BELARUSIAN STATE UNIVERSITY

UNIVERSITÉ DU LUXEMBOURG

VSM+U

LUXEMBOURG INSTITUTE OF HEALTH

Multiomics Data Science, Quantitative Biology Unit

Deconvolution of "Big Data" in Cancer Genomics: from Pan-cancer Level to Single Cells

Maryna Chepeleva, Yibioa Wang, Aliaksandra Kakoichankava, Arnaud Muller, Tony Kaoma, Petr V. Nazarov

petr.nazarov@lih.lu

GCB2020

Fonds National de la Recherche Luxembourg

Background

“Big data” in Genomics

TCGA

The Cancer Genome Atlas

Over 11k tumors:

- **DNA** (CNV and mutations)
- **Methylation** (~450k features)
- **RNA** (~20k coding, ~1k non-coding, >300k exonic features)
- **Tissue images** (hematoxylin/eosin staining)
- **Clinical data** (age, gender, survival...)

https://portal.gdc.cancer.gov/

GTEx

The Genotype-Tissue Expression

Over 15k normal tissues:

- **DNA** (SNPs)
- **RNA** (~20k coding, >300k exonic)
- **Tissue images** (hematoxylin/eosin staining)
- **Clinical data** (age, gender, tissue characterization)

https://gtexportal.org/

GEO

Gene Expression Omnibus

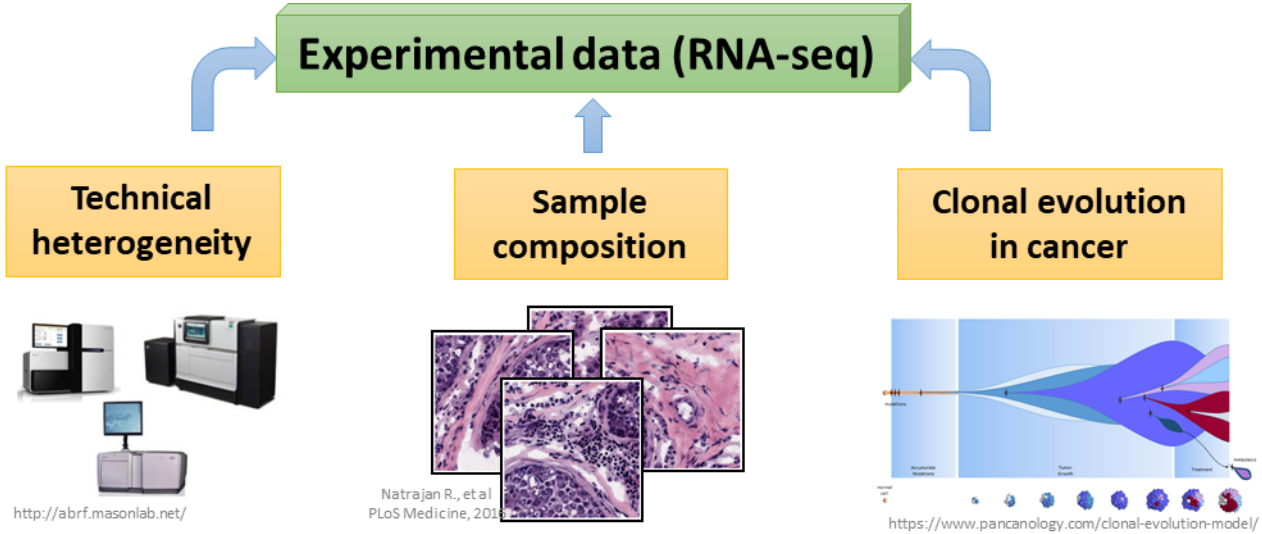
Repository Browser	
DataSets:	4348
Series:	128420
Platforms:	20821
Samples:	3549647

https://www.ncbi.nlm.nih.gov/geo/

Why can’t we still understand and treat cancer?

Background

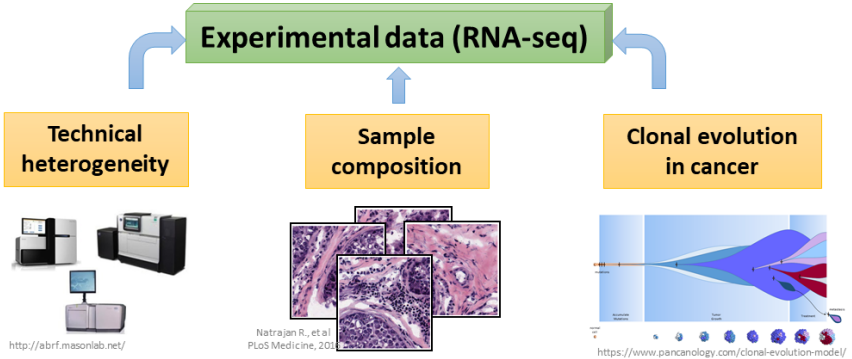
Heterogeneity

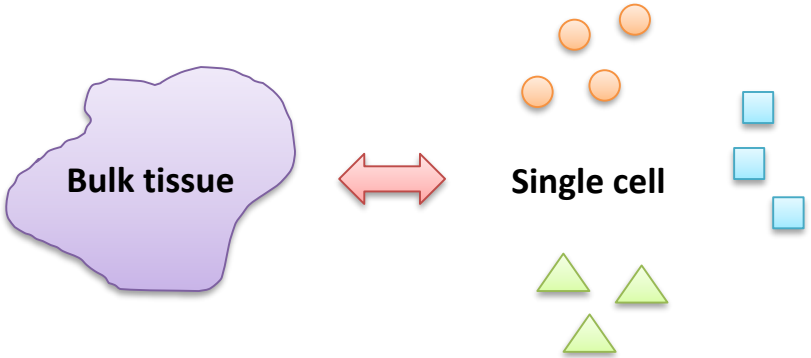


- Technical heterogeneity
- Native heterogeneity of biological tissues
- Inter/intra tumor heterogeneity due to clonal evolution

