BELARUSIAN STATE UNIVERSITY

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Proteome and Genome Research Unit

Data Analysis and Modelling in Transcriptomics:

Petr Nazarov

petr.nazarov@lih.lu

edu.sablab.net

Fonds National de la Recherche Luxembourg

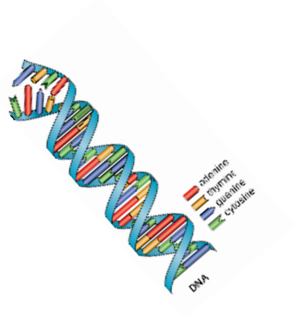
2018-11-06, Belarus State University, Minsk

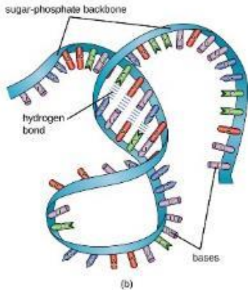
Outline

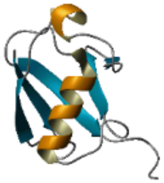
- **Concept & Data**
  - Central dogma of information transfer…
  - …and how it is implemented (to the current knowledge)
  - Data examples
- **Models**
  - Original question of 2003: “can a biologist fix a radio?”
  - Some models frequently used
- **Methods in transcriptomics**
  - Statistics and linear models
  - Dimensionality reduction: PCA and tSNE
- **Example**
  - independent component analysis for signal separation

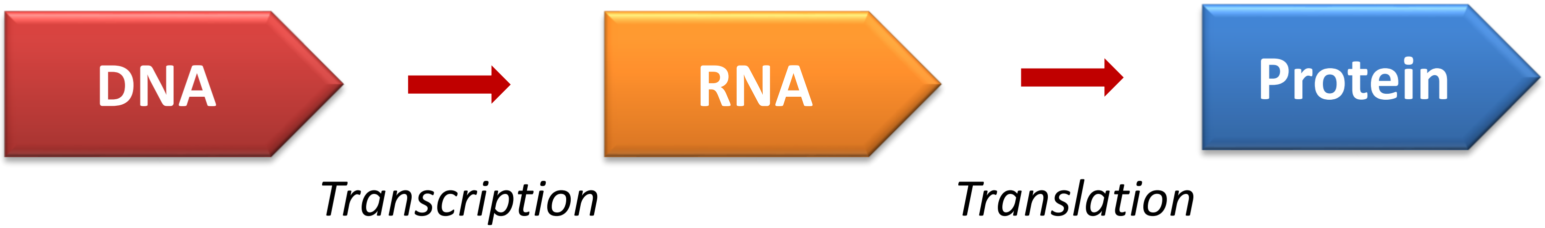
Concept and the Data

Central Dogma of Biology





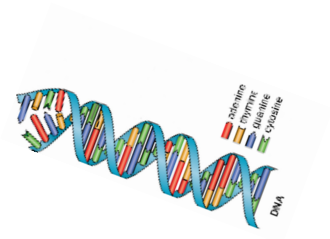




Adapted from:

- 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

More Realistic Central Dogma



DNA

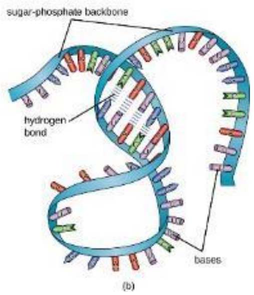
...ATGCTATATGC...
| | | | | | | | | | |
...TACGATATACG...

→

Methylation and chromatin remodeling

Is the sequence accessible for reading?
Cell differentiation

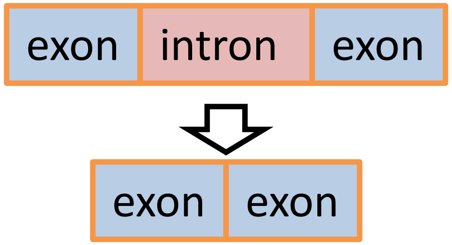
→



RNA

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

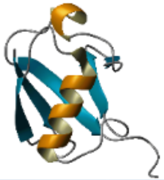
→



RNA splicing & RNA maturation

The source of protein diversity: 20k genes → 80k proteins.
(?) Data preservation

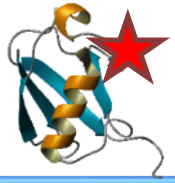
→



Protein

The main building blocks & nano-machinery

→

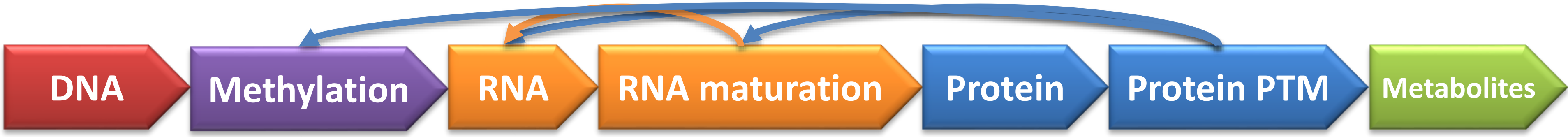


Protein post translational modification

Activation of proteins

Central Dogma and the Data

Even More Realistic



DNA

Methylation

RNA

RNA maturation

Protein

Protein PTM

Metabolites

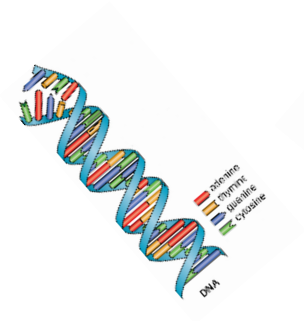
Genetic data

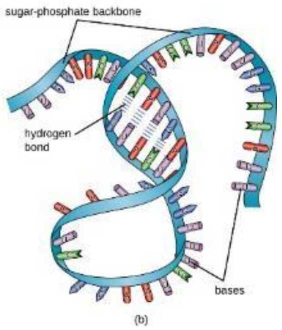
Epigenetic data

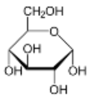
Transcriptomic data

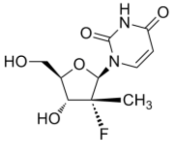
Proteomic data

Metabolomic data









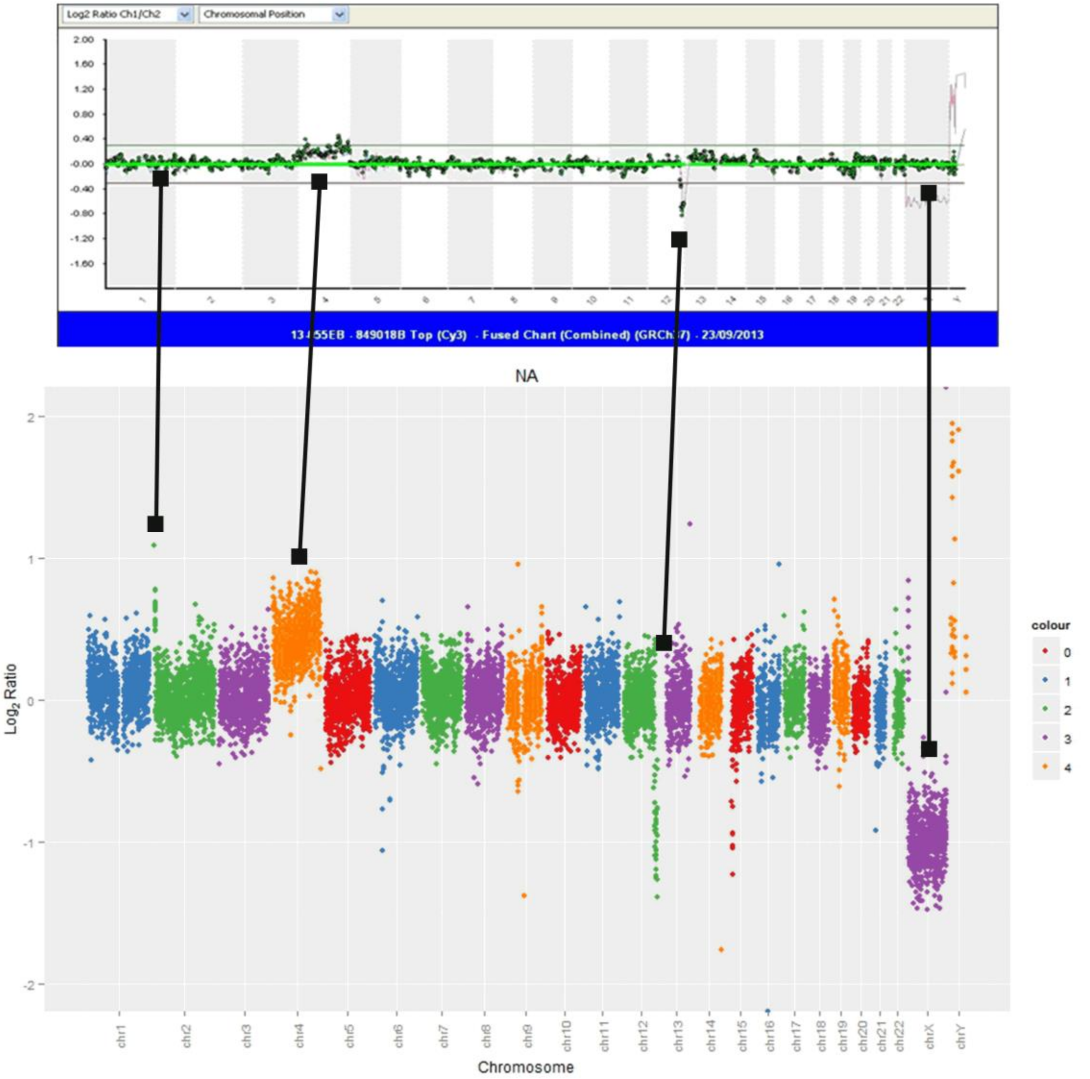
It is impossible to use these levels of data without proper: **Clinical & histological data**

Data

DNA: Copy Number Variation (CNV) Data

DNA level data can be presented as:

- a single mutation (e.g. SNP)
- copy number variation (CNV)



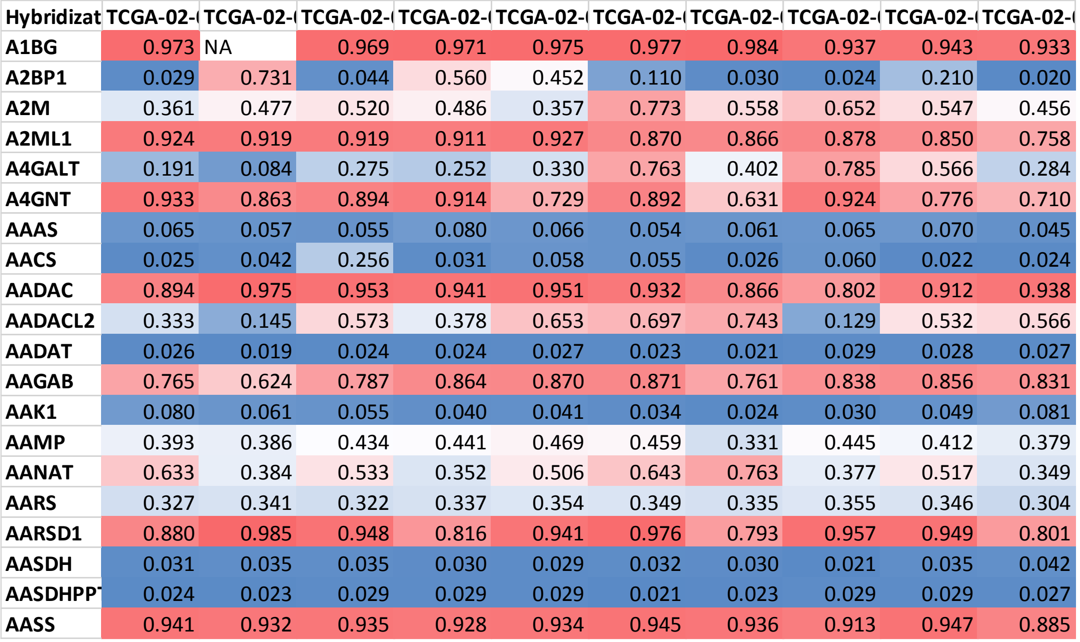
Data

Epigenetic: Methylation Data

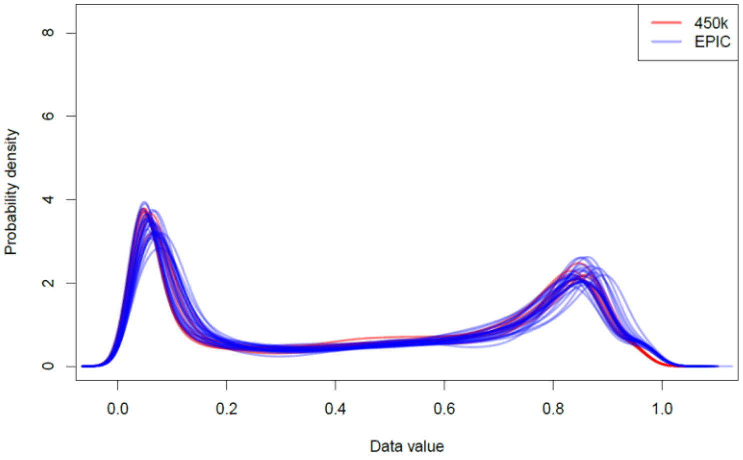
Epigenetic level:

- DNA methylation (cytosine)
- histone modifications (a lot!)

