Fonds National de la Recherche Luxembourg

LUXEMBOURG INSTITUTE OF HEALTH

Multiomics Data Science, Quantitative Biology Unit

Deconvolution of Mixed Transcriptomes Improves Characterization of Cancer Patients

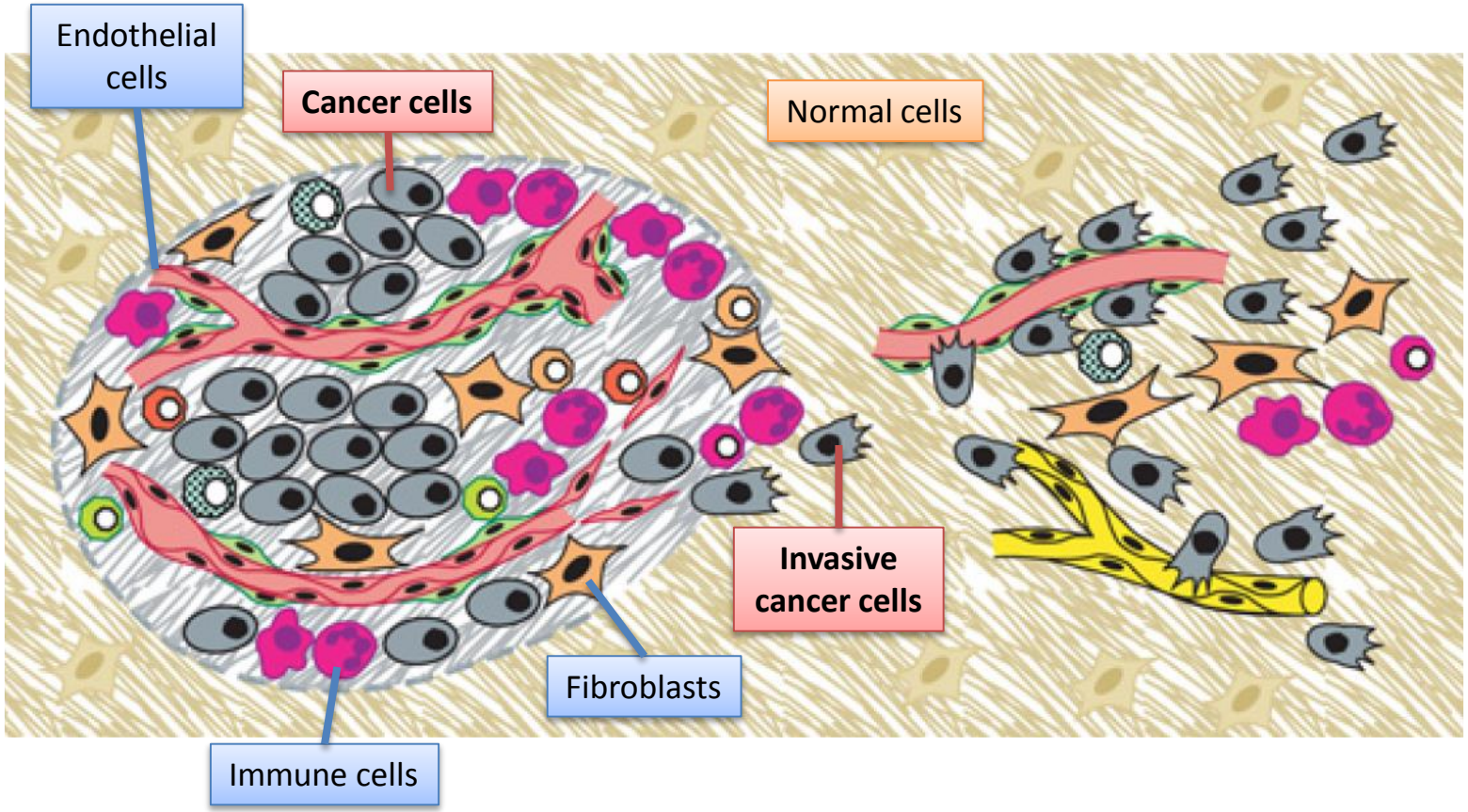
Petr V Nazarov, Thomas Eveno, Maryna Chepeleva, Tony Kaoma, Arnaud Muller, Francisco Azuaje

petr.nazarov@lih.lu

DKFZ, Heidelberg

Issues

Imagine we are going to analyze RNA from a tumor biopsy (sample):



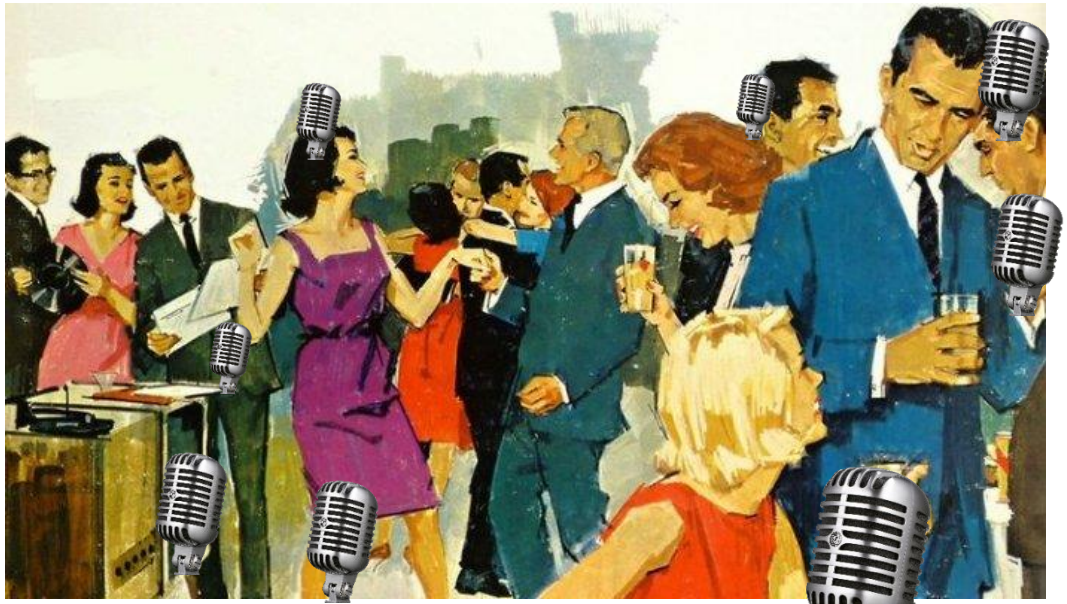
- Sample native heterogeneity
- Inter/intra tumour heterogeneity
- Technical heterogeneity

Hanahan D, Weinberg RA. ***Cell* 2011**, 144, 646-74

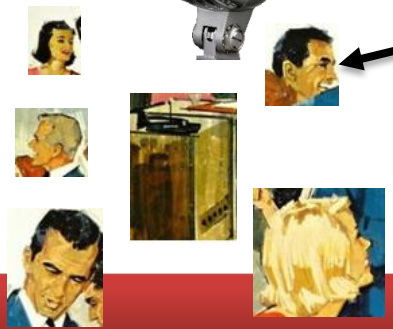
Independent Component Analysis

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

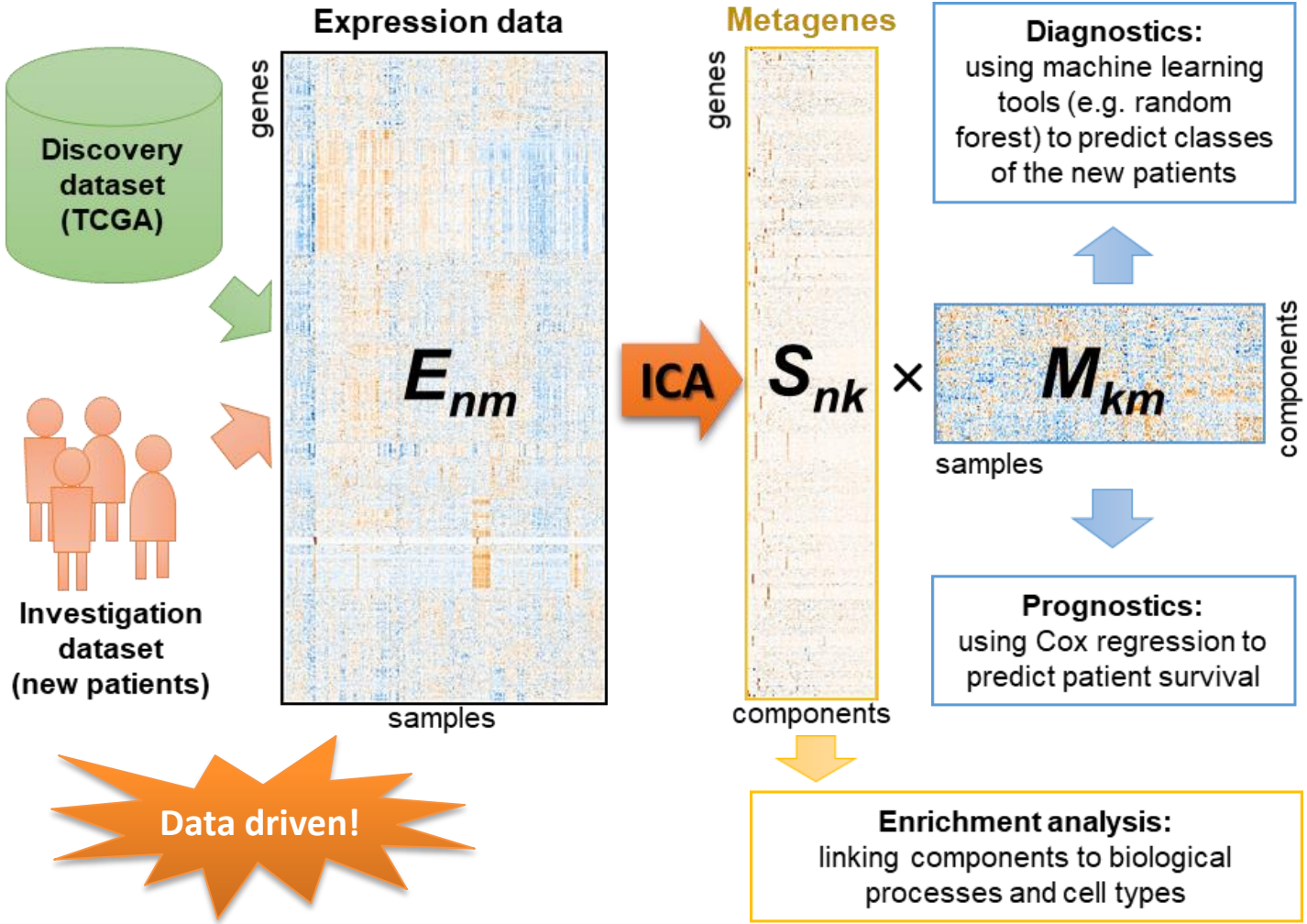
One of the methods to solve cocktail party problem…