Background

Questions / Needs





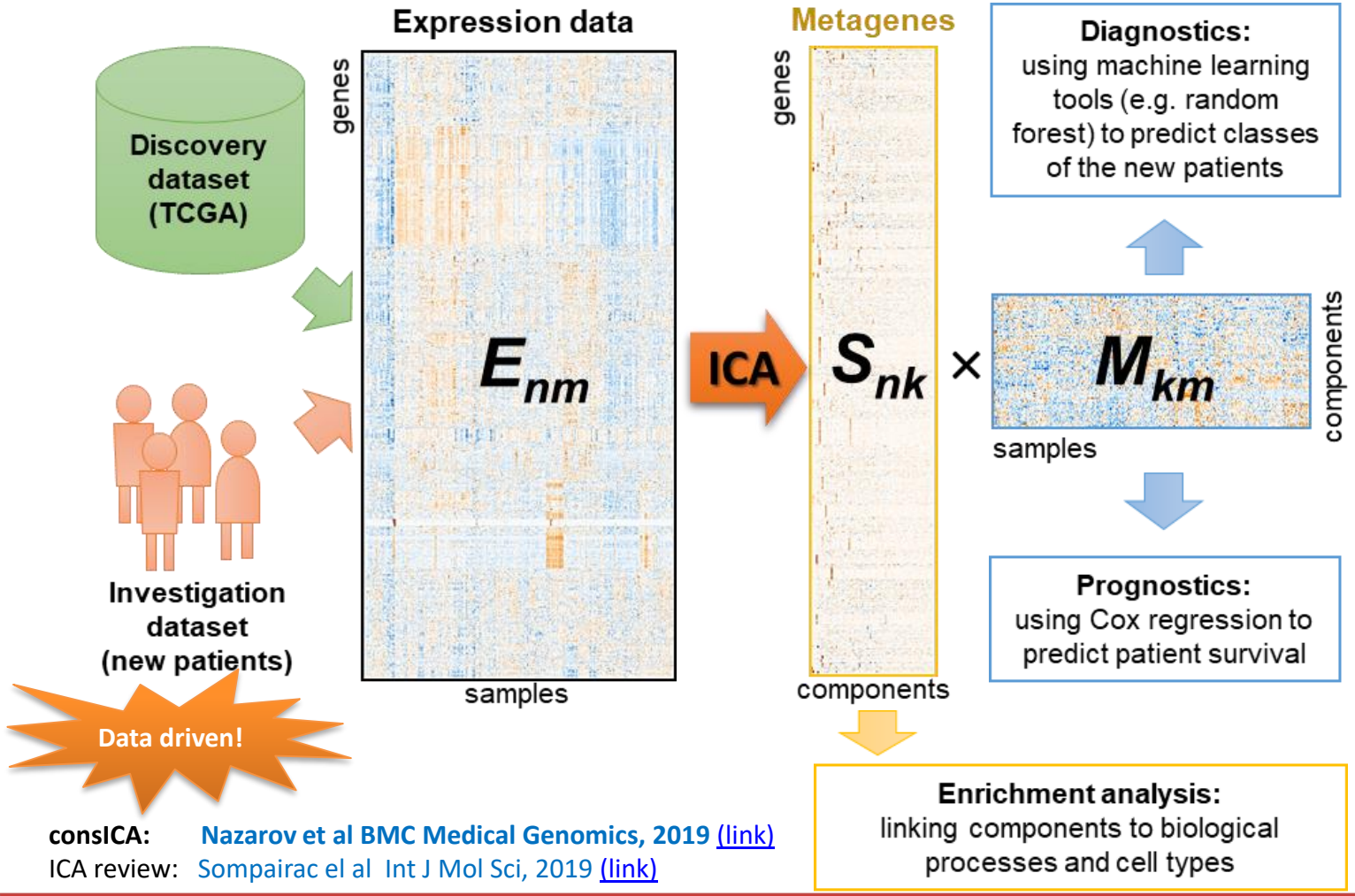
Can we come out with a method, that would :

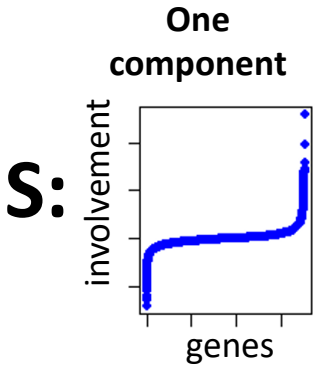
1. Disentangle cell composition heterogeneity
2. Corrects technical biases in bulk sample and in single cell data
3. Make use of single cell data for bulk sample

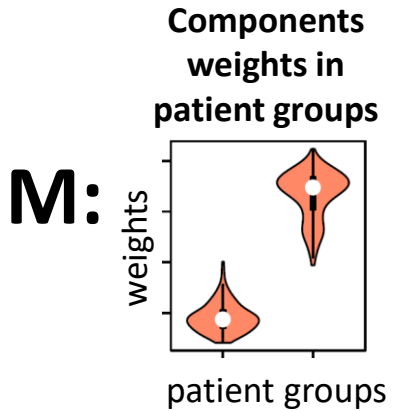
?

Consensus ICA (**consICA**) and its validation

Consensus Independent Component Analysis (consICA)

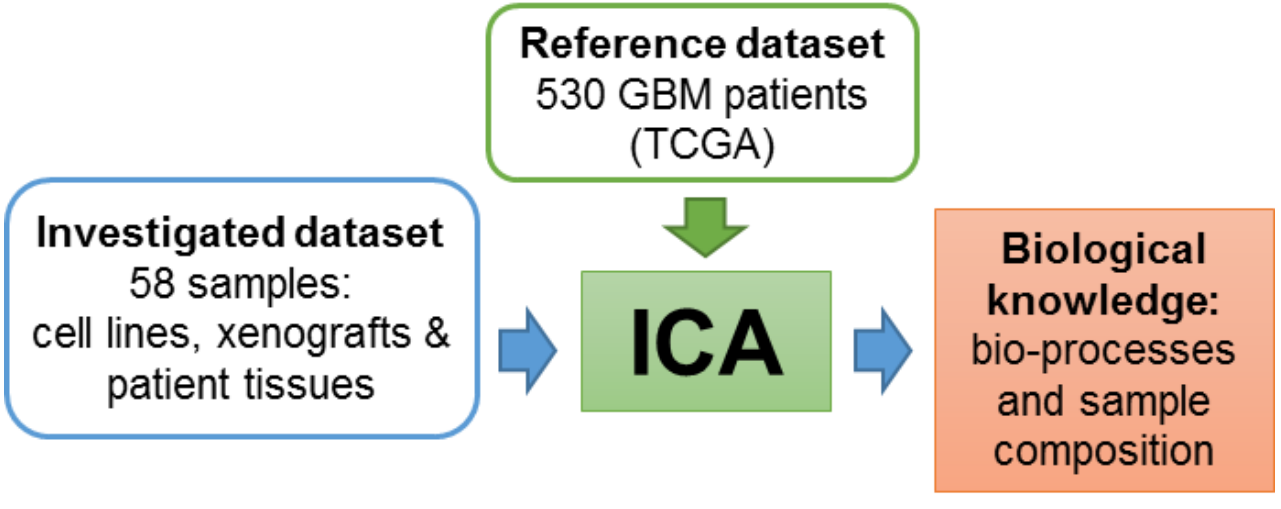




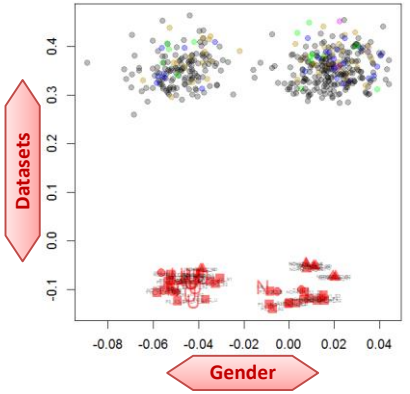


consICA: Nazarov et al BMC Medical Genomics, 2019 (link)
ICA review: Sompairac el al Int J Mol Sci, 2019 (link)