DNA methylation data



Hybridizat	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(	TCGA-02-(
A1BG	0.973	NA	0.969	0.971	0.975	0.977	0.984	0.937	0.943	0.933
A2BP1	0.029	0.731	0.044	0.560	0.452	0.110	0.030	0.024	0.210	0.020
A2M	0.361	0.477	0.520	0.486	0.357	0.773	0.558	0.652	0.547	0.456
A2ML1	0.924	0.919	0.919	0.911	0.927	0.870	0.866	0.878	0.850	0.758
A4GALT	0.191	0.084	0.275	0.252	0.330	0.763	0.402	0.785	0.566	0.284
A4GNT	0.933	0.863	0.894	0.914	0.729	0.892	0.631	0.924	0.776	0.710
AAAS	0.065	0.057	0.055	0.080	0.066	0.054	0.061	0.065	0.070	0.045
AACS	0.025	0.042	0.256	0.031	0.058	0.055	0.026	0.060	0.022	0.024
AADAC	0.894	0.975	0.953	0.941	0.951	0.932	0.866	0.802	0.912	0.938
AADACL2	0.333	0.145	0.573	0.378	0.653	0.697	0.743	0.129	0.532	0.566
AADAT	0.026	0.019	0.024	0.024	0.027	0.023	0.021	0.029	0.028	0.027
AAGAB	0.765	0.624	0.787	0.864	0.870	0.871	0.761	0.838	0.856	0.831
AAK1	0.080	0.061	0.055	0.040	0.041	0.034	0.024	0.030	0.049	0.081
AAMP	0.393	0.386	0.434	0.441	0.469	0.459	0.331	0.445	0.412	0.379
AANAT	0.633	0.384	0.533	0.352	0.506	0.643	0.763	0.377	0.517	0.349
AARS	0.327	0.341	0.322	0.337	0.354	0.349	0.335	0.355	0.346	0.304
AARSD1	0.880	0.985	0.948	0.816	0.941	0.976	0.793	0.957	0.949	0.801
AASDH	0.031	0.035	0.035	0.030	0.029	0.032	0.030	0.021	0.035	0.042
AASDHPP	0.024	0.023	0.029	0.029	0.029	0.021	0.023	0.029	0.029	0.027
AASS	0.941	0.932	0.935	0.928	0.934	0.945	0.936	0.913	0.947	0.885



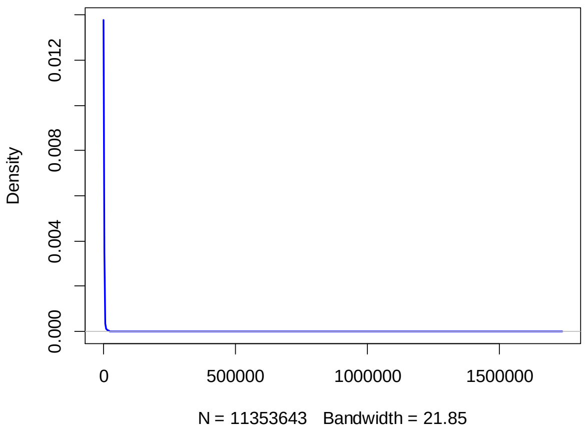
Data

RNA: Gene Expression Data

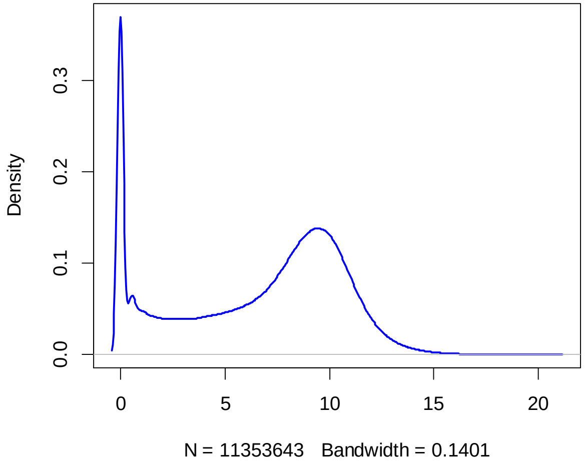
RNA abundance (expression) in counts

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

Distribution of counts

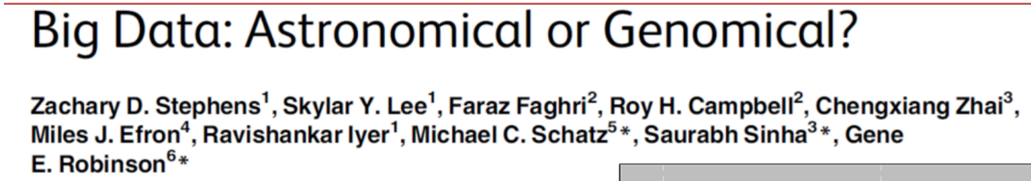


Distribution of log-counts



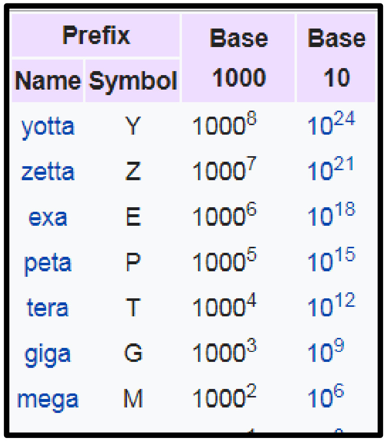
The most straight-forward data ☺

Exponential Growth



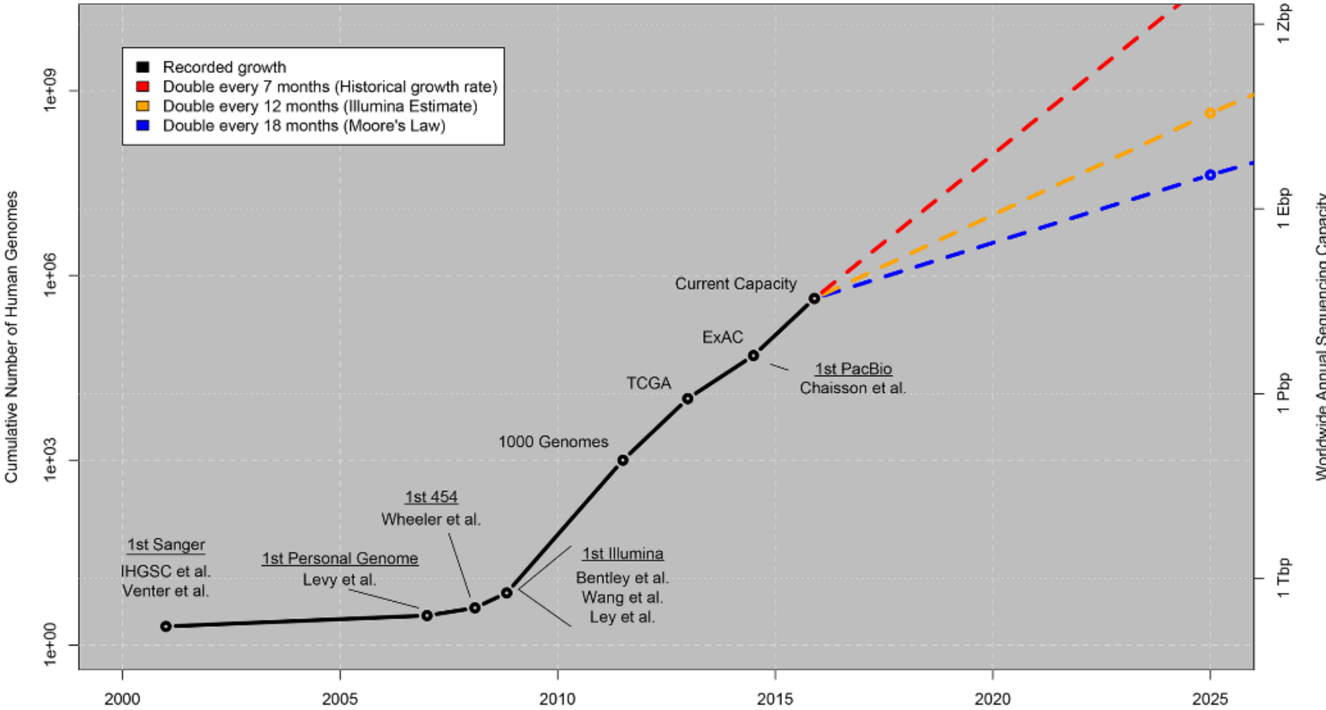
Big Data: Astronomical or Genomical?

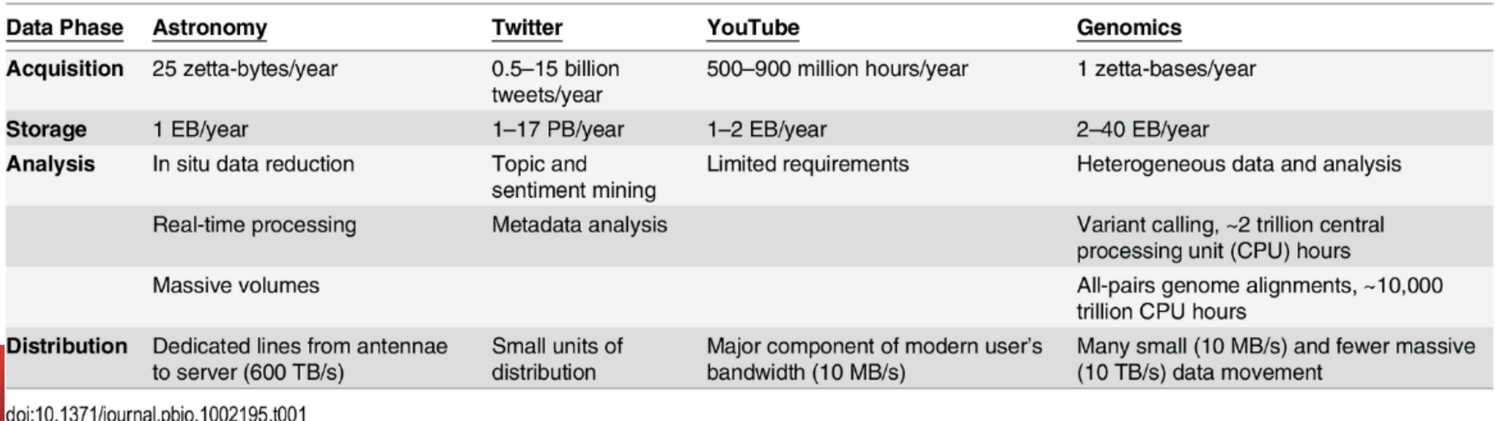
Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*



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

Growth of DNA Sequencing



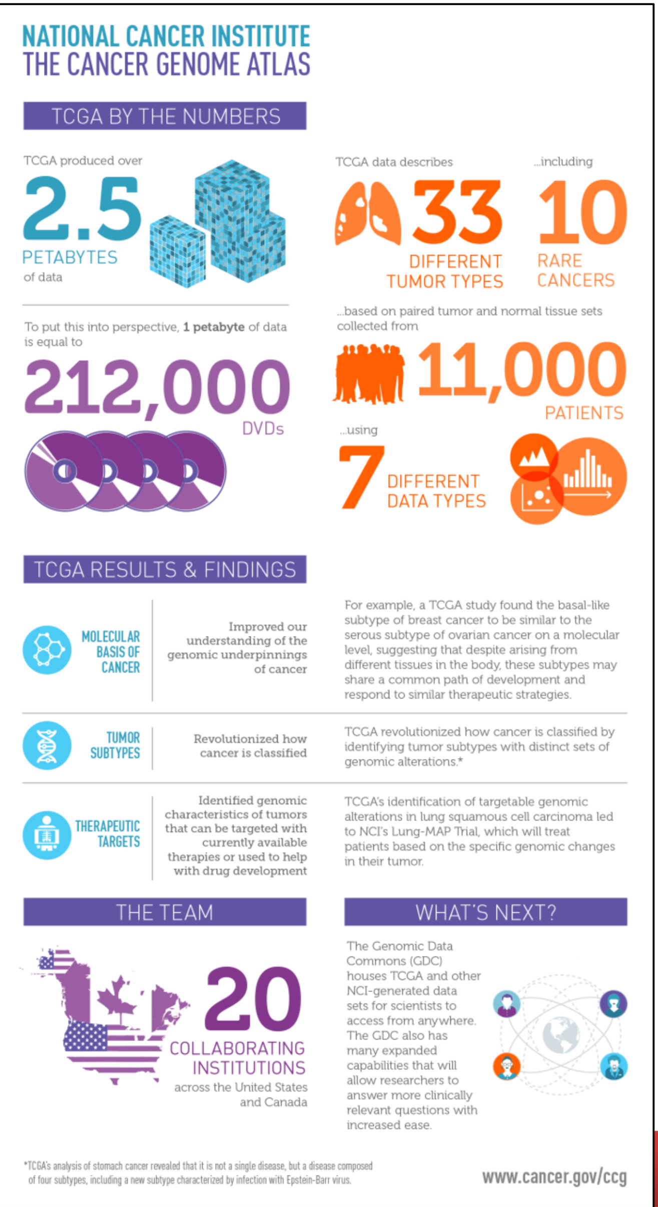


Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous 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

Data Repositories

Examples of Open Repositories



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, 1 **petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER — Improved our understanding of the genomic underpinnings of cancer — For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TUMOR SUBTYPES — Revolutionized how cancer is classified — TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*

THERAPEUTIC TARGETS — Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development — TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

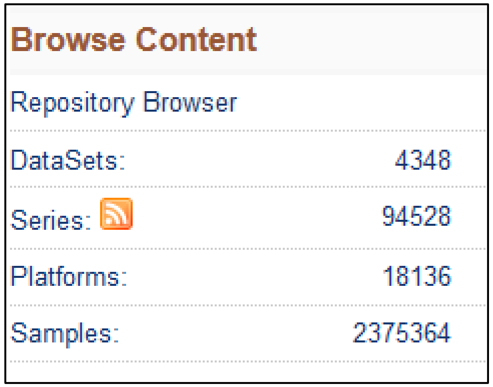
WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

