Fonds National de la Recherche Luxembourg

LUXEMBOURG INSTITUTE OF HEALTH

Multiomics Data Science Research Group

Deconvolution and machine learning reveal oncogenic processes and characterize cancer patients

Petr Nazarov

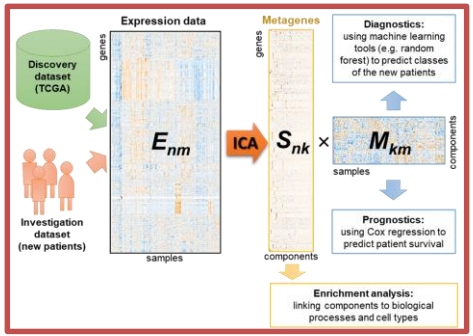
petr.nazarov@lih.lu

LIH PI meeting

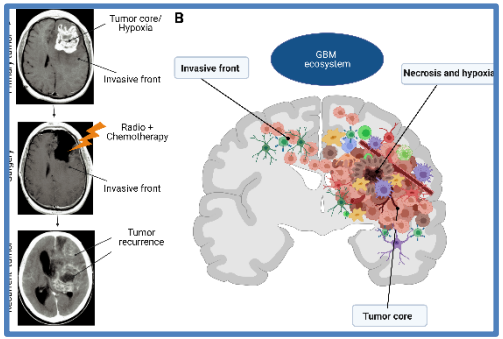
Outline

FNR CORE Project and Project Proposals:

DEMICS → DIOMEDES + MEDEA

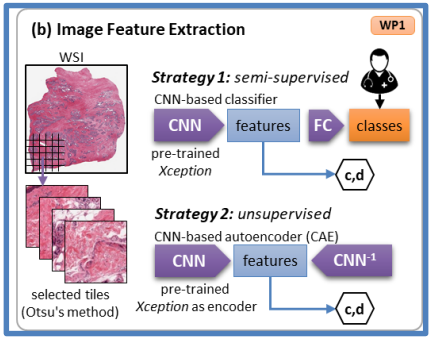


Work started: **2015** (with Laurent)
Financed: **2018-2019** (with Francisco)

submitted **2021**

Anna – PI

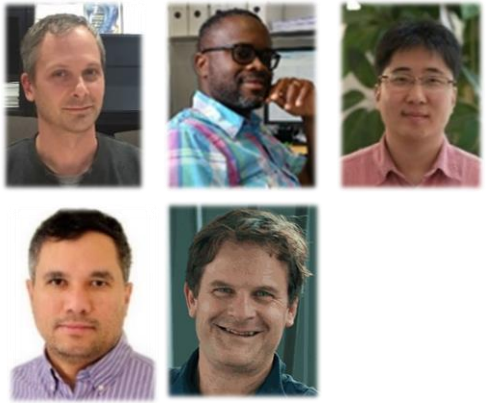


submitted **2021**

Michel – Co-PI

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Colleagues / co-authors



DEMICS

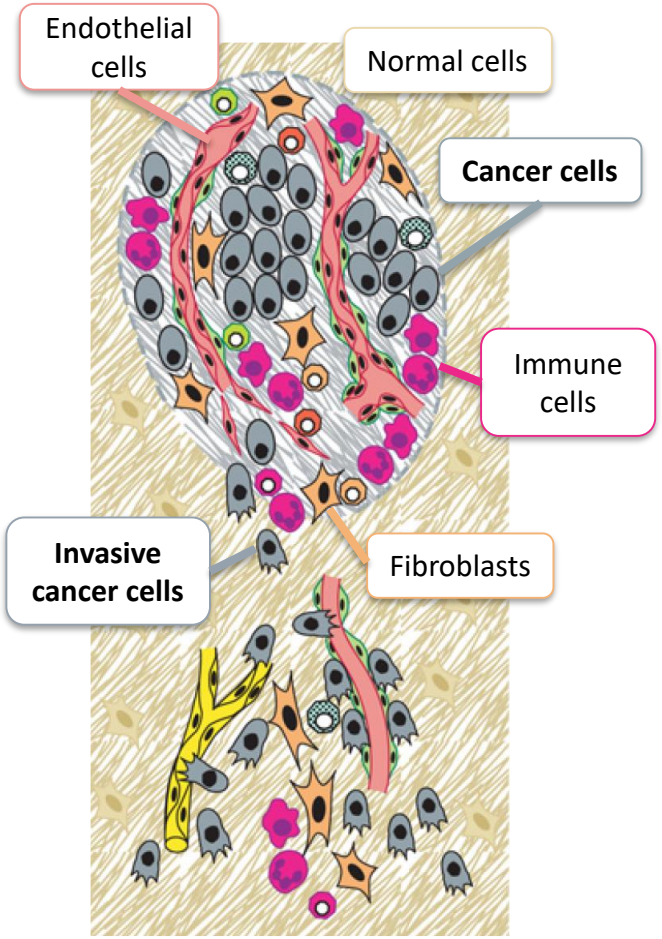
Colleagues / collaborators

Interns / students

Fonds National de la Recherche Luxembourg

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

Methods



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

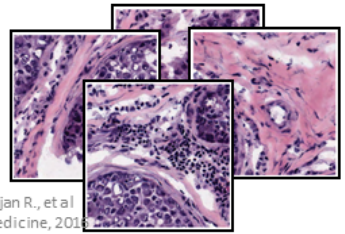
Levels of Heterogeneity

Experimental data (RNA-seq)

Technical heterogeneity

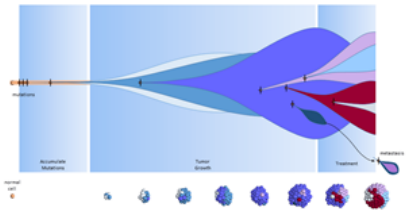
http://abrf.masonlab.net/

Sample composition



Natrajan R., et al
PLoS Medicine, 201

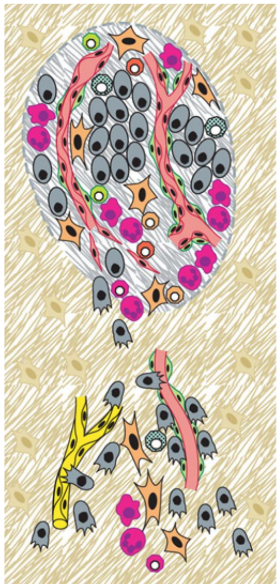
Clonal evolution in cancer



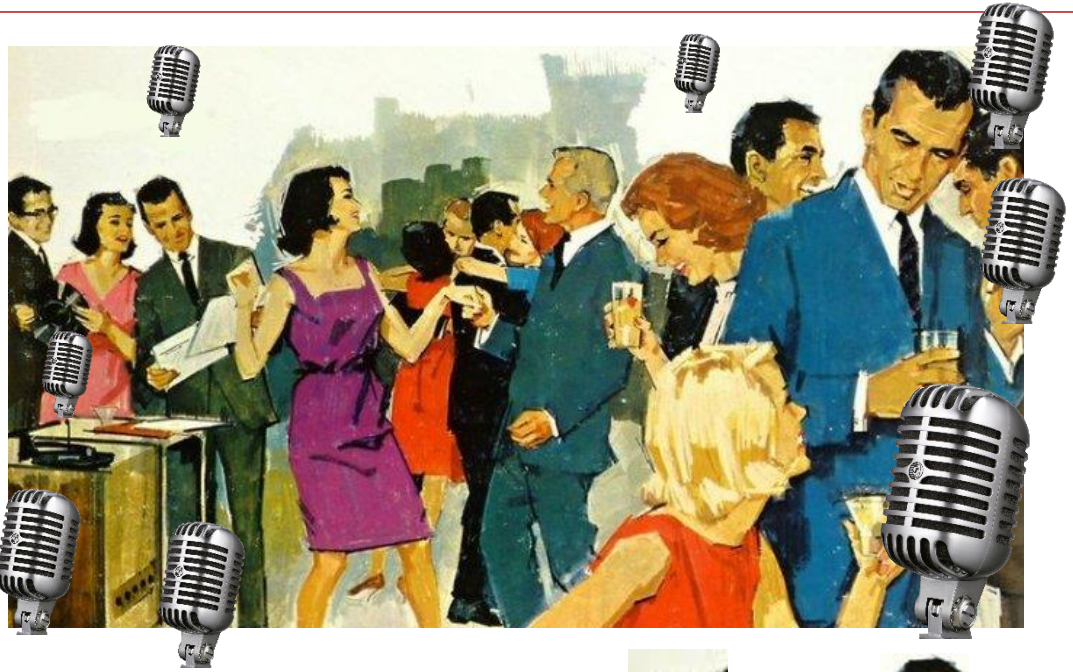
https://www.pancanology.com/clonal-evolution-model/

- Technical heterogeneity
- Native heterogeneity of biological tissues
- Inter/intra tumor heterogeneity due to clonal evolution

Methods



~



Cocktail Party Problem

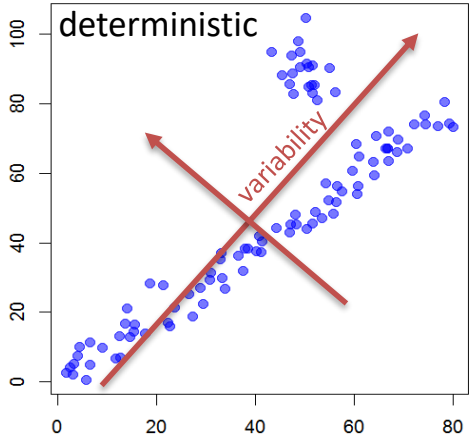
Mixed transcriptomes are similar mixed audio signals in a "cocktail party problem"

ICA: extracts statistically independent signals

Independent
Component
Analysis

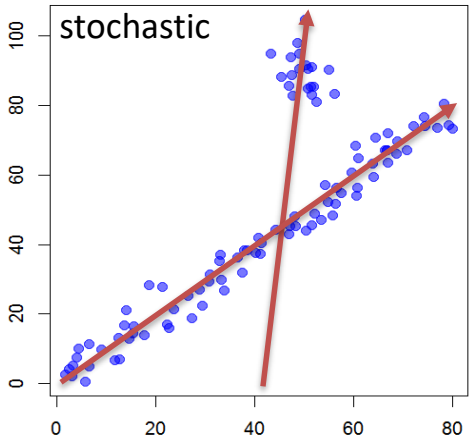
Methods

PCA



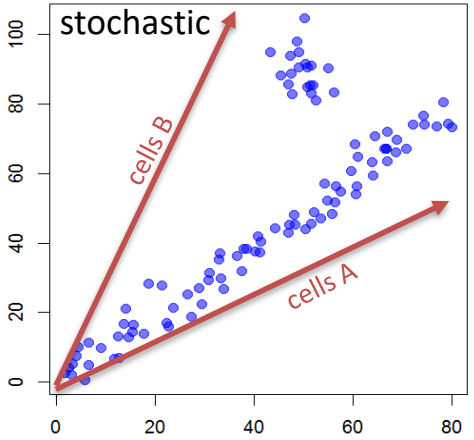
+ deterministic & fast
+ any number of samples
+ unsupervised
– often biological factors are presented by a sum of several components
– positive and negative values

ICA



+ **correlates with biology**
+ **unsupervised (agnostic)**
+ **quite stable**
– stochastic
– needs a lot of samples
– positive and negative values

NMF

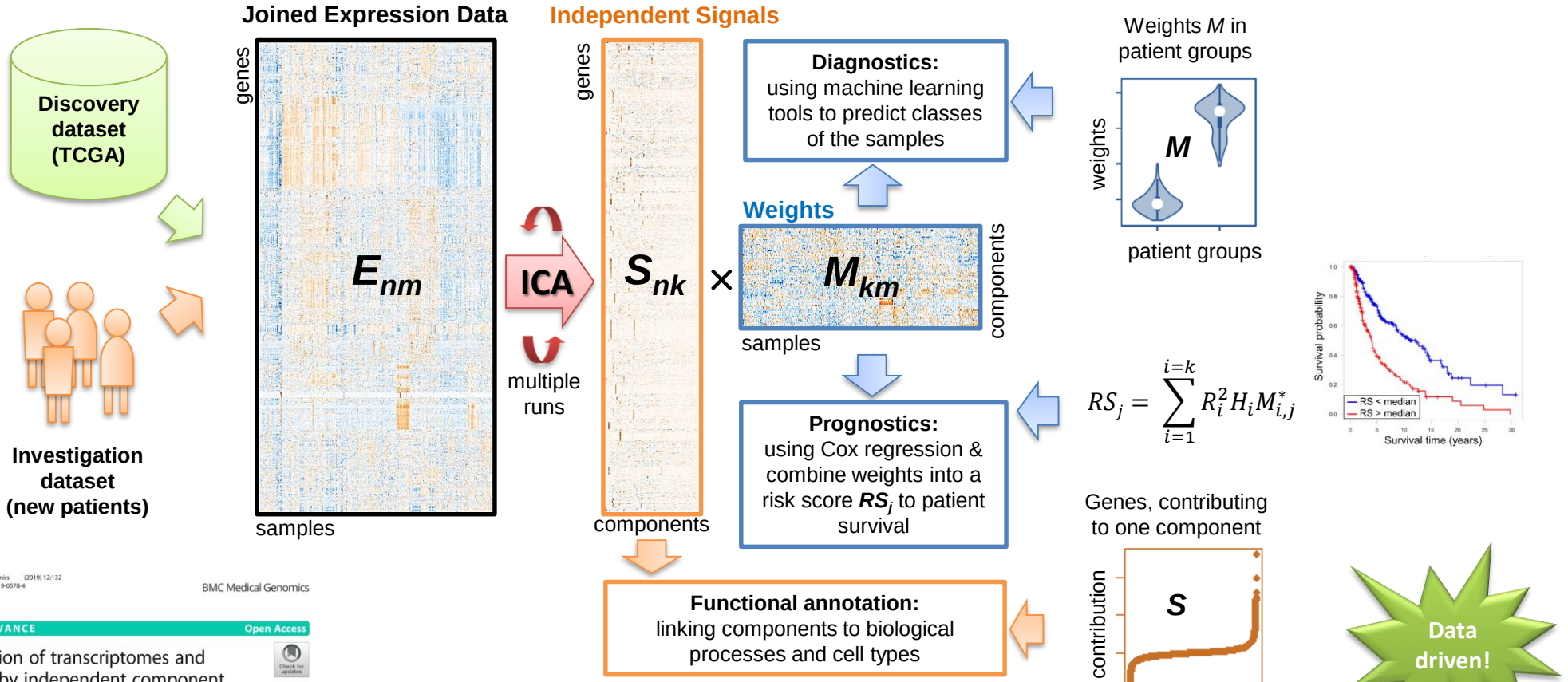


+ semi-unsupervised
+ easy to interpret
– stochastic
– unstable

Sompairac, Nazarov, el al, Int J Mol Sci, 2019 (link)
Cantini el al, Bioinformatics, 2019 (link)

Consensus Independent Component Analysis (*consICA*)

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE



Nazarov et al. BMC Medical Genomics (2019) 12:132
https://doi.org/10.1186/s12920-019-0578-4

BMC Medical Genomics

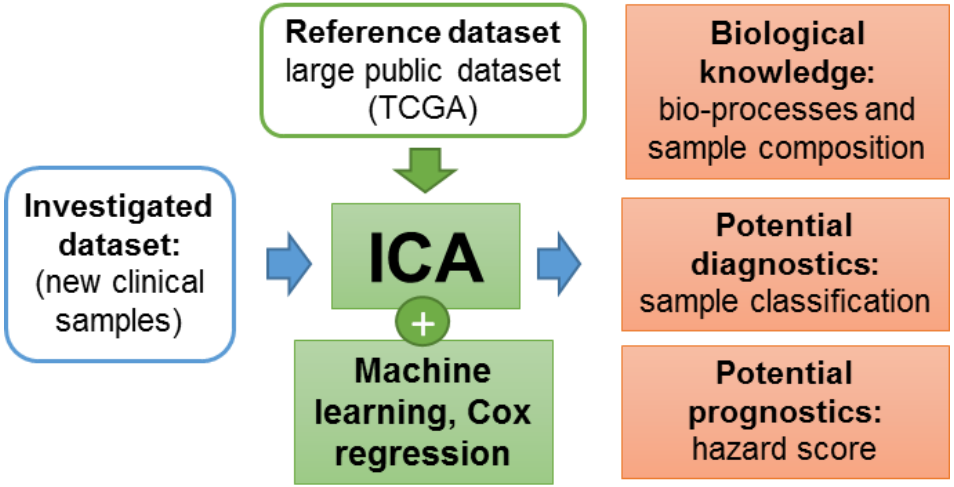
TECHNICAL ADVANCE Open Access

Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients

Petr V. Nazarov[1*], Anke K. Wienecke-Baldacchino[2,3†], Andrei Zinovyev[4,5], Urszula Czerwińska[4,5,6], Arnaud Muller[1], Dorothée Nashan[1], Gunnar Dittmar[1], Francisco Azuaje[1] and Stephanie Kreis[2]

