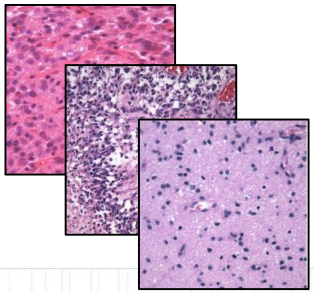
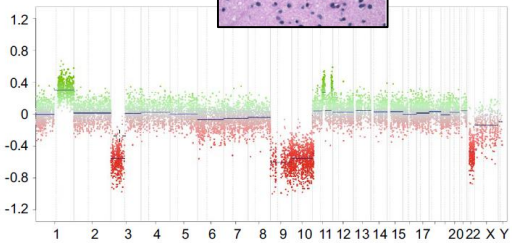
LUXEMBOURG INSTITUTE OF HEALTH

lih.lu

Multi-modal Data Analysis in Cancer Research





Petr V. Nazarov

modas.lu

Open lecture, Department of Mathematical Problems of Control and Cybernetics,

Chernivtsi National University, Ukraine, 2023-03-23

Fonds National de la Recherche Luxembourg

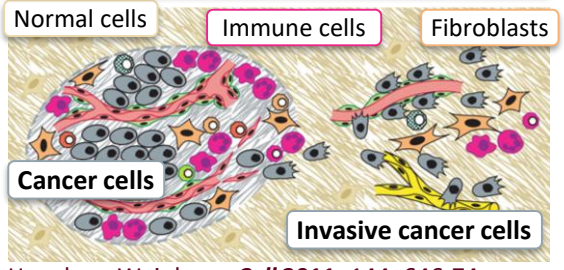
Outline

- **Challenges and Methods**
  - Heterogeneity in Cancer Research
  - Histopathology and molecular methods
  - Data integration
- **Multi-omics data deconvolution and integration**
  - Single omics data deconvolution and integration
  - Multi-omics data deconvolution and integration
- **Multi-modal data integration**
  - Combining histopathology and molecular methods

Challenges and Methods

Levels of Heterogeneity in Samples of Cancer Patients

Native h. of biological tissues



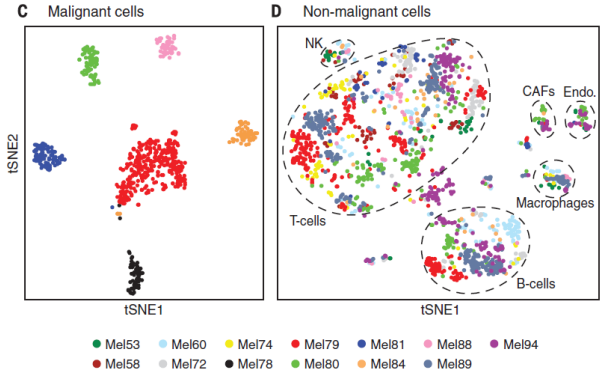
Hanahan, Weinberg. ***Cell* 2011**, 144, 646-74

Cancer sample

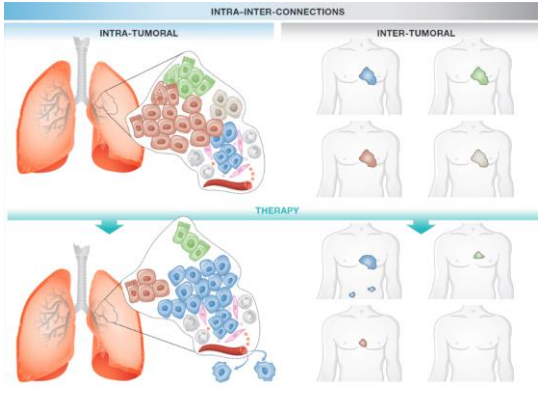
Data

Technical heterogeneity

Inter- and intra-tumor heterogeneity

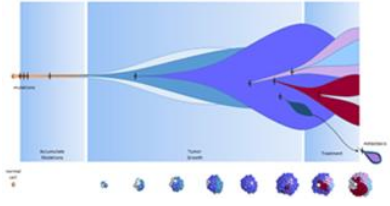


Tirosh, et al. ***Science* 2016**, 352(6282):189

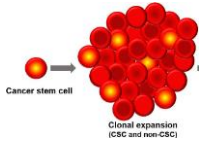


De Sousa E Melo, et al. ***EMBO* 2013**, 14(8):686

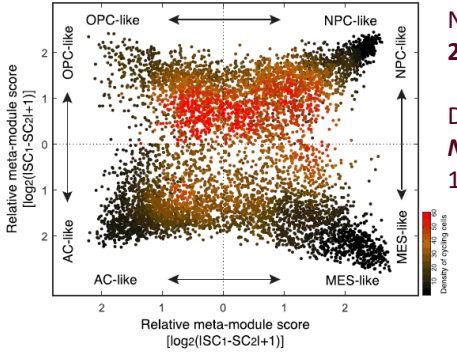
Clonal evolution



Cancer stem-cells



Cell plasticity



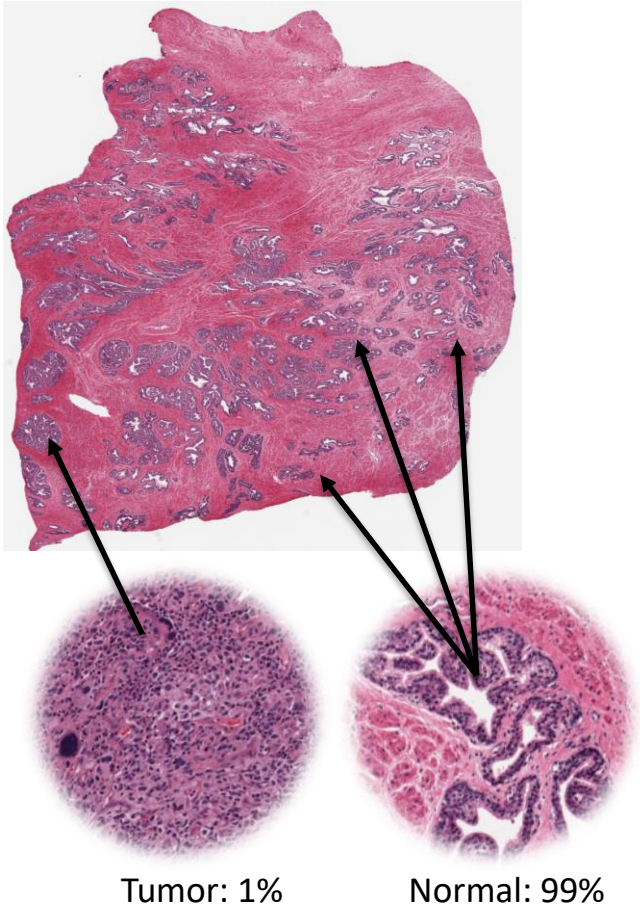
Neftel, et al. ***Cell* 2019**, 178:835

Dirkse, et al. ***Nat Commun* 2019**, 10:1787

Approach 1: Histopathology

LUXEMBOURG INSTITUTE OF HEALTH

Hematoxylin and Eosin (H&E) stain



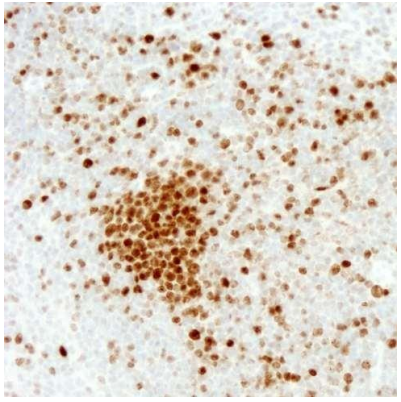
Features of histopathology

- Gold standard!
- Cheap (H&E or 2-3 antibodies in IHC)
- Captures native heterogeneity of tissues
- Shows inter/intra tumor heterogeneity
- Often allows precise diagnostics

Issues in histopathological image analysis:

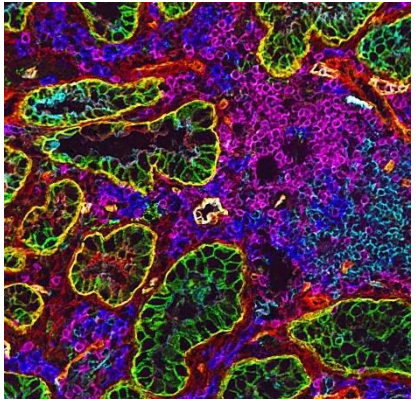
- Tedious analysis
- In some cancers (e.g. prostate) < 1% of the image is cancer-related
- For some cancers, it does not allow precise diagnostics (e.g. some astrocytomas vs oligodendrogliomas)
- Gives non-structured data
- Invasive

Immunohistochemistry (IHC)



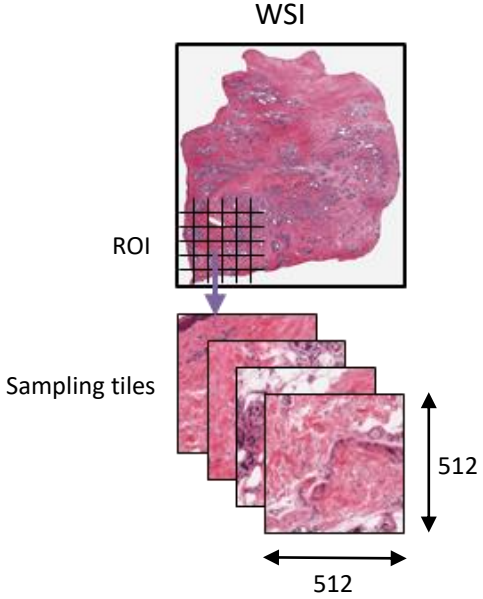
Ki-67 - proliferation marker

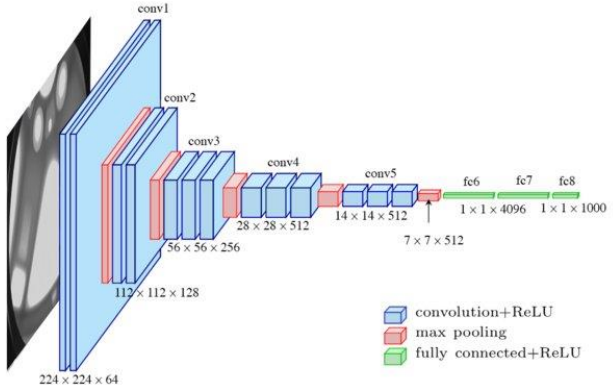
Multicolor IHC



Approach 1: Histopathology

Deep Learning for Tumor Identification / Classification



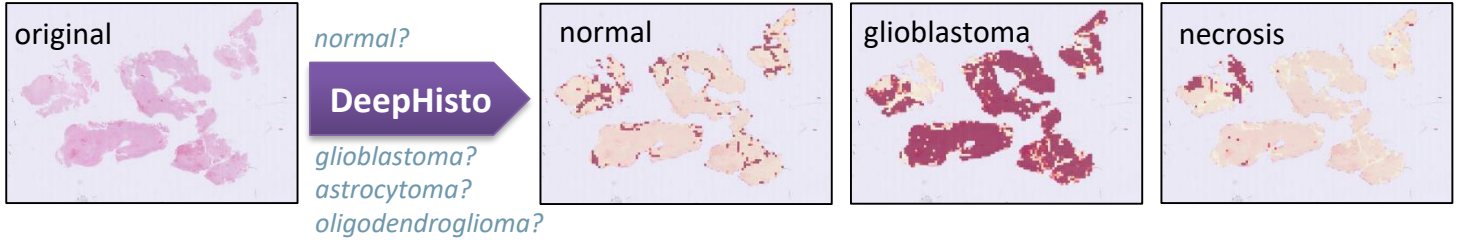


classes:

- Astrocytoma
- Oligodendroglioma
- Glioblastoma
- Normal
- Necrosis

First practical outcome:

DeepHisto tool for automatic detection/classification of gliomas (LIH)



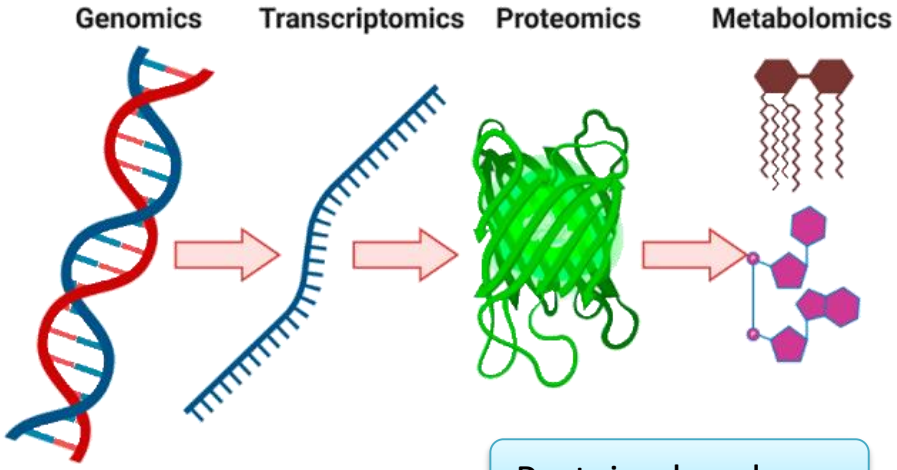
Approach 2: Molecular Profiling

Mutations

Copy-number variation (CNV)

Chimeric genes

Methylation



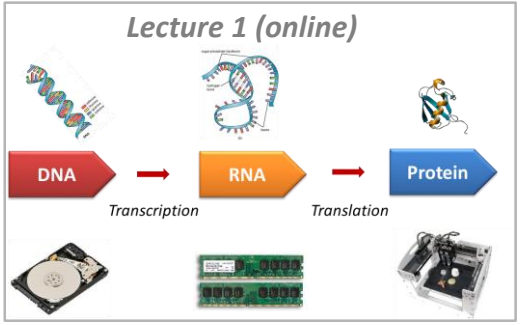
Gene expression

Protein abundance

Chimeric proteins

Isoform detection

Abundance of metabolites



Features of molecular approach

- Very specific
- Generate a lot of data
- Generate structured data

Issues of molecular approach

- Quite expensive
- Is sensitive to heterogeneity of samples
- Is sensitive to a technique

