Automatic quantification of subtle cellular phenotypes in microscopy-based high-throughput experiments

Vebjorn Ljosa
Broad Institute of MIT and Harvard
Cambridge, MA, USA

With other members of the Imaging Platform and numerous collaborators

Fourth International Workshop on Machine Learning in Systems Biology (MLSB)
Edinburgh, Scotland 2010-10-16
van Leeuwenhoek’s microscope (late 1600s)

Nikon’s First Microscope (circa early 1900s)

more relevant for systems biology

modern robotic microscope

more in need of machine learning
Image data are exciting.

You should join us in working on them.
1. The Imaging Platform at the Broad Institute

2. CellProfiler – image-analysis software

3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype

4. Large-scale training of a classifier for subtle phenotype changes

5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition

6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
The Broad Institute’s unusual organization

Faculty member + lab

Imaging platform

Chemical biology platform

Sequencing platform

Faculty member + lab

[phdcomics.com]
The Imaging Platform at the Broad Institute

Anne Carpenter
Director

Methods development

Vebjorn Ljosa
Carolina Wahlby

Image assay development

Mark Bray
David Logan
Kate Madden

Software engineering

Lee Kamentsky
Adam Fraser
1. The Imaging Platform at the Broad Institute

2. CellProfiler – image-analysis software

3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype

4. Large-scale training of a classifier for subtle phenotype changes

5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition

6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
CellProfiler—why was new software needed?

Assay-specific pre-packaged commercial software

Fast

Poor results on crowded cells or unusual cell types

Designed for standard assays

- Cell adhesion
- Neurite outgrowth
- Micronucleus formation
- Protein translocation
- Cell cycle analysis
- Adipogenesis
- Reporter gene analysis
- Cell viability
- Apoptosis
- Cell migration

Often inflexible

Proprietary methods

Expensive

Published raw source code

Advanced algorithms

Heavily customized, not generalizable

Requires programming skills + a lot of time to adapt to new situations

Rarely applied outside the originating lab

```matlab
function [rgOut, varargout] = ImDAPI2Rg(imDAPIin, LoGDim, LoGHW, MinArea)

wiendim=[5 5];

rgLoG=fspecial('log',LoGDim,LoGHW);
imLoGout=imfilter(double(imDAPIin),rgLoG);
imLoGoutW=wiener2(imLoGout,wiendim);

rgNegCurve=imLoGoutW<-1;

%set outsides
rgNegCurve([1 end],[1 end])=1;
rgNegCurve([1:end],[1 end])=1;

%Throw out noise, label regions
rgArtOpen=bwareaopen(rgNegCurve,MinArea,4);
```
Convenient for biologists, convenient for algorithm comparisons.
Typical CellProfiler pipeline

1. Image processing modules
2. Illumination correction modules
3. Object identification modules
4. Measurement modules

Measurements for every cell in every image (number, location, size, shape, intensity, texture) can be analyzed by:
1. Built-in CellProfiler data tools
2. Exporting to spreadsheet
3. Exporting to database
Measure everything first, ask questions later

“Cytological profile”: collection of measurements describing the appearance of a cell

Data plot: Noa Shefi
Successful image-based assays

Phenotypes:
- Cell count
- Cell size
- DNA content
- Nuclear speckles
- Cytoplasm/nucleus localization
- Membrane localization
- Protein or phospho-protein levels
- Metastasis
- Wound healing
- Metaphase
- Anaphase/telophase
- Prophase
- Shape/texture
- Crescent-shaped nuclei
- Peas-in-a-pod
- Cells-on-the-move
- Long projections
- Crooked projections
- Hyphae-like fingers
- Actin at contractile ring/cell junctions
- Internal actin
- Actin circles
- Large spread cells
- Phospho-histone H3 nuclear dots
- Bi/multinucleate

Cell types:
- Drosophila Kc167 cells
- Drosophila S2R+ cells
- Drosophila epithelial tissue
- Drosophila embryo
- Human HT29 cells
- Human A549 cells
- Human TOV21G cells
- Human H1299 lung carcinoma cells
- Human biopsied prostate gland tissue
- Human adult mesenchymal stem cells
- Mouse NIH/3T3 cells
- Mouse neural precursor cells derived from embryos
- Mouse lung tissue sections
- Mouse isolated germ cells
- Rat H9c2 cells
- C. elegans worms (preliminary)
- Neurons (preliminary)
- S. cerevisiae cells
- Yeast colonies
- Yeast growth patches
- Array grids
CellProfiler is free open-source software designed to enable biologists without training in computer vision or programming to quantitatively measure phenotypes from thousands of images automatically. See our papers on analyzing cell images and non-cell images.

Examples

The best way to learn how to use CellProfiler and CellProfiler Analyst is to download our examples and try it out!

CellProfiler example images and pipelines (jump to CellProfiler Analyst example)

How do I get started?

Download example images along with pipelines so you can get immediate hands-on experience in using CellProfiler.

1. Select an example from the pipeline list below where the cell type (or object type) resembles yours. You can move your mouse over an image below to see an example of the CellProfiler analysis, or click to see an expanded view.
2. Download the example images and pipeline from the list below and run it in CellProfiler to see how it works.
3. Try the pipeline on your own images.
4. Adjust the pipeline to identify objects properly in your images. This most often includes changing the size range of the objects. See the Tutorial page for step by step instructions.
5. Still stuck? See if your question has been answered on the forum.

Please note that each example links to a compressed ZIP file containing the following:

1. CellProfiler pipelines, in both version 1.0 (.mat) and version 2.0 (.cp) formats.
2. Example images which are to be used as input for the pipeline.

Basic Pipelines

These pipelines are made for simple cellular and tissue image assays, and include some basic measurements.

- **Human cells**: Human HT29 cells are fairly smooth and elliptical. This pipeline demonstrates how to accurately identify these cells and how to measure important features such as morphology, count, intensity and texture.
  - [Download](0.3 MB)

- **Fruit flies**: In contrast to the HT29 cells, Drosophila Kc167 cells are a highly textured and clumpy cell type. This pipeline demonstrates how to identify these clumpy cells and obtain morphological, intensity and texture measurements.
  - [Download](4 MB)

- **Tumors**: A simple pipeline that identifies and counts tumors in a mouse lung, and then measures their size.
  - [Download](0.9 MB)
CellProfiler around the world

CellProfiler is downloaded 400x/month, 11,000x total (~50% USA, 90% non-profit institutions).
The CellProfiler project

free, at www.CellProfiler.org

Selected high-throughput screens using CellProfiler

<table>
<thead>
<tr>
<th>Lab</th>
<th>Publication</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root lab</td>
<td>Cell, 2006</td>
<td>Screen for cell cycle regulators</td>
</tr>
<tr>
<td>Alon lab</td>
<td>Nature Methods, 2006</td>
<td>High-throughput analysis of protein dynamics</td>
</tr>
<tr>
<td>Neefjes lab</td>
<td>Nature, 2007</td>
<td>Screen for levels of Salmonella typhimurium infection</td>
</tr>
<tr>
<td>Raff lab</td>
<td>PLoS Biology, 2008</td>
<td>Screen for centriole duplication and mitotic PCM recruitment</td>
</tr>
<tr>
<td>Carpenter lab</td>
<td>PNAS 2009 &amp; BMC Bioinformatics 2008</td>
<td>Screens for &gt; 15 diverse phenotypes in human and Drosophila cells</td>
</tr>
<tr>
<td>Shokat lab</td>
<td>Cancer Cell, 2008</td>
<td>Screen for PI3K inhibitor resistance mutations in S. cerevisiae</td>
</tr>
<tr>
<td>Pelkmans lab</td>
<td>Nature, 2009</td>
<td>High-throughput infection assay</td>
</tr>
<tr>
<td>Ausubel lab</td>
<td>ACS Chem Bio, 2009</td>
<td>Screen for inhibitors of infection by E. faecalis</td>
</tr>
</tbody>
</table>

CellProfiler’s is the 5th most-accessed Genome Biology paper of all time
### Experiments we have completed recently

<table>
<thead>
<tr>
<th>Biological process/phenotype</th>
<th>Laboratory</th>
<th>Samples tested</th>
<th>Number of fields of view (images) processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meiosis</td>
<td>Terry Orr-Weaver (Whitehead)</td>
<td>RNAi</td>
<td>84,000</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>Vamsi Mootha (HMS/MGH)</td>
<td>chemicals</td>
<td>100,000</td>
</tr>
<tr>
<td>Morphology</td>
<td>AstraZeneca</td>
<td>chemicals</td>
<td>109,200</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>AstraZeneca</td>
<td>chemicals</td>
<td>109,200</td>
</tr>
<tr>
<td>Breast cancer/Heregulin</td>
<td>Eric Lander (Broad)</td>
<td>RNAi</td>
<td>144,798</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>Deb Hung (Broad)</td>
<td>chemicals</td>
<td>164,000</td>
</tr>
<tr>
<td>Glioma</td>
<td>David Sabatini &amp; Bill Hahn (Whitehead, Harvard, Dana-Farber)</td>
<td>RNAi</td>
<td>286,000</td>
</tr>
<tr>
<td>Polyploidy: AMKL</td>
<td>John Crispino (Northwestern University)</td>
<td>chemicals, some RNAi</td>
<td>530,000</td>
</tr>
<tr>
<td>Hematopoietic stem cells</td>
<td>David Scadden and Stuart Schreiber (HMS, MGH, Broad)</td>
<td>chemicals, some RNAi</td>
<td>465,448</td>
</tr>
<tr>
<td>Leukemic stem cells</td>
<td>Gary Gilliland (BWH/HMS)</td>
<td>chemicals, some RNAi</td>
<td>1,040,098</td>
</tr>
<tr>
<td>Hepatotoxicity</td>
<td>Sangeeta Bhatia (MIT)</td>
<td>chemicals</td>
<td>1,135,093</td>
</tr>
</tbody>
</table>
1. The Imaging Platform at the Broad Institute

2. CellProfiler – image-analysis software

3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype

4. Large-scale training of a classifier for subtle phenotype changes

5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition

6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
Thousands of samples

Add thousands of chemicals or RNAi agents, each one in a different sample

Hit
Phosphorylated histone H3 – Mean

G0, G1, S, G2

M?

?
DAPI and mean pHH3
“Simple” phenotypes: one or two features is enough

X-axis: DNA content

Y-axis: phospho-H3 staining

Anaphase/telophase

Late prophase/metaphase
Example of complex phenotype: motile T47D cells

Normal T47D cells

Features associated with cell motility: lamellipodia, filopodia, polarized cell shape, F-actin nucleation at filapodia, less clumping

Inducable by HRG
Challenges

- How to get and crossvalidate with rare phenotypes?
- How to make classifier interpretable by biologist?
- Normalization of features?
- Dimensionality reduction?
- Prevent overfitting?
- Avoid having to tune parameters?
Rare phenotypes: HT29 colon cancer cells

- Crescent-shaped nuclei
- Peas-in-a-pod
- Cells-on-the-move
- Long projections
- Crooked projections
- Actin at contractile ring/cell junctions
- Hyphae-like fingers

[Jones et al., PNAS, 2009]
Iterative machine learning

Rule

Iteration

Yes

No

[Jones et al., PNAS, 2009]

Using gentle boosting [Friedman et al., 1998]
Incorporated into CellProfiler Analyst
Rules for distinguishing HRG-stimulated T47D cells

(IF MeanSpeckles_AreaShape_Area > 12.000000, 0.827550, -0.350258) +
(IF Cells_AreaShape_FormFactor > 0.449767, -0.331746, 0.706321) +
(IF CellMembrane_Texture_3_CorrGreen_DifferenceVariance > 0.718124, 0.593955, -0.198424) +
(IF Cells_Intensity_CorrGreenSpeckle_MaxIntensity > 0.370382, 0.787301, -0.204062) +
(IF CellMembrane_Intensity_CorrGreenSpeckle_MinIntensityEdge > 0.001284, 0.275866, -0.500179) +
(IF Cells_Texture_3_CorrGreen_SumEntropy > 1.710600, -0.199515, 0.700658) +
(IF Nuclei_AreaShape_Perimeter > 96.669000, -0.177882, 0.788582) +
(IF Cells_Texture_3_CorrGreen_GaborY > 0.147318, -0.215618, 0.613682) +
(IF Cells_Intensity_CorrGreenSpeckle_StdIntensity > 0.033557, 0.742323, -0.169222) +
(IF Cells_Intensity_CorrGreenSpeckle_StdIntensityEdge > 0.009328, -0.118830, 0.956272)
Multiple classes
QC and Tracking Down Hits

### Enrichments grouped by Gene

<table>
<thead>
<tr>
<th>gene</th>
<th>Counts positive</th>
<th>Counts negative</th>
<th>p(Enriched) positive</th>
<th>p(Enriched) negative</th>
<th>Enriched Score positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME1</td>
<td>179</td>
<td>2248</td>
<td>0.96984534</td>
<td>0.03015463</td>
<td>1.5073480914</td>
</tr>
<tr>
<td>PRPS2</td>
<td>140</td>
<td>2002</td>
<td>0.91686196</td>
<td>0.08313715</td>
<td>1.04250419841</td>
</tr>
<tr>
<td>TK1</td>
<td>146</td>
<td>2358</td>
<td>0.84006921</td>
<td>0.15993923</td>
<td>0.720383014599</td>
</tr>
<tr>
<td>PMS1</td>
<td>86</td>
<td>1356</td>
<td>0.83575721</td>
<td>0.16424677</td>
<td>0.706593839064</td>
</tr>
<tr>
<td>GALK1</td>
<td>119</td>
<td>1938</td>
<td>0.82593326</td>
<td>0.17408002</td>
<td>0.676229179724</td>
</tr>
<tr>
<td>MAPK13</td>
<td>148</td>
<td>2908</td>
<td>0.64246398</td>
<td>0.35755690</td>
<td>0.25452898308</td>
</tr>
<tr>
<td>Gabra2</td>
<td>112</td>
<td>2208</td>
<td>0.63344543</td>
<td>0.36661946</td>
<td>0.237570577506</td>
</tr>
<tr>
<td>PHKG2</td>
<td>72</td>
<td>1431</td>
<td>0.61455659</td>
<td>0.38561178</td>
<td>0.202601261265</td>
</tr>
<tr>
<td>MAP2K3</td>
<td>181</td>
<td>3784</td>
<td>0.56842729</td>
<td>0.43162143</td>
<td>0.119620956956</td>
</tr>
<tr>
<td>STK19</td>
<td>105</td>
<td>2377</td>
<td>0.46921976</td>
<td>0.53066779</td>
<td>-0.0535384459292</td>
</tr>
<tr>
<td>Gabra1</td>
<td>92</td>
<td>2086</td>
<td>0.46803014</td>
<td>0.53249026</td>
<td>-0.0556131897801</td>
</tr>
<tr>
<td>MAP2K6</td>
<td>91</td>
<td>2169</td>
<td>0.41092623</td>
<td>0.58938132</td>
<td>-0.15640582072</td>
</tr>
<tr>
<td>Gpr12</td>
<td>79</td>
<td>1978</td>
<td>0.35990666</td>
<td>0.64030450</td>
<td>-0.2500534158</td>
</tr>
<tr>
<td>PDXK</td>
<td>112</td>
<td>2829</td>
<td>0.34267214</td>
<td>0.65696491</td>
<td>-0.282903238922</td>
</tr>
<tr>
<td>MAPK11</td>
<td>48</td>
<td>1360</td>
<td>0.26241467</td>
<td>0.73790298</td>
<td>-0.448824138167</td>
</tr>
<tr>
<td>Gabra3</td>
<td>84</td>
<td>2504</td>
<td>0.19606696</td>
<td>0.80500784</td>
<td>-0.6128154381</td>
</tr>
</tbody>
</table>
1. The Imaging Platform at the Broad Institute

2. CellProfiler – image-analysis software

3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype

4. Large-scale training of a classifier for subtle phenotype changes

5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition

6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
The phenotype of motile T47D cells

Features associated with cell motility: lamellipodia, filopodia, polarized cell shape, F-actin nucleation at filopodia, less clumping

Normal T47D cells

Unstimulated

Stimulated by heregulin
Built training set of ~300 cells
Why cut out the human?

HRG stimulation

Motility

Captured by the human-trained classifier

Metastasis

Tumor growth

Not captured
Labeling for automatic training set

Replicate 1  Replicate 2  Replicate 3

Unstimulated  Stimulated by heregulin

45% motile cells  55% motile cells
Two ways to improve the classifier

Accuracy

Training set size

See [Banko & Brill, 2001]
Random Fourier features

$$\cos (\omega' x + b)$$

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>$k(\Delta)$</th>
<th>$p(\omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$e^{-\frac{|\Delta|_2^2}{2}}$</td>
<td>$(2\pi)^{-\frac{D}{2}} e^{-\frac{|\omega|_2^2}{2}}$</td>
</tr>
<tr>
<td>Laplacian</td>
<td>$e^{-|\Delta|_1}$</td>
<td>$\prod_d \frac{1}{\pi(1+\omega_d^2)}$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$\prod_d \frac{2}{1+\Delta_d^2}$</td>
<td>$e^{-|\Delta|_1}$</td>
</tr>
</tbody>
</table>

[Rahimi and Recht, NIPS, 2007]
Random features

[Rahimi and Recht, NIPS, 2007]

Datapoints in a fairly low-dimensional space (a few hundred dimensions) spanned by random Fourier bases

The kernel trick

Original data

Kernel function

Inner product used, e.g., by SVM.

Random features

Kernel function

Inner product used, e.g., by SVM.
Linear discriminant on random features

7.6 million training cells, 130 measurements

Mapped into 250-dimensional random feature space

Trained Fisher’s linear discriminant

source: Cooley & Lohnes ((1971))
Automatic vs. hand training

- Projection on linear discriminant in random feature space
  - Unstimulated
  - Stimulated

- Boosting score
  - Unstimulated
  - Stimulated
1. The Imaging Platform at the Broad Institute

2. CellProfiler – image-analysis software

3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype

4. Large-scale training of a classifier for subtle phenotype changes

5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition

6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
Expected likelihood kernel

\[ k(A, B) = \int \Pr(x | A) \Pr(x | B) \, dx \]

[Jebara et al., 2004]
$z(x)$ projection of $x$ into random-feature space

\[ \langle z(a), z(b) \rangle \approx k(a, b) \]

\[
v_A = \frac{1}{|A|} \sum_{x \in A} z(x)
\]

\[
\langle v_A, v_B \rangle \approx \frac{1}{|A||B|} \sum_{x \in A, y \in B} e^{-(x-y)^2/2}
\]

\[
\approx \int \Pr(x|A) \Pr(x|B) \, dx
\]
Correlation coefficient of dot products

AZ-A
AZ-H
Aphidicolin
Cytochalsin B
DMSO
Nocodazole
Taxol

0 10 20 30 40 50
Outline

1. The Imaging Platform at the Broad Institute
2. CellProfiler – image-analysis software
3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
4. Large-scale training of a classifier for subtle phenotype changes
5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
Histone deacetylases

• 11 enzymes
• Component of chromatin-regulating complexes
• Also target many non-histone proteins
• Broad relevance to cell signaling and cell state
• Induce differentiation and inhibit proliferation in cancer models
• Inhibitors used clinically for cutaneous T-cell lymphoma, others in trials for other cancers
Screening question: find specific HDAC inhibitors

Inhibition potency data from biochemical assays

HDAC1
HDAC2
HDAC3
HDAC4
HDAC5
HDAC6
HDAC7
HDAC8
HDAC9

1354, APHA, Apicidin, A.analog, nDiv 5058, CI-994, Depudecin, HCToxin, JNJ, LAO824, MCGD, Merck60, MS-275, nitubacin, nistascript, PTACh, PXD101, yroxamide, SAHA, Scriptaid, SuberoHA, Trapoxin, TSA, Tubacin, Vproic Acid, WT161
Training

\[ T = K X^{-1} \]
Testing

\[ T \times X = K' \]

Schematic
Convert scores to soft labels by logistic transform

9-simplex with the 9 HDACs and DMSO at the vertices
Proportions of cells for each topic
Correlations between classes match HDAC phylogeny

Class IIa HDACs
Class IIb HDACs
Class I HDACs
Summary

1. The Imaging Platform at the Broad Institute
2. CellProfiler – image-analysis software
3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
4. Large-scale training of a classifier for subtle phenotype changes
5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
6. Discovering latent “phenotypes” by learning to scale image features using linear regression and a topic model
Thank you!

Imaging Platform and alumni:
Mark Bray
Anne Carpenter
Adam Fraser
Thouis (Ray) Jones
Lee Kamentsky
David Logan
Kate Madden
Tejas Shah
Carolina Wähly

Collaborators:
Peter Caie
Neil Carragher
Paul Clemons
Emma Cooke
Christopher Denz
Piyush Gupta
Sigrun Gustafsdottir
Tom Houslay
Melissa Kemp
Angela Koehler
Mijung Kwon
Melissa Passino
Eric Lander
Aly Shamji