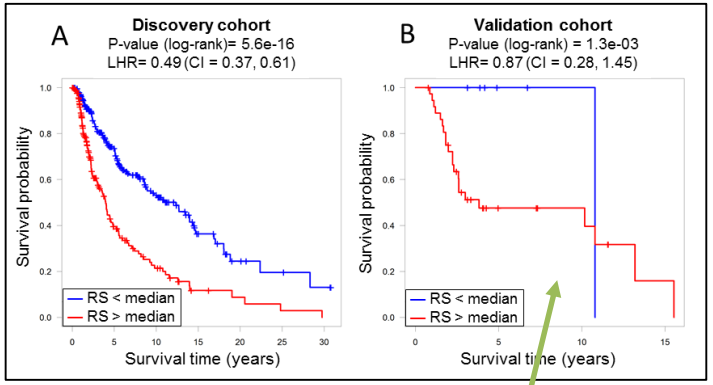
consICA: **Nazarov et al BMC Medical Genomics, 2019** (link)
ICA review: Sompairac, Nazarov, el al Int J Mol Sci, 2019 (link)

Melanoma



Cross-validation on the same cohort

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



Independent cohort, different platform

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

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

Collaboration with Dr S. Kreis (UL)
Nazarov et al, BMC Medical Genomics, 2019 (link)

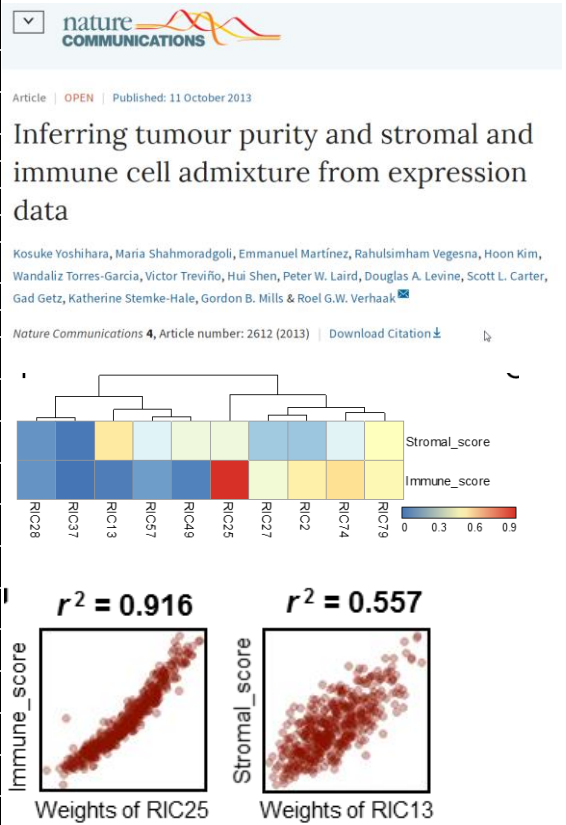
- In addition to diagnostics and prognostics, ICA allowed ranking patients based on activity of biological processes: cell cycle, signals of leukocytes, etc.

Melanoma

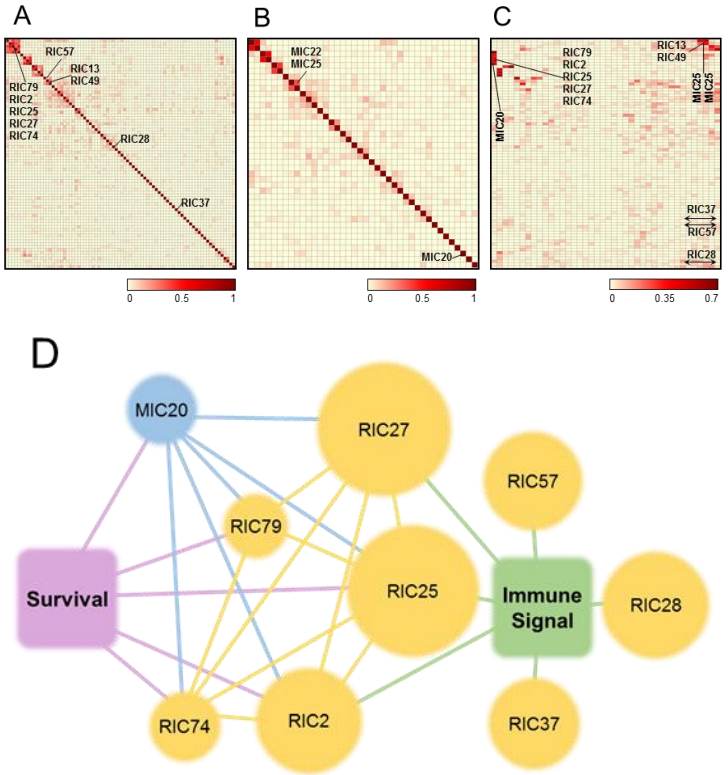
Deciphering biological processes and cell types

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

ESTIMATE



Data integration: mRNA + miRNA + …

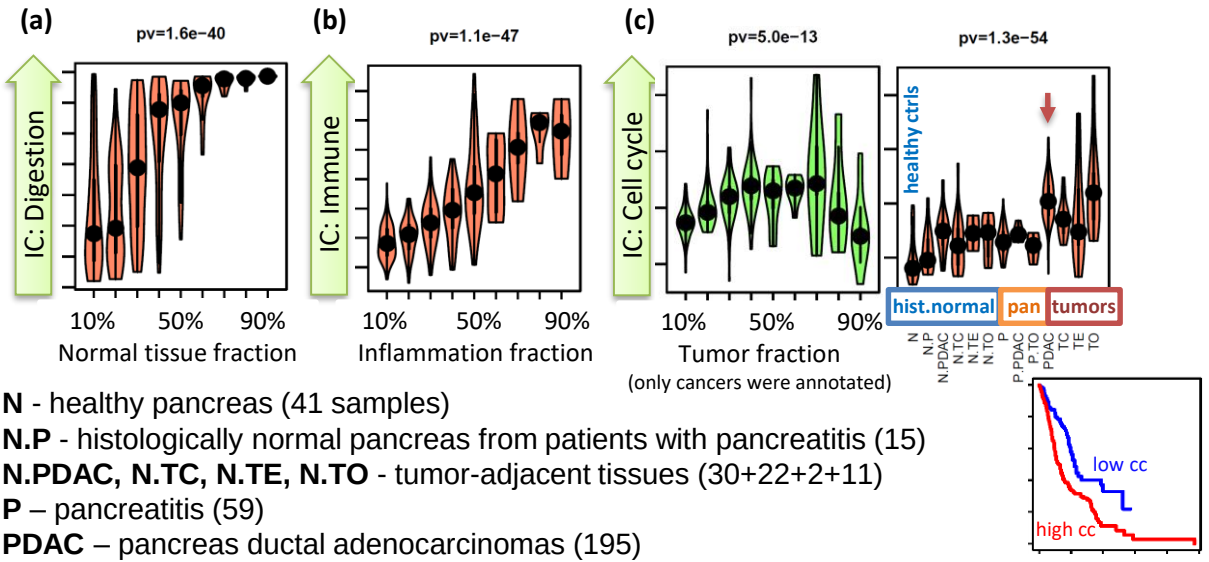


← New samples are mapped to the space defined by reference data.

Pancreatic Tumors → MEDEA

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Pancreatic cancers: ICA results of mRNA expression data from DKFZ cohort



N - healthy pancreas (41 samples)
N.P - histologically normal pancreas from patients with pancreatitis (15)
N.PDAC, N.TC, N.TE, N.TO - tumor-adjacent tissues (30+22+2+11)
P – pancreatitis (59)
PDAC – pancreas ductal adenocarcinomas (195)
TC – cystic tumors (24)
TE – neuroendocrine tumors (18)
TO – other tumors (31)

Components identified by ICA were annotated by biological functions (GO) and linked to survival using Cox regression.

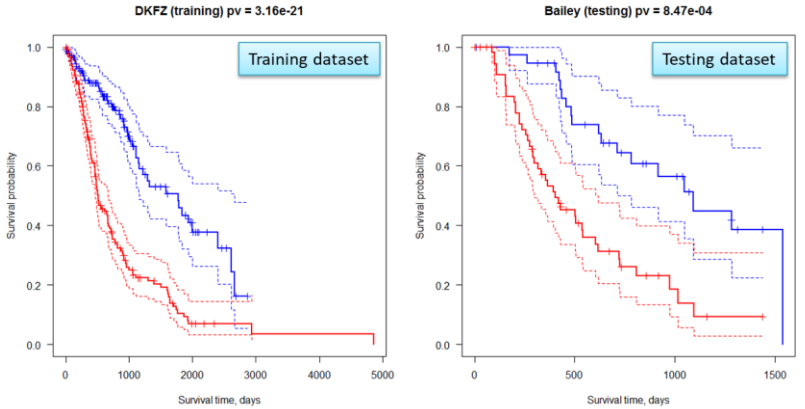
Increased risk:
- **keratinization**
- **cell cycle**
- response to hypoxia
- neoangiogenesis
- activation of ERK-signaling

No effect:
- immune response
- gender
- axon development

Reduced risk:
- secretion activity (normal)
- digestion
- antigen binding

Unlike in melanoma, no direct link was found between immune response and survival: perhaps due to a dual / antinomic effect.

Prognostic markers between 2 cohorts



Acc: 0.83	N	N.PDAC	P	PDAC
pred.N	32	2.6	1.8	2
pred.N.PDAC	0.6	1.7	2	1.6
pred.P	4.7	17.3	51.8	5.4
pred.PDAC	3.7	8.4	3.4	186

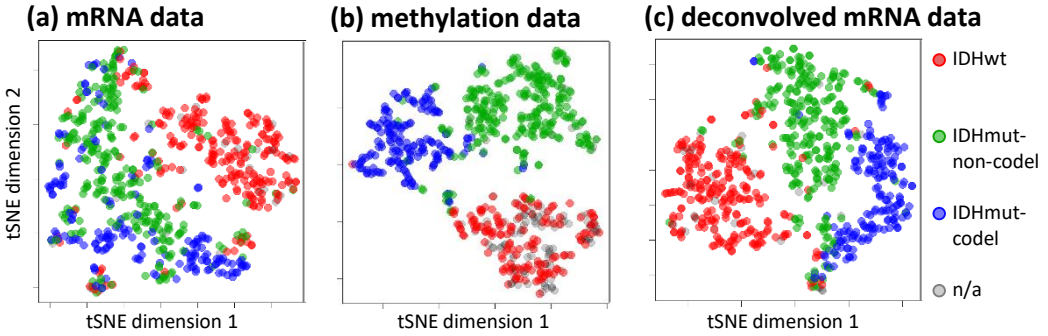
Aliaksandra Kakoichankava
(MD student)

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

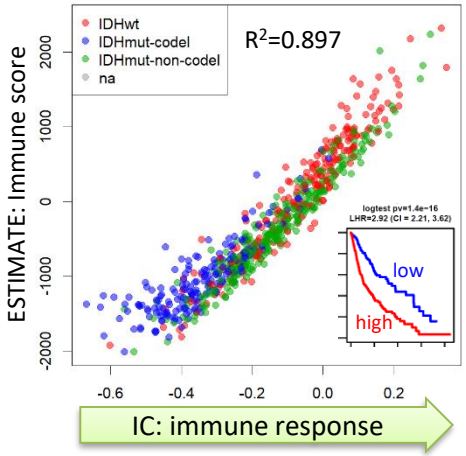
- **Prof. Jörg Hoheisel**
- Dr. Andrea Bauer
- PD Dr. med. Nathalia Giese

Gliomas → DIOMEDES + MEDEA

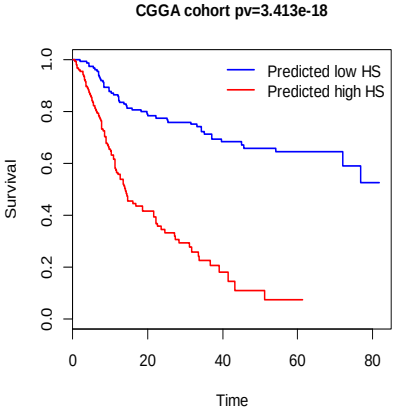
Gliomas in TCGA:



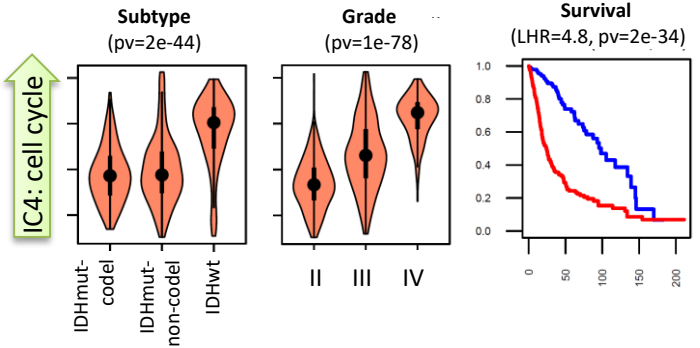
(d) Histopathological differences



Cross-cohort prognosis



Training data - TCGA
Testing data - CGGA



GBM Cell Lines

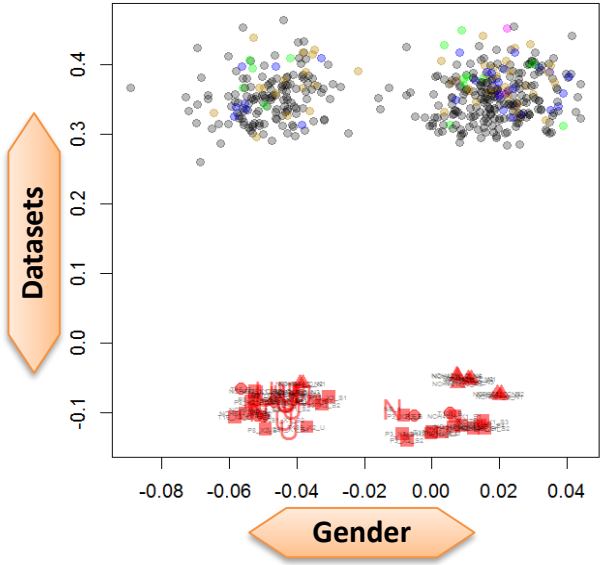
Reference dataset
530 GBM patients
(TCGA)

Investigated dataset
58 samples:
cell lines, xenografts & patient tissues

ICA

Biological knowledge:
bio-processes and sample composition

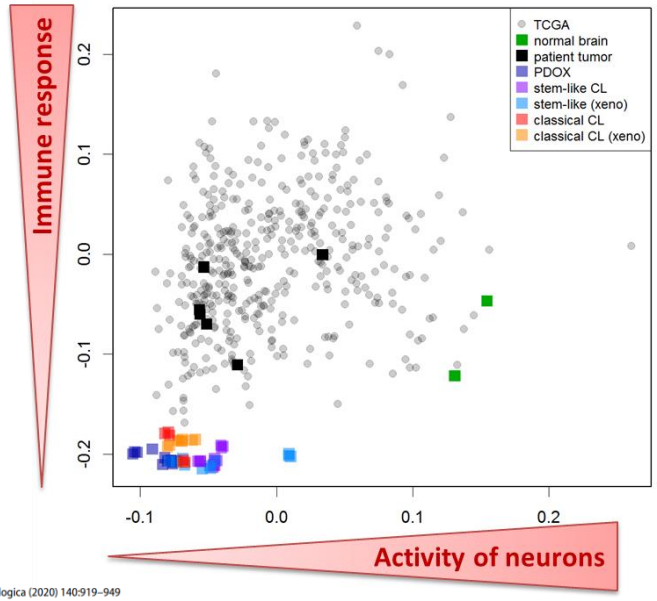
Technical/trivial components: gender and platforms



- We were able to map in-house cell line data onto TCGA dataset (GBM)
- Some components captured *technical factors* →
  (and thus clean other components from them)
- Other – relevant *biological information*: cell cycle, cell migration, presence of stromal and immune cells. We were able to predict phenotype of cell lines using their transcriptomes.