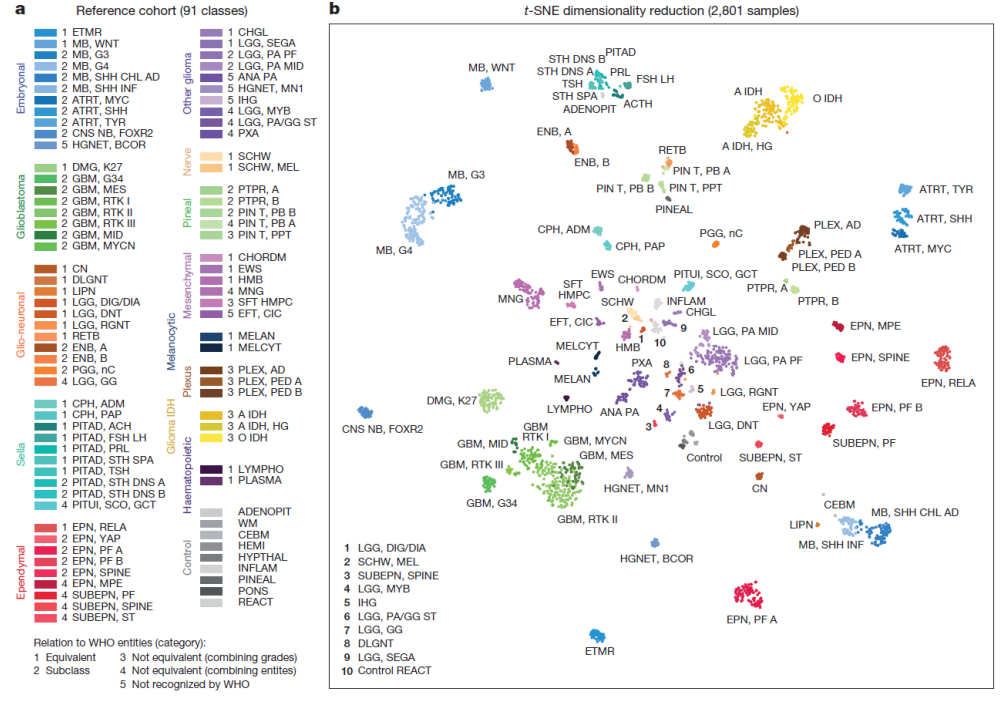
Image: Raja et al https://www.mdpi.com/1422-0067/22/3/1160/htm

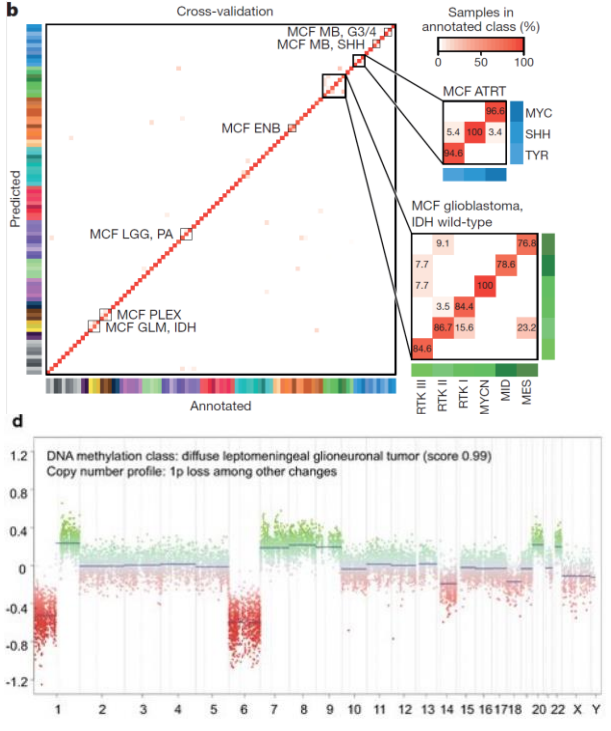
Approach 2: Molecular Profiling

Heidelberg Brain Tumor Classifier

Capper et al. ***Nature* 2018**, 555(7697):469
Capper et al. ***Acta Neuropathologica* 2018**, 136:181

DNA methylation–based classification of central nervous system tumours

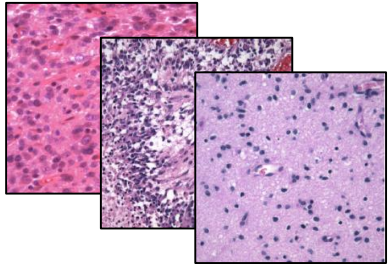




- Methylation showed more specificity than histopathology identifying types of brain tumors
- A highly standardized pipeline allowed analysis across many cohorts worldwide
- **Result**: "Heidelberg classifier" is used by pathologists ☺

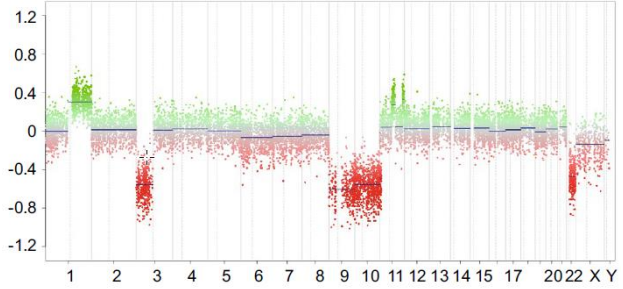
Improvements via Integration

1. Histopathology



- Automate analysis
- Transform unstructured data (images) to structured (features)

2. Molecular methods



- Deconvolute mixed signals
- Integrate various molecular data

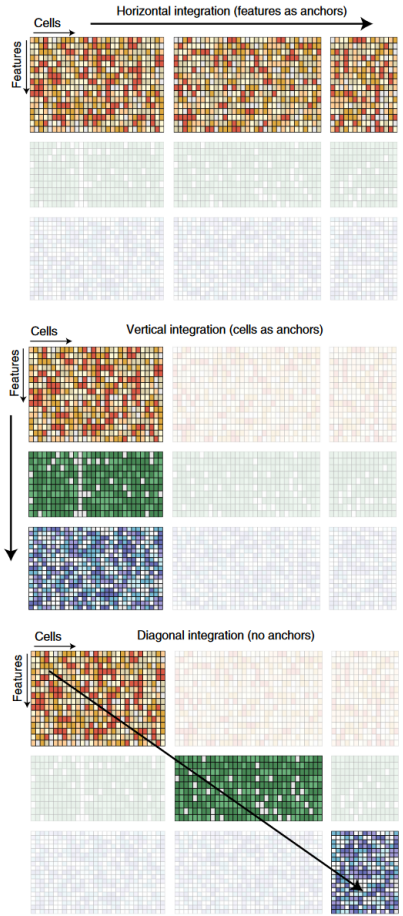
Integrate both approaches for better patient diagnostics and studying molecular processes

- ~~Tedious analysis~~
- ~~< 1% of the image is cancer-related~~
- ~~For some cancers, it does not allow precise diagnostics~~
- ~~Gives non-structured data~~

- ~~Quite expensive~~
- ~~Is sensitive to the heterogeneity of samples~~
- ~~Is sensitive to a technique~~

Multi-modal Data Integration

Data integration tasks



Horizontal integration

- Batch correction
- Normalization
- ANOVA

Vertical integration:

- Correlation analysis
- Canonical correlation analysis
- Matrix factorization

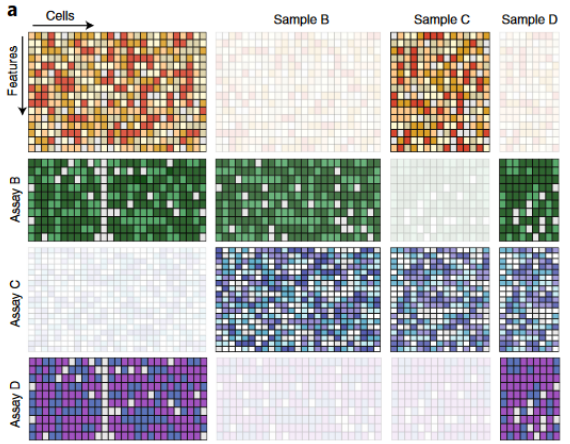
Diagonal integration:

- Latent manifold (*многовид / многообразие*)
- Simplify to H. or V. by labelling similar subsets
- Use deep-learning (e.g. variational autoencoders)

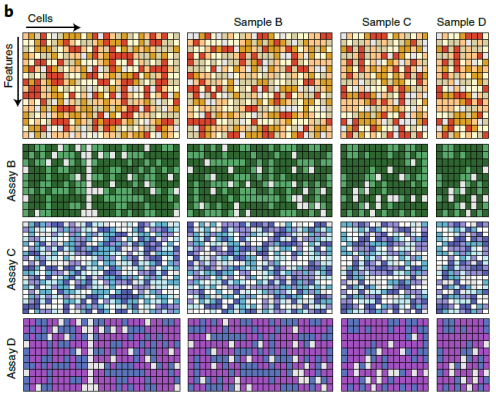
Table 1 | Overview of common data integration methods classified according to their anchor choice

Integration task	Method
Vertical (global)	CCA
Vertical (global)	JIVE
Vertical (global)	PLS
Vertical (global)	MCIA
Vertical (global)	MOFA+
Vertical (global)	scAI
Vertical (global)	iNMF
Vertical (global)	Seurat v4
Vertical (local)	Spearman's rank correlation coefficient
Vertical (local)	LMM
Horizontal	MNN
Horizontal	Seurat v3
Horizontal	LIGER
Horizontal	Harmony
Horizontal	Scanorama
Horizontal	BBKNN
Horizontal	scVI
Horizontal	scmap
Horizontal	conos
Diagonal	MATCHER
Diagonal	MMD-MMA
Diagonal	SCIM
Diagonal	UnionCom
Diagonal	coupledNMF

Mosaic integration:

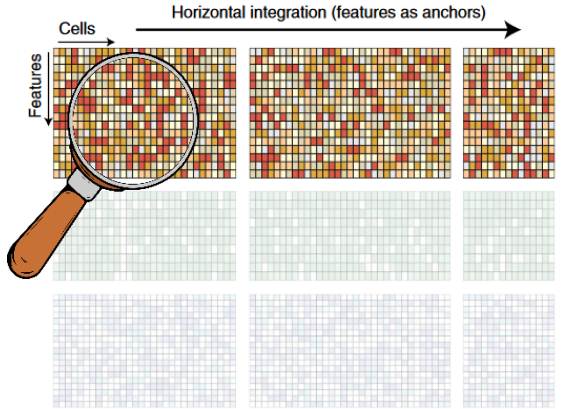


imputation | deep learning?

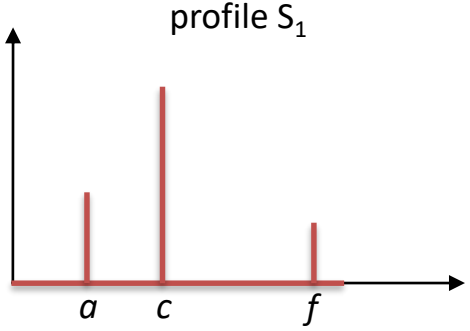


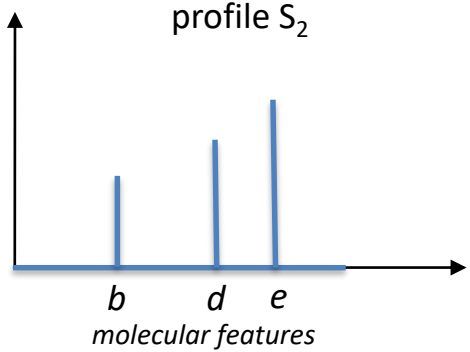
Argelaguet et al. ***Nature Biotechnology* 2021**, 39, 1202

Deconvolution



Deconvolution: Concept





Mixing

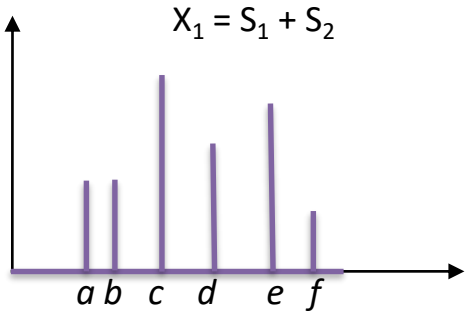
X = S x M

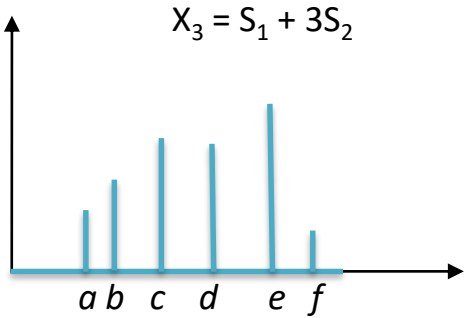
M =

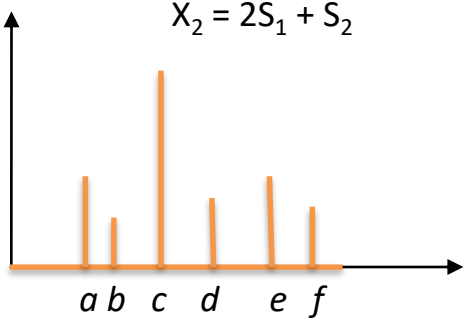
1	2	1
1	1	3

De-mixing

Often called:
- **decomposition**
- **deconvolution**

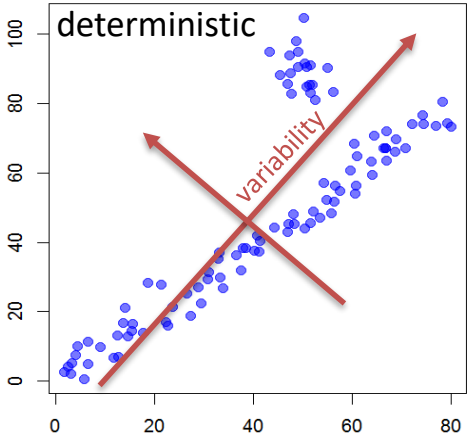






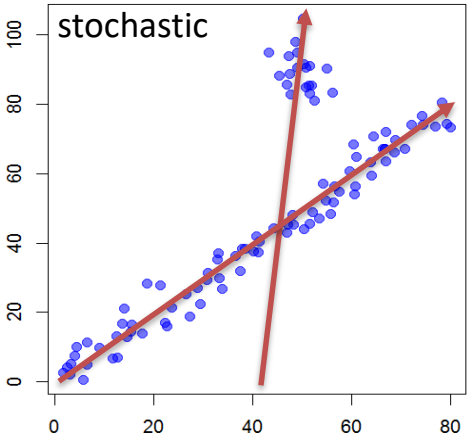
Methods

PCA



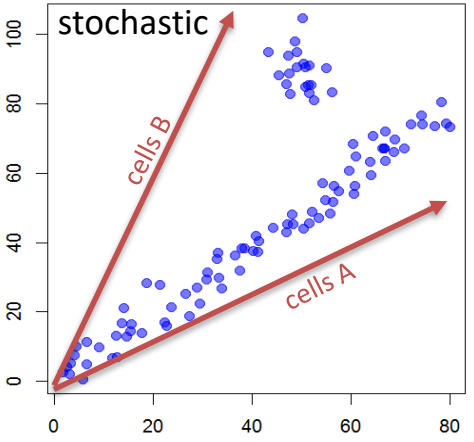
+ deterministic & fast
+ any number of samples
+ unsupervised
– often biological factors are presented by a sum of several components
– positive and negative values

