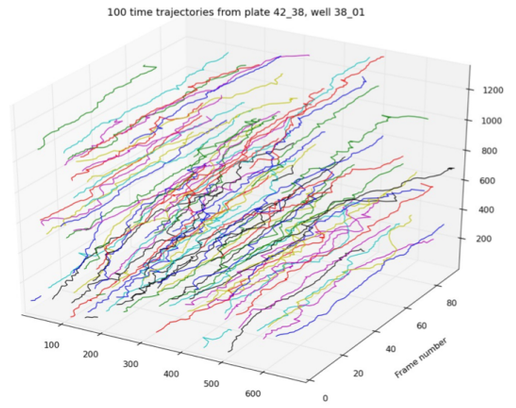
Big data approaches for computational phenotyping: current challenges

Cours ENSMP
October 23, 2017

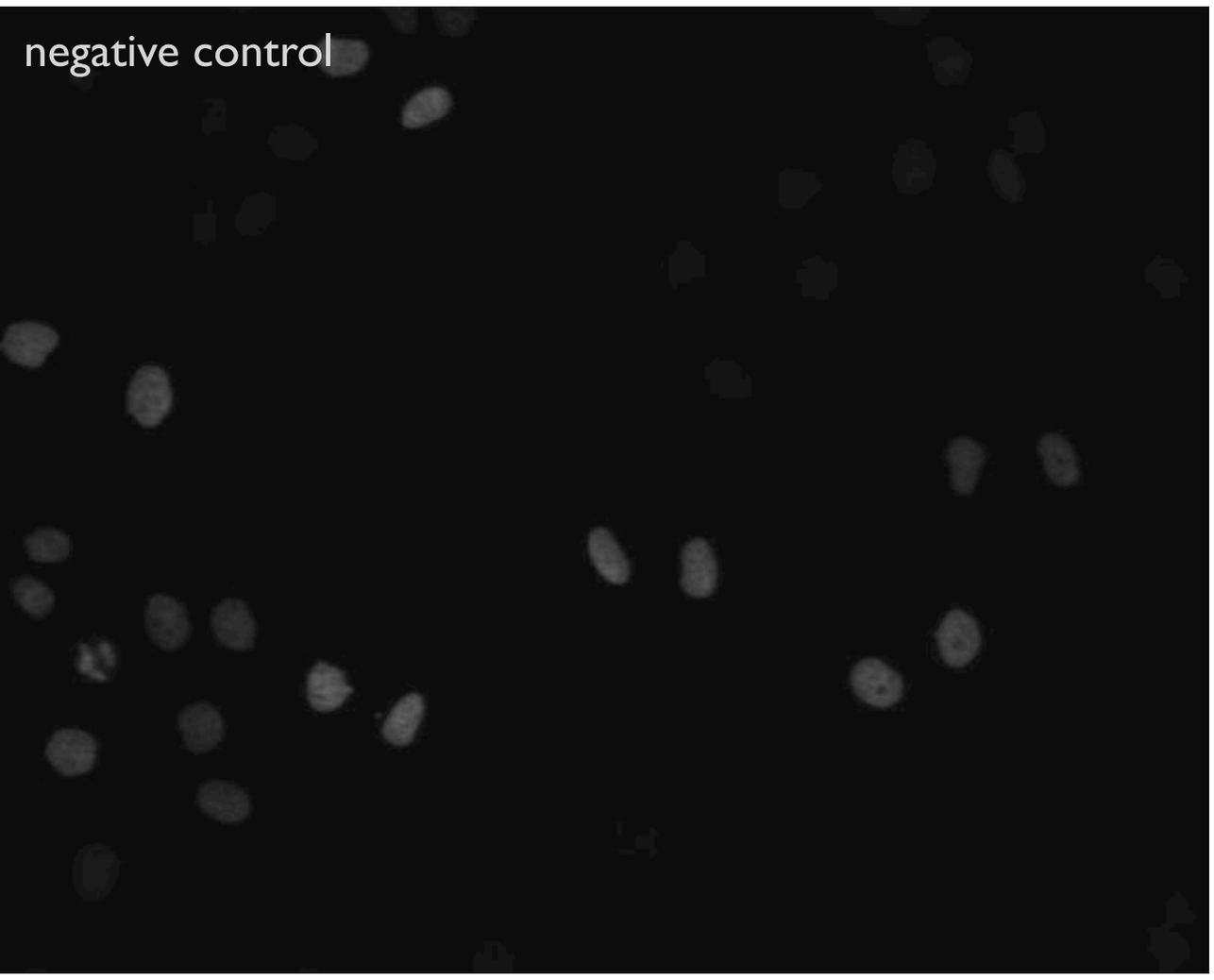
Thomas Walter, Centre for Computational Biology
Mines ParisTech, Institut Curie, Paris

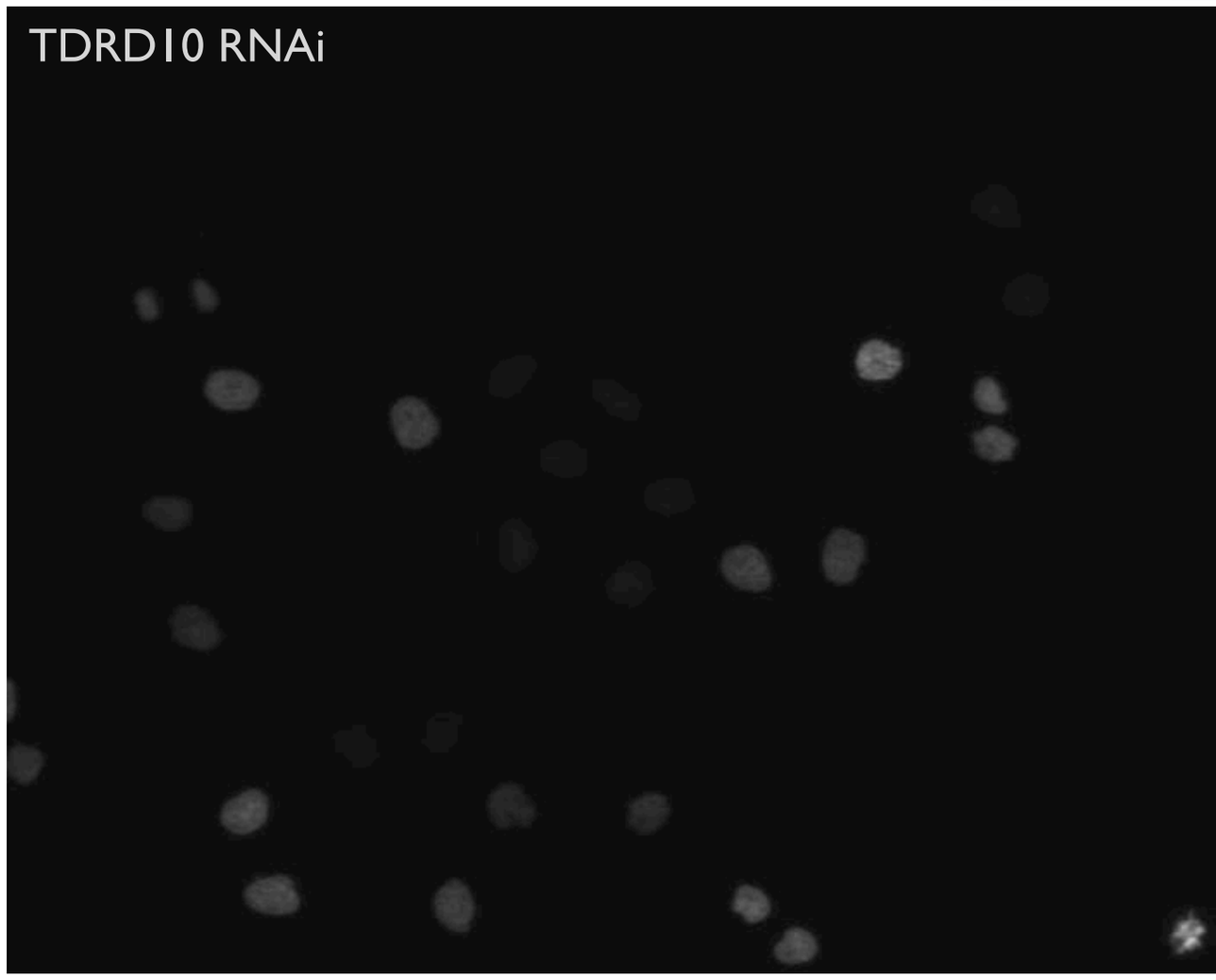
Current challenges

HCS as a scientific resource
Remining existing data sets



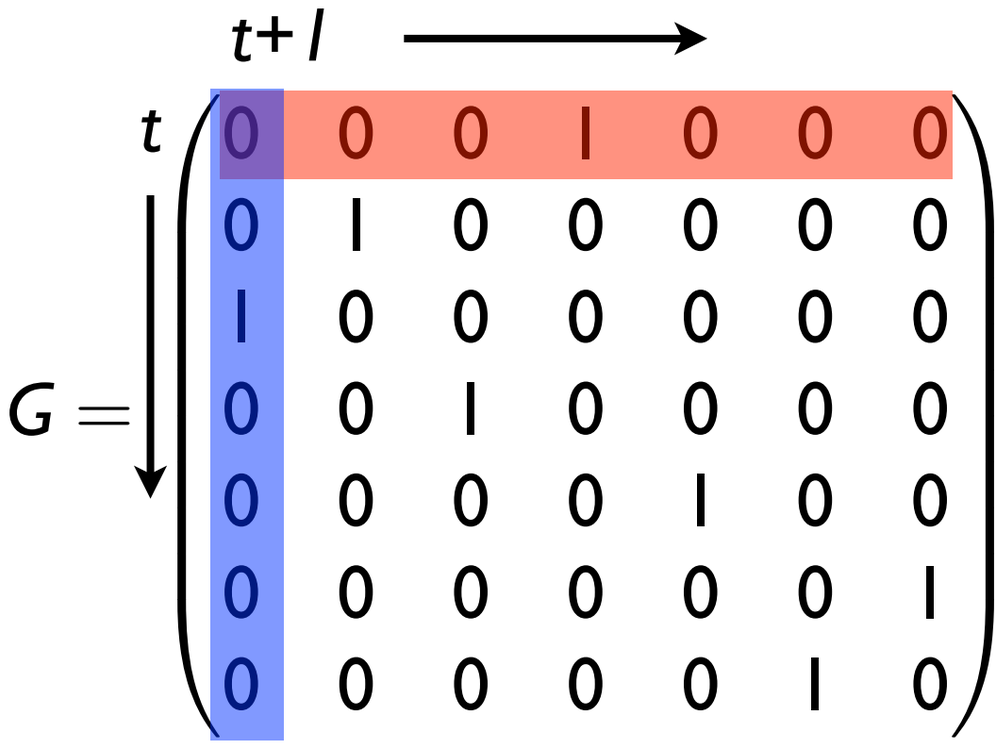
Analysis of cell motility phenotypes





- Live cell imaging data is informative about nuclear motility.
- Are there different types of cell/nuclear movements?
- Which gene knockdowns lead to changed motility?

Particle Tracking: principle



Cost function

$$\phi_{ij} = \|p_i - q_j\|^2 + \Delta(f(p_i), f(q_j))$$

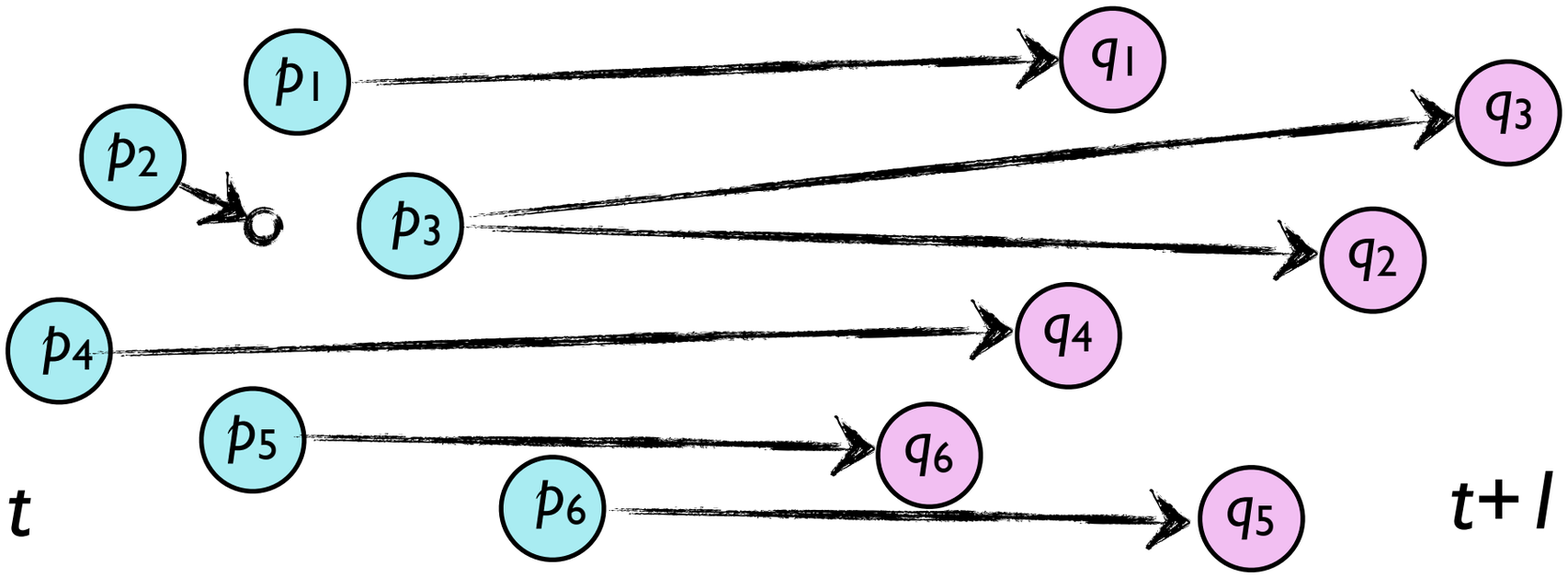
$$\phi = \sum_{i=1}^{N_t} \sum_{j=1}^{N_{t+k}} \phi_{ij} g_{ij}$$

This cost function is optimized.

From particle tracking to cell tracking

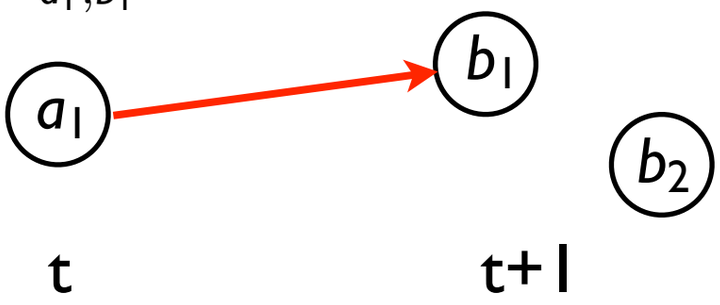
The idea is to formulate the particle tracking problem as an optimization problem. Each association of particles at t and t+1 have an associated cost (typically a mixture of Euclidean and grey level distance). In the case of cells, this cost could potentially involve more and more complex features, such that formulating this cost function ad hoc might be challenging.

Cell Tracking by structured learning

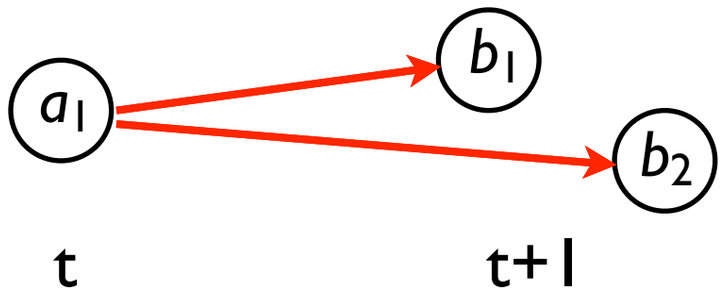


Indicator Variable: $z^e_{i,j} \in \{0, 1\}$

$z^m_{a_1,b_1} = 1$



$z^s_{a_1,\{b_1,b_2\}} = 1$



Constrained optimization problem:

$$\hat{z}(t) = \arg\max_{z} \sum_{\substack{e \\ Obj_{i,t} \\ Obj_{i,t+1}}} < w^e, f_{i,j} > z^e_{i,j}$$

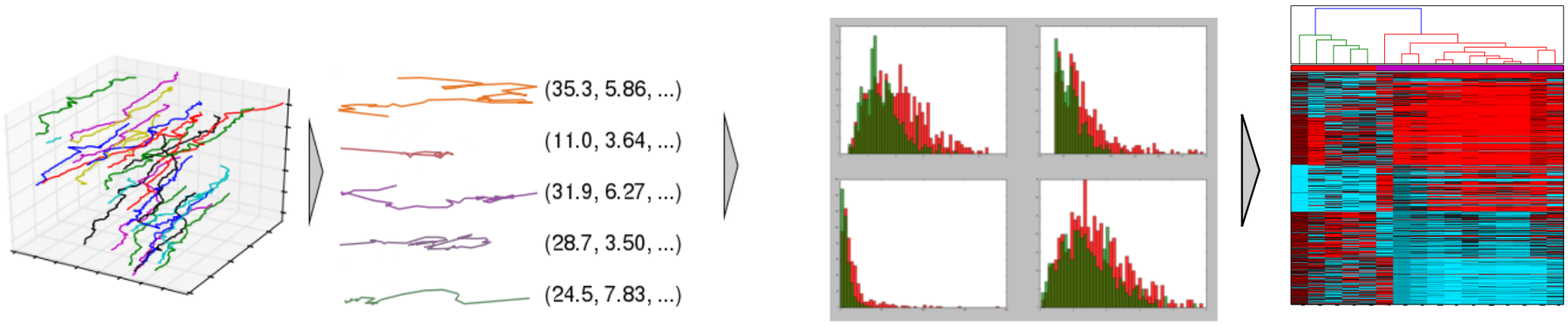
subject to

$$\forall i \in \{0, \ldots, N(t)\} \sum_{Obj_{j,t+1}} z^e_{i,j} = 1$$

$$\forall j \in \{0, \ldots, N(t+1)\} \sum_{Obj_{i,t}} z^e_{i,j} = 1$$

based on Lou and Hamprecht, 2011

Motility analysis



Cell Tracks

Track features (trajectories)

Statistical test (experiment level)

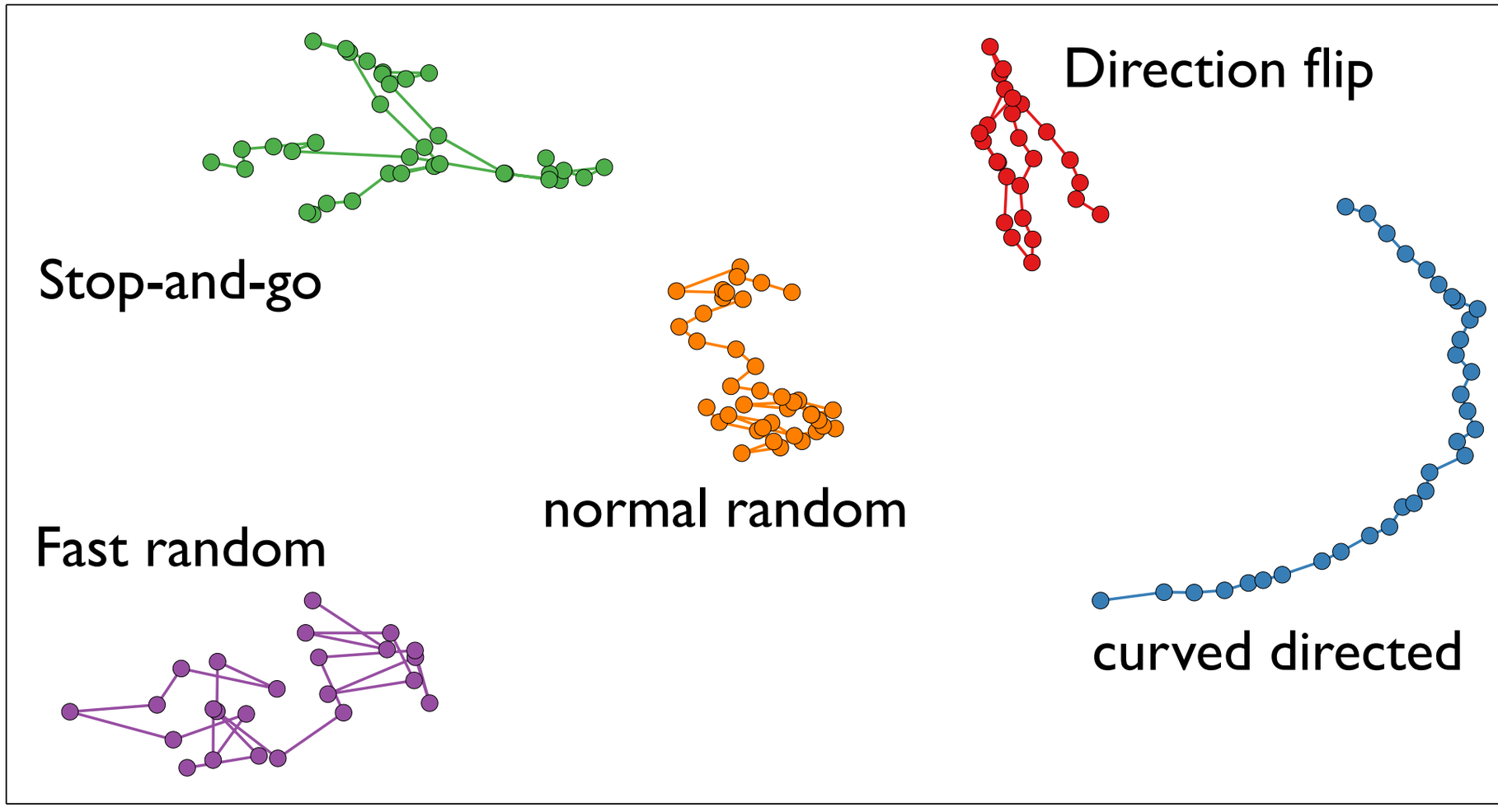
Clustering

- No classes a priori available
- Strategy:
  - Subset: hits and negative controls
  - Clustering of the trajectories from the extreme experiments and some negative controls
- Validation by analysis of simulated trajectories

Schoenauer Sebag et al., Bioinformatics 2015, 31(12), i320–i328.

Simulated trajectories

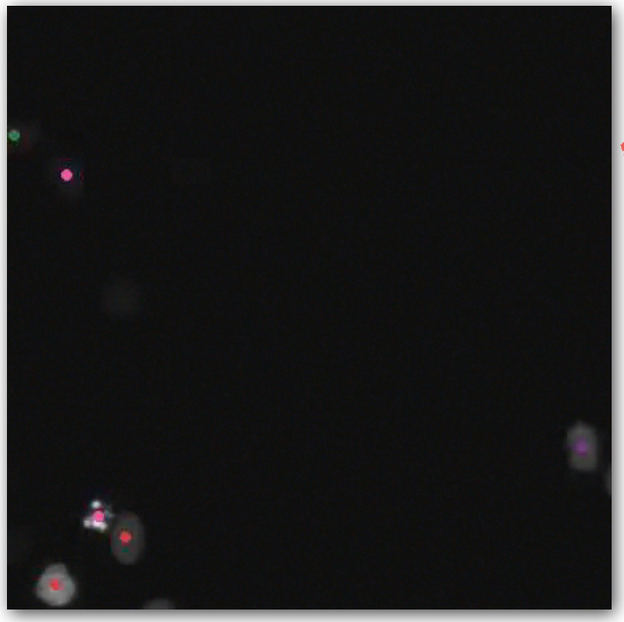
Simulated trajectory types



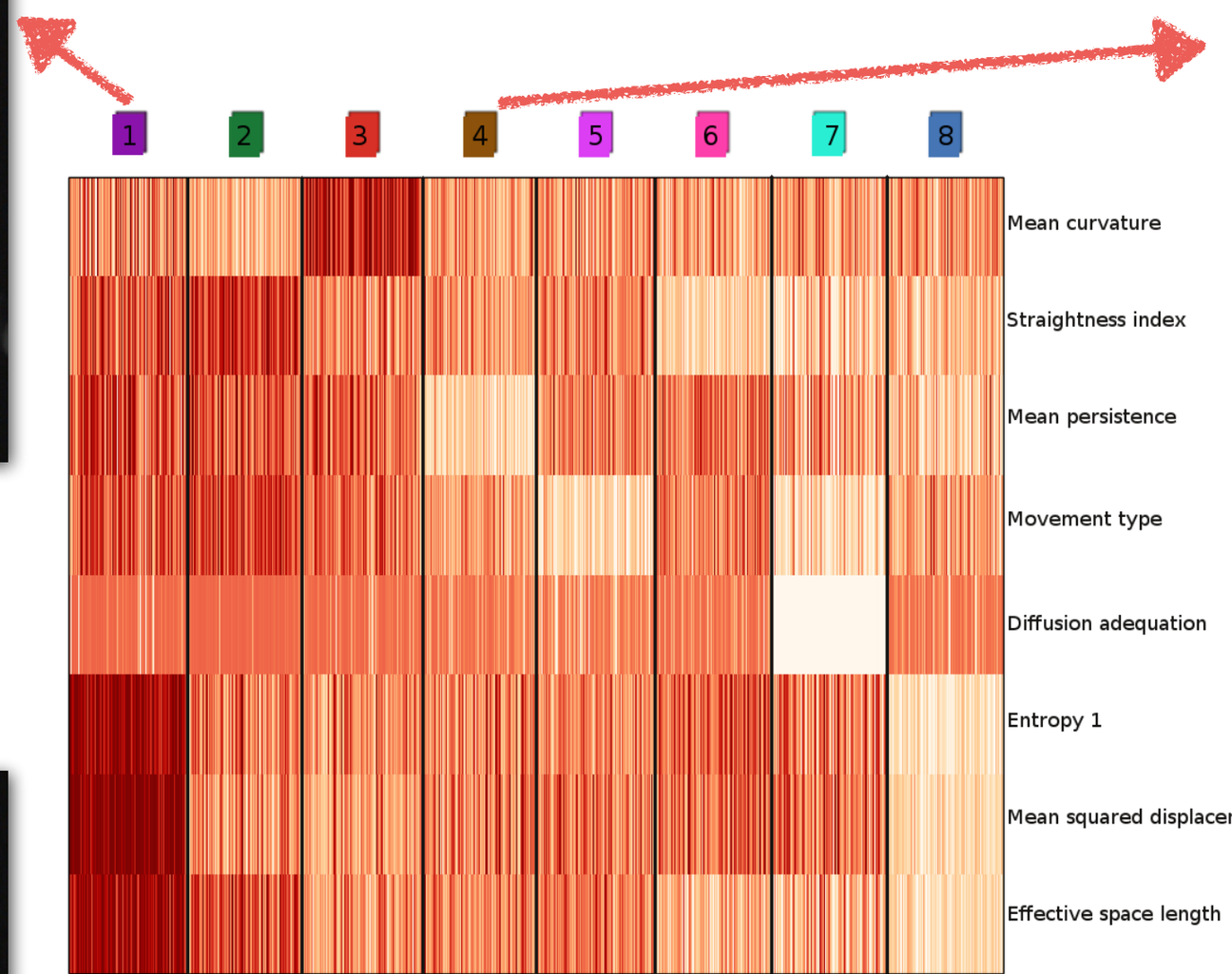
Simulated screen (50000 experiments on 130 plates, in triplicates)

Recall: 91.4% Precision: 89.4%

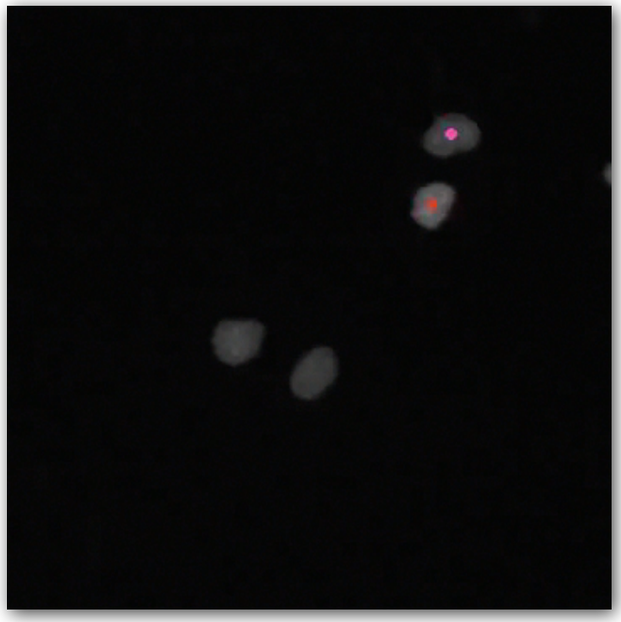
Results on the mitocheck data

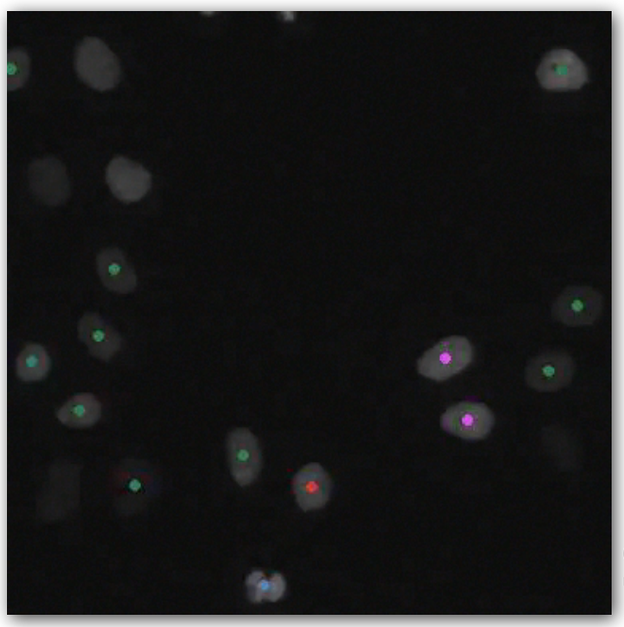


Fast random



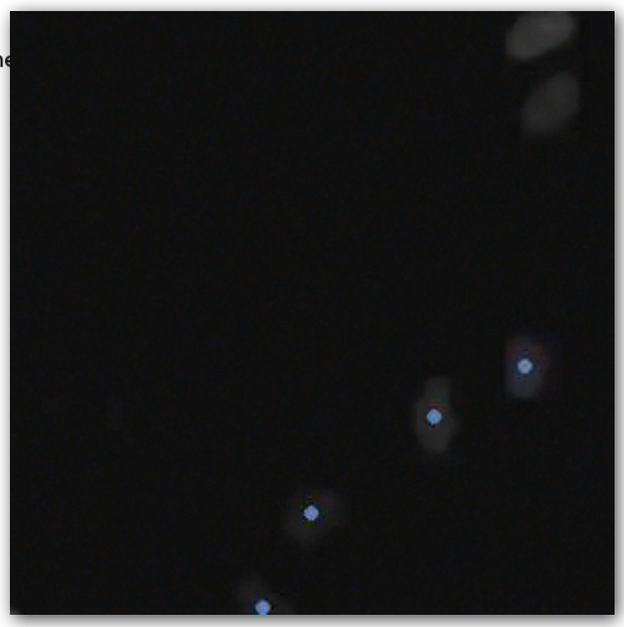
Shiver





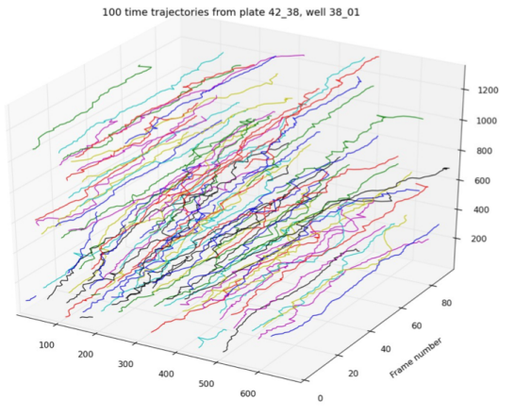
Slow directed

Slow random

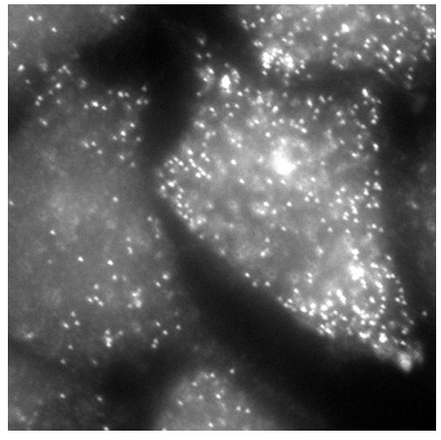


Current challenges

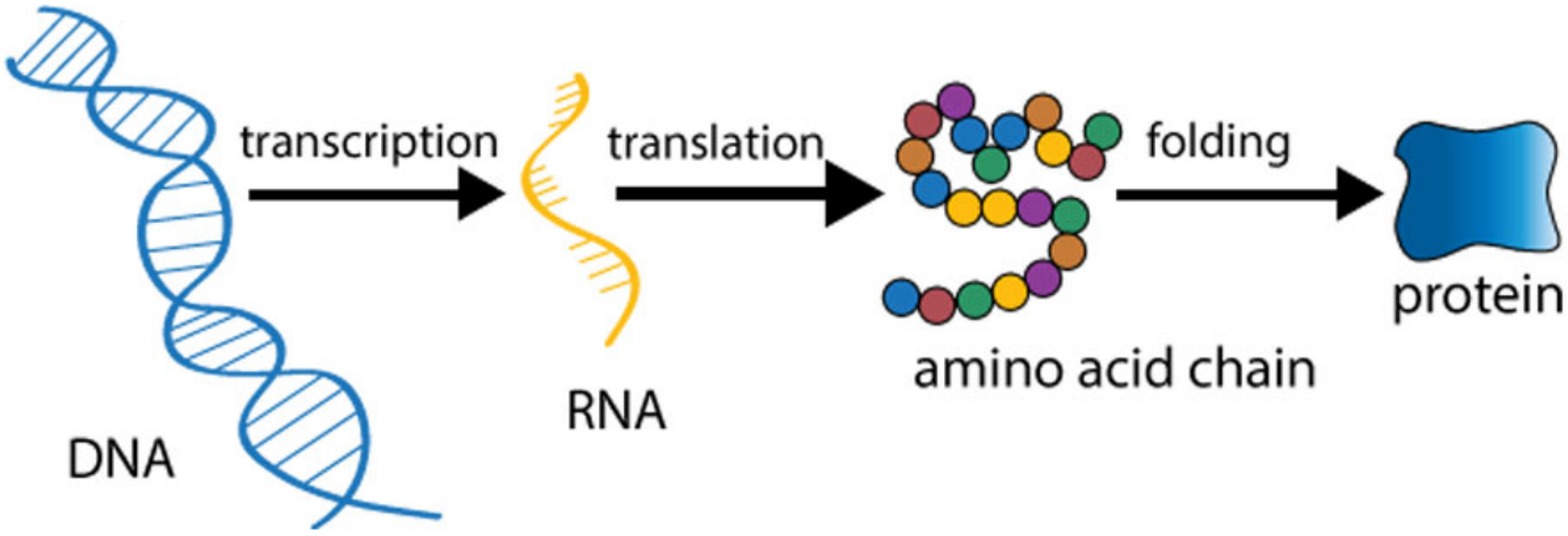
HCS as a scientific resource
Remining existing data sets



Exploring the spatial dimension
Spatial transcriptomics

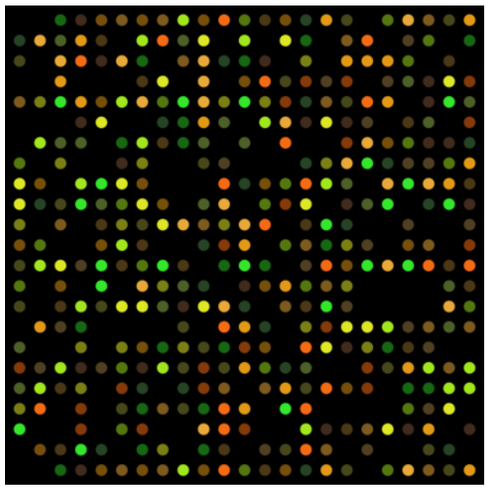


Gene expression



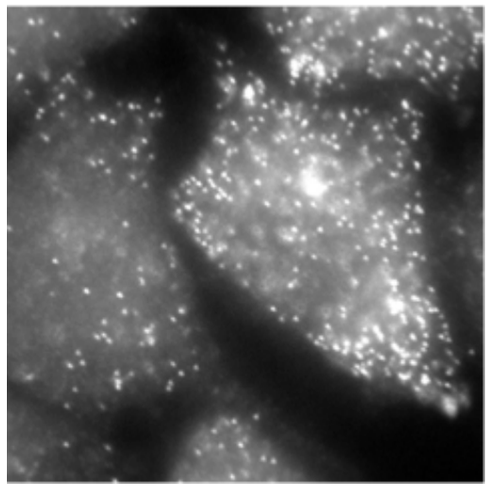
- All cells of an individual share the same genetic code.
- Cells have very different properties and fulfil very different functions.
- This is achieved by “expression” of different sets of genes.
- A gene is expressed, if its information is copied to an RNA molecule.

The study of gene expression



Microarray

- Gene expression is traditionally quantified by the number of mRNA.
- The number of mRNA can be measured by microarrays or RNA sequencing.



smFISH

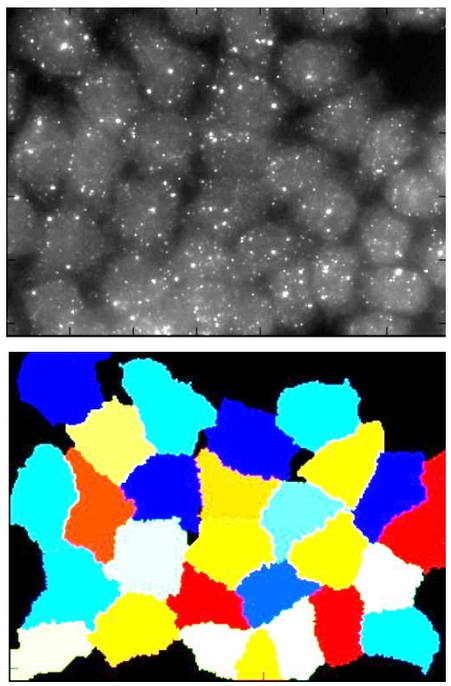
- Today, we can also analyze the spatial distribution of transcripts.
- It turns out that many transcripts are not uniformly distributed.

Tsanov, N., Samacoits, A. et al.,
Nucleic Acids Research. September 2016.

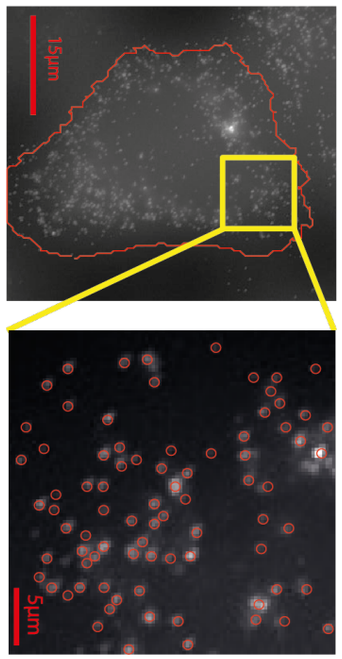
Bertrand group (IGMM)
Zimmer group (Pasteur)

Analysis of smFISH data

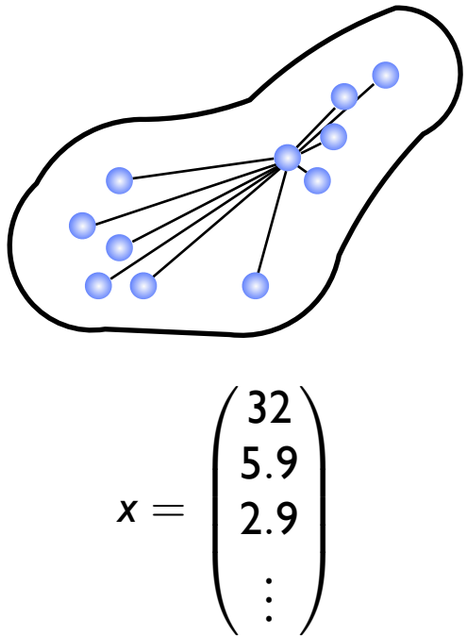
Aubin Samacoits



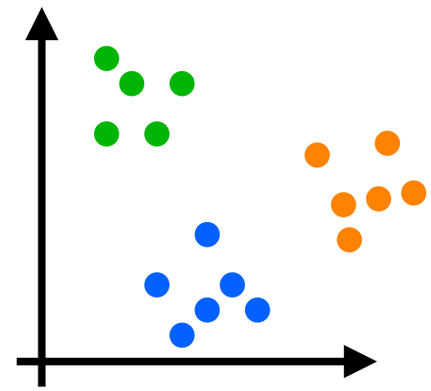
Cell Segmentation



Spot detection



Features



Clustering

Problem:
How can we assess the quality of the strategy?

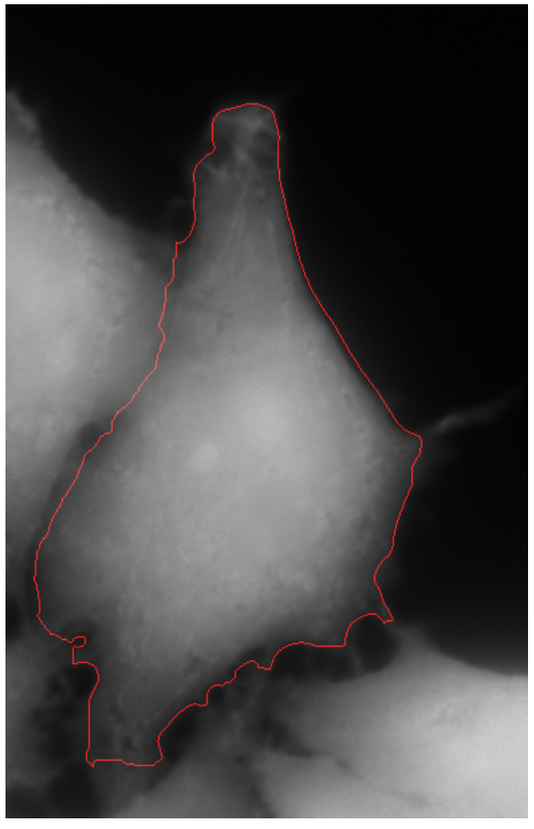
Battich et al., Nature Methods, October 2013.

Tsanov, N., Samacoits, A. et al., Nucleic Acids Research. September 2016.

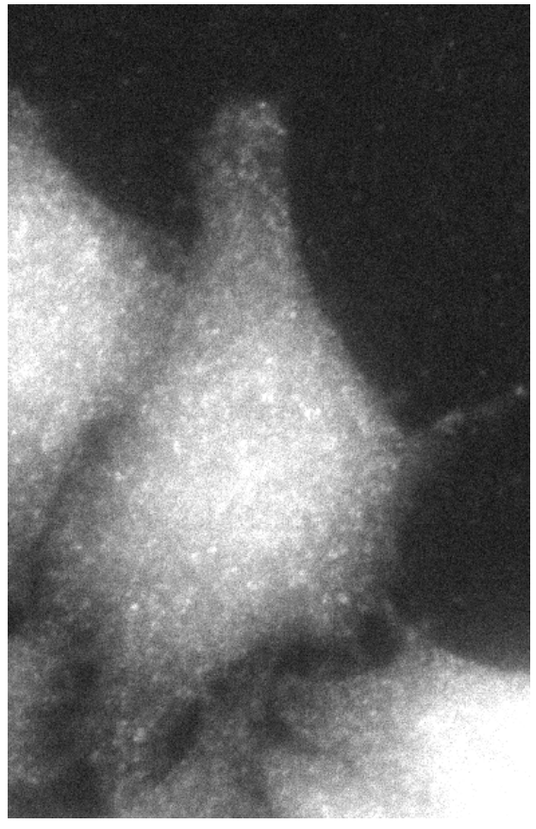
Generating experimental data as input for simulations

Available markers for the simulation environment:
- GAPDH (abundant housekeeping gene)
- mock FISH (background simulation)

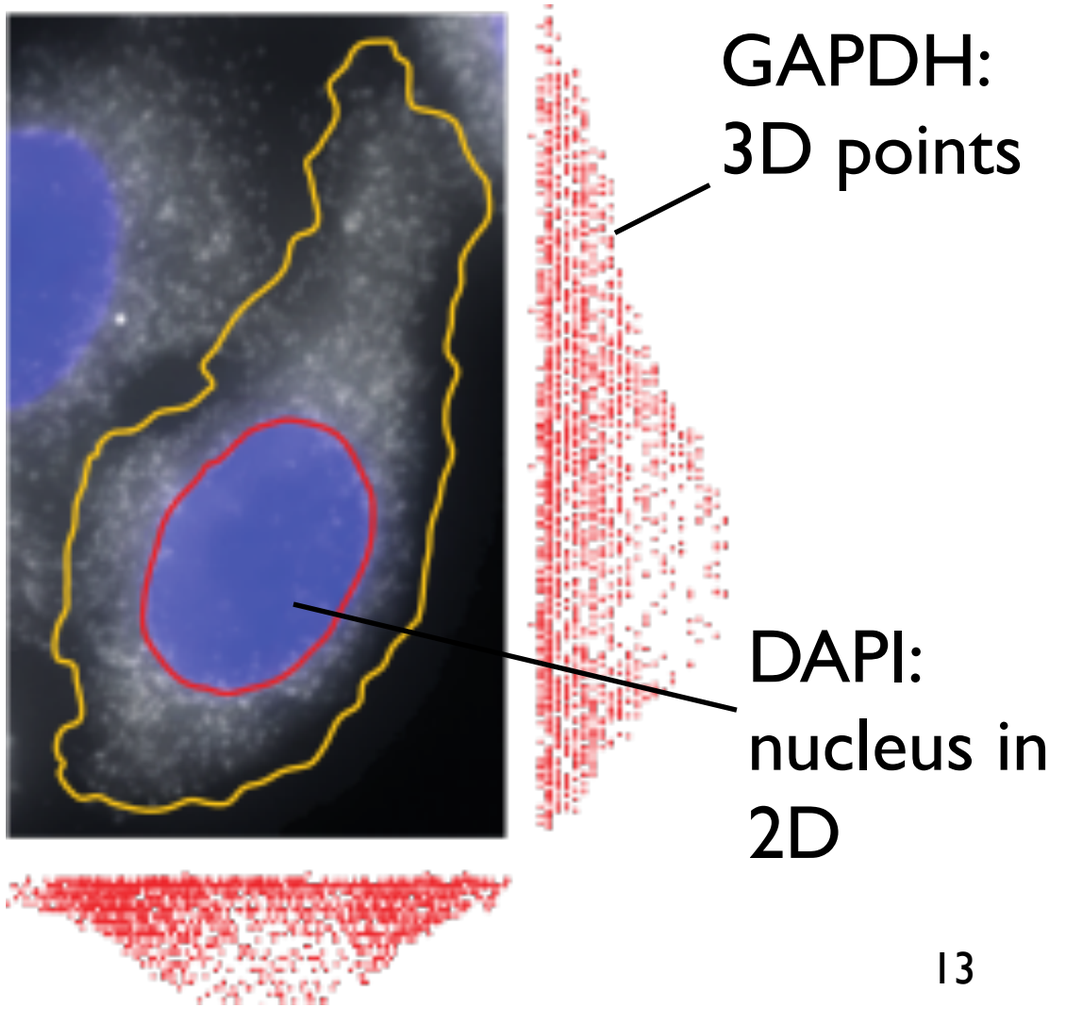
Markers available in standard screening applications:
- Cell Mask (cytoplasm)
- DAPI (nucleus)



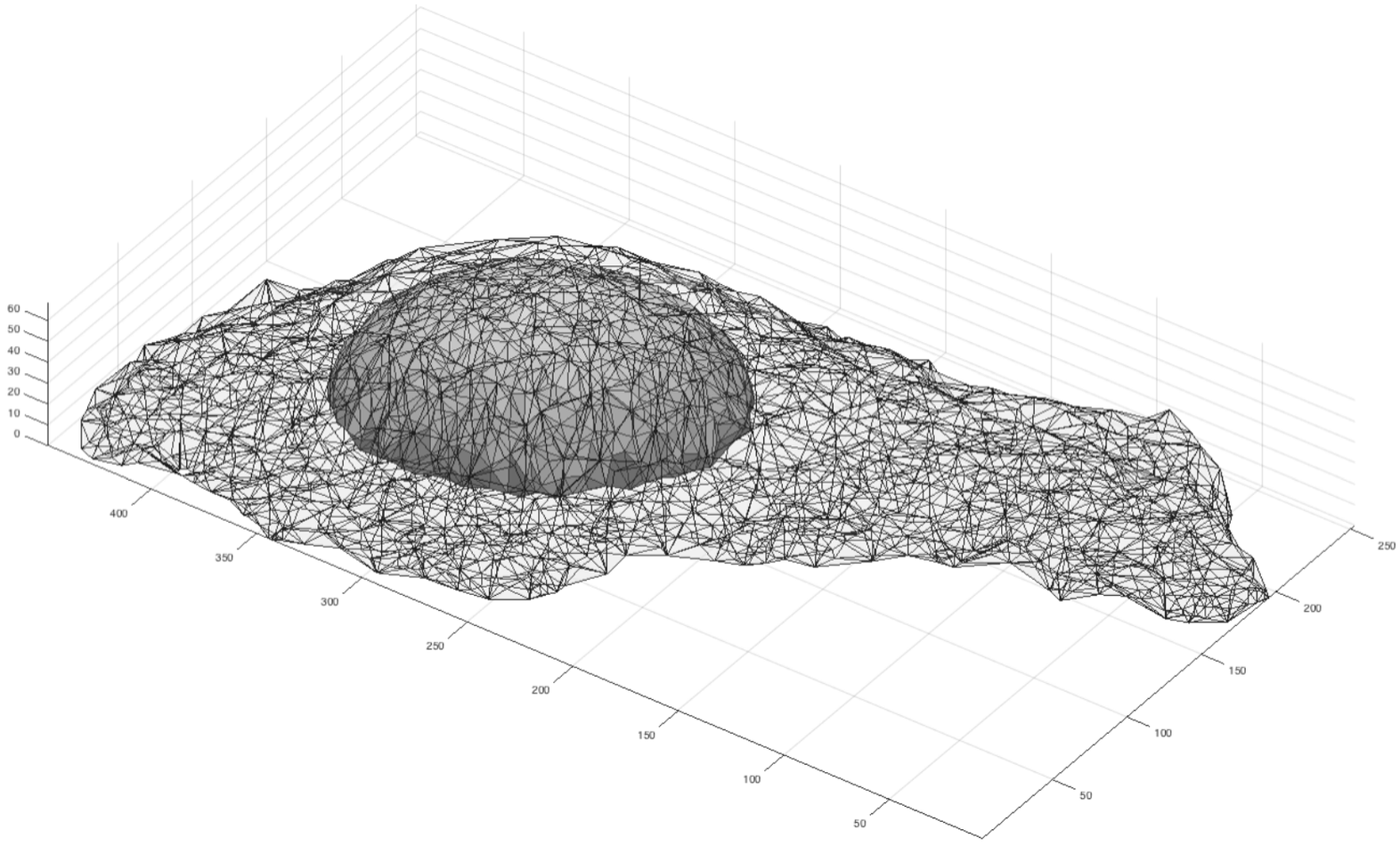
CELLMASK:
cytoplasm in 2D



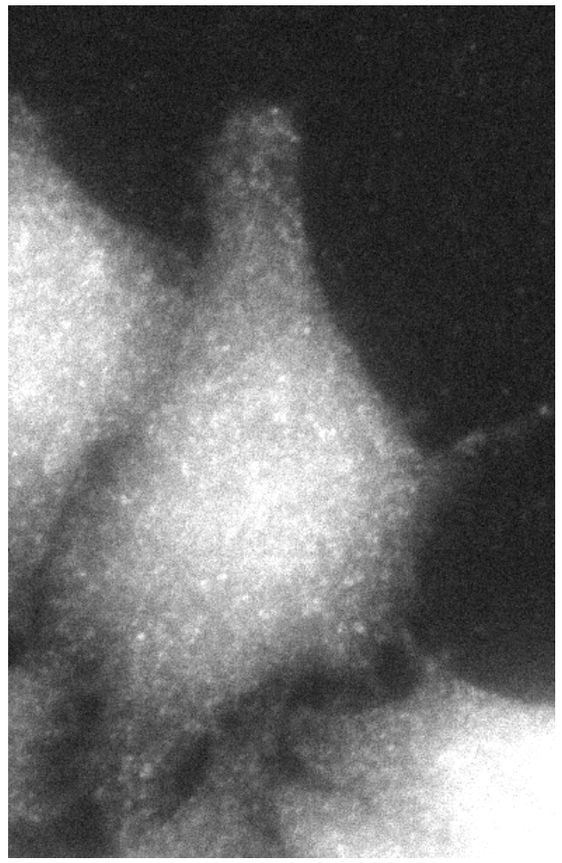
mock FISH:
background



A cell library for simulations



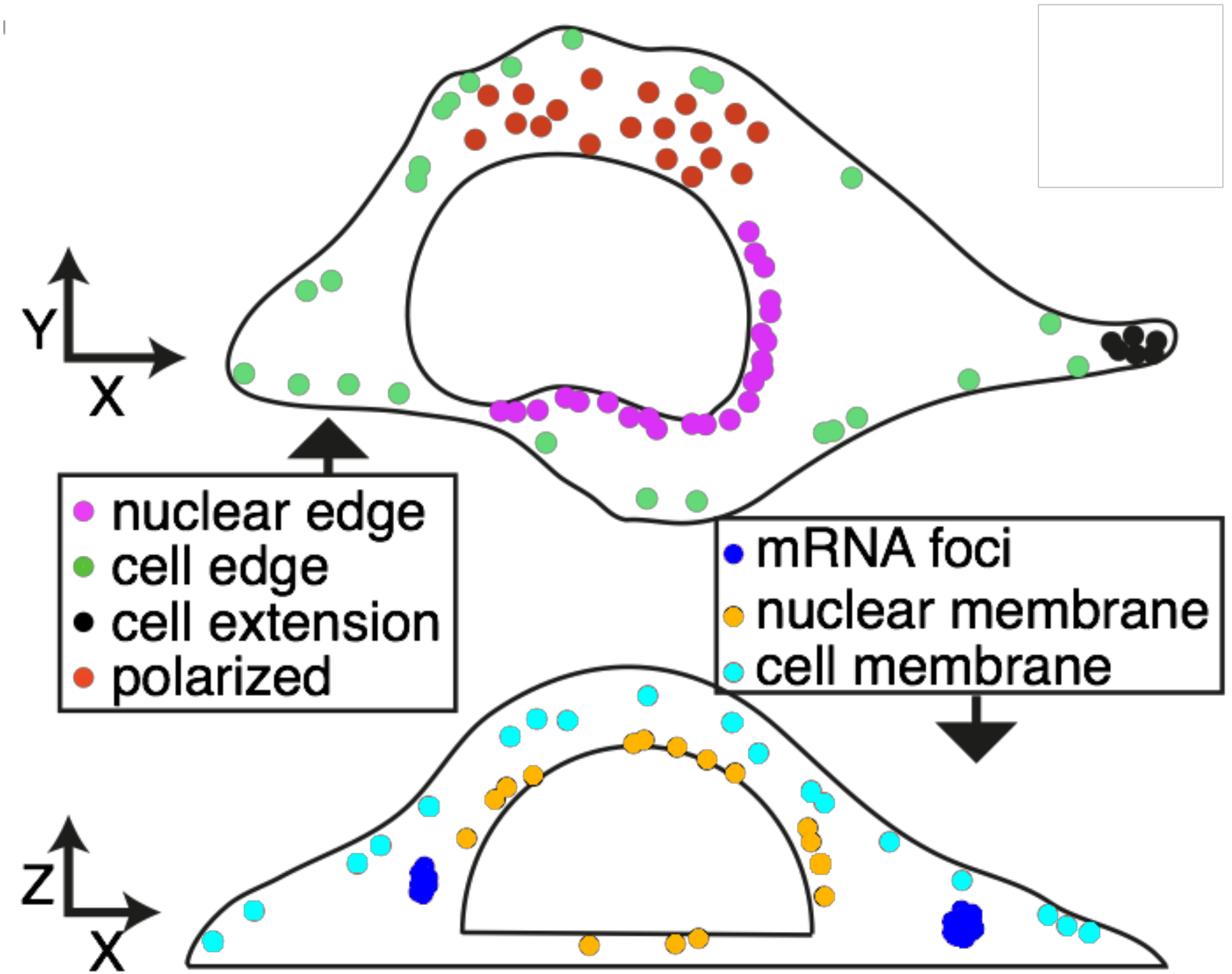
3D reference volume, measured from real cells



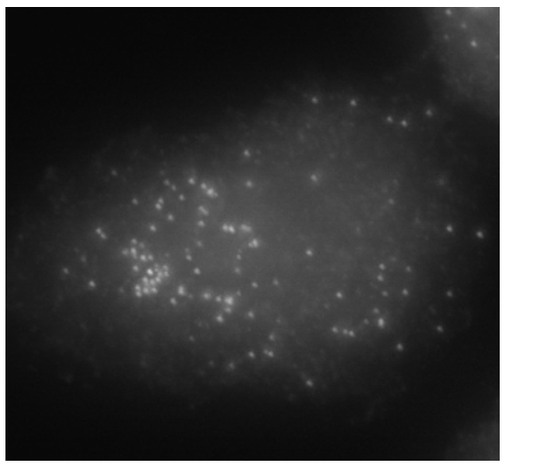
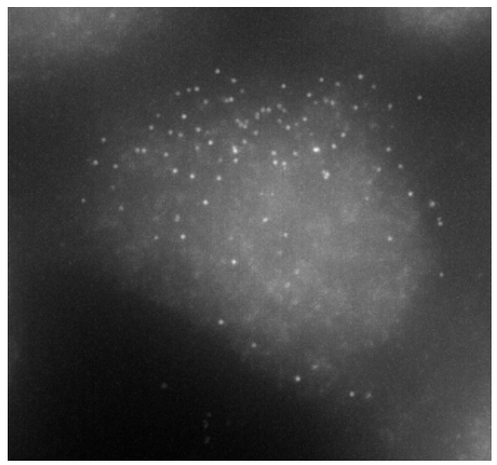
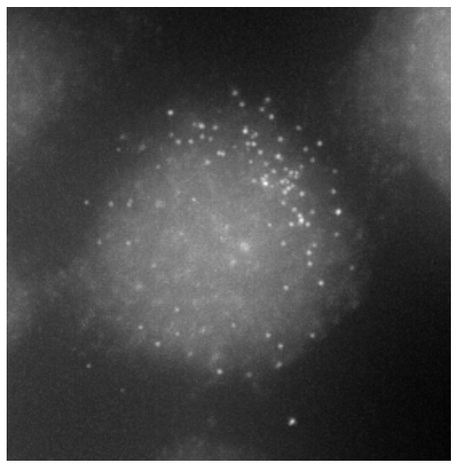
mock-FISH: typical background

- Cell and nuclear shape in 3D, measured from microscopy images.
- Background signal acquired by MOCK-FISH.
- We can now simulate spot distributions (with a realistic PSF) according to a priori distributions.

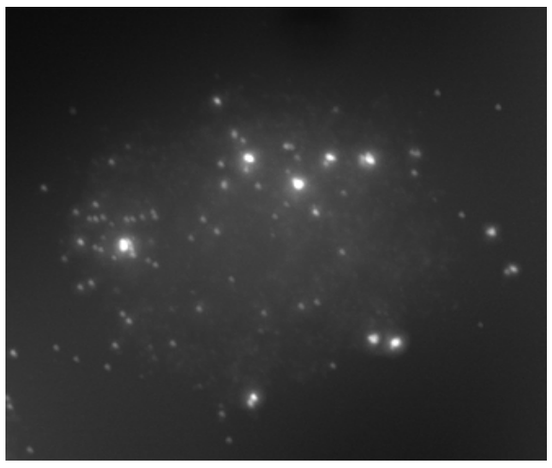
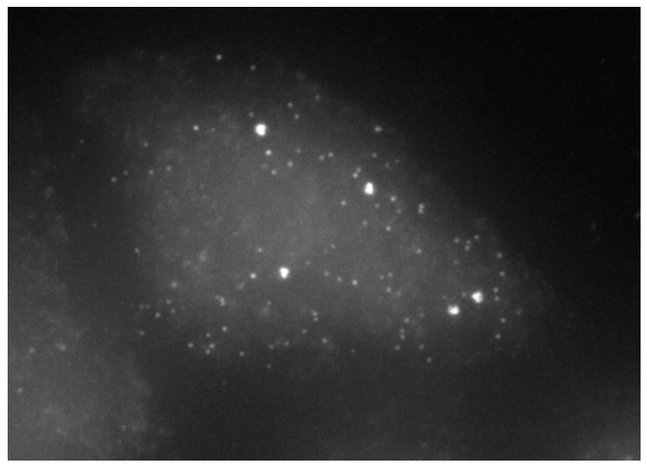
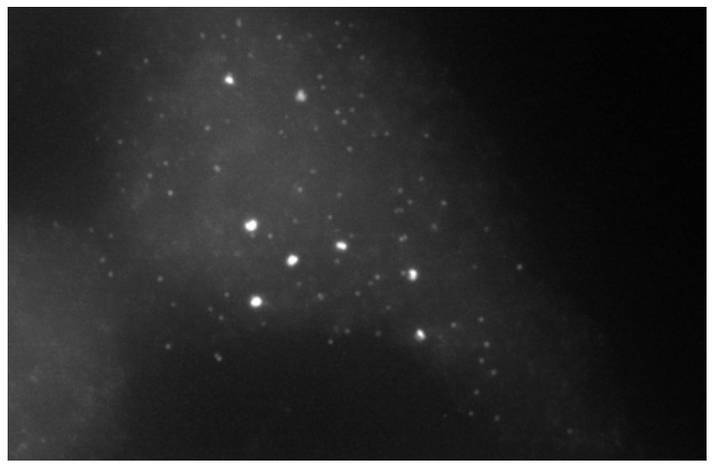
Simulation of different localization patterns



Simulation of localization patterns: examples

Polarized

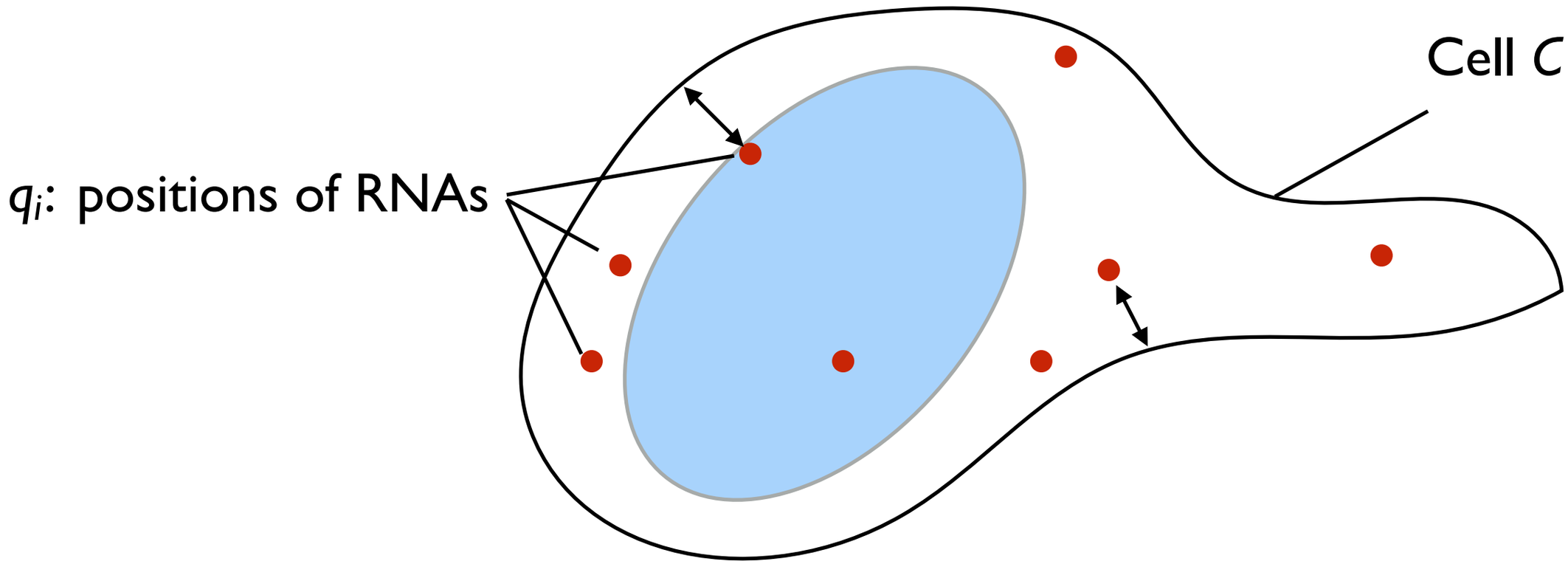




Blobs

Original *Simulated*

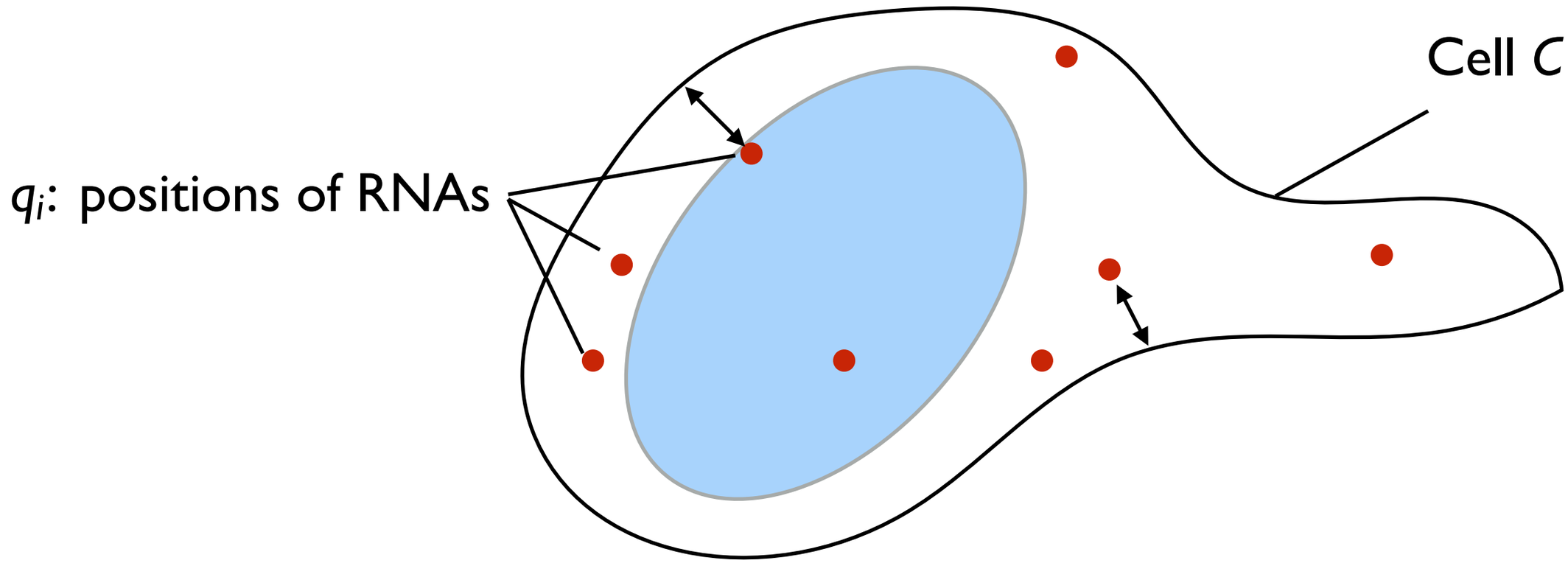
- With this we can now easily generate data sets of cells with defined localization patterns.
- We can also control the strength of the pattern.

Features for the description of RNA localization



- In each segmented cell *C*, we have segmented a number *N(C)* of RNAs at positions q_i.
- We can now calculate a number of features that describe the distribution of these positions with respect to
  - Landmarks (nucleus, plasma membrane, extensions)
  - Each other (based on Ripley's L-function)

Features for the RNA distribution (examples)

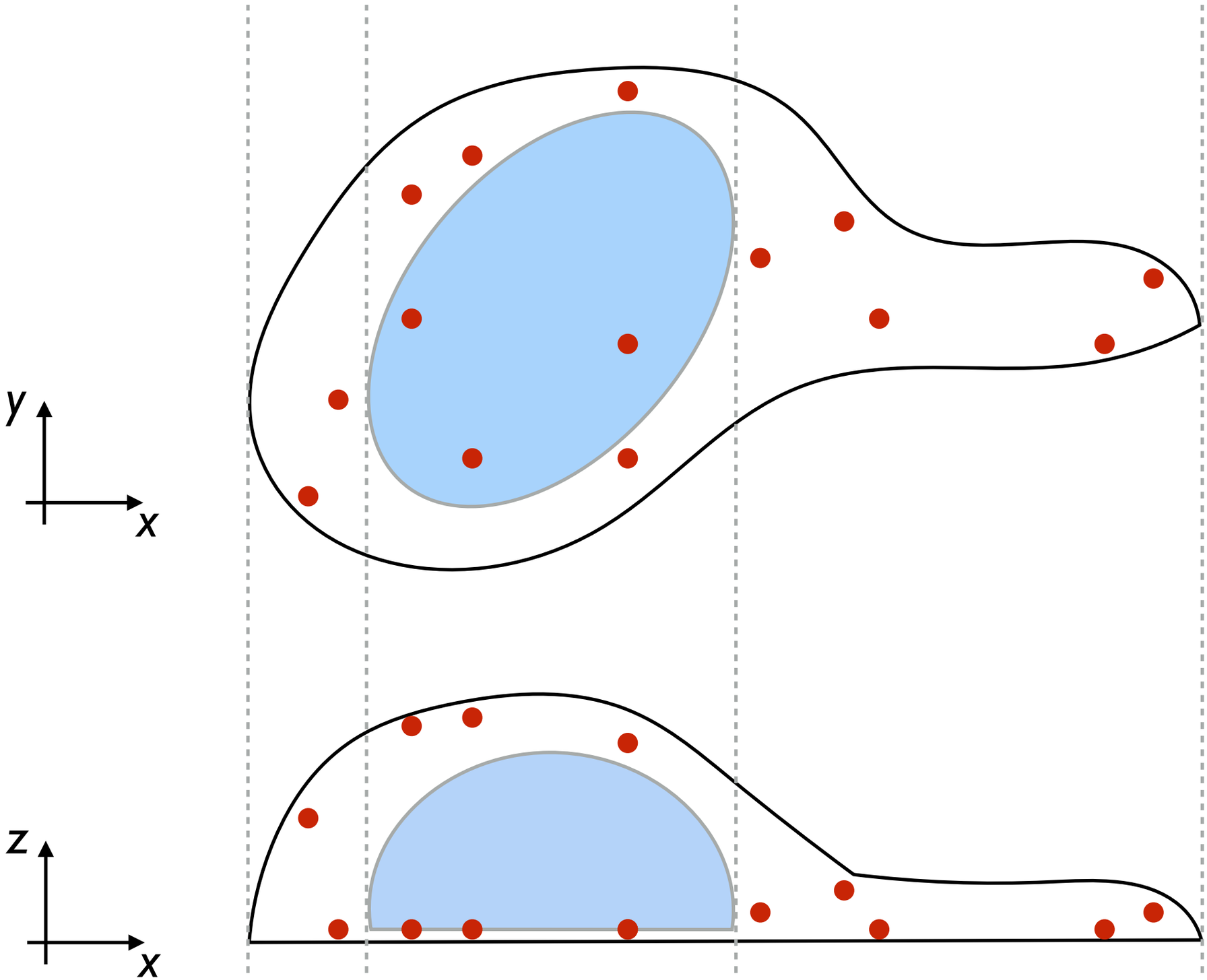


- Average distance to the nuclear and cytoplasmic membrane
- Importantly, the distance is normalised (dividing by its expectation under complete spatial randomness).

$$\overline{d_{cyto}}(C) = \frac{1}{N}\sum_{i}^{N}\min_{y\notin C}\|q_i - y\|_2$$

$$\overline{d'_{cyto}}(C) = \frac{d_{cyto}}{\mathbb{E}(d_{cyto}|CSR)} = \frac{|C|}{N}\frac{\sum_i^N \min_{y\notin C}\|q_i - y\|_2}{\sum_{x\in C}\min_{y\notin C}\|x - y\|_2}$$

Features for the RNA distribution (examples)

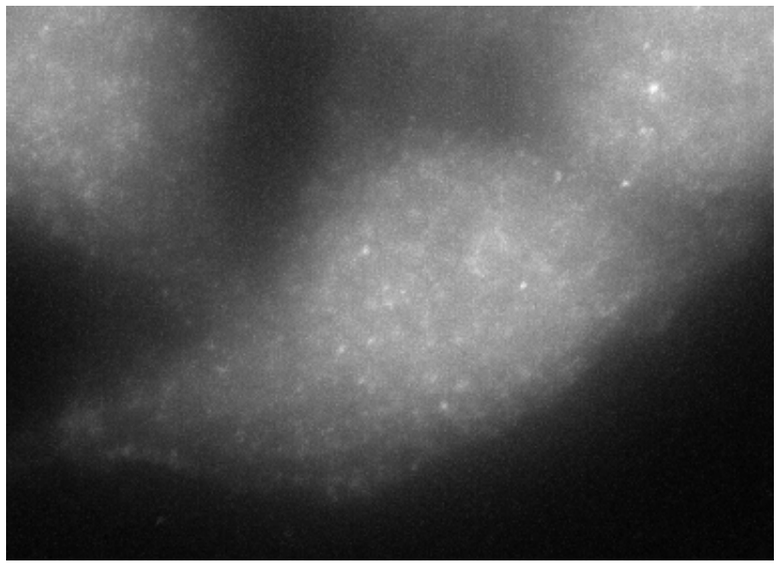


Standard screens: no marker for the plasma membrane

How can we construct a feature informative on the proximity to the membrane ?

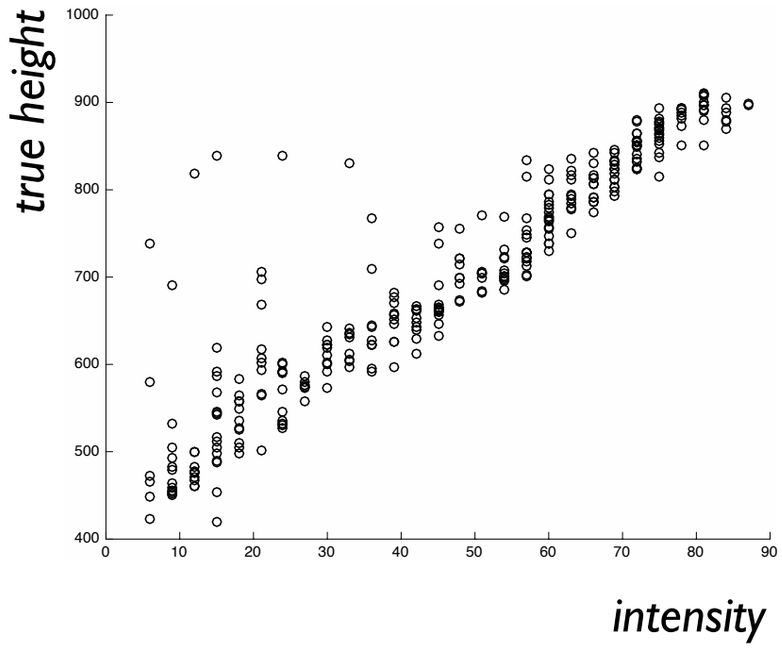
Features to describe RNA localisation (examples)

- Localization at the plasma membrane is difficult to quantify in the absence of a membrane marker.
- Can we estimate the cell membrane from the background signal?
- Is the background signal intensity proportional to the height of the cell?



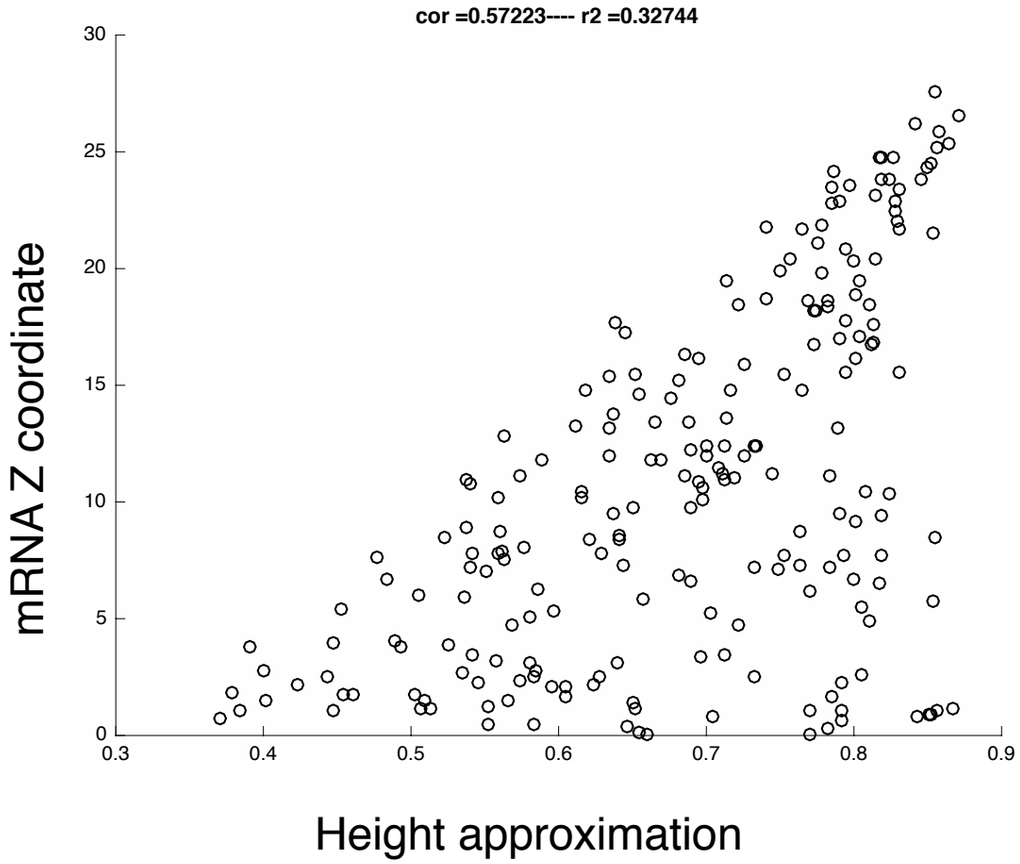
smFISH

Background estimation

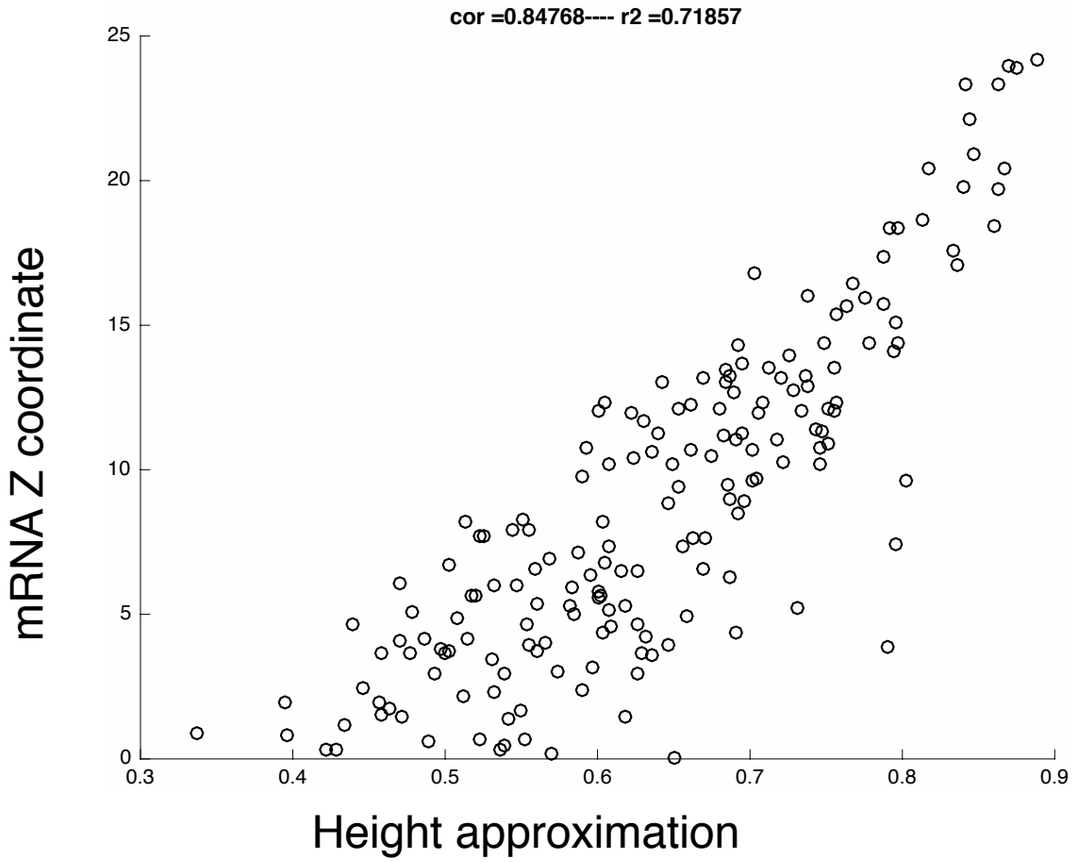


Features to describe RNA localisation (examples)

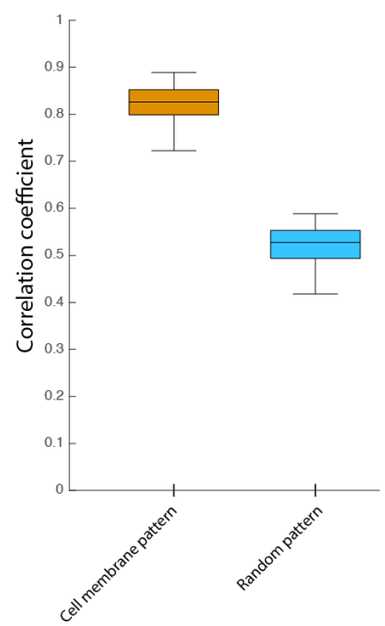
Simulation: Random localization in the cytoplasm



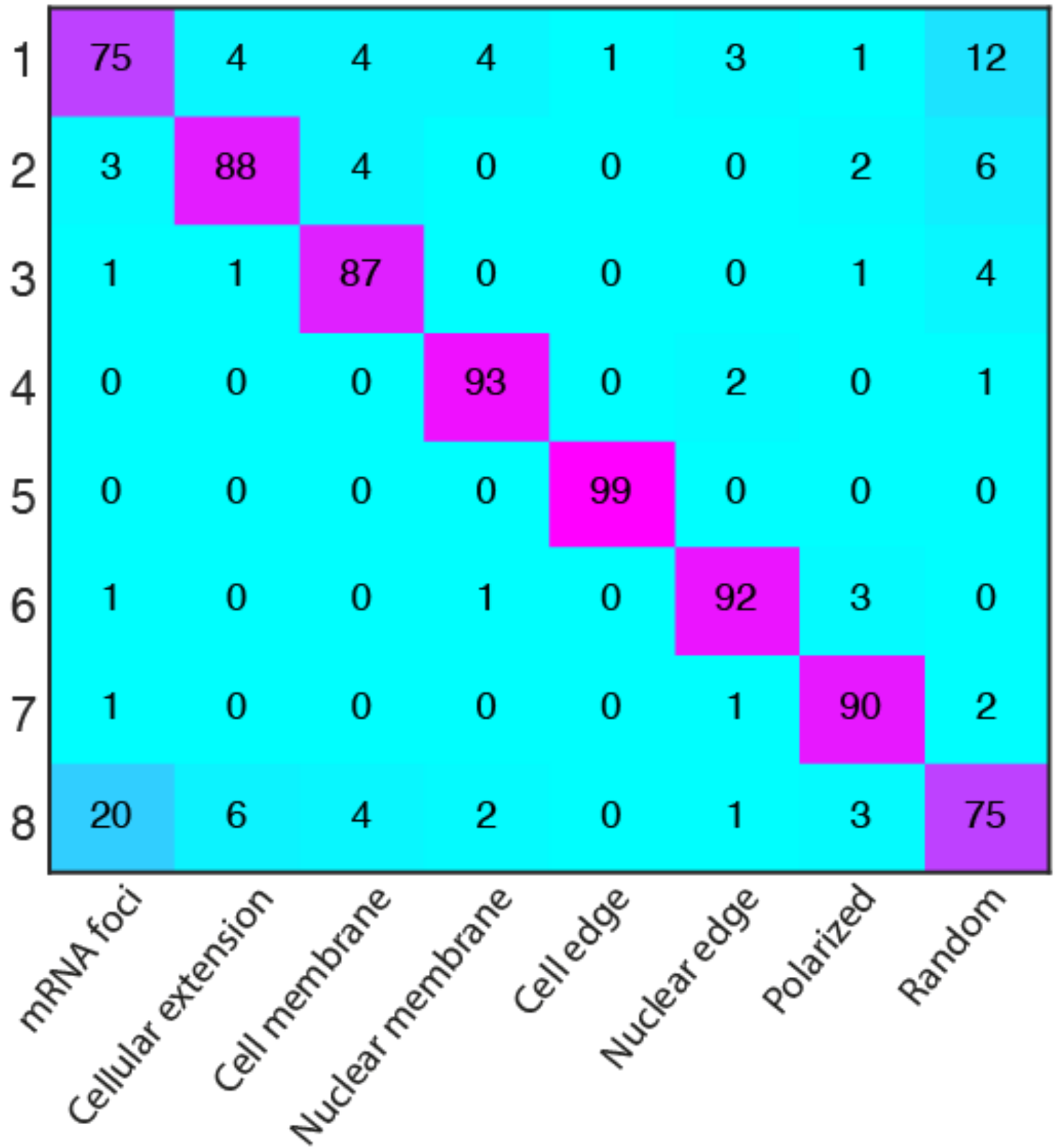
Simulation: localization close to the membrane



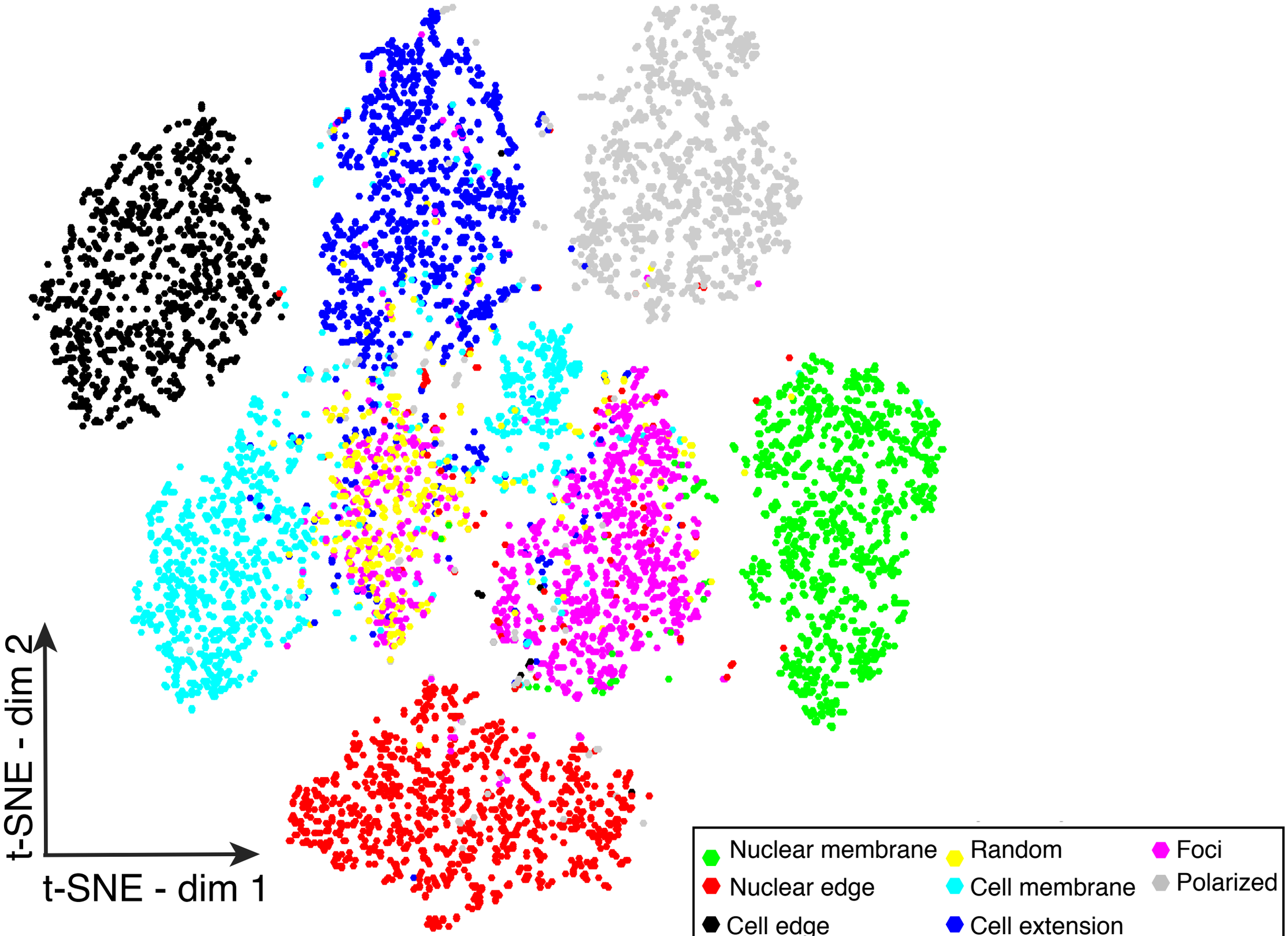
Feature: Pearson correlation between intensities of the background and z-coordinates of the RNA.



Confusion matrix k-means



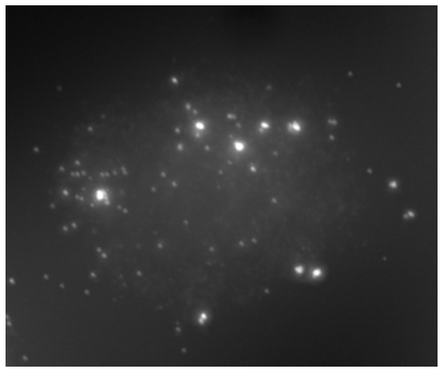
Benchmark of a new feature set (t-SNE projection)

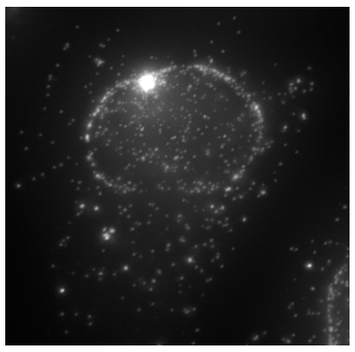


Application to real data

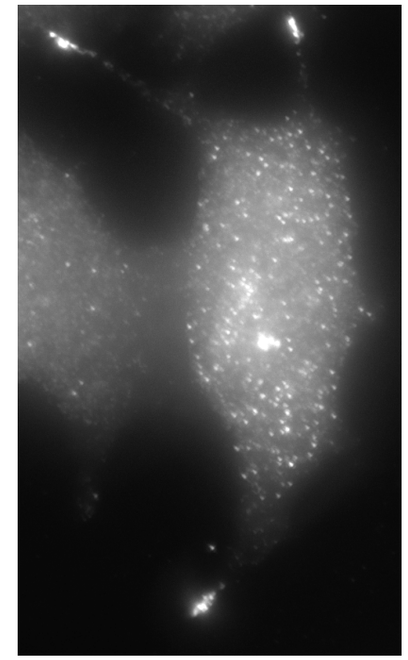
Small data set of 10 genes with several localization patterns.

Foci

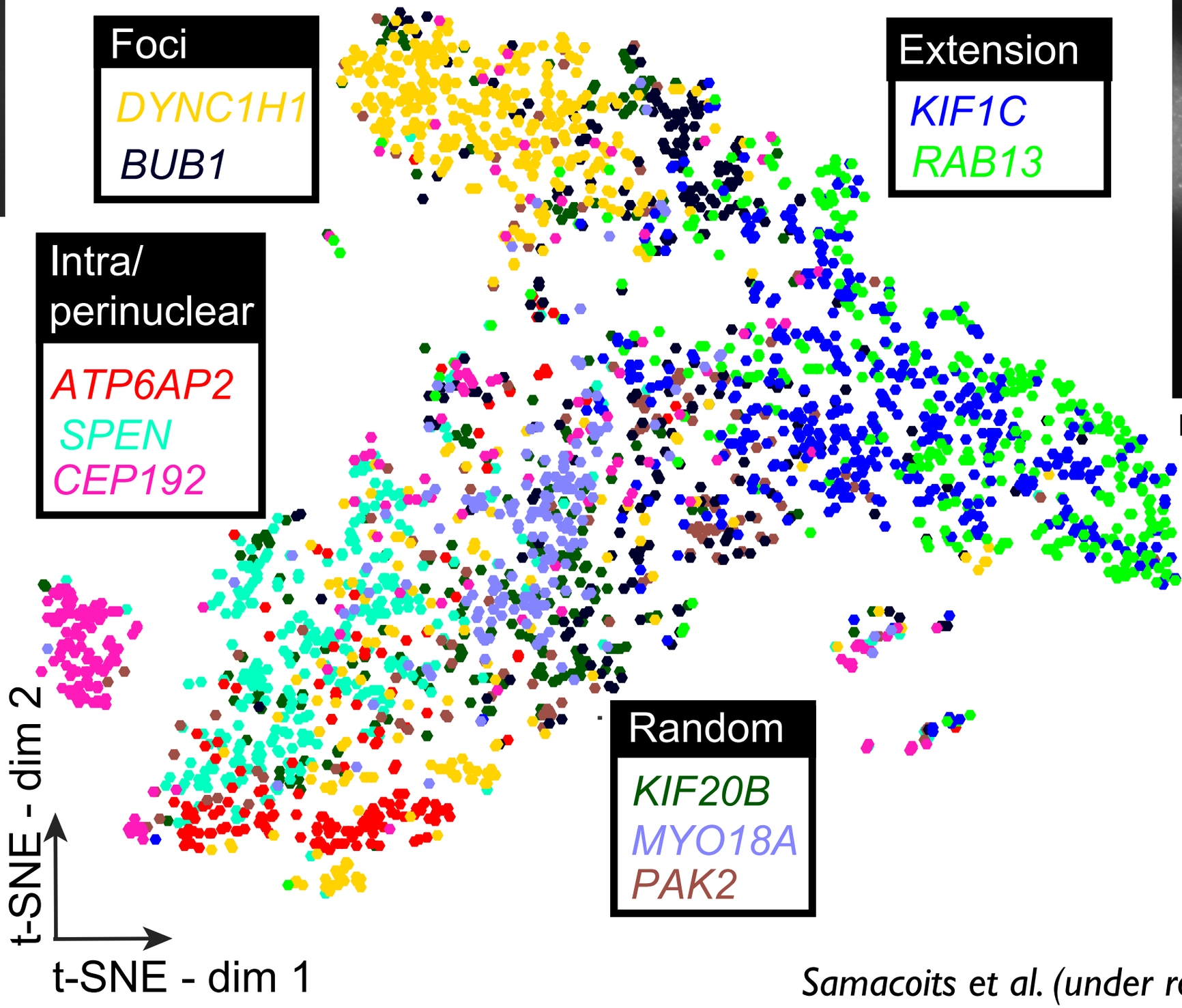




Perinuclear



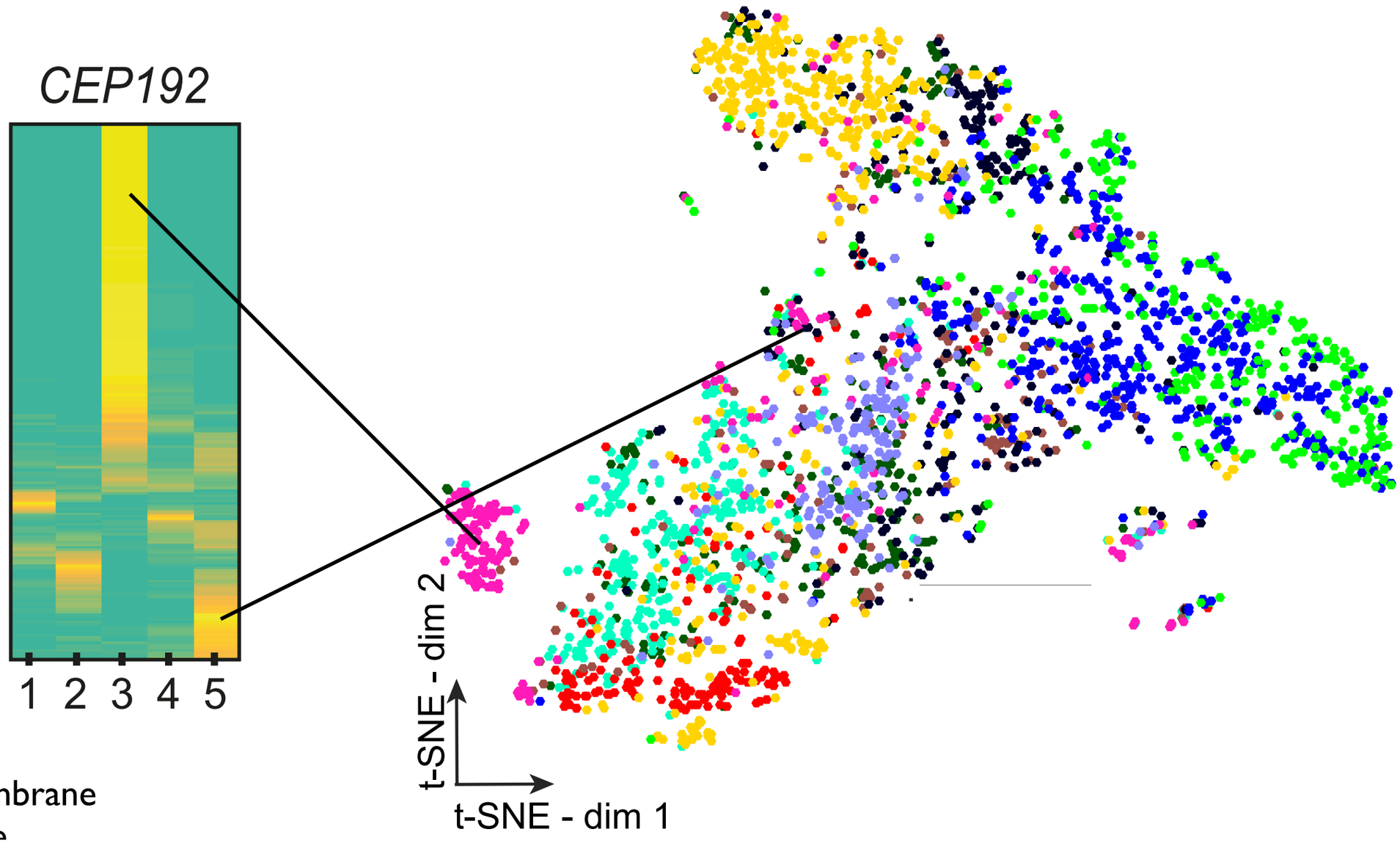
Extension



Samacoits et al. (under review)

Quantifying localisation heterogeneity

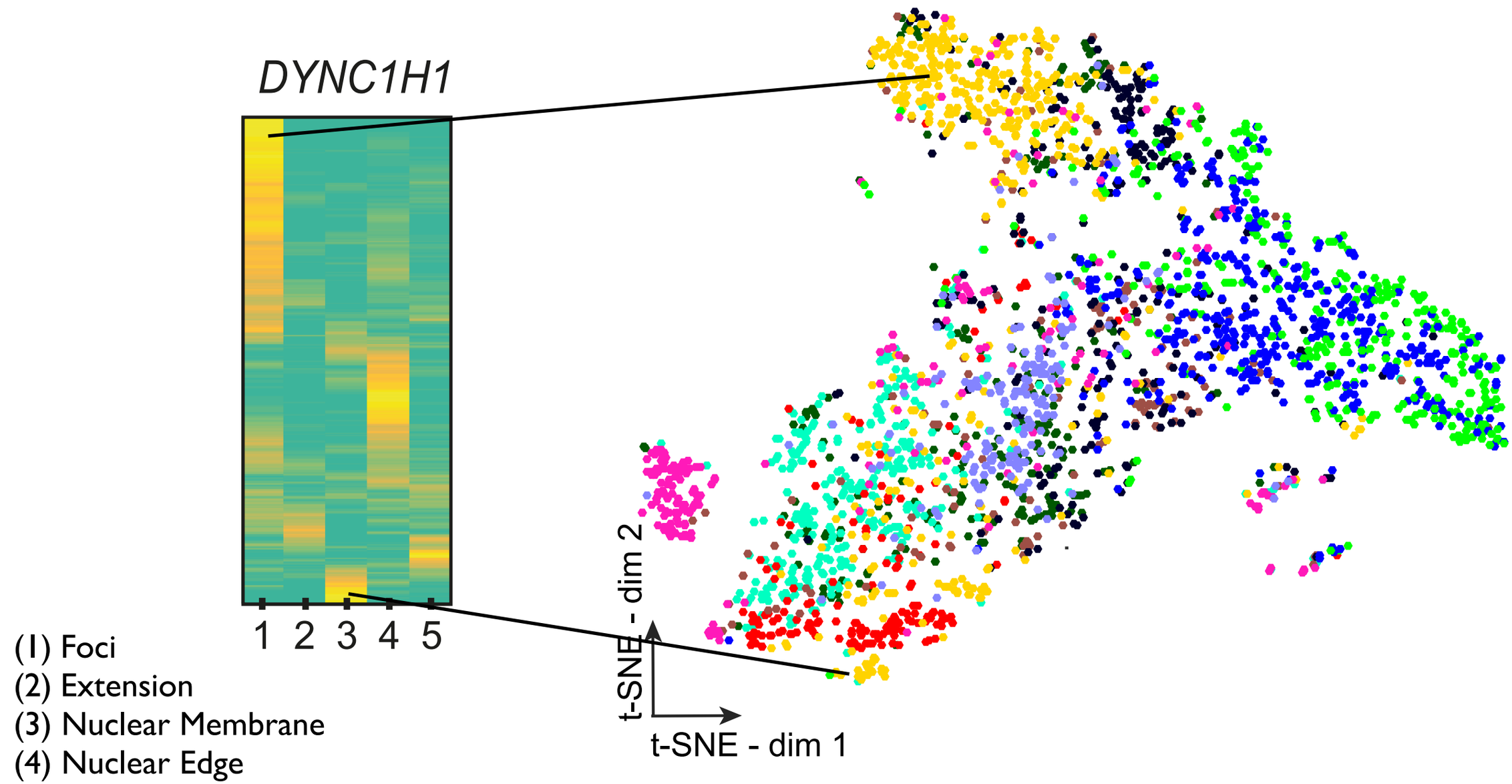
- Gene expression is a stochastic process, leading to heterogeneous expression levels.
- It is unknown to which extent this also holds for transcript localisation.



(1) Foci
(2) Extension
(3) Nuclear Membrane
(4) Nuclear Edge
(5) Random

Quantifying localisation heterogeneity

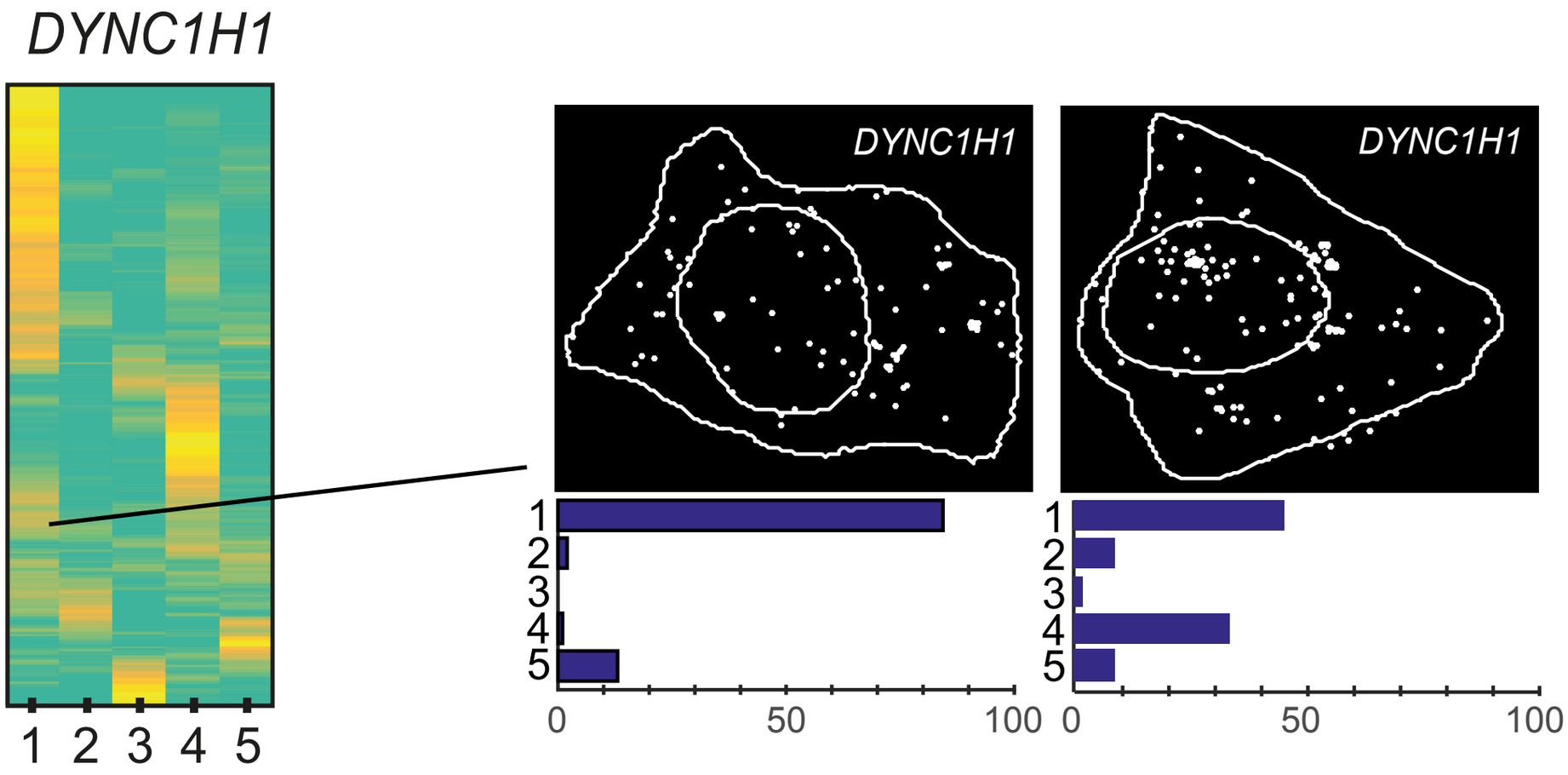
- Furthermore, we can distinguish two types of heterogeneity:
  - Population level heterogeneity corresponding to different subpopulations of cells with different localisation signatures.



(1) Foci
(2) Extension
(3) Nuclear Membrane
(4) Nuclear Edge
(5) Random

Quantifying localisation heterogeneity

- Furthermore, we can distinguish two types of heterogeneity:
  - Population level heterogeneity corresponding to different subpopulations of cells with different localisation signatures.
  - Intra-cellular heterogeneity corresponding to mixtures of pure patterns



(1) Foci
(2) Extension
(3) Nuclear Membrane
(4) Nuclear Edge
(5) Random

Quantifying localisation heterogeneity

- Furthermore, we can distinguish two types of heterogeneity:
  - Population level heterogeneity corresponding to different subpopulations of cells with different localisation signatures.
  - Intra-cellular heterogeneity corresponding to mixtures of pure patterns.
- We can quantify heterogeneity with the GINI impurity:

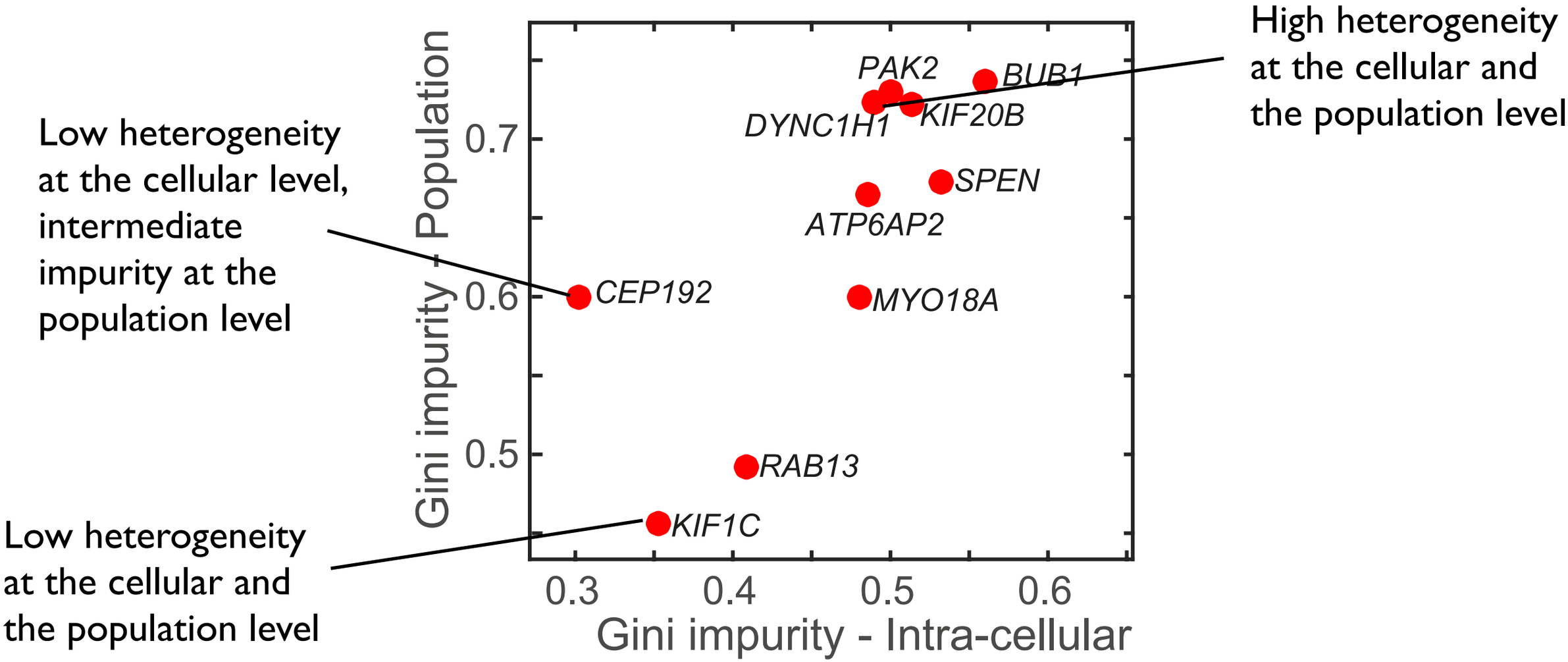
$$GINI(cell) = \sum_{i=1} p_i(1 - p_i)$$

$$GINI(population) = \sum_{i=1}^{K} q_i(1 - q_i)$$

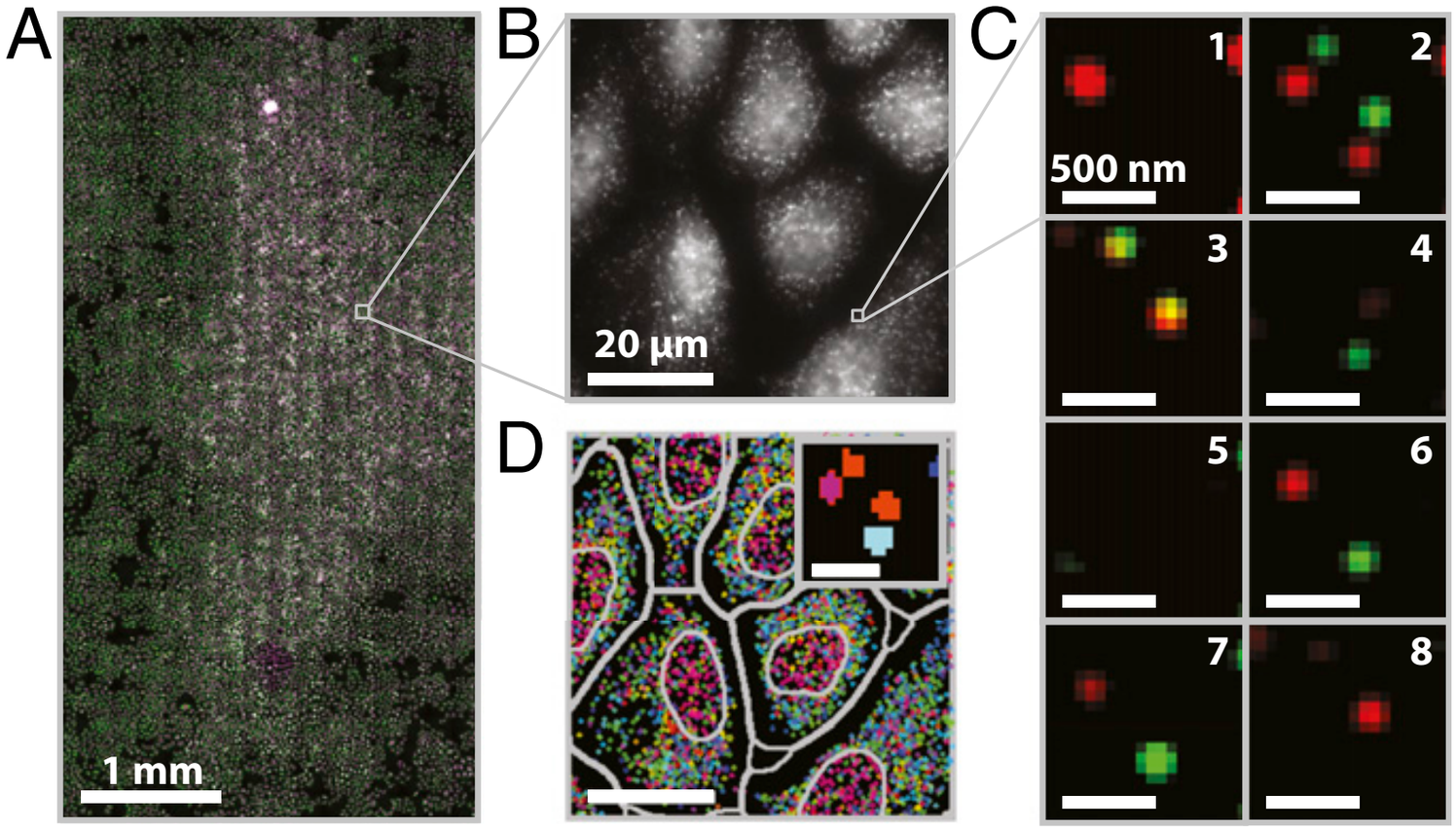
p_i : posterior probability of class *I*
q_i : fraction of samples classified as class *I*
K: number of classes

Quantifying localisation heterogeneity

- Furthermore, we can distinguish two types of heterogeneity:
  - Population level heterogeneity corresponding to different subpopulations of cells with different localisation signatures.
  - Intra-cellular heterogeneity corresponding to mixtures of pure patterns.
- We can quantify heterogeneity with the GINI impurity.



Next steps …



MERFISH

- 100.000 cells
- 1000 RNA species

Moffit et al., PNAS, 2016

Challenges:

- Visualization of stacks with hundreds of channels, projection techniques
- Analysis of co-expression: inference of regulation networks by causal inference

Conclusion

- Object tracking by machine learning.

- Machine learning for the recognition of localization patterns: application to spatial transcriptomics.

Acknowledgements

Institut Curie: CBIO
- Alice Schoenauer Sebag
- Peter Naylor
- Joseph Boyd
- Jean-Philippe Vert

Institut Curie: Reyal group
- Marick Lae
- Fabien Reyal

Institut Pasteur: Zimmer group
- Aubin Samatcoits
- Florian Müller
- Christoph Zimmer

IGMM: Bertrand group
- Nikolay Tsanov
- Racha Chouaib
- Marion Peter
- Edouard Bertrand

EMBL: Ellenberg group
- Jean-Karim Hériché
- Jan Ellenberg

EMBL: screening facility
- Beate Neumann
- Rainer Pepperkok

EMBL: Huber group
- Grégoire Pau
- Wolfgang Huber

CellCognition Tean
- Michael Held
- Christoph Sommer
- Rudolf Höfler
- Daniel Gerlich