Active Learning for Drug Discovery

Joshua Kangas

CMU-CB-13-100

February 2013

Pittsburgh, Pennsylvania

Committee: Robert F. Murphy (Chair)
Gustavo K. Rohde
Jeffrey Schneider
D. Lansing Taylor, University of Pittsburgh

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Carnegie Mellon University

Copyright © 2013 Joshua Kangas

This research was supported in part by National Institutes of Health grant R01 GM075205 and by a generous gift from Ray and Stephanie Lane in support of graduate education in computational biology. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: drug discovery, active learning, machine learning, computational biology, high-throughput screening, high-content screening, protoplasts, bioimage informatics

To my sweet wife and dear children.

Abstract

The use of high throughput screening methods has aided the drug discovery process allowing for the testing of numerous compounds for effects on a single target protein. However, by focusing primarily on a single target during high throughput screening, undesirable secondary effects are often detected late in the development process after substantial investment has been made. In order to better detect effects on a system, high content screening methods have been developed utilizing imaging technology in conjunction with machine learning methods to detect effects on living systems as a result of exposure to a drug stimulus. These have primarily been applied in animal systems, and we therefore explored approaches to extending high content screening methods to plant cells. A pilot high content screening approach was developed and used to test the effects of nine compounds on protoplasts from six lines of *Arabidopsis thaliana* expressing different fluorescently-tagged proteins. Various image analysis and machine learning techniques were used to determine which compounds affected the subcellular distributions of the proteins.

Both high throughput and high content screening methods are primarily limited in that very few target proteins are measured directly in these experiments. An alternative approach would be to do a more global screen against many undesired effects early in the process, but the number of possible secondary targets makes this prohibitively expensive due to the number of combinations of potential drugs and secondary targets. Methods for making this global approach feasible through active machine learning were therefore developed. The active learning approach iteratively constructs models to predict the results of unobserved experiments and utilizes these models to guide experimentation efforts. Such methods were developed and applied to screening data for 20,000 compounds on 177 assays. It was shown through simulations that nearly 60% of all hits (compounds that have an effect on a particular assay) could be identified after exploring only 3% of the experimental space. Finally, an automated approach to creating NIH 3T3 cell lines expressing fluorescently-tagged proteins via CD-tagging and identifying the tagged protein was developed. This was used to create a set of lines used to test active learning for detection of compound effects on the location patterns of the tagged proteins.

Our results suggest that active learning can be used to enable more complete characterization of compound effects across a diverse set of assays than otherwise affordable. The methods described are also likely to find widespread application in biomedical research.

Acknowledgements

I would like to express my gratitude to the members of my thesis committee for their involvement in these research efforts: Drs. Robert F. Murphy, Gustavo K. Rohde, Jeff Schneider and D. Lansing Taylor. I greatly appreciate the excellent research direction and guidance offered by the members of the committee. I would also like to thank Dr. Robert F. Murphy for the interesting and challenging research opportunities he provided. As a result of his guidance, I was able to work on projects that encompassed both the computational and experimental opportunities I hoped to experience during my research time. I am also grateful for Armaghan Naik who has been an excellent collaborator and friend during these years of research. I would also like to acknowledge Dr. Luis Coelho whose python packages PySLIC (image processing), Mahotas (image processing) and Milk (machine learning) were utilized in some way in all of these chapters.

I would like to express my appreciation to Ray and Stephanie Lane, whose gift to establish the Lane Center in part supported my work. My work was also supported in part by National Institutes of Health grant GM075205.

For the development of Chapter 2, Drs. Alexander (Sasha) Dovzhenko and Klaus Palme gave advice in guidance for the experimental processes to be used as well as access to laboratory resources at the University of Freiburg. Karsten Voigt and Rüdiger Trojok assisted me with the preparation of protoplasts for experimentation. I would also like to thank Dr. Anke Becker for the use of the liquid handling robot and Dr. Roland Nitschke for the use and guidance in the use of the automated microscope and computational facilities.

For the development of Chapter 3, I am thankful for the efforts of Armaghan Naik in the development of the active learning method tested in this chapter.

The results in Chapter 4 are an extension of a project developed for an interesting and informative course on active learning taught by Drs. Robert F. Murphy, Jeff Schneider and Jaime Carbonell. Additionally, I want to thank Dr. D. Lansing Taylor for his advice and guidance in this effort.

Chapter 5 was built upon the RandTag project. This work could not have been completed without the efforts of Armaghan Naik. While I worked on this project, I had useful interactions with Drs. Elvira Highley (neé Garcia-Osuna), Jon Jarvik and Peter Berget. Dr. Margaret Furhman executed most of the infections used during the RandTag project and her experience with tissue culture was an invaluable resource as the project developed. Jennifer Bakal was instrumental in preparing the RandTag sequencing results for publication. The efforts from Ling-Yi Kung, Stephanie Luck, Vilcya Wirantana and Jessika Louissaint are greatly appreciated also greatly appreciated.

I would like to thank past and current members of the Murphy lab for their input and friendship during this research endeavor: Dr. Dechang Xu, Dr. Aabid Shariff, Dr. Tao Peng, Dr. Luis Pedro Coelho, Dr. Arvind Rao, Dr. Estelle Glory-Afshar, Taráz E. Buck, Armaghan Naik, Devin Sullivan, Gregory Johnson, Aparna Kumar, Ivan Cao-Berg and Jennifer Bakal.

The faculty and students of the Combined Carnegie Mellon-University of Pittsburgh PhD program in Computational Biology have provided an excellent resource in terms of learning experiences in classes and program activities.

My undergraduate research advisors at Truman State Univsersity, Drs Jon Beck, Diane Janick-Buckner and Brent Buckner, gave excellent direction, advice and support allowing me to do research which directly led to this opportunity at Carnegie Mellon University.

Finally, I would like to thank my wife and family for their support. The feedback from my sister, Linda Hutchinson, father, Dr. Donald A. Kangas and mother, Jeanne C. Kangas, was invaluable during the final editing of this thesis. I would also like to thank my wife, Hilary, and children, Ryan and Dylan, for their patience and sacrifice while I worked toward accomplishing this goal.

Table of Contents

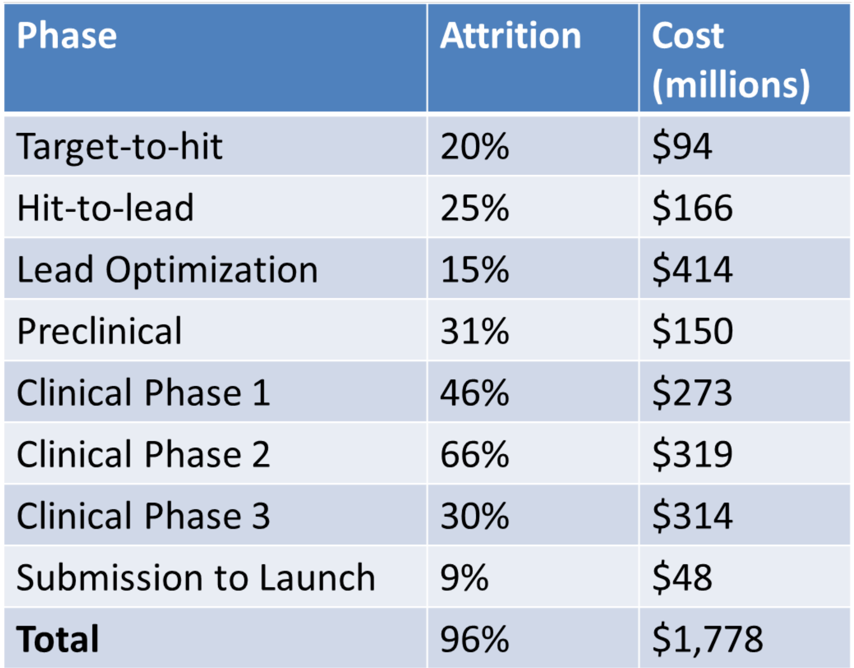
List of Figures

List of Tables

Chapter 1: Introduction

Drug development is a lengthy, risky and expensive process initialized by the identification of compounds which affect disease-associated targets and ending after testing in clinical trials. Current estimates for the costs of the process are shown in Table 1. The process is initialized by the identification of druggable targets. In general, these targets are proteins that are integral parts of mechanisms hypothesized to be involved with disease. Much initial information about these targets is gathered through basic science studies. Once a target protein has been identified, the goal is to discover compounds which increase or decrease the function of a target protein. One common method for testing large numbers of compounds for effects on a single target is referred to as high throughput screening.

Table 1: The attrition rate and costs for each phase of the drug discovery process are shown (Paul, et al., 2010).



Phase	Attrition	Cost (millions)
Target-to-hit	20%	$94
Hit-to-lead	25%	$166
Lead Optimization	15%	$414
Preclinical	31%	$150
Clinical Phase 1	46%	$273
Clinical Phase 2	66%	$319
Clinical Phase 3	30%	$314
Submission to Launch	9%	$48
Total	96%	$1,778

High-throughput screening (HTS) is a process by which compounds are tested for effects on protein function resulting from compound exposure. The first step in a HTS process is to develop an assay to be used to detect the effects of compounds on the target protein. In general, the goal for HTS assays is to detect the effects of many drugs on a single protein by running multiple experiments rapidly. There are multiple types of experimental assays which can be used for HTS. These screens are commonly performed by adding drugs to a protein in a microtiter plate or to a well of live cells. In either case, fluorescent readouts are commonly used. A plate reader can be used to measure absorbance, fluorescence or luminescence of an experimental well (Inglese, et al., 2007). Flow cytometry can be used to measure changes in cellular populations that result from the addition of drugs (Sklar, et al., 2007). Fluorescence microscopy can be used to gather more information in a screen utilizing live cells (Trask, et al., 2009) in a process referred to as high-content screening. The development of liquid handling robots has made it possible to test millions of compounds in a single pass for a single experimental configuration.

Imaging is often used in high throughput screening. High-content screening (HCS) is used to describe experiments which require the combination of image analysis and machine vision methods to characterize large amounts of biological image data. HCS experiments are typically executed to describe some physical characteristic of the cells. Machine vision methods can be used in the context of HCS experiments to describe various physical characteristics of a biological system such as protein location patterns (Boland & Murphy, 2001), object motion (Meijering, et al., 2006) and gene expression levels (Peng & Myers, 2004). Images can be captured at many different scales depending on the goals of the experiment. These experiments can vary greatly in time scale depending on the organism being studied. Images as small as single cells or as large as whole organisms including zebrafish (Pardo-Martin, et al., 2010) and mice (Brommage, et al., 2008) can be used in high-content screening.

In the context of drug discovery, HCS can be used to capture phenotypic changes which result from the addition of compounds. Often these experiments are designed to detect perturbations in phenotypes resulting from the addition of potential drugs (Perlman, et al., 2004). For the analysis of cell-based imaging assays, multiple methods have been used to detect the desired characteristics. Generally these involve calculating some sort of feature describing the images or subsets of the image and then running a machine learning algorithm on the resulting measurements. Many methods have been developed for the analysis of HCS results; these include (but are not limited to): field and cell level feature calculation, segmentation (Duda & Hart, 1972), patch-based methods, interest point detection, and SURF (Bay, et al., 2008). For some experiments, specific custom image analysis techniques for the experimental condition have been developed. For example, image analysis for the study of neurons poses a complex problem that has been partially addressed using image tracing algorithms. Once features have been calculated, a machine learning method is typically used. There are three major classes of machine learning algorithms commonly used in high-content screening processes and other image analysis problems. When phenotypes are known for certain experimental conditions, supervised classification techniques (Breiman, 2001; Burges, 1998) can be used to detect effects from drug treatment. When some information is known about experimental conditions, such as the fact that duplicated conditions should have identical phenotypes, and positive controls are available, semi-supervised methods (Zhang, et al., 2008) can be used. When no information is known about the "correct" phenotypes for experimental conditions or information may be unreliable such as in large scale location proteomics projects (Garcia-Osuna, et al., 2007; Barbe, et al., 2008), unsupervised methods (MacQueen, 1967) can be utilized.

Once a compound has been identified as having an effect on a target, the compound may be advanced further into the drug development pipeline. It is not uncommon, however, for previously unknown effects to be discovered after significant investment into a potential drug (**Figure** 1). These are not discovered earlier because these processes are primarily designed to detect the effects of compounds on a single target protein without considering whether or not it has effects on other targets. One solution to this problem is to test every small molecule compound (~10^6) against every possible protein target (~10^4). However, an estimated 10^{10} experiments would be required to explore this experimental space completely, the cost of which would likely be prohibitive. A cellular systems biology approach to this problem would be to develop many assays to capture information about cell and tissue behaviors and exhaustively test compounds on those assays (Giuliano, et al., 2009) as shown in **Figure** 2. By considering many effects of each potential drug earlier in the drug discovery process, the attrition rates in early phases would likely increase. By rejecting the truly undesirable compounds earlier in the process, the attrition rates in the later more costly phases of the drug discovery process would decrease reducing downstream research costs.

Given the prohibitively large number of experiments needed to exhaustively explore these experimental spaces, only a subset of experiments could actually be selected and executed. The results of the remaining untested experiments can then only be predicted. There are two clear approaches to this problem: experiments can be selected that are predicted to have the highest activity based on some model or experiments can be selected that are predicted to give the most information about the experimental space. The latter allows for more accurate predictions on the remainder of the experimental space.

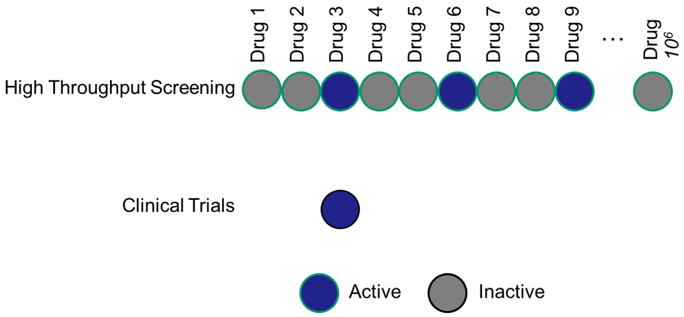


Figure 1: High-throughput screening can be used to test for the effects of a large number of compounds on a single protein target. Often compounds are allowed to progress through the drug discovery process (ex. Drug 3) only to have deleterious off-target effects discovered during a later phase after significant costly effort to develop the drug.

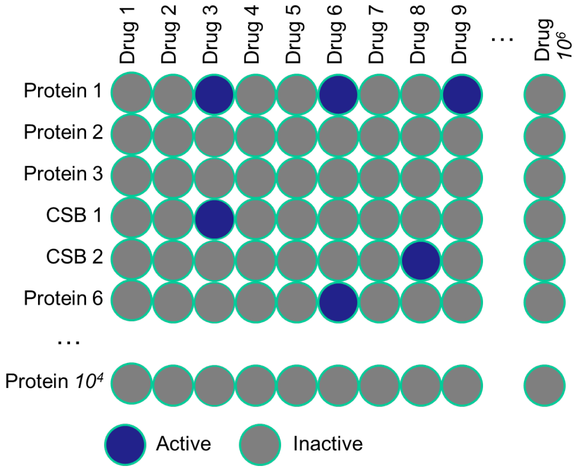


Figure 2: To avoid allowing potential drugs which might have off-target effects to progress through the drug discovery pipeline, one approach would be to test every potential drug against every protein and against a large number of cellular systems biology assays.

Active learning is a machine learning method specifically developed to permit efficient exploration of such large experimental spaces. Active learning consists of three phases performed in a loop. A thread of experiments can be initialized either using prior results from literature or databases or by selecting a batch of random experiments from an experimental space. **(1)** A model is generated to represent the currently available data. **(2)** From that model, experiments which are expected to most improve the model are selected for execution. **(3)** The set of experiments is executed and the data gathered is agglomerated with previously collected experimental data. The loop then continues from Step 1 until either a desired accuracy of predictions is achieved or a specified budget has been exhausted.

The overall goal of an active learning process is to efficiently learn a predictive model or, as in the case of drug discovery, to find all possible hits. Through simulations, one can compare different methods first by selecting an initial set of experiments and then using different selection methods to select the following batches of experiments to be executed until some budget is exhausted or a desired accuracy is reached. If one has a fixed experimental budget, the resulting accuracies from each selection method can be compared. The difference between these two results after the budget has been exhausted is the improvement from active learning to select experiments. Alternatively, if one desires to reach a specific accuracy, a successful active learning method would reach that accuracy before the model learned using the alternative selection method. These are illustrated in Figure 3.

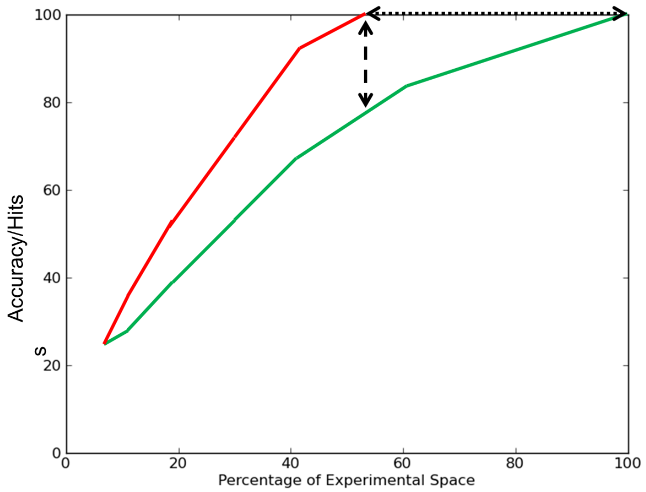


Figure 3: The resulting accuracy or hits discovered as a function of percentage of experimental space explored is shown for an example active learning method (green) and an alternative selection method such as random selection (red). With a fixed budget, the improvement from using active learning is shown by the dashed line. When perfect accuracy is desired, the experimental savings using active learning is shown by the dotted line.

The overall goal of an active learning process is to improve a predictive model or to optimize an output. These predictions could be of the discrete form in which one is trying to predict the discrete class of an observation or one is trying to predict the value of an observation in a continuous fashion. There are various formulations of this problem. Sometimes, knowledge of a set of features describing each experimental unit is available ahead of time. In the context of drug discovery, an example of this sort of problem could be the availability of a set of features describing each compound that could be used in an experiment. Sometimes, there may not be information available to describe the experiments to be executed. This might be the case when information is unavailable about the compounds to be tested. The final measure for an experiment may also take different forms. The result could be a discrete class such as a phenotype discovered using high-content screening. Alternatively, the result may be continuous, such as the measured activity from a high-throughput screening experiment. A single experiment may also result in numerous measurements.

Another consideration for the application of active learning is the selection of the appropriate batch size. If a single experiment is chosen each round, this is called myopic selection. Most active learning work using batch selection has used a fixed batch size. There has been some work showing a reduction in the total number of batches needed to reach a specific accuracy by selecting dynamically-sized batches (Chakraborty, et al., 2010). In the case of HTS and HCS there are per-batch and per-experiment costs which need to be considered when selecting a batch size. For example, it is likely for a set of HTS experiments, the costs of doing 10 experiments is comparable to the costs of doing 96 or 384 experiments because the per-experiment costs are relatively low compared to the per-batch costs. In proactive learning, experimental costs and reliability of labeling are considered to optimally select experiments (Yang & Carbonell, 2009).

Application of active learning to biological problems has been limited (Liu, 2004; Mohamed, et al., 2010; Stegle, et al., 2009). In the area of drug discovery, very few applications of active learning have been published. In these efforts, compound activity was considered to be binary in nature (active or inactive) and these efforts focused on only a single target at a time (Fujiwara, et al., 2008; Warmuth, et al., 2003; Cui & Schneider, 2010). Work in the field of chemogenomics attempts to make predictions for the effect of compounds on protein targets using various methods (Koutsoukas, et al., 2011; Bredel & Jacoby, 2004; Keiser, et al., 2009). To the best of my knowledge, active learning has not been used in any chemogenomic context.

In context of active learning for directing high-content screening efforts, the value of active learning increases with the cost (computational or financial) of experimentation as active learning attempts to select for execution only those experiments from which the most benefit can be gained while avoiding experiments for which the results can be accurately predicted.

The successful implementation of an active-learning directed experimental process incorporating results from experiments involving multiple targets could result in a predictive model that can more accurately predict potential side effects of potential drugs reducing waste in their further experimentation. Additionally, a model that is capable of predicting side effects may also be able to predict alternative uses for compounds currently in use pharmaceutically.

Thesis Tasks

The overall goal of this work is to demonstrate the utility of active learning for the efficient exploration of large experimental spaces and specifically for high-content screening experiments.

To that end, we will begin by initially discussing the implementation, execution and subsequent analysis of a high-content screening campaign to identify drugs (27 treatments + control) that affect *Arabidopsis thaliana* protoplasts (six tagged proteins).

Next, using simulations, we will assess the effect on model accuracy that would result from the utilization of active learning to explore a blindly duplicated experimental space (12 tagged proteins x 54 treatments + control).

Next we will demonstrate through simulations the potential improvement in lead discovery efficiency resulting from the utilization of various active learning methods when applied across an experimental space consisting of diverse assays testing for effects on diverse target proteins and a large number of compounds.

Finally, we will extend the RandTag project in such a way as to establish an efficient framework for the detection of phenotypic changes in location patterns of endogenously expressed proteins in NIH 3T3 resulting from drug exposure and execute a campaign directed using active learning to do so.

Chapter 2: High Content Screening Using **Arabidopsis thaliana** *Protoplasts*

Protoplast Experimentation

Protoplasts can be generated through the enzymatic or physical removal of the cell wall. After the removal of the cell walls, turgor pressure causes the cells to assume a spherical shape. After a short recovery period, the cell wall regenerates. The *Arabidopsis* protoplasts used for this experiment were generated by enzymatically removing the cell walls from all above ground tissues of the plant. Because source tissues included stem, leaf and vasculature, the resulting populations of protoplasts were highly heterogeneous. It has been shown in some systems that different subpopulations react differently to stimuli (Balaban, et al., 2004; Çağatay, et al., 2009). Additionally, it has been shown that the location pattern of a single protein can vary from cell type to cell type (Faraco, et al., 2011). The vacuole also provides a unique challenge in the analysis of the location patterns of proteins in protoplasts. As most of the volume within the cell is taken by the vacuole, unless a protein is found within the vacuole itself, there is relatively limited volume remaining for the protein to be found. This can make distinguishing between two location patterns very difficult. Protoplasts were generated from six cell lines expressing different tagged proteins (AHA-ATPase, ER-GK, Talin, 313-YFP, Rab-F2a, M4). The location and function of each tagged protein are shown in Table 2. Samples of these protoplasts were counted to ensure numerical consistency from well to well and across plates. Sample false color images from each of the tagged lines are shown in Figure 4 through Figure 9. Three concentrations (2.0 μM, 0.5 μM, 0.1 μM) of nine drugs (benzylphosphonic acid, Brefeldin-A, Damnacanthal, endothall, N9-isopropyl olomoucine, oryzalin, tyrphostin, ZM-449829) were used. Information about each drug can be found in Table 3. The drugs were robotically added to samples of the protoplasts to yield a total of 168 unique experimental conditions including untreated controls. Each well was imaged in a single central slice for 16 fields per experimental well. The location of this slice was determined as a distance relative to the bottom of the glass in the 96-well plate. The robotic addition of compounds to each experimental well was synchronized with the timing of imaging on the microscope such that drug exposure times for all experiments were approximately equal (4 hours) across all wells in the multiwell plate. At least two experimental wells were available for each combination of tagged protein and drug at each concentration. Due to microscope equipment constraints, only 60 wells were available per 96-well plate. Of these, 12 wells were used for a pair of untreated wells per cell line and the remaining 48 were used for pairs of different experimental conditions.

Table 2: This table shows the cellular location and protein function of tagged protein where available. All information was gathered from Uniprot.

Tagged Protein	**Location (Uniprot)**	**Function (Uniprot)**
AHA-ATPase	Cell Membrane	Hydrogen ion transport
ER-GK	Endoplasmic Reticulum	NA
Talin	Cytoskeleton	Cytoskeletal
313-YFP	NA	NA
Rab-F2a	Cell Membrane	Transport
M4	Cell Membrane	Transport

Table 3: The actions of each compound utilized according to PubChem are shown.

Drug	**Effect (PubChem)**
benzylphosphonic acid	tyrosine phosphatase inhibitor
Brefeldin-A	Transport inhibitor
Damnacanthal	p56lck tyrosine kinase inhibitor
endothall	Phosphatase inhibitor
N9-isopropyl olomoucine	Kinase inhibitor
oryzalin	Tubulin modulator
tyrphostin	Kinase inhibitor
ZM-449829	Kinase inhibitor

Each captured image contained four channels. The GFP channel depicted the location of the tagged protein within the protoplasts. The second channel showed the location of DAPI, a fluorescent stain that binds strongly to DNA. In this experiment, DAPI served as an indicator of the health of the protoplast. A single bright spot within the protoplast indicated that DAPI reached the DNA and the cell was dead or dying. DAPI can also attach to the vacuole in healthy cells. Because the vacuole is the largest organelle in the cell by volume, most of the cell can appear to be filled with DAPI fluorescence. The third channel showed the autofluorescence of the chloroplasts within the protoplasts. Across different types of protoplasts, the number and distribution of chloroplasts can vary dramatically. The fourth channel contained a DIC image of the cells.

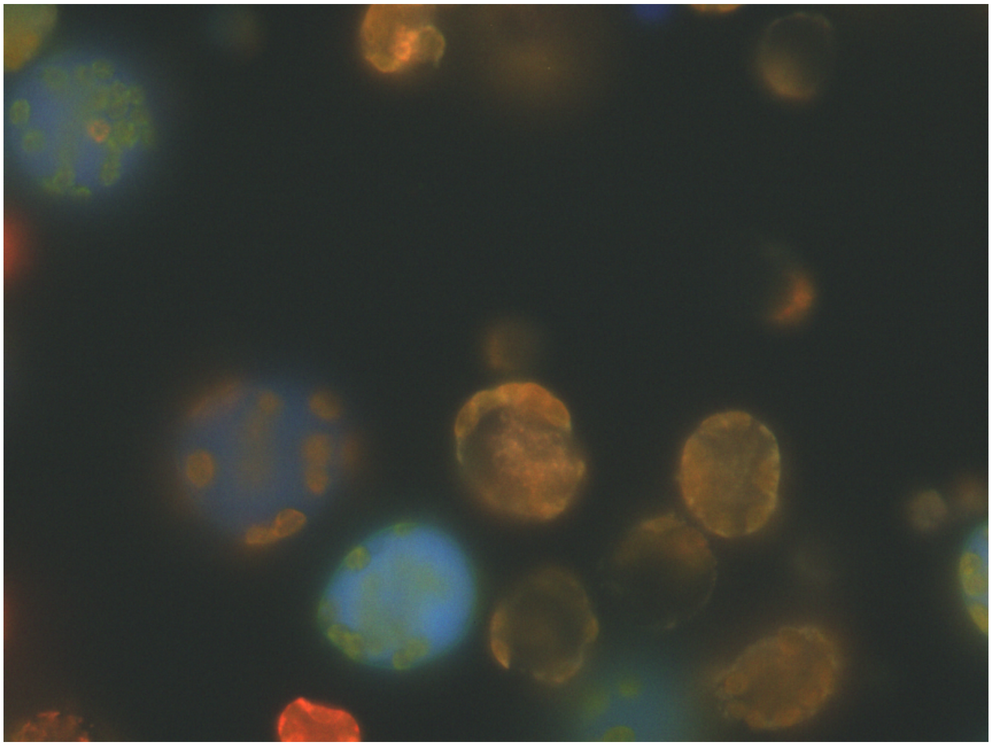


Figure 4: This false color image shows a sample image from the set of protoplast images collected. AHA-ATPase (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. AHA-ATPase is known to be found within the plasma membrane and is important for ion transport.

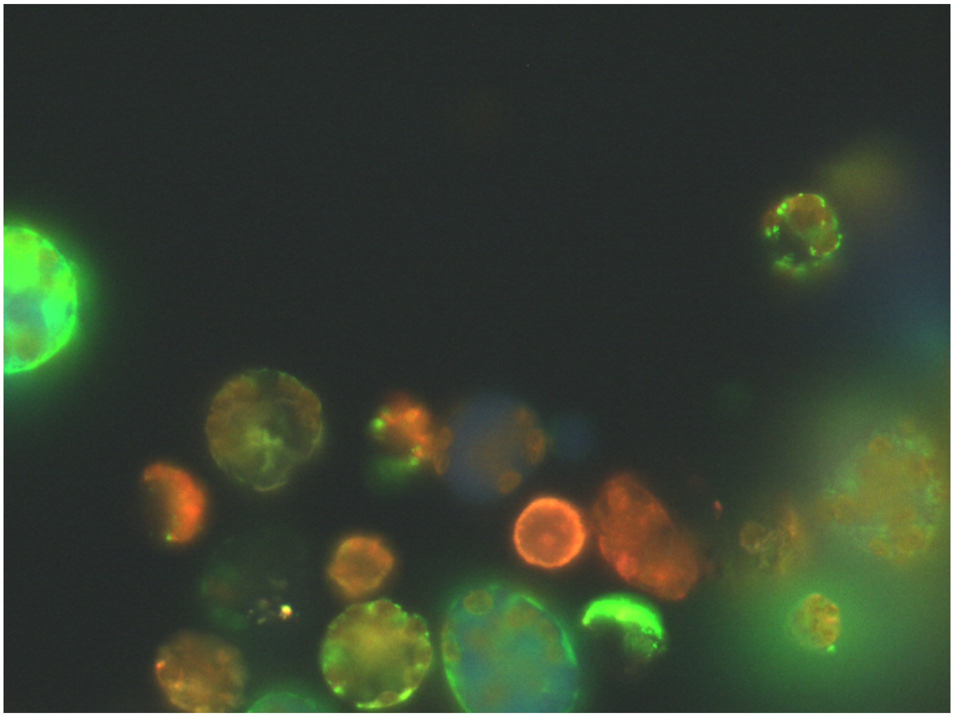


Figure 5: This false color image shows a sample image from the set of protoplast images collected. ER-GK (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. ER-GK is known to be found in the endoplasmic reticulum.

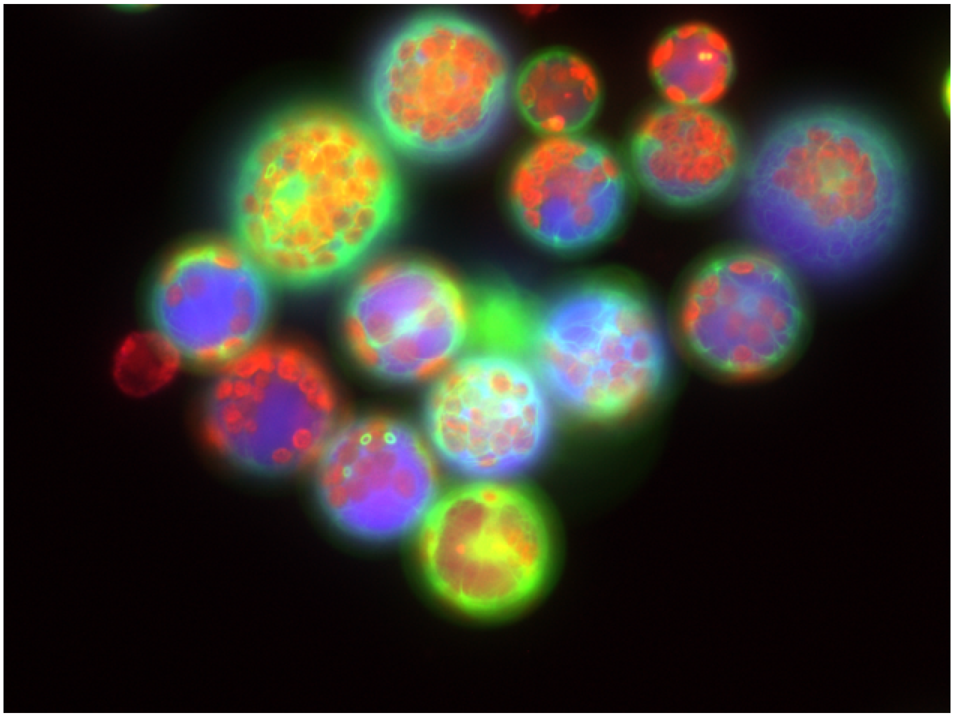


Figure 6: This false color image shows a sample image from the set of protoplast images collected. Talin (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. Talin is a cytoskeletal protein.

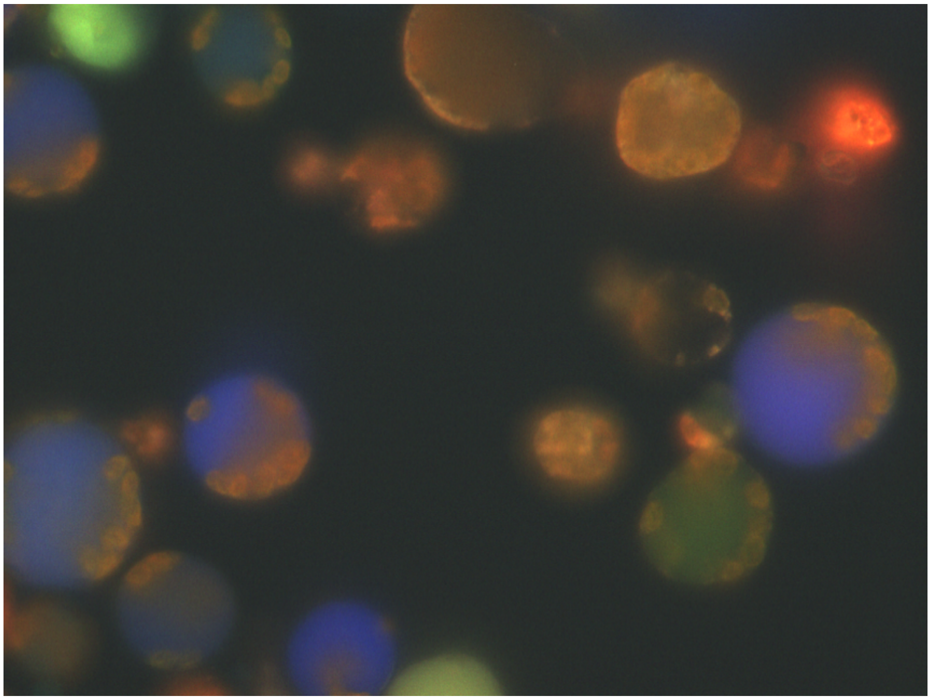


Figure 7: This false color image shows a sample image from the set of protoplast images collected. 313-YFP (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. No information is available about the function or location of cellular location of 313-YFP.

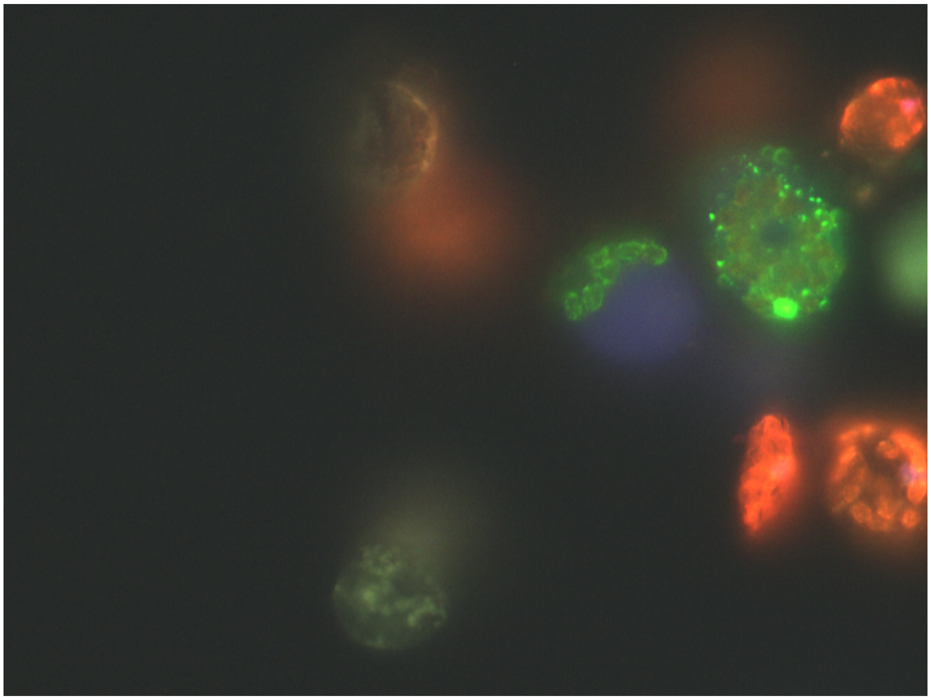


Figure 8: This false color image shows a sample image from the set of protoplast images collected. Rab-F2a (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. Rab-F2a is involved in protein transport and is normally found in the membrane or surface of the vacuole.

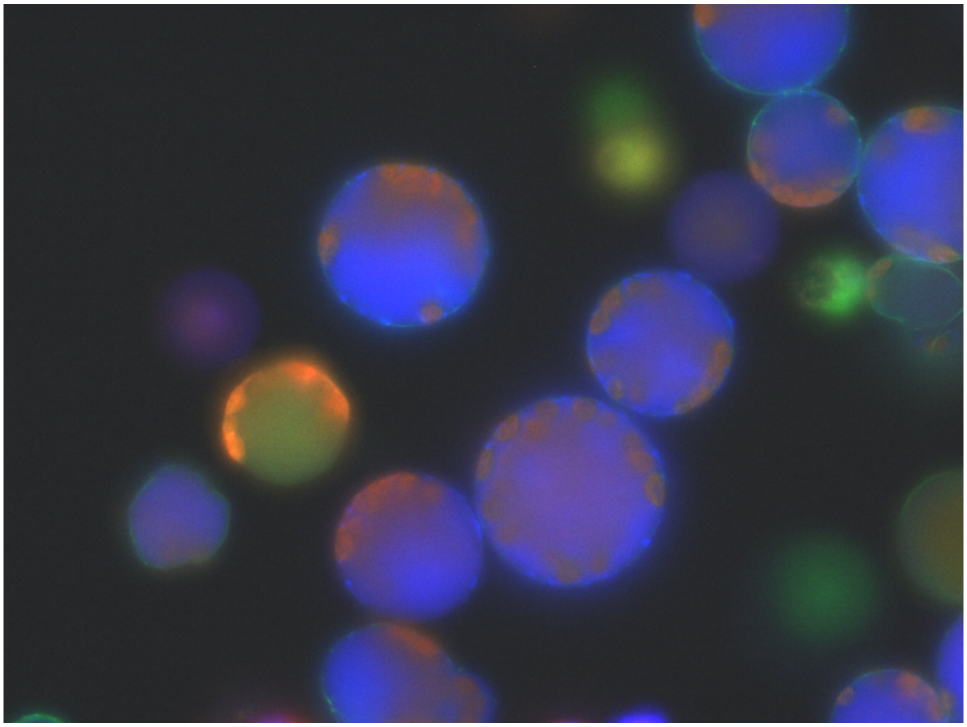


Figure 9: This false color image shows a sample image from the set of protoplast images collected. M4 (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. M4 is a transport-related protein found in the cell wall and other locations.

Use of Hough Transform to Identify Cells

The goal of the image analysis process is to detect and characterize changes in the population phenotypes. In these particular *Arabidopsis* protoplasts, there was significant variation within the population of protoplasts. As a result of the protoplasts being generated from above-ground tissues, cells were sourced from the stems, leaves and vasculature. The diversity of the cells can be seen clearly in Figure 6. Because of this significant variation and the possibility that different cell types might react differently to a compound, it would likely have been inappropriate to calculate field level features. As a result of the cells being generally spherical in shape, we chose to utilize a circular Hough transform (Duda & Hart, 1972) to identify the circles within the images. The Hough transform works by convolving a circle over a thresholded image. The result is that centers of circles within the image tend to have high values after the convolution with circles of matching radii. In order to identify cells in these images, edges were detected using a Canny edge detector in the DIC channel. The Hough transform was utilized on these images (Canny, 1986). A single pass of the Hough transform can be utilized to detect circles of a single size. Because of the diversity in the sizes of the circles, multiple passes of the Hough circle detector were used for various radii from 60 (smaller than the smallest protoplast observed) to 240 pixels (larger than the largest protoplast observed). Each pass resulted in a matrix that was then normalized according to the circumference of the circle used for detection. This resulted in a three dimensional matrix (image width × image height × radii of circles) with each cell (x,y,r) representing the confidence of having a circle of radius r centered at position x,y.

Circles were greedily accepted based on the probabilities. The cells generally did not overlap significantly, so when a cell was accepted a cone was removed from consideration in the confidence matrix preventing subsequent selected protoplasts from overlapping with the selected protoplast. The removed cone was centered at the detected center point (x,y), and at each r-slice had a radius equal to twice the r of that layer. This minimized the number of overlapping cells selected. A threshold was manually chosen such that the quality of segmentations was qualitatively maximized (15%). Sometimes cells were identified within regions where there were no cells present. It was hypothesized that local variation within the DIC channel allowed for the detection of “edge” pixels which were then detected as the edges of cells. These incorrectly identified cells were removed by filtering circles with very low intensity variation in the DIC channel. For each of these accepted circles, SLF34 (http://murphylab.web.cmu.edu/services/SLF/) features were calculated with the GFP channel as the primary protein channel and with autofluorescence and DAPI separately as reference channels. These features take into consideration the spatial distribution of color intensities within the protoplasts. This resulted in two feature vectors that were largely identical. For the resulting two feature vectors, the duplicated features were unduplicated to yield a single feature vector to describe each of the circles.

The result of this effort was a set of feature vectors describing each found cell under each experimental condition as well as a the confidence that the found cell was actually a circle. A histogram of resulting confidences and associated sample images is shown in Figure 10. It is clear that as the confidence decreases, it is less likely that a complete circle will be found within that block. This is hypothesized to be due to noise in the DIC channel images.

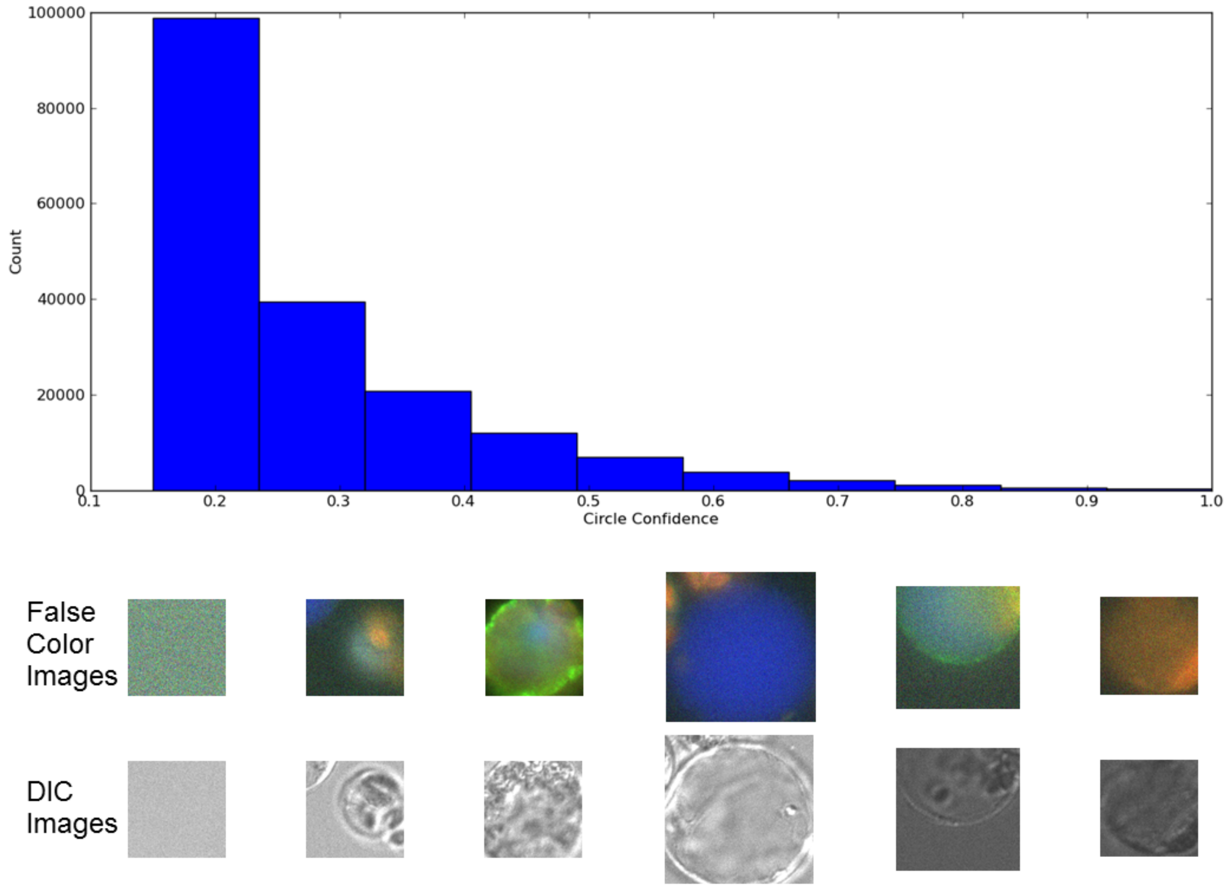


Figure 10: A histogram of Hough circle confidences and associated patches from the images is shown. False color images are shown in the same manner as in the previous figure along with the DIC image from the same patch.

Detecting Phenotype Differences using Mixture Models

In order to test for differences in phenotypes, an approach to measure the overlap between distributions was utilized. First, the protoplasts were clustered based on their features into clusters using K-means (MacQueen, 1967). Using the resulting labels for all protoplasts, a mixture model was formed in which each experiment was represented by a vector of length K containing the proportion of protoplasts from that experiment belonging to each cluster. The results of this process for K=10 are shown in Figure 11. The mixture proportions are represented by the height of the blocks and they are sorted based on their deviation from cell line to cell line. Examples for four clusters are shown in each of the color coded boxes. The first cluster (blue box, bottom left) showed the most variation across cell lines. Protoplasts belonging to this cluster were largely absent from talin and present to varying degrees in the other cell lines. The third cluster (green, bottom right) is present in significant amounts only in talin tagged cells. Cells across all lines have approximately equal proportions of protoplasts belonging to the tenth cluster (light purple, top right).

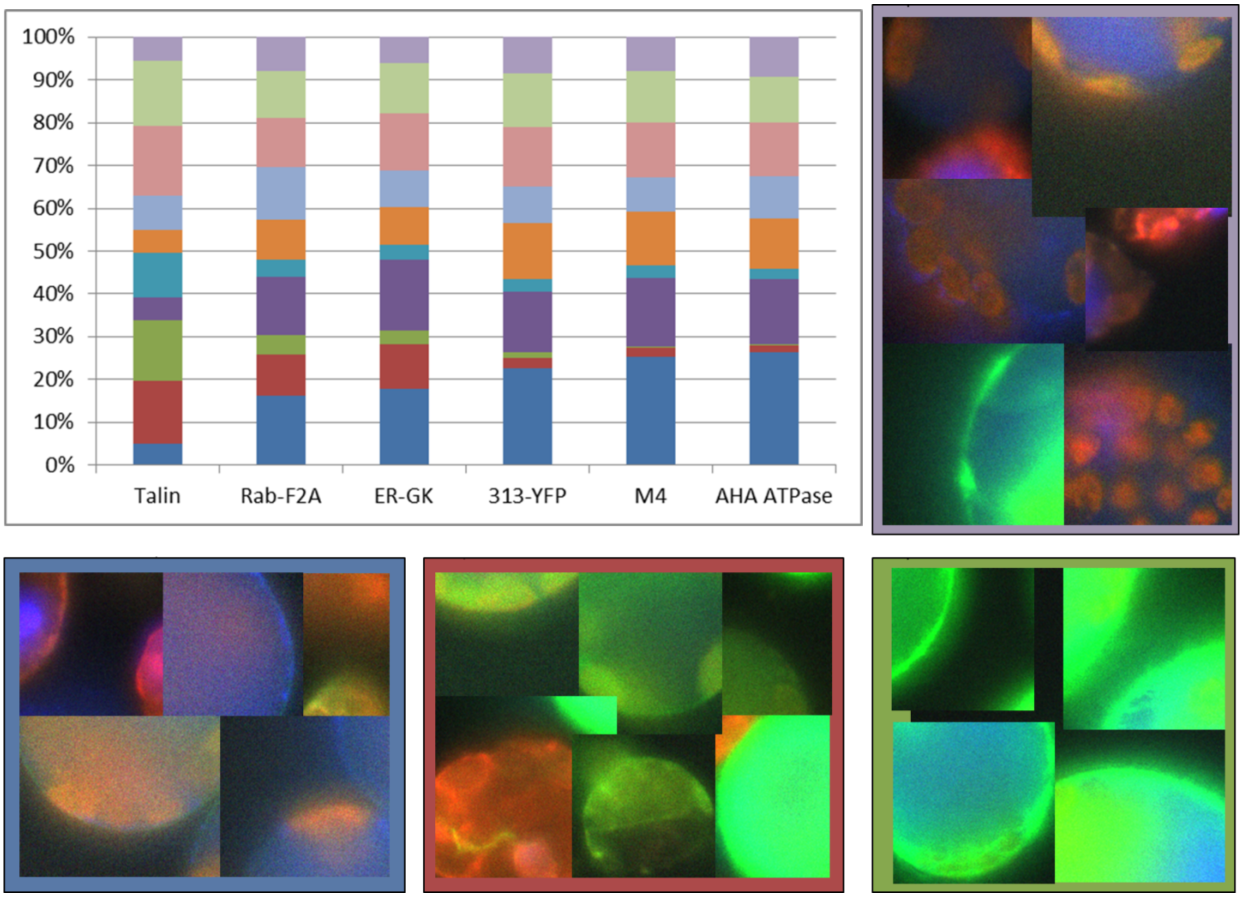


Figure 11: After clustering protoplasts, the proportion of cluster membership for each cluster was calculated across all images for each untreated cell line. Samples images from four clusters are shown. The mixtures are ordered by their variance across lines.

Two parameters should be considered when building a model in this manner: an appropriate K and an appropriate confidence threshold for the acceptance of circles as protoplasts. In order to select a model, it was assumed that replicated control experiments across should have similar mixtures (must link constraints). Conversely, control experiments from different cell lines were assumed to be different (cannot link constraints). The Manhattan distance was used to calculate the difference between two mixtures. A range of K from 2 to 49 and confidence thresholds from 5% to 65% in increments of 1% were tested. Above 65%, some experiments were found to have no protoplasts and were thus immeasurable. At each pair of parameters, the ratio of the mean cannot-link distance to the mean must-link distance was calculated. We accepted parameters which maximized this ratio. These results are shown in Figure 12. The best pair of parameters was K=3 and circle confidence > 0.17.

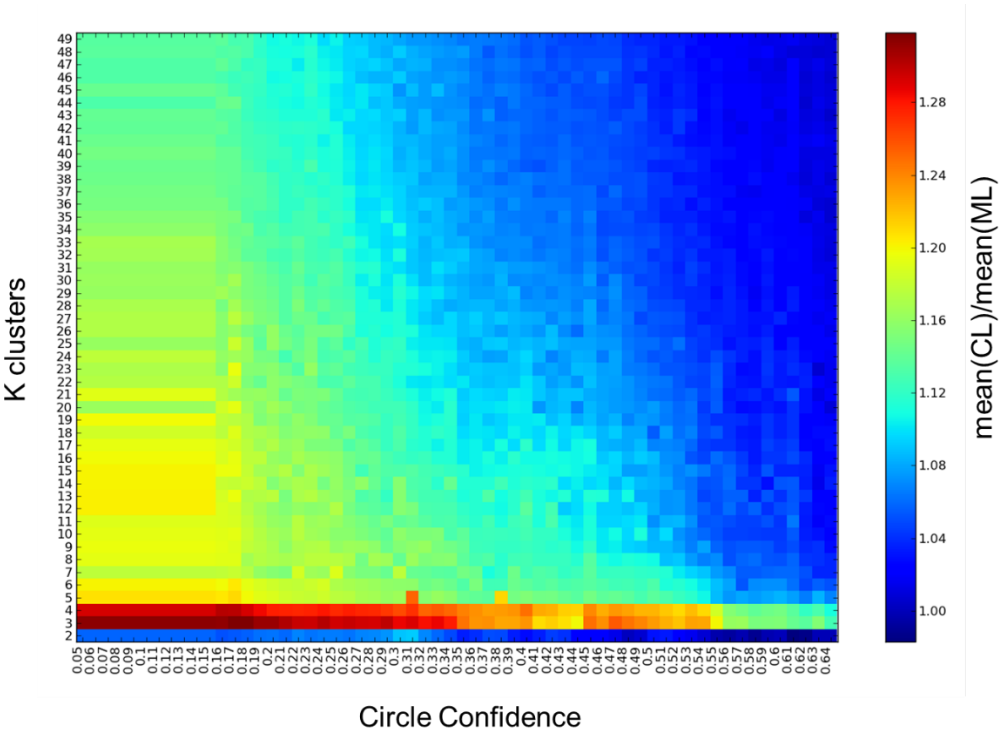


Figure 12: Protoplasts with a circle confidence above a threshold were clustered into K clusters that were used to calculate mixtures for all wells. The ratio of the mean distance between duplicated unlike control conditions (cannot link conditions) to the mean distance between like control conditions (must link conditions) is shown. A higher ratio indicates a better model. The best pair of parameters was found at K=3 and confidence > 0.17.

All protoplasts for each control condition were grouped together into a single experiment. Using the parameters discovered determined from Figure 12, (K=3 and confidence > 0.17) the pairwise distance between each control condition was calculated. As shown in Figure 13, these results indicate that the distribution of protoplasts for the talin line was quite different from that for other tagged lines. Based on a qualitative comparison of the images, this was not a surprising result. These results also indicate that AHA-ATPase, 313 YFP and M4 are largely indistinguishable from one another and Rab-F2a and ER-GK were indistinguishable from one another.

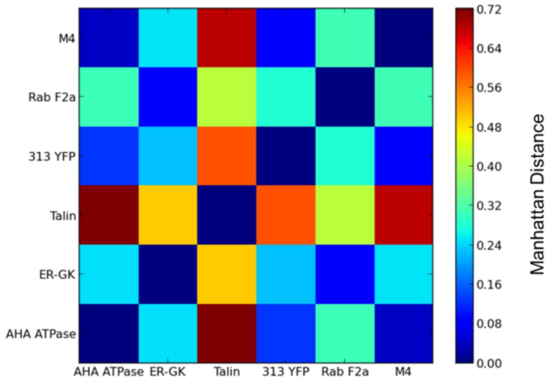


Figure 13: Pairwise distances for mixtures were calculated using the aforementioned parameters. The distributions of protoplasts for the Talin line appear to be the most different from the other lines.

Detecting Phenotype Differences using PhenoRipper

Alternative methods for identifying regions of interest within fields have been developed. PhenoRipper is a software system designed for the large scale analysis of high-content screening data (Rajaram, et al., 2012). It was specifically designed to operate in experimental contexts in which heterogeneity provides a challenge for analysis. Instead of attempting to carefully segment cells or calculate field-level features, this method looks for patterns in the colors found within the image. First, images were thresholded based on a hand-selected threshold (20%). Then a random selection of pixels across numerous images was clustered using K-means based on their color intensities. The images were then projected into the space of these discrete colors formed from the clustering of colors. Then the image was broken into small non-overlapping blocks (100 x 100 pixels). For each block, the mixture of colors within that block was calculated. Blocks that were more than half "background color" were discarded. A random selection of these blocks was clustered using K-means based on their color mixtures. Then the images were projected into the block-type space. Finally, superblocks were found by calculating the mixture of block types found in every 3 × 3 patch of blocks (superblocks were allowed to overlap). The mixtures of block types within the superblocks were used to cluster superblocks using K-means as well. The final phenotype for an experiment was represented by the experiment's mixture of superblock types. For each run of K-means using PhenoRipper, a K of 10 was used as that was the recommended default for the program. This program was implemented in MATLAB and worked nicely for small sets of images. However, with nearly 60,000 images, there were stability issues, so the process was re-implemented in Python to alleviate those problems and take advantage of the cluster architecture to process the images more quickly. The resulting mixtures were then analyzed in the same manner as the previous mixtures. The Manhattan distance between the control mixtures is shown in Figure 14. In this case, the talin line was again relatively different from other lines, but M4 was also found to be somewhat different from the other lines.

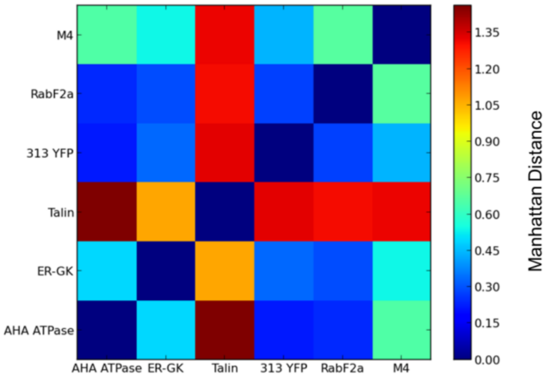


Figure 14: Pairwise distances for mixtures were calculated using mixtures resulting from executing the PhenoRipper method on the collection of images. The mixtures for the Talin line and M4 appear to be the most different from the other lines.

Detecting Phenotype Differences Using 1-Nearest Neighbor Classifiers

A third alternative method for this analysis was also developed. Two distributions can be easily compared using a classifier trained to recognize each class. If the accuracy of that classifier is high, then the confusion between the two classes is low and thus the distributions are highly dissimilar. In this case, we utilized a 1-nearest-neighbor classifier. For two samples (of equivalent size) from classes with identical distributions, as the number of observations in each class increases, the accuracy is expected to approach 50%. For two samples (of equivalent sizes) from classes with different distributions, as the number of observations in each class increases, the error of the classifier should approach a value directly proportional to the overlap of the two original distributions.

Because the number of protoplasts between experiments varied, there would be significant effects on the resulting accuracy for the nearest neighbor classifier. Thus, for each comparison to be made, a nearest neighbor classifier was "learned" on subsets of the data such that the number of samples from each class was balanced. For each experimental condition this was executed multiple times and an average accuracy was calculated.

In this case, the selection of a K was unnecessary, but the selection of a circle confidence for protoplasts was required. For this analysis, we wanted to maximize the accuracy of classifiers trained on different control experiments and minimize the accuracy of classifiers trained on identical experiments. We desired to maximize the ratio of the mean of these two values such that the distributions of different controls are most separated while like control distributions are most similar. The resulting ratios as a function of circle confidence are shown in Figure 15.

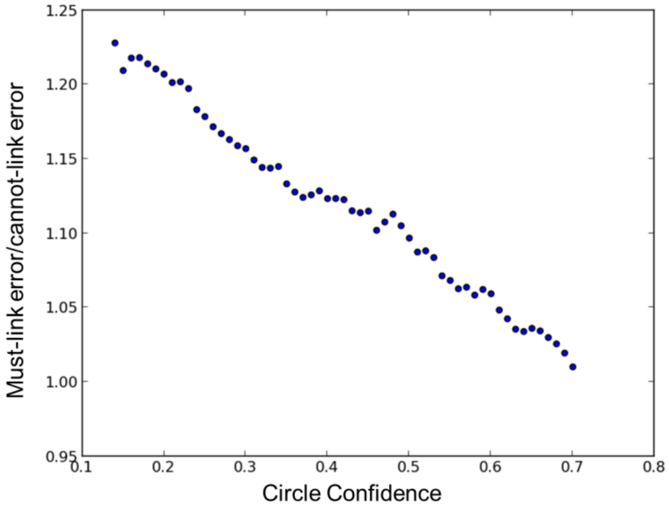


Figure 15: The ratio of the mean error of classifiers trained with like control conditions to the error of a classifier trained with different control conditions is shown. The best circle confidence is the value that maximizes the ratio (confidence > 0.15).

Once the best circle confidence threshold was determined, the control experiments were compared. The results of these comparisons are shown in Figure 16. These results indicate again that talin is easily distinguishable from the other lines. Relative to the differences between talin and the other lines, no other differences appear to be very significant.

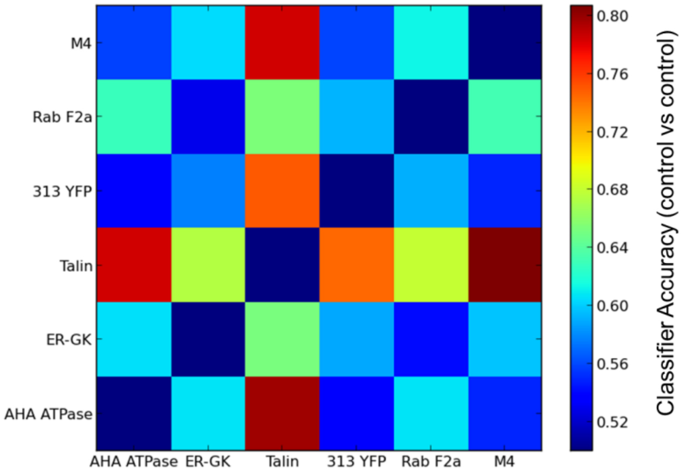


Figure 16: The accuracy of 1-nearest neighbor classifiers "trained" to recognize each control line is shown. A higher accuracy indicates less similarity between the distributions of the protoplasts for the lines.

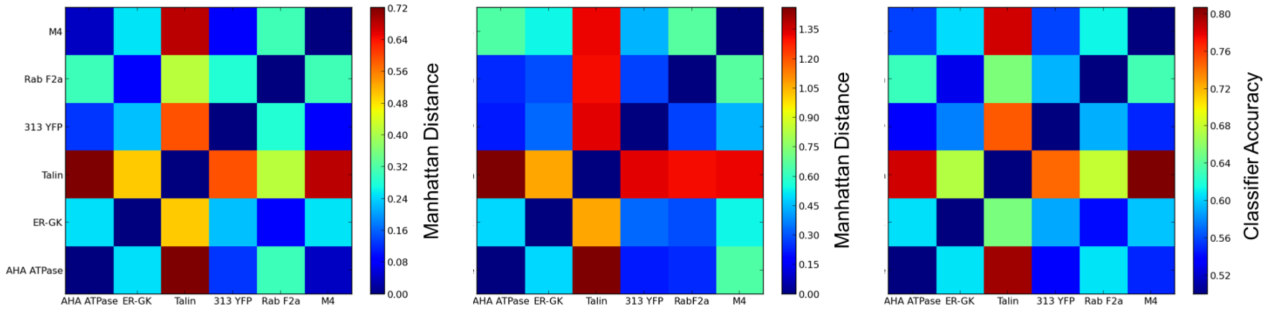


Figure 17: The results for the comparison of the control lines across all three methods are shown together for more direct comparison. From left to right: protoplast mixture model, PhenoRipper assessment and protoplast classifier accuracy-based assessment.

When comparing across the different phenotype comparison methods (Figure 17), it is clear that the protoplast mixture-based method (left) and classifier accuracy-based method (right) have very similar results. This is not surprising as the basic unit (the Hough detected protoplast and the associated features) was identical in both analyses. However, the model selection for these two methods yielded different cutoffs for circle confidence. This probably contributes to the fact that the mixture based model appears to detect a slightly larger relative difference between Rab-F2a and the other lines than does the classifier accuracy based method. When considering the six lines and all methods, the population distributions appear to fall into three groups. Talin, a cytoskeletal protein, exhibited a unique phenotype when compared to all other lines. Rab-F2a and ER-GK exhibited similar phenotypes to one another. The phenotypes of the remaining three lines appeared very similar to one another.

Detection of Drug Effects using Mixture Models

To detect the effects of compounds from the distributions of protoplasts, a mixture based approach was utilized. For each line, the distribution of Manhattan distances between all control experiments on separate plates was calculated (must-link distances). All protoplasts across all plates coming from common conditions were grouped together, and a single mixture was calculated for each experimental condition (cell line, drug). Then the distances between mixtures from experimental conditions and their controls were calculated. These results are shown in Figure 18. By themselves, these results are not useful for making direct conclusions as they give no indication of the variance of a cell line. To determine the probability of an effect, the intrinsic variation of the cell line without treatment must be considered. In order to determine the significance of these distance calculations, the proportion of must-link distances derived from the same control lines that were greater than the measured difference between the controls and experimental conditions were calculated, giving a p-value. The resulting p-values are shown in Figure 19.

Overall, the results of the analysis shown in Figure 18 indicate that there may be somewhat large differences between experimental conditions and control experiments. However, Figure 19 indicates that these differences were not statistically significant. Normally in this situation, some form of multiple hypothesis testing might be utilized, but in this case there were no distances which had a p-value < 0.05. Some experimental conditions appeared to have relatively low p-values, but their corresponding distances did not match any larger trend amongst drugs across concentrations that would be consistent with first-order concentration dependent effects. There was weak evidence for effect from Damnacanthal across most lines except Talin. Unfortunately, there was also weak evidence that media in the middle "concentration" had an effect on Rab F2a. As this experimental condition is equivalent a control, no hits were expected for media at any concentrations.

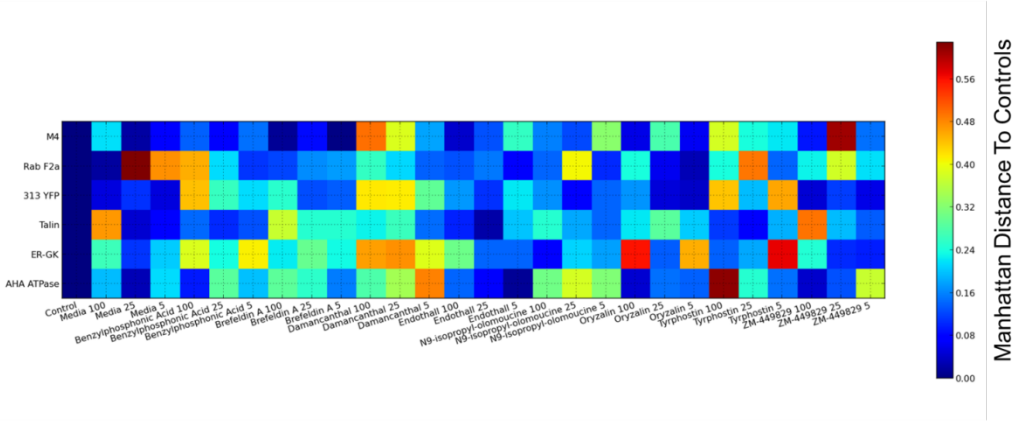


Figure 18: Mixtures were calculated based on the clustering of protoplasts for each unique condition (cell line, drug). The distances between each condition and the associated control condition were calculated and are shown. Numbers accompanying the compound names indicate the relative dose of that compound for the experiment.

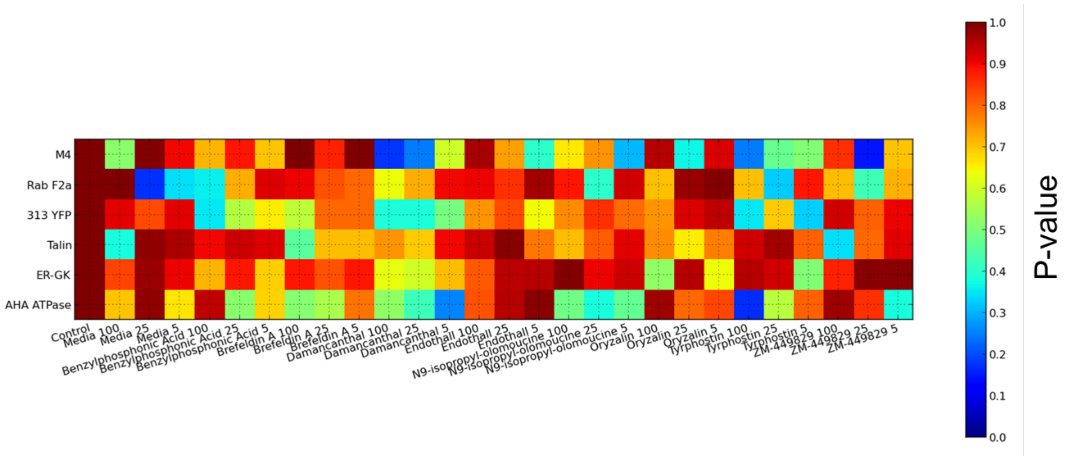


Figure 19: For each experimental condition, the distance between its mixture and the corresponding control mixture was calculated. A p-value was determined by calculating how many pairwise within control distances were greater than the experiment to control distance. No distances were found to be statistically significant.

Detection of Effects Utilizing PhenoRipper

Likewise, the distances between experimental conditions and the corresponding control conditions were calculated (Figure 20). The p-values for these distances were calculated as well (Figure 21). Based on this analysis, Damnacanthal and Tyrphostin appeared to have statistically significant effects on M4, Rab-F2a, 313-YFP and AHA-ATPase. At the highest concentration N9-isopropyl-olomoucine appeared to have an effect on AHA-ATPase. Tyrphostin appeared to have an effect on all lines, but not all effects were statistically significant. There was no strong evidence for any of the media based "drugs" having any effect on the tagged lines.

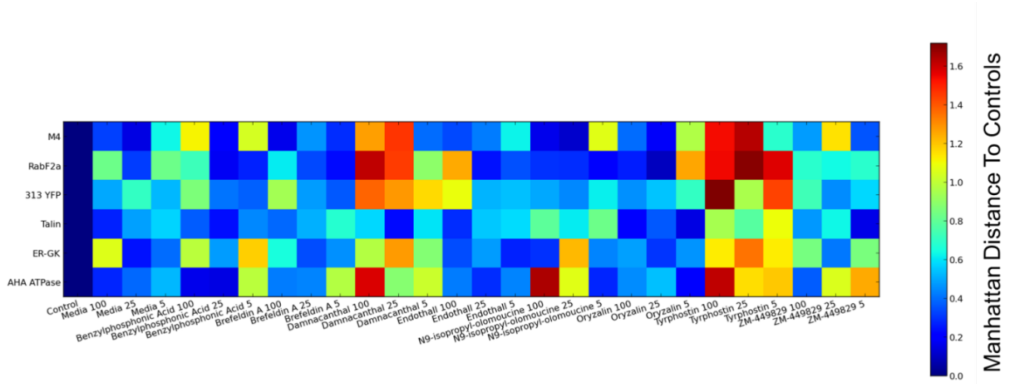


Figure 20: Mixtures were calculated using the PhenoRipper method. The distances between each condition and the associated control condition were calculated and are shown.

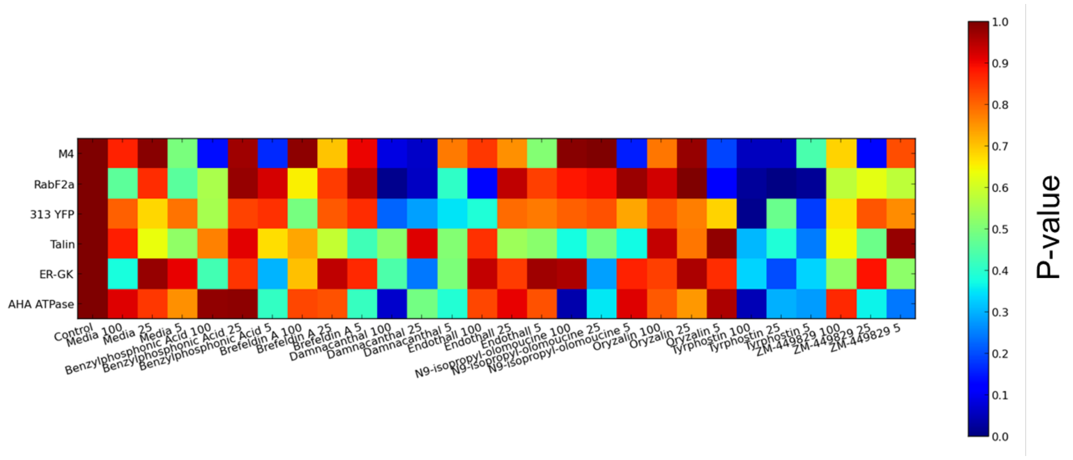


Figure 21: Using the PhenoRipper devised mixtures, for each experimental condition, the distance between its mixture and the corresponding control mixture was calculated. A p-value was determined by calculating how many pairwise within control distances were greater than the experiment to control distance.

Detection of Effects from Drug Treatment Using 1-Nearest Neighbor Classifiers

For each experimental condition, the accuracy of the classifiers when trained to recognize the experimental condition and the control condition was calculated. These results are shown in Figure 22. Confidence was assessed by calculating how often pairs of control wells from the same cell line had a higher accuracy than that calculated between the control condition and the experimental condition. The resulting p-values are shown in Figure 26. None of the effects were below a cutoff of 0.05, but there were some trends. Damnacanthal showed evidence for having affected all lines except talin. There was weak evidence for a low concentration of Tyrphostin having an effect on 313 YFP. N9-isopropyl-olomoucine showed evidence for affecting AHA-ATPase with this method as well. As expected, there was no strong evidence for any of the media drugs having any effect on the protoplasts.

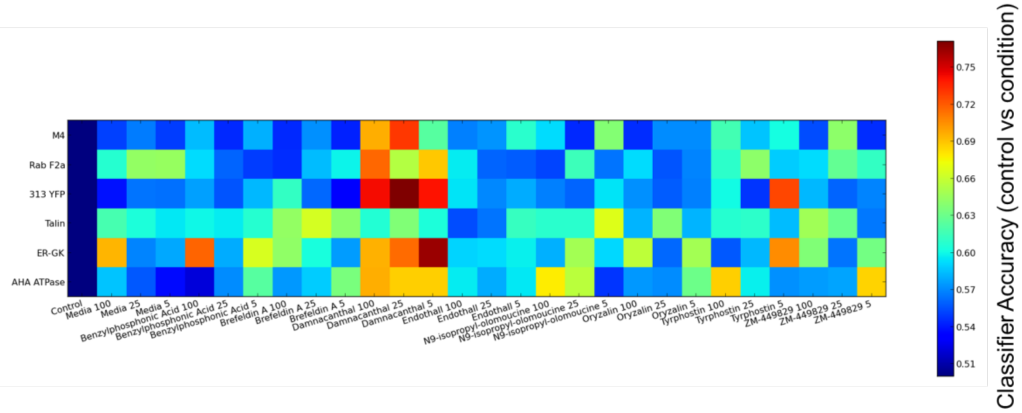


Figure 22: The mean errors of 1-nearest neighbor classifier are shown for when each classifier was "trained" to recognize the protoplasts from the experimental condition and protoplasts from the corresponding control condition. A greater error indicates more similarity between the two distributions of protoplasts.

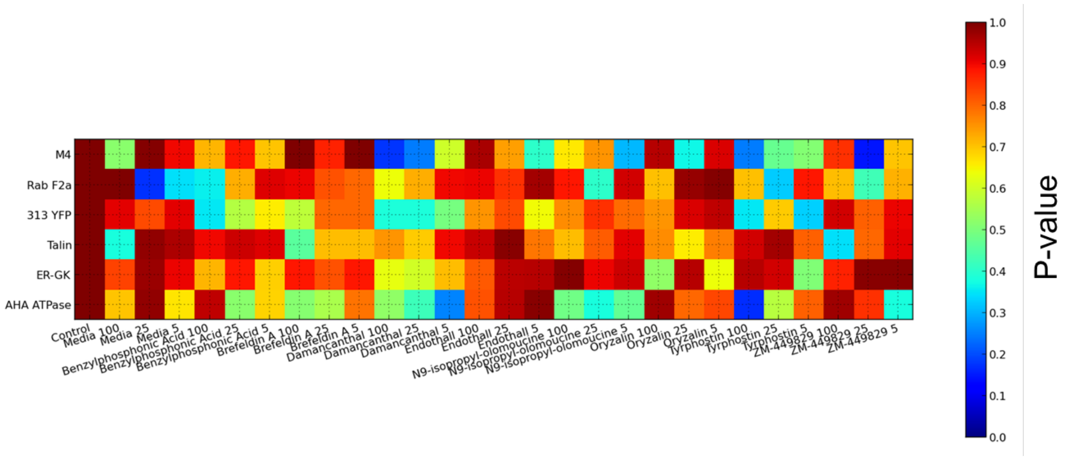


Figure 23: For each experimental condition, the distance between its mixture and the corresponding control mixture was calculated. A p-value was determined by calculating how many pairwise within control distances were greater than the experiment to control distance. No distances were found to be statistically significant.

When comparing the p-values for the drug affects across different analysis methods (shown together for convenience in Figure 24 to Figure 26), some trends were observed. Across all analysis methods, there was only weak evidence for any of the drugs having an effect on talin. Damnacanthal showed effects across most lines to some degree. When considering Damnacanthal effects against M4, one sees that there may be a concentration dependent effect whose relative strength was detected across all analysis methods.

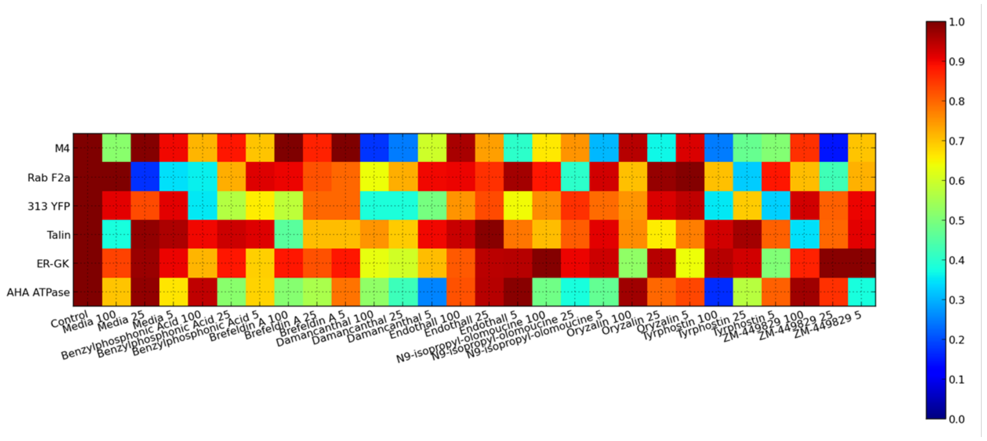


Figure 24: For each experimental condition, the distance between its mixture and the corresponding control mixture was calculated. A p-value was determined by calculating how many pairwise within control distances were greater than the experiment to control distance. No distances were found to be statistically significant.

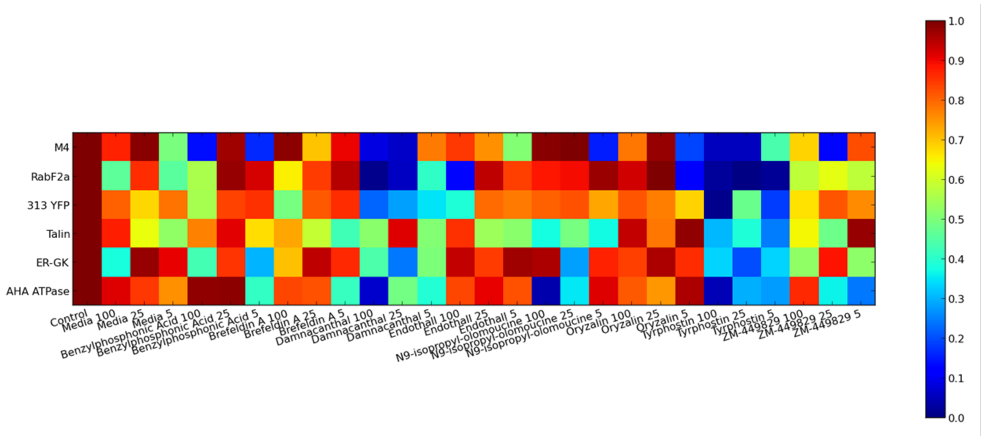


Figure 25: Using the PhenoRipper devised mixtures, for each experimental condition, the distance between its mixture and the corresponding control mixture was calculated. A p-value was determined by calculating how many pairwise within control distances were greater than the experiment to control distance.

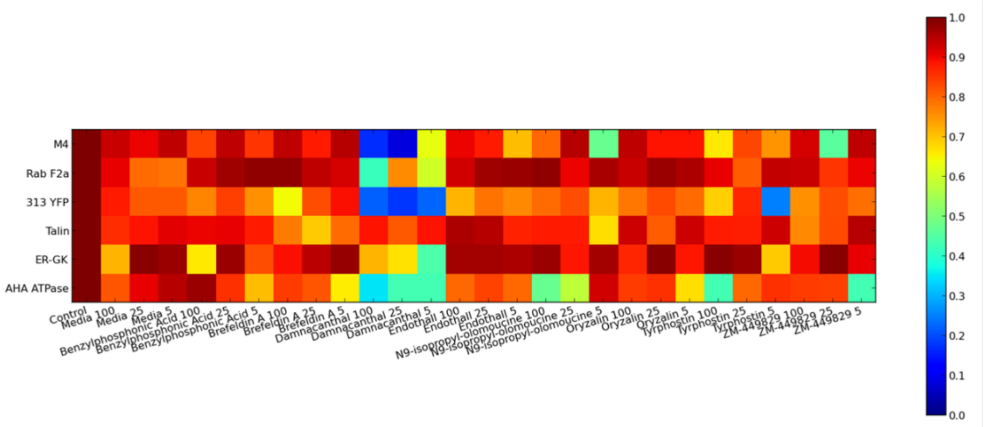


Figure 26: The probability of the condition-control classifiers having an error drawn from the same distribution as the control-control errors.

Full Assessment of Effects Across Concentrations and Methods

We desired to completely assess each drug across all concentrations. Because it is possible for drugs to exhibit non-linear effects, we opted to calculate the probability that all observations for the effects of a single compound across concentrations on a single tagged line came about randomly. For each drug and each line, this is simply the product of the p-values from each of Figure 23, Figure 25 and Figure 26. These results are shown in Figure 27. This method reduced our sensitivity to variations in experiments which had few observations. Firstly, across all analysis methods, there was no strong evidence for media having any effect on the protoplasts. There were no effects detected at all using the protoplast-based mixture model detection method. Using PhenoRipper, evidence for effects was detected for Damnacanthal across all lines except for talin. Again it appeared that the probability for a false-positive detection of effects from N9-isopropyl-olomoucine on AHA-ATPase was small. Using the classifier error-based assessment, Damnacanthal showed the strongest effects across all lines except for talin.

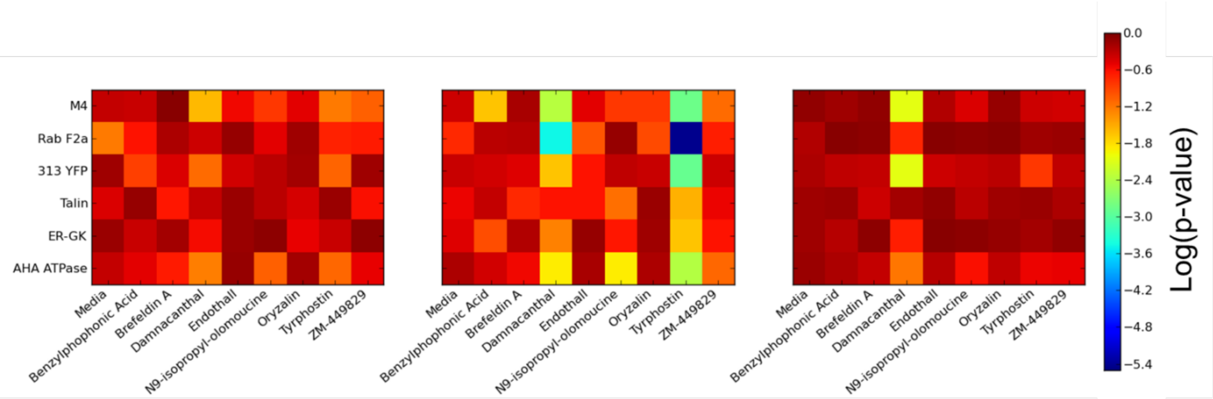


Figure 27: For each drug, all concentrations were combined such that the resulting p-value is the probability that the detected differences between the drug conditions and the control conditions arose randomly. The heat maps from left to right represent the following analysis methods: protoplast mixture model, PhenoRipper mixture model, classifier error-based detection.