GBM Cell Lines

ICA correctly predicts sample composition & phenotype

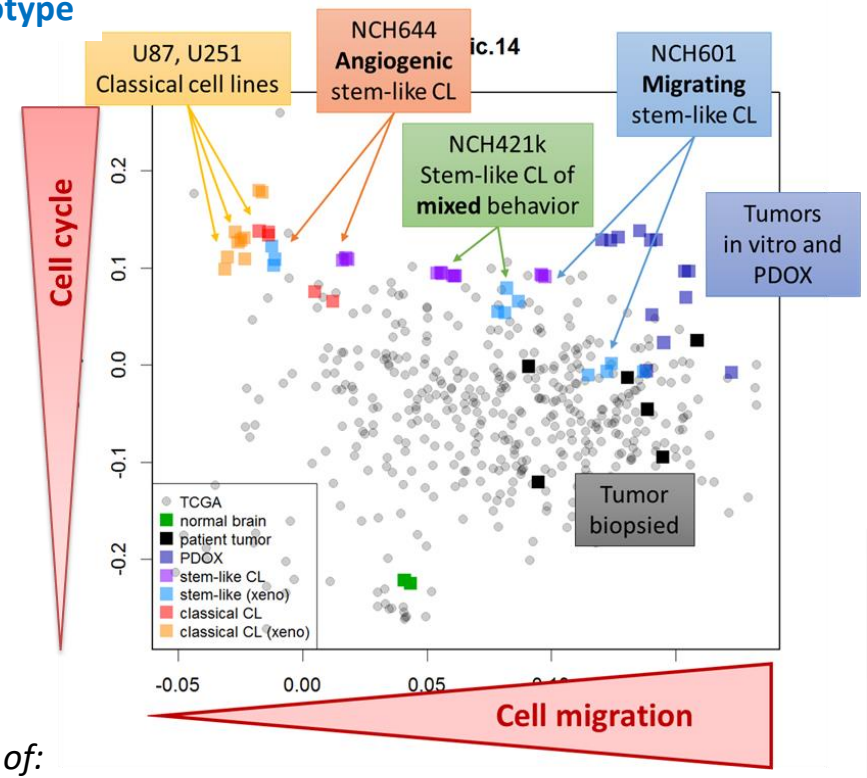


Acta Neuropathologica (2020) 140:919–949
https://doi.org/10.1007/s00401-020-02226-7

ORIGINAL PAPER

Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology

Anna Golebiewska[1] · Ann-Christin Hau[1] · Anaïs Oudin[1] · Daniel Stieber[1,2] · Yahaya A. Yabo[1,3] · Virginie Baus[1] · Vanessa Barthelemy[1] · Eliane Klein[1] · Sébastien Bougnaud[1] · Olivier Keunen[1,4] · May Wantz[1] · Alessandro Michelucci[1,5,6] · Virginie Neirinckx[1] · Arnaud Muller[4] · Tony Kaoma[4] · Petr V. Nazarov[4] · Francisco Azuaje[4] · Alfonso De Falco[2,3,7] · Ben Flies[2] · Lorraine Richart[3,7,8,9] · Suresh Poovathingal[6] · Thais Arns[6] · Kamil Grzyb[6] · Andreas Mock[10,11,12,13] · Christel Herold-Mende[10] · Anne Steino[14,15] · Dennis Brown[14,15] · Patrick May[6] · Hrvoje Miletic[16,17] · Tathiane M. Malta[18] · Houtan Noushmehr[18] · Yong-Jun Kwon[9] · Winnie Jahn[19,20] · Barbara Klink[2,9,19,20,21] · Georgette Tanner[22] · Lucy F. Stead[22] · Michel Mittelbronn[6,7,8,9] · Alexander Skupin[6] · Frank Hertel[6,23] · Rolf Bjerkvig[1,16] · Simone P. Niclou[1,16]

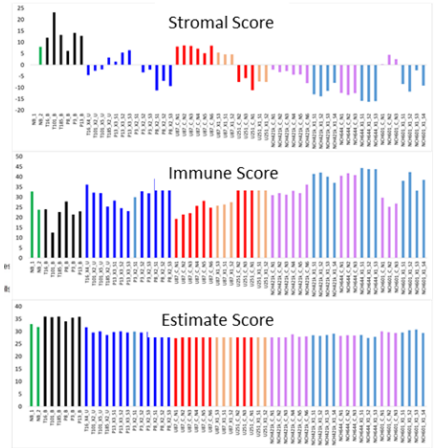


part of:
Golebiewska A. et al, **Acta Neuropathologica, 2020** (link)

Phenotype of cell lines were predicted using unsupervised deconvolution of their transcriptomes!

- ICA deconvolution is reasonable and predicts phenotypic behavior of cell lines
- Tumor cells show higher mobility in xenografts

ESTIMATE was confused



GBM Bulk-Sample Datasets → DIOMEDES

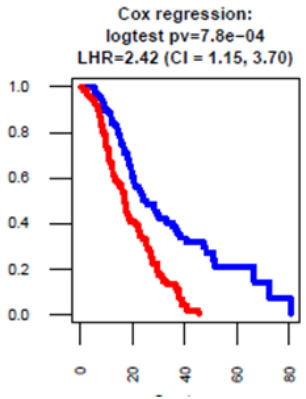
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

GLIOTRAIN

152 samples, IDHwt only

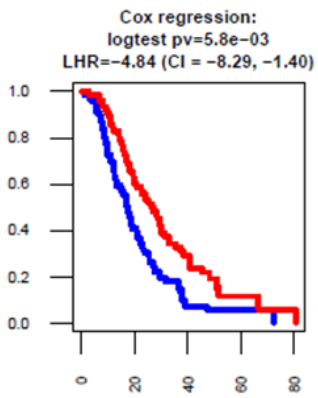
⇒ *consICA* identified relevant & **prognostic** independent signals in a "homogenous" IDHwt datasets,

IC22– B cell activation and phagocytosis



GO (BP/CC/MF)	FDR
phagocytosis	1e-5
humoral immune response	1e-5
B cell receptor signaling pathway	4e-4
immunoglobulin complex	2e-9
immunoglobulin receptor binding	6e-7

IC39 – oxygen transport



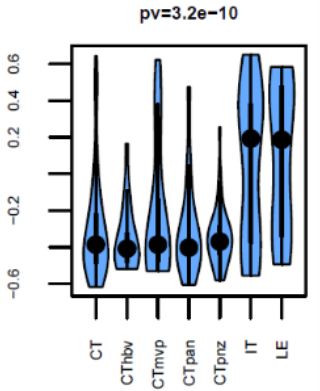
GO (BP/CC/MF)	FDR
oxygen transport	1e-2
integral complex of plasma membrane	2e-2
oxygen carrier activity	2e-3
haptoglobin binding	6e-3
oxygen binding	1e-1

IVY GAP

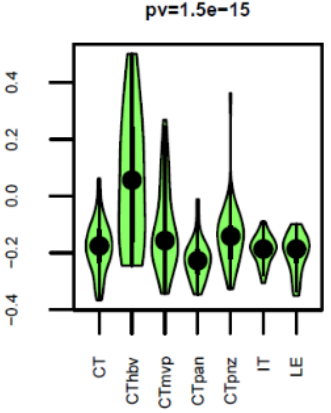
262 samples, 37 patients (~7 regions)

⇒ *consICA* characterize signals overrepresented in different tumor **niches**

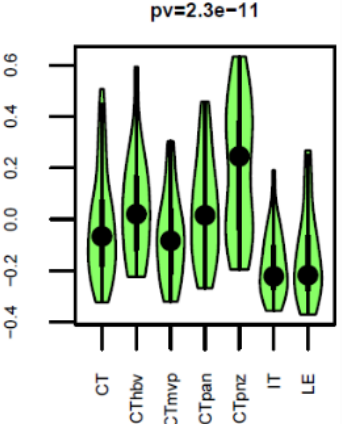
IC2 – activity of neurons



IC3 – angiogenesis



IC7 – inflammation



CT = Cellular tumor
CThbv = Hyperplastic blood vessels
CTmvp =Microvascular proliferation
Ctpan = Around necrosis
CTpnz = Perinecrotic zone
IT = Infiltraitng tumor
LE = Leading edge

Pan-cancer & Multi-omics

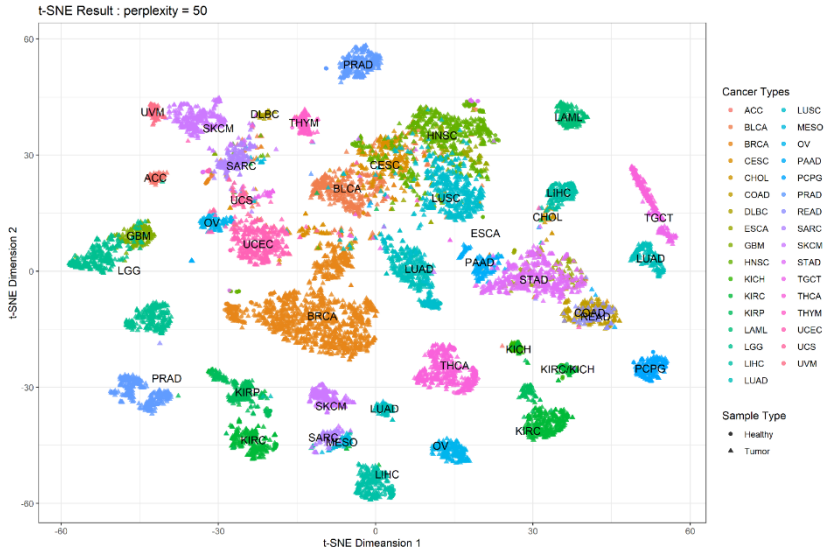
Yibioa Wang
(MSc)

TCGA

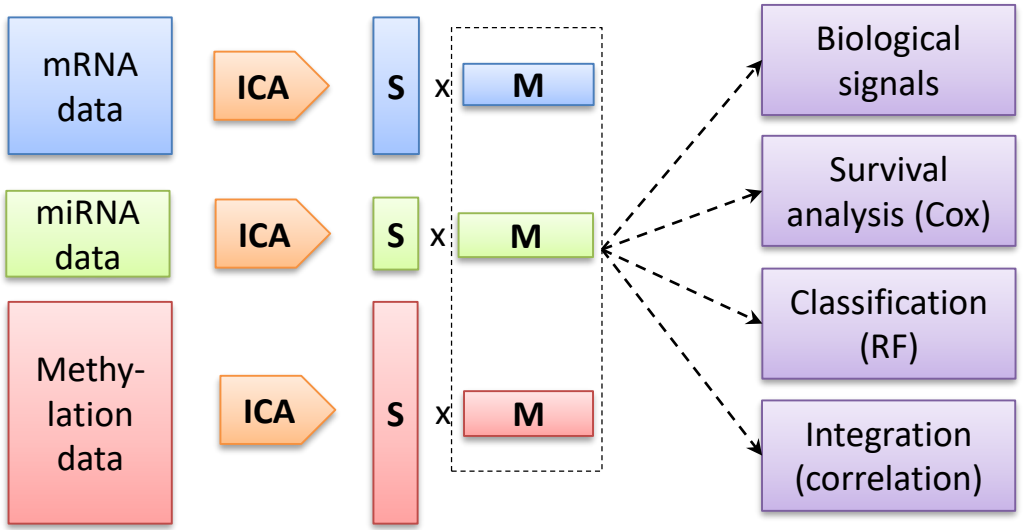
The Cancer Genome Atlas

>11k patients, 33 types of tumors

- **clinical data** (age, gender, survival...)
- **mRNA** (10k samples, 20k features)
- **miRNA** (> 9k samples, ~1k features)
- **methylation** (>9k samples, 450k features)

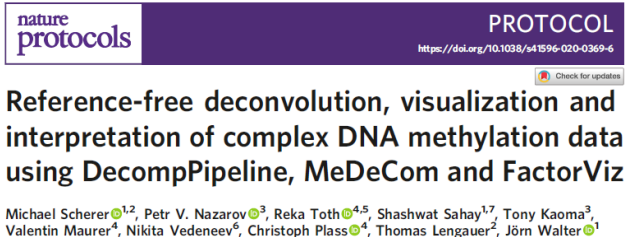


Approach



Here we used *consICA* with 100 components & 40 runs

another example of ICA for methylation data:
Scherer M, Nazarov P, et al. **Nature Protocols, 2020** (link)

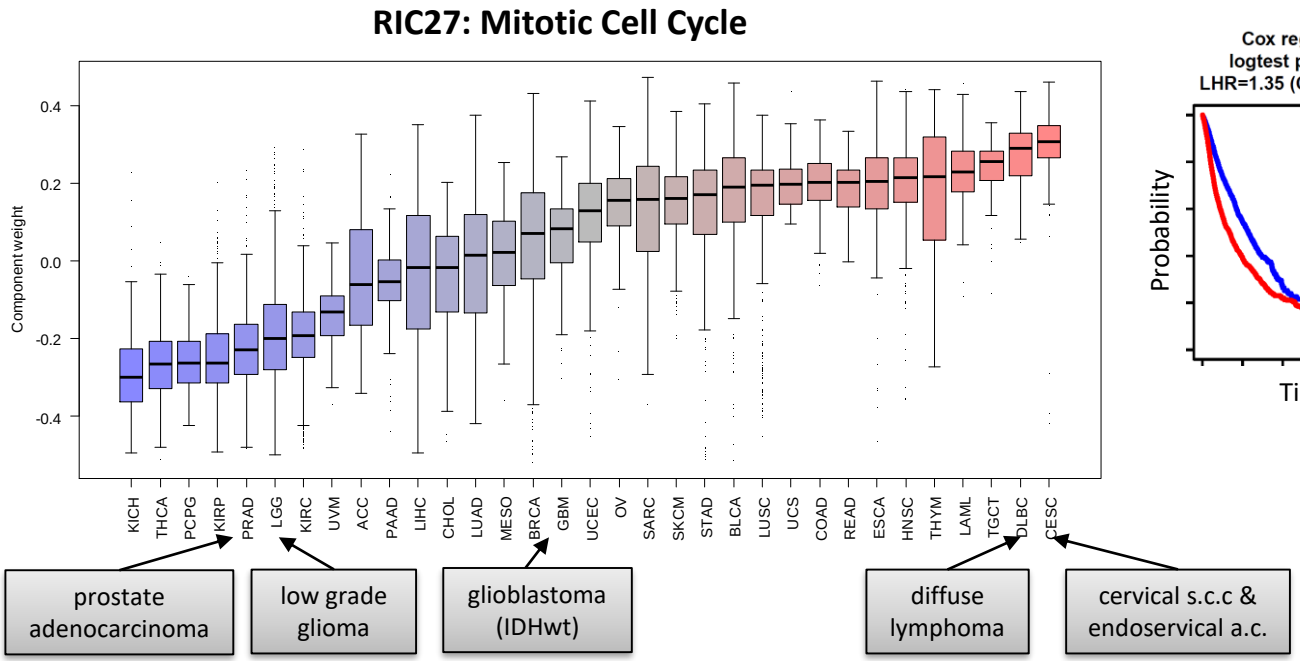

nature protocols
PROTOCOL
https://doi.org/10.1038/s41596-020-0369-6
Check for updates

Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz

Michael Scherer[1,2], Petr V. Nazarov[3], Reka Toth[4,5], Shashwat Sahay[1,7], Tony Kaoma[3], Valentin Maurer[4], Nikita Vedeneev[6], Christoph Plass[4], Thomas Lengauer[2], Jörn Walter[1] and Pavlo Lutsik[4]✉

