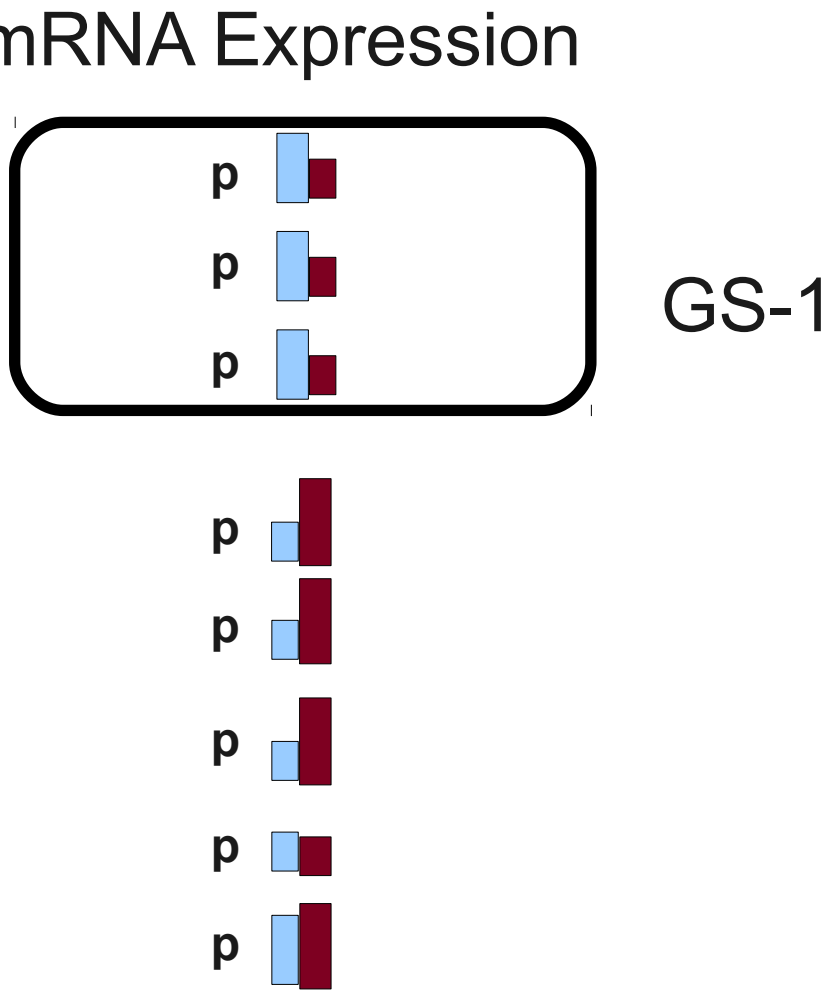
mRNA Expression



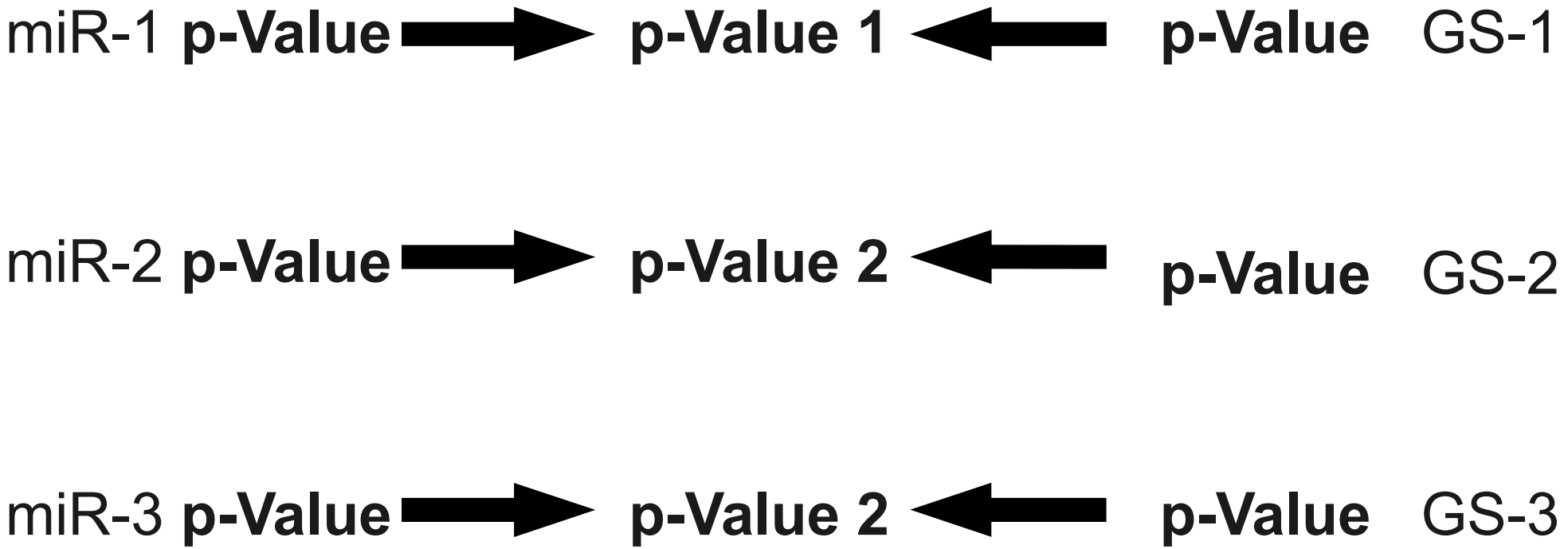
Enrichment Tests

- Fisher's Exact
- Kolm. Smirnov
- Wilcoxon
- Romer
  - Limma
  - Mean-Statistik
  - Repeated as Random Rotation

(Majewski, 2010)



p-Value-Combination



p-Value-Combination

- Fisher-Method → **Globaltests, E. Tests**
  - High Power (one-sided Test)
  - p-Value dependent on tested direction

$$p^{up} = -2\left(\ln\left(p^{up}_{micro}\right) + \ln\left(p^{down}_{gene\,set}\right)\right) \qquad p^{down} = -2\left(\ln\left(p^{down}_{micro}\right) + \ln\left(p^{up}_{gene\,set}\right)\right)$$

$$p = 2 \cdot min\left(p^{up}, p^{down}\right)$$

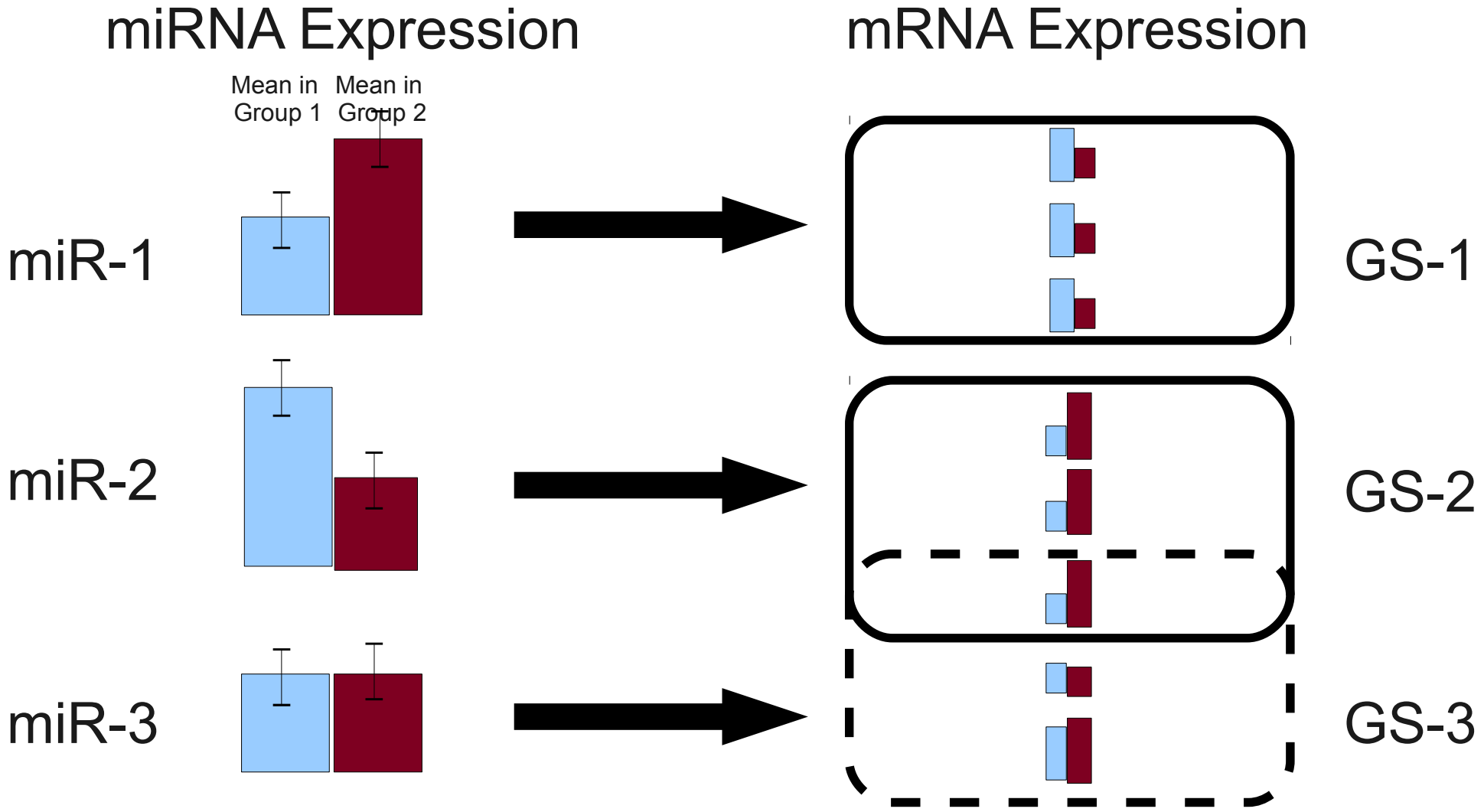
(Fisher et al, 1970)

- Invers-Normal Method → **Wilcox., RTs**
  - Lower Power (one-sided Test)
  - Only when $p = p^{up}_{gene\,set} = 1 - p^{down}_{gene\,set}$

$$p = \frac{\Phi^{-1}\left(p^{up}_{micro}\right) + \Phi^{-1}\left(p^{down}_{gene\,set}\right)}{\sqrt{2}} = \frac{\Phi^{-1}\left(p^{down}_{micro}\right) + \Phi^{-1}\left(p^{up}_{gene\,set}\right)}{\sqrt{2}}$$

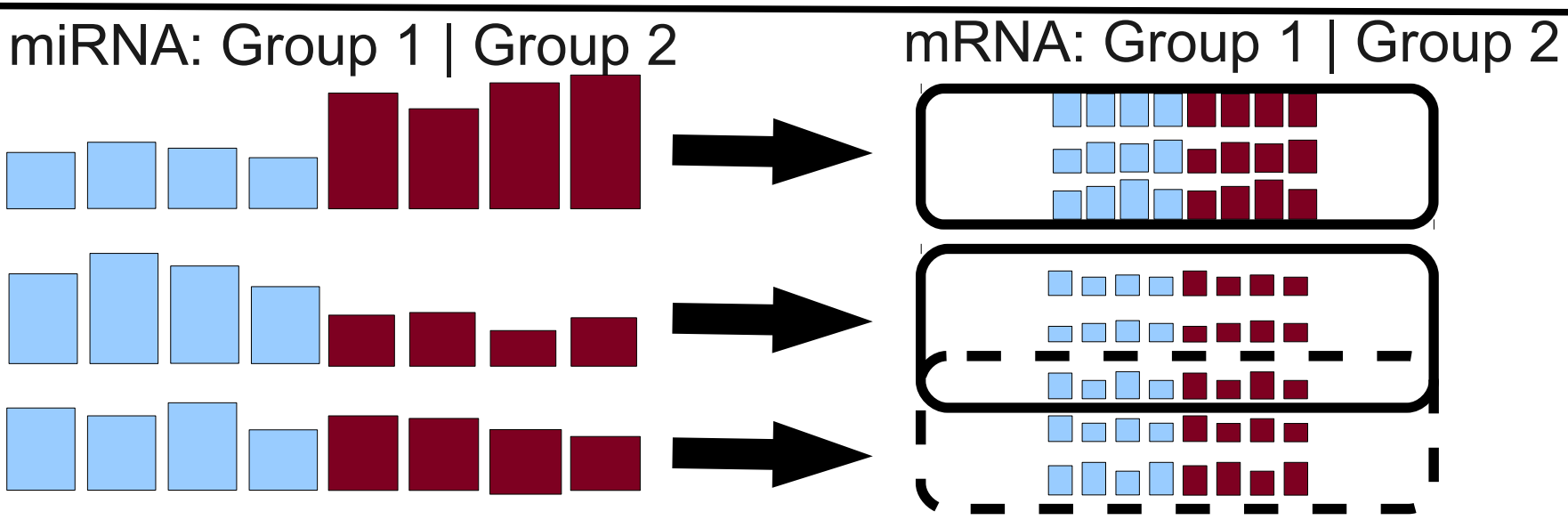
(Stouffer et al, 1949)

Simulation


miRNA Expression
mRNA Expression
Mean in Group 1
Mean in Group 2
miR-1
miR-2
miR-3
GS-1
GS-2
GS-3

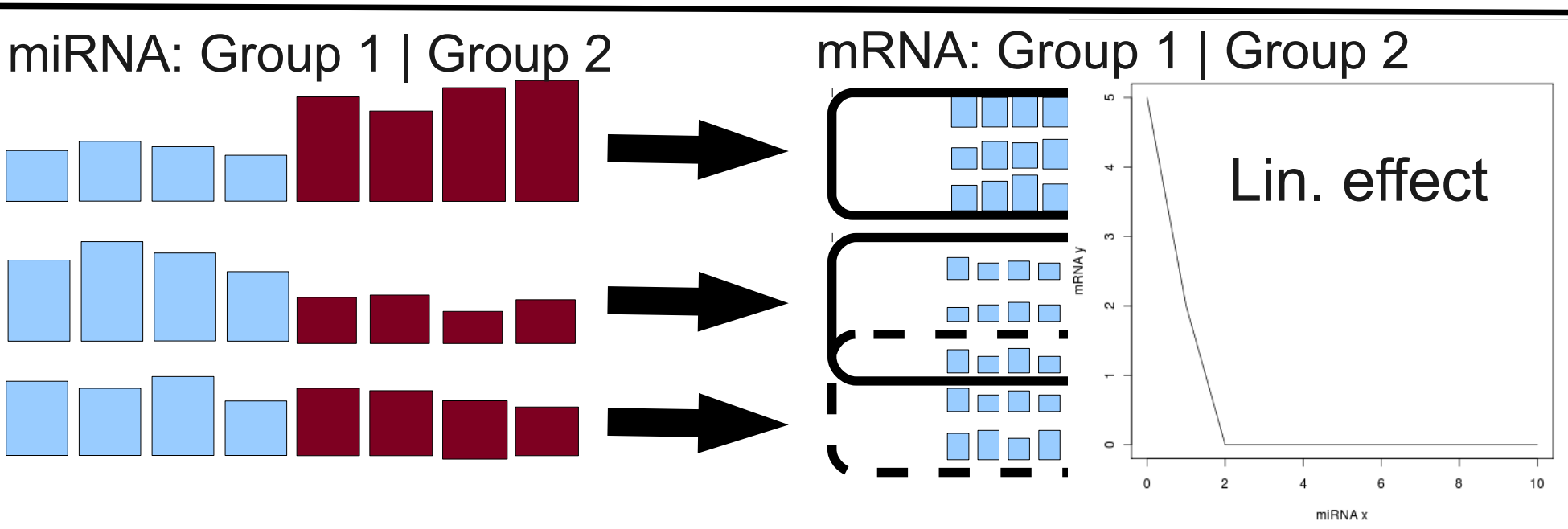
Simulation

- $X1 \sim N(\mu,\Sigma)$
- $X2 \sim N(\mu \pm effect,\Sigma)$
- $Y1 \sim N(\nu 1,T)$
- $Y2 \sim N(\nu 2,T)$
- Allocation Matrix: $A \sim Bernoulli$