ICA



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

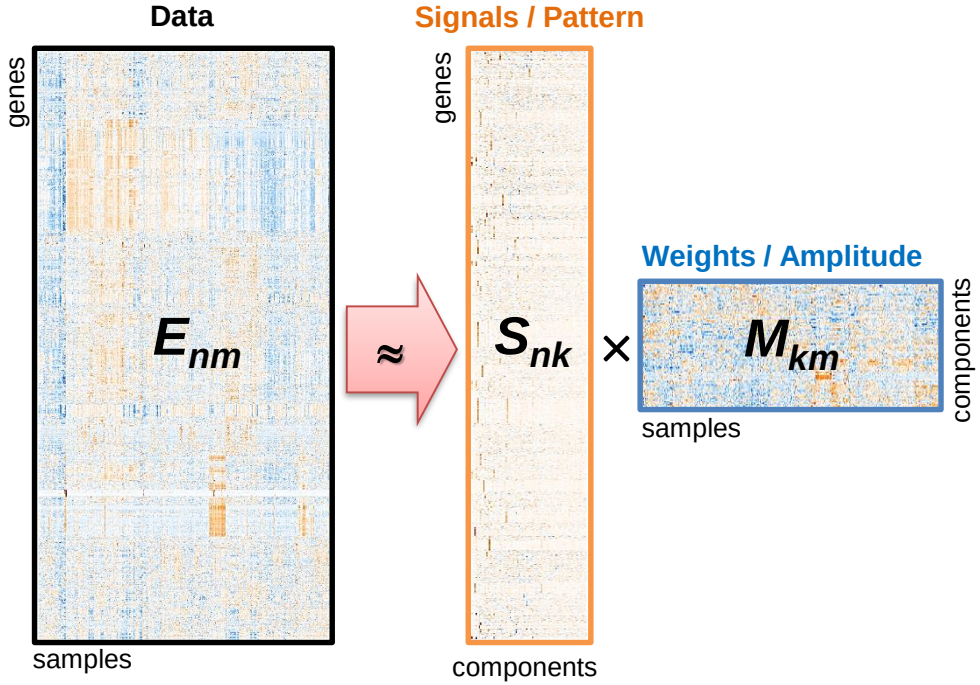
NMF



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

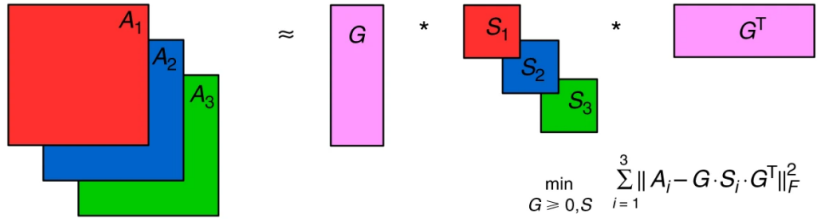
Sompairac et al, Int J Mol Sci, 2019 (link)
Cantini el al, Bioinformatics, 2019 (link)

Deconvolution via Matrix Factorization



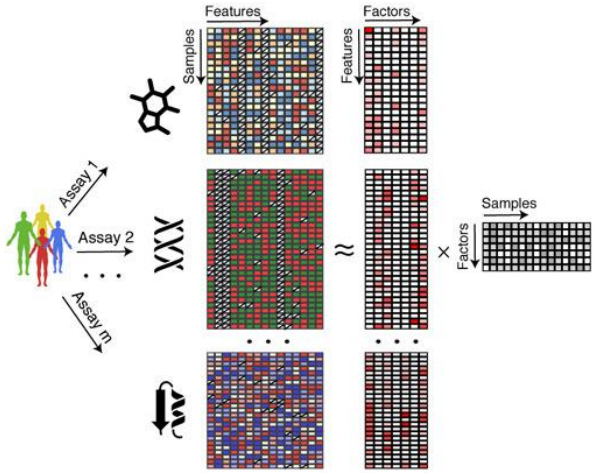
PCA: principal component analysis
NMF: non-negative matrix factorization
ICA: independent component analysis
etc.

Matrix tri-factorization



Malod-Dognin et al. ***Nat Commun* 2019**, 10:805

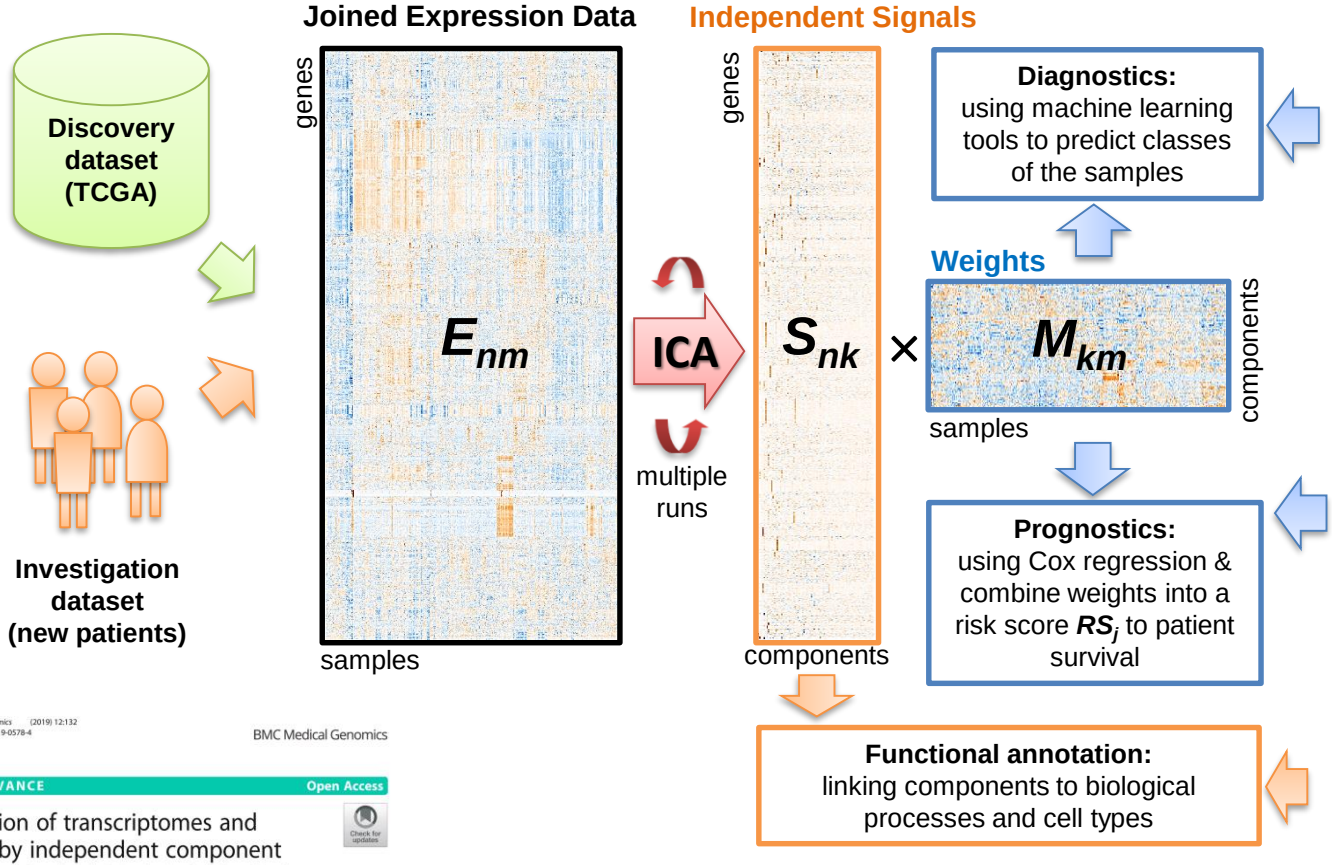
Multi-omics Factor Analysis



Argelaguet et al. ***Mol Syst Biol* 2018**, 14:e8124

Stein-O'Brien et al. ***Trends in Genetics* 2018**, 34(10):790

Research Focus: Deconvolution of Omics Data




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[2], Gunnar Dittmar[1], Francisco Azuaje[1] and Stephanie Kreis[2]

consICA: Nazarov et al **BMC Medical Genomics**, 2019 (link)
ICA review: Sompairac, et al **Int J Mol Sci**, 2019 (link)
Application: Golebiewska et al**, Acta Neuropathol**, 2020
Scherer, Nazarov et al, **Nat Protoc**, 2020

Deconvolution for Horizontal Integration: GBM

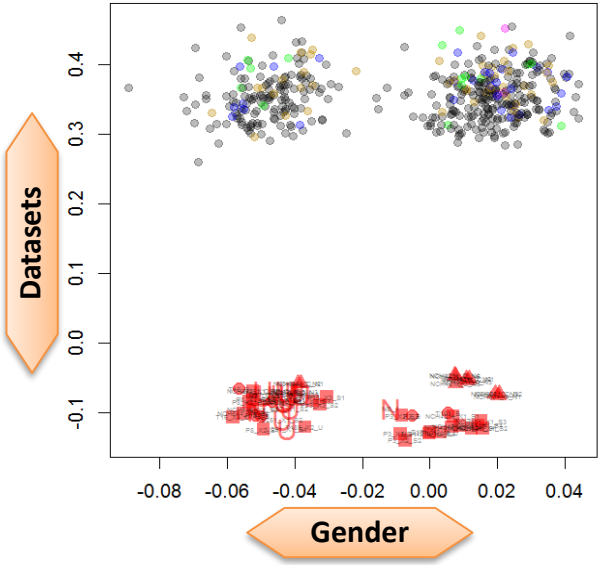
Reference dataset
530 GBM patients
(TCGA)

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

ICA

Biological knowledge:
bio-processes and sample composition

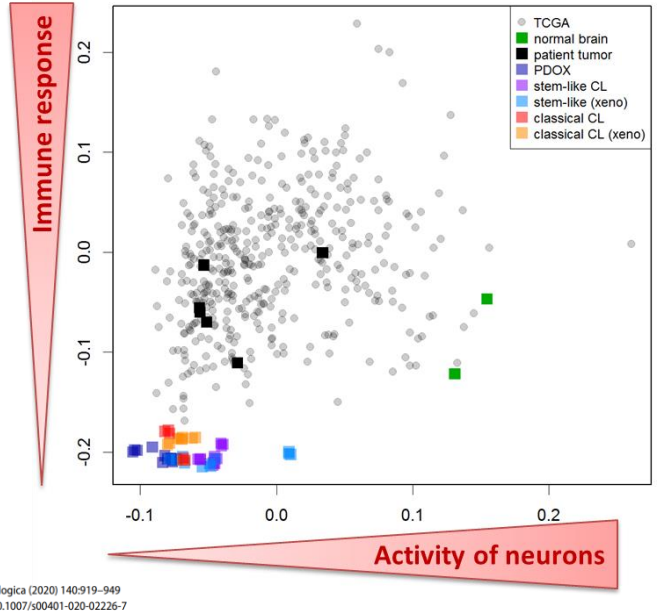
Technical/trivial components: gender and platforms



- 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.

Deconvolution for Horizontal Integration: GBM

ICA correctly predicts sample composition & phenotype

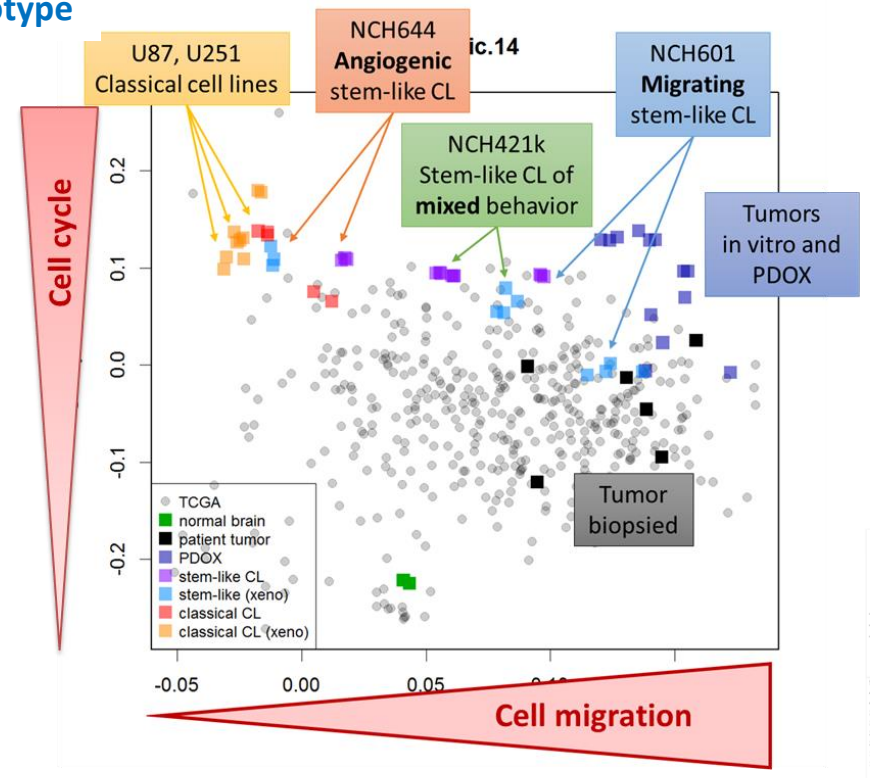


Acta Neuropathologica (2020) 140:919–949
https://doi.org/10.1007/s00401-020-02226-7

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]

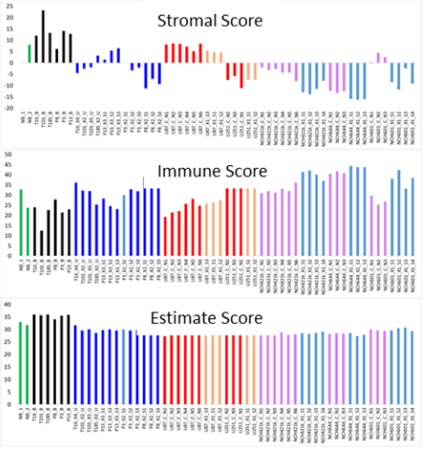


Golebiewska A. et al, **Acta Neuropathologica, 2020** (link)

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



Deconvolution for Horizontal Integration: PDAC

TCGA
183 samples
mRNA, miR, methylation

Bailey et al
96 samples
mRNA

DKFZ
457 samples (268 tumors)
mRNA, miR

Reference dataset
large public dataset (TCGA)