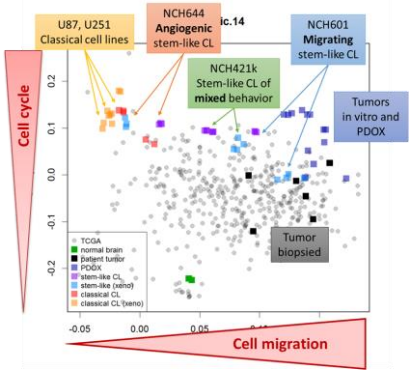
Validation with Cell Lines



Technical/trivial components: gender and platforms



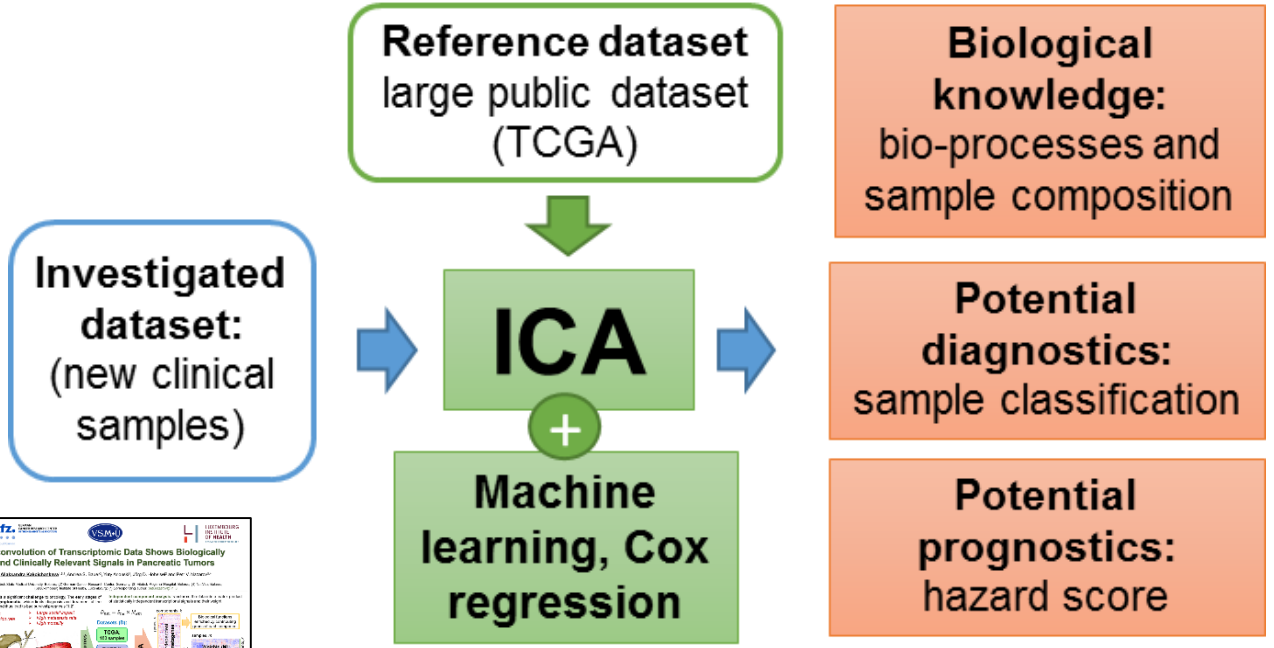
Relevant components: immune signal, stroma, cancer cells



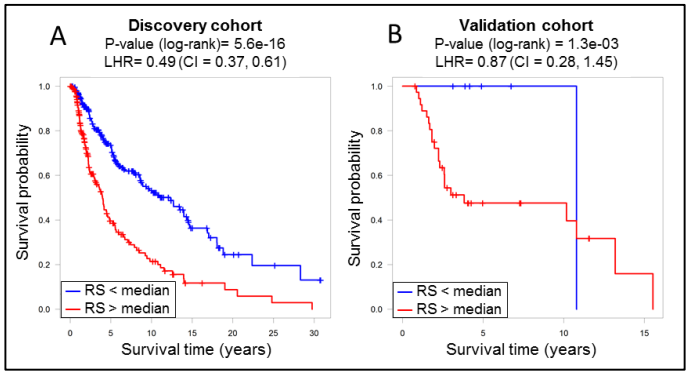
- We were able to map in-house cell line data onto TCGA dataset (GBM)
- Some IC components captured technical factors
- Other components – relevant biological information: cell cycle, cell migration, presence of stromal and immune cells

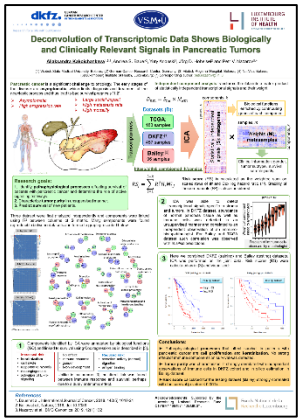
part of:
Golebiewska A. et al, **Acta Neuropathologica, 2020** (accepted)
https://www.biorxiv.org/content/10.1101/2020.04.24.057802v1.full

Patient Characterization



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

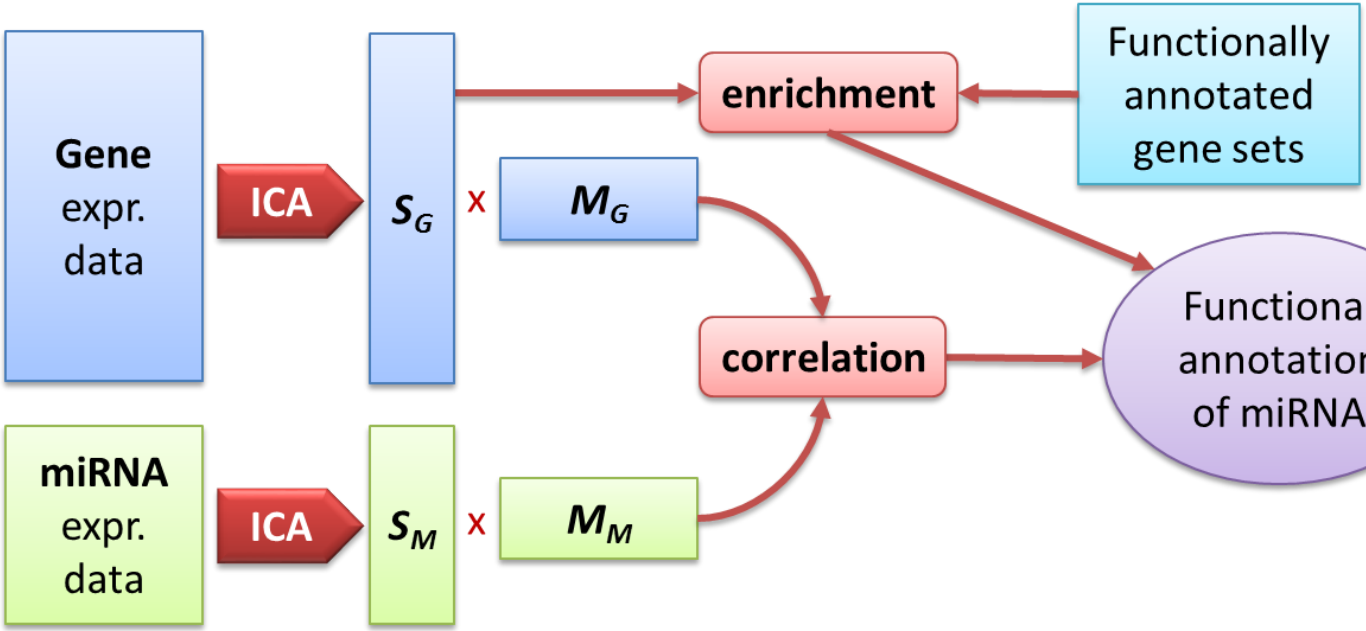


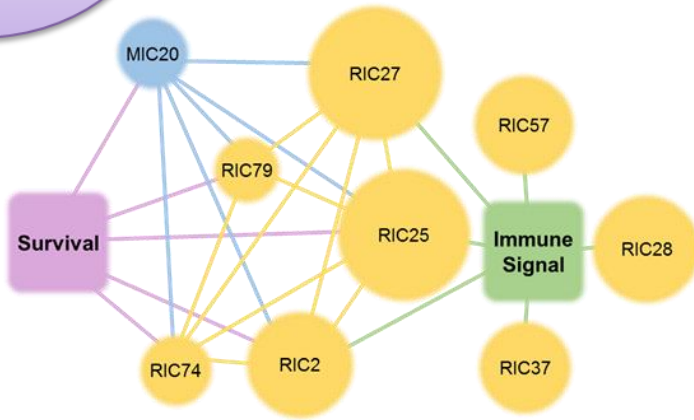


Aliaksandra Kakoichankava
(MD student)

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

ICA for Data Integration





Collaboration with Dr S.Kreis (University of Luxembourg)

Nazarov et al, BMC Medical Genomics, 2019 (link)

Application to **Pan-Cancer** Data

Yibioa Wang
(ex-MSc student)

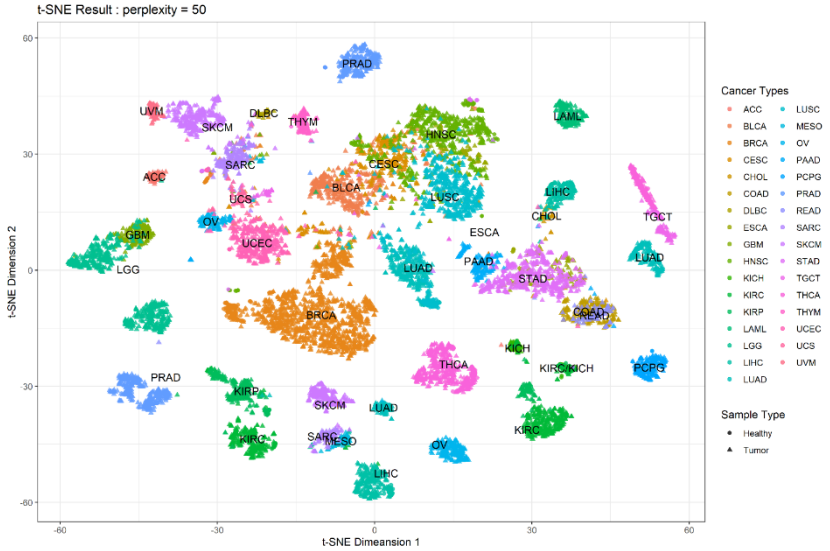
ICA for Multi-omics Pan-cancer Data

TCGA

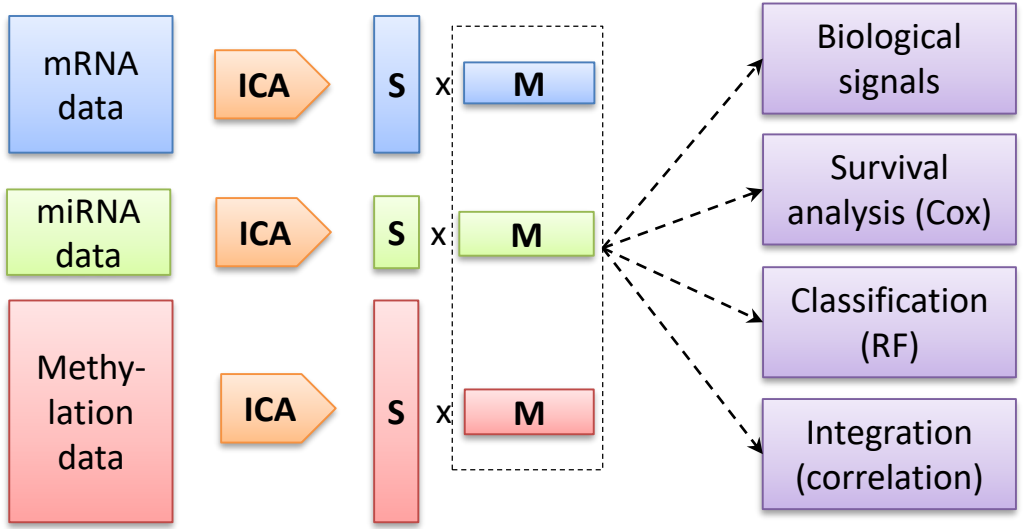
The Cancer Genome Atlas

>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)



Scheme

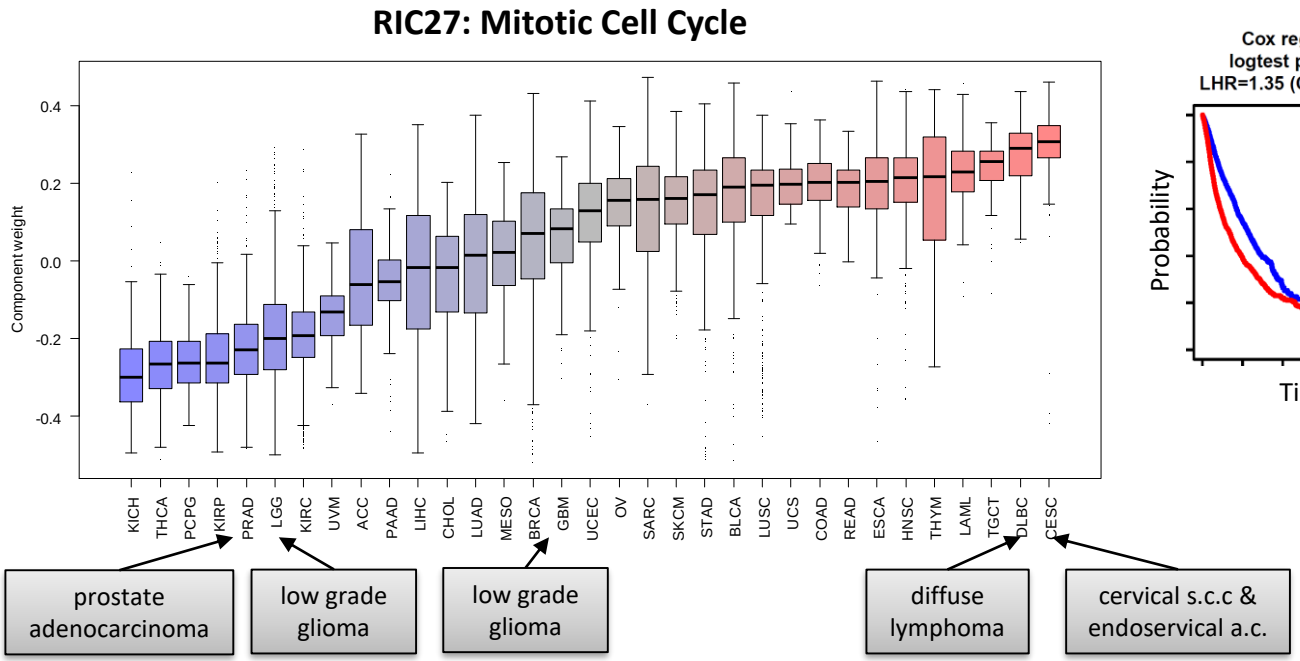


Here we used consICA with 100 components & 40 runs

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

Pan-cancer: Signals

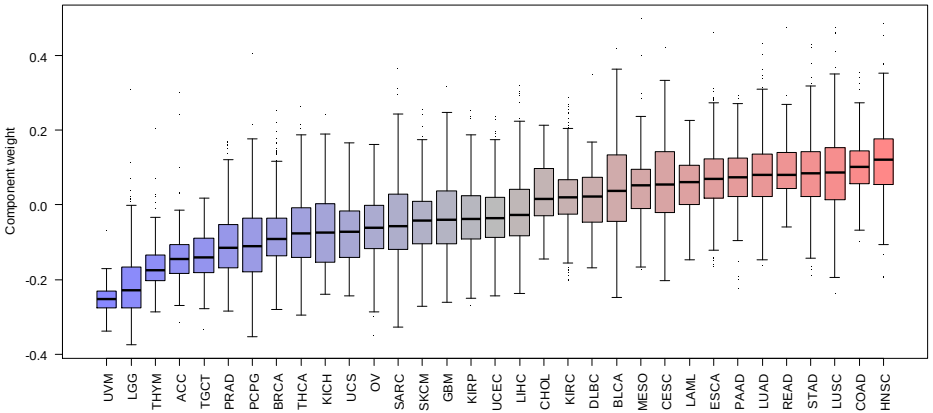
ICA Results: Cell Cycle

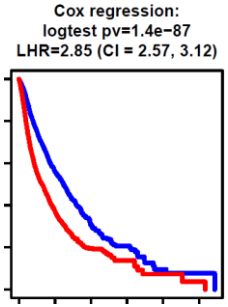


Code	Study Name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical sq. cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large b-cell lymphoma
ESCA	Esophageal carcinoma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

RIC17: Signal of Mast Cells*

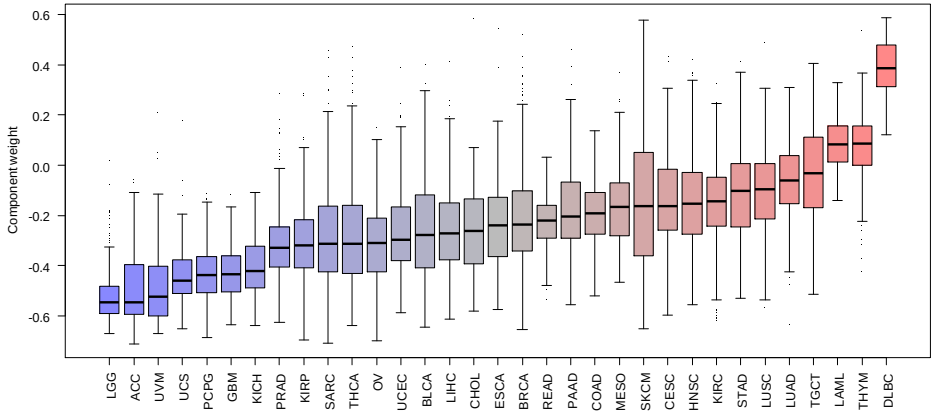




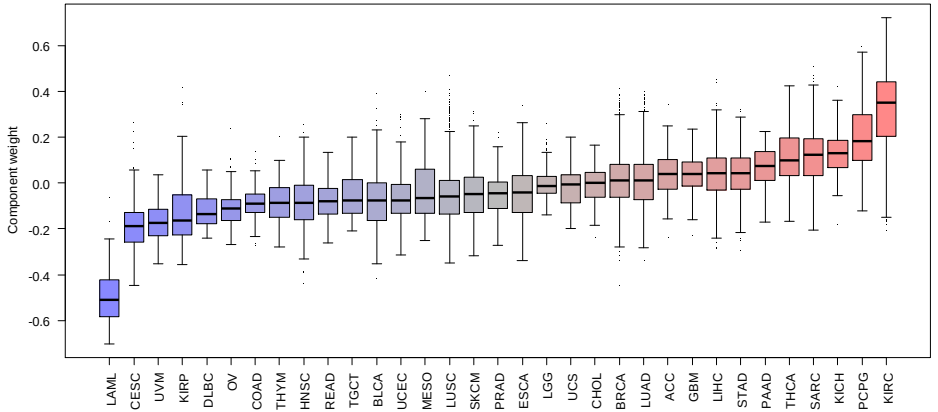
Tumor-associated mast cells (TAMCs) ?

(*) assigned based on LM22 signature (CIBERSORT)

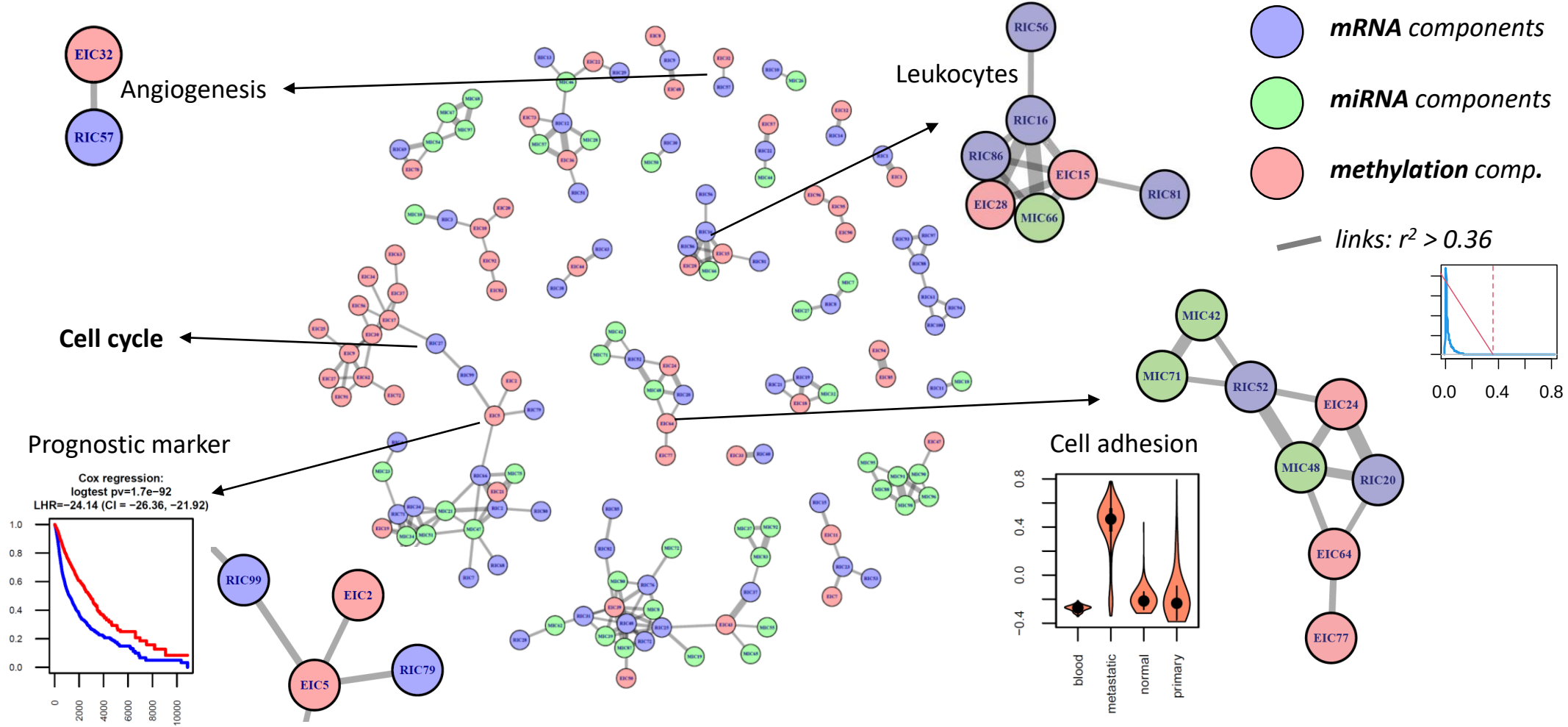
RIC16: Signal of T-Cells*



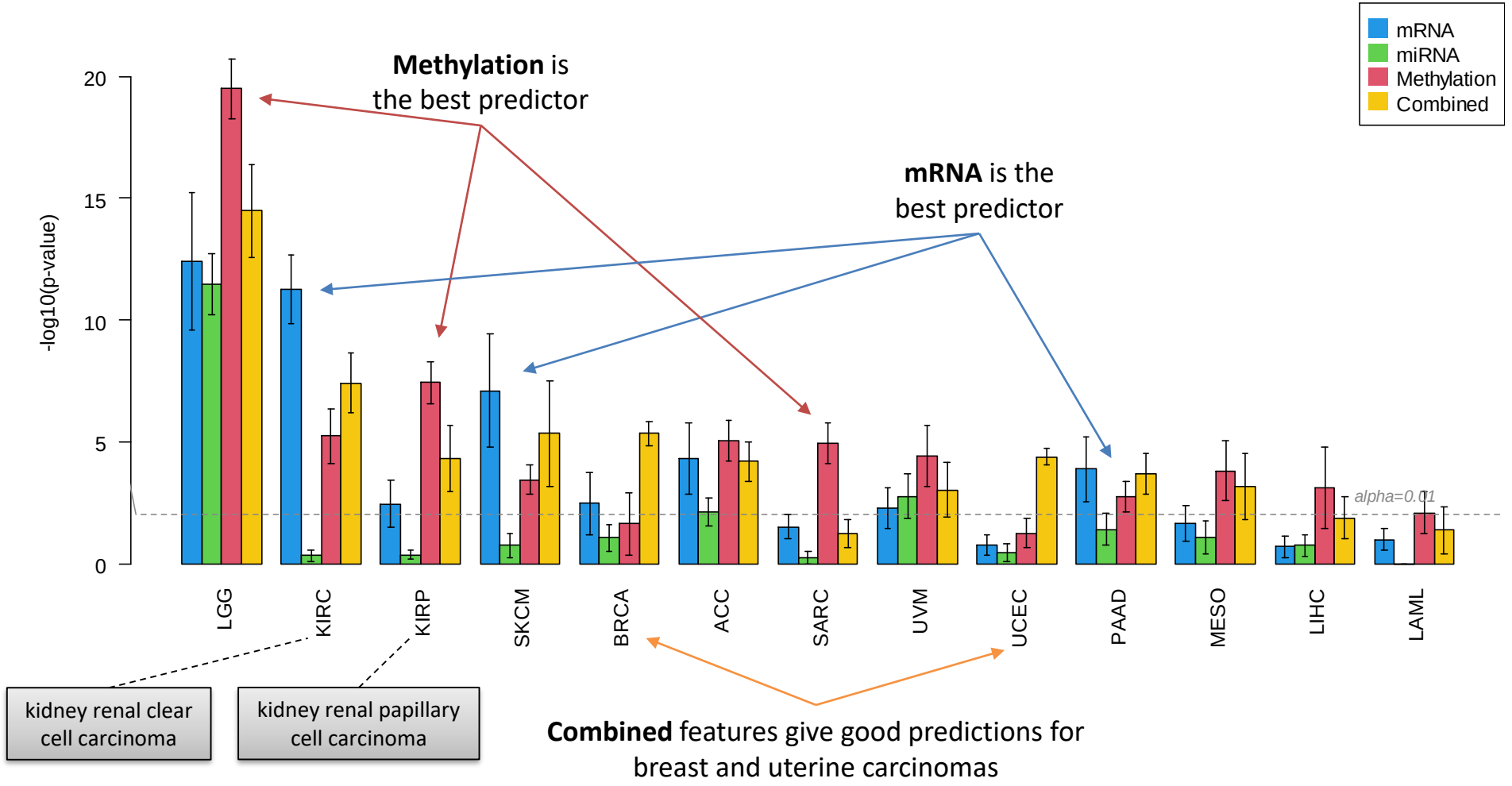
RIC57: Angiogenesis



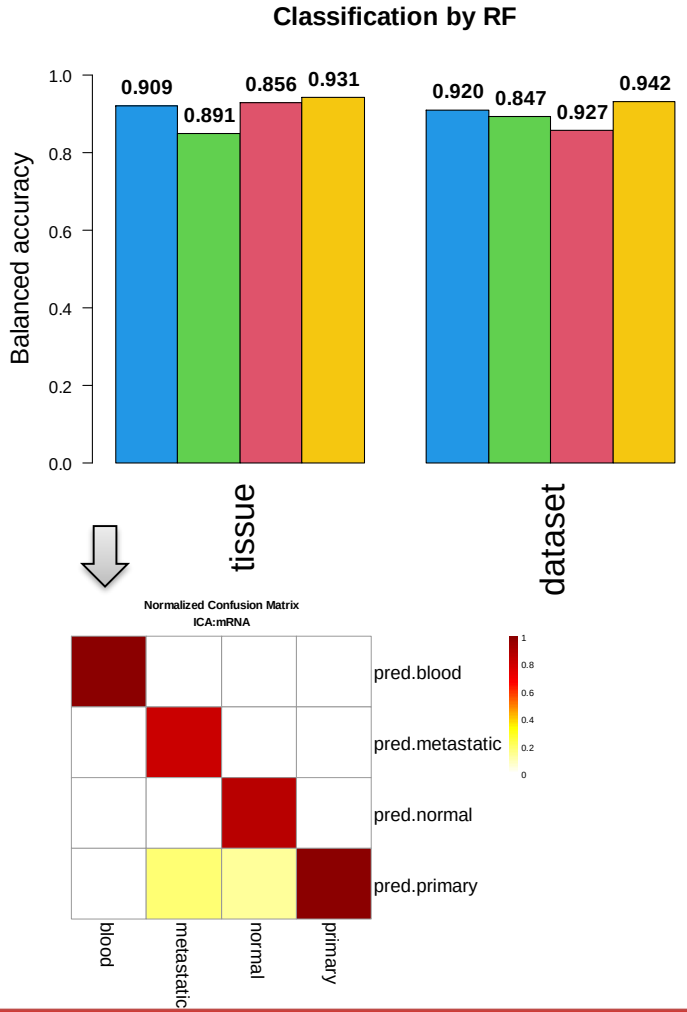
Pan-cancer: Data Integration

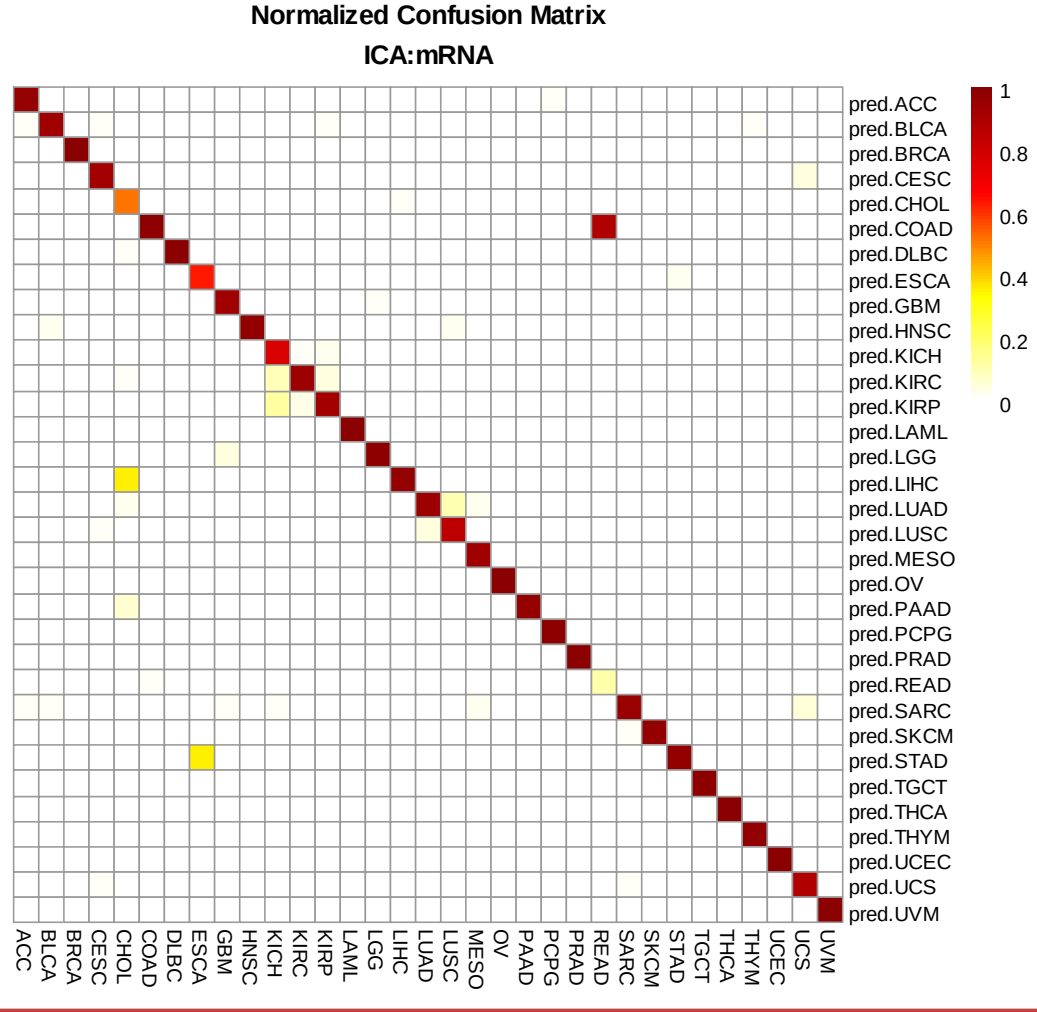


Pan-cancer: Prognosis

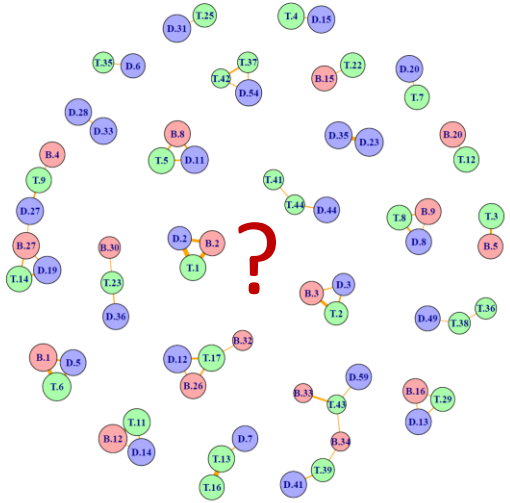


Pan-cancer: Classification





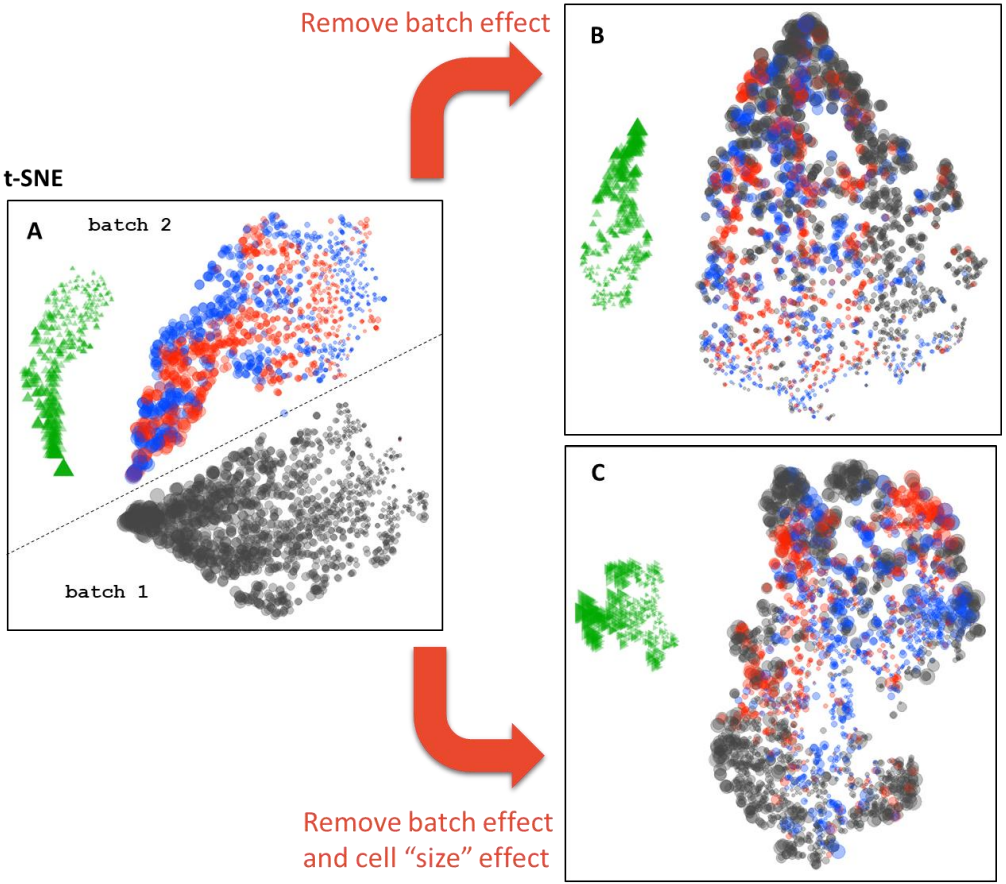
ICA on single-cell data: application and interpretation of bulk-sample data