GEO Gene Expression Omnibus



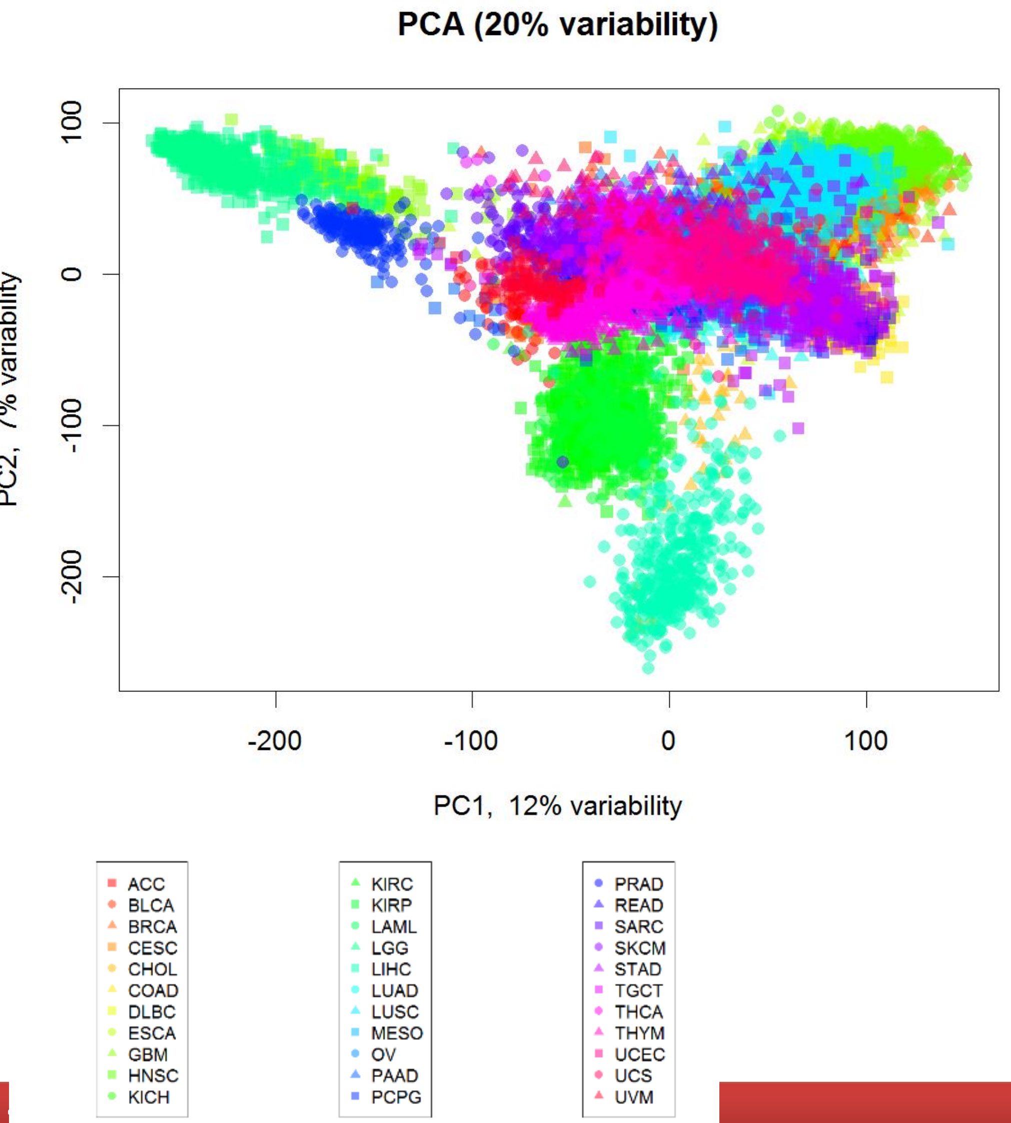
Browse Content

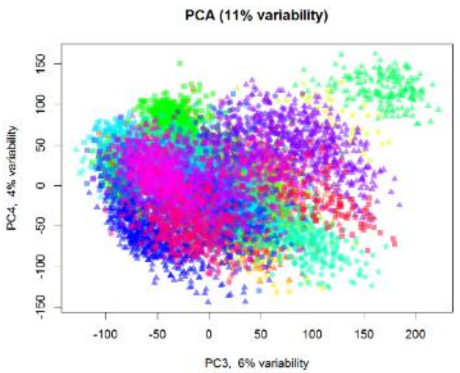
Repository Browser

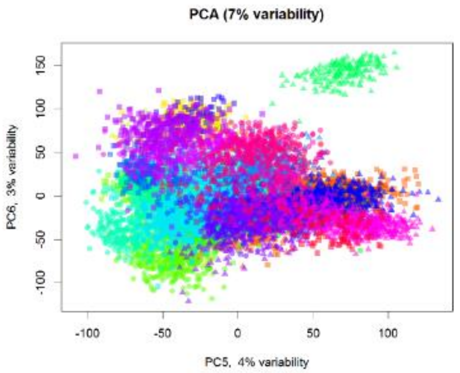
DataSets:	4348
Series:	94528
Platforms:	18136
Samples:	2375364

Data Repositories

Example: pan-cancer TCGA data analysis







- 11k samples
- 20k genes
- 300k exons
- 250k junctions

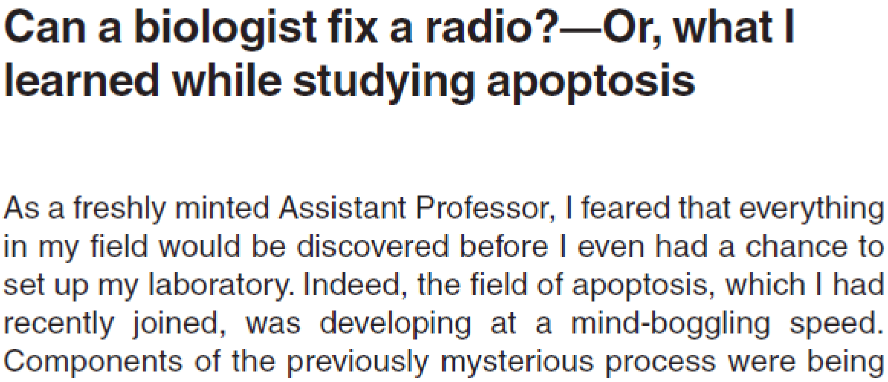
Models

All models are wrong, but some are useful

George E.P. Box

A Bit of History

Yuri Lasebnik, **Cancer Cell**, 2002



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

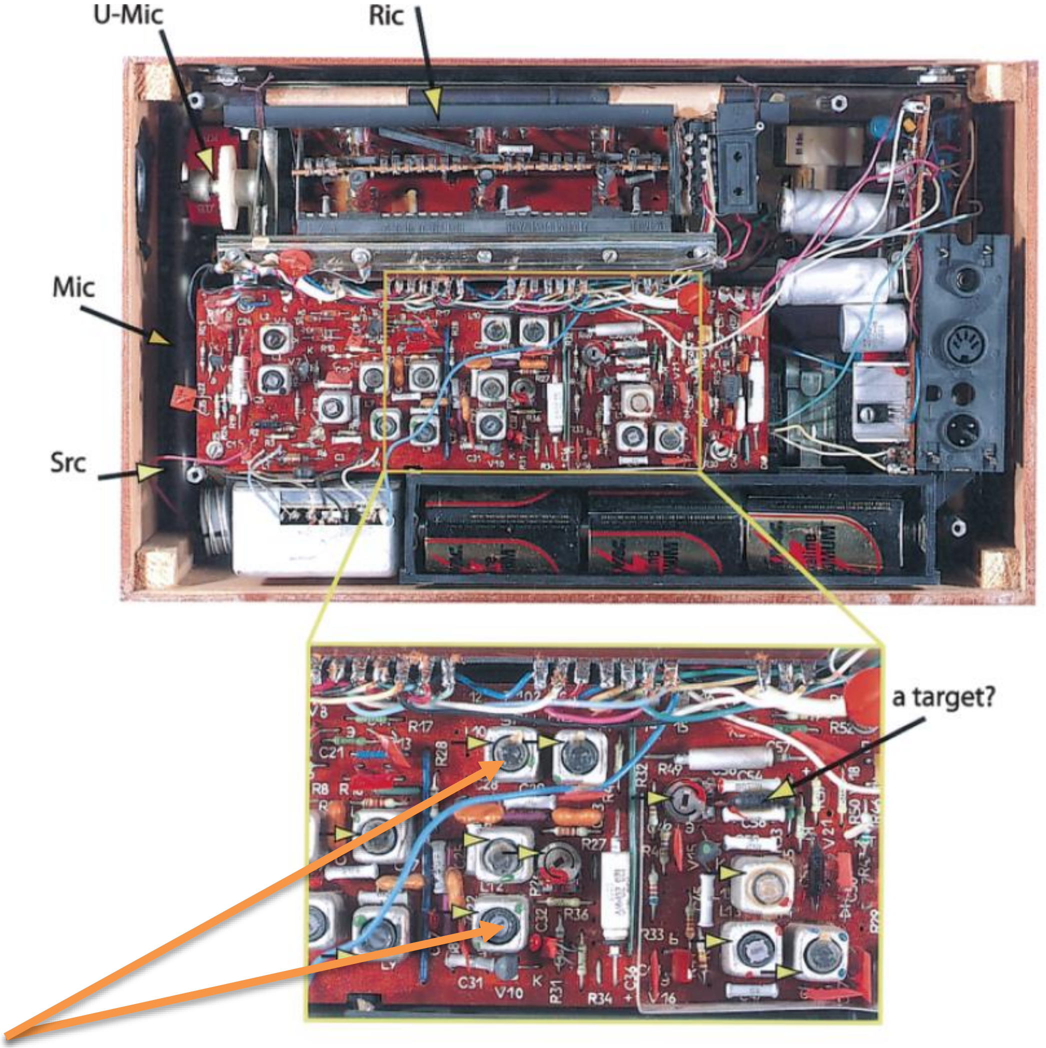
- If you want to see whether your method works, apply to a task with already known solution

- As an example, let’s see how the “standard” approach to modeling could help us to understand a complex system - radio

“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 !!!



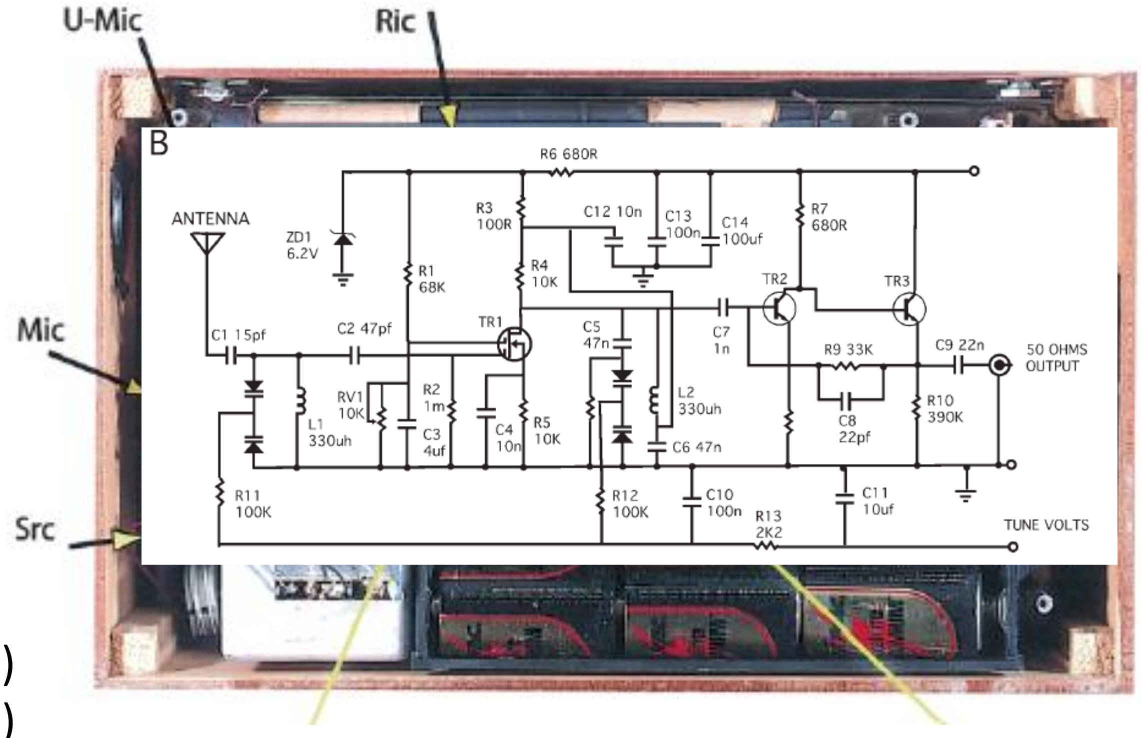
However it worked (if it work) only for this radio.
And what if the problem is in the tunable elements?

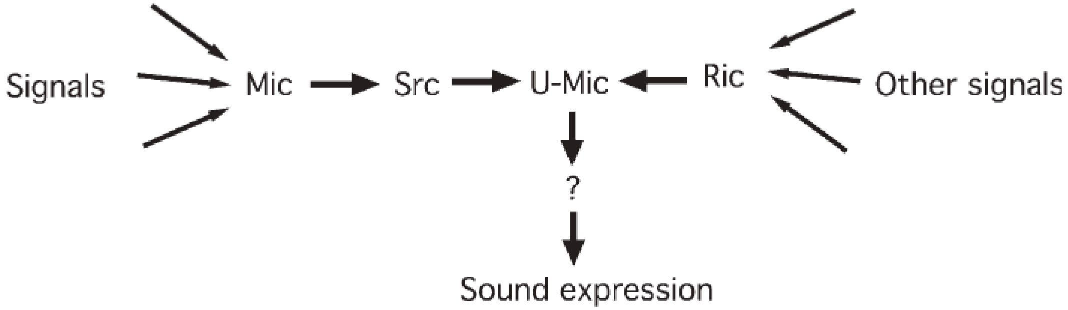
Two Models of the Same Radio

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).




Signals

Mic

Src

U-Mic

Ric

Other signals

?

Sound expression

Some Types of Models

Kinetic modeling: sets of ODE describing concentrations

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

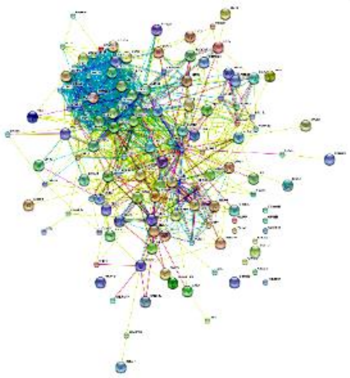
$$\frac{d[S]}{dt} = -k_1[E][S] + k_{-1}[ES]$$

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

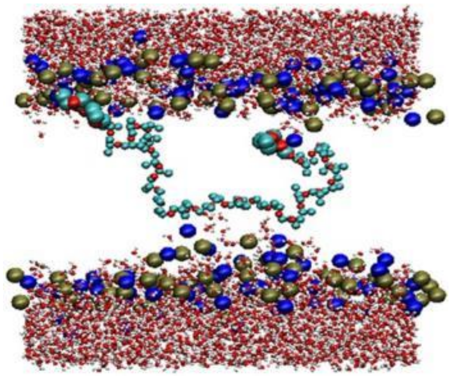
$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] + k_{-2}[E][P]$$

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

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 group by the gene expression

GWAS – genome-wide association studies

Methods

Methods Overview (biased)

Statistical methods:

Linear models

- normal
- Poisson
- negative binomial

Rank product (non-parametrical)

Enrichment analysis

Dimensionality reduction:

PCA

ICA

NMF

MDS

tSNE

Clustering:

Hierarchical clustering

K-means

NMF

Fuzzy methods

Survival:

Cox regression

Classification & Predictions:

Linear models

Random Forest

SVM

LASSO

Neural networks

Dependencies & Networks:

Correlation

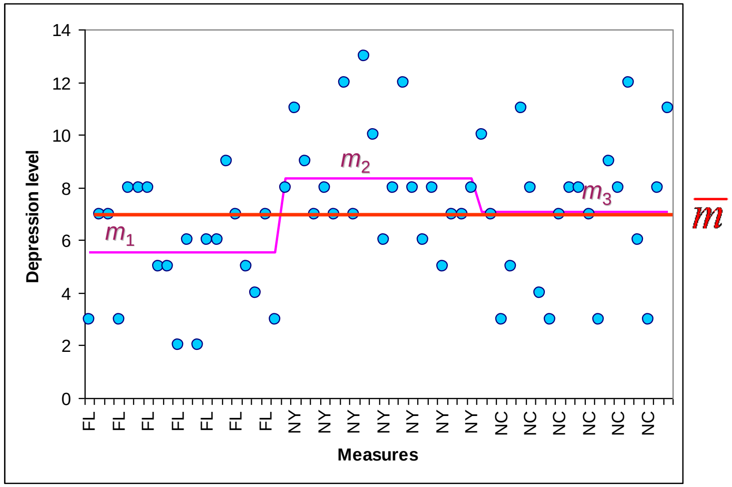
DCEA

Mutual information

Methods of topological analysis

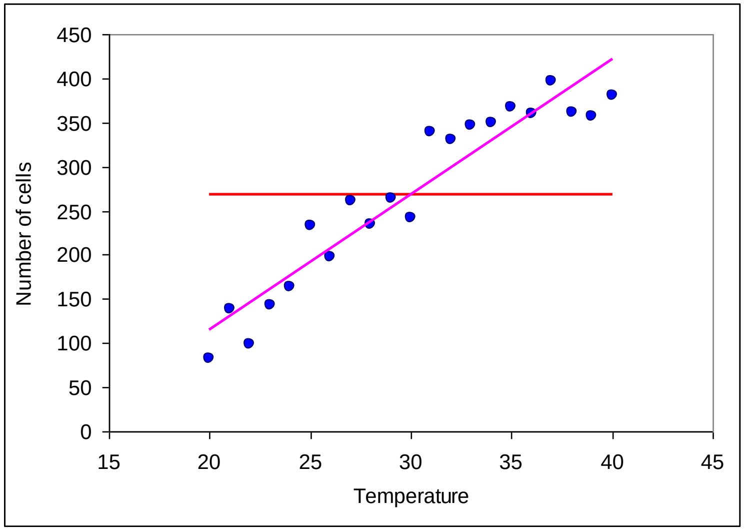
Linear Models

ANOVA



$$SST = SSTR + SSE$$

Linear Regression



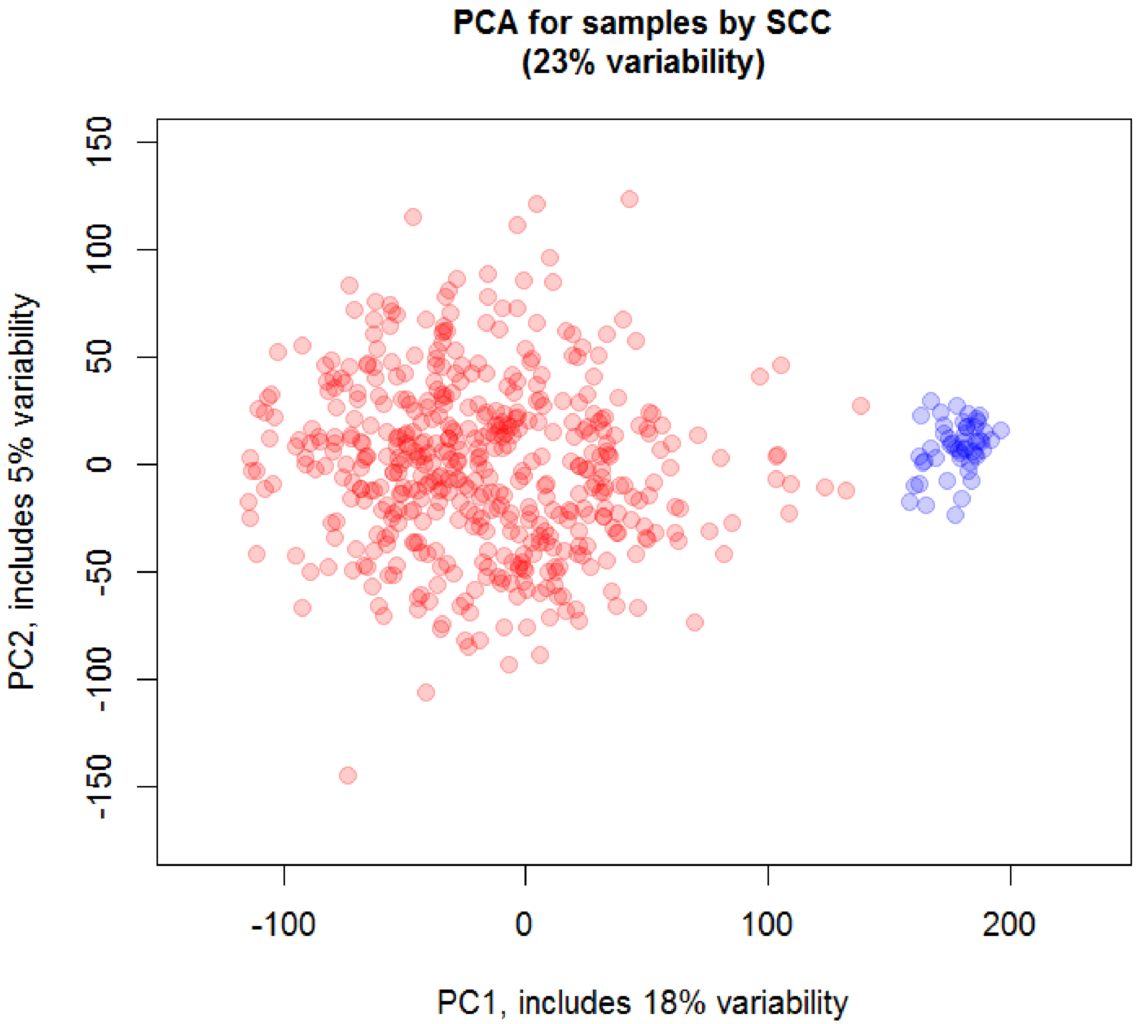
$$SST = SSR + SSE$$

Depression = μ + Location + ε

Number = b_1 * Temperature + b_0 + ε

Dimension Reduction

Principal Component Analysis (PCA)