Simulation

- $X1 \sim N(\mu,\Sigma)$
- $X2 \sim N(\mu \pm effect,\Sigma)$
- $Y1 \sim N(\nu 1,T)$
- $Y2 \sim N(\nu 2,T)$
- Allocationsmatrix A ~ Bernoulli
- Lin. effect: $\nu = (A \cdot B)^* \mu$, mit $B \sim N(-1,0.1)$



Simulations

Parameter	Simulation 1	Simulation 2	Simulation 3
Repetitions	1000	1000	1000
# samples per group	4	4	4
# mRNAs	5000	5000	5000
# miRNAs	100	100	100
Var(miRNA),Var(mRNA)	1 bis 2	1 bis 2	1 bis 2
Covariance Structure	autoregressive	autoregressive	autoregressive
# differential miRNAs	10 %	10 %	10 %
up- / down-regulated	50 / 50	50 / 50	50 / 50
Effect strength	0, 1, 2, 4, 6	0, 1, 2, 4, 6	0, 1, 2, 4, 6
μ and ν	~ logN(1,0.1)	~ logN(1,0.1)	~ logN(1,0.1)
Allocation Matrix A	**Structure w.o. overlapp**	**a ~ binom(0.04-0.08)**	
Modification Factor B	**~ N(1,0.1)**	**~ N(1,0.1)**	**~ N(10,0.1)**
# mRNAs per miRNA	**50**	**variable**	**variable**

Simulation 1

Test	FDR	Power
Globaltests		
Globaltest	≥ 0.05	Limma < GST < Combi.
GlobalAncova	≥ 0.05	Limma < GST < Combi.
RepeatedHighDim	>> 0.05	Limma < GST < Combi.
Enrichment Tests		
Kolm. Smirnov	± 0.05	Limma < GST < Combi.
Wilcoxon	± 0.05	Limma < GST < Combi.
Fisher	<< 0.05	Limma < GST < Combi.
Rotation Tests		
ROAST	± 0.05	Limma < Combi. < GST
Romer	± 0.05	Limma < Combi. < GST

Simulation 2

Test	FDR	Power
Globaltests		
Globaltest	>>> 0.05	Limma < GST < Combi.
GlobalAncova	>>> 0.05	Limma < GST < Combi.
RepeatedHighDim	>>> 0.05	Limma < GST < Combi.
Enrichment Tests		
Kolm. Smirnov	± 0.05	Limma < GST < Combi.
Wilcoxon	± 0.05	Limma < GST < Combi.
Fisher	± 0.05	Limma < GST < Combi.
Rotationstests		
ROAST	~ Effect	Limma < Combi. < GST
Romer	± 0.05	Limma < Combi. < GST

Simulation 3

Test	FDR	Power
Globaltests		
Globaltest	>>> 0.05	Limma < GST < Combi.
GlobalAncova	>>> 0.05	Limma < GST < Combi.
RepeatedHighDim	>>> 0.05	Limma < GST < Combi.
Enrichment Tests		
Kolm. Smirnov	± 0.05	Limma < GST < Combi.
Wilcoxon	± 0.05	Limma < GST < Combi.
Fisher	± 0.05	Limma < GST < Combi.
Rotationstests		
ROAST	~ effect	Limma < Combi. < GST
Romer	± 0.05	Limma < Combi. < GST

Simulation Results Summary

		Global Tests	**ET**	RT
Simulation 1	no Overlap	FDR not controlled	Fisher too conservative	ok Low Power
Simulation 2	Overlap, varying gene set size	FDR not controlled	ok	FDR ok for Romer Low Power
Simulation 3	Overlap, varying gene set size very strong gene set effect	FDR not controlled	ok	FDR ok for Romer Low Power

Data Example

- Rats: early neuronal progenitors
  - Embryonic day 11 (E11) vs. day 13 (E13)

BMC Neuroscience

BioMed Central

Research article

Open Access

Integrating microRNA and mRNA expression profiles of neuronal progenitors to identify regulatory networks underlying the onset of cortical neurogenesis

Joseph A Nielsen*†[1,2], Pierre Lau†[1], Dragan Maric[3], Jeffery L Barker[3] and Lynn D Hudson[1]

Address: [1]Section of Developmental Genetics, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA, [2]Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA and [3]Laboratory of Neurophysiology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA

Email: Joseph A Nielsen* - joseph.nielsen@fda.hhs.gov; Pierre Lau - laup@ninds.nih.gov; Dragan Maric - maricd@ninds.nih.gov; Jeffery L Barker - jeffery.barker@nih.hhs.gov; Lynn D Hudson - hudsonl1@od.nih.gov

* Corresponding author †Equal contributors

Data Example

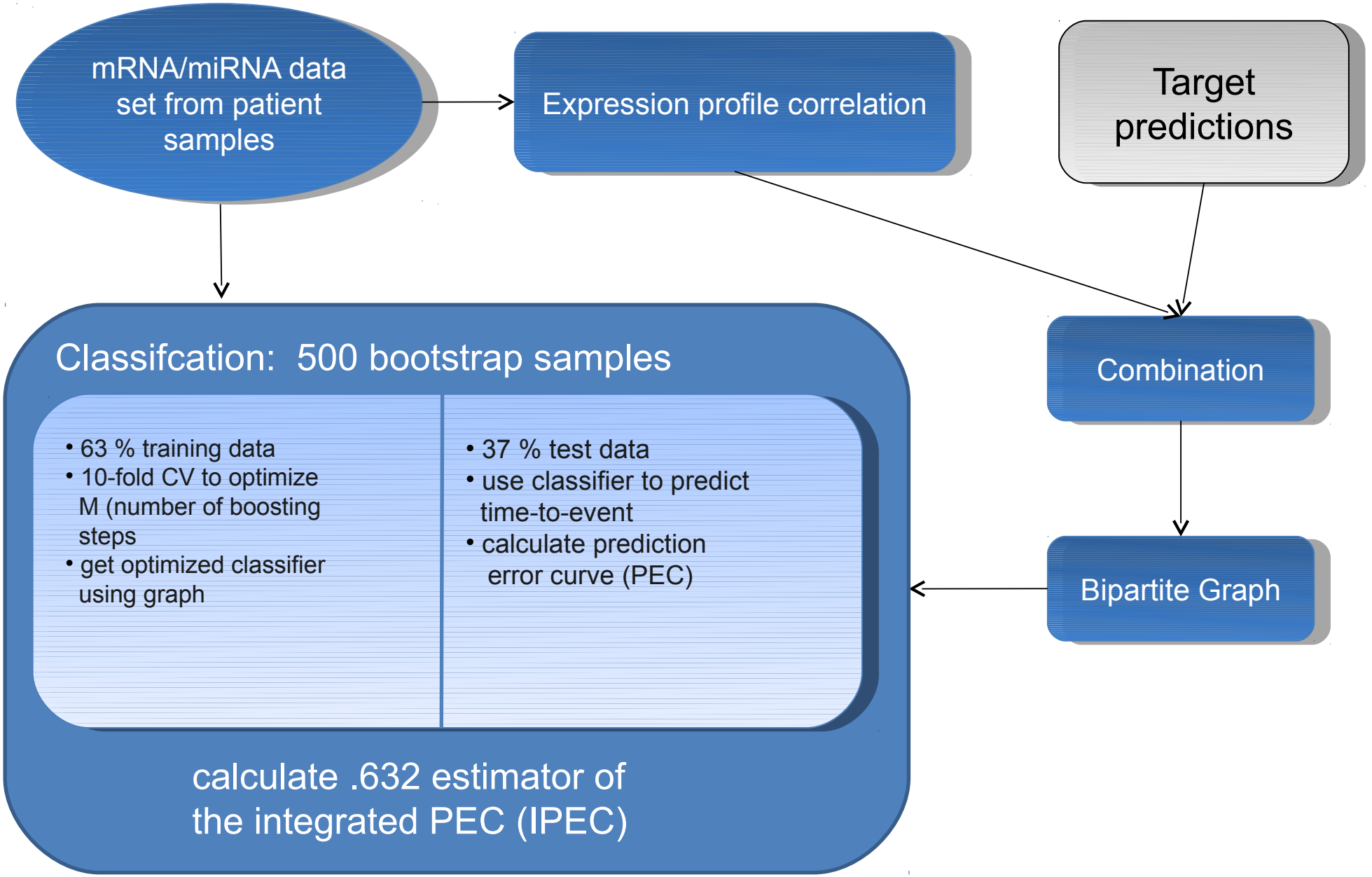
miRNAs	Regulation in E13	Gene Set	Global Tests	ET	RT
13 miRNAs	**down-reg.**	**v.a. up-reg.**	***q < 5 %***	***q < 5 %***	***q < 5 %***
8 miRNAs	**up-reg.**	**v.a. down-reg.**	***q < 5 %***	***q < 5 %***	***q < 5 %***
5 miRNAs	differential	not sign. / weak corr. with miRNA	***q < 5 %***	***q < 5 %***	q > 5 %
miR-19a & -210	slightly down-reg.	slightly upreg. And not sign.	***q < 5 %***	***q < 5 %***	***q < 5 % in ROAST***
miR-126	down-reg.	not sign., lower p for down-reg.	***q < 5 %***	***q < 5 %***	***q < 5 % in ROMER***
miR-290	**down-reg.**	**mainly down-reg.**	***q < 5 %***	***q < 5 % in KS und F***	*q > 5 %*
18 miRNAs	not mentioned, but differential	not mentioned	***q < 5 %***	***q < 5 %***	***q < 5 %***