Pan-cancer: ICA Components

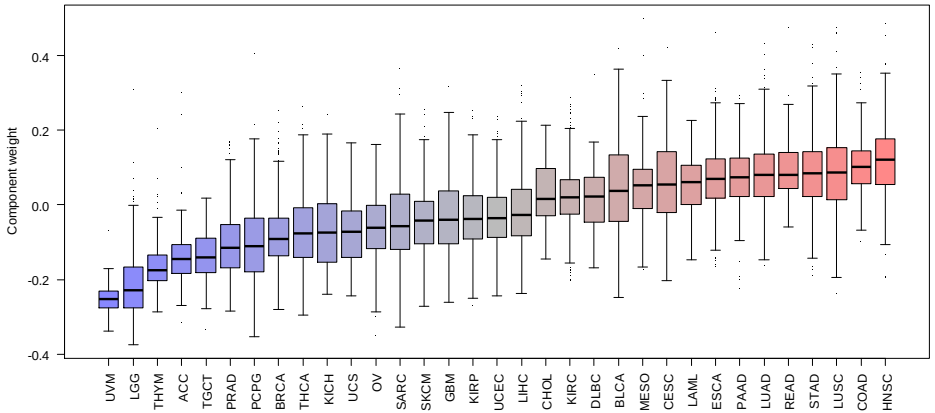
ICA Results: Cell Cycle

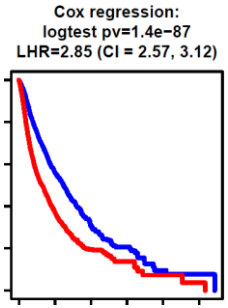


Code	Study Name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical sq. cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large b-cell lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid leukemia
LCML	Chronic myelogenous leukemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma

Pan-cancer: ICA Components

RIC17: Signal of Mast Cells*



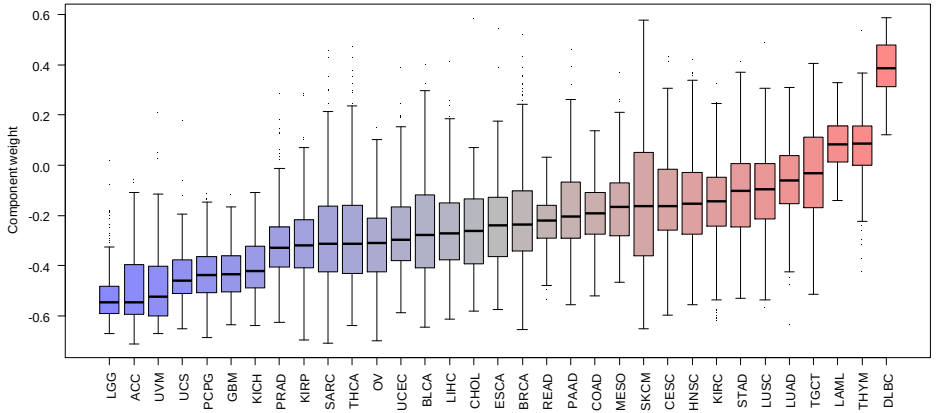


Tumor-associated mast cells (TAMCs) ?

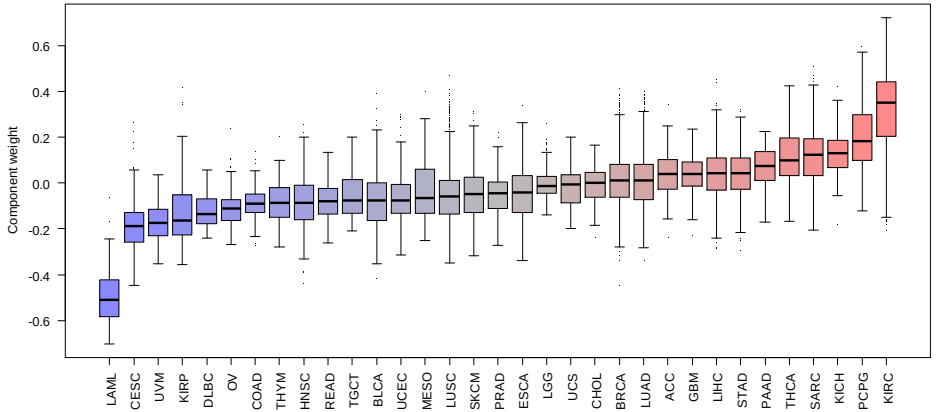
This can we wrong – we need a review from a biologist! ☺

(*) assigned based on LM22 signature (CIBERSORT)

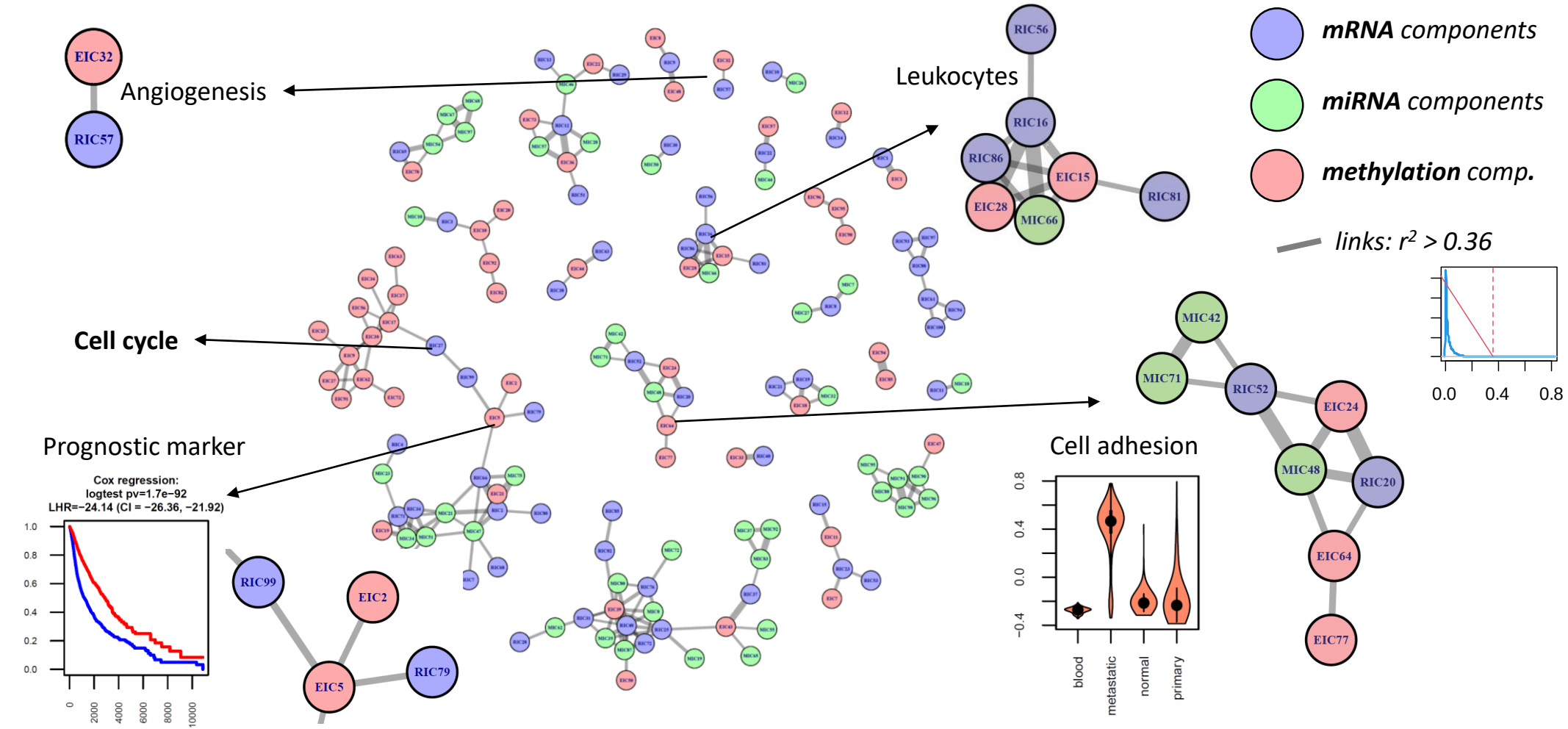
RIC16: Signal of T-Cells*



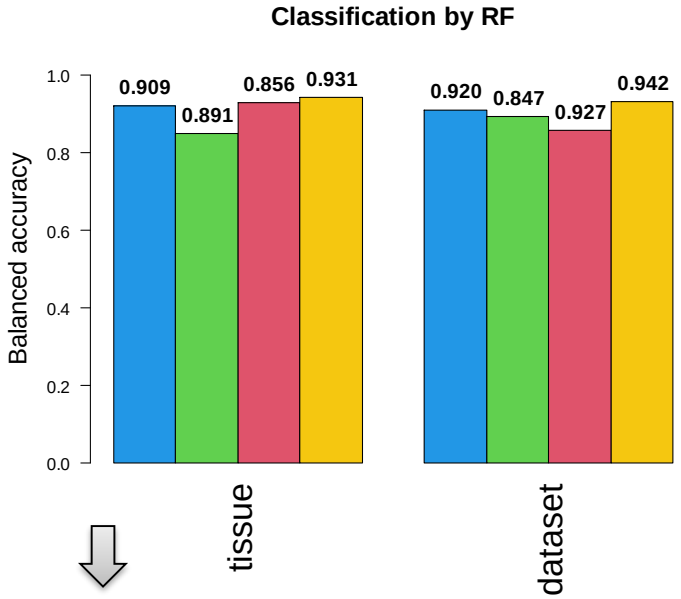
RIC57: Angiogenesis

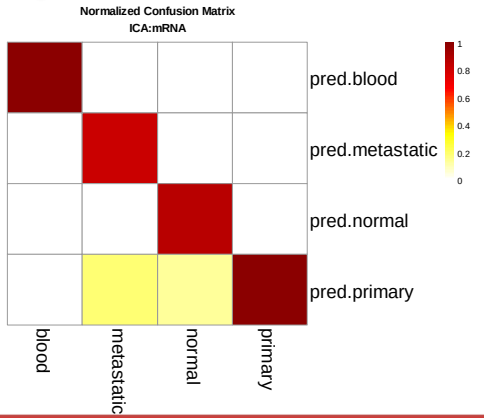


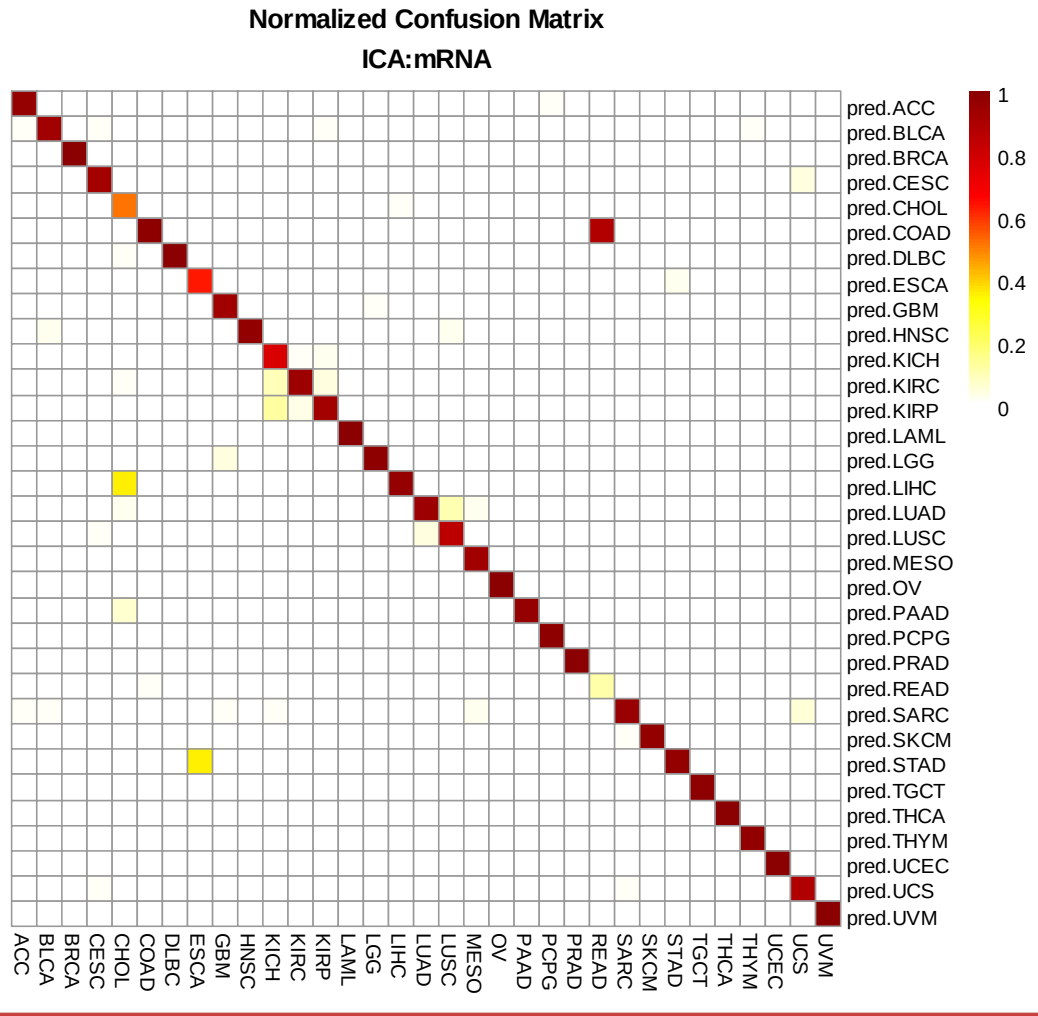
Pan-cancer: ICA-based Data Integration



Pan-cancer: Classification

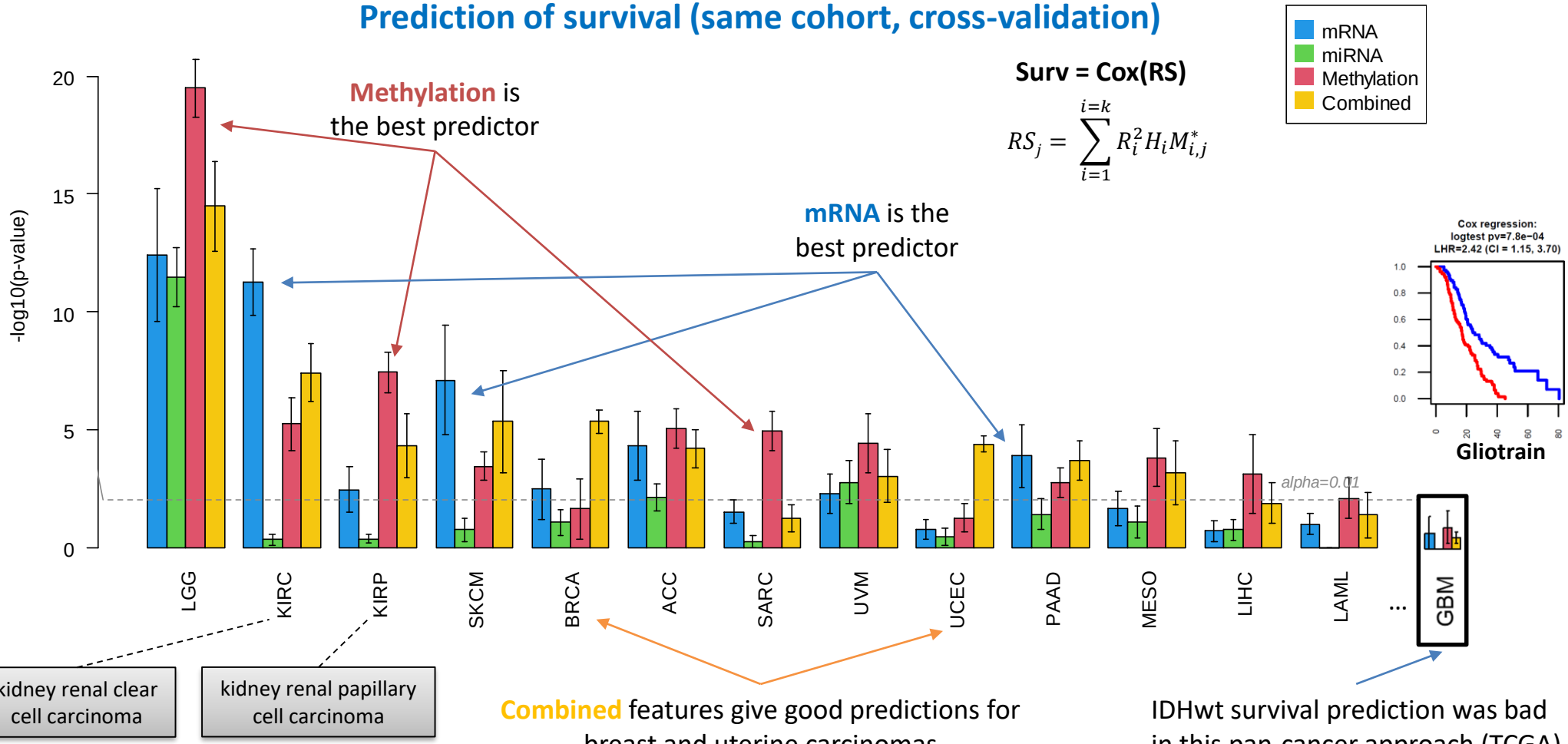






Pan-cancer: Prognosis

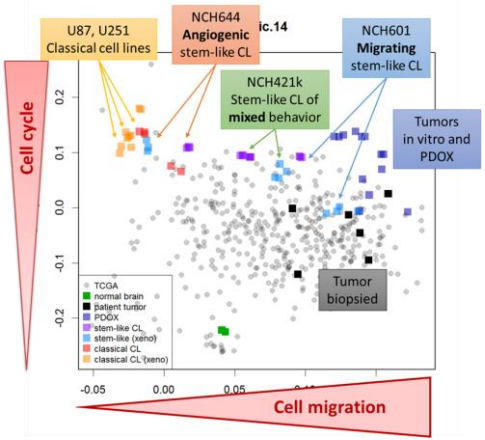
Prediction of survival (same cohort, cross-validation)



Conclusions I: Bulk Sample Data

- Our modification of ICA deconvolution (*consICA*) :
  - Corrects technical biases
  - Extracts "cleaned" biological signals from bulk-sample data
  - Maps new samples into the space of biologically meaningful components
  - Extracts prognostic features and features with classification power
  - Can be used to integrate multi-omics data
  - Diagnostic & prognostic properties could be expected for many cancers

- Was validated:
  - Using acceptable computational methods (cross-validation)
  - On cell lines
  - Independent cohorts of patients



DIOMEDES*

* Single-cell part

Maryna Chepeleva
(PhD student)

Arnaud MULLER
(Bioinf.)

