LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Proteome and Genome Research Unit

An Update on DEMICS Project and Future Research Plans

Petr Nazarov
petr.nazarov@lih.lu

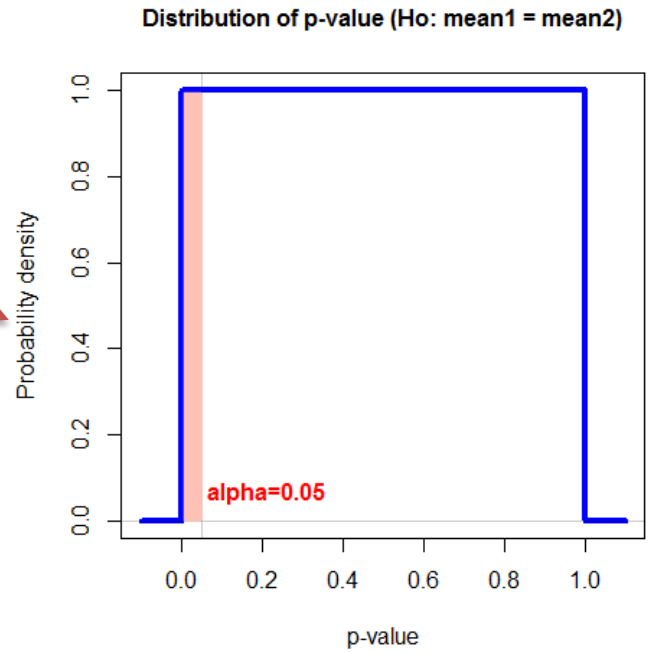
2018-09-04

Concept: Multiple Hypothesis Testing

Let's generate a completely random experiment:

6 "samples" = 3 "group A" and 3 "group B", 100 "genes". Then run a Student t-test.

Gene	A1	A2	A3	B1	B2	B3	|	p-value
gene001	-0.95	-1.43	-0.66	0.983	0.978	0.89	|	0.0009
gene002	-0.96	-1.97	1.376	-0.31	0.005	-0.48	|	0.8085
gene003	0.542	1.569	0.479	-1.01	2.602	-1.06	|	0.6164
gene004	-1.71	-0.38	0.07	-0.26	-1.1	1.24	|	0.5069
gene005	-0.78	-0.79	0.039	-0.08	-0.41	-1.01	|	0.9807
gene006	-0.49	-0.26	-0.53	-1.52	-1.81	0.687	|	0.5996
gene007	0.354	0.469	0.256	0.759	0.571	-0.91	|	0.7004
gene008	-1.34	-0.36	0.753	0.623	-1.14	1.678	|	0.529
			. . .					
gene097	-1.26	-1.46	-1.45	-0.25	0.636	-0.5	|	0.0183
gene098	0.783	1.004	-1.03	0.088	0.314	0.32	|	0.9845
gene099	-0.23	-0.66	1.171	-1.61	-0.72	-0.08	|	0.2766
gene100	0.874	-2.03	-1.31	0.067	-0.23	-0.47	|	0.5287



Some p-values < 0.05. For 100 genes you should expect 5 genes with p-value<0.05

If you repeat this experiment, you discover another ~5 genes. But they will be different!

Concept: Multiple Hypothesis Testing

Multiple Hypotheses: False Discovery Rate

False discovery rate (FDR)

FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

Conclusion	Population Condition: H_0 is TRUE	H_0 is FALSE	Total
Accept H_0 (non-significant)	U	T	$m - R$
Reject H_0 (significant)	V	S	R
Total	m_0	$m - m_0$	m

False Positives, (α error)

$$FDR = E\left(\frac{V}{V+S}\right)$$

Concept: Multiple Hypothesis Testing

False Discovery Rate (FDR): Benjamini & Hochberg

Assume we need to perform ***m*** comparisons and select acceptable **FDR = α = 0.05**

1. Run ***m*** t-tests and sort "genes" by p-value ***P***
2. Assign rank ***k*** : smallest p-value gets ***k*** = 1, largest gets ***k*** = ***m***

Expected value for FDR < α if $P_{(k)} < \frac{k}{m}\alpha$ ⇒ $\frac{mP_{(k)}}{k} < \alpha$

$$FDR = E\left(\frac{V}{V+S}\right)$$

Theoretically, the sign should be "≤".
But for practical reasons it is replaced by "<"

Familywise Error Rate (FWER)

Probability of making at least one mistake

Bonferroni – simple, but too stringent, *not recommended* $mP_{(k)} < \alpha$

Holm-Bonferroni – a more powerful, less stringent but still universal $(m+1-k)P_{(k)} < \alpha$

Lab-meeting Outline

- **DEMICS**
  - short reminder
  - challenges and achievements
  - comparing to BIODICA (the tool developed in Paris & Astana)

- **MelanomICA: an application of ICA to melanoma**
  - correcting technical biases
  - patient group prediction
  - prognosis for the new patients
  - biological processes in the new samples

- **Ideas for future**
  - ICA as a tool for “omics” data integration
  - miRNA functional annotation based on ICA

DEMICS Project

Fonds National de la Recherche Luxembourg

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

People

Dr. Francisco AZUAJE

Prof. Inge JONASSEN

Arnaud MULLER

Dr. Anna GOLEBIEWSKA

Tony KAOMA

Dr. Victoria EL KHOURY

Laurene PICANDET

Dr. Andrei ZINOVYEV

MSc student 2

Plans & Reality ☺

WP1. Improvement of the ICA method

1st January – 30th September

- pipelines implemented at HPC
- transcription signals investigated
- ICA results linked to clinical data

WP2. Development of ICA-based classifier

1st October – 28th February

- the best classifier selected
- patient classes on validation dataset 1 are predicted