	Roast	Romer	KS	W	F	GT	GA	RHD
# miRNAs	3	25	31	45	76	202-31	202-34	202-35

Summary

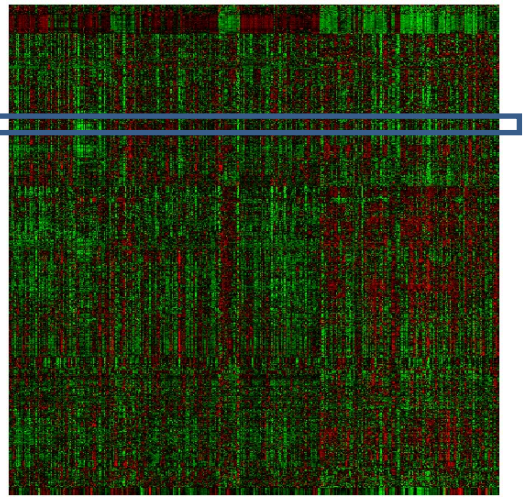
- Combination of test results increases Power to detect differential miRNAs in comparison to testing the miRNAs alone.
- Combination of test results leads to less “false positive” results than using Gene-Set tests for the mRNA targets of the miRNAs alone.
- Enrichment Tests in these combinations are more conservative than Globaltests.
- Suggestion: The Romer method appears to be the best compromise. For much computationally faster but almost as accurate results use Wilcoxon-Test.
- All methods are implemented and available in the R package miRtest on CRAN or R-Forge.

Approach 2: Combination miRNA and mRNA data to train classification model



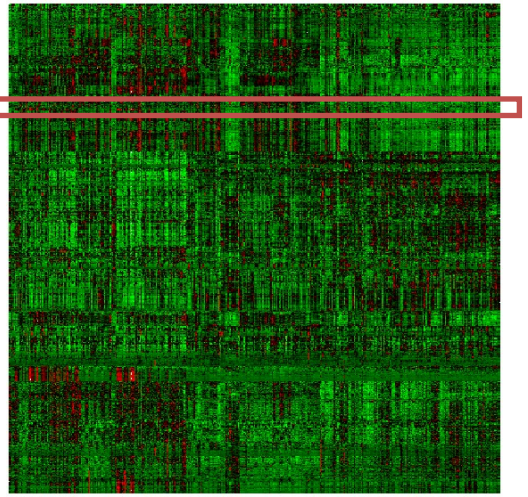
mRNA-miRNA Correlations

mRNA expression data



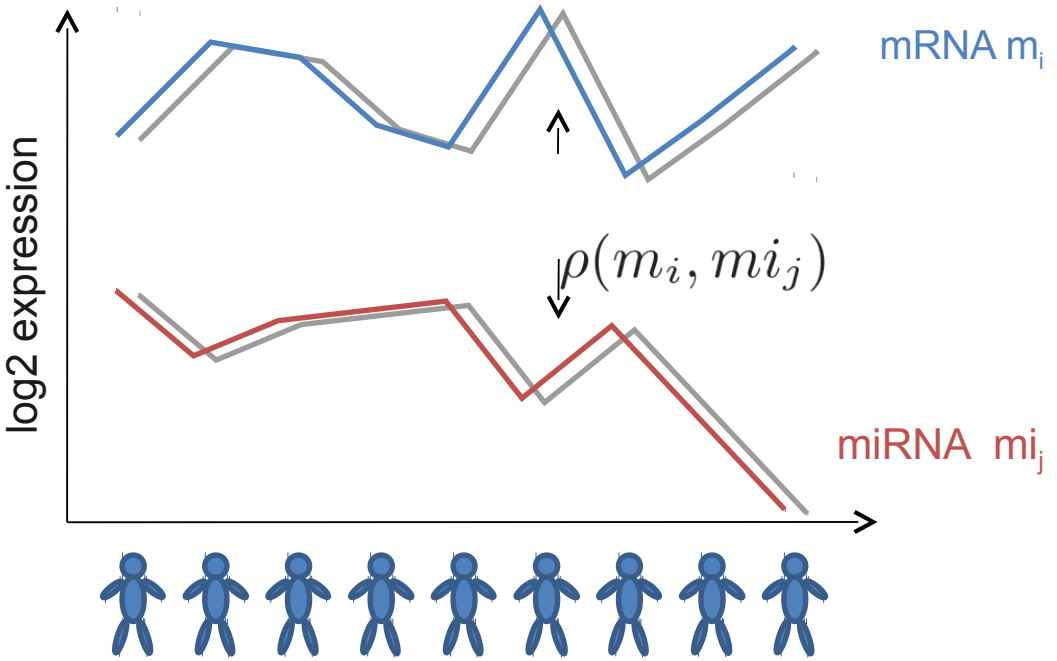
mRNAs

miRNA expression data



miRNAs

(Glioblastoma multiforme, TCGA)

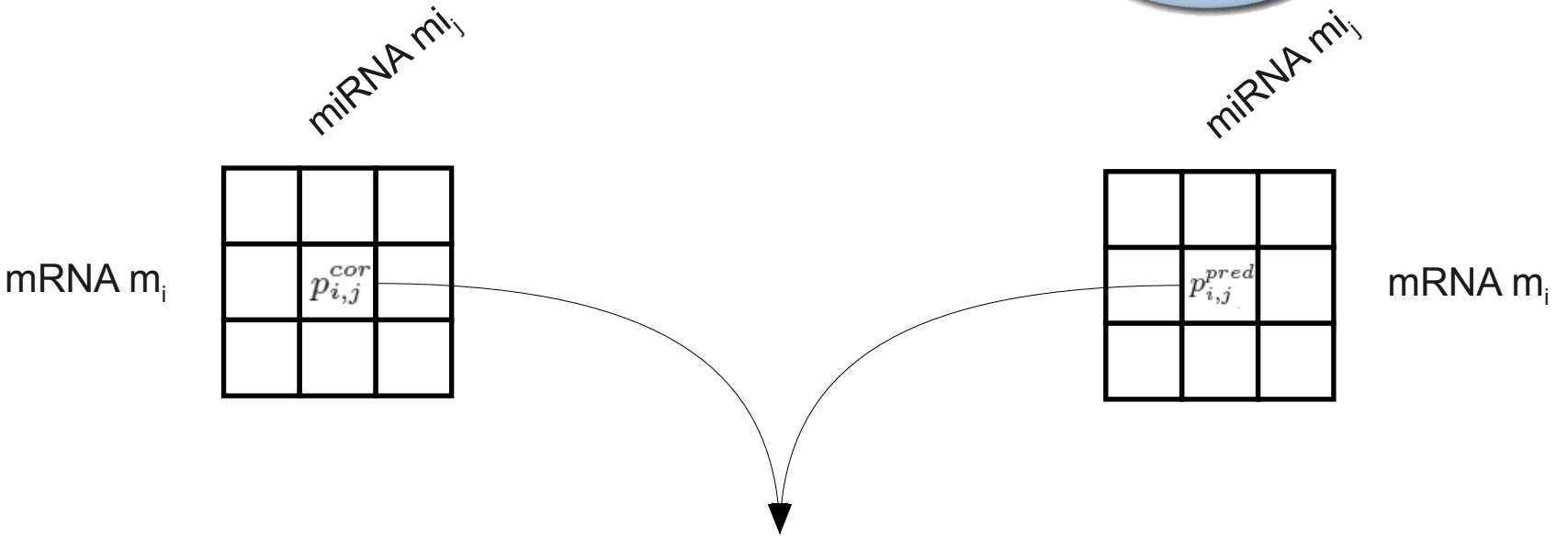


$$p_{i,j}^{cor} = P(H_0 : \rho(m_i, mi_j) = 0)$$

$$\forall i \in \{1, n_1\}, j \in \{1, n_2\}$$

Combination of miRNA-mRNA correlation and target prediction

Correlation p-values
(from patient samples)



$$p_{i,j} = 1 - \Phi\left(\frac{1}{\sqrt{2}}\left(\Phi^{-1}(1 - p_{i,j}^{cor}) + \Phi^{-1}(1 - p_{i,j}^{pred})\right)\right)$$

