

An Interdisciplinary Summer School on Mining of Biological Data for MSc and PhD students

Invited Lecture: **Methods in Single Cell Transcriptomics**

Petr Nazarov

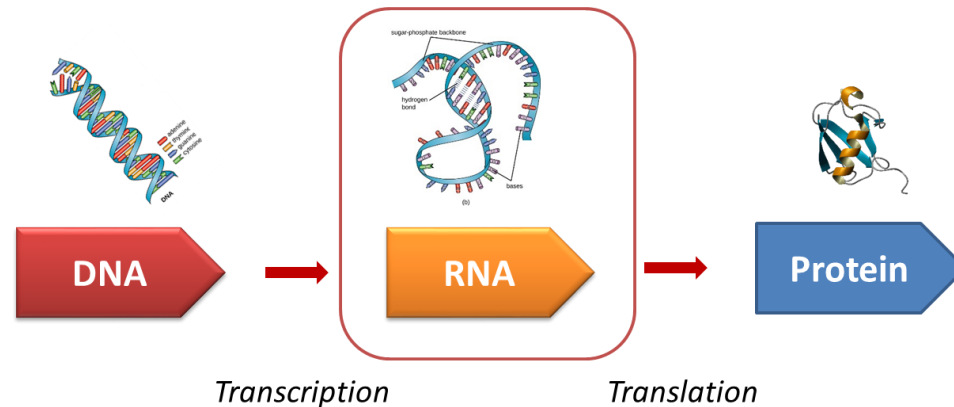
petr.nazarov@lih.lu

2018-08-14, Festsalen

<http://edu.sablab.net/nmbu2018/>

Outline

- The problem of heterogeneity
- Method 1: ICA
- Single-cell (SC) transcriptomics
- SC data properties
- Method 2: t-SNE (t-distributed stochastic neighbor embedding)
- Some examples

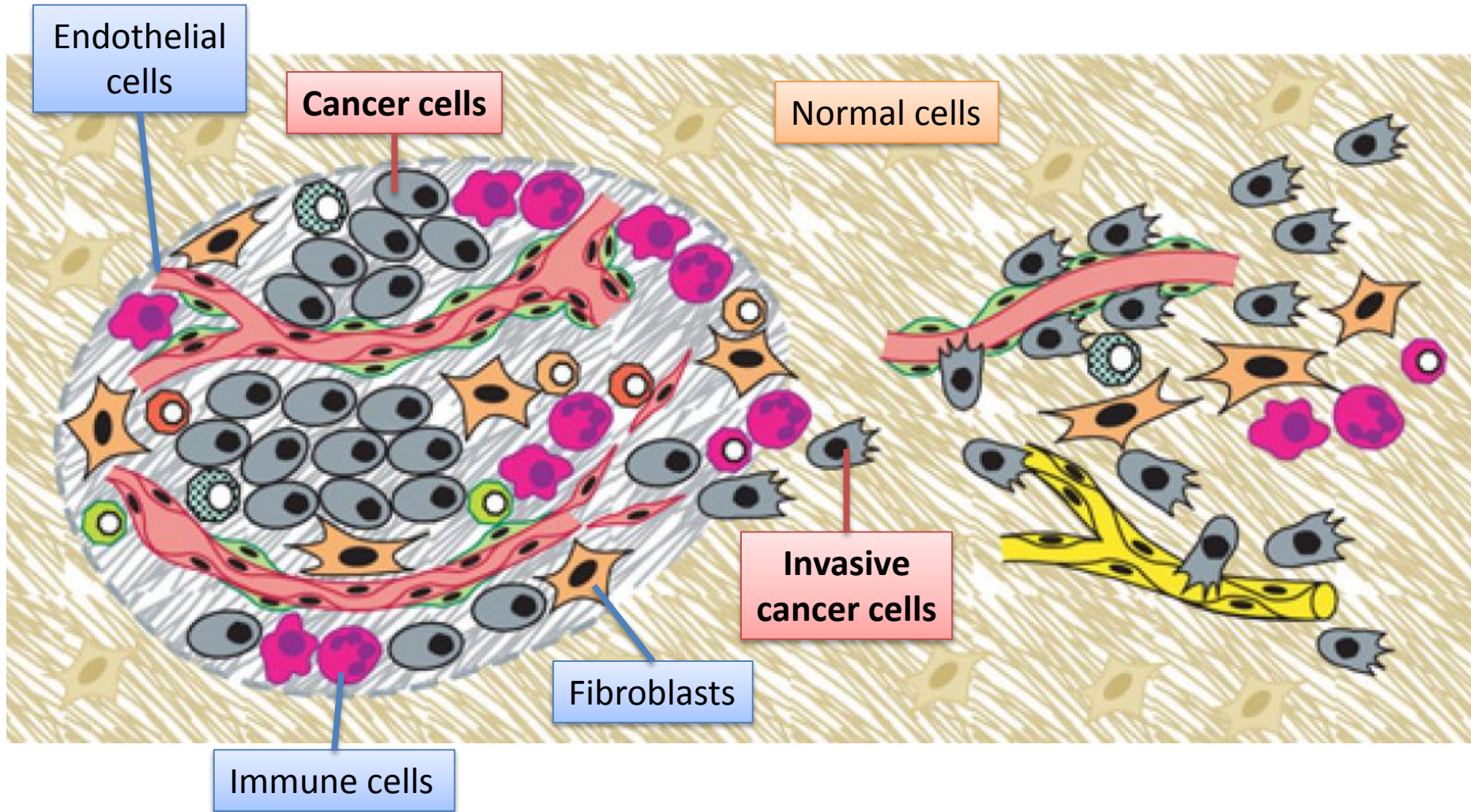


<http://edu.sablab.net/nmbu2018/>

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](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.bmrw.wisc.edu/featuredSys/ubiquitin/ubiquitin1.shtml>

Introduction

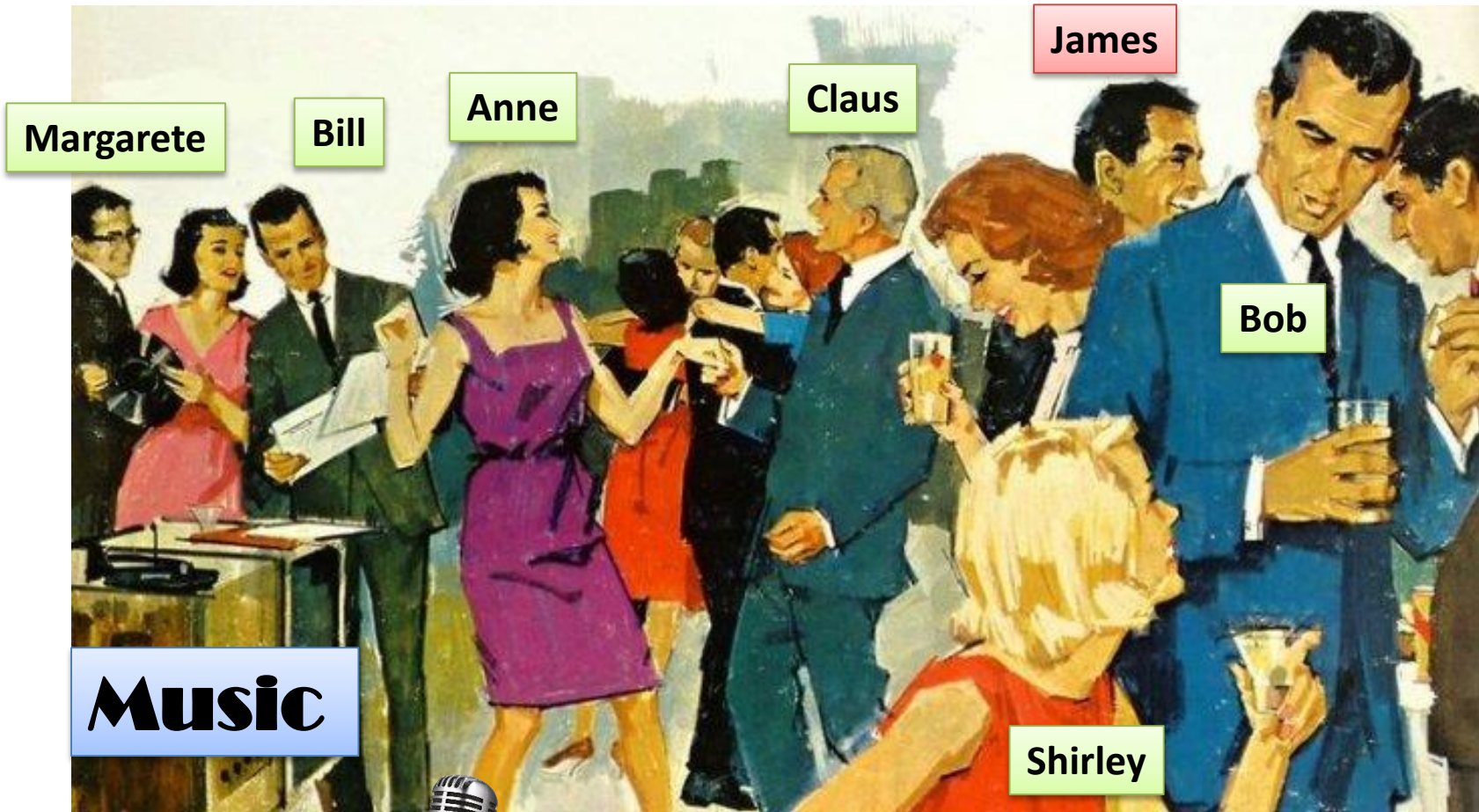
Imagine we are going to analyze RNA from a tumor biopsy (sample):



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

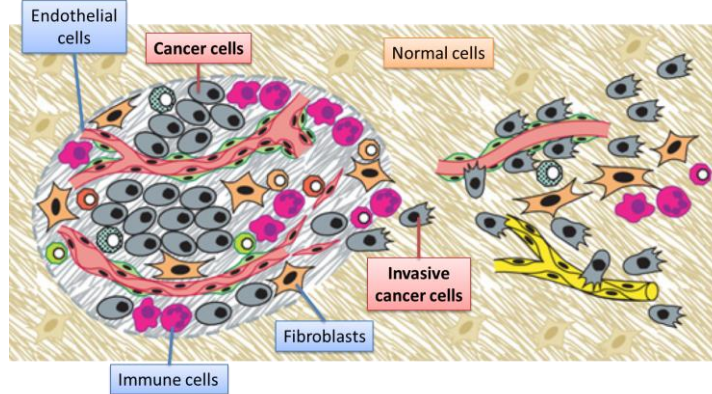
Introduction

This is like recording a cocktail party:



What did James say?..

Introduction



Physical separation of cells

Computational separation of signals

- Laser microdissection (time consuming)
- Cell dissociation and single cell analysis

- Non-negative matrix factorization (NMF)
- Independent component analysis (ICA)

Let's consider first **ICA method** and then move to **single cell transcriptomics** (as we shall use the method there) 😊

Independent Component Analysis

One of the methods to solve cocktail party problem...