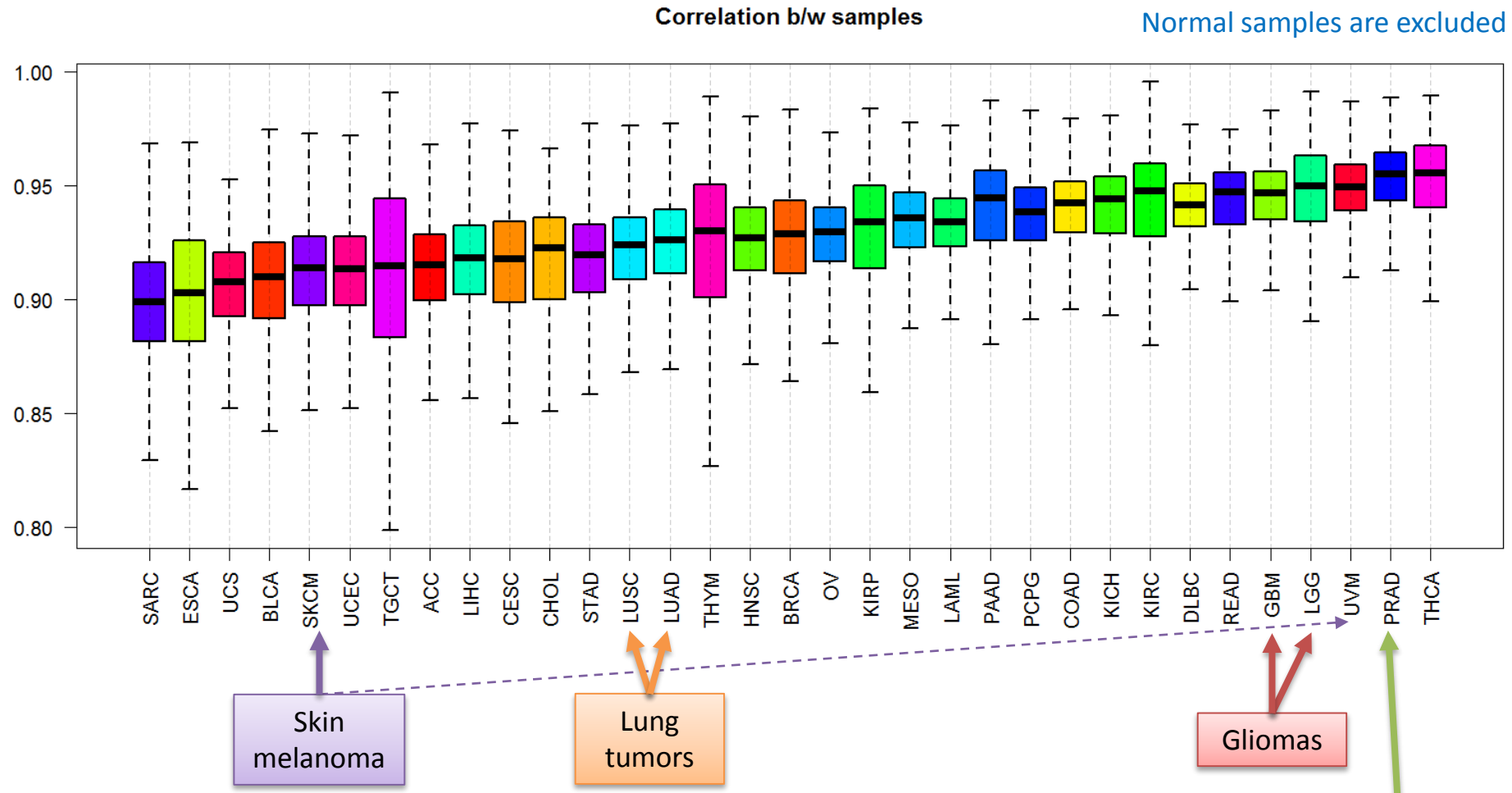
Challenges

- LGG and GBM are not very exciting tumors: IDH1 mutation and 1p/19q co-deletion are the 2 main factors affecting the data.
- As shown by Lorena, even random 100 genes can classify the samples (Acc ≈ 0.94)
- LGG and GBM are not so heterogeneous, compared to other tumors
- Exon-exon junction counts did not improve the classification (low coverage)

Achievements

- Our method of consensus **ICA works** (better than the one from collaborators so far)
- We obtained nice results on SKCM (melanoma) and **submitted a manuscript**
- In principle, 90% of WP1 and 50% of WP2 are **done**
- Some **new ideas** – to be discussed

Heterogeneity in Cancers



SARC – sarcoma (multiple tissues)
ESCA – esophageal carcinoma
UCS – uterine carcinoma

THCA – thyroid carcinoma
PRAD – prostate carcinoma
UVM – uveal melanoma

Comparing Consensus ICA Algorithms

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

consICA LIH

- Using R-package *fastICA*
- Consensus = mean
- Multiple runs excluding one sample, with different initial estimations
- Multiplatform
- Multicore
- No GUI

BIODICA Institut Curie

- Using *fastICA* implemented in MATLAB
- Consensus = "centroid"
- Multiple runs with different initial estimations
- Multiplatform
- Multicore (?) (CPU load as for 4x cores)
- User-friendly