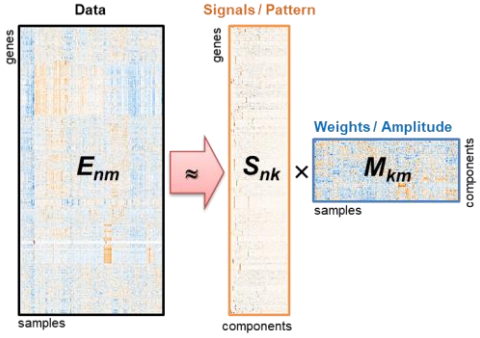
Investigated dataset:
(new clinical samples)

ICA + **Machine learning, Cox regression**

Biological knowledge:
bio-processes and sample composition

Potential diagnostics:
sample classification

Potential prognostics:
hazard score



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

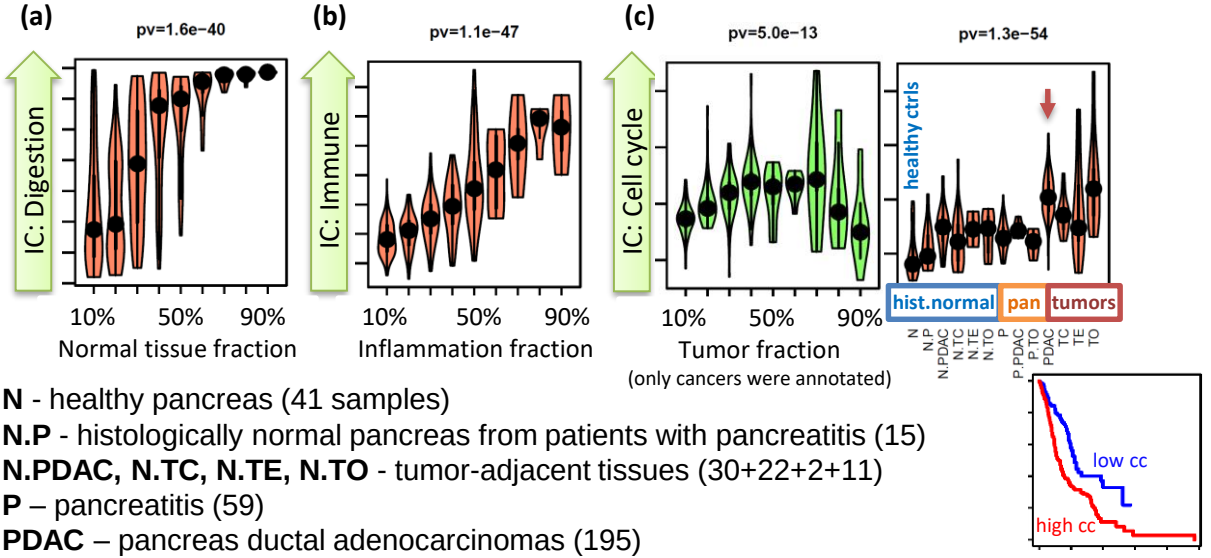
j – patient index
i – component index
R^2_i – stability of *i*-th component (from 0 to 1)
H_i – Cox' log hazard ratio calculated on **training set**
$M^*_{i,j}$ – element of centered & scaled M-matrix

Nazarov et al, BMC Medical Genomics, 2019

➢ In addition to diagnostics and prognostics, ICA allowed ranking patients based on the activity of biological processes

Deconvolution for Horizontal Integration: PDAC

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.

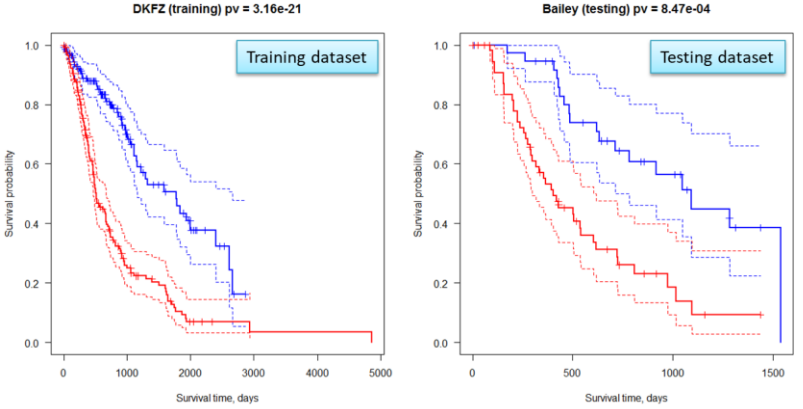
Increased risk:
- **keratinization**
- **cell cycle**
- response to hypoxia
- neoangiogenesis
- activation of ERK-signaling

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.

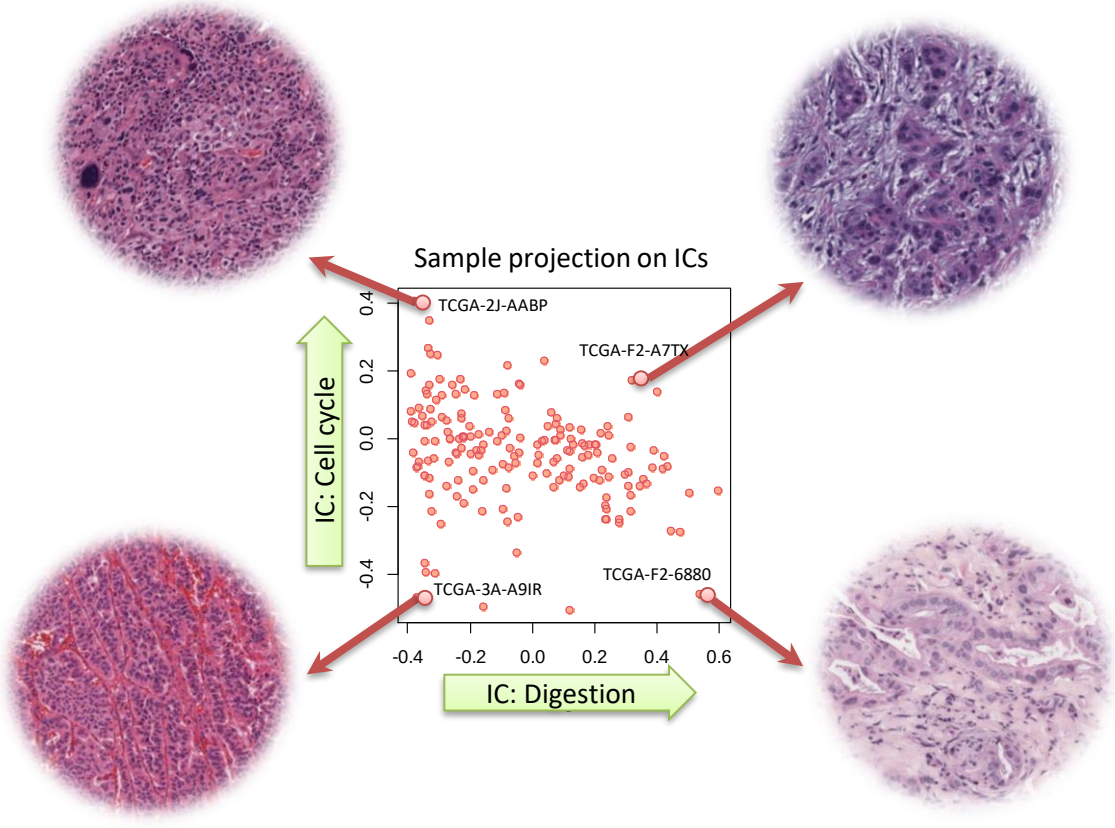
Prognostic markers between 2 cohorts



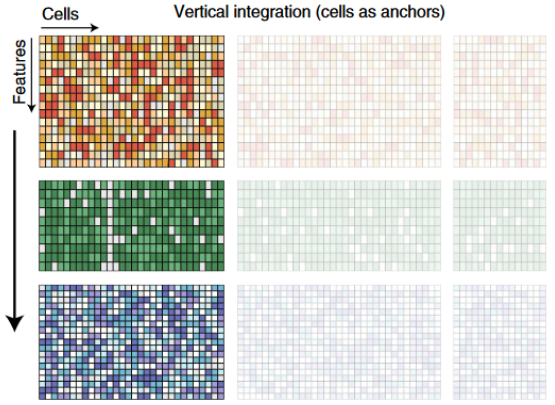
Acc: 0.83	N	N.PDAC	P	PDAC
pred.N	32	2.6	1.8	2
pred.N.PDAC	0.6	1.7	2	1.6
pred.P	4.7	17.3	51.8	5.4
pred.PDAC	3.7	8.4	3.4	186

Observation

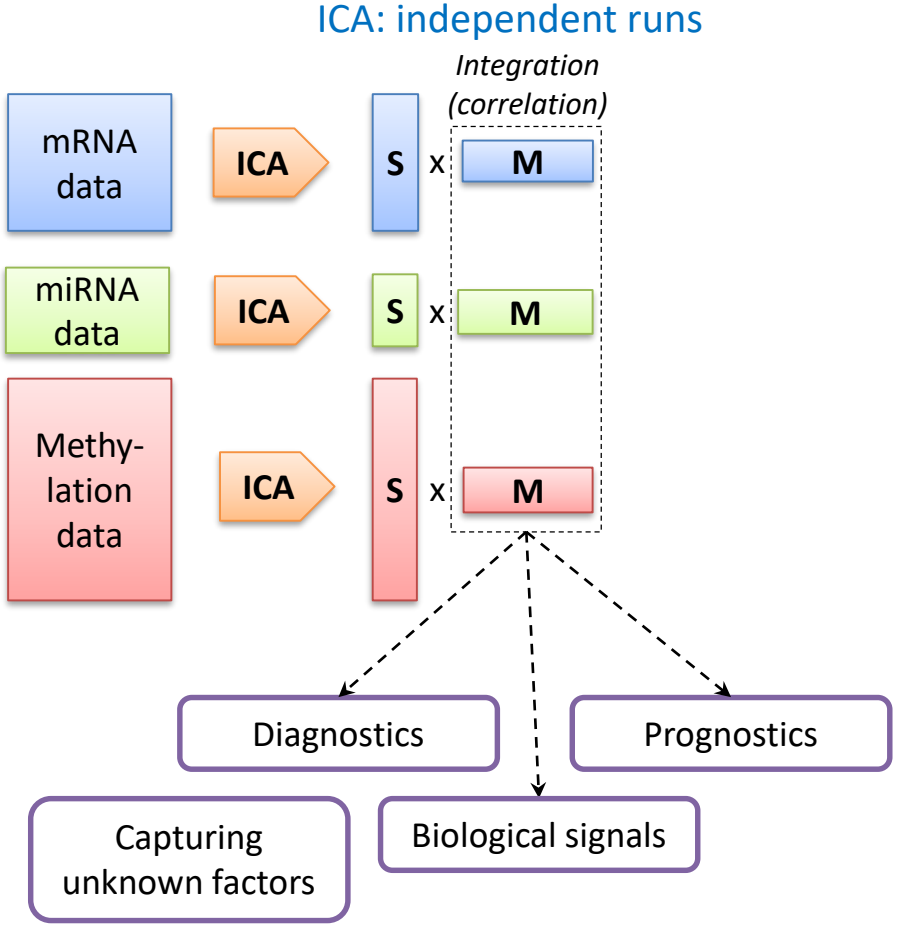
ICA results of mRNA expression data from TCGA-PAAD cohort

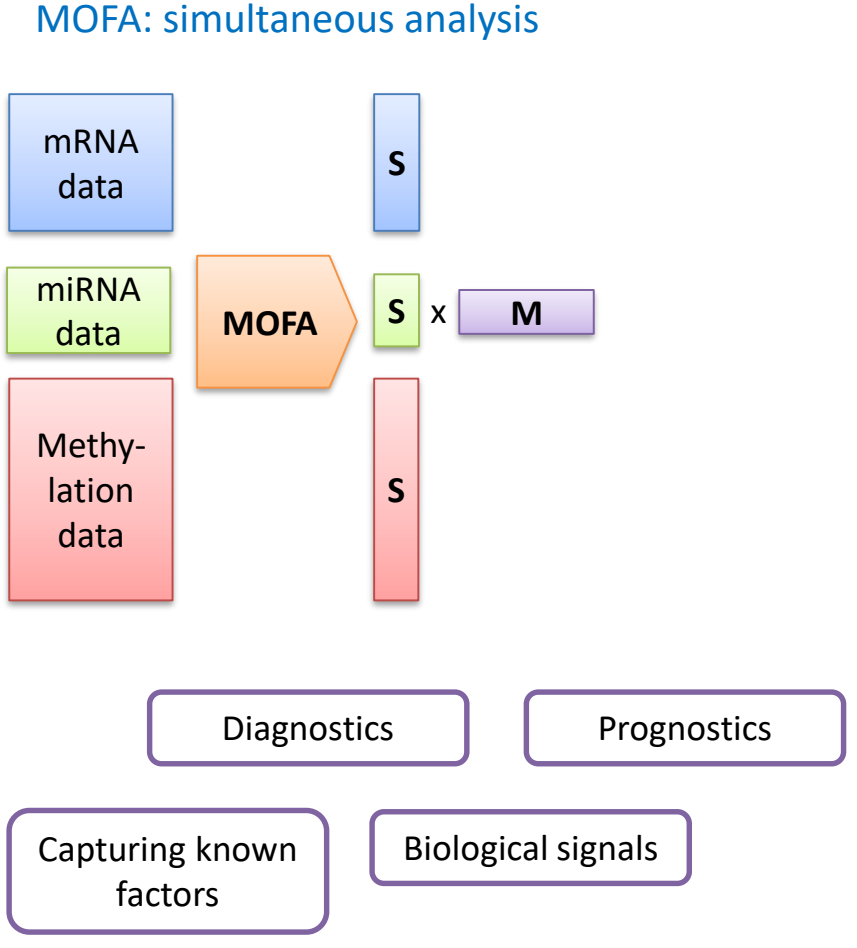


Integration
(multi-omics)



Multi-omics Data Integration via Deconvolution





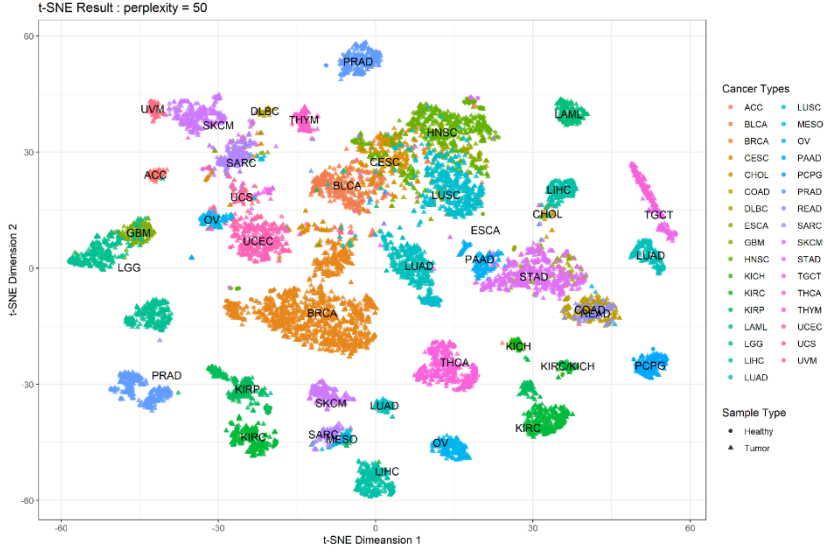
Pan-Cancer Data Integration

TCGA

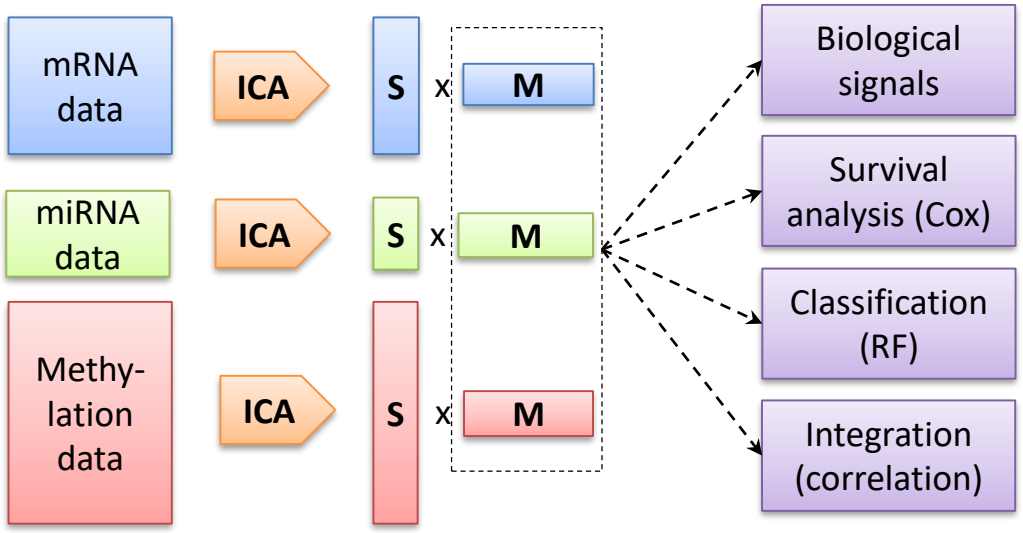
The Cancer Genome Atlas

>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)



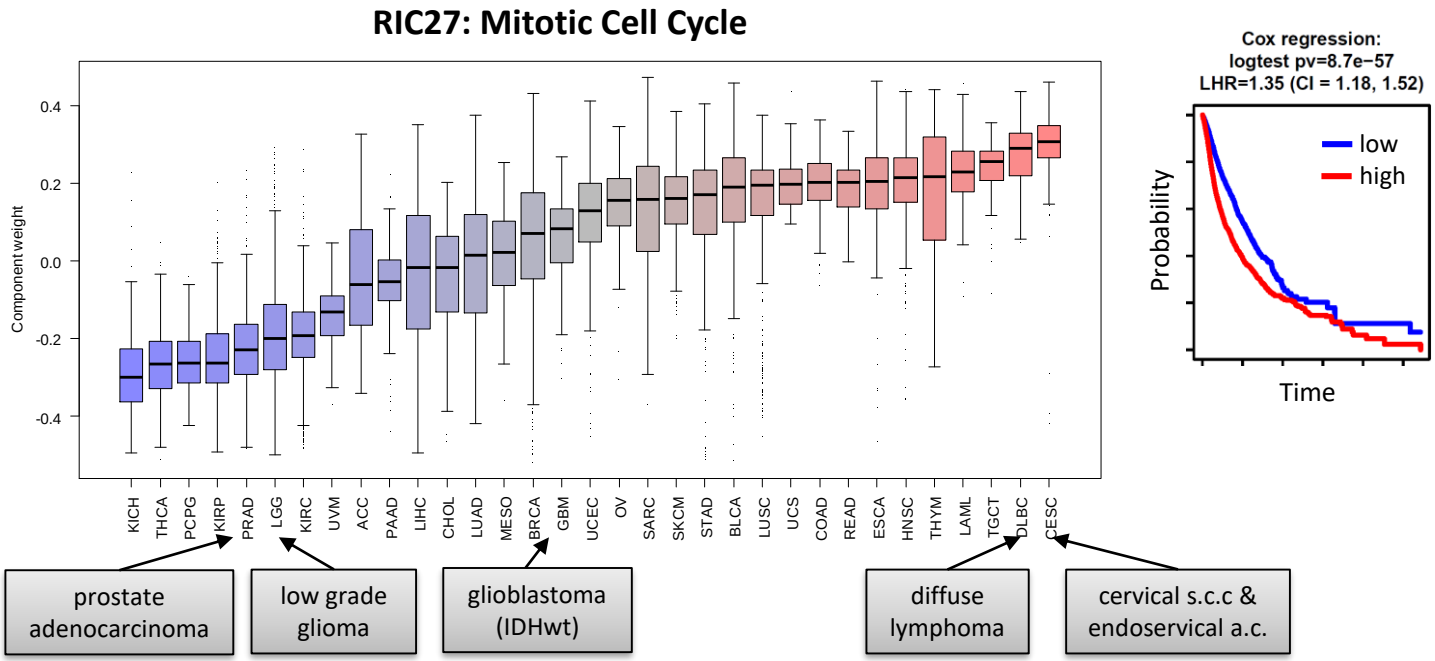
Approach



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

Pan-cancer: ICA Components (unimodal)

ICA Results: Cell Cycle

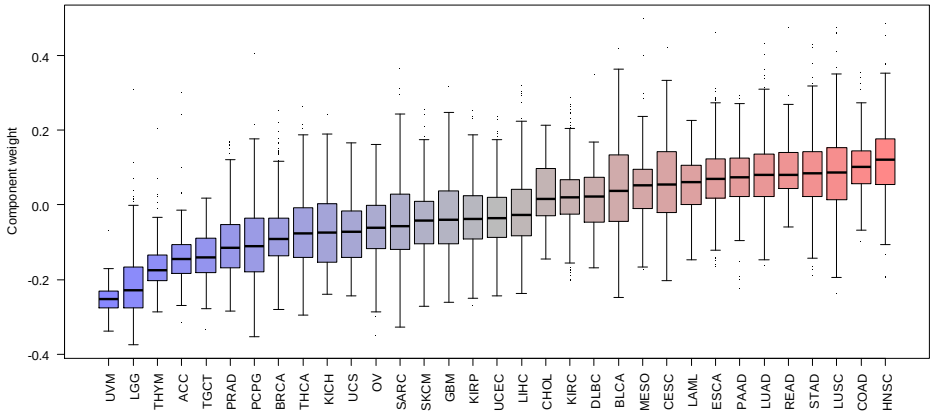


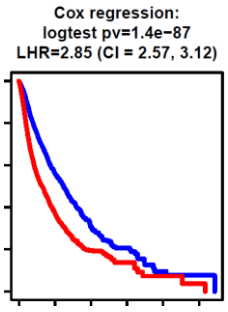
Code	Study Name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical sq. cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large b-cell lymphoma
ESCA	Esophageal carcinoma
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 (unimodal)

LUXEMBOURG INSTITUTE OF HEALTH

RIC17: Signal of Mast Cells*

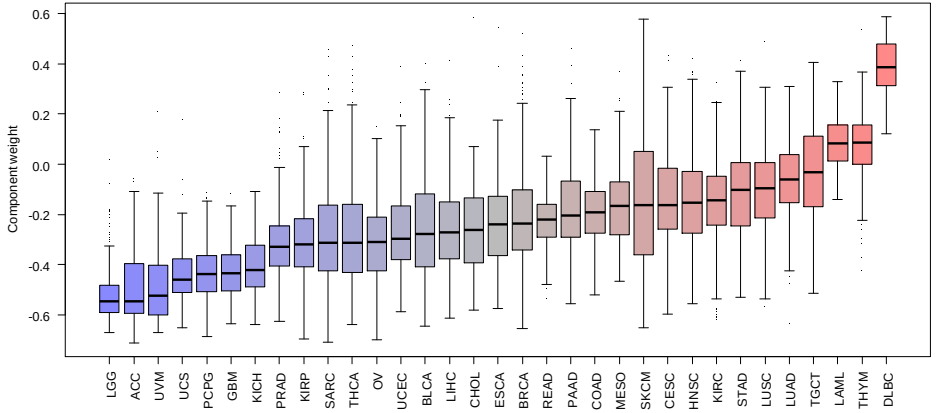




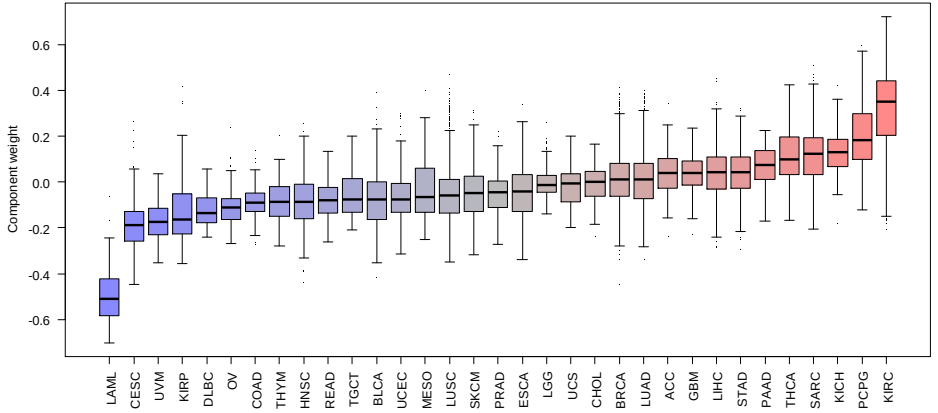
Tumor-associated mast cells (TAMCs) ?☺?

(*) assigned based on LM22 signature (CIBERSORT)

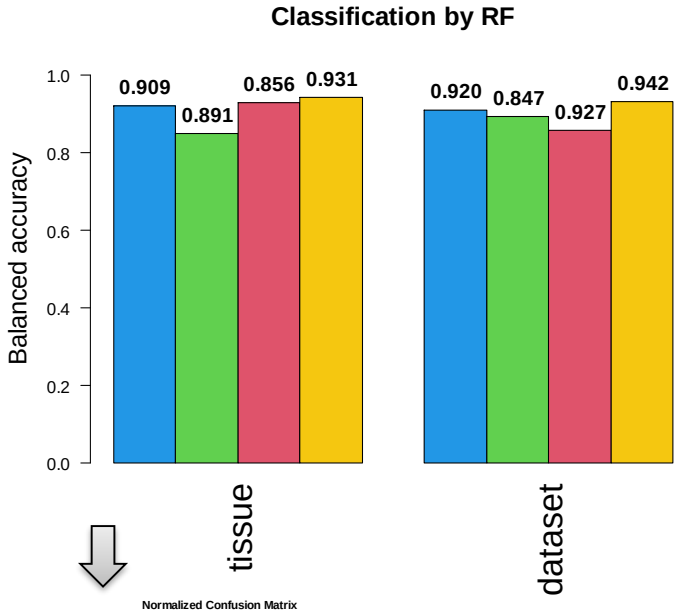
RIC16: Signal of T-Cells*

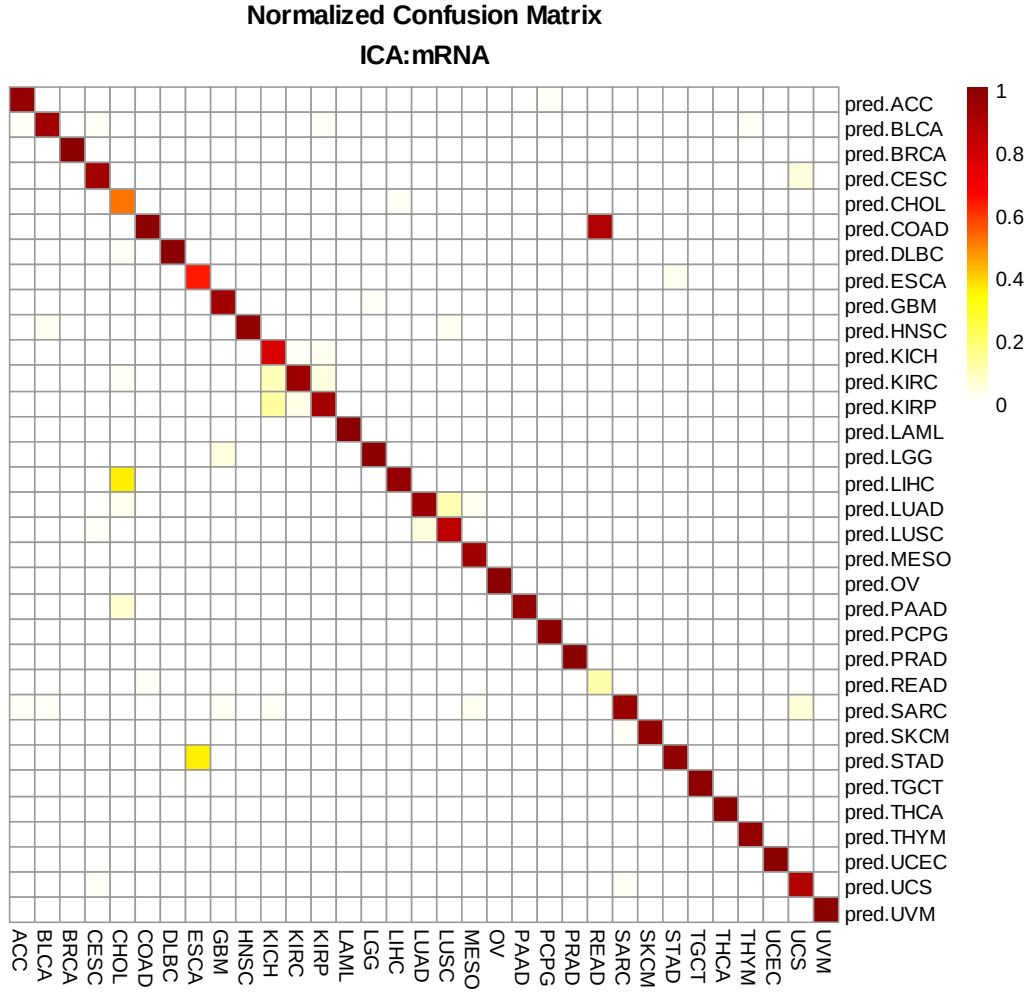


RIC57: Angiogenesis



Pan-cancer: Classification (unimodal)