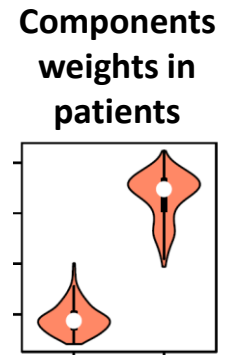
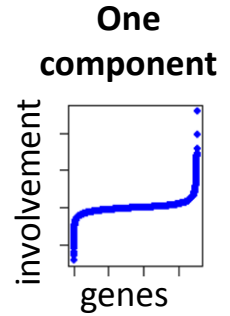
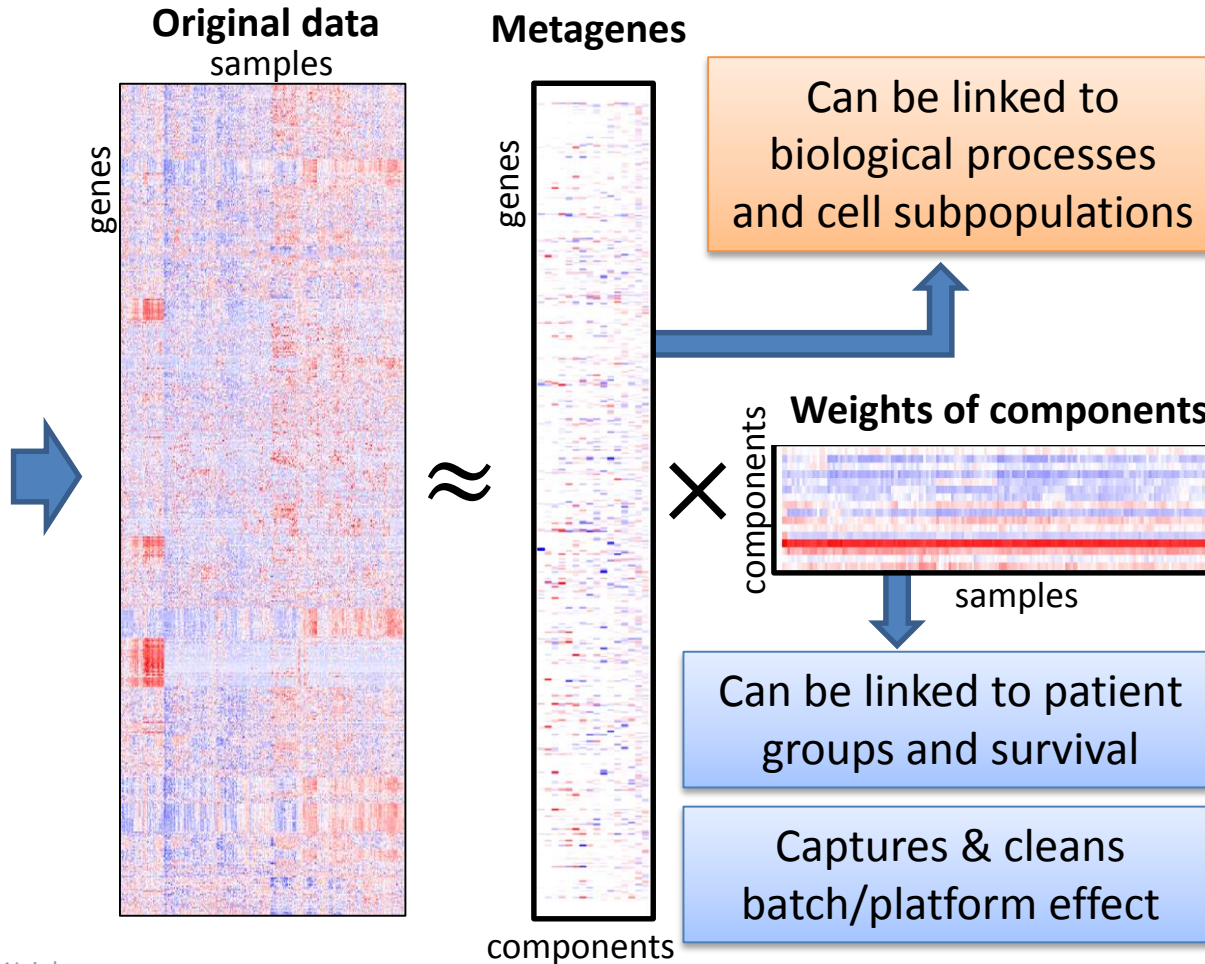


**Independent
Component
Analysis**



Independent Component Analysis

Deconvolution of Cell Ensemble



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

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

What ICA does and does not

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

g – genes

s – samples

k - components

Pro:

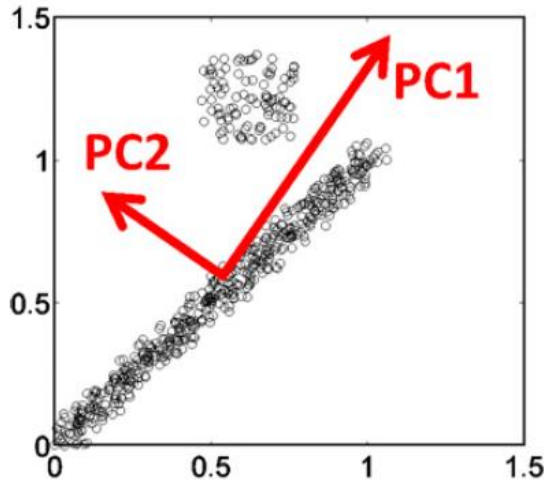
1. Finds **statistically-independent signals** (components) in the expression profiles
2. Identifies the **most important genes** in each component
3. Tells what is the weight of **each component in the samples**
4. Works on data *per se*, **without any additional knowledge**
5. Gives quite **robust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

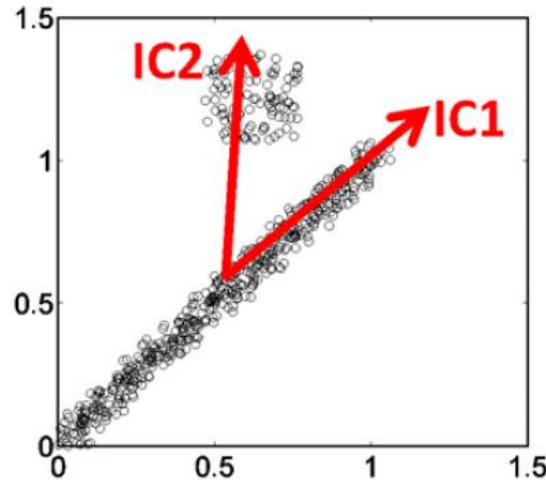
Geometrical view ☺

PCA



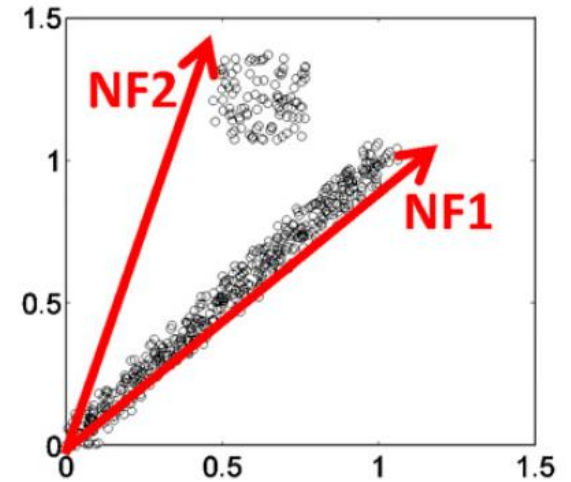
Orthogonal
Captures major variation
(well, on average...)

ICA



Linear combination of
independent sources.
Positive and negative.

NMF

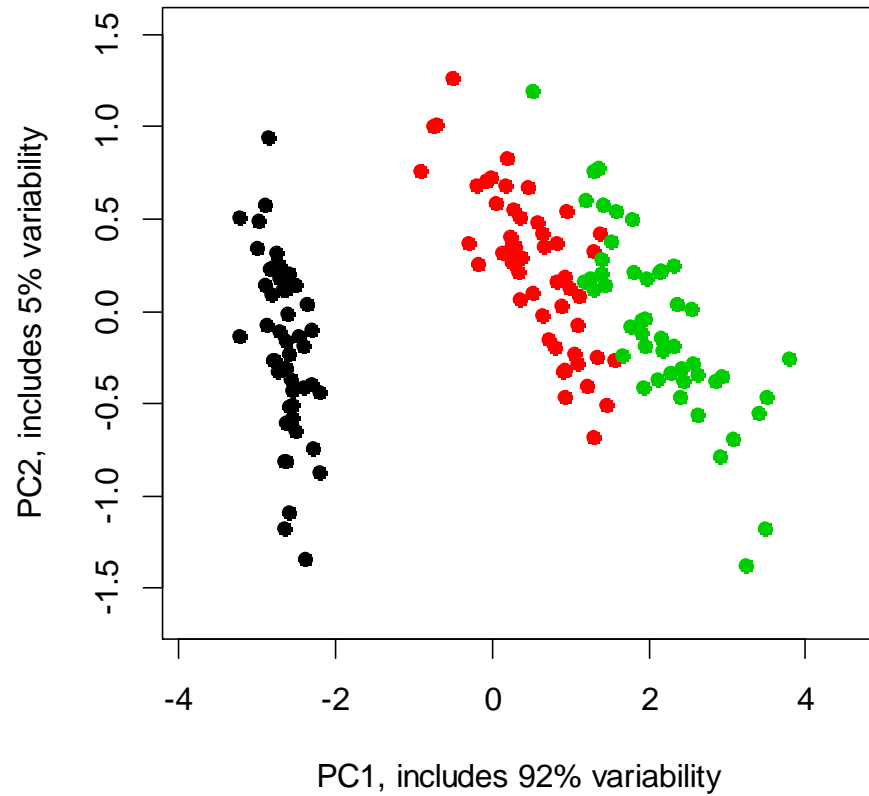


Each point can be
represented as a vector sum
of NF1, NF2. Strictly positive.

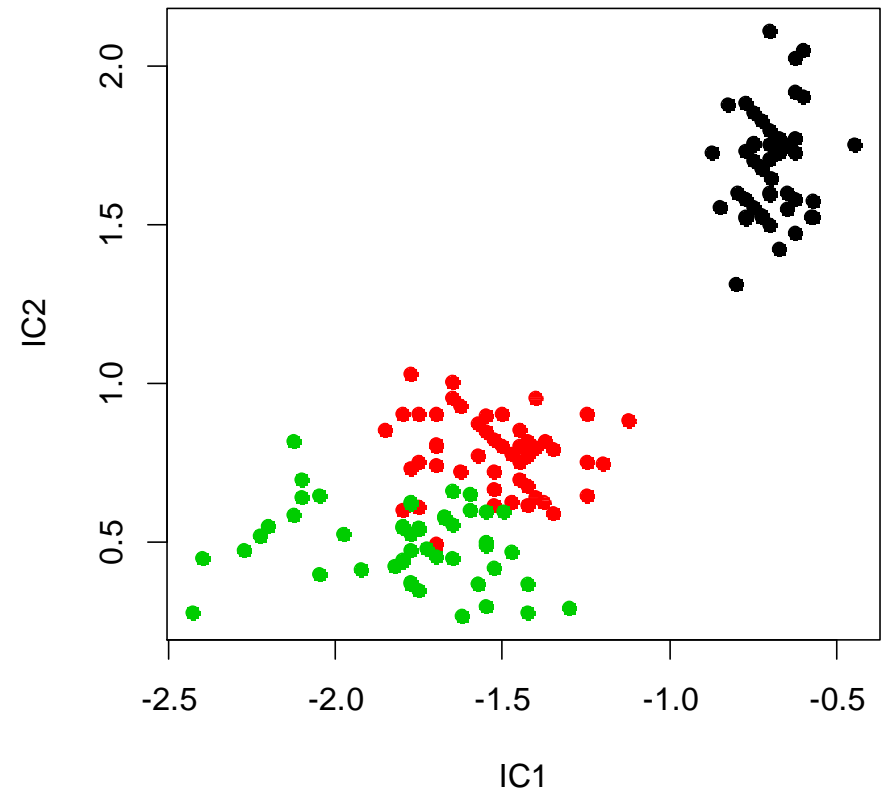
from A. Zinovyev, et al, Biochem Biophys Res Commun. 2013,18;430(3):1182-7
<https://www.ncbi.nlm.nih.gov/pubmed/23261450>

