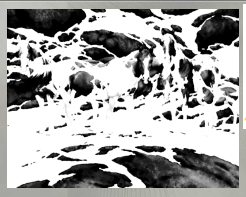
Life beyond the pixels: Image analysis and machine learning methods for high-content cancer drug discovery

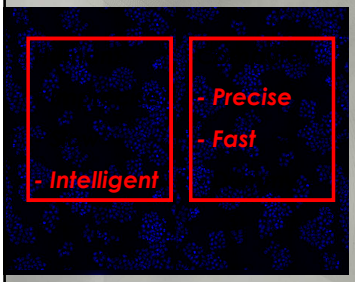
Peter Horvath
FIMM, Institute for Molecular Medicine Finland, Helsinki
BRC, Hungarian Academy of Sciences

Szeged 2015

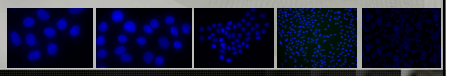
Where…? What…? How many…?



How fast can you count?



#	Human	Computer
9	4 sec	0.1 sec
18	7 sec	0.15 sec
51	17 sec	0.25 sec
210	60 sec	1 sec
3610	☺	13 sec

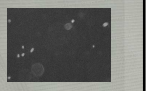


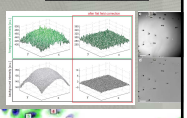
Throughput vs. human knowledge

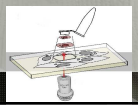
- **Video**

Research activities

- Image segmentation and tracking
  - Cell segmentation
  - Label free reconstruction
  - Tracking (2-3D and phase contrast)
- Image quality improvement
  - Flat field, photo bleaching
- Machine learning
  - Phenotypic profiling
  - Regression models
- CL²M (correlative light-light microscopy)
  - High-throughput single cell isolation



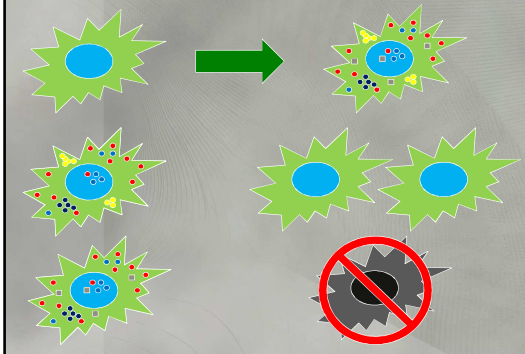




Needle meet haystack

SCREENING

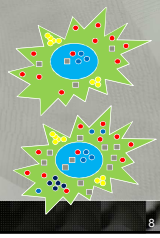
The question



7

What is screening?

- Systematic test of the effect of one or more parameters in a complex system
- Aim:
  - local or global optimum
  - Network/correlation/connection between different parameter values
- Further challenges:
  - Multiple knock-out
  - Knock-in



8

Human Genome Project (HGP)

- Analysis of the whole human genome up to nucleotides (approx 3 billion base pairs)
- 1990-2006
- Identification of all human genes (~22.000 different)

Challenge in cell biology

- Approx. 22.000 human genes
- These ~22.000 genes encodes ~100.000 proteins
- Deeper knowledge ~1.500 proteins
- How to analyze unknown proteins?
  - Animal models are expensive
  - Possible solution: RNAi technology

10

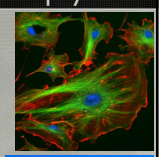
RNAi and the siRNA technology

- RNAi = RNA interference
- siRNA = small interfering RNA
  - RNA molecule consists of 20-25 nucleotides
  - Interferes with the expression of a specific gene
- siRNA against the whole human genome is available

11

Fluorescence microscopy

- Optical microscopy technique to study properties of substances using the phenomena of fluorescence
- Specifically labeled specimen with a fluorophore molecule
  - GFP (green fluorescent protein, Jellyfish, 1996; 2008 Nobel Prize)
  - DAPI



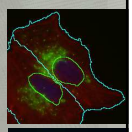
1. Specimen illuminated with a specific wavelength→ absorbed by the fluorophore
2. Emit light of longer wavelength

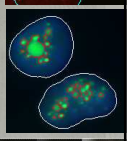
Source: Wikipedia

12

Single Cell-Based fluorescent microscopic screening

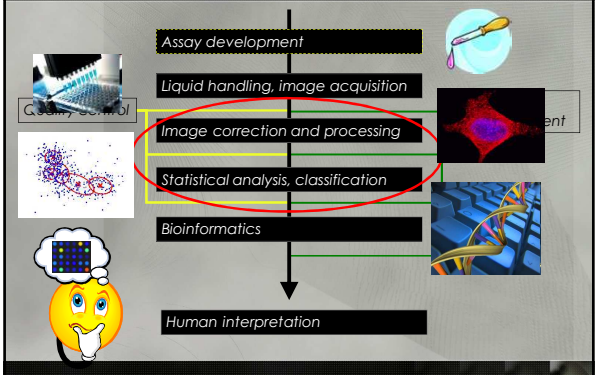
- The newest acquisition technologies:
  - fast imaging
  - high-resolution
- Possible to analyze single cells, or intracellular structures
- Computer aided data analysis:
  - accurate
  - fast
  - intelligent and reliable





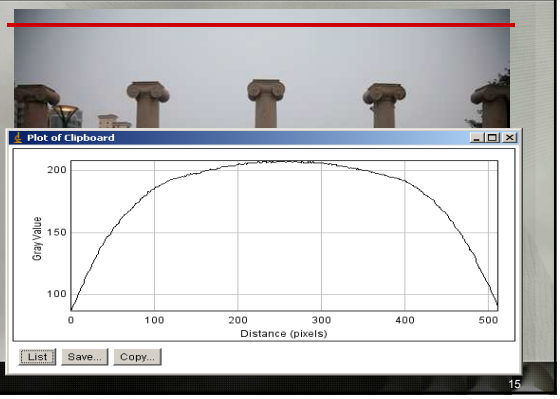
13

The data processing pipeline



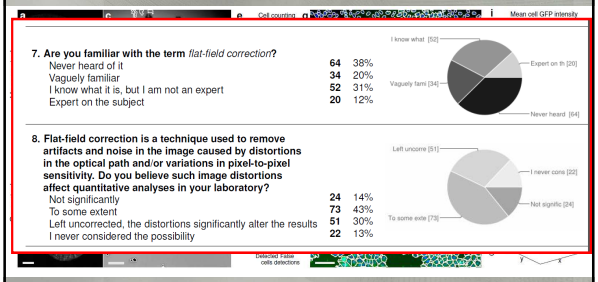
14

Vignetting



15

Illumination correction



7. **Are you familiar with the term *flat-field correction*?**

Never heard of it	64	38%
Vaguely familiar	34	20%
I know what it is, but I am not an expert	52	31%
Expert on the subject	20	12%

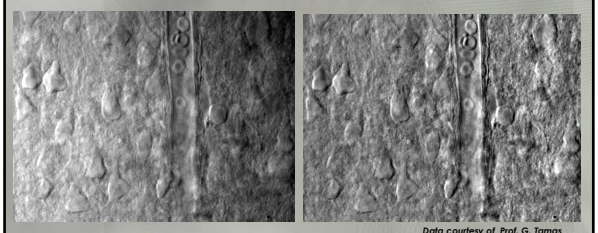
8. **Flat-field correction is a technique used to remove artifacts and noise in the image caused by distortions in the optical path and/or variations in pixel-to-pixel sensitivity. Do you believe such image distortions affect quantitative analyses in your laboratory?**

Not significantly	24	14%
To some extent	73	43%
Left uncorrected, the distortions significantly alter the results	51	30%
I never considered the possibility	22	13%

K. Smith, F. Piccinini, C. Balazs, G. Csucs, A. Bevilacqua, P. Horvath; *CIDRE: A general illumination correction method for optical microscopy*, ***Nature Methods 2015***

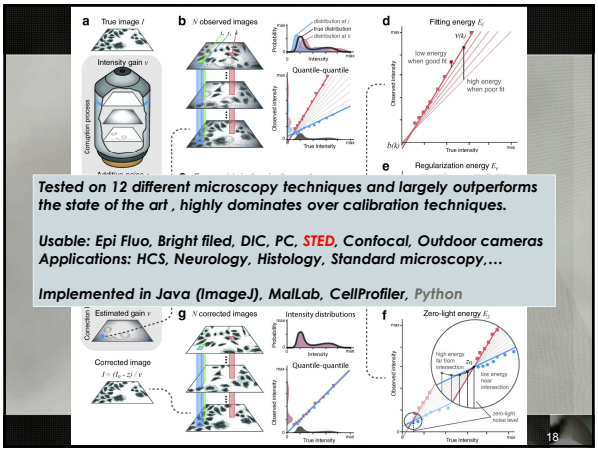
16

Flat field correction



Data courtesy of Prof. G. Tamas

K. Smith, F. Piccinini, C. Balazs, G. Csucs, A. Bevilacqua, P. Horvath; *CIDRE: A general illumination correction method for optical microscopy*, ***Nature Methods 2015***



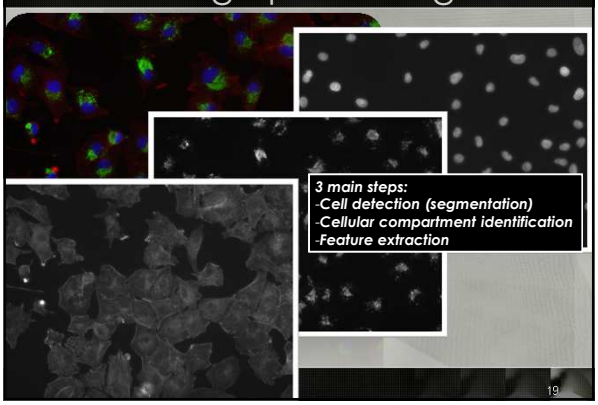
Tested on 12 different microscopy techniques and largely outperforms the state of the art , highly dominates over calibration techniques.

Usable: Epi Fluo, Bright filed, DIC, PC, STED, Confocal, Outdoor cameras
Applications: HCS, Neurology, Histology, Standard microscopy,…

Implemented in Java (ImageJ), MalLab, CellProfiler, Python

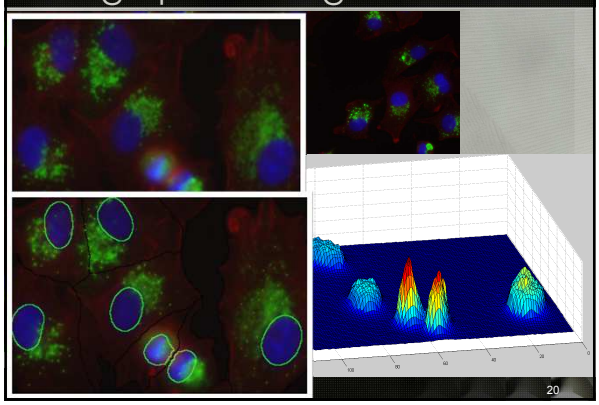
18

Image processing



19

Image processing I. – cell detection



20

Image processing II. – compartments

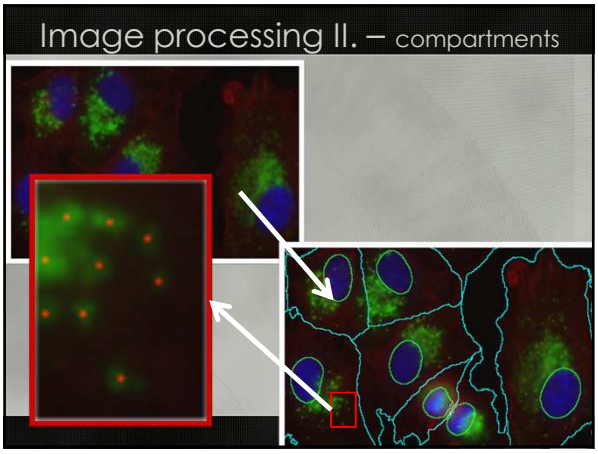


Image processing II. – compartments

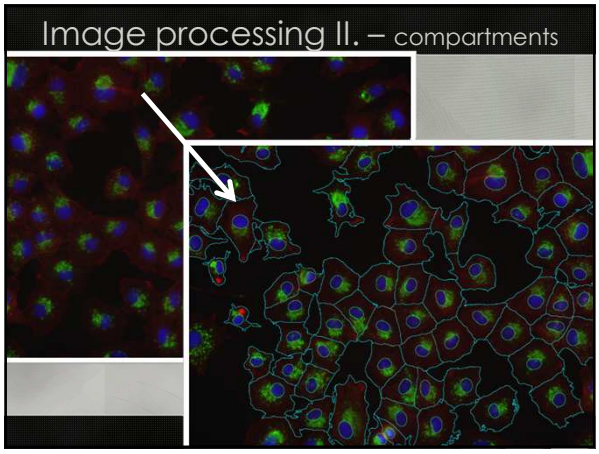
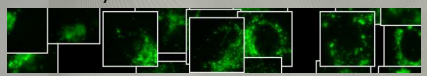


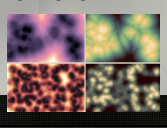
Image processing III. – feature extraction

- Morphology

- Intensity-based



- Texture



23

Image processing - Summary

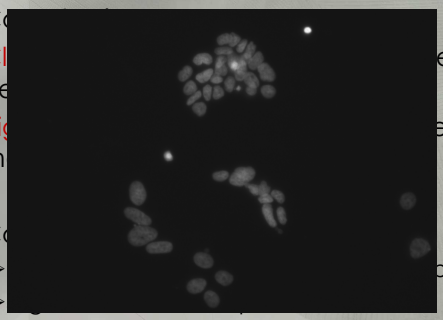
The challenge

IMAGE SEGMENTATION

25

Image segmentation with shape priors

- Co
- Cl … e
  ce
- Rig … e
  m
- Co
  - …
  - …
- Variational methods

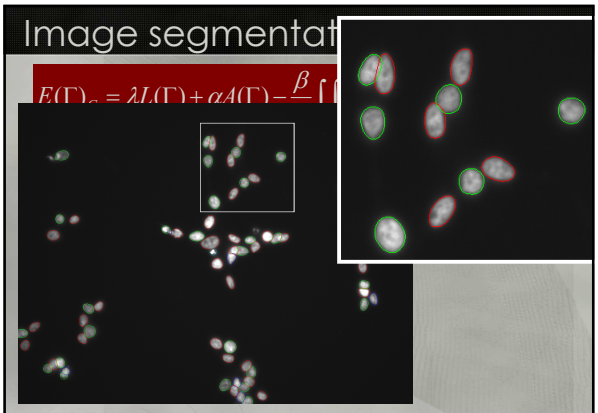


Data courtesy: Precision medicine; prostate cancer
Päivi Östling, Taija af Hällström, Vilja Pietiäinen, Piia Mikkonen FIMM Helsinki

26

Image segmentat…

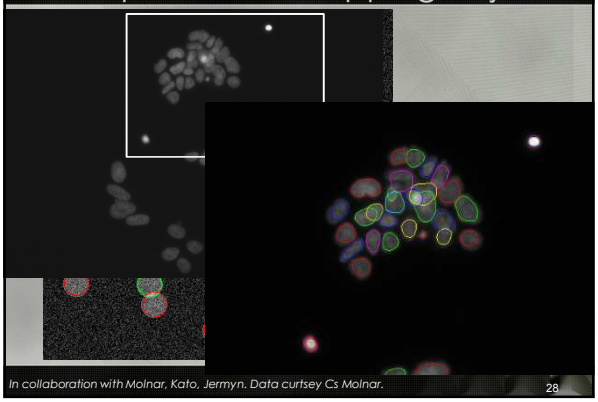
$$E(\Gamma)_c = \lambda L(\Gamma) + \alpha A(\Gamma) - \frac{\beta}{\dots}\int \dots$$



Horvath, Jermyn, Kato, Zerubia; Gas of Circles model - *Pattern Recognition*
Molnar, Kato, Jermyn. A Multi-Layer Phase Field Model for Extracting Multiple Near-Circular Objects (ICPR)

27

Examples for overlapping objects



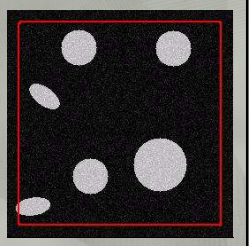
In collaboration with Molnar, Kato, Jermyn. Data curtsey Cs Molnar.

28

Describing more general shapes

- **Simplest size/shape priors**
  - **Area fixed, or**
  - **Perimeter fixed, or**
  - **Area/perimeter fix**
  - **Curvature**

Large circular
Small circular
Small elliptical



Jozsef Molnar, Adam Szucs, Peter Horvath : Active contours for selective object segmentation *(work in progress)*

29

Energy minimizing functional

- The usual structure of energy minimizing functional for object segmentation:

$$E = E_{data} + E_{smoothness} + E_{prior}$$

- We propose composite prior term with
  - Size and
  - Shape priors

$$E_{prior} = E_{size} + E_{shape}$$

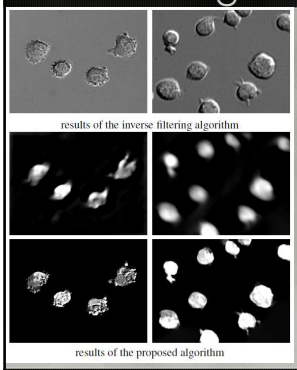
Size priors

- Two basic types considered
  - Preferred perimeter-length: $F(\dot{\mathbf{r}}) \propto f\left(L_0 - \oint ds\right)$
  - Preferred internal area: $F(\mathbf{r},\dot{\mathbf{r}}) \propto f\left(A_0 - \iint dA\right)$
- The function *f* should have inflection at zero
  - eg. Cubic function is a simple choice

Shape priors

- One of the simplest prior is the perimeter-area ratio: $q = \frac{\left(\oint ds\right)^2}{\iint dA}$
- Describes "Amoeba"-invariant shapes
- Note that the same terms were used to define size and shape priors
- Fast and capable to describe a family of objects

DIC image reconstruction



DIC microscopy:
DIC is label free and low phototoxic +

DIC is not quantitative but qualitative
Difficult:
-Measurements
-Large imaging scenarios

Goal:
Automated image reconstruction
(+GPU and 3D):

$$E \doteq \frac{1}{2}\iint_{\Omega}\left[\underbrace{\mathbf{u}\cdot\nabla\hat{I}-(G-c)}_{S}\right]^2 d\Omega + \lambda\iint_{\Omega}|\nabla I|\,d\Omega$$

Data fit — Smoothness

Krisztian Koos, Jozsef Molnar, Peter Horvath : *DIC microscopy image reconstruction using a novel variational framework*

CellTracker

- **Video**

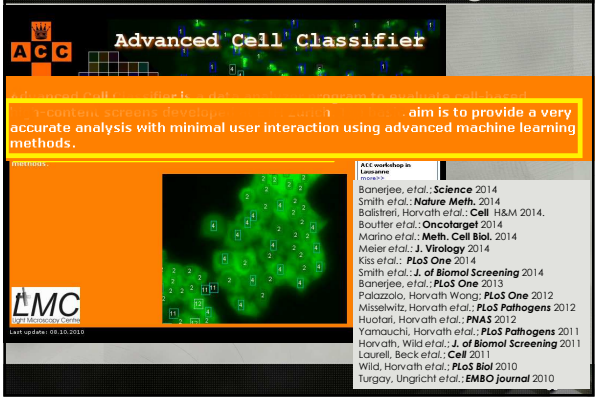
34

Multi-parametric non-linear – Follow the fashion!

SINGLE CELL-BASED CLASSIFICATION FOR HCA

35

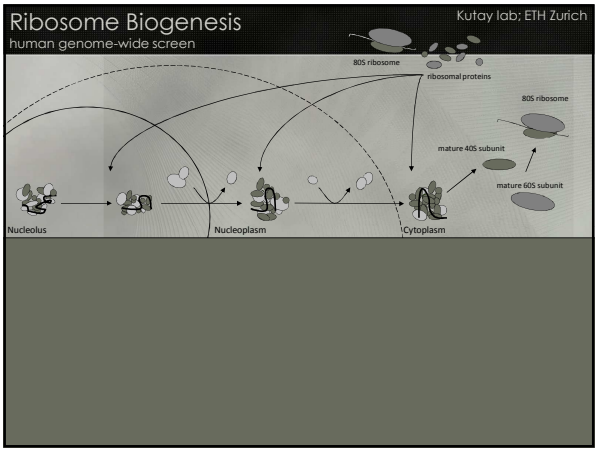
www.cellclassifier.org



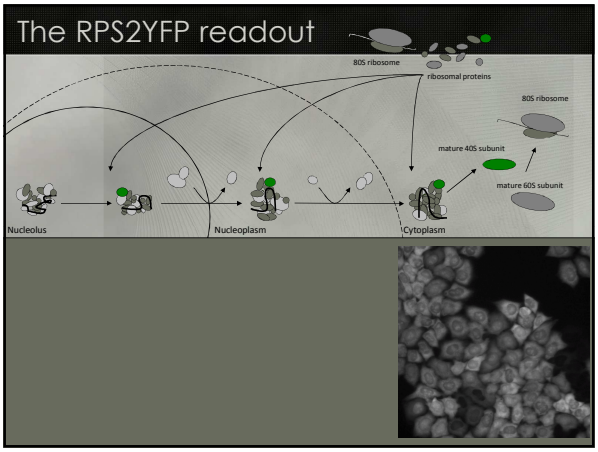
Ribosome Biogenesis

human genome-wide screen

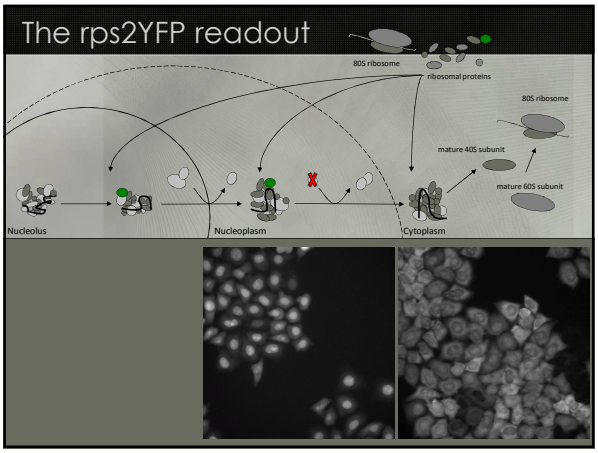
Kutay lab; ETH Zurich



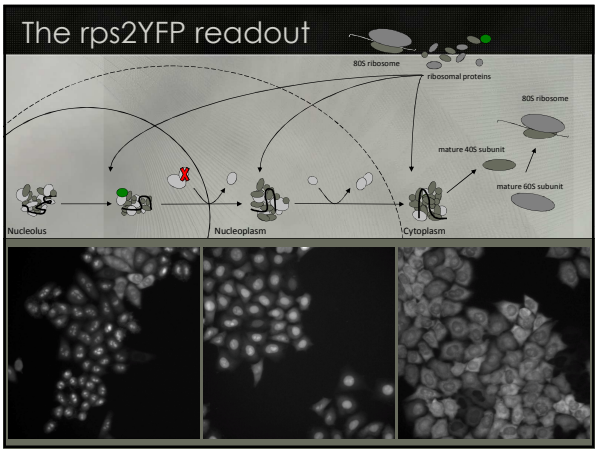
The RPS2YFP readout



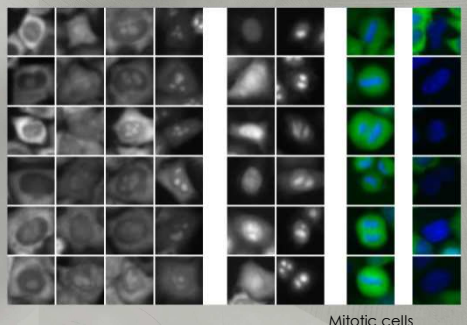
The rps2YFP readout



The rps2YFP readout



Advanced Cell Classifier I.



Normal cells (1-4) Hit cells (1, 2) Mitotic cells No YFP signal

Thomas Wild, Peter Horvath, et.al. A protein inventory of human ribosome biogenesis reveals an essential function of Exportin 5 in 60S subunit export. ***PLoS Biol*** *(2010).*

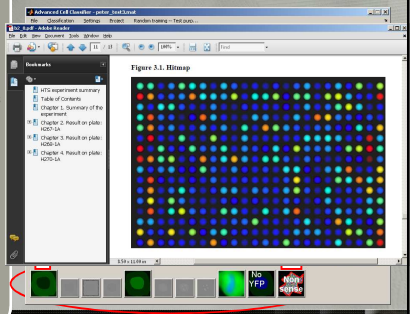
41

Training, classification

Advanced Cell Classifier

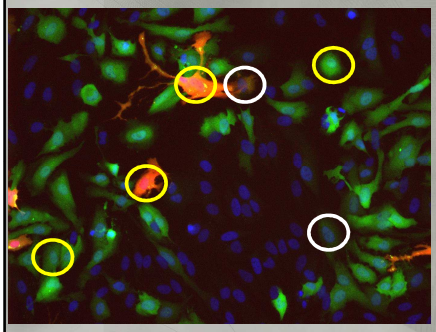
www.cellclassifier.org

- Custom written software
- Easy training by clicking
- Predefined phenotypes and subtypes
- Quick prediction
- Machine learning and simple feature-based statistics
- Prediction of the entire screen and quick report (pdf, html, xml, csv)
- Available learning methods:
  - Neural network
  - Support vector machine
  - Random forest
  - Logistic
  - 20+ more



42

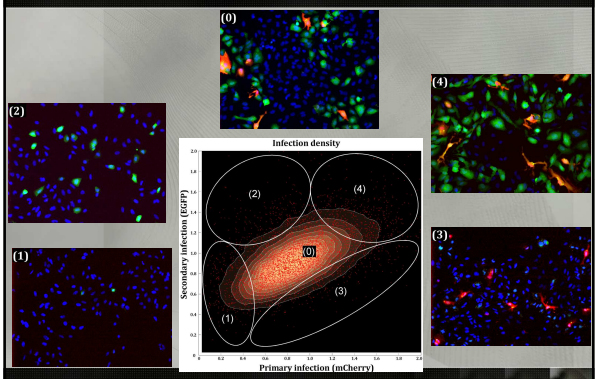
Further results – Vaccinia genome-wide screen



Jason Mercer, ETH Zurich

Readout: % of primary and secondary infected cells

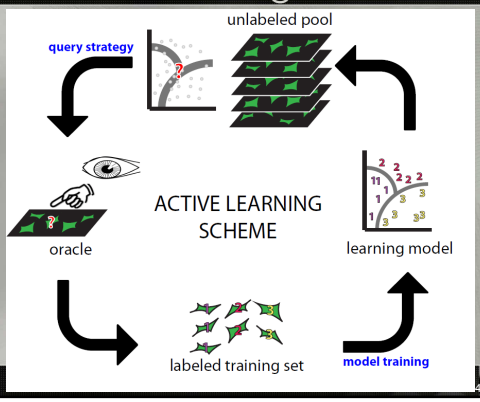
Further results – Vaccinia genome-wide screen



Jason Mercer, ETH Zurich

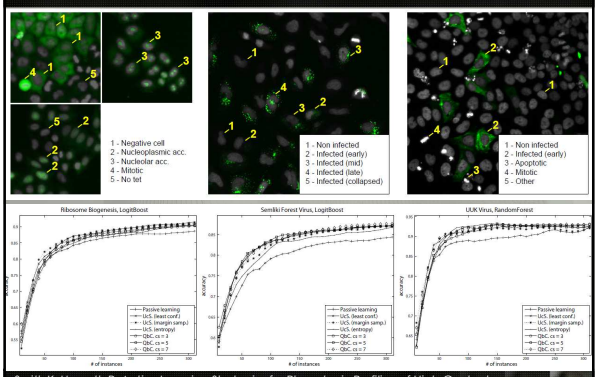
44

Active learning for HCS



45

Results



Smith K, Horvath P: Active Learning Strategies for Phenotypic Profiling of High-Content Screens. Journal of Biomolecular Screening. 2014.

46

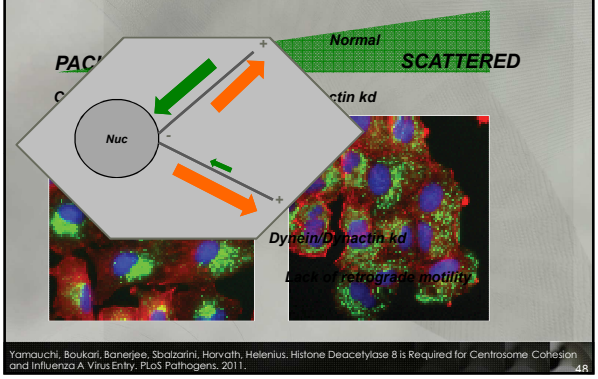
Beyond single cell-based discrete decisions

REGRESSION MODELS FOR HCA

47

Localization of Late Endosome/Lysosomes

Helenius lab; ETH Zurich

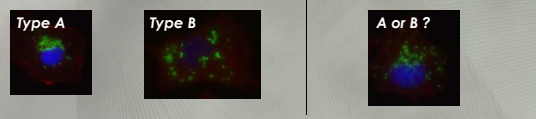


Yamauchi, Boukari, Banerjee, Sbalzarini, Horvath, Helenius. Histone Deacetylase 8 is Required for Centrosome Cohesion and Influenza A Virus Entry. PLoS Pathogens. 2011.

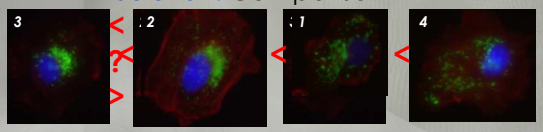
48

Regression models for HCA

➢Problem: Discrete decisions

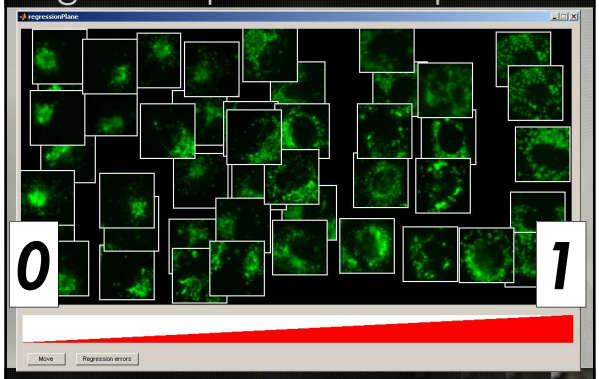


➢Solution: Comparison

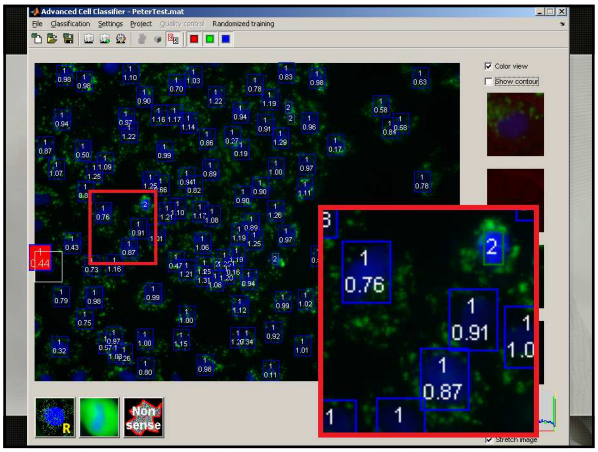


49

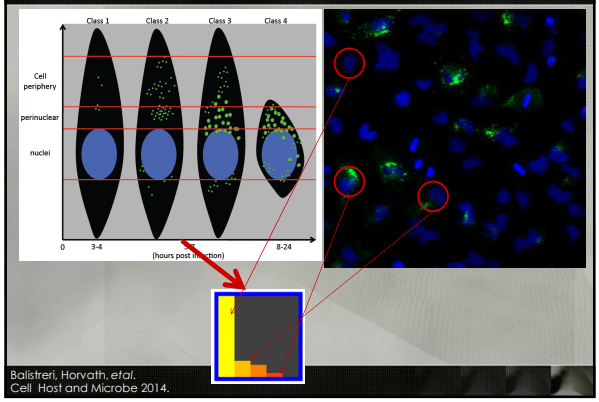
Regression plane concept



50

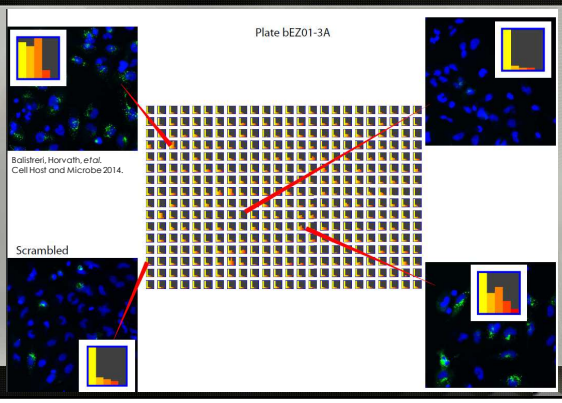


Further results - Semliki Forest virus genome-wide screen



Balistreri, Horvath, *et al.*
Cell Host and Microbe 2014.

Results

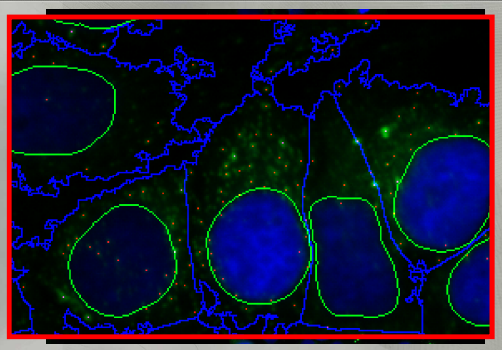


Lipid droplet phenotype analysis

MORE COMPLEX DECISIONS

54

Lipid droplet phenotype analysis



Data curtsey: Surakka, ..., S. Timonen, V. Pietiäinen,... *etal:* The impact of low-frequency and rare variants on lipid levels. **Nature Genetics** in Revision

55

Using supervised classification

56

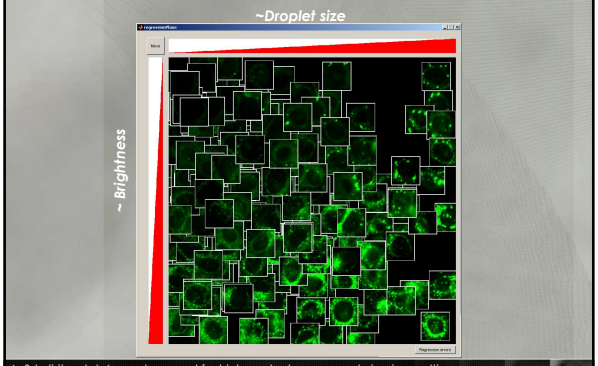
2D regression plane concept

- Can we predict/learn more than one property of the cells?
  - Yes, we can!
- I do not want to talk/analyze in terms of image processing quantities (jargon) but based on visual observations

A. Szkalisity, *etal:* A novel concept for high-content screen analysis using multi-parametric regression models to predict continuous cellular processes; KEPAF15

57

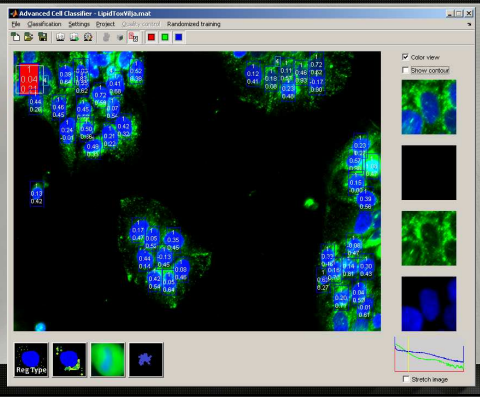
Example of an annotated plane



A. Szkalisity, *etal:* A novel concept for high-content screen analysis using multi-parametric regression models to predict continuous cellular processes; KEPAF15

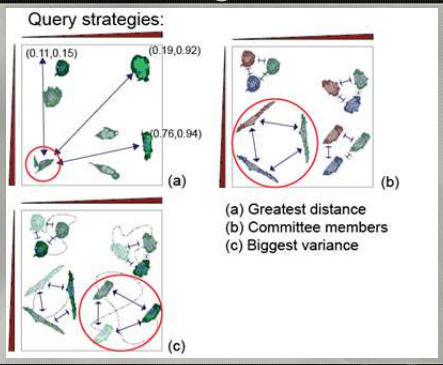
58

Results



59

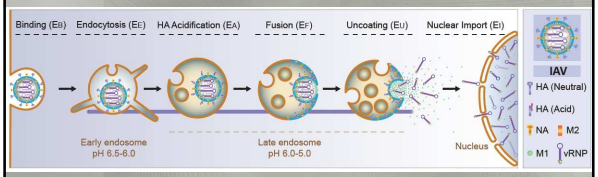
Active regression



A. Szkalisity, *J. Peltonen, P. Horvath – Work in progress*

60

Sequential assays



- Consecutive HCS assays to more precisely localize genetic or drug effects and determine timing
- Reverse subsets

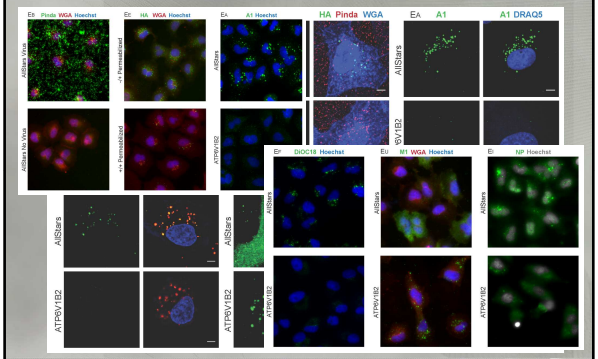
$Event_1$ -> $Event_2$...-> $Event_N$

Detected actors N
Detected actors 2
Detected actors 1

Banerjee, …, Horvath, Yamauchi, Helenius; ***Science 2014.***
Banerjee, Yamauchi, Helenius, Horvath; ***PLoS One 2013.***

61

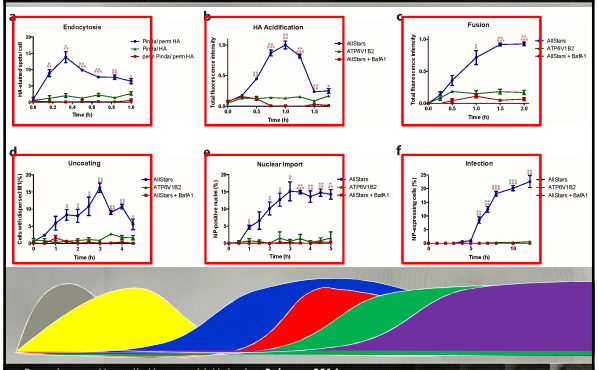
High-content analysis of sequential events during influenza A virus infection



Banerjee, …, Horvath, Yamauchi, Helenius; ***Science 2014.***
Banerjee, Yamauchi, Helenius, Horvath; ***PLoS One 2013.***

62

Results using time course data

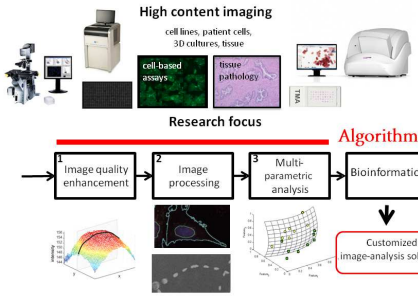


Banerjee, …, Horvath, Yamauchi, Helenius; ***Science 2014.***
Banerjee, Yamauchi, Helenius, Horvath; ***PLoS One 2013.***

63

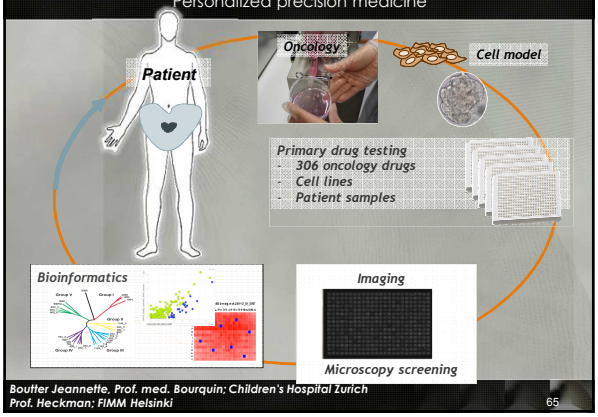
Personalized and translational medicine

Research plan: ***Next generation image analysis solutions – towards image-based diagnostics***



64

Childhood acute lymphoblastic leukemia Personalized precision medicine

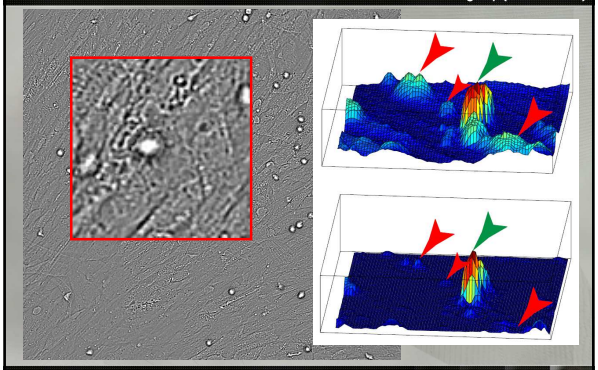


Boutter Jeannette, Prof. med. Bourquin; Children's Hospital Zurich
Prof. Heckman; FIMM Helsinki

65

Acute myeloid leukemia: drug testing using live cell microscopy

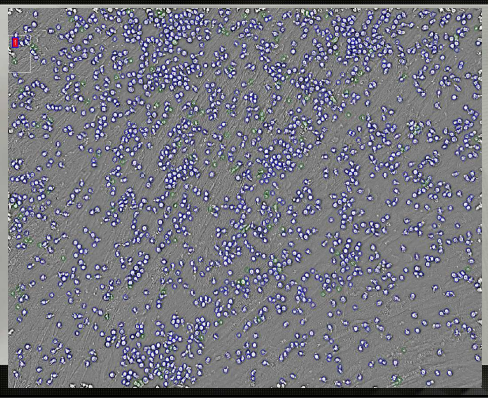
In collaboration with Prof. Caroline Heckman's group (FIMM Helsinki)



Based on: *Boutter, Horvath, Bourquin;* ***Oncotarget 2014.***

66

Results: image segmentation and machine learning



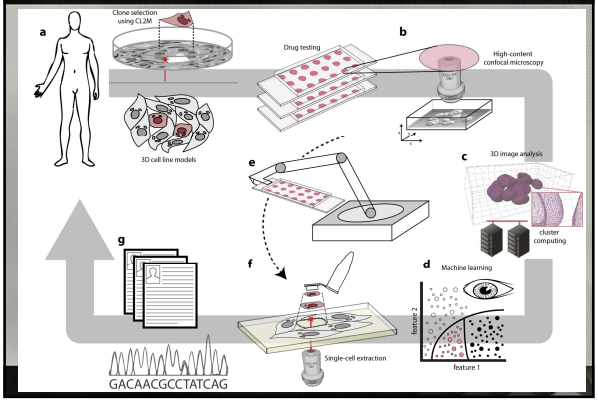
67

Can we learn more? How?

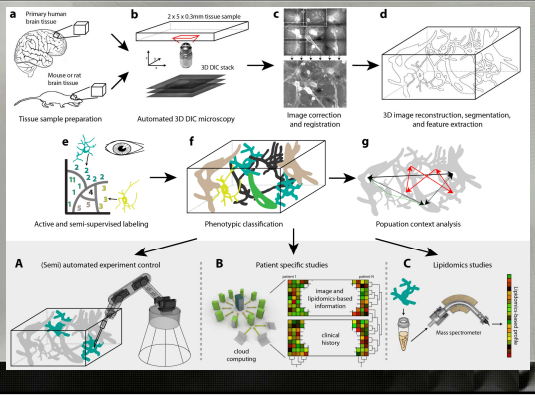
SINGLE CELLS

68

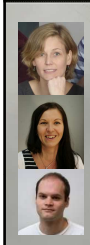
The CL²M system (correlative light-light microscopy)

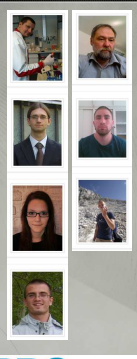


Current work, National Brain Program



Thank you!





www.brc.hu/sysbiol

FIDIPRO Finland Distinguished Professor Programme

MARIE CURIE ACTIONS

We are hiring

PhD and PostDoc positions in:

- Image analysis using shape priors
- Machine learning for image-based phenotypic discovery

72