Comparison on melanoma tumors (SKCM TCGA) : 477 samples, 16579 genes

Observation 1: **BIODICA** is ~6 times faster than **consICA**

consICA -> 72 min

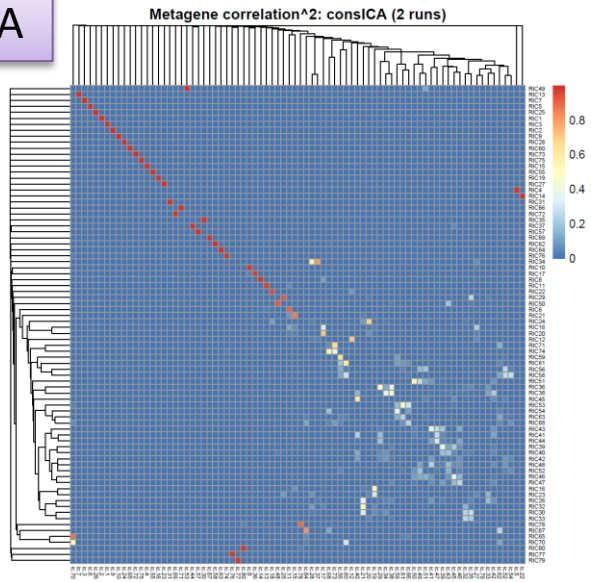
BIODICA -> 12 min

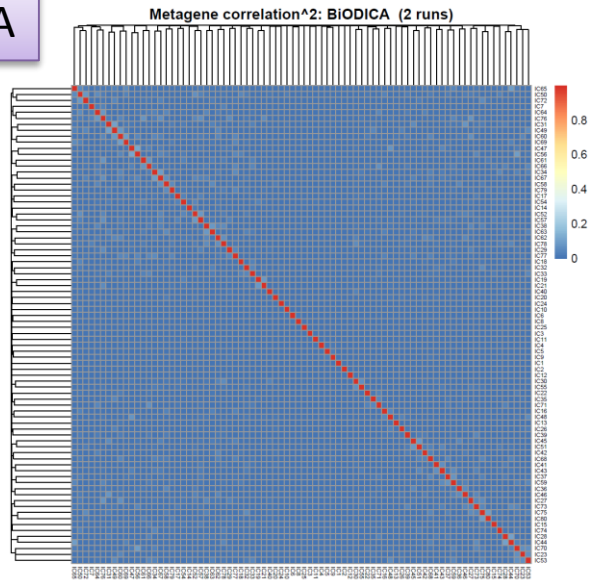
Comparing Consensus ICA Algorithms

consICA

BIODICA

S-matrix

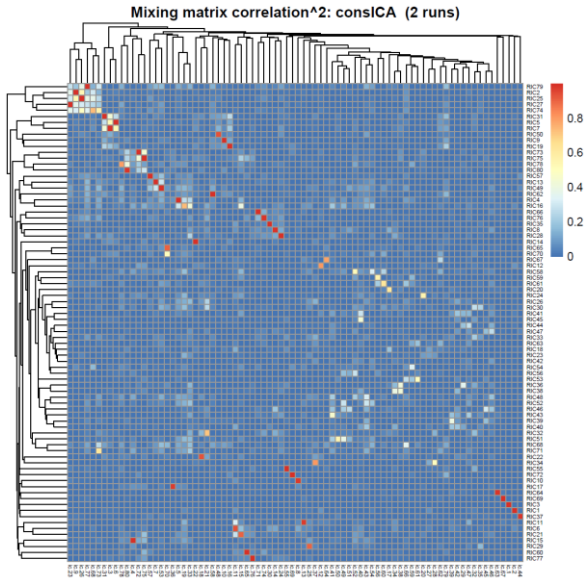


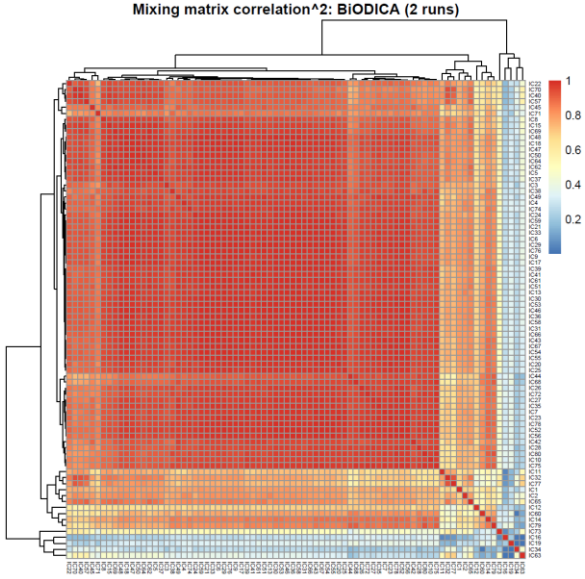


BIODICA: more stable

consICA: more realistic, imho ☺

M-matrix



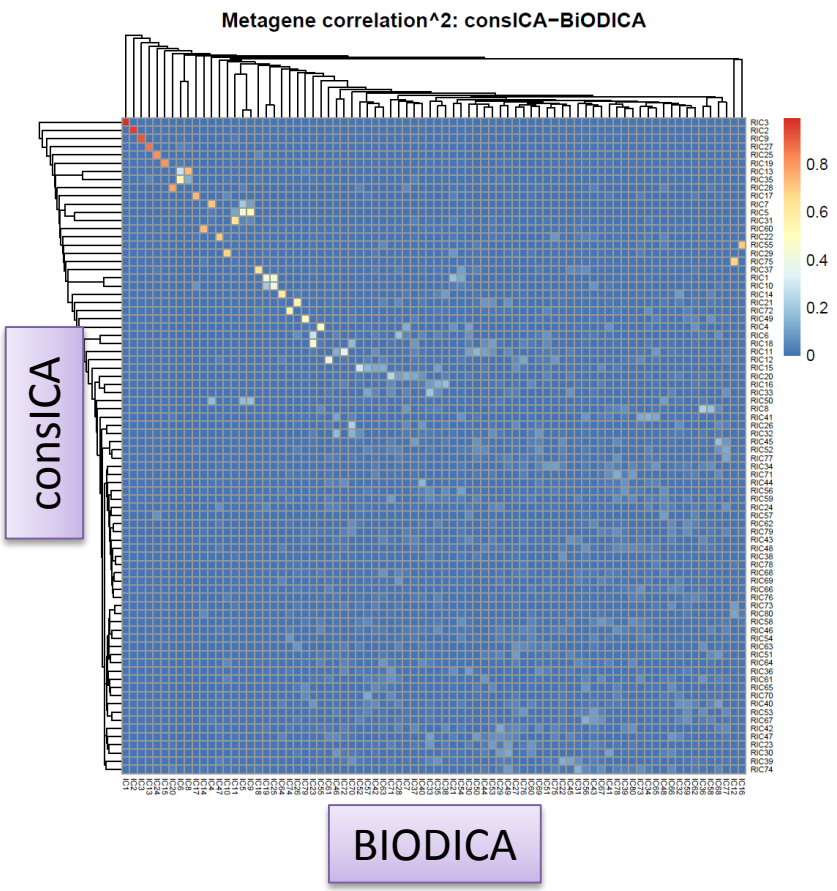


BIODICA: extremely correlated weight matrix!

consICA: more realistic ☺

Comparing Consensus ICA Algorithms

S-matrix

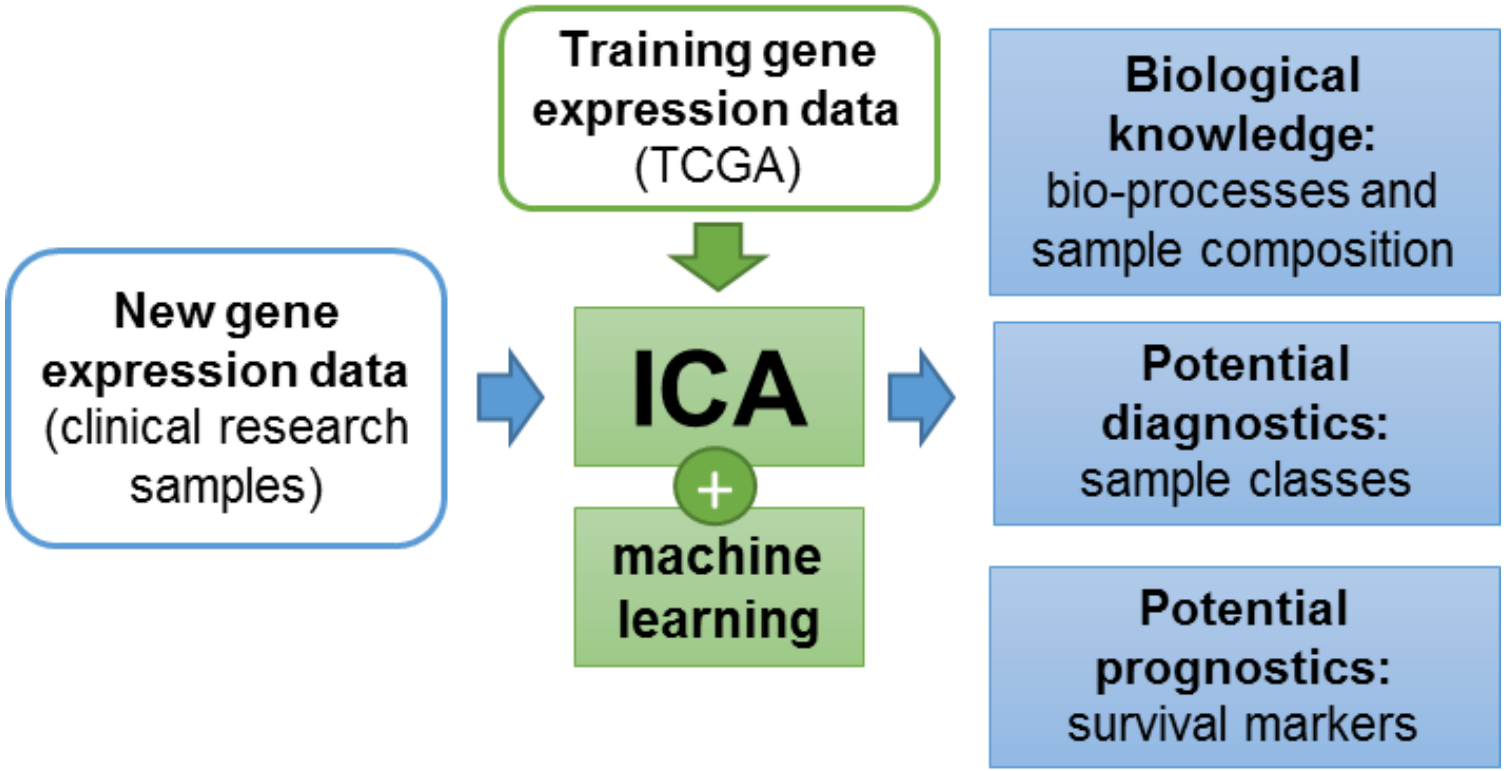


- The strongest signals are recovered by both algorithms
- The discrepancy for mixing matrix of **BIODICA** is under investigation by Paris and Astana teams now
- => we aim at our **consICA** method for the moment.