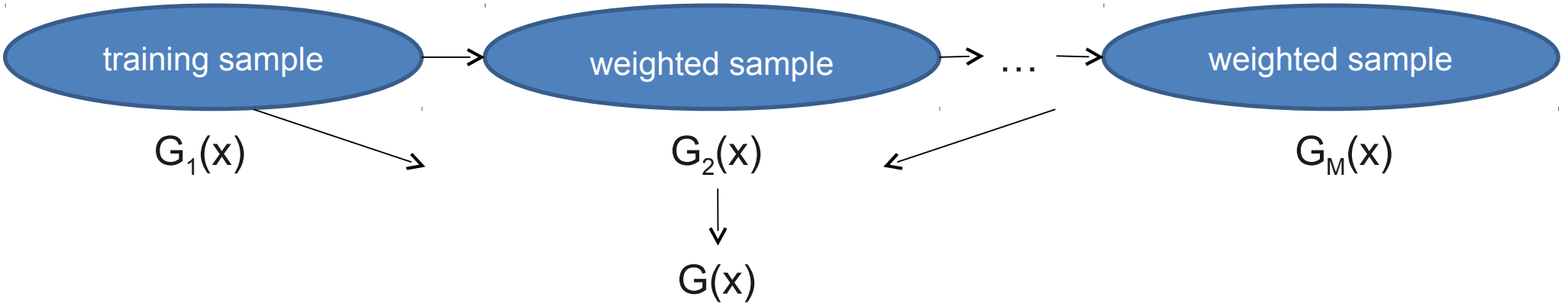
- p-value combination according to Stouffer et. al, 1949
- result: matrix of new **combined p-values** $p_{i,j}$

Bipartite graph

- matrix W of $1\text{-}p_{i,j}$ can be seen as the adjacency matrix of a bipartite graph
- describes the relations between miRNAs and mRNAs in the data set
- can be used to „guide“ a classifier during feature selection

Boosting

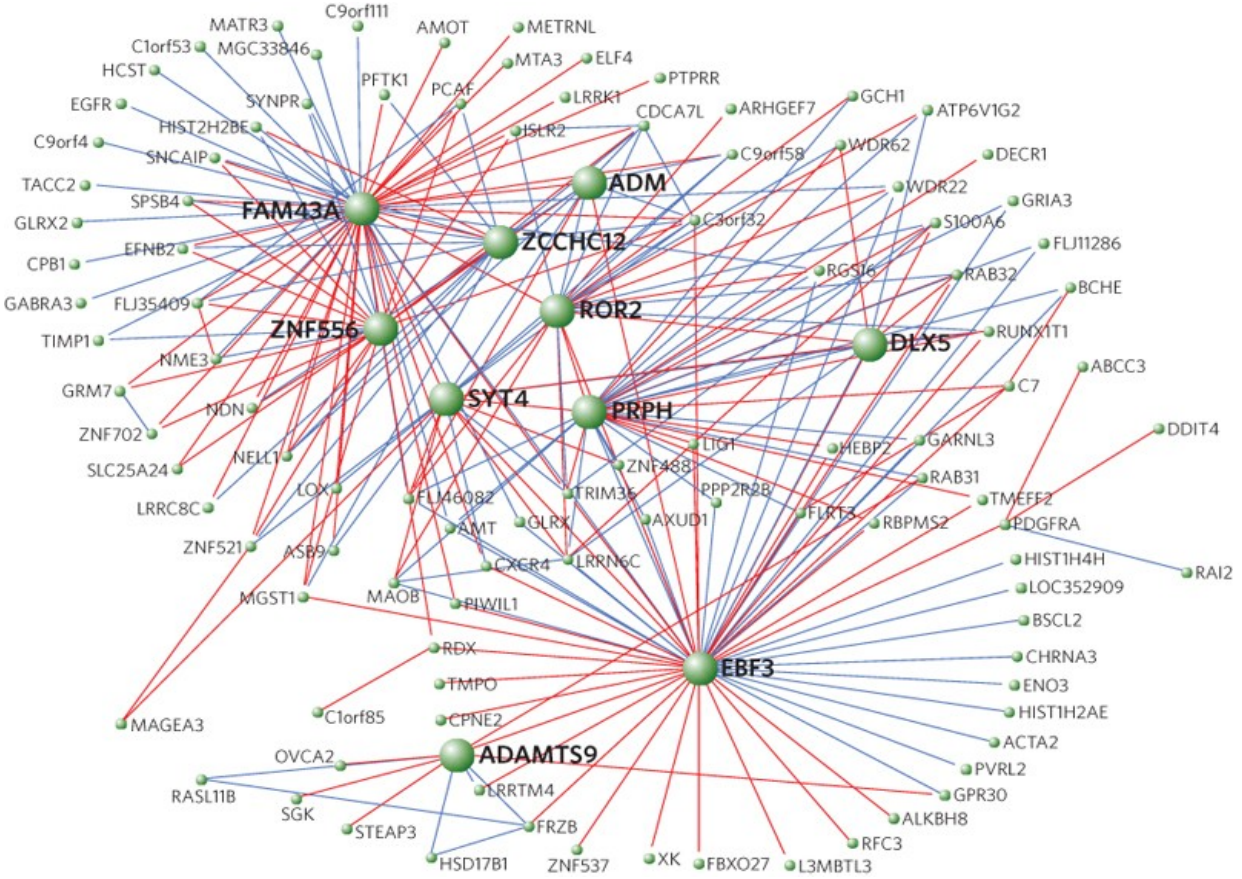
- belongs to the class of ensembl learners
- first introduced by Freund and Schapire, 1996
- weighted combination of several weak classifiers to build one strong classifier



(modified from Hastie et.al, 2009)