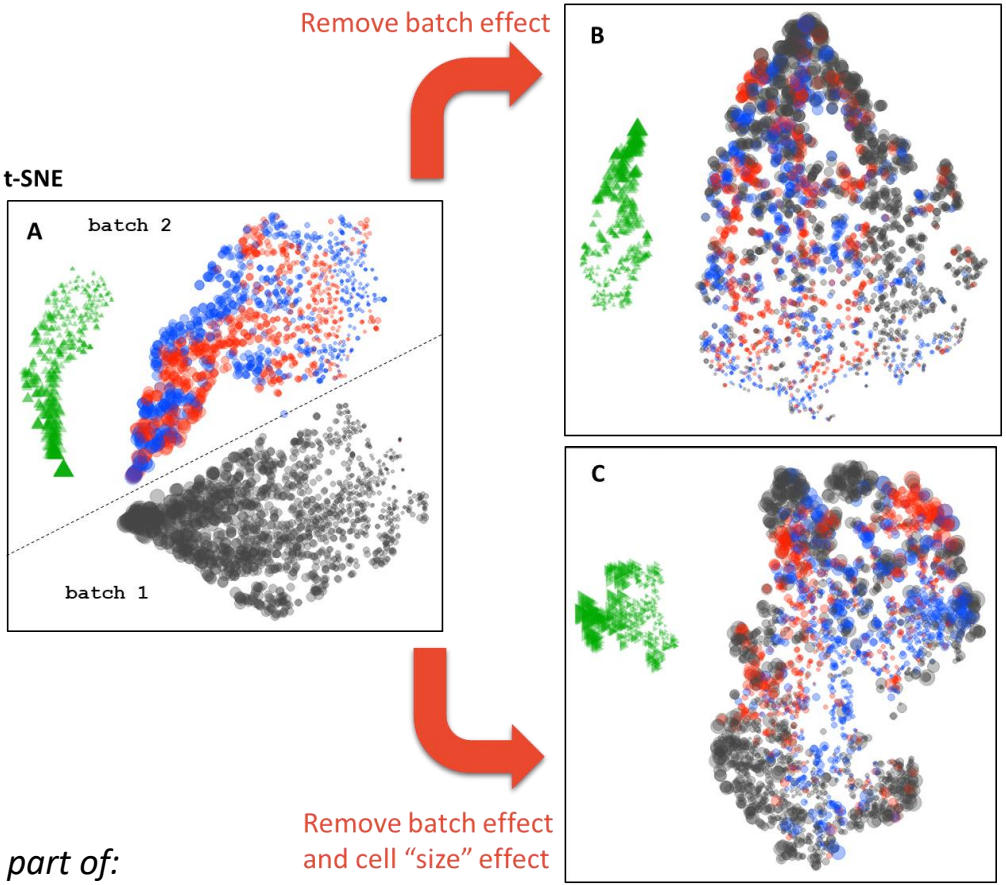
Tony KAOMA
(Bioinf.)

Anna GOLEBIEWSKA (PI)

Simone NICLOU

ICA for Single Cell RNA-seq Data

Correction of technical effects



$$E_{nm} \Rightarrow S_{nk} \times M_{km}$$
$$M'_{7,m} \leftarrow 0$$
$$E'_{nm} \leftarrow S_{nk} \times M'_{km}$$

t-SNE representations of original data (A) and ICA-recovered data, after excluding batch effect (B) or several (C) components linked to technical factors ("cell size").

part of:
Dirkse, Golebiewska, et al. **Nature Communications, 2019** (link)

nature COMMUNICATIONS

ARTICLE

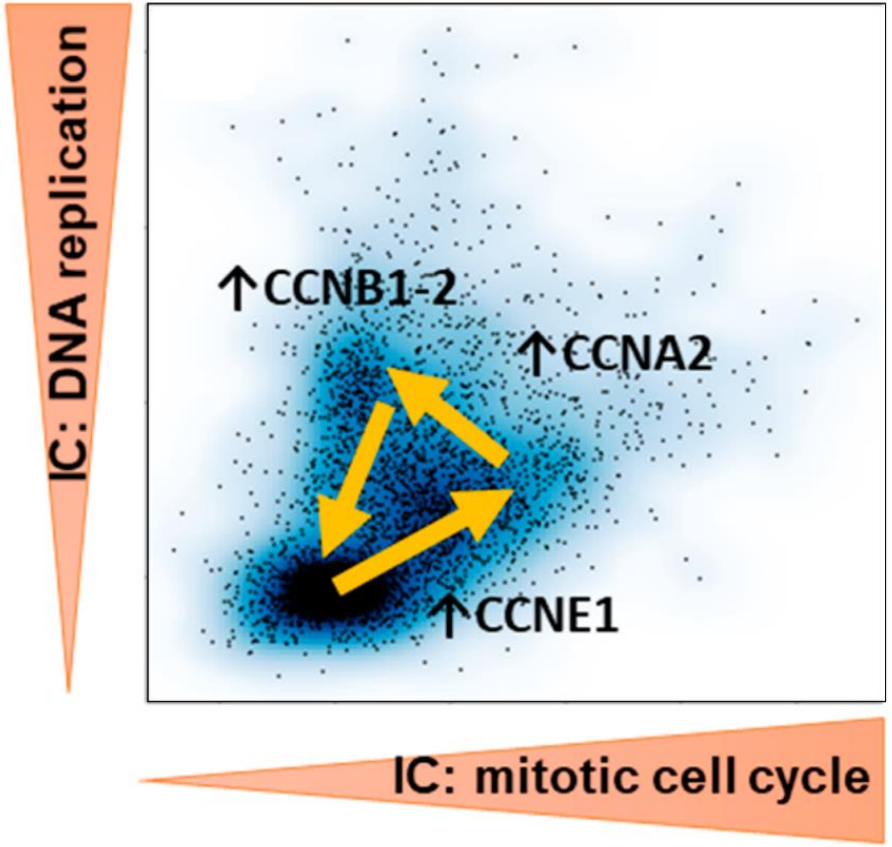
https://doi.org/10.1038/s41467-019-09853-z OPEN

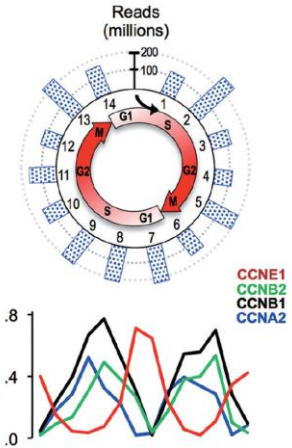
Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment

Anne Dirkse[1,2,12], Anna Golebiewska[1,12], Thomas Buder[3,4], Petr V. Nazarov[5], Arnaud Muller[5], Suresh Poovathingal[6], Nicolaas H.C. Brons[7], Sonia Leite[8], Nicolas Sauvageot[8], Dzjemma Sarkisjan[1], Mathieu Seyfrid[1], Sabrina Fritah[1], Daniel Stieber[1], Alessandro Michelucci[1,6], Frank Hertel[9], Christel Herold-Mende[10], Francisco Azuaje[5], Alexander Skupin[6], Rolf Bjerkvig[1,11], Andreas Deutsch[3], Anja Voss-Böhme[3,4] & Simone P. Niclou[1]

ICA for Single Cell RNA-seq Data

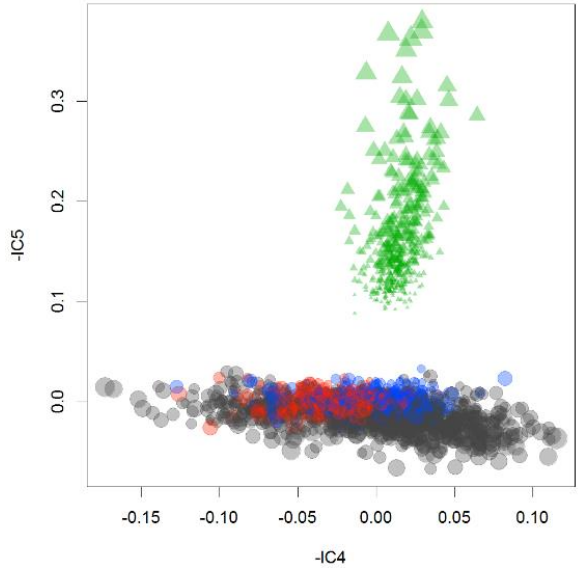
Cell Cycle in Single Cells





Dominiguez (2016) Cell Research

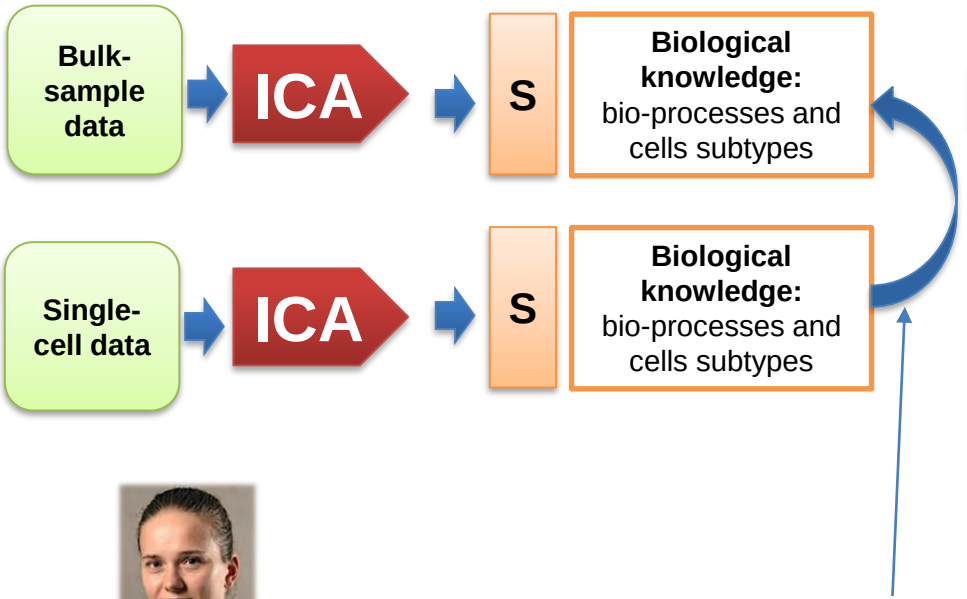
Cell Types



Dirkse, Golebiewska et al. **Nature Communications, 2019** (link)
Sompairac, Nazarov el al. **Int J Mol Sci, 2019** (link)

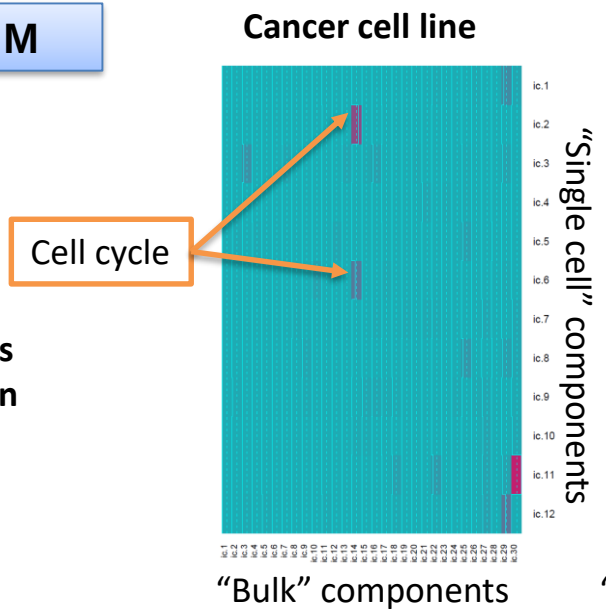
Combining Bulk and Single Cell

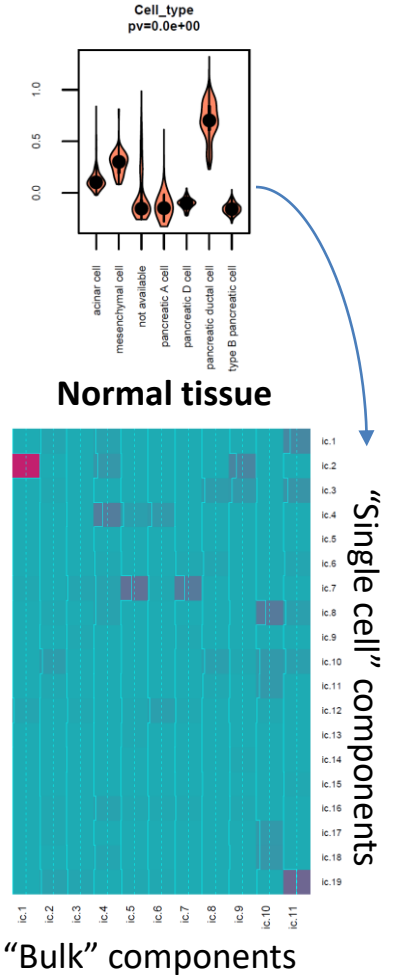
Direct comparison of ICA results on bulk-sample data with single-cell data does not work!