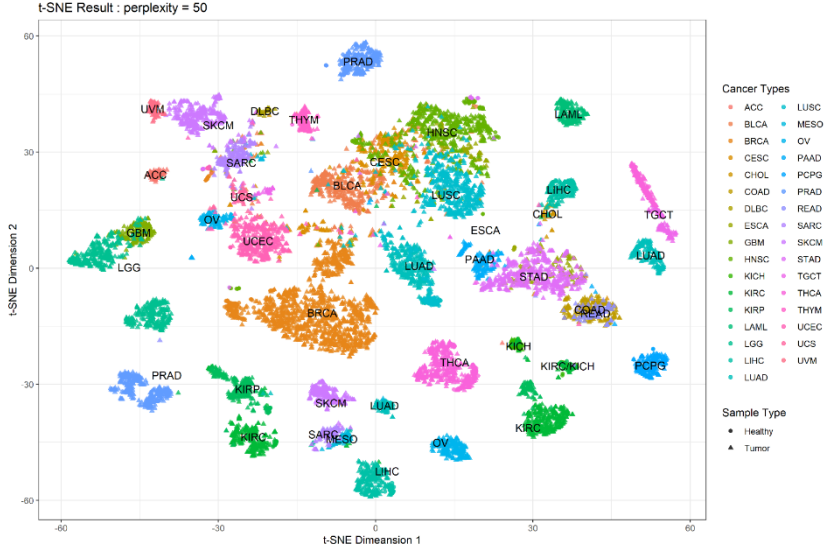
Pan-Cancer Data Integration

TCGA

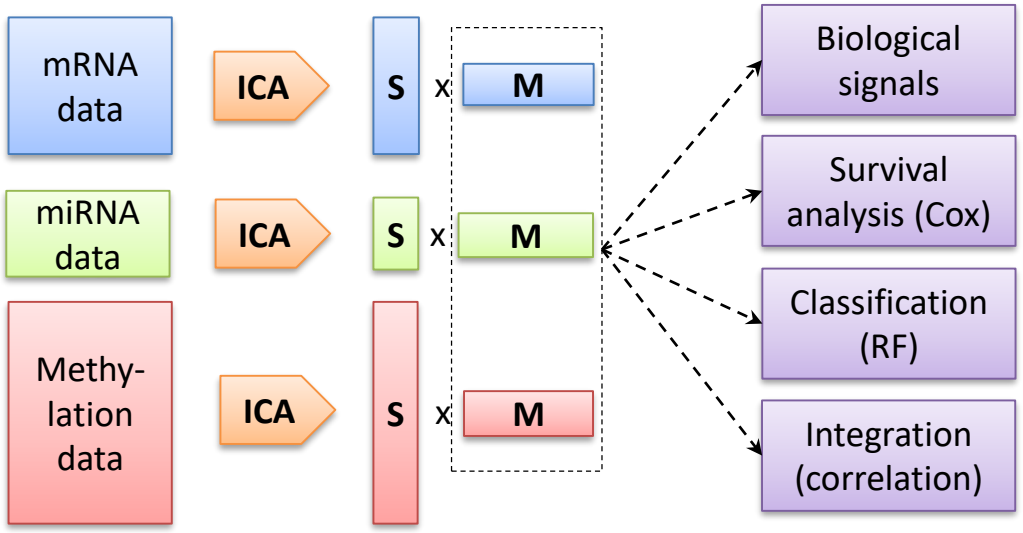
The Cancer Genome Atlas

>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)

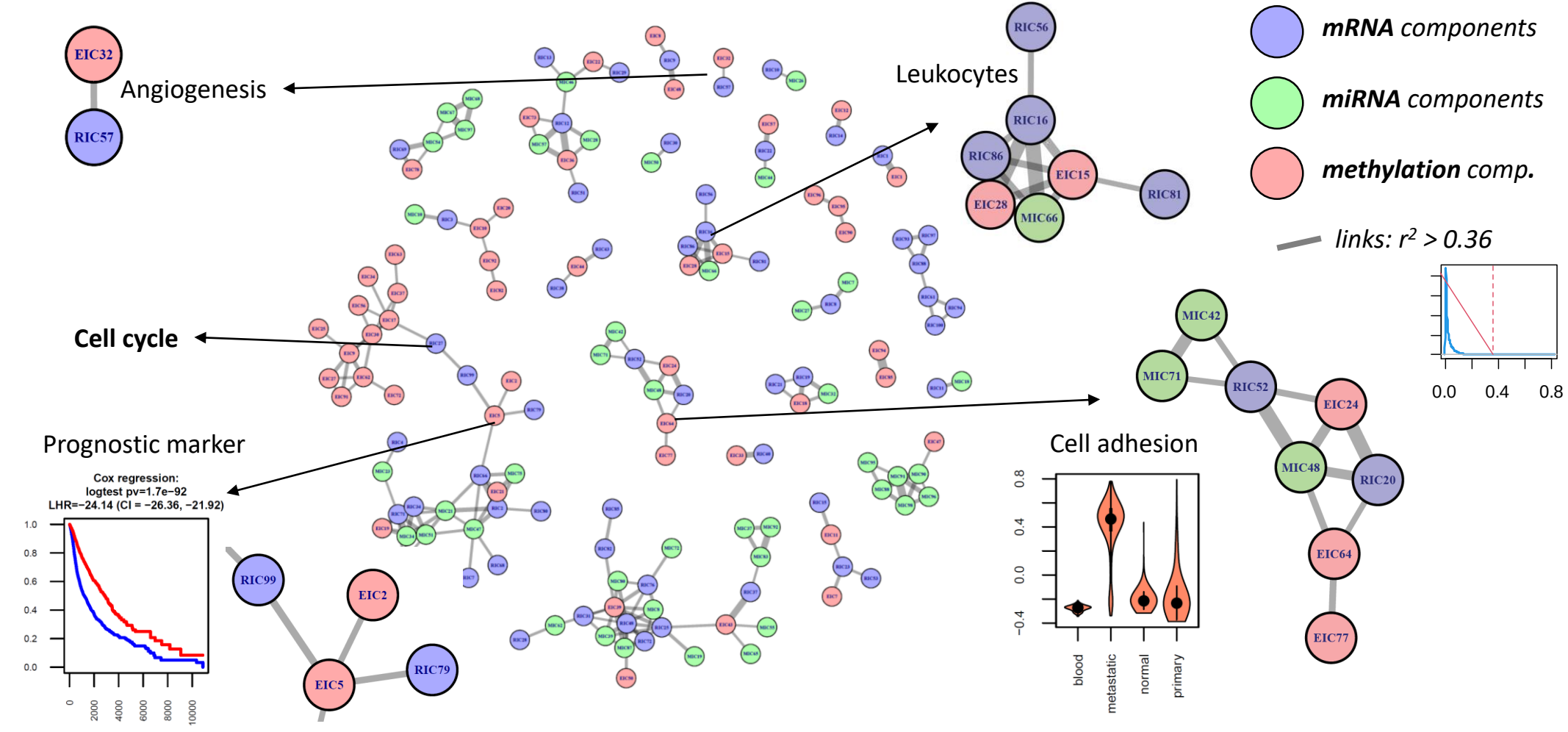


Approach



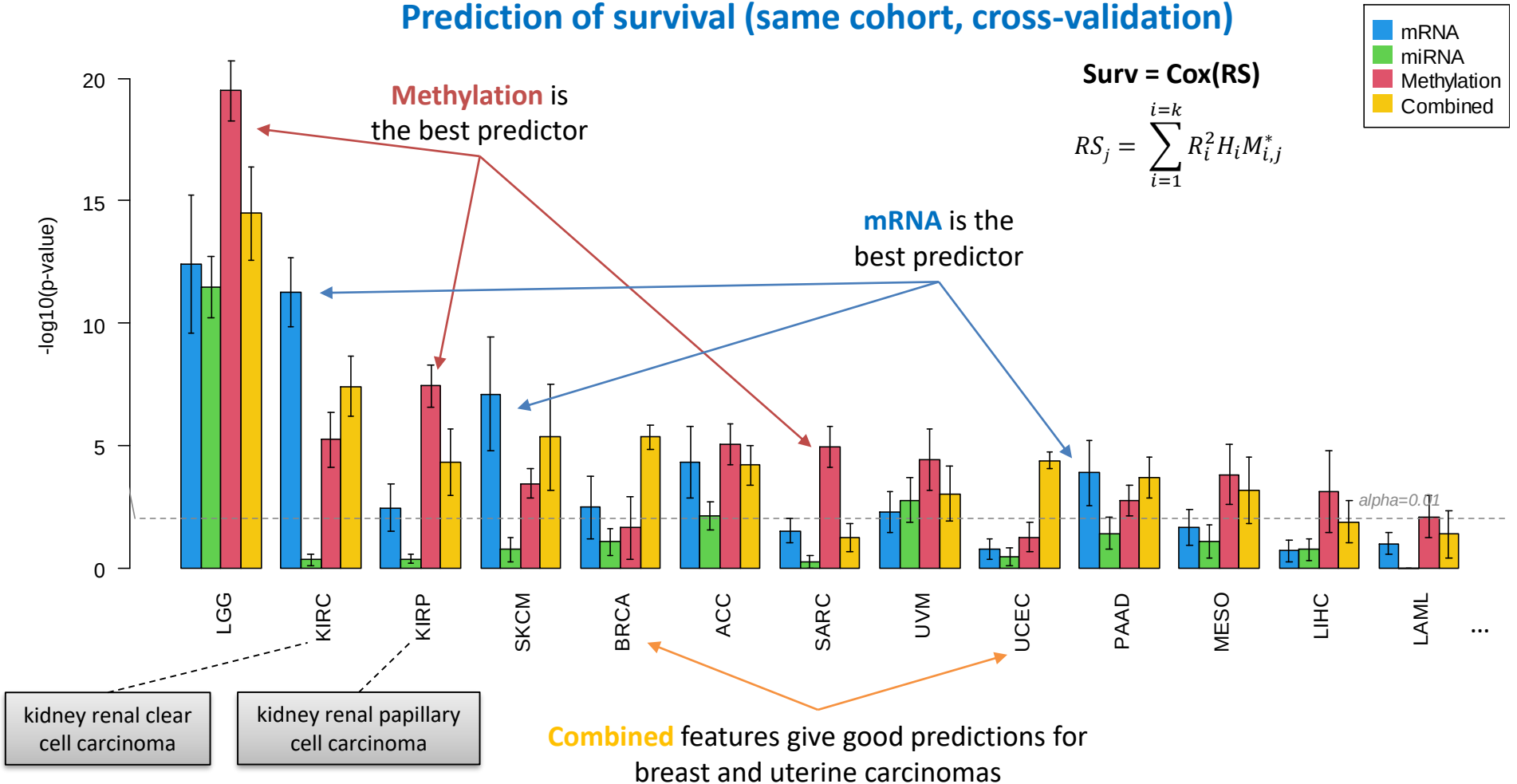
Here we used *consICA* with 100 components & 40 runs

Pan-cancer: ICA-based Data Integration



Pan-cancer: Prognosis

Prediction of survival (same cohort, cross-validation)



Take Home Message 1

- ICA-based deconvolution:
  - 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
  - Reduce dimensionality

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

Integration
(multi-modal)

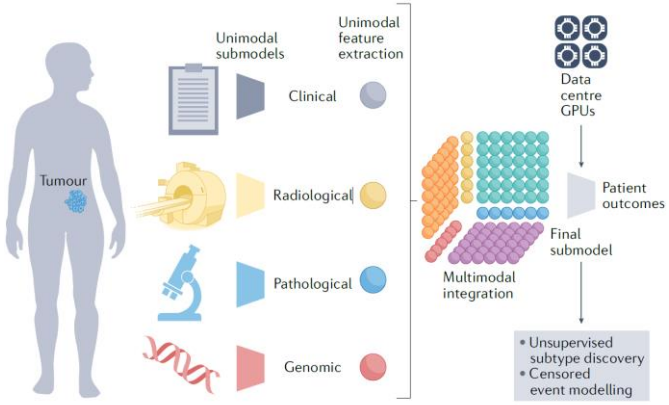


Fig. 2 | **Multimodal models integrate features across modalities.** Submodels extract unimodal features from each data modality. Next, a multimodal integration step generates intermodal features — a tensor fusion network (TFN) is indicated here[86]. A final submodel infers patient outcomes. GPU, graphics processing unit.

Boehm, et al. ***Nature Reviews Cancer*** **2021**, 22, 114-126

Digital Histopathology and Feature Extraction

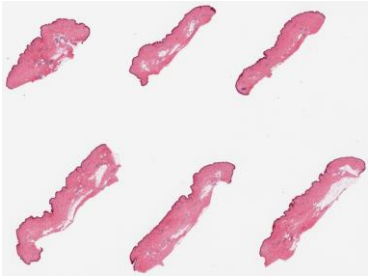
How can we work with unstructured data (images)? Extract features!

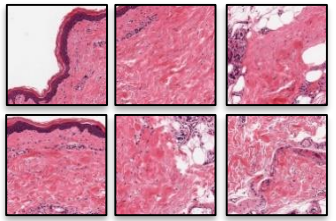
The Task

N slides:
10^4 x 10^5 pixels

N x M tiles / patches:
256 x 256 pixels

1 patient





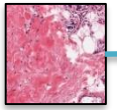
~ 10^3-10^4 tiles

AI

1 label

Type

Classical image analysis approaches

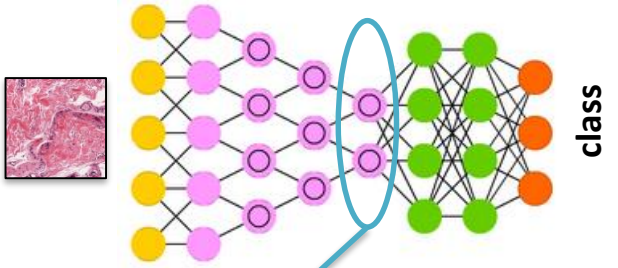


Counting nuclei
Edge selection
Cell shape
Cell graph
....