PathBoost

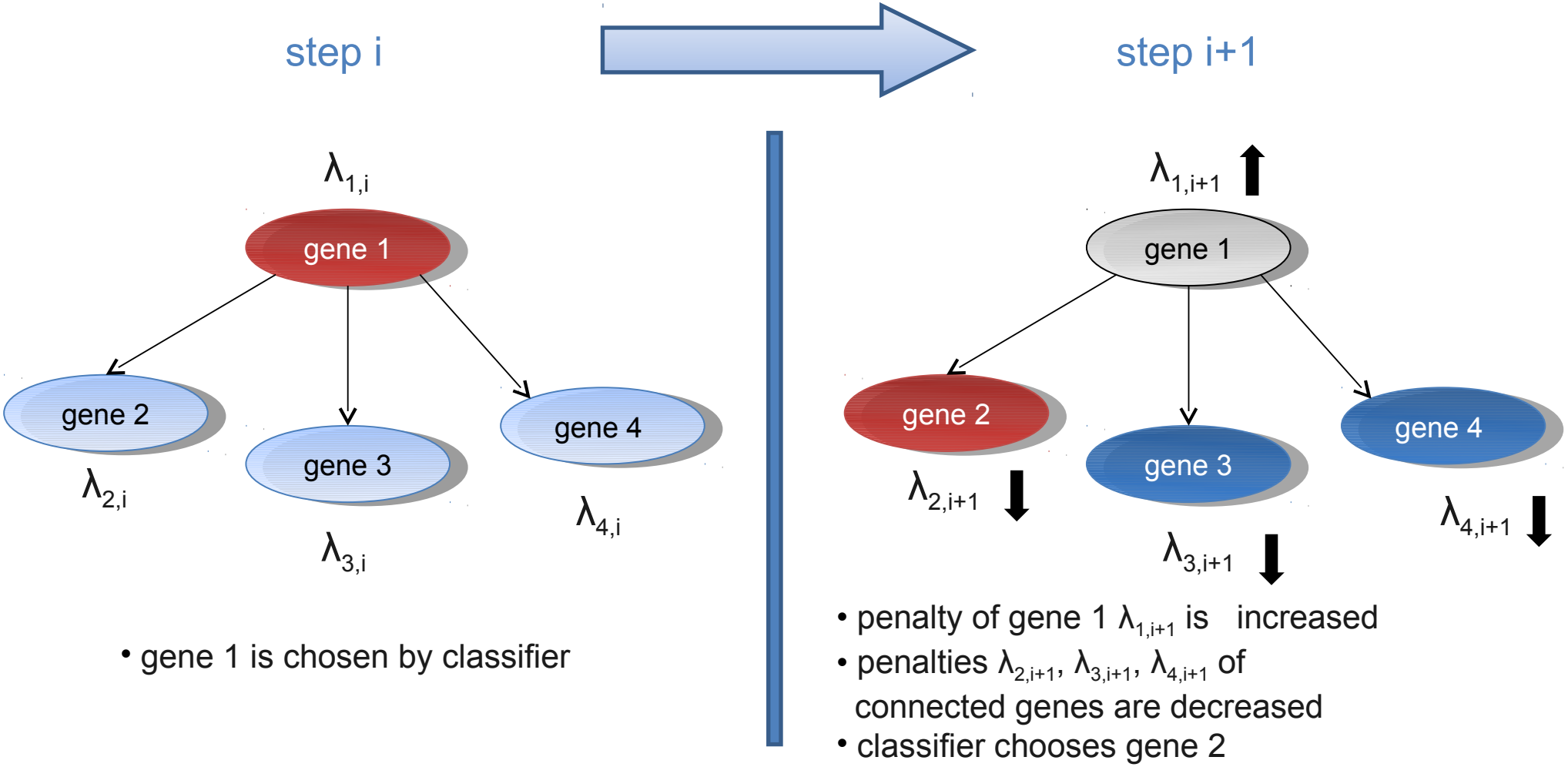
Motivation: A gene with a low fold-change should have an increased influence on the classifier if it is connected to differentially expressed genes.



PathBoost

Basic Idea (Binder and Schumacher, 2009):

Increase penalties λ of chosen genes and decrease penalties of genes connected with those.



- gene 1 is chosen by classifier

- penalty of gene 1 $\lambda_{1,i+1}$ is increased
- penalties $\lambda_{2,i+1}$, $\lambda_{3,i+1}$, $\lambda_{4,i+1}$ of connected genes are decreased
- classifier chooses gene 2

mRNA-miRNA Fusion

Motivation: if a gene is chosen, the regulating miRNAs of this gene might be important for the outcome as well or vice versa

- use graph W=1-pi,j as graph information between genes and miRNAs
- decrease penalties of connected miRNAs according to weights in W

Example Taylor Data

- 98 prostate cancer patients with mRNA and miRNA expression data
  - 18 with event → biochemical relapse
  - 80 censored

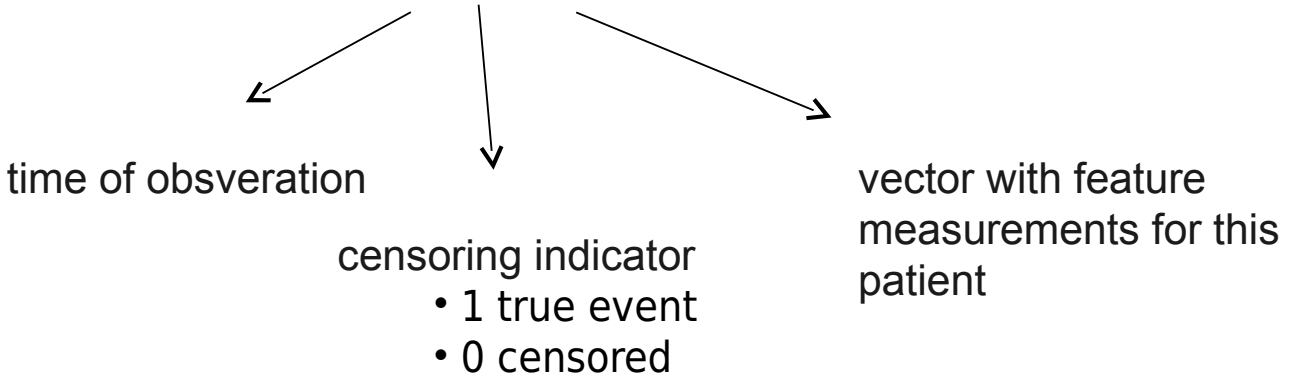
Cancer Cell
Article

Integrative Genomic Profiling of Human Prostate Cancer

Barry S. Taylor,[1,8] Nikolaus Schultz,[1,8] Haley Hieronymus,[2,8] Anuradha Gopalan,[3] Yonghong Xiao,[3] Brett S. Carver,[4] Vivek K. Arora,[2] Poorvi Kaushik,[1] Ethan Cerami,[1] Boris Reva,[1] Yevgeniy Antipin,[1] Nicholas Mitsiades,[5] Thomas Landers,[2] Igor Dolgalev,[2] John E. Major,[6] Manda Wilson,[6] Nicholas D. Socci,[6] Alex E. Lash,[6] Adriana Heguy,[2] James A. Eastham,[4] Howard I. Scher,[5] Victor E. Reuter,[3] Peter T. Scardino,[4] Chris Sander,[1] Charles L. Sawyers,[2,7,*] and William L. Gerald[2,3,9]
[1]Program in Computational Biology
[2]Program in Human Oncology and Pathogenesis (HOPP)
[3]Department of Pathology
[4]Department of Urology
[5]Department of Medicine
[6]Bioinformatics Core
Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA
[7]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, USA
[8]These authors contributed equally to this work
[9]Deceased
*Correspondence: sawyersc@mskcc.org
DOI 10.1016/j.ccr.2010.05.026

Time-to-event Data

- Observations: $(t_i, \delta_i, \mathbf{x}_i)$ for n patients and a given endpoint



- Cox porportinal hazards model:

$$h(t|\mathbf{x_i}) = h_0(t)\ \exp(\eta)$$

baseline hazard

linear predictor

$$\eta = \mathbf{x_i}^T \boldsymbol{\beta}$$

estimated by classifier

The Brier score

- from the estimates $\hat{\beta}$ the risk for a single patient can be calculated

$$\hat{r}(t|\mathbf{x_i}) = \exp(-\hat{H}_0(t)) \exp(\mathbf{x_i}^T \hat{\boldsymbol{\beta}})$$

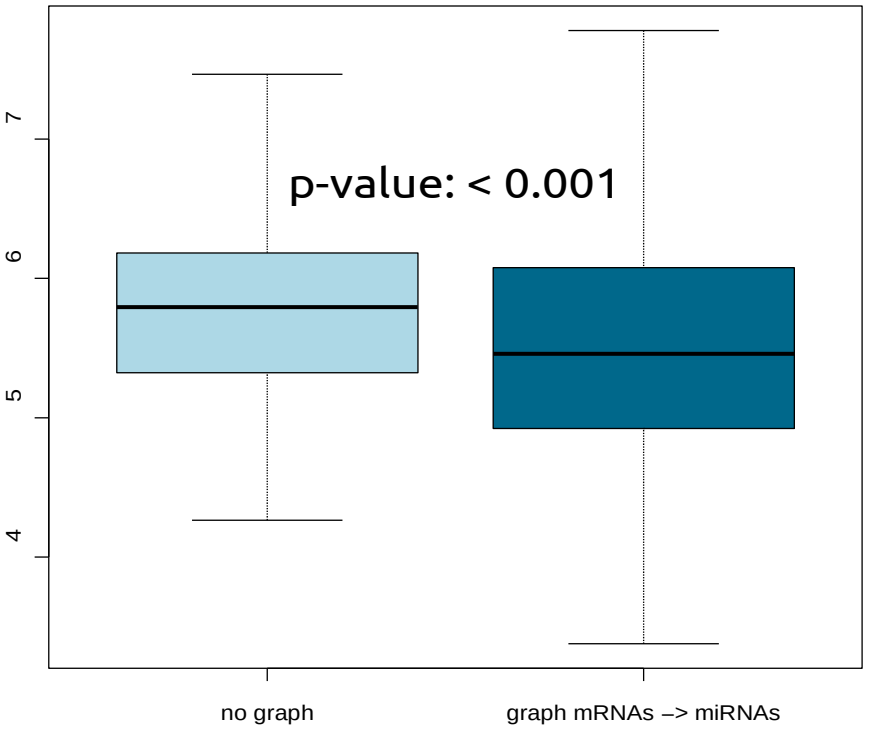
- the Brier score tracked over time can be calculated