Maryna Chepeleva
(PhD student)

Single-cell data helps interpreting signals in bulk-sample deconvolution





https://f1000research.com/slides/9-1025

GBM: Single Cell Data

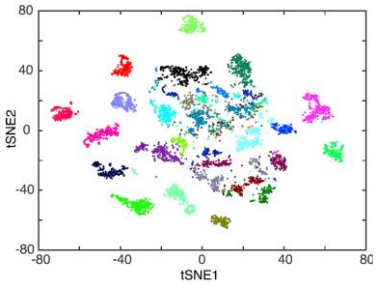
Neftel,...,Suvà, **Cell, 2019** (link)

Article

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

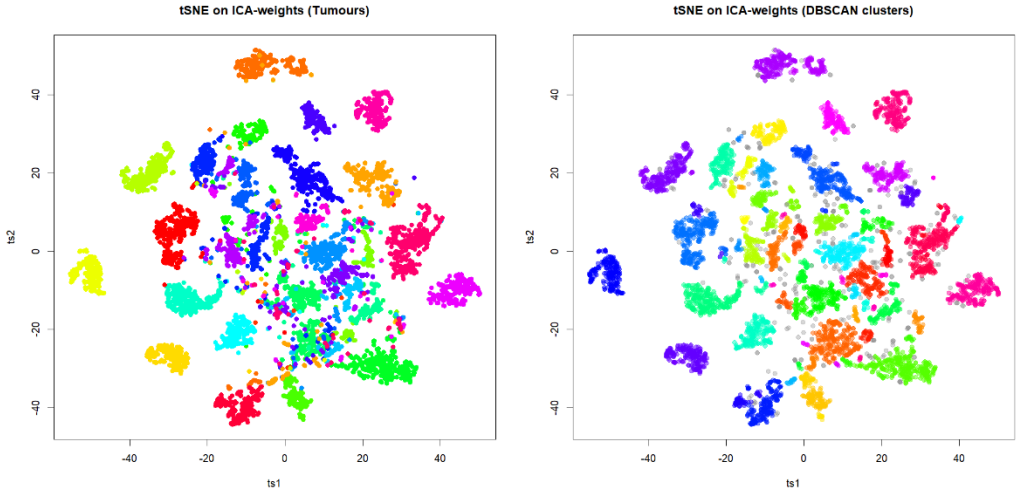
Cyril Neftel,[1,2,3,4,19] Julie Laffy,[5,19] Mariella G. Filbin,[1,2,3,6,19] Toshiro Hara,[1,2,3,7,19] Mami E. Shore,[1,2,3] Gilbert J. Rahme,[1,2,3]

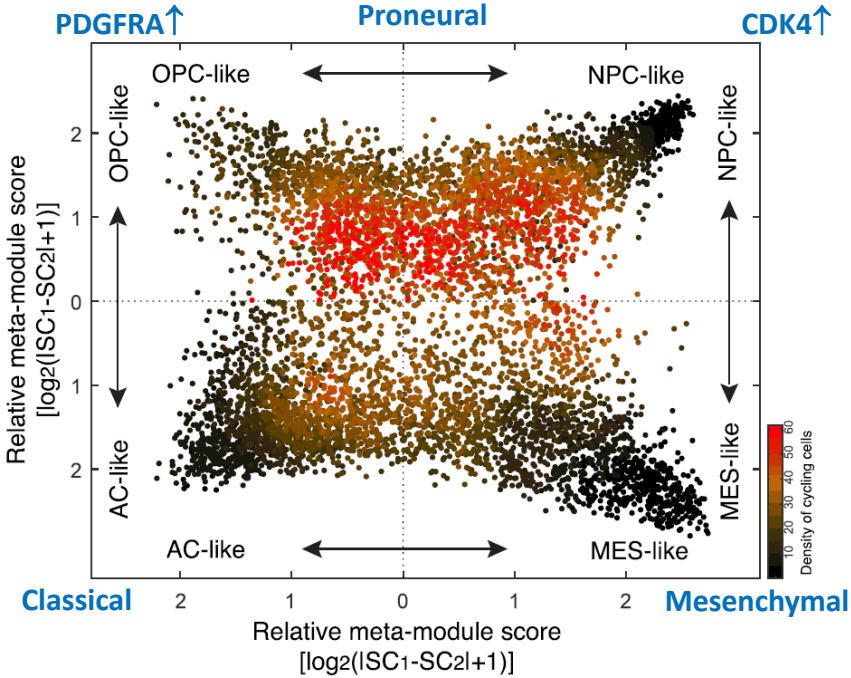
Cell



SC RNA-seq Data → ICA → S-matrix X M-matrix

2 datasets were reported performed with **SMART-Seq** and **10x** platform:
no discrete tumor cell subtypes!
Instead, cells show gradient of states





(1) Topologically similar to tSNE on original data

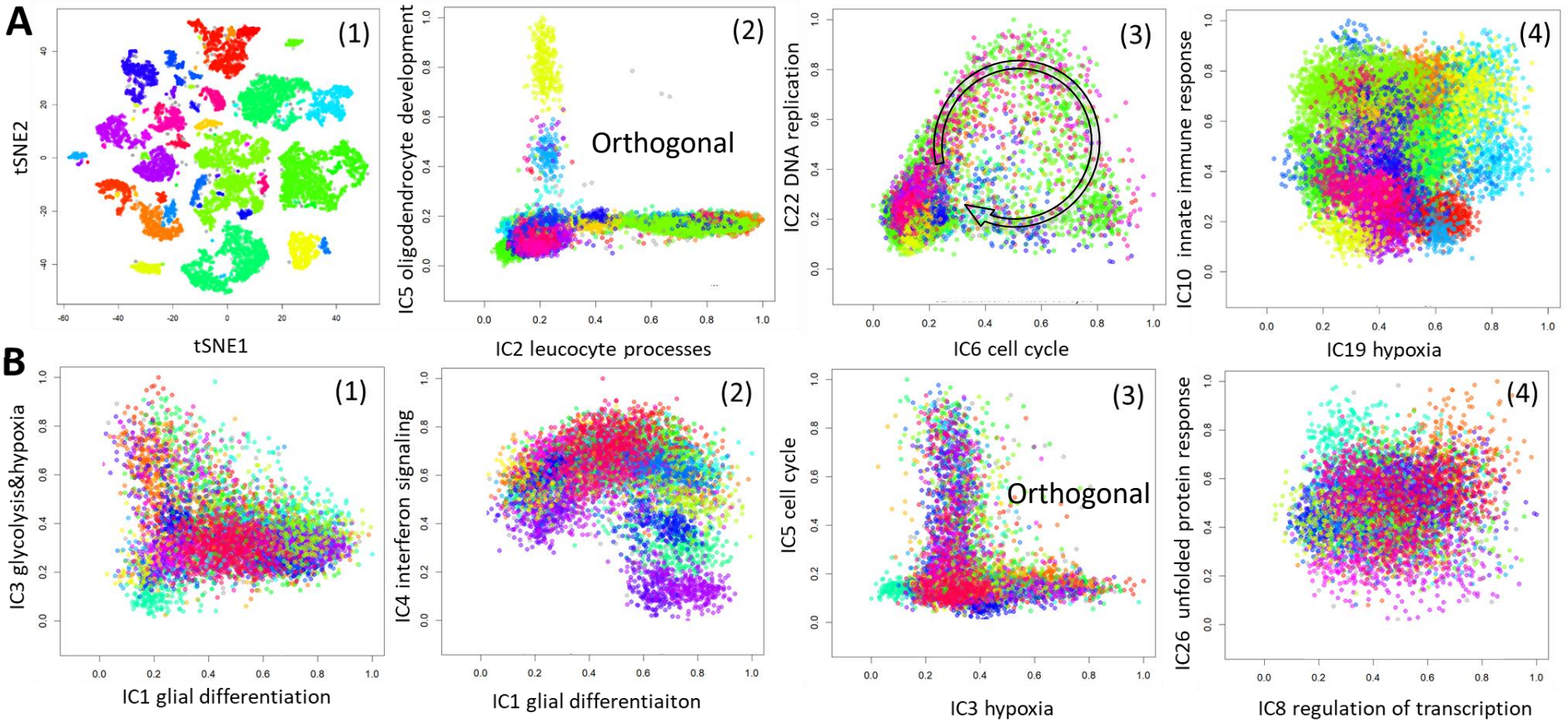
(2) Large inter-tumor variability (both SMART-Seq and 10x) !

GBM: Single Cell Data Deconvolution

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

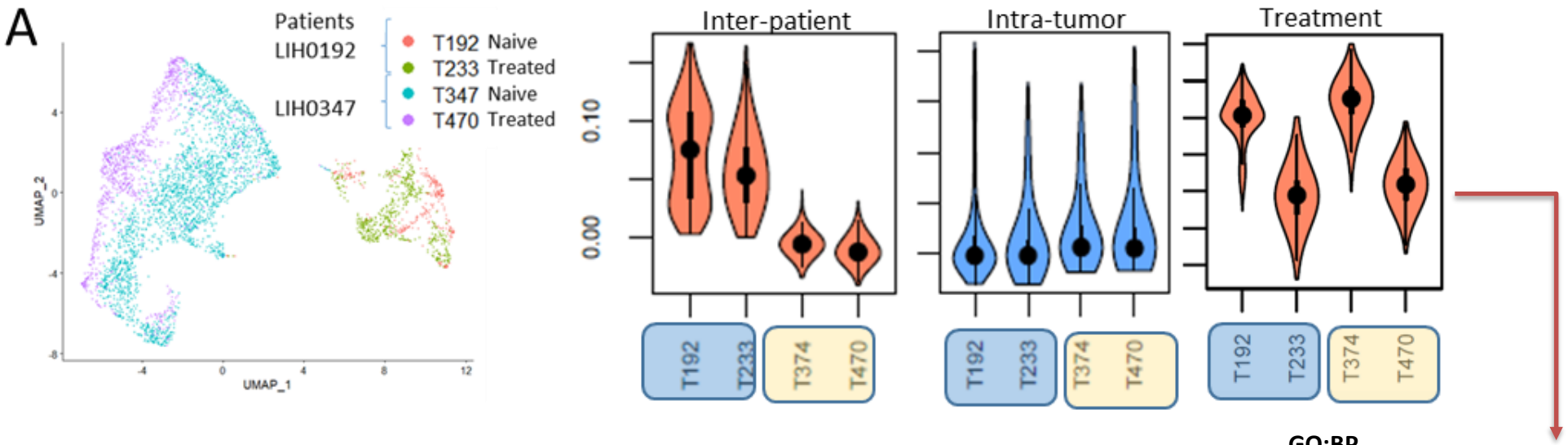
10X

Smart-Seq



consICA can assess cellular subpopulations and phenotypic states associated with specific biological processes

Longitudinal PDOX: TMZ Resistance



Examples of independent components that represent tumor variability:

- between patients (left)
- within each patient regardless treatment (middle)
- within each patient before and after treatment (right)

Deconvolution captured signal related to treatment
(no genes using standard DEA approach)

GO:BP
translation initiation
mRNA catabolic process
cytoplasmic translation
GO:CC
cytosolic ribosome
GO:MF
structural constituent of ribosome
RNA binding

Conclusions II: Single Cell Data

- Deconvolution with *consICA* on single cell data:
  - Corrects technical biases
  - Extracts signals of biological processes (or cell types) from single-cell data
  - Could detect weak signals that are masked by other processes (e.g. TMZ resistance masked by cell cycle and inter-tumor variability)
  - Can be used to interpret results of bulk-sample data deconvolution

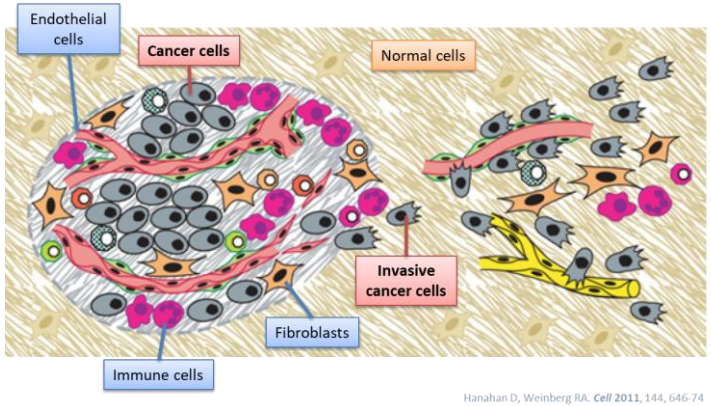
MEDEA

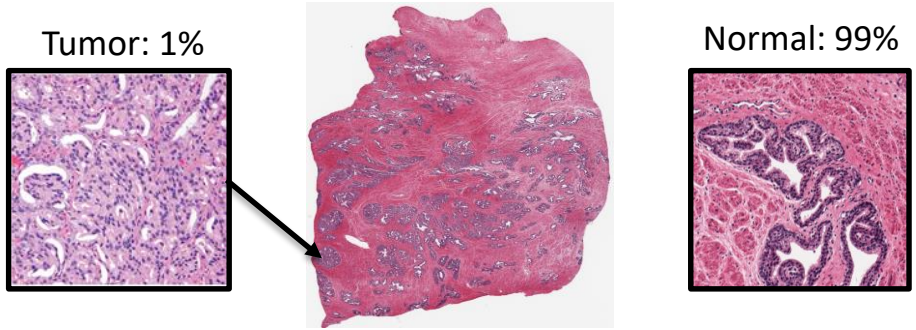
Sang Yoon KIM

Olivier KEUNEN

Michel MITTELBRONN (Co-PI)

Background

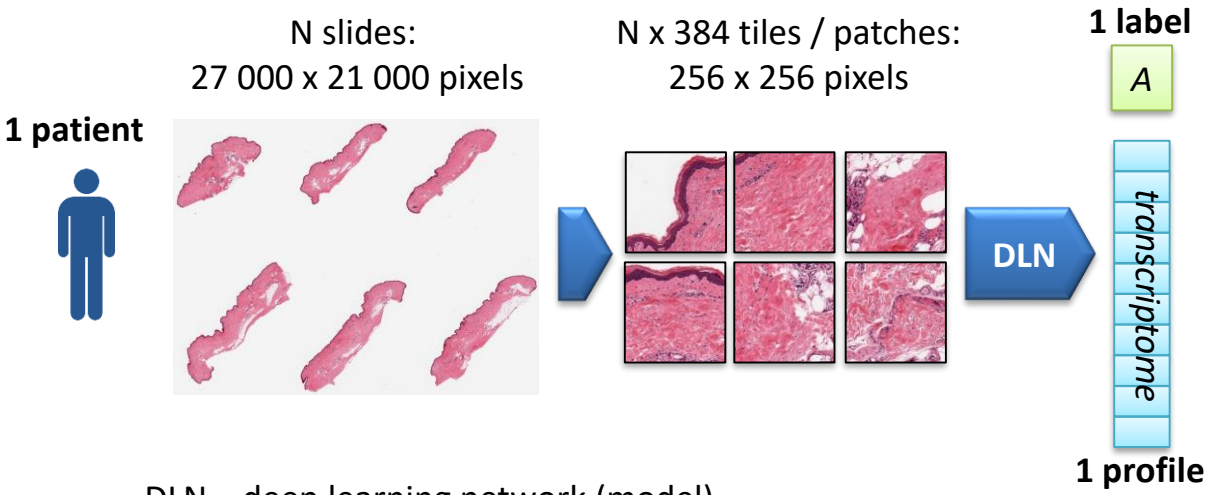




- Native heterogeneity of tissues
- Inter/intra tumor heterogeneity

Issues in histopathological image analysis:

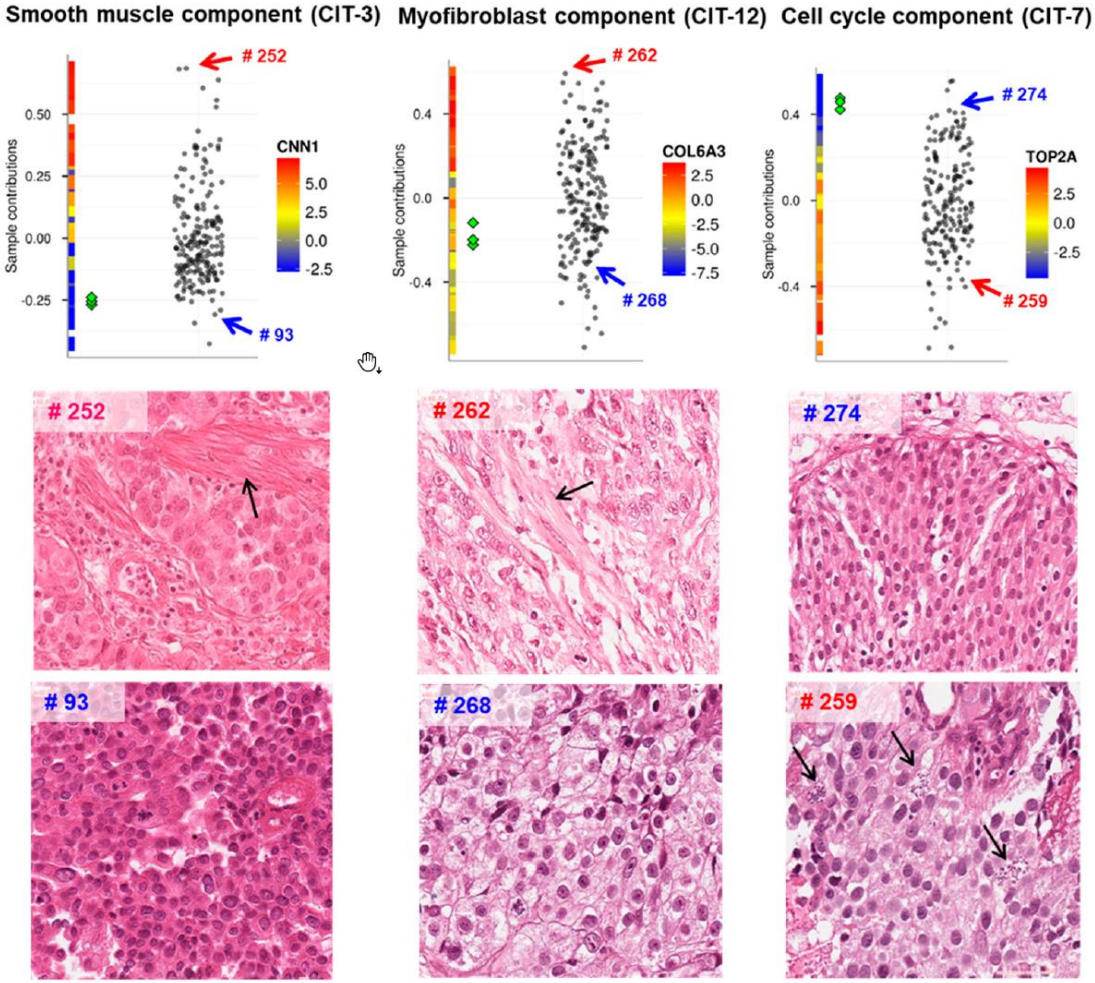
- Tedious analysis
- In some cancers (e.g. prostate) < 1% of the image is cancer-related
- Standard approaches require supervised "pixel-wise" labelling - unrealistic