Maryna Chepeleva
(MSc)

ICA for Single-cell Data

Correction of technical effects



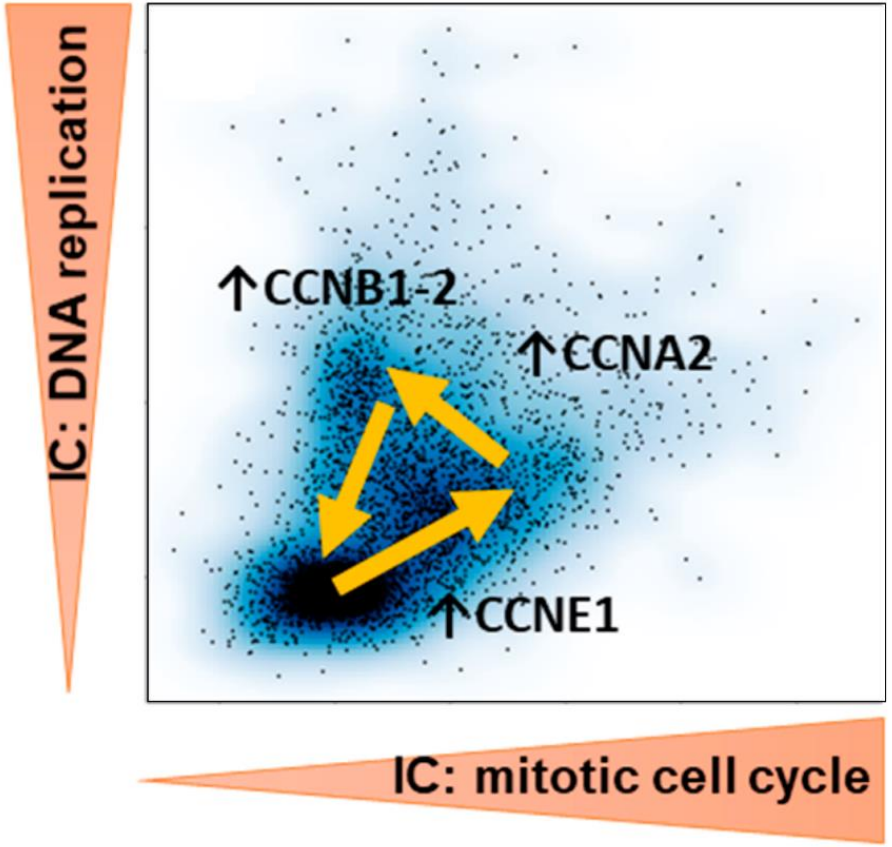
$$E_{nm} \Rightarrow S_{nk} \times M_{km}$$
$$M'_{7,m} \leftarrow 0$$
$$E'_{nm} \leftarrow S_{nk} \times M'_{km}$$

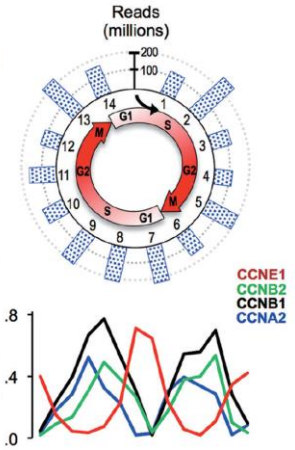
t-SNE representations of original data (A) and ICA-recovered data, after excluding batch effect (B) or several (C) components linked to technical factors ("library" size).

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

ICA for Single-cell Data

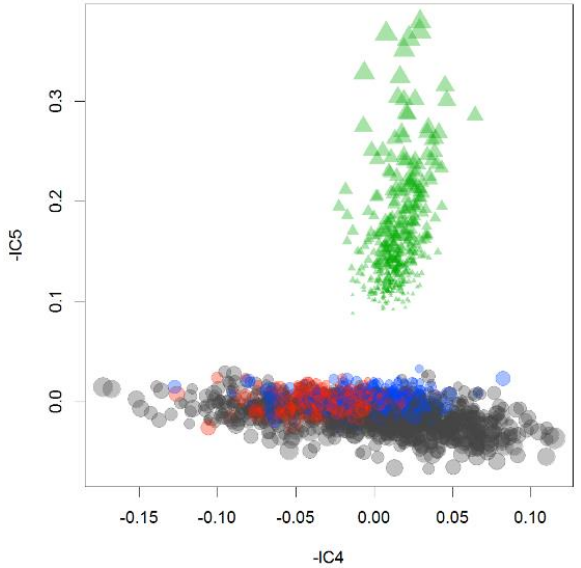
Cell Cycle in Single Cells





Dominiguez (2016) Cell Research

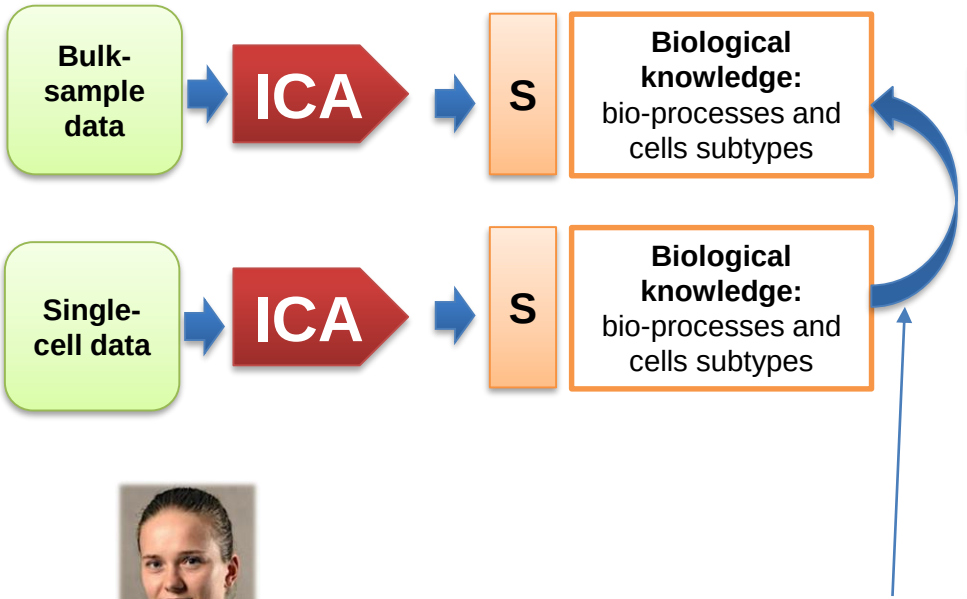
Cell Types



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

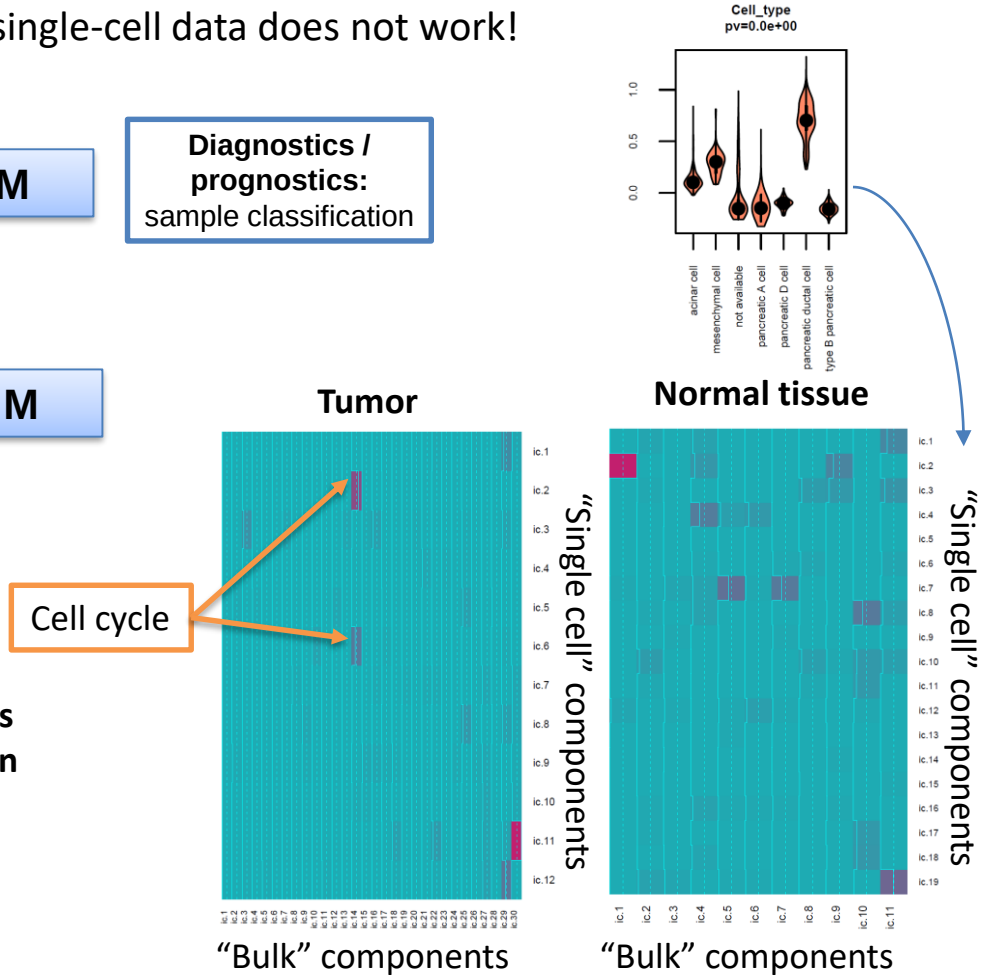
Combining Bulk and Single Cell

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



Maryna Chepeleva
(MSc student)

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



Conclusions

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- ICA on large datasets:
  - Corrects technical biases
  - Extracts "cleaned" biological signals from bulk-sample data
  - Map 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
  - Different cancers are better "presented" by different omics data

- Combining results of ICA on large bulk-sample datasets and ICA results on single-cell data strongly improves interpretability
  - normal cells helps identifying signal from stroma

Acknowledgements

Quantitative Biology Unit, LIH

Arnaud Muller (Ing.Bioinf.)

Tony Kaoma (Ing.Bioinf.)

Yibioa Wang (MSc)

Maryna Chepeleva (MSc)

Aliaksandra Kakoichankava (MD student)

Sang Yoon Kim (Ing.Comp. Sci.)

Prof. **Gunnar Dittmar**

Fonds National de la Recherche Luxembourg

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

BSU, Belarus
Dr. Mikalai Yatskou

DKFZ, Heidelberg
Dr. Pavlo LUTSIK

Max-Planck-Institut für Informatik
Michael SCHERER

DKFZ, Heidelberg
Dr. Andrea Bauer
Prof. Jörg Hoheisel

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

LSRU, Uni Luxembourg
Dr. Stephanie KREIS

UCB Celltech, UK
Dr. Francisco AZUAJE

Institute Curie, France
Dr. Andrei ZINOVYEV