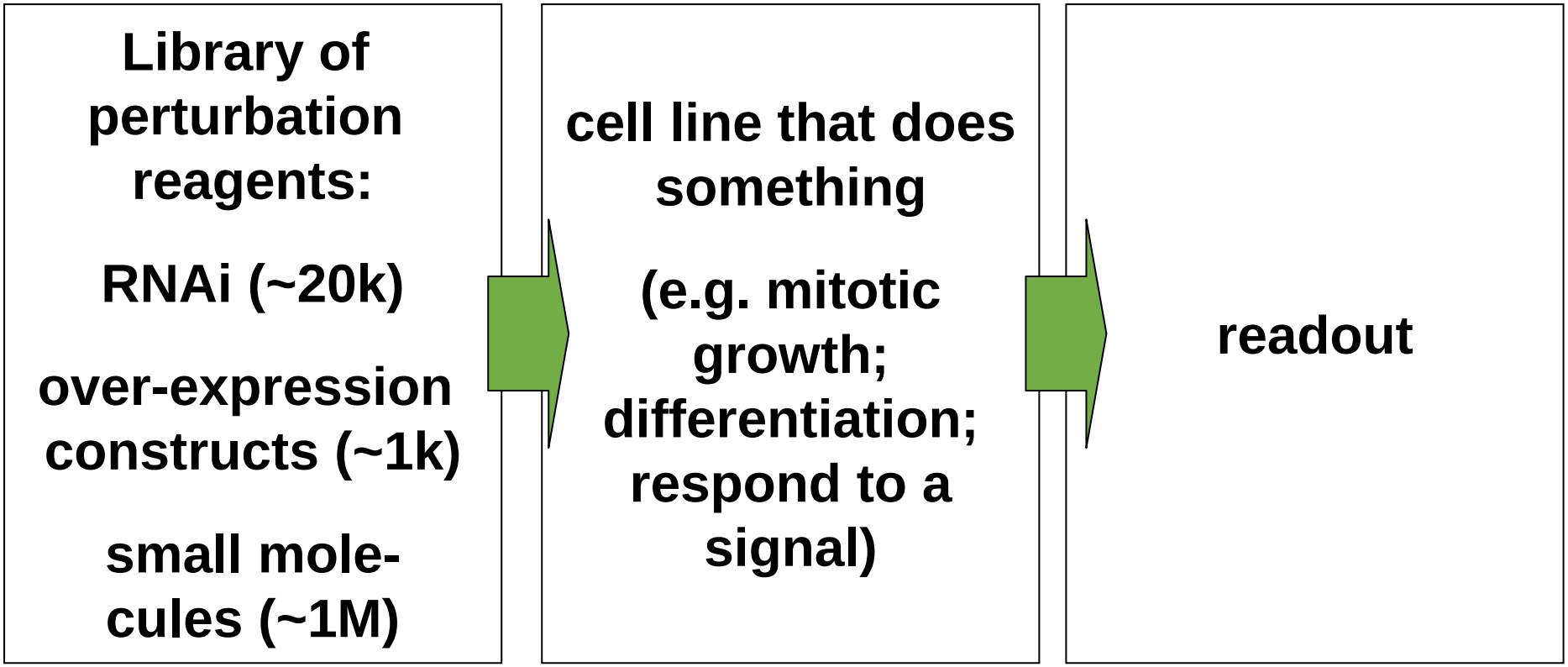
Image analysis and modelling of high-throughput cell based assays

Wolfgang Huber

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Cellular Phenotype Assays

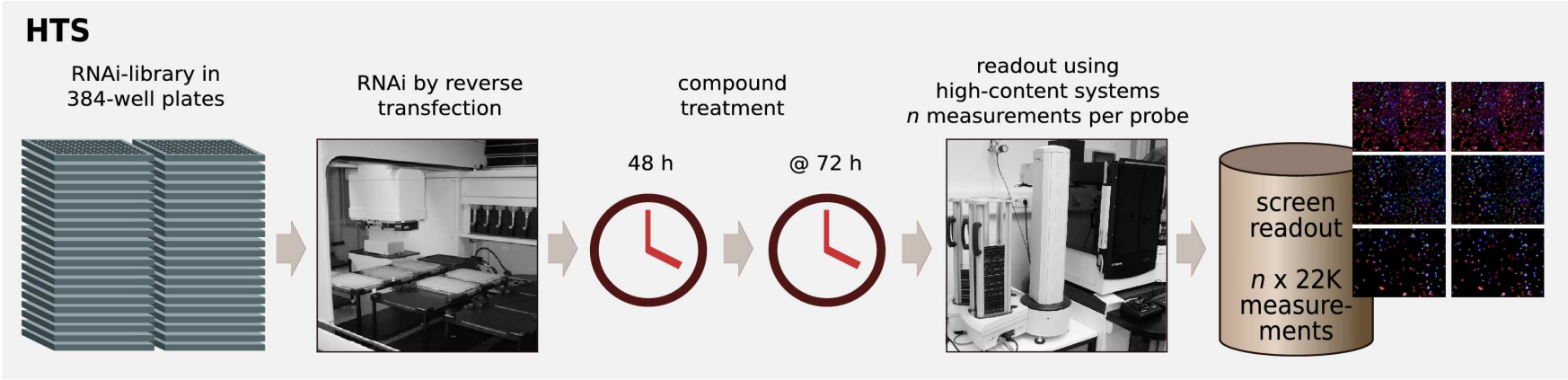


What is a phenotype? It all depends on the assay.

Any cellular process can be probed.
- (de-)activation of a signaling pathway
- cell differentiation
- changes in the cell cycle dynamics
- morphological changes
- activation of apoptosis
Similarly, for organisms (e.g. fly embryos, worms)

Phenotypes can be registered at various levels of detail
- yes/no alternative
- single quantitative variable
- tuple of quantitative variables
- image
- time course

High-throughput microscopy screening



quality control

image processing and analysis

Computational analysis

statistical analysis
determination of phenotypes

quality control

candidates

in vivo analysis

secondary assays

integration with biological databases

Genetic interactions

- in yeast, ~73% of genes are "non-essential"
  (Glaever et al. Nature 418 (2002))

- synthetic lethality phenotypes are prevalent (Tong et al. Science (2004))

- in drosophila, ~95% no viability phenotype
  (Boutros, Kiger, et al. Science 303 (2004))

- association studies for most human genetic diseases did not produce single loci with high penetrance

- evolutionary pressure for robustness

Two types of unspecificity effects

➔ **because the phenotype assay may lump together a number of different underlying mechanisms (e.g. viability assay)**

➔ **because the reagents are not as specific to their target as intended**

What are the implications for designing functional studies?

- need specific phenotypes: multiple assays, complex readout, over time
- use combinatorial perturbations (co-RNAi, small molecules, different genetic backgrounds)
- good preprocessing (normalisation/transformation, QA just as important as for μarrays)
- graph-type models to relate the data to gene-gene and gene-phenotype interactions, detect patterns and estimate modules

Monitoring tools

Plate reader

96 or 384 well, 1…4 measurements per well

FACS

4…8 measurements per cell, thousands of cells per well

Automated Microscopy

unlimited

Bioconductor packages for cell-based assays

cellHTS (Ligia Bras, M. Boutros)
genome-wide screens with scalar (or low-dimensional) read-out
data management, normalization, quality assessment, visualization,
hit scoring, reproducibility, publication
raw data → annotated hit list

prada (Florian Hahne); **flowCore, -Utils** et al. (B. Ellis, P. Haaland, N. Lemeur, F. Hahne)
flow cytometry
data management

EBImage (O. Sklyar)
image processing and analysis
construction of feature extraction workflows for large sets of similar images

cellHTS

Bioconductor package for the analysis of cell-based high-throughput screening (HTS) assays

Manage all data and metadata relevant for interpreting a cell-based screen

Data cleaning, preprocessing, primary statistical analysis

Raw data -> annotated hit list

Boutros, Bras, Huber. Analysis of cell-based RNAi screens. Genome Biology (2006)

The *cellHTS* package

per plate quality assessment

- **Dynamic range**
- **Distribution of the intensity values for each replicate**
- **Scatterplot between replicates and correlation coefficient**
- **Plate plots for individual replicates and for standard deviation between replicates**

per experiment quality assessment

- **Boxplots grouped by plate**
- **Distribution of the signal in the control wells, Z'-factor**

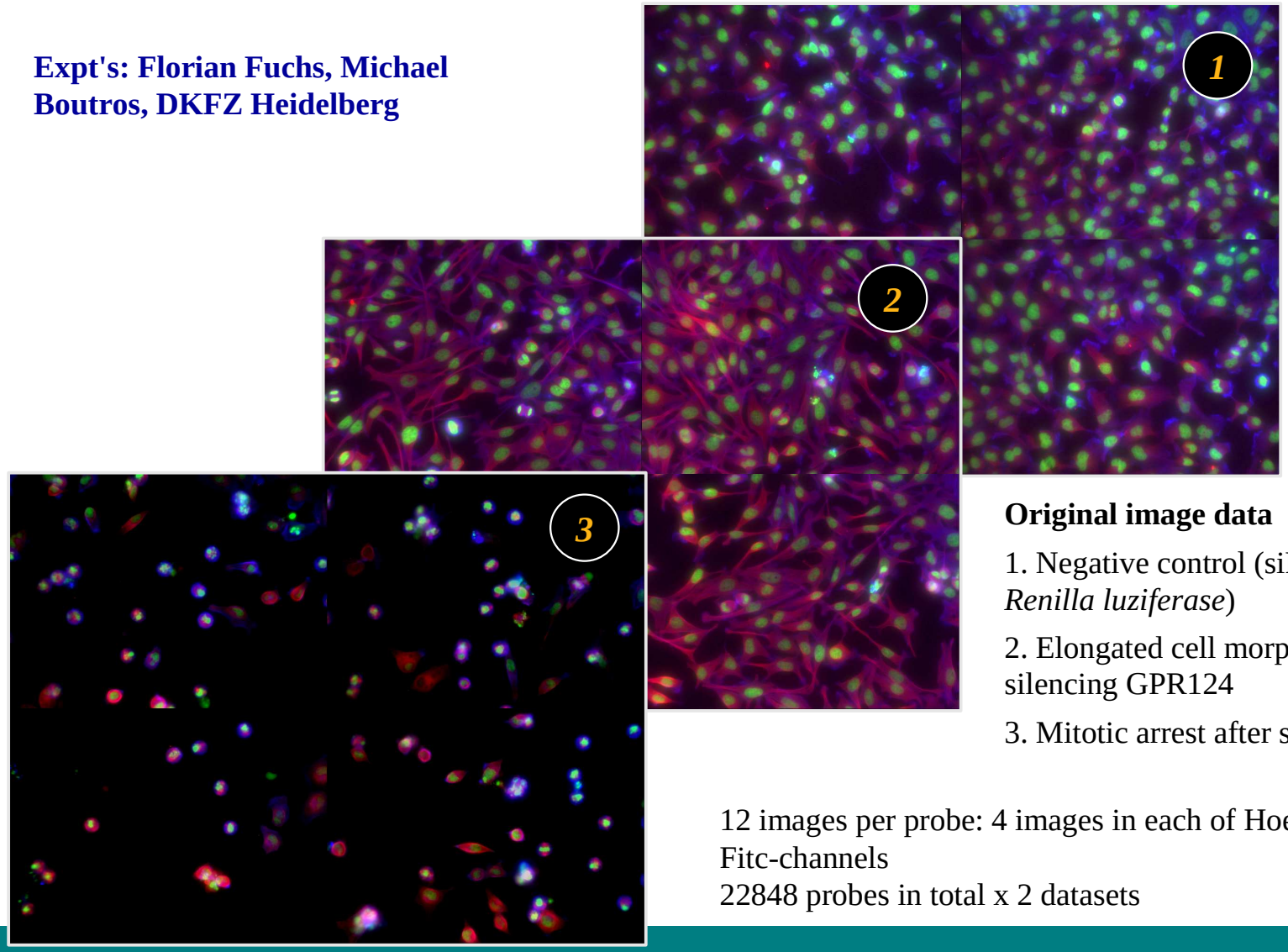
whole screen visualization

KcViab Analysis Report rendered in HTML

A genome-wide siRNA screen

on HEK293 cells to identify modulators of cell morphology (apoptosis, cell cycle, …)

Expt's: Florian Fuchs, Michael Boutros, DKFZ Heidelberg



Original image data

1. Negative control (siRNA against *Renilla luziferase*)
2. Elongated cell morphology after silencing GPR124
3. Mitotic arrest after silencing CDCA1

12 images per probe: 4 images in each of Hoechst-, Tritc- and Fitc-channels
22848 probes in total x 2 datasets

EBImage

- ***Image processing*** **and** ***analysis*** **on large sets of images in a programmatic fashion**
- **A package of R functions - to construct workflows that integrate statistic analysis and quality assessment, using a "real" modern language**
- **Number crunching uses C (easy to add your own C/C++ modules)**
- **Based on ImageMagick and other C/C++ image processing libraries**
- **Free and open source (LGPL), distributed with Bioconductor**

Collaboration with Michael Boutros, Florian Fuchs (DKFZ)

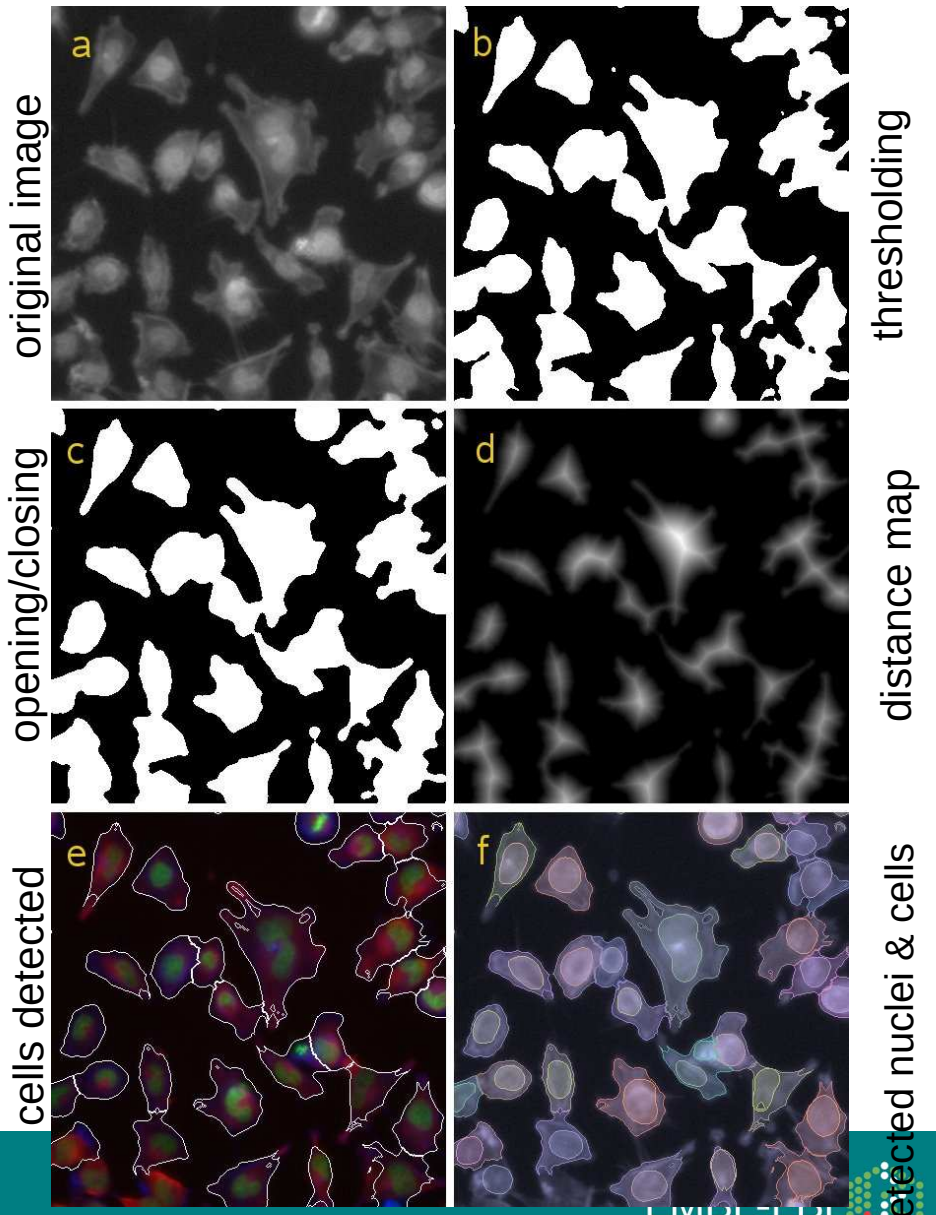


Image processing with R: simple operations

I/O

```
files = c(“im1.tif”, “im2.tif”)
im = read.image(files)
```

Subsetting

```
w = dim(im)[1]/2 - 1
h = dim(im)[2]/2 - 1
r1 = im[1:w, 1:h, ]
w1 = r1[,, 1]
```

Image stacks

```
combine(w1, r1[,,2], r1[,,3])
```

Logical indexing

```
x[ x > 0.5 & w1 > 0.7 ] = 1
```

Colour channels, greyscale

```
ch1 = channel(w1, “asred”)
ch2 = channel(res[,,2], “asgreen”)
ch3 = channel(res[,,3], “asblue”)
rgb = ch1 + ch2 + ch3
```

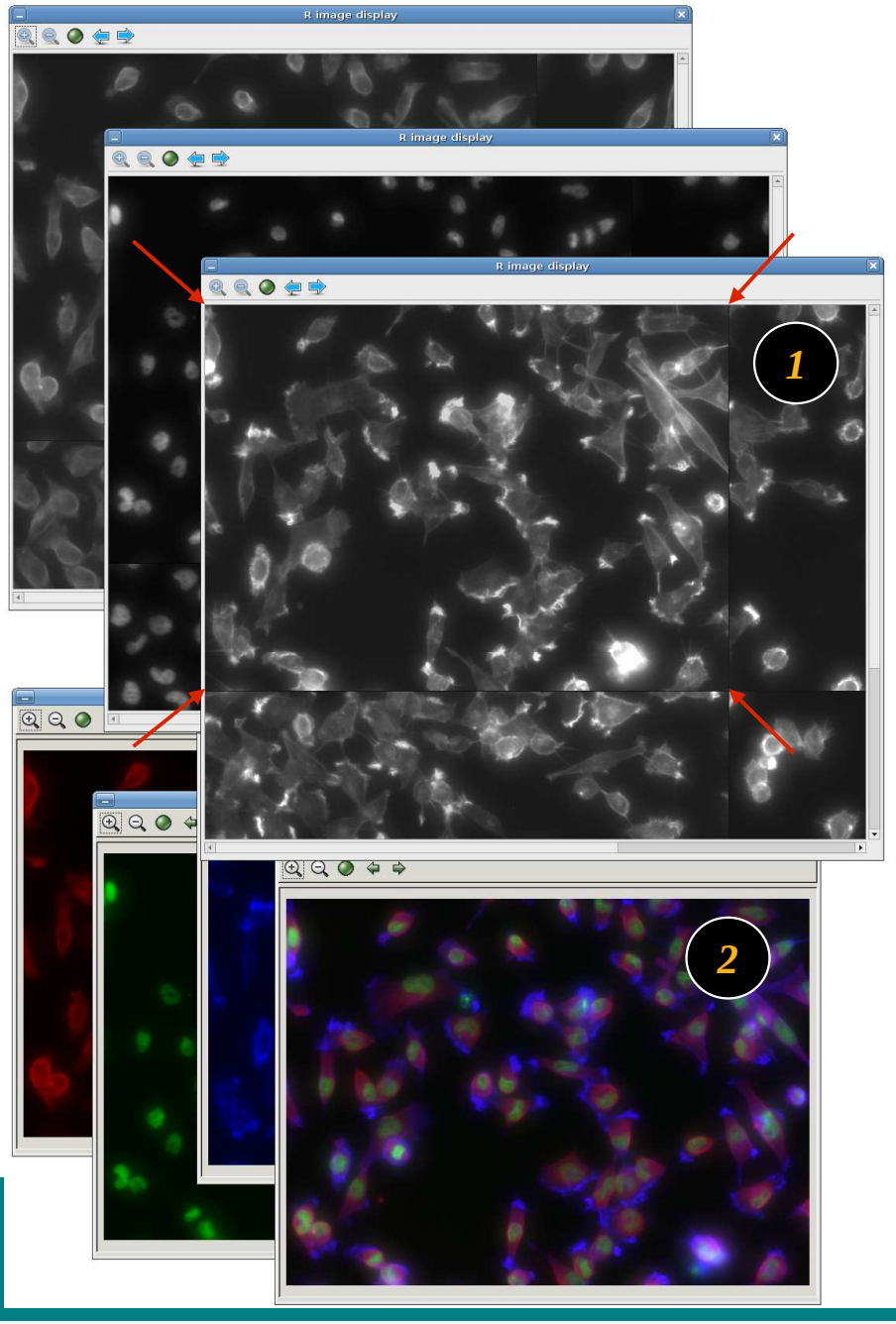
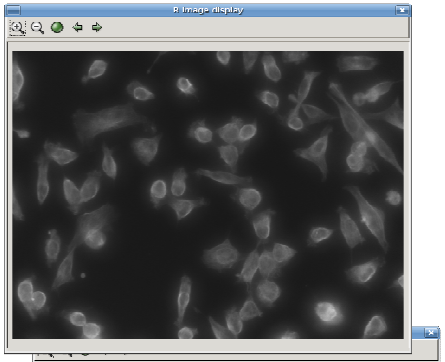
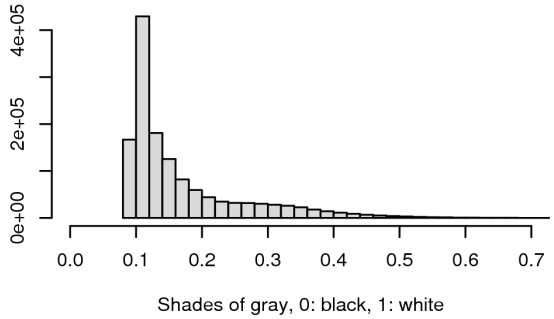
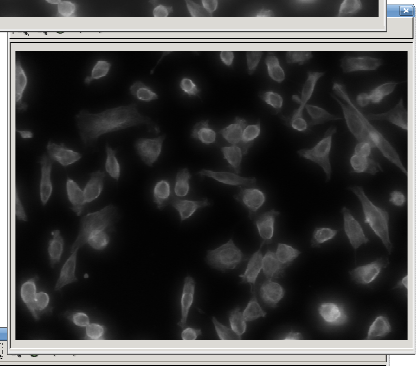


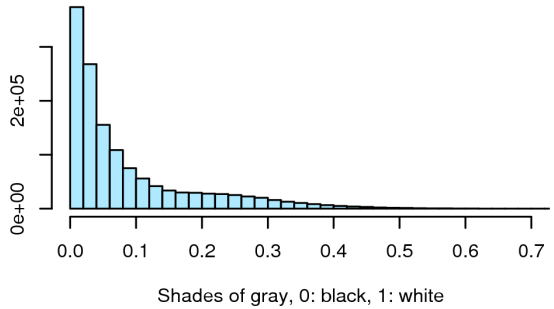
Image processing: arithmetic and visualization



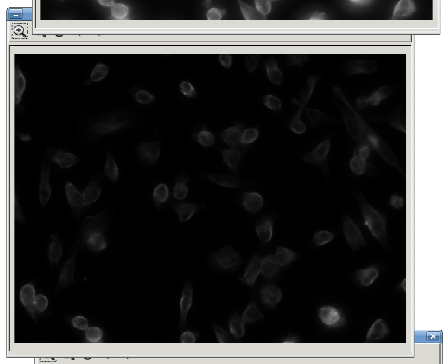


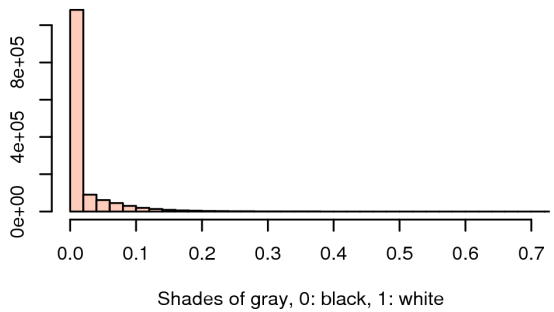

```
display(x)
hist(x, xlim=c(0,.7), col="gray")

nx = (x-min(x))/diff(range(x))
```

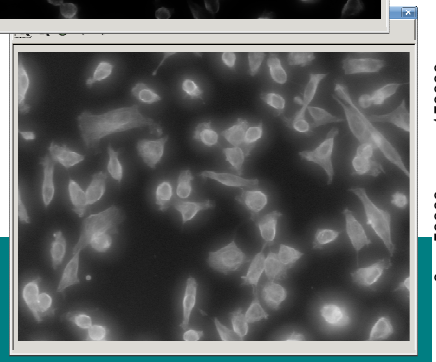





```
## naïve high pass filter
fx = fft(x)
fx[ 1:10, 1:10 ] = 0
x1 = normalize(Re(fft(fx, inv=TRUE)))
```







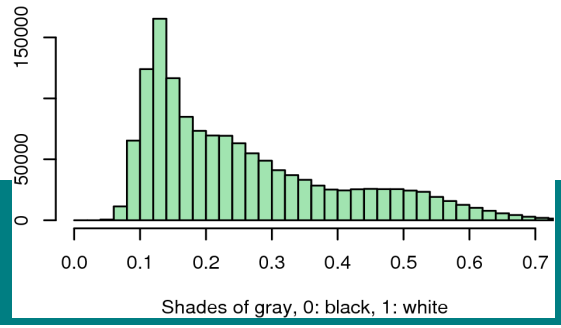
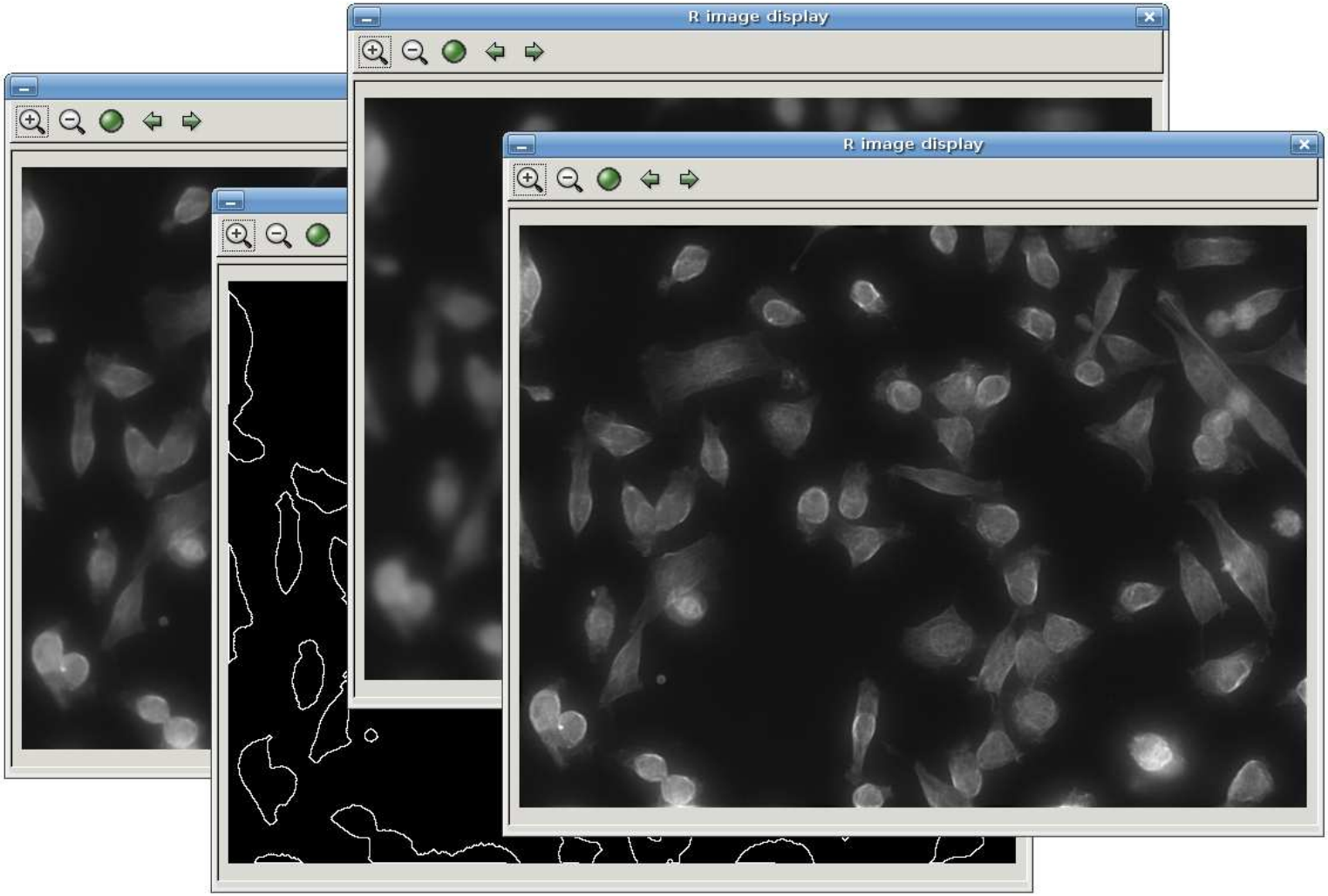


Image processing: filters from *ImageMagick*

```
display( x             )
display( edge(x, 1)    )
display( blur(x, 6, 2) )
display( sharpen(x)    )
```

others

```
normalize2
enhance
contrast
cgamma

denoise
despeckle
umask
mediansmooth

resize
resample
flip
flop
rotate

segment
athresh
cthresh

modulate
negate

etc
```



Basic tools for segmentation

Locally adaptive thresholding

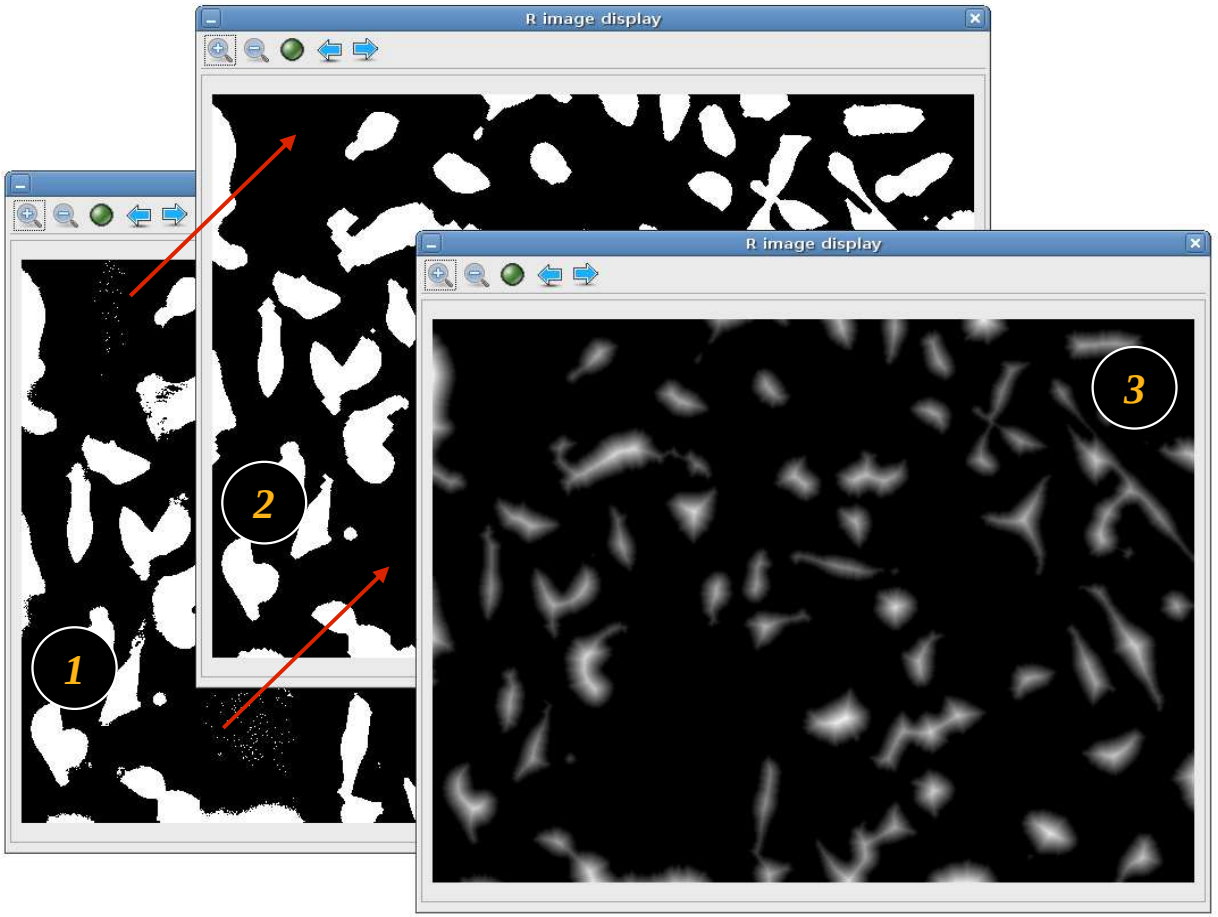
Mathematical Morphology

Distance map transformation

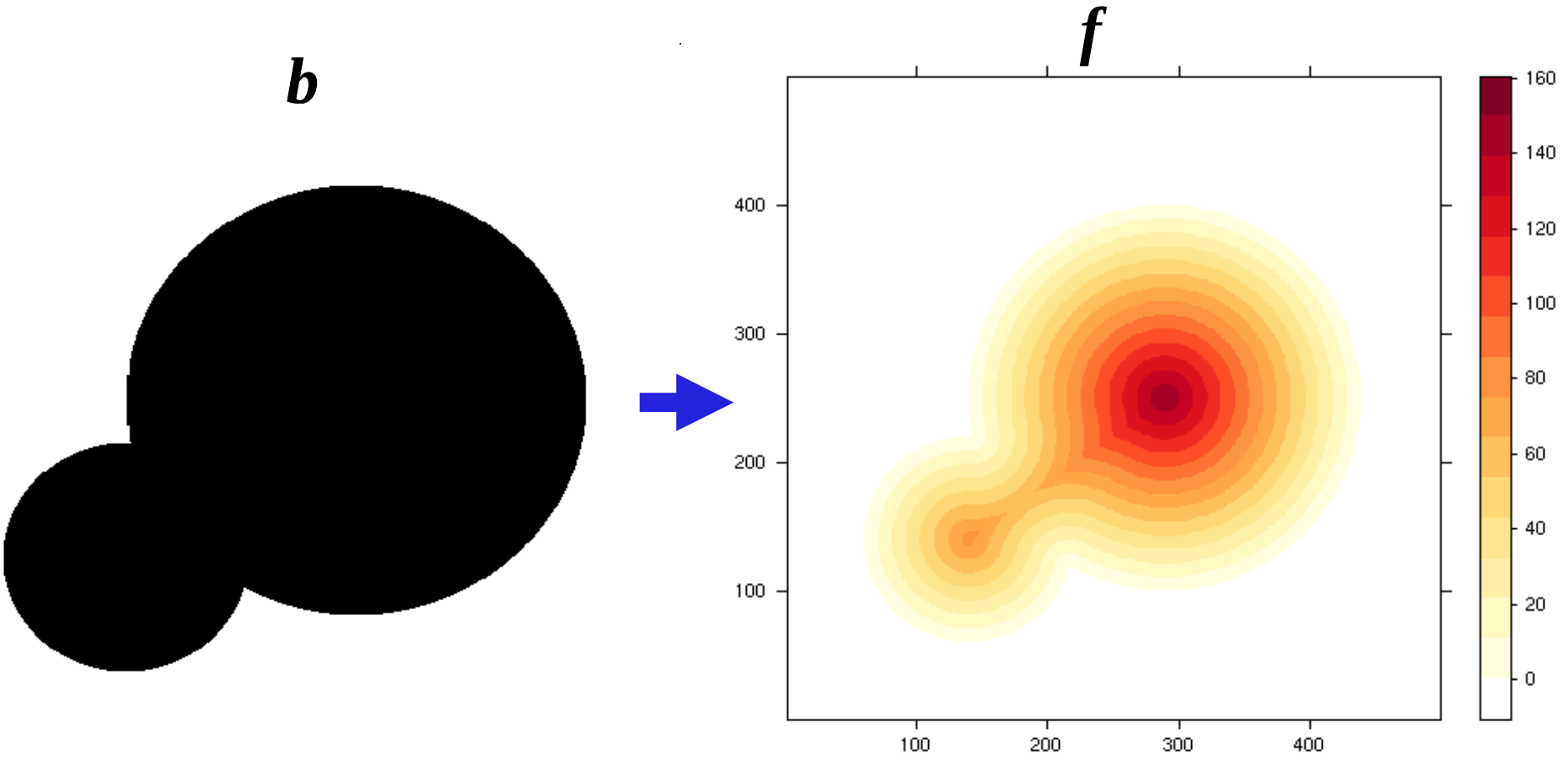
binary image -> greyscale

each pixel is given the value of its distance to the nearest background pixel

1.
```
t = thresh(w0, 40, 40, 0.001)
mask = closing(t, morphKern(5))
```
2.
```
mask = opening(mask, morphKern(5))
```
3.
```
dm = distmap(mask)
range(dm)
[1] 0 87
```

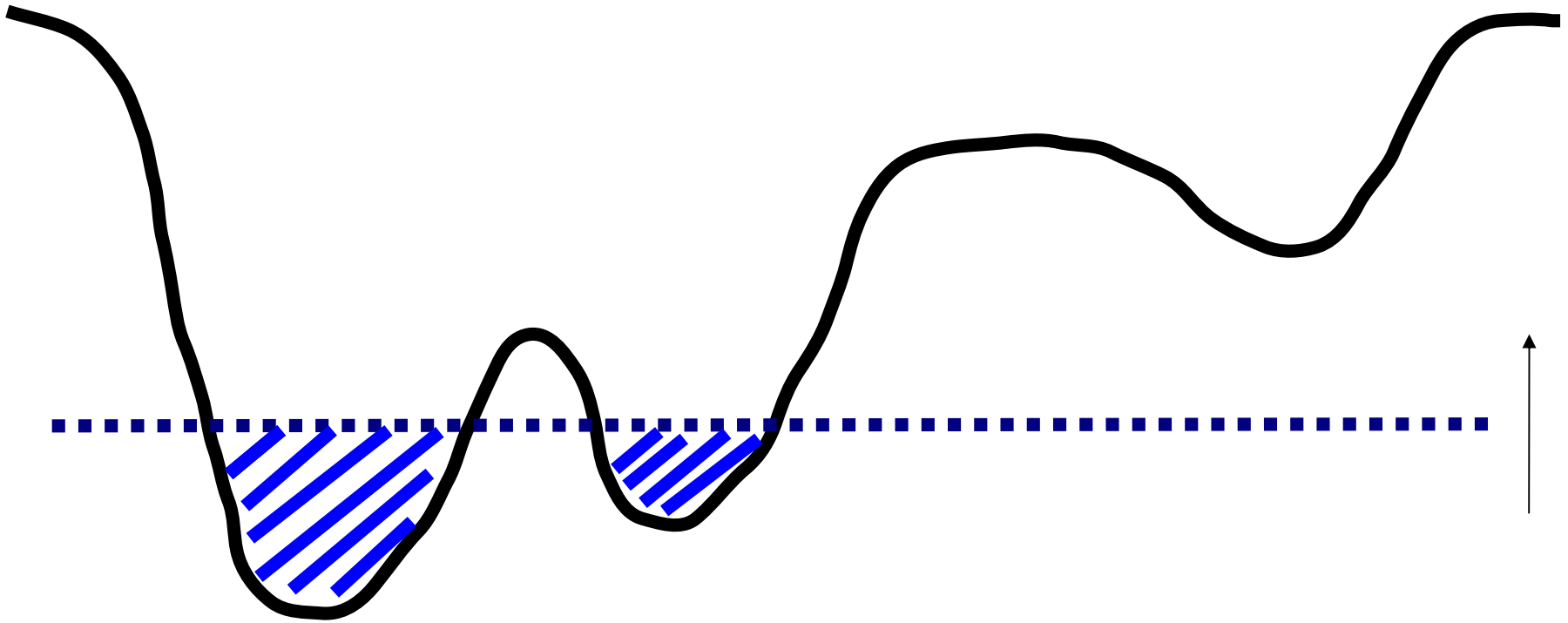


Distance map transformation



$$f(\vec{x}) = min\{d(\vec{x}', \vec{x}) \mid b(\vec{x}') = 0\}$$

Watershed segmentation



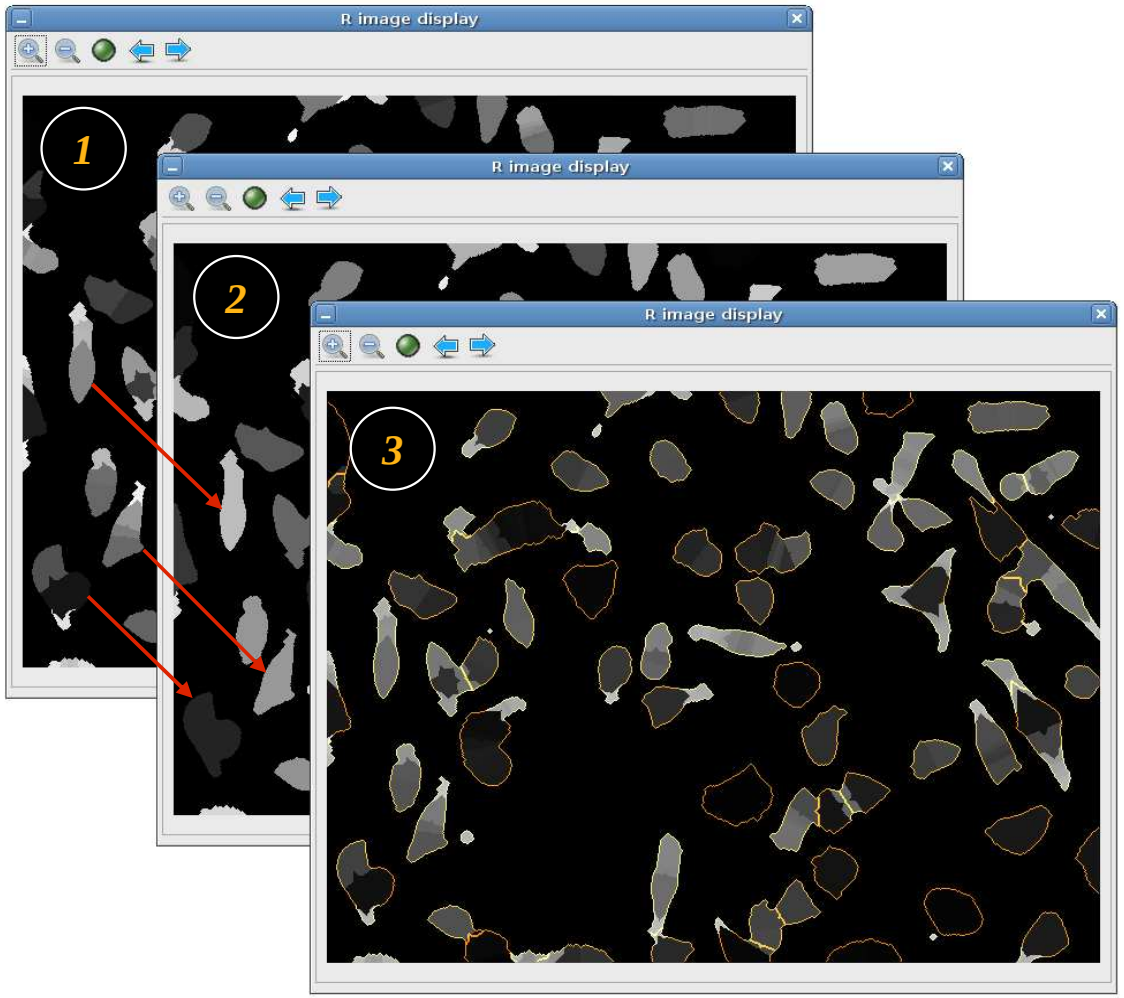
distance map/ watershed segmentation **can be very effective, but...:**

- susceptible to spurious local minima

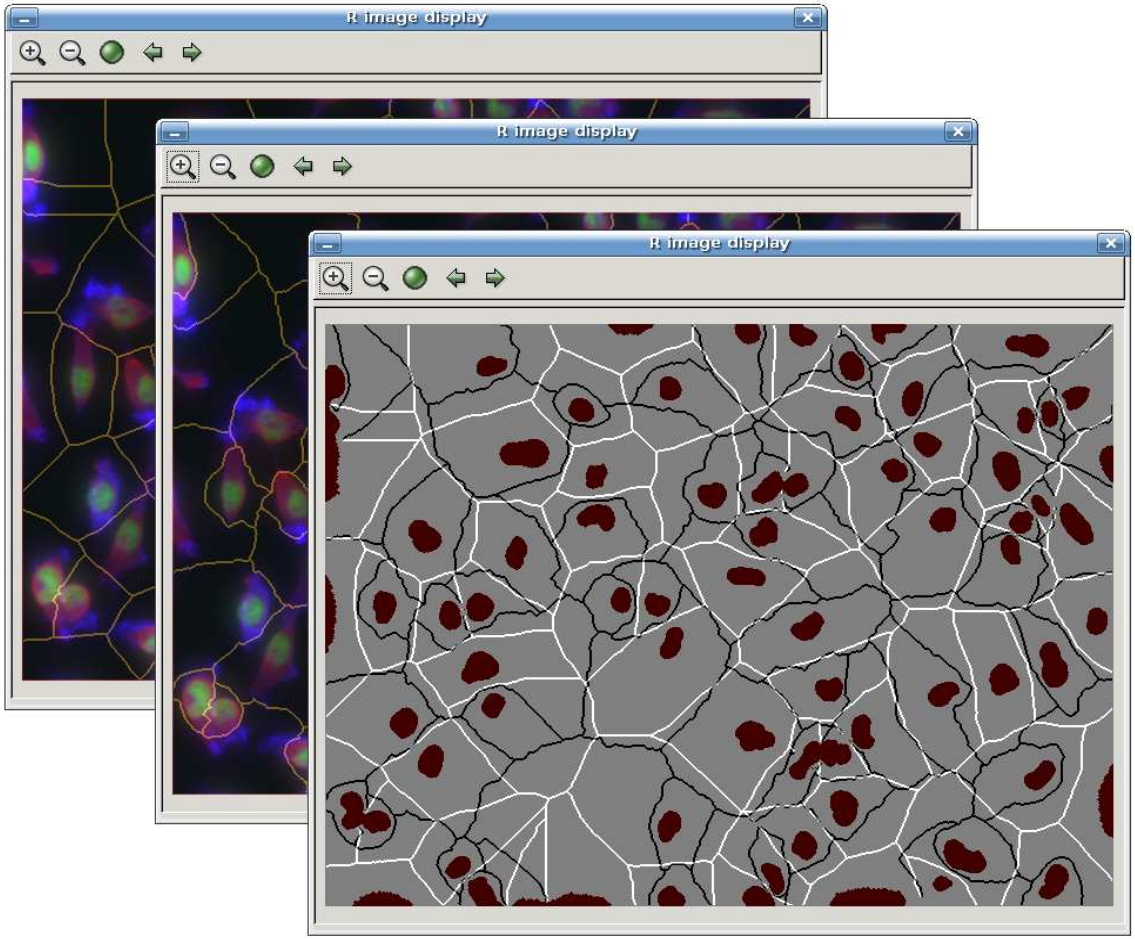
- potentially unstable around flat ridges

- does not use shape or distance criteria

1. `w1 = watershed(dm, 0, 1)`
   `range(w1)`
   `[1] 0 189`
2. `w2 = watershed(dm, 2, 1)`
   `range(w2)`
   `[1] 0 61`
3. `x = paintObjects(w2,`
   `channel(w1, “rgb”))`



Voronoi diagrams



partitioning of a plane with *n* convex seed sets into *n* convex polygons

such that each polygon contains only one seed

and every point in a polygon is closer to its seed than to any other

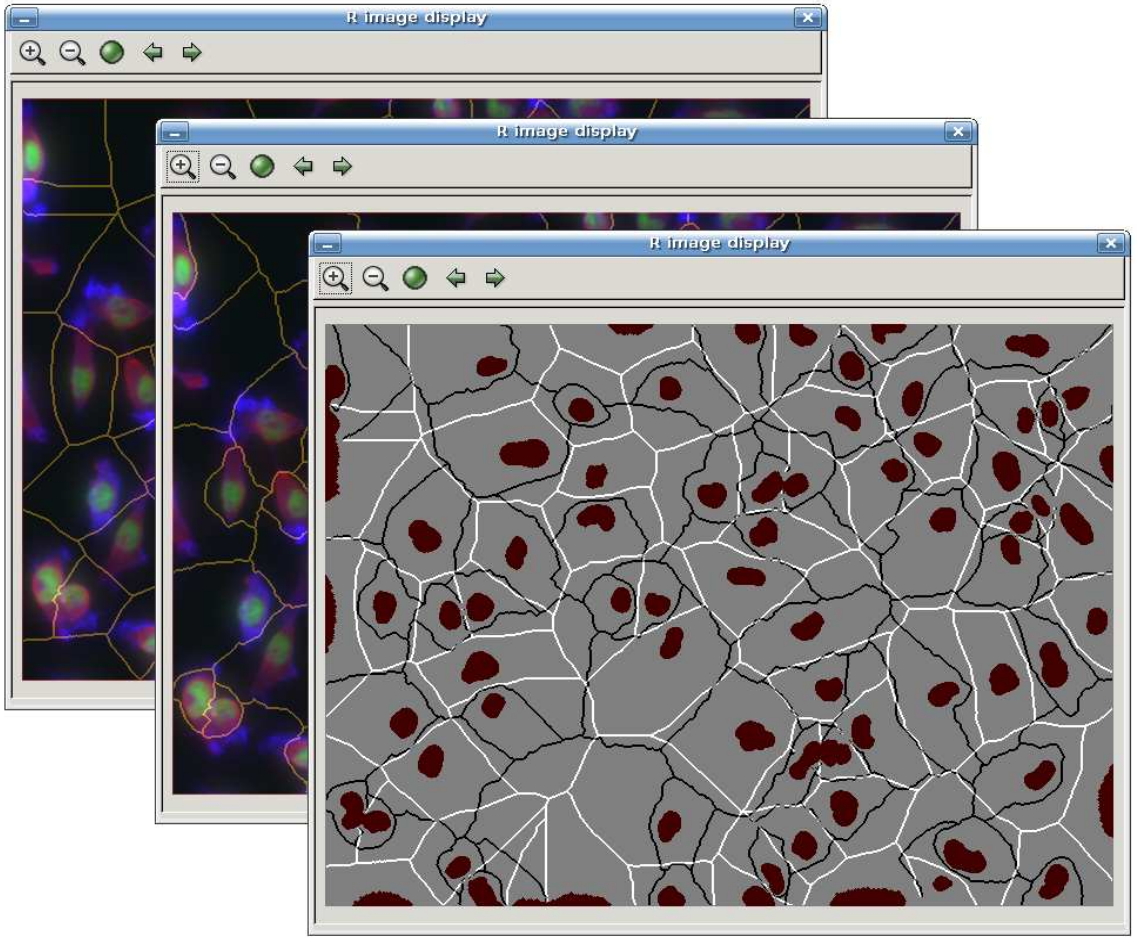
Example:

segment nuclei (easy)

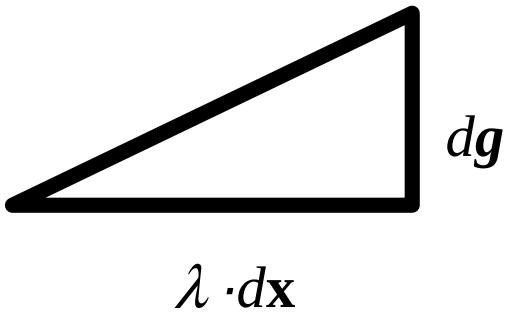
use them as seed points

Voronoi sets: estimates of cell shapes

Voronoi diagrams on image manifolds



Instead of Euclidean distance in (x,y)-plane, use geodesic distance on the image manifold

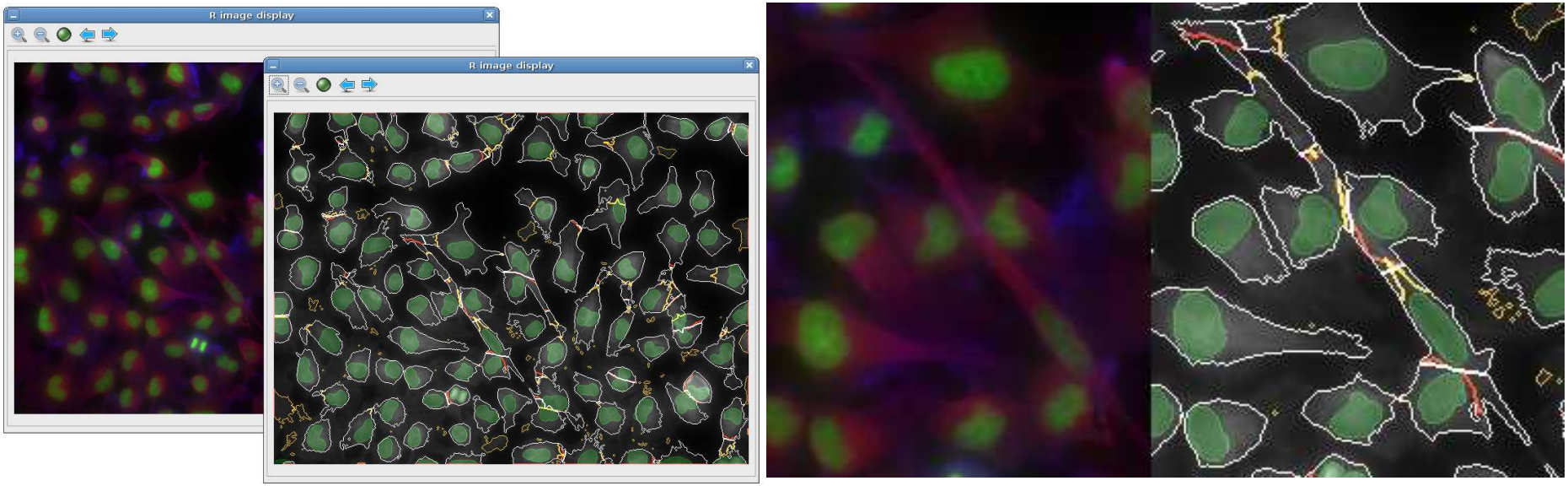


$$\mathbf{G} = \frac{\nabla \mathbf{g}(\mathcal{I}) \nabla \mathbf{g}^T(\mathcal{I}) + \lambda \mathbf{I}}{1 + \lambda}$$

$$||d\mathbf{x}||^2_{\mathbf{G}} \equiv d\mathbf{x}^T \mathbf{G} d\mathbf{x} = \frac{(d\mathbf{x}^T \nabla \mathbf{g}(\mathcal{I}))^2 + \lambda (d\mathbf{x}^T d\mathbf{x})^2}{\lambda + 1}$$

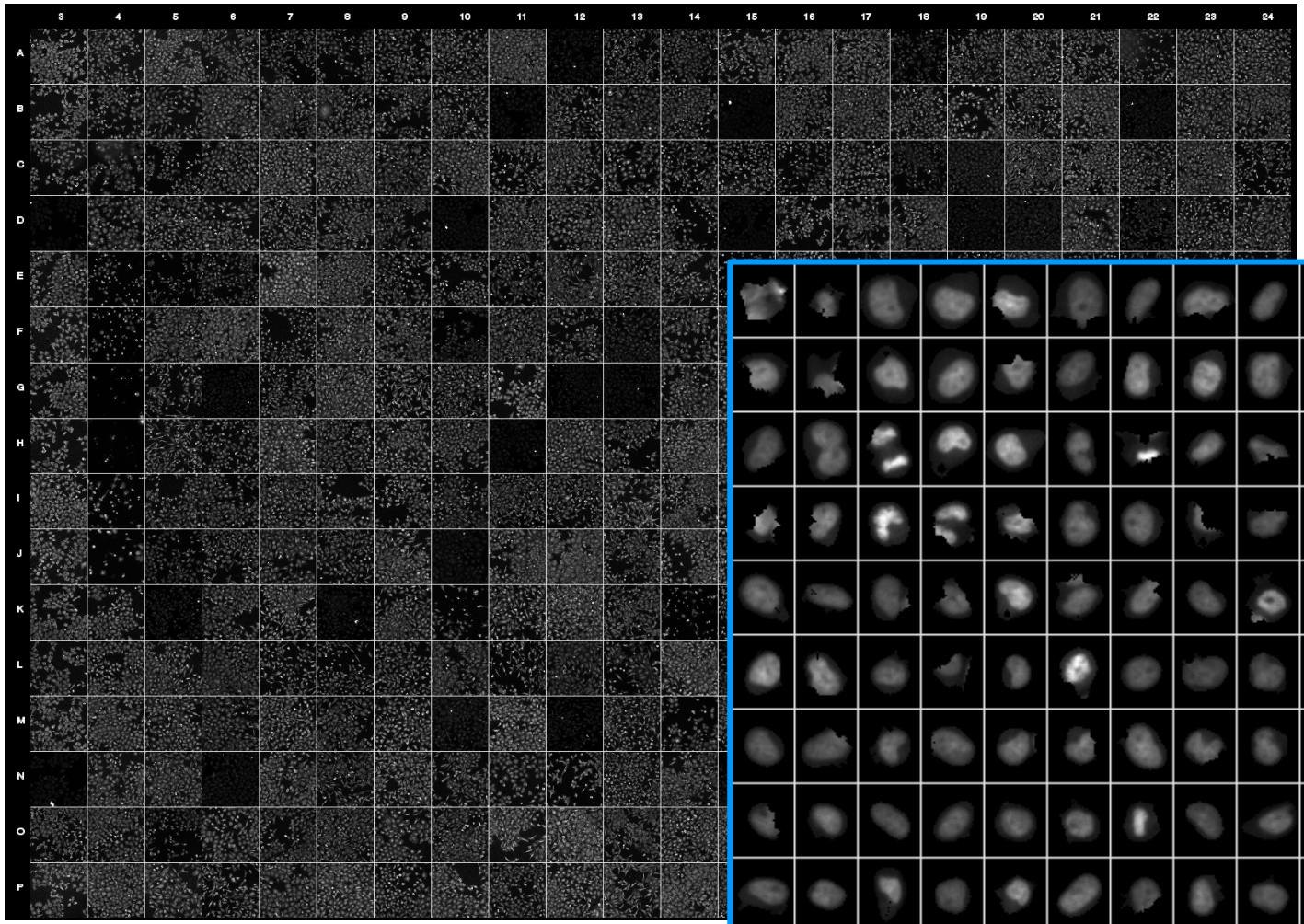
T. Jones, A. Carpenter et al.: CellProfiler

Voronoi diagrams on image manifolds

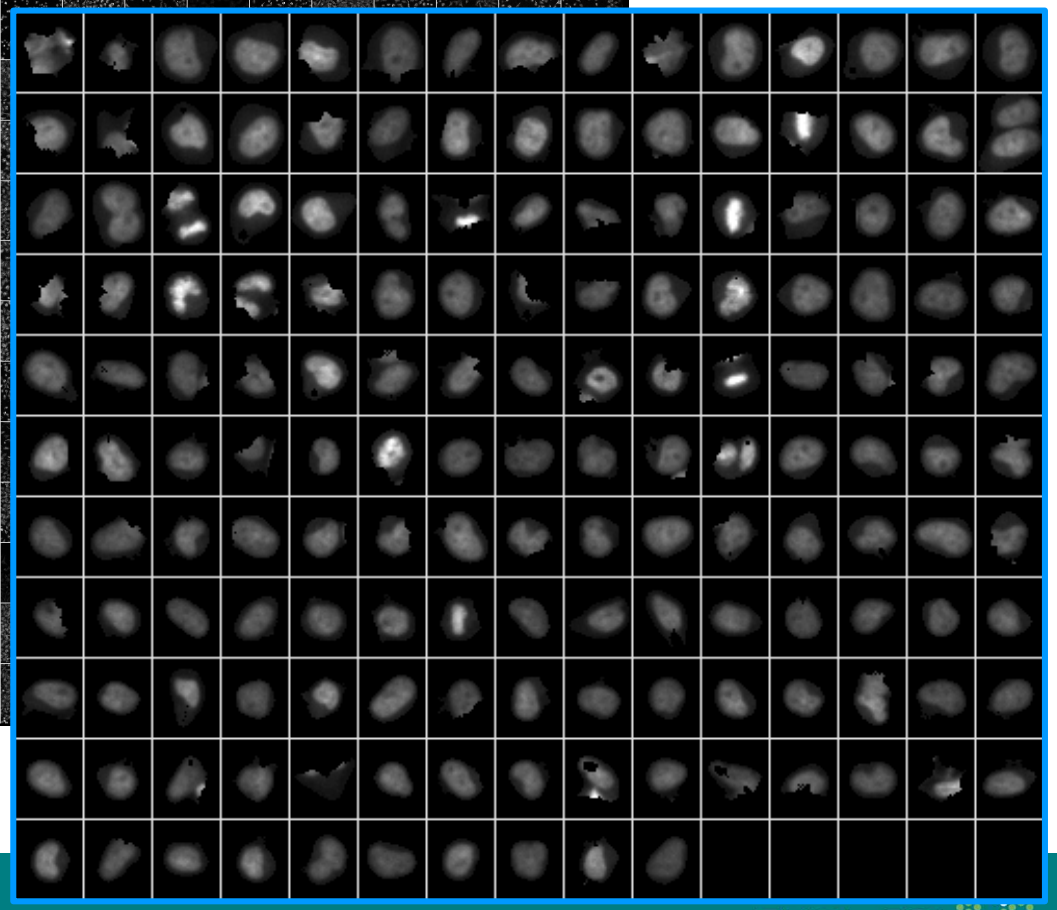



```
dm = distmap( thresh(nucl, 30, 30) )
seeds = watershed(dm, 1, 1)
mask = thresh(cell, 60, 60)
w  = watershed(distmap(mask), 2, 1)            ## yellow
vi = propagate(cell, seeds, mask, lambda=0)    ## red
v  = propagate(cell, seeds, mask, lambda=2e16) ## white
```

Some visualisation before we continue with the analysis



Thumbnail overview of one plate's images



Gallery view of segmented objects of one well

EMBL-EBI

Object features

number of objects

Generic

Moments: area, mass (=intensity), center of mass, elements of the covariance matrix and its eigenvalues, rotation angle, Hu's 7 rotation invariants

Haralick texture features

Zernike rotation invariant moments

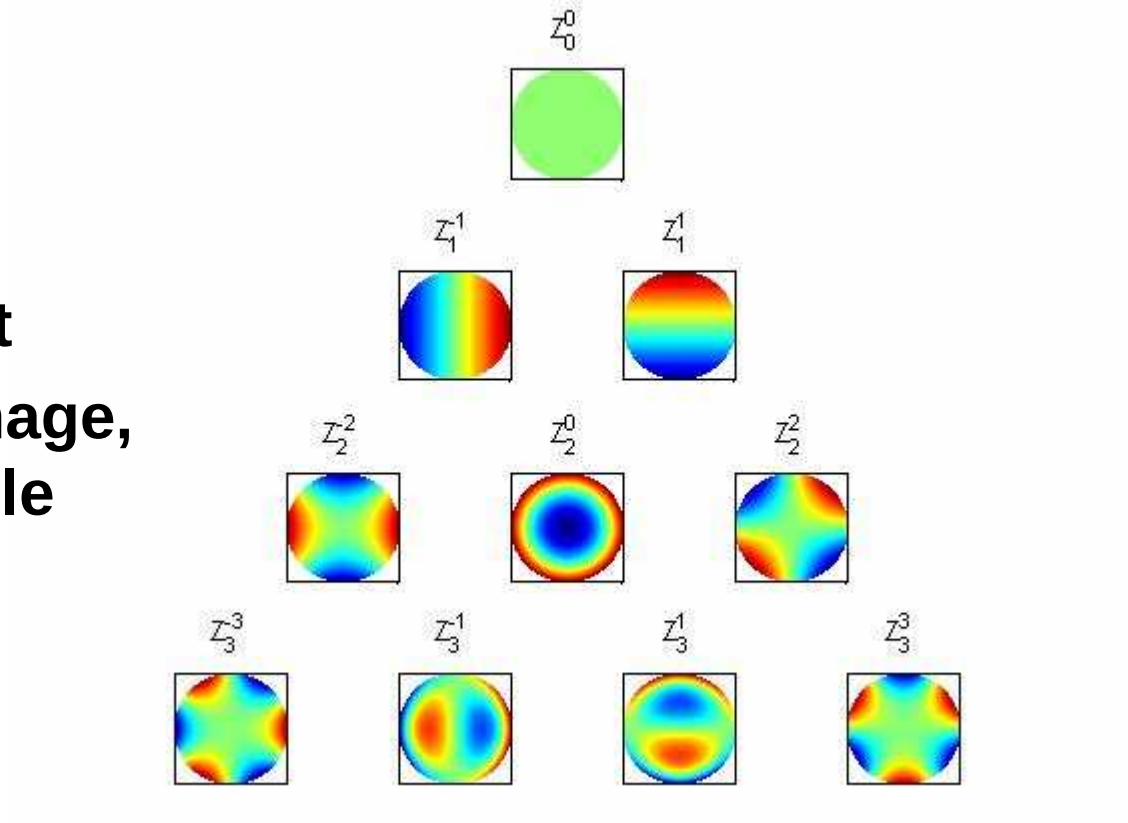
Application-adapted

measures of acircularity or relative overlap between different stain channels

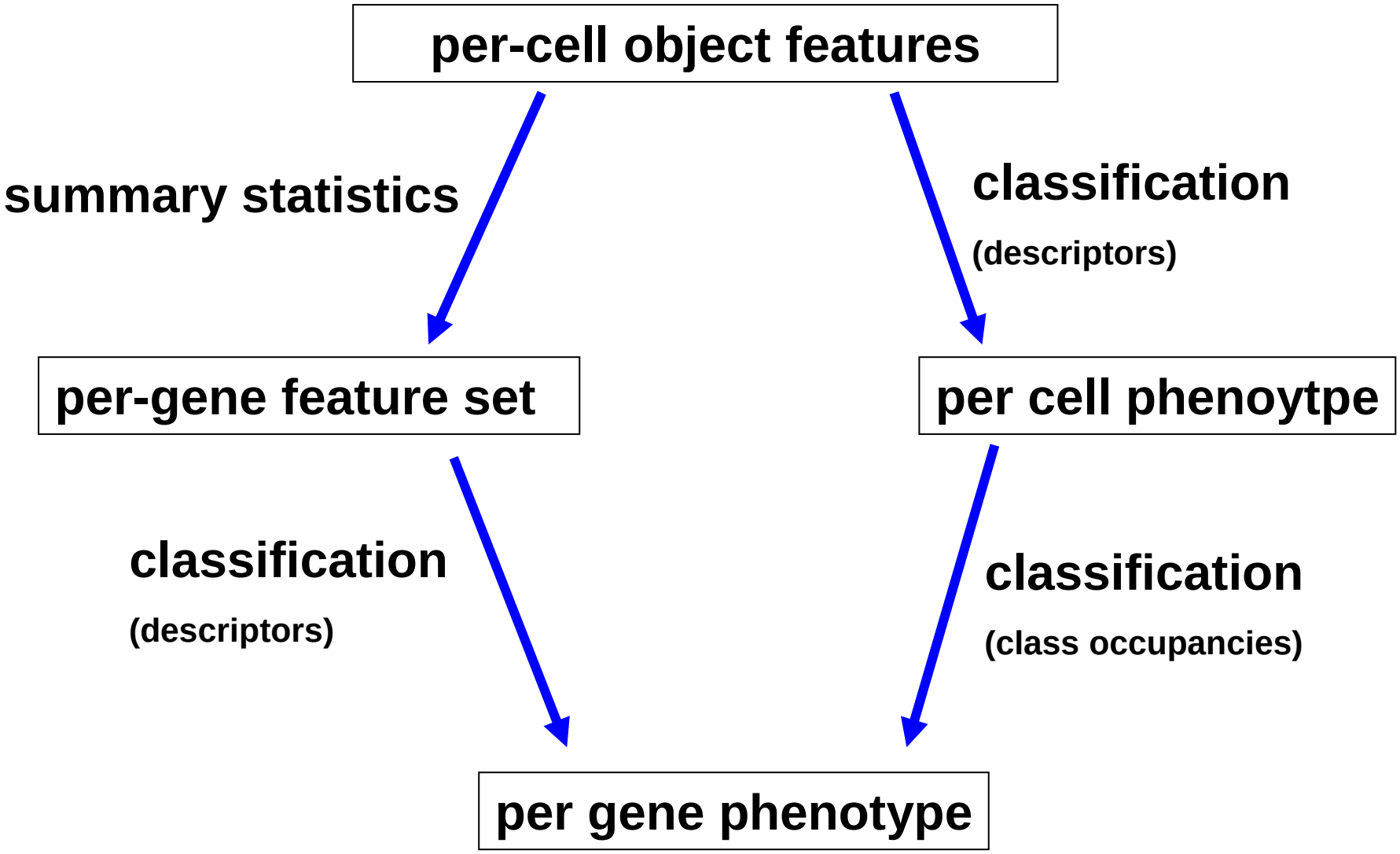
Zernike Moments

$$A_{mn} = \frac{m+1}{\pi} \int_{\text{unit circle}} e^{-in\theta} Z_{mn}(r,\theta)\, f(r,\theta)\, d\theta dr$$

- **|n|<=m, m-|n| even**
- **$|A_{mn}|$ rotation invariant**
- **careful: f a discrete image, pixelisation of the circle**



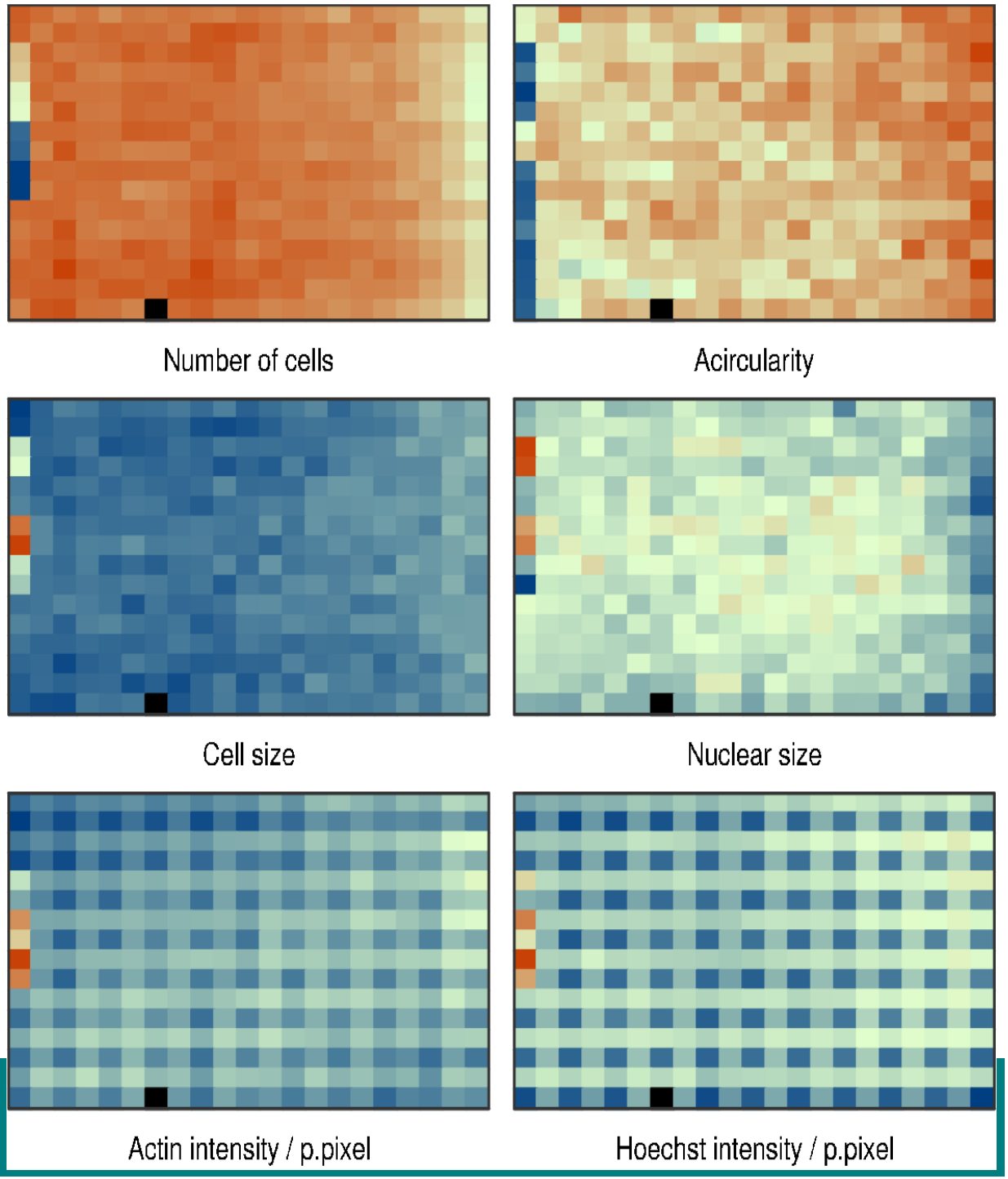
From object features to phenotypes



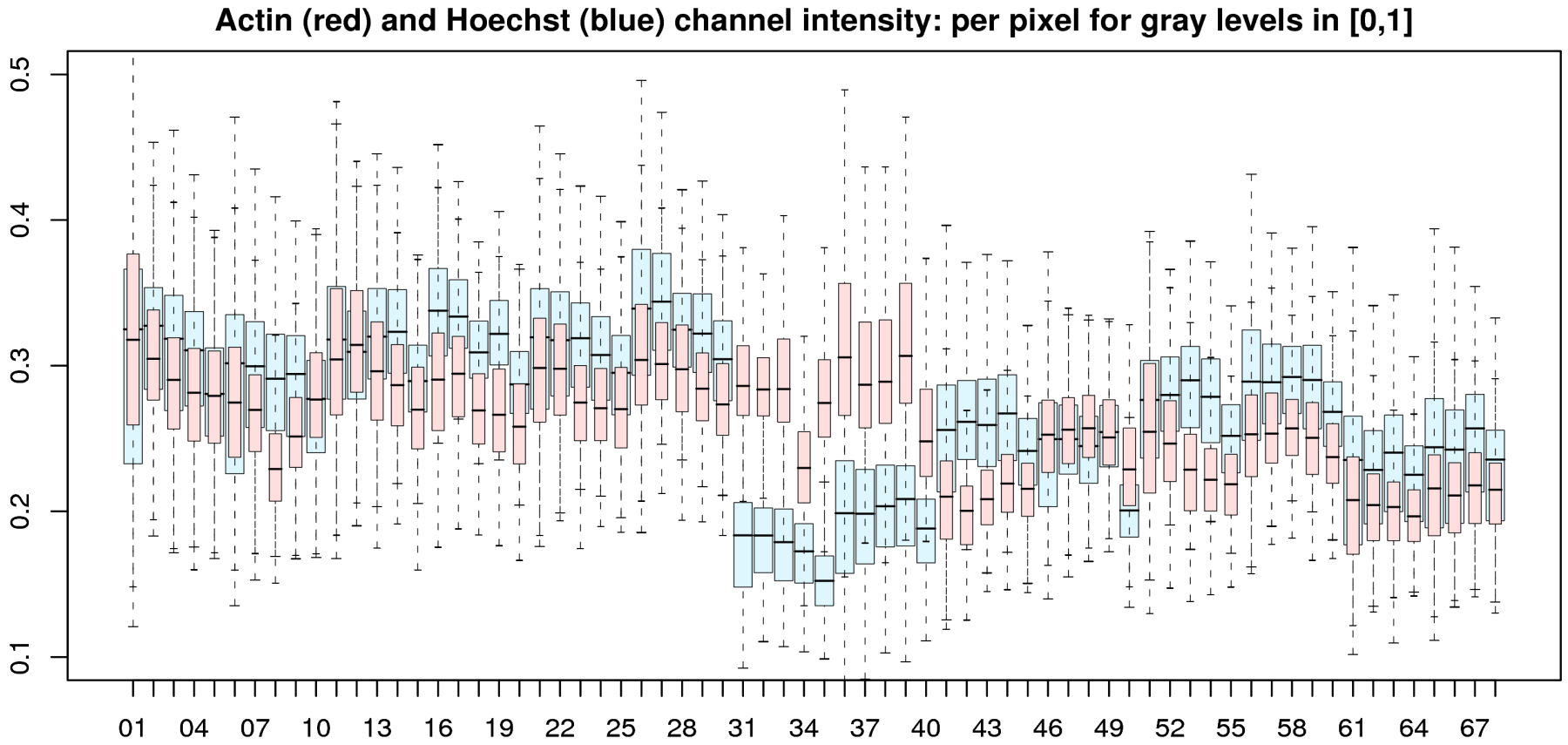
Back to reality:

within plate spatial trends -

normalization and quality assessment



Batch effects



Normalization: Plate effects

Percent of control

$$x'_{ki} = \frac{x_{ki}}{\mu_i^{pos}} \times 100$$

k-th well
i-th plate

Normalized percent inhibition

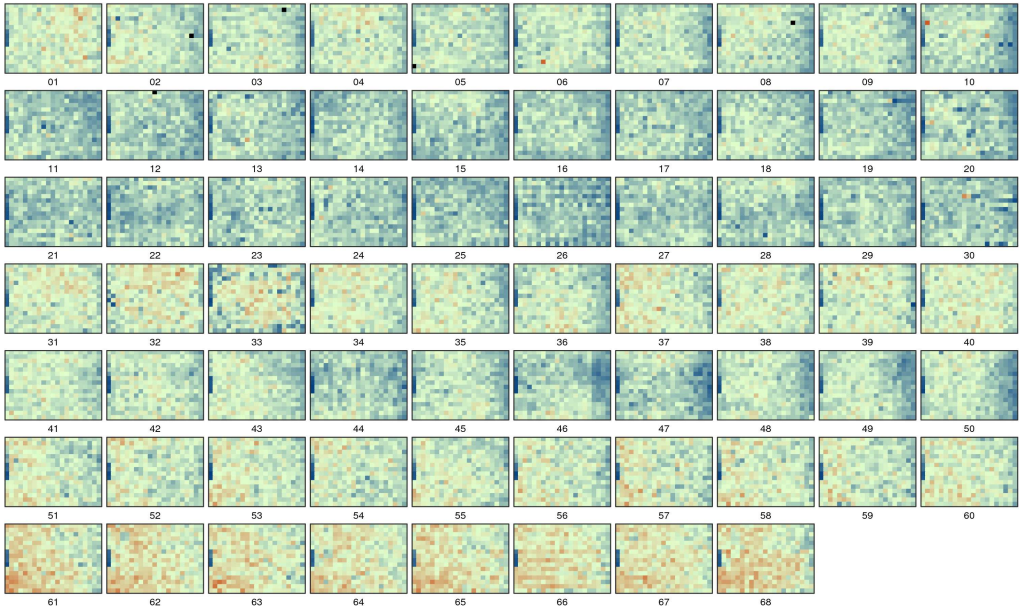
$$x'_{ki} = \frac{\mu_i^{pos} - x_{ki}}{\mu_i^{pos} - \mu_i^{neg}} \times 100$$

***z*-score**

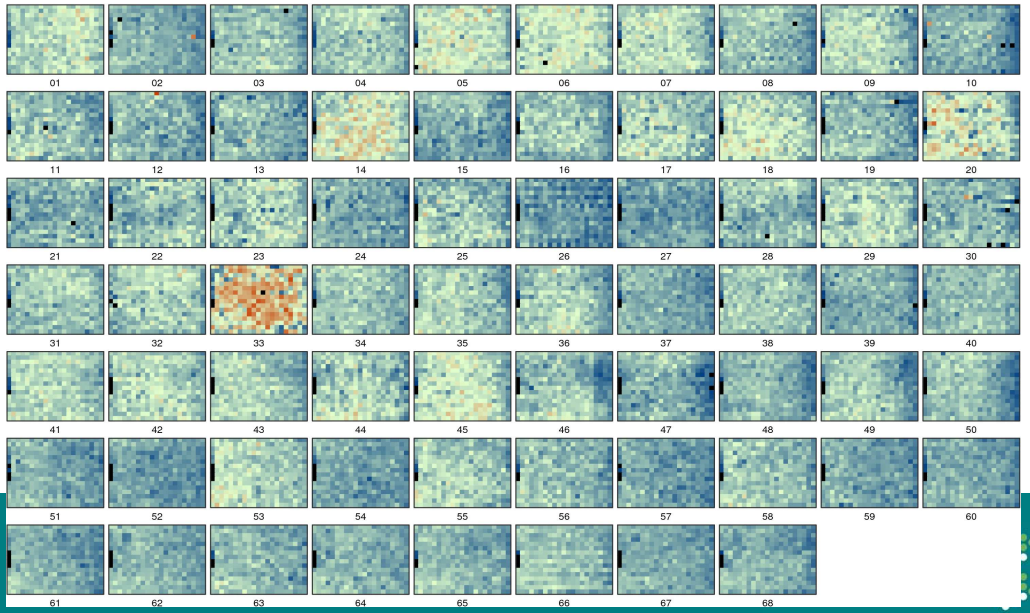
$$x'_{ki} = \frac{x_{ki} - \mu_i}{\sigma_i}$$

Long term drifts

Number of cells

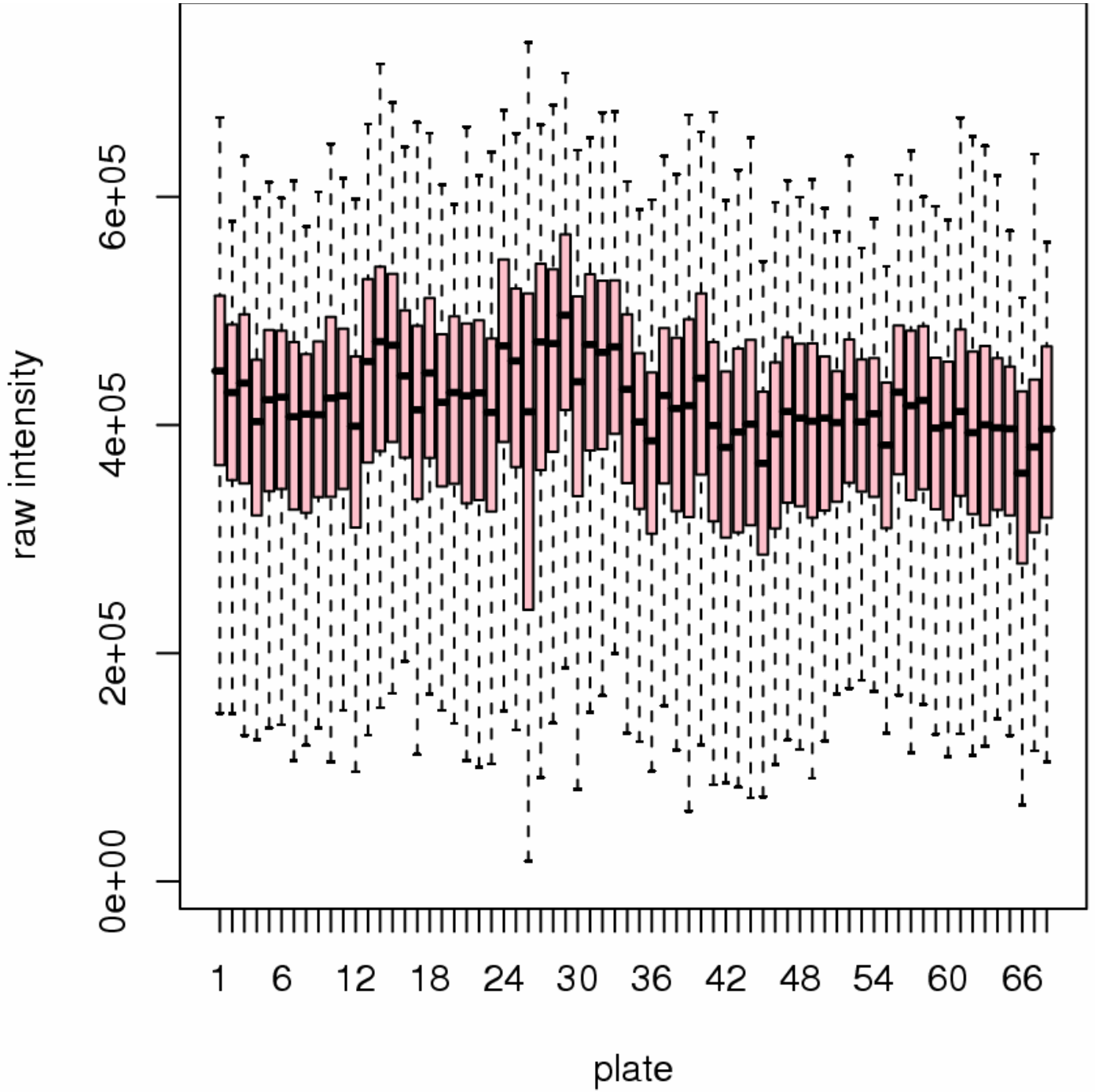


Number of cells / no. cells in negative controls in same plate

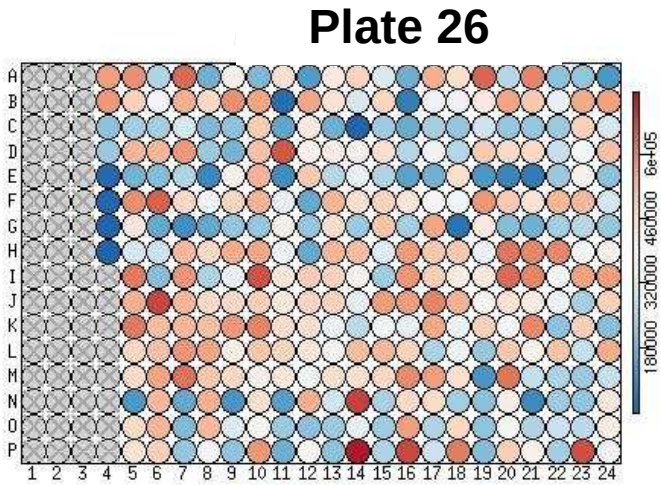


Dharmacon siARRAY library

Hek293 cells
viability screen
Boutros Lab
DKFZ

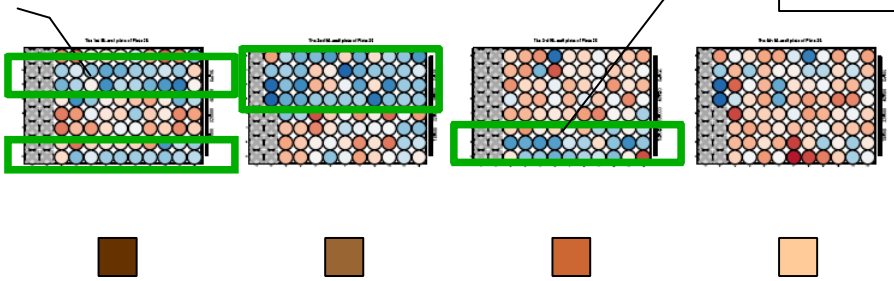


Normalization problem…
Too many hits

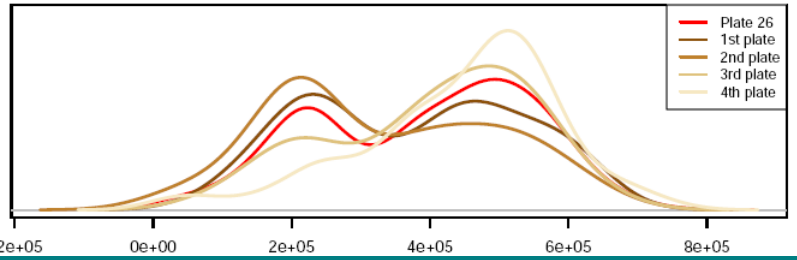


proteasome subunits
or components;
ATP/GTP-binding site motifs

like-Sm nucleoproteins and
ribosomal proteins



ribosomal proteins



Show imageHTS^3

EMBL-EBI

Phenotype of interest: elongated cells

67 / F13

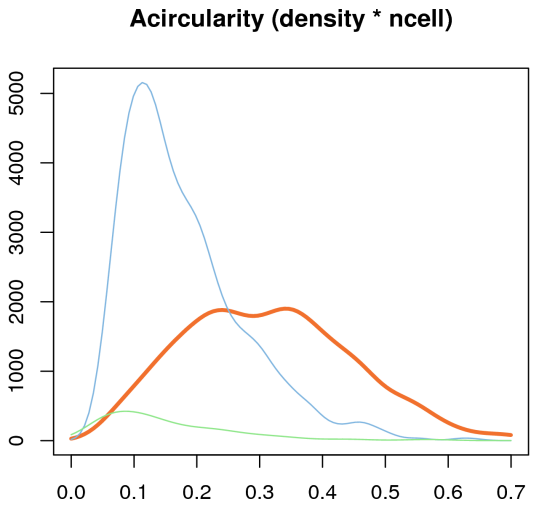
GPR124

Homo Sapiens probable G protein-coupled receptor 124 precursor (tumor endothelial marker 5)

Number of cells

```
Run 1: 357 / NC:473.5
Run 2: 357 / NC:474
```

```
Wilcoxon test for acirc:
p= 0,         W= 1078176
Z-test acirc:
p= 4.9e-105, t= 24.5806
```



01 / A08

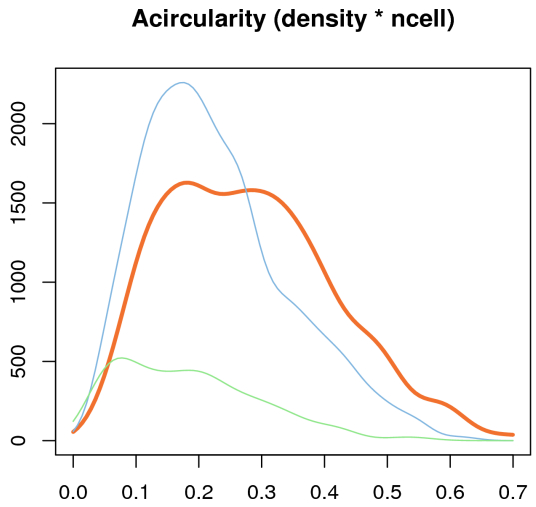
AZU1

Homo Sapiens azurocidin precursor (cationic antimicrobial protein CAP37), heparin-binding protein) (HBP)

Number of cells:

```
Run 1: 302 / NC:308
Run 2: 312 / NC:305
```

```
Wilcoxon test for acirc:
p=1.11022e-16, W= 465024
Z-test acirc:
p=1.87601e-17, t= 8.5637
```



54/ F13

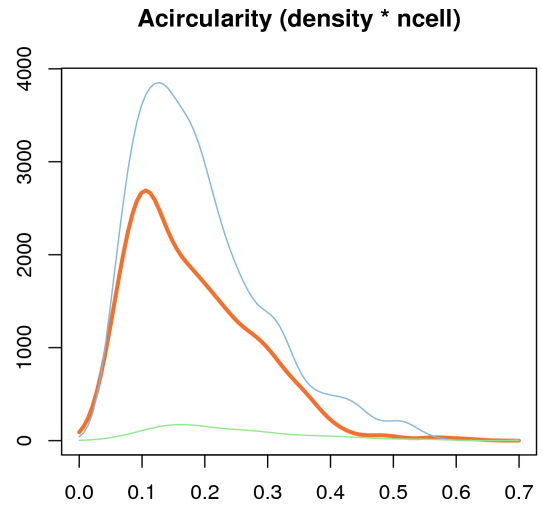
FLJ41238

Homo sapiens family with sequence similarity 79, member B (FAM79B), mRNA

Number of cells:

```
Run 1: 281 / NC:417.5
Run 2: 274 / NC:432.5
```

```
Wilcoxon test for acirc:
p=0.990294, W= 440619
Z-test acirc:
p=0.994775, t=-2.56542
```



Wilcox: Wilcoxon rank sum test with continuity correction. One sided with alternative hypothesis: shift > 0
Z-test: Two-sample Welch t-test. One sided with alternative hypothesis of diff(means) > 0

Gene info obtained from ensembl using biomaRt

EMBL-EBI

Phenotype of interest: elongated cells

67 / F13

GPR124

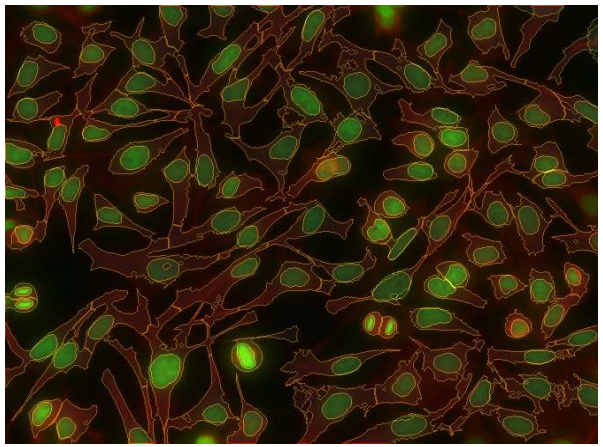
Homo Sapiens probable G protein-coupled receptor 124 precursor (tumor endothelial marker 5)

Number of cells

Run 1: 357 / NC:473.5
Run 2: 357 / NC:474

Wilcoxon test for acirc:
p= 0, W= 1078176

Z-test acirc:
p= 4.9e-105, t= 24.5806



01 / A08

AZU1

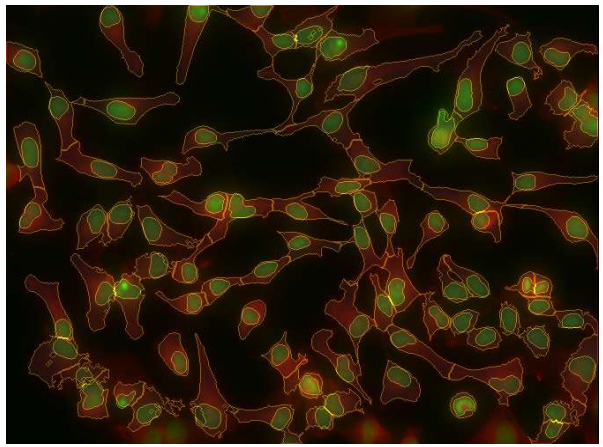
Homo Sapiens azurocidin precursor (cationic antimicrobial protein CAP37), heparin-binding protein) (HBP)

Number of cells:

Run 1: 302 / NC:308
Run 2: 312 / NC:305

Wilcoxon test for acirc:
p=1.11022e-16, W= 465024

Z-test acirc:
p=1.87601e-17, t= 8.5637



54/ F13

FLJ41238

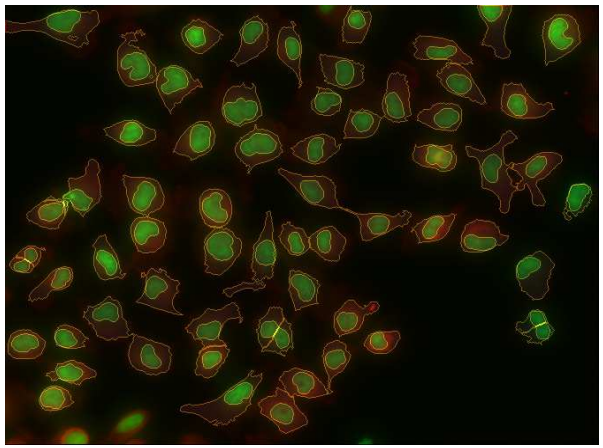
Homo sapiens family with sequence similarity 79, member B (FAM79B), mRNA

Number of cells:

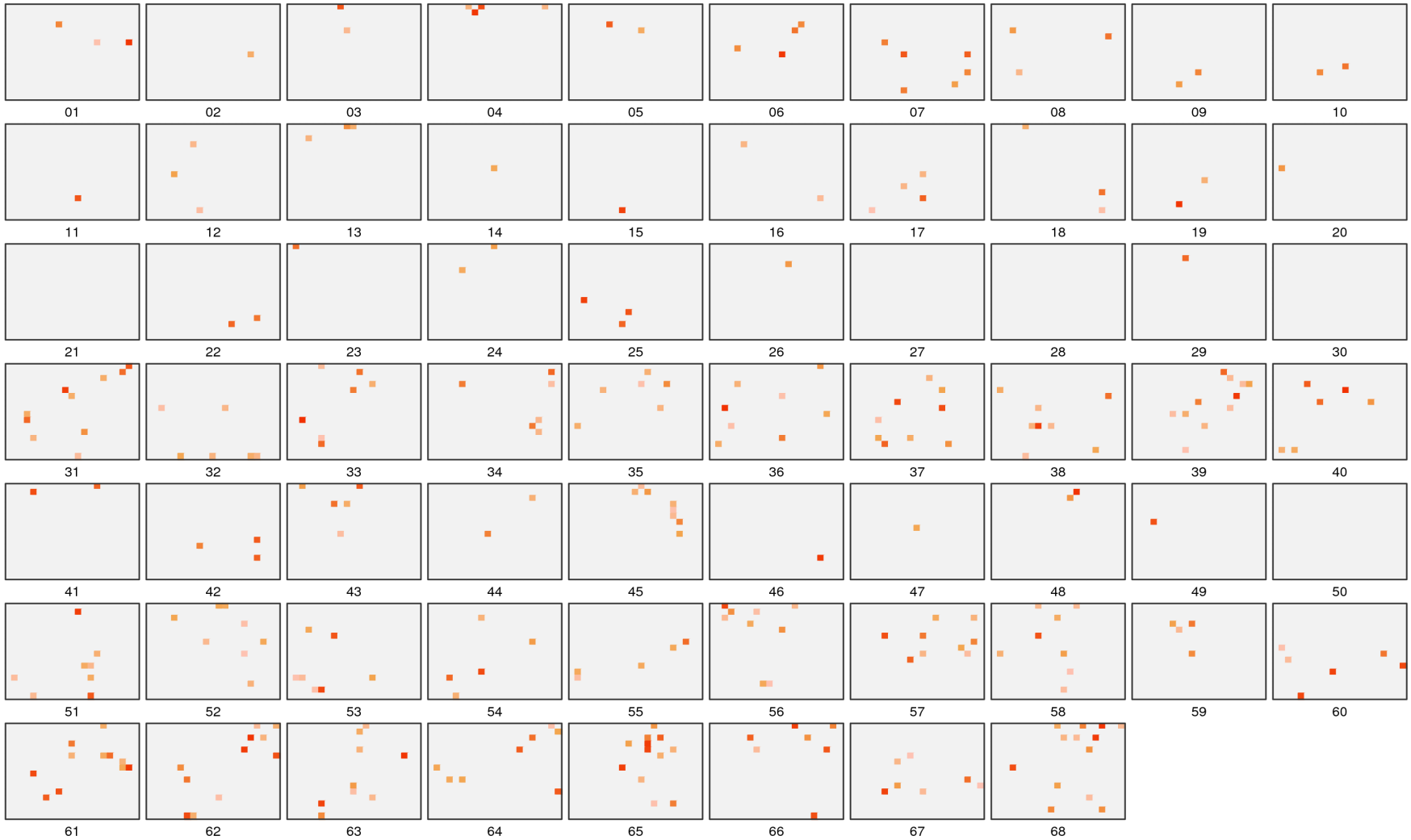
Run 1: 281 / NC:417.5
Run 2: 274 / NC:432.5

Wilcoxon test for acirc:
p=0.990294, W= 440619

Z-test acirc:
p=0.994775, t=-2.56542



Phenotype of interest: elongated cells – hit list visualisation



acircularity T-test: `acirc.T > 12 & 250 < n < 450`

EMBL-EBI

Mitocheck:
dynamic modeling of live cell populations for clustering and classification of genes and phenotypes

Gregoire Pau (EBI)

with
Thomas Walter
Beate Neumann
Jan Ellenberg (EMBL)

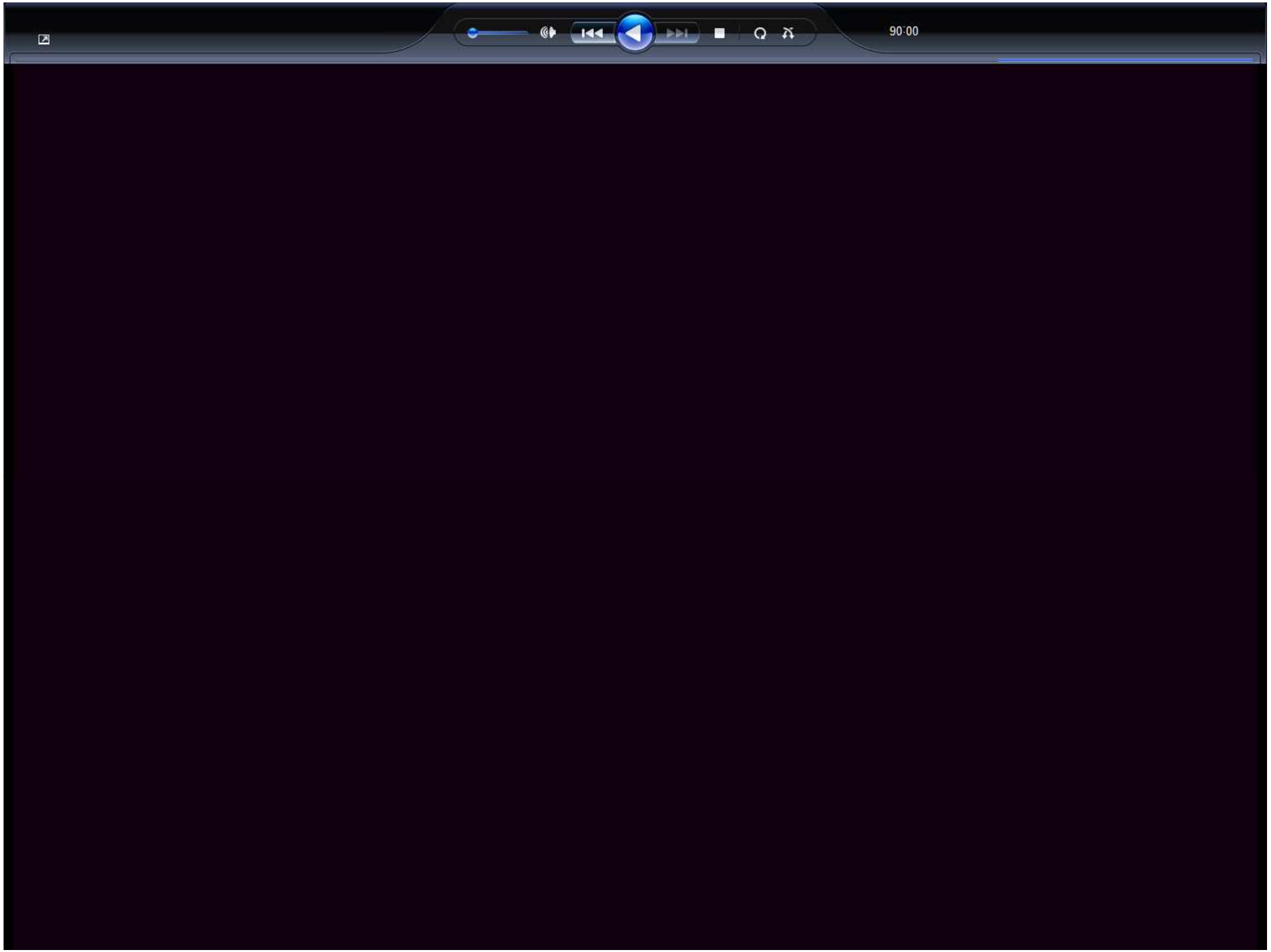
Mitocheck time lapse data

Live cell time-lapse imaging

- HeLa cell line expressing H2B GFP
- seeded on siRNA spots and grown during ~48h
- fluorescence time-lapse live imaging (sampling rate=30 min)

Experimental output

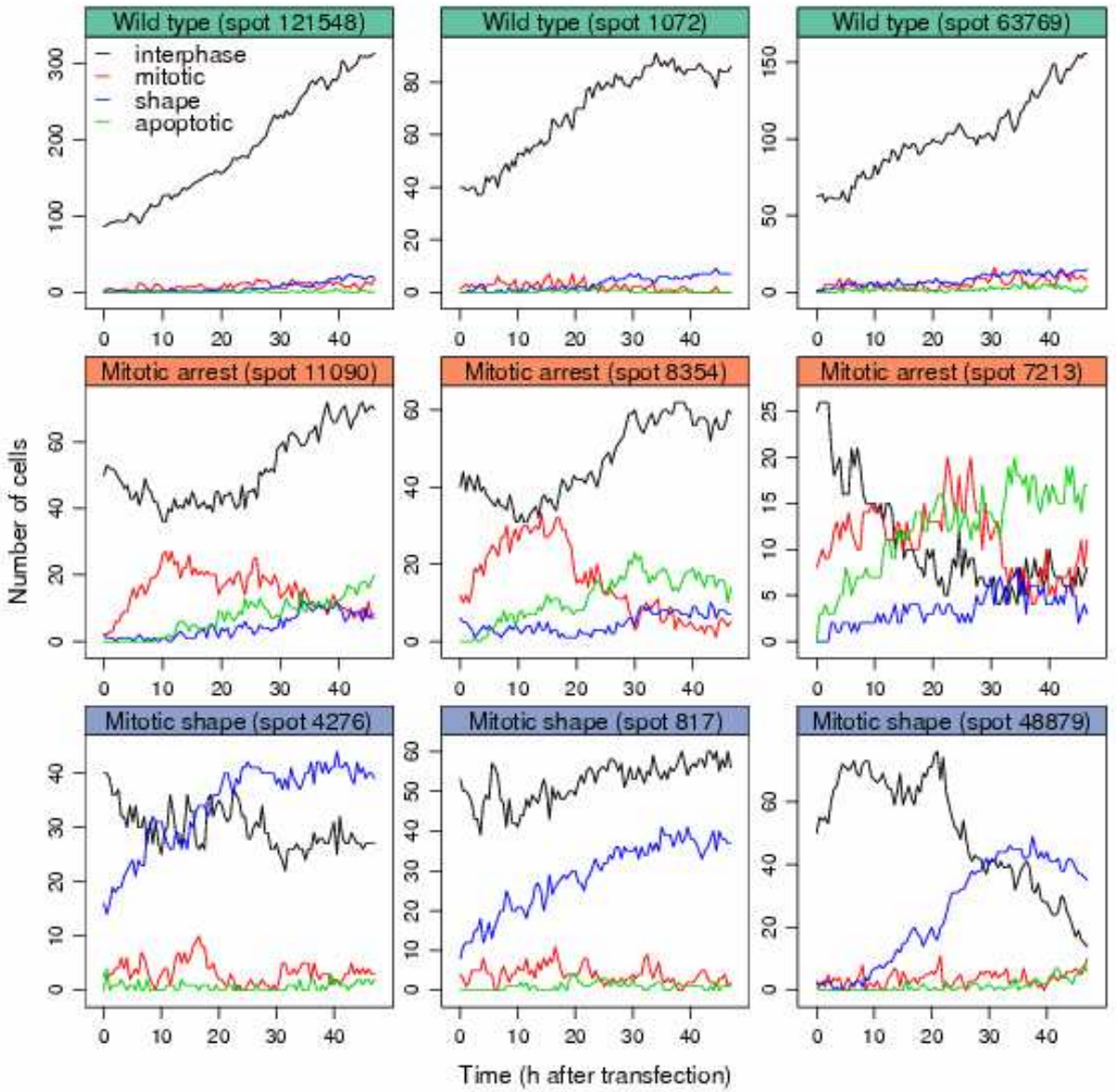
- video sequences of 96 images (1024x1024)
- 100 MB per spot
- ~200,000 spots (20 TB)


90:00

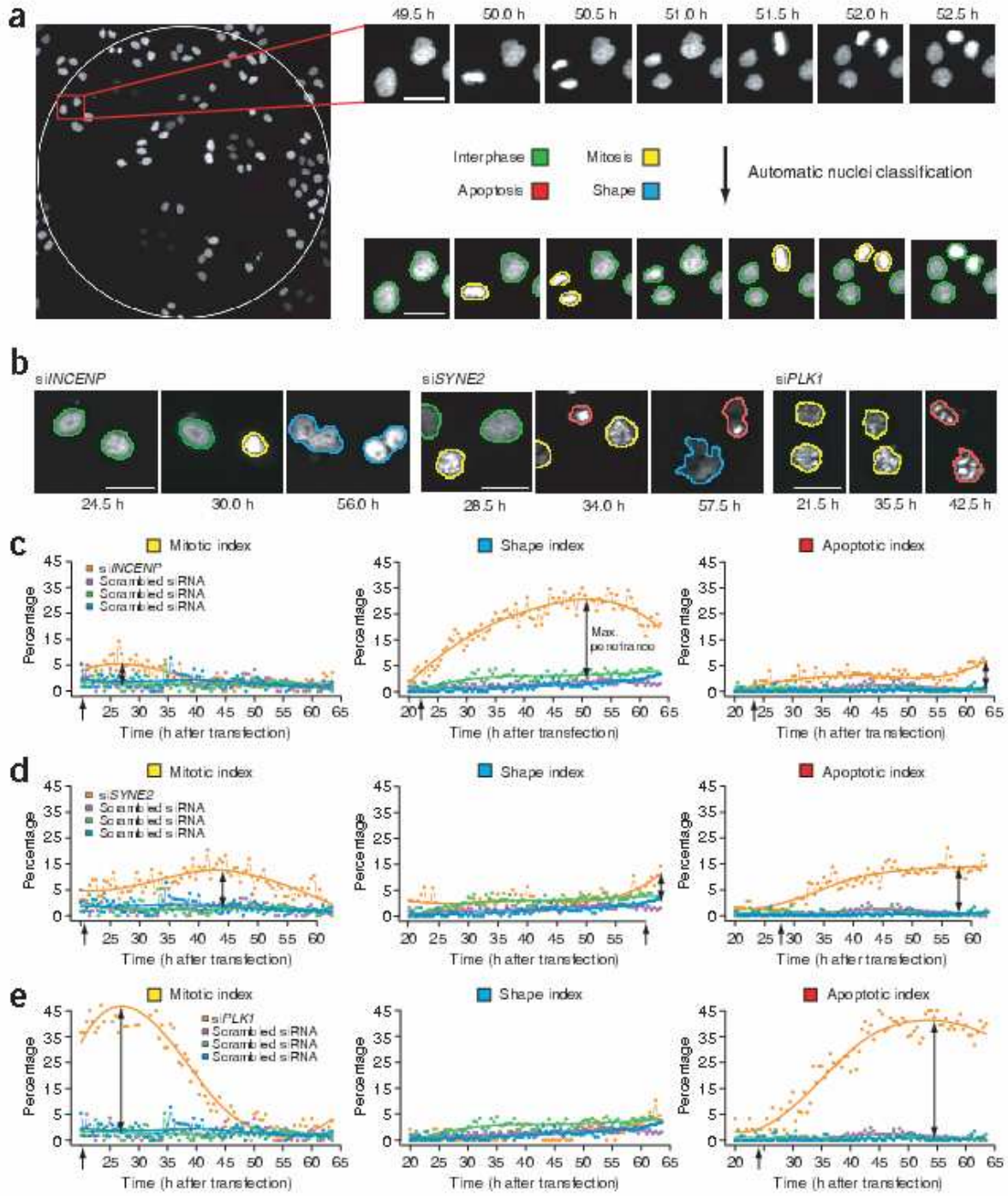
Examples

Kif11

Incenp



	Name	Description
■	Wild type	Cells are dividing and growing normally.
■	Mitotic arrest	Accumulation of cells blocked in prometaphase, followed by apoptosis.
■	Mitotic shape	Constant increase of multi-nucleated cells.



Neumann et al.
Nature Methods
2006

Conclusions

HT microscopy of biological systems is becoming a rich source of such data

Tools in Bioconductor (et al.)
Reproducible research
Feature extraction, variable selection, machine learning

mitoODE

Parameters of a biologically motivated model of the data are a more useful phenotype for classification than the raw time courses

EBI
Elin Axelsson
Richard Bourgon
Alessandro Brozzi
Ligia Bras
Tony Chiang
Audrey Kauffmann
Gregoire Pau
Oleg Sklyar
Mike Smith
Jörn Tödling

DKFZ
Florian Fuchs
Thomas Horn
Dierk Ingelfinger
Sandra Steinbrink
Michael Boutros

Cristina Cruciat

Florian Hahne
Stefan Wiemann

UCSD
Amy Kiger

EMBL
Lars Steinmetz
Eugenio Mancera
Zhenyu Xu
Julien Gagneur

Jan Ellenberg
Thomas Walter
Beate Neumann

Bioconductor
Robert Gentleman
Seth Falcon
Martin Morgan
Rafael Irizarry
Vince Carey
… & many others

EMBL-EBI

2007 Call for Applications

EMBL Interdisciplinary Postdocs - EIPOD

This new EMBL initiative promotes cross-disciplinary research. EIPODs are supported by at least two labs at the five EMBL sites in Heidelberg and Hamburg (Germany), Grenoble (France), Hinxton (UK) and Monterotondo (Italy). EIPOD projects connect scientific fields that are usually separate, or transfer techniques to a novel context.

For a list of possible projects and further information please visit: **www.embl.org/eipod**

You are also encouraged to propose your own interdisciplinary project.

Online application until **31st August 2007**

EMBL