High-throughput image analysis with EBImage

Gregoire Pau

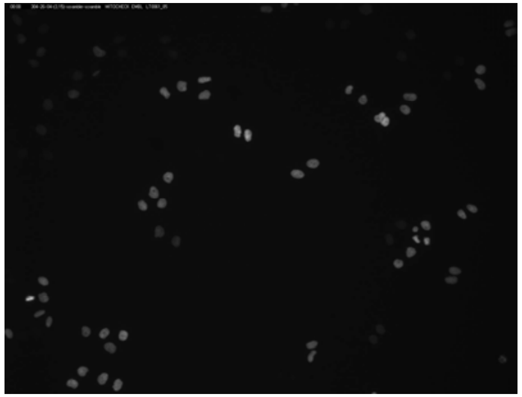
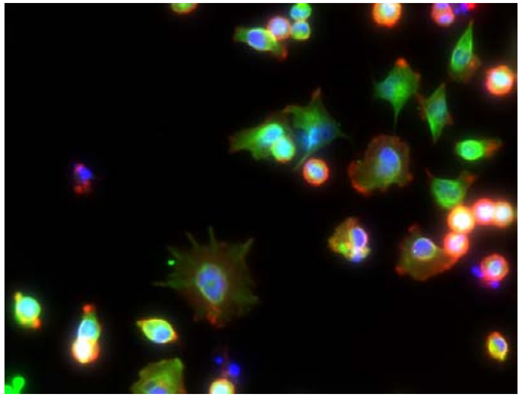
EMBL Heidelberg

gregoire.pau@embl.de

EMBL European Molecular Biology Laboratory

EBImage

- Fast and user-friendly image processing toolbox for R
- Provides functionality for
  - Reading/writing/displaying images
  - Image processing (pixel arithmetic, filtering, geometric transform)
  - Object segmentation
- Goal
  - Preprocess multidimensional images
  - Automated extraction of quantitative descriptors from microscope images

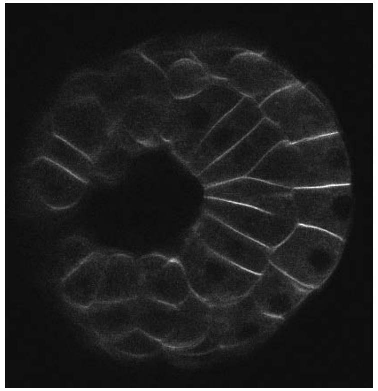
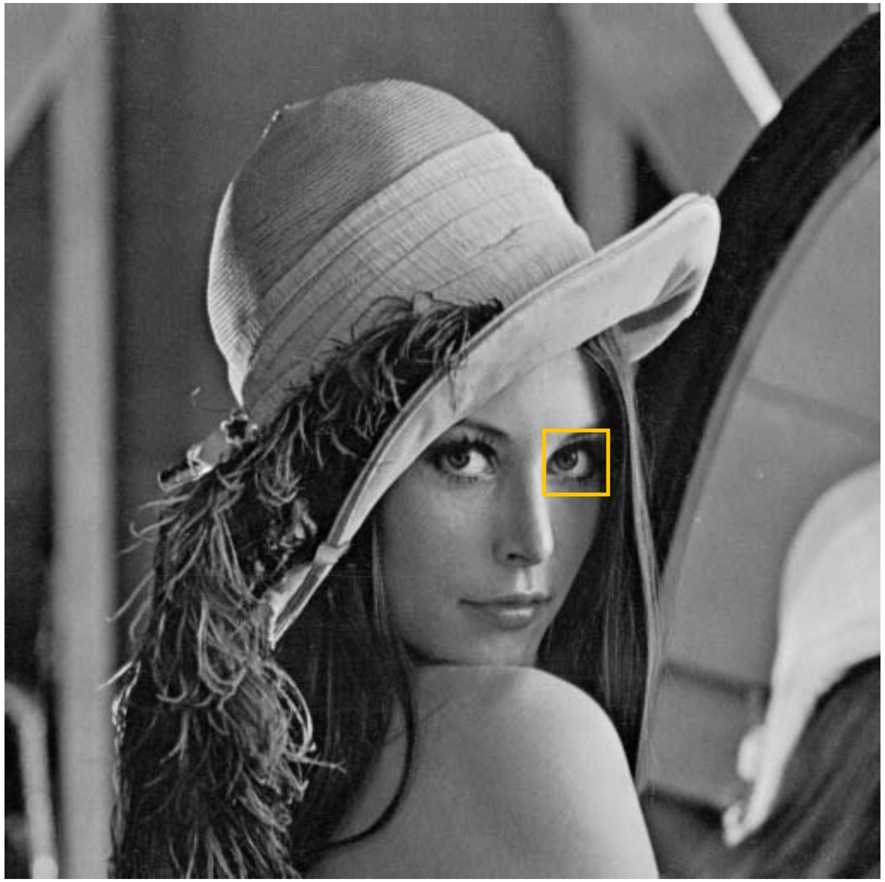


Image representation

- Multidimensional array of intensity values
- Seamless integration with R's native arrays

21	20	21	28	43	53	67	54
12	31	30	41	52	71	98	78
11	14	33	49	72	110	133	144
12	19	29	39	57	74	121	100
16	21	28	31	59	74	98	74
18	23	27	38	50	61	62	49
17	19	24	39	42	48	47	52
16	15	23	37	41	38	36	41

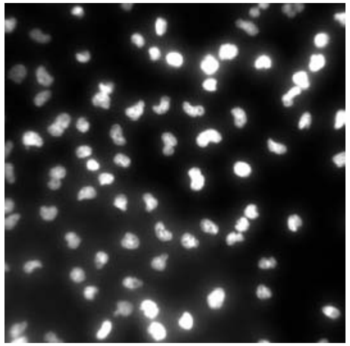


Lena: 512x512 matrix

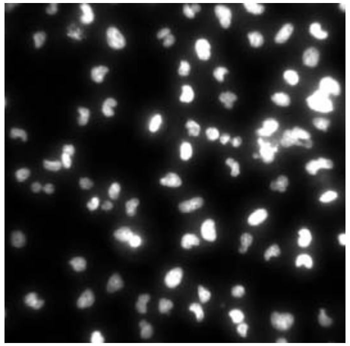
Image representation

- Multidimensional array
  - 2 first dimensions: spatial dimensions
  - Other dimensions: replicate, color, time point, condition, z-slice…

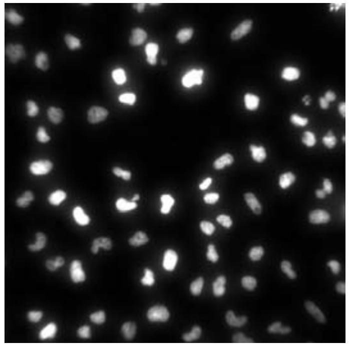
Nuclei
4 replicates



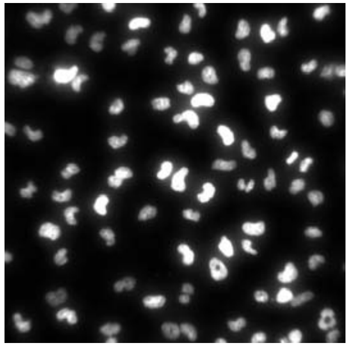
r0



r1



r2



r3

Lena
3 color
channels

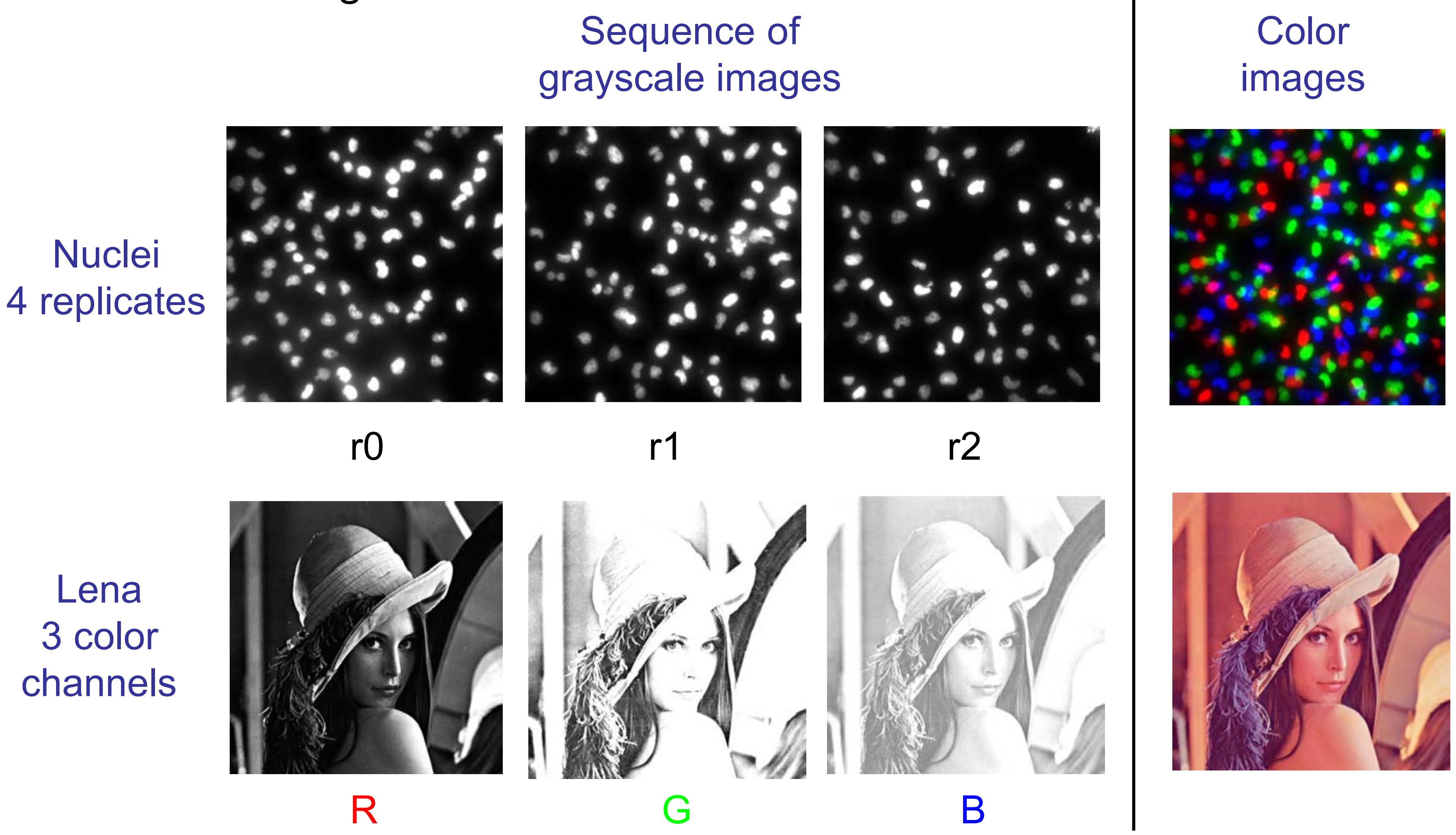
R

G

B

Image rendering

- Rendering dissociated from representation
- 2 rendering modes



Sequence of grayscale images

Color images

Nuclei 4 replicates

r0

r1

r2

Lena 3 color channels

R

G

B

IO

- Functions readImage(), writeImage()
  - Reads an image, returns an array
  - Supports more than 80 formats (JPEG, TIFF, PNG, GIF, …)
  - Supports HTTP, sequences of images

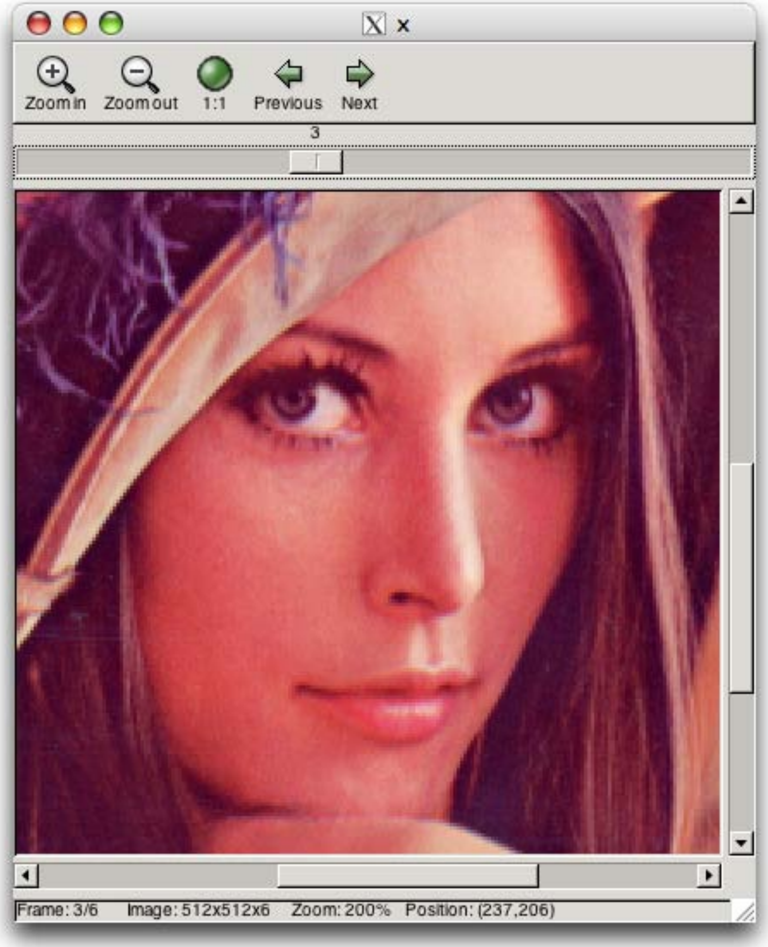
- Example: format conversion

```
library(EBImage)
x = readImage('sample-001-02a.tif')
writeImage(x, 'sample-001-02a.jpeg', quality=95)
```

Display

- Function display()
  - GTK+ interactive: zoom, scroll, animate
  - Supports RGB color channels and sequence of images

```
x = readImage('lena.png')
display(x)
```

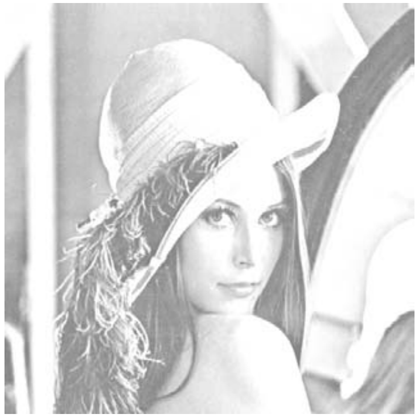


Frame: 3/6 Image: 512x512x6 Zoom: 200% Position: (237,206)

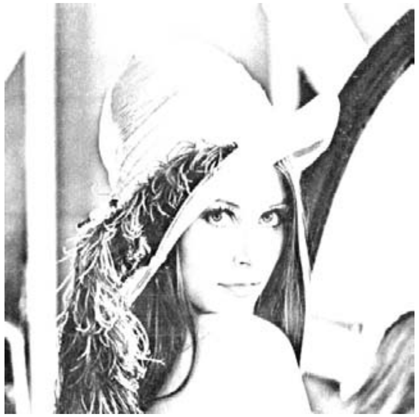
Pixel arithmetic

- Seamless integration with R's native arrays
- Adjust brightness, contrast and gamma-factor

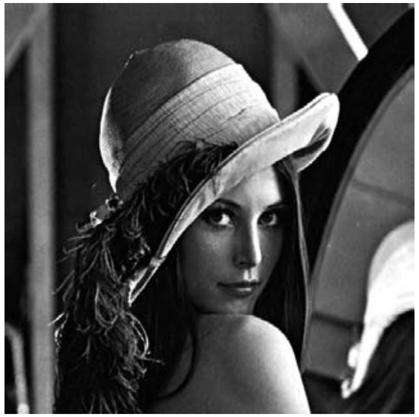
`x`



`x+0.5`



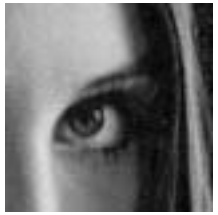
`3*x`



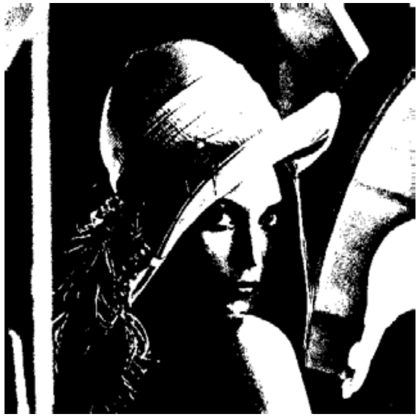
`(x+0.2)^3`

Spatial transformations

- Cropping, thresholding, resizing, rotation

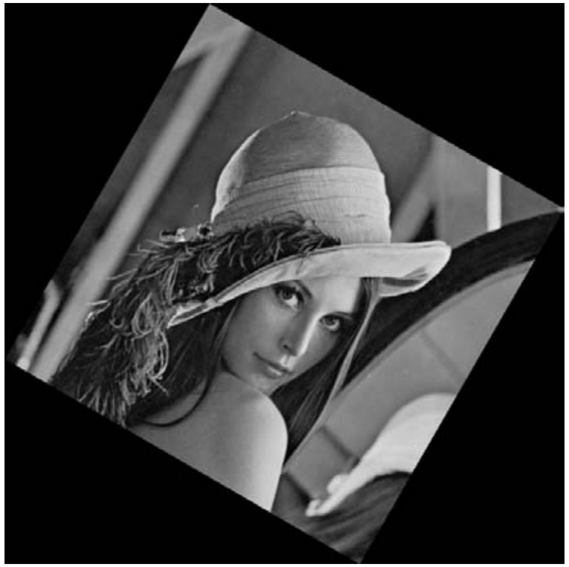


x[45:90, 120:165]



x>0.5

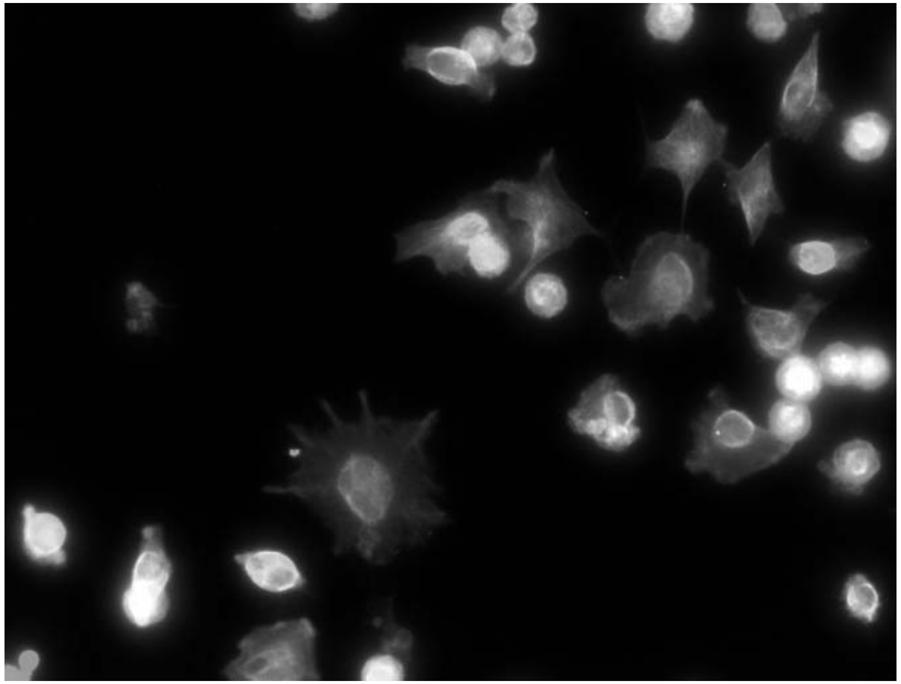
resize(x, w=128)



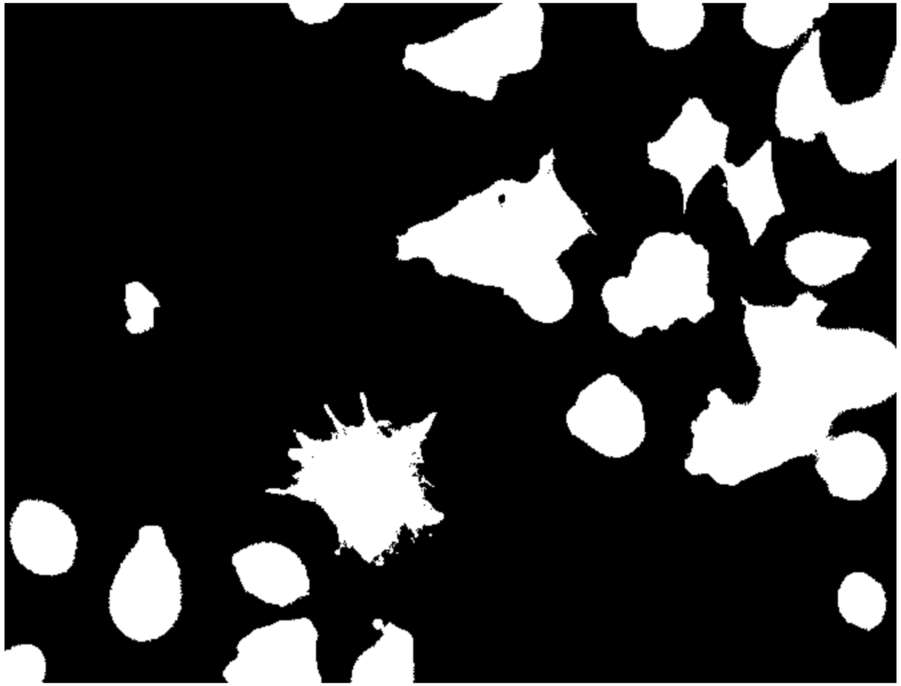
rotate(x, 30)

Thresholding

- Global thresholding
- Building block tool to segment cells



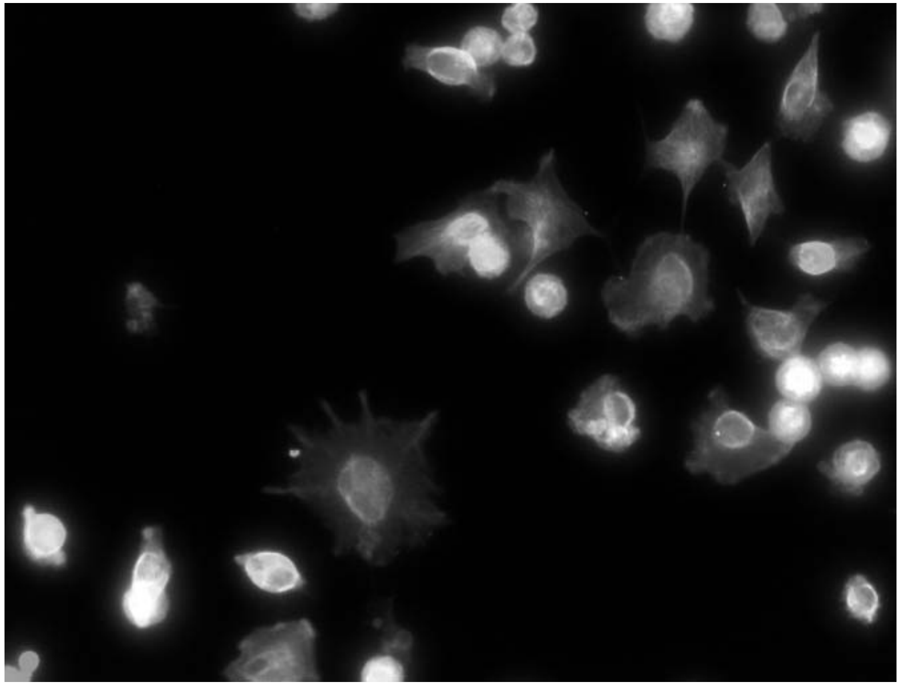
`x`



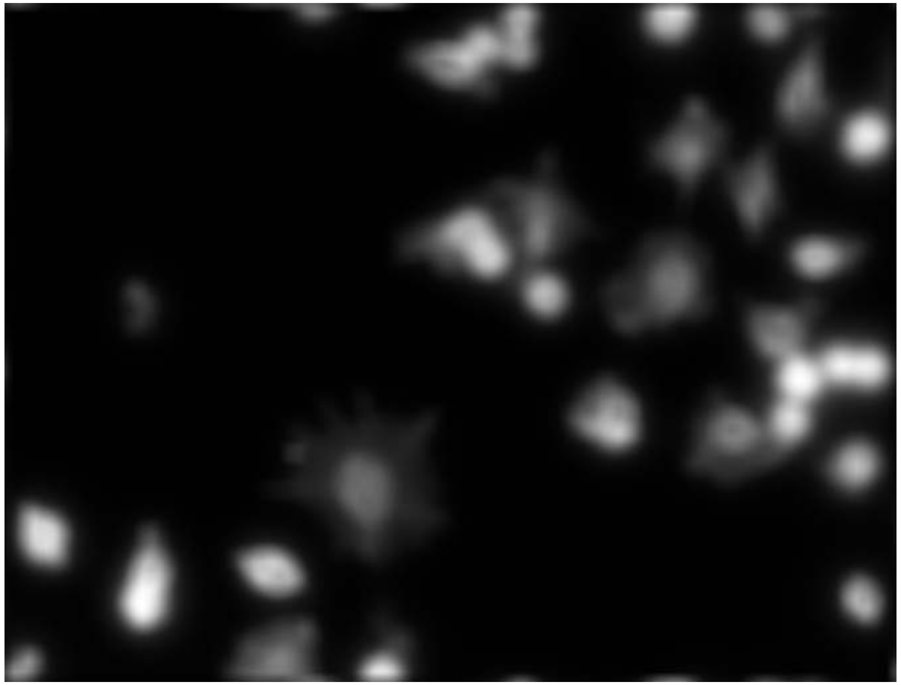
`x>0.3`

Linear filter

- Fast 2D convolution with filter2()
- Low-pass filter: smooth images, remove artefacts



```
x
```

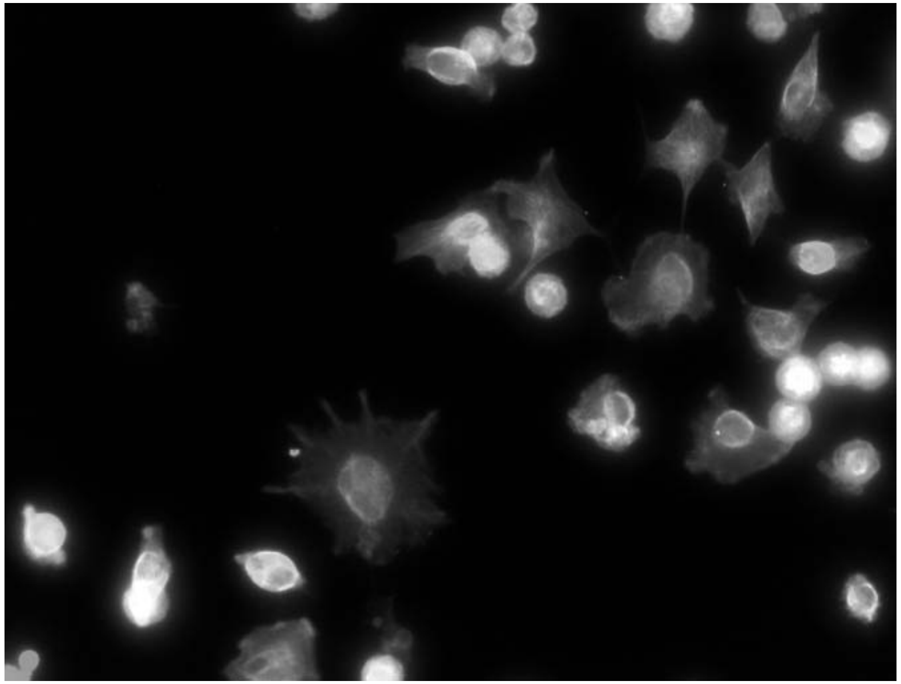


```
f = array(1, dim=c(9, 9))
f = f/sum(f)
y = filter2(x, f)
```

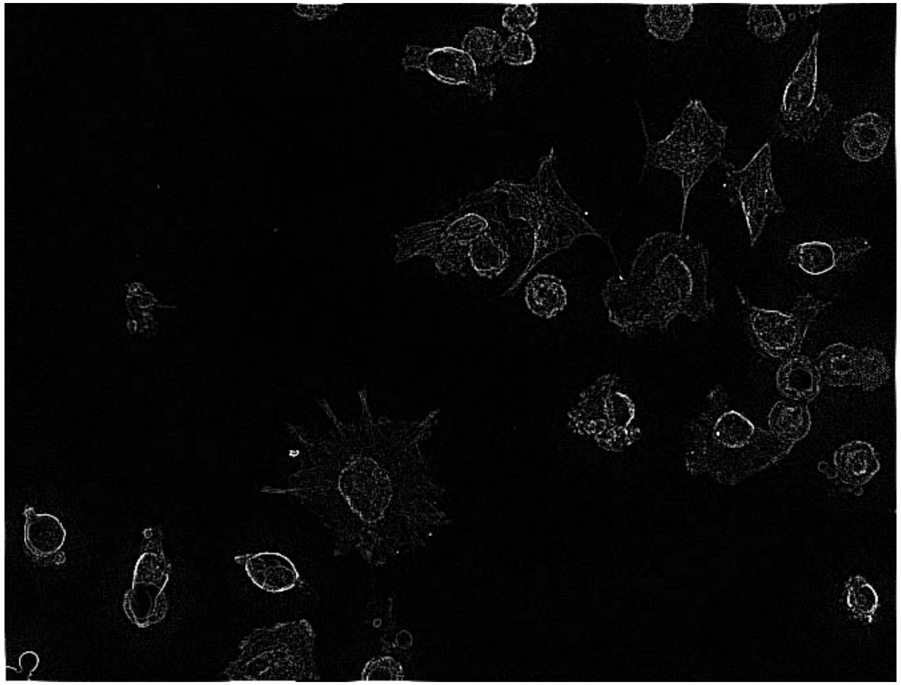
$$x \star \begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$$

Linear filter

- Fast 2D convolution with filter2()
- High-pass filter: detect cell edges



```
x
```

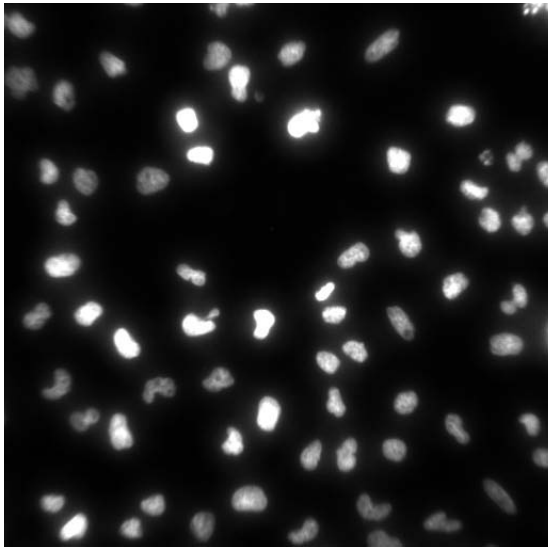


```
f = array(1, dim=c(9, 9))
f[3, 3] = -8
y = filter2(x, f)
```

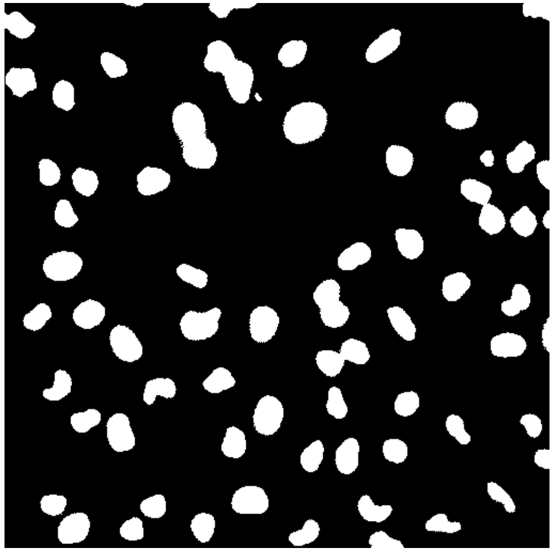
$$x \star \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Nucleus segmentation

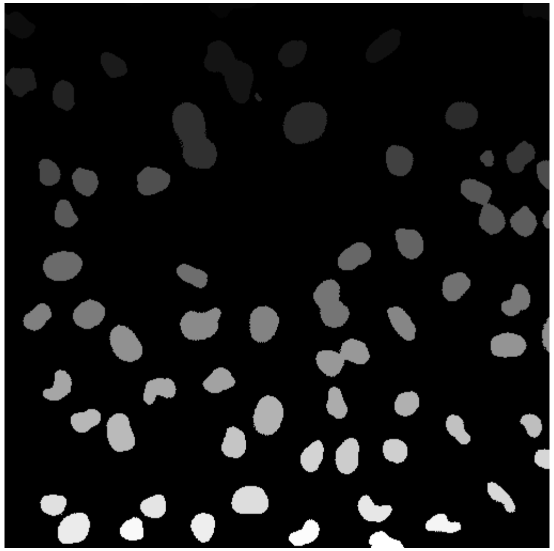
- Global thesholding + labelling
- Function bwlabel()
  - Labels connected sets (objects) from a binary image
  - Every pixel of each connected object is set to an unique integer value
  - max(bwlabel(x)) gives the number of detected objects



`x`



`x>0.2`



`bwlabel(x>0.2)`

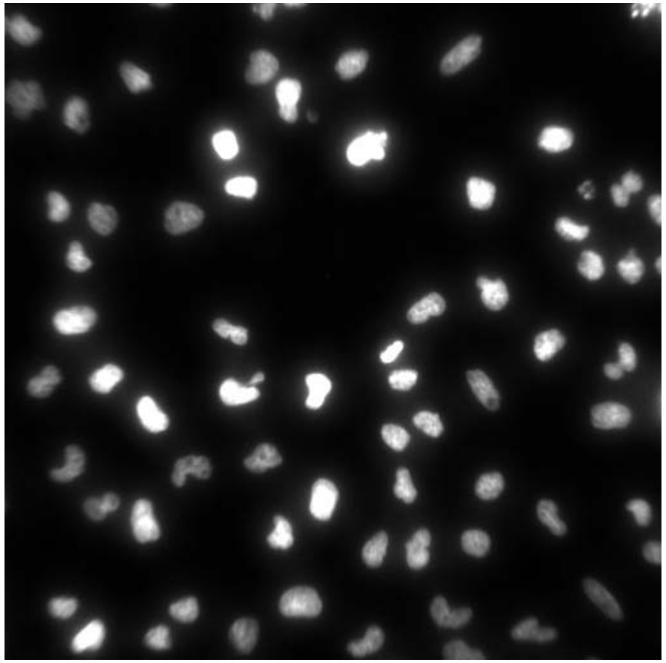
Nucleus morphology quantification

- Function getFeatures()
  - Extracts object features
  - Geometric, image moment based features
  - Texture based features (Zernike moments, Haralick features)
  - Direct interpretation (ex: DNA content) or for classification/clustering

41 features

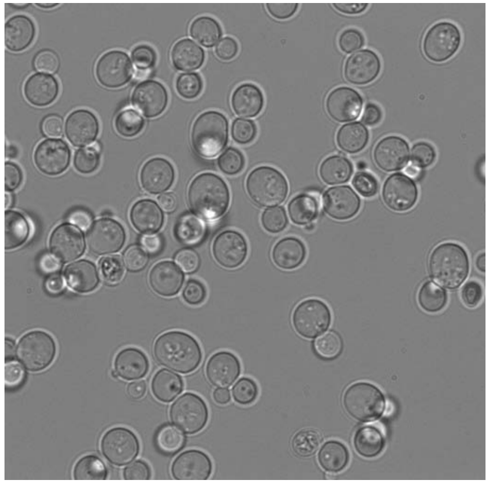
76 nuclei

```
           g.x        g.y    g.s g.p     g.pdm    g.pdsd    g.effr   g.acirc
 [1,] 123.1391   3.288660  194  67  9.241719  4.165079  7.858252 0.417525
 [2,] 206.7460   9.442248  961 153 20.513190  7.755419 17.489877 0.291363
 [3,] 502.9589   7.616438  219  60  8.286918  1.954156  8.349243 0.155251
 [4,]  20.1919  22.358418 1568 157 22.219461  3.139197 22.340768 0.116709
 [5,] 344.7959  45.501992 2259 233 35.158966 15.285795 26.815332 0.501106
 [6,] 188.2611  50.451863 2711 249 28.732680  6.560911 29.375808 0.168941
 [7,] 269.7996  46.404036 2131 180 26.419631  5.529232 26.044546 0.193805
 [8,] 106.6127  58.364243 1348 143 21.662879  6.555683 20.714288 0.264836
 [9,] 218.5582  77.299007 1913 215 25.724580  6.706719 24.676442 0.243073
[10,]  19.1766  81.840147 1908 209 26.303760  7.864686 24.644173 0.304507
[11,]   6.3558  62.017647  340  68 10.314127  2.397136 10.403142 0.188235
[12,]  58.9873  86.034128 2139 214 27.463158  6.525559 26.093387 0.207106
[13,] 245.1087  94.387405 1048 123 18.280901  2.894758 18.264412 0.112595
[14,] 411.2741 109.198678 2572 225 28.660816  7.914664 28.612812 0.224727
[15,] 167.8151 107.966014 1942 160 24.671533  2.534342 24.862779 0.084963
[16,] 281.7084 121.609892 2871 209 31.577270  6.470767 30.230245 0.128874
[17,] 479.2334 143.098241 1649 183 23.913630  6.116630 22.910543 0.248635
[18,] 186.5930 146.693122 2079 199 27.280908  6.757808 25.724818 0.195286
[19,] 356.7303 148.253418 3145 285 34.746206 11.297632 31.639921 0.313513
[20,] 449.2436 147.798319  119  37  5.873578  1.563250  6.154582 0.243697
...
```

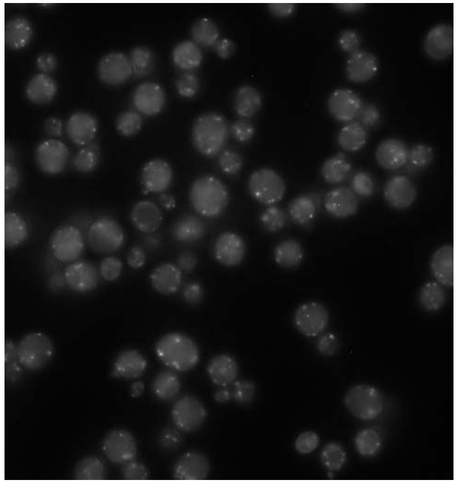


EBImage

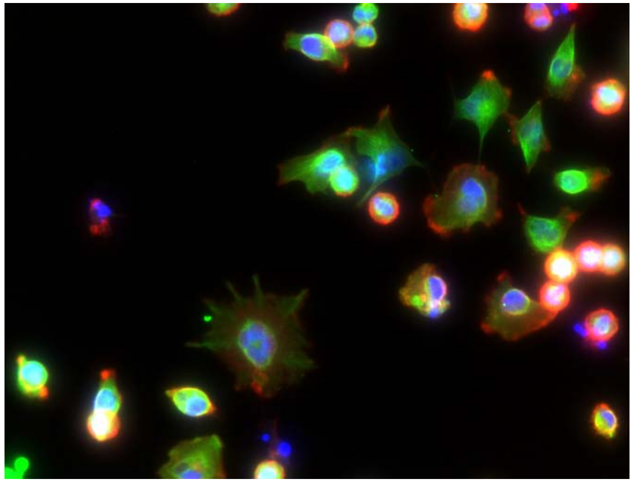
- Powerful and fast package to process images in R
- Diverse use cases
  - Counting objects
  - Detection/quantification of structures of interest
  - High-throughput/high-content batch phenotyping



Yeast, BF



Yeast, GFP



HeLa, Hoetsch+Actin+Tubulin

Clustering phenotype populations by genome-wide RNAi and multiparametric imaging

Gregoire Pau, Oleg Sklyar, Wolfgang Huber
EMBL, Heidelberg

Florian Fuchs, Dominique Kranz, Christoph Budjan,
Thomas Horn, Sandra Steinbrink, Angelika Pedal, Michael Boutros
DKFZ, Heidelberg

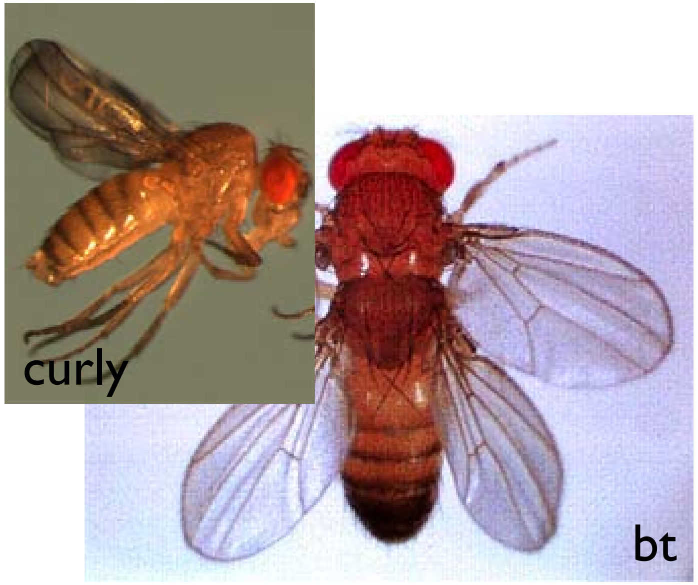
Molecular Systems Biology, 2010

EMBL-EBI

dkfz. GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

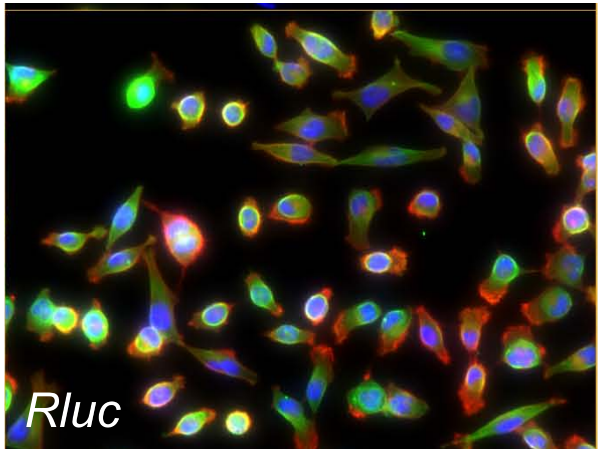
Genome annotation

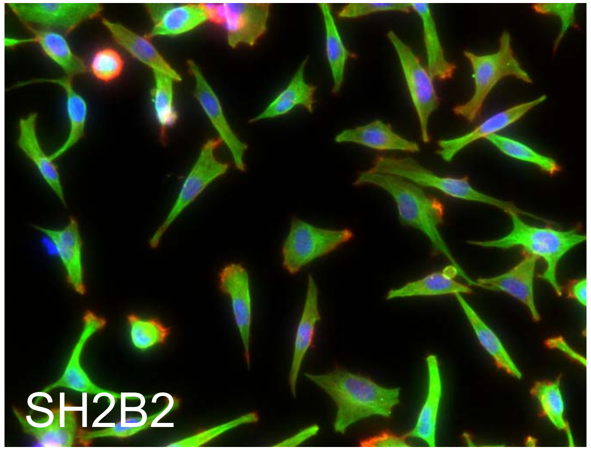
- 60 % of the ~22839 human genes have no known function
- Key techniques to annotate genes are based on similarity
  - Ex: screening random *Drosophila* mutants [Nüsslein-Volhard, 1980]
  - Genes aggregation by loss-of-function phenotype similarity
  - Reverse genetics: from phenotypes to genes

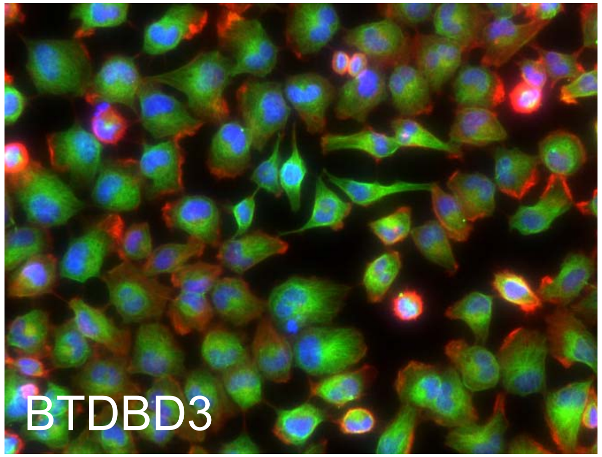


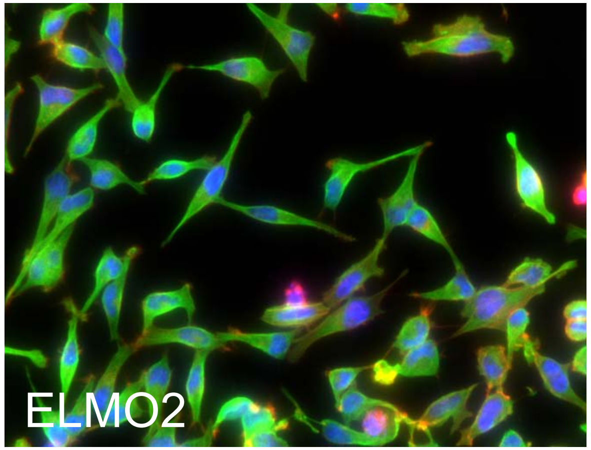
RNA interference & cell morphology

- Selective transcript depletion with siRNA
- Cell morphology is a broad reflector of biological processes
- Gene annotation by loss-of-function phenotypic similarity



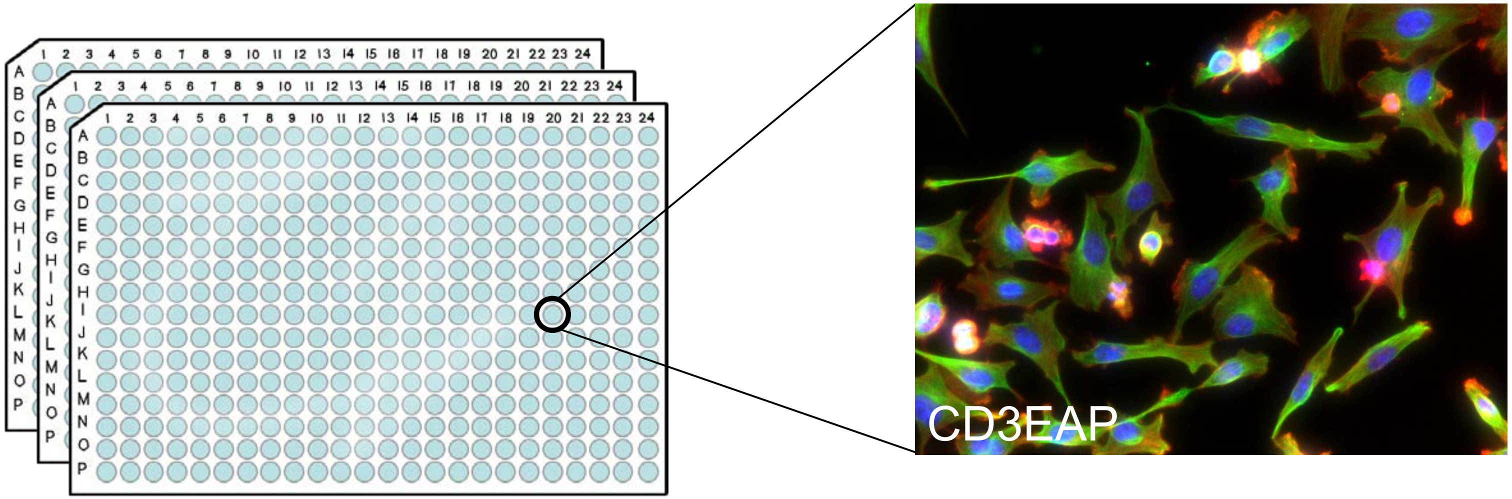




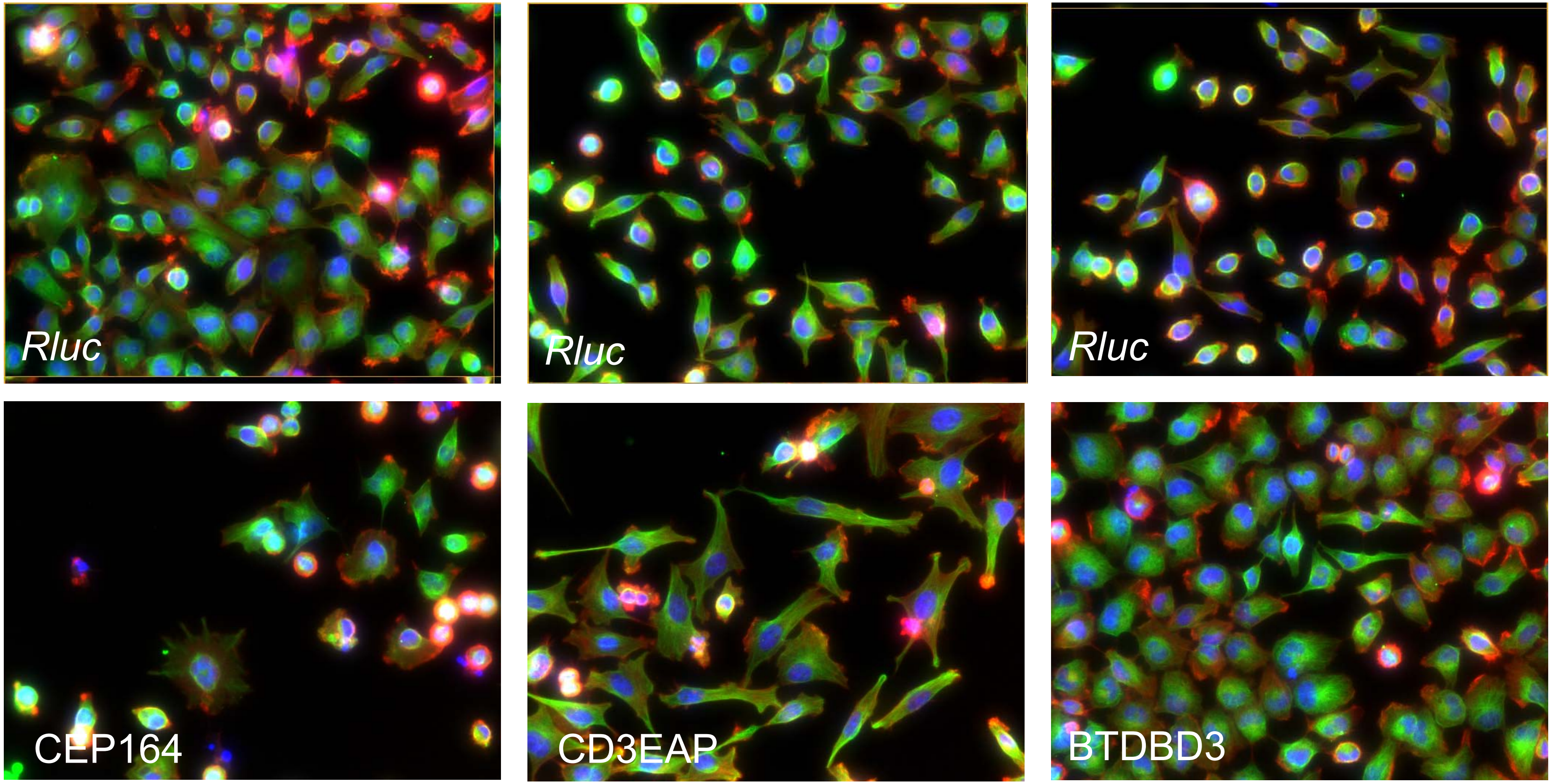


Experimental setup

- Human cervix carcinoma HeLa cells
- Genome-wide RNAi screen, testing 22839 genes
- Cells are incubated for 48 h and fixed
- Staining using DNA (DAPI), Tubulin (Alexa), Actin (TRITC)
- Readout: microscopy images



Examples of cellular phenotypes



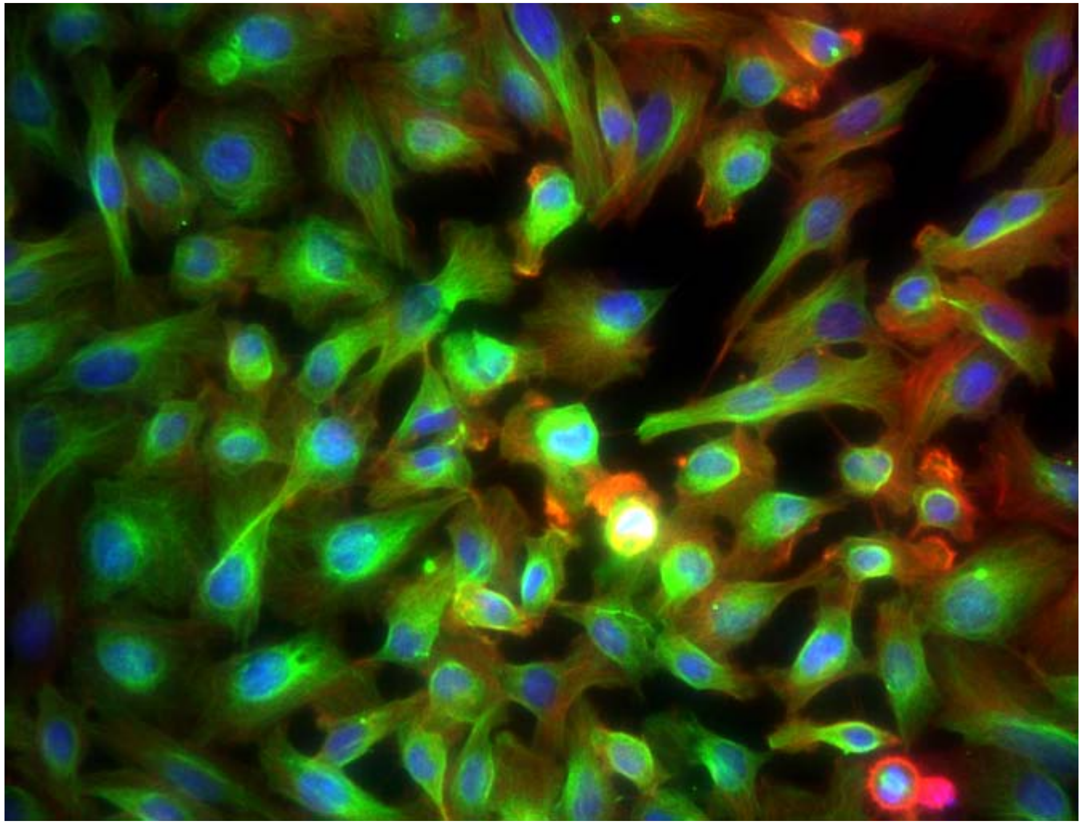
Motivation

- Biological questions
  - Gene perturbations leading to similar phenotypes
  - Gene association by loss-of-function phenotypic similarity
  - Extreme phenotypes

- Approach
  - Phenotype quantification: image → R^p
  - Definition of a similarity measure in the phenotypic space
  - Generation of hypotheses about gene functions
  - Testing hypotheses using secondary assays

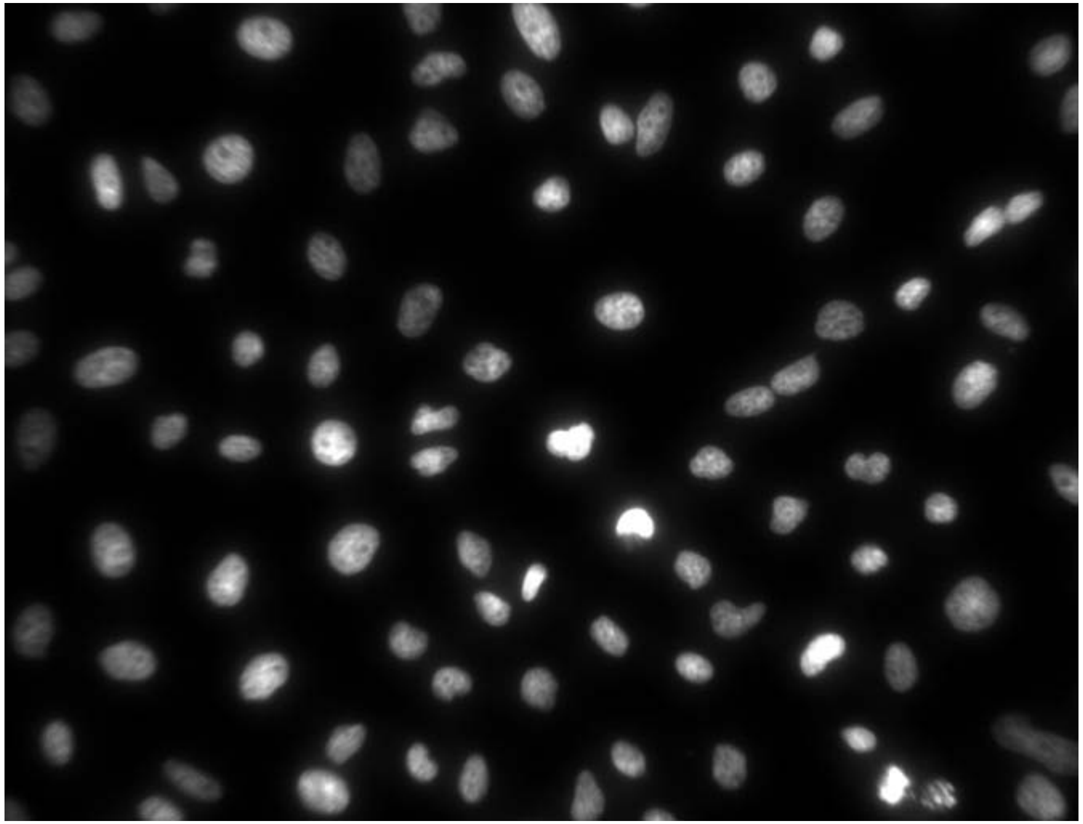
Nucleus segmentation

- Nucleus are extracted from the DNA channel H
- Adaptative thesholding: Nmask = $H \star w > \sigma_H$
- Connected set labelling + morphological opening



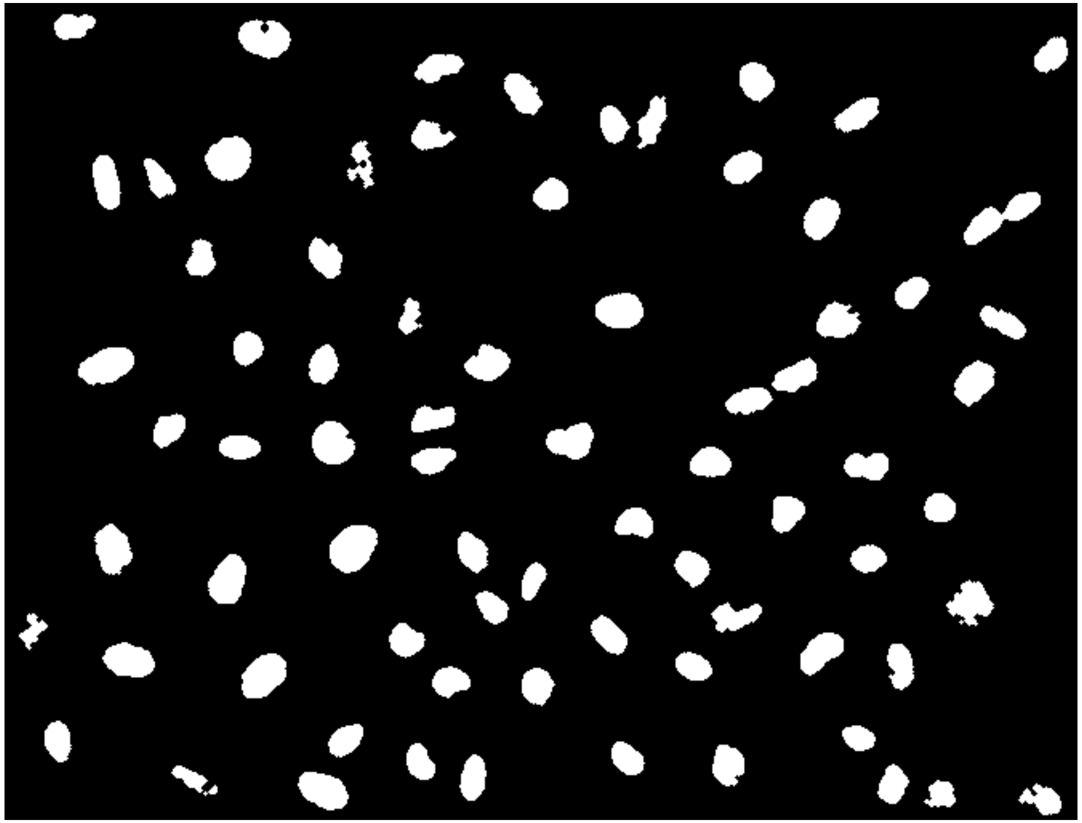
Nucleus segmentation

- Nucleus are extracted from the DNA channel H
- Adaptative thesholding: Nmask = $H \star w > \sigma_H$
- Connected set labelling + morphological opening



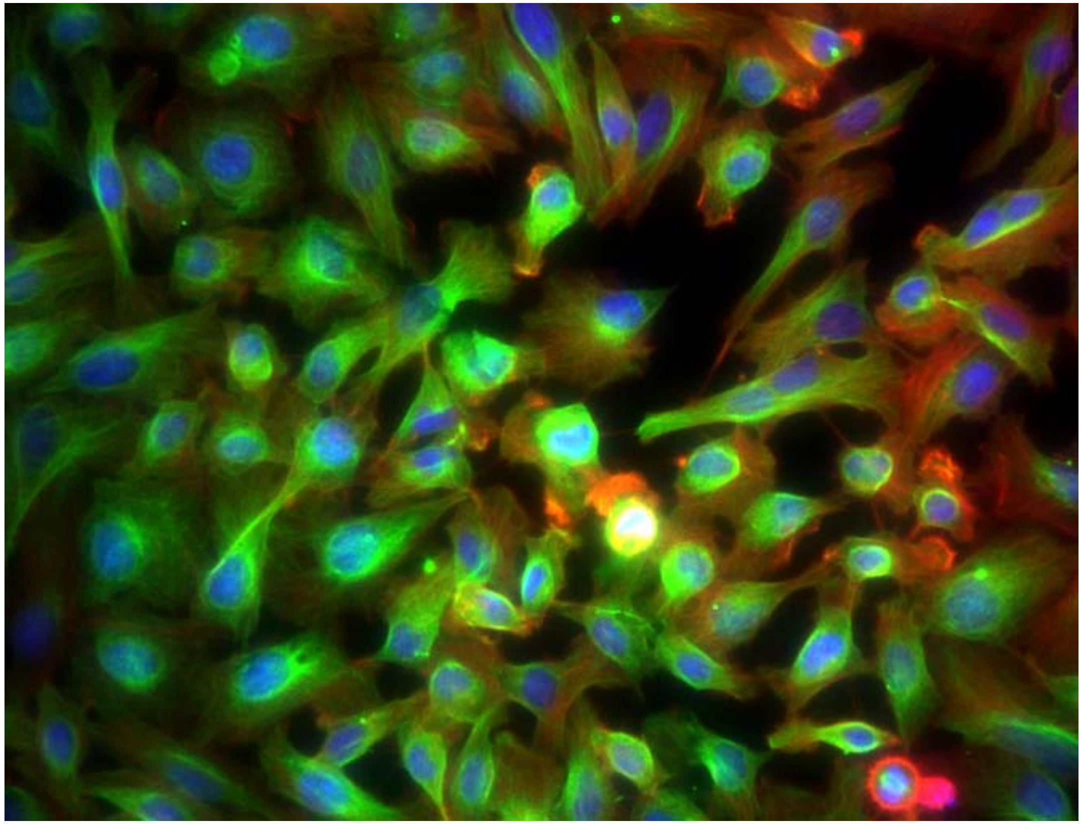
Nucleus segmentation

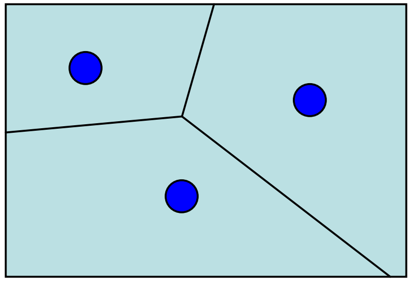
- Nucleus are extracted from the DNA channel H
- Adaptative thesholding: Nmask = $H \star w > \sigma_H$
- Connected set labelling + morphological opening

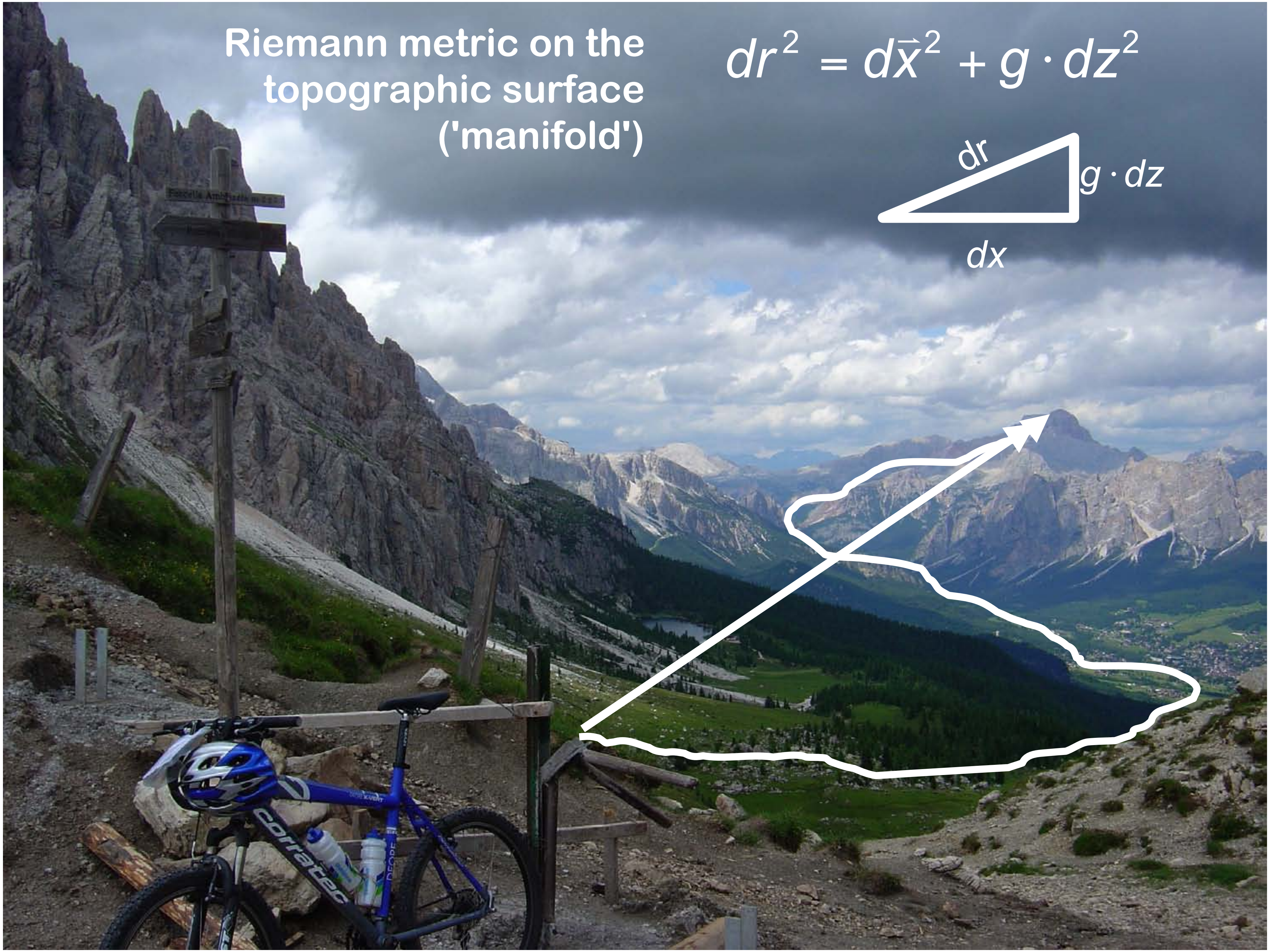


Cell membrane determination

- Using nuclei as seeds
- Voronoi segmentation using an image gradient based metric





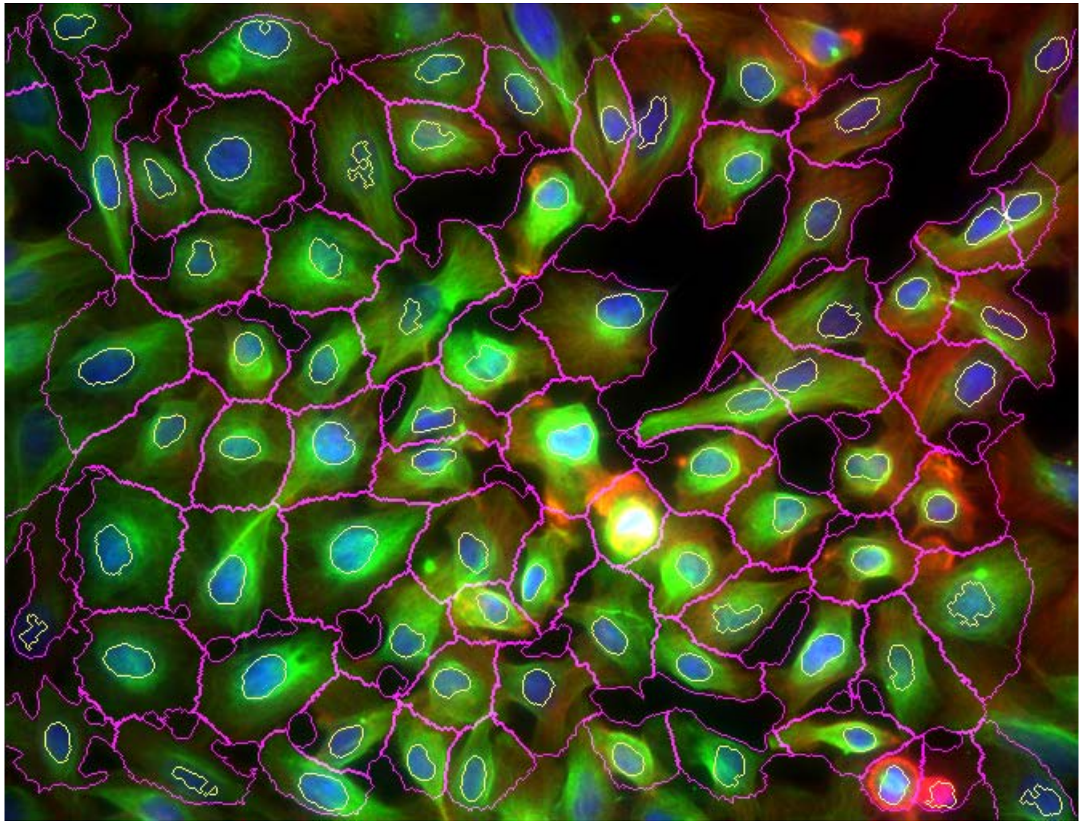


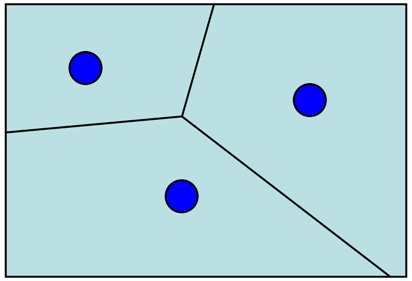
Riemann metric on the topographic surface ('manifold')

$$dr^2 = d\vec{x}^2 + g \cdot dz^2$$

Cell membrane determination

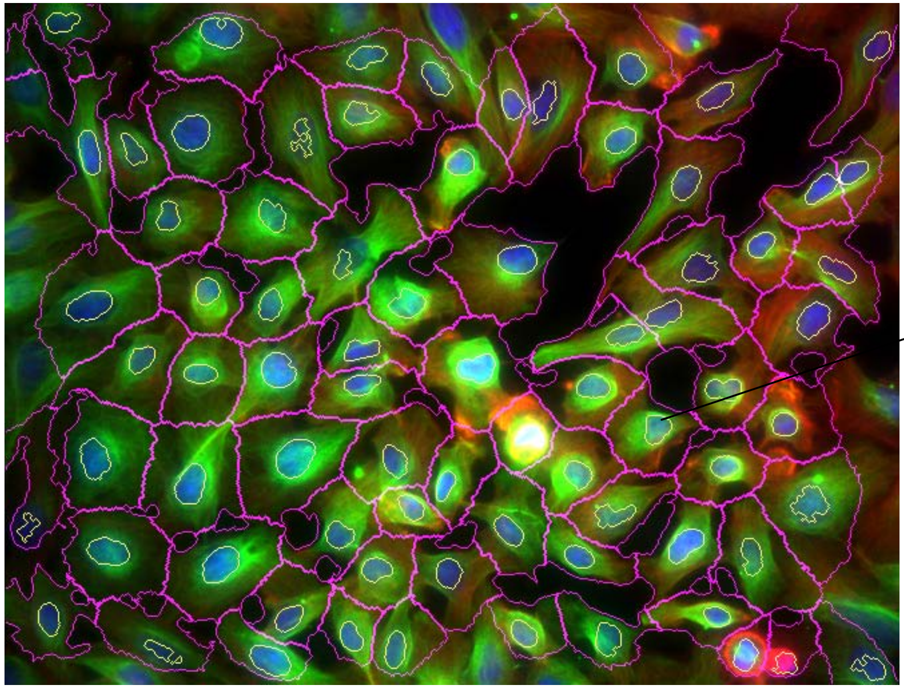
- Using nuclei as seeds
- Voronoi segmentation using an image gradient based metric





Cell descriptors

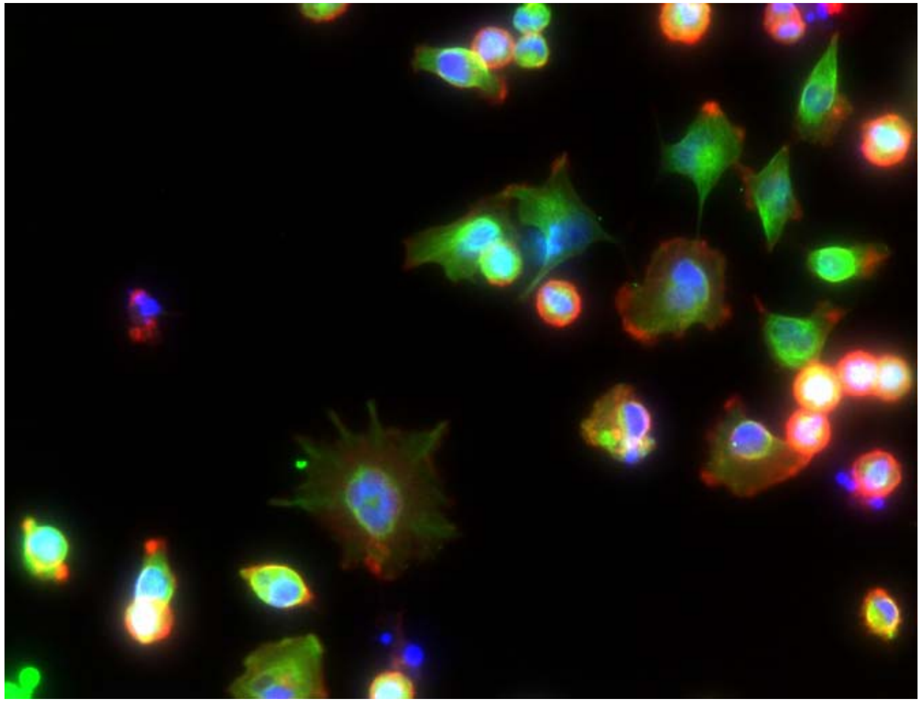
- Quantitative characterization of cells
- q = 181 rotation and translation invariant descriptors
  - Geometric (intensity, size, perimeter, eccentricity…)
  - Texture (Haralick, Zernike moments…)
  - $y_k = \text{sum}_{xy}\, w^k_{xy} p_{xy}$ where w^k_{xy} is rotation invariant

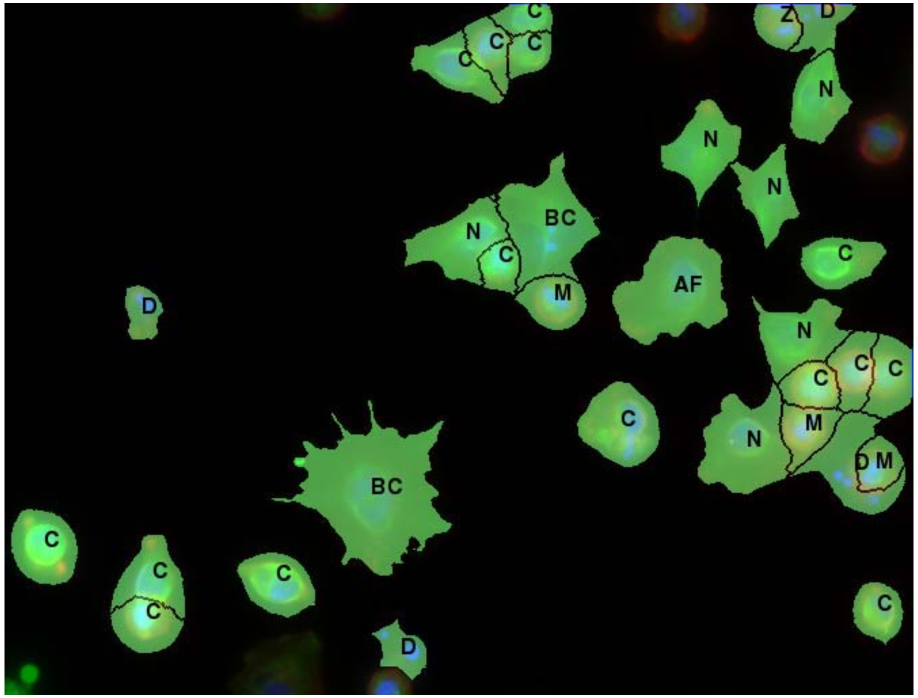


c.a.m.l1	0.587605
c.a.m.l2	0.033118
c.a.m.ec	0.472934
c.a.m.ss	2857.35619
c.t.m.int	485.271057
c.t.m.l1	0.828876
c.t.m.l2	0.098647
c.t.m.ec	0.549594
c.t.m.ss	2338.817467
c.h.m.int	219.588177
c.h.m.l1	0.779339
c.h.m.l2	0.009249
c.h.m.ec	0.219697
c.h.m.ss	1067.046085
c.m.m.int	966.307719
c.m.m.l1	0.475141
c.m.m.l2	0.02463
c.m.m.ec	0.496583
c.m.m.ss	2722.903987
n.a.m.int	202.842021
…	

Cell classification

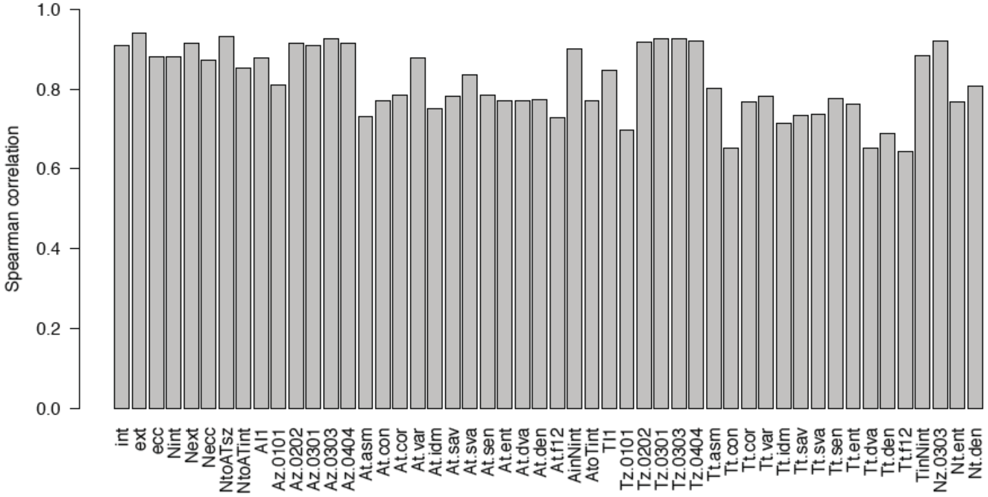
- Using cell descriptors as input
- SVM with radial kernel + 8 classes + training set of ~3000 cells
- Classification performance (5-fold CV) on TS: ~85 %

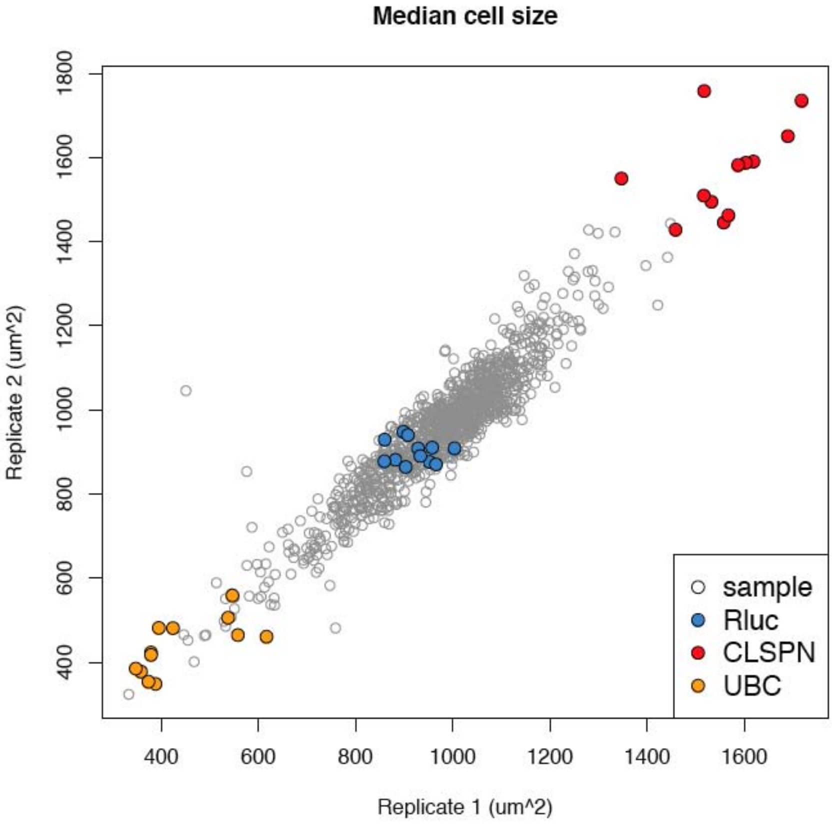


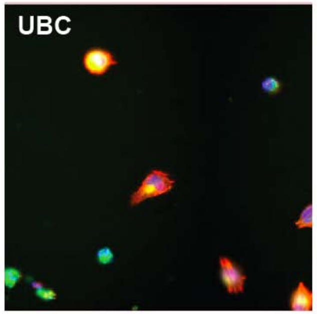


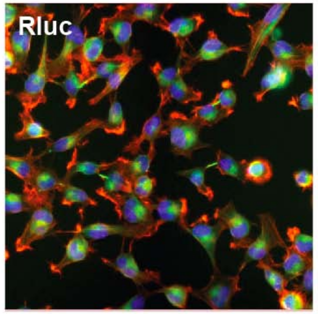
Cellular descriptor statistics

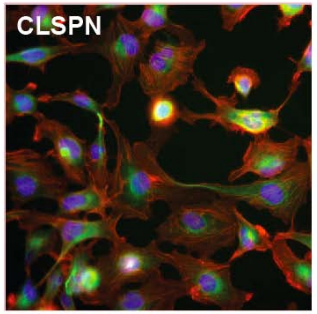
- Out of the 181 descriptors, 50 are highly reproducible
- Single descriptors can discriminate controls





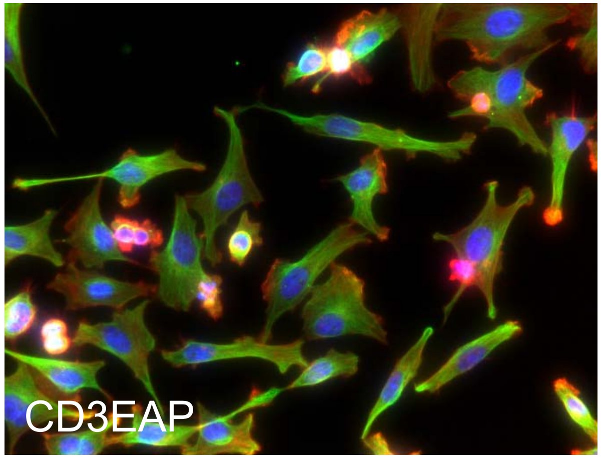






Phenotypic profile

- Phenotype expressed by a population of cells
- Phenotypic profile, vector of p = 13 parameters
  - Number of cells
  - Statistics on cell features (size, eccentricity, …)
  - Cell types distribution (normal, metaphase, condensed, protruded…)



$$x = \left[\begin{array}{ll}
\text{n} & 289 \\
\text{ext} & 34.33118 \\
\text{ecc} & 0.472934 \\
\text{Next} & 2857.356 \\
\text{Nint} & 485.2710 \\
\text{a2i} & 0.828876 \\
\text{Next2} & 0.098647 \\
\text{AF \%} & 0.049594 \\
\text{BC \%} & 0.081746 \\
\text{C \%} & 0.158817 \\
\text{M \%} & 0.179339 \\
\text{LA \%} & 0.009249 \\
\text{P \%} & 0.219697
\end{array}\right.$$

Preliminary conclusion

- Automated phenotype quantification of cellular populations
  - Multiparametric imaging
- High-throughput batch processing by EBImage
  - ~92000 images: 22 h of processing time with 30 CPUs
- Phenotypic screens
  - RNAi + HeLa +morphology
  - RNAi + U2OS + morphology
  - Drugs + yeast + tagged GFP proteins

- ImageHTS
  - Automated analysis of cell-based imaging screens
  - Distributed and hierarchical (well, cell, features) web data access
  - Upcoming !