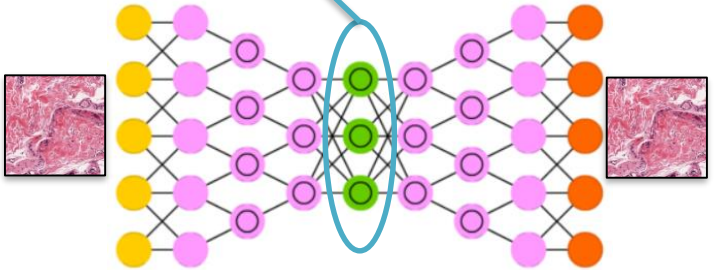
features

Deep Artificial Neural Networks

Deep convolutional neural network (CNN)

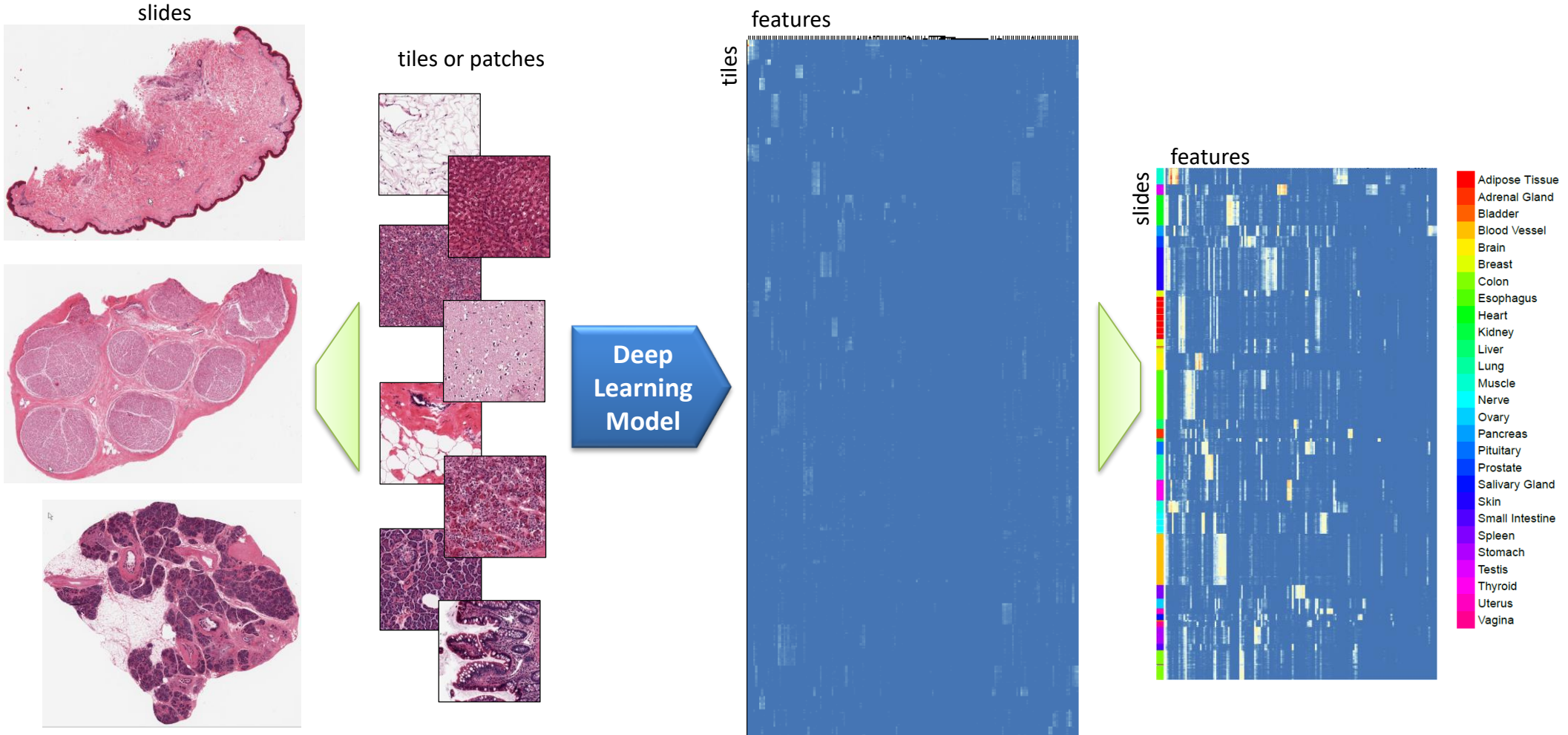


features



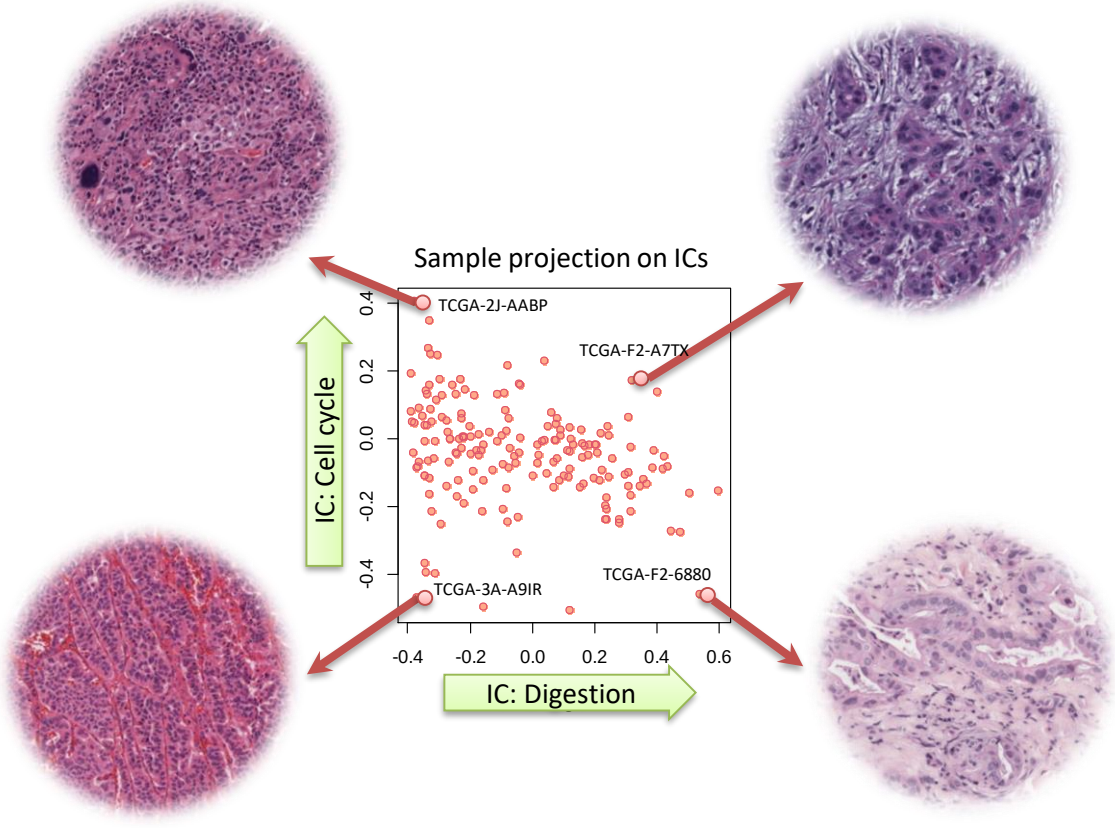
Convolutional Autoencoder (CAE)

Digital Histopathology and Feature Extraction

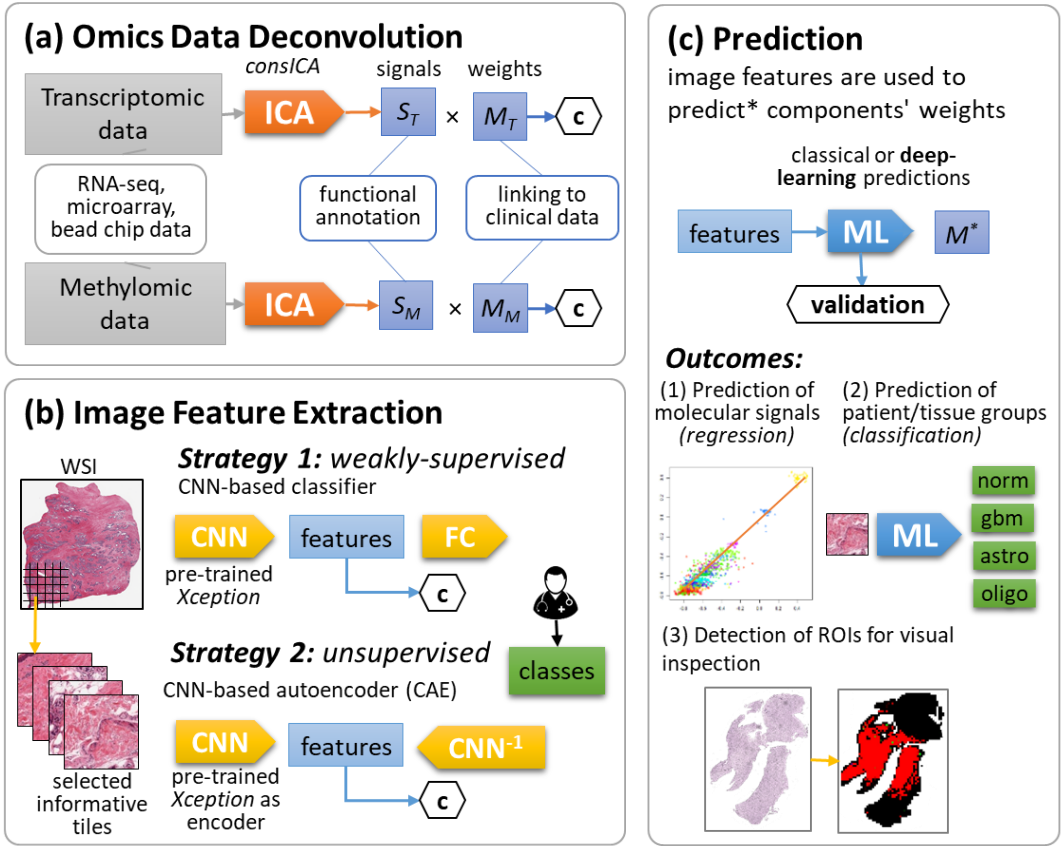


Observation

ICA results of mRNA expression data from TCGA-PAAD cohort



MEDEA: Project Overview



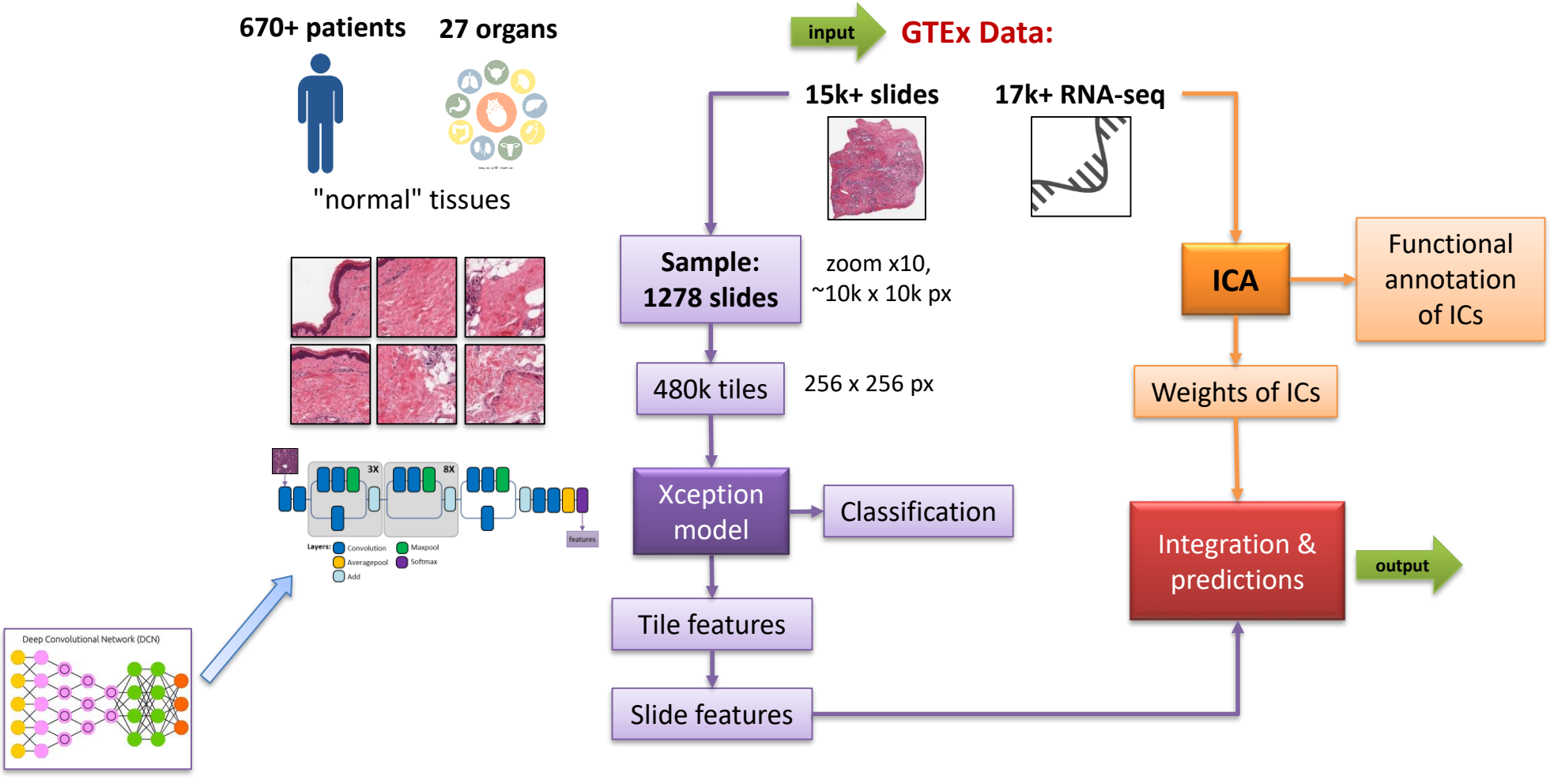
(a) Deconvolution of the omics data using developed tool *consICA*. This method was already developed and applied to entire GTEx (mRNA), TCGA (mRNA and meDNA), and DKFZ (mRNA) cohorts.

(b) Image analysis and feature extraction starts with a pre-trained DLN 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 – is completely unsupervised using CAE. Pretrained DLN can be used as an initial estimation of the encoder's parameters.