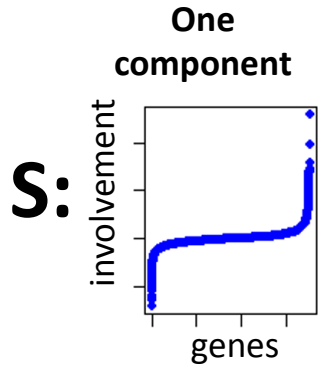


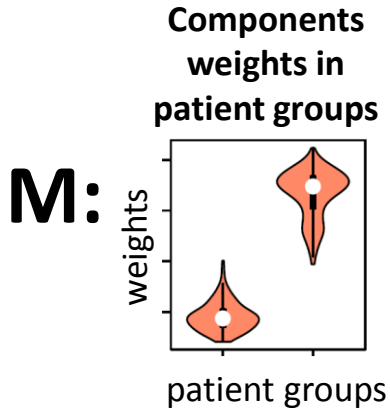
Independent Component Analysis



Method


Expression data
Discovery dataset (TCGA)
genes
E_{nm}
samples
Investigation dataset (new patients)
ICA
Metagenes
genes
S_{nk}
components
×
M_{km}
samples
components
Diagnostics: using machine learning tools (e.g. random forest) to predict classes of the new patients
Prognostics: using Cox regression to predict patient survival
Enrichment analysis: linking components to biological processes and cell types
Data driven!

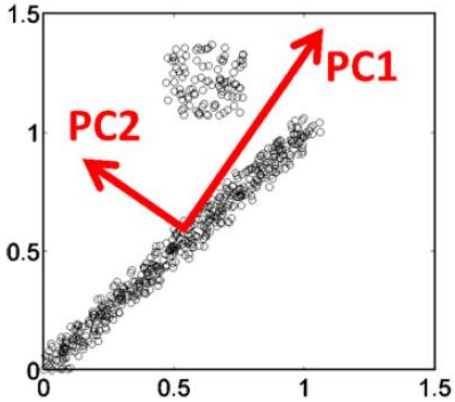

One component
S:
involvement
genes


Components weights in patient groups
M:
weights
patient groups

Independent Component Analysis

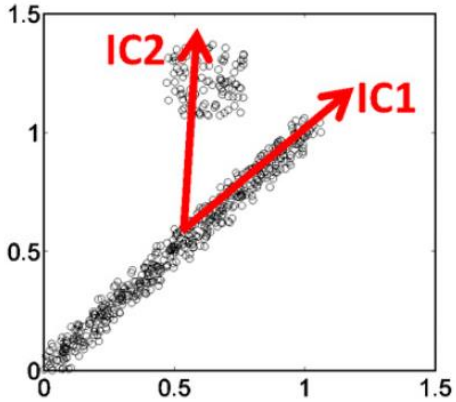
Geometrical view on differences b/w matrix factorization methods

PCA



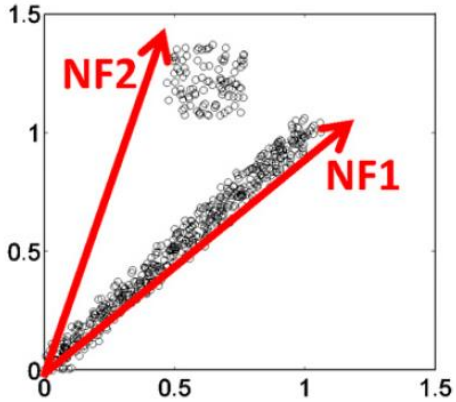
Orthogonal
Captures major variation (well, on average…)

ICA



Linear combination of independent sources.
Positive and negative.

NMF



Each point can be represented as a vector sum of NF1, NF2. Strictly positive.

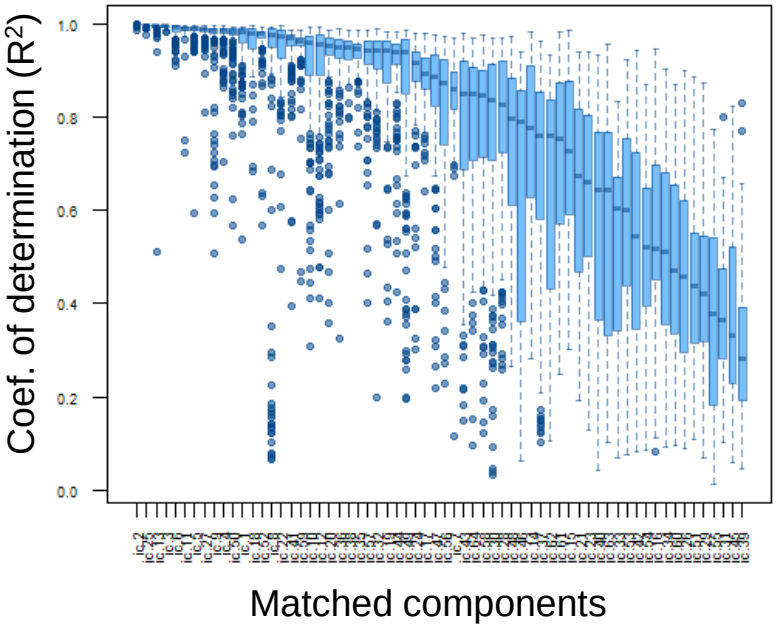
from A. Zinovyev, et al, Biochem Biophys Res Commun. 2013,18;430(3):1182-7
https://www.ncbi.nlm.nih.gov/pubmed/23261450

More details: Sompriac, Nazarov, … Zinovyev, **Int J Mol Sci, 2019**, 20(18)

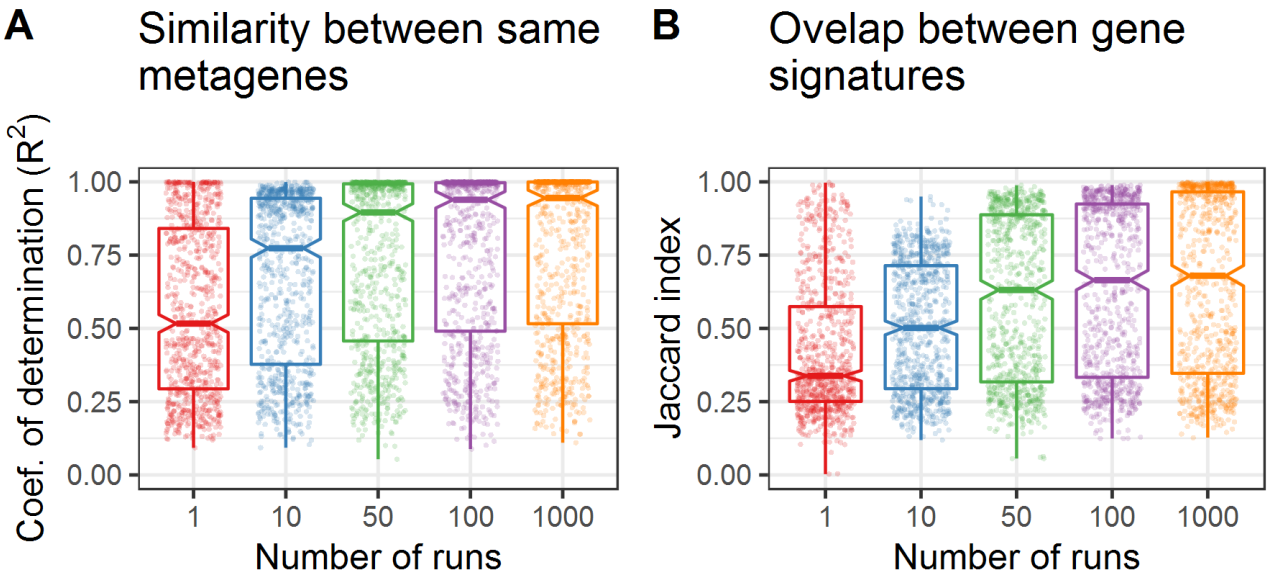
Consensus ICA -vs- ICA

Why consensus ICA is better than a simple ICA?

Reproducibility between metagenes (S) in a single run

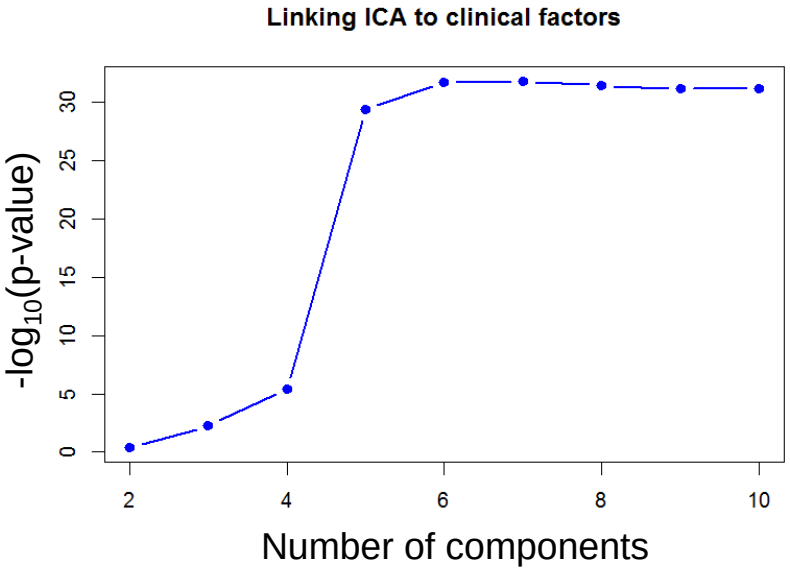


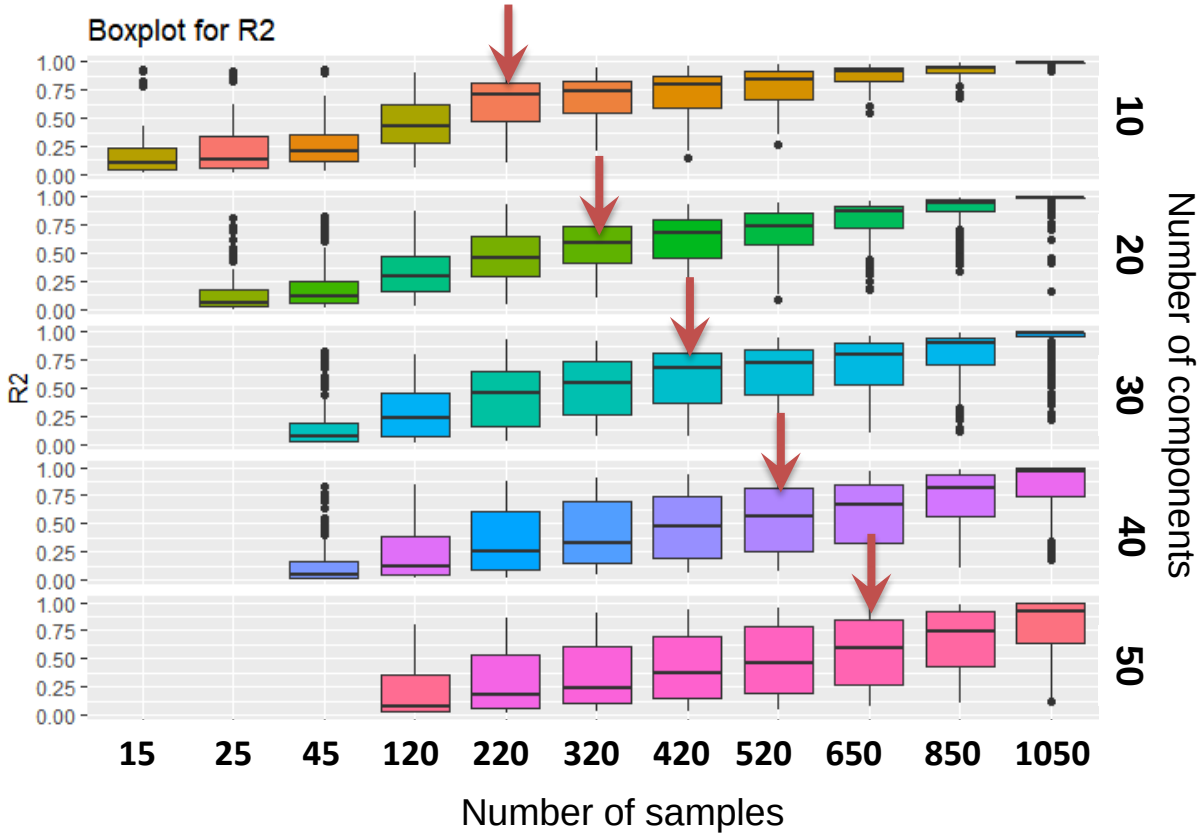
Reproducibility between metagenes (S) with many runs



Number of Components

A balance between sensitivity and reproducibility



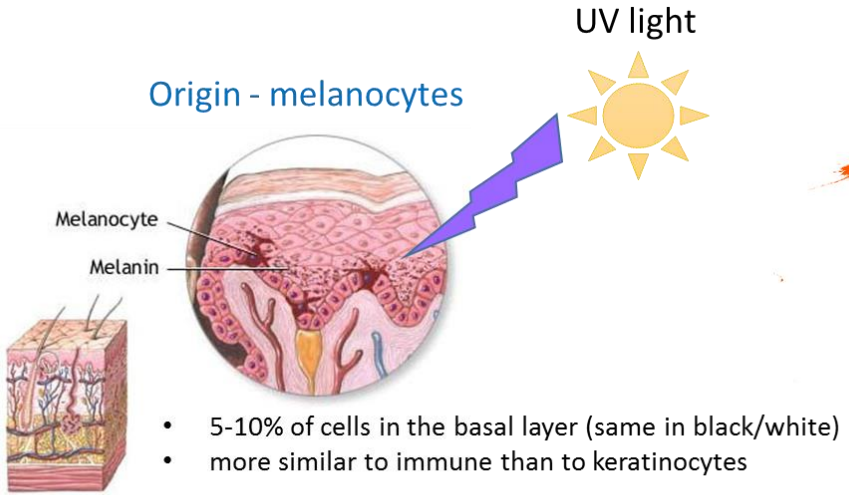


LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

MelanomICA: mRNA and miRNA of melanoma

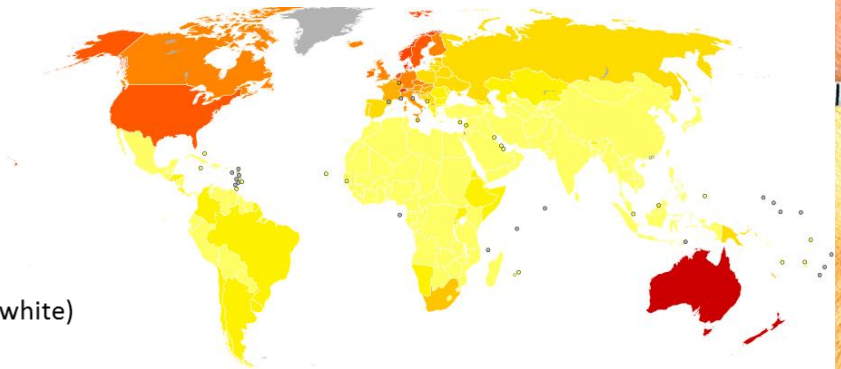
Nazarov, Wieneke-Baldacchino et al **BMC Medical Genomics, 2019**

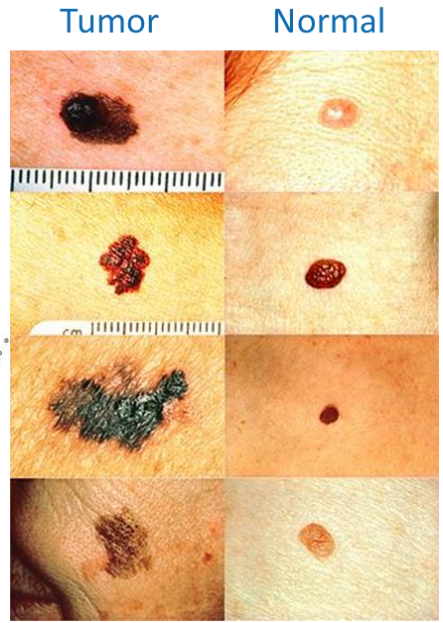
MelanomICA: ICA for Melanoma Patients



- 5-10% of cells in the basal layer (same in black/white)
- more similar to immune than to keratinocytes

Only 25% of tumors originate from nevi (moles)!





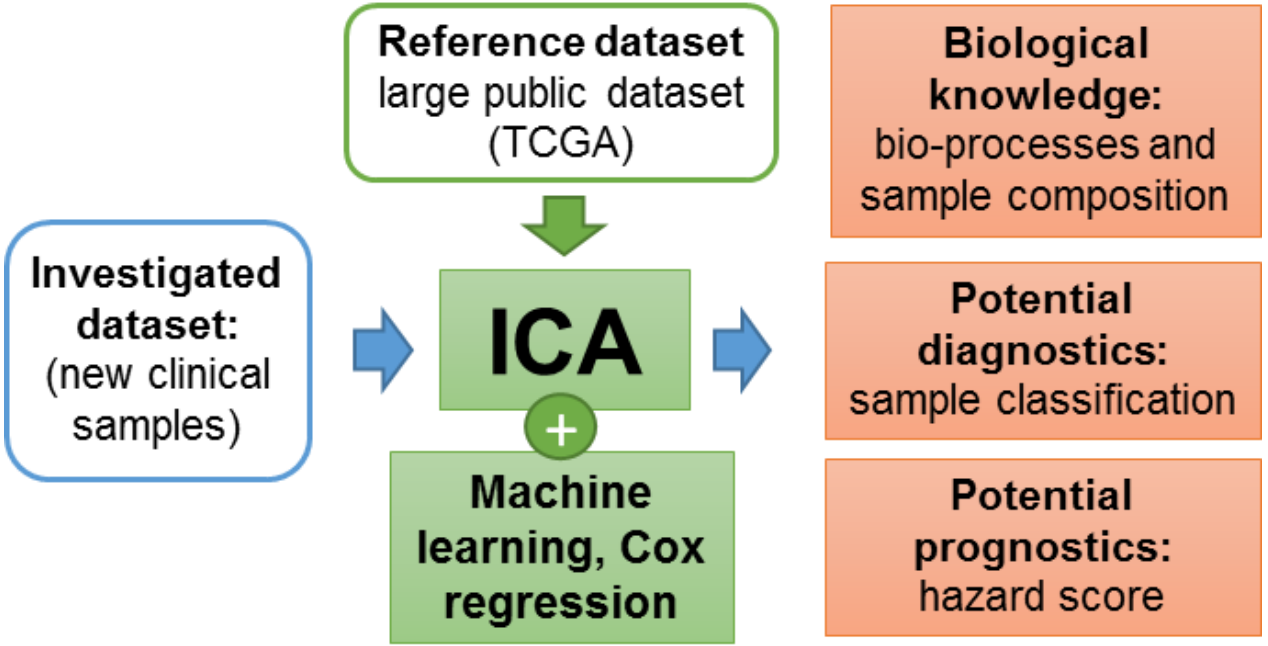
Properties:

- **Rapid** progression
- Early metastasis
- Highest mutation load
- Immune response +/-
- 5-year survival:
  - 98% when primary
  - 17% after spread

Data

- **Discovery set** – 473 primary and metastatic samples
- **Validation set** – 44 independent metastatic samples
- **Investigation set** – 3 clinical and 2 control samples

Proposed Concept



We developed

consICA

https://gitlab.com/biomodlih/consica

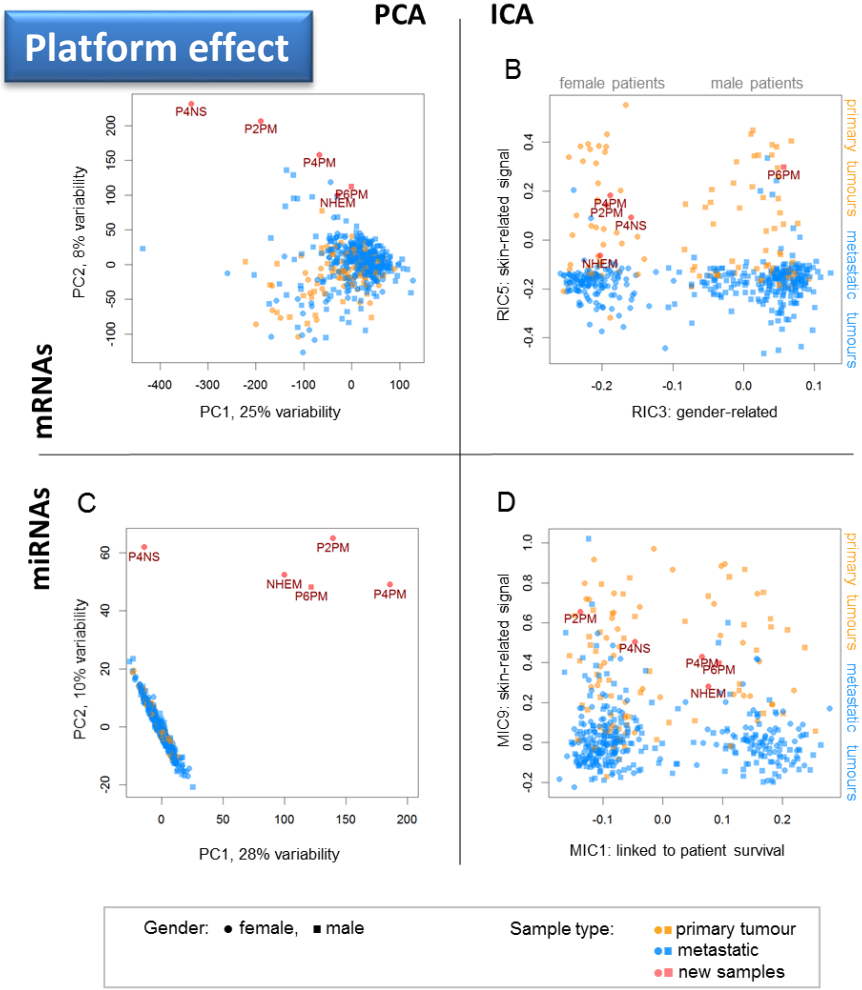
- Using R-package *fastICA*
- Consensus = mean
- Multiple runs excluding one sample, with different initial estimations
- Multiplatform
- Multicore
- Automatic report generator
- No GUI

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

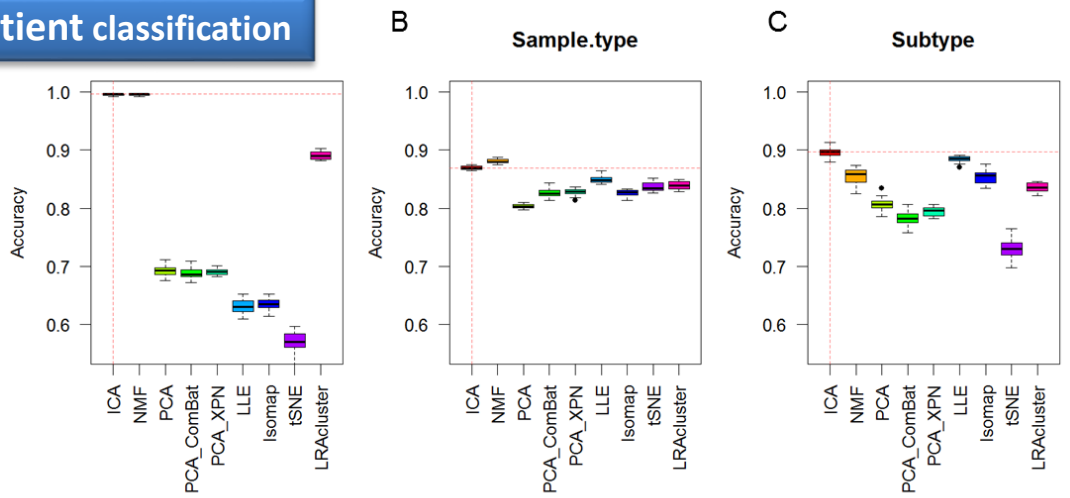
j – patient index
i – component index
R^2_i – stability of *i*-th component (from 0 to 1)
H_i – Cox' log hazard ratio calculated on **training set**
$M^*_{i,j}$ – element of centered & scaled M-matrix

MelanomICA

Platform effect



Patient classification



Gender		
Accuracy	Actual gender	
99.6%	female	male
female	177	0
male	2	293

Type		
Accuracy	Actual sample type	
78.9%	metastatic	primary
metastatic	177	54
primary	7	51

Cluster			
Accuracy	Actual cluster		
90.0%	immune	keratine	MITF-low
immune	160	9	6
keratine	9	91	6
MITF-low	1	2	47

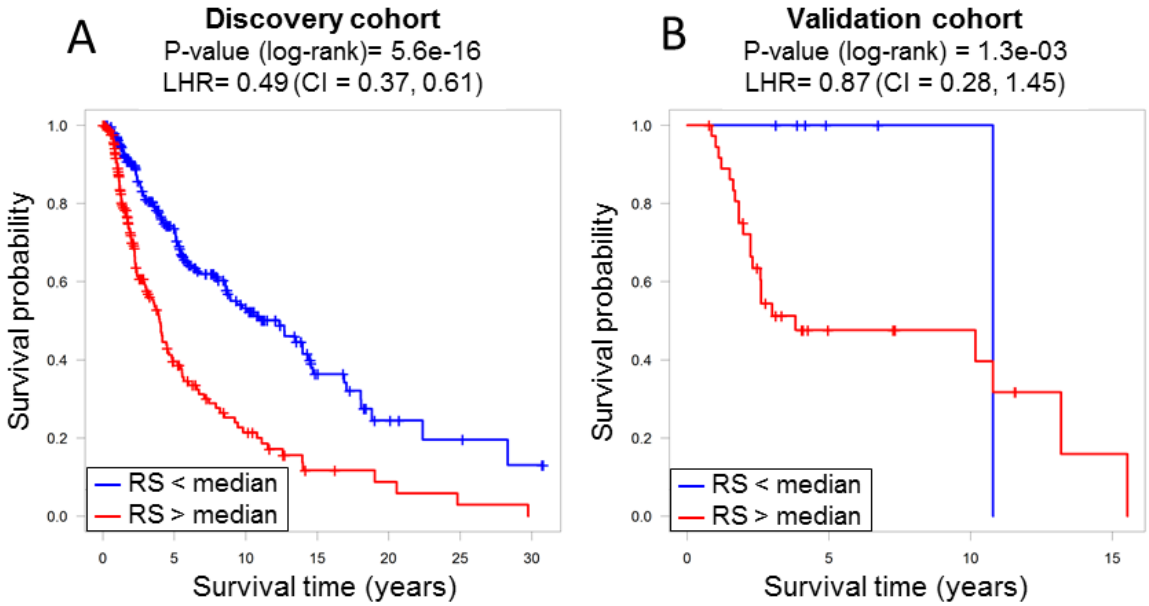
Prognostics ICA-based risk score

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

j – patient index
i – component index
R^2_i – stability of i-th component (from 0 to 1)
H_i – Cox' log hazard ratio calculated on **training set**
$M^*_{i,j}$ – element of centered & scaled M-matrix

Cross-validation

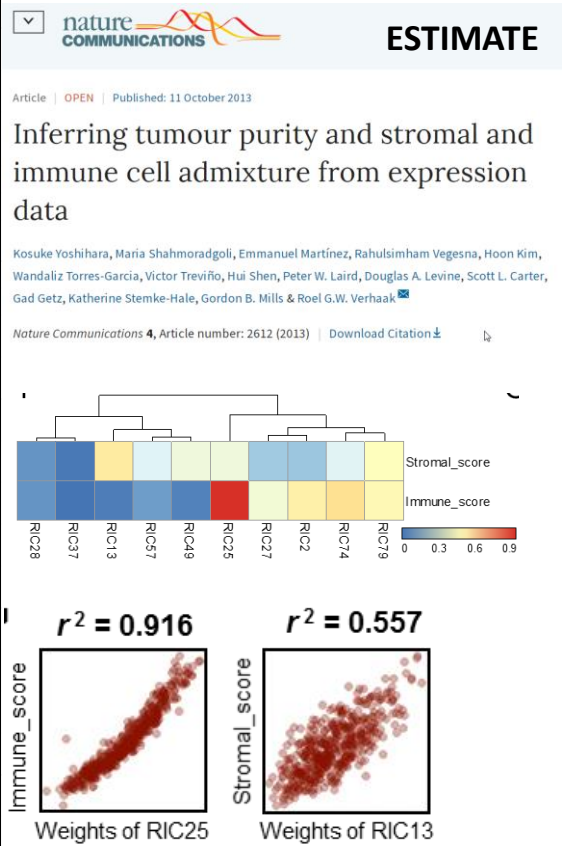
Independent cohort, different platform



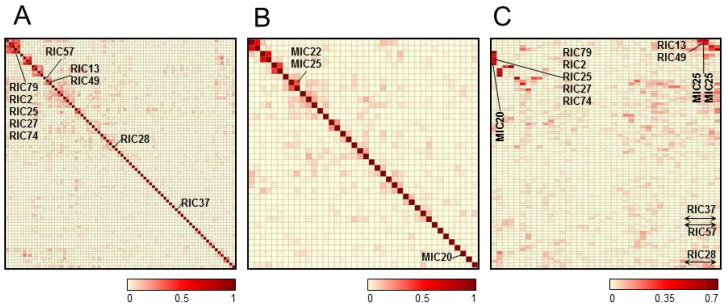
MelanomICA

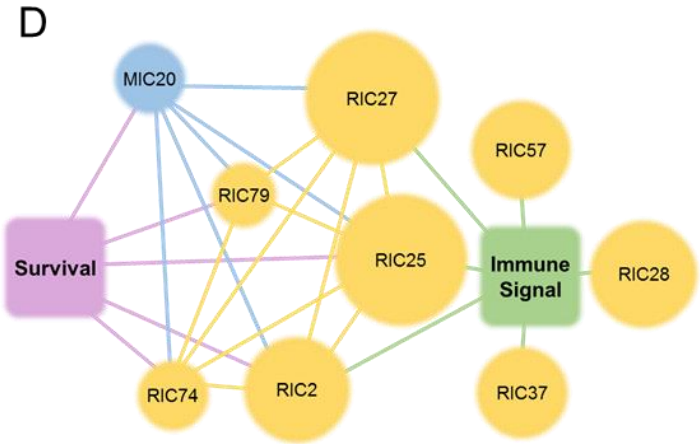
Deciphering biological processes and cell types

Cluster	Compo-nent	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	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



Data integration: mRNA + miRNA + …



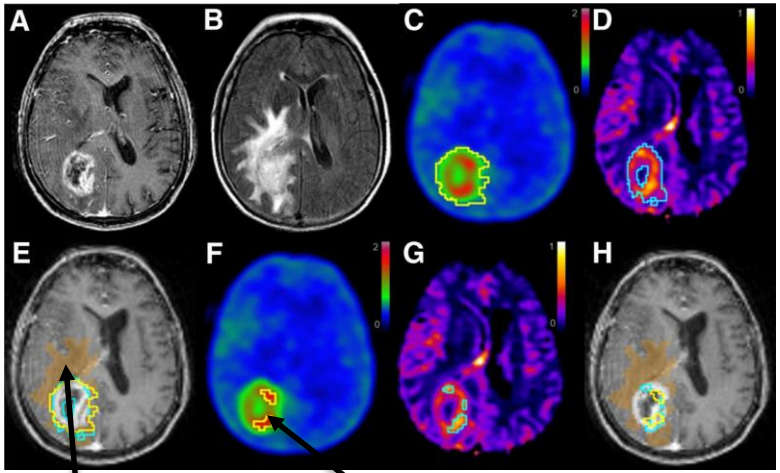


Nazarov, Wieneke-Baldacchino et al **BMC Medical Genomics, 2019**

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

DEMICS: Gliomas

DEMICS: brain tumours



Tumor cell infiltration

Core tumor (hypoxia, necrosis)

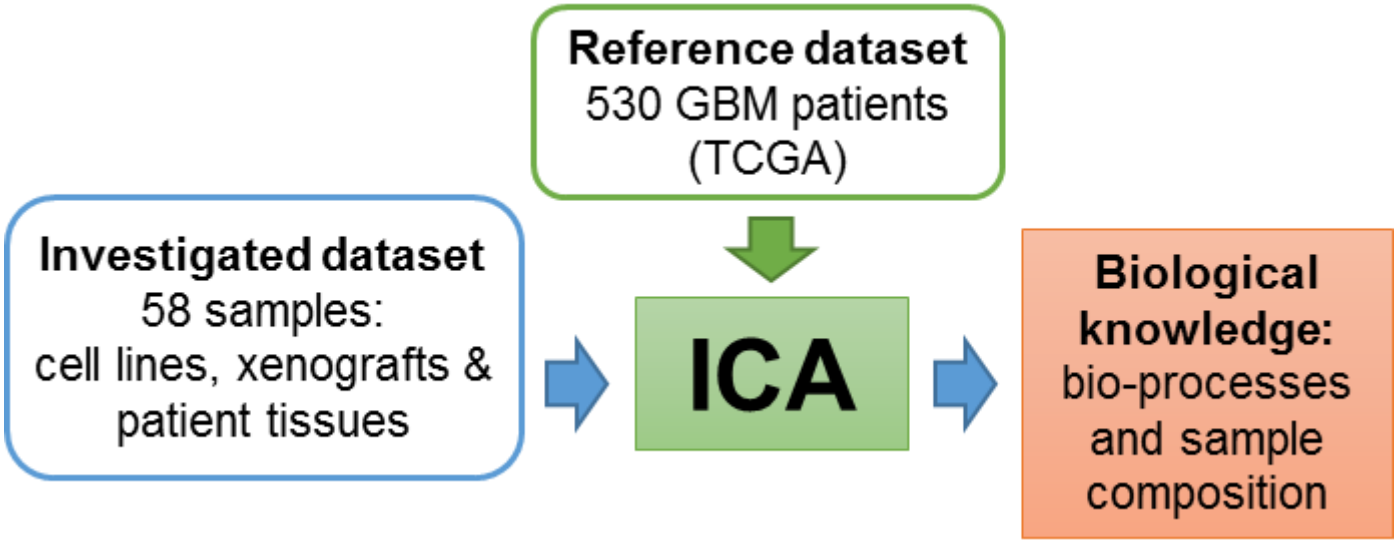
- Glioblastoma multiform (GBM) is the 4th grade glioma
- No known carcinogens
- Poor prognosis for GBM, good for low grade gliomas (LGG)
- Some GBMs originate from LGG
- GBM and LGG biopsies should share some cell types

Datasets tested:

- **TCGA-GBM**: 171 RNA-seq, 441 microarrays
- **TCGA-LGG**: 530 RNA-seq
- **CGGA** (LGG+GBM): 325 RNA-seq
- **LRNO** cell lines & PDX (A.Golebiewska, S.Fritah)

Ferreira da Ponte et al, JNM, http://jnm.snmjournals.org/content/58/10/1574.abstract

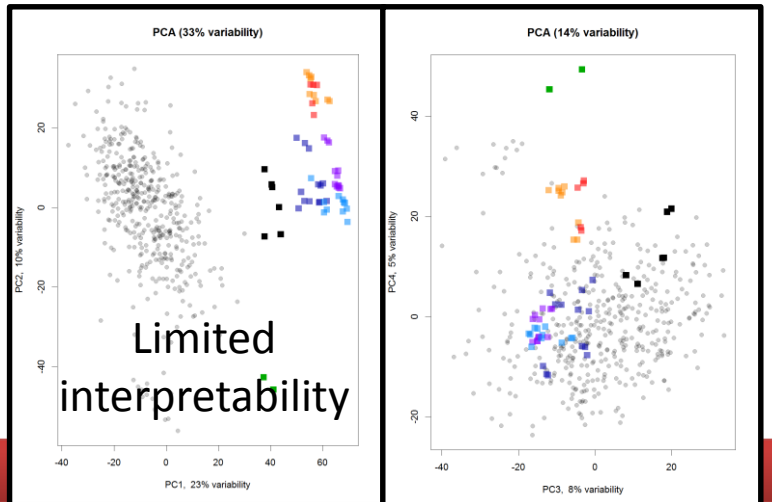
DEMICS: Validation with Cell Lines



DEMICS: Validation with Cell Lines

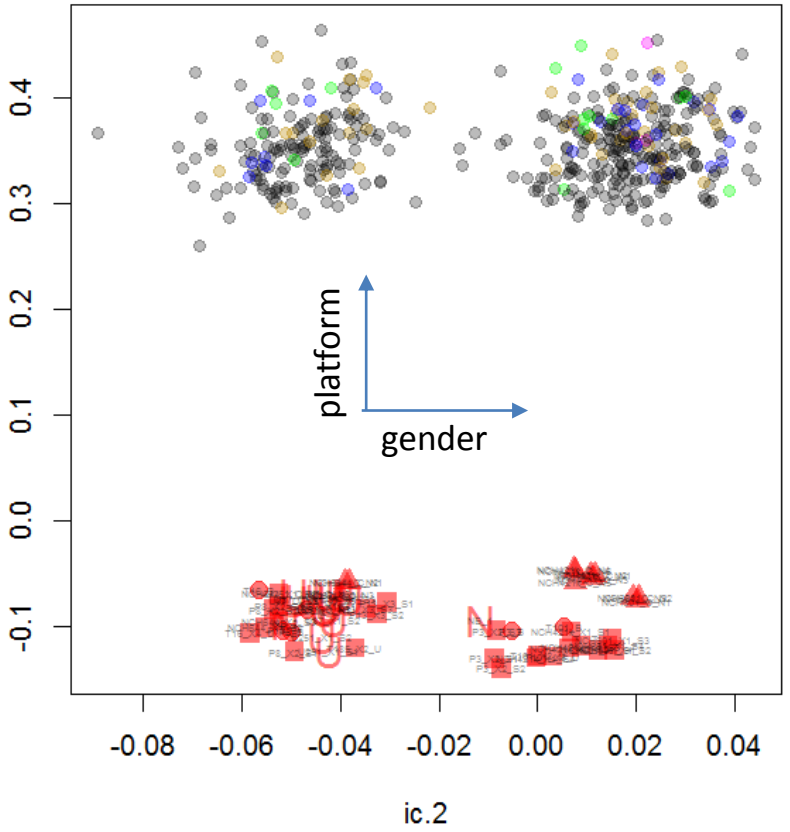
NORLUX cell line data from
Anna Golebiewska, Sabrina Fritah, Simone Niclou and other

Anna Golebiewska (microarrays):
- 2 normal tissues samples
- 6 patient biopsy samples
- 24 xenografts
- 12 patient derived cell lines
- 14 stable cell lines & their xenografts

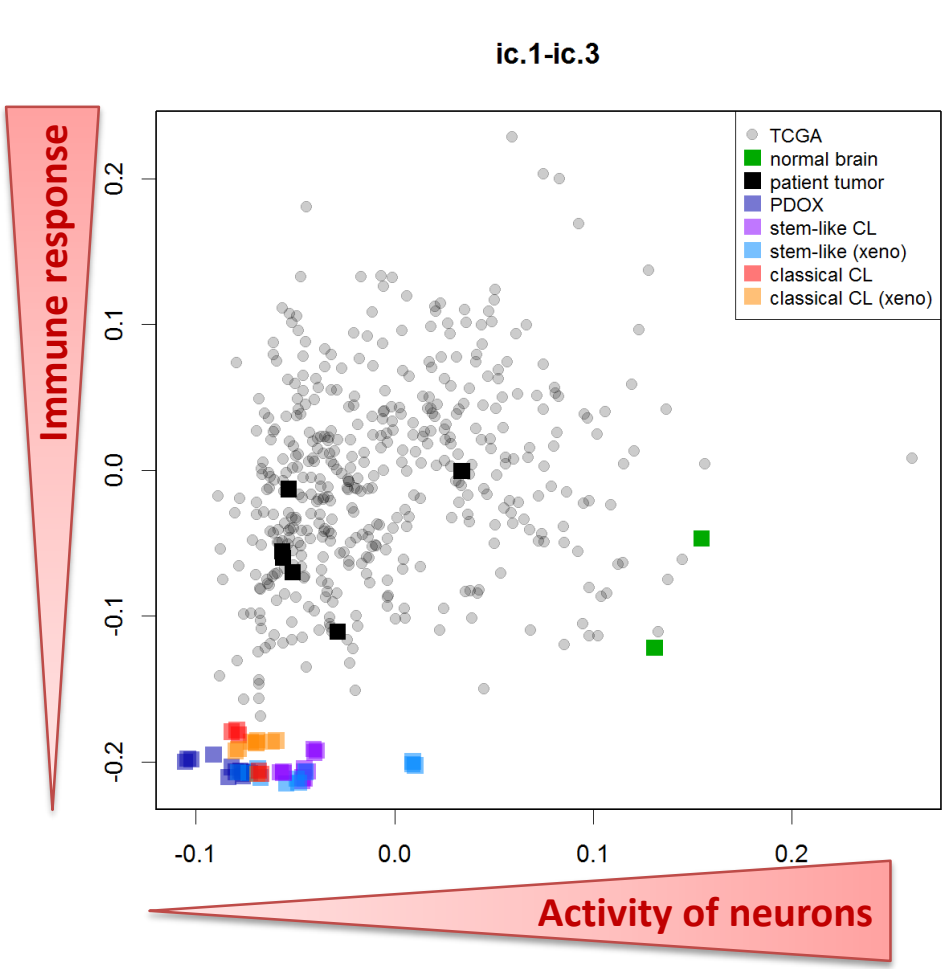


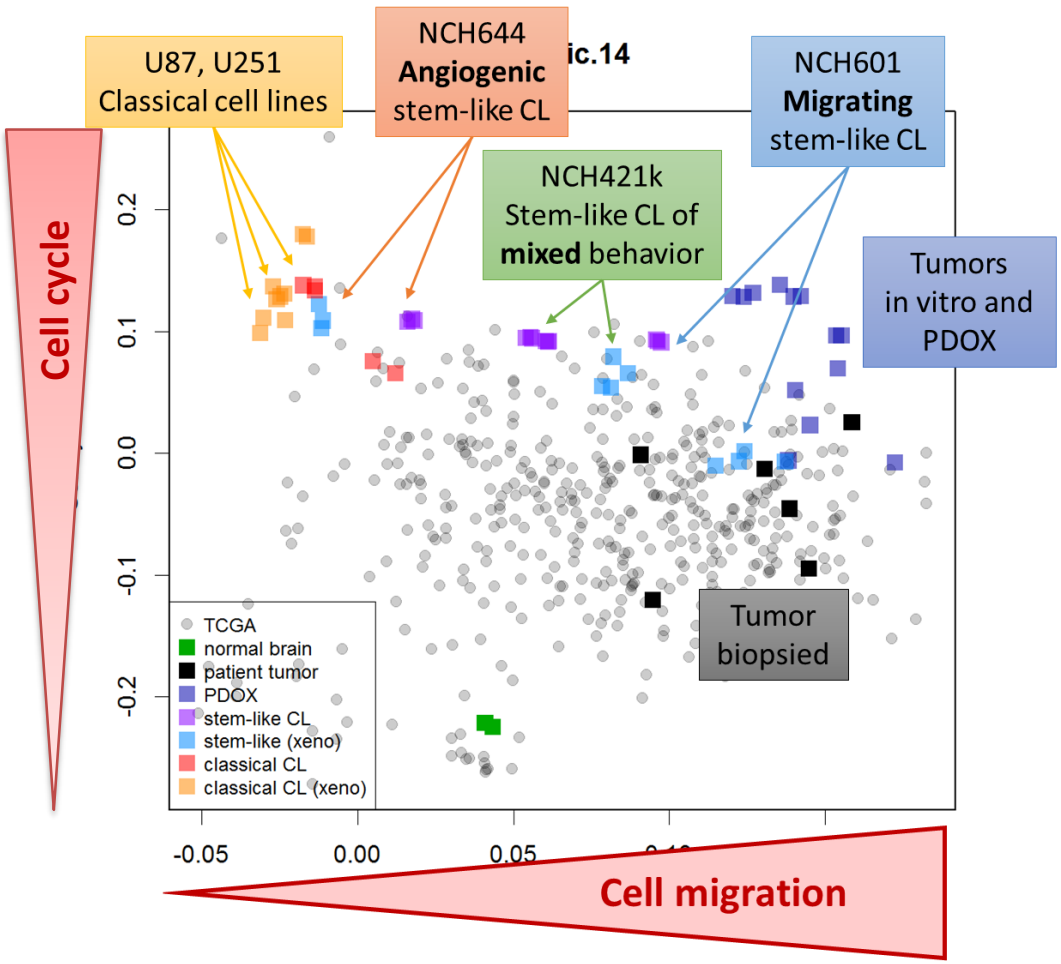
- TCGA classical
- TCGA mesenchimal
- TCGA neural
- TCGA proneural
- N normal brain
- patient biopsy
- xenograft
- patient-derived cell line
- U immortal cell line

Technical/trivial components: gender and platforms

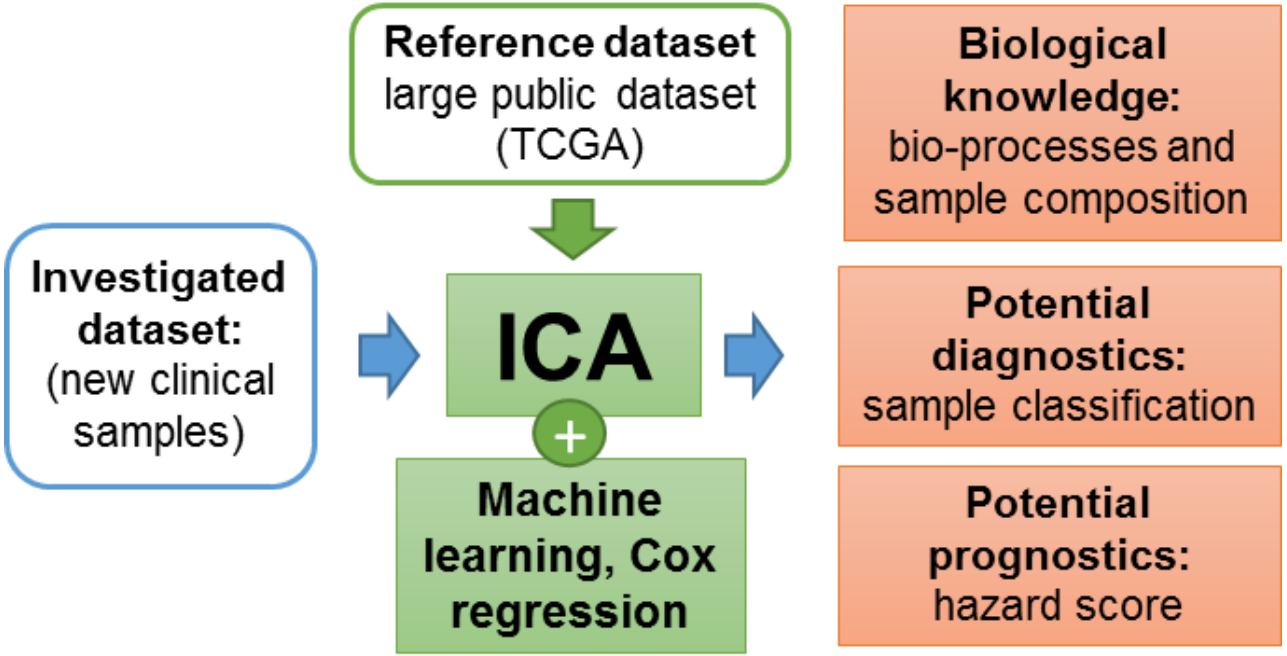


DEMICS: Validation with Cell Lines





DEMICS: Validation on Independent Ddataset

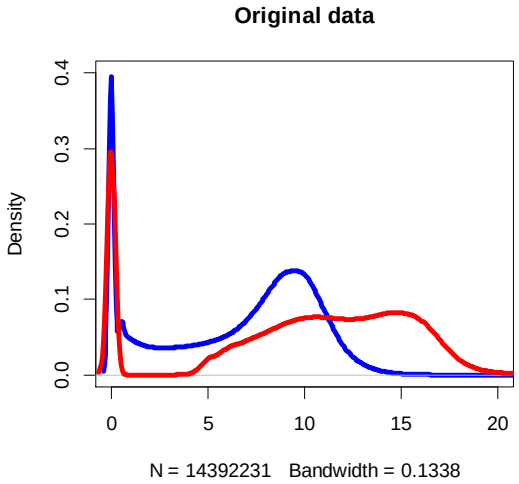


DEMICS: CGGA Cohort

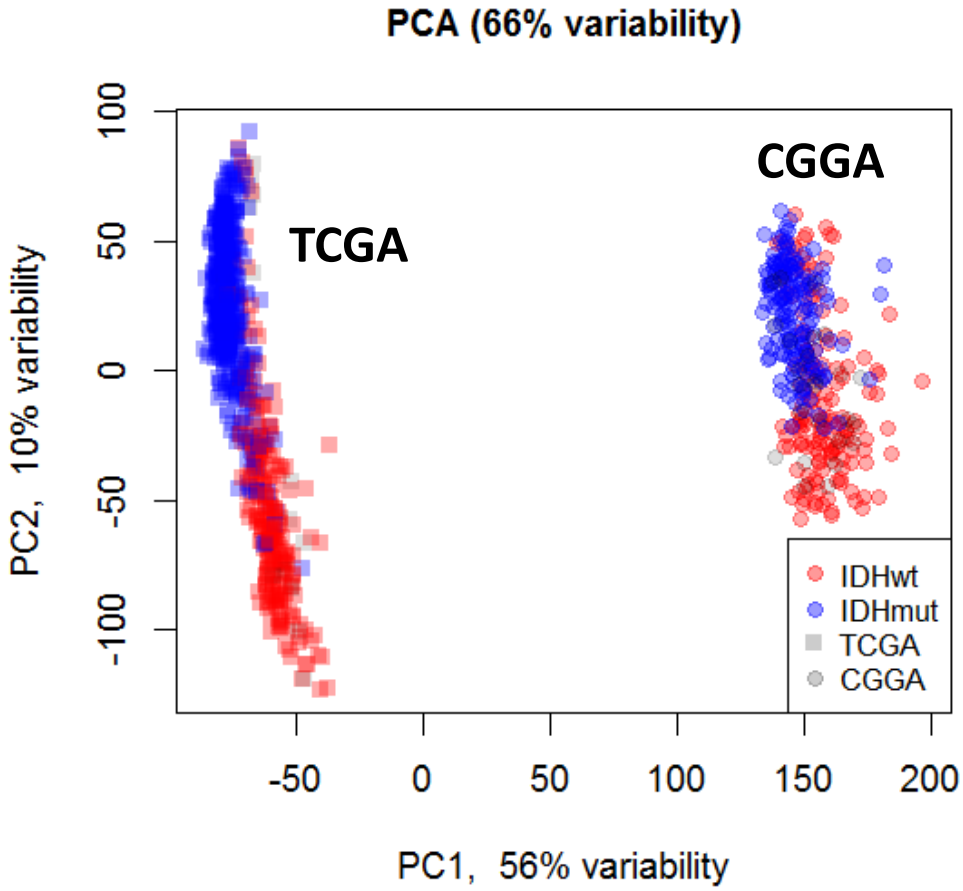
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Chinese Glioma Genome Altas (CGGA)

- 325 glioma patients
- 148 IDHwt, 152 IDHmut
- Some mutations, Verhaak's classes, histology and survival data are provided
- but **FPKM** only: gene-length correction does not help



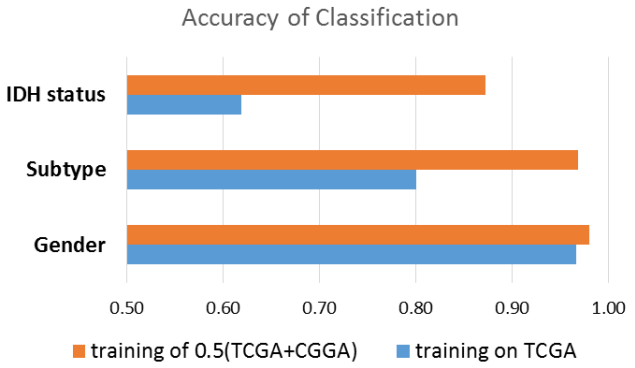
Strong differences in the data between TCGA and CGGA



Independent Validation: CGGA Cohort

Mixing CGGA & TCGA

Gender: acc = **0.97** **-> 0.98**
Subtype: acc = **0.62** **-> 0.87**
IDH1 status: acc = **0.80** **-> 0.97**



Training on TCGA and predicting CGGA

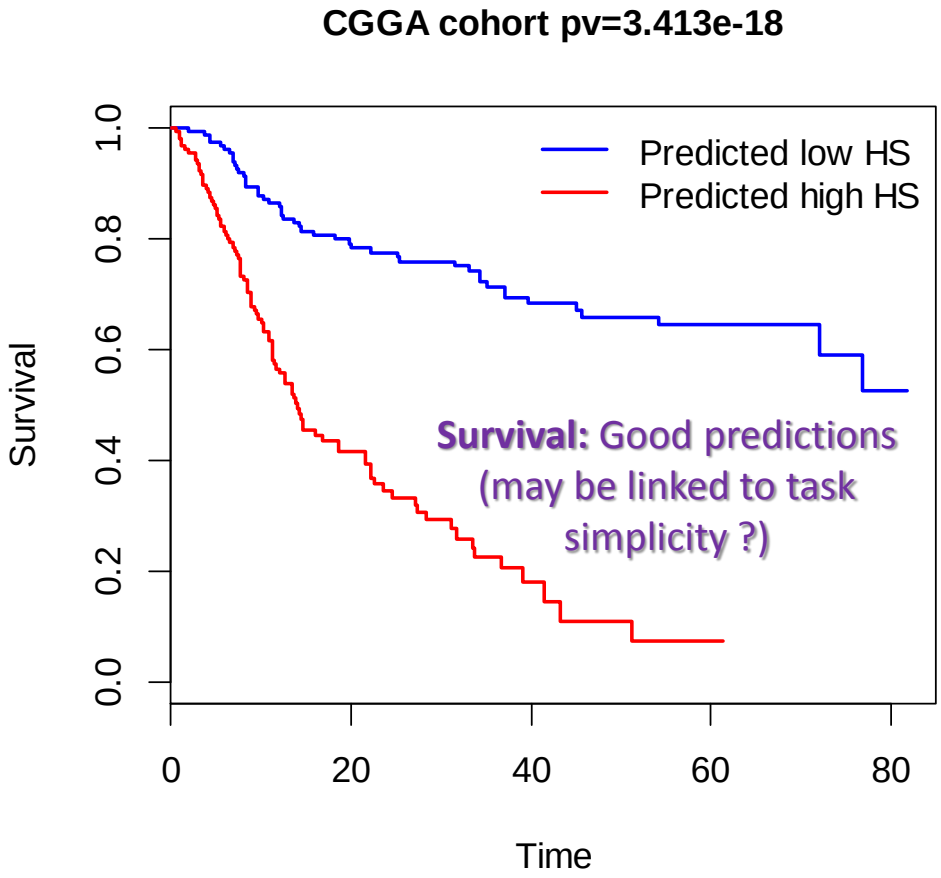
Gender	Female	Male
pred.Female	117	6
pred.Male	5	197

Subtype	CL	ME	NE	PN
pred.CL	6	0	0	0
pred.ME	66	68	21	22
pred.NE	1	0	52	5
pred.PN	1	0	8	75

IDH1	Mutant	WT
pred.Mutant	101	9
pred.WT	51	139

Classification:
Much lower accuracy then expected, when datasets are strictly separated.

Adding a part of CGGA data to training set significantly improved classification



Conclusions

- ICA **corrects technical biases** in the data
- ICA captures **biologically-relevant signals** of cell populations and biological processes
- ICA provides **good features for patient classification**
- ICA-based features can be **united in a risk score**, predicting patient survival
- We hope that ICA decomposition can be used for better **data integration** (under investigation)

Acknowledgements

Luxembourg Institute of Health

Quantitative Biology Unit

Gunnar Dittmar
(Head of the Unit)

Computational Biomedicine

Francisco Azuaje (PI)

Katharina Baum (Postdoc)

Arnaud Muller (Bioinf.)

Tony Kaoma (Bioinf.)

Yue Zhang (PhD student)

Multiomics Data Science

Petr Nazarov (PI)

Sang Yoon Kim (Bioinf.)

Thomas Eveno (MSc student)

Maryna Chepeleva (MSc student)

LSRU, UniLu
Dr. Anke Wieneke
Dr. Stephanie KREIS

NORLUX Neuro-Oncology, LIH
Dr. Anna GOLEBIEWSKA
Prof. Simone NICLOU

Institute Curie, France
Dr. Andrei ZINOVYEV

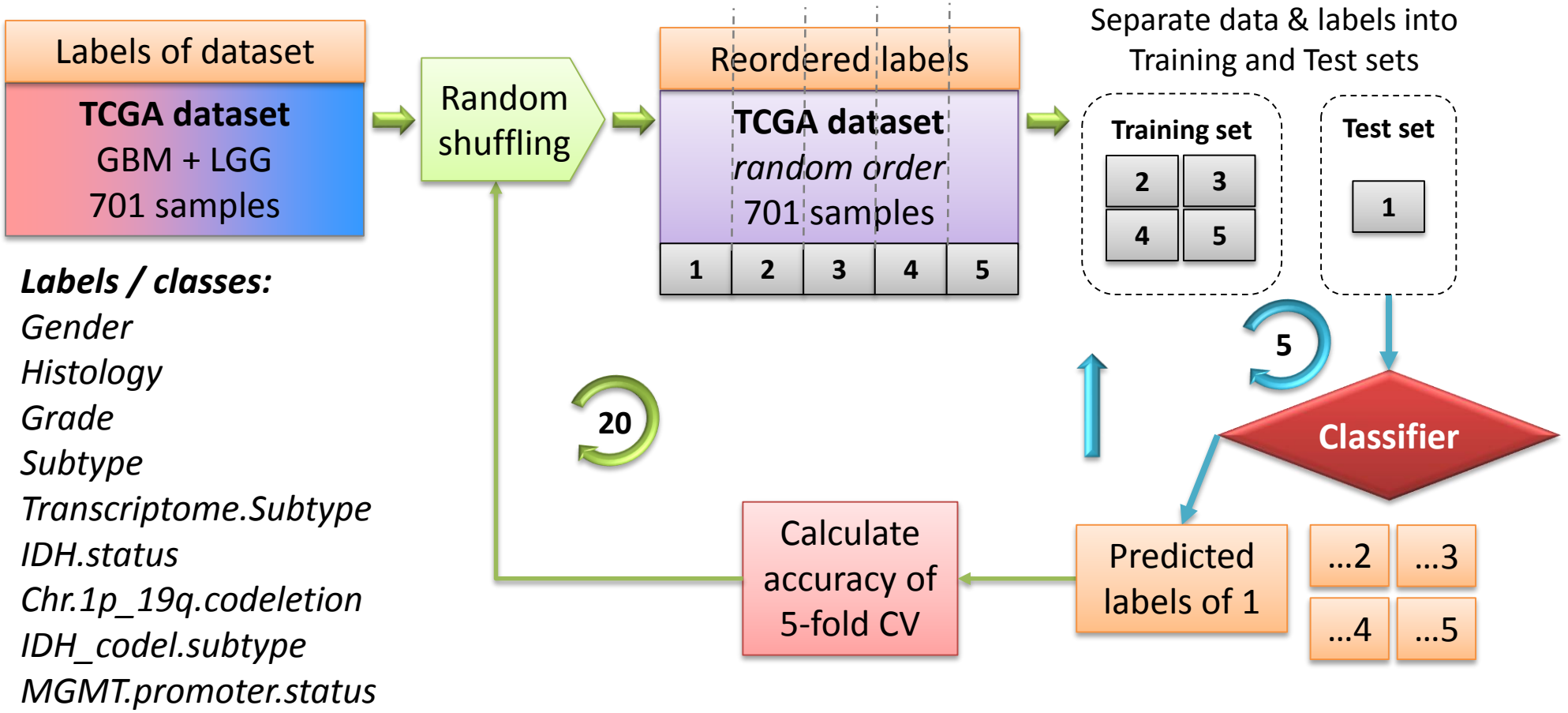
Fonds National de la Recherche Luxembourg

Supported by Luxembourg National Research Fund
C17/BM/11664971/**DEMICS**

University of Bergen, Norway
Prof. Inge JONASSEN *(Mentor)*

Supplementary Slides

Classification: Cross-validation Scheme



Labels / classes:

Gender
Histology
Grade
Subtype
Transcriptome.Subtype
IDH.status
Chr.1p_19q.codeletion
IDH_codel.subtype
MGMT.promoter.status

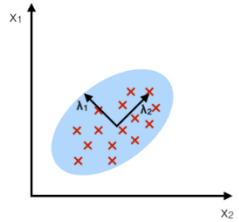
Classification methods

Classifiers can perform differently on different features. What is the most optimal for ICA?

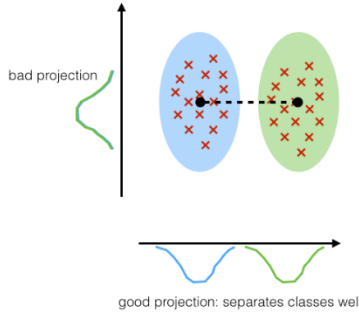
Tested methods

- **LR** - logistic regression
- **PLSDA** - partial least square discriminant analysis
- **LDA** - linear discriminant analysis
- **MDA** - mixture discriminant analysis
- **RDA** - regularized discriminant analysis
- **FDA** - flexible discriminant analysis
- **kSVM** - support vector machine (kernlab)
- **eSVM** - support vector machine (e1071)
- **KNN** - k-nearest neighbors
- **RF** - random forest (randomForest)
- **RRF** - regularized random forest
- **NB** - naive Bayes classifier
- **GBM** - gradient boosting model
- **ABM** - AdaBoost model

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation

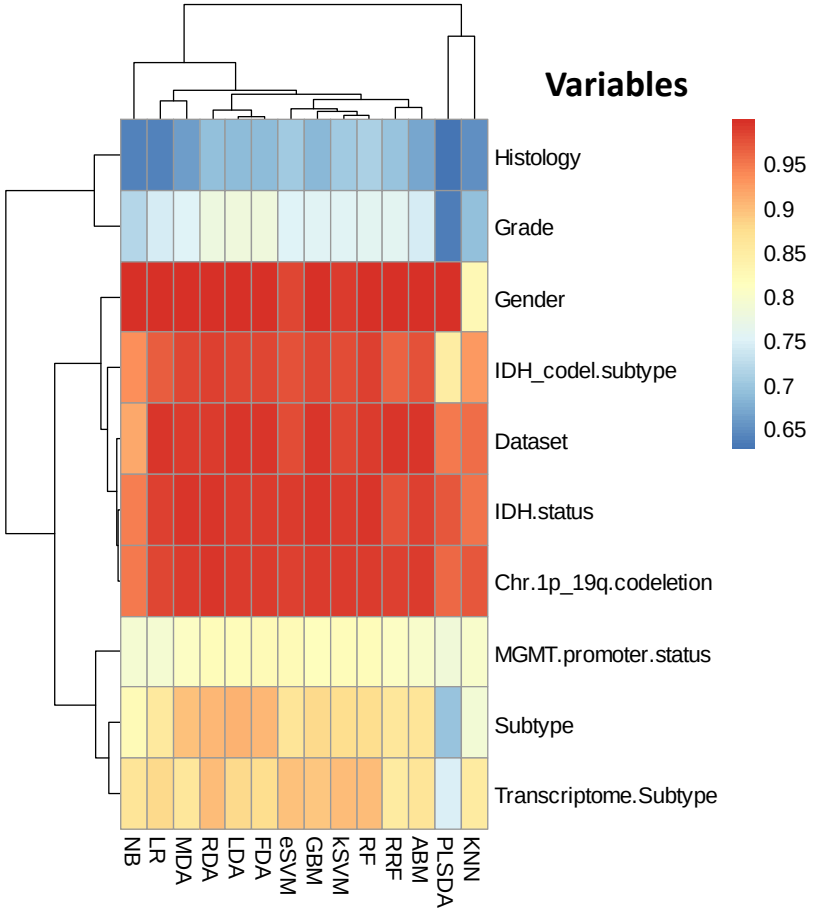


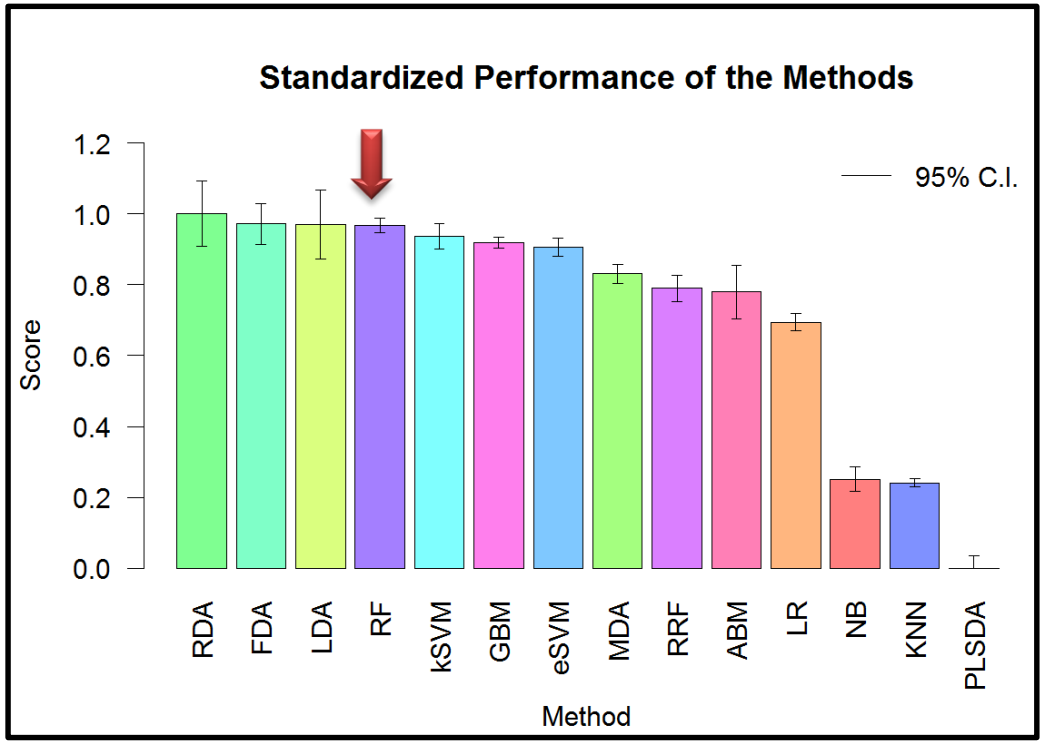
Sebastian Raschka

LDA is a generalization of Fisher's linear discriminant, a method used to find a **linear combination of features** that separates several classes of objects. The resulting combination may be used as a linear classifier or for dimensionality reduction before later classification.

Classification Method Selection

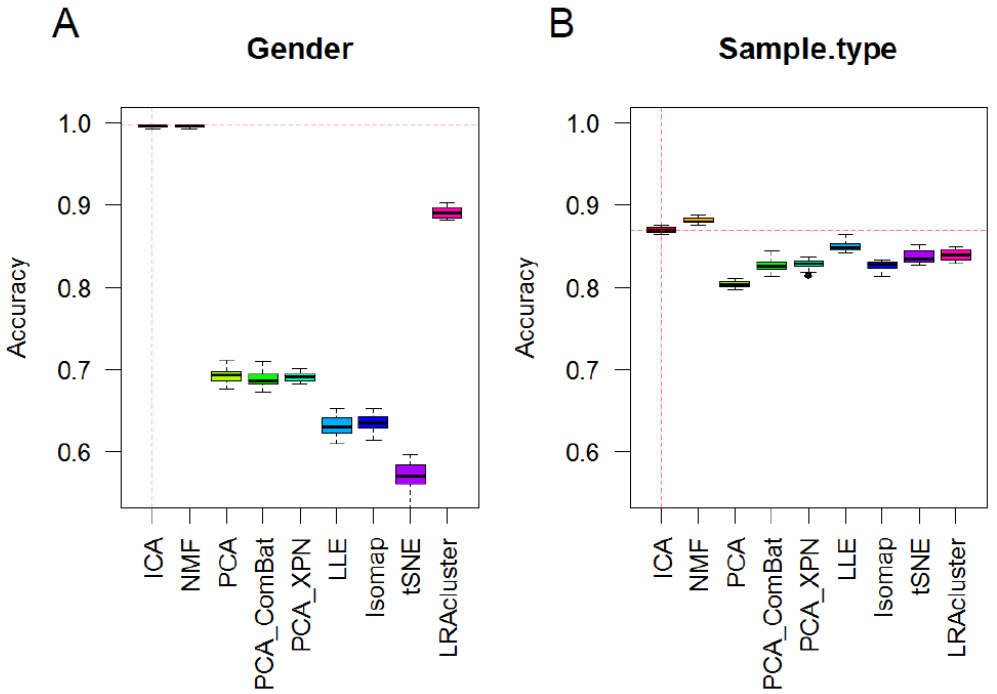
Mean Accuracy (5-fold cross-validation)

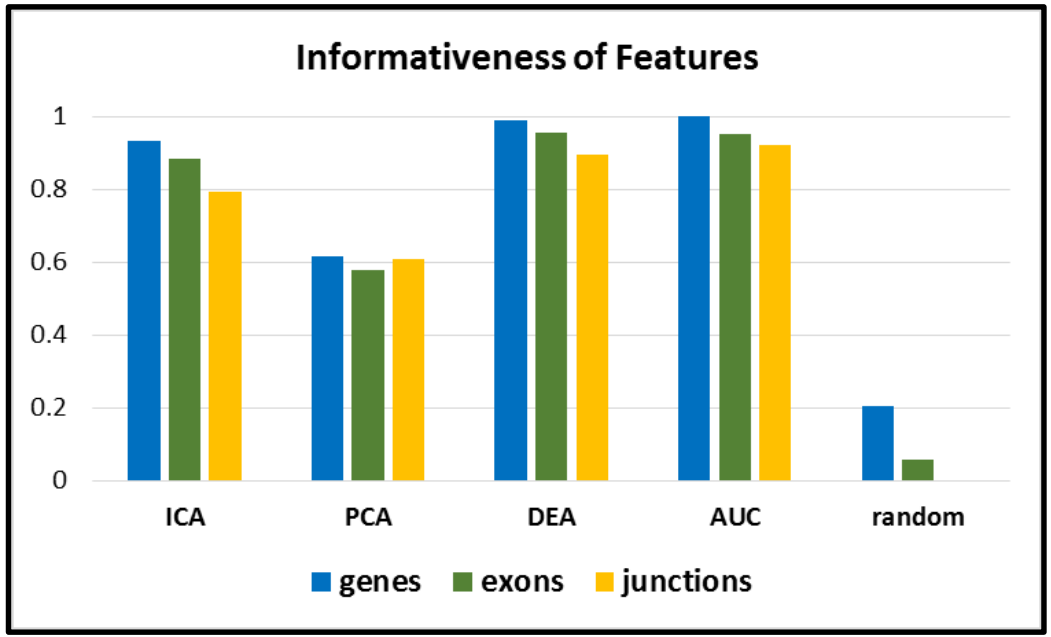




(constant effect of variables was removed)

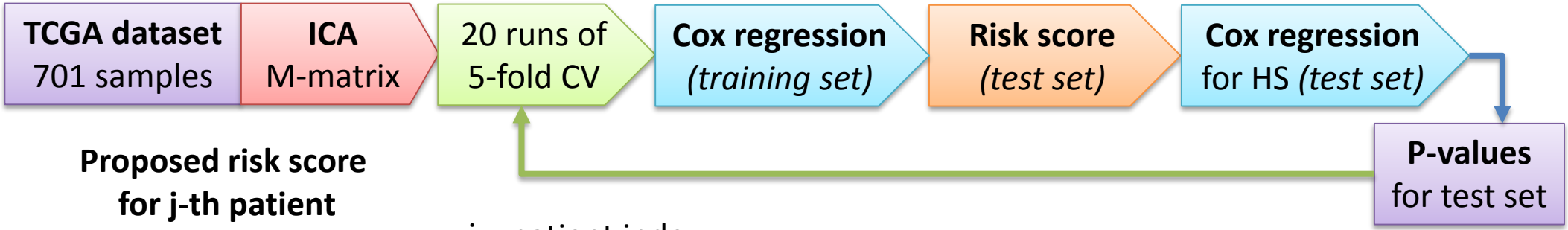
Other Feature Selection Methods / Data Sources





Nazarov et al https://www.biorxiv.org/content/10.1101/395145v1

Predicting Survival: Approach for Cross-validation



Proposed risk score for j-th patient

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

j – patient index
i – component index
R^2_i – stability of i-th component (from 0 to 1)
H_i – Cox' log hazard ratio calculated on **training set**
$M^*_{i,j}$ – element of centered & scaled M-matrix