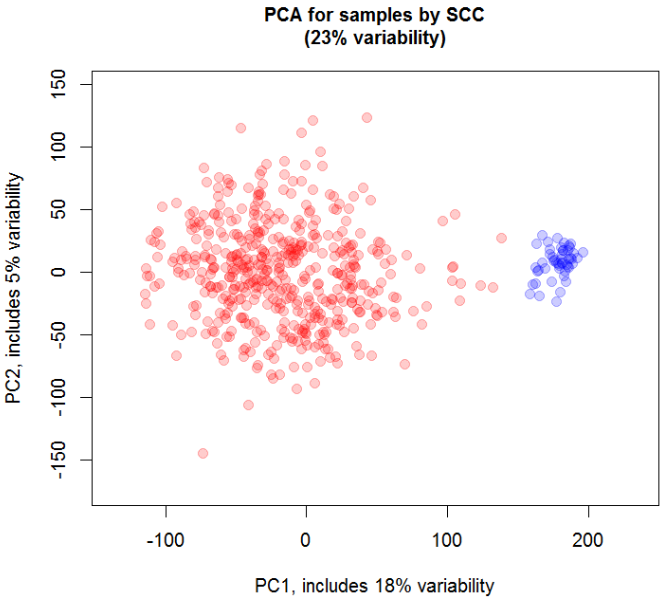


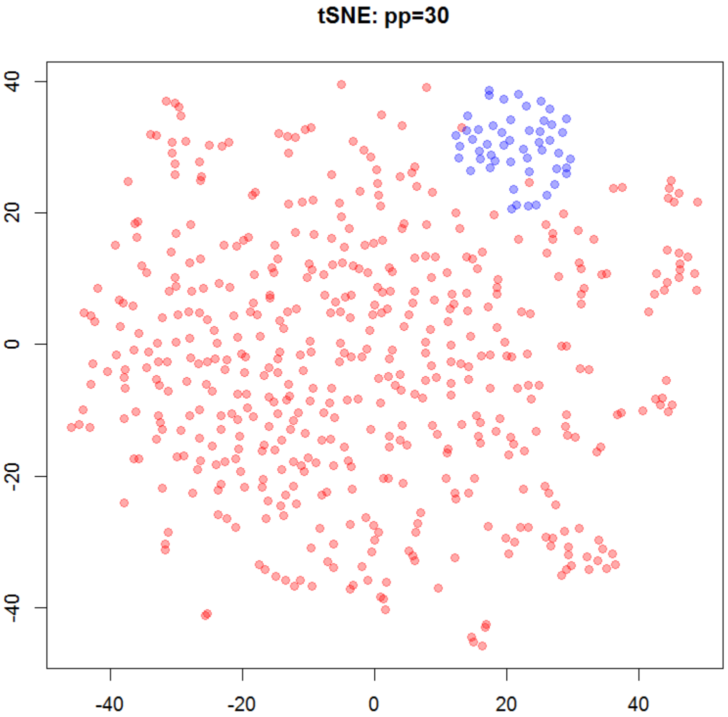
Dimension Reduction

t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE

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



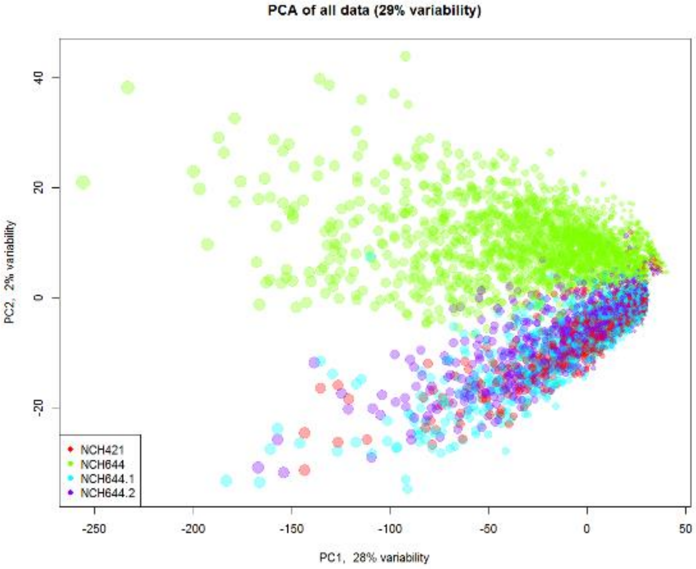


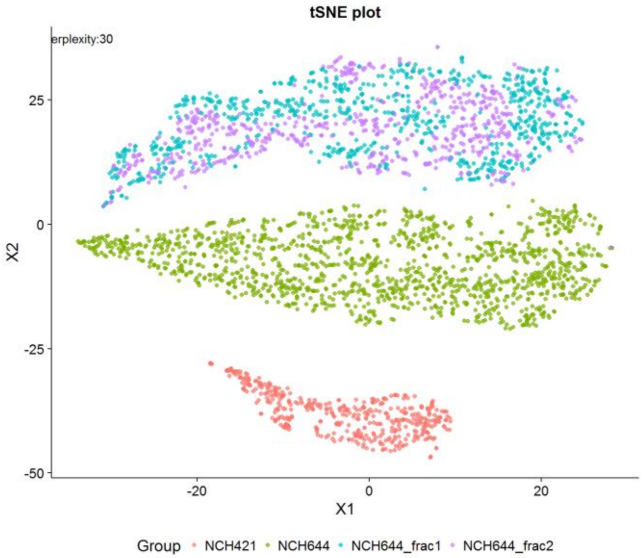
Dimension Reduction

t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE

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

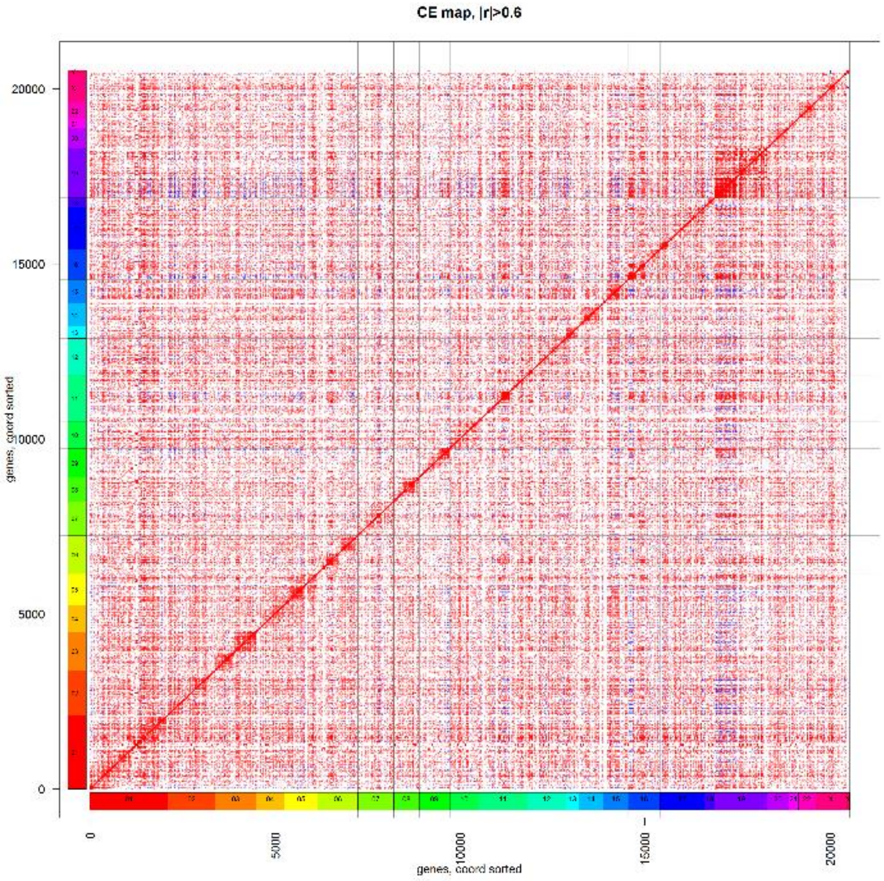




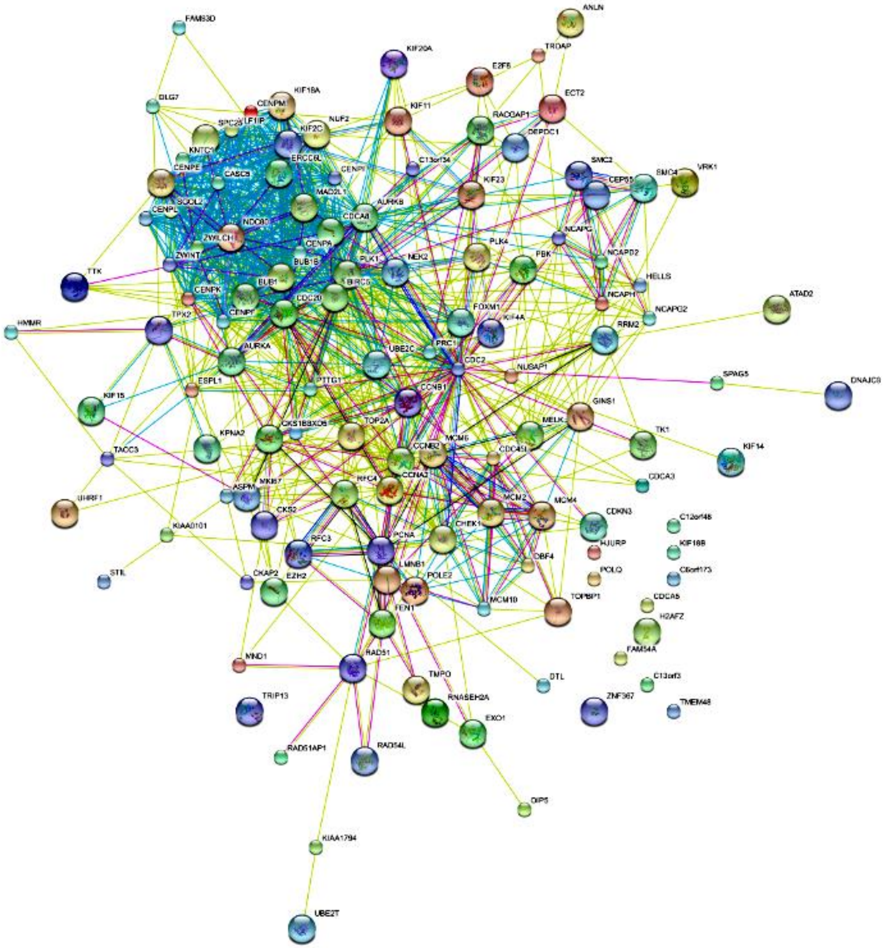
Correlation and Networks

Building Networks of Genes

Example: TCGA data, all genes, 9k tumors

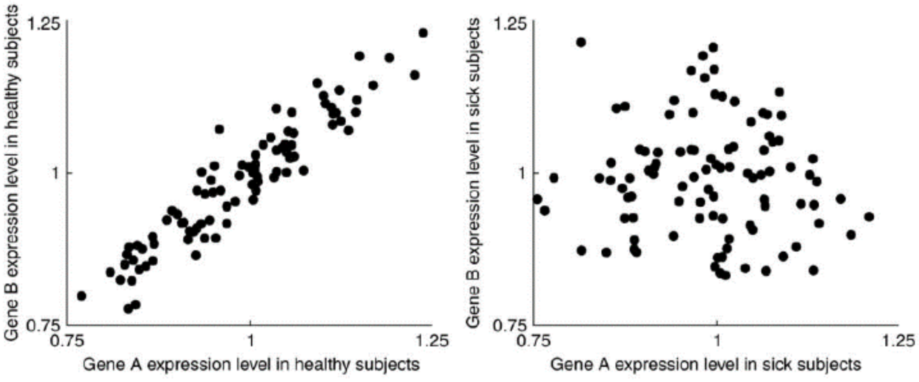


Example: network in String.DB

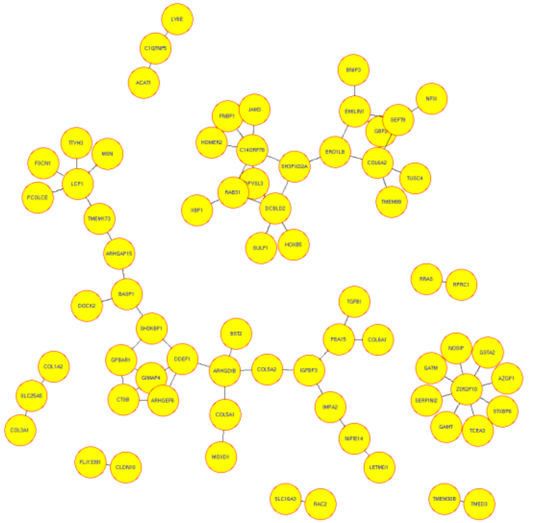


Correlation and Networks

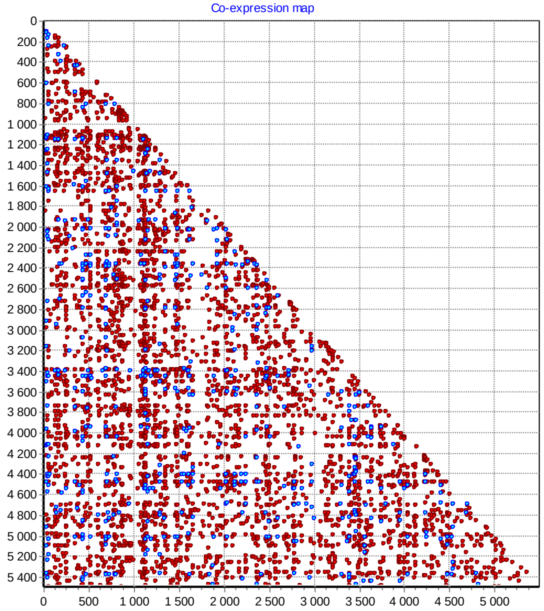
- Differential Co-expression Analysis



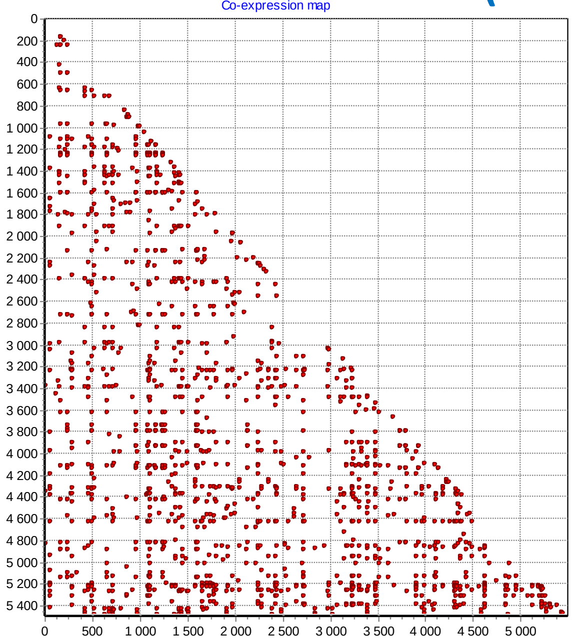
TRENDS in Genetics



44 normal pancreas (NP)



44 ductal adencarcinoma (PDAC)



Methods Overview (biased)

Statistical methods:

Linear models

- normal
- Poisson
- negative binomial

Rank product (non-parametrical)

Enrichment analysis

Dimensionality reduction:

PCA

ICA

NMF

MDS

tSNE

Clustering:

Hierarchical clustering

K-means

NMF

Fuzzy methods

Survival:

Cox regression

Classification & Predictions:

Linear models

Random Forest

SVM

LASSO

Neural networks

Dependencies & Networks:

Correlation

DCEA

Mutual information

Methods of topological analysis

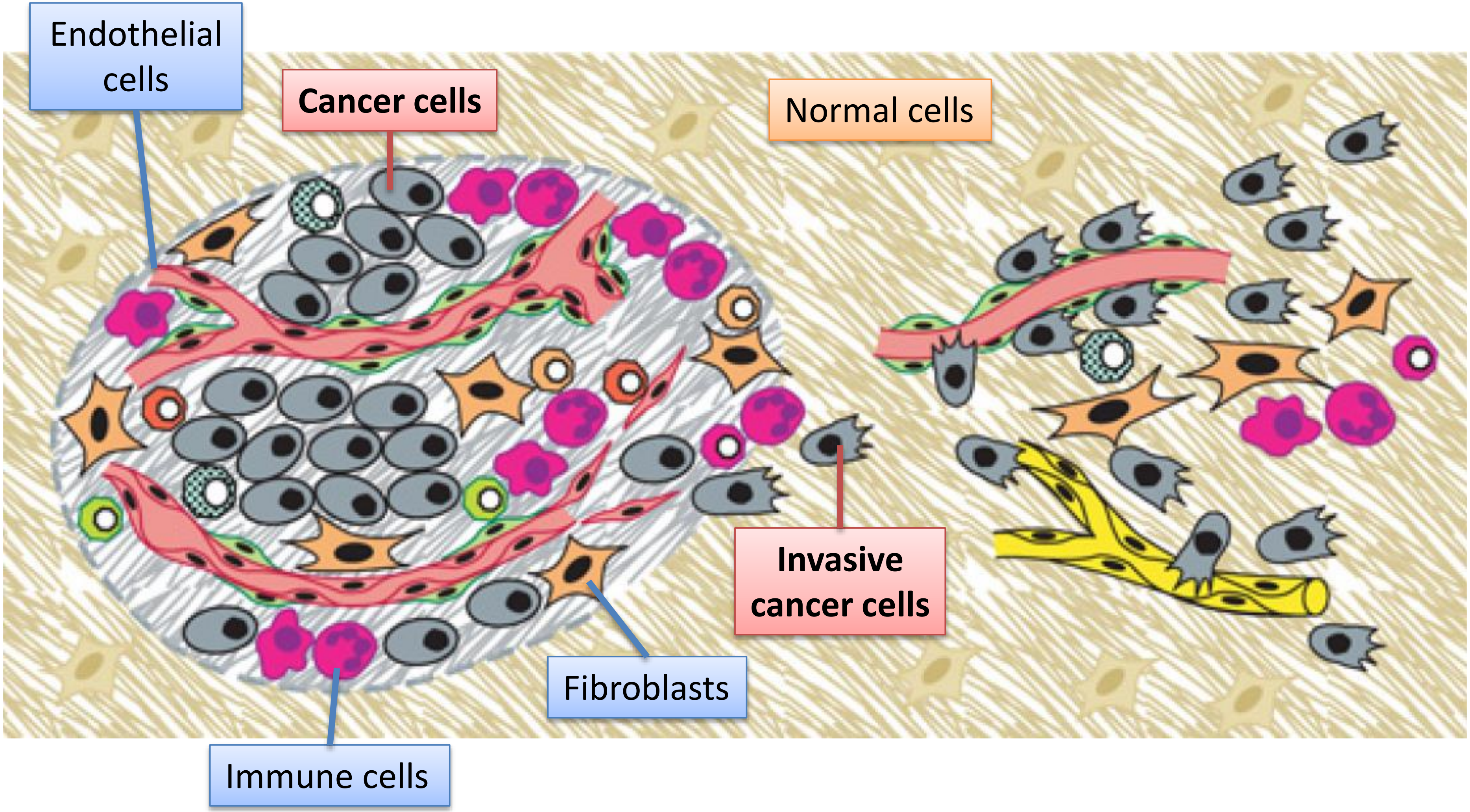
Questions?

Example:

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

Introduction

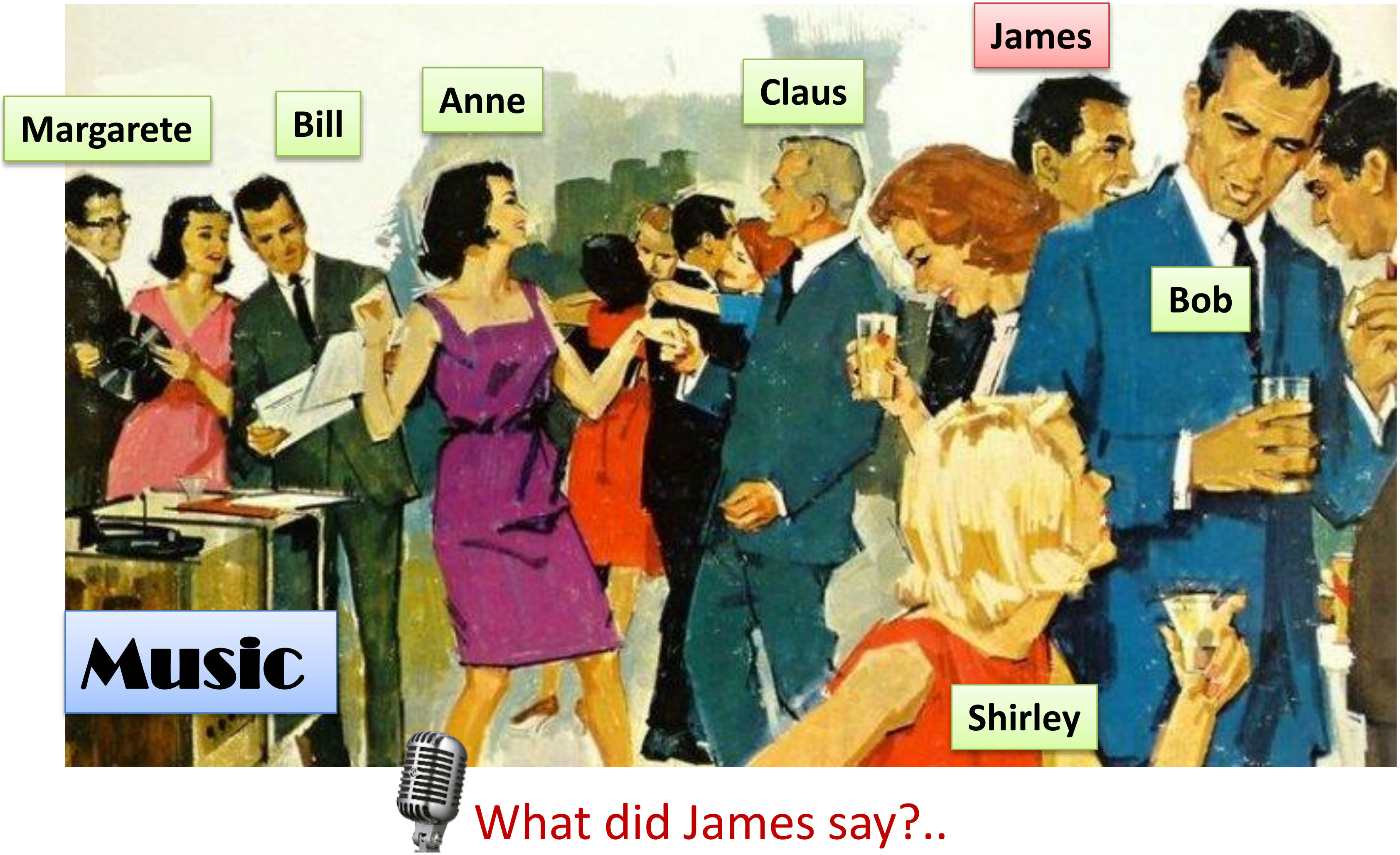
Imagine we are going to analyze RNA from a tumor biopsy (sample):



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

Introduction

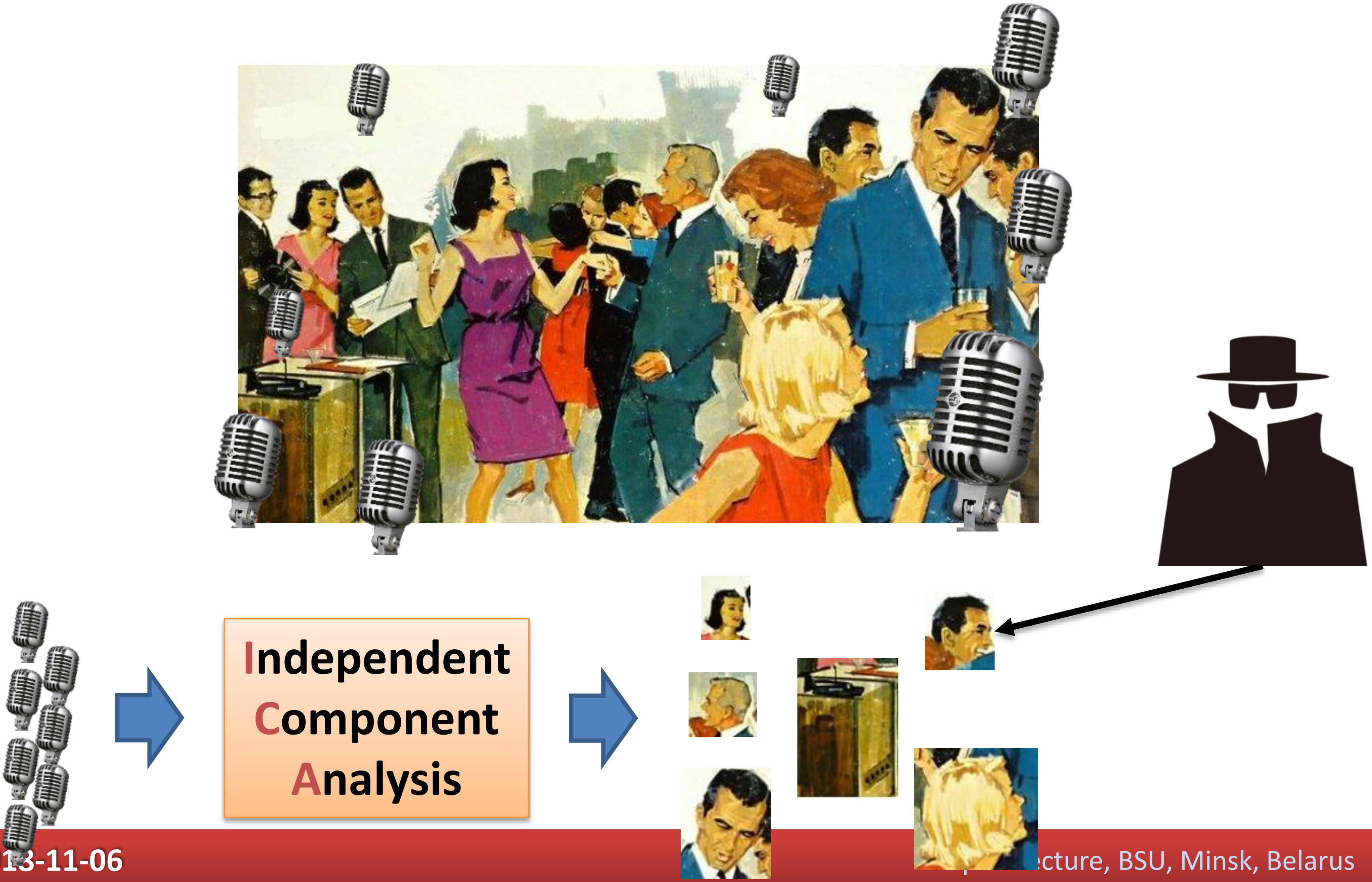
This is like recording a cocktail party:



What did James say?..

Independent Component Analysis

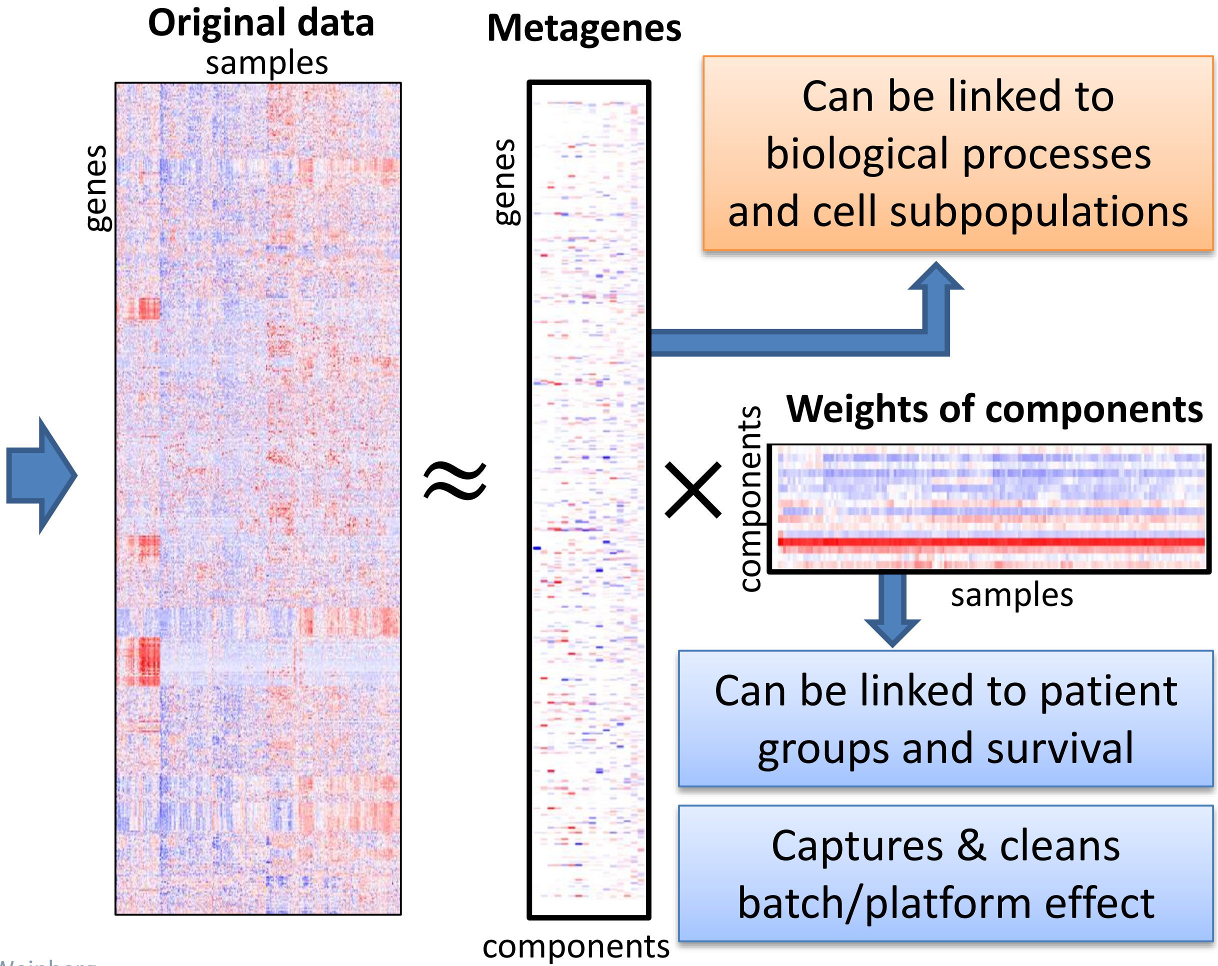
One of the methods to solve cocktail party problem…



Independent Component Analysis

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

Deconvolution of Cell Ensemble



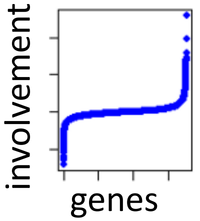
Can be linked to biological processes and cell subpopulations

Can be linked to patient groups and survival

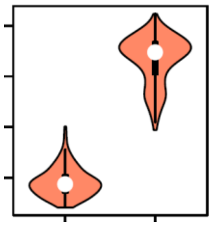
Captures & cleans batch/platform effect

$$X_{gs} \approx S_{gk} \times M_{ks}$$

One component



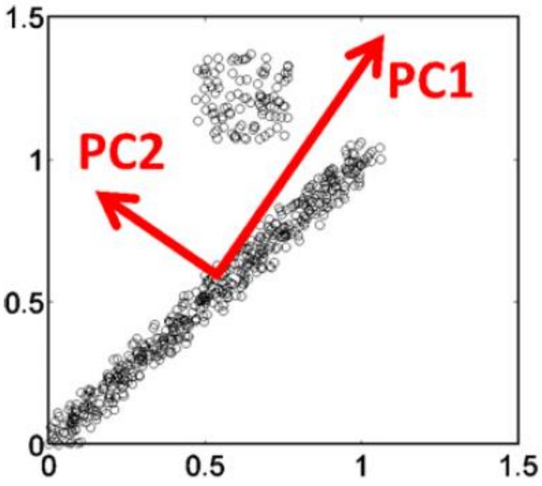
Components weights in patients



Independent Component Analysis

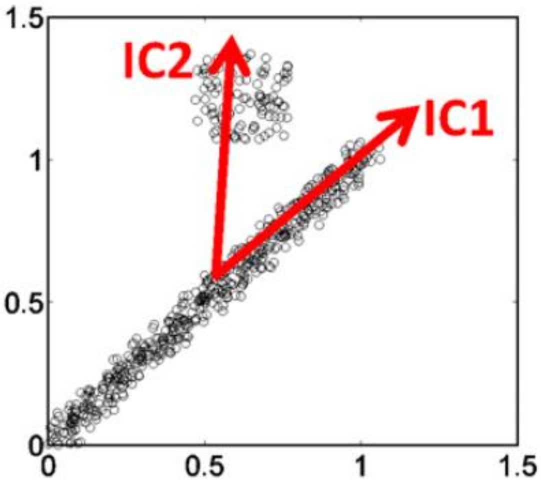
Geometrical view ☺

PCA



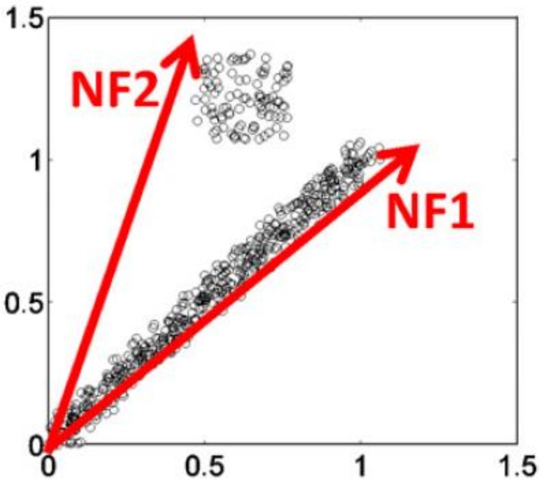
Orthogonal
Captures major variation
(well, on average…)

ICA



Linear combination of
independent sources.
Positive and negative.

NMF

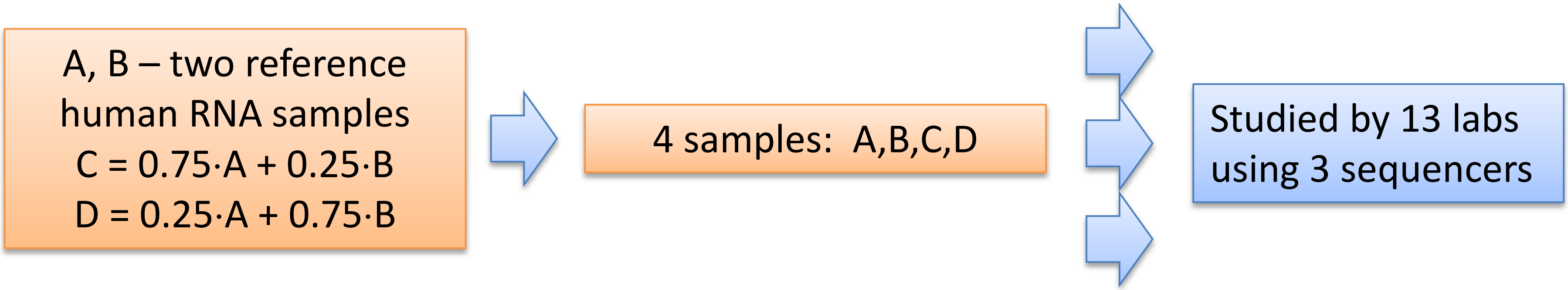


Each point can be
represented as a vector sum
of NF1, NF2. Strictly positive.

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

Independent Component Analysis

SEQC Data

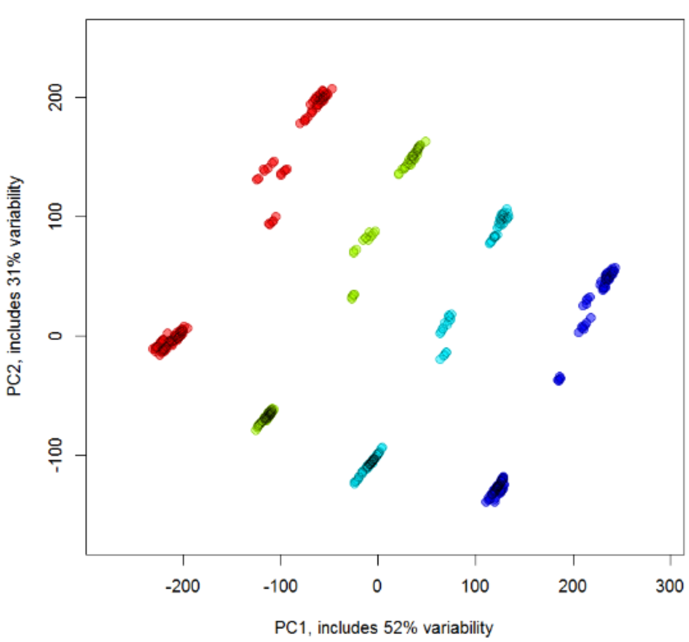


A, B – two reference human RNA samples
C = 0.75·A + 0.25·B
D = 0.25·A + 0.75·B

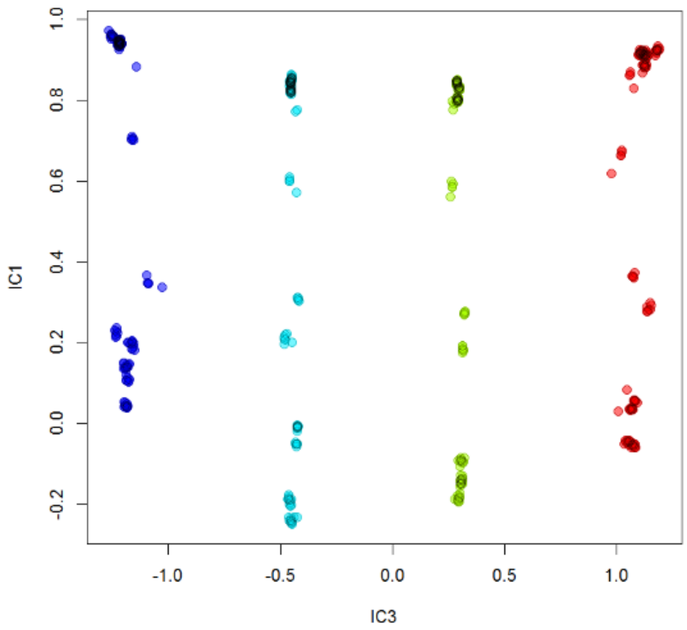
4 samples: A,B,C,D

Studied by 13 labs using 3 sequencers

Principle component analysis (PCA) (83% variability)



Independent Component Analysis (ICA)



The effect of sample mixing is captured by **two PCs** and **single IC_3** !

See `library(seqc)` in R if you want to play with the data

Independent Component Analysis

What ICA does and does not

$$\boxed{X_{gs} \approx S_{gk} \times M_{ks}}$$

g – genes
s – samples
k - components

Pro:

1. Finds **statistically-independent signals** (components) in the expression profiles
2. Identifies the **most important genes** in each component
3. Tells what is the weight of **each component in the samples**
4. Works on data *per se*, **without any additional knowledge**
5. Gives quite **robust answer**… just… reshuffled

Contra:

1. Needs **a lot of data**. The original data should not be too skewed.
2. **No ranking of the components** by importance (not like PCA)
3. Results are **not deterministic** and can to some extent depends on the run => multiple run / consensus approach is needed!
4. **Orientation of the signal is arbitrary** from one run to another
5. If you look for precise estimation of cell fraction – not a good idea (results will be qualitative not quantitative)

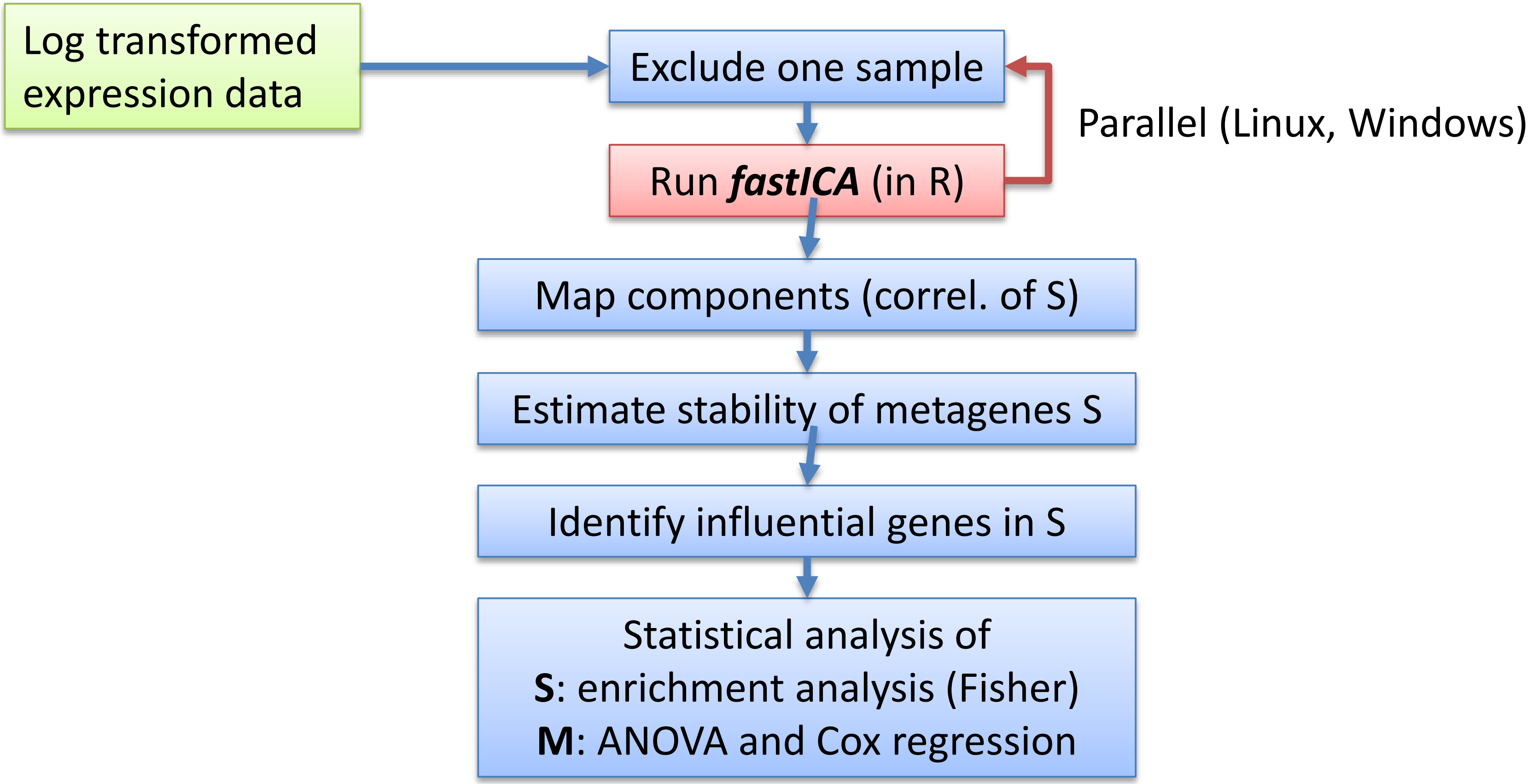
Methods

Consensus ICA

$$X_{gs} \approx < S_{gk} > \times < M_{ks} >$$

g – genes
s – samples
k - components

<S>, <M> – mean over multiple runs, excluding random samples



Log transformed expression data

Exclude one sample

Run ***fastICA*** (in R)

Parallel (Linux, Windows)

Map components (correl. of S)

Estimate stability of metagenes S

Identify influential genes in S

Statistical analysis of
S: enrichment analysis (Fisher)
M: ANOVA and Cox regression

Results

Positively and negatively contributing genes

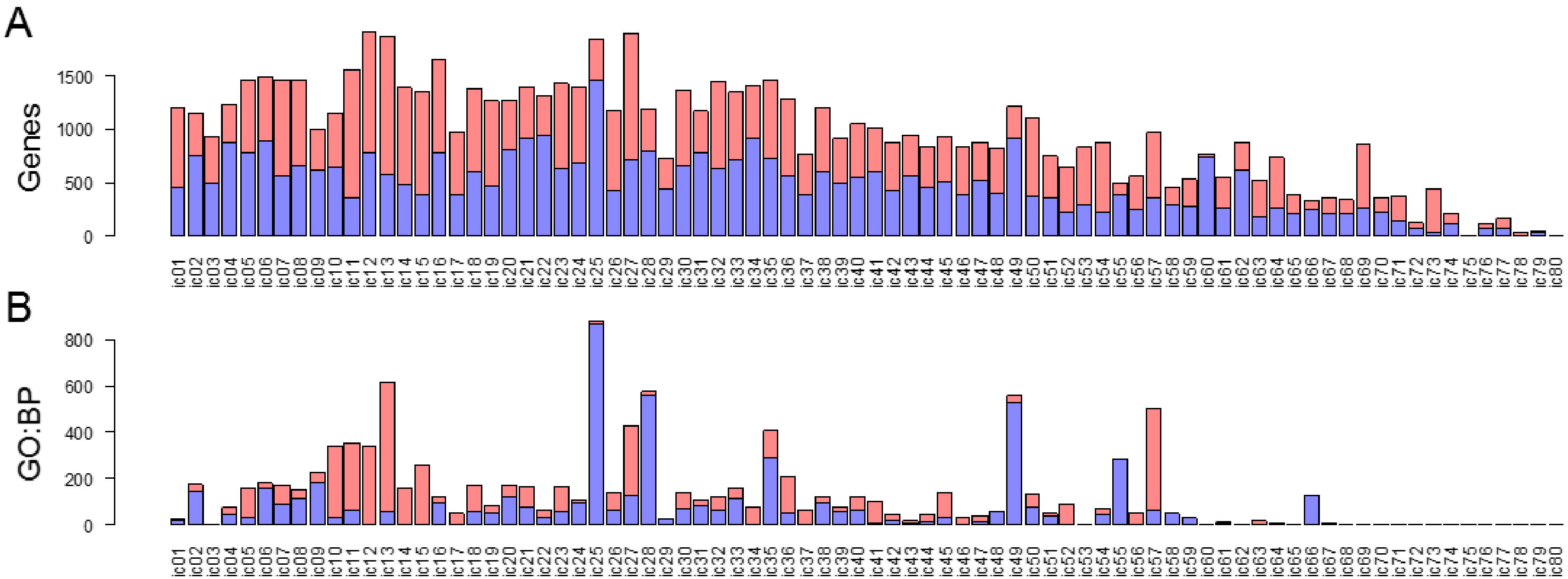
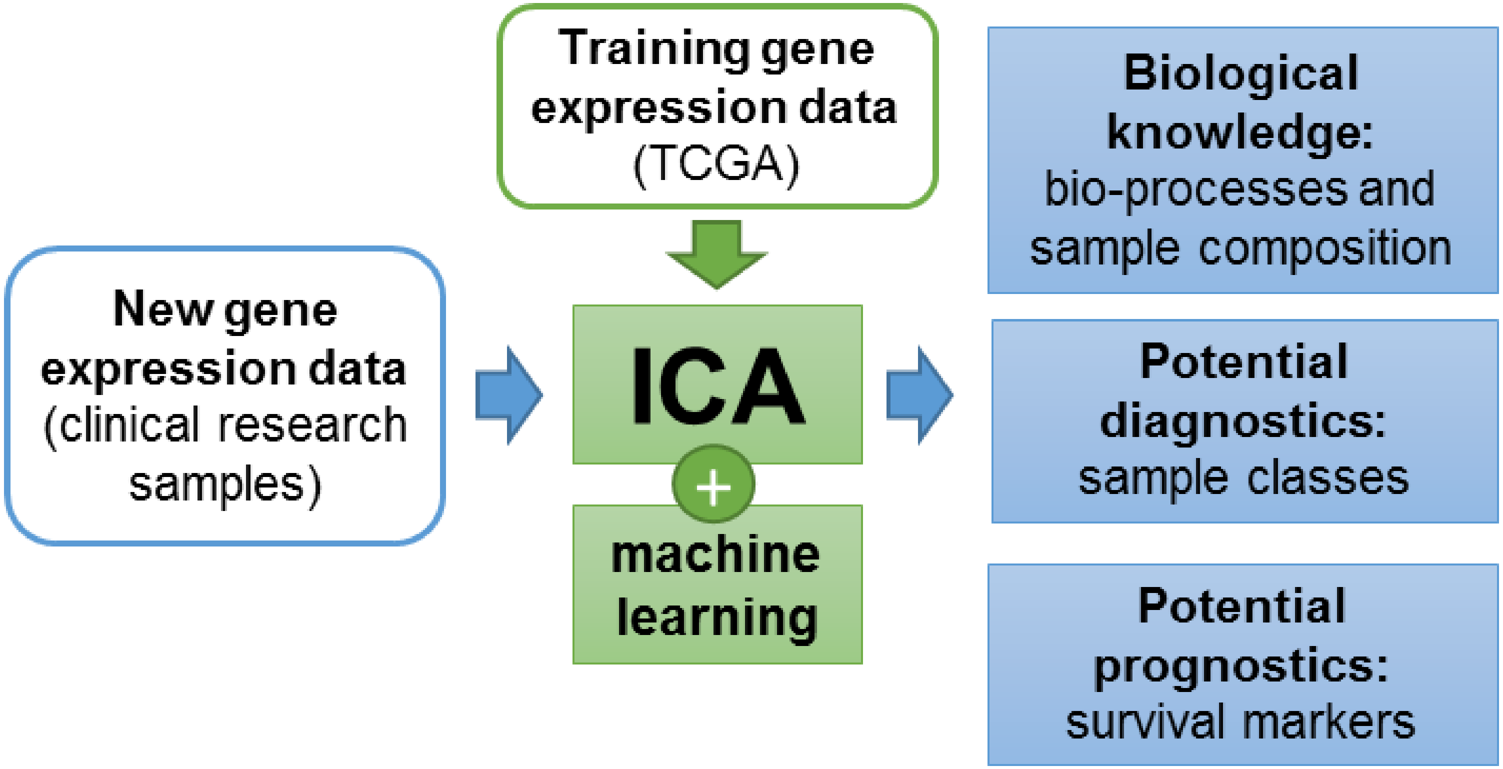


Figure S6. (A) Number of significant positively (red) and negatively (blue) involved genes in metagene of each of the components. (B) Number of enriched GO biological processes found for these genes. For the most cases, only one list of genes is biologically meaningful: either positive (e.g. ic10-ic15) or negative (e.g. ic25, ic28, ic49, ic55).

Methods

ICA to study new patients


Training gene expression data (TCGA)
New gene expression data (clinical research samples)
ICA
+
machine learning
Biological knowledge: bio-processes and sample composition
Potential diagnostics: sample classes
Potential prognostics: survival markers

We use our **parallel consensus ICA** that provides quite **robust estimation of the matrices** (based on fastICA package in R)

Results

Patient classification in SKCM

SKCM
(skin cutaneous melanoma)

472 samples

- SVM & RF work both fine when n_{comp} is small
- For large n_{comp} – RF gives much better predictions (SVM is overtrained)

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

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

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

Here accuracy was estimated using LOOCV

Results

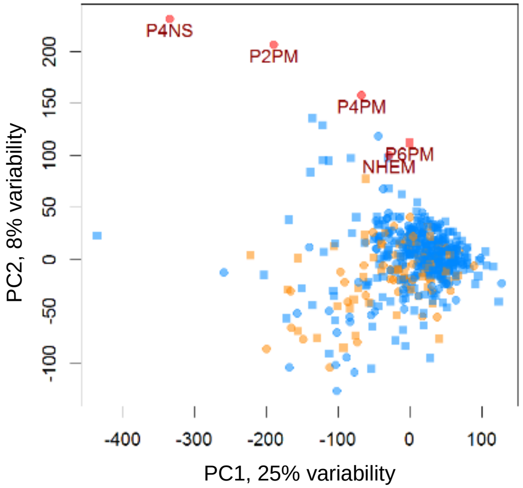
New samples: mRNA

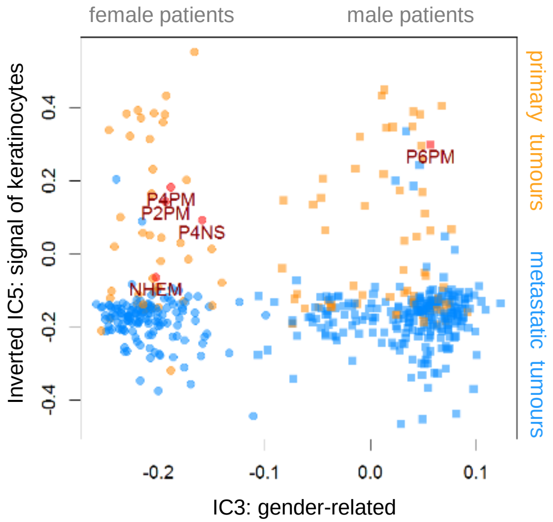
5 new samples: 3 primary tumors (PM), 1 normal skin (NS), 1 cell line (SKCM)

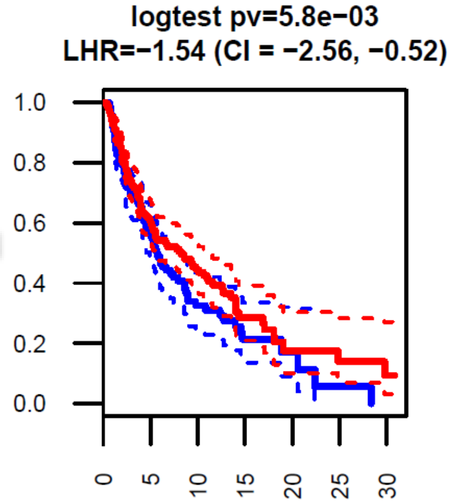
mRNA level: RNA-seq + RNA-seq

Gender: ● female, ■ male

Sample type: ●■ primary tumour
●■ metastatic
●■ new samples







When ICA is run over new samples and training samples together, it corrects for platform bias.

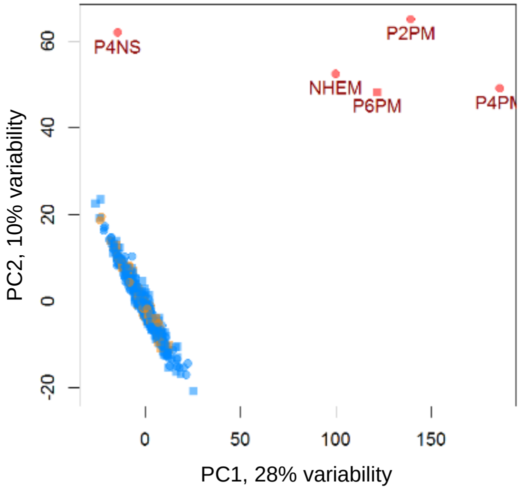
Results

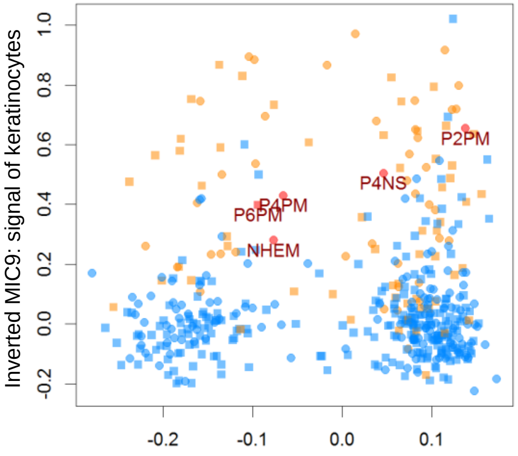
New samples: mRNA

Gender: ● female, ■ male

Sample type: ●■ primary tumour
●■ metastatic
●■ new samples

miRNA level: RNA-seq + qPCR

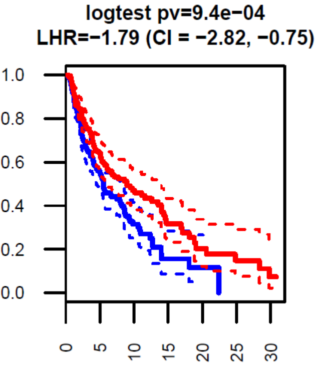




MIC1: unknown segregation of samples

miR−146a−3p
miR−338−5p
miR−551b−3p
miR−598−3p
miR−206
miR−34a−5p
miR−338−3p
miR−146a−5p
miR−1269a
miR−573

miR−205−5p
miR−199b−5p
miR−876−5p
miR−1266−5p
miR−301b−3p
miR−3690
miR−365a−3p
miR−125b−1−3p



Conclusion 1:

Consensus ICA can correct technical biases between platforms

MelanomICA

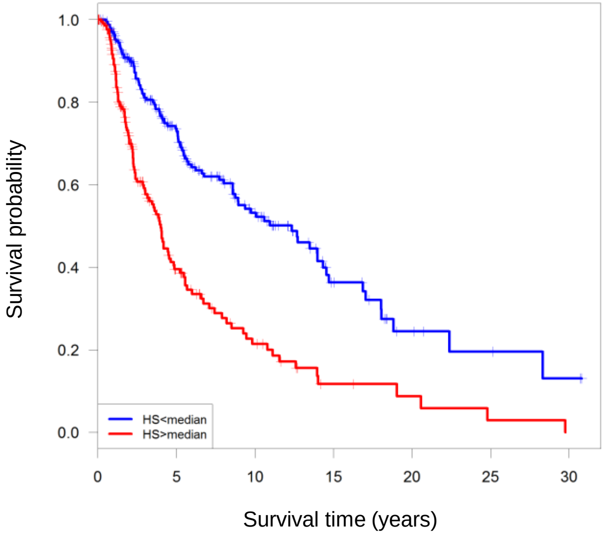
Hazard score

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

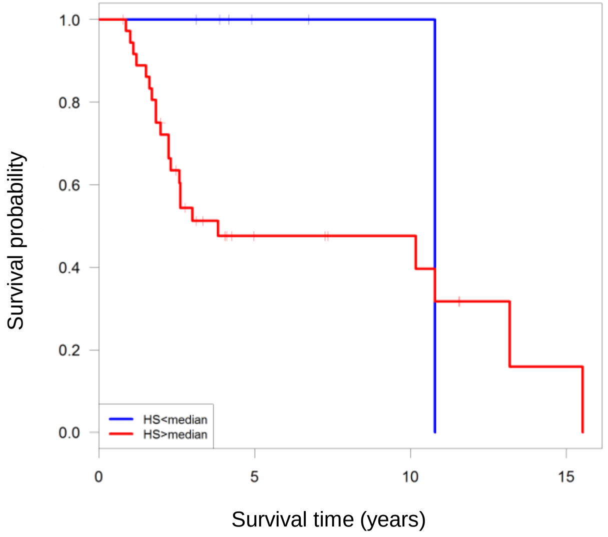
$$H_i = \begin{cases} LHR & \text{for significant components} \\ 0 & \text{for non-significant components} \end{cases}$$

44 metastatic patients

Training / Reference set
Log-rank test p-value= 5.6e-16
LHR= 0.49 (CI = 0.37, 0.61)



Validation set
Log-rank test p-value= 1.3e-03
LHR= 0.87 (CI = 0.28, 1.45)



Conclusion 2:

Consensus ICA can be used to predict cancer subtype and patient survival

MelanomICA

Immune

Stromal and angiogenic

Skin related

Melanocytes

Other

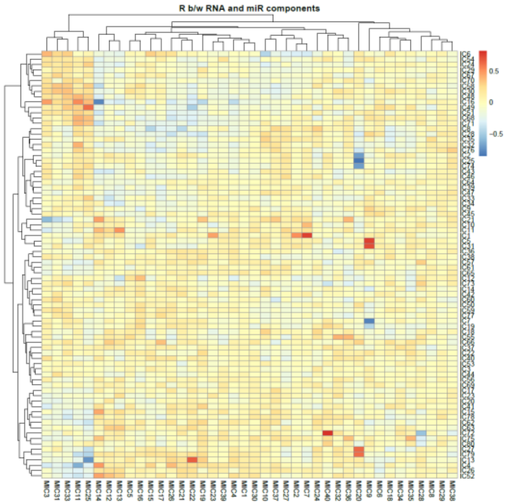
Cluster	Component	Risk (p-value)	Meaning	P2PM	P4PM	P6PM	P4NS	NHEM
Immune	RIC2	decreased (1.8e-4)	B cells	0.11	0.07	0.02	0.19	0.01
	RIC25	decreased (2.8e-7)	T cells	0.26	0.06	0.24	0.18	0.00
	RIC27	no effect	B cells	0.80	0.37	0.31	0.80	0.00
	RIC28	no effect	response to wounding	0.34	0.57	0.78	0.43	0.84
	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	potentially related to 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
Melanocyte s	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

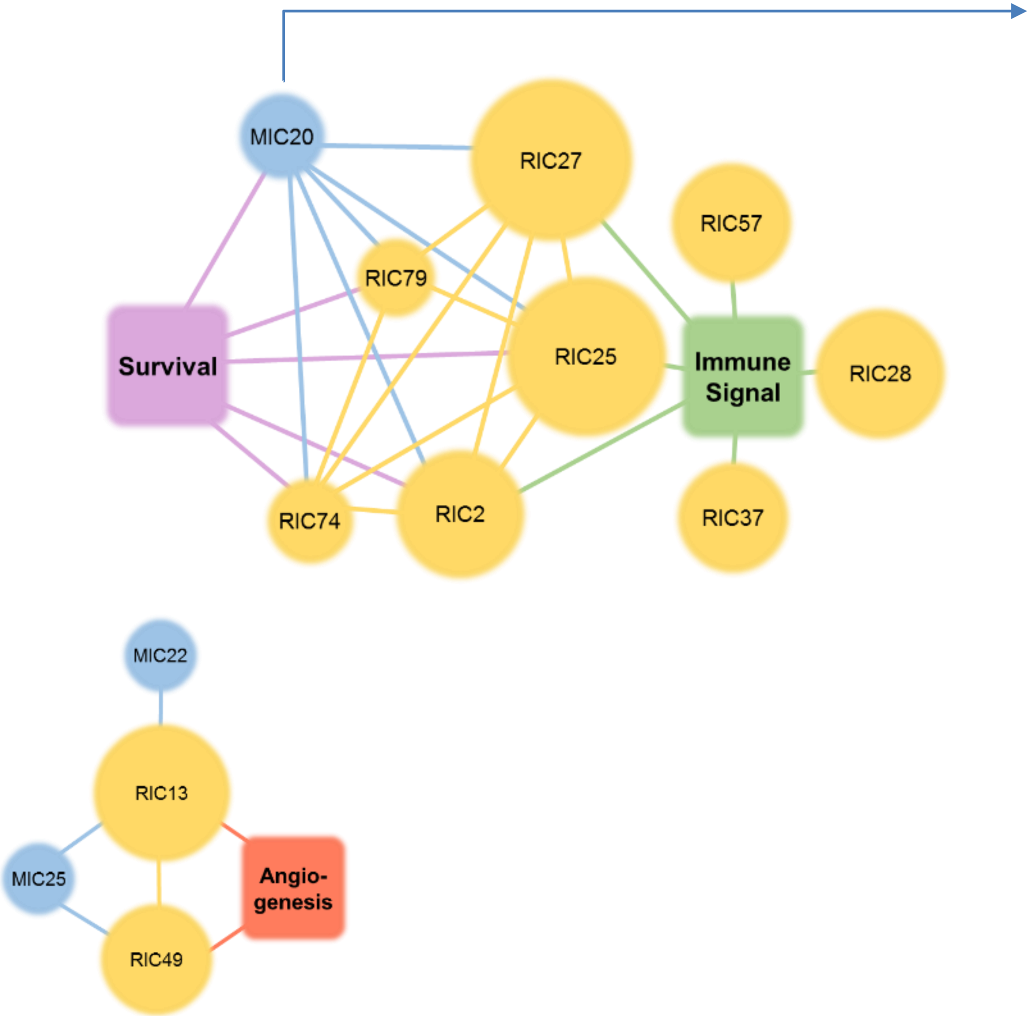
Conclusion 3:

Consensus ICA can be used to get biological knowledge about the new samples

MelanomICA

Correlation of weights:
mRNA-miRNA





hsa-miR-1247-5p
hsa-miR-142-3p
hsa-miR-142-5p
hsa-miR-146b-3p
hsa-miR-150-5p
hsa-miR-155-3p
hsa-miR-155-5p
hsa-miR-1976
hsa-miR-223-3p
hsa-miR-223-5p
hsa-miR-29c-3p
hsa-miR-342-3p
hsa-miR-342-5p
hsa-miR-3614-3p
hsa-miR-511-5p
hsa-miR-642a-5p
hsa-miR-146b-5p
hsa-miR-766-3p
hsa-miR-625-3p
hsa-miR-653-5p

Conclusion 4:

Consensus ICA can be used to integrate the data and assign functions to miRNAs

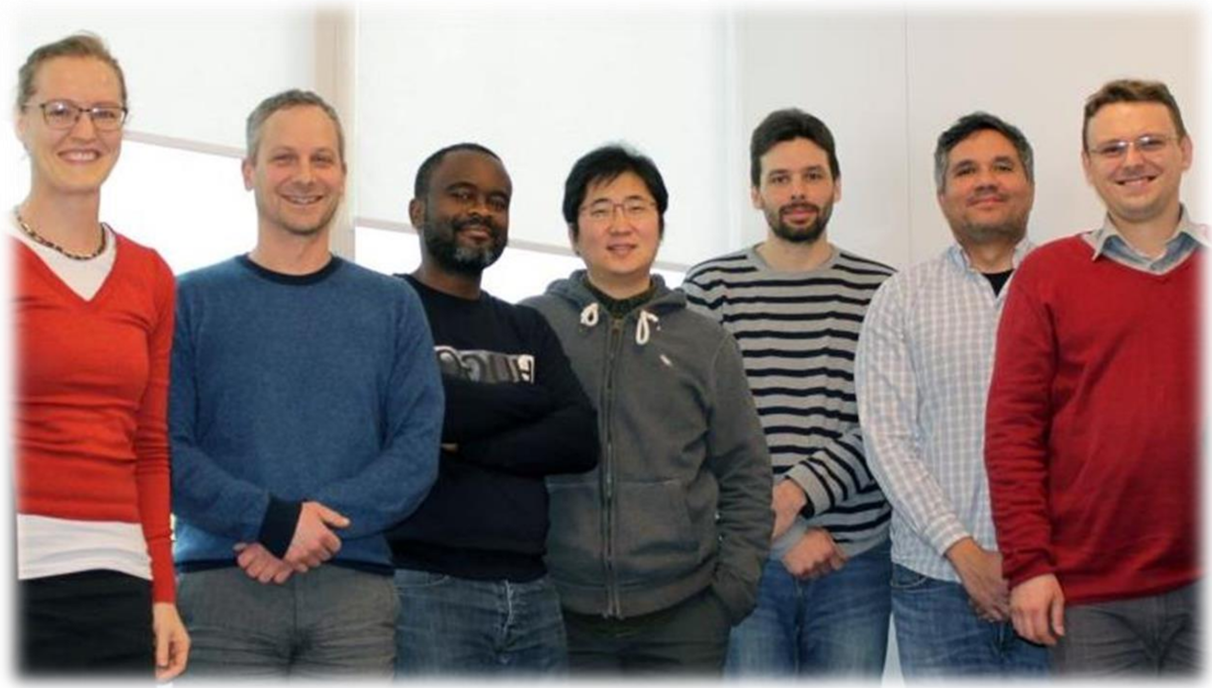
Conclusions

- We tested our implementation of consensus ICA
  (before publication, the script is available upon request)
- ICA decomposes large bulk data set into meaningful signals
- New samples are properly mapped in IC-space
- The method allows classifying and scoring new patients (clinical research studies)
- The method allows linking miRNA to mRNA and thus predicting miRNA functions

Acknowledgements

Dr. Gunnar DITTMAR

BIOMOD team of Proteome and Genome Research Unit, LIH



Dr. Francisco AZUAJE

This lecture was supported by Luxembourg National Research Fund (C17/BM/11664971/DEMICS)