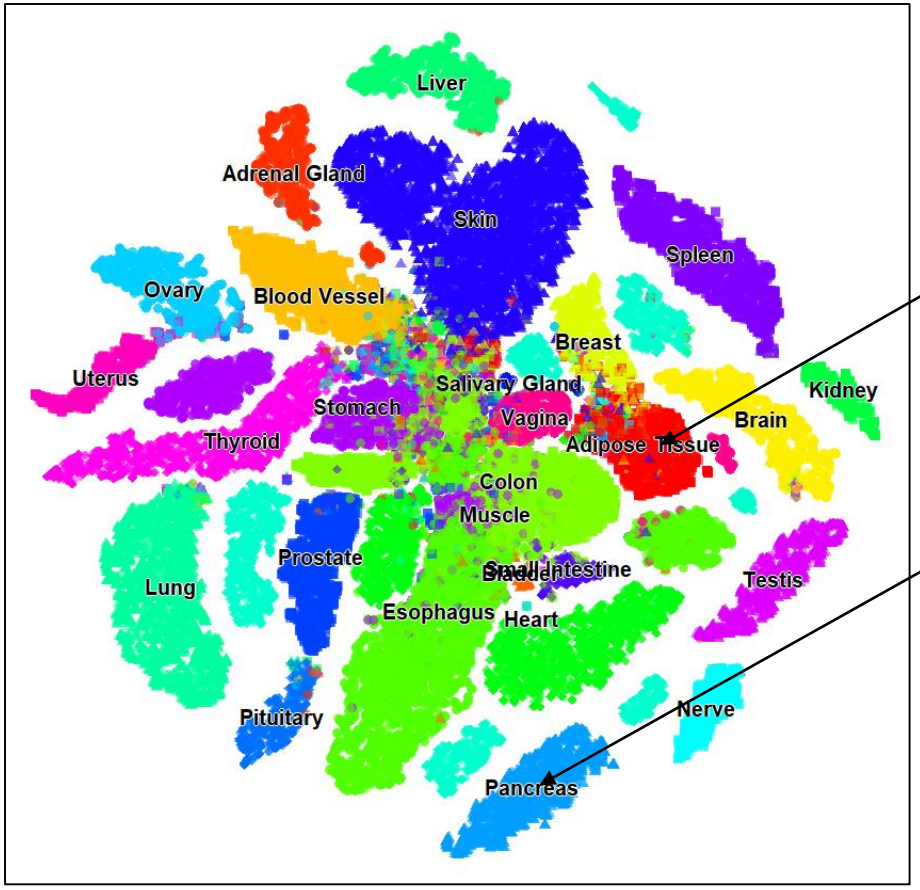
(c) Integration of ICA-weights and image features is done either by a classical ML-approach (linear regression or random forest regression) or by an FC neural network. A thorough validation of the results include (i) validation of an external pancreatic cancer cohort (DKFZ) and collection and (ii) in-depth 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.

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

Preliminary Results at GTEx Dataset

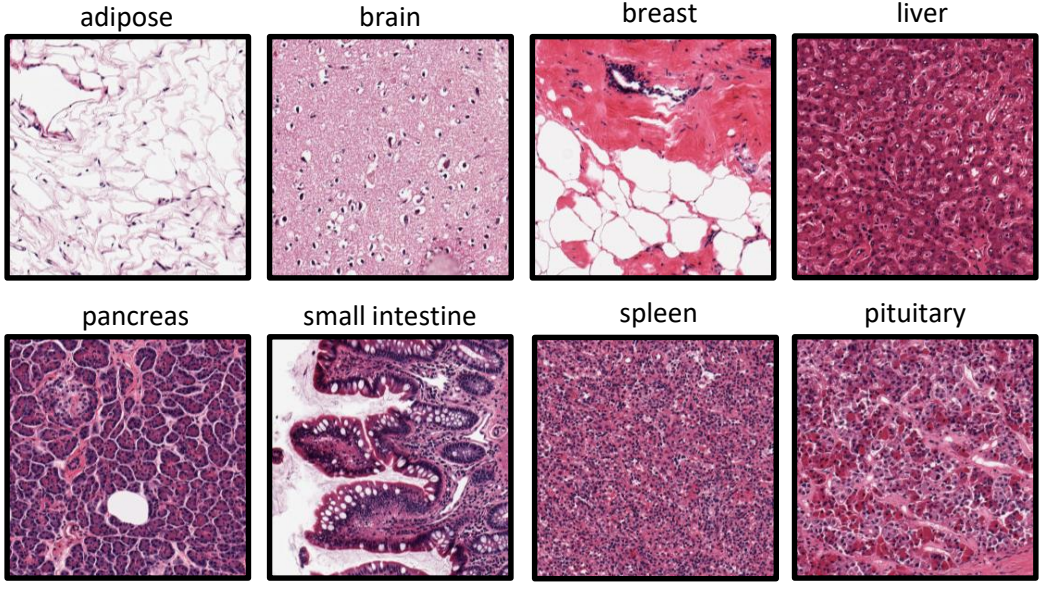


Tile-level Feature Extraction



tSNE of tile features

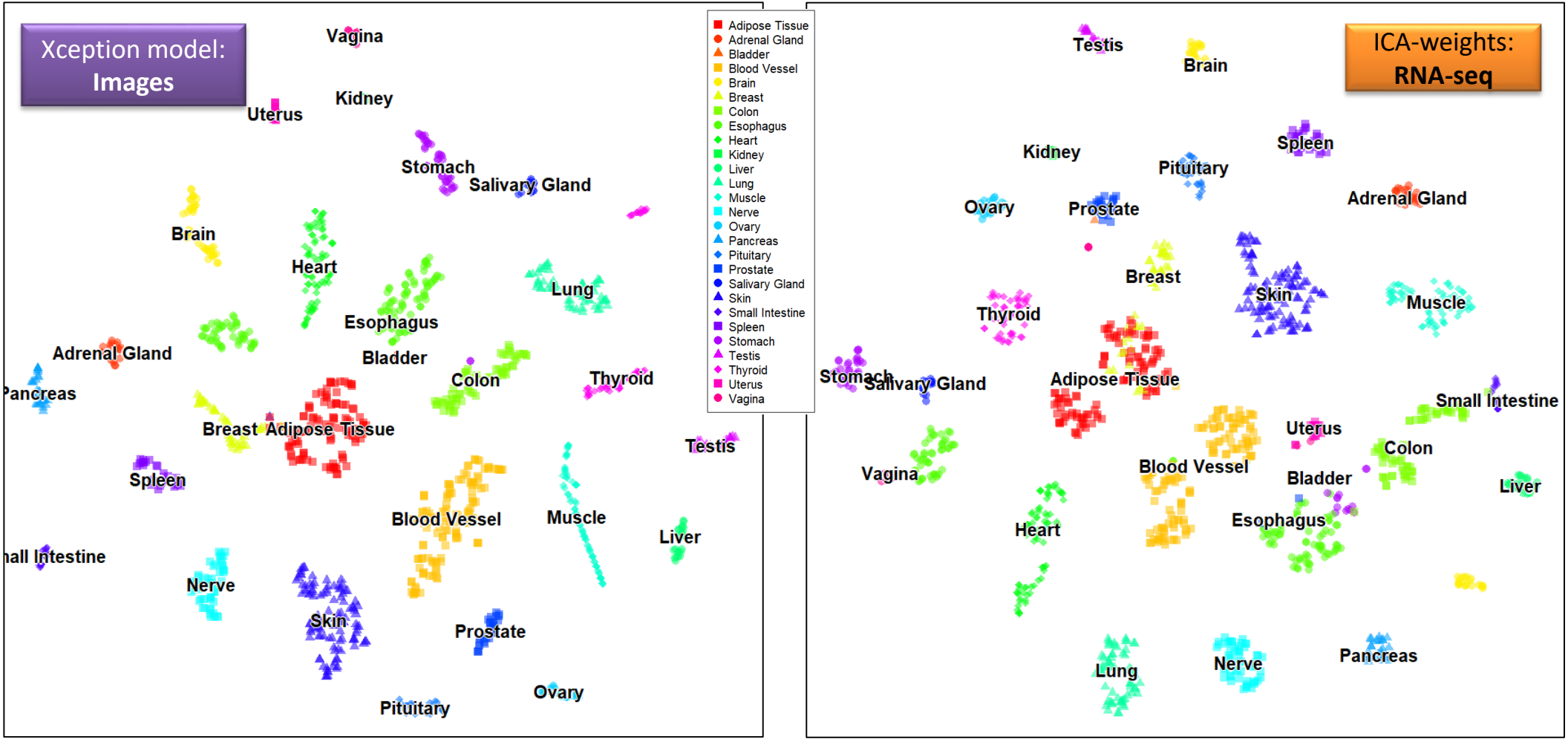
Examples of tiles classified with top certainty and co-localized with class medoids



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

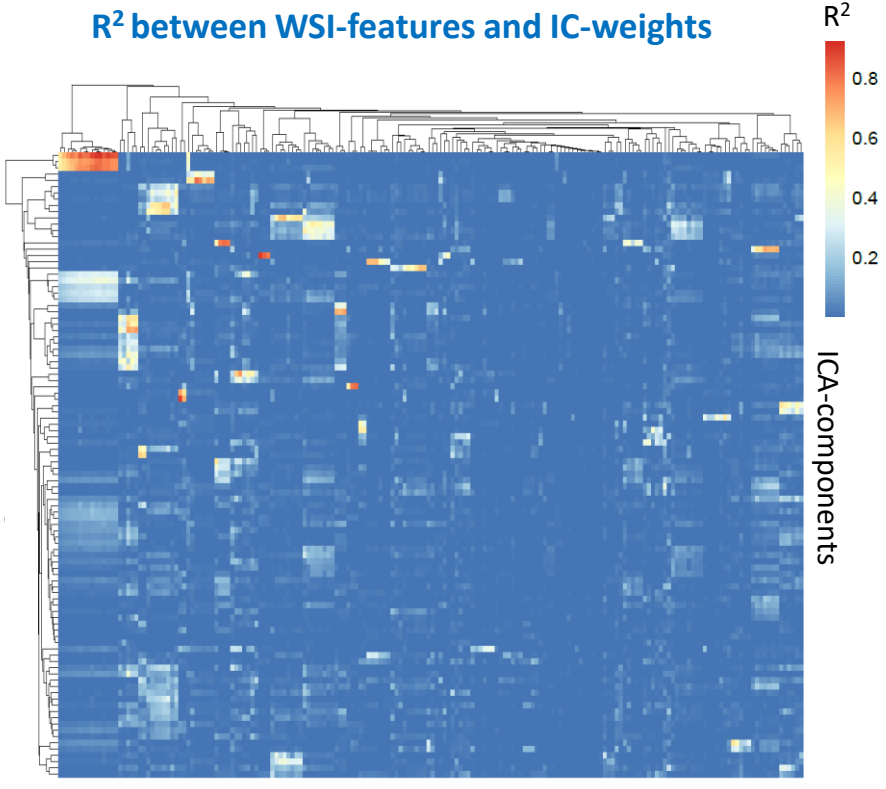
Further analysis:
These features were summarized to slide-level. Only 50% top-correlated tiles were preserved (can be further improved later…)

Slide-level Analysis and ICA

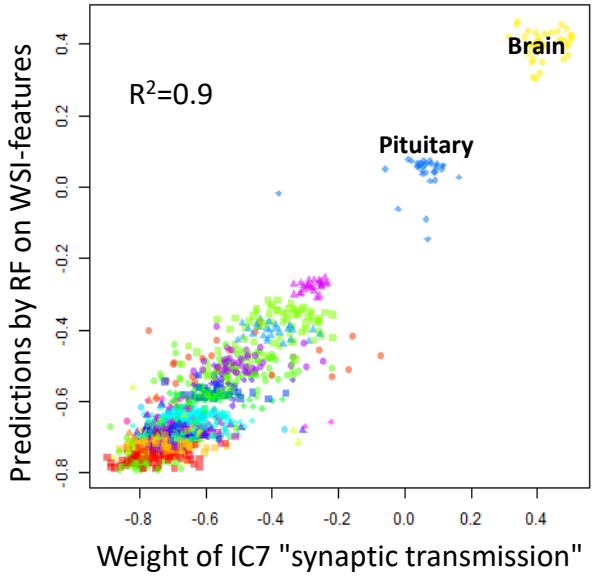


Predictions

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 $R^2>0.5$

nature COMMUNICATIONS

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✉]

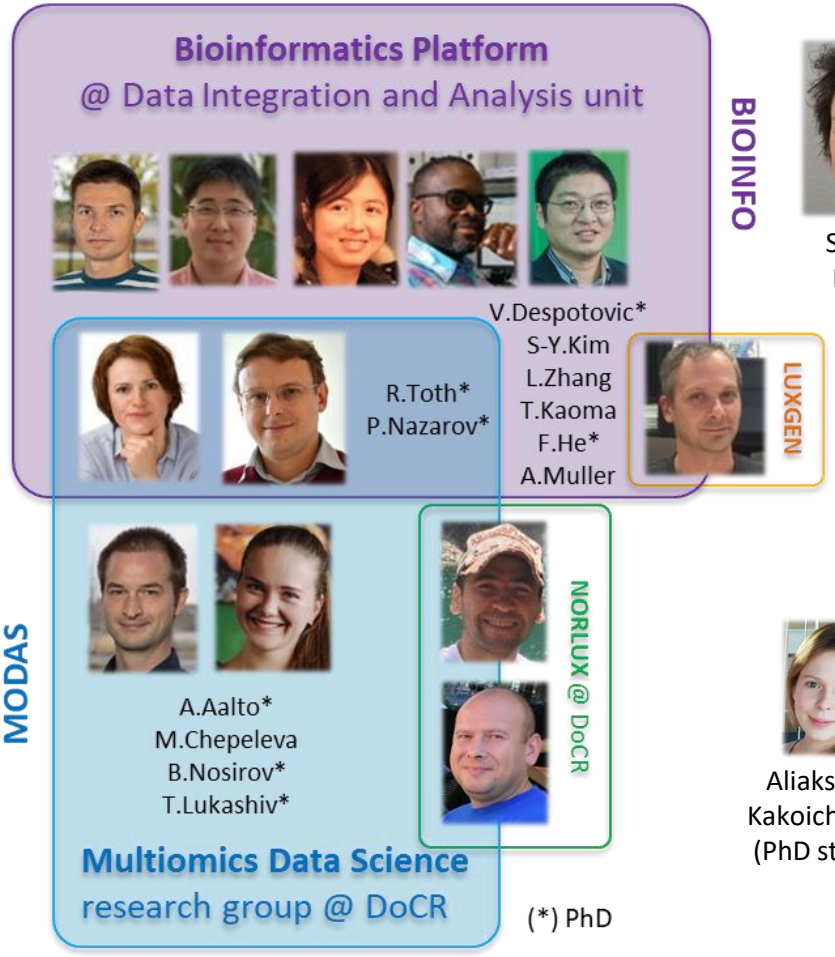
Predicting genes

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

Take Home Message 2

- 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
- Combining molecular and histopathological data may:
  - Help pathologists faster and more accurately classify samples
  - Improve the accuracy of automatic data analysis

Acknowledgements



Key internal collaborators


Simone Niclou


Anna Golebiewska


Michel Mittelbronn

Key external collaborators


LSRU, Uni Luxembourg
Stephanie Kreis


Institute Curie, France
Andrei Zinovyev




DKFZ, Heidelberg
Jörg Hoheisel
Andrea Bauer
Nathalia Giese

Interns / students


Aliaksandra Kakoichankava (PhD student)


Yibioa Wang (MSc)


Thomas Eveno (MSc)


Laurene Picandet (MSc)

Fonds National de la Recherche Luxembourg

Supported by FNR Luxembourg. Grants:
- C17/BM/11664971/**DEMICS**
- C21/BM/15739125/**DIOMEDES**

Finally: