Journal Club 15/05/18

Joseph Boyd JOSEPH.BOYD@MINES-PARISTECH.FR

May 27, 2018

A Multi-Scale Convolutional Neural Network for Phenotyping High-Content Cellular Images

William J. Godinez [1,*], Imtiaz Hossain [1], Stanley E. Lazic [1,3], John W. Davies [2] and Xian Zhang [1,*]

[1]Novartis Institutes for BioMedical Research Inc., Basel, Switzerland.
[2]Novartis Institutes for BioMedical Research Inc., Cambridge, Massachusetts, USA.
[3]Current address: Quantitative Biology, AstraZeneca, Cambridge, CB4 0WG, UK.

*To whom correspondence should be addressed.

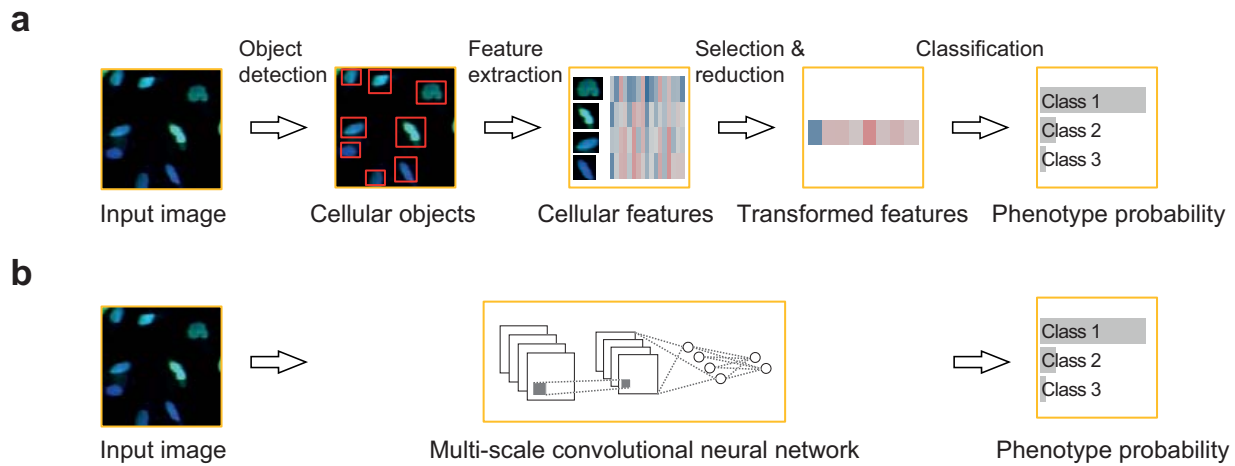
Associate Editor: Prof. Robert Murphy

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

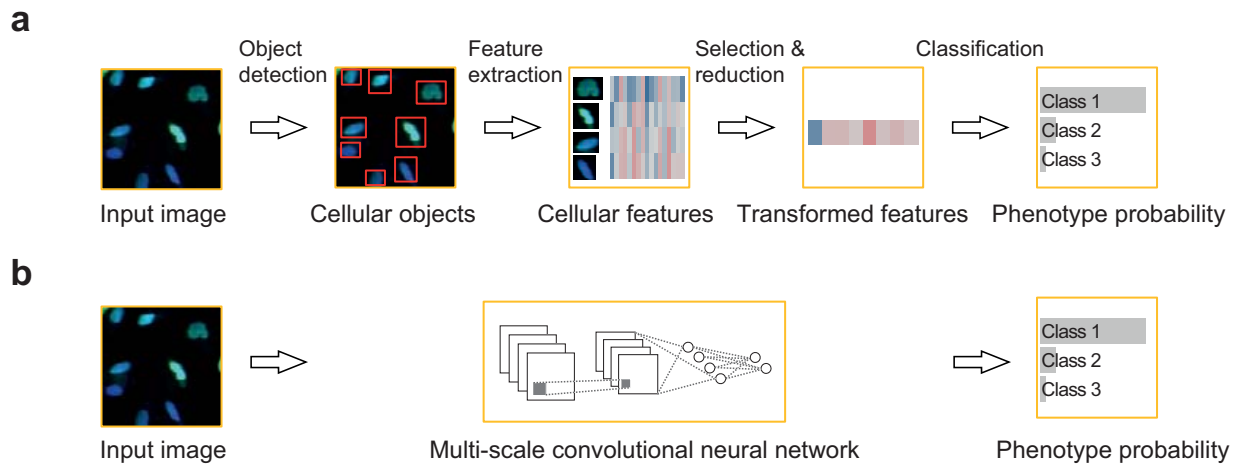
Motivation: Identifying phenotypes based on high-content cellular images is challenging. Conventional image analysis pipelines for phenotype identification comprise multiple independent steps, with each step requiring method customization and adjustment of multiple parameters.
Results: Here we present an approach based on a multi-scale convolutional neural network (M-CNN) that classifies, in a single cohesive step, cellular images into phenotypes by using directly and solely the images' pixel intensity values. The only parameters in the approach are the weights of the neural network, which are automatically optimized based on training images. The approach requires no a priori knowledge or manual customization, and is applicable to single- or multi-channel images displaying single or multiple cells. We evaluated the classification performance of the approach on eight diverse benchmark datasets. The approach yielded overall a higher classification accuracy compared to state-of-the-art results, including those of other deep CNN architectures. In addition to using the network to simply obtain a yes-or-no prediction for a given phenotype, we use the probability outputs calculated by the network to quantitatively describe the phenotypes. Our study shows that these probability values correlate with chemical treatment concentrations. This finding validates further our approach and enables chemical treatment potency estimation via convolutional neural networks.

Per-cell vs. per-image analysis



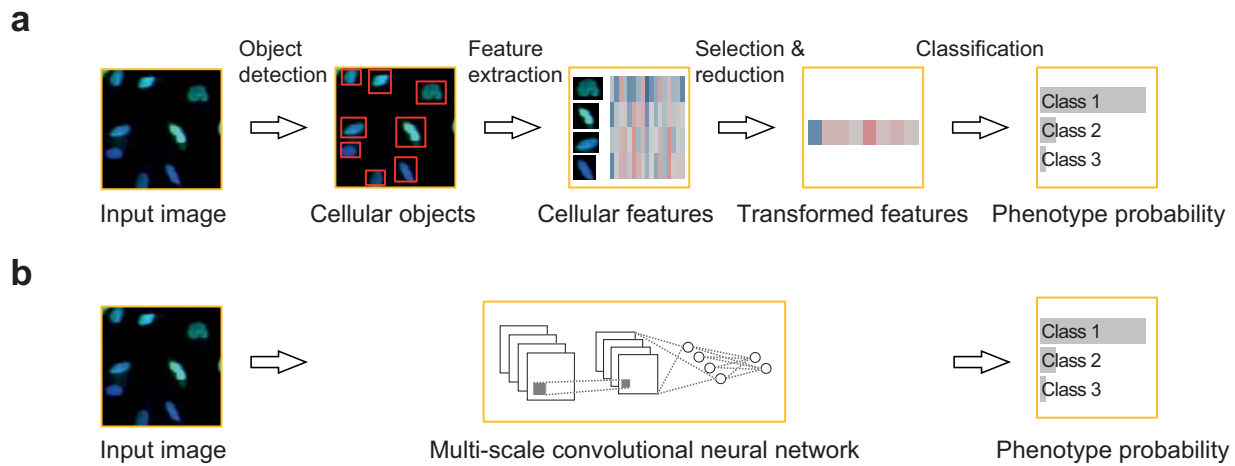
- The joint optimisation of all parameters across the entire [classical] analysis pipeline remains challenging (page 2)

Per-cell vs. per-image analysis



- The joint optimisation of all parameters across the entire [classical] analysis pipeline remains challenging (page 2)
- A CNN-based approach that [...] in one cohesive step, classifies cellular images into phenotypes (page 2)

Per-cell vs. per-image analysis



- The joint optimisation of all parameters across the entire [classical] analysis pipeline remains challenging (page 2)
- A CNN-based approach that [...] in one cohesive step, classifies cellular images into phenotypes (page 2)
- See also Orlov et al. (2008) or Uhlmann and Singh (2016)

Data and learning task

- Network tested on 8 datasets spanning 6 cell lines (versatility)

[1] Broad Institute Bioimage Benchmark Collection

Data and learning task

- Network tested on 8 datasets spanning 6 cell lines (versatility)
- Tasks differ by dataset, for example CHO cell dataset used in Boland et al. 2001 for organelle classification

[1] Broad Institute Bioimage Benchmark Collection

Data and learning task

- Network tested on 8 datasets spanning 6 cell lines (versatility)
- Tasks differ by dataset, for example CHO cell dataset used in Boland et al. 2001 for organelle classification
- Of particular interest is the image set **BBBC21v2**[1] (MCF-7 breast cancer cell line)

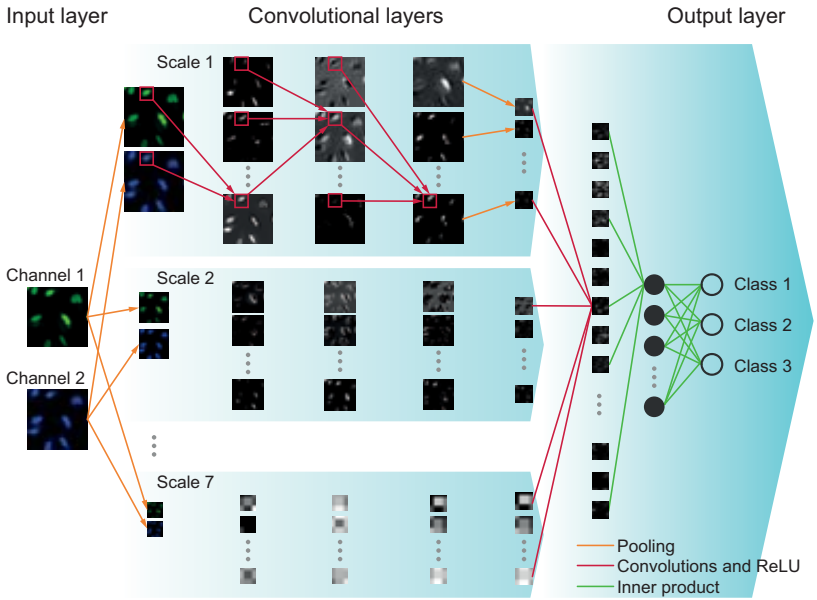
[1] Broad Institute Bioimage Benchmark Collection

Data and learning task

- Network tested on 8 datasets spanning 6 cell lines (versatility)
- Tasks differ by dataset, for example CHO cell dataset used in Boland et al. 2001 for organelle classification
- Of particular interest is the image set **BBBC21v2**[1] (MCF-7 breast cancer cell line)
- 38 compounds at 8 concentrations annotated into 13 mechanisms of action (MOA), including negative (DMSO)

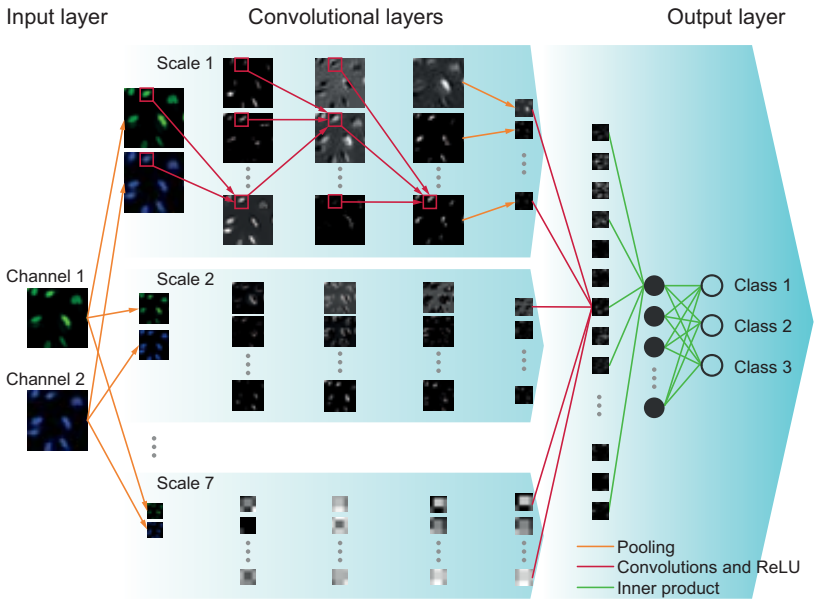
[1] Broad Institute Bioimage Benchmark Collection

Multi-scale CNN architecture



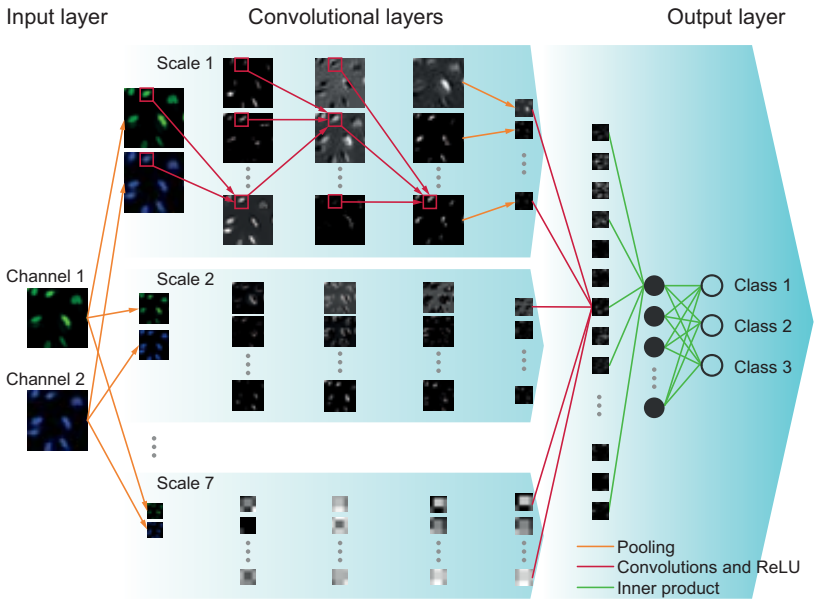
- MCNN consists of 7 paths of three 5×5 convolutional layers

Multi-scale CNN architecture



- MCNN consists of 7 paths of three 5×5 convolutional layers
- Each path processes a downsampled version of the input $(1, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^6})$ before aggregation

Multi-scale CNN architecture



- MCNN consists of 7 paths of three 5×5 convolutional layers
- Each path processes a downsampled version of the input $(1, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^6})$ before aggregation
- Network designed to be robust to cell line, magnification, etc.

Model evaluation - hard and soft

Network produces $N_p = 13$ probabilities per replicate / drug / concentration / field

Model evaluation - hard and soft

Network produces $N_p = 13$ probabilities per replicate / drug / concentration / field

- **Hard classification**: Predict MOA in a leave-one-out cross validation scheme

$$\hat{y} = \arg\max_k p(y = k|\mathbf{x})$$

Model evaluation - hard and soft

Network produces $N_p = 13$ probabilities per replicate / drug / concentration / field

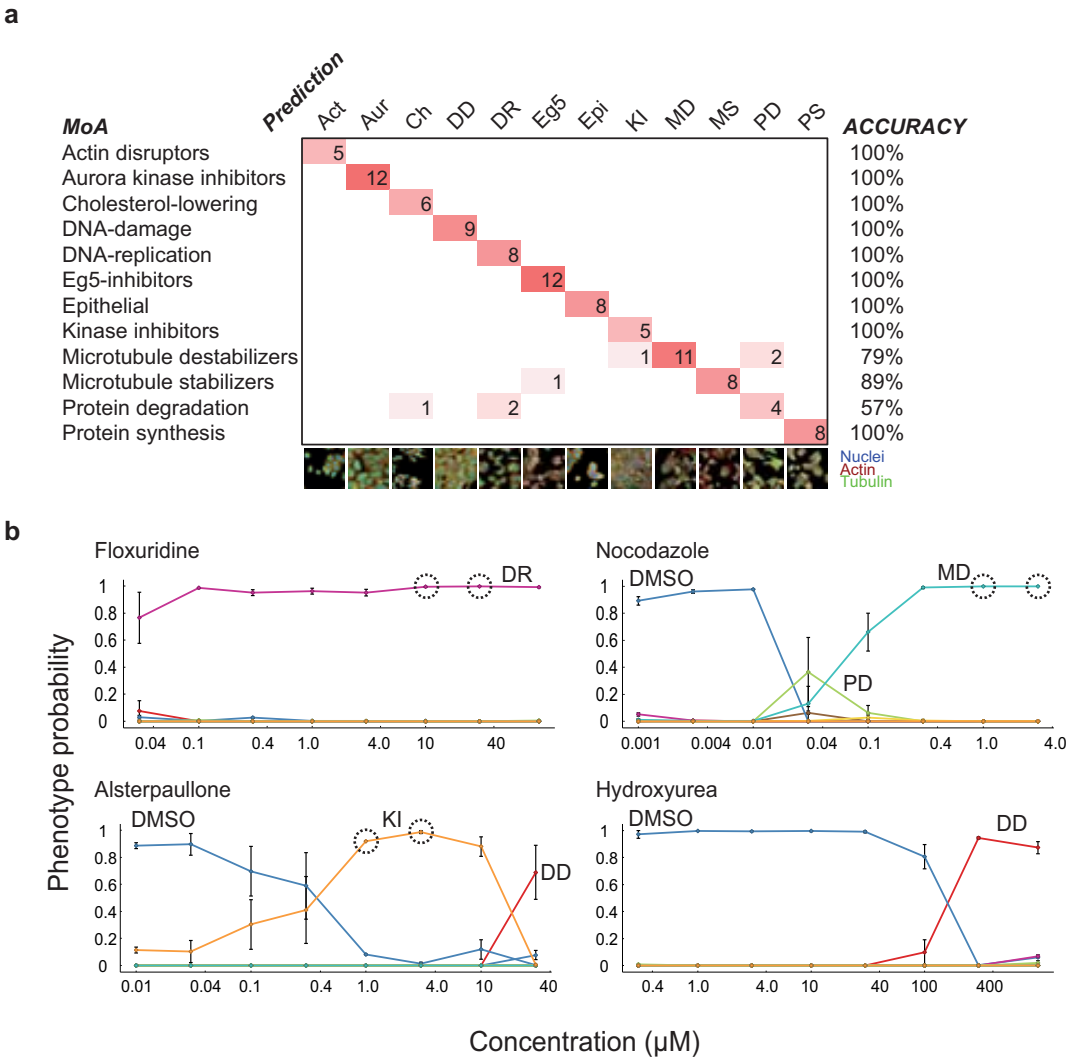
- **Hard classification**: Predict MOA in a leave-one-out cross validation scheme

$$\hat{y} = \arg\max_k p(y = k|\mathbf{x})$$

- **Soft classification**: Track probabilities $p_{rf}(y = k)$ over titration series as median over replicate r of median over field f of view

$$\rho_{rfk} = med\{med\{p_{rf}(y = k)\}_f\}_r$$

Model evaluation - hard and soft



MOA prediction with multiple cell lines

Original Report

SLAS Technology
1–9
© 2017 Society for Laboratory Automation and Screening
DOI: 10.1177/2472630317740831
journals.sagepub.com/home/jla
SAGE

Compound Functional Prediction Using Multiple Unrelated Morphological Profiling Assays

France Rose[1*], Sreetama Basu[1*], Elton Rexhepaj[1,2], Anne Chauchereau[3], Elaine Del Nery[2], and Auguste Genovesio[1]

Abstract

Phenotypic cell–based assays have proven to be efficient at discovering first-in-class therapeutic drugs mainly because they allow for scanning a wide spectrum of possible targets at once. However, despite compelling methodological advances, posterior identification of a compound's mechanism of action (MOA) has remained difficult and highly refractory to automated analyses. Methods such as the cell painting assay and multiplexing fluorescent dyes to reveal broadly relevant cellular components were recently suggested for MOA prediction. We demonstrated that adding fluorescent dyes to a single assay has limited impact on MOA prediction accuracy, as monitoring only the nuclei stain could reach compelling levels of accuracy. This observation suggested that multiplexed measurements are highly correlated and nuclei stain could possibly reflect the general state of the cell. We then hypothesized that combining unrelated and possibly simple cell-based assays could bring a solution that would be biologically and technically more relevant to predict a drug target than using a single assay multiplexing dyes. We show that such a combination of past screen data could rationally be reused in screening facilities to train an ensemble classifier to predict drug targets and prioritize a possibly large list of unknown compound hits at once.

Keywords

target prediction, high-content screening, mechanism of action, ensemble classifier

Premise

- The contribution of phenotypic screening to the discovery of first-in-class small-molecule drugs exceeded that of target-based approaches [in recent years] (page 1, paraphrasing Swinney & Anthony (2011))

Premise

- The contribution of phenotypic screening to the discovery of first-in-class small-molecule drugs exceeded that of target-based approaches [in recent years] (page 1, paraphrasing Swinney & Anthony (2011))

- While a complex painting assay on an optimised cell line represents a compelling approach [...] simple image-based assays on several cell lines may be more relevant biologically. (page 2)

Data and prediction task

- 9 (distinct) pleural mesothilioma + 1 prostate cancer cell lines

Data and prediction task

- 9 (distinct) pleural mesothilioma + 1 prostate cancer cell lines
- Prestwick dataset (1200 compounds), of which 614 have an MOA (DrugBank)

Data and prediction task

- 9 (distinct) pleural mesothilioma + 1 prostate cancer cell lines
- Prestwick dataset (1200 compounds), of which 614 have an MOA (DrugBank)
- Predict MOA (from 113 target classes) in leave-one-out cross validation scheme from Ljosa et al. (2013)

Methodology

Following a classical pipeline:

- Average per-cell features over well to create *phenotypic profile* for each drug k:

$$profile_k = \left[\frac{1}{N} \sum_{n=1}^{N} x_{n1}^{k}, \ldots, \frac{1}{N} \sum_{n=1}^{N} x_{iD}^{k} \right]$$

where x_{nd}^{k} is feature d of D for cell n of N for the kth drug.

Methodology

Following a classical pipeline:

- Average per-cell features over well to create *phenotypic profile* for each drug k:

$$profile_k = \left[\frac{1}{N} \sum_{n=1}^{N} x_{n1}^k, \ldots, \frac{1}{N} \sum_{n=1}^{N} x_{iD}^k \right]$$

  where x_{nd}^k is feature d of D for cell n of N for the kth drug.

- Train one random forest (30 trees) per cell line (10 forests)

Methodology

Following a classical pipeline:

- Average per-cell features over well to create *phenotypic profile* for each drug k:

$$profile_k = \left[\frac{1}{N} \sum_{n=1}^{N} x_{n1}^k, \ldots, \frac{1}{N} \sum_{n=1}^{N} x_{iD}^k \right]$$

  where x_{nd}^k is feature d of D for cell n of N for the kth drug.

- Train one random forest (30 trees) per cell line (10 forests)

- Predict MOA based on combining probabilities (ensemble prediction)

Results

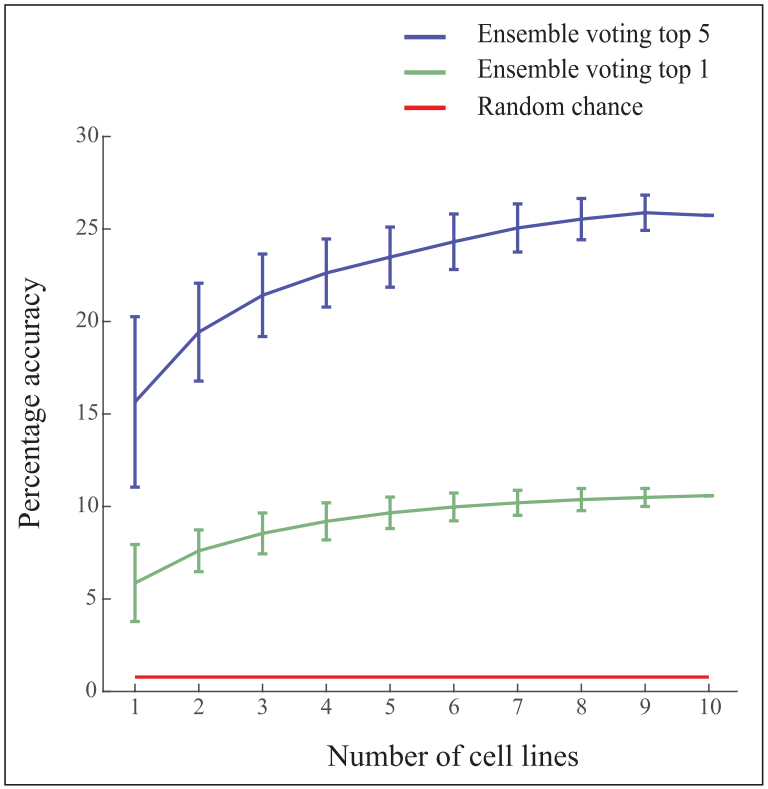


Figure: Top-5 and top-1 classification accuracy improves with the addition of cell lines

nature biotechnology

Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines

Jungseog Kang[1,2,6], Chien-Hsiang Hsu[1,3,4,6], Qi Wu[1], Shanshan Liu[1], Adam D Coster[1], Bruce A Posner[5], Steven J Altschuler[1,4] & Lani F Wu[1,4]

High-content, image-based screens enable the identification of compounds that induce cellular responses similar to those of known drugs but through different chemical structures or targets. A central challenge in designing phenotypic screens is choosing suitable imaging biomarkers. Here we present a method for systematically identifying optimal reporter cell lines for annotating compound libraries (ORACLs), whose phenotypic profiles most accurately classify a training set of known drugs. We generate a library of fluorescently tagged reporter cell lines, and let analytical criteria determine which among them—the ORACL—best classifies compounds into multiple, diverse drug classes. We demonstrate that an ORACL can functionally annotate large compound libraries across diverse drug classes in a single-pass screen and confirm high prediction accuracy by means of orthogonal, secondary validation assays. Our approach will increase the efficiency, scale and accuracy of phenotypic screens by maximizing their discriminatory power.

Opening remarks

- Molecular screens expensive for large-scale drug screening

Opening remarks

- Molecular screens expensive for large-scale drug screening
- Image-based high-content screening (HCS) a promising alternative

Opening remarks

- Molecular screens expensive for large-scale drug screening
- Image-based high-content screening (HCS) a promising alternative
- Main objective of *identifying lead compounds in a single-pass screen* (page 1) while respecting scalability

A three-step process

In a *live-cell assay*, an optimal cell line is chosen by brute force to be used in a large-scale screen:

A three-step process

In a *live-cell assay*, an optimal cell line is chosen by brute force to be used in a large-scale screen:

1. Construct library of reporter cell lines

A three-step process

In a *live-cell assay*, an optimal cell line is chosen by brute force to be used in a large-scale screen:

1. Construct library of reporter cell lines
2. Identify optimal reporter cell line for compound libraries (ORACL)

A three-step process

In a *live-cell assay*, an optimal cell line is chosen by brute force to be used in a large-scale screen:

1. Construct library of reporter cell lines
2. Identify optimal reporter cell line for compound libraries (ORACL)
3. Use ORACL to identify lead compounds in a single-pass screen

Step 1: library of reporter cell lines

A549 non-small cell lung cancer *parent* cell line. 93 reporter clones were produced by:

- randomly labeling proteins with yellow fluorescent protein as an extra exon via viral transfection (*central dogma tagging*).

Step 1: library of reporter cell lines

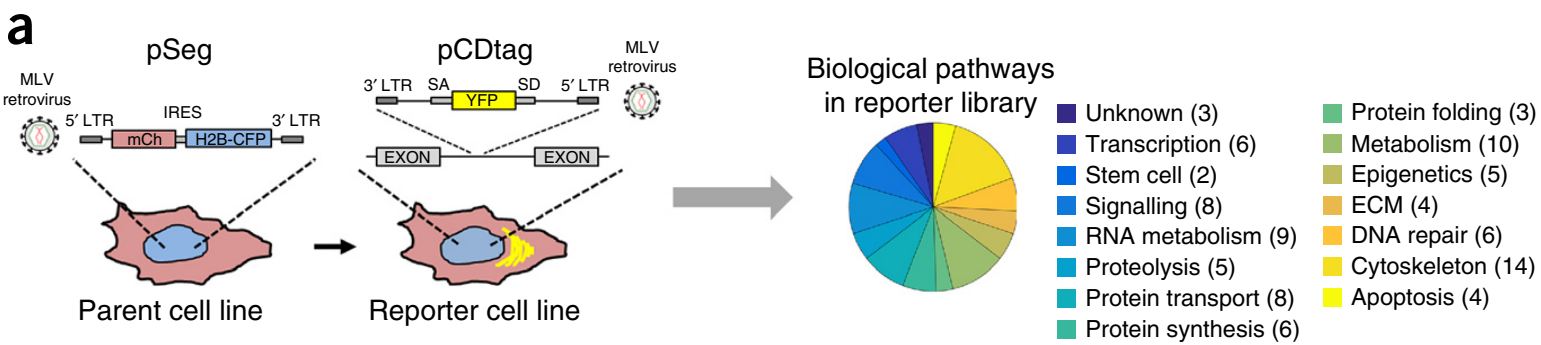
A549 non-small cell lung cancer *parent* cell line. 93 reporter clones were produced by:

- randomly labeling proteins with yellow fluorescent protein as an extra exon via viral transfection (*central dogma tagging*).
- The genes can then be identified using 3'RACE (duplicates discarded)

Step 1: library of reporter cell lines

A549 non-small cell lung cancer *parent* cell line. 93 reporter clones were produced by:

- randomly labeling proteins with yellow fluorescent protein as an extra exon via viral transfection (*central dogma tagging*).
- The genes can then be identified using 3'RACE (duplicates discarded)



Step 2: identification of the ORACL

- The 93 reporter cell lines were treated by a small 6 classes $\times$ 5 drugs + DMSO *reference set*

Step 2: identification of the ORACL

- The 93 reporter cell lines were treated by a small 6 classes $\times$ 5 drugs + DMSO *reference set*
- Phenotypic profile created as in Perlman et al. (2004):

$$profile_k = \left[KS(X^k_{:,1}, X^-_{:,1}), \ldots, KS(X^k_{:,D}, X^-_{:,D}) \right]$$

where $X^k_{:,d}$ and $X^-_{:,d}$ is feature d of D of all cells for drug k and the negative control respectively.

Step 2: identification of the ORACL

- The 93 reporter cell lines were treated by a small 6 classes $\times$ 5 drugs + DMSO *reference set*
- Phenotypic profile created as in Perlman et al. (2004):

$$profile_k = \left[KS(X^k_{:,1}, X^-_{:,1}), \ldots, KS(X^k_{:,D}, X^-_{:,D}) \right]$$

where $X^k_{:,d}$ and $X^-_{:,d}$ is feature d of D of all cells for drug k and the negative control respectively.

- One drug was removed at random from each class and assigned a class as the nearest centroid of the remaining four drugs per class

Step 2: identification of the ORACL

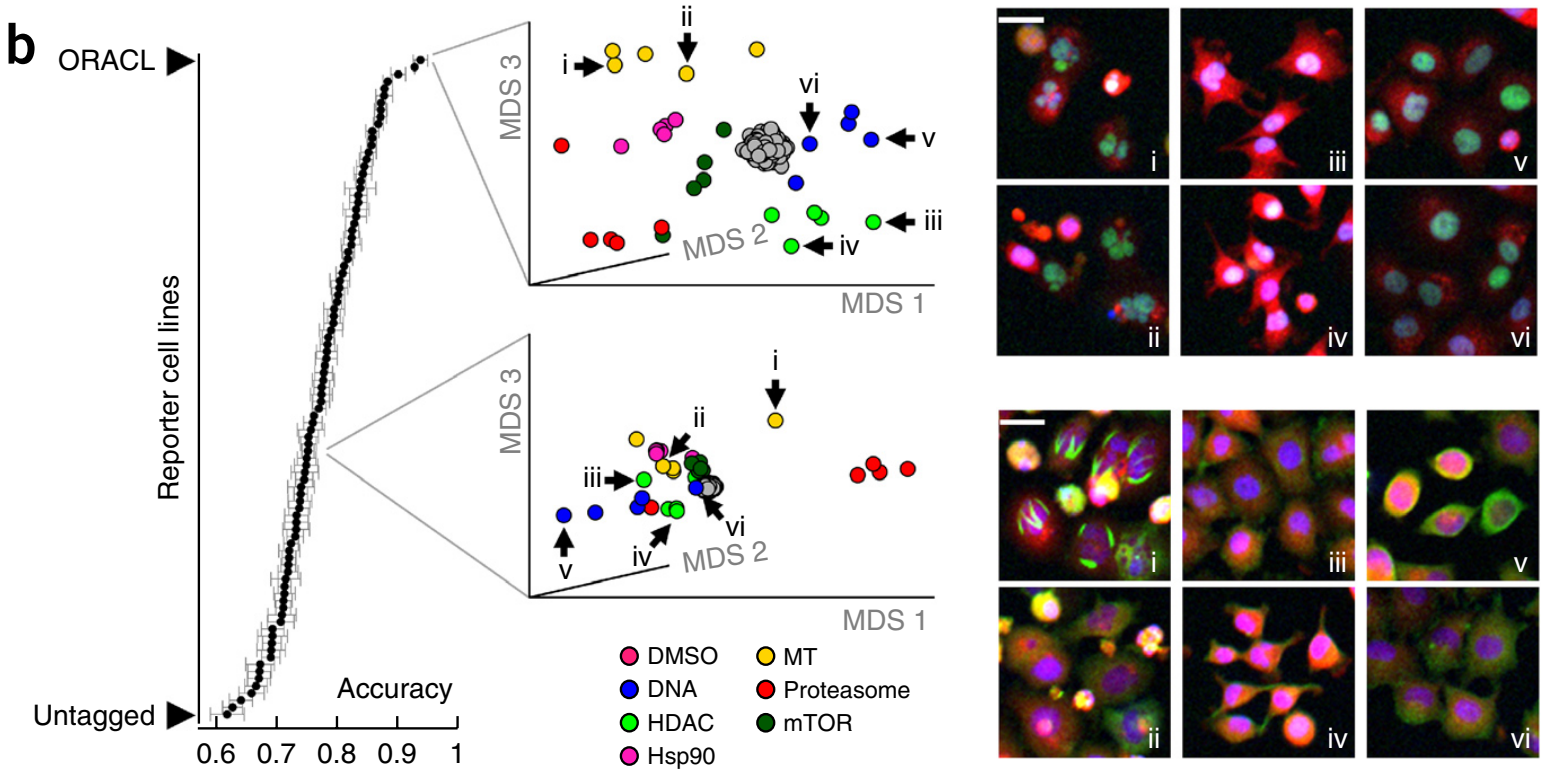


Figure: The reporter cell line maximising prediction accuracy (94% over 100 repetitions was chosen as the ORACL (XRCC5 DSB-repair protein)

Step 3: identification of multi-classes hits

- The selected ORACL is then tested on a large suite of drug libraries (10483 drugs total)

Step 3: identification of multi-classes hits

- The selected ORACL is then tested on a large suite of drug libraries (10483 drugs total)
- This yields $\sim$ 62000 live-cell images after 24h and 48h and $\sim$ 60 million cells with $\sim$ 230 morphological features apiece

Step 3: identification of multi-classes hits

- The selected ORACL is then tested on a large suite of drug libraries (10483 drugs total)
- This yields $\sim$ 62000 live-cell images after 24h and 48h and $\sim$ 60 million cells with $\sim$ 230 morphological features apiece
- Linear discriminant analysis to reduce dimensionality, in which reference drug clusters are used as nearest neighbour classifier

Step 3: identification of multi-classes hits

- The selected ORACL is then tested on a large suite of drug libraries (10483 drugs total)
- This yields $\sim$ 62000 live-cell images after 24h and 48h and $\sim$ 60 million cells with $\sim$ 230 morphological features apiece
- Linear discriminant analysis to reduce dimensionality, in which reference drug clusters are used as nearest neighbour classifier
- Confidence scores (based on distance, calibrated using reference set) calculated for each classification. Compounds left unclassified when confidence low ($p < 0.1$)

Step 3: identification of multi-class hits

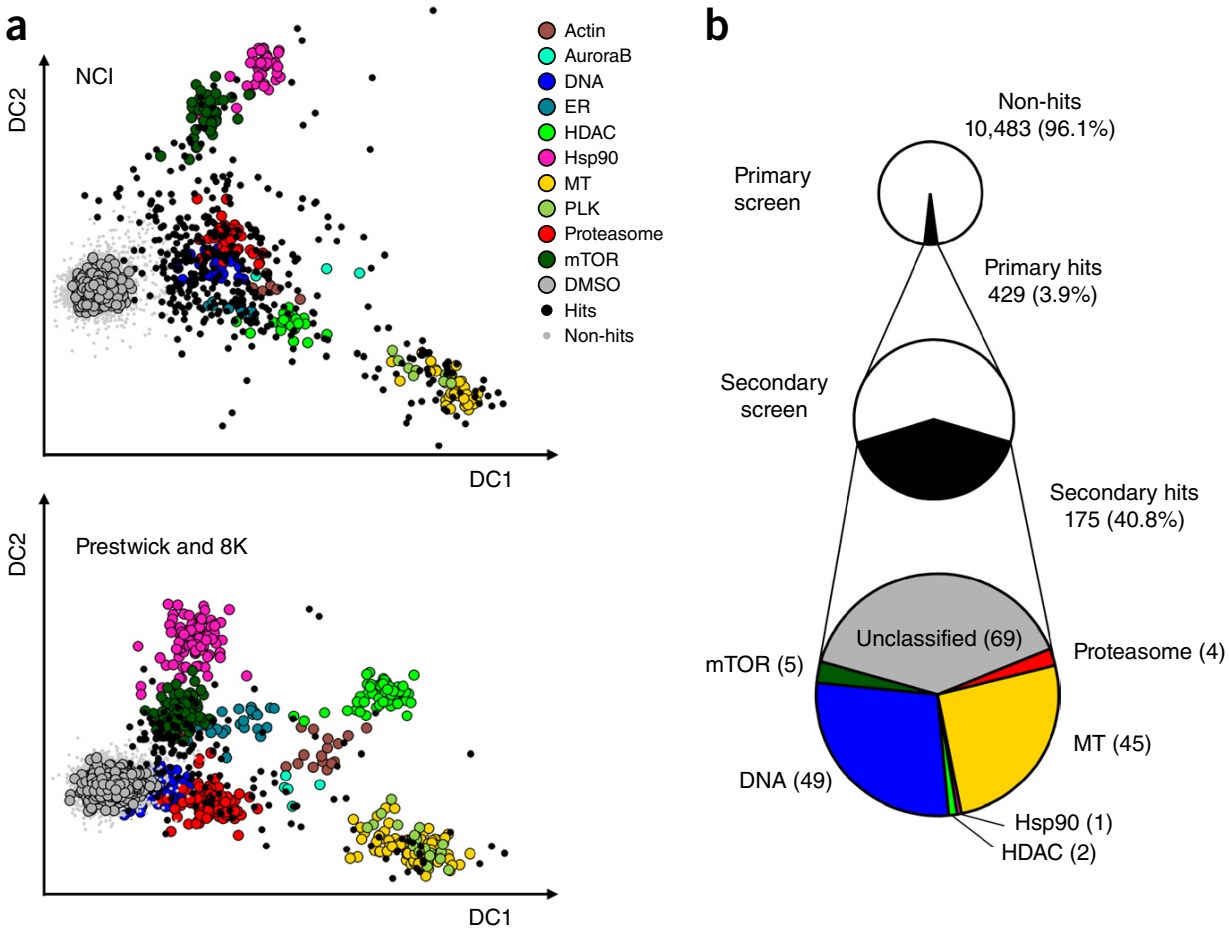


Figure: *Primary hits* (429 total) defined as non-DSMO classified (whether to a known class or not)

Experimental validation

- A secondary screen on the primary hits was performed to eliminate weak phenotypes (precision over recall) leaving 175 *secondary hits*

Experimental validation

- A secondary screen on the primary hits was performed to eliminate weak phenotypes (precision over recall) leaving 175 *secondary hits*
- They are then able to validate HDAC directly through literature (gratifyingly)

Experimental validation

- A secondary screen on the primary hits was performed to eliminate weak phenotypes (precision over recall) leaving 175 *secondary hits*
- They are then able to validate HDAC directly through literature (gratifyingly)
- The others are validated through a series of secondary experiments