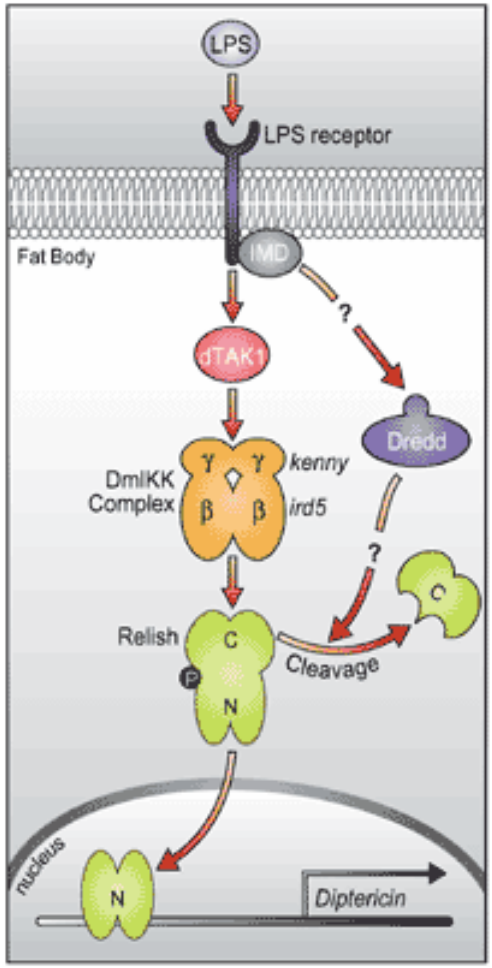
Image analysis and modelling of high-throughput cell based assays

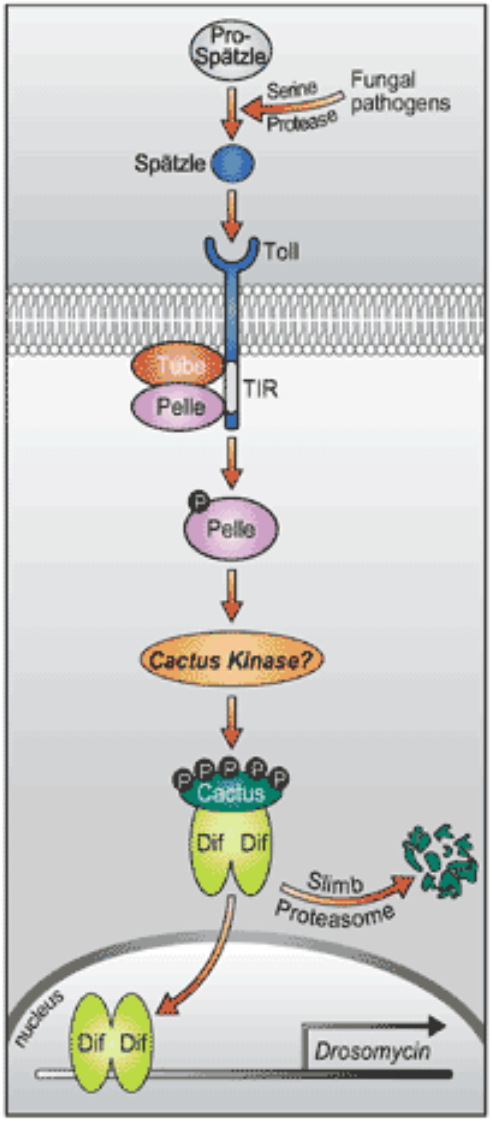
Wolfgang Huber

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Signaling pathways



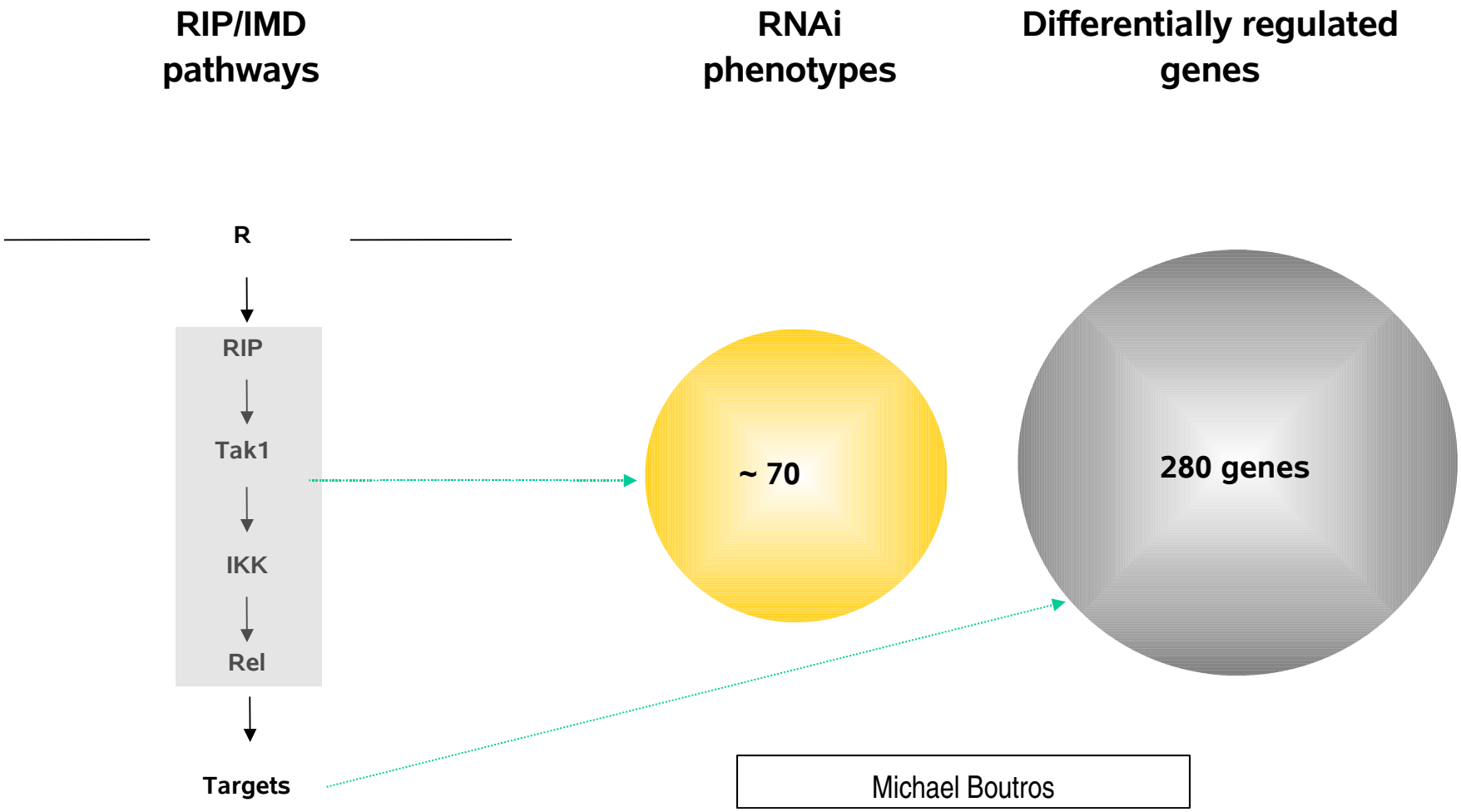


Drosophila **antibacterial signalling**

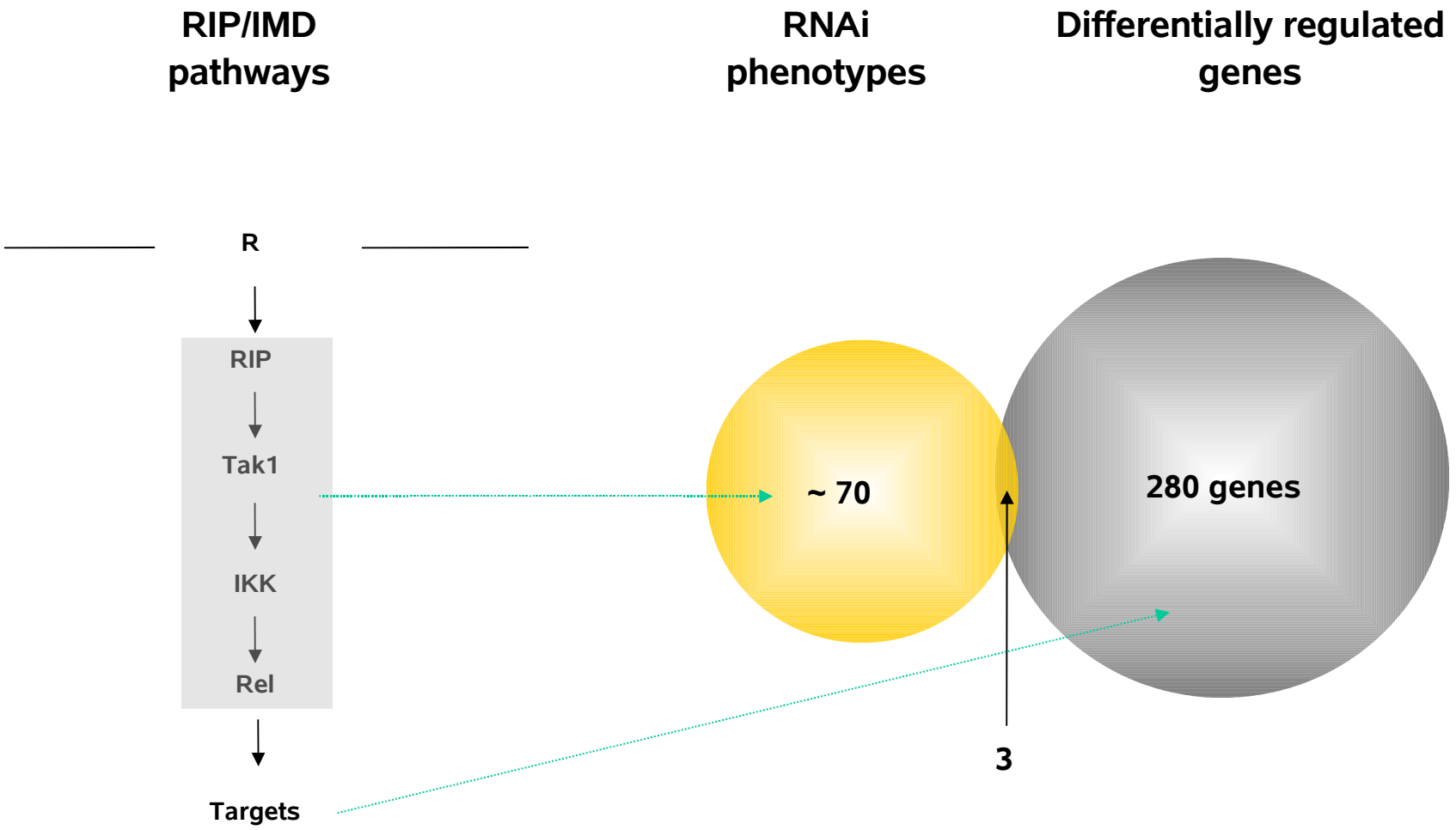
Drosophila **Toll/anti-fungal signaling**

EMBL-EBI

Differential Expression vs Signaling Function



Most pathway targets are not required for pathway function



Genetic interactions

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

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

➔ in drosophila, ~95% no viability phenotype

(Boutros et al. Science 303 (2004))

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

➔ evolutionary pressure for robustness

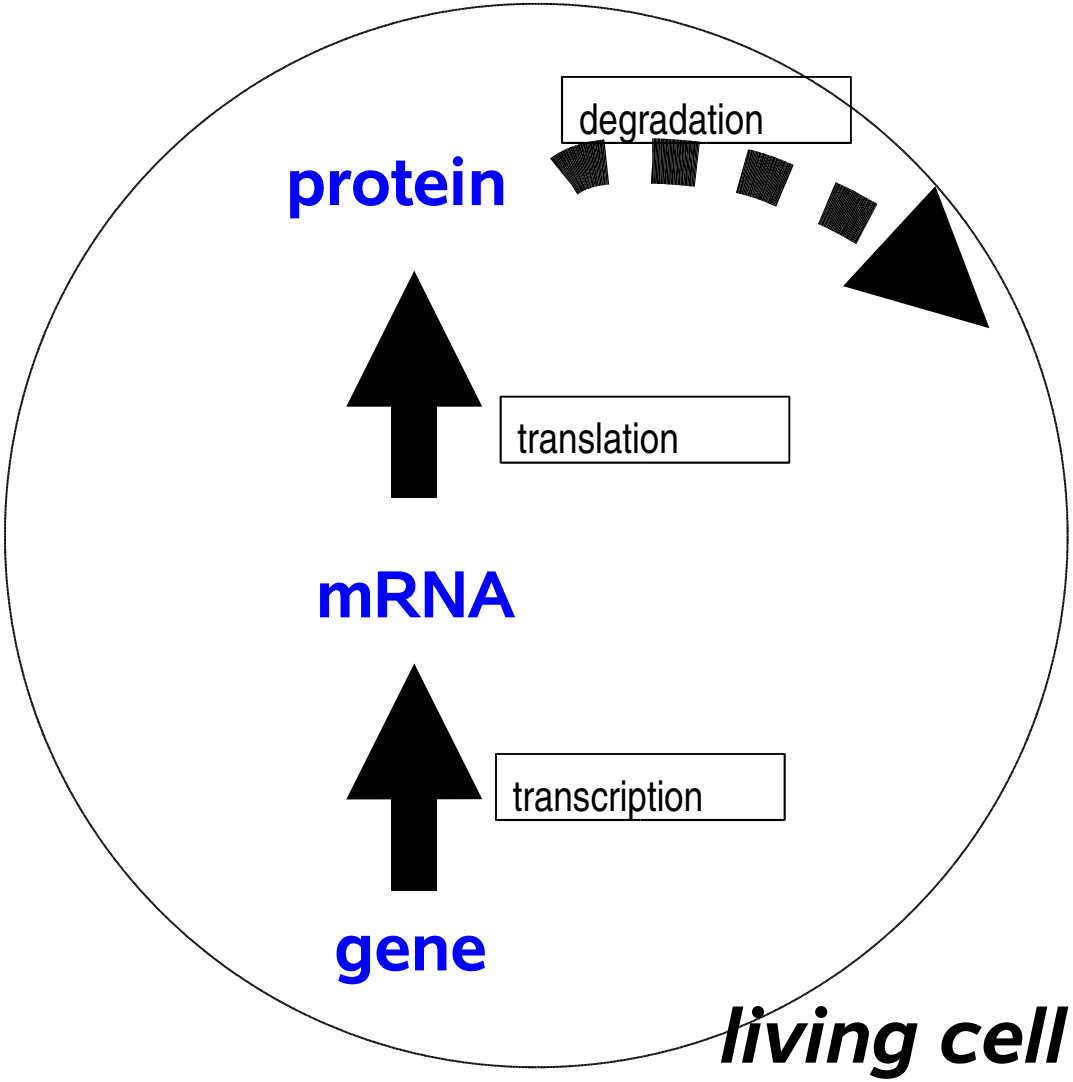
What are the implications for functional studies?

- use combinatorial perturbations (co-RNAi, small molecules, different genetic backgrounds)
- observe multiple and complex phenotypes over time with high sensitivity
- complex analyses (e.g. graph-like models) to relate the data to gene-gene and gene-phenotype interactions

RNAi as a loss of function perturbator

gene-sequence specific reagents (eg siRNAs)

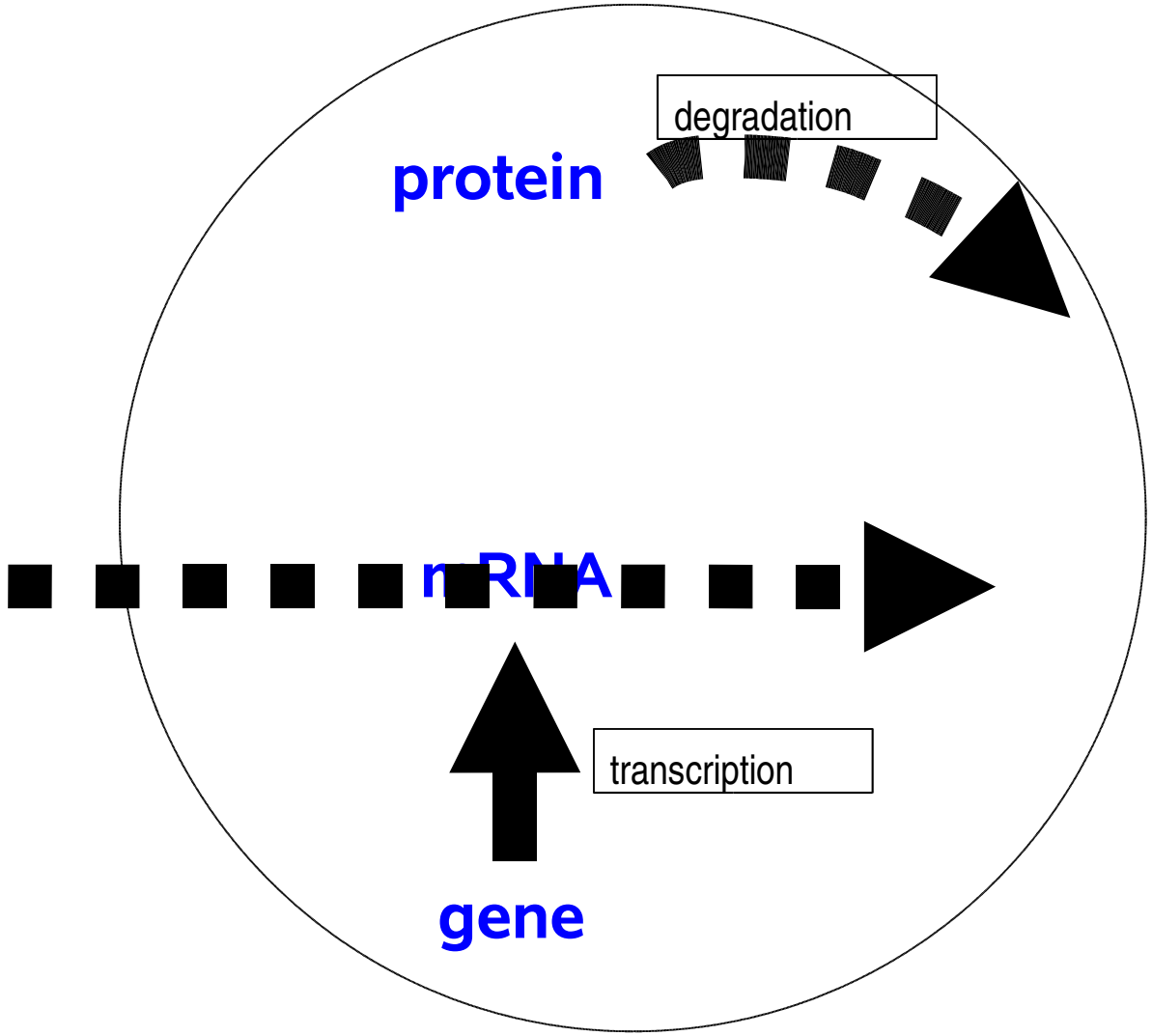
easy to make for any gene (there are caveats...)



RNAi as a loss of function perturbator

gene-sequence specific reagents (eg siRNAs)

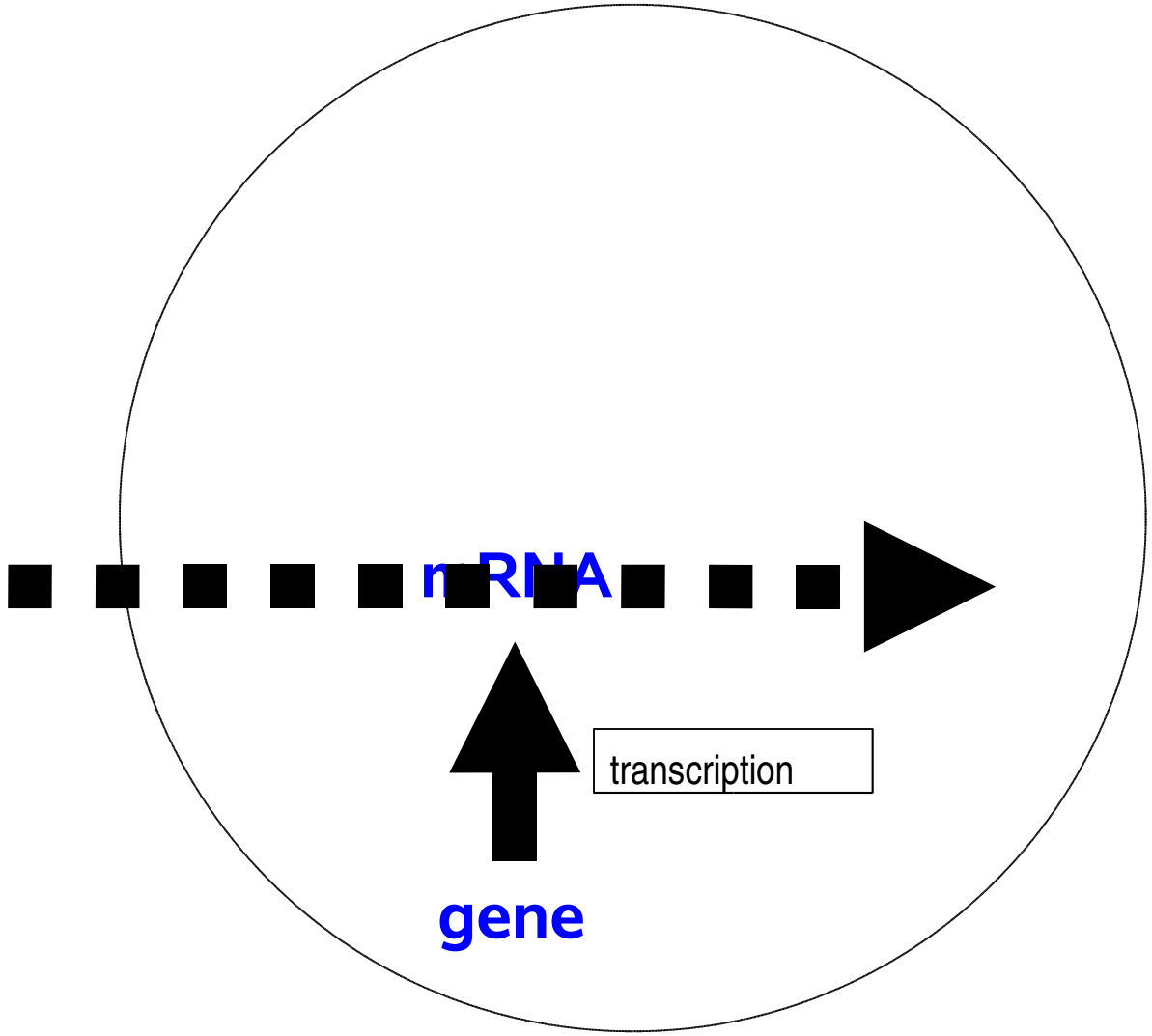
easy to make for any gene (there are caveats...)



RNAi as a loss of function perturbator

gene-sequence specific reagents (eg siRNAs)

easy to make for any gene (there are caveats...)



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

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 systematic phenotype analysis with cellbased 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

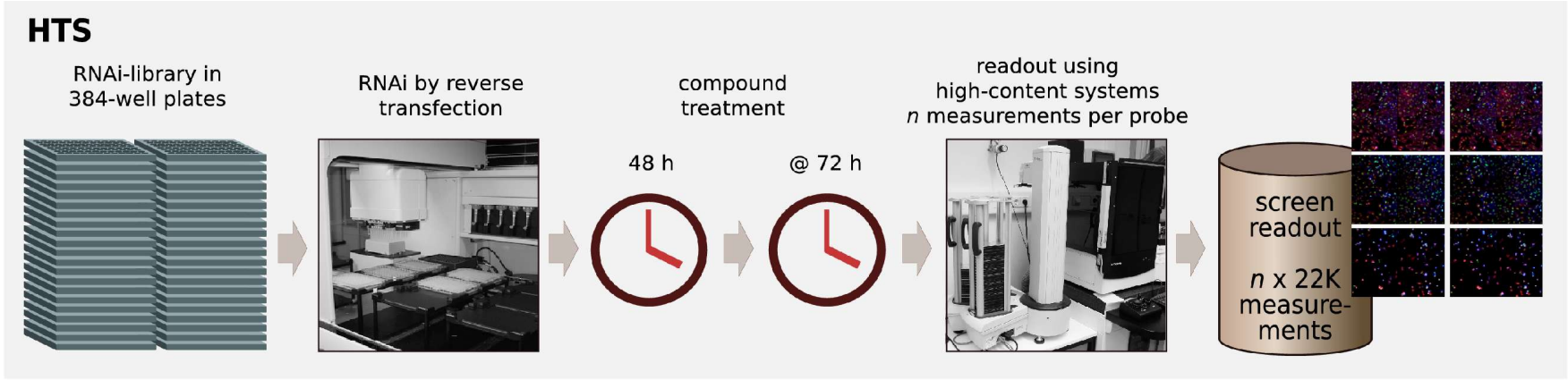
data management

EBImage (O. Sklyar)

image processing and analysis

construction of feature extraction workflows for large sets of similar images

High-throughput microscopy screening



quality control

image processing and analysis

Computational analysis

statistical analysis
determination of phenotypes

quality control

candidates

in vivo analysis

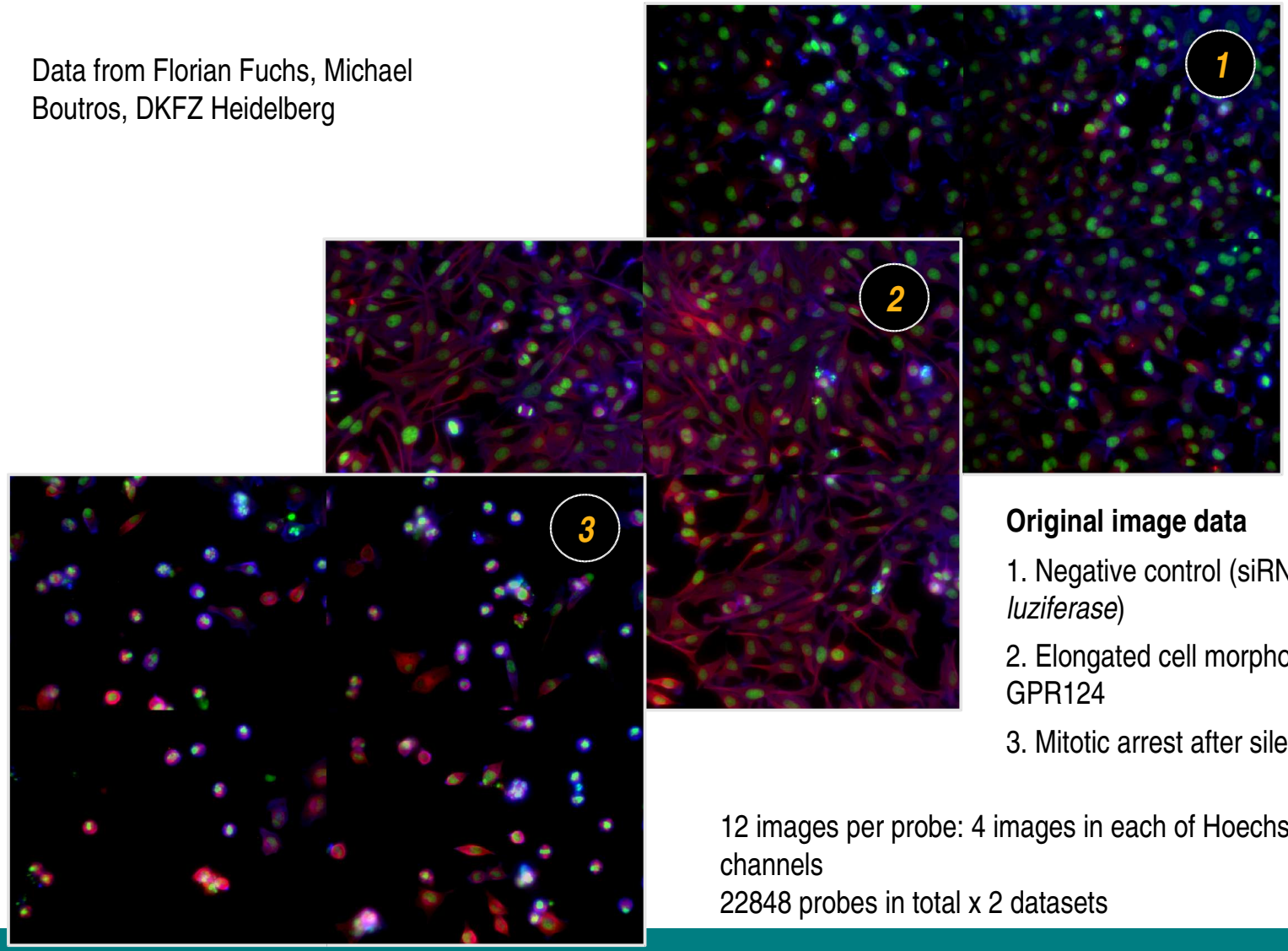
secondary assays

integration with biological databases

A genome-wide siRNA screen

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

Data from 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)

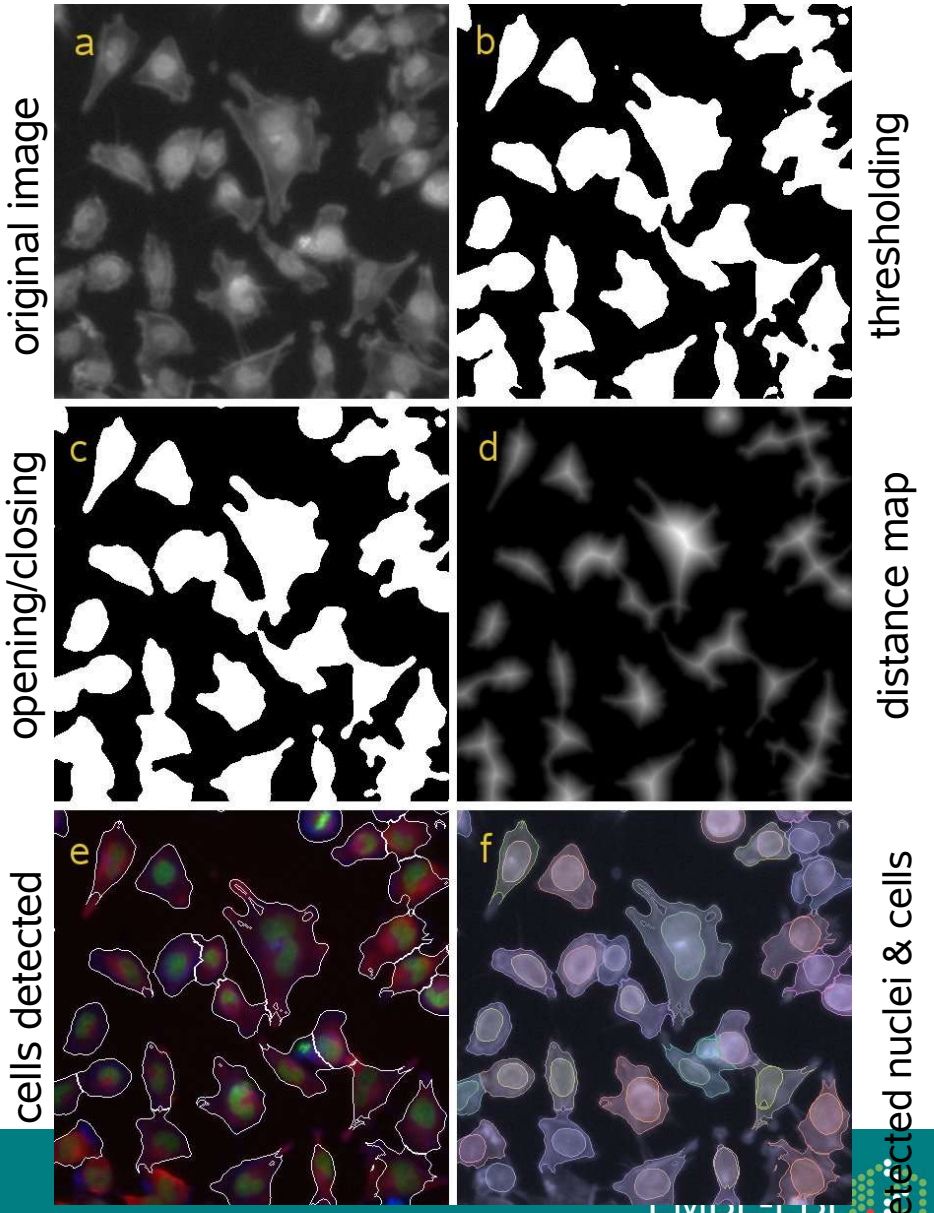


Image processing with R: simple operations

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
```

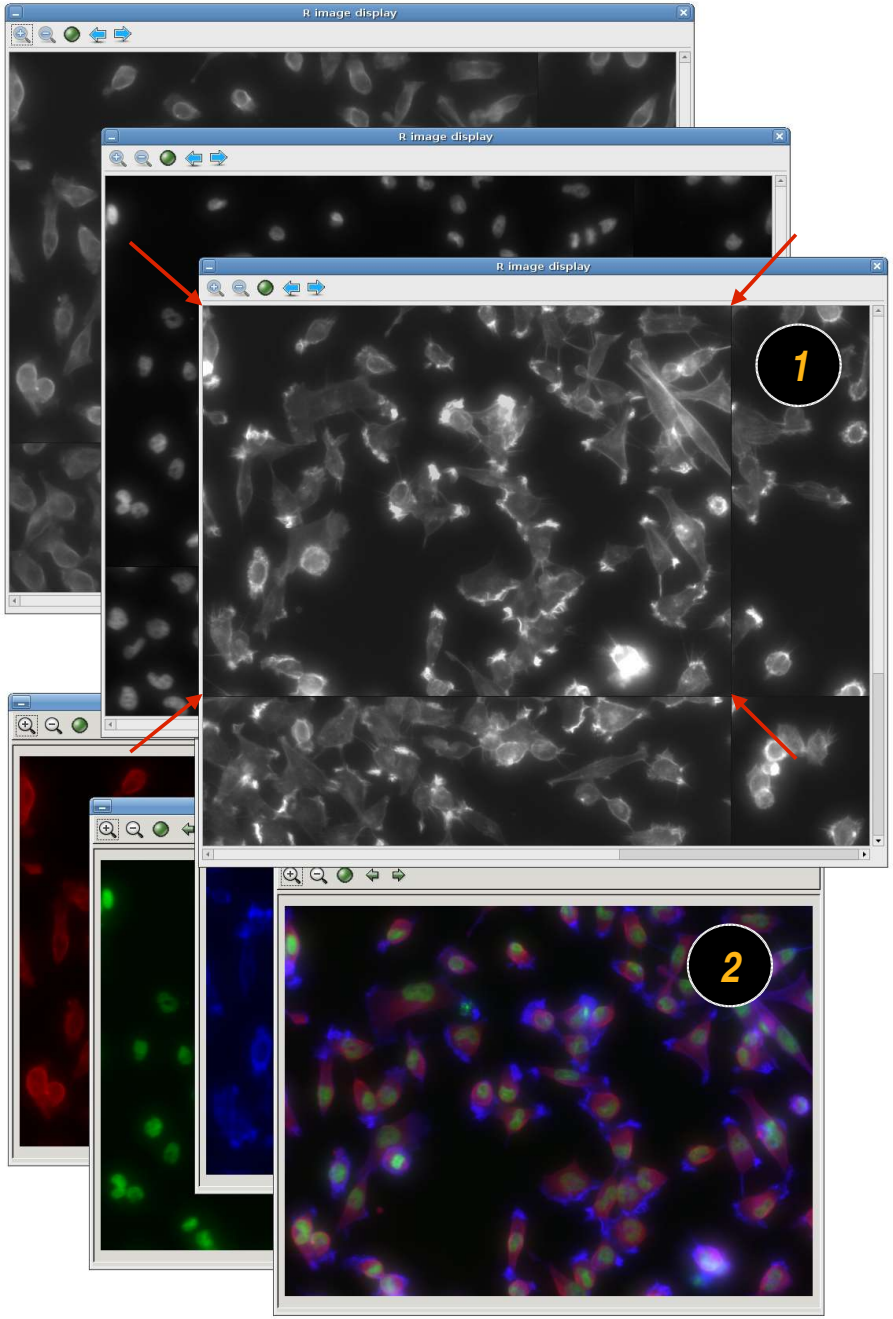
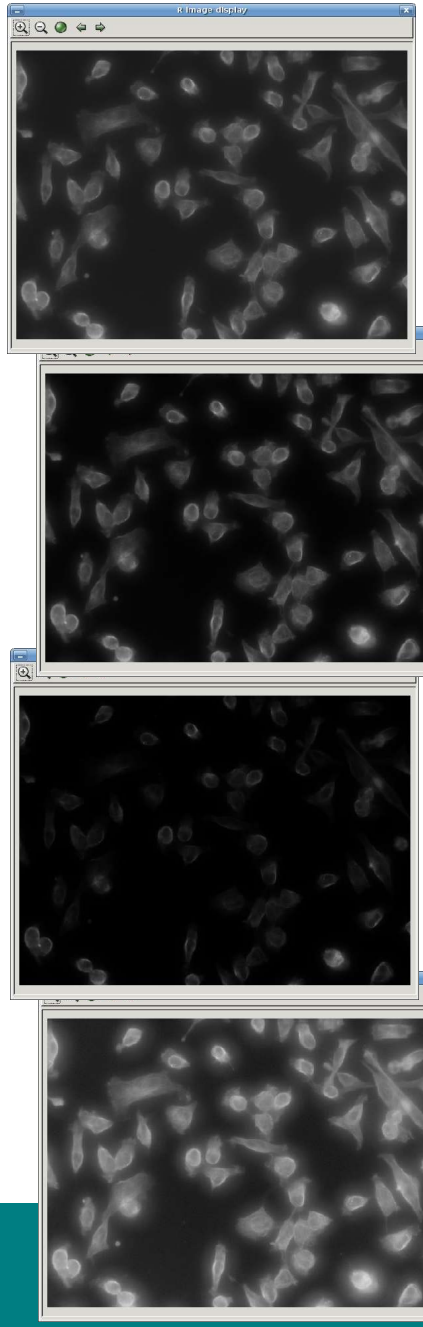
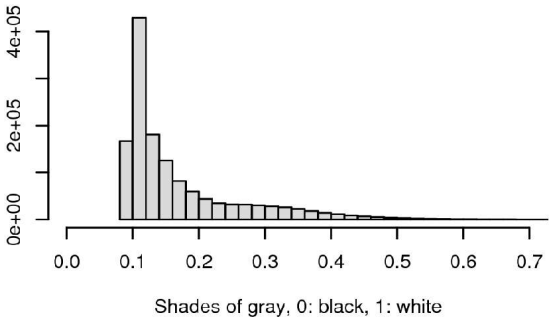
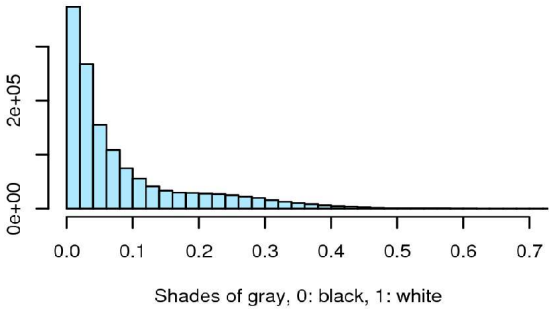
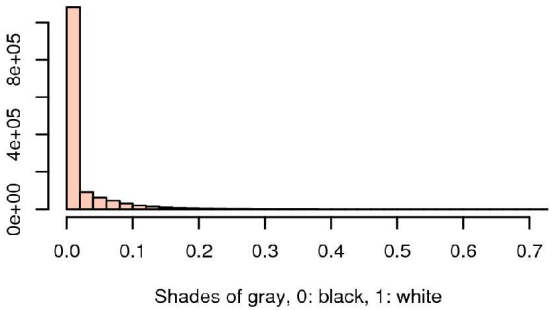


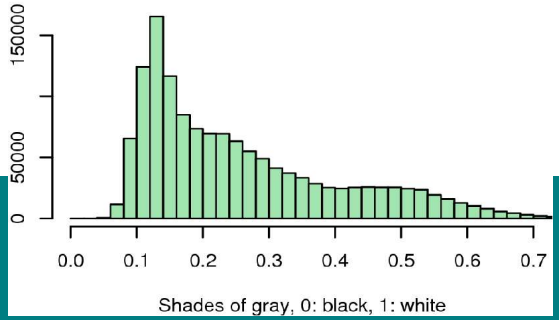
Image processing: arithmetic and visualization












```
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)))
```

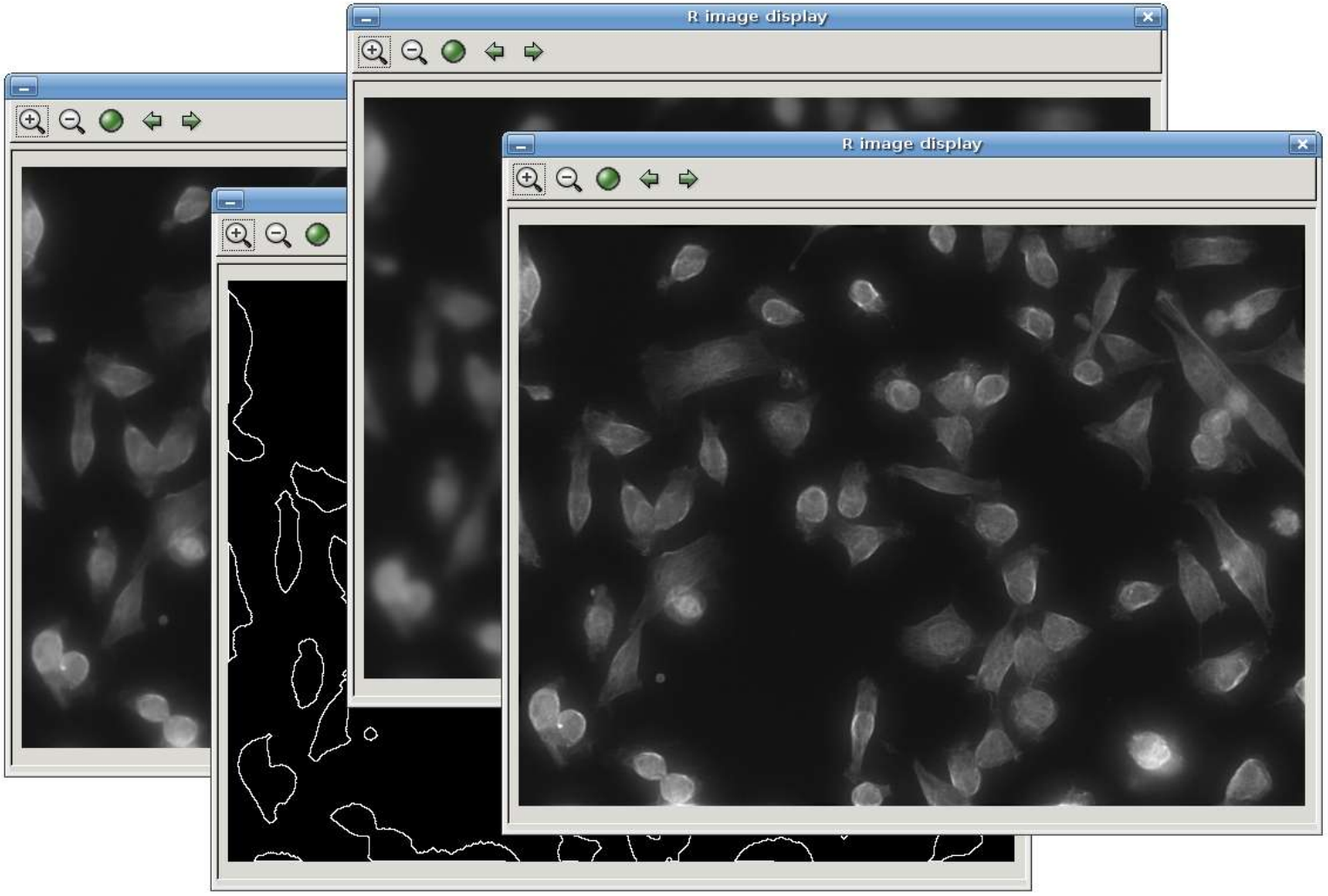
EMBL-EBI

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

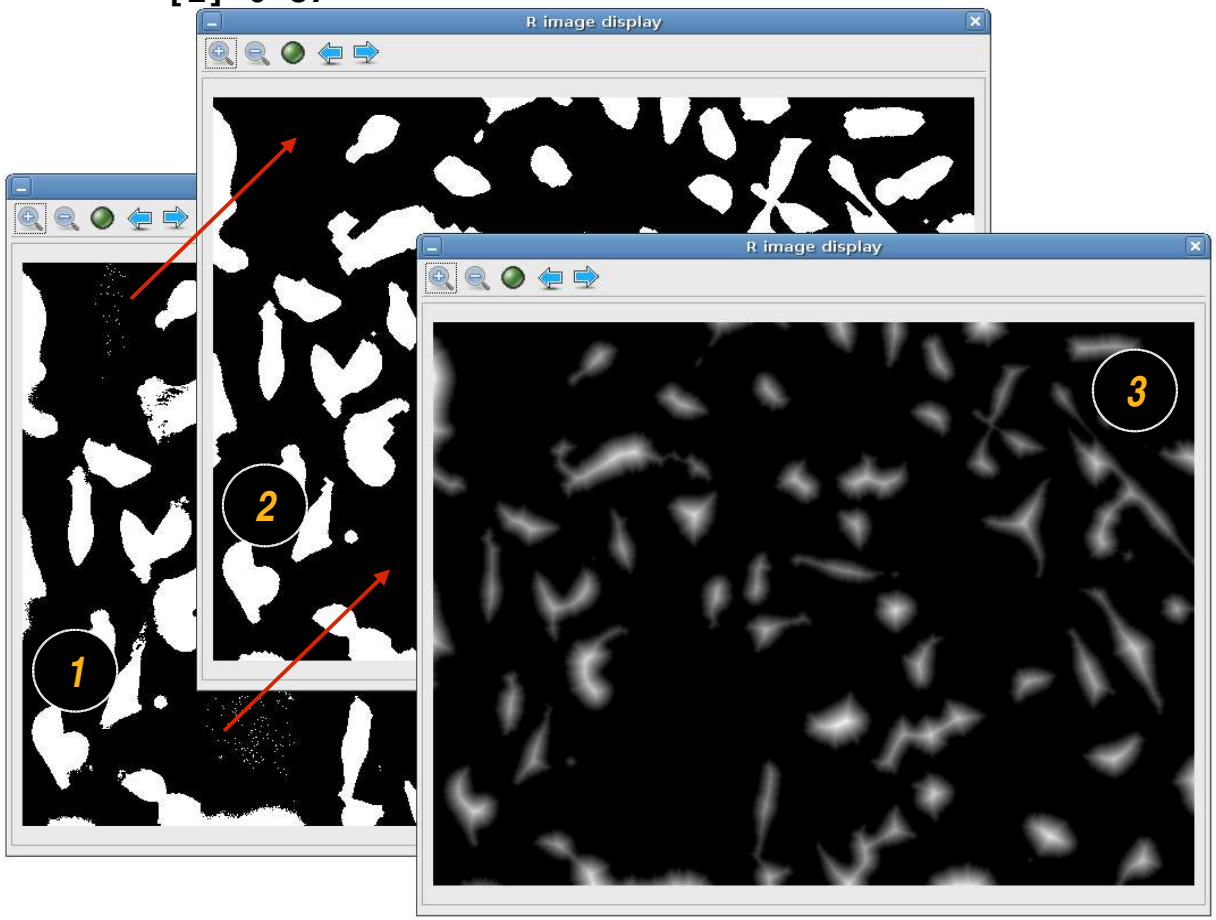
Mathematical Morphology

Distance map transformation

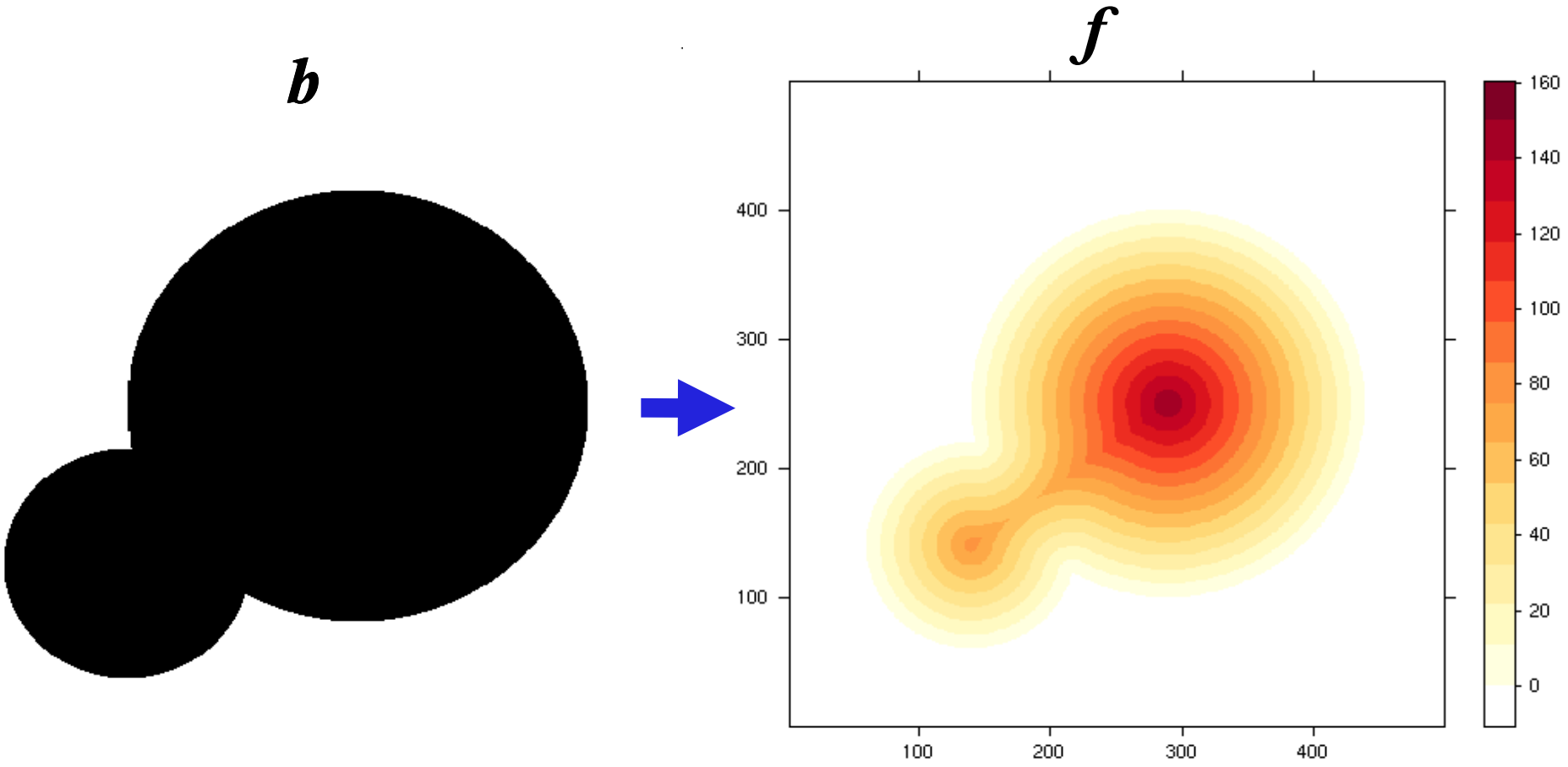
binary image -> greyscale

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`

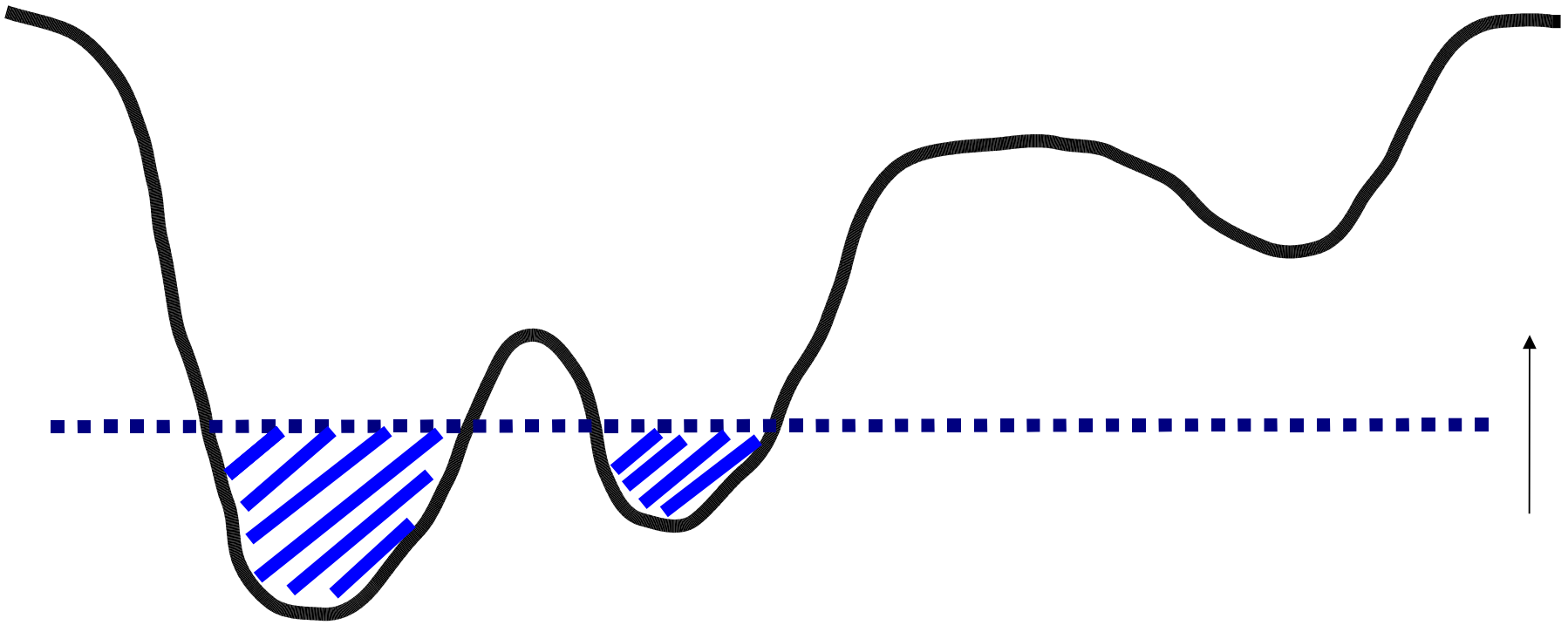


Distance map transformation



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

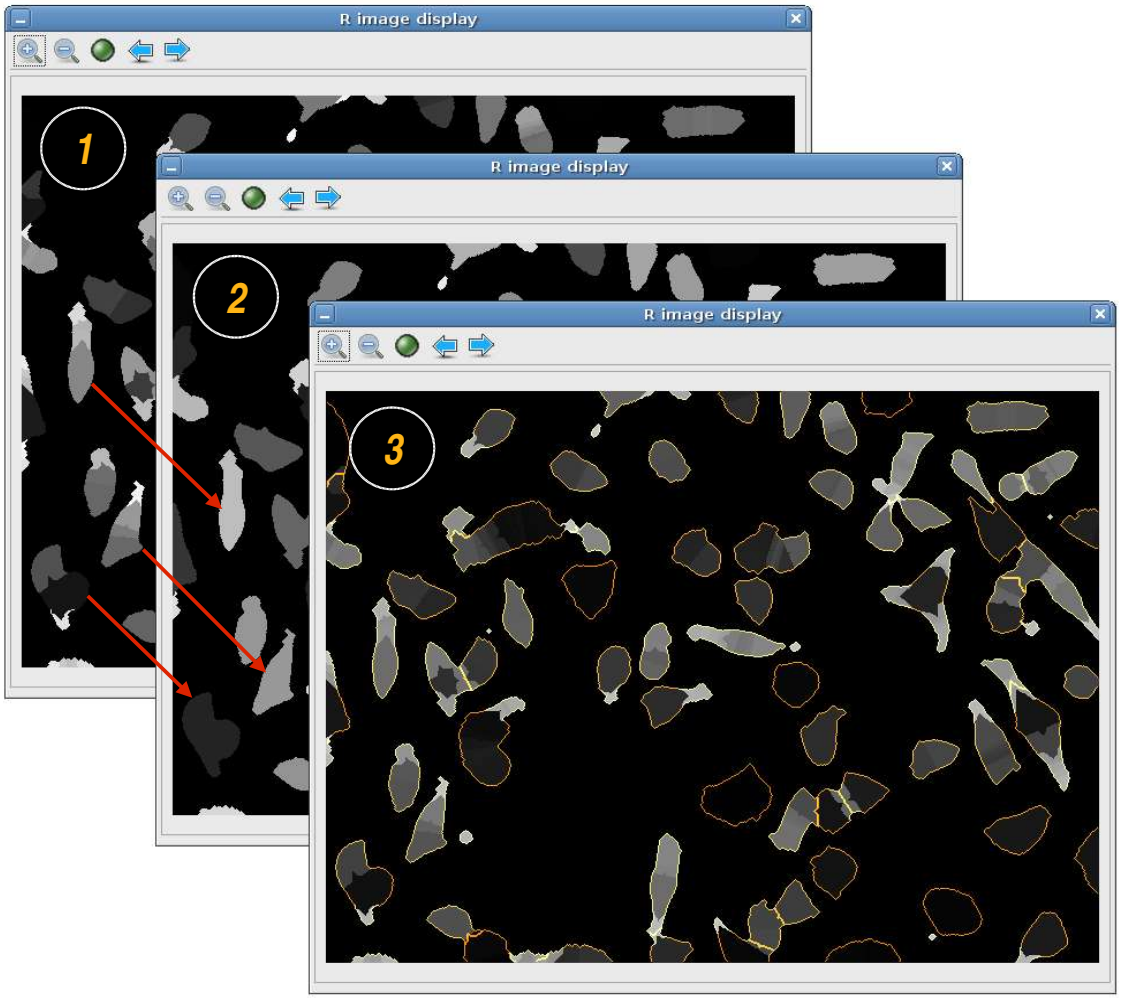
Watershed segmentation



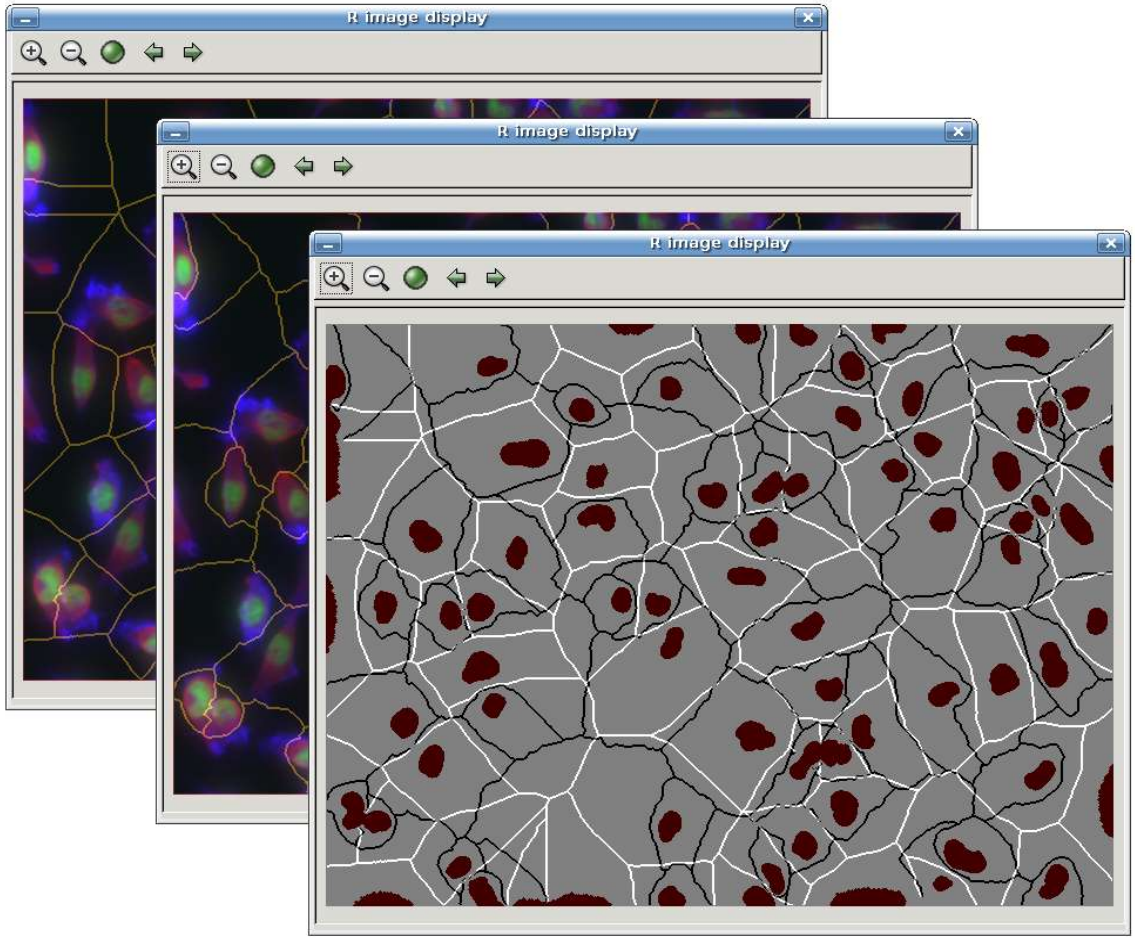
watershed segmentation can be very effective, but…:

- 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

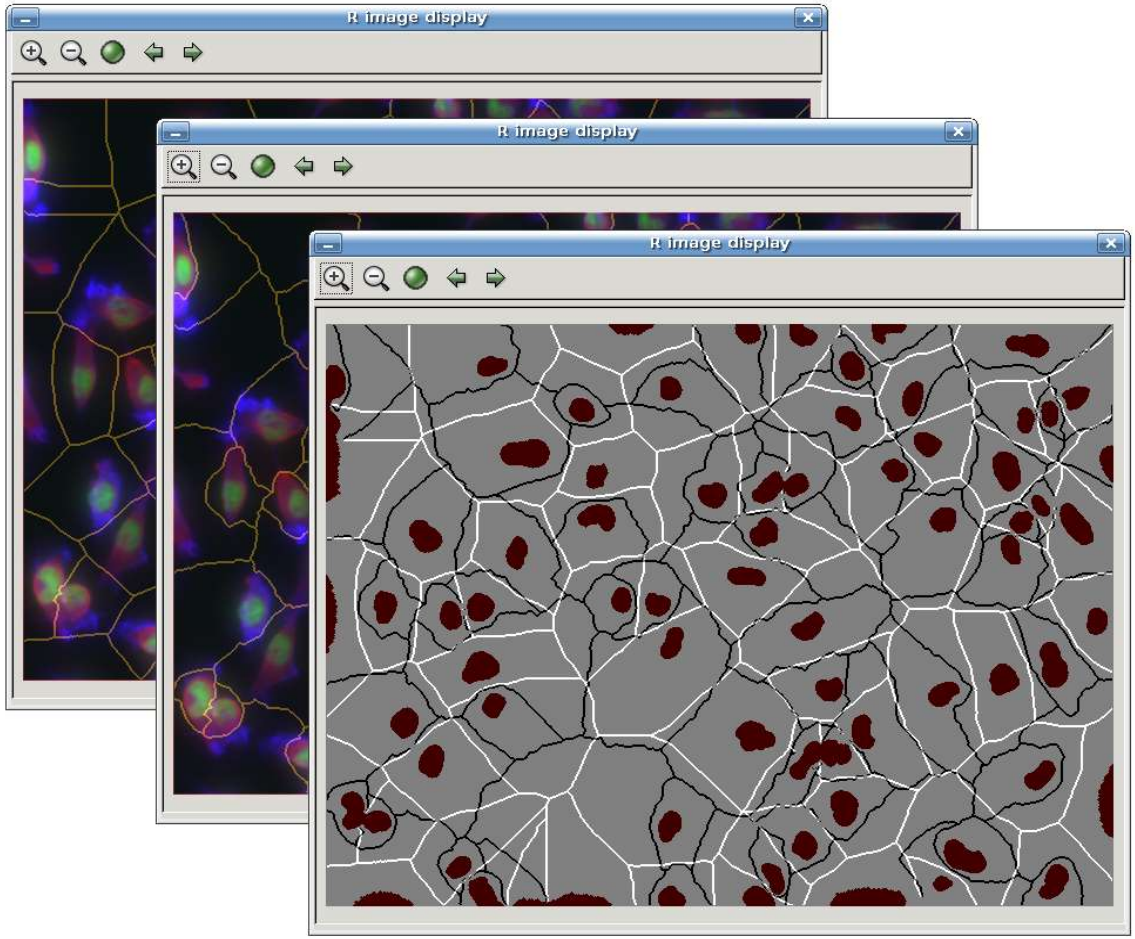
Example:

segment nuclei (easy)

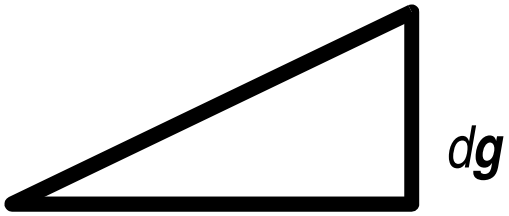
use them as seed points

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



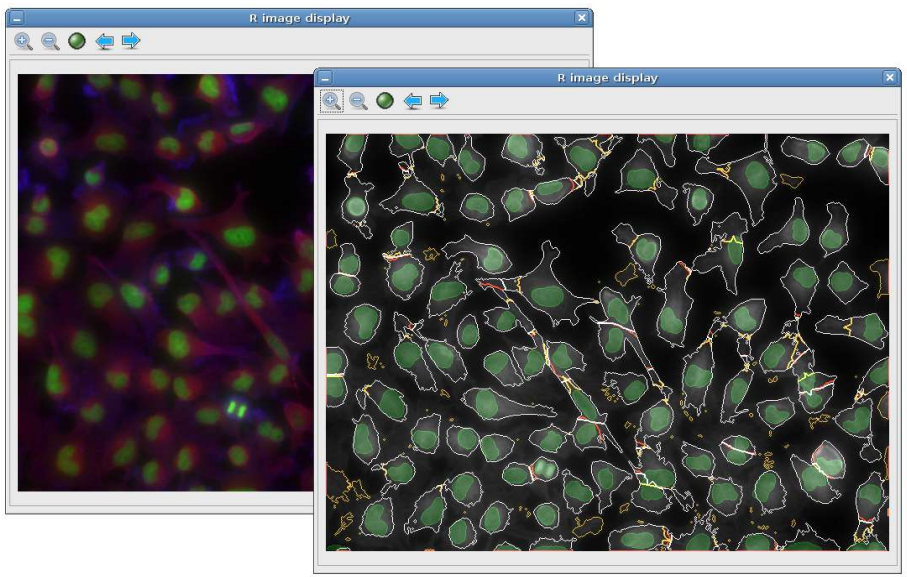
$\lambda \cdot d\mathbf{x}$

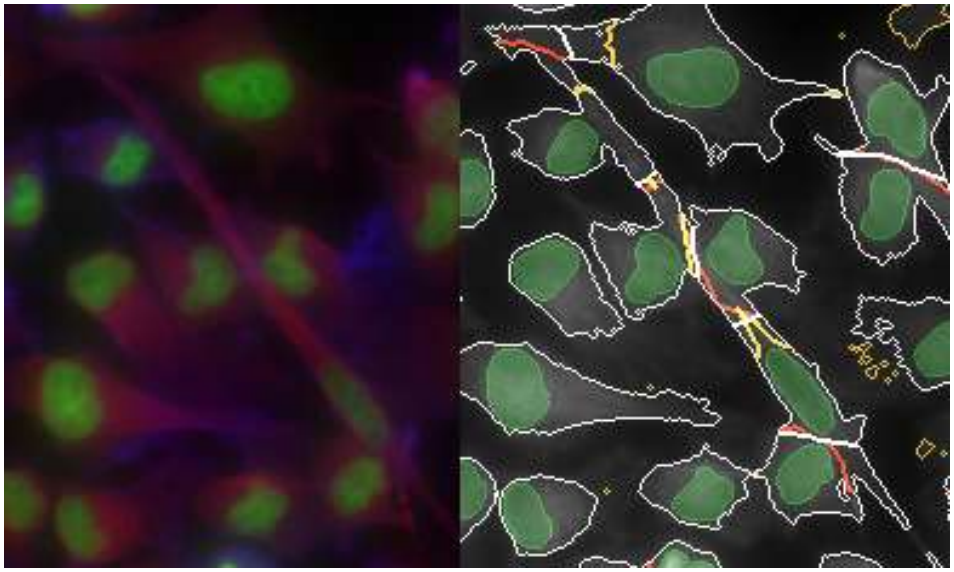
$$\mathbf{G} = \frac{\nabla \mathbf{g}(\mathcal{I}) \nabla \mathbf{g}^T(\mathcal{I}) + \lambda \mathbf{I}}{1 + \lambda}$$

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

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

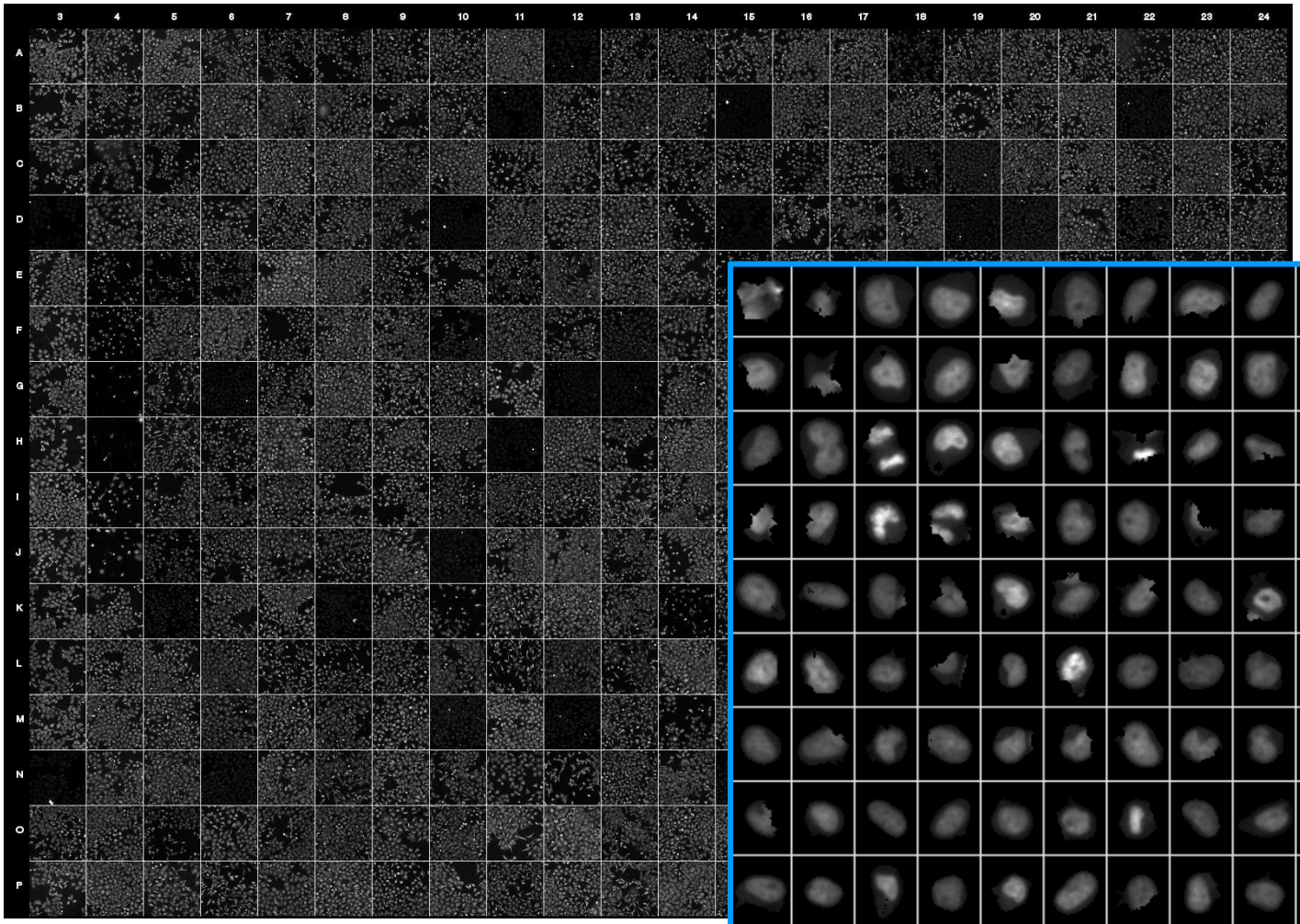
Voronoi diagrams on image manifolds



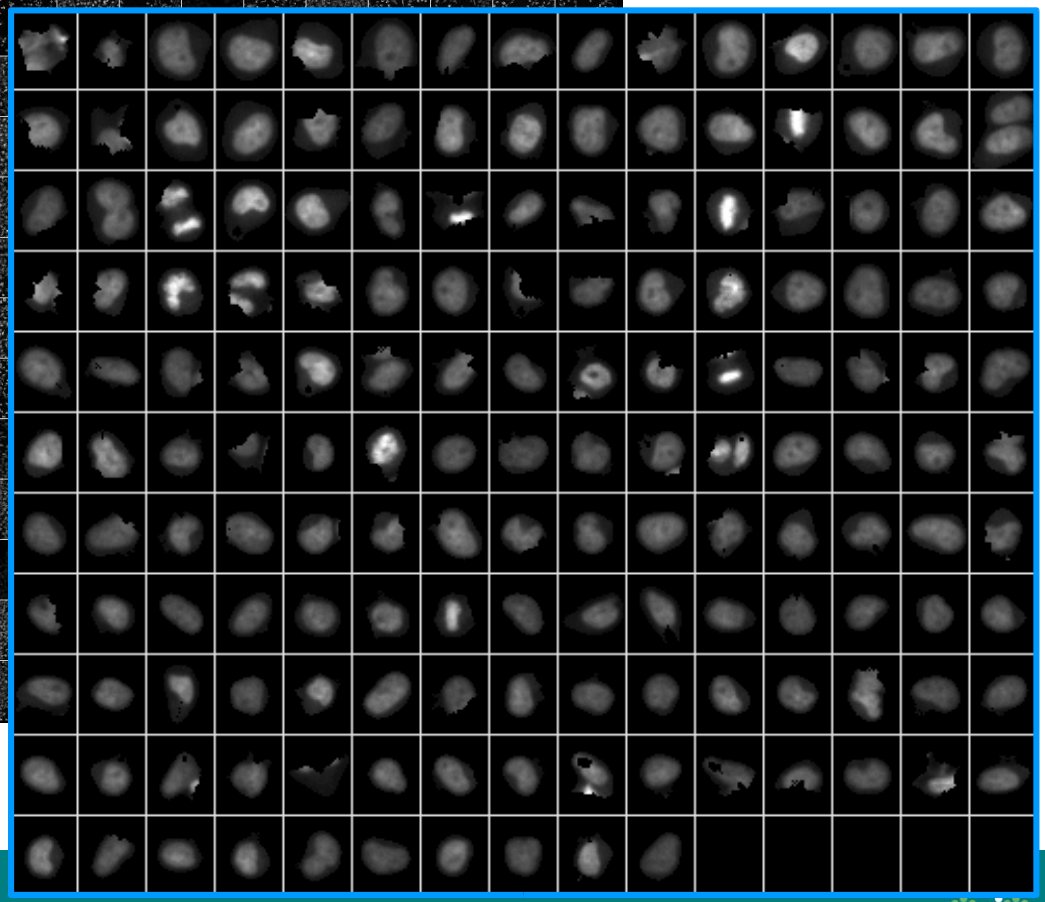


```
dm = distmap( thresh(nucl, 30, 30) )
seeds = watershed(dm, 1, 1)
mask = thresh(cell, 60, 60)
w  = watershed(distmap(mask), 2, 1)             ## yellow
vi = propagate(cell, seeds, mask, lambda=0)     ## red
v  = propagate(cell, seeds, mask, lambda=2e16) ## white
```

Some visualisation before we continue with the analysis



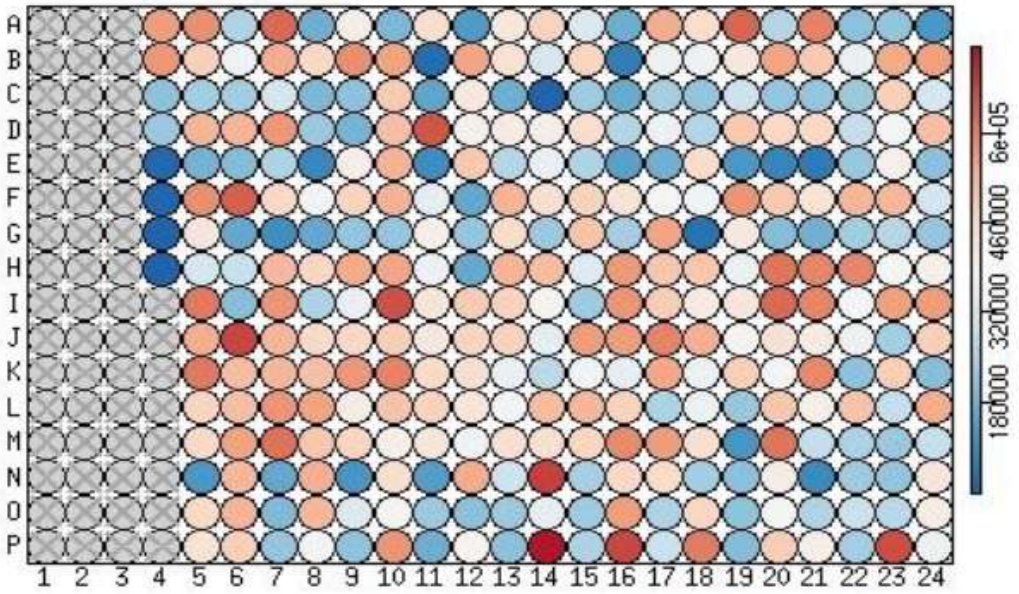
Thumbnail overview of one plate's images

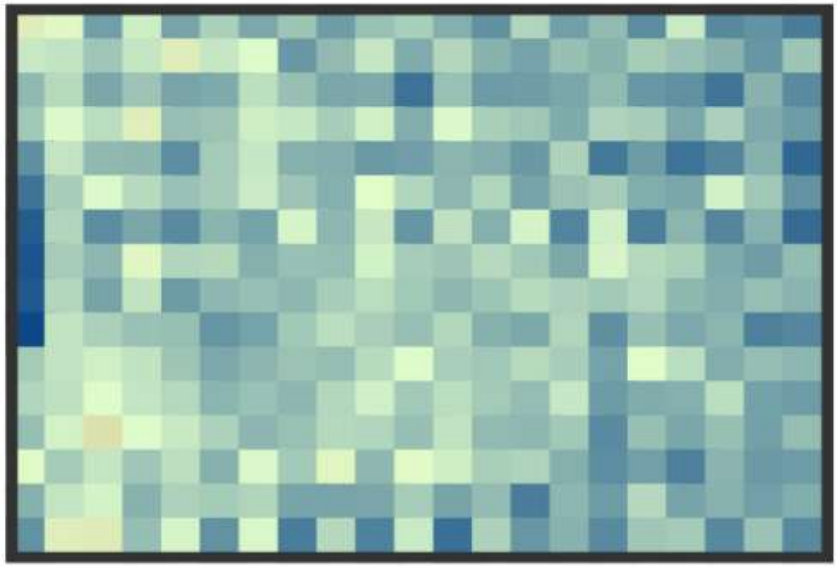


Gallery view of segmented objects of one well

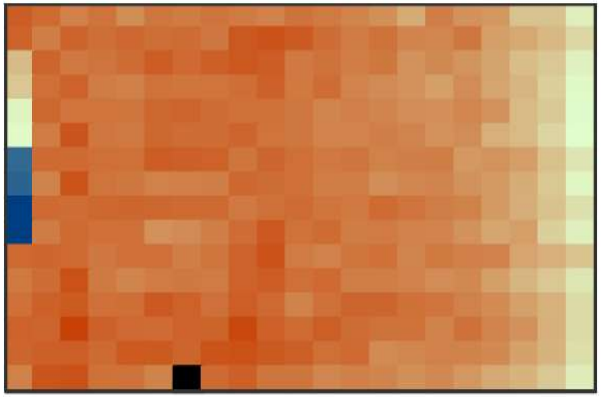
EMBL-EBI

Cell counts: comparable to luminenscence readings from the plate reader?

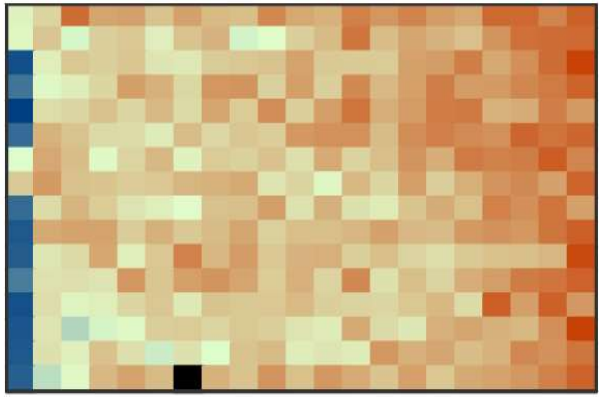




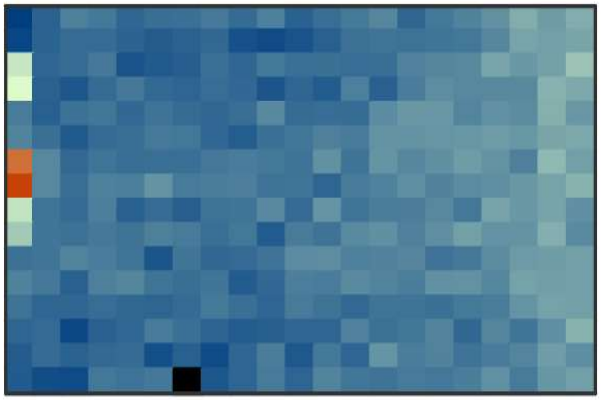
Within plate spatial trends - normalization and quality assessment



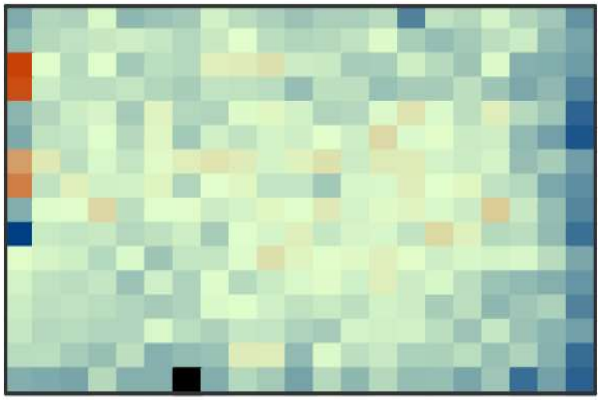
Number of cells



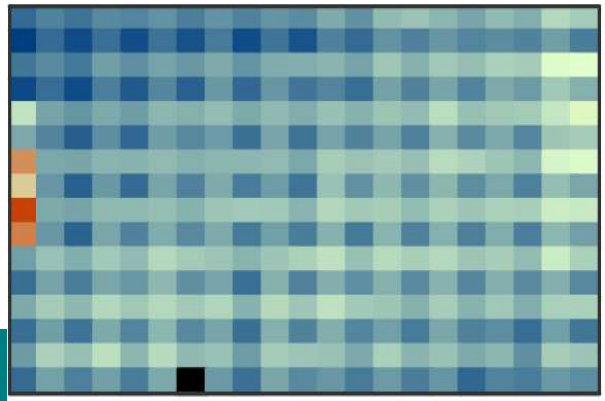
Acircularity



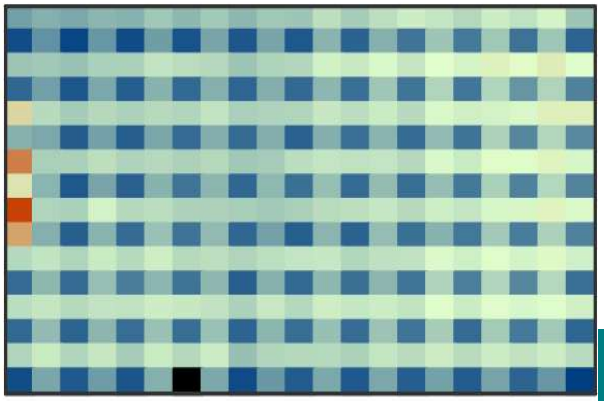
Cell size



Nuclear size

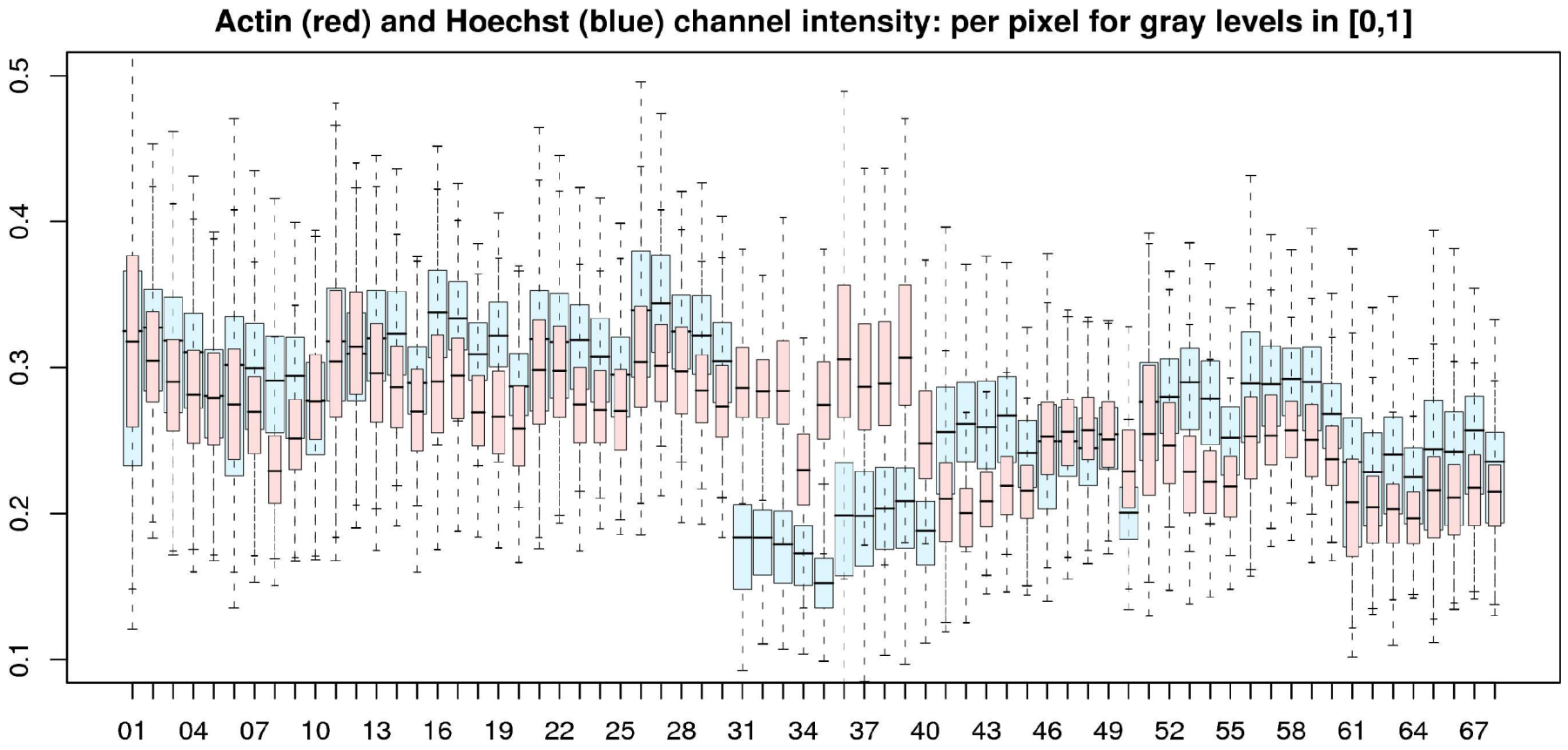


Actin intensity / p.pixel



Hoechst intensity / p.pixel

Batch effects



Normalization: Plate effects

Percent of control

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

k-th well
i-th plate

Normalized percent inhibition

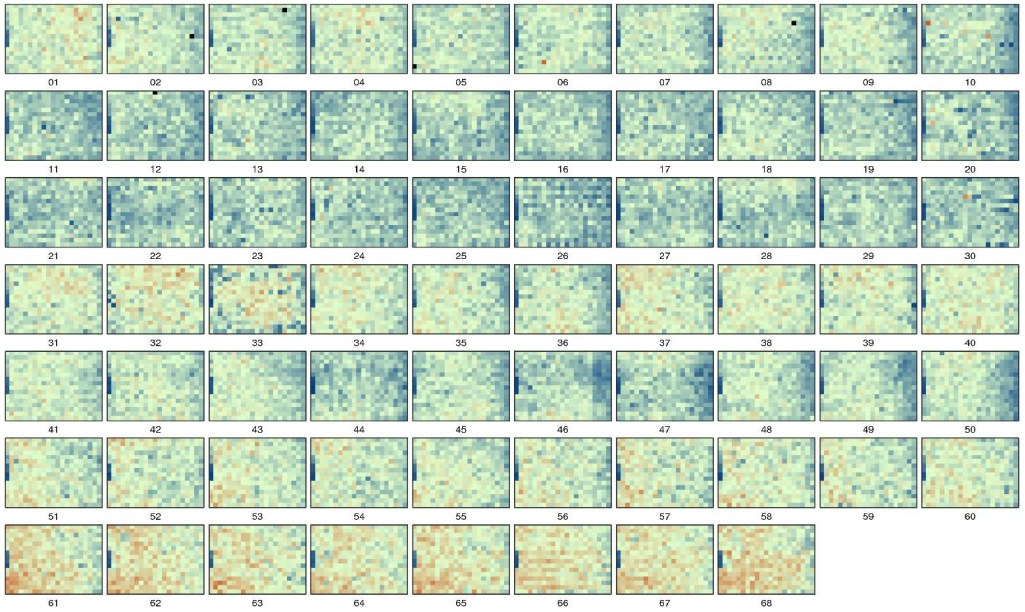
$$x'_{ki} = \frac{\mu_i^{pos} - x_{ki}}{\mu_i^{pos} - \mu_i^{neg}} ? 100$$

***z*-score**

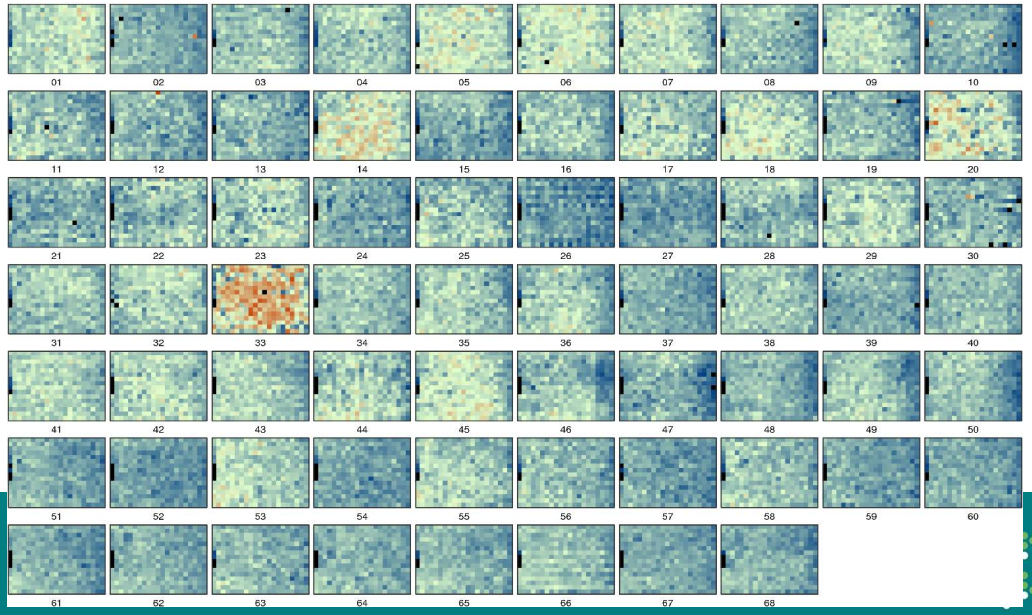
$$x'_{ki} = \frac{x_{ki} - \mu_i}{\sigma_i}$$

Long term drifts

Number of cells

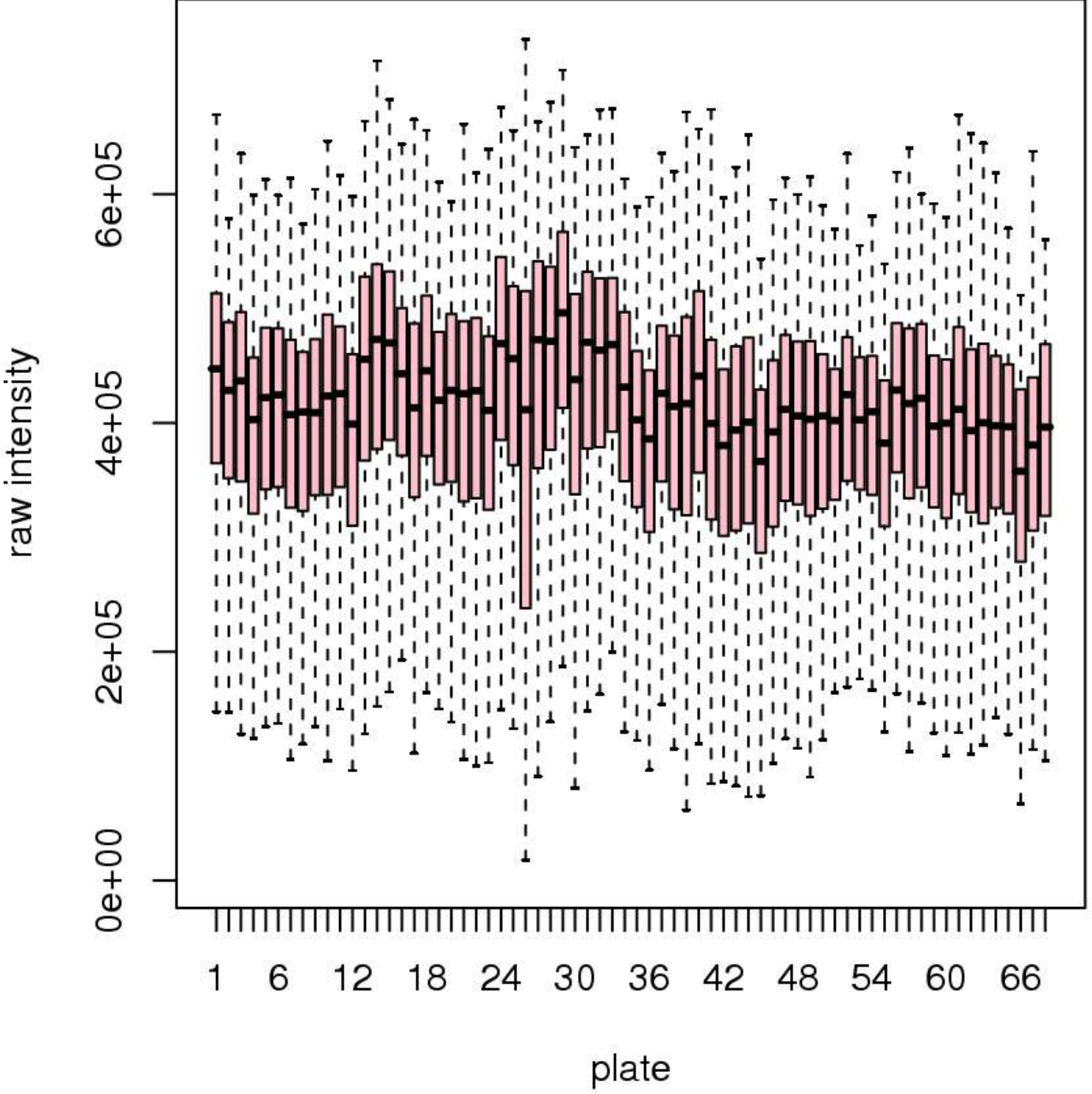


Number of cells / no. cells in negative controls in same plate

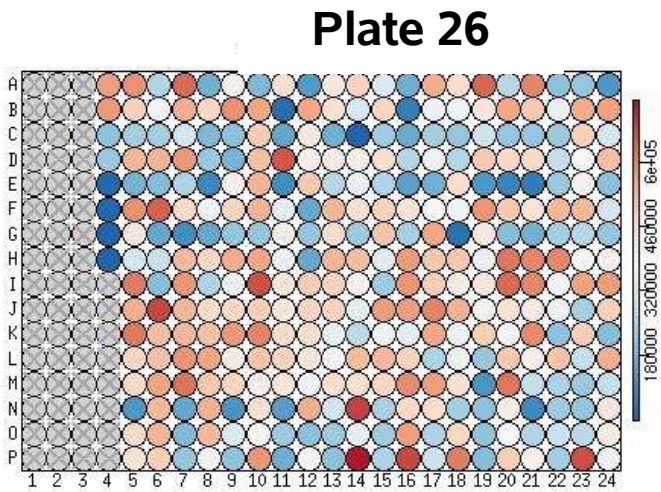


Dharmacon siARRAY library

Hek293 cells viability screen
Boutros Lab
DKFZ

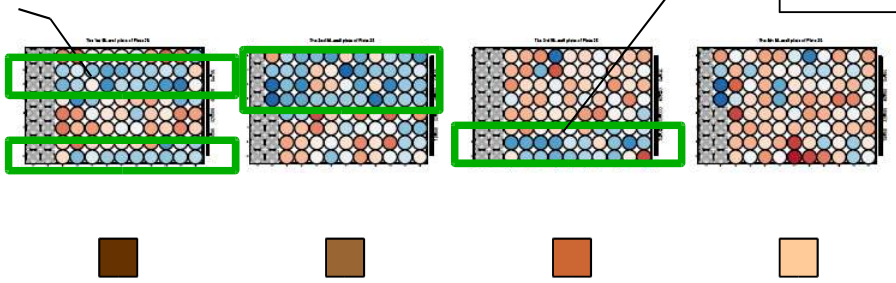


Normalization problem…
Too many hits

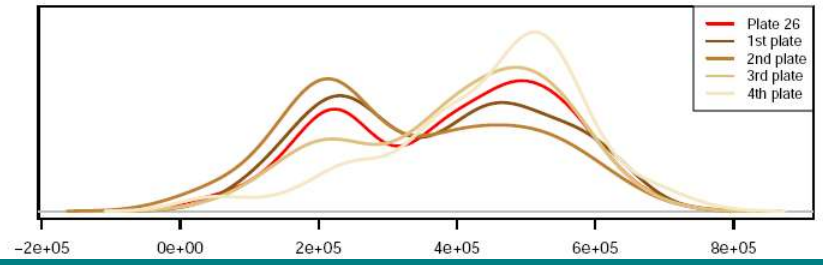


proteasome subunits
or components;
ATP/GTP-binding site motifs

like-Sm nucleoproteins and
ribosomal proteins



ribosomal proteins



How to estimate the normalization parameters?

From which data points:

- **Based on the intensities of the controls**

  if they work uniformly well across all plates

- **Based on the intensities of the samples**

  invoke assumptions such as "most genes have no effect", or "same distribution of effect sizes"

Which estimator:

mean *vs* median *vs* shorth

standard deviation *vs* MAD vs IQR

No universally optimal answer, it depends on the data.
In the best case, it doesn't matter.

Show imageHTS[3]

EMBL-EBI

Phenotype of interest: elongated cells

67 / F13

GPR124

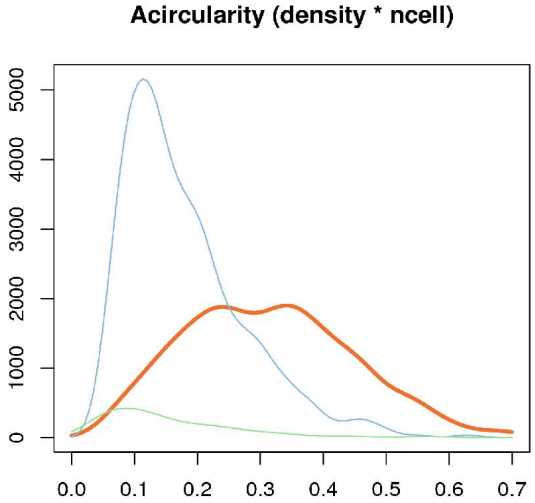
Homo Sapiens probable G protein-coupled receptor 124 precursor (tumor endothelial marker 5)

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

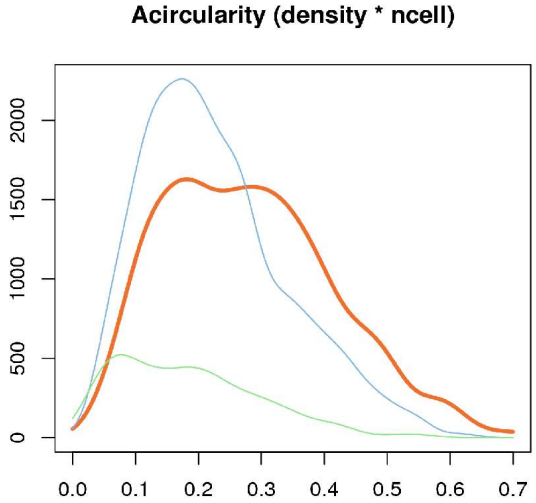
Homo Sapiens azurocidin precursor (cationic antimicrobial protein CAP37), heparin-binding protein) (HBP)

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

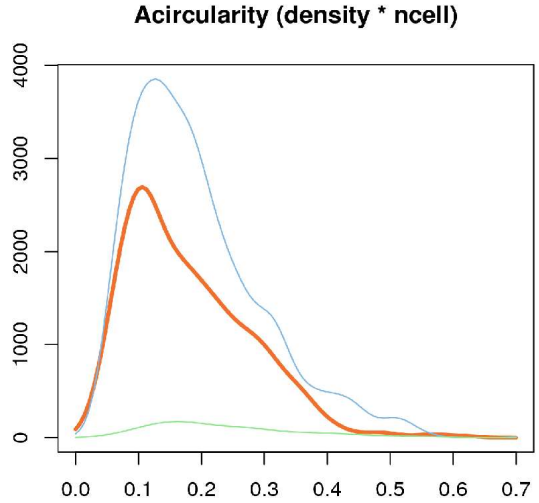
Homo sapiens family with sequence similarity 79, member B (FAM79B), mRNA

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

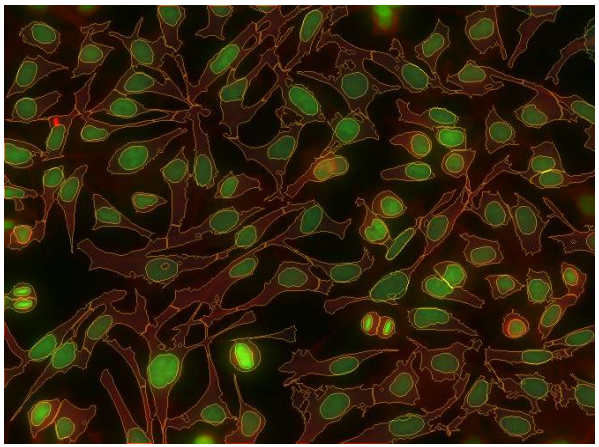
Homo Sapiens probable G protein-coupled receptor 124 precursor (tumor endothelial marker 5)

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

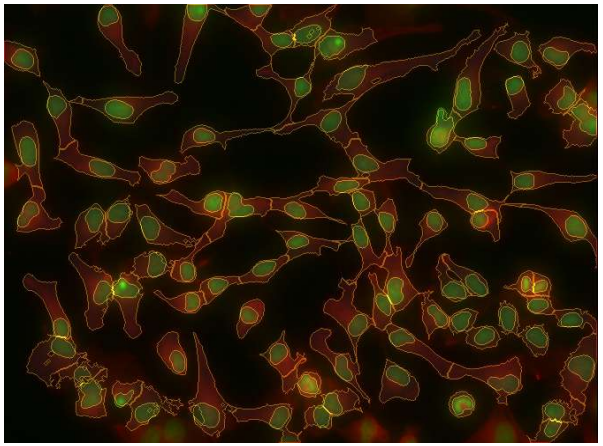
Homo Sapiens azurocidin precursor (cationic antimicrobial protein CAP37), heparin-binding protein) (HBP)

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

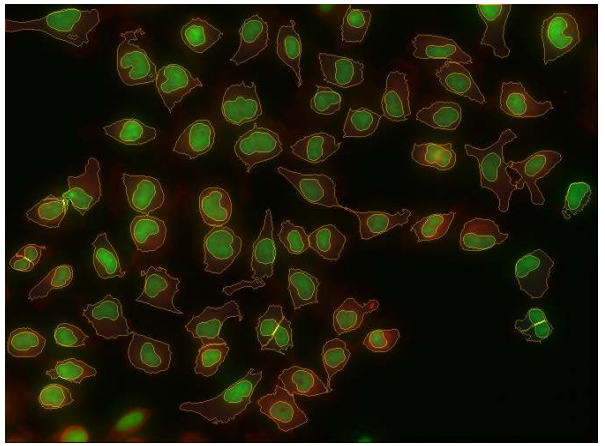
Homo sapiens family with sequence similarity 79, member B (FAM79B), mRNA

Number of cells:

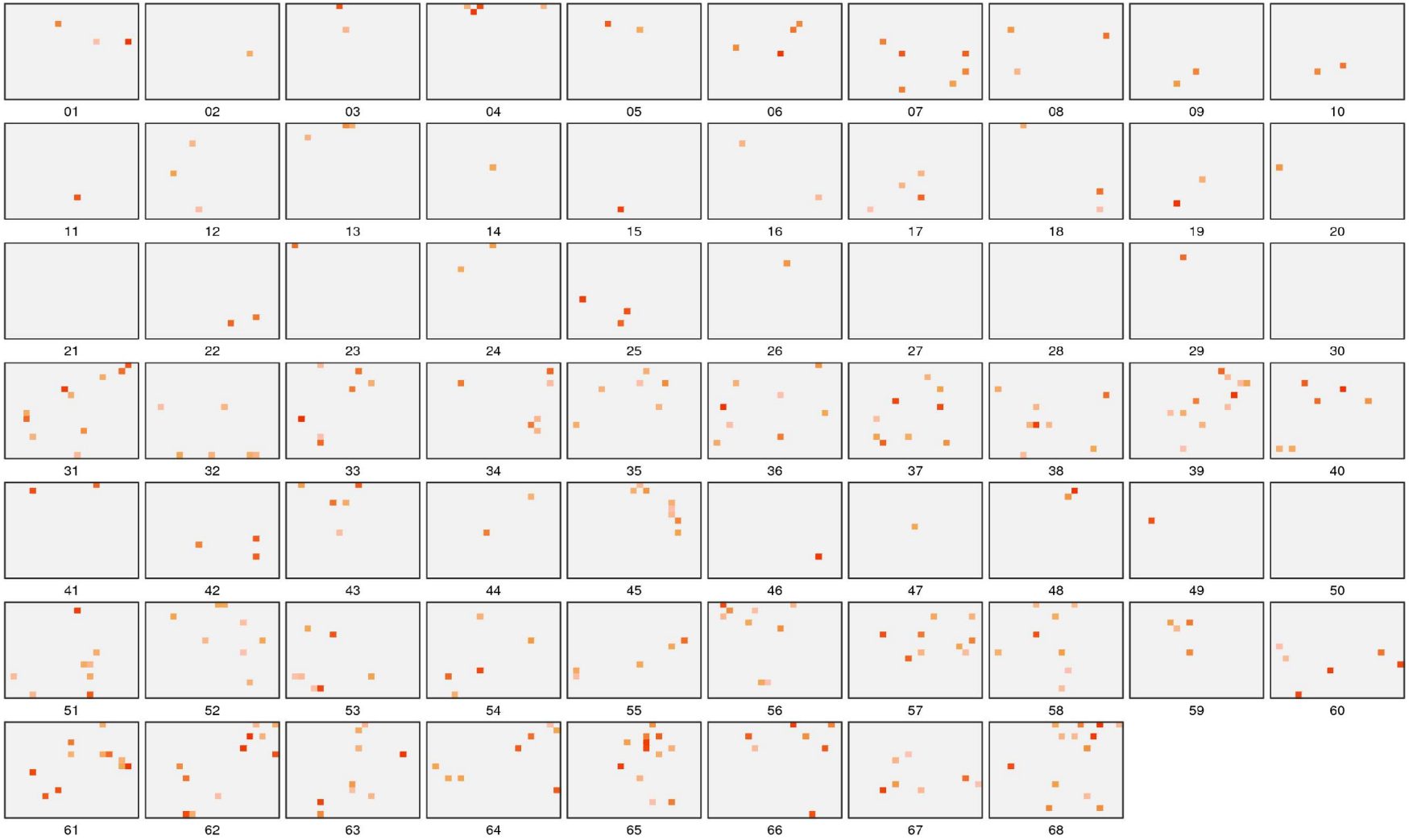
Run 1: 281 / NC:417.5
Run 2: 274 / NC:432.5

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: cell population dynamics models for clustering and classification of genes and phenotypes

Gregoire Pau (EBI)

with
Thomas Walter
Beate Neumann
Jan Ellenberg (EMBL)

Mitocheck time lapse data

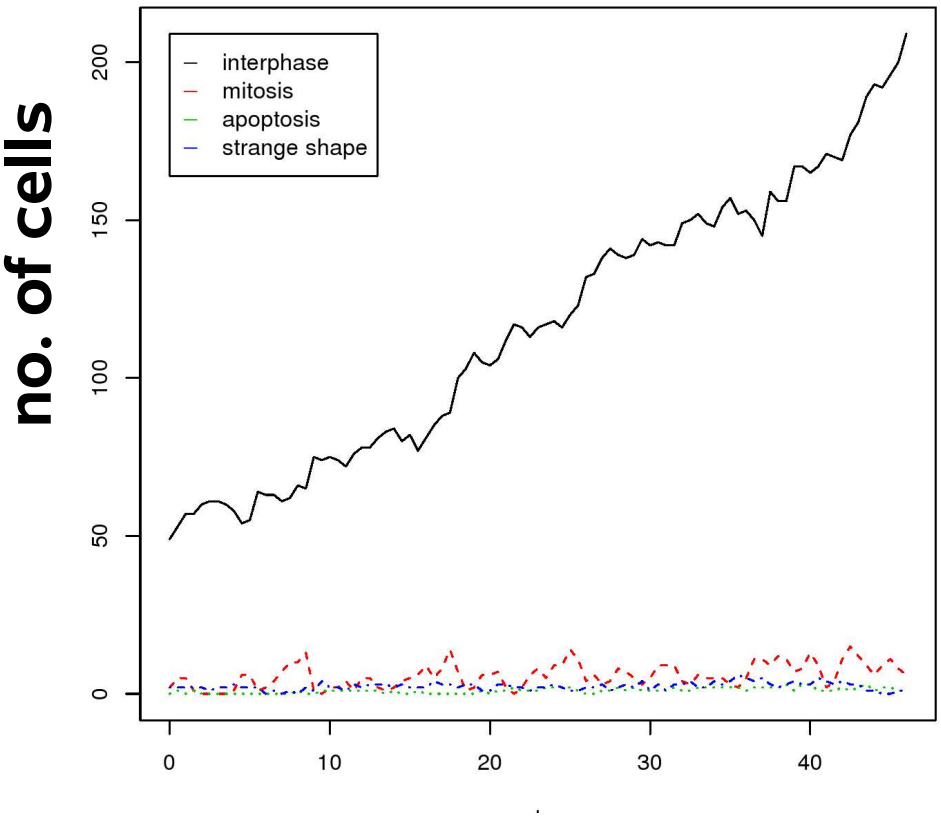
Live cell time-lapse imaging

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

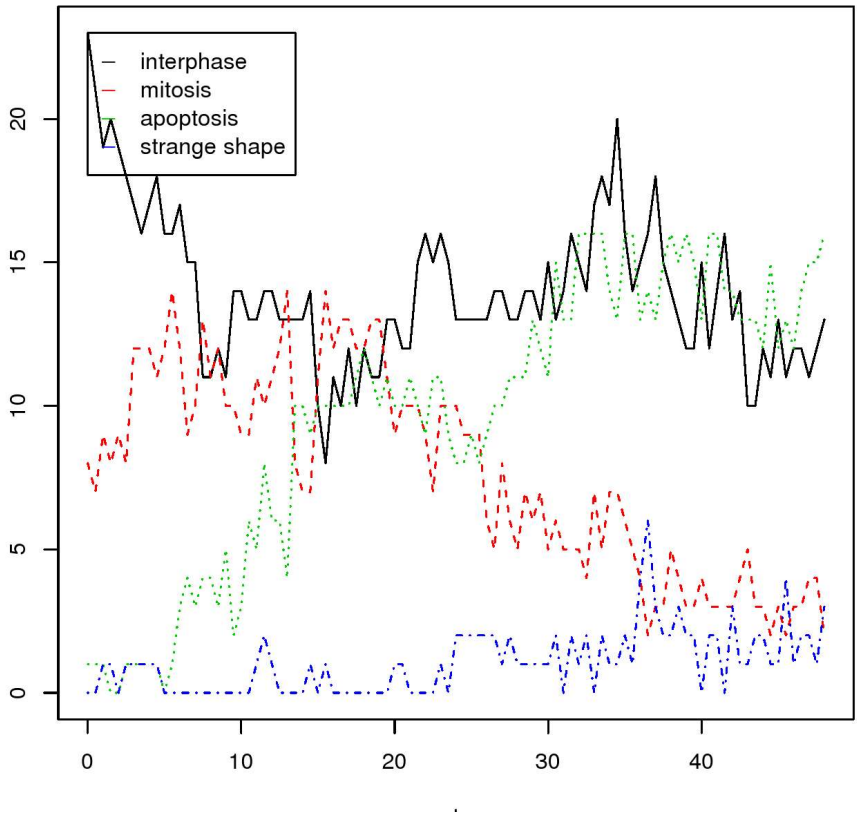
Experimental output

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

Examples



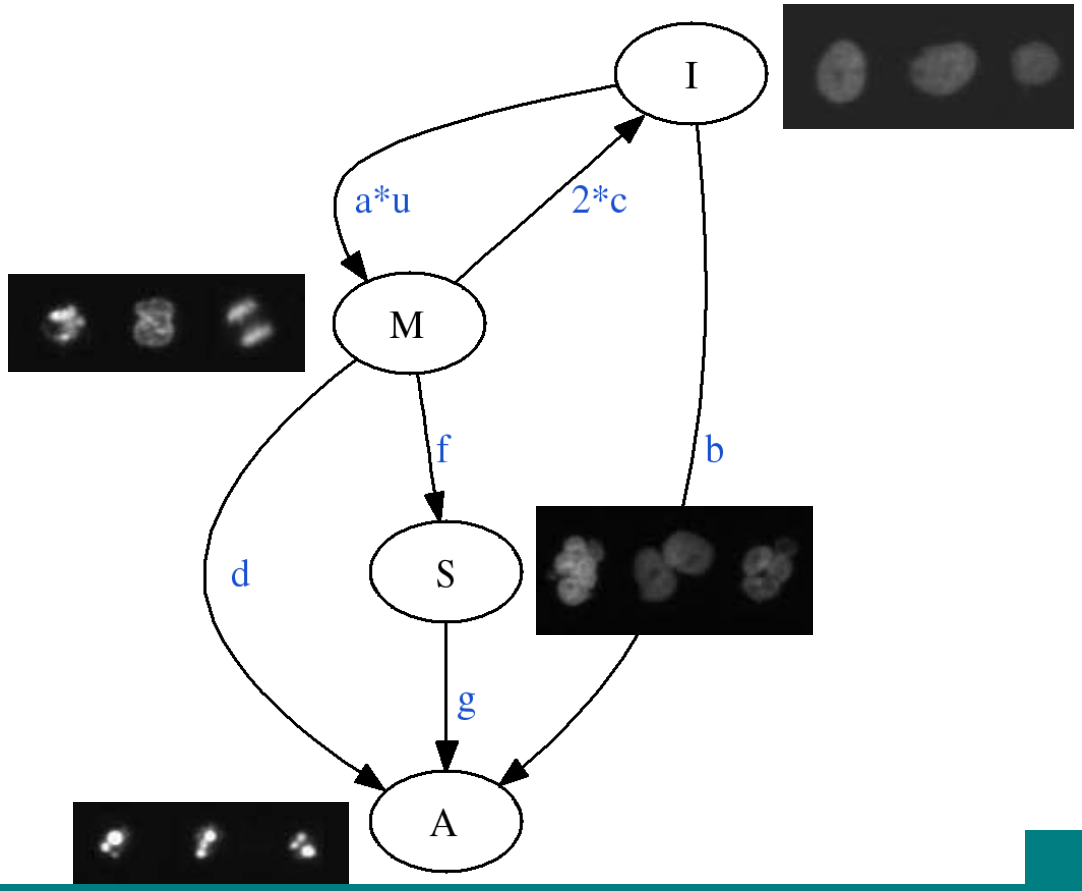
Wild type (no knockdown)



Mitotic arrest phenotype (KIF11 knockdown)

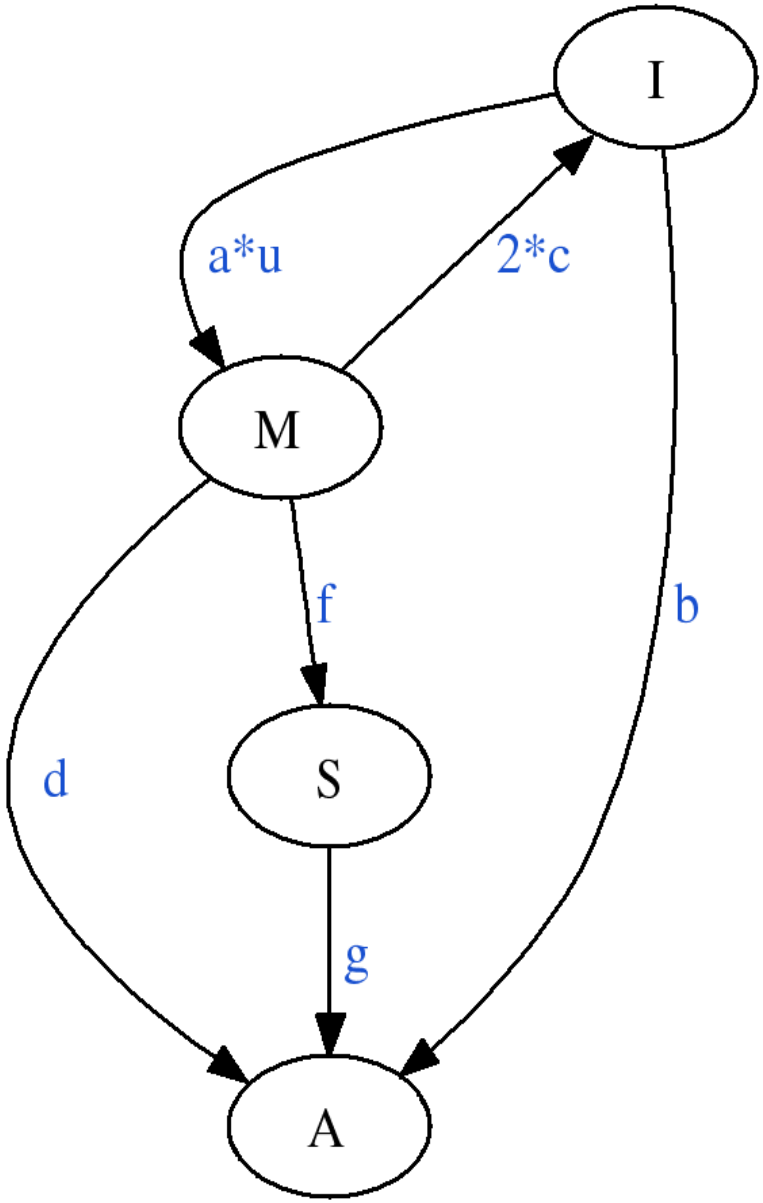
Non-linear Ordinary Differential Equation Model

- Temporal dynamics of cell state change on the population average level
- 7 kinetic parameters *a, b, c, d, e, f, g* and 4 initial conditions



I – Interphase
M – Mitosis
A – Apoptosis
S – Shape

Non-linear Ordinary Differential Equation model

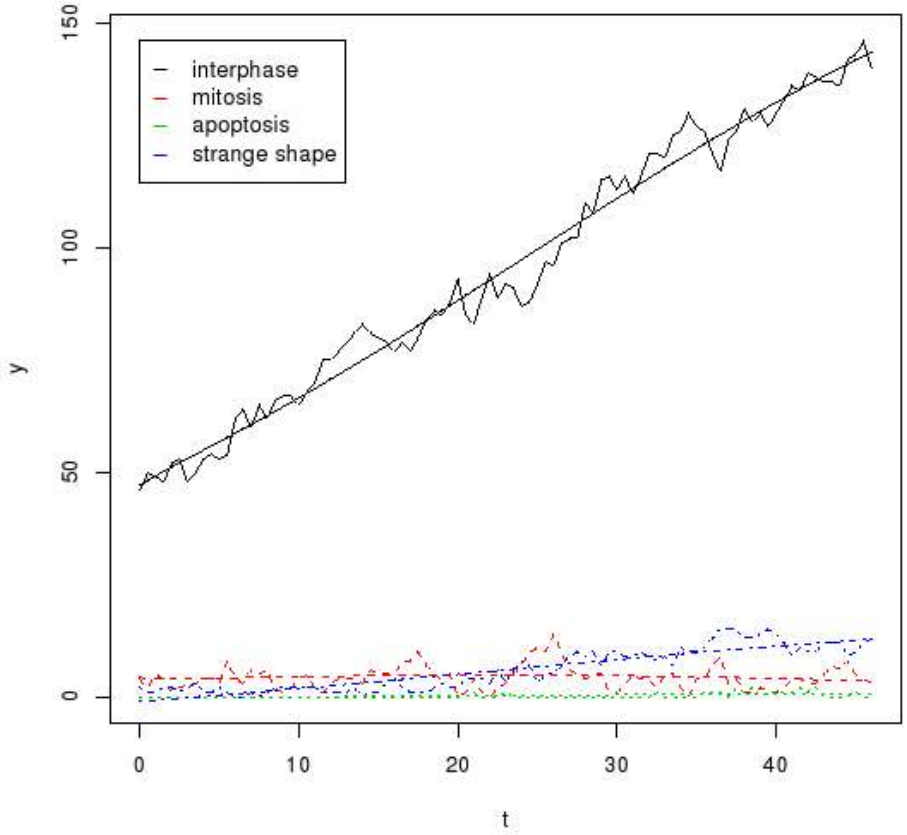


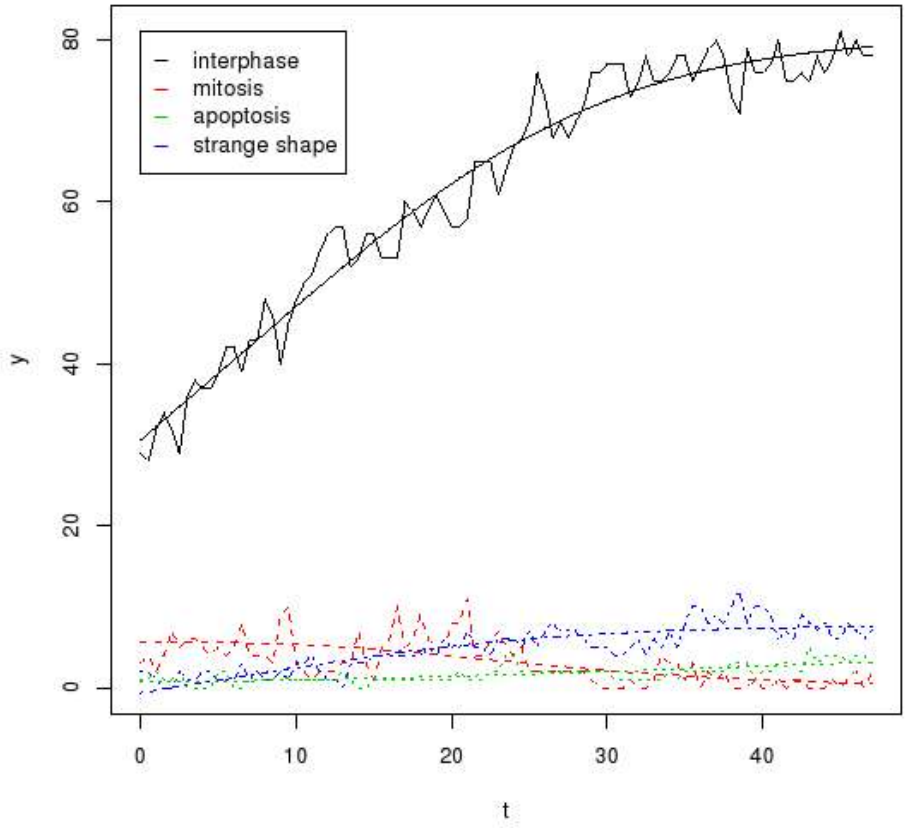
$$
\begin{cases}
\dfrac{dn_I}{dt} = -aun_I - bn_I + 2cn_M \\
\dfrac{dn_M}{dt} = aun_I - cn_M - dn_M - fn_M \\
\dfrac{dn_S}{dt} = fn_M - gn_S \\
\dfrac{dn_A}{dt} = bn_I + dn_M + gn_S \\
\dfrac{du}{dt} = eun_I
\end{cases}
$$

+ 4 initial conditions

Fitting examples

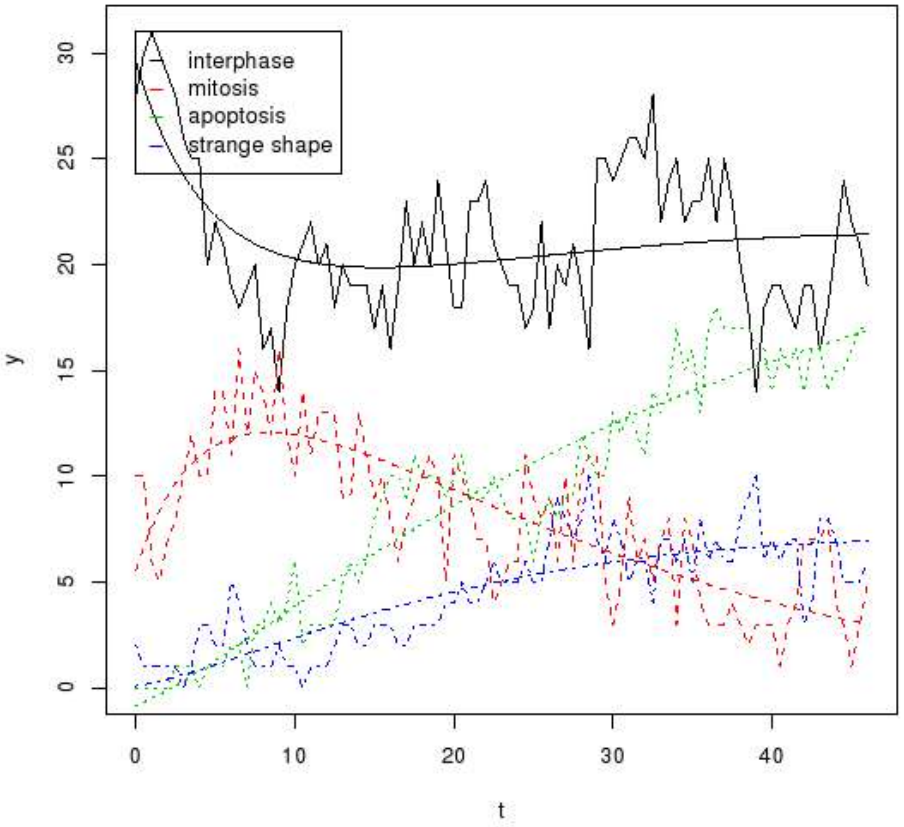
No knockdown

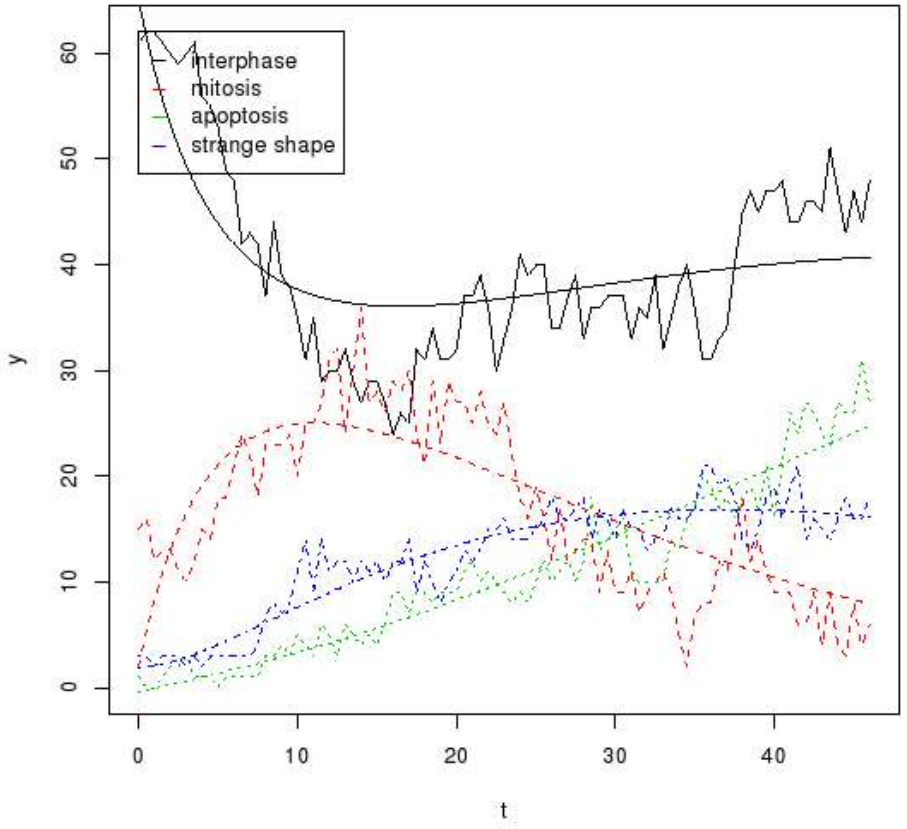




Fitting examples

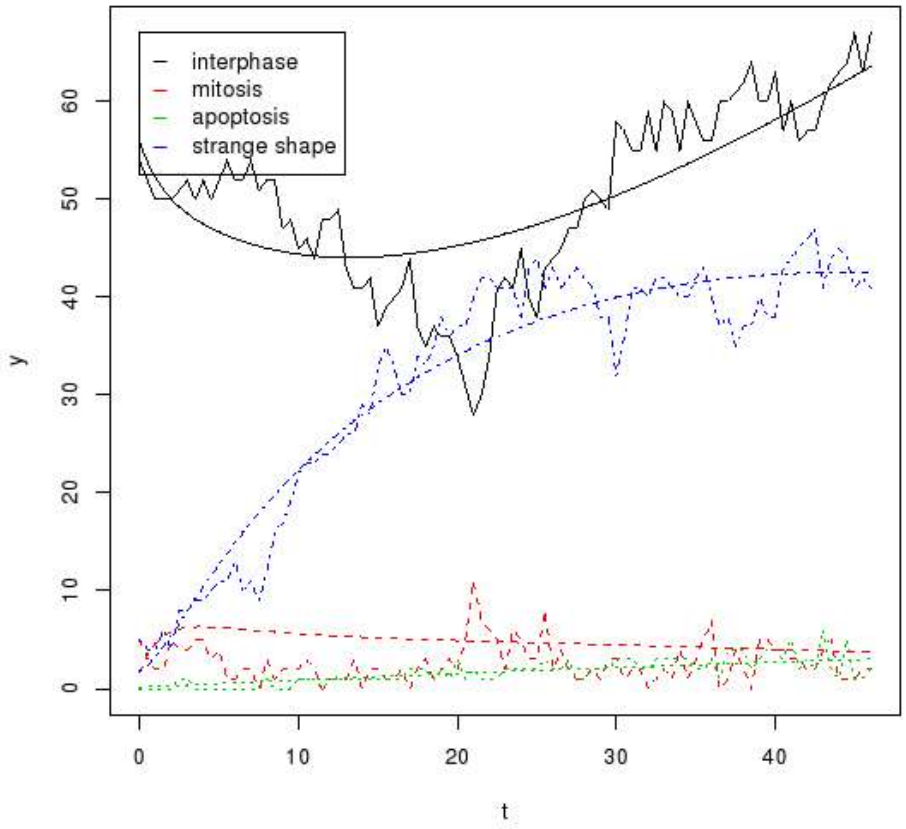
KIF11 knockdown (mitotic arrest phenotype)

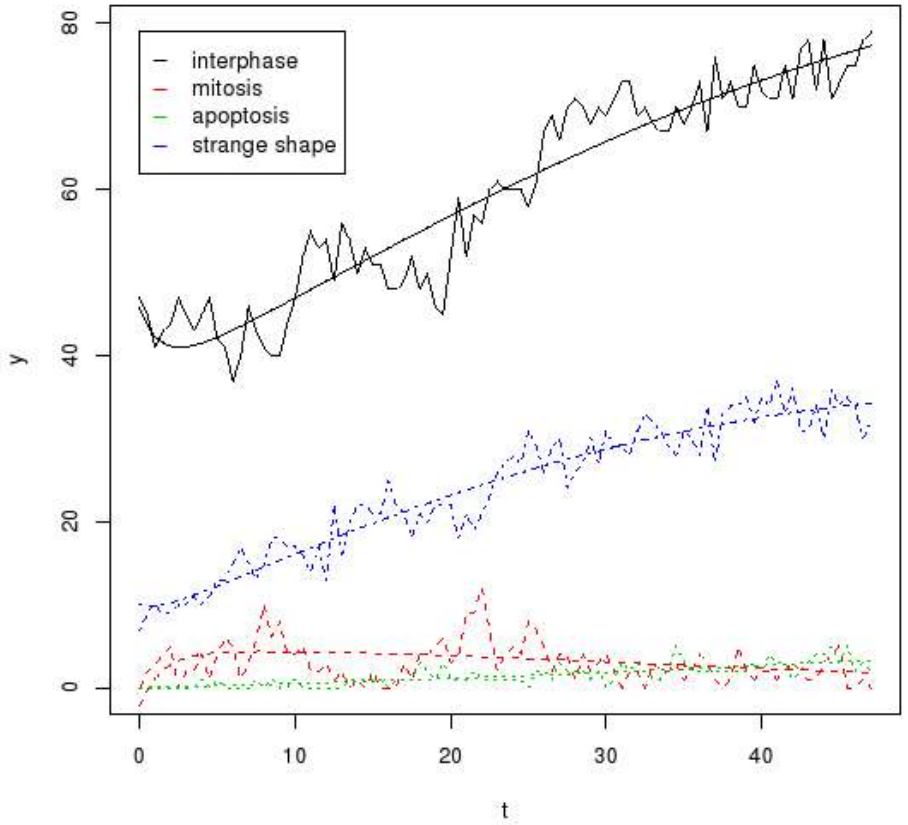




Fitting examples

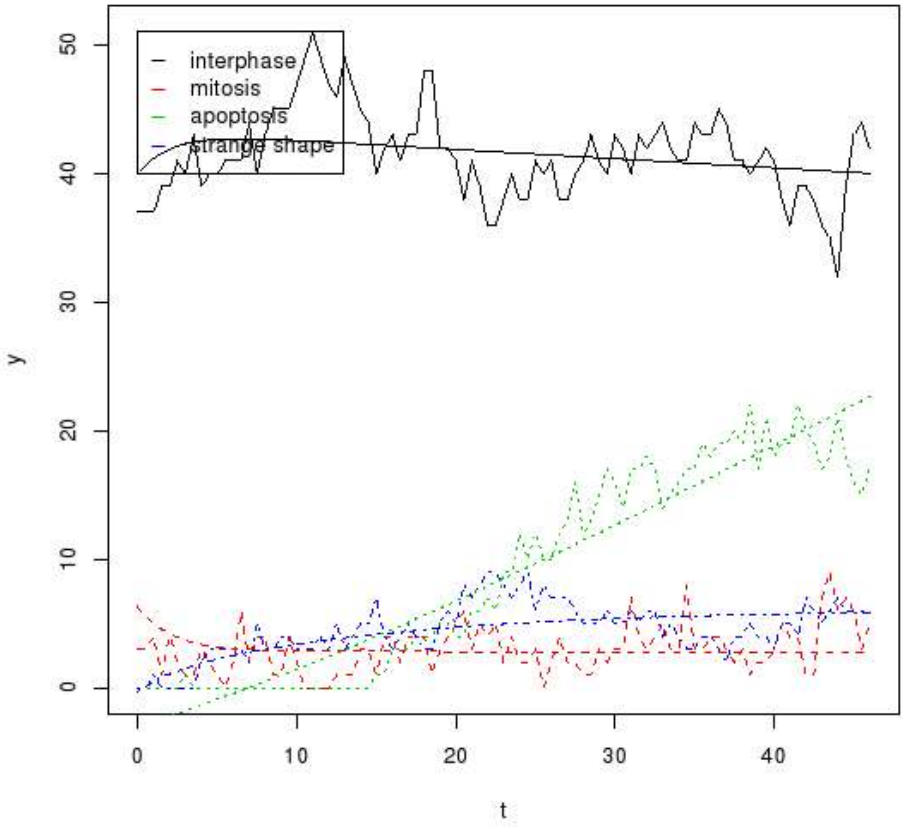
INCENP knockdown (shape phenotype)

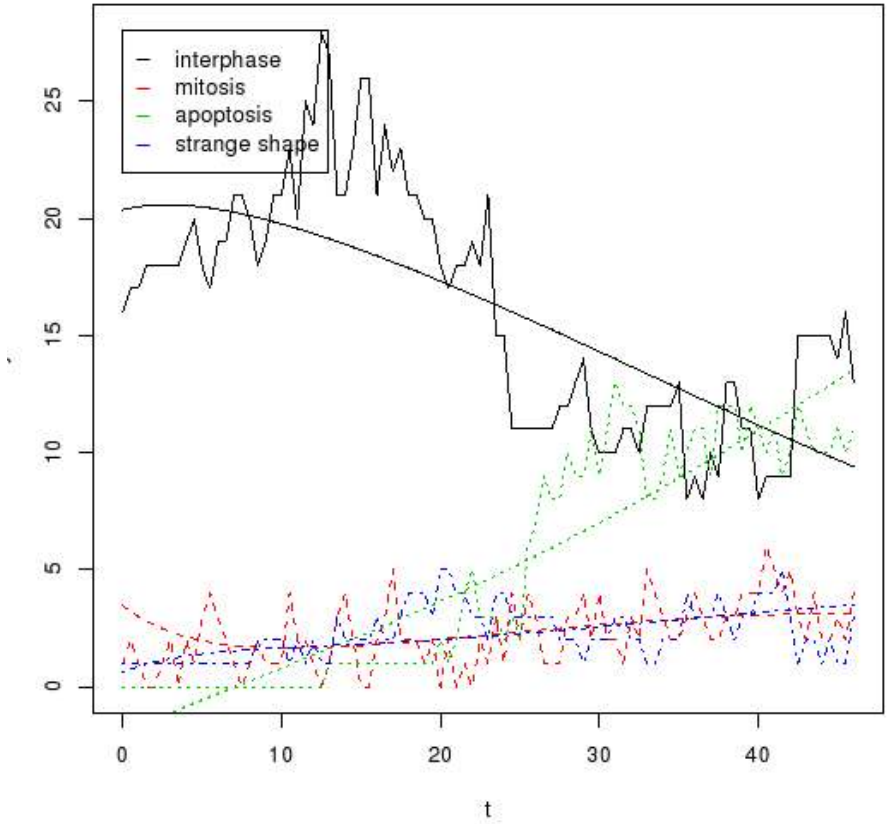




Fitting examples

bCOP knockdown (apoptotic phenotype)





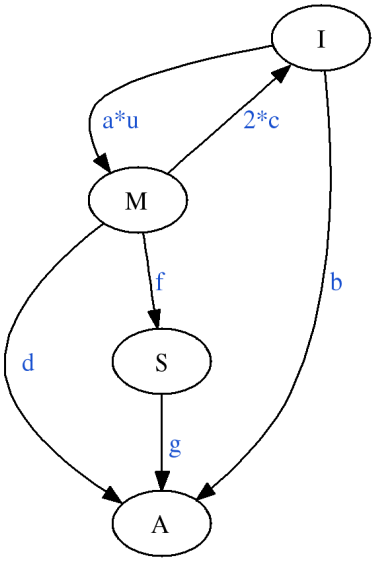
β-coatomer protein

Expected parameters for known phenotypes

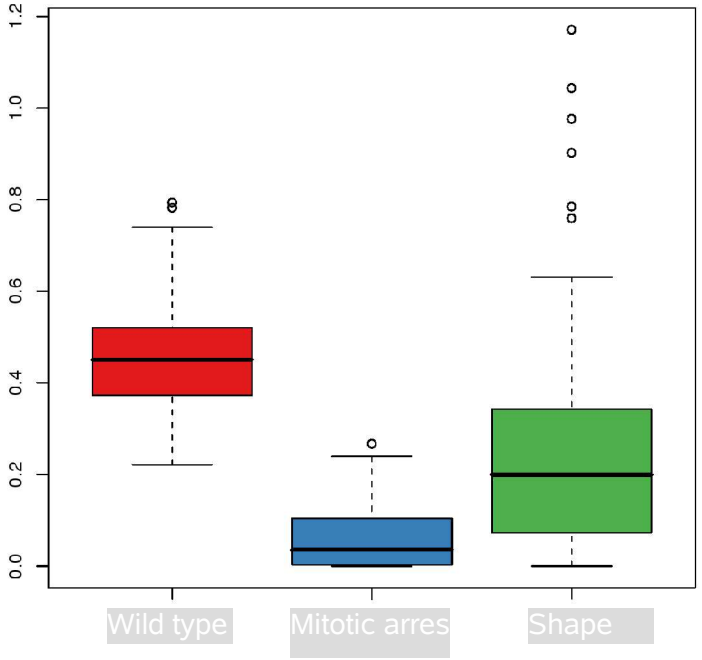
Mitotic arrest: low c & high d

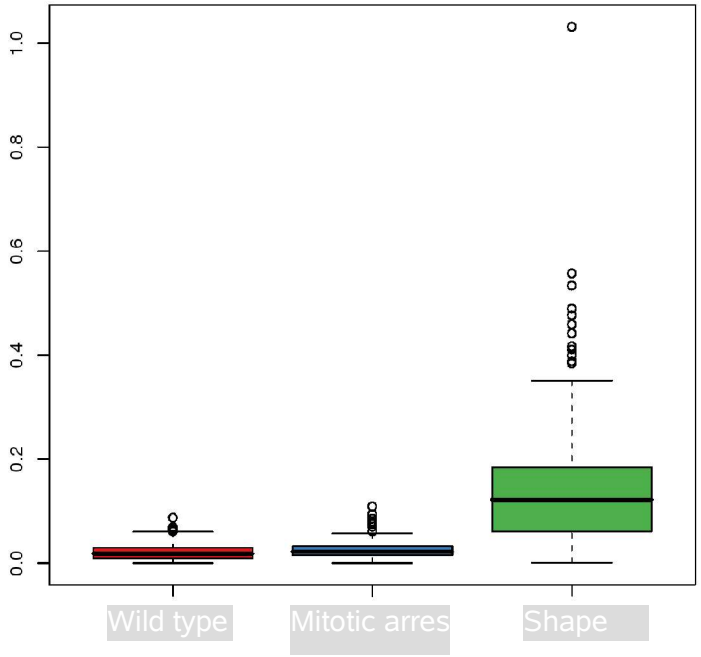
Shape: high f

Apoptotic: high b, d or g

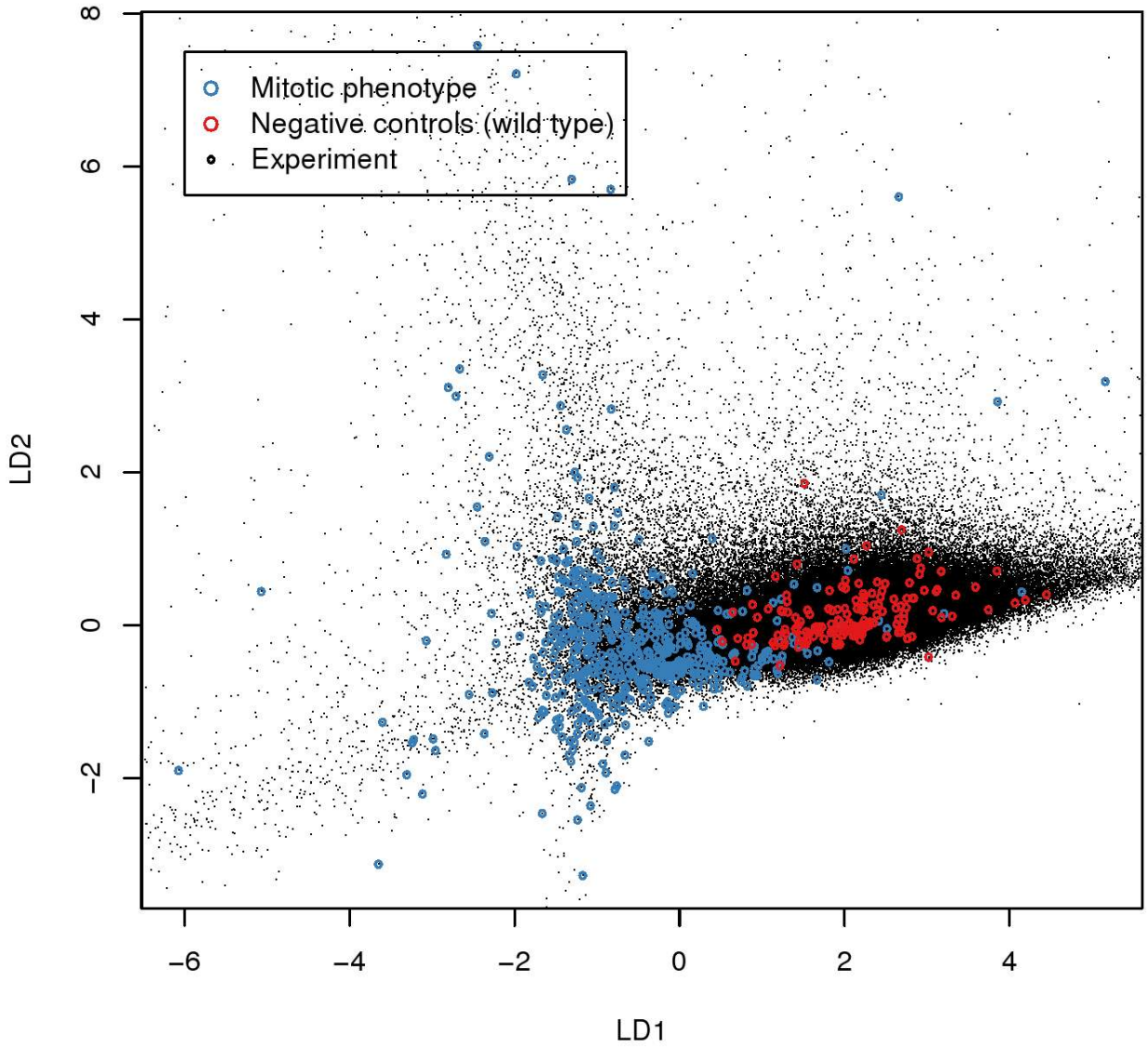


c -marginal distributions- f





Data visualisation in 2D linear discriminant component space

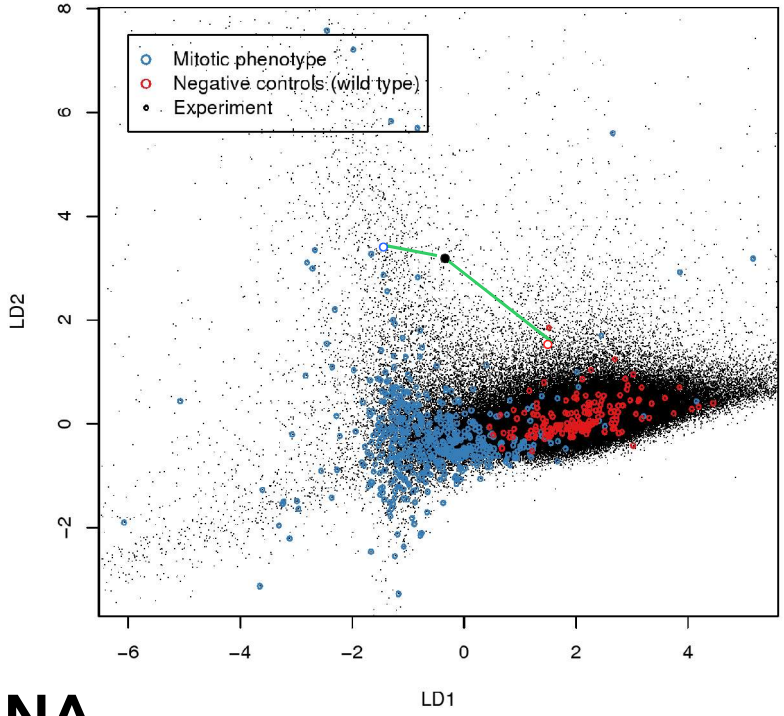


EMBL-EBI

Automatic phenotyping: mitotic genes

Per spot

$$d = \frac{\text{distance to nearest mitotic spots}}{\text{distance to nearest non-mitotic spots}}$$



Per each siRNA

- median of replicates

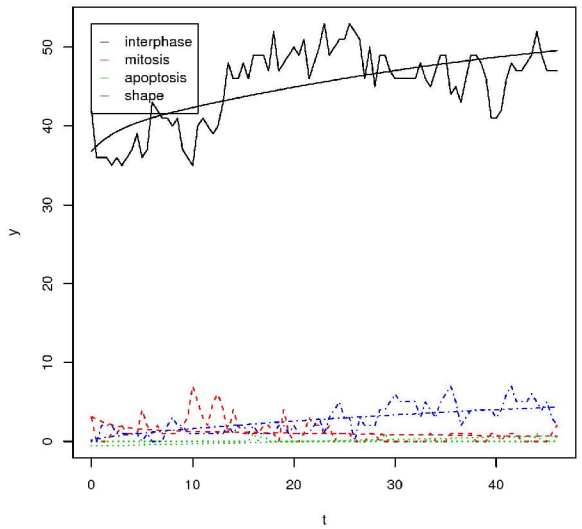
Ranking by mitotic score

New mitotic hits:

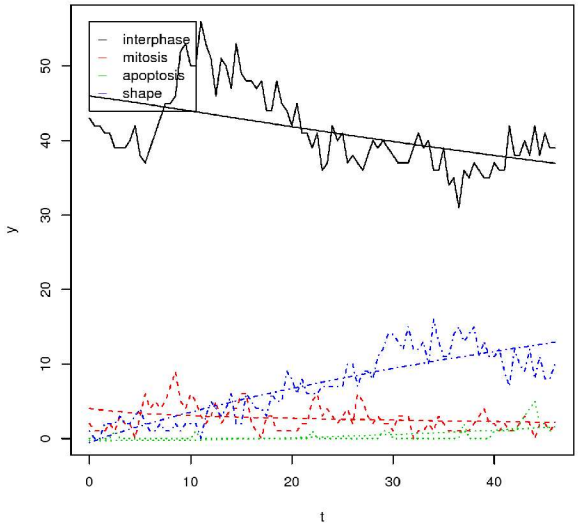
- ENSG00000197652
- ENSG00000178882
- TXNL4A
- ASB2
- ENSG00000188329
- MTP18
- GPR19
- MGAT4B
- ENSG00000140006

ENSG00000197652

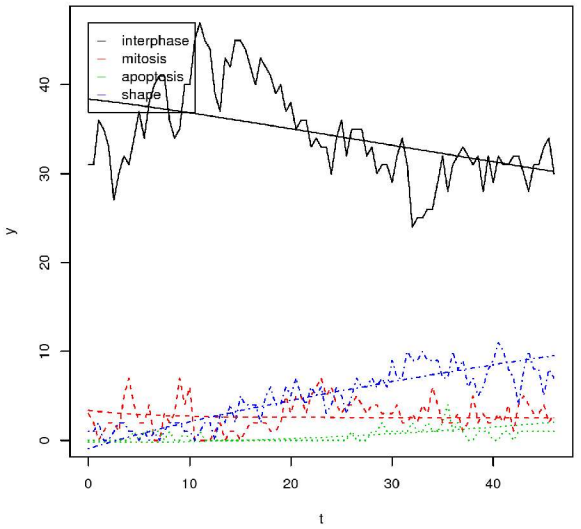
3 siRNAs



score=2.324



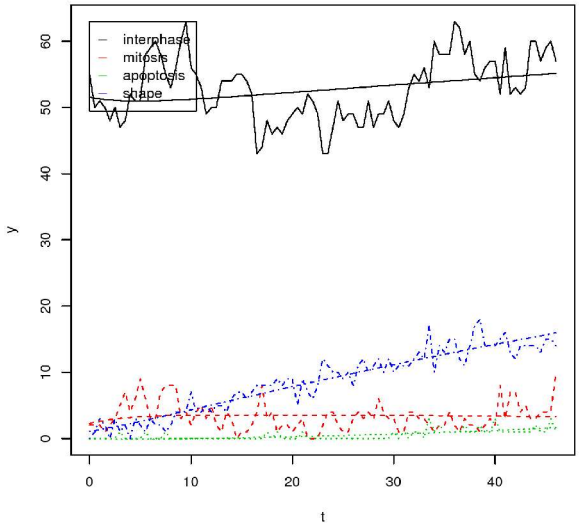
score=0.1041



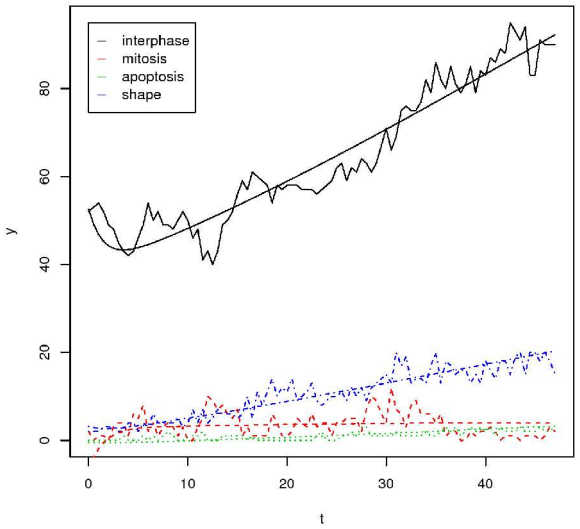
score=0.1424

GPR19

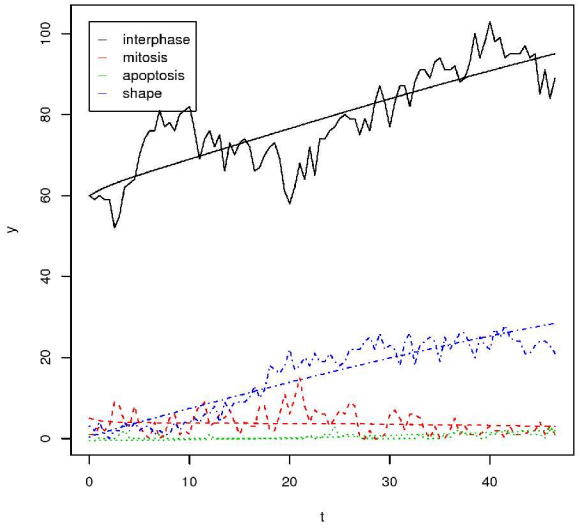
3 siRNAs



score=0.2836



score=0.1869



score=0.078

Conclusions

Specificity and biological significance of gene perturbation-phenotype associations **can be increased by:**

Combinatorial perturbation

Directed variations of assay conditions

Specific phenotype readouts

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

Tools in Bioconductor (et al.)

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
Ligia Bras
Tineke Casneuf
Tony Chiang
Audrey Kauffmann
Gregoire Pau
Oleg Sklyar
Jörn Tödling

DKFZ

Florian Fuchs
Thomas Horn
Dierk Ingelfinger
Sandra Steinbrink
Michael Boutros

Cristina Cruciat

Florian Hahne
Stefan Wiemann

EMBL

Lars Steinmetz
Eugenio Mancera
Zhenyu Xu
Julien Gagneur
Fabiana Perocchi

Jan Ellenberg
Thomas Walter
Beate Neumann

Bioconductor

Robert Gentleman
Seth Falcon
Martin Morgan
Rafael Irizarry
Vince Carey

EMBL-EBI