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} (I(t_i > t) - \hat{r}(t|\mathbf{x_i}))^2$$

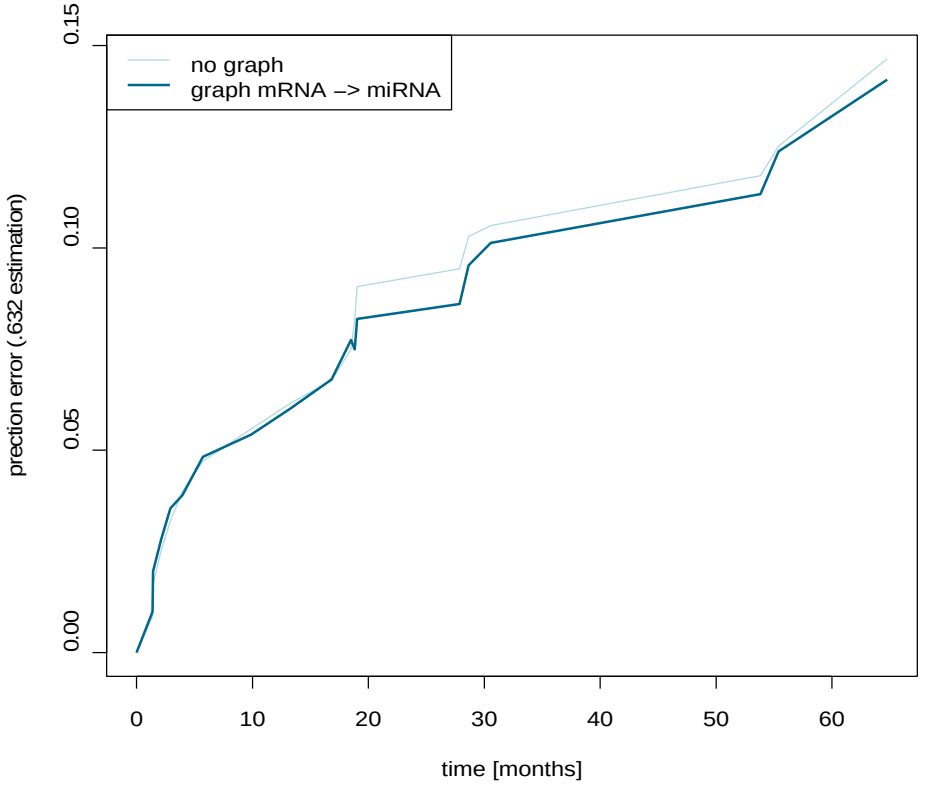
- in presence of censoring the Brier score has to be reweighted yielding the prediction error curve (PEC)

- integration over time gives the integrated prediction error curve (IPEC)

CoxBoost with and without graph

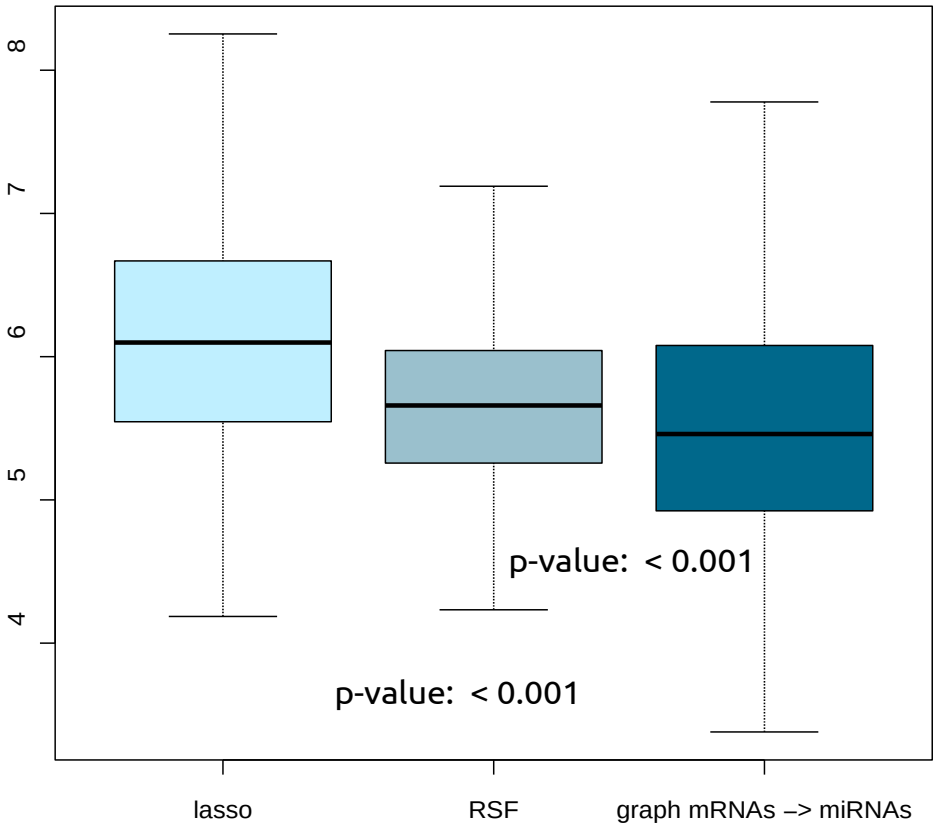


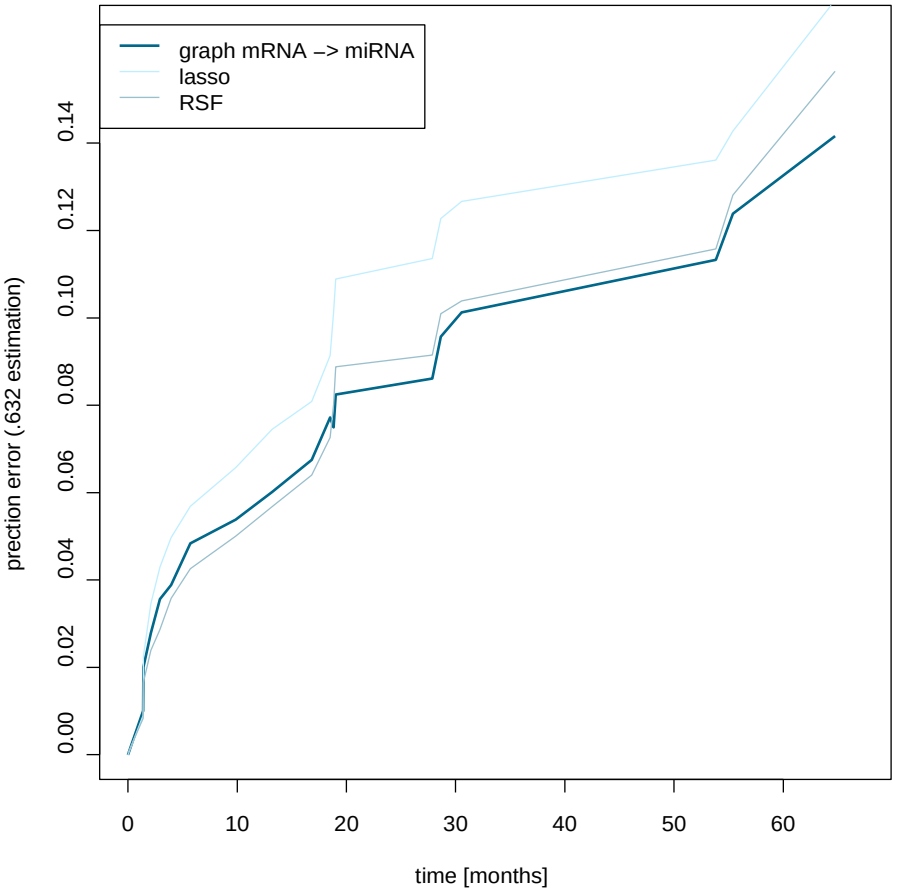
- no graph: CoxBoost with mRNA and miRNA data but no graph
- 500 IPECs of both classifiers
- Wilcox test with alternative “greater” to compare IPECs



- PECs from CoxBoost with and without graph
- prediction errors from 500 bootstrap samples are averaged

Comparison to other Methods





- compared to Lasso and Random Survival Forests
- mRNA and miRNA data given
- Wilcox test with alternative “greater” to test for differences in the 500 IPECs of all three classifiers

Summary Approach 2

- Combination of miRNA and mRNA profiles in a graph based approach.
- Feature selection is influenced in consecutive boosting steps by transferring weight from mRNAs to connected miRNAs.
- On a Prostate Cancer data set it could be demonstrated that this procedure may help to improve classification.

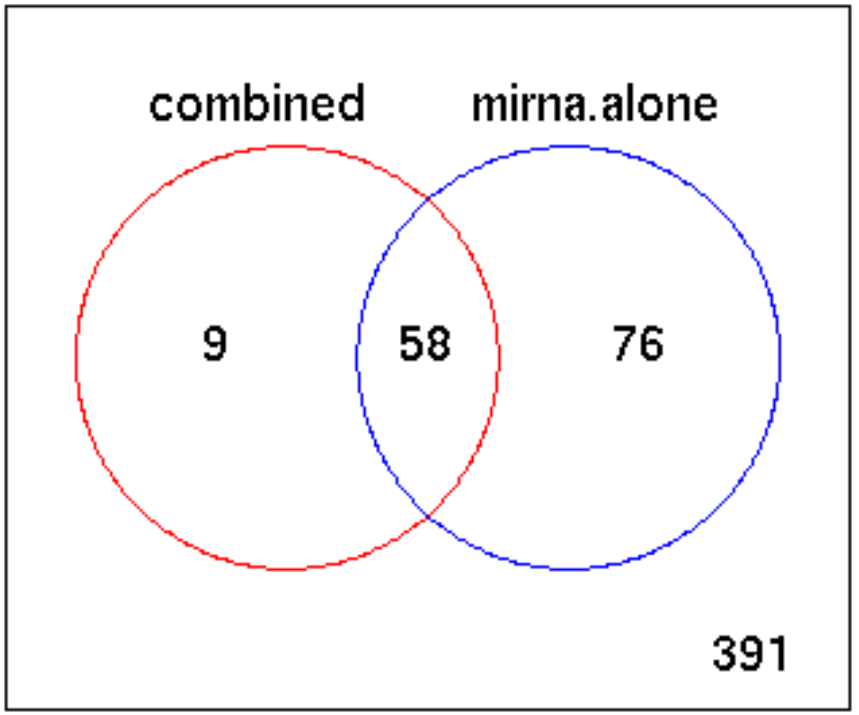
Application to CAMDA 2011 data set

- Glioblastoma data from The Cancer Genome Atlas.
  - gene transcript expression (435 cancer patients versus 11 controls)
  - miRNA expression (426 tumour samples versus 10 controls)
  - genomic DNA methylation (256 tumour samples versus a control)
  - copy number variation (465 tumour samples versus 430 controls [402 matched normals])
  - a variety of clinical parameters and survival outcomes
- Downloaded miRNA expression (Agilent) + mRNA expression (Agilent) => 418 matched samples.

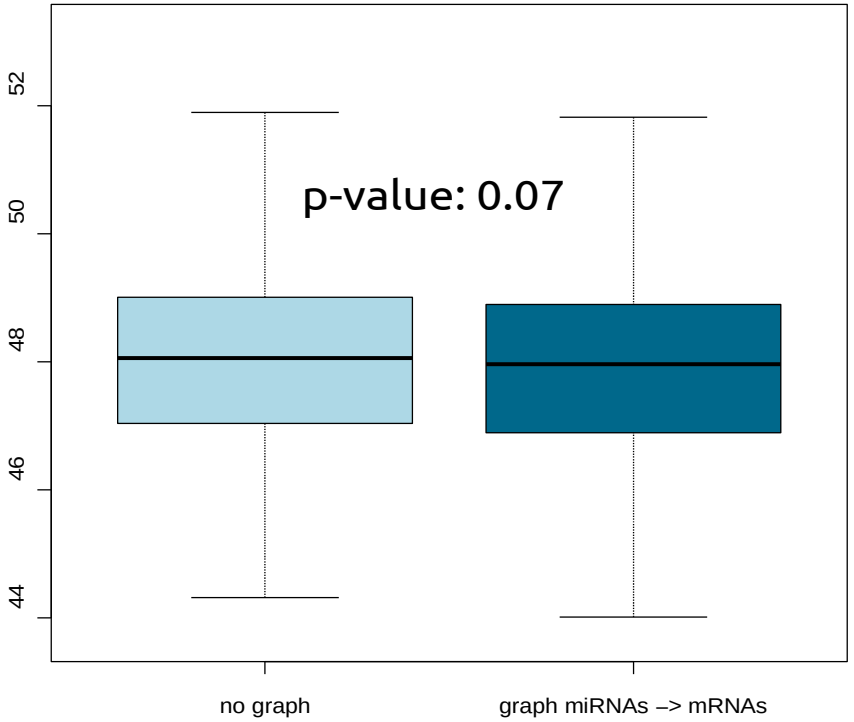
wget -r -l1 -nd -np -erobots=off --wait 8 -A.tar.gz http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/cgcc/unc.edu/agilentg4502a_07_2/transcriptome/

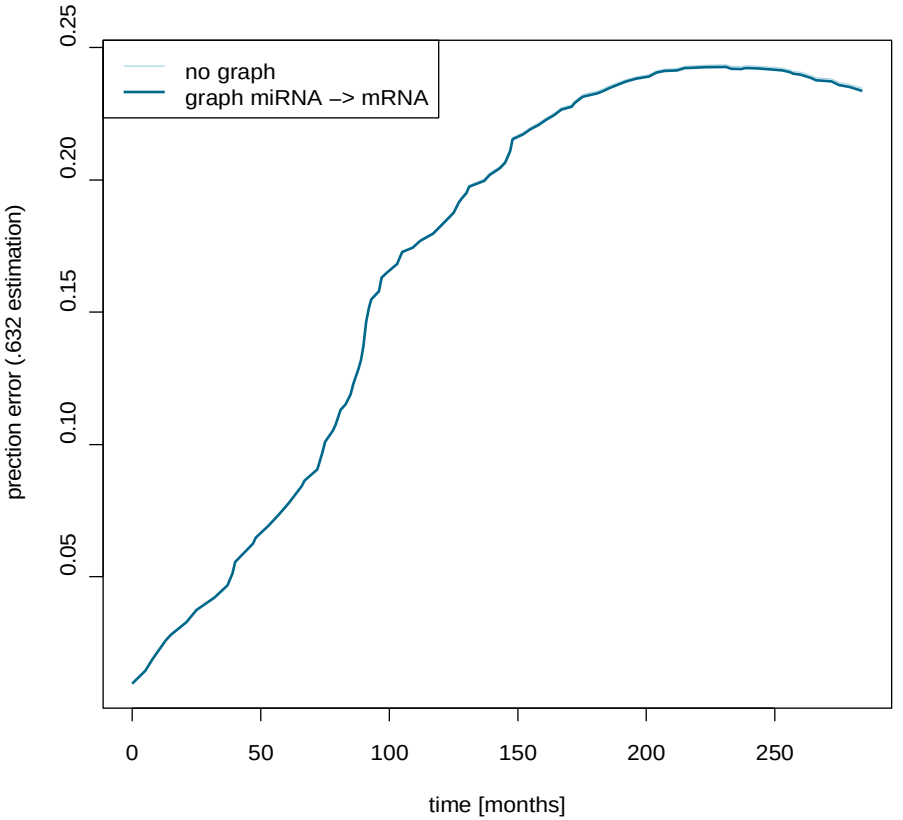
Approach 1: differential miRNAs

- Progression [201] vs. no progression [217]
- Differential miRNAs (here Wilcoxon Tests):



Approach 2: Classifier





- disease free survival as clinical endpoint
  - 291 patients with event (relapse or recurrence)
  - 127 patients censored
- 500 IPECs with every classifier
- Wilcox test with alternative “greater”