https://gitlab.com/biomodlih/consica

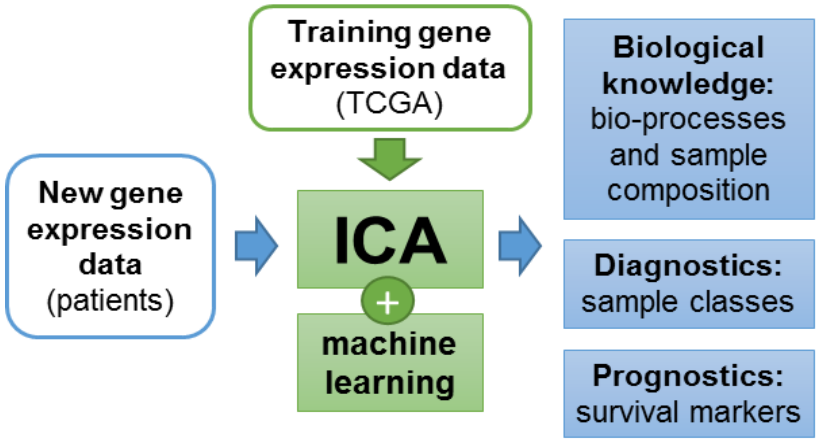
MelanomICA: Method

ICA to study new patients



Preprint is available at
https://www.biorxiv.org/content/early/2018/08/20/395145

MelanomICA

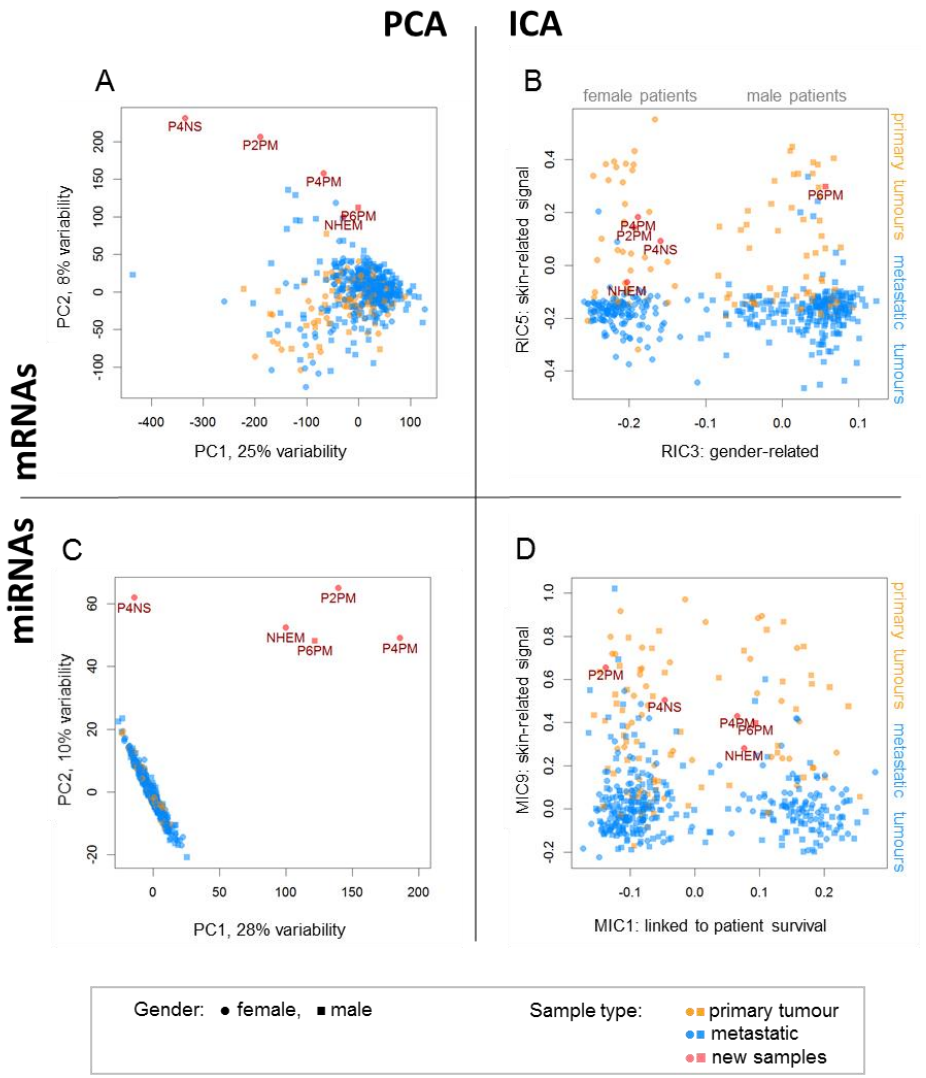


RNA-seq + miRNA

Reference data: 472 samples

Validation data: 44 samples

Investigation data: 5 samples



Conclusion 1:

Consensus ICA can correct technical biases between platforms

MelanomICA

Accuracy: 90.9%	Actual tumour cluster:		
	immune	keratin	MITF-low
immune	158	4	8
keratin	9	98	6
MITF-low	3	0	45

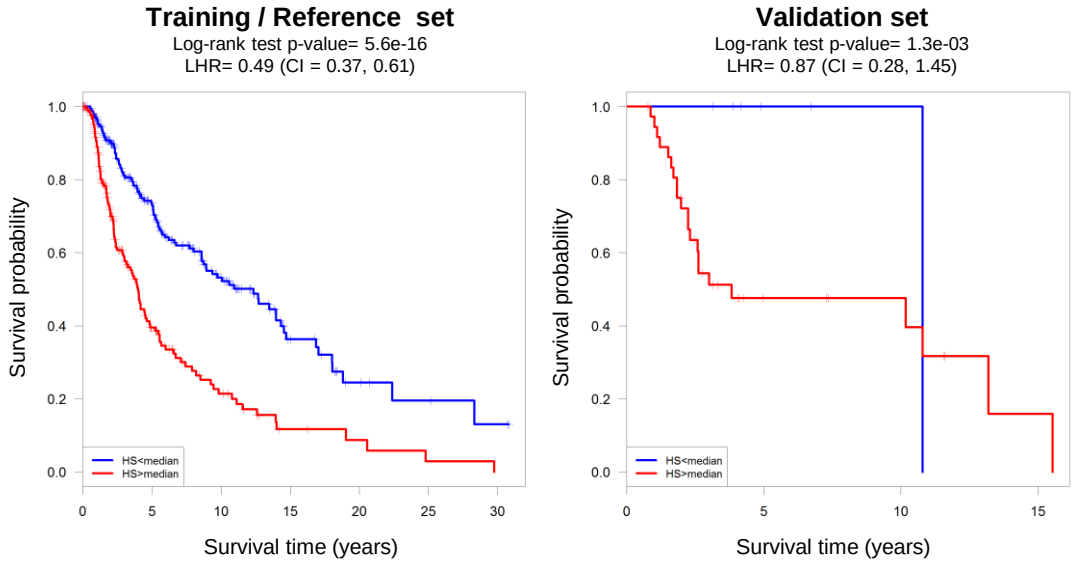
Accuracy: 91.3%	Actual sample type:	
	metastatic	primary
metastatic	364	38
primary	3	67

Hazard score

$$HS_j = \sum_{i=1}^{k} H_i \, R_i^2 M_{i,j}^*$$

$$H_i = \begin{cases} LHR & for\ significant\ components \\ 0 & for\ non-significant\ components \end{cases}$$

44 metastatic patients



Conclusion 2:

Consensus ICA can be used to predict cancer subtype and patient survival

MelanomICA: Results

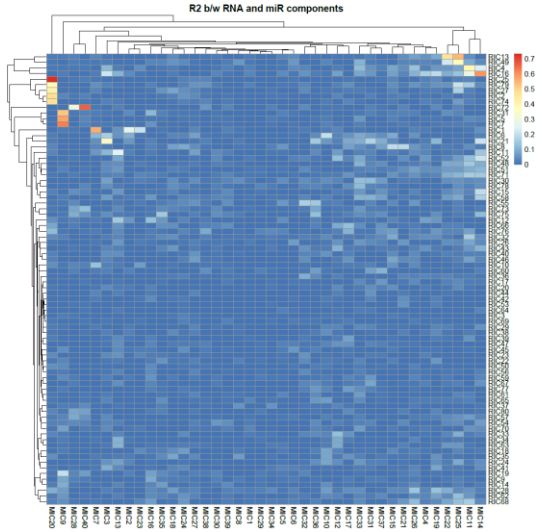
Cluster	Component	Risk (p-value)	Meaning	P2PM	P4PM	P6PM	P4NS	NHEM
Immune	RIC2	decreased (1.8e-4)	B cells	0.11	0.07	0.02	0.19	0.01
	RIC25	decreased (2.8e-7)	T cells	0.26	0.06	0.24	0.18	0.00
	RIC27	no effect	B cells	0.80	0.37	0.31	0.80	0.00
	RIC28	no effect	response to wounding	0.34	0.57	0.78	0.43	0.84
	RIC37	no effect	IFN signalling pathway	0.97	0.66	0.99	0.90	1.00
	RIC57	no effect	monocytes	0.00	0.25	0.24	0.02	0.00
	MIC20	decreased (1.2e-4)	T cells, chr1q32.2	0.14	0.08	0.37	0.02	0.19
Stromal and angiogenic	RIC13	no effect	cells of stroma	0.81	0.40	0.50	0.86	0.03
	RIC49	no effect	endothelial cells	0.73	0.12	0.29	0.84	0.00
	MIC22	no effect	miR-379/miR-410 cluster, chr14q32.2,14q32.31	0.29	0.20	0.27	0.38	0.16
	MIC25	no effect	potentially related to stromal cells; clusters: chr1q24.3, 5q32, 17p13.1, 21q21.1	0.97	0.85	0.76	0.80	0.26
Skin-related	RIC5	increased (5.8e-3)	epidermis development and keratinisation	0.92	0.93	0.96	0.92	0.87
	RIC7	increased (8.9e-6)	epidermis development and keratinisation	0.94	0.93	0.93	0.95	0.57
	RIC19	increased (4.0e-2)	epidermis development and keratinisation	1.00	0.62	0.22	1.00	0.93
	RIC31	increased (2.2e-2)	epidermis development and keratinisation	0.98	0.85	0.89	0.99	0.28
	MIC9	increased (2.9e-2)	skin-specific miRNAs	0.95	0.88	0.87	0.91	0.83
Melanocytes	RIC4	increased (5.4e-3)	melanin biosynthesis	0.62	0.77	1.00	0.21	0.96
	RIC16	decreased (5.1e-4)	melanosomes (negative gene list)	0.68	0.77	0.54	0.75	0.39
	MIC11	no effect	potential regulators of malignant cells, chrXq27.3	0.21	0.96	0.62	0.13	0.48
	MIC14	decreased (1.5e-2)	potential regulators of melanocytes, chrXq26.3	0.01	0.29	0.67	0.29	0.38
Other	RIC55	increased (3.0e-2)	cell cycle	0.48	0.46	0.88	0.00	0.53
	RIC6	decreased (5.5e-3)	potentially linked to neuron differentiation	0.43	0.73	0.59	0.46	0.01
	MIC1	increased (9.4e-4)	regulators of EMT	0.11	0.07	0.02	0.19	0.01

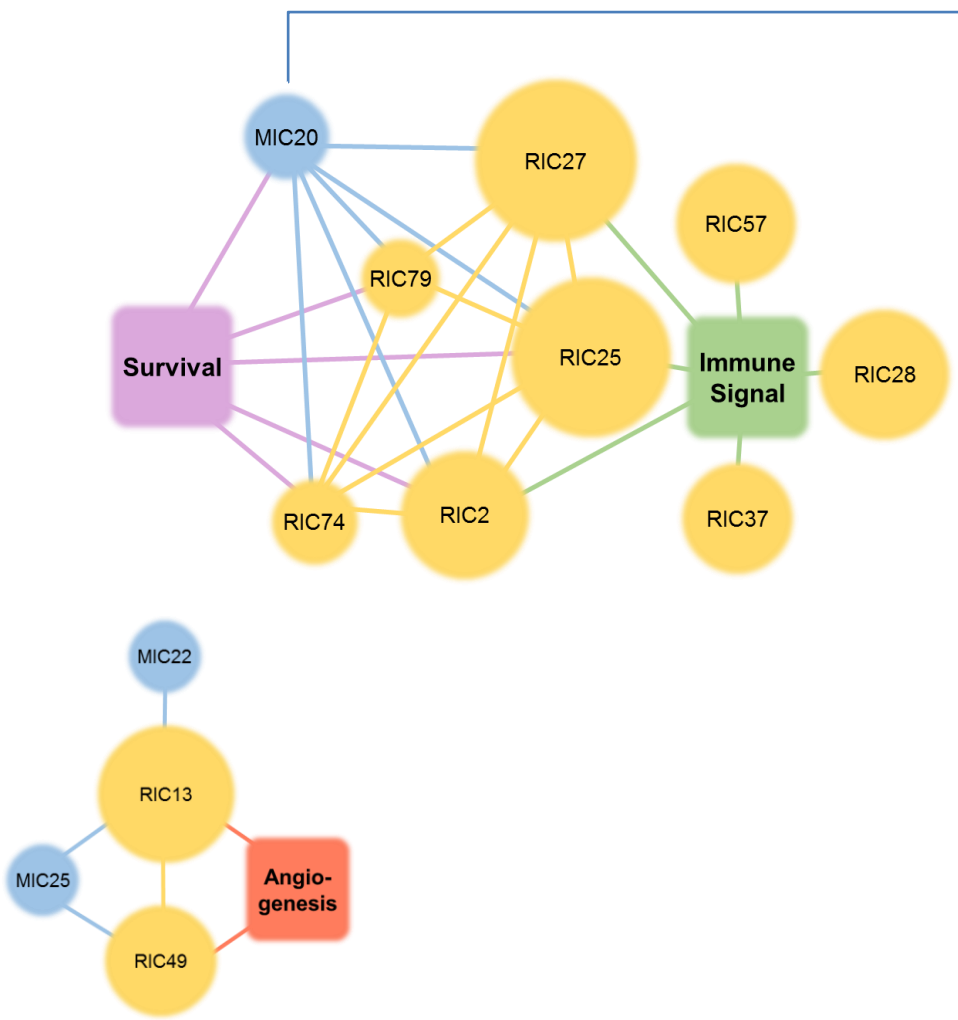
Conclusion 3:

Consensus ICA can be used to get biological knowledge about the new samples

MelanomICA: Results

Correlation of weights:
mRNA-miRNA





hsa-miR-1247-5p
hsa-miR-142-3p
hsa-miR-142-5p
hsa-miR-146b-3p
hsa-miR-150-5p
hsa-miR-155-3p
hsa-miR-155-5p
hsa-miR-1976
hsa-miR-223-3p
hsa-miR-223-5p
hsa-miR-29c-3p
hsa-miR-342-3p
hsa-miR-342-5p
hsa-miR-3614-3p
hsa-miR-511-5p
hsa-miR-642a-5p
hsa-miR-146b-5p
hsa-miR-766-3p
hsa-miR-625-3p
hsa-miR-653-5p

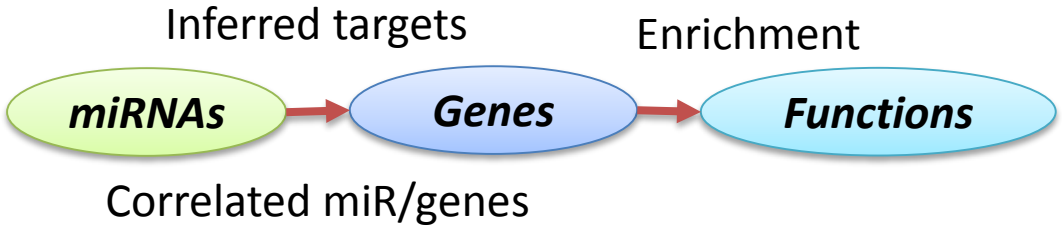
Conclusion 4:

Consensus ICA can be used to integrate the data and assign functions to miRNAs

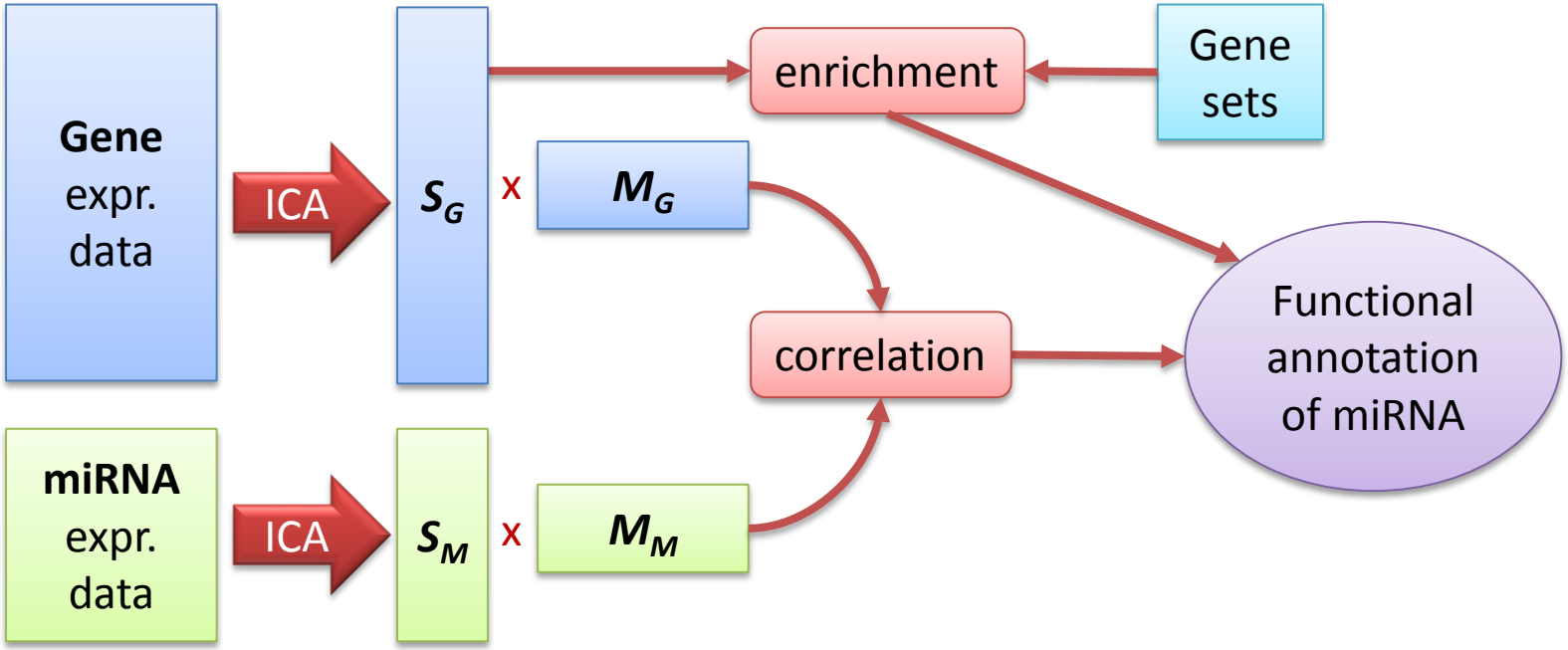
ICA-based Data Integration

ECCB poster: Multi-omics data integration using parallel consensus independent component analysis

"Classical" approach:

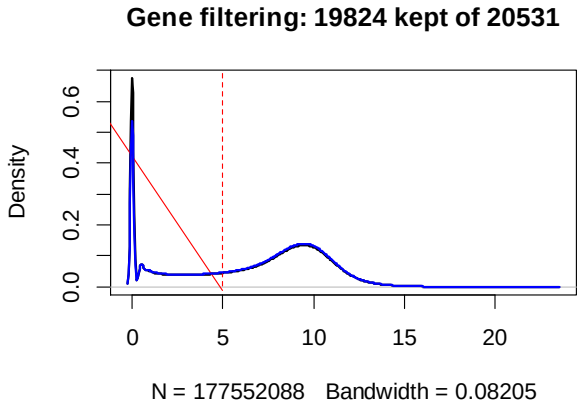


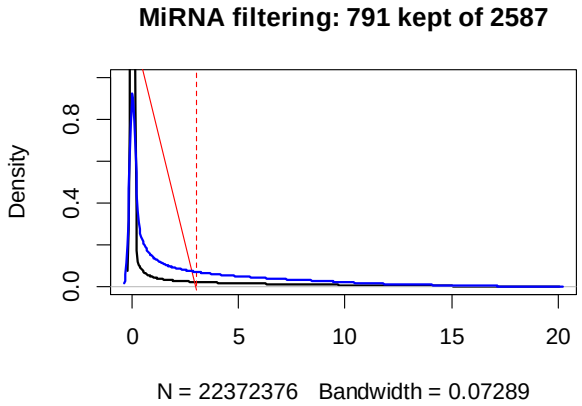
Proposed approach:



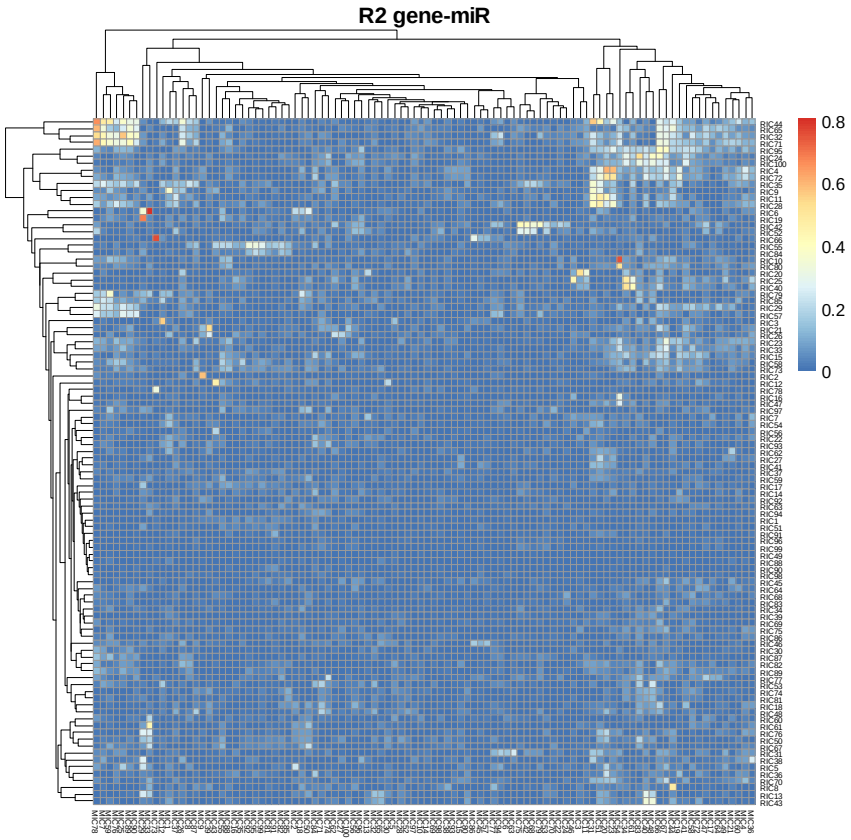
ICA-based Data Integration

TCGA, paired mRNA / miRNA data: **8648** samples, 20531 genes, 2587 miRNAs
After filtering uninformative: 19824 genes, 791 miRNAs





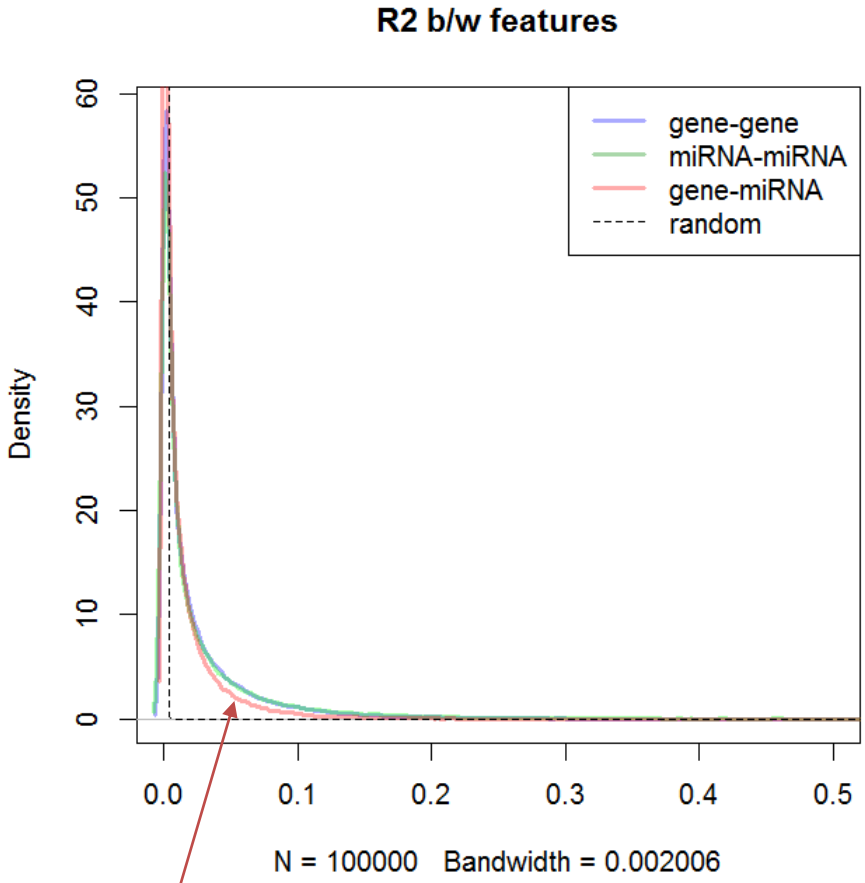
ICA: 100 runs, 100 components

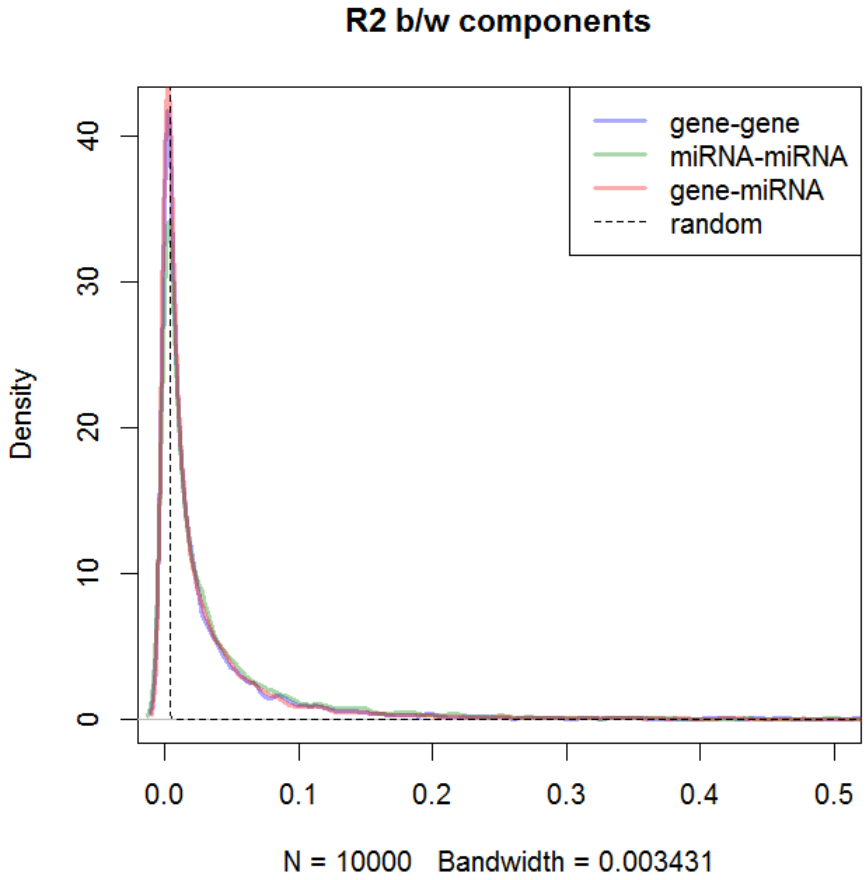


Observation: RAM is limiting factor

ICA-based Data Integration

Correlation properties

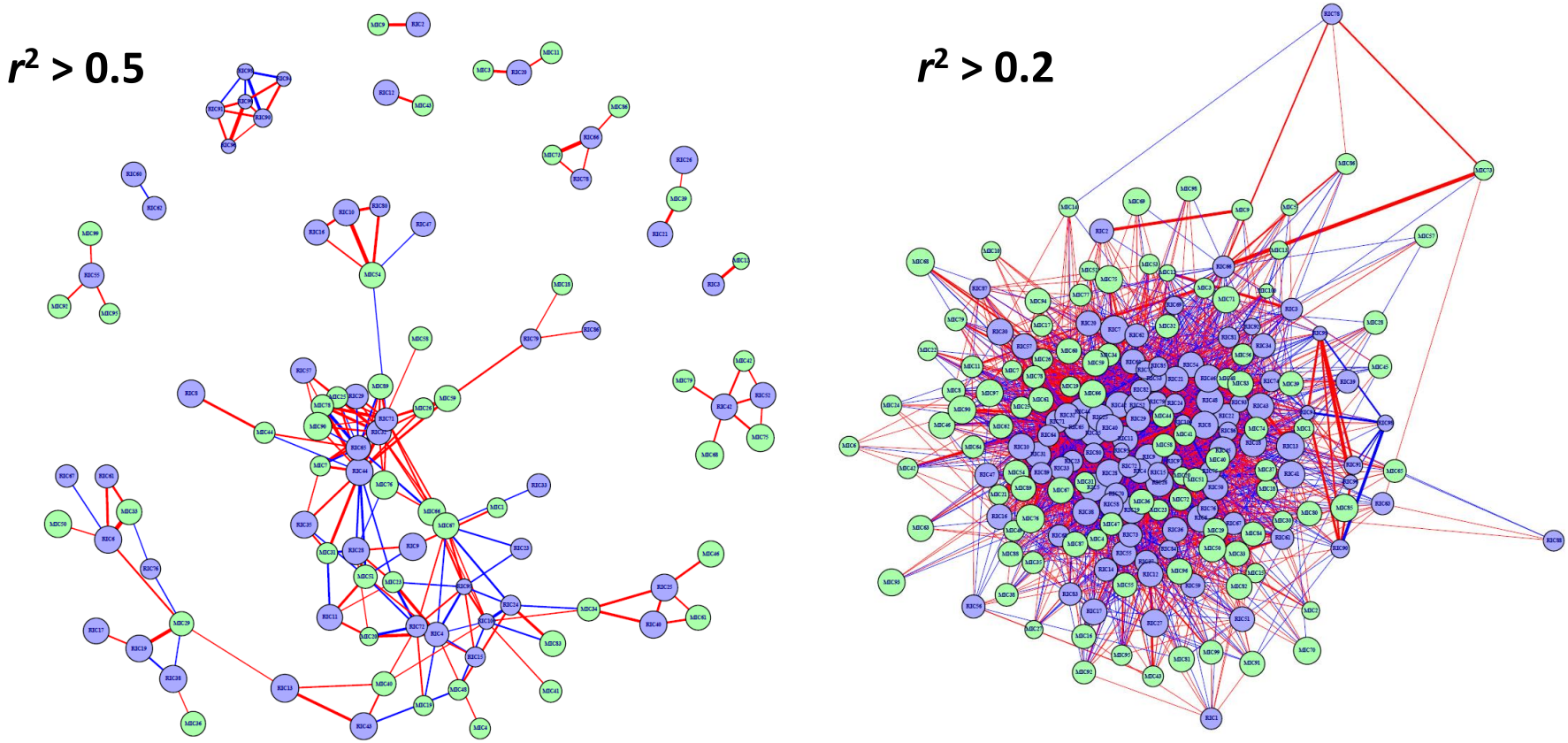




Gene-miR shows lower correlation, as sample effect is removed. Not seen in ICA results