Data visualization: PCA & ICA

PCA (98% variability)



ICA



Independent Component Analysis

SEQC Data

A, B – two reference
human RNA samples
 $C = 0.75 \cdot A + 0.25 \cdot B$
 $D = 0.25 \cdot A + 0.75 \cdot B$

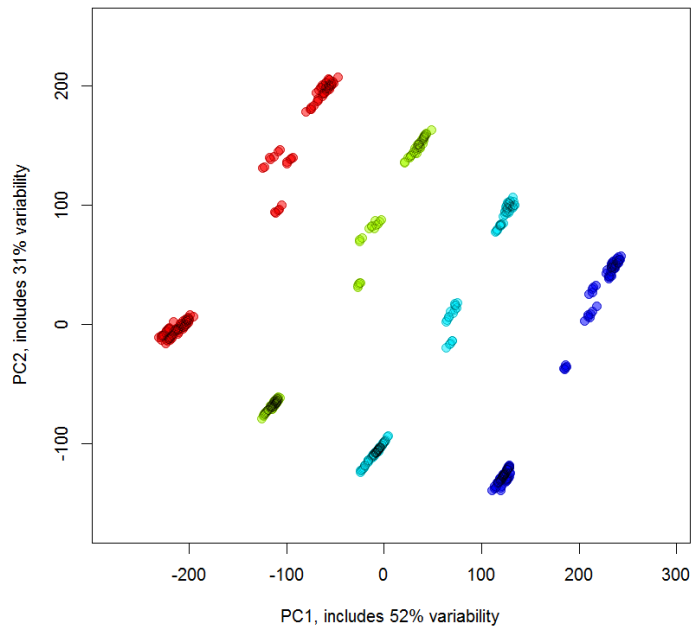


4 samples: A,B,C,D

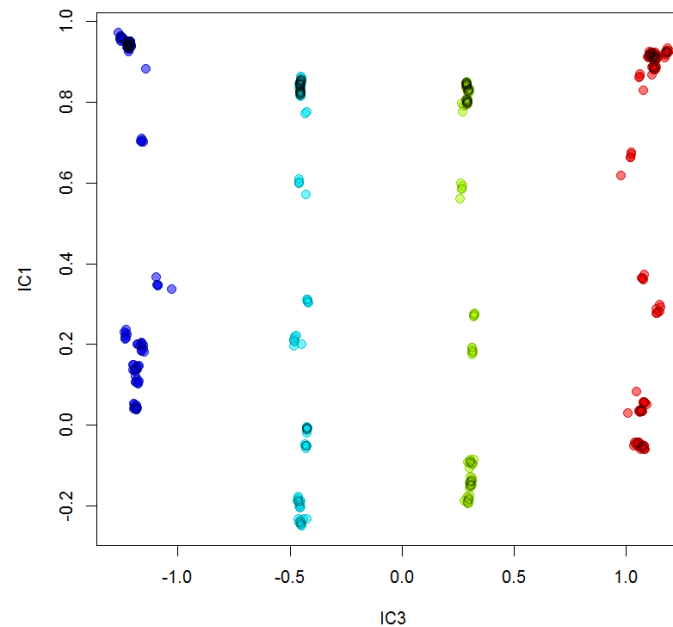


Studied by 13 labs
using 3 sequencers

Principle component analysis (PCA) (83% variability)



Independent Component Analysis (ICA)



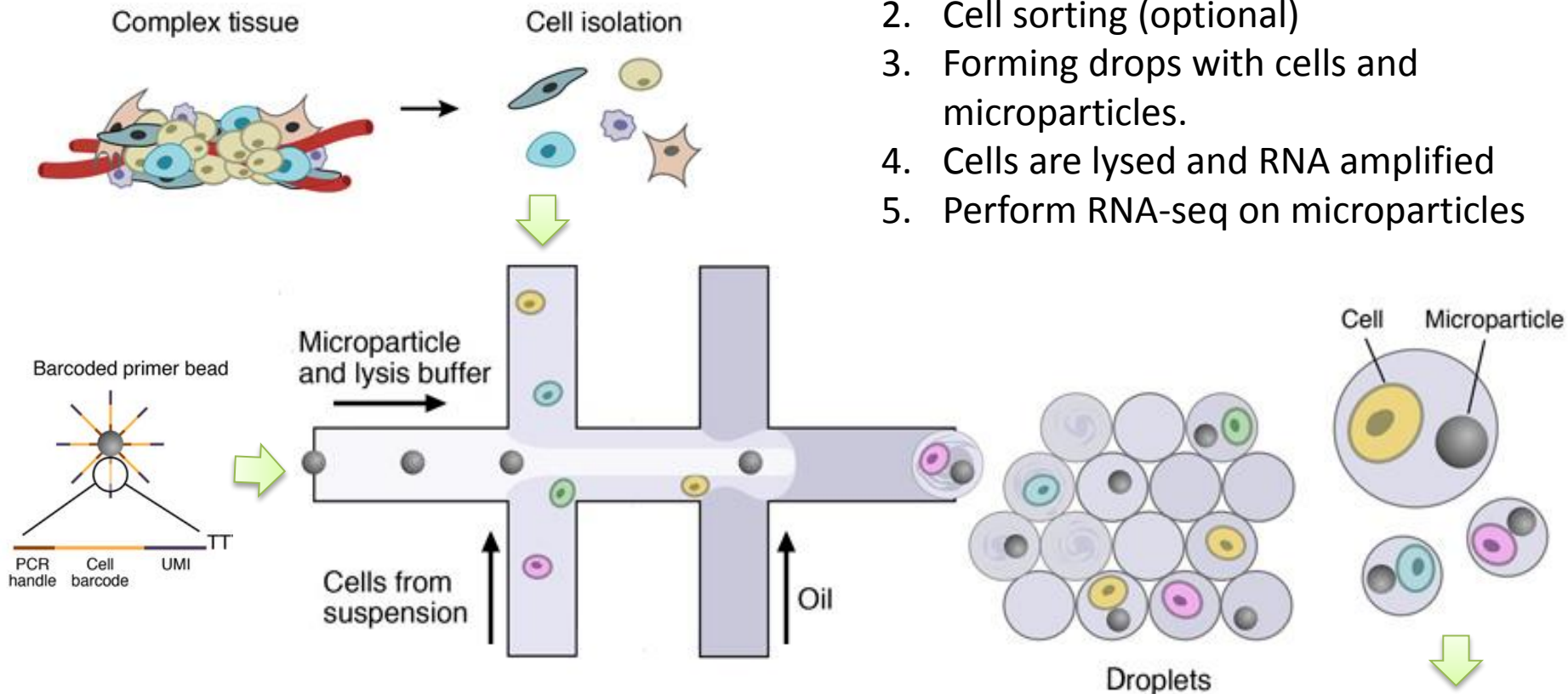
The effect of sample
mixing is captured
by **two PCs** and
single IC₃ !

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

Single Cell Transcriptomics

Single Cell Transcriptomics – one of the method to handle the tissue heterogeneity problem.

1. Cell dissociation
2. Cell sorting (optional)
3. Forming drops with cells and microparticles.
4. Cells are lysed and RNA amplified
5. Perform RNA-seq on microparticles



Each microparticle contains more than 10^8 individual primers that share the same “PCR handle” and “cell barcode”, but have different unique molecular identifiers (UMIs).

<https://www.elveflow.com/microfluidic-tutorials/microfluidic-reviews-and-tutorials/drop-seq/>

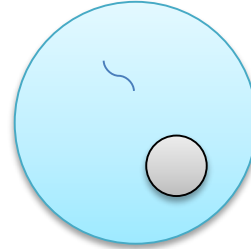


Single Cell Data Properties

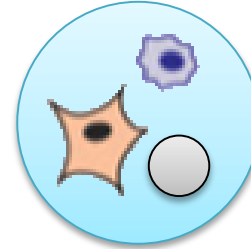
Ideal: one bead - one cell



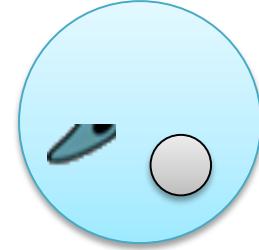
What you have in practice:



no cell,
floating RNA

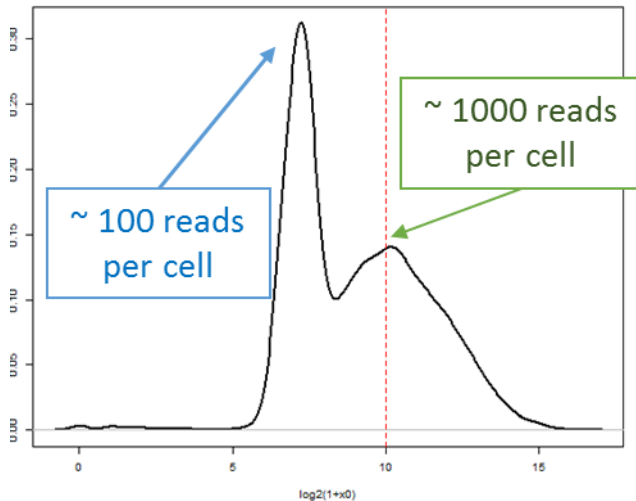


two cells



some cellular
debris: often
mitochondria

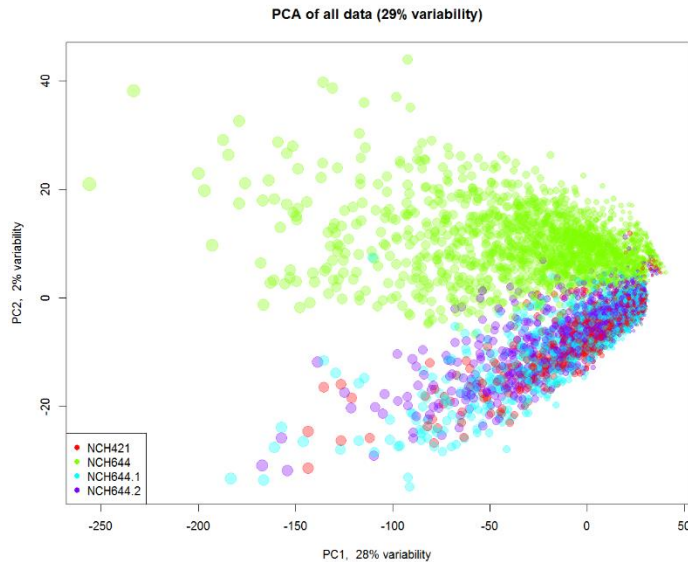
Number of “reads” (detected RNA fragments) per cell



Therefore:

1. Single-cell RNA-seq data are sparse (many zeros) and large (expect to have 10^2 - 10^4 cells x 10^3 - 10^4 genes).
2. Filtering is unavoidable and often remove majority of “cells”.
3. Standard normalization methods are questionable.