Observation 1

ICA:

Pathology:



International Journal of *Molecular Sciences*

MDPI

Review

Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets

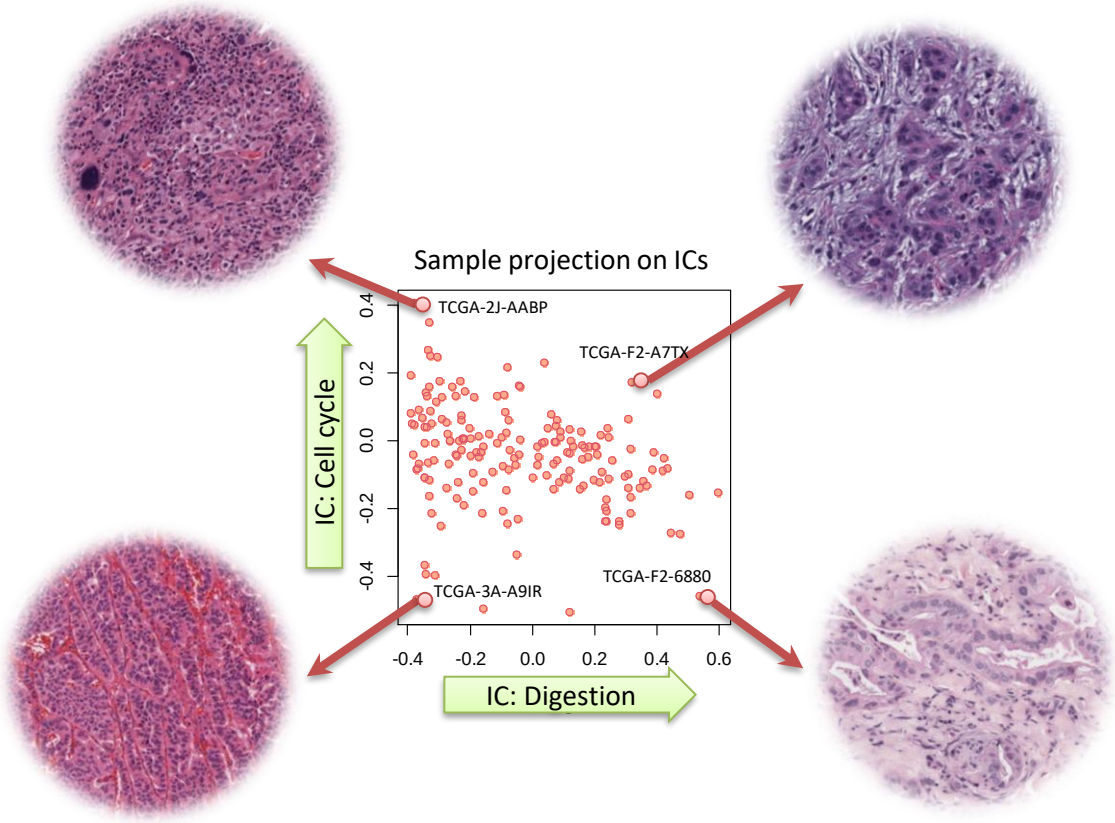
Nicolas Sompairac [1,2,3,4], Petr V. Nazarov [5], Urszula Czerwinska [1,2,3], Laura Cantini [6], Anne Biton [7], Askhat Molkenov [8], Zhaxybay Zhumadilov [8,9], Emmanuel Barillot [1,2,3], Francois Radvanyi [1,10], Alexander Gorban [11,12], Ulykbek Kairov [8] and Andrei Zinovyev [1,2,3,*]

Observation

ICA results are confirmed by H&E histopathology: smooth muscles, fibroblasts, cell cycle were observed.

Observation 2

ICA results of mRNA expression data from TCGA-PAAD cohort



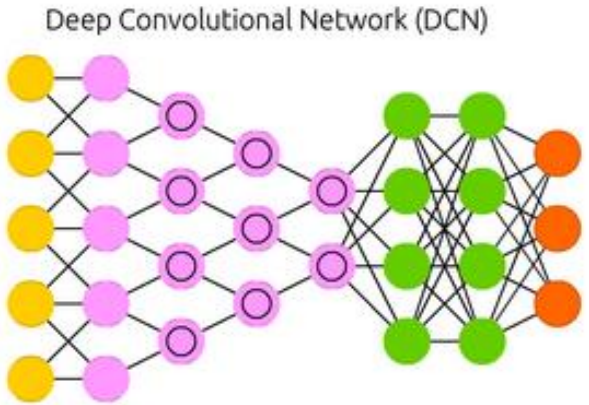
Idea for Approach

Patient classification using weakly supervised DLN

nature medicine
ARTICLES
https://doi.org/10.1038/s41591-019-0508-1

Clinical-grade computational pathology using weakly supervised deep learning on whole slide images

Gabriele Campanella[1,2], Matthew G. Hanna[1], Luke Geneslaw[1], Allen Miraflor[1], Vitor Werneck Krauss Silva[1], Klaus J. Busam[1], Edi Brogi[1], Victor E. Reuter[1], David S. Klimstra[1] and Thomas J. Fuchs[1,2*]



RNA-seq prediction

nature COMMUNICATIONS
ARTICLE
https://doi.org/10.1038/s41467-020-17678-4 OPEN
Check for updates

A deep learning model to predict RNA-Seq expression of tumours from whole slide images

Benoît Schmauch[1✉], Alberto Romagnoni[1,4], Elodie Pronier[1,4], Charlie Saillard[1], Pascale Maillé[2,3], Julien Calderaro[2,3], Aurélie Kamoun[1], Meriem Sefta[1], Sylvain Toldo[1], Mikhail Zaslavskiy[1], Thomas Clozel[1], Matahi Moarii[1], Pierre Courtiol[1,5] & Gilles Wainrib[1,5✉]

Idea

Instead of predicting mixed, bulk-sample mRNA or DNA-methylation signals, we will predict already deconvolved, clean signals, linked to biological processes & cell subpopulations.

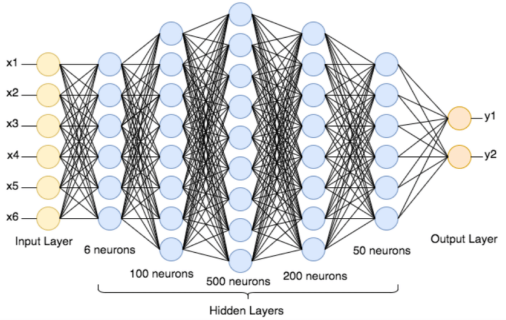
Artificial Neural Networks → Deep Learning Networks

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

MLP

Deep Feed Forward (DFF)

Multilayer perceptron, a.k.a.
(Deep) feed-forward network,
back-propagation network
fully-connected layers,
etc…

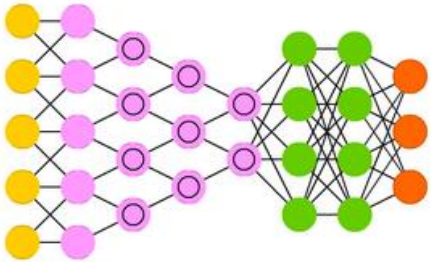


My first "love"… ☺

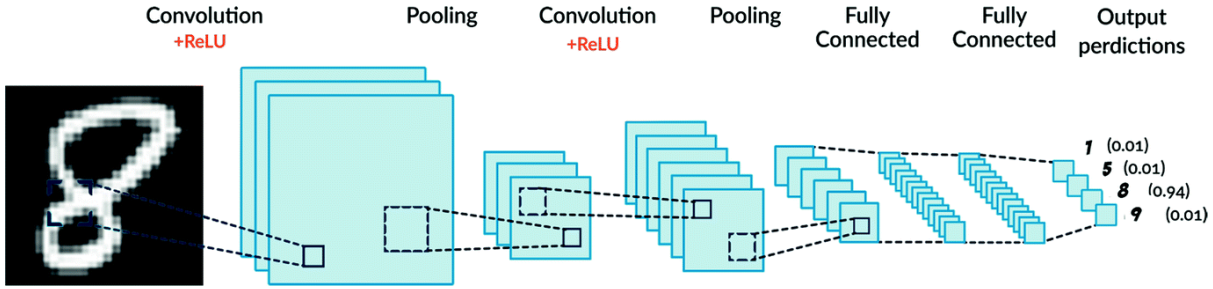
Nazarov et al (2004)
J Chem Inf Comput Sci

CNN

Deep Convolutional Network (DCN)

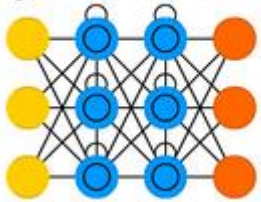


Convolutional networks

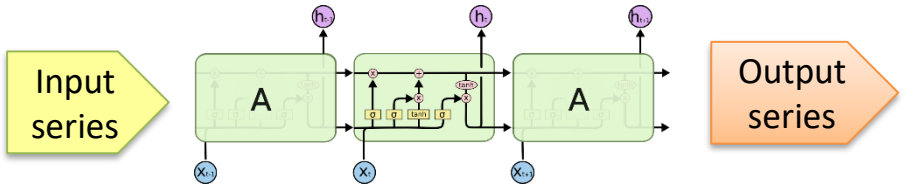


RNN

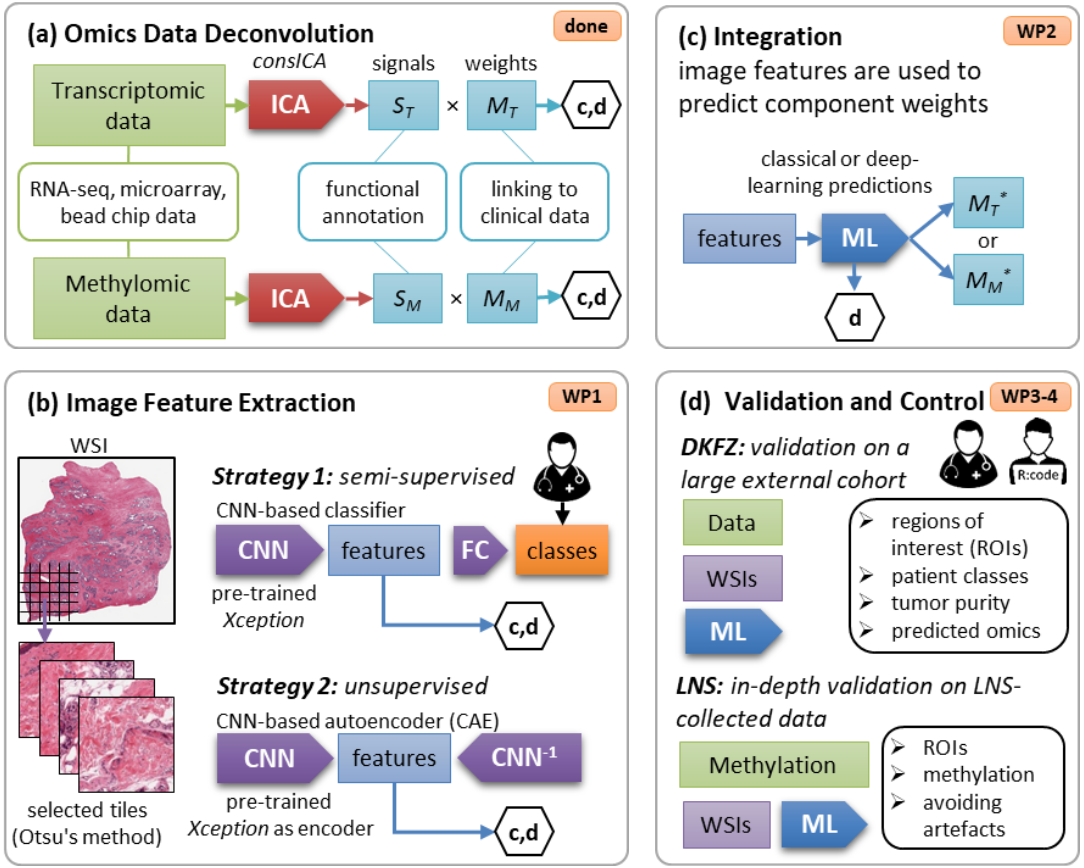
Long / Short Term Memory (LSTM)



Recurrent networks



MEDEA: Project Overview



CAE: convolutional autoencoder; **CNN**: convolutional neural network; **FC**: fully-connected network or layer; **ICA**: independent component analysis; **ML**: machine learning; **ROI**: region of interest; **WSI**: whole slide image.

(a) Deconvolution of the omics data using developed tool *consICA*. This method was already developed and applied to entire GTEx (mRNA), TCGA (mRNA and meDNA), and DKFZ (mRNA) cohorts.

(b) Image analysis and feature extraction starts with a pre-trained *Xception* model and uses weakly supervised training to fine-tune model's parameters. Two strategies will be compared in the project: strategy 1 is a semi-supervised one using CNN-based classifier and strategy 2 – completely unsupervised using CAE. *Xception* will be used as an initial estimation of the encoder's parameters.

(c) Integration of ICA-weights and image features will be done either by a classical ML-approach (linear regression or random forest regression) or by a FC neural network.

(d) A thorough validation of the results include (i) validation of an external pancreatic cancer cohort (DKFZ) and collection and (ii) in-depth analysis of in-house (LNS) samples of glioma patients. The expertise of the Co-PI (pathologist) will be used to validated predictions and the PI and his team will control that the WSI-features are sensible and not artefacts.

