LUXEMBOURG INSTITUTE OF HEALTH

lih.lu

Data Analysis and Modeling in Multi-Omics

Petr V. Nazarov

modas.lu

Open lecture, Department of Math Modelling

Chernivtsi National University, Ukraine, 2023-03-14

Fonds National de la Recherche Luxembourg

Luxembourg Institute of Health

LUXEMBOURG INSTITUTE OF HEALTH

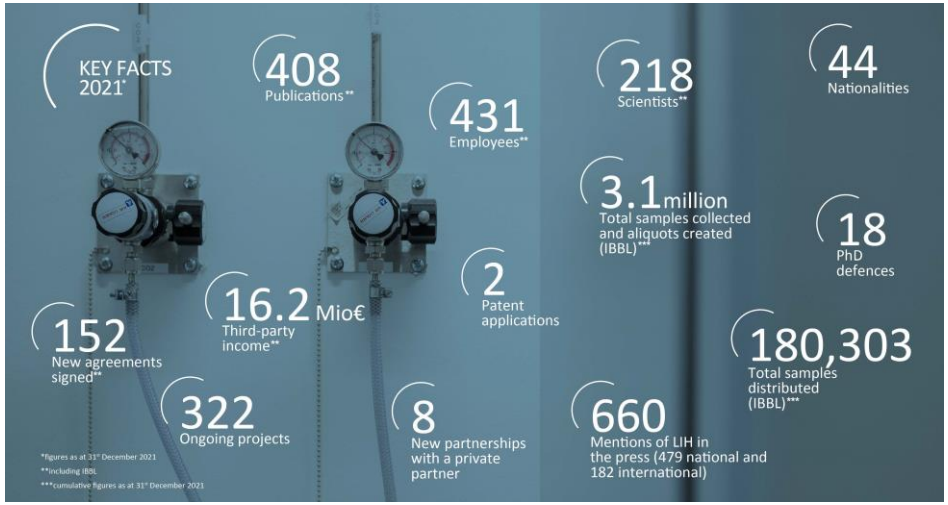
- **OUR MISSION**
- The Luxembourg Institute of Health is a precision medicine institute performing patient-based research with the aim of generating a direct and meaningful impact on people's health.

DEPARTMENTS:

- Cancer Research
- Immunology
- Precision Health

Support Teams:

- Biobank
- Experimental Platforms
- Data Analysis and Integration



Team & Competences

LUXEMBOURG INSTITUTE OF HEALTH

12 colleagues: 7 CDI, 4 CDD, 1 student

Bioinformatics Platform
@ Data Integration and Analysis unit

BIOINFO

V.Despotovic*
S-Y.Kim
L.Zhang
T.Kaoma
F.He*
A.Muller

R.Toth*
P.Nazarov*

LUXGEN

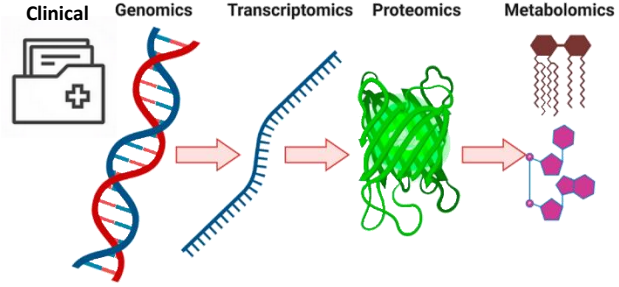
MODAS

A.Aalto*
M.Chepeleva
B.Nosirov*
T.Lukashiv*

NORLUX @ DoCR

Multiomics Data Science
research group @ DoCR

(*) PhD



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

Methods

- planning / study design
- data cleaning & sharing
- **statistical methods:**
  - linear modeling
  - survival analysis
  - enrichment analysis
- **dimensionality reduction**
- **deconvolution**
- **classical machine learning**
- databases & programming
- visualization
- pipeline development
- **deep-learning**

Bioinformatics domains

- genomics
- **epigenomics**
- **transcriptomics (bulk-sample)**
- **transcriptomics (single cell)**
- proteomics
- clinical data
- drug sensitivity
- video & **audio analysis**
- image analysis

Preference to open-source solutions over commercial

Training @ LIH / University of Luxembourg

"**Biostatistics**", 60hrs (PN) *2011-2021*
"**Multiomics Data Science**", 20hrs (PN)
"**Introduction to R Programming**" 30hrs (RT)
"**Machine Learning**", 24hrs (VD)

Output 2020-2023

- 51 journal publications (IF>2)
- Contributors to 2 EU grants
- Co-PI or WP leaders in 5 other projects

Outline

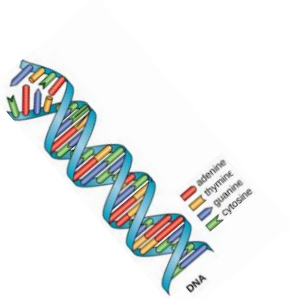
- **Concept & Data**
  - The central dogma of information transfer…
  - …and how it is implemented (to the current knowledge)
  - Data examples
- **Models**
  - The original question of 2003: “Can a biologist fix a radio?”
  - Some models frequently used
- **Methods**
  - Linear models
  - Dimensionality reduction: PCA and tSNE/UMAP
- **Example**
  - independent component analysis (ICA) for signal separation in cancer research

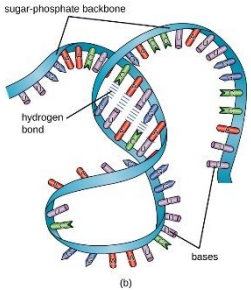
Data

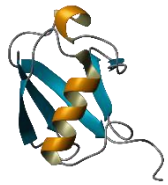
"Data is the new oil"

Clive Humby

Central Dogma of Biology







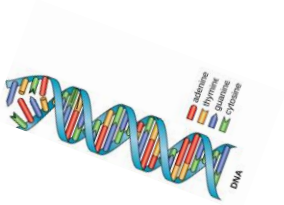
DNA → **RNA** → **Protein**

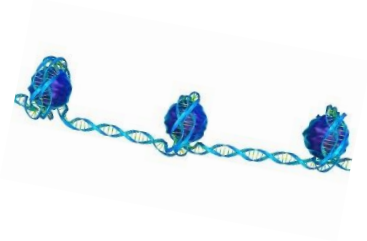
Transcription *Translation*

Adapted from:

- https://bio.libretexts.org/TextMaps/Map%3A_Microbiology_(OpenStax)/10%3A_Biochemistry_of_the_Genome/10.3%3A_Structure_and_Function_of_RNA
- http://www.bmrb.wisc.edu/featuredSys/ubiquitin/ubiquitin1.shtml

A Bit More Realistic Central Dogma…



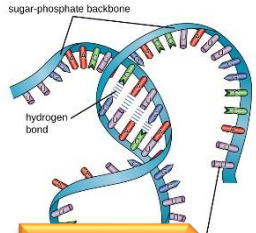


DNA

…ATGCTATATGC…
| | | | | | | | | | |
…TACGATATACG…

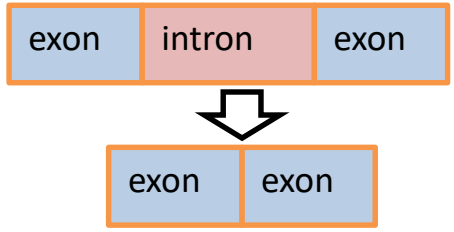
Methylation and chromatin remodeling

Is the sequence accessible for reading?
Cell differentiation



RNA

…UACGAUAUACG…
70% of HS genome is transcribed, only 2%-protein coding

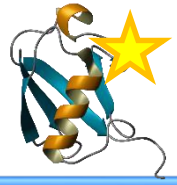


RNA splicing & RNA maturation

The source of protein diversity:
20k genes → 80k proteins

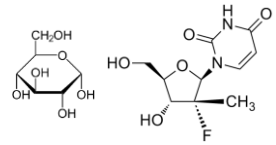
Protein

The main building blocks & nano-machinery



Protein post translational modification

Activation of proteins by (e.g.) phosphorylation



Metabolome

Biochemical compounds, small molecules

Central Dogma and the Data

Data Types

Genomic data	Epigenomic data	Transcriptomic data	Proteomic data	Metabolomic data
• DNA-seq • SNPs • indels • CNV	• CpG methylation • ChIP-Seq • ATAC-seq	• qPCR • RNA-seq • gene expression • exon/junction expression • small RNA expression	• protein arrays • mass spectrometry data	• mass spectrometry

It is impossible to use these levels of data without proper: **Clinical & histological data**

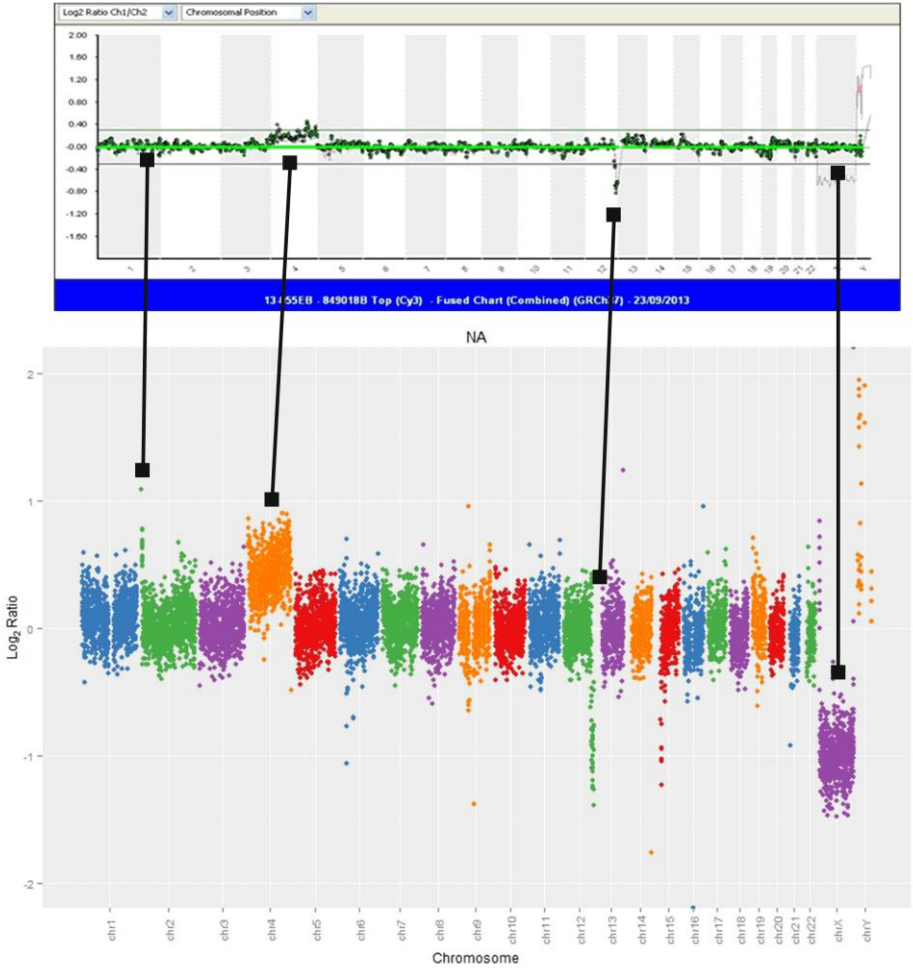
DNA: Copy Number Variation (CNV) Data

Changes at **DNA** level data can be presented as:

- a single mutation (e.g. SNP)

  ATGATTGGCA
  ATGATAGGCA

- copy number variation (CNV)



Epigenomics: Methylation Data

Epigenetic level:

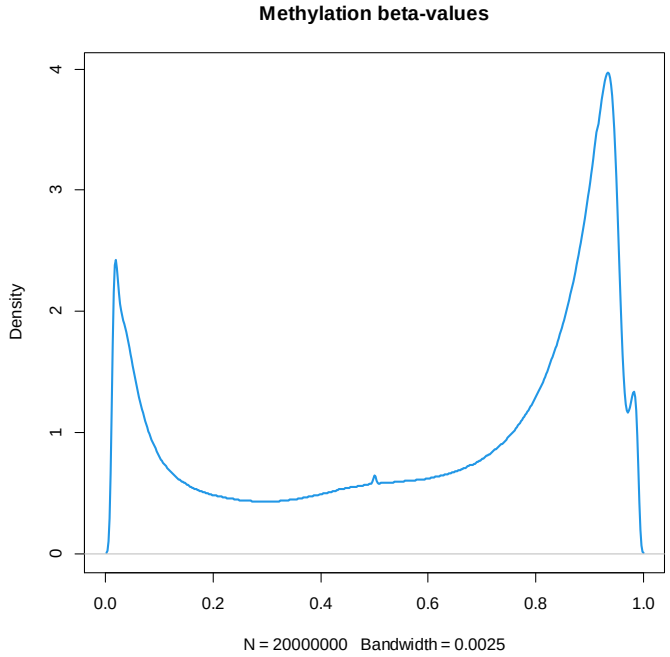
- DNA methylation (cytosine) in CG pairs
- histone modifications