ICA-based Data Integration

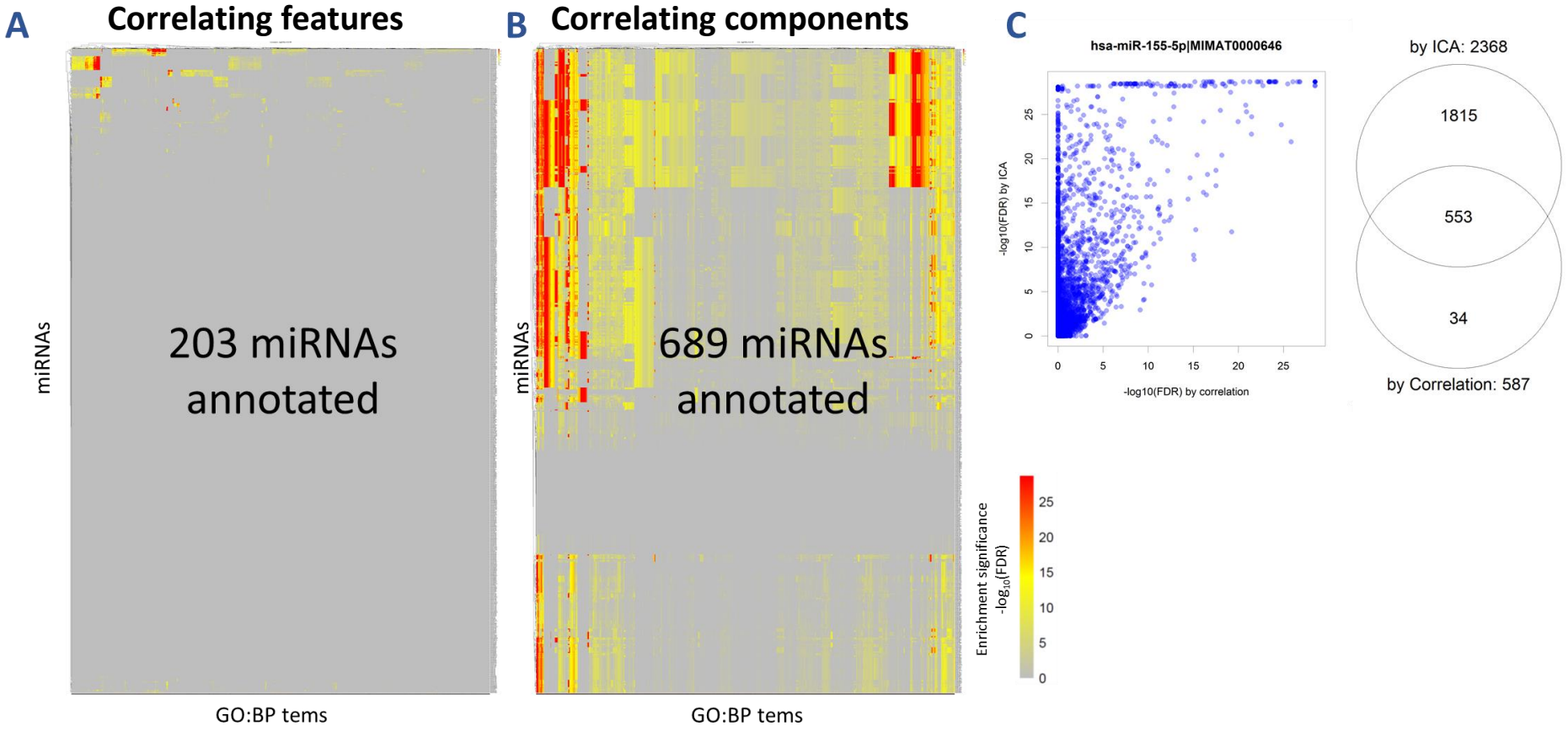
Networks of the ICA components



Networks composed of correlated miRNA (● MIC) and mRNA components (● RIC) for two correlation cut-offs. Edge colour represents correlation (– positive, – negative). Size of a node represents relative number of contributing genes and miRNAs in it.

ICA-based Data Integration

GO annotation



Results of miRNA annotation using a direct approach (A) and proposed method (B). Heatmap colour represents $-\log_{10}$(FDR) of the hypergeometric test used in enrichment analysis. (C) Scatter of $-\log_{10}$(FDR) for miR-155-5p and comparison of enriched GO terms (FDR<0.001).

Future Plans

DEMICS

- Finalize LGG/GBM part for the annual FNR report
- Optional: try exon level data instead of junctions?
- Work on WP2: prediction / classification task. Include a new cohort (Chinese)
- Hire a MSc student for 2019. But 💰 can be an issue (only 400-500 per month).

Data Integration

- Can we aim at a publication: ***ICA-based miRNA function prediction*** ?
  - It could be a DB or software note
- We need to prove that our predictions are relevant and are not composed of false hits
  - How? Literature search?
- In addition to gene-miR correlation, we should consider miR-target approach. This is the most accepted method (however I was not impressed, when I tried)

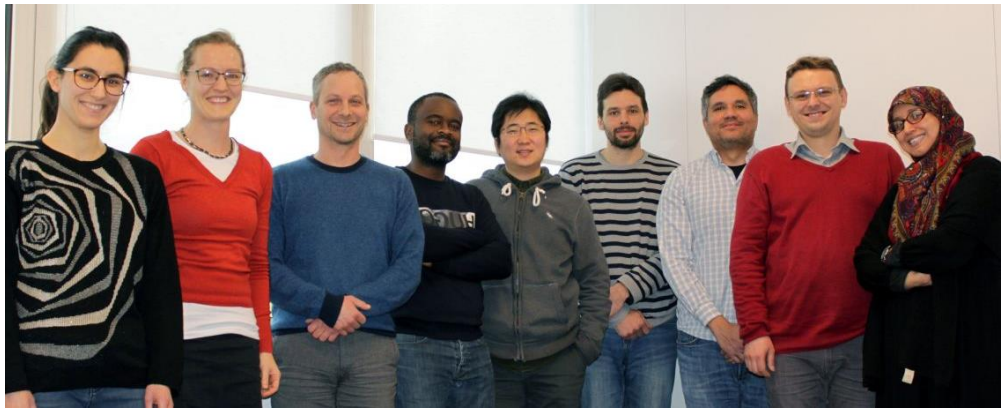
Conclusions

- We tested our implementation of consensus ICA, that decomposes large bulk data set into meaningful signals
- The hypothesis of “junctions” is not supported. However other hypotheses of DEMICS are.
- New samples are properly mapped in IC-space
- The method allows classifying and scoring new patients => can be used for diagnostics and building prognosis.
- The method allows linking miRNA to mRNA and thus predicting miRNA functions

Acknowledgements

Proteome and Genome Research Unit, Luxembourg Institute of Health (LIH)

Tony KAOMA
Arnaud MULLER
and other BIOMOD members
Dr. Francisco AZUAJE
Dr. Gunnar DITTMAR



LSRU, University of Luxembourg

Dr. Anke WIENECKE
Dr. Stephanie KREIS

Institute Curie, France

Urszula Czerwinska
Dr. Andrei ZINOVYEV

This work was supported by Luxembourg National Research Fund (C17/BM/11664971/DEMICS)