PCA of SC RNA-seq data



- PCA captures variability => distant data points have larger effect
- PC1 always captures number of reads per cell – this is the largest effect (even after normalization)
- Biologists do not like it as the density of points is not constant 😊

We need a method that is going to:

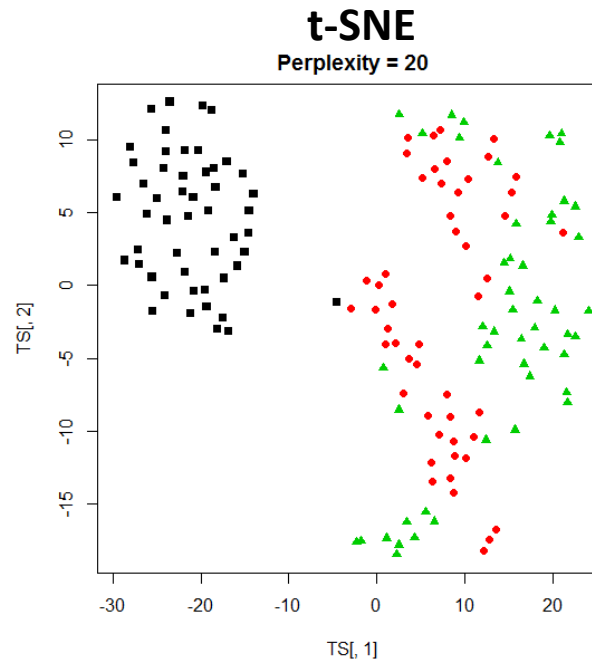
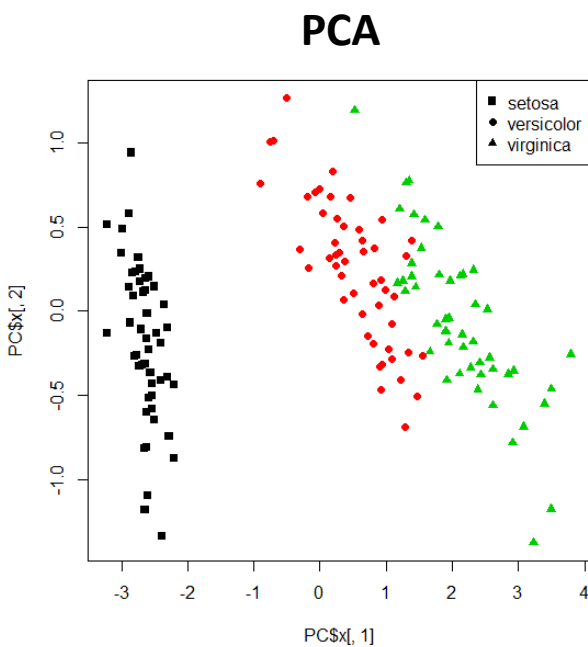
- puts the similar objects together
- produces the picture with constant density
- is easy to understand 😊

Visualization of large datasets

t-SNE is an iterative non-linear transformation that search for objects representation in 2D space by:

- 1) placing the similar objects together
- 2) controlling the density of the obtained clusters

Unlike PCA, distant objects are not influencing t-SNE!



Pro:

- easy to understand
- no effect of outliers

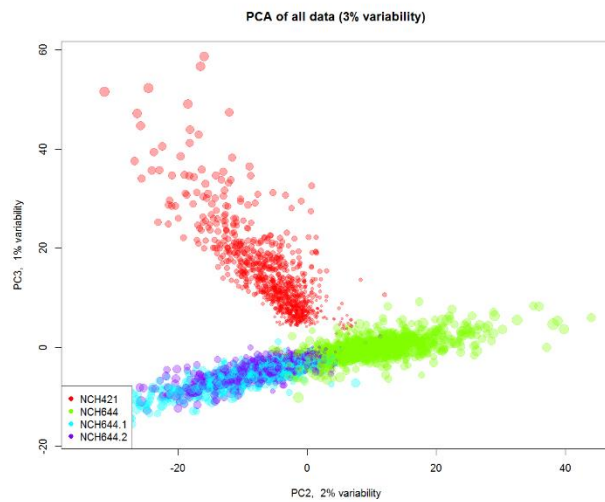
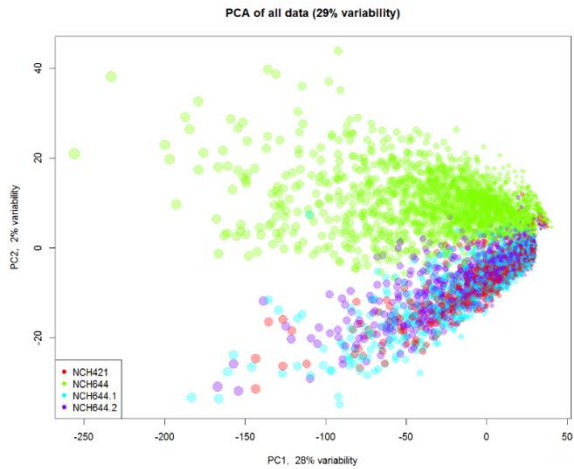
Con:

- depends on init.estim.
- can be over-interpreted !
- depends on *perplexity* parameter

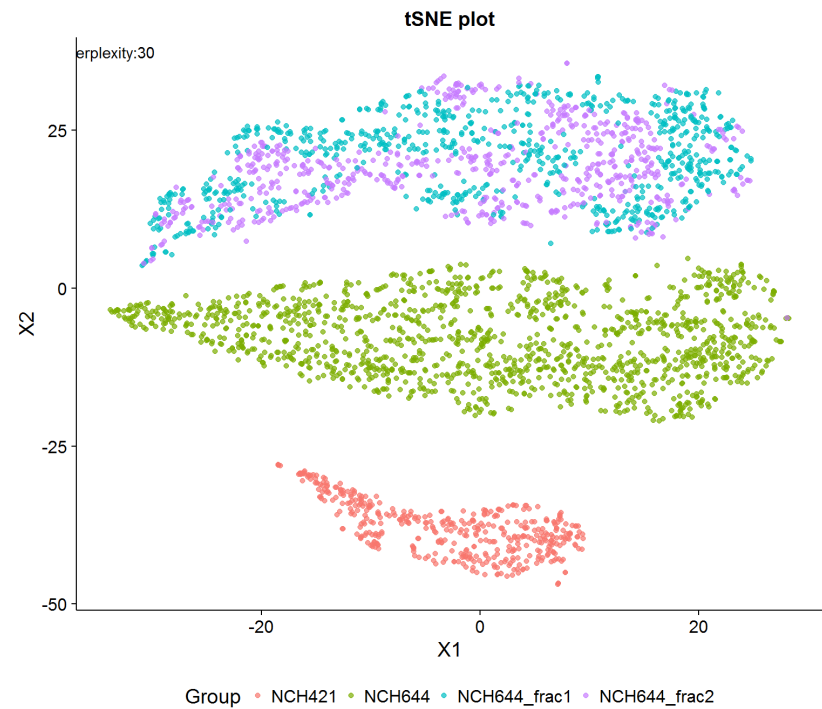
Play with t-SNE here: <https://distill.pub/2016/misread-tsne/>

t-SNE for single cell transcriptomics

PCA plots



t-SNE plot



t-SNE for single cell transcriptomics

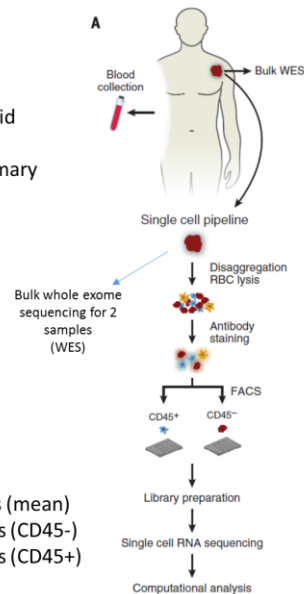
RESEARCH ARTICLES

CANCER GENOMICS

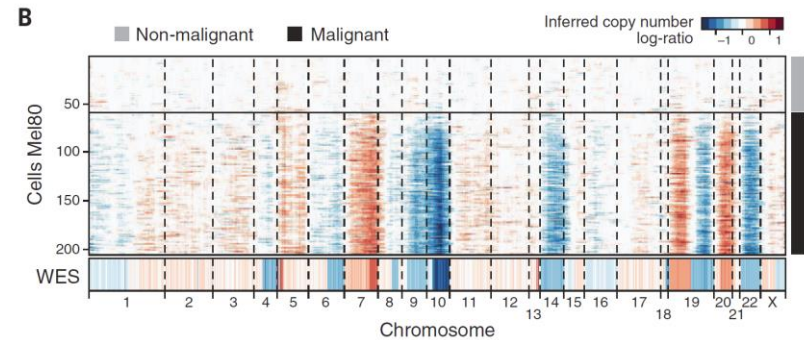
Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

Itay Tirosh,^{1*} Benjamin Izar,^{1,2,3*} Sanjay M. Prakadan,^{1,4,5,6}
Mara H. Wadsworth II,^{1,4,5,6} Daniel Treacy,¹ John J. Trombetta,¹ Asaf Rotem,^{1,2,3}

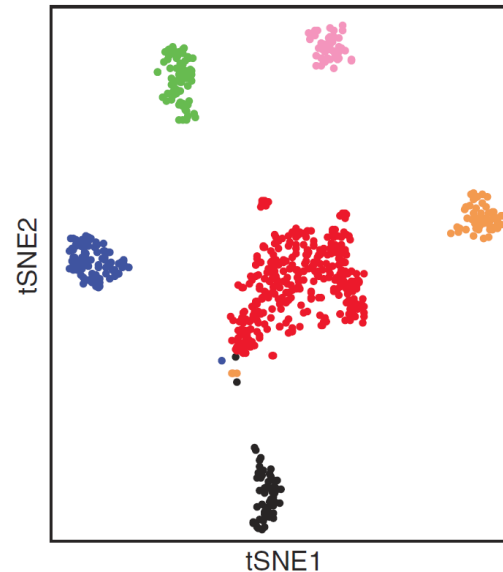
- 19 tumors:
 • 10 lymphoid
 • 8 distant
 • 1 acral primary



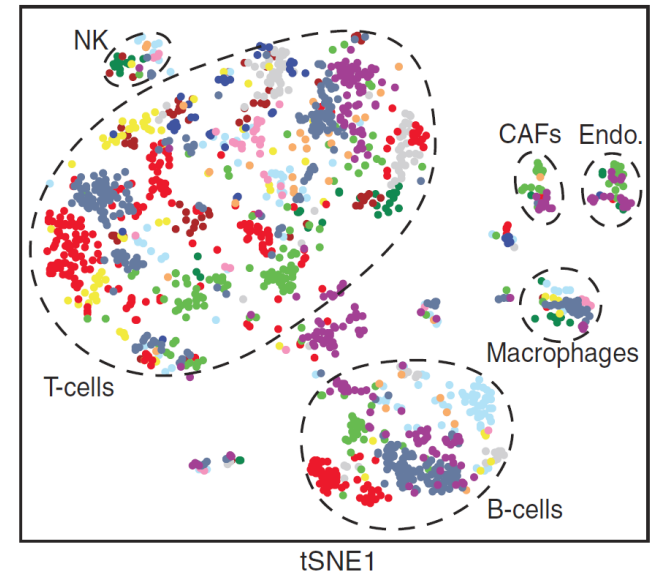
- 4645 cells:
 • 150k reads (mean)
 • 4659 genes (CD45-)
 • 3438 genes (CD45+)