Preliminary Results

670+ patients

27 organs

"normal" tissues

input **GTEx Data:**

15k+ slides

17k+ RNA-seq

Sample: 1278 slides

zoom x10, ~10k x 10k px

480k tiles

256 x 256 px

Xception model

Classification

Tile features

Slide features

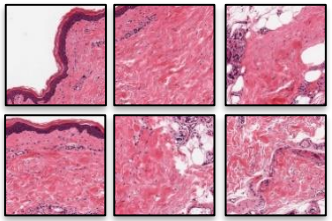
ICA

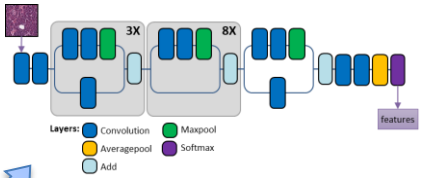
Functional annotation of ICs

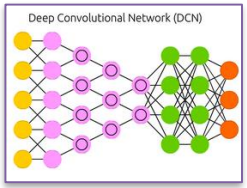
Weights of ICs

Integration & predictions

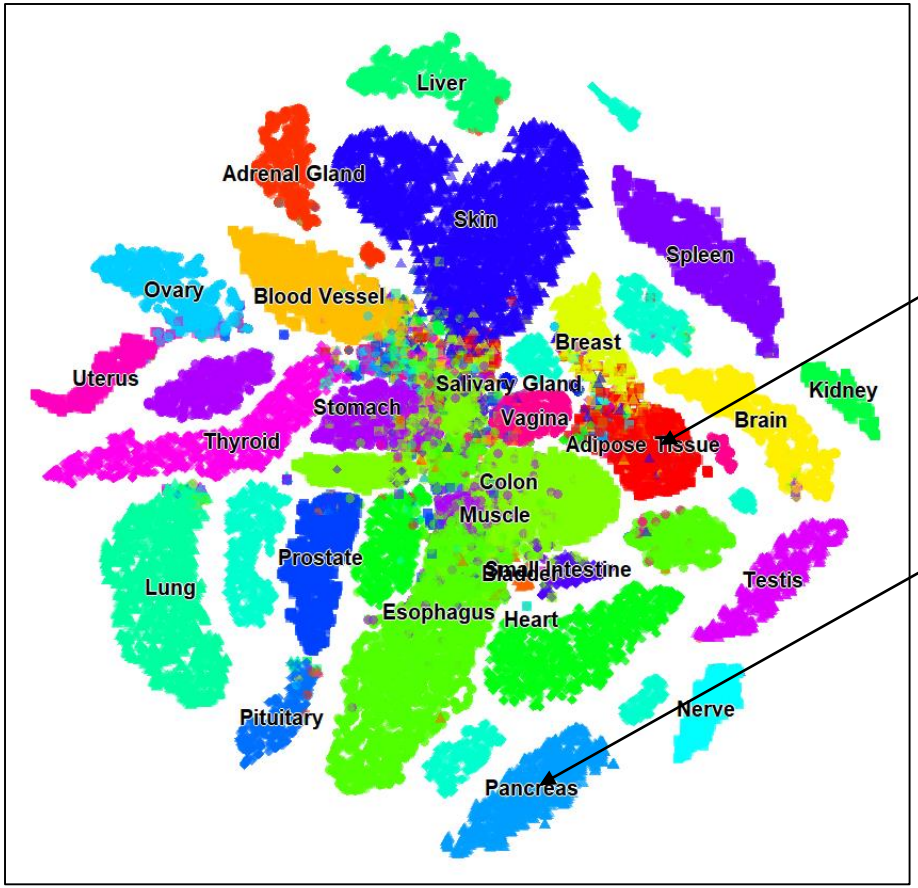
output



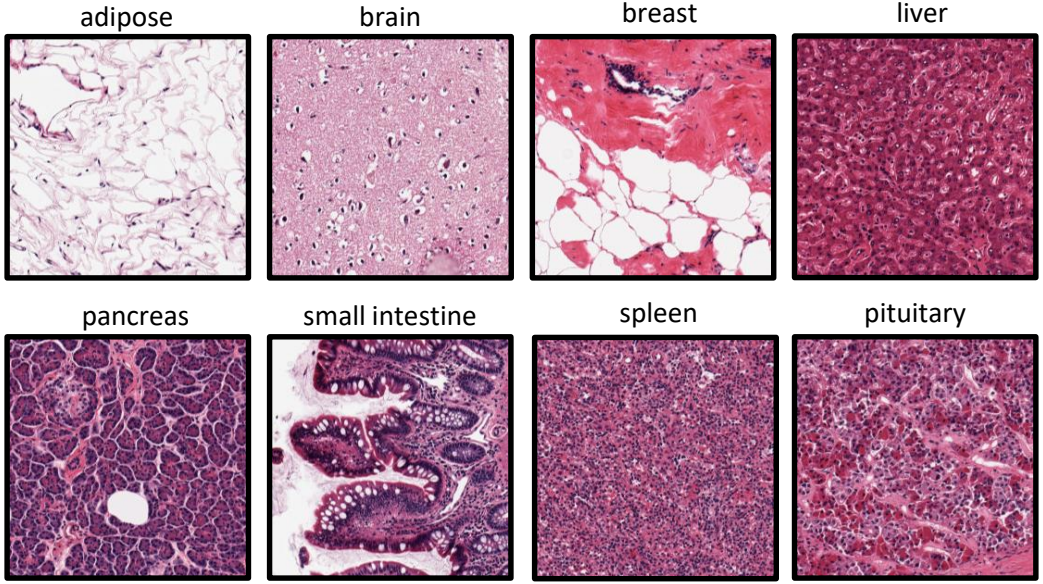




Tile-level Feature Extraction



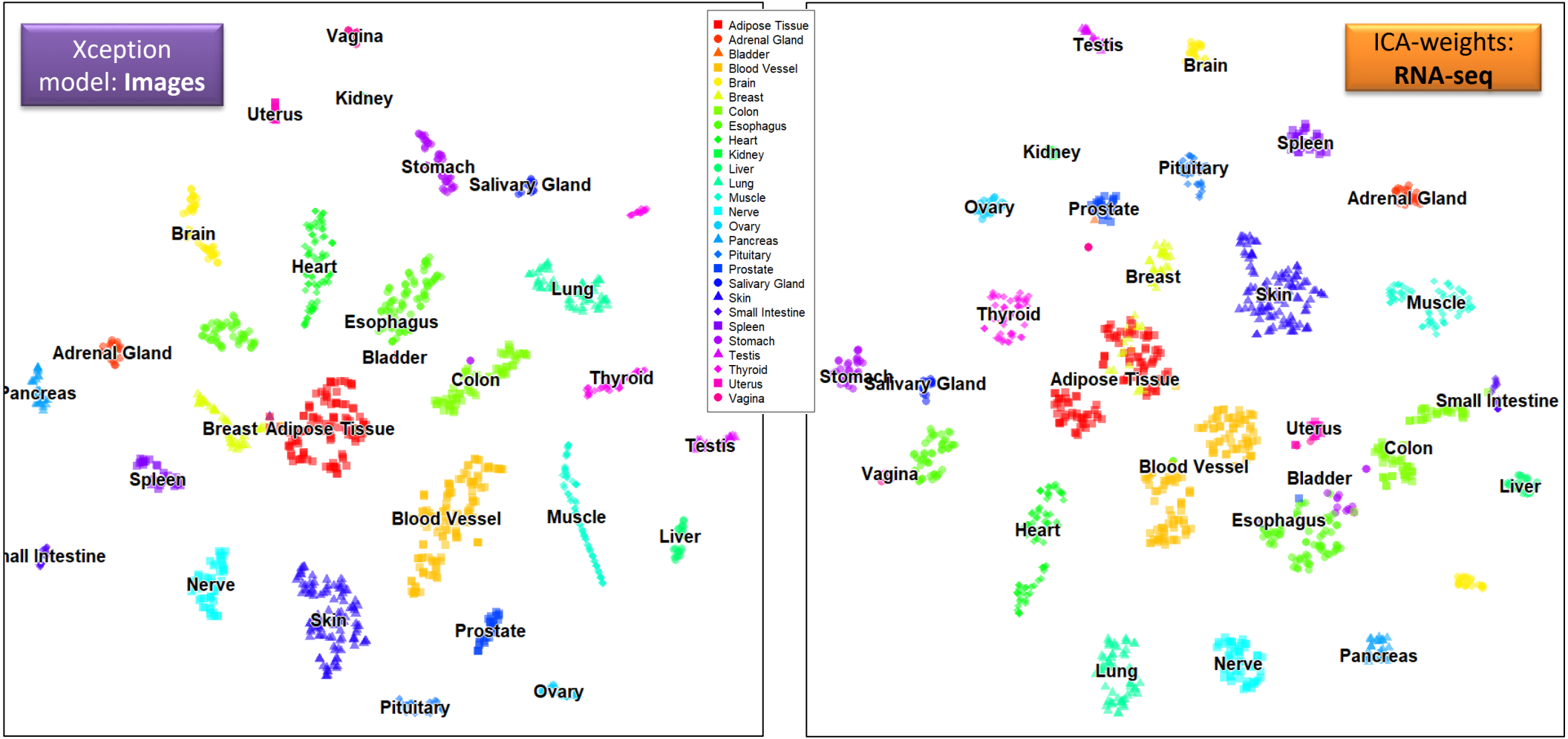
Examples of tiles classified with top certainty and co-localized with class medoids



Xception, after parameter fine-tuning on organ classification task, transform each tile to ~150 non-zero features.

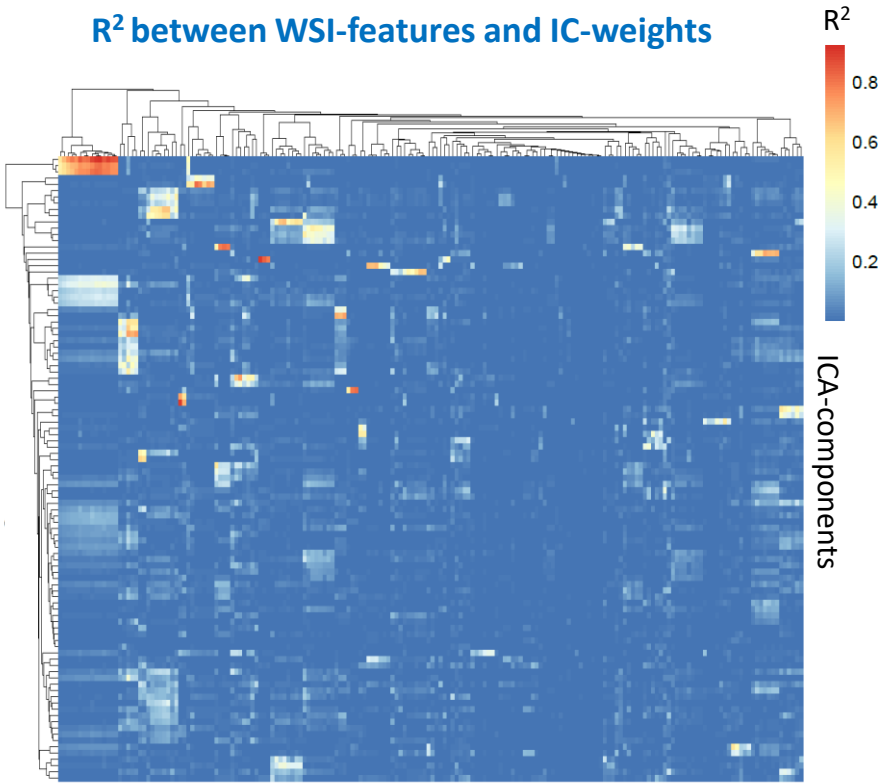
Further analysis:
These features were summarized to slide-level. Only 50% top-correlated tiles were preserved (can be further improved later…)

Slide-level Analysis and ICA

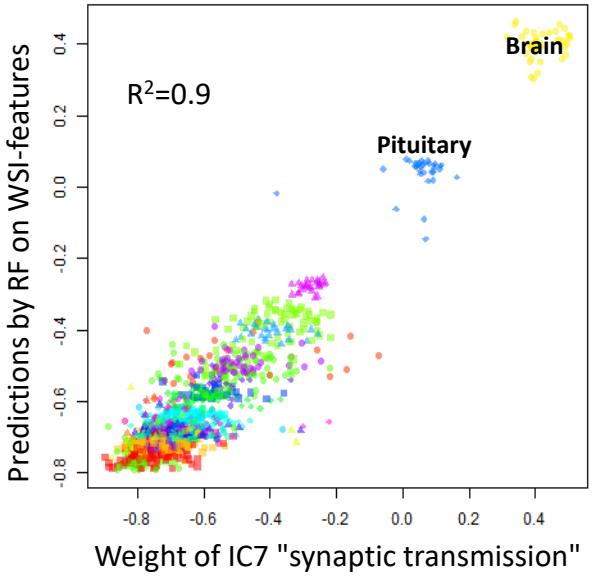


Predictions

R^2 between WSI-features and IC-weights



Predicting IC-weight



GO:BP linked to IC7	FDR
chemical synaptic transmission	8e-28
regulation of membrane potential	8e-28
behavior	4e-22
regulation of ion transport	6e-22
synaptic vesicle cycle	3e-20
cognition	7e-20

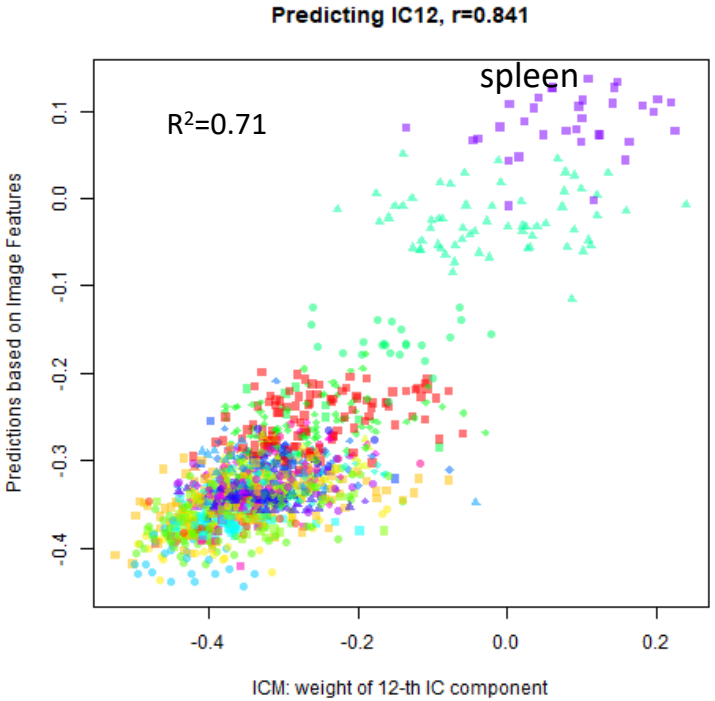
Predicting ICA-components

- 20% of the components were predicted with $R^2>0.9$
- 89% – with $R^2>0.5$

Predicting genes

- 0.4% of the genes showed $R^2>0.9$
- 28% – $R^2>0.5$

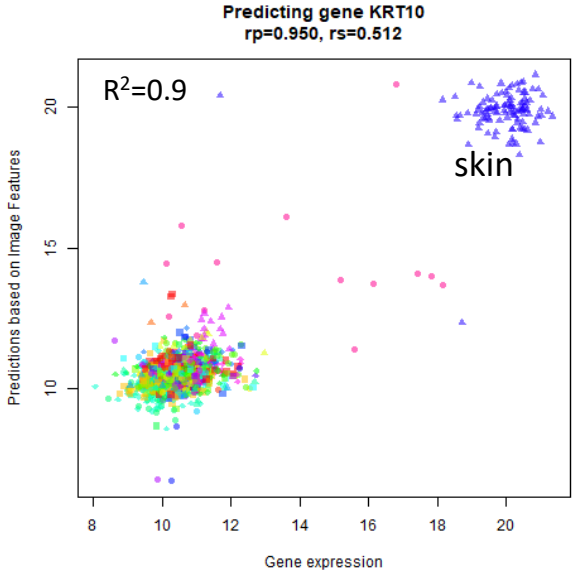
Predicting ICs or Genes?

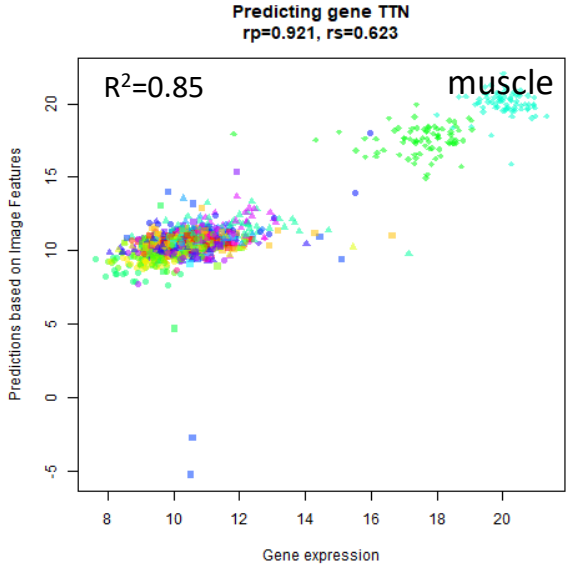


GO:BP pos : 32 terms(FDR<0.01)

Term	FDR
neutrophil degranulation	1.58e−27
defense response to other organism	4.36e−20
immune response	1.48e−15
neutrophil chemotaxis	8.71e−12

Two best-predicted coding genes





Majority of the predicted genes: tissue-specific non-coding

No GO enriched!

Conclusions III: Transcriptomics + Histopathology

- Deep Learning Networks could be used for feature extraction
- Image features could be used to predict deconvolved signals
- Deconvolved ("clean") signals are better predicted than genes (and related GO gene sets)
- Approach was validated on a part of GTEx data