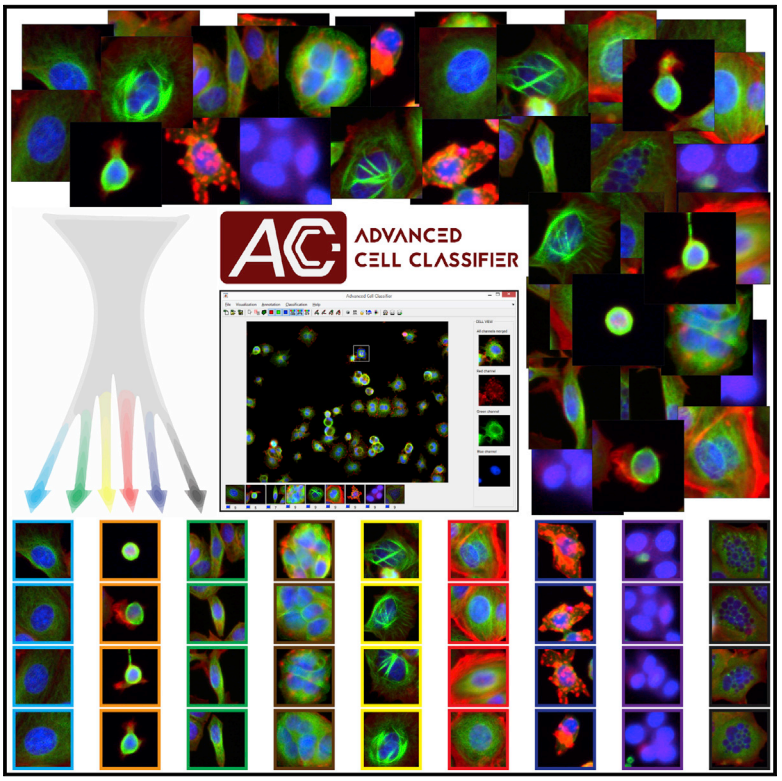
Tool

Cell Systems

Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data

Graphical Abstract



Authors

Filippo Piccinini, Tamas Balassa, Abel Szkalisity, ..., Ulrike Kutay, Kevin Smith, Peter Horvath

Correspondence

horvath.peter@brc.mta.hu

In Brief

We have developed new algorithms to mine microscopic image-based screening data, discover new phenotypes, and improve recognition performance. These methods are implemented into a user-friendly, free open-source tool that improves phenotype classification accuracy and helps users to quickly uncover hidden phenotypes from large datasets.

Highlights

- A powerful new suite of algorithms for phenotype discovery in image-based screens
- A visual tool for exploring and annotating phenotypes within large datasets
- New methods to find similar looking cells and suggest useful annotations
- Faster and more complete discovery of phenotypes, more accurate classification

Piccinini et al., 2017, Cell Systems *4*, 651–655
June 28, 2017
http://dx.doi.org/10.1016/j.cels.2017.05.012

CellPress

Cell Systems
Tool

Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data

Filippo Piccinini,[1,8] Tamas Balassa,[2,8] Abel Szkalisity,[2] Csaba Molnar,[2] Lassi Paavolainen,[3] Kaisa Kujala,[3] Krisztina Buzas,[2,4] Marie Sarazova,[5] Vilja Pietiainen,[3] Ulrike Kutay,[5] Kevin Smith,[6,7] and Peter Horvath[2,3,9,*]
[1]Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) S.r.l., IRCCS, Via Piero Maroncelli 40, 47014 Meldola (FC), Italy
[2]Synthetic and Systems Biology Unit, Hungarian Academy of Sciences, Biological Research Center (BRC), Temesvári körút 62, 6726 Szeged, Hungary
[3]Institute for Molecular Medicine Finland, University of Helsinki, Tukholmankatu 8, 00014 Helsinki, Finland
[4]University of Szeged, Faculty of Dentistry, Tisza Lajos körút 64, 6720 Szeged, Hungary
[5]Insitute of Biochemistry, ETH Zurich, Otto-Stern-Weg 3, 8093 Zurich, Switzerland
[6]KTH Royal Institute of Technology, School of Computer Science and Communication, Lindstedtsvägen 3, 10044 Stockholm, Sweden
[7]Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden
[8]These authors contributed equally
[9]Lead Contact
*Correspondence: horvath.peter@brc.mta.hu
http://dx.doi.org/10.1016/j.cels.2017.05.012

SUMMARY

High-content, imaging-based screens now routinely generate data on a scale that precludes manual verification and interrogation. Software applying machine learning has become an essential tool to automate analysis, but these methods require annotated examples to learn from. Efficiently exploring large datasets to find relevant examples remains a challenging bottleneck. Here, we present Advanced Cell Classifier (ACC), a graphical software package for phenotypic analysis that addresses these difficulties. ACC applies machine-learning and image-analysis methods to high-content data generated by large-scale, cell-based experiments. It features methods to mine microscopic image data, discover new phenotypes, and improve recognition performance. We demonstrate that these features substantially expedite the training process, successfully uncover rare phenotypes, and improve the accuracy of the analysis. ACC is extensively documented, designed to be user-friendly for researchers without machine-learning expertise, and distributed as a free open-source tool at www.cellclassifier.org.

INTRODUCTION

In the past, limits in automation, processing, and storage technology imposed practical restrictions on the number of images we could analyze. Typical research projects were limited to a few dozen to a few hundred. On that scale, it was possible to verify entire experiments manually. Today, high-content screening (HCS) experiments easily produce over 10,000 times more data (Neumann et al., 2010). It is no longer feasible to visually interpret and verify every data point. As we depend more and more on automated computational methods for complex and large-scale image analysis, we run the risk of only partially understanding the data. We must ask ourselves "Have I understood the data?" and "Is my analysis as accurate as possible?" Without the right analysis tools, our answers to these questions may be unsatisfactory.

In this paper, we introduce Advanced Cell Classifier (ACC), a machine-learning software designed to give a quicker and more complete understanding of large datasets and to train predictive models as accurately as possible. In 2011, we released ACC version 1.0 (ACC v1.0), a graphical image analysis software tool that offers access to a variety of machine-learning methods and provides accurate analysis (Horvath et al., 2011). Several large-scale cell-based phenotypic HCS studies have made use of ACC v1.0, including at least 15 human genome-wide RNAi screens and numerous extensive drug screens. These studies cover a wide variety of biological topics ranging from the studies of influenza A virus (Banerjee et al., 2014) to studies of acute lymphoblastic leukemia (Fischer et al., 2015).

ACC v1.0 shared a drawback with other similar machine-learning HCS analysis software (Orlov et al., 2008; Uhlmann et al., 2016; Held et al., 2010; Sommer et al., 2011; Laksameethanasan et al., 2013; Ogier and Dorval, 2012; Rämö et al., 2009; Misselwitz et al., 2010; Jones et al., 2008; Dao et al., 2016); it lacked discovery and data visualization tools to enable the user to fully explore and understand their data. We view this as a crucial shortcoming. Little attention has been given to the methods used to explore the data, understand it efficiently, or to ensure the quality of the annotations. The cost of collecting expert annotations is high and categorizing cells into strict classes is often ambiguous. Experts are often unsure if they have uncovered all the important phenotypes buried within the data because they lack the tools to fully explore it. There are also limitations in the annotation process. Existing software packages force the user to manually select cells to label or randomly select

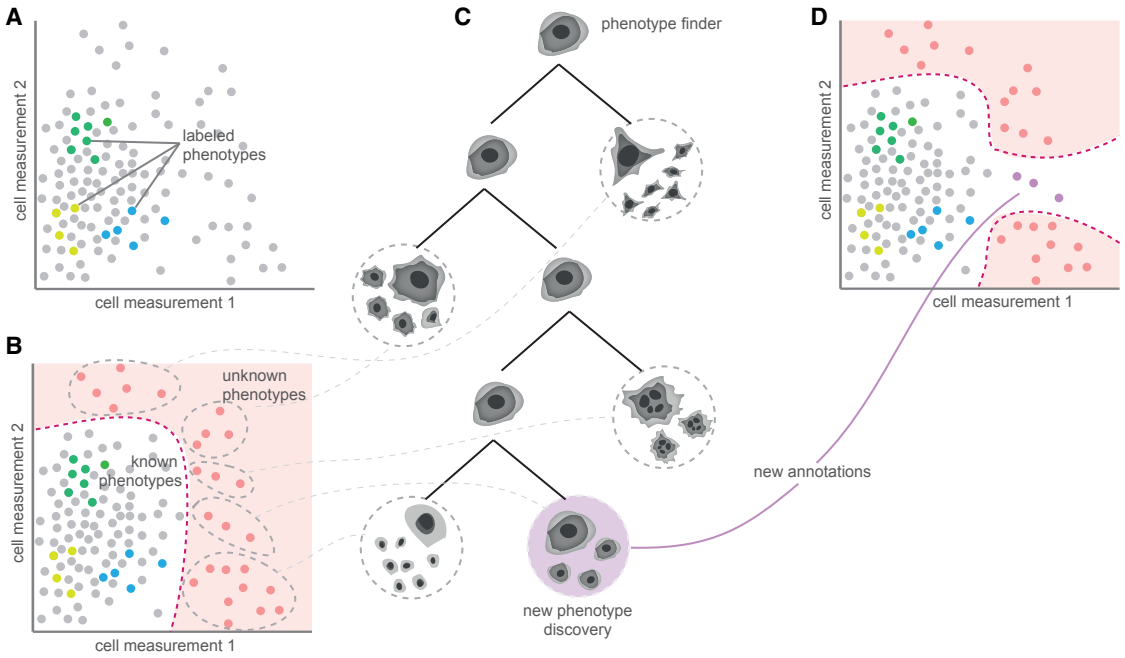


Figure 1. Phenotype Finder Tool

The phenotype finder tool organizes cells into a browsable hierarchy, which facilitates the discovery of new classes.

(A) Cells are represented by dots embedded in a two-dimensional synthetic feature space. Sets of cells annotated by the expert are shown in green, yellow, and blue. Unannotated cells are shown in gray.

(B) A one-class classifier is used to automatically determine which cells are least similar to the known cell types (pink region).

(C) Cells with the least similarity to known examples are sampled and clustered to construct a dendrogram. The expert can browse and analyze the representative cells shown in the tree, create a new phenotype class from these cells, or add cells to an existing one. In addition, irrelevant cells can be discarded to improve the classification results to be obtained in the next run.

(D) If a new phenotype is discovered, the region of non-annotated cells changes in the multidimensional feature space.

cells. These methods are inefficient and generate many wasteful annotations that ultimately do not prove useful for the classifier. Furthermore, these procedures make it difficult to discover new phenotypes or to intelligently refine decision boundaries (Smith and Horvath, 2014).

ACC v2.0 is a completely re-designed and user-friendly software tool with the goal of improving the collection and understanding of image data and the accuracy of the analysis (Figure S1). It allows researchers, even those without computer vision or machine-learning knowledge, to efficiently characterize and exploit their cell-based and image-based HCS experiments, leading to new discoveries. The main differences between ACC v2.0 and its previous versions include: (1) intelligent methods to explore and annotate large single-cell image data, including an active learning approach to improve the accuracy of the classifier, and similar cell search, an algorithm to find similar cells and increase the number of annotations for rare phenotypes; (2) an easy-to-use report generator to automatically obtain statistics on cell distribution and class incidence; (3) a new, re-designed and user-friendly interface; (4) detailed documentation, video tutorials, and online resources; and (5) improved data visualization methods. Finally, and most importantly, (6) we have implemented phenotype finder, a novel method to automatically discover new and biologically relevant cell phenotypes (Figure 1). The source code of ACC v2.0 is freely distributed as an open-source tool at www.cellclassifier.org.

RESULTS

To evaluate the effectiveness of the discovery and annotation tools included in ACC v2.0, we generated a synthetic dataset simulating images of cells from a high-content screen. This provided us with completely accurate knowledge concerning every cell and phenotype in every image, something that is impossible with images of actual cells. Using this information, we compared ACC v1.0 with ACC v2.0. A group of annotators were able to reach much higher recognition accuracy and found even extremely rare phenotypes in significantly less time using the tools we developed. A description of the synthetic dataset is given in the STAR Methods section along with the analysis, which is summarized in Figure S2.

We also applied ACC v2.0 to a drug screen and a small interfering (siRNA) screen to identify phenotypic classes using real data. In each case, the annotators were able to discover relevant phenotypic classes, including rare types, within a short time using our new methods. Details of the drug and the siRNA screens are provided in the STAR Methods section. Example images from the phenotypic classes identified using ACC v2.0 are shown in Figures 2 and S3.

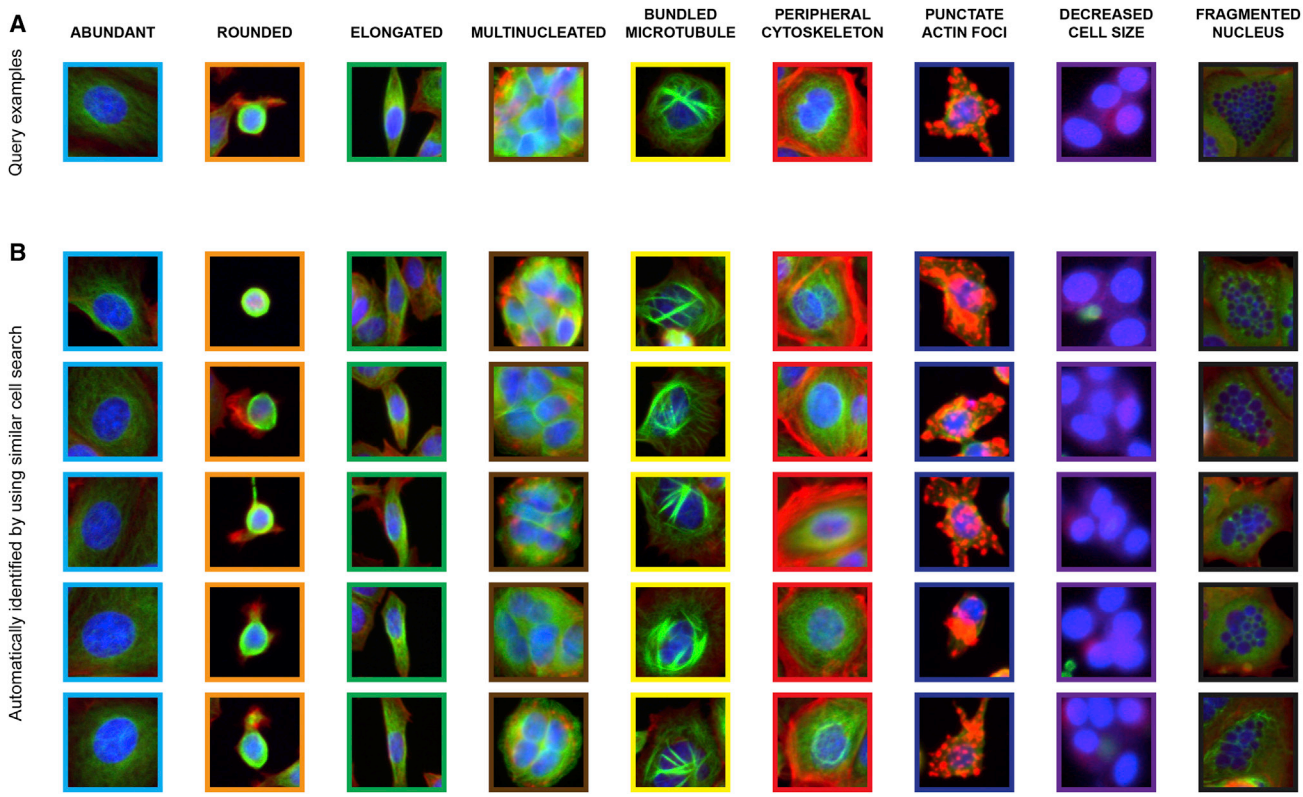


Figure 2. Efficient Example-Based Mining of Relevant Cell Phenotypes

(A) Cells from nine relevant cell phenotypes identified using the phenotype finder on data from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012). Searching through thousands of images for more examples of a rare phenotype can be cumbersome, but the similar cell search function makes it simple.

(B) Results of the similar cell search given the query examples above. Similar cells were determined by computing the cosine similarity on image feature vectors between the query example and other cells in the dataset.

DISCUSSION

ACC v2.0 provides several new innovative tools designed to explore and collect the data necessary to train classifiers more efficiently and effectively. The ability of the classifier to correctly recognize cell types ultimately depend on the quality of data provided. While there is no universally accepted recipe for generating quality training data, many principles and techniques can be applied in practice to improve the efficiency and quality of the annotation process.

Perhaps the most fundamental principle is to ensure that the training data are complete in the sense that they includes examples of all the important phenotypes present in the screen. Although it may seem obvious, practically speaking, this can be tedious when the amount of data is very large. Another common issue is imbalance between classes. Often, interesting phenotypes are in the minority or occur very infrequently. If the data are imbalanced due to the presence of a rare class, the lack of representative data will make learning difficult (He and Garcia, 2009). Given a single example of a rare cell, ACC v2.0 can quickly identify additional, previously unidentified examples using the similar cell search feature, thereby helping balance the dataset and improve classification performance. Another way to improve data collection is to avoid redundant annotations and to prioritize annotations that are most useful for boosting classification performance. ACC v2.0 uses active learning to carefully select the most informative examples for labeling, which avoids irrelevant examples and refines regions where the classifier is uncertain.

Data quality can also be improved through iterative refinement of the classifier. During data collection, ACC v2.0 can train a classifier on existing annotated data and display predicted annotations on unlabeled data. By correcting erroneous predictions, the user adds valuable data points to the training set, which can help correct predictive errors.

In total, ACC v2.0 includes powerful new methods to mine microscopic image data, discover new phenotypes, and improve recognition performance. While no single method can be regarded as a silver bullet that solves all annotation problems, in our experience, the most effective strategy is to alternate between discovery tools as the biological task demands. ACC v2.0 gives the user access to a large variety of state-of-the-art machine-learning algorithms, has an intuitive user interface with advanced visualization, and allows for efficient navigation of image data. It is easy to use, well documented, and comes with helpful video tutorials. Using synthetic data and existing screens, we demonstrated that the discovery tools in ACC v2.0 improve the quality of training datasets and ultimately create classifiers with better phenotype recognition. Using our software, it is possible to discover interesting cell phenotypes hidden in large datasets.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Synthetic Dataset
  - Drug HCS Dataset
  - siRNA HCS Dataset
- METHOD DETAILS
  - Overview of the Software
  - Phenotype Finder
  - Similar Cell Search
  - Active Learning Methods to Prioritize Useful Annotations
  - Manual Correction of Predictions
  - Annotation of Manually Selected Cells
  - Annotation of Randomly Selected Cells
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures, one table, four files, two data files, and five movies and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2017.05.012.

AUTHOR CONTRIBUTIONS

P.H. conceived and led the project. F.P. wrote the documentation, designed the GUI, prepared the figures, and co-supervised the project. U.K. and M.S. provided one of the HCS datasets and the knowledge behind it. P.H., K.S., T.B., V.P., and K.B. designed the experiments and analyzed the data. T.B. was the lead developer of the implementation and the phenotype finder module. A.S. implemented the active learning module. L.P and K.K. implemented the find similar cells module. K.S. implemented the suggest a classifier module. C.M. implemented the image visualization module and tested the code. F.P., K.S., T.B., K.B., V.P., and P.H. wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors wish to thank Beverley Isherwood, Neil Carragher, and Anne E. Carpenter for the information they provided about the HCS image dataset; ETH Zurich Microscopy Center (ScopeM) for HCS imaging; Arpad Balind for suggestions and feedback; Gabriella Tick and Máté Görbe for their help with the documentation; Dóra Bokor for proofreading the manuscript. F.P. was supported by a Short-Term Fellowship awarded by the Federation of European Biochemical Societies (FEBS). F.P., B.T., CS.M., A.SZ., and P.H. acknowledge support from the Hungarian National Brain Research Program (MTA-SE-NAP B-BIOMAG). U.K. received support from the SNF (313003A_166565). L.P., V.P., and P.H. acknowledge support from the Finnish TEKES FiDiPro Fellow Grant 40294/13 and European Union and the European Regional Development Funds (GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037).

Received: October 19, 2016
Revised: January 17, 2017
Accepted: May 19, 2017
Published: June 21, 2017

REFERENCES

Badertscher, L., Wild, T., Montellese, C., Alexander, L.T., Bammert, L., Sarazova, M., Stebler, M., Csucs, G., Mayer, T.U., Zamboni, N., et al. (2015). Genome-wide RNAi screening identifies protein modules required for 40S subunit synthesis in human cells. Cell Rep. *13*, 2879–2891.

Banerjee, I., Miyake, Y., Nobs, S.P., Schneider, C., Horvath, P., Kopf, M., Matthias, P., Helenius, A., and Yamauchi, Y. (2014). Influenza A virus uses the aggresome processing machinery for host cell entry. Science *346*, 473–477.

Caie, P.D., Walls, R.E., Ingleston-Orme, A., Daya, S., Houslay, T., Eagle, R., Roberts, M.E., and Carragher, N.O. (2010). High-content phenotypic profiling of drug response signatures across distinct cancer cells. Mol. Cancer Ther. *9*, 1913–1926.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Cooper, J.A. (1987). Effects of cytochalasin and phalloidin on actin. J. Cell Biol. *105*, 1473–1478.

Dao, D., Fraser, A.N., Hung, J., Ljosa, V., Singh, S., and Carpenter, A.E. (2016). CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. Bioinformatics *32*, 3210–3212.

Fischer, U., Forster, M., Rinaldi, A., Risch, T., Sungalee, S., Warnatz, H.J., Bornhauser, B., Gombert, M., Kratsch, C., Stütz, A.M., et al. (2015). Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. Nat. Genet. *47*, 1020–1029.

Fujikawa-Yamamoto, K., Teraoka, K., Zong, Z.P., Yamagishi, H., and Odashima, S. (1994). Apoptosis by demecolcine in V79 cells. Cell Struct. Funct. *19*, 391–396.

He, H., and Garcia, E.A. (2009). Learning from imbalanced data. IEEE Trans. Knowledge Data Eng. *21*, 1263–1284.

Held, M., Schmitz, M.H., Fischer, B., Walter, T., Neumann, B., Olma, M.H., Peter, M., Ellenberg, J., and Gerlich, D.W. (2010). CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. Nat. Methods *7*, 747–754.

Horvath, P., Wild, T., Kutay, U., and Csucs, G. (2011). Machine learning improves the precision and robustness of high-content screens using nonlinear multiparametric methods to analyze screening results. J. Biomol. Screen. *16*, 1059–1067.

Jones, T.R., Kang, I.H., Wheeler, D.B., Lindquist, R.A., Papallo, A., Sabatini, D.M., Golland, P., and Carpenter, A.E. (2008). CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC Bioinformatics *9*, 482.

Jordan, M.A., and Wilson, L. (2004). Microtubules as a target for anticancer drugs. Nat. Rev. Cancer *4*, 253–265.

Kavallaris, M. (2010). Microtubules and resistance to tubulin-binding agents. Nat. Rev. Cancer *10*, 194–204.

Laksameethanasan, D., Tan, R.Z., Toh, G.W., and Loo, L.H. (2013). cellXpress: a fast and user-friendly software platform for profiling cellular phenotypes. BMC Bioinformatics *14* (*Suppl 16*), S4.

Laplante, M., and Sabatini, D.M. (2009). mTOR signaling at a glance. J. Cell Sci. *122*, 3589–3594.

Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H., and Yli-Harja, O. (2007). Computational framework for simulating fluorescence microscope images with cell populations. IEEE Trans. Med. Imaging *26*, 1010–1016.

Liu, L., Chen, L., Chung, J., and Huang, S. (2008). Rapamycin inhibits F-actin reorganization and phosphorylation of focal adhesion proteins. Oncogene *27*, 4998–5010.

Ljosa, V., Sokolnicki, K.L., and Carpenter, A.E. (2012). Annotated high-throughput microscopy image sets for validation. Nat. Methods *9*, 637.

Ljosa, V., Caie, P.D., Horst, R.T., Sokolnicki, K.L., Jenkins, E.L., Daya, S., Roberts, M.E., Jones, T.R., Singh, S., Genovesio, A., et al. (2013). Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J. Biomol. Screen. *18*, 1321–1329.

MacLean-Fletcher, S., and Pollard, T.D. (1980). Mechanism of action of cytochalasin B on actin. Cell *20*, 329–341.

Misselwitz, B., Strittmatter, G., Periaswamy, B., Schlumberger, M.C., Rout, S., Horvath, P., Kozak, K., and Hardt, W.D. (2010). Enhanced CellClassifier: a multi-class classification tool for microscopy images. BMC Bioinformatics *11*, 30.

Mortensen, K., and Larsson, L. (2003). Effects of cytochalasin D on the actin cytoskeleton: association of neoformed actin aggregates with proteins involved in signaling and endocytosis. Cell. Mol. Life Sci. *60*, 1007–1012.

Moya, M.M., Koch, M.W., and Hostetler, L.D. (1993). One-class Classifier Networks for Target Recognition Applications (No. SAND-93-0084C) (Sandia National Labs), Technical Report.

Mühlradt, P.F., and Sasse, F. (1997). Epothilone B stabilizes microtubuli of macrophages like taxol without showing taxol-like endotoxin activity. Cancer Res. *57*, 3344–3346.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Ogier, A., and Dorval, T. (2012). HCS-Analyzer: open source software for high-content screening data correction and analysis. Bioinformatics *28*, 1945–1946.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., and Goldberg, I.G. (2008). WND-CHARM: multi-purpose image classification using compound image transforms. Pattern Recognit. Lett. *29*, 1684–1693.

Rämö, P., Sacher, R., Snijder, B., Begemann, B., and Pelkmans, L. (2009). CellClassifier: supervised learning of cellular phenotypes. Bioinformatics *25*, 3028–3030.

Schölkopf, B., Smola, A.J., Williamson, R.C., and Bartlett, P.L. (2000). New support vector algorithms. Neural Comput. *12*, 1207–1245.

Smith, K., and Horvath, P. (2014). Active learning strategies for phenotypic profiling of high-content screens. J. Biomol. Screen. *19*, 685–695.

Sommer, C., Straehle, C., Köthe, U., and Hamprecht, F.A. (2011). Ilastik: Interactive Learning and Segmentation Toolkit. Proceeding of the 2011 IEEE International Symposium on Biomedical Imaging: from Nano to Macro (ISBI 2011) (IEEE), pp. 230–233.

Uhlmann, V., Singh, S., and Carpenter, A.E. (2016). CP-CHARM: segmentation-free image classification made accessible. BMC Bioinformatics *17*, 51.

Witten, I., and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Advanced Cell Classifier version 2.0	Current manuscript	www.cellclassifier.org/download/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents or computational resources may be directed to, and will be fulfilled by, the Lead Contact Peter Horvath (horvath.peter@brc.mta.hu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Synthetic Dataset

To evaluate the effectiveness of the discovery and annotation tools included in ACC v2.0 (Figure S1), we designed and generated a synthetic dataset simulating images of cells from a high content screen. This provided us with completely accurate knowledge concerning every cell and phenotype in every image, something that is impossible with images of actual cells. Using this information, we compared ACC v1.0 to ACC v2.0 and evaluated how the tools we introduced improved the ability of annotators to discover phenotypes and train accurate classifiers.

The dataset was organized into 32 plates containing 384 images each. Every image contained human cells simulated using SIMCEP, a software tool for simulating fluorescence microscopy images of cell populations (Lehmussola et al., 2007). The amount of cells in each image was sampled uniformly between 5 and 40. In total, 12,288 images and 310,929 cells were generated. While this is a relatively small dataset by HCS standards, it provides sufficient statistical power to judge the efficacy of our tools. Eight distinct phenotypic classes were designed by varying the size and shape of the nucleus and cytoplasm, and by varying the number and size of subcellular vesicles (Figure S2A). Each cell is assigned a phenotype sampled from these classes, ranging in frequency from very common (appearing with 49.7% probability), to less common (24.8%), to extremely rare (0.01%). Furthermore, each image was generated with a dominant phenotype (approximately 80% of the cells in the image), and random cell classes made up the remainder of the phenotypes in the image. The synthetic images were processed with a CellProfiler pipeline to segment the cells and extract features. The synthetic data and pipeline are available in Data S1.

Three experts used ACC v1.0 and ACC v2.0 to annotate the dataset. They were given no prior information about the number of phenotypes or their locations. In each trial, the expert was given 30 minutes to annotate the synthetic data. The goals were (*1*) to discover all the phenotypes within the given time and (*2*) to create a high quality dataset which, when used to train a classifier, results in the highest accuracy on the whole dataset. The results of the experiment appear in Figure S2. All three experts were able to discover all eight phenotypes within the time limit using ACC v2.0 (Figure S2B). Using ACC v1.0, the three rarest phenotypes were only discovered by 2/3 of the experts. The experts were able to find phenotypes faster using ACC v2.0. Using the annotations collected by the experts, classifiers were trained and used to predict phenotypes for the entire dataset. The phenotypes predicted by trained classifiers were compared to the true phenotypes and normalized prediction accuracy was computed. The results show a substantial improvement (a 27% increase in recognition) with annotations collected using ACC v2.0 (Figure S2C). The last pane of the figure shows timelines of how many annotations were collected by each annotator using each version of the software (Figure S2D). The bold line shows the mean number of collected annotations.

Drug HCS Dataset

To demonstrate the capabilities of ACC v2.0 on real data, we analysed a dataset of MCF-7 breast cancer cells (BBBC021v1 (Caie et al., 2010), available from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012): *https://www.broadinstitute.org/bbbc/BBBC021/*) treated with a collection of 113 small molecules in 8 different concentrations for 24 hours. A subset of this dataset is formatted for ACC and supplied as a test dataset with the software (*www.cellclassifier.org/download/*). The molecule set consists of a mechanistically distinct set of targeted and cancer-relevant cytotoxic compounds inducing a broad range of gross and subtle phenotypes. Cells were fixed, labelled for DNA, F-actin, and β-tubulin, and imaged by fluorescent microscopy acquiring multiple images of all data points and technical repeats (Caie et al., 2010). This resulted in 39,600 images containing approximately 2,000,000 cells.

The images were segmented and features were extracted with CellProfiler 2.2.0 (CellProfiler pipeline provided as File S1). To demonstrate the phenotype discovering capabilities of ACC v2.0, we started by manually annotating a few cells from a single class (standard abundant cells from the most common phenotype). We then used the phenotype finder tool to identify interesting cell

phenotypes. Representative cells from each phenotypic class appear in Figure 2A. Names of the phenotypic classes were selected by using Cellular Microscopy Phenotype Ontology (CMPO; *http://www.ebi.ac.uk/cmpo/*), which provides a species-neutral controlled vocabulary for cellular phenotypes. To increase the number of annotations for rare classes such as multinucleated cells, we used the 'find similar cells' tool (Figure 2B). To improve the classifier's ability to distinguish between classes, we used the active learning module to refine decision boundaries. After approximately one hour we obtained around 1,500 annotated cells from 9 phenotype classes.

Finally, we assessed the predictive capability of the classifiers trained using the data we collected. We applied several different popular supervised classification methods and chose the one with best 10-fold cross validation performance. The Logistic Boost classifier achieved 88.4%, the highest recognition accuracy (Figure S4), and was applied to the entire screen. The final results of the analysis showing the number of cells in every phenotypic group for the different drug treatments can be found in Data S2.

We evaluated whether the identified phenotypes correlated to any known effects of the used drugs (Table S1). While most of the identified phenotypes were determined to be non-drug specific, we highlight two phenotypes showing strongly correlated characteristics to specific well-known drug effects: (*a*) multinucleated; and (*b*) bundled microtubule, cells with collapsed microtubule. The Cytochalasin B treatment leads to the classification of 20% of cells in the multinucleated group. Cytochalasin B is known to inhibit both the rate of actin polymerization and the interaction of the actin filaments in solution (MacLean-Fletcher and Pollard, 1980), thus preventing the formation of contractile microfilaments. This can result in a disturbed cell cycle, yielding an increased number of nuclei with variable shape and size within the cell (Cooper, 1987). The bundled microtubules group contains samples of a characteristic phenotype with microtubular bundles crossing through the centre of the cell. Most of the cells in this phenotypic class were treated with Taxol (Paclitaxel), Docetaxel, or Epothilone B. All of these agents bind the β-tubulin, and stabilize microtubules. They inhibit the microtubule function and alter their dynamics, as well as enhance the polymerization of tubulin, whereby they have antimitotic effect (Mühlradt and Sasse, 1997). In this case, the software identified a characteristic phenotype linked to different agents with the same mechanism of action.

We also identified phenotypes that correlate with well-known physiological cell status: (*1*) increased amount of punctate actin foci; and (*2*) fragmented nucleus, often the hallmark of apoptosis. The cells identified in the punctate actin class received Cytochalasin D as top hit. Cytochalasin D is an actin depolymerizing drug and can also induce the actin aggregation (Mortensen and Larsson, 2003). Interestingly, some of the cells treated with Rapamycin correlated strongly with this phenotype, while most of the Rapamycin–treated cells were classified to abundant or elongated cells classes. Rapamycin targets the mTOR (a mammalian target of Rapamycin-complex) with an essential role in the cell cycle and responses to changing nutrient levels. Inhibition of mTOR signalling by Rapamycin leads to defects in mitochondrial function, cell proliferation, cytoskeletal organization, protein synthesis, and can result in cell death in many ways: apoptosis, necrosis or autophagy (Laplante and Sabatini, 2009). Thanks to the new tools in ACC v2.0, we were also able to find the cells in a rarer actin-related phenotype of Rapamycin, reported to affect the F-actin reorganization by blocking the kinase activity of mTOR (Liu et al., 2008). Demecolcine (Colcemid) and Vincristine treatment resulted in an overrepresented number of cells of the phenotype fragmented nucleus. By inhibiting the polymerization of microtubules, they can arrest the cells in the metaphase, which can ultimately lead to apoptosis (Fujikawa-Yamamoto et al., 1994; Jordan and Wilson, 2004). These drugs are known to bind the tubulin and destabilize the microtubules, in contrast to the stabilizing agents, such as Docetaxel and Paclitaxel described above with a strong bundled microtubules phenotype (Kavallaris, 2010). This demonstrates the power of the ACC v2.0 to correctly categorize the drugs with opposite biological functions although they would have a potentially similar end-point outcome (cell death). As these results show, ACC v2.0 is an effective tool to find the novel, unknown biological effects of drugs or silencing/overexpression of certain genes, as well as for mining previously described phenotypes of interest.

siRNA HCS Dataset

We analysed images derived from a pilot siRNAs screen targeting selected hit factors as well as a suite of positive and negative controls for a genome-wide screen on 60S ribosomal subunit biogenesis in HeLa cells. The assay relies on an inducible RPL29-GFP reporter as a read-out, similar to our recent genome-wide analysis of 40S synthesis using RPS2-YFP (Badertscher et al., 2015). In brief, cells were transfected with the respective siRNAs by reverse transfection in 384 well plates. Reporter construct expression was induced by addition of tetracycline after 44 hours. 8 hours later, the culture medium was replaced by medium lacking tetracycline and cells were incubated for another 20 hours. Then, cells were fixed, DNA stained with Hoechst, and images taken by fluorescent microscopy acquiring 9 images per well.

Images were segmented based on Hoechst staining of cell nuclei, and 150 diverse features were extracted, including nuclear and cytoplasmic fluorescence intensities. We started by manually annotating a few cells from a single class of a mock-treated well. We then used the phenotype finder tool to identify interesting cell phenotypes. Representative cells from each phenotypic class appear in Figure S3A. To increase the number of annotations for rare classes such as large nuclei, we used the find similar cells tool (Figure S3B). We acquired ~500 annotated cells from 7 phenotype classes.

Finally, we assessed the predictive capability of the classifiers trained using the data we collected. We applied several different popular supervised classification methods and chose the one with best 10-fold cross validation performance. The Artificial Neural Network classifier achieved 95%, the highest recognition accuracy (Figure S5), and was applied to the entire screen.

METHOD DETAILS

Overview of the Software

ACC v2.0 is a software tool to apply sophisticated machine learning analysis to microscopic image data. Particular attention has been given to user-friendliness and tools to help non-experts explore their data, understand it, and train machine learning algorithms to be as accurate as possible. Figure S1A shows the main interface which consists of a menu bar, a toolbar, the main window, and cell view windows focusing on details of the currently highlighted cell. A separate image selector window allows the user to quickly switch between images and plates from the experiment (S1B). ACC v2.0 works in Windows, Linux, and Macintosh environments, and takes images as input with support for most common image formats (*e.g.* tif, bmp, png). In addition to the original image data, ACC v2.0 requires features, the image measurements extracted from the segmented objects (in txt, csv, or HDF5 format), and it supports contours of segmented sub-cellular/cellular objects (*e.g.* cells membranes and nuclei), such as those produced by CellProfiler (Carpenter et al., 2006).

Detailed documentation (File S2 and File S3) provides step-by-step instructions to use the software, and a series of video tutorials with how-to examples are also available (Movies S1–S5). A detailed flowchart diagram of the usage and services is presented on Figure S6. Typically, analysis begins with the user starting a project (''Getting started'', Movie S1) and importing data (''ACC and CellProfiler'', Movie S2, and ''Input data structure'', Movie S3). We provide a CellProfiler module designed to export necessary segmentation and feature data directly to ACC v2.0 (File S4). The next step is to begin exploring the data and assigning annotations to cells as members of a phenotypic class. Instructions on how to define classes, customize the visualization, and navigate the data are provided in Movie S1. Several innovative annotation strategies are available in ACC v2.0 which greatly improve the efficiency and quality of data annotation. These include the phenotype finder (Figure S1C, Figure 1), similar cell search, active learning, as well as manual and random selection. Examples on how to apply these strategies are shown in Movie S4. Once the annotated data is collected, up to sixteen different classifiers may be chosen to predict phenotypes for unannotated cells (Figure S1E) including well-known methods such as support vector machine (SVM), multilayer perceptron (MLP), and random forest using the Weka framework (Witten and Frank, 2005). Initial predictions may contain errors, but these can be corrected by replacing the incorrect prediction with the correct annotation and re-training the classifier. This process can be repeated until satisfactory results are achieved. Finally, the improved classifier is applied to the entire experiment and user selects the desired formats of the output including (*a*) cell-by-cell classification; (*b*) incidence and distribution of the different classes; (*c*) phenotype-based statistics of any selected features (Movie S5).

Phenotype Finder

Screening datasets often consist of tens of thousands of images, or orders of magnitude more. The amount of data is often so substantial that it exceeds the capabilities of a human expert to observe everything. Therefore, it is difficult for the annotator to know whether the training data he/she collected contains examples representing all the important phenotypes present in the screen. To address this, ACC v2.0 includes a phenotype finder tool which helps find and define new phenotypes efficiently, without requiring a priori knowledge about the underlying dataset (Figure 1). It does this by hierarchically grouping cells based on their appearance. Images of the cells are organized into a browsable tree-like structure (Figure 1C) which allows the user to quickly identify previously unseen phenotypes or subpopulations within a known phenotype (Movie S4). The phenotype finder operates on the assumption that true phenotypes will cluster together in a space of relevant image features. Using a bottom-up approach, a dendrogram can be constructed by first defining each observation (cell) as its own cluster, and by making pairwise similarity comparisons between each cluster. The dendrogram is constructed by greedily merging the most similar clusters first, and proceeding up the hierarchy until the entire set belongs to a single cluster (Figure S1D). In ACC v2.0, the similarity metric is defined as the Euclidean distance between the mean of the feature vectors associated with cells belonging to each cluster.

The computational complexity and memory consumption associated with hierarchical clustering means it cannot be directly applied to very large data sets (in practice, 25,000-50,000 cells can be clustered on a standard computer with 16 GB memory in 15-80 s). Because of this limitation, unless the dataset is small, the phenotype finder must be applied to subsets of the data - cells belonging to a known phenotype or cells that are hypothesized to belong to an undiscovered phenotype. In the first case, existing annotations are used to train a classifier that predicts which cells belong to a selected known phenotype. These are sampled and clustered using hierarchical clustering. In the second case, a one-class classifier is used to identify cells that do not fall within any currently known phenotype. One-class classification, also known as unary classification (Moya et al., 1993), is a method that can perform this task by estimating the support of a high-dimensional distribution given a set of positive samples (Figures 1A and 1B). A one-class SVM (Schölkopf et al., 2000) predicts the probability that a given cell belongs to any of the known phenotypes in this manner. Cells are sorted based on this probability, and the cells least likely to belong to an existing phenotype are clustered using hierarchical clustering (Figure 1C). The one-class SVM can be sensitive to the features it is provided. Oftentimes, HCS data contains some redundant features or features with little discriminative power. Feature selection methods can help by reducing the feature vector to the most informative features. We have tested three methods on synthetic and real HCS data: information gain, principal component analysis (PCA), and factor analysis. We found that information gain proves to be the most useful method, and the optimal number of selected features is in the range of 10-20 (Figure S7).

Once a subset of cells has been selected and the dendrogram has been constructed, it is displayed as a collapsible tree interface which allows the user to view the member cells for each cluster in the hierarchy and to quickly create new annotations from a cluster in

the dendrogram (Figure 1C). If the user determines that the cluster shows signs of novelty it can be annotated as a new class, otherwise it can be inserted into an existing phenotype. If a new phenotype is discovered, the class boundaries will change (Figure 1D). Iteratively applying the phenotype finder in this manner will allow the user to quickly identify rare phenotypes or subpopulations within an existing phenotype.

Similar Cell Search

Often, important or interesting phenotypes occur very infrequently. While the phenotype finder can help initially discover the phenotype, without sufficient examples of the phenotype to train with, the classifier may struggle to identify it reliably. ACC v2.0 offers a solution to this issue. Given a single example of a rare cell, it can quickly identify new examples with similar appearance to the selected cell. This is accomplished by computing the cosine similarity on a reduced feature vector between the query cell and all other cells in the screen (Ljosa et al., 2013). The feature set is reduced in a pre-processing step to remove highly correlated features. The cells are sorted according to their similarity score, and the most similar cells are presented to the user (Figures 2B and S3B). The user may then select these cells and add them to the corresponding classes. In this manner, rare and important phenotypes can be quickly populated with reliable annotations.

Active Learning Methods to Prioritize Useful Annotations

Expert annotations are costly to acquire, but typically no guidance is provided as to which cells are useful to annotate. The user is often left to select cells manually, or is presented with a random sampling of cells. But there is no guarantee on the usefulness of the labels provided with these strategies. It can be advantageous to avoid redundant annotations and prioritize annotations that will be most useful for increasing classification performance. ACC v2.0 uses active learning to carefully select the cell which, if annotated, will be most useful in improving the classification accuracy. It avoids uninformative examples and refines regions where the classifier is uncertain. Using the currently annotated set (Figure S8A), a classifier is trained and makes predictions on unlabeled data and uses a query strategy to estimate which example will be most useful to improve classification performance (Figure S8B). We use two popular query strategies: uncertainty sampling and query by committee (Smith and Horvath, 2014). Uncertainty sampling selects the instance it is least certain how to label based on the classifier prediction probabilities. Query by committee selects the instance with the most disagreement between different classification algorithms. The expert annotates the requested cell, and the annotation is added to the training set. A new predictive model is trained and a new cell is selected for annotation according to the query strategy (Figure S8C). This process is repeated and iteratively improves classification performance. Potentially, new phenotypes can be discovered by exploring the boundary between classes (Figures S8D–S8E).

Manual Correction of Predictions

Another practical method of improving the quality of training data is to iteratively refine classifier performance by correcting errors in its predictions. ACC v2.0 can train a classifier on existing annotated data and visualize predicted annotations on unlabeled data. The user can visually inspect these predictions and add new annotations by confirming correct labels or by correcting erroneous predictions. By adding corrective annotations where the classifier made a mistake, the user adds valuable data points to the training set to improve performance.

Annotation of Manually Selected Cells

This annotation method is standard in most cell classification software. The expert is free to navigate the data and select which cells he/she wishes to annotate. Manual exploration has its merits, as the expert may often have a good intuition about which cells are useful to annotate early in process. However, there are several dangers to relying solely on this method. The user may be biased towards easy or redundant examples, may not balance the number of examples between classes, and may fail to discover rare phenotypes. When performing manual annotation, it is recommended that cells are selected from several different images. We also note that is it always possible to override other annotation modes and switch to a manual annotation if the user notices something interesting in the image.

Annotation of Randomly Selected Cells

To ensure a representative sampling of the data, it can be advantageous to randomly sample from the screen. This is a forced choice mode: the expert is required to annotate the phenotype class of a randomly selected cell from a randomly chosen image. This annotation mode thereby avoids selection bias by the user (Misselwitz et al., 2010). However, with this method and with manual annotation, luck may be required to discover rare phenotypes in large datasets.

DATA AND SOFTWARE AVAILABILITY

In this work we used the HCS image set BBBC021v1 (Caie et al., 2010), available from the Broad Bioimage Benchmark Collection (Ljosa et al., 2012). It is composed of 39,600 image files (13,200 fields of view imaged in three channels) in TIFF format. It can be downloaded at: *https://www.broadinstitute.org/bbbc/BBBC021/*. For any copyright issues regarding the dataset, follow the instructions provided at the referenced website.

ACC v2.0 is freely distributed at *www.cellclassifier.org* as an open-source tool. It is written in MATLAB R2015a and requires the Image Processing Toolbox 9.2. The source code, standalone versions, video tutorials, help documentation files, and additional modules are available at the official repository for all future software releases *www.cellclassifier.org*. Further developing the source code requires Matlab license. However, Windows, Linux, and Macintosh standalone compiled versions— that do not require license— are also available online. All the ACC materials are copyright protected and distributed under GNU General Public License version 3 (GPLv3).