← samples →

site	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10
cg01542719	0.840	0.502	0.875	0.680	0.609	0.678	0.715	0.858	0.745	0.738
cg01949461	0.116	0.137	0.075	0.053	0.129	0.122	0.085	0.052	0.047	0.117
cg02466016	0.900	0.786	0.878	0.539	0.862	0.889	0.899	0.885	0.929	0.892
cg02657012	0.898	0.602	0.787	0.731	0.770	0.784	0.882	0.836	0.924	0.863
cg03741155	0.835	0.774	0.925	0.422	0.672	0.479	0.919	0.911	0.889	0.897
cg05799169	0.956	0.870	0.943	0.959	0.934	0.956	0.973	0.966	0.945	0.975
cg06939852	0.862	0.733	0.913	0.694	0.622	0.865	0.901	0.862	0.877	0.809
cg07080864	0.710	0.533	0.659	0.680	0.406	0.835	0.382	0.602	0.566	0.755
cg07743730	0.861	0.684	0.888	0.221	0.700	0.667	0.907	0.858	0.859	0.841
cg08856118	0.478	0.375	0.513	0.379	0.295	0.499	0.332	0.196	0.476	0.473
cg10904070	0.934	0.509	0.893	0.846	0.815	0.897	0.918	0.897	0.945	0.808
cg13491584	0.633	0.335	0.443	0.818	0.333	0.711	0.170	0.275	0.455	0.175
cg18854666	0.858	0.369	0.628	0.191	0.687	0.470	0.812	0.746	0.835	0.788
cg20019985	0.634	0.089	0.019	0.509	0.112	0.514	0.116	0.012	0.556	0.027
cg20534287	0.827	0.570	0.678	0.396	0.561	0.721	0.861	0.615	0.843	0.704
cg21089930	0.795	0.582	0.370	0.560	0.484	0.764	0.645	0.762	0.649	0.548
cg21521758	0.755	0.679	0.765	0.543	0.697	0.738	0.783	0.736	0.557	0.791
cg21913888	0.641	0.389	0.793	0.859	0.283	0.575	0.761	0.649	0.606	0.340
cg22335490	0.018	0.050	0.025	0.024	0.020	0.013	0.023	0.022	0.026	0.023
cg23038813	0.386	0.030	0.070	0.034	0.072	0.180	0.137	0.212	0.029	0.033

← features = CpG sites →

↑ β-values ∈[0,1]



Examples of Data

RNA: Gene Expression Data

The most straight-forward data ☺

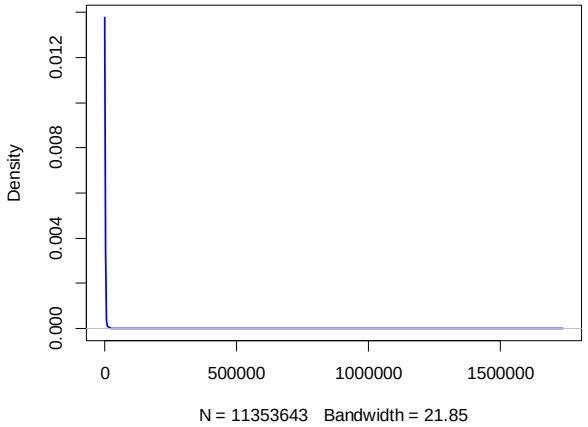
← samples →

ID	Gene.Symbol	A1	A2	A3	A4	B1	B2
ENSG00000135899	SP110	32	31	33	33	136	136
ENSG00000154451	GBP5	0	0	0	0	395	383
ENSG00000226025	LGALS17A	0	0	0	0	217	196
ENSG00000213512	GBP7	0	0	0	0	44	47
ENSG00000260873	SNTB2	198	193	195	196	483	502
ENSG00000063046	EIF4B	552	546	548	550	428	429
ENSG00000102524	TNFSF13B	0	0	0	0	16	17
ENSG00000107201	DDX58	79	81	82	77	296	310
ENSG00000010030	ETV7	2	2	2	0	93	85
ENSG00000125347	IRF1	22	24	27	22	234	236
ENSG00000180616	SSTR2	0	0	0	0	19	21
ENSG00000155962	CLIC2	2	2	1	1	71	65
ENSG00000153944	MSI2	55	54	54	54	37	37
ENSG00000197646	PDCD1LG2	0	0	0	0	58	60
ENSG00000108771	DHX58	5	4	4	5	26	25
ENSG00000100336	APOL4	9	8	11	8	130	135
ENSG00000182551	ADI1	88	86	88	89	59	60
ENSG00000128284	APOL3	14	14	14	13	85	94
ENSG00000153989	NUS1	214	216	212	214	167	167
ENSG00000131979	GCH1	57	61	57	56	172	167

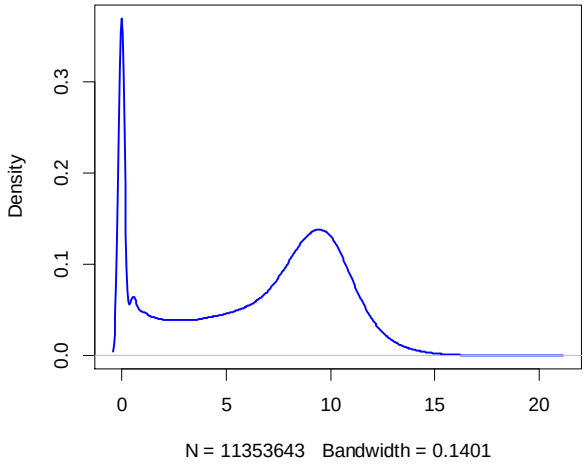
← features = CpG sites →

↑ counts (here - integer)

Distribution of counts



Distribution of log-counts

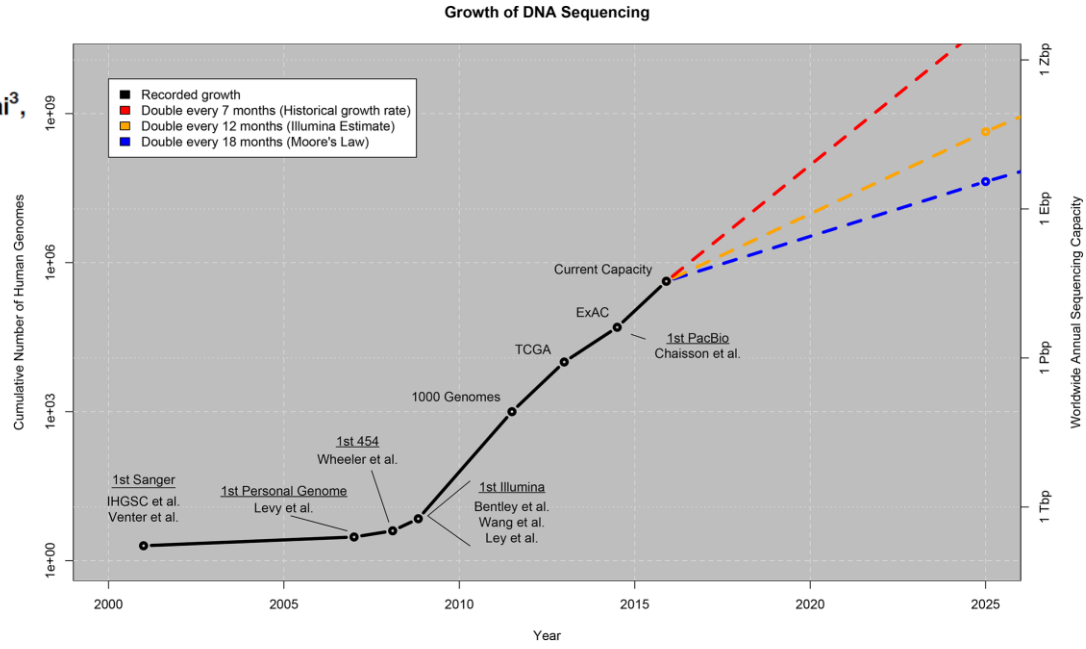


Exponential Growth

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

Prefix		Base	Base
Name	Symbol	1000	10
yotta	Y	1000^8	10^{24}
zetta	Z	1000^7	10^{21}
exa	E	1000^6	10^{18}
peta	P	1000^5	10^{15}
tera	T	1000^4	10^{12}
giga	G	1000^3	10^9
mega	M	1000^2	10^6



Current growth estimation:

40 Exabytes / year

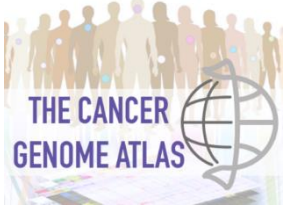
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

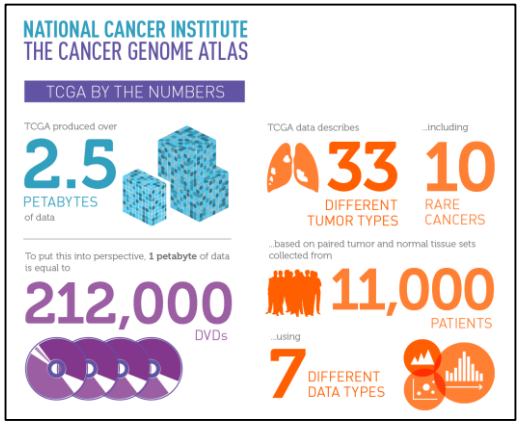
doi:10.1371/journal.pbio.1002195.t001

Open Data Repositories

LUXEMBOURG INSTITUTE OF HEALTH

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





- **Cancers**

DNA, Epi, RNA, miRNA, some proteins
Images

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

V8 Release	# Tissues	# Donors	# Samples
Total	54	948	17382
With Genotype	54	838	15253
Has eQTL Analysis*	49	838	15201

- **Normal donors**

RNA, genotype
Images

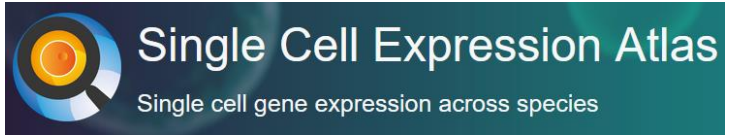
GEO https://www.ncbi.nlm.nih.gov/geo

Browse Content	
Repository Browser	
DataSets:	4348
Series:	195515
Platforms:	24867
Samples:	5575349

- **Various**

RNA, DNA, Epi., Prot….

Single cell https://www.ebi.ac.uk/gxa/sc/home



- 21 species
- 355 studies
- 10 505 726 cells

- **Various**

RNA

Models

All models are wrong, but some are useful

George E.P. Box

A Bit of History

Yuri Lasebnik, Cancer Cell, 2002 (link)

Can a biologist fix a radio?—Or, what I learned while studying apoptosis

As a freshly minted Assistant Professor, I feared that everything in my field would be discovered before I even had a chance to set up my laboratory. Indeed, the field of apoptosis, which I had recently joined, was developing at a mind-boggling speed. Components of the previously mysterious process were being

Figure 1. The radio that has been used in this study

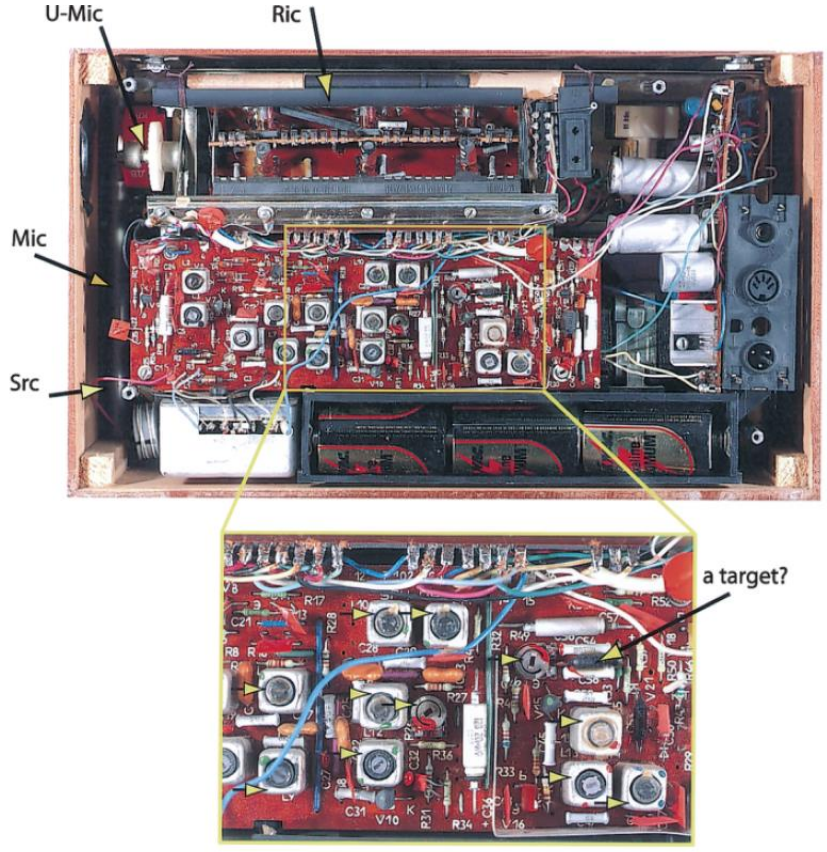
- If you want to see whether your method works, apply it to a task with an already-known solution

- As an example, let's see how the "standard" approach to modelling could help us to understand a complex system - radio

The “Standard” Approach

1. Get money to buy enough radios
2. Learn how to open a radio
3. Try to recolor the elements -> fail
4. Record and classify all elements
5. Finally, you find an element that is red in working radios but is black and smelly in the broken one ☺

??? TARGET !!!



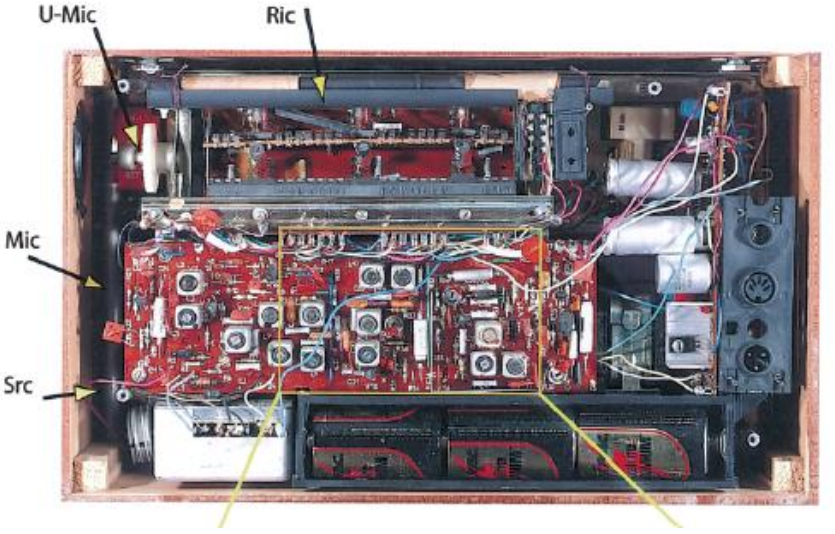
However it worked (if it work) only for this radio.
And what if the problem is in the tunable elements?

The “Standard” Approach

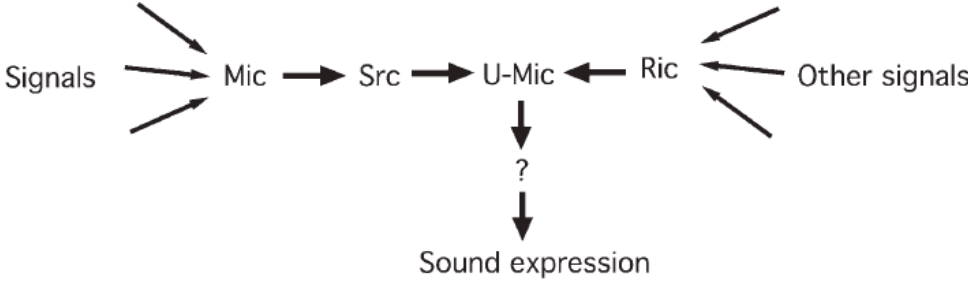
6. Try to remove elements one by one or use a short-gun over a number of radios

7. You name some discovered elements that influence radio performance as :

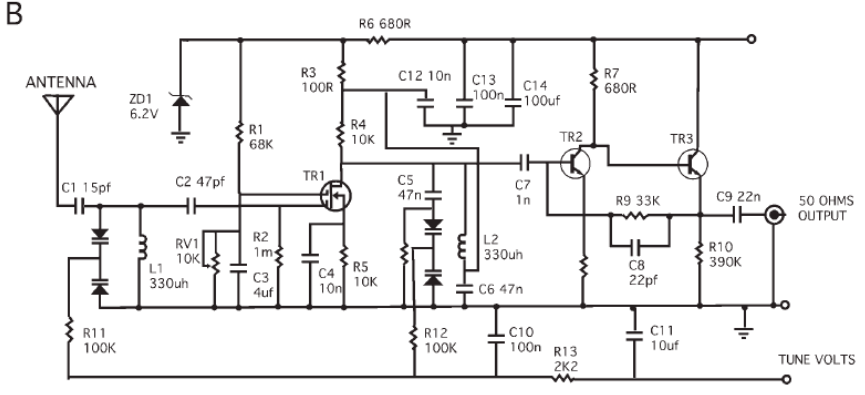
- Serendipitously Recovered Component (Src)
- Most Important Component (Mic)
- Really Important Component (Ric)
- Undoubtedly Most Important Component (U-Mic).



Our model:



Reality:



Some Types of Models Used

Kinetic modelling:
sets of ODE describing concentrations

$$S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_{+2}}{\rightleftharpoons}} P + E$$

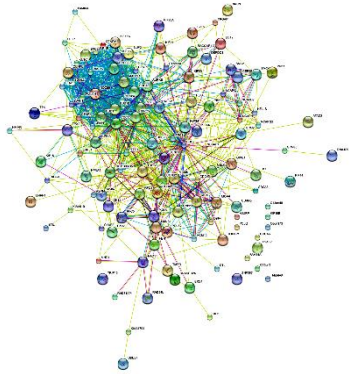
$$\frac{d[S]}{dt} = -k_1[E][S] + k_{-1}[ES]$$

$$\frac{d[E]}{dt} = -k_1[E][S] + (k_{-1} + k_2)[ES] - k_{-2}[E][P]$$

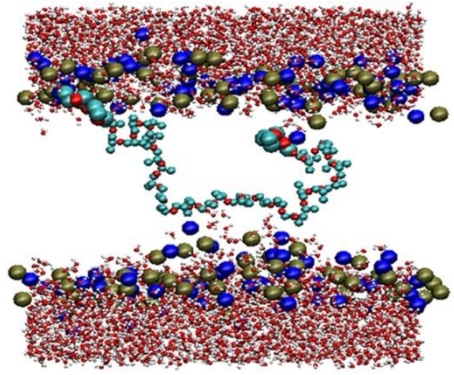
$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] + k_{-2}[E][P]$$

$$\frac{d[P]}{dt} = k_2[ES] - k_{-2}[E][P]$$

Network models:
Protein-protein interactions, Boolean networks, correlation networks, etc. Easy to build, difficult to use for explanation



Molecular dynamics simulation:
Simulate location of each atom in the system



Statistical models:
Estimation of the factor effects on gene/protein expression

GWAS:
Estimation of the mutation effects on disease

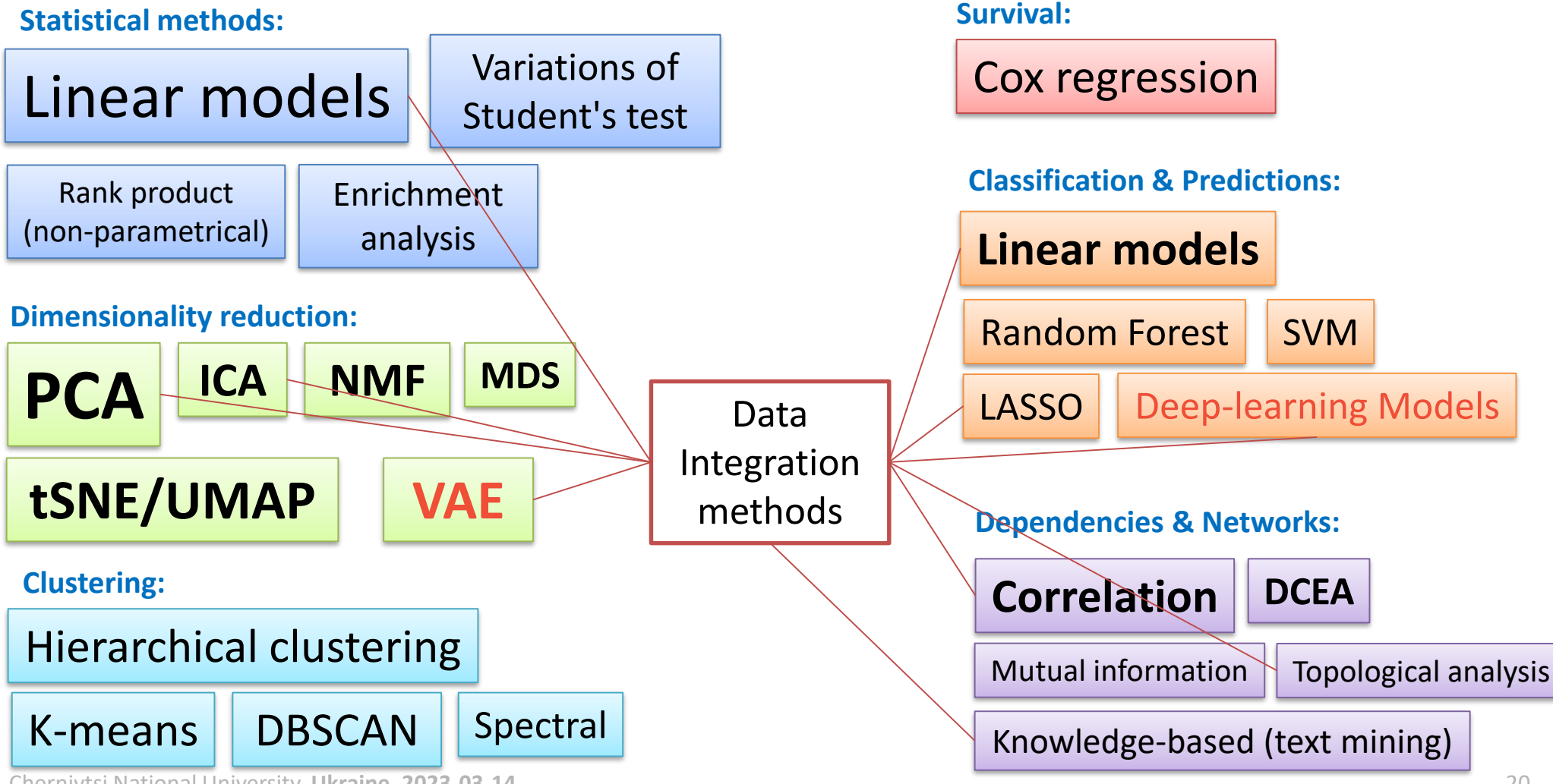
Predictive systems:
Classifiers able to predict patient groups by the gene expression

Methods

"No matter how many instances of white swans we may have observed, it does not justify the conclusion that all swans are white."

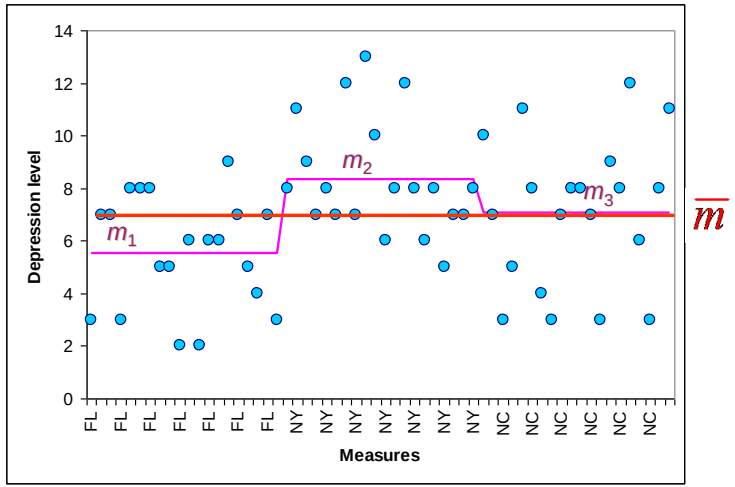
Sir Karl Popper

Methods Overview *(biased)*



Example: Linear Models

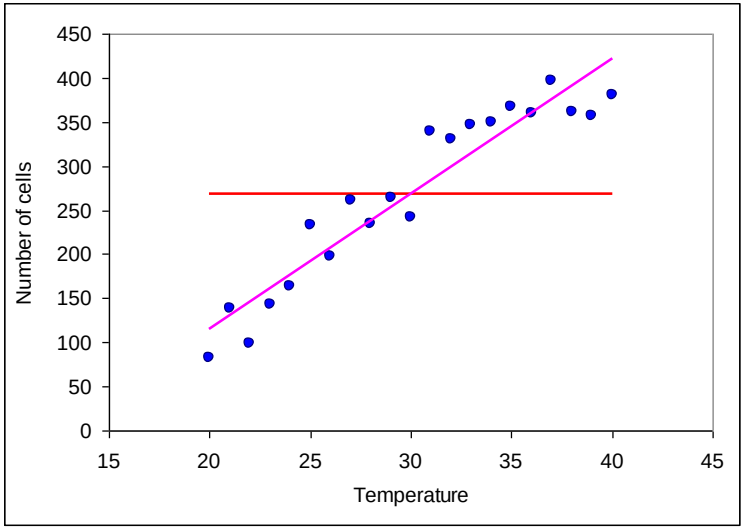
ANOVA



$$SST = SSTR + SSE$$

Observation = μ + Factor + ε

Linear Regression



$$SST = SSR + SSE$$

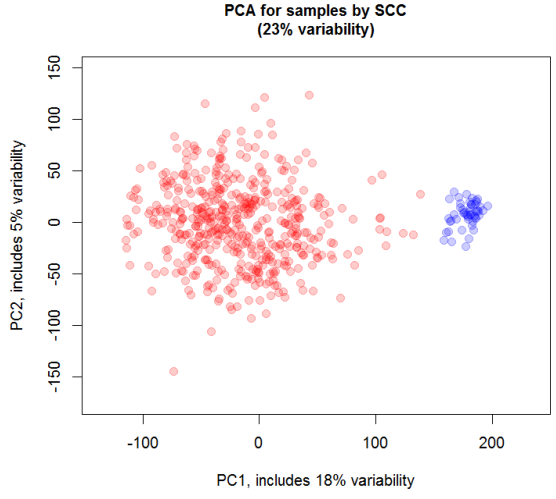
Observation = b_1 * Variable + b_0 + ε

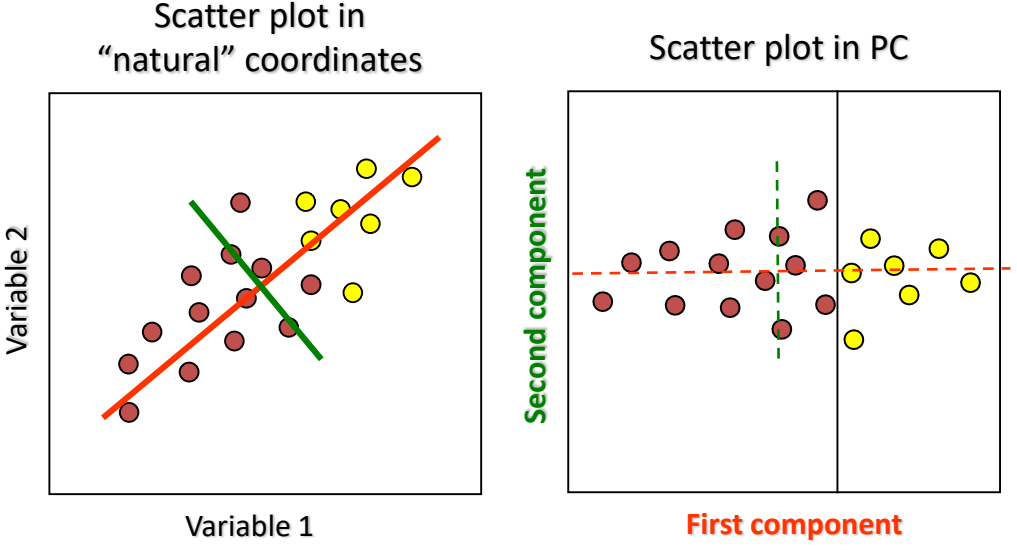
Example: Dimension Reduction

Principal Component Analysis (PCA)

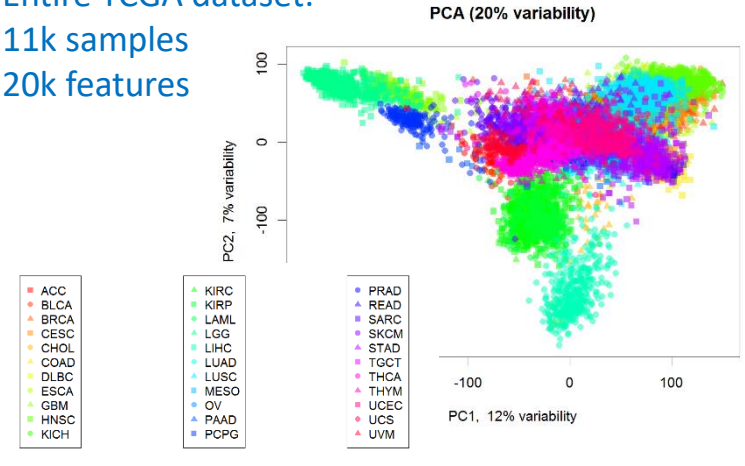
- Deterministic
- Linear (rotation in *n*-dimensional space)
- Captures dimensions with the highest variability
- => sees large differences in the data

Tumor and normal lung tissues





Entire TCGA dataset:
11k samples
20k features



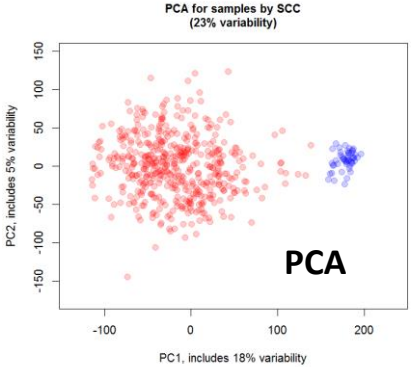
Play with PCA here: https://setosa.io/ev/principal-component-analysis

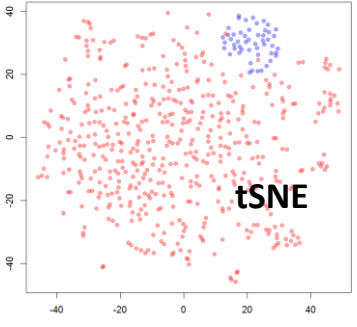
Example: Dimension Reduction

t-distributed Stochastic Neighbor Embedding (tSNE)

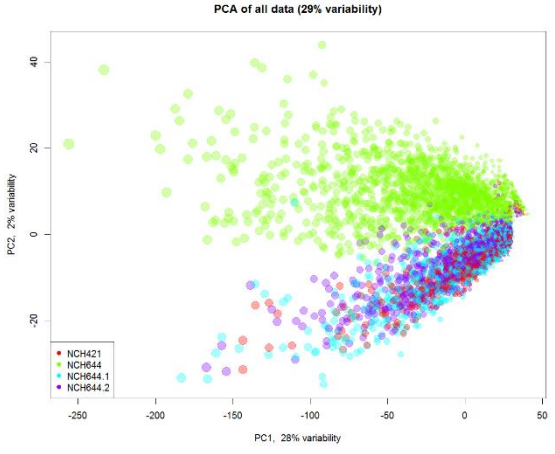
tSNE
nonlinear dimensionality reduction technique that uses local distance instead of global one: similar objects must be close, and distant at any distance above a certain threshold.

- Stochastic
- Non-linear
- Captures similarities, not differences

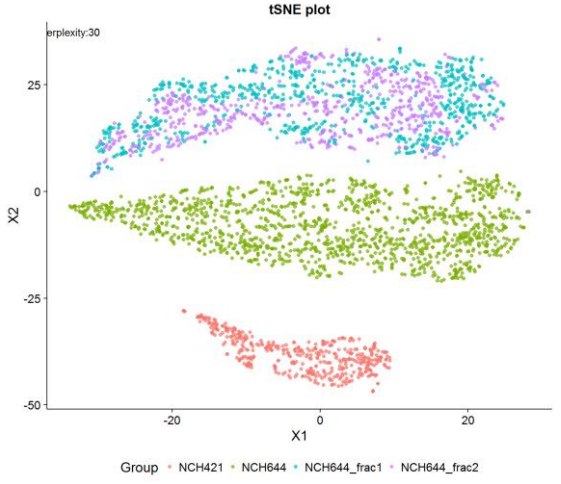




PCA



tSNE



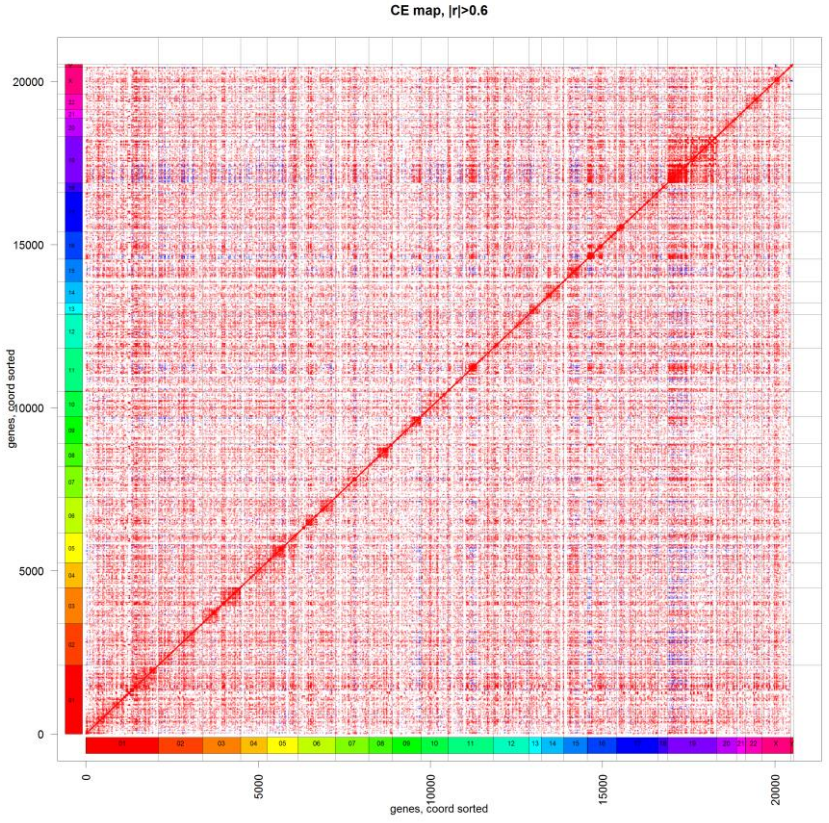
Note: Currently **UMAP** (*uniform manifold approximation and projection for dimension reduction*) is used more often than tSNE. It preserves the topology of the data.

Play with UMAP here: https://pair-code.github.io/understanding-umap/

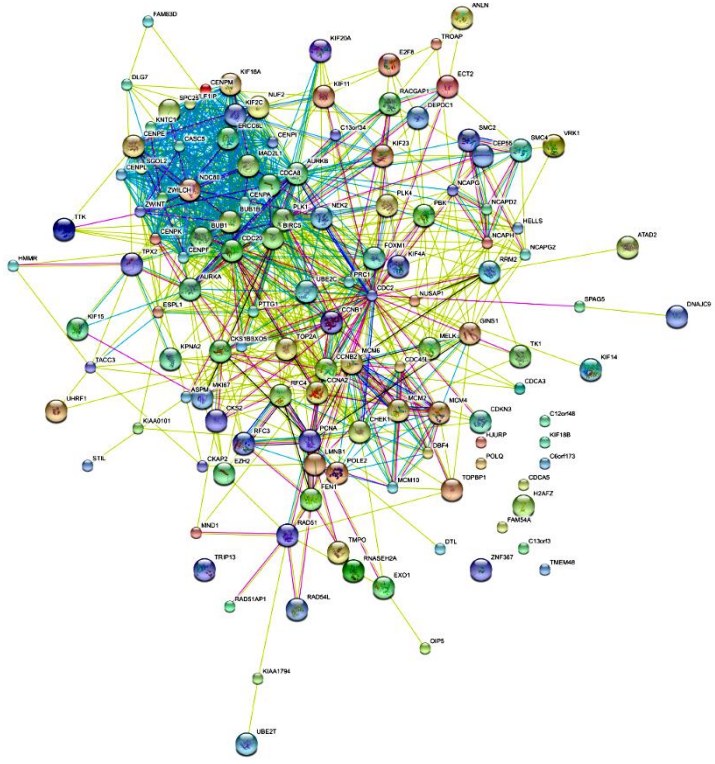
Example: Correlation Analysis

Building Networks of Genes / Proteins

Example: TCGA data, all genes, 9k tumors. Correlation matrix:



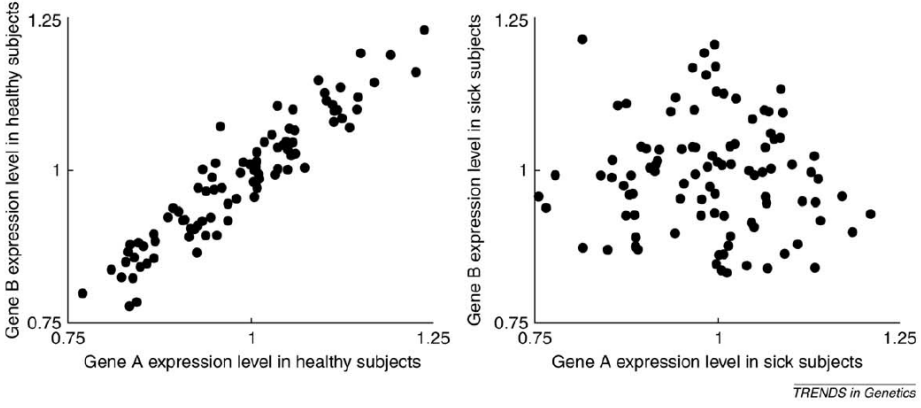
Example: network in [String.DB]

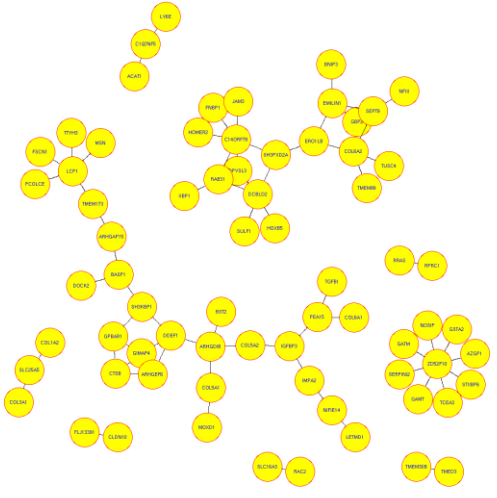


Protein-protein interaction detected in studies (text mining) or predicted by co-expression of genes

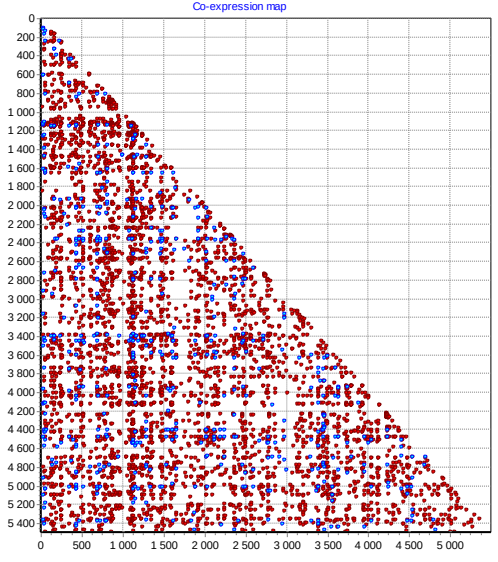
Example: Correlation Analysis

Differential Co-expression Analysis

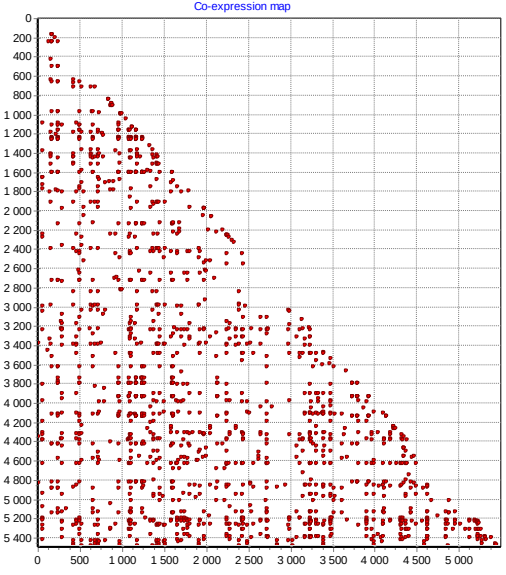




44 normal pancreas (NP)

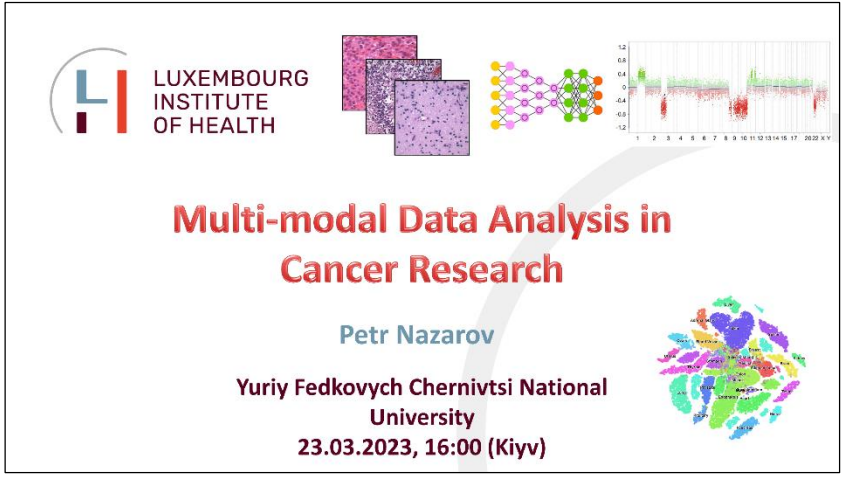


44 ductal adencarcinoma (PDAC)

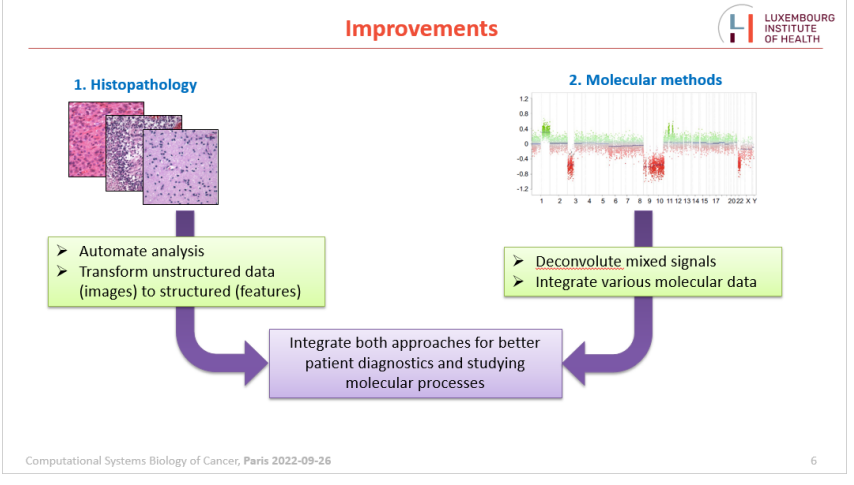


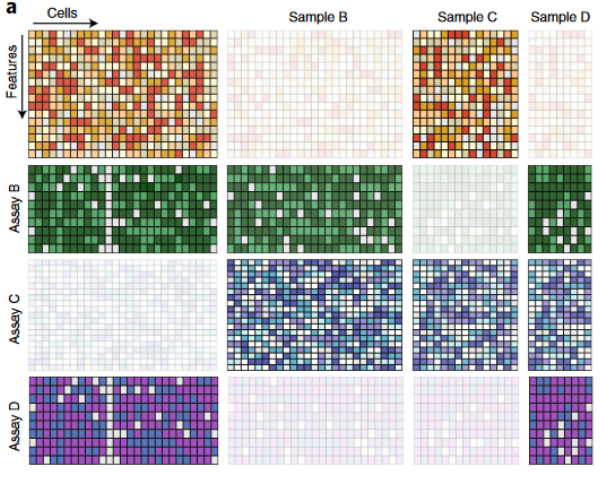
Data Integration

Next Session: 2023-03-23, 16:00 (Kiyv)



https://meet.google.com/bid-cgzt-kag



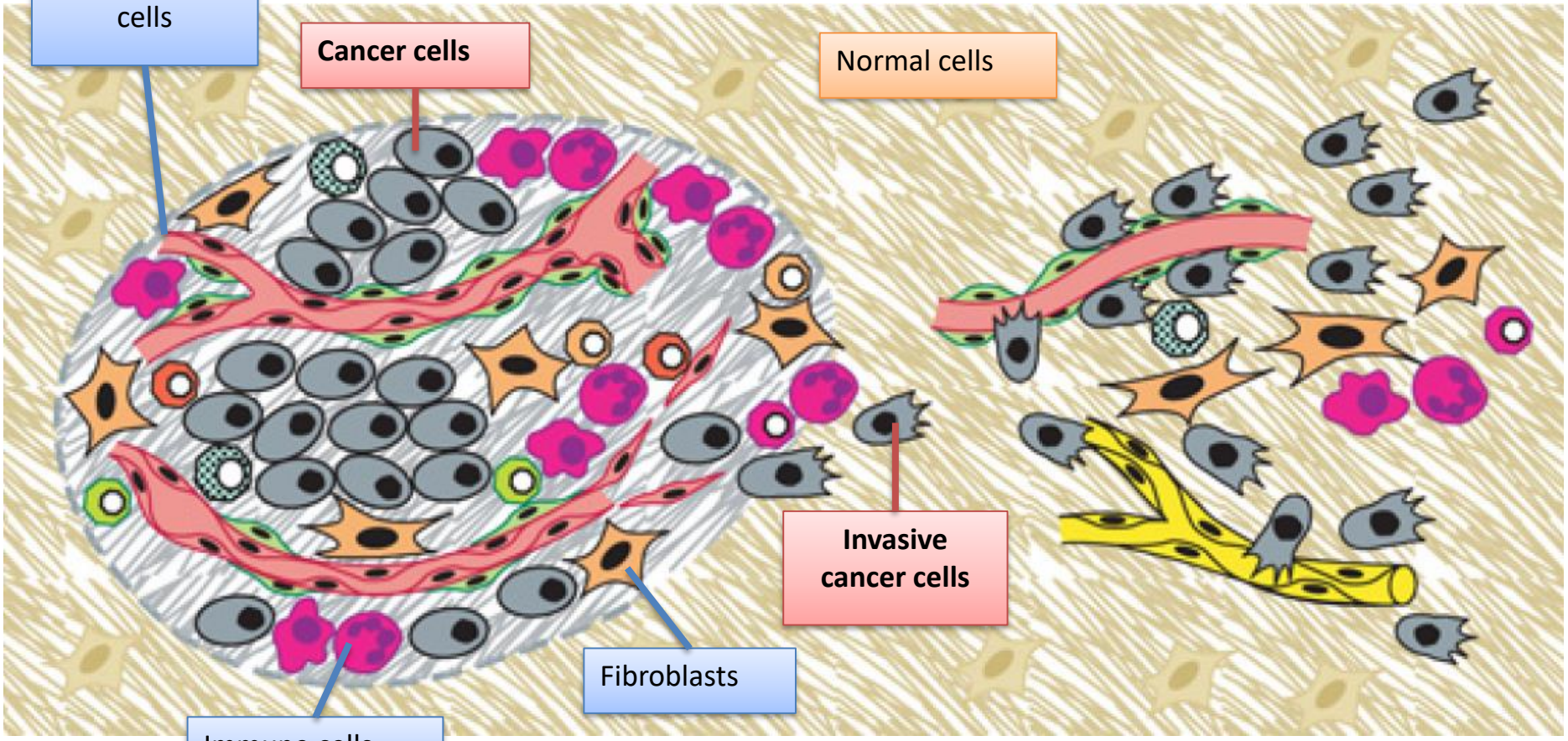


Example:

Independent component analysis (ICA) provides insights into biological processes and clinical outcomes for cancer patients

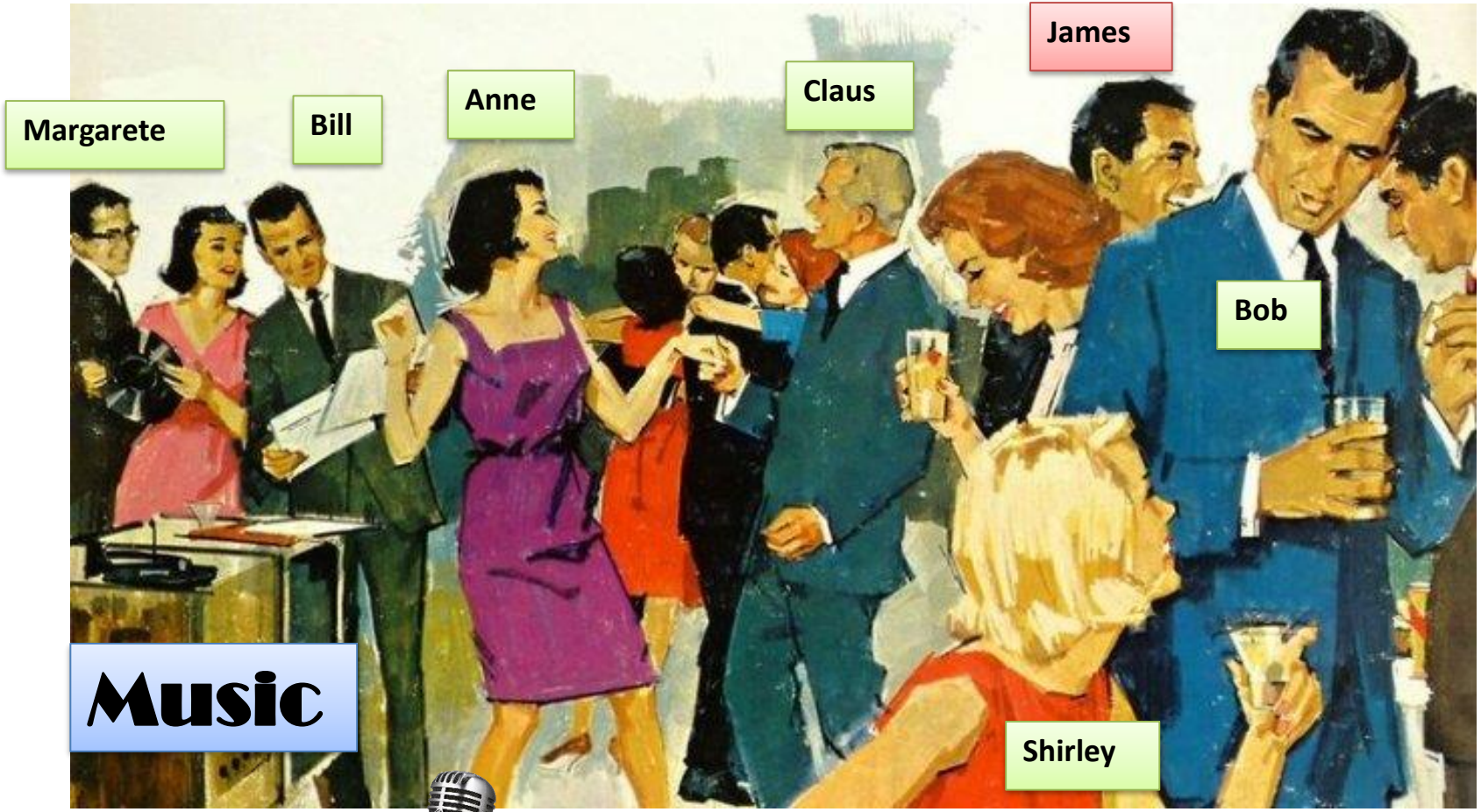
Introduction

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



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

Cocktail Party Problem

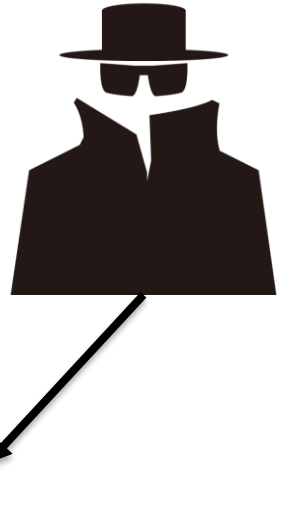


What did James say?..

Independent Component Analysis

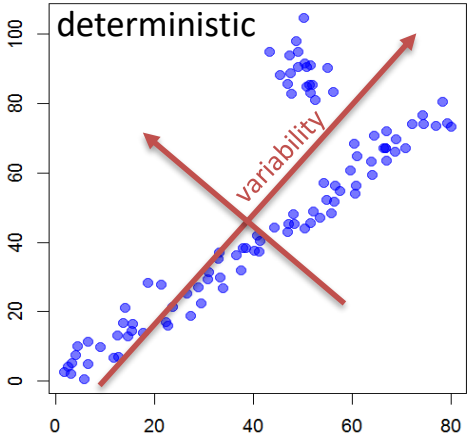
ICA is one of the methods to solve cocktail party problem

Independent Component Analysis



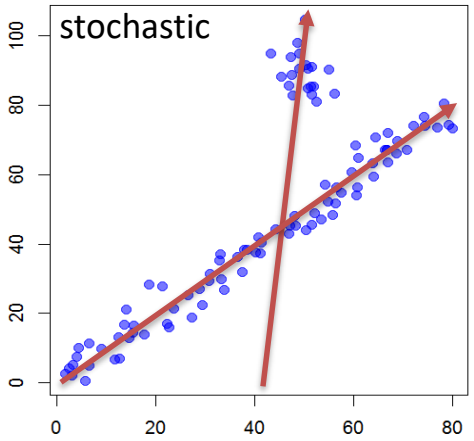
Matrix Factorization Methods Compared

PCA



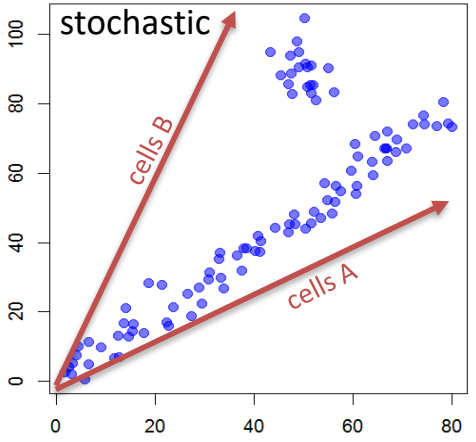
+ deterministic & fast
+ any number of samples
+ unsupervised
– often biological factors are presented by a sum of several components
– positive and negative values

ICA



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

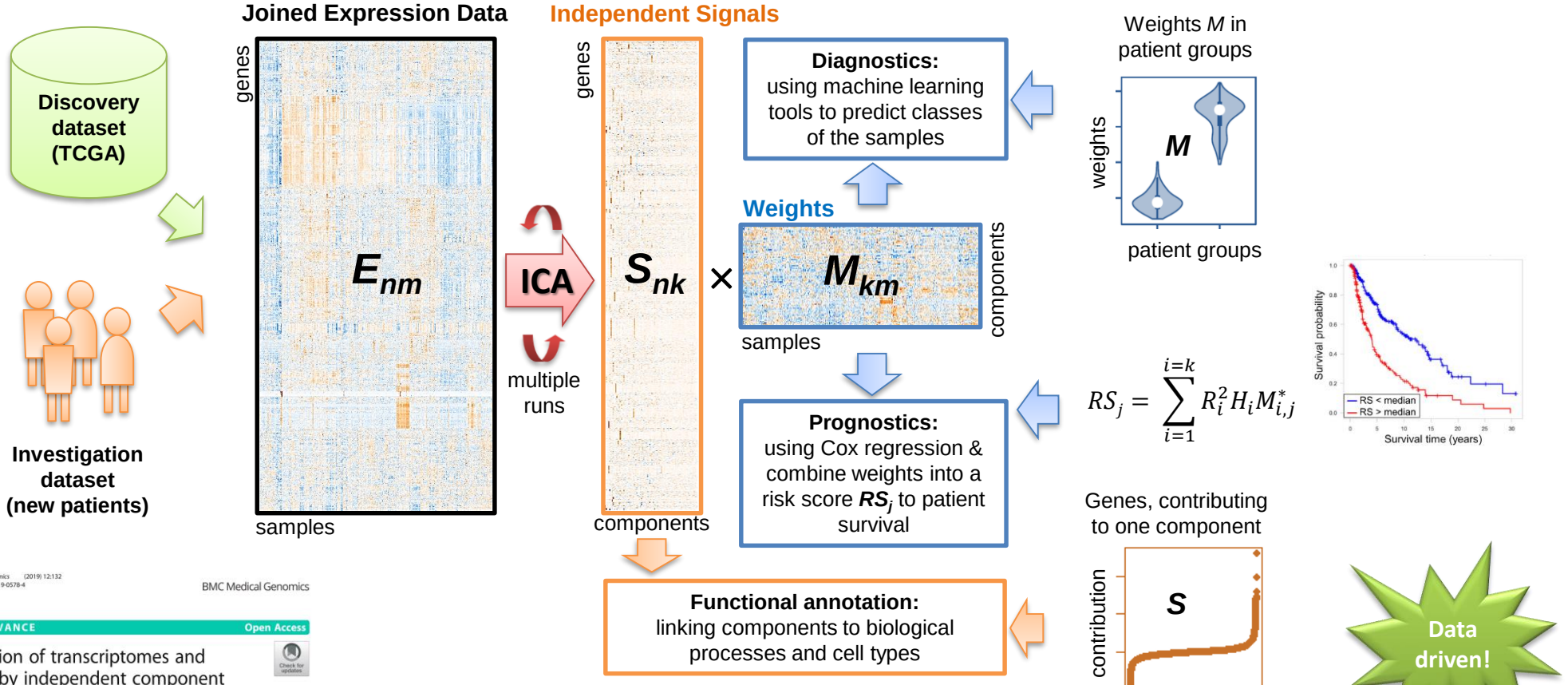
NMF



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

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

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[1], 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

Glioblastoma Cell Lines

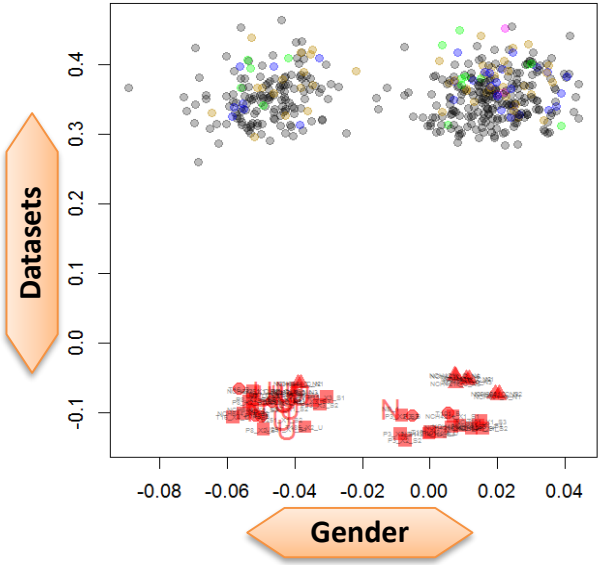
Reference dataset
530 GBM patients
(TCGA)

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

ICA

Biological knowledge:
bio-processes and sample composition

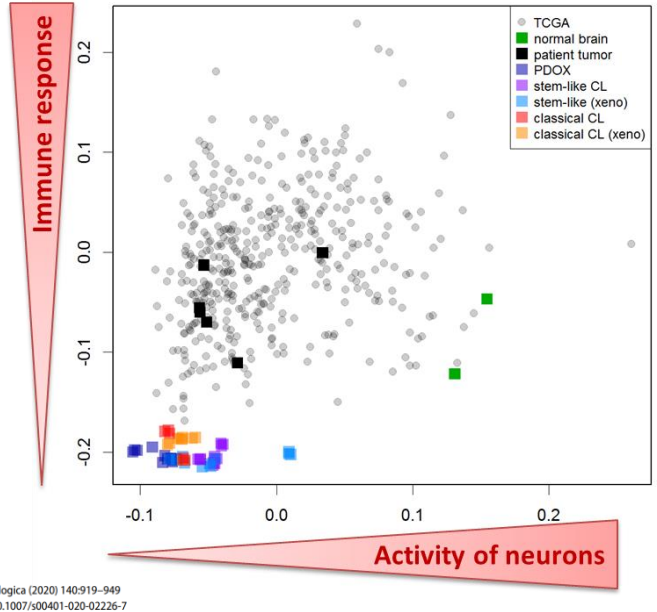
Technical/trivial components: gender and platforms



- 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.

Glioblastoma Cell Lines

ICA correctly predicts sample composition & phenotype

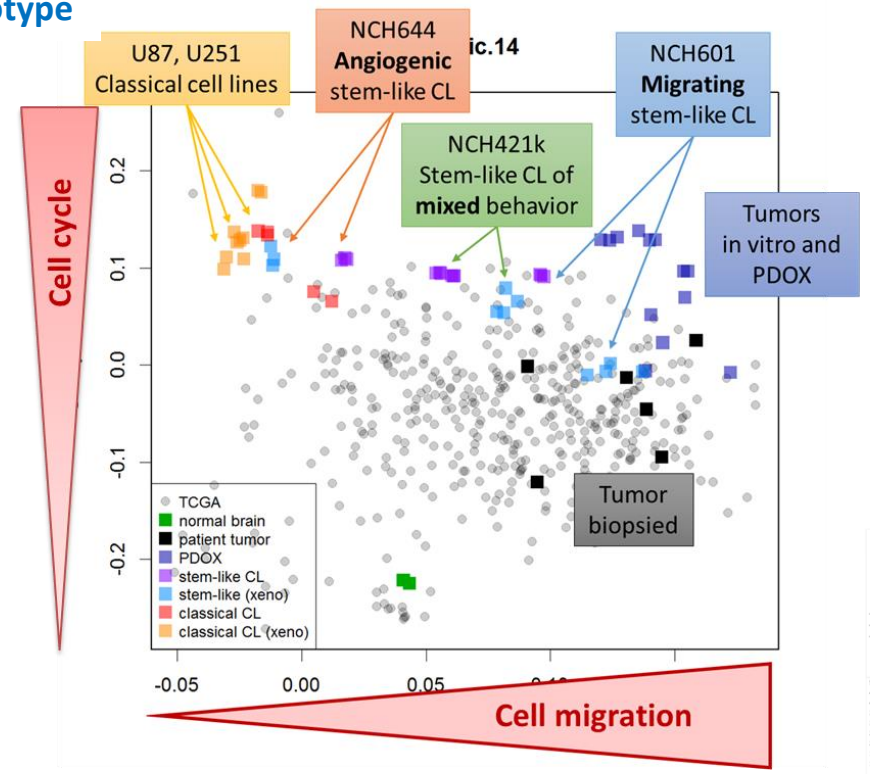


Acta Neuropathologica (2020) 140:919–949
https://doi.org/10.1007/s00401-020-02226-7

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]

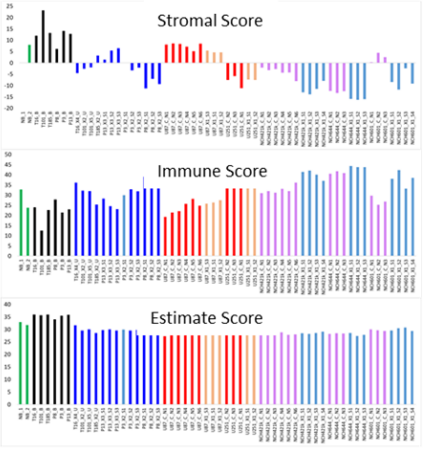


Golebiewska A. et al, **Acta Neuropathologica, 2020** (link)

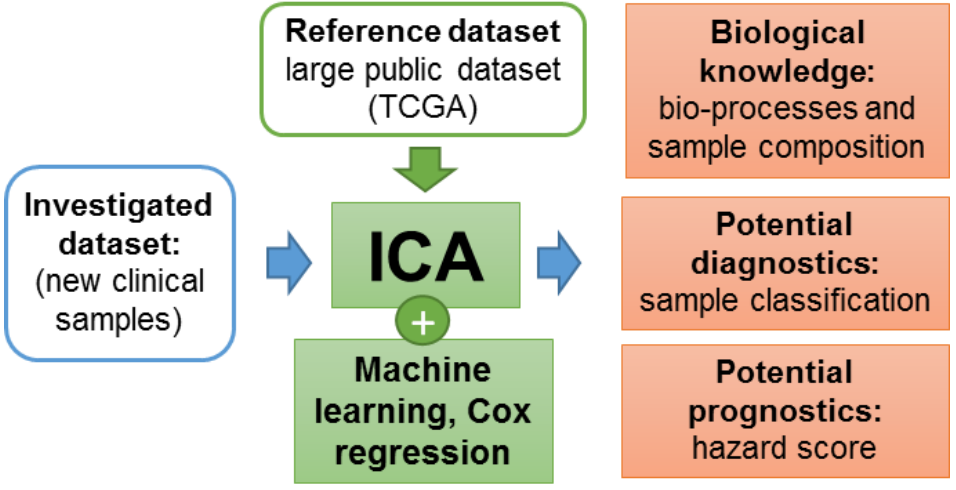
Phenotype of cell lines were predicted using unsupervised deconvolution of their transcriptomes!

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

ESTIMATE was confused

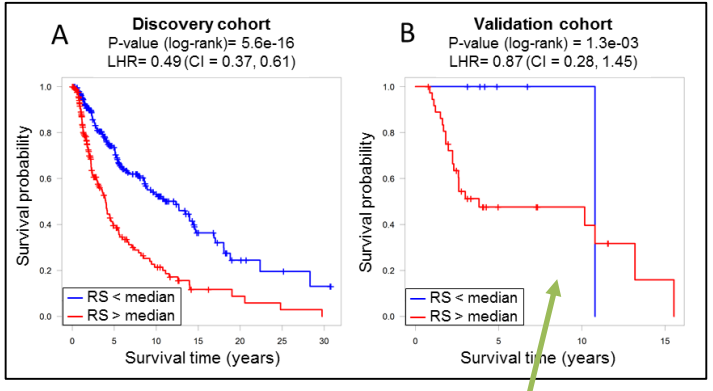


Melanoma



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

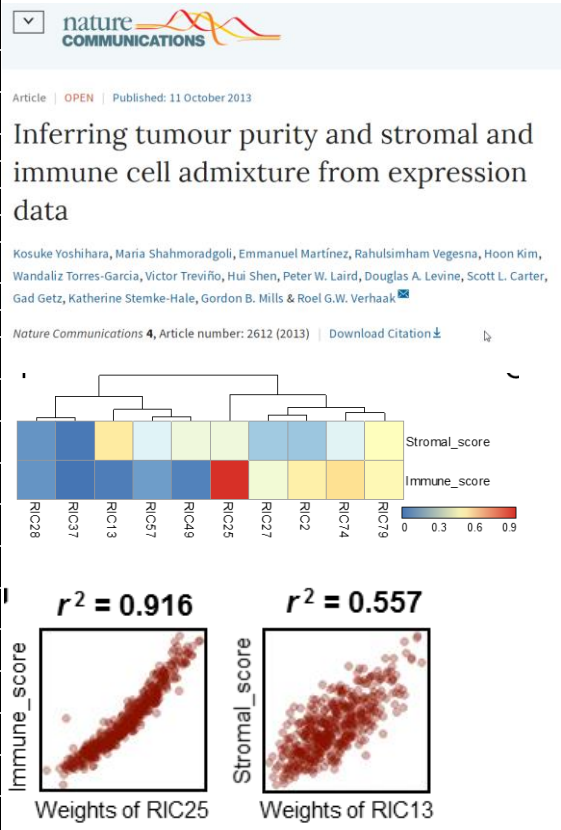
- In addition to diagnostics and prognostics, ICA allowed ranking patients based on the activity of biological processes: cell cycle, signals of leukocytes, etc.

Nazarov et al, BMC Medical Genomics, 2019

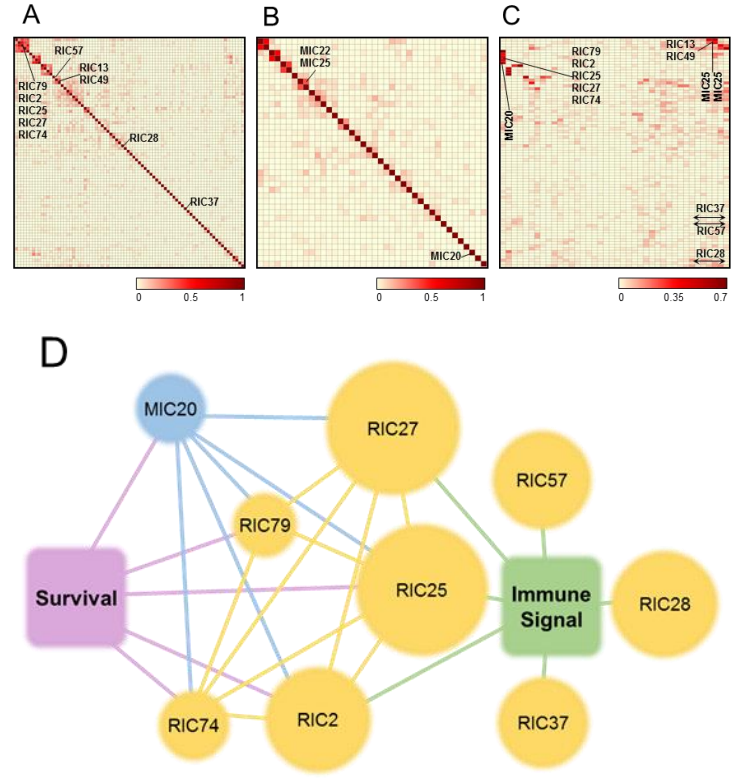
Deciphering biological processes and cell types

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.

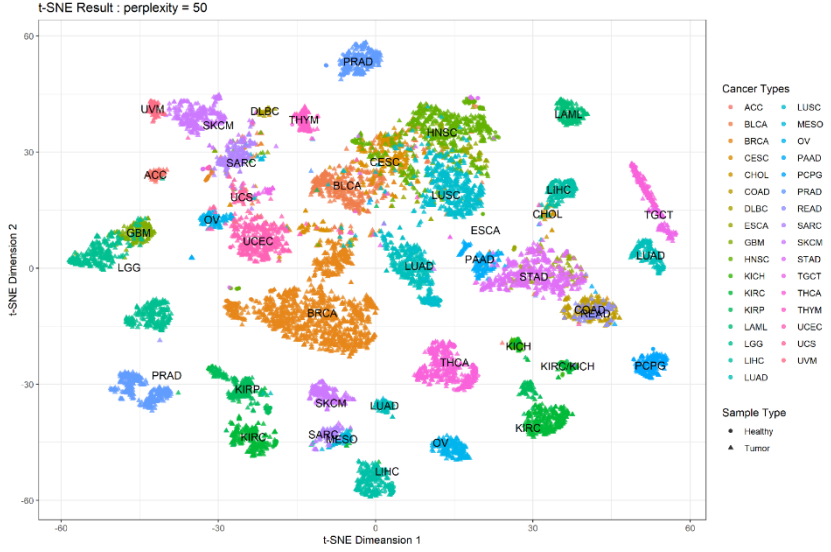
Pan-Cancer Data Integration

TCGA

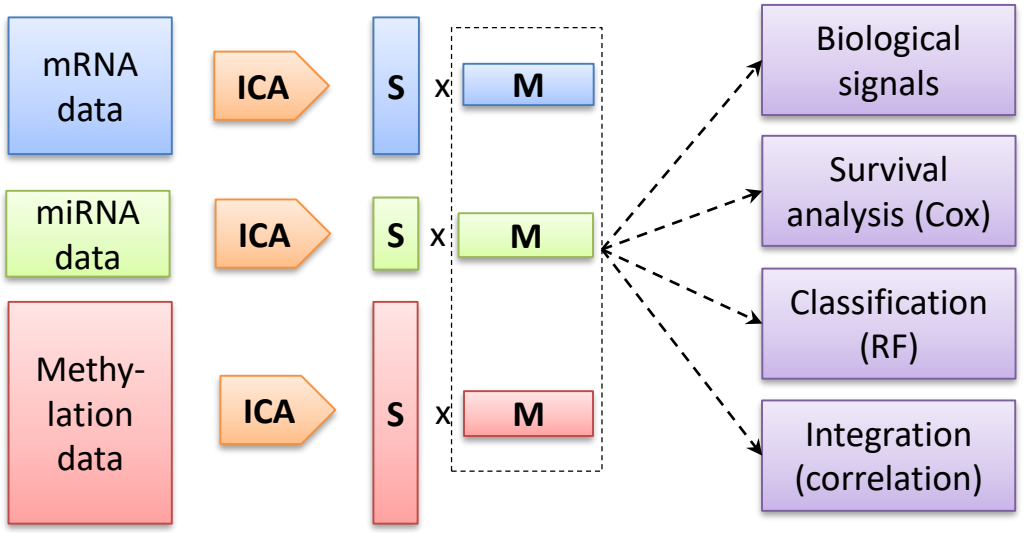
The Cancer Genome Atlas

>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)



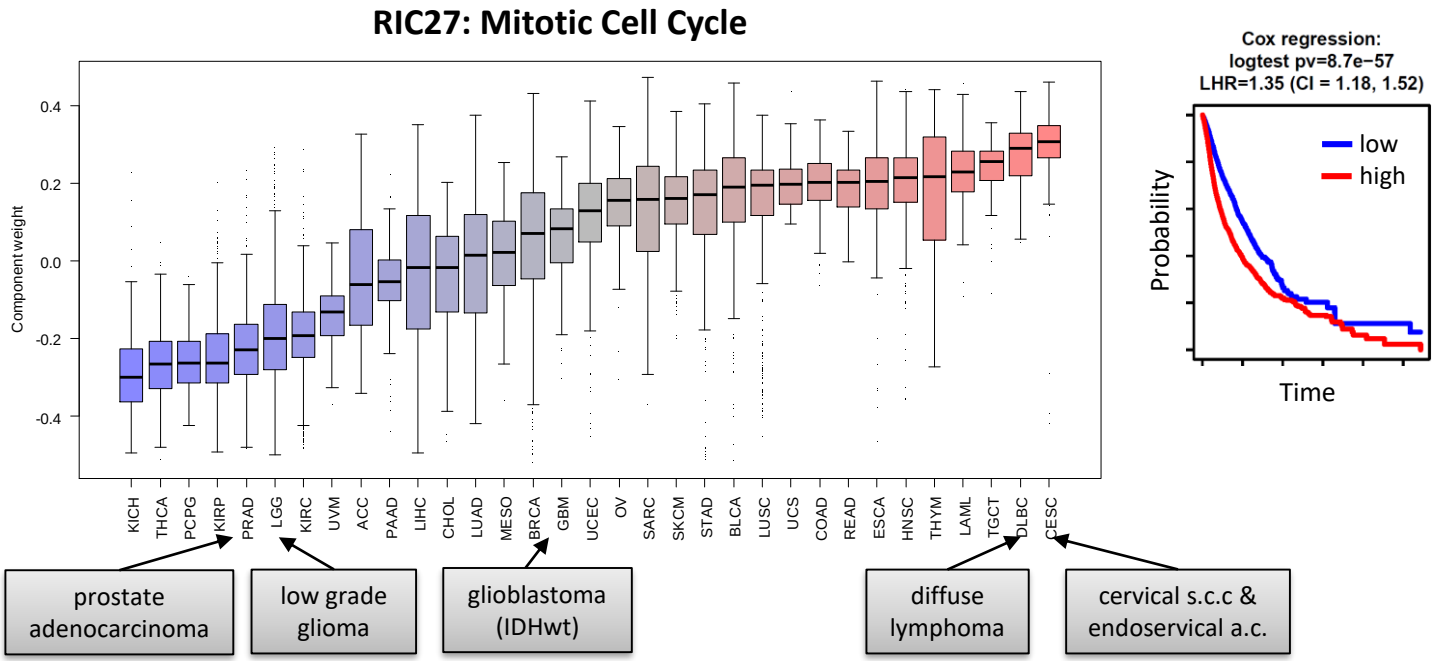
Approach



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

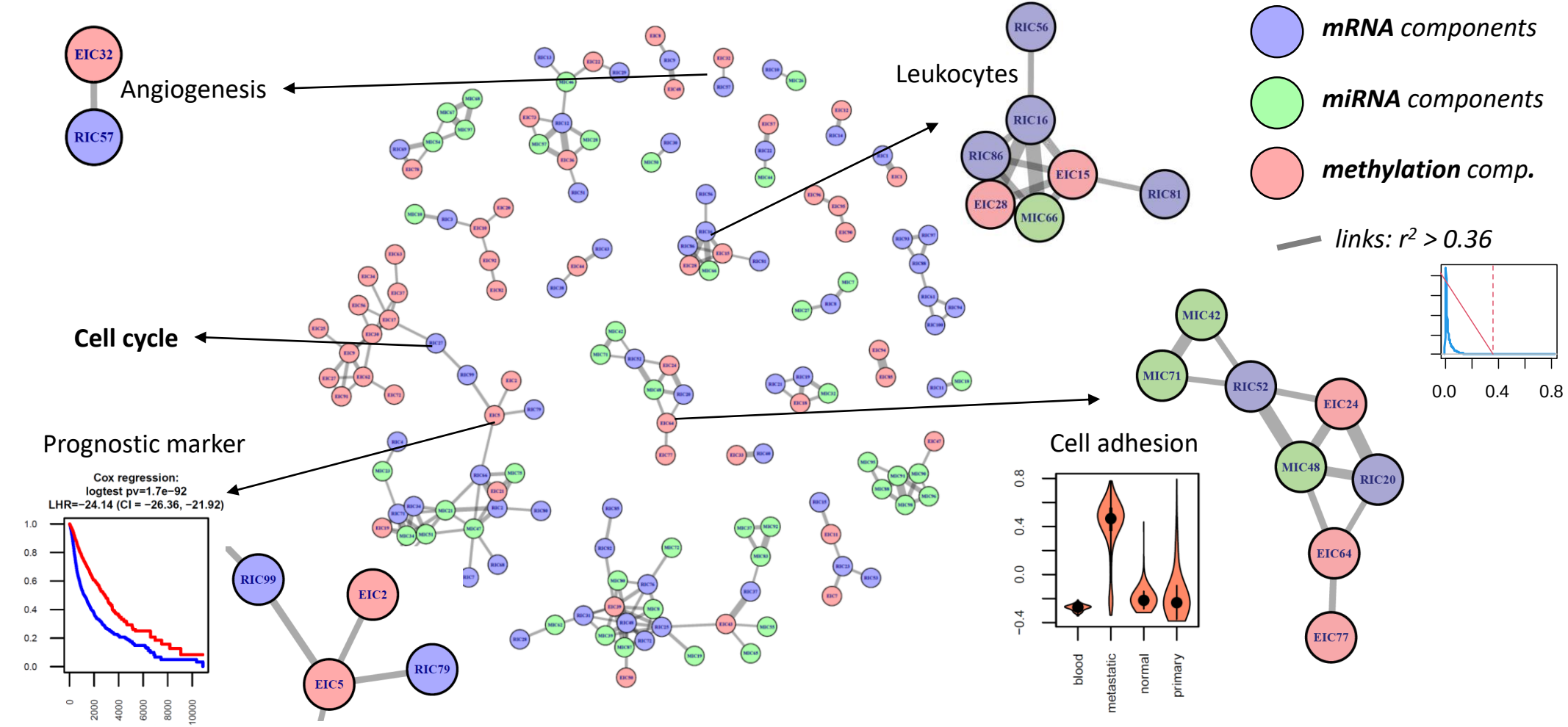
Pan-cancer: ICA Components

ICA Results: Cell Cycle

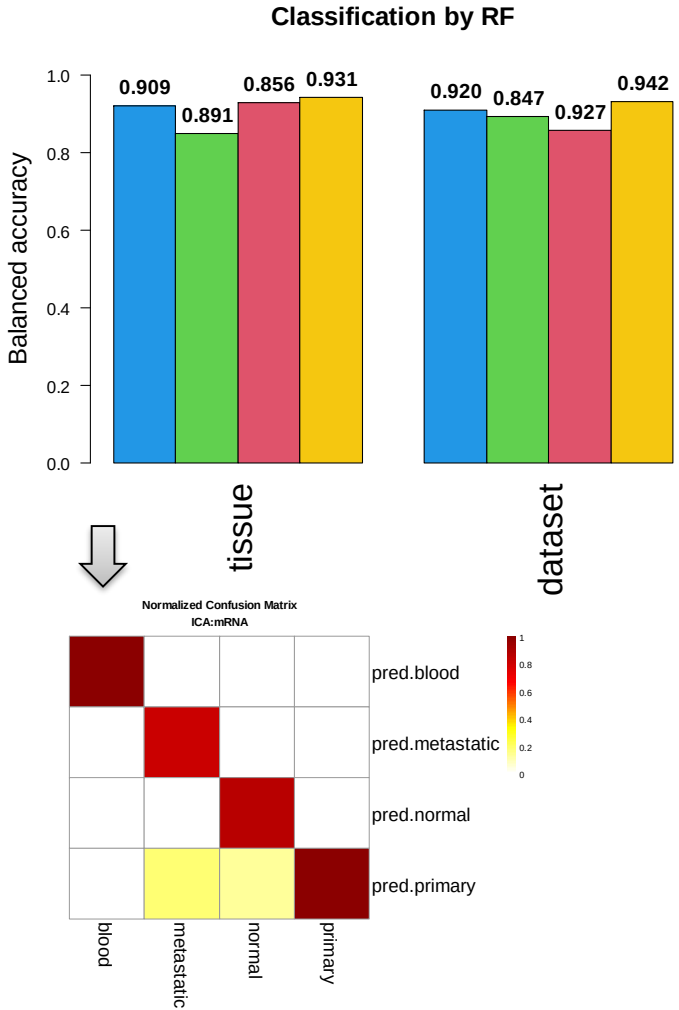


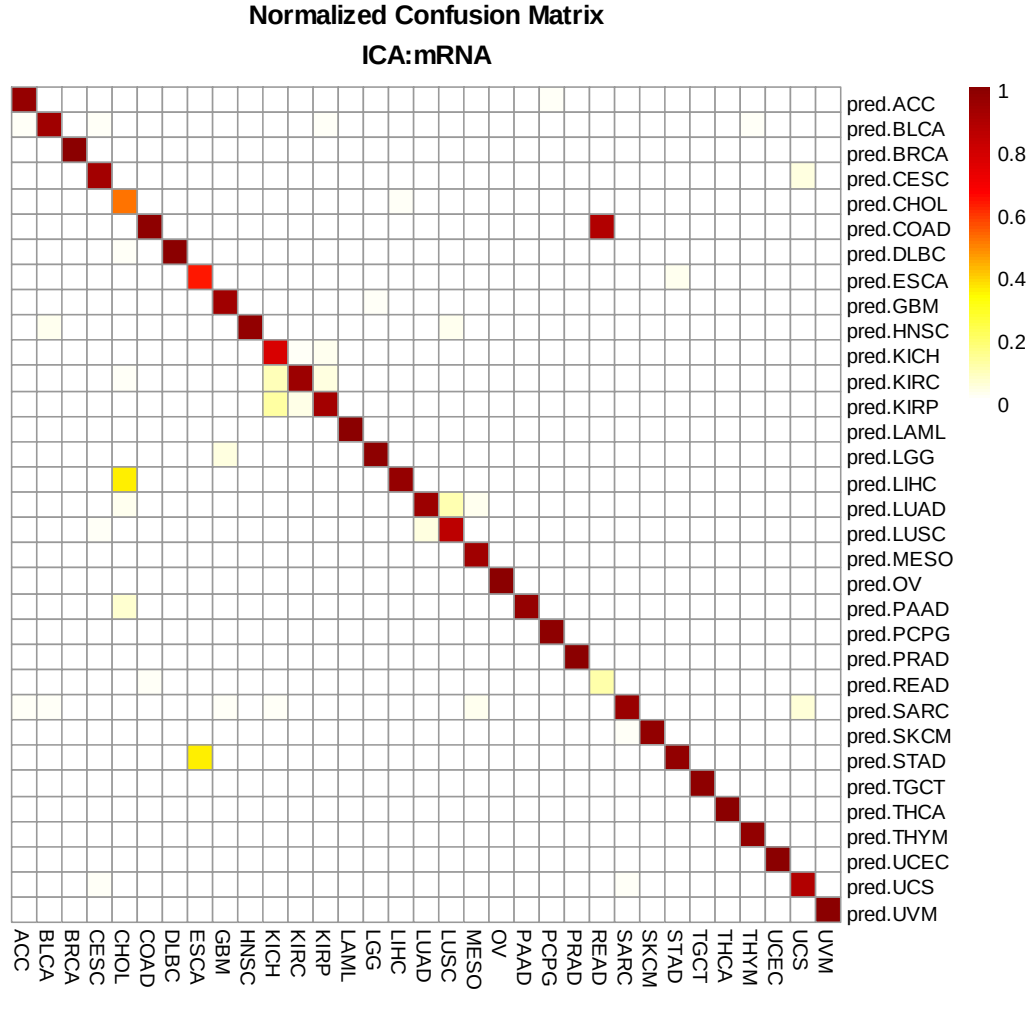
Code	Study Name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical sq. cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large b-cell lymphoma
ESCA	Esophageal carcinoma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



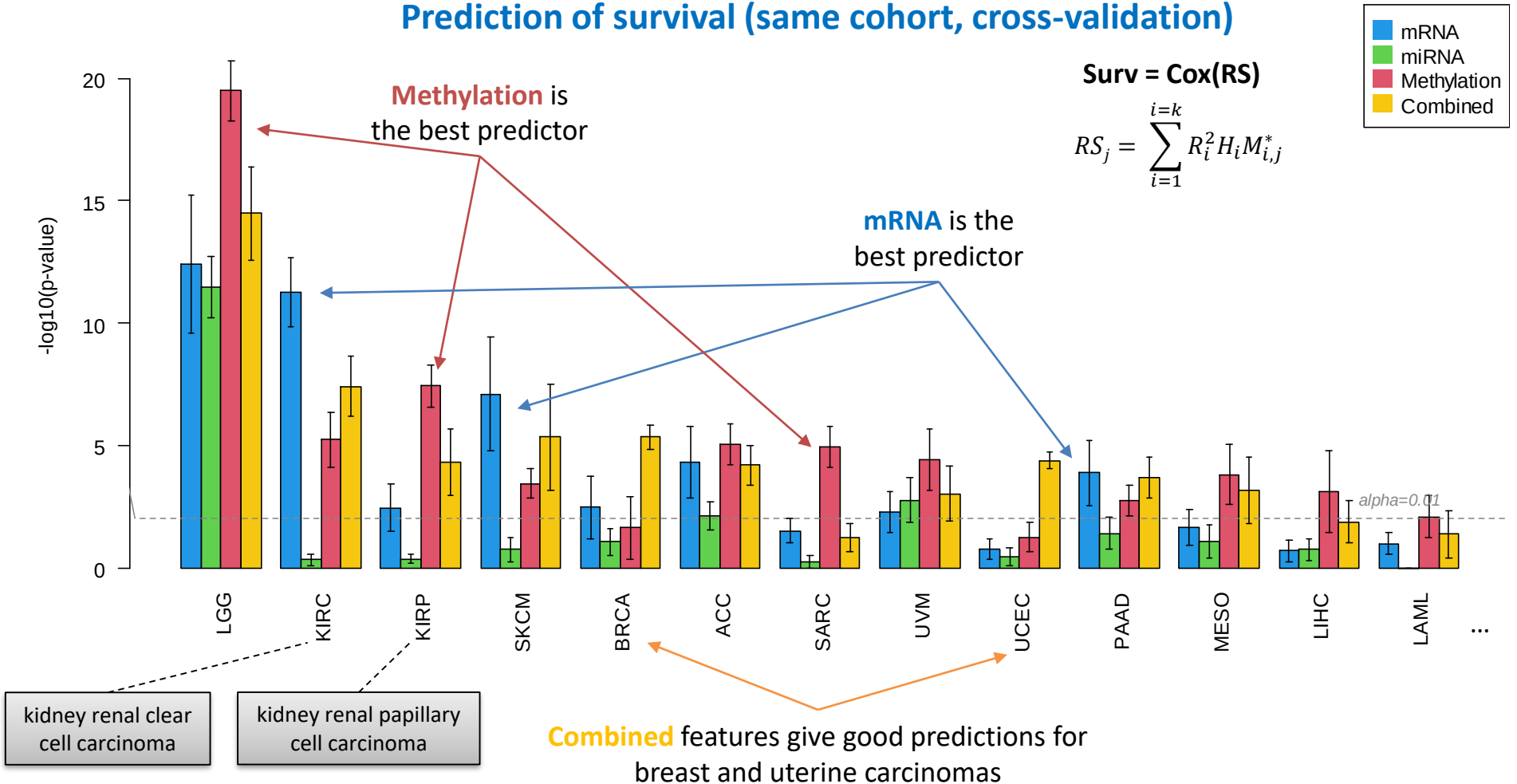
Pan-cancer: Classification





Pan-cancer: Prognosis

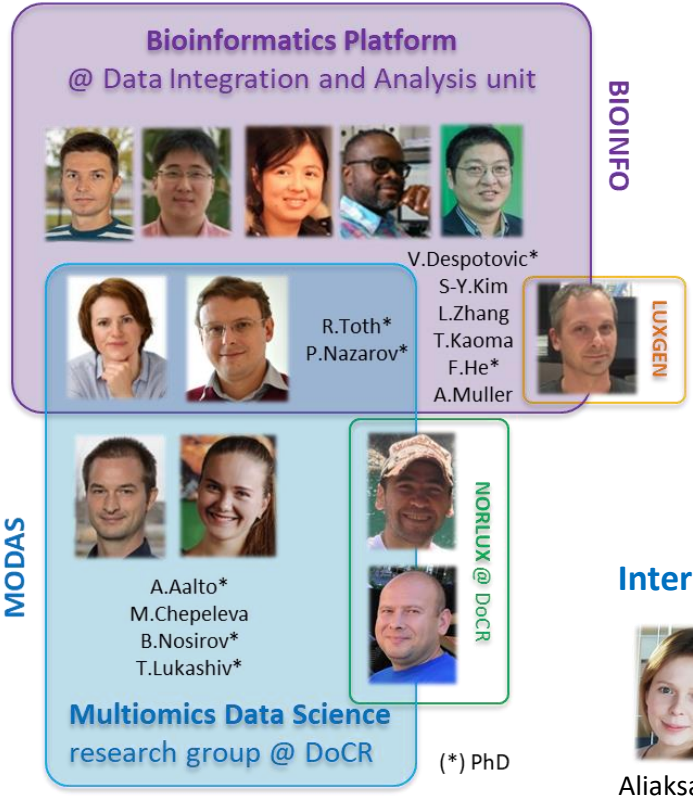
Prediction of survival (same cohort, cross-validation)



Take Home Messages from This Example

- 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

Acknowledgements



Key internal collaborators


Prof. Simone Niclou


Dr. Anna Golebiewska


Prof. Michel Mittelbronn


Prof. Gunnar Dittmar

Key external collaborators


LSRU, Uni Luxembourg
Prof. Stephanie Kreis


Institute Curie, France
Dr. Andrei Zinovyev




DKFZ, Heidelberg
Dr. Jörg Hoheisel
Dr. Andrea Bauer
Prof. 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: