LUXEMBOURG
INSTITUTE
OF **HEALTH**
RESEARCH DEDICATED TO LIFE

Performance Assessment of RNA Sequencing and Expression Arrays for Transcriptome Analysis in Cancer Research

Petr V. Nazarov

petr.nazarov@lih.lu

Luxembourg Institute of Health

Fonds National de la
Recherche Luxembourg

Outline

Part I. Comparison of RNA-seq and microarray performance

- Similar and specific features of the platforms
- Protein coding and long non coding genes
- Gene expression analysis and analysis of alternative splicing

Part II. Independent Component Analysis (ICA) in transcriptomics

- The brief introduction to the method
- Deconvolution of biological signals and cell subtypes
- Potential for patient diagnostics in future

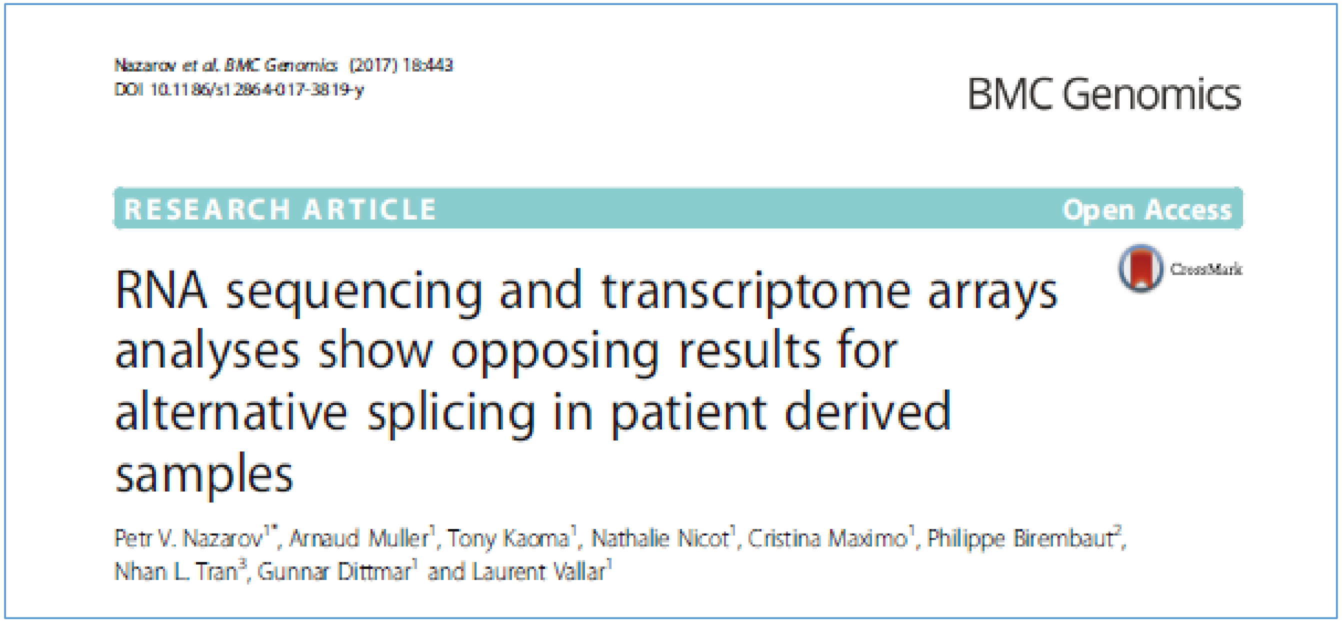
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Fonds National de la Recherche Luxembourg

IBBL
INTEGRATED BIOBANK OF LUXEMBOURG
FOR NEXT GENERATION HEALTHCARE

Part I. Comparison of RNA-seq and microarray performance

Based on Nazarov et al ***BMC Genomics*, 2017**;18(1):443.



Nazarov *et al. BMC Genomics* (2017) 18:443
DOI 10.1186/s12864-017-3819-y

BMC Genomics

RESEARCH ARTICLE Open Access

CrossMark

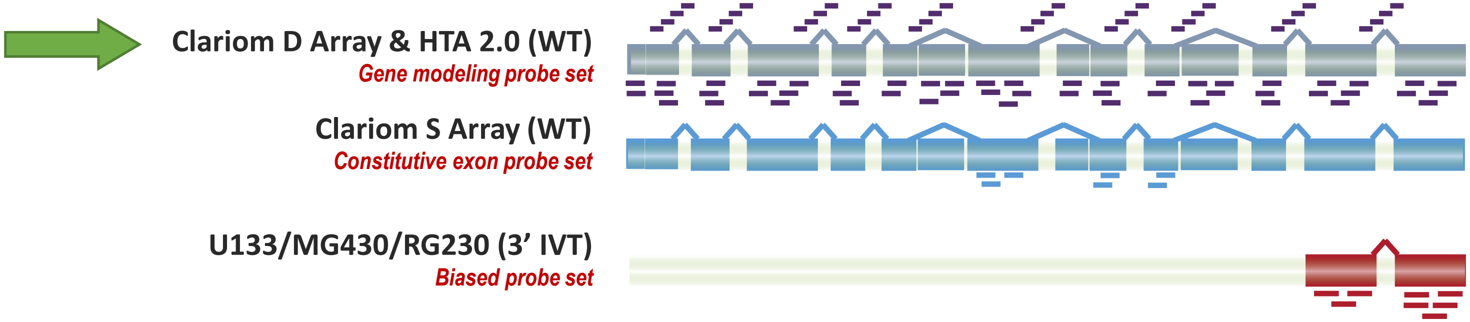
RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples

Petr V. Nazarov[1*], Arnaud Muller[1], Tony Kaoma[1], Nathalie Nicot[1], Cristina Maximo[1], Philippe Birembaut[2], Nhan L. Tran[3], Gunnar Dittmar[1] and Laurent Vallar[1]

Supported by Fond National de la Recherché Luxembourg (FNR) with the **grant C08/BM/05** and by the Luxembourg Ministry of Higher Education and Research. Integrated Biobank of Luxembourg (IBBL) sponsored RNA-seq experiments and shared their computational infrastructure for RNA-seq analysis.

Questions

Majority of comparisons in literature claim that RNA-seq outperforms microarrays.
However, comparing RNA-seq with old 3’ microarrays… not too fair.
Currently more advanced arrays are available: HTA and its successor Clariom.



Clariom D Array & HTA 2.0 (WT)
Gene modeling probe set

Clariom S Array (WT)
Constitutive exon probe set

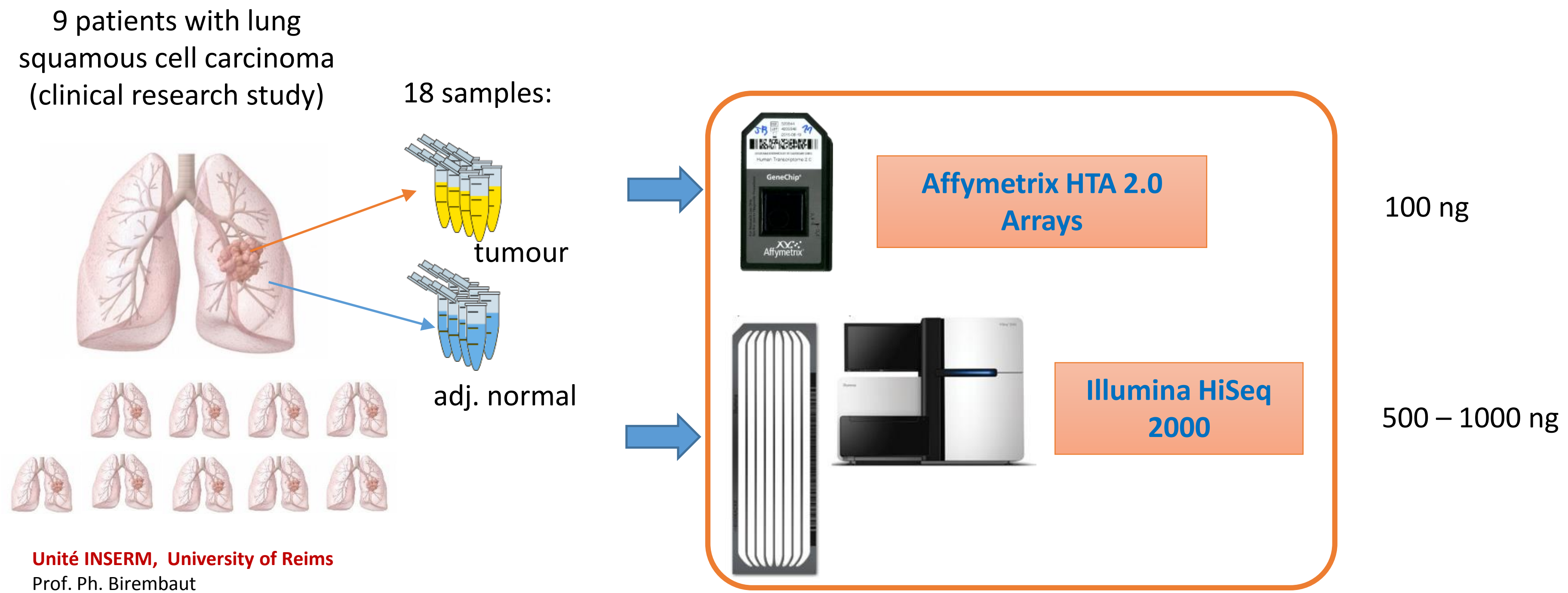
U133/MG430/RG230 (3’ IVT)
Biased probe set

- How similar are the results obtained by last version arrays and RNA-seq ?
  - protein coding / other biotypes, genes / exons
- What are the differences between platforms?
- Which platform should one use

Image is provided by ThermoFisher Scientific

Design

Research includes: 1 cancer, 9 patients, 18 samples, 2 platforms



9 patients with lung squamous cell carcinoma (clinical research study)

18 samples:

tumour

adj. normal

Affymetrix HTA 2.0 Arrays

100 ng

Illumina HiSeq 2000

500 – 1000 ng

Unité INSERM, University of Reims
Prof. Ph. Birembaut

- Total RNA extracted using miRNeasy Mini Kit
- **Arrays:** GeneChip® WT Plus Reagent Kit
- **Sequencing:** TruSeq total RNA Sample Preparation Kit v.1.0, polyA selection

Images:
http://www.qmedicine.co.in/top%20health%20topics/L/Lung%20Cancer.html

http://www.biopticsworld.com/articles/print/volume-5/issue-06/features/dna-sequencing-technologies-the-next-generation-and-beyond.html

Sequencing Results

Sequencing depth

RNA-seq result: 120-280 M paired reads with 77 bp/read.

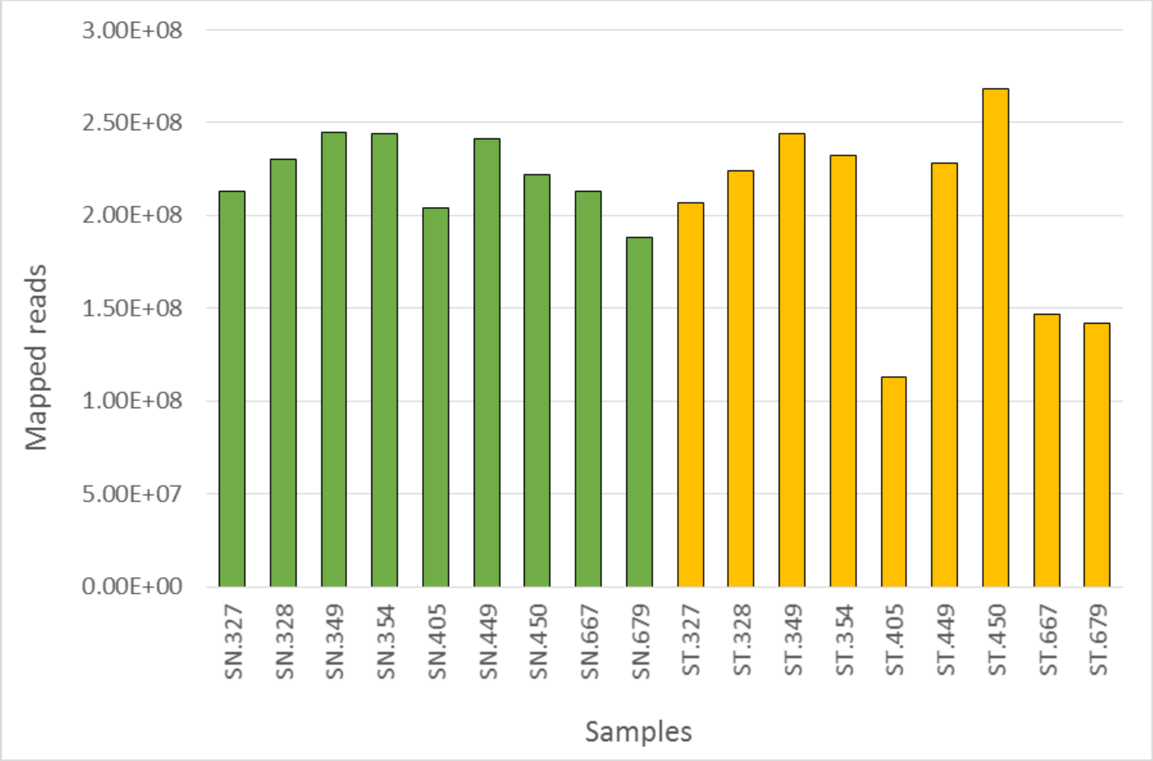
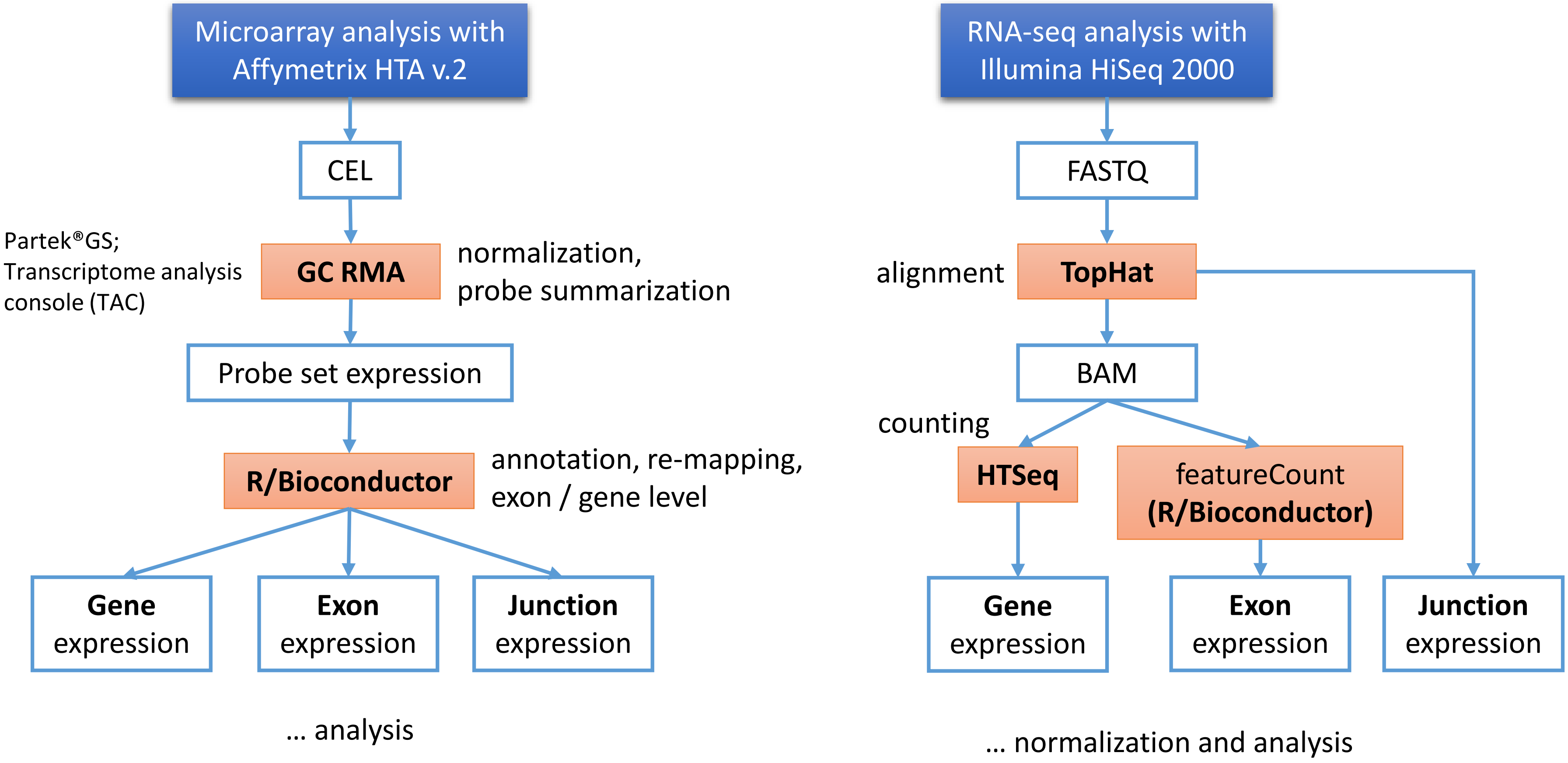


Figure. Number of mapped reads after RNA-seq analysis of the samples, including *TopHat* alignment. In general, normal tissues (green) show more reproducible mapping results than tumours (yellow).

Analysis Overview

Data acquisition



Microarray analysis with Affymetrix HTA v.2

CEL

Partek®GS; Transcriptome analysis console (TAC)

GC RMA

normalization, probe summarization

Probe set expression

R/Bioconductor

annotation, re-mapping, exon / gene level

Gene expression

Exon expression

Junction expression

… analysis

RNA-seq analysis with Illumina HiSeq 2000

FASTQ

alignment

TopHat

BAM

counting

HTSeq

featureCount **(R/Bioconductor)**

Gene expression

Exon expression

Junction expression

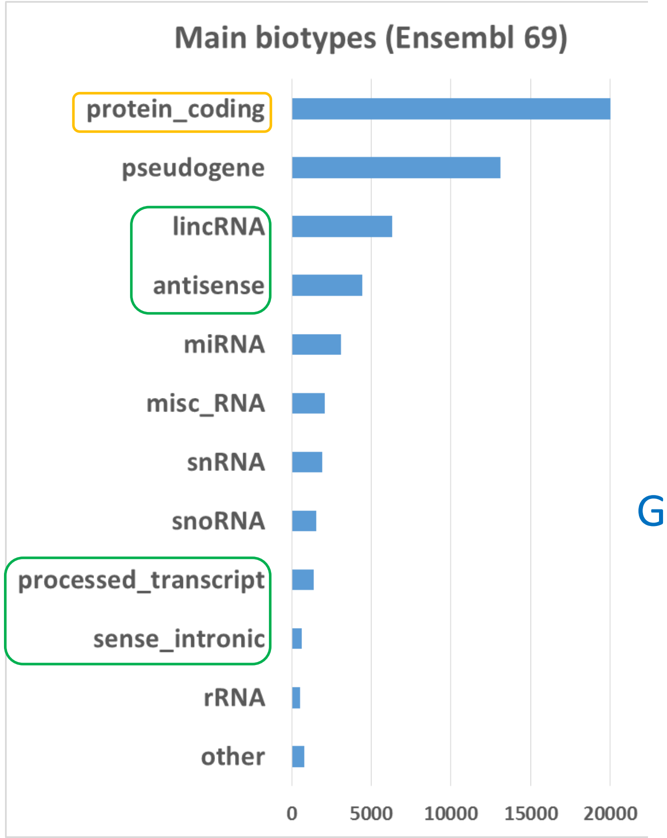
… normalization and analysis

Important:

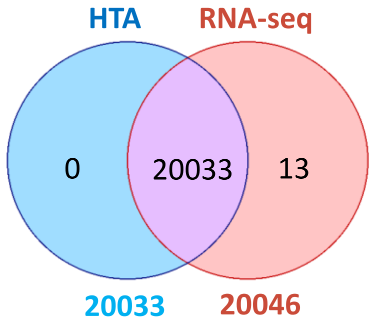
in order to compare the platforms, we re-mapped Affymetrix probesets onto the Ensembl 69 genome using GenomicRanges package of R.

Feature Lists

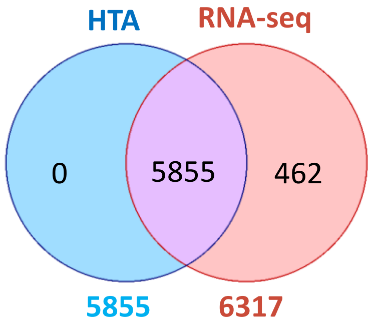
Overlap of features is high


Main biotypes (Ensembl 69)

Protein coding genes

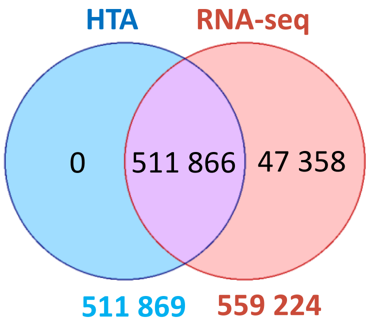


lncRNA genes

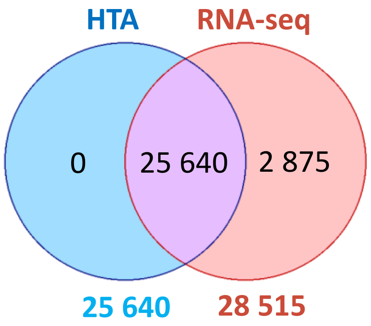


Good overlap of the genes and exons

Protein coding exons



lncRNA

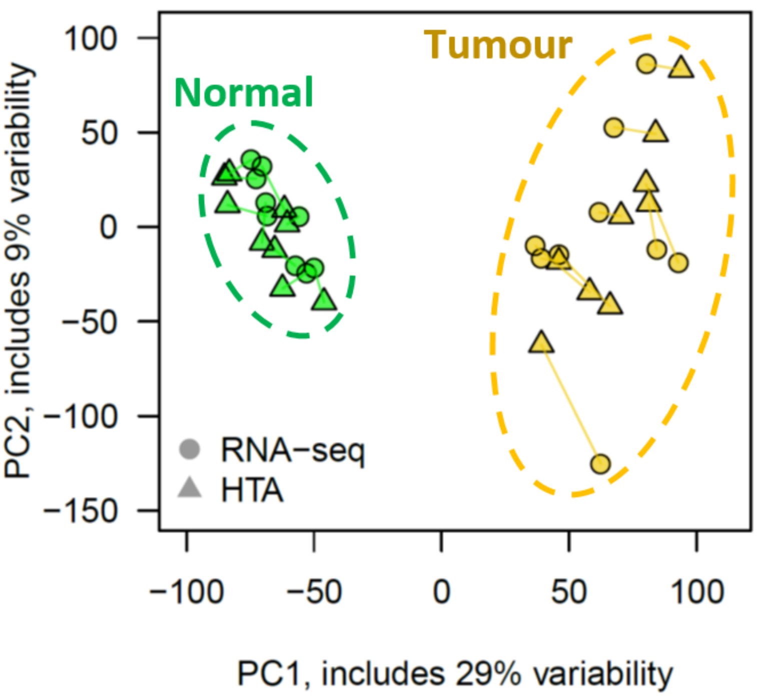


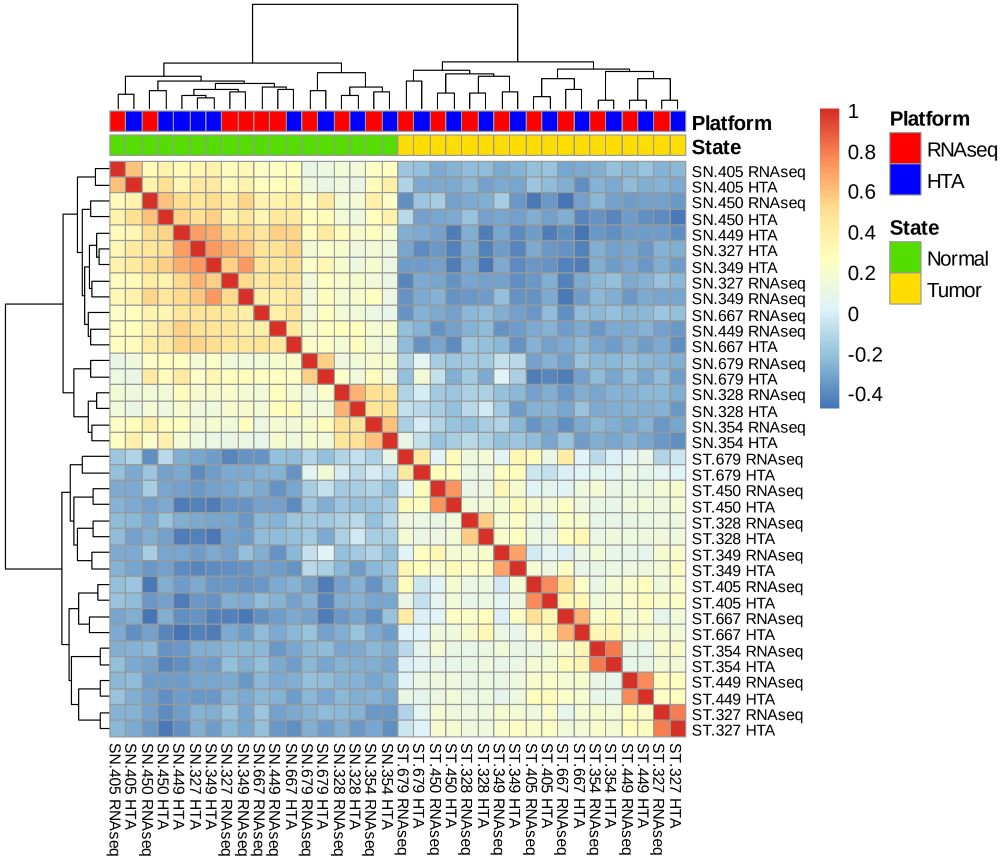
Important:
in order to compare the platforms, we re-mapped Affymetrix probesets onto the Ensembl 69 genome using GenomicRanges package of R.

Clustering

Coding genes: removable platform effect

Linearly scaled/centered data



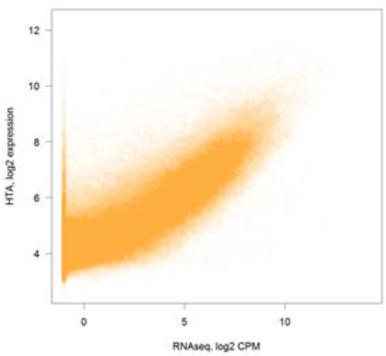


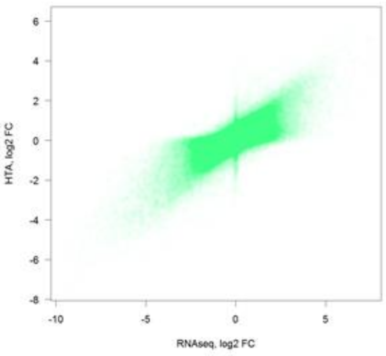
- Strong effect of tumour/normal condition
- Platform-specific effect can be reduced by simple centring-scaling (standardization)
- lncRNA show similar behavior with, with higher variability

Correlations

Coding genes are more correlated than lncRNA

Correlation	coding mRNA	lncRNA
log signal	0.76	0.319
logFC	0.743	0.349





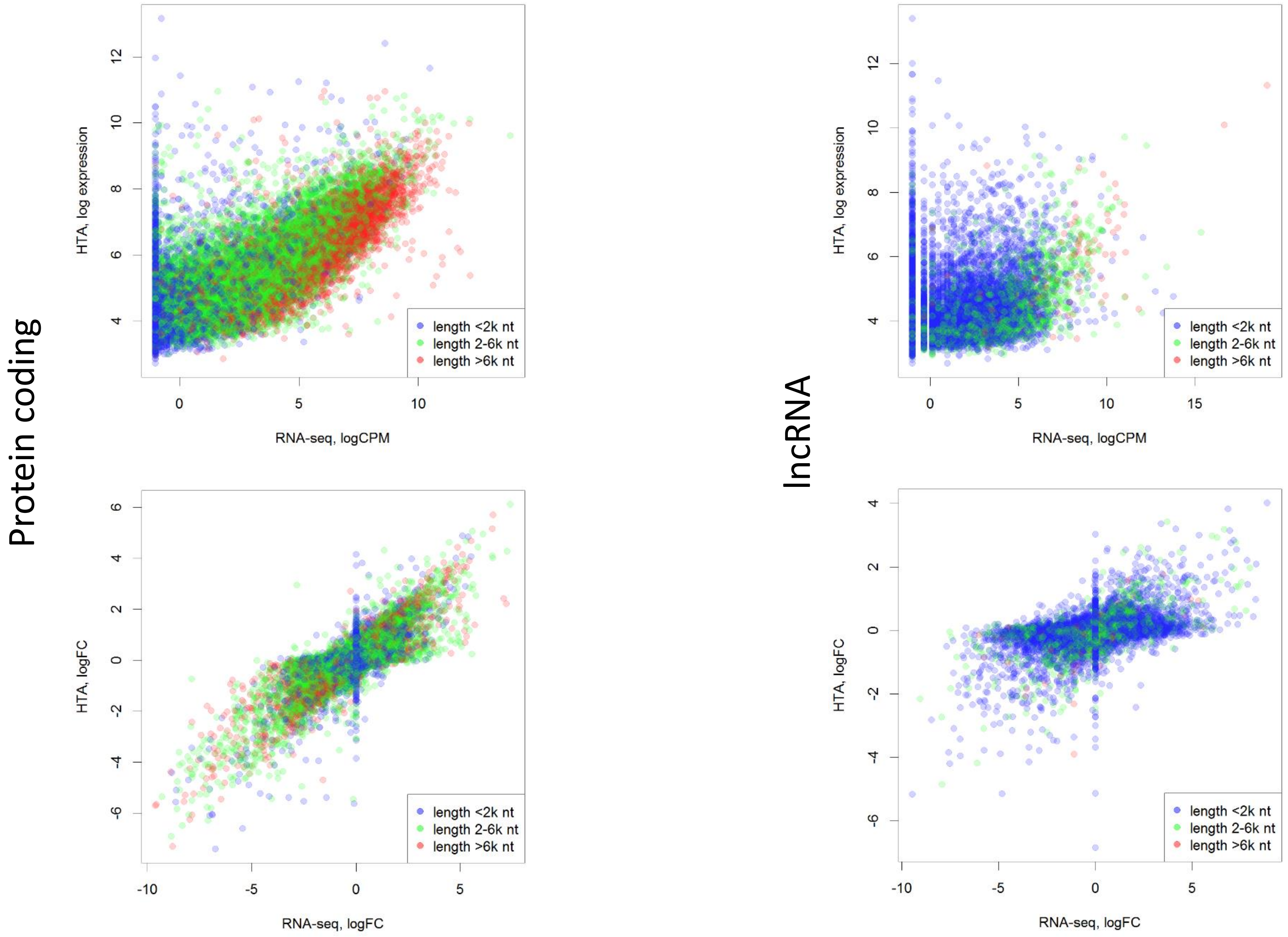
Scatter plots showing general tendency in RNA-seq and HTA protein coding gene expressions (orange) and logFC (green). Scatter plots are built by overlap of all available data for SCC patients.

- Correlation for protein coding genes is in range of values reported in literature
- lncRNA are not so nicely correlated. Reason? ↓

Correlations

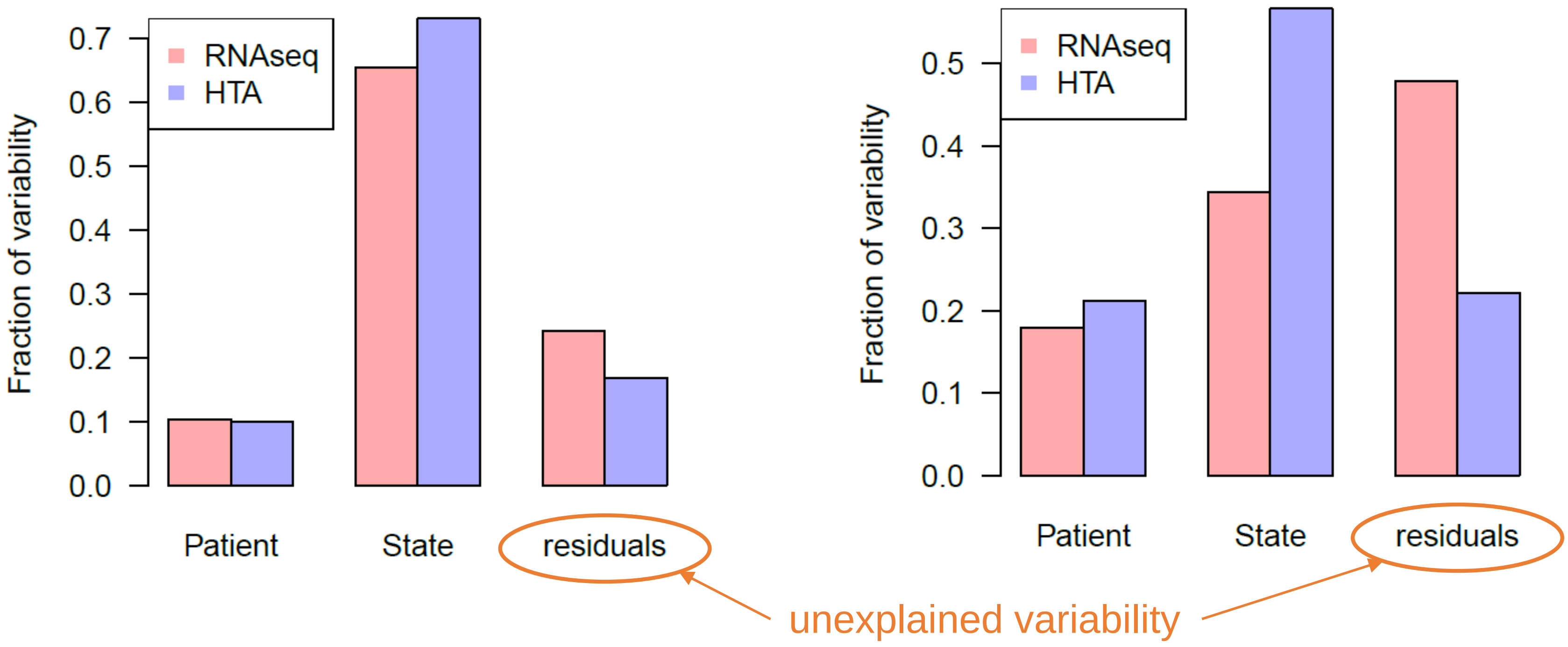
Clinical research study

Gene length matters!



Variability

Explained variability in the data: better for HTA



protein coding

lncRNA

unexplained variability

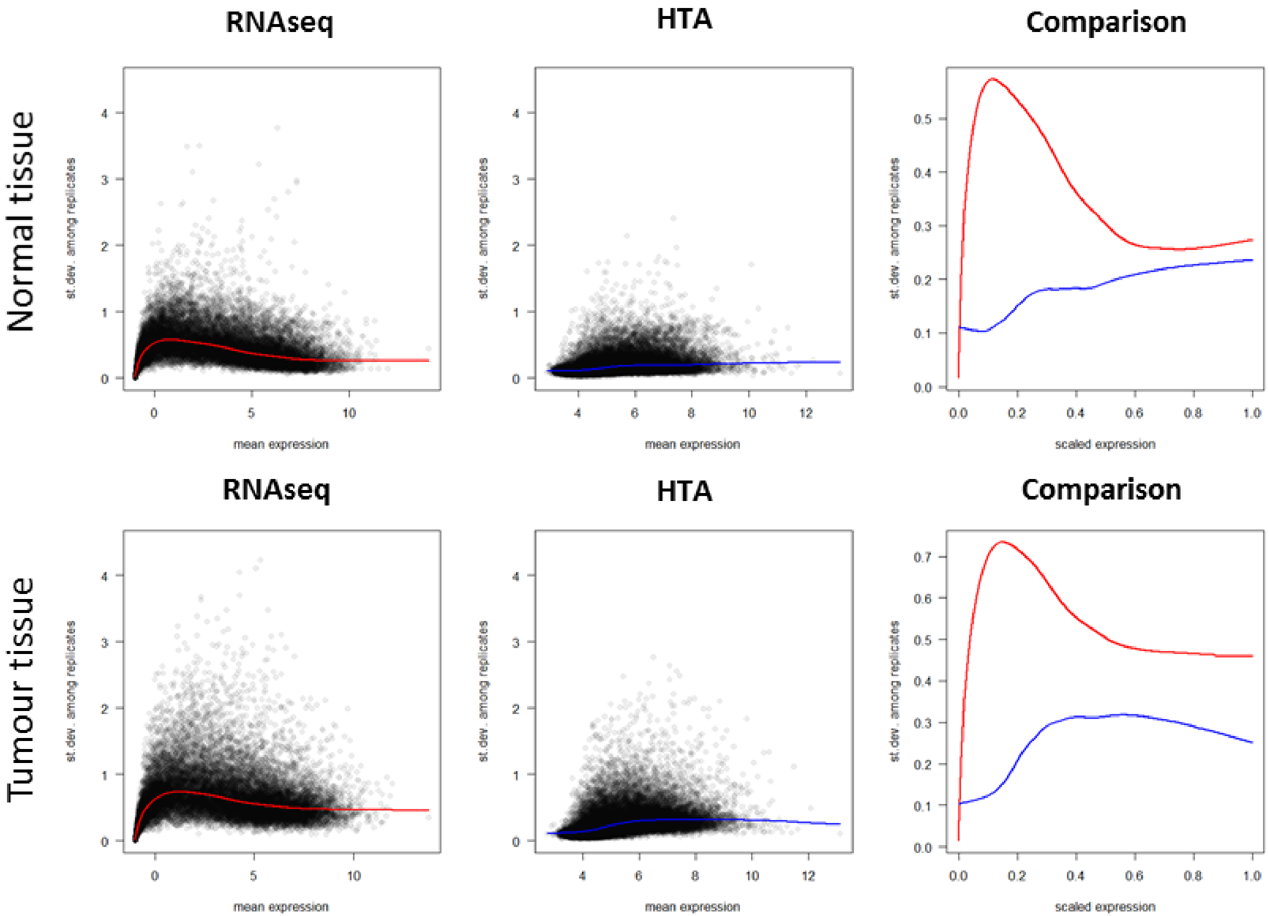
- HTA show less unexplained variability and higher cancer-associated variability

Principal Variance Component Analysis (PVCA) was described in:

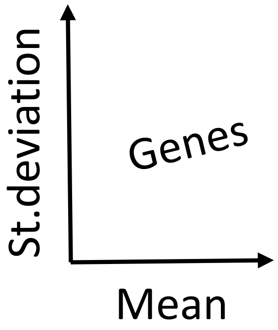
Li, J., Bushel, P., Chu, T.-M., and Wolfinger, R.D. (2009) Principal Variance Components Analysis: Estimating Batch Effects in Microarray Gene Expression Data, Batch Effects and Noise in Microarray Experiments: Sources and Solutions, ed. A. Scherer, John Wiley & Sons.

Variability

St.deviation in biological replicates is higher in RNA-seq



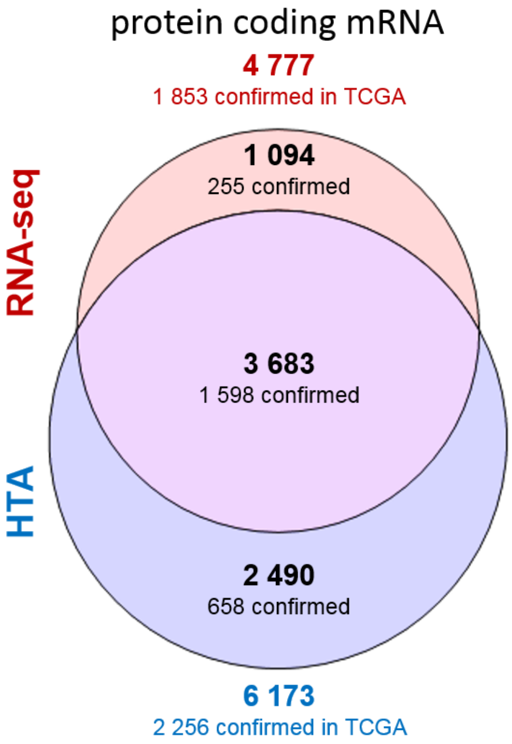
RNA-seq
HTA



Variability between biological replicates is higher for RNA-seq data for both normal and tumour samples, especially for lowly abundant transcripts

Differential Expression Analysis

DE gene lists vs TCGA: similar level of confirmation



protein coding mRNA

4 777
1 853 confirmed in TCGA

RNA-seq

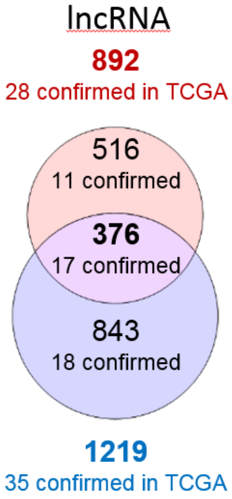
1 094
255 confirmed

3 683
1 598 confirmed

2 490
658 confirmed

HTA

6 173
2 256 confirmed in TCGA



lncRNA

892
28 confirmed in TCGA

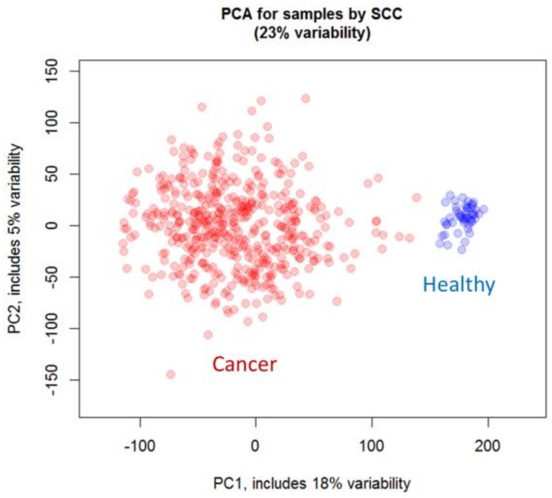
516
11 confirmed

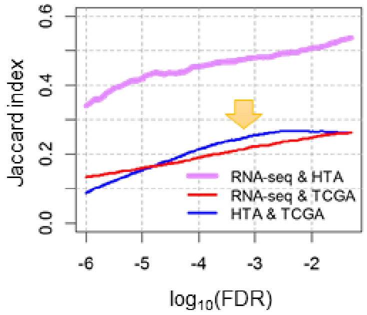
376
17 confirmed

843
18 confirmed

1219
35 confirmed in TCGA

TCGA LUSC data series: 502 -vs- 31





- More DEG for HTA with FDR<0.01
- Comparing with TCGA – similar confirmation rate
- Overlapping genes: 1598 of 3683 are found in the top 25% of TCGA

Detection Limits

How to compare “pears” with “apples”?

We proposed considering only significant genes, in order to make the analysis more fair.

Measure	**RNA-seq**	**HTA**
Lower limit of log expression	-0.80	3.83
Higher limit of log expression	9.20	8.89
Dynamic range of log expression	**10.00**	**5.06**
Lower limit of absolute logFC	0.67	0.17
Lower limit of absolute logFC	7.55	3.58
Dynamic range of absolute logFC	**6.87**	**3.41**

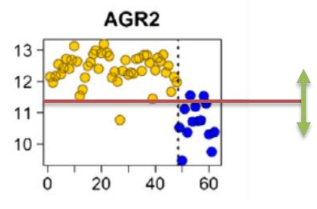
Values are in log_2

- As expected, dynamic range of RNA-seq is higher. But taking into account that HTA allow for detecting genes with smaller fold change - it still can be related to difference in scales.

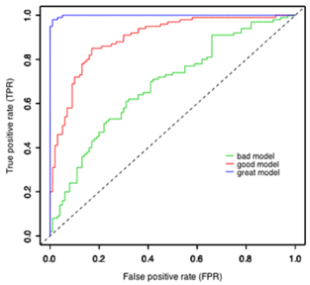
Prediction Analysis

More predictive genes were observed with arrays

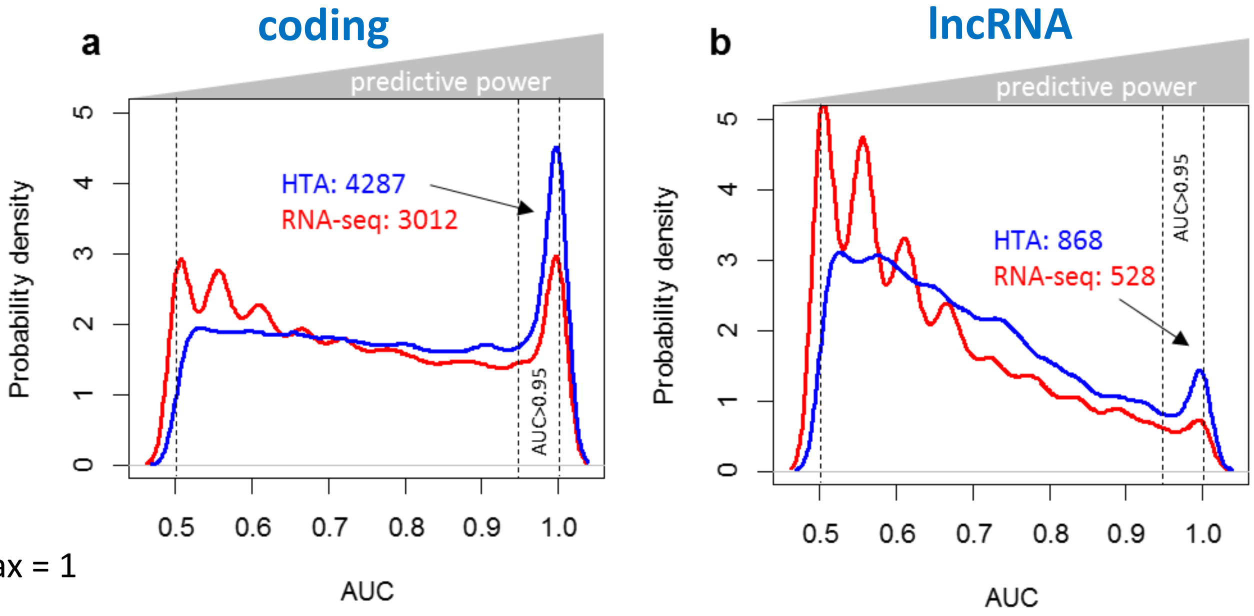
Area under ROC curve (AUC) characterizes applicability of a gene to distinguish between 2 groups of samples and, therefore, tells whether a gene can be used as a marker to predict the group.



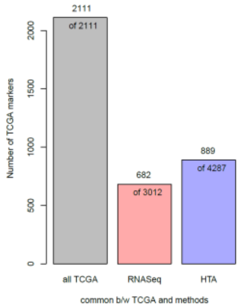
ROC curves



AUC – area
min 0.5, max = 1



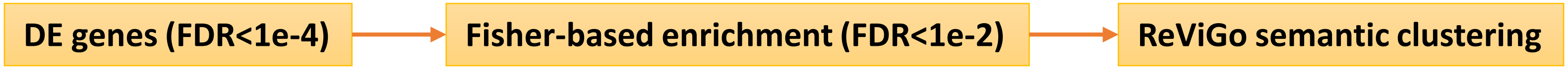
TCGA validation AUC>0.95



- ➢ AUC constantly shows better values for HTA data

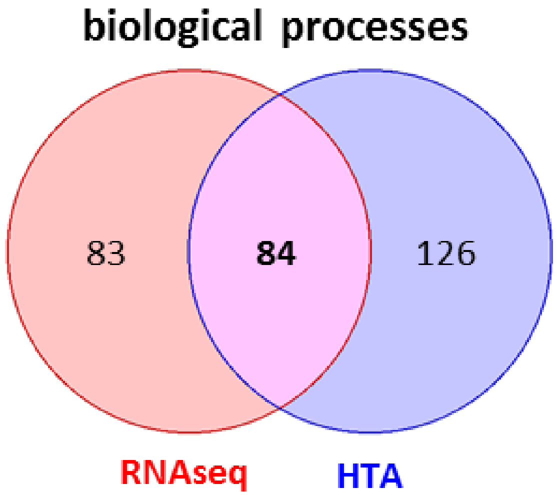
Gene Set Analysis

Biological processes (GO:BP) enriched with DE genes



DE genes (FDR<1e-4) → Fisher-based enrichment (FDR<1e-2) → ReViGo semantic clustering

topGO package of R/Biocondictor



-Σlog(FDR) > 100

-Σlog(FDR) > 10

-Σlog(FDR) > 2

RNAseq

- tissue development
- collagen catabolism
- extracellular matrix organization
- positive regulation of mitotic cell cycle
- cellular component movement
- developmental process
- single-organism cellular process
- single-organism process
- cell proliferation
- multicellular organismal process
- reproduction
- response to alcohol

common

- cell cycle process
- cilium organization
- DNA metabolism
- microtubule-based movement
- microtubule-based process
- cell cycle
- cellular component organization or biogenesis
- cell division
- chromosome segregation
- regulation of cell division
- anatomical structure homeostasis
- protein localization to chromosome
- response to ionizing radiation

HTA

- protein-DNA complex assembly
- DNA integrity checkpoint
- cellular response to DNA damage stimulus
- RNA transport regulation of ligase activity
- epithelial cilium movement involved in determination of left/right asymmetry
- single-organism metabolism

- ➢ GO:BP biases are found: extracellular in RNA-seq , DNA-related in HTA
- ➢ More GO:BP in with HTA analysis

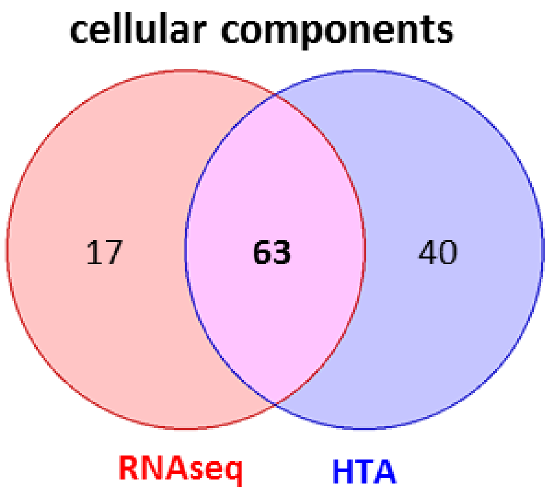
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Gene Set Analysis

Cellular components (GO:CC) enriched with DE genes

DE genes (FDR<1e-4) → **Fisher-based enrichment (FDR<1e-2)** → **ReViGo semantic clustering**

topGO package of R/Biocondictor



cellular components

$-\Sigma\log(FDR) > 100$

$-\Sigma\log(FDR) > 10$

$-\Sigma\log(FDR) > 2$

RNAseq

- proteinaceous extracellular matrix
- extracellular region
ciliary tip
- extracellular matrix
- cell-cell junction
- cornified envelope
- intraciliary transport particle
- chaperonin-containing T-complex
- collagen trimer
- intraciliary transport particle B

common

- microtubule cytoskeleton
- cilium
- extracellular vesicular exosome
- non-membrane-bounded organelle
- organelle part
- membrane-enclosed lumen
- organelle lumen
- organelle
- protein complex
- cytosol
- cytoplasm
- macromolecular complex
- cell projection
- proteasome accessory complex
- vesicle
- midbody
- MCM complex
- desmosome

HTA

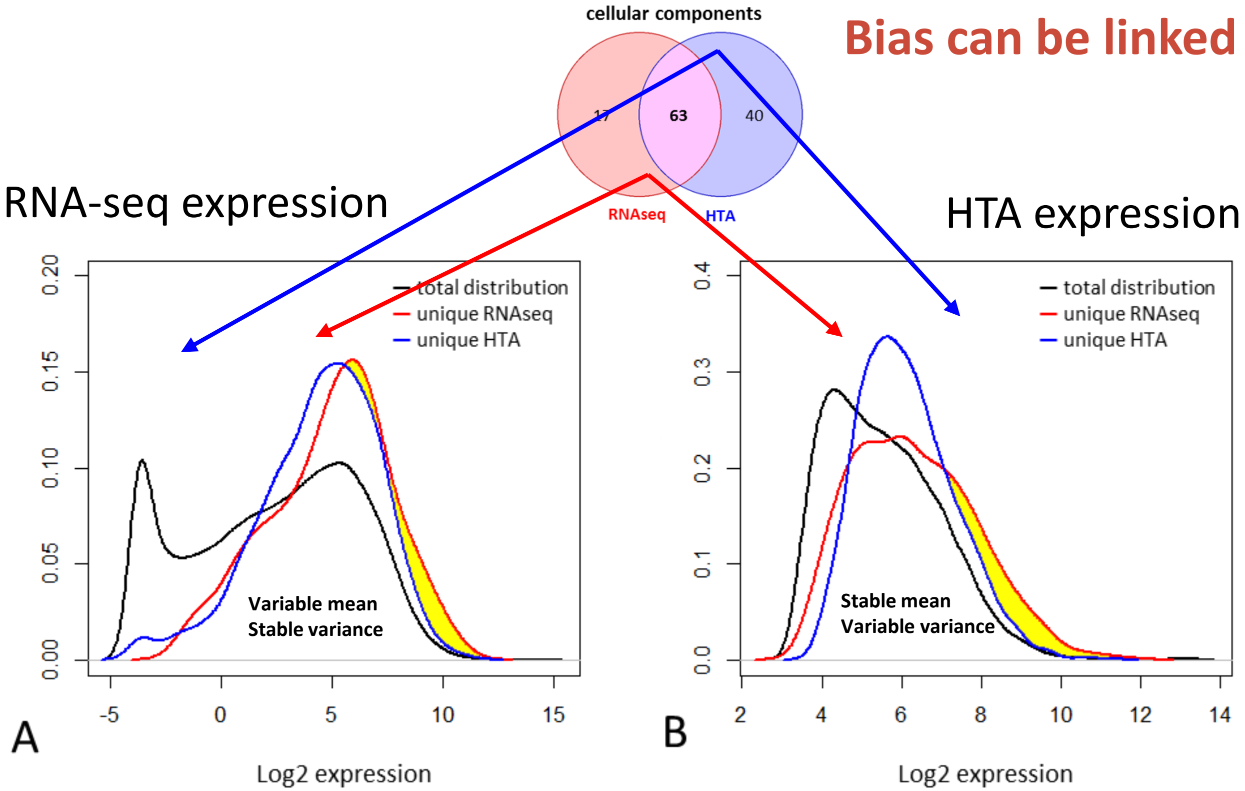
- nucleoplasm
- intracellular part
- intracellular
- intracellular organelle
- membrane-bounded organelle
- DNA packaging complex
- protein-DNA complex
- DNA bending complex
- DNA polymerase complex
- cell
- cell part
- pore complex
- proteasome complex
- envelope

➢ GO:CC biases are found: extracellular in RNA-seq , nucleus in HTA

➢ More GO:CC in with HTA analysis, again

Gene Set Analysis

Bias can be linked to RNA abundance



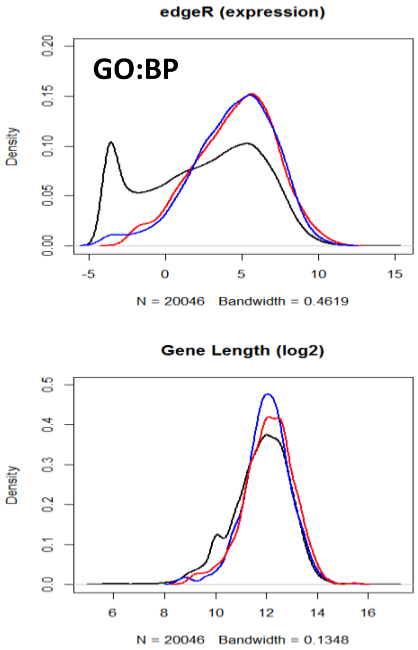
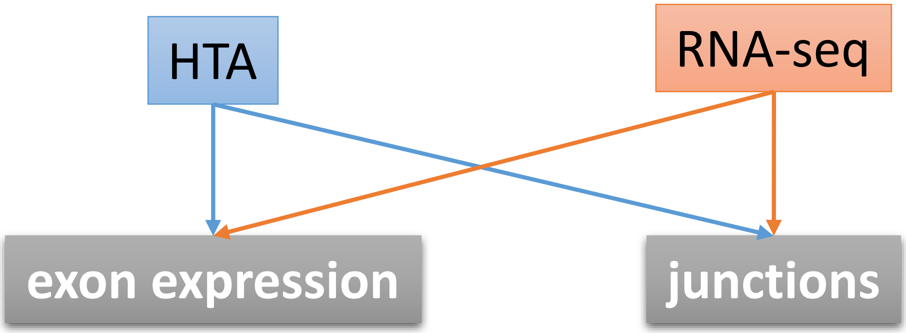


Figure S6. Expression of the genes related to cellular component ontologies uniquely identified by RNA-seq (red lines) and HTA (blue lines). The distributions of gene expressions are based on sequencing (A) and microarray (B) data. Both data agree, that genes participating in the functions uniquely found in RNA-seq analysis show higher expression than one of HTA analysis (yellow area).

- Abundance of the genes participation in extracellular biofunctions is higher then for nucleus-related genes.
- Small bias of the length was seen as well, but it cannot explain the expression differences: checked with ***goseq*** package (correcting for gene length)
- Strong bias is seen only for CC. Only minor for BP

Analysis of Splicing Events

Methods



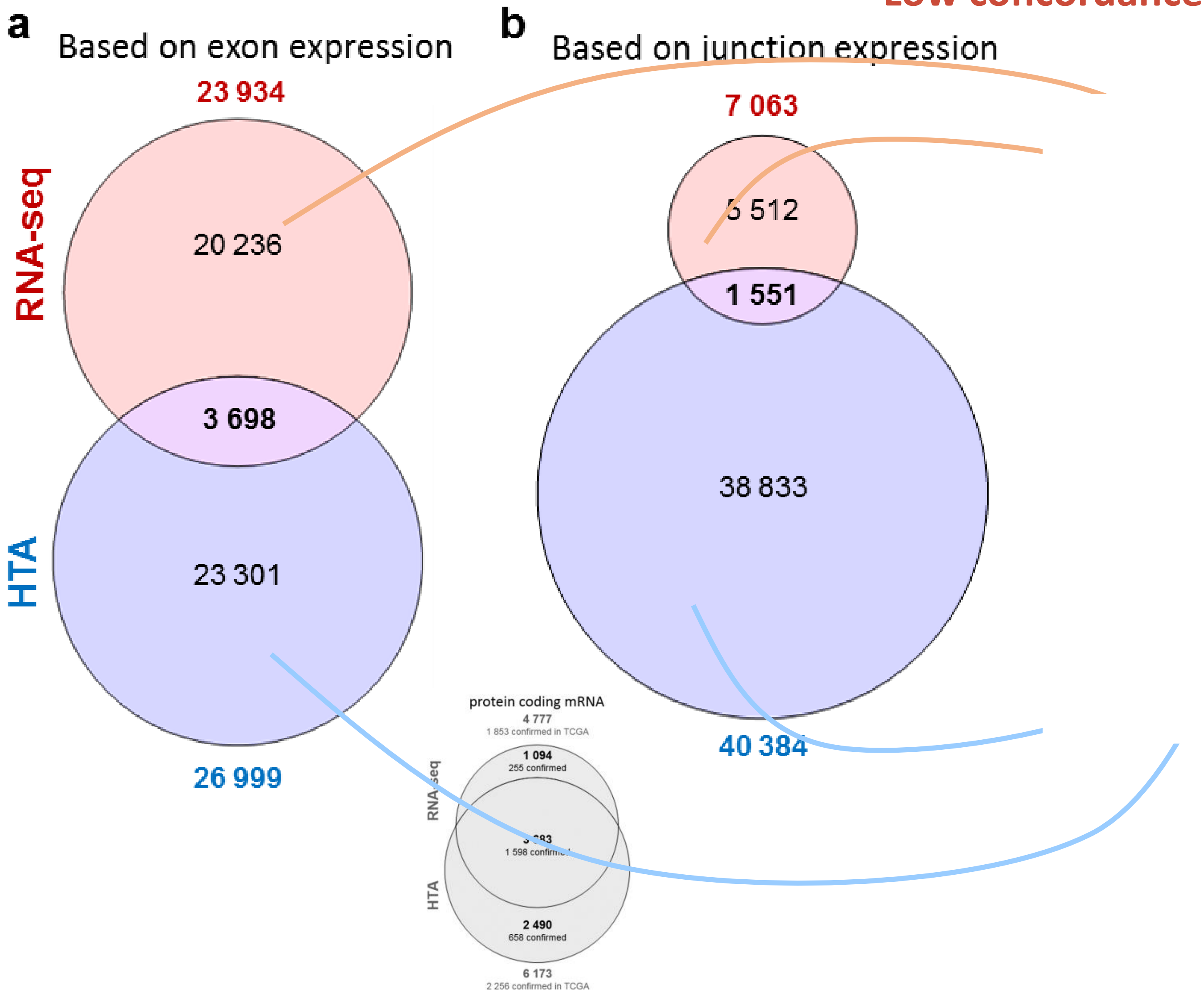
- ➢ Linear models are used
- ➢ HTA: **DiffSplice** from **limma** package
- ➢ RNA-seq: **DEXSeq**

Challenge: HTSeq tool does not work for exons – too many overlapped entities (correlation b/w platforms ≈ **0.2**)

Solution: Changing counting tool to *featureCount (Rsubread)* improved concordance b/w HTA and RNA-seq: correlation ≈ **0.6-0.7**

Analysis of Splicing Events

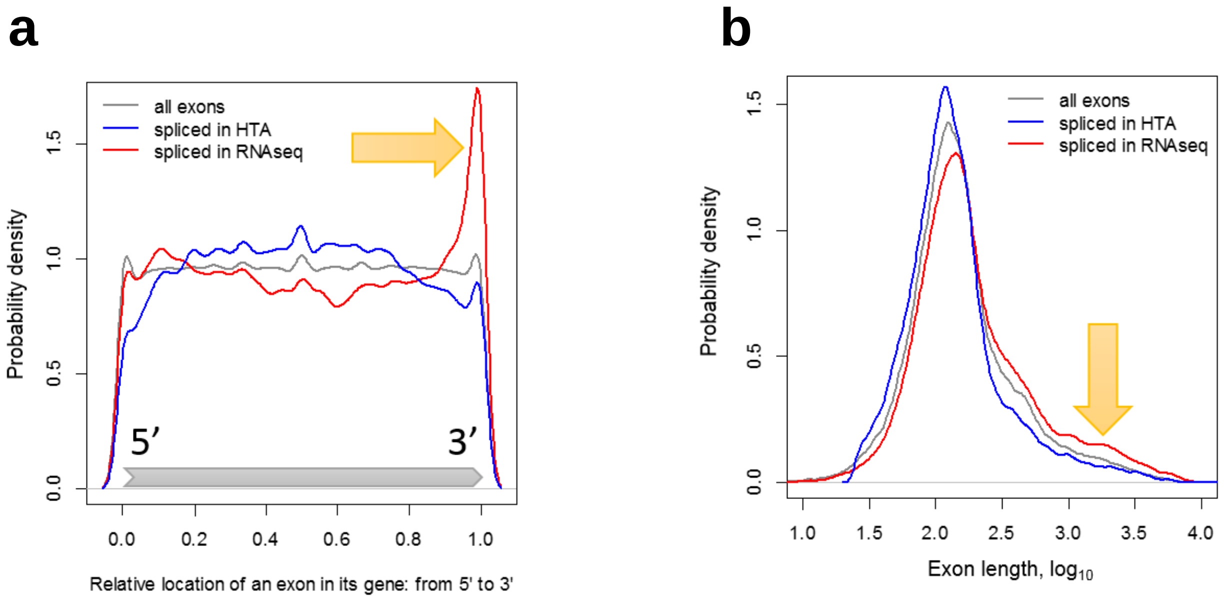
Low concordance of the results



Analysis of Splicing Events

The 3’-exons and long exons show-up in RNA-seq

The exon parameters distribution among differentially used exons detected by the two platforms

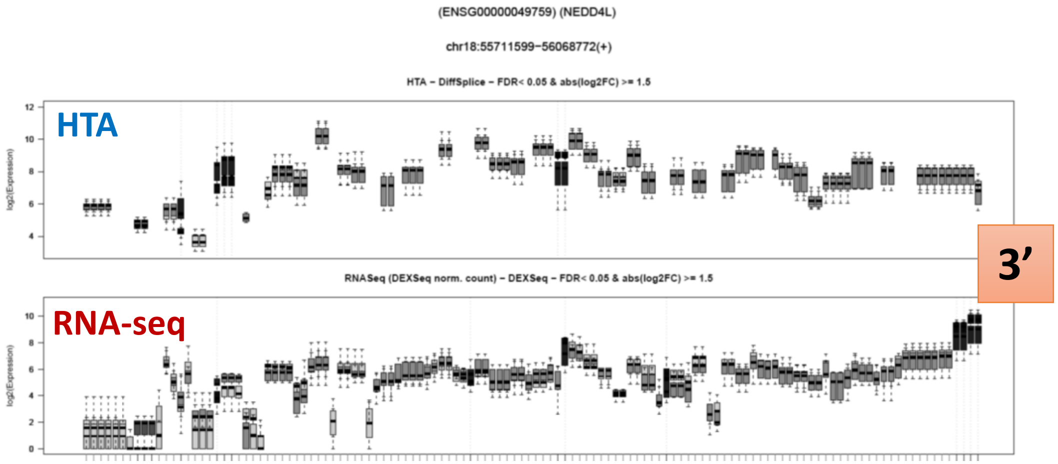


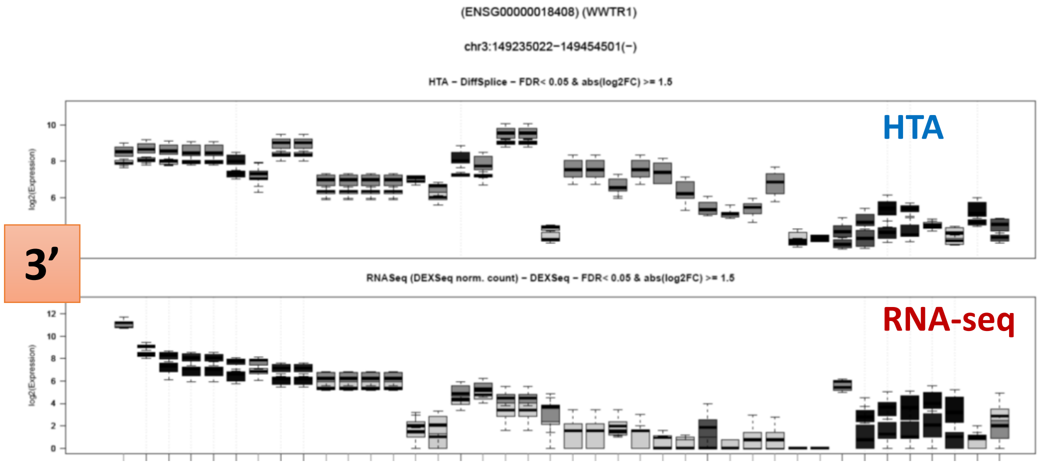
The relative position of the exons within their genes, varying from 5’ end (relative position = 0) to 3’ end (relative position = 1), shows a 3’ bias in RNA-seq (**a**).

Exon length shows that RNA-seq tends to find more significantly splice events among long exons than HTA (**b**).

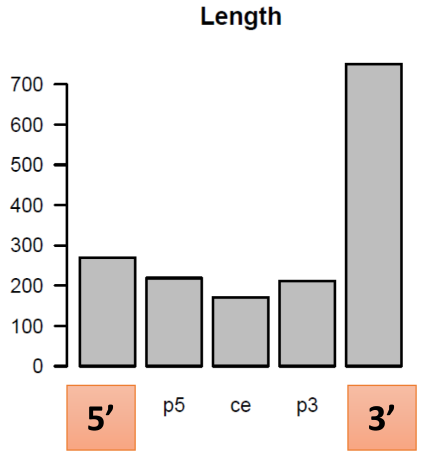
Analysis of Splicing Events

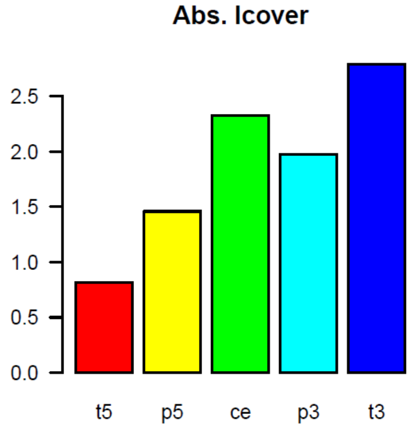
3’ bias or length-related bias?





The RNA-seq data show tendency to increase expression at 3’-end…





Probably 2 effects play role: the length of 3’ exon and poly-A selection. The length bias cannot explain 100% of expression bias

Conclusions I

- In our study, HTA showed more reliable results than RNA-seq with 200M reads.

- Length sensitivity makes RNA-seq a difficult technique for non-coding RNA and requires high coverage.

- RNA-seq is very good as a discovery tool!

- Be careful when doing isoform study with any platform!

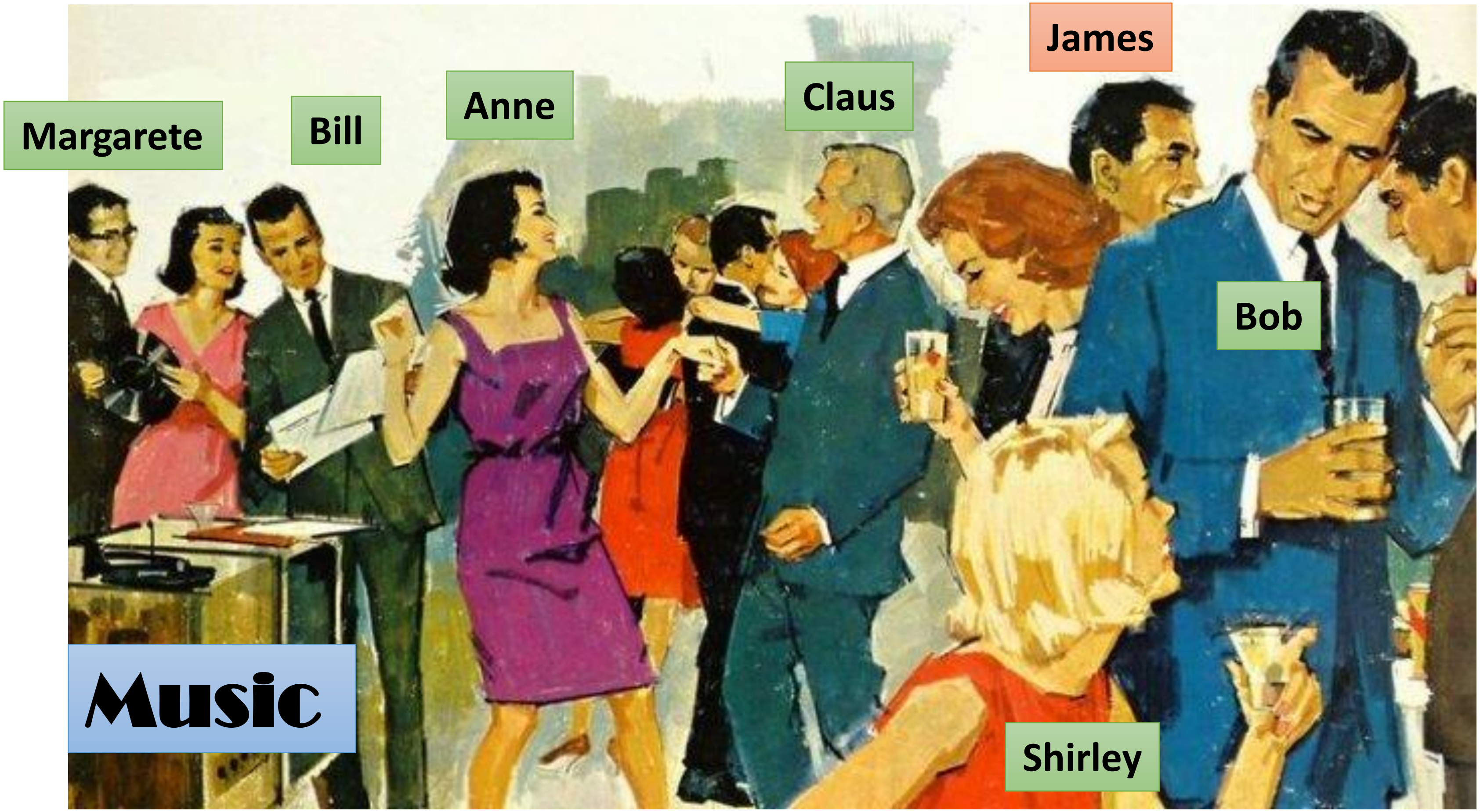
Part II. Independent Component Analysis in Transcriptomics

In collaboration with
Dr. Anke Wienecke and **Dr. Stephanie Kreis,**
Life Science Research Unit, University of Luxembourg

Recently was supported by the Luxembourg National Research (FNR) Fund C17/BM/11664971/DEMICS

Introduction

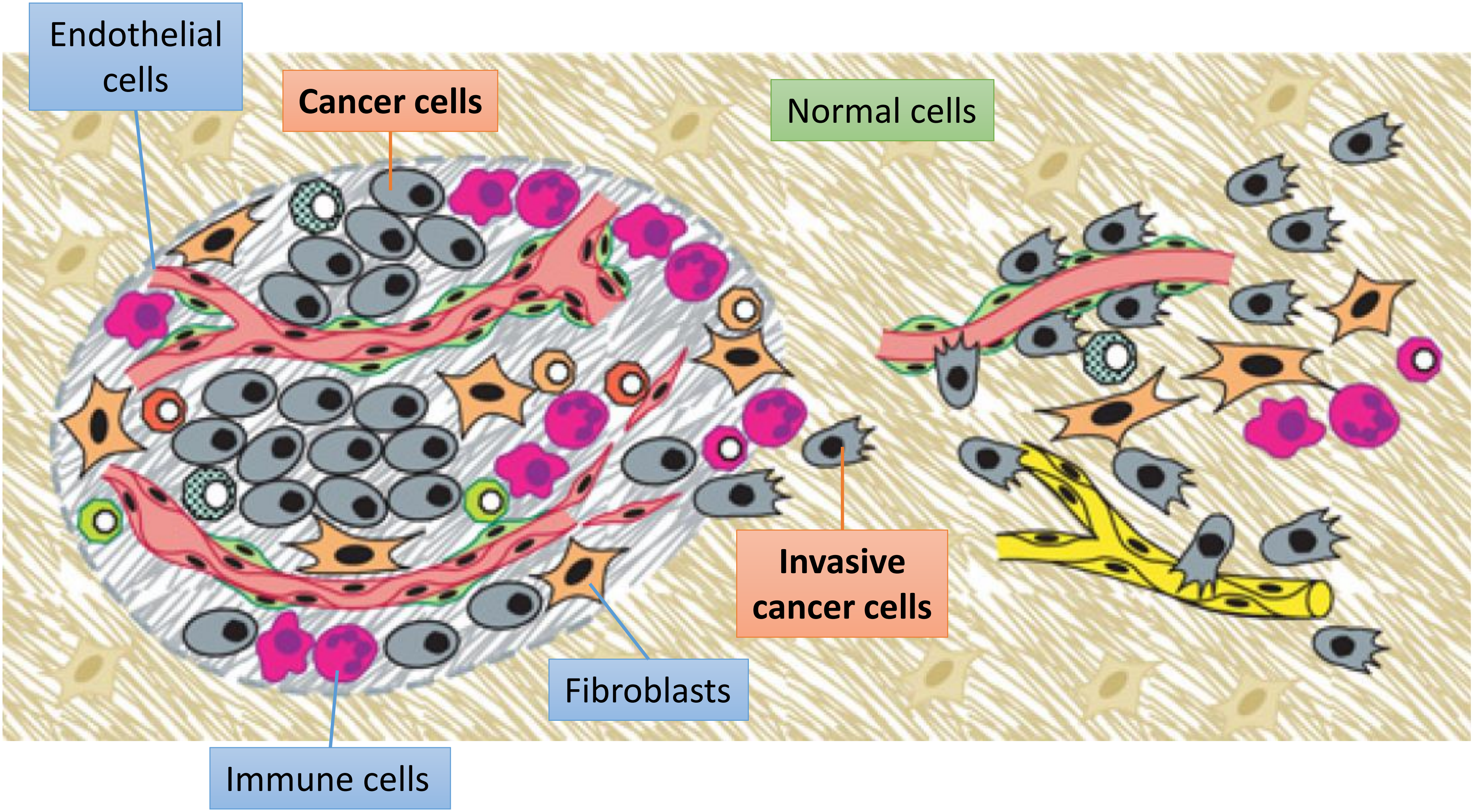
Cocktail party problem



What did James say?..

Introduction

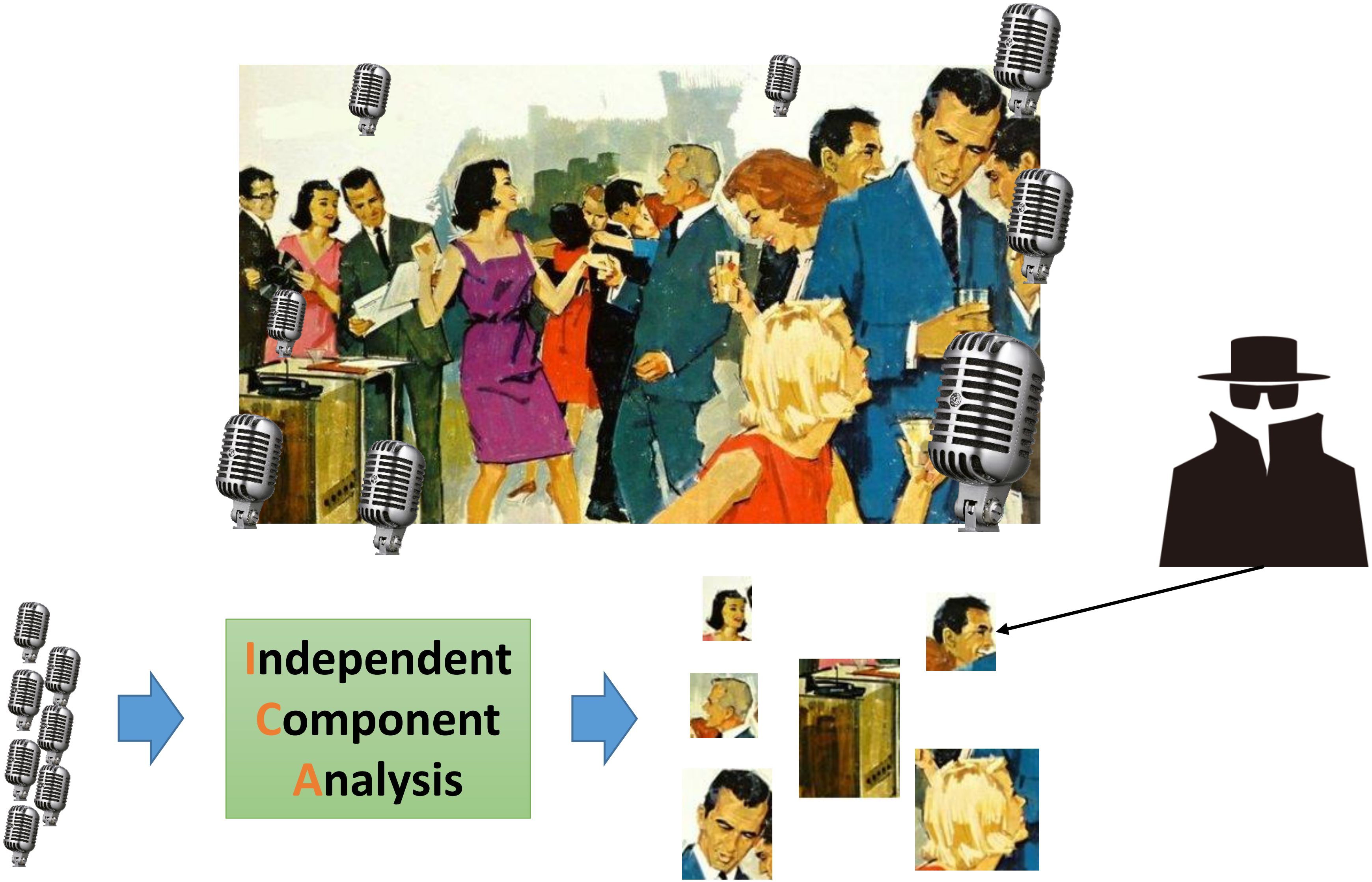
Cell ensemble is as well a “cocktail party”



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

Introduction

The method to solve it…



Independent Component Analysis

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

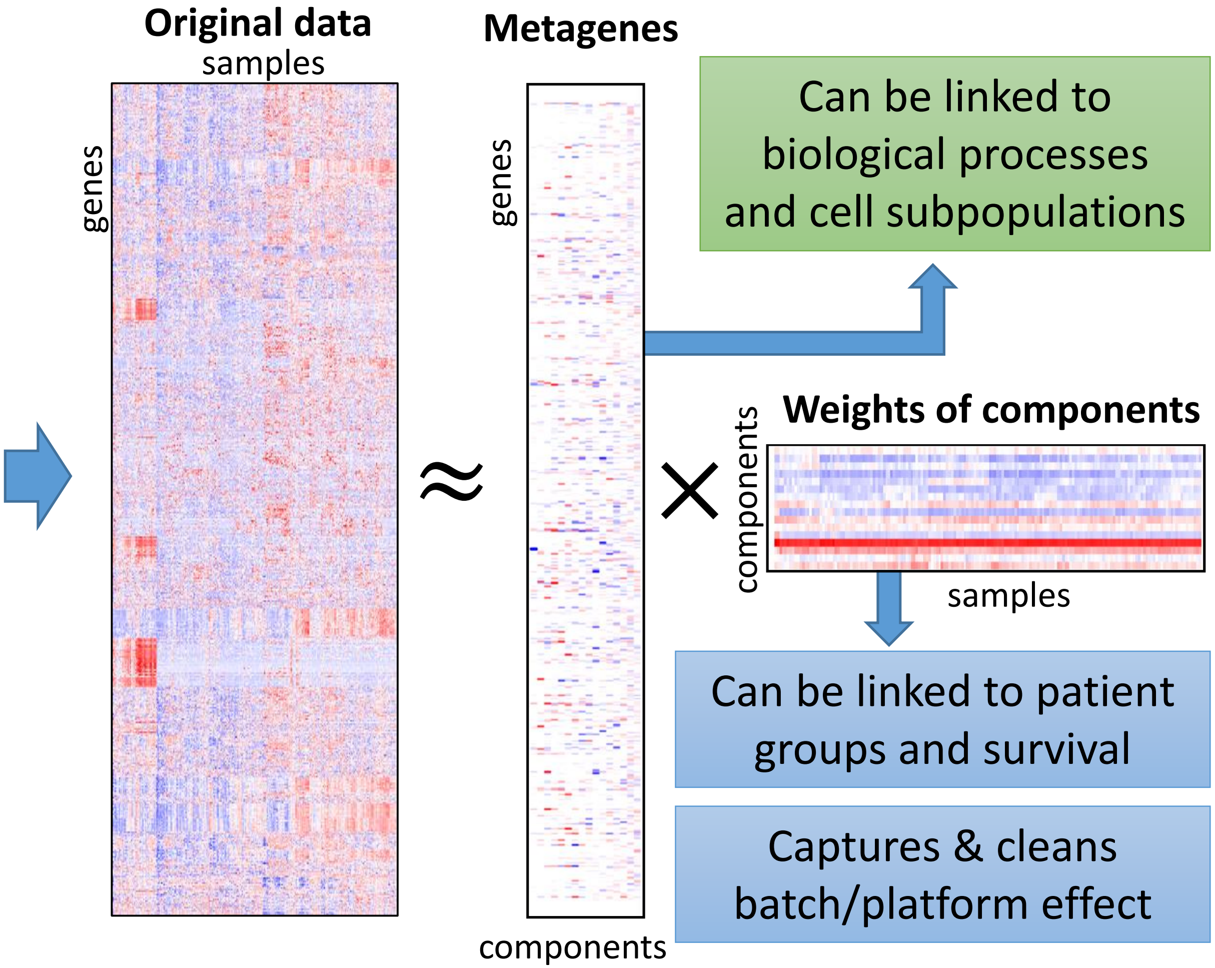
Methods

Independent Component Analysis

Translational research study:

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

Deconvolution of Cell Ensemble



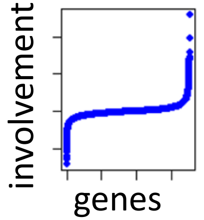
Can be linked to biological processes and cell subpopulations

Can be linked to patient groups and survival

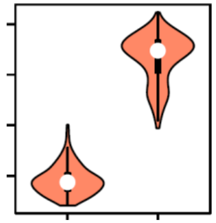
Captures & cleans batch/platform effect

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

One component



Components weights in patients



A. Biton et al, Cell Reports 9, 2014
A. Zinovyev et al, Biochem Biophys Res Commun. 2013

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Methods

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. **No ranking of the components** by importance (not like PCA)
2. Results are **not deterministic** and can to some extent depends on the run
3. **Orientation of the signal is arbitrary** from one run to another
4. If you look for precise estimation of cell fraction – not a good idea (results are qualitative not quantitative)

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

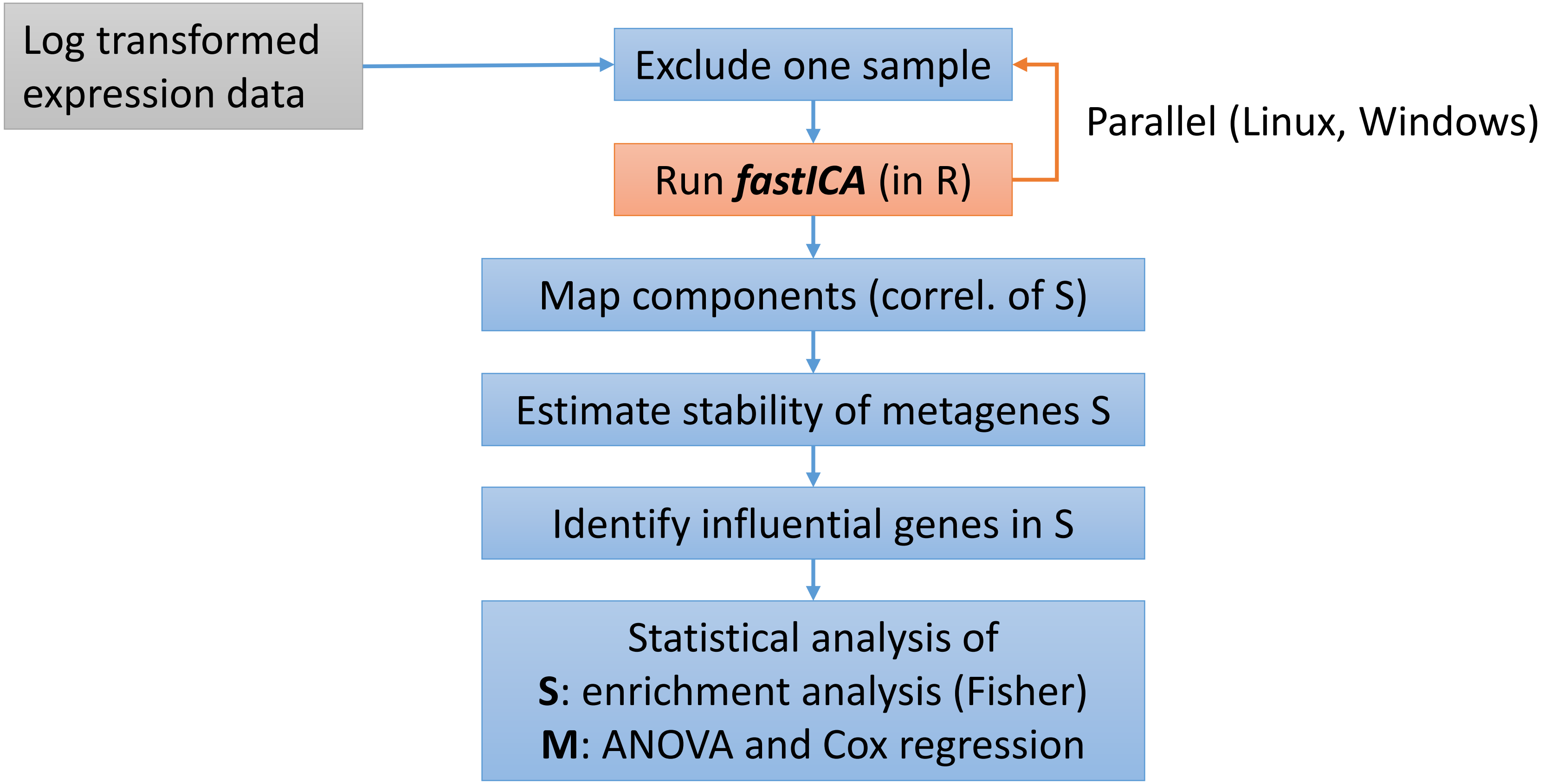
Methods

Consensus ICA

$$\boldsymbol{X}_{gs} \approx < \boldsymbol{S}_{gk} > \times < \boldsymbol{M}_{ks} >$$

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

Positive and Negative Genes within Components

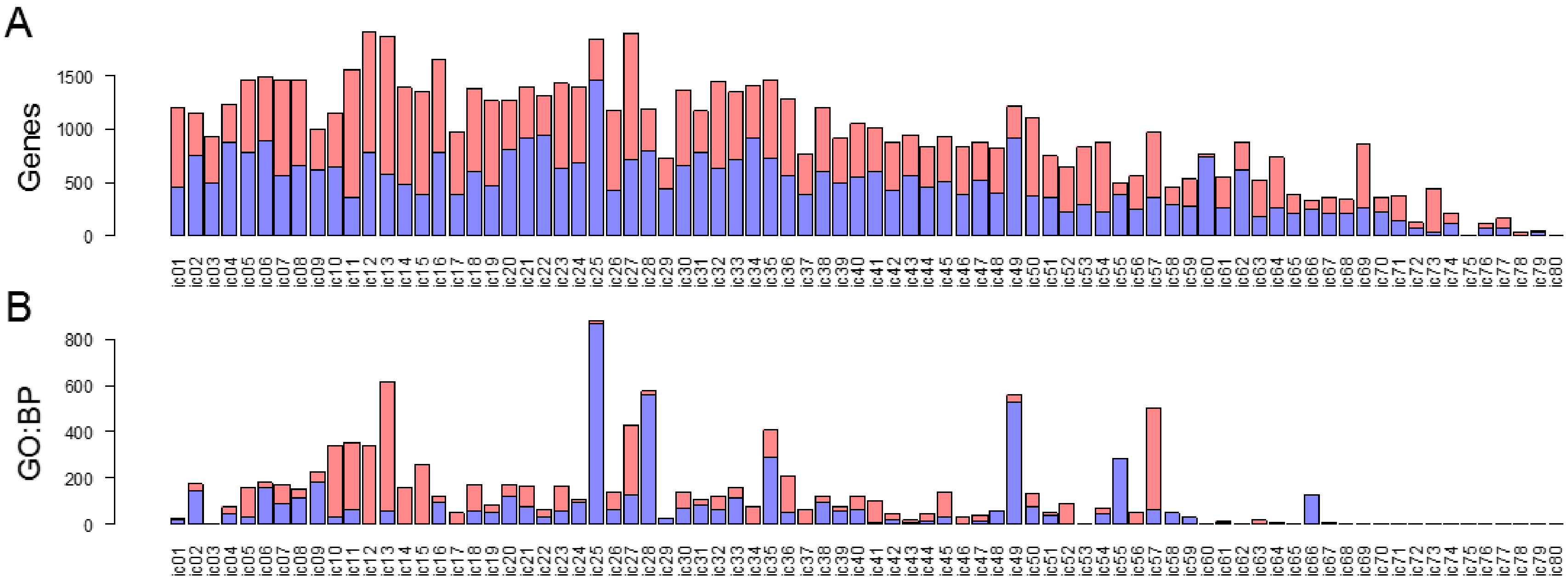
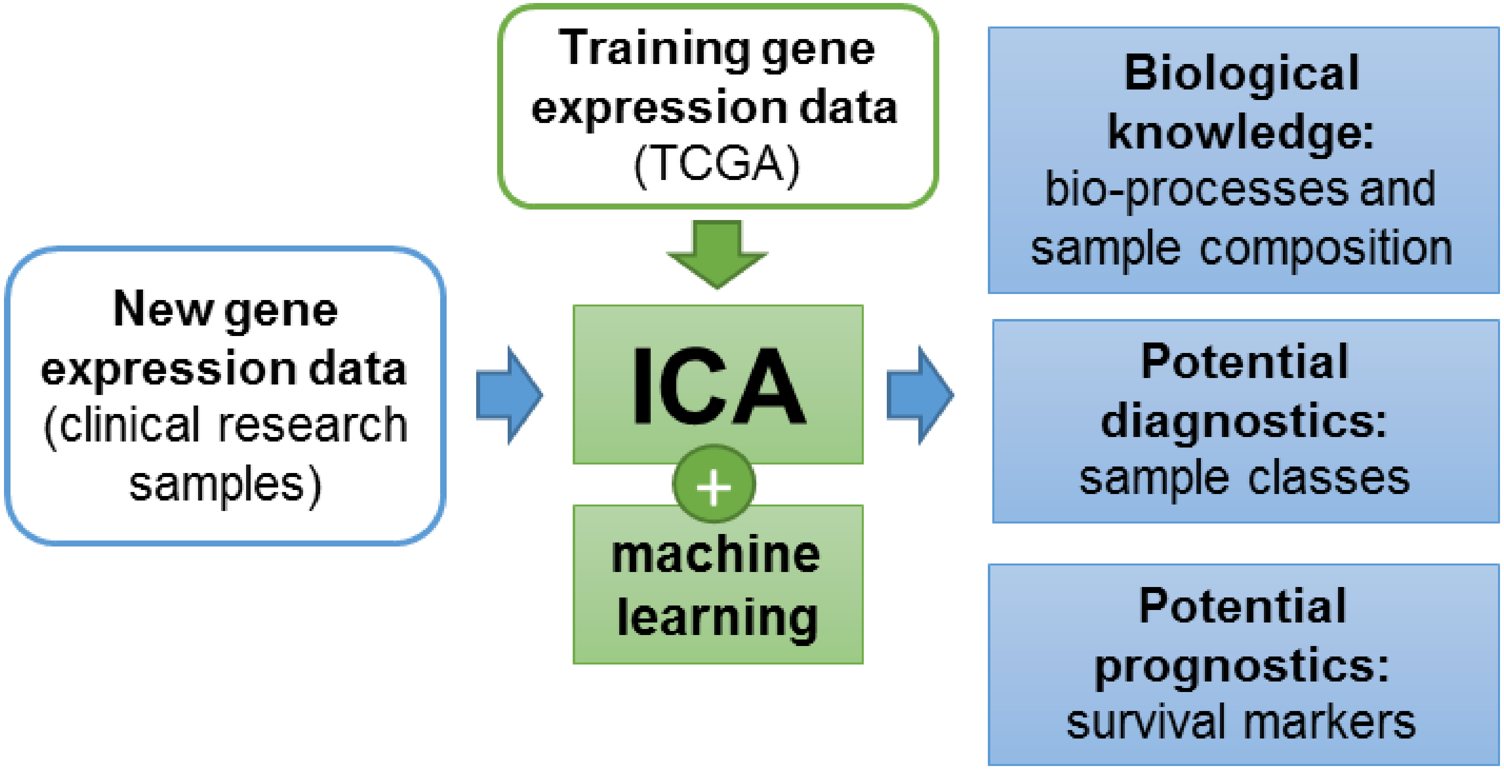


Figure S6. (A) Number of significant positively (red) and negatively (blue) involved genes in metagene of each of the components. (B) Number of enriched GO biological processes found for these genes. For the most cases, only one list of genes is biologically meaningful: either positive (e.g. ic10-ic15)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

Methods

ICA for patient classification

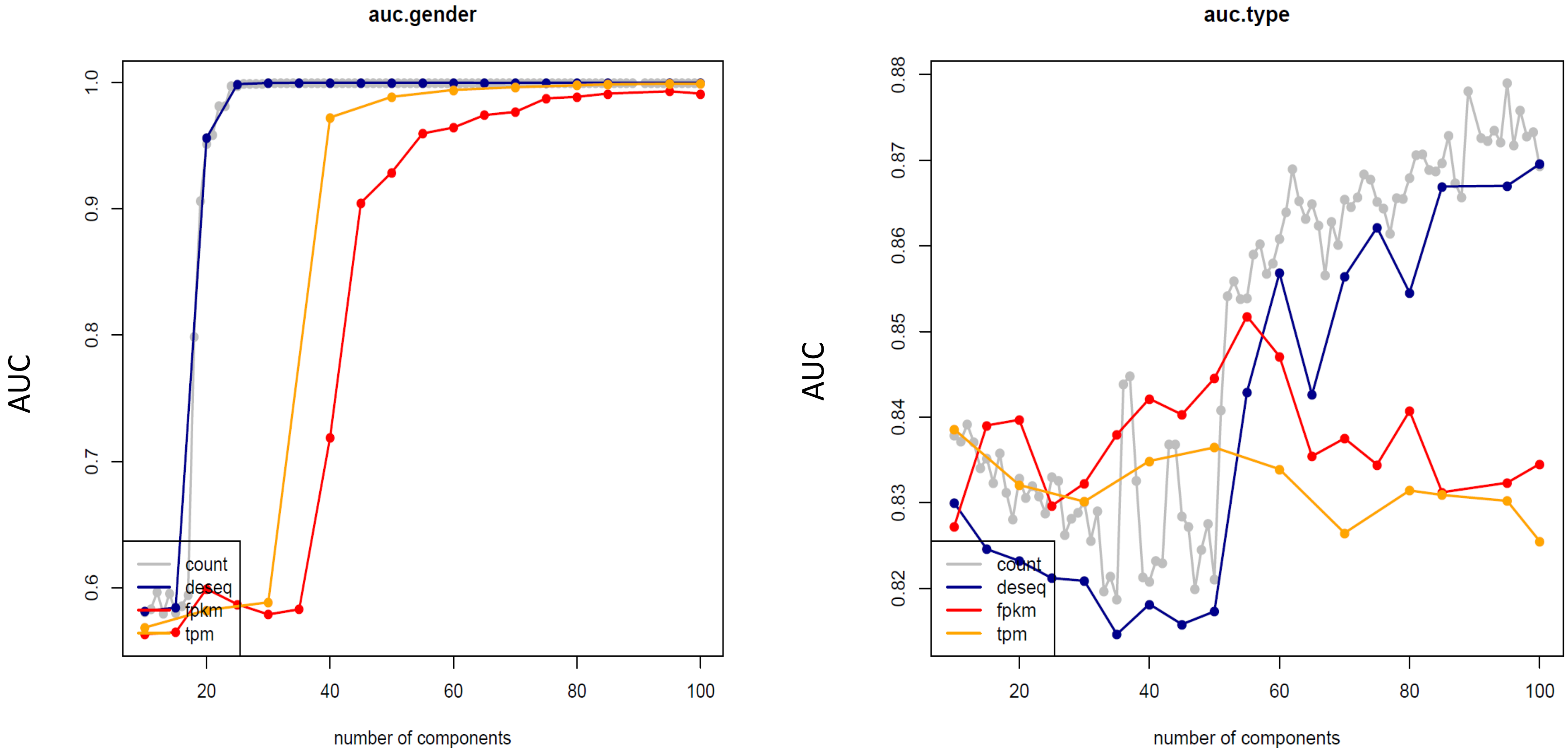
clinical research study


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

Results

Optimal measure for RNA-seq



Raw count
DESeq norm
FPKM
TPM

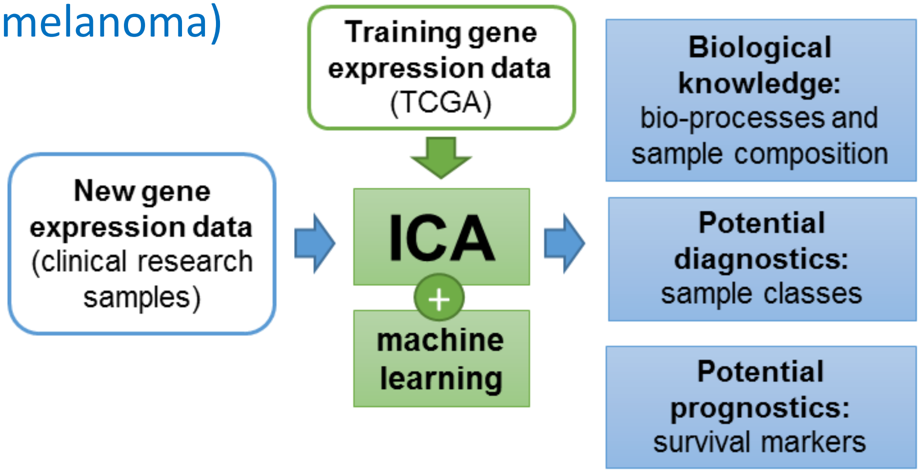
Results

Patient classification in SKCM

SKCM
(skin cutaneous melanoma)

472 samples

5 samples


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

keratinocytes
immune response
melanin metabolic process

immune, keratin, MITF-low

low risk / high risk

- 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

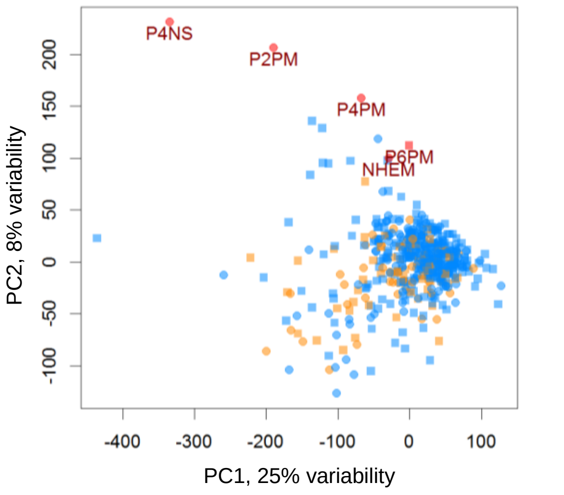
Results

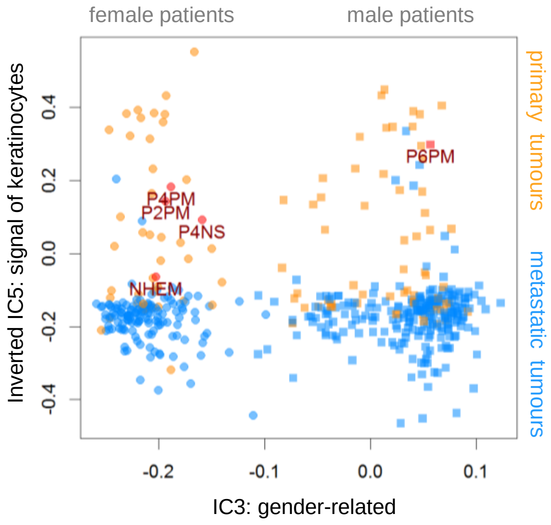
New samples: mRNA and miRNA

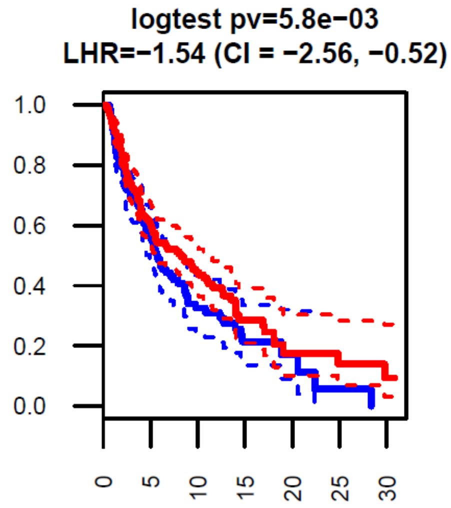
mRNA level: RNA-seq + RNA-seq

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.

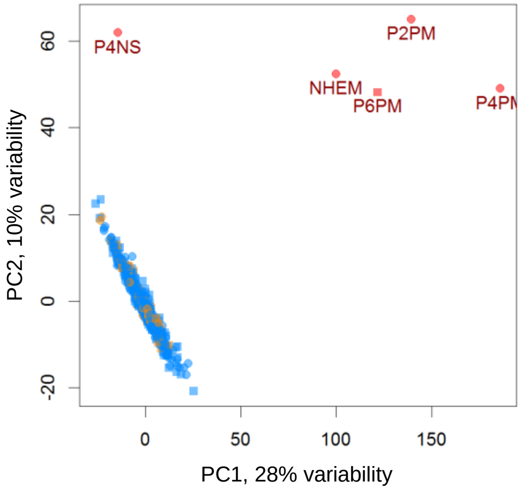
Results

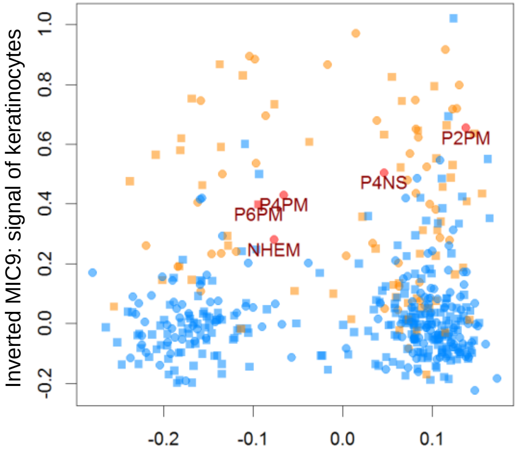
New samples: mRNA and miRNA

Gender: ● female, ■ male

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

miRNA level: RNA-seq + qPCR



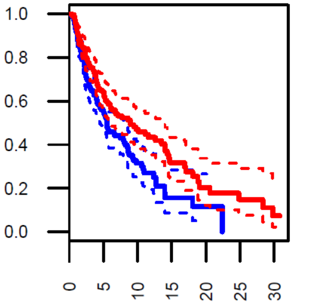


MIC1: unknown segregation of samples

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

logtest pv=9.4e−04
LHR=−1.79 (CI = −2.82, −0.75)

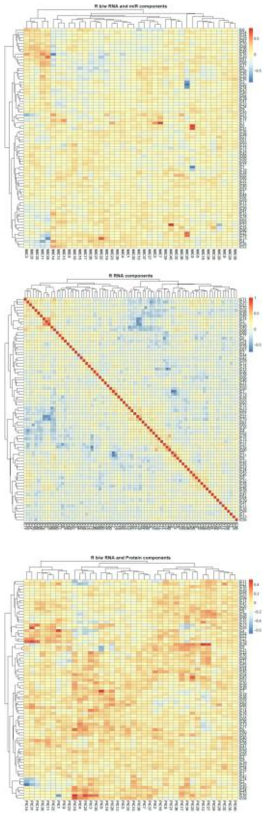


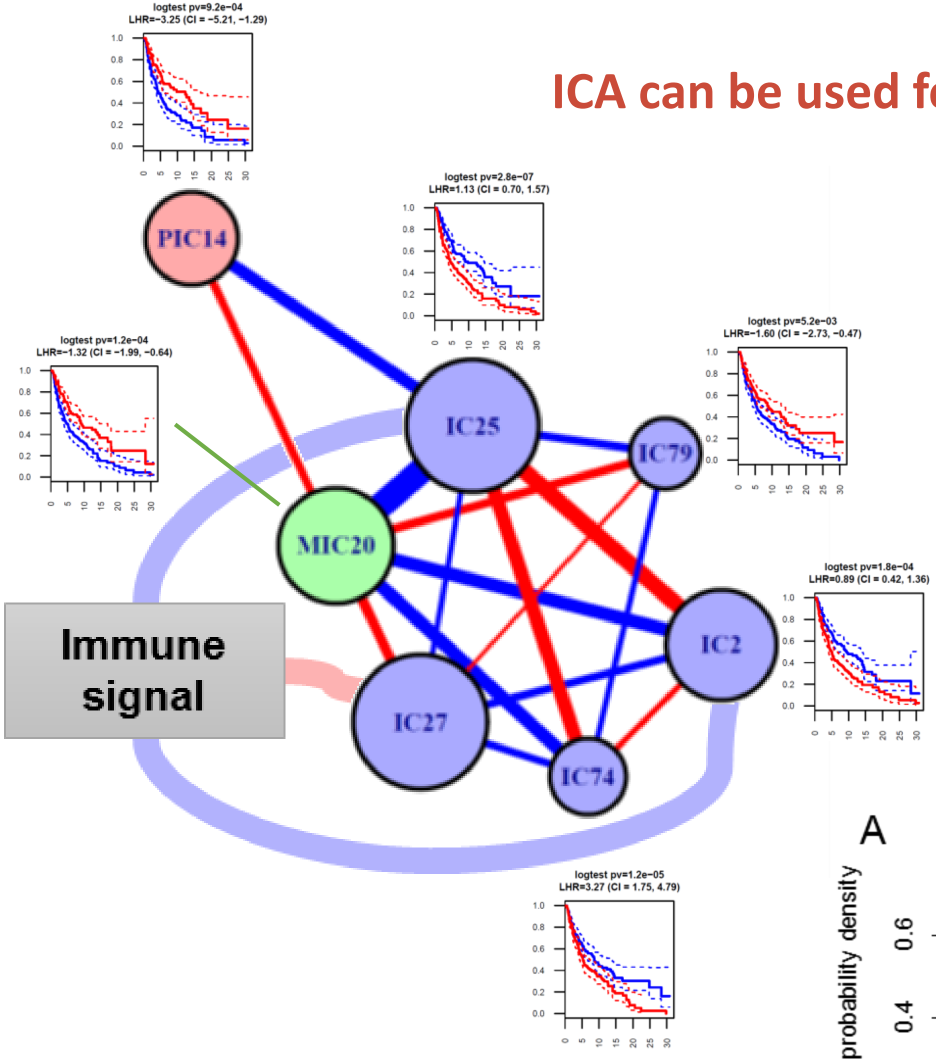
When ICA is run over new samples and training samples together, it corrects for platform bias.

Results

ICA can be used for data integration

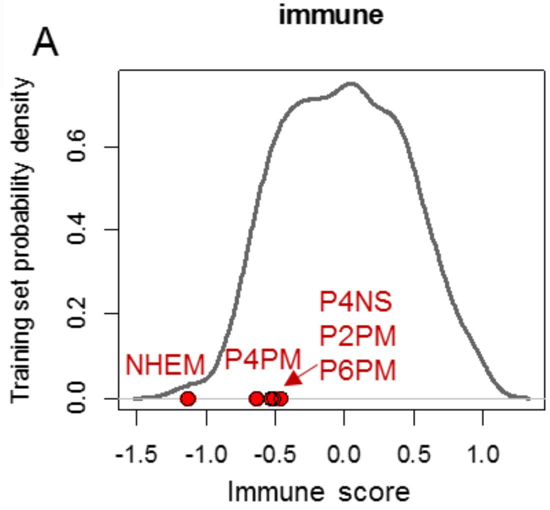
Correlation of weights: mRNA-miRNA-Proteins




logtest pv=9.2e−04
LHR=−3.25 (CI = −5.21, −1.29)
PIC14
logtest pv=2.8e−07
LHR=1.13 (CI = 0.70, 1.57)
logtest pv=5.2e−03
LHR=−1.60 (CI = −2.73, −0.47)
logtest pv=1.2e−04
LHR=−1.32 (CI = −1.99, −0.64)
IC25
IC79
MIC20
IC2
IC27
IC74
logtest pv=1.8e−04
LHR=0.89 (CI = 0.42, 1.36)
logtest pv=1.2e−05
LHR=3.27 (CI = 1.75, 4.79)

Immune signal

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



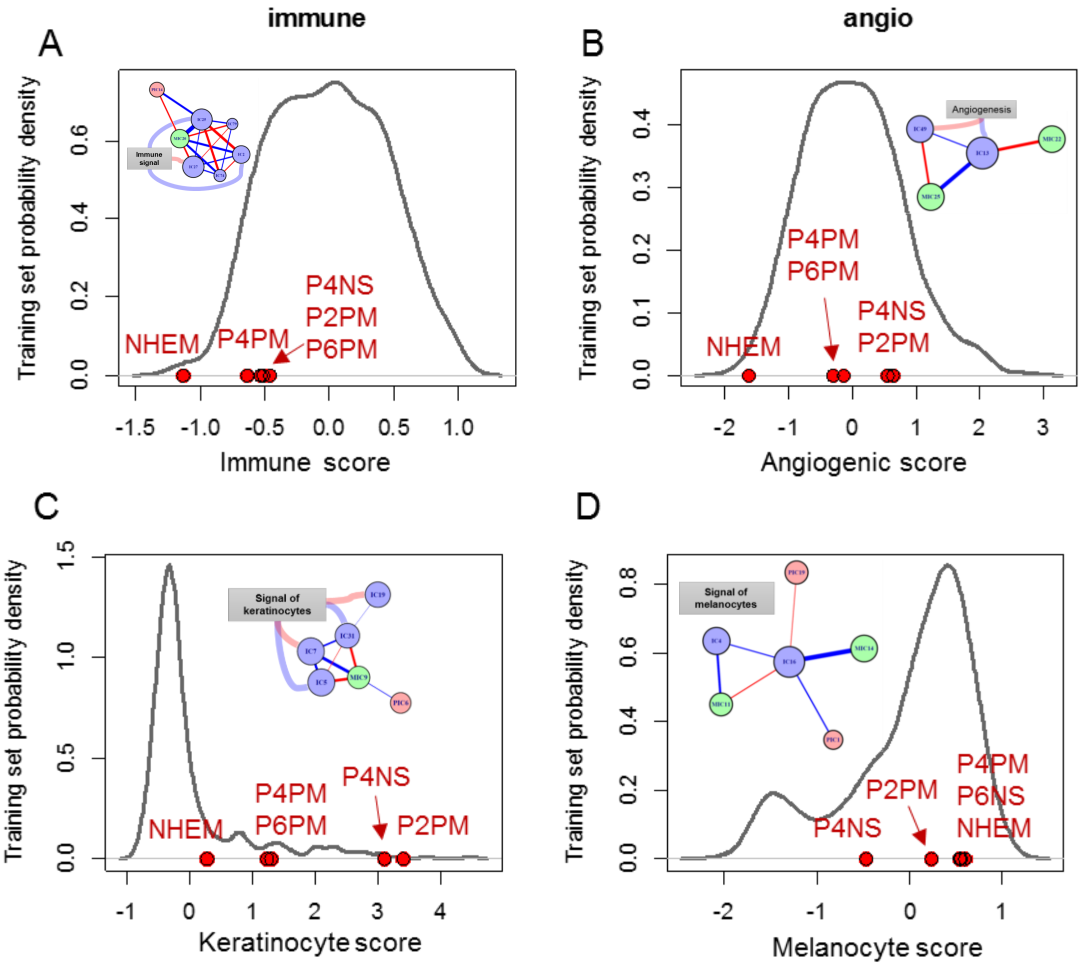
Results

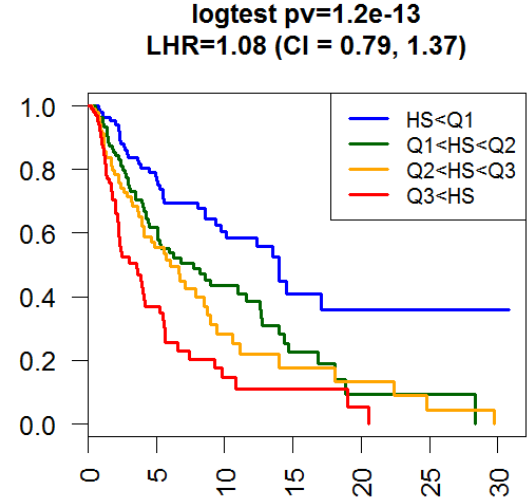
ICA helps establishing scores for new samples

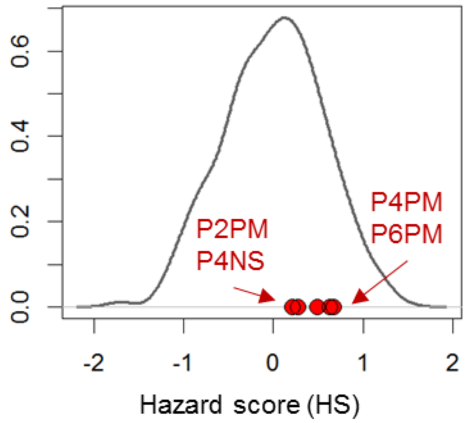
$$score_j = \sum_{i=1}^{k} d_i \, R_i^2 M_{i,j}^*$$

d_i – direction of the component (pos/neg)
H_i – log-hazard of Cox regression
R^2_i – stability of the *i*-th component
$M^*_{i,j}$ – weight of *i*-th component in sample *j*

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







Conclusions II

- 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)

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Proteome and Genome Research Unit, Luxembourg Institute of Health (LIH)

Arnaud MULLER
Tony KAOMA
Nathalie NICOT
Dr. Victoria EL KHOURY
Christina MAXIMO
Dr. Laurent VALLAR
Dr. Francisco AZUAJE
Dr. Gunnar DITTMAR

LSRU, University of Luxembourg

Dr. Anke WIENECKE
Dr. Stephanie KREIS

NORLUX Neuro-Oncology, LIH

Dr. Anna GOLEBIEWSKA
Dr. Sabrina FRITAH
Dr. Eric Van DYCK
Prof. Simone NICLOU

Institute Curie, France

Dr. Andrei ZINOVYEV

Unité INSERM, University of Reims

Prof. Ph. BIREMBAUT

TGen, USA

Prof. Nhan TRAN

Fonds National de la Recherche Luxembourg