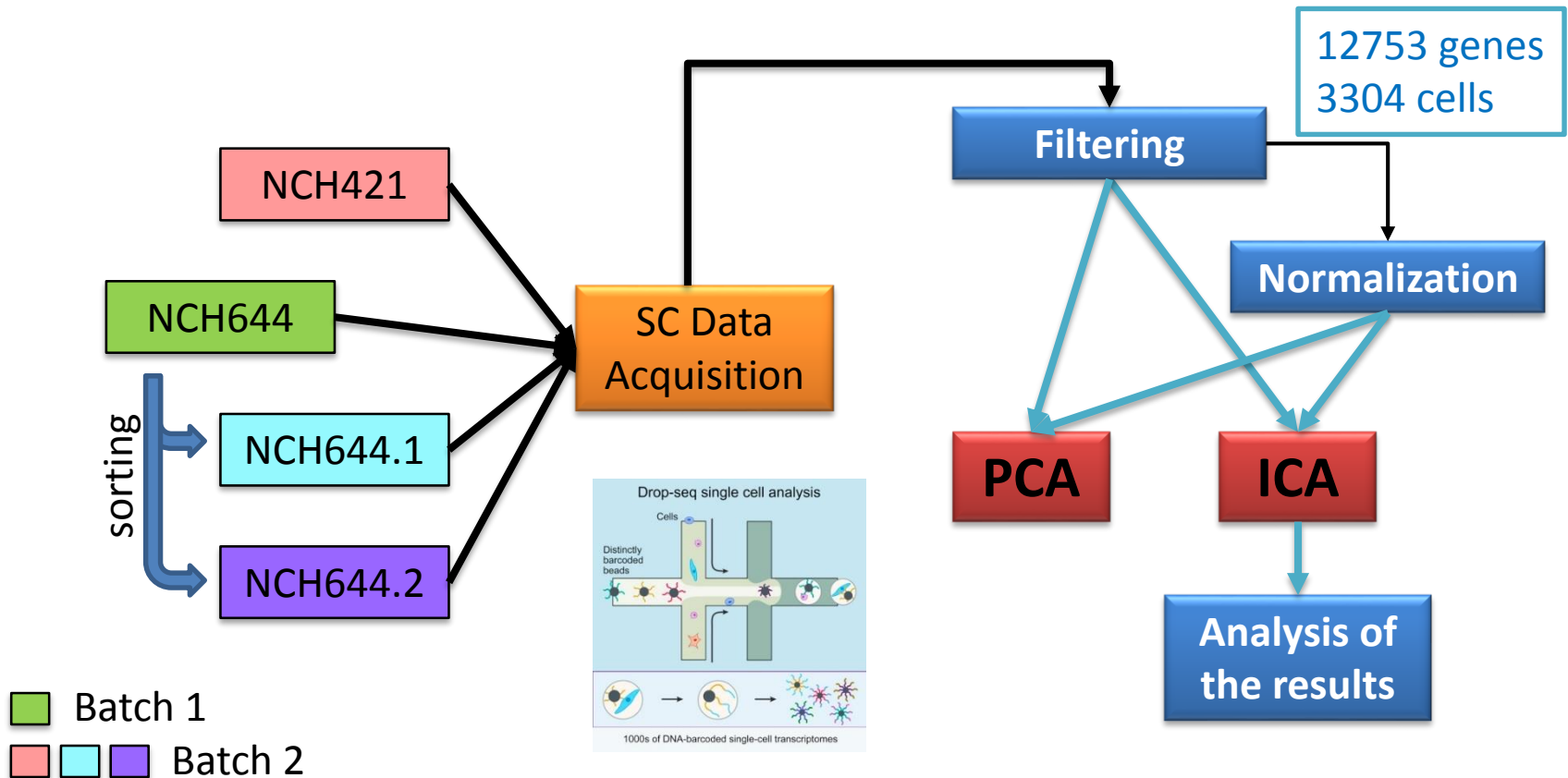
C Malignant cells



D Non-malignant cells



Example: Design



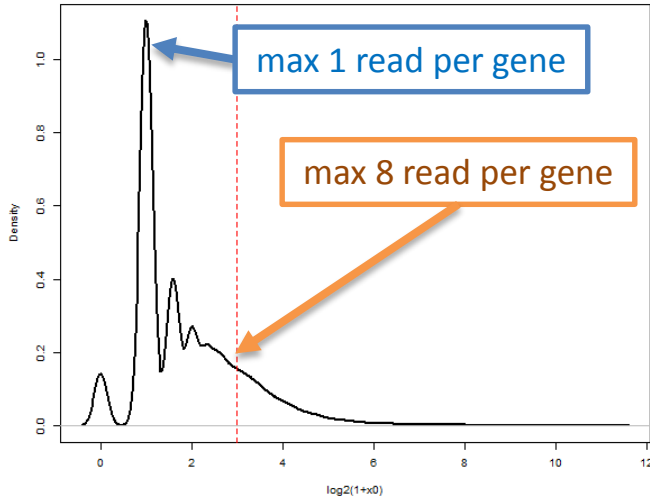
Application of ICA to SC data is a strange idea. But why not ;)

Expectations:

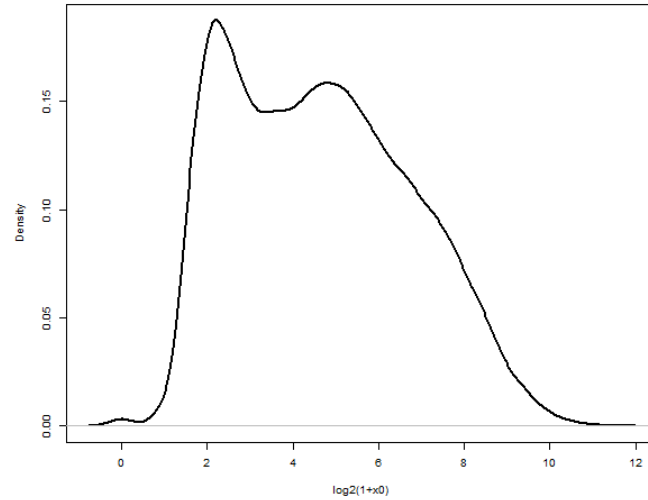
- See cellular process
- Get visualization within the coordinates, that can have biological meaning

Example: Data

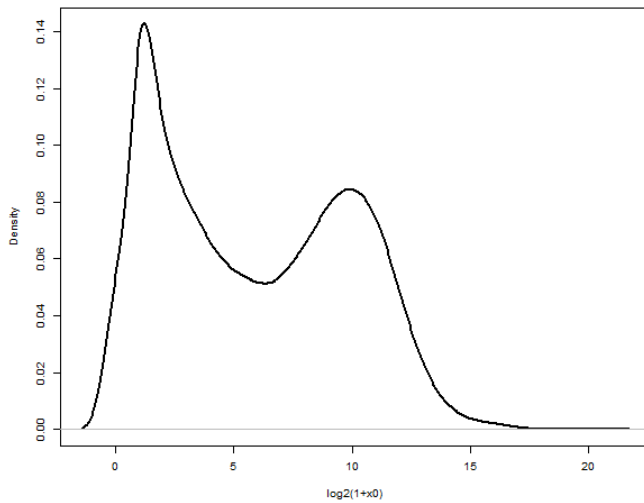
Max expression for genes



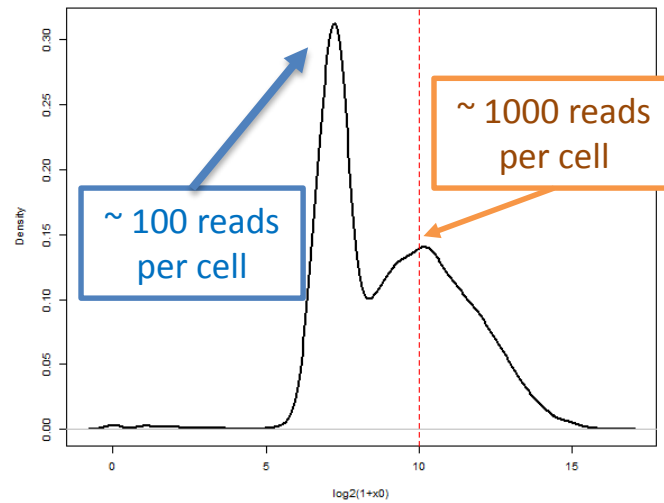
Max for cells



Sum for genes



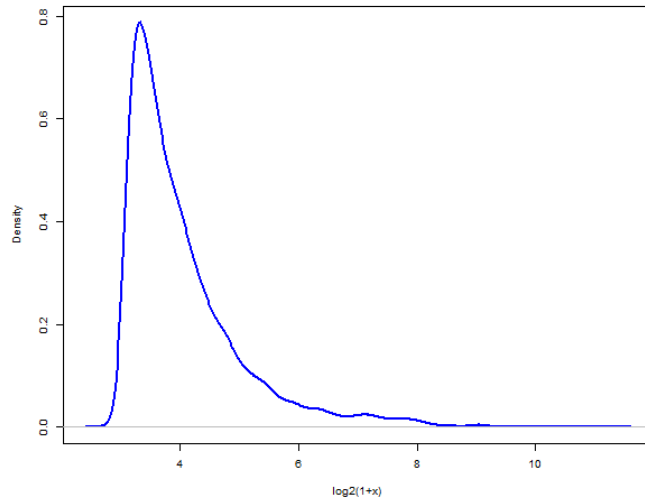
Number of reads for genes



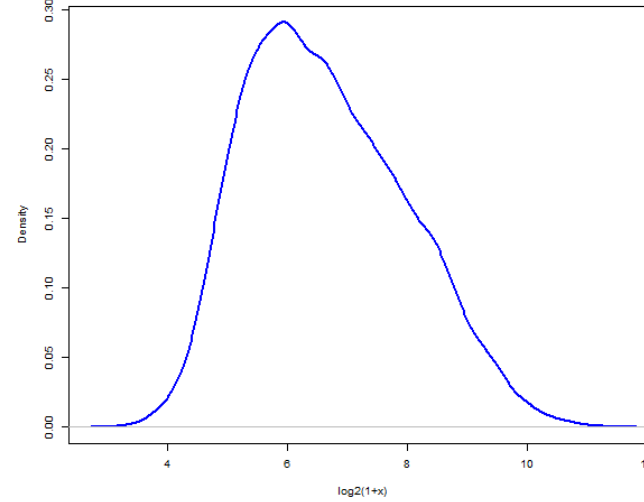
Example: Data

Filtered data

Max expression for genes

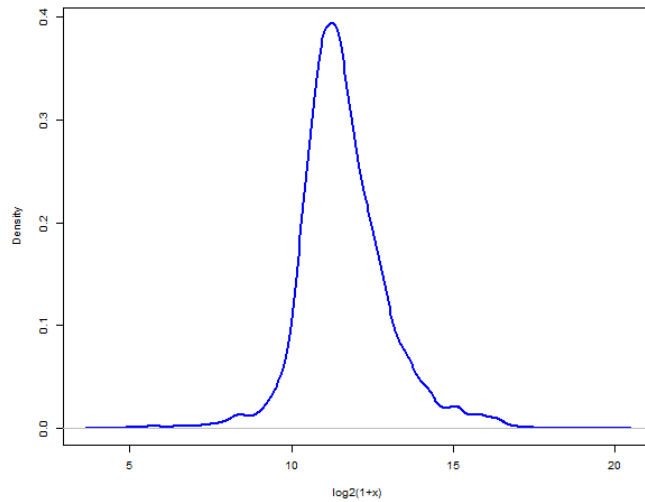


Max for cells (cleaned)

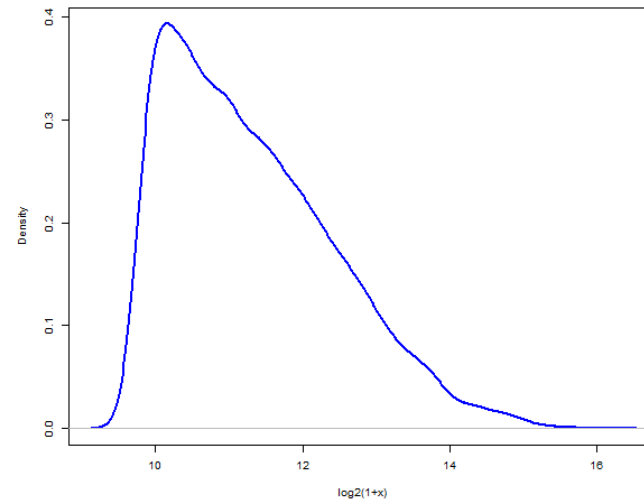


4087 genes
6290 cells

Sum for genes (cleaned)



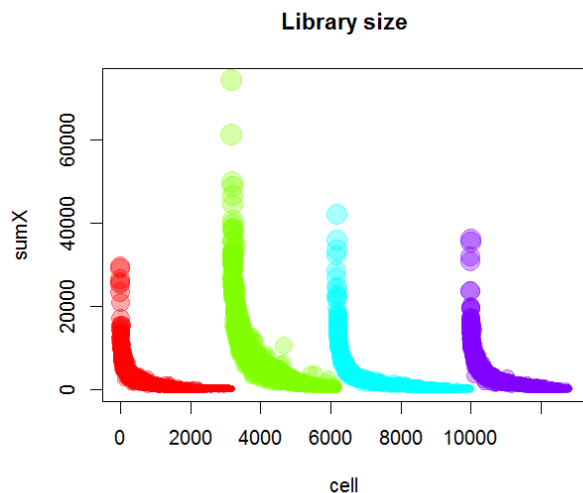
Number of reads for genes



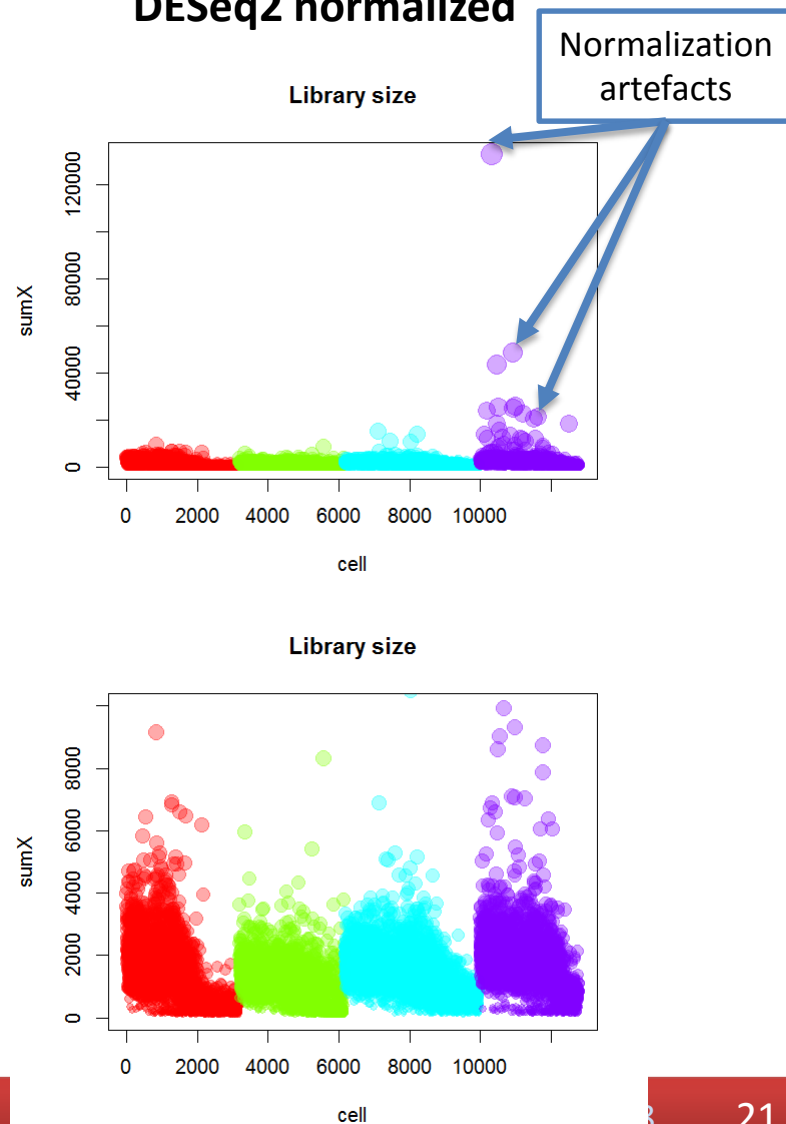
Example: Normalization Effect

Normalization Issues

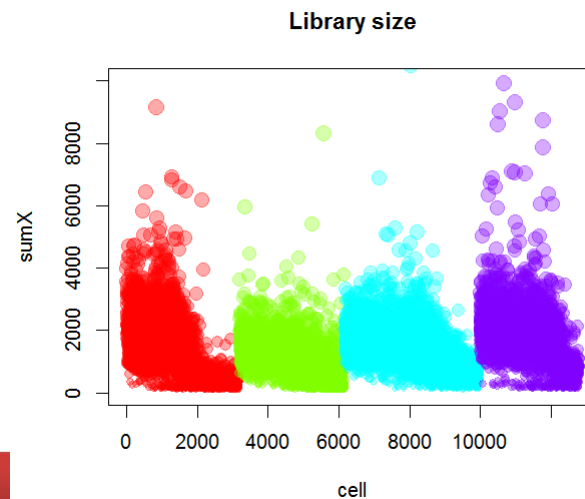
Not normalized



DESeq2 normalized



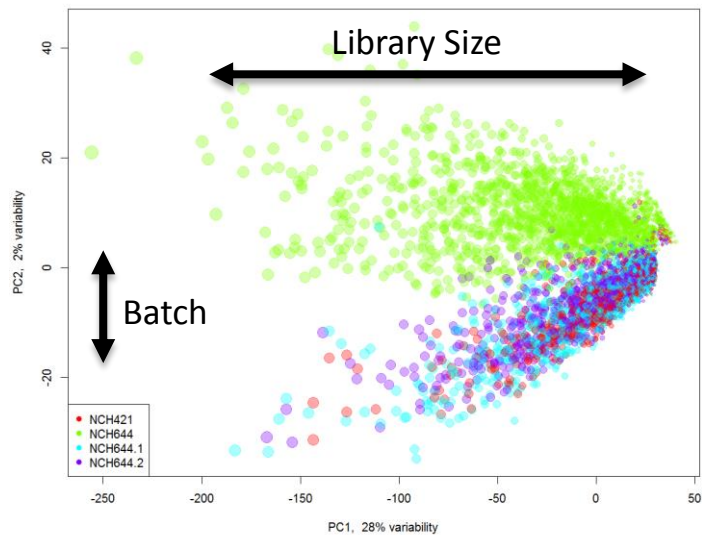
- **DESeq2** homogenizes number of reads per cell, but introduces technical artefacts for some cells



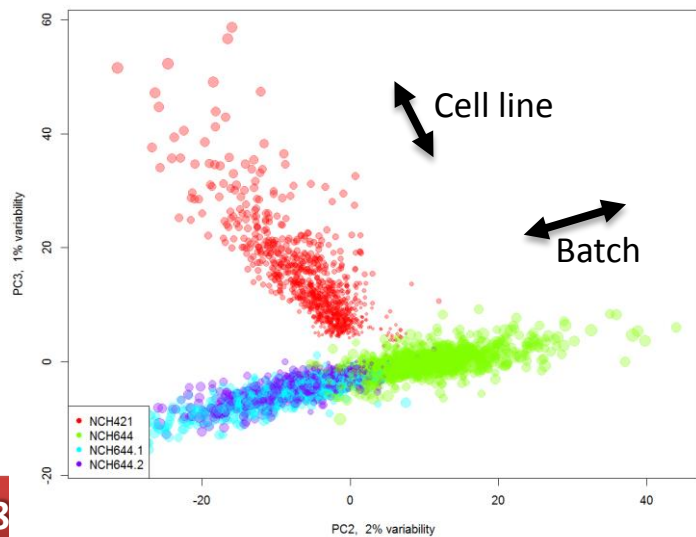
Example: Normalization Effect

Raw

PCA of all data (29% variability)



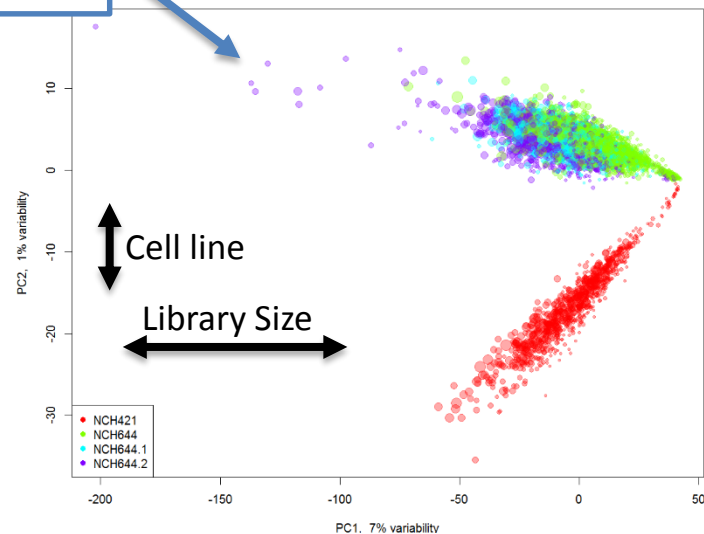
PCA of all data (3% variability)



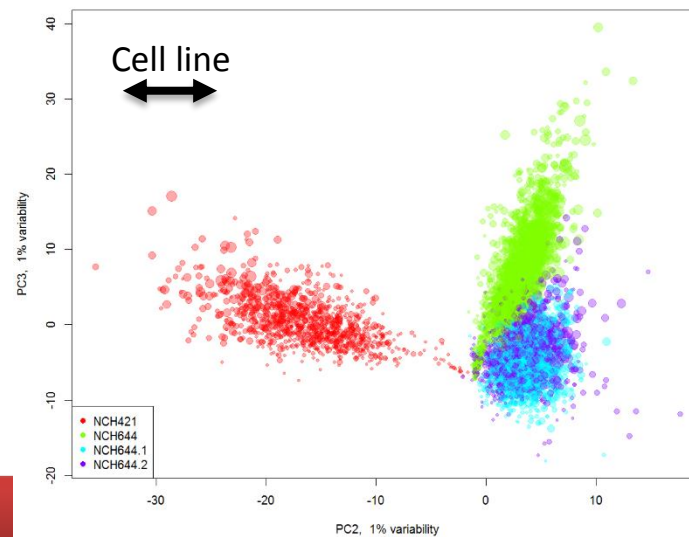
Normalized

Normalization artefacts ?

PCA of all data (8% variability)

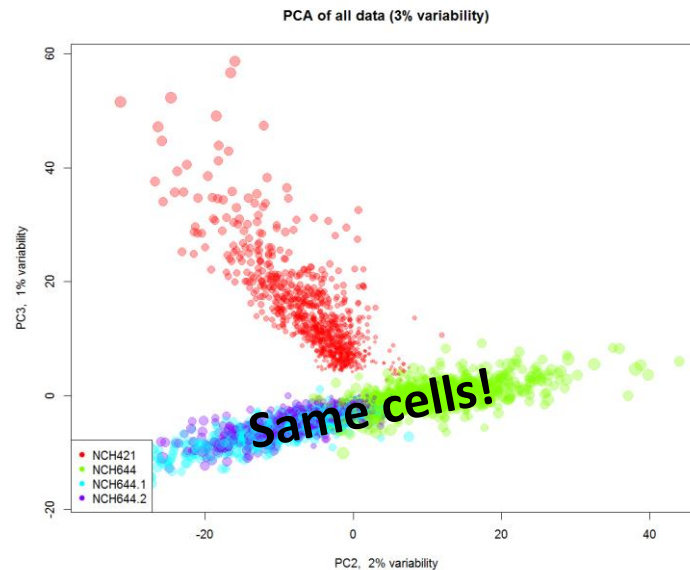


PCA of all data (3% variability)



Example: the Question to Answer

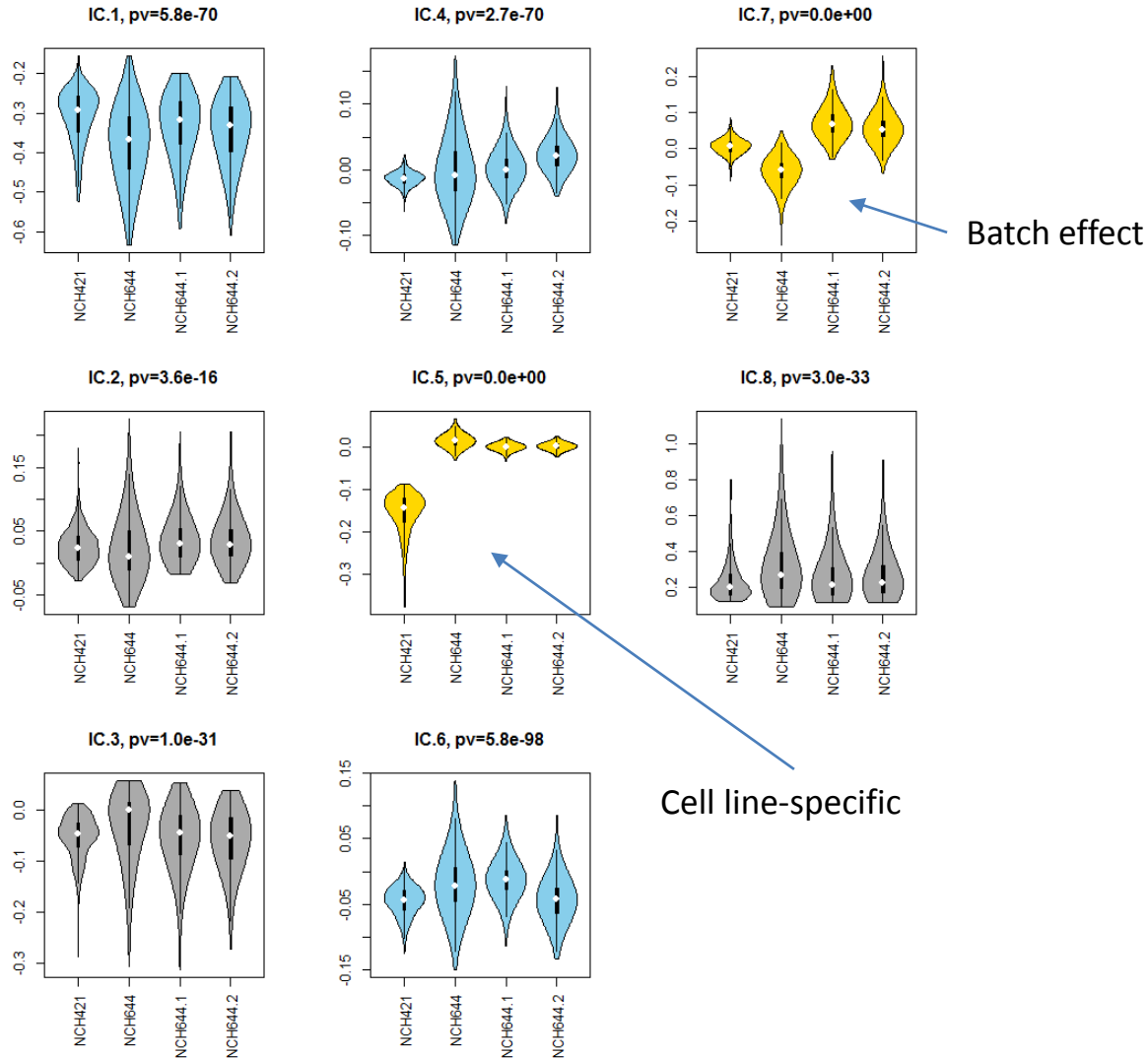
Do we stop the project and lose over 30k euro ?



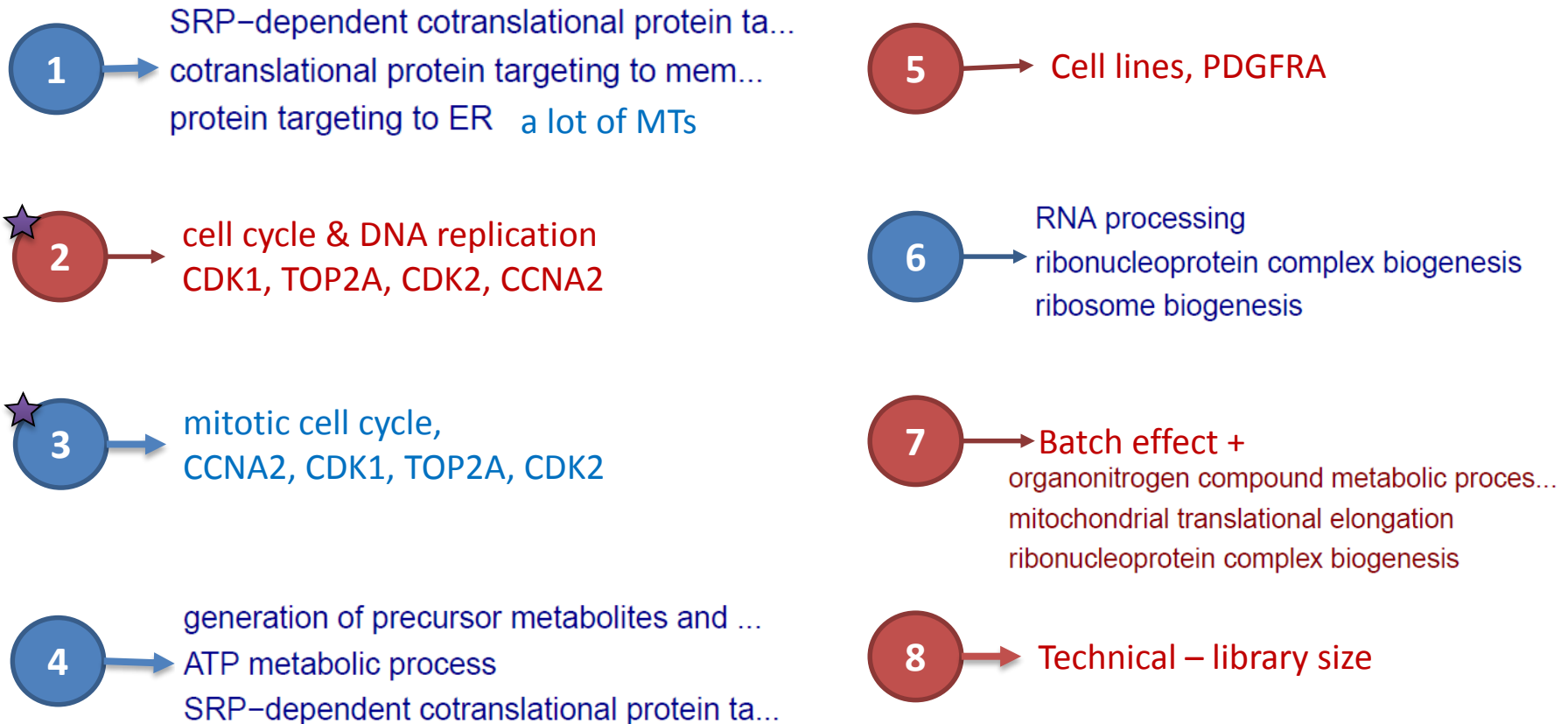
No. Let's have some fun with the data 😊

Example: ICA

ICA with 8 components: M (weights) over experiments



Analysis of contributing genes (S): biological processes

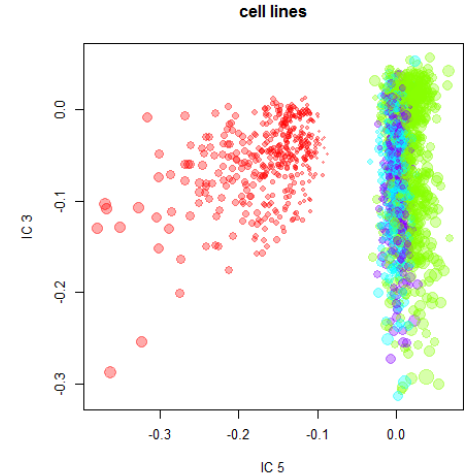
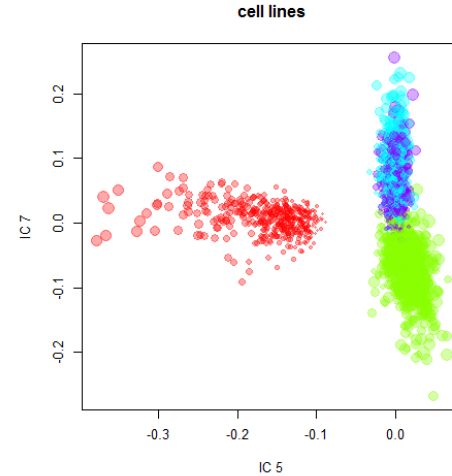
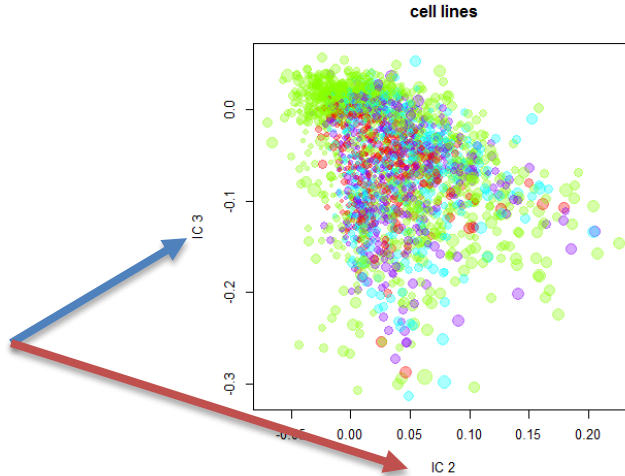


Example: ICA

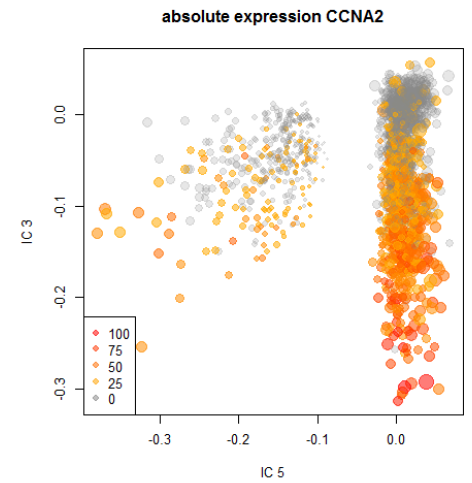
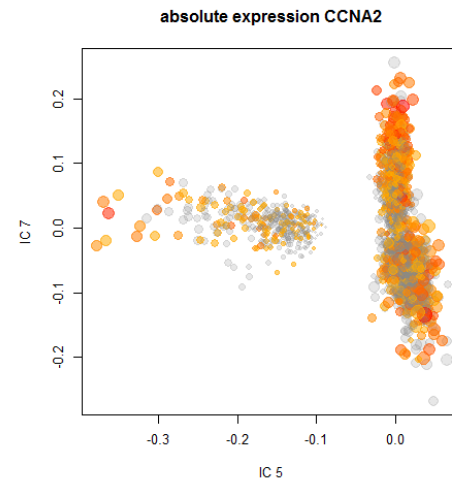
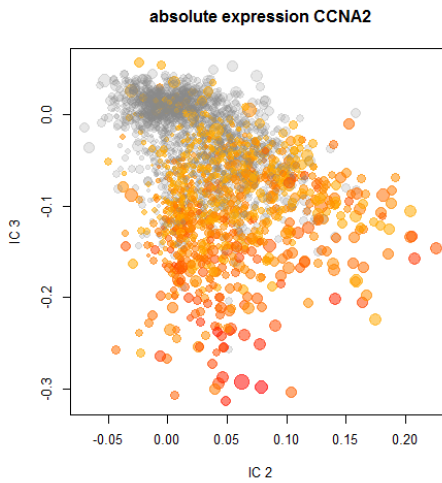
Gene expression: CCNA2

- NCH421
- NCH644
- NCH644.1
- NCH644.2

IC2, -IC3:
linked to cell
cycle



IC5:
predictor of cell
lines

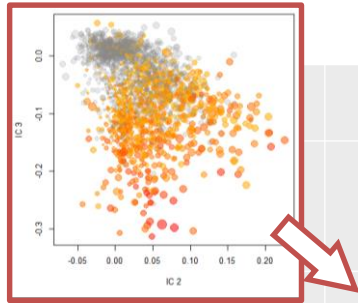


IC7:
Is linked to time
or batch

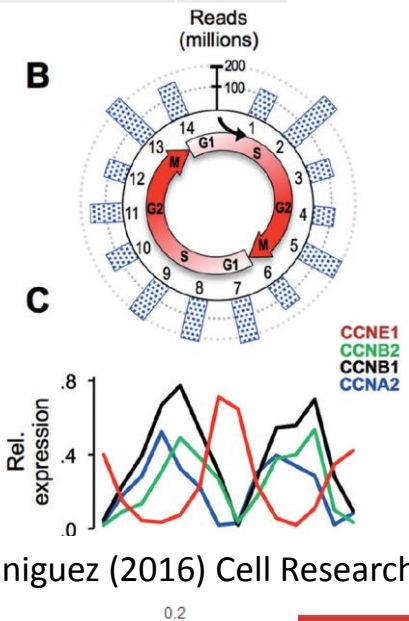
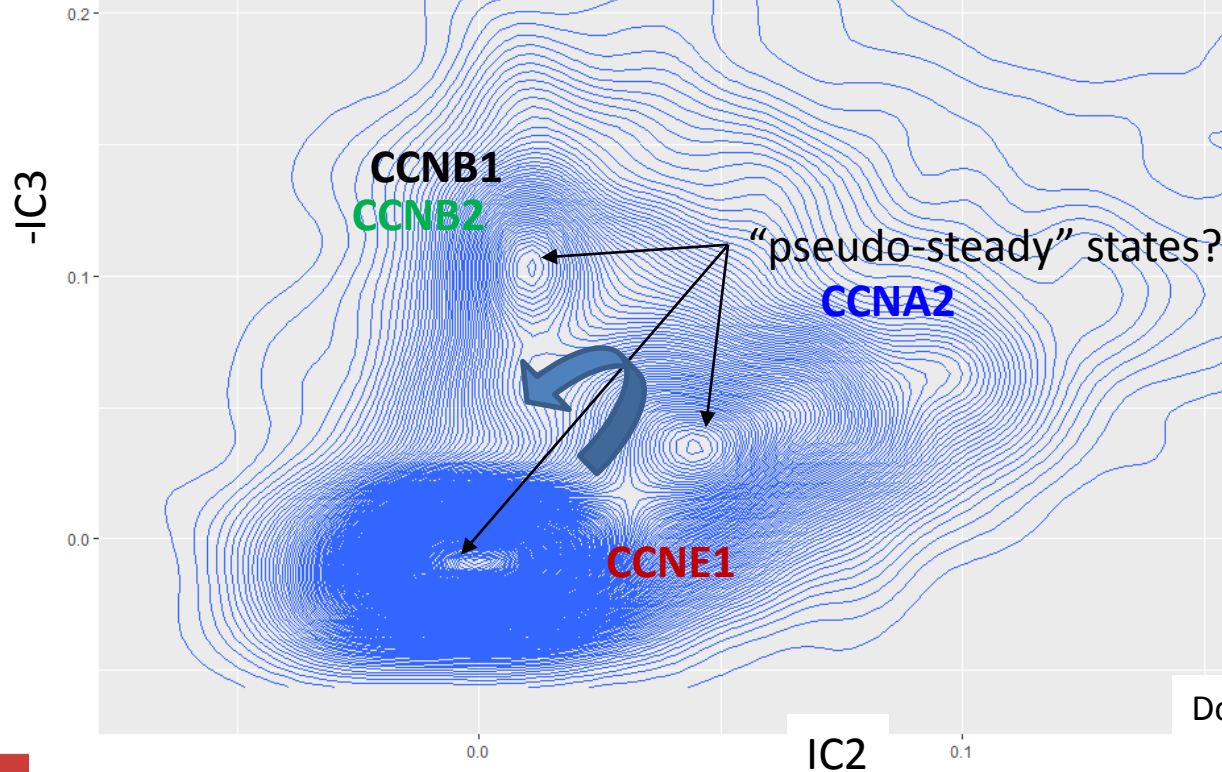
IC8:
strongly linked
to library size

Single Cell ADAPT: ICA

Same cell cycle in depth



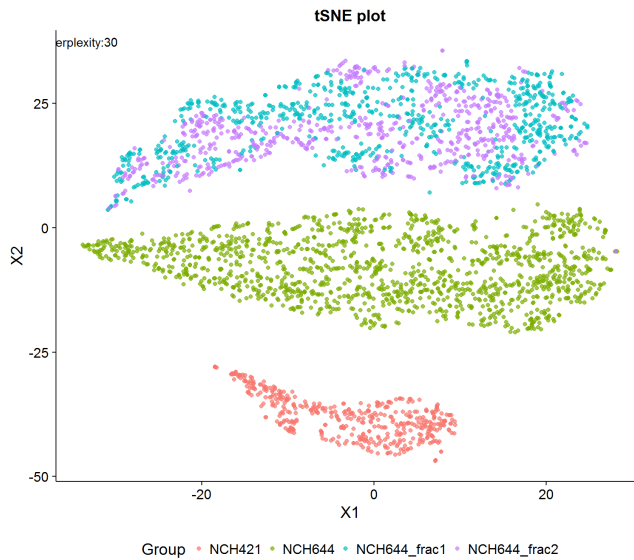
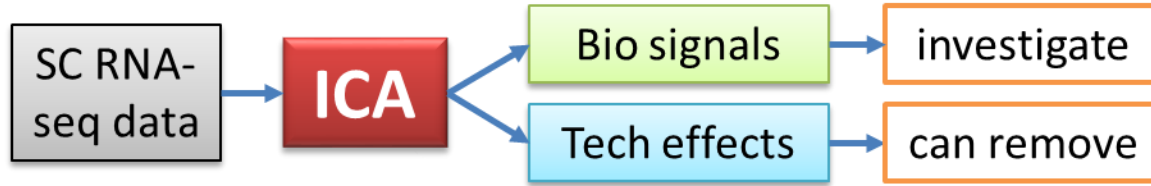
by Arnaud



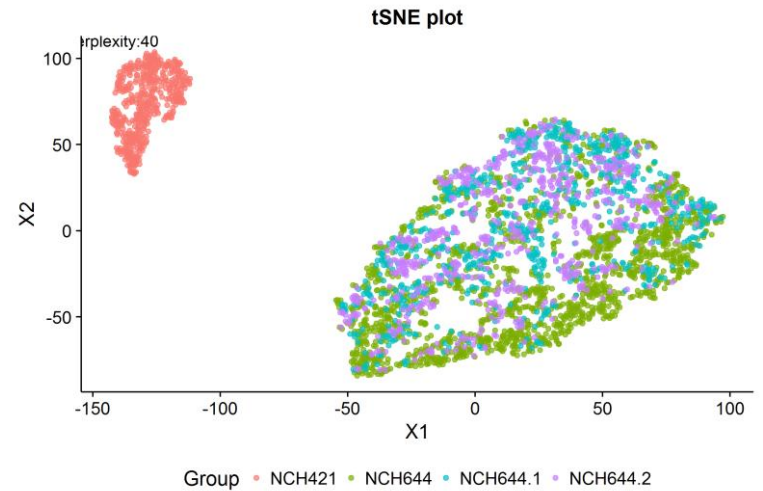
Dominiguez (2016) Cell Research

Single Cell ADAPT: ICA

Correction of batch effect

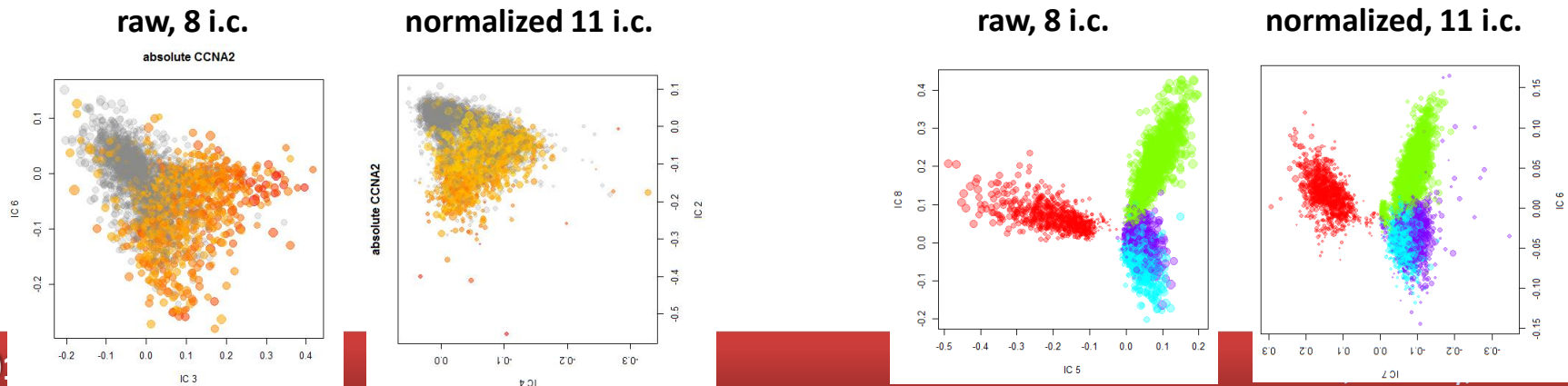


Remove IC7
 →
 set corresponding
 row of \mathbf{M} to 0,
 then $\mathbf{S} \times \mathbf{M}^{\odot} = \mathbf{X}^{\odot}$



Example: Conclusions

- **PCA can capture 2 differences:**
 - b/w NCH421 and NCH644.x cells
 - batch (time?) effect: NCH644 -vs- NCH421 + NCH644.1 + NCH644.2
- **ICA can capture the same as PCA, and in addition:**
 - Cell cycle and other bio-relevant processes
 - Technical bias
- The SC normalization can be omitted. ICA results are similar with or without normalization: **biologically-relevant components are reproducible** in raw and normalized datasets.



Acknowledgements

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

Arnaud MULLER

Tony KAOMA

Dr. Francisco AZUAJE

Dr. Gunnar DITTMAR



NORLUX Neuro-Oncology, LIH

Dr. Anna GOLEBIEWSKA

Prof. Simone NICLOU



Institute Curie, France

Dr. Andrei ZINOVYEV



Fonds National de la
Recherche Luxembourg