To globally assess the effects of drugs on all proteins across all analysis methods, a similar approach was taken to measure the probability of all measurements across all methods having come about as a result of random chance. The results of this final assessment are shown in Figure 28. These results indicate that Damnacanthal is likely to have actually affected all proteins, except for talin. Also Tyrphostin is likely to have affected all proteins except for talin. The effect of N9-isopropyl-olomoucine appears to be confirmed based on the combination of all analyses.

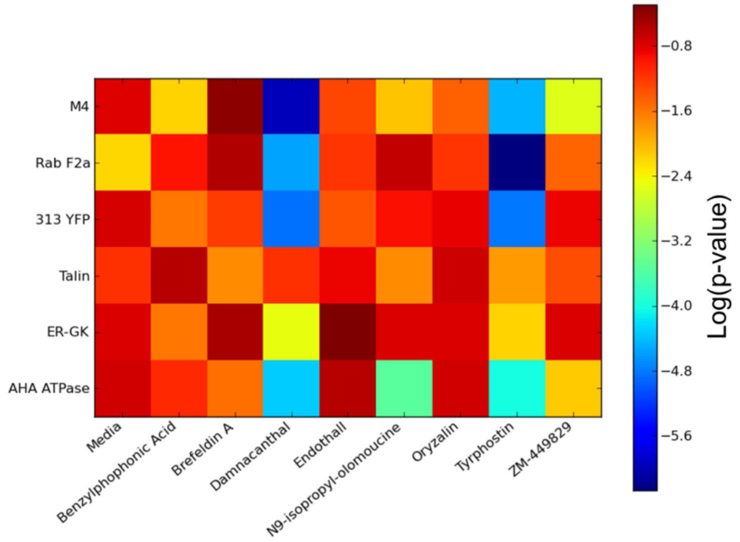


Figure 28: Across all three analysis methods, for each drug, all concentrations were combined such that the resulting p-value is the probability that the detected differences between the drug conditions and the control conditions arose randomly.

Discussion

Biological Insights

When considering the six lines, the population distributions appear to fall into three groups. Talin, a cytoskeletal protein, exhibited a unique phenotype when compared to all other lines. Rab-F2a and ER-GK exhibited similar phenotypes to one another. The phenotypes of the remaining three lines appeared very similar to one another. As a result of the similarities in these phenotypes, one might expect to see that, if affected at all, similar proteins would be affected by the same drugs. Talin, in a phenotypic cluster by itself appeared to be largely unaffected by any of the drugs. AHA-ATPase, 313-YFP and M4 which could be clustered based on their phenotypic similarity appeared to each be affected by Damnacanthal and Tyrphostin. Rab-F2a and to a lesser degree ER-GK also appeared to be affected Damnacanthal and Tyrphostin as well. AHA-ATPase showed some evidence of having been affected by N9-isopropyl-olomoucine. The similarity of the effects detected can also be seen in a hierarchical clustering of the lines based on their observed effects in Figure 29.

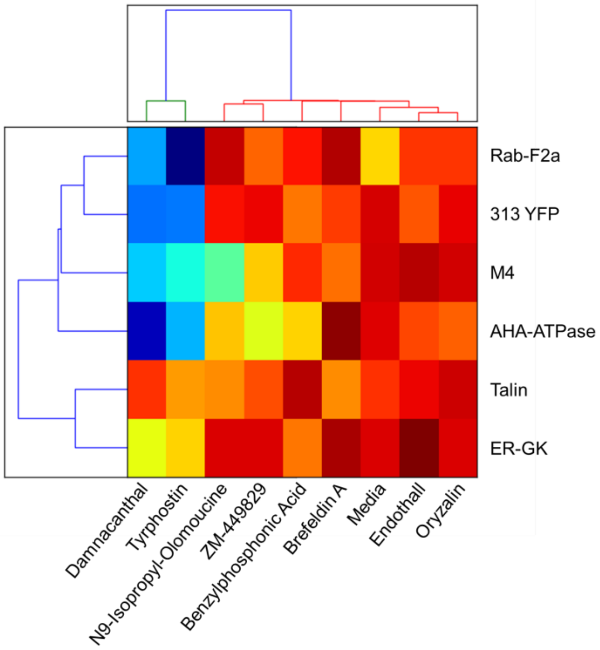


Figure 29: The tagged proteins and drugs were sorted using hierarchical clustering along both axes.

Conversely, if a drug affects some tagged protein, it would be reasonable to expect that a similar drug might affect the same protein. That appears to be the case for these observed effects. Damnacanthal and Tyrphostin are both tyrosine kinase inhibitors and there was significant evidence that both affected Rab-F2a, 313-YFP, M4 and AHA-ATPase. N9-

isopropyl-olomoucine and ZM-449829 are both cyclin dependent kinase inhibitors. There was not strong evidence that they affected Rab-F2a or 313 YFP, but there was some evidence that they both affected M4 and AHA-ATPase. Of the remaining drugs, Brefeldin A and Endothall are both phosphatase inhibitors and did not appear to have any significant effects. Oryzalin is a tubulin modulator and as such, one might have expected it to have an effect on a cytoskeletal protein such as talin, but an affect was not observed. Perhaps the concentration was too low or the exposure duration was too short to observe an effect. Using mixture based methods, the distances between mixtures were calculated. These distances represented the changes in relative populations of different types of protoplasts as a result of drug exposure. Given the heterogeneity of the populations, it is conceivable that only a subset of the population might have been affected. Relatively small, but statistically significant differences might give evidence for either small effects on large populations or significant effects on small populations. Because of the relatively poor segmentation of these protoplasts, it was difficult to determine the types of protoplasts affected by these drugs. An improved segmentation method may have allowed for a more thorough assessment of subtle population changes which could also have given evidence for biological reason underlying these effects.

One common task in location proteomics is to predict the subcellular location pattern of a protein from images. In the case of protoplasts, this task is made more challenging as many proteins are found outside of the vacuole. As showed previously, our cell lines easily clustered into three groups and members of each group were indistinguishable. With the inclusion of drug effect data, we can now separate these proteins from one another. Talin can be separated from the other lines using only on the control images. Using only control images, Rab-F2a and ER-

GK were easy to distinguish from the rest of the cell lines, but impossible to distinguish from one another. With the addition of information about how each protein is affected by Tryphostin or Damnacanthal, these two lines are easy to distinguish because ER-GK shows only a weak effect as a result of these drugs while Rab-F2a is strongly affected. For the remaining group of 313 YFP, M4 and AHA-ATPase, it is easy to distinguish 313 YFP when testing with N9-isopropyl-olomoucine because it is not affected. A test with Benzylphosphonic acid would then allow for the separation of M4 and AHA-ATPase as M4 is unaffected.

The biological interpretations of the classifier error-based assessment and the PhenoRipper methods are also difficult to determine. For that reason, sample images from lines under conditions which showed some effect in these analyses are shown. Sample images are shown for M4 untreated (Figure 30) and treated with Damnacanthal (Figure 31) and Rab-F2a untreated Figure 32and treated with Tyrphostin (Figure 33). The sample images shown were selected based on the focus and the number of cells in the images in order to best give an impression of the population.

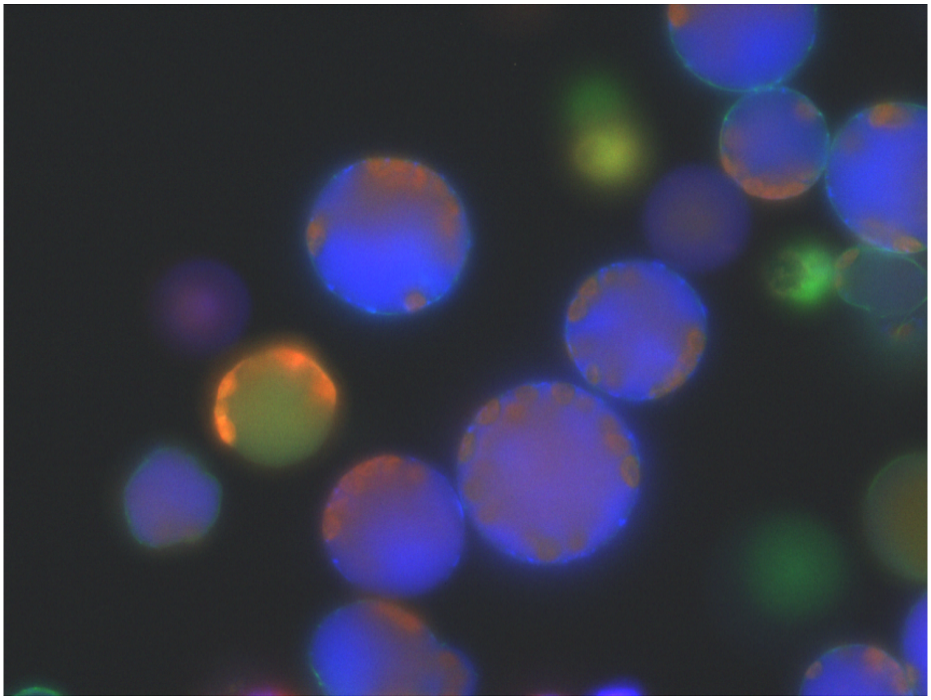


Figure 30: This false color image shows a sample image from the set of protoplast images collected. M4 (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. M4 is a transport-related protein found in the cell wall and other locations.

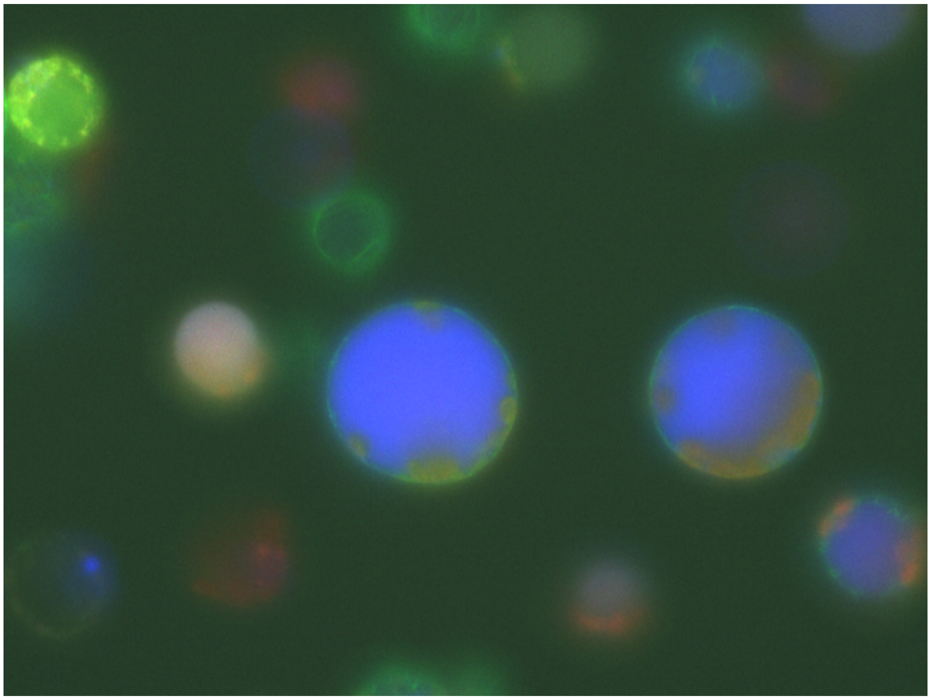


Figure 31: This image was selected from a well that contained cells tagged for M4 and treated with Damnacanthal. An image of untreated cells taken from the same plate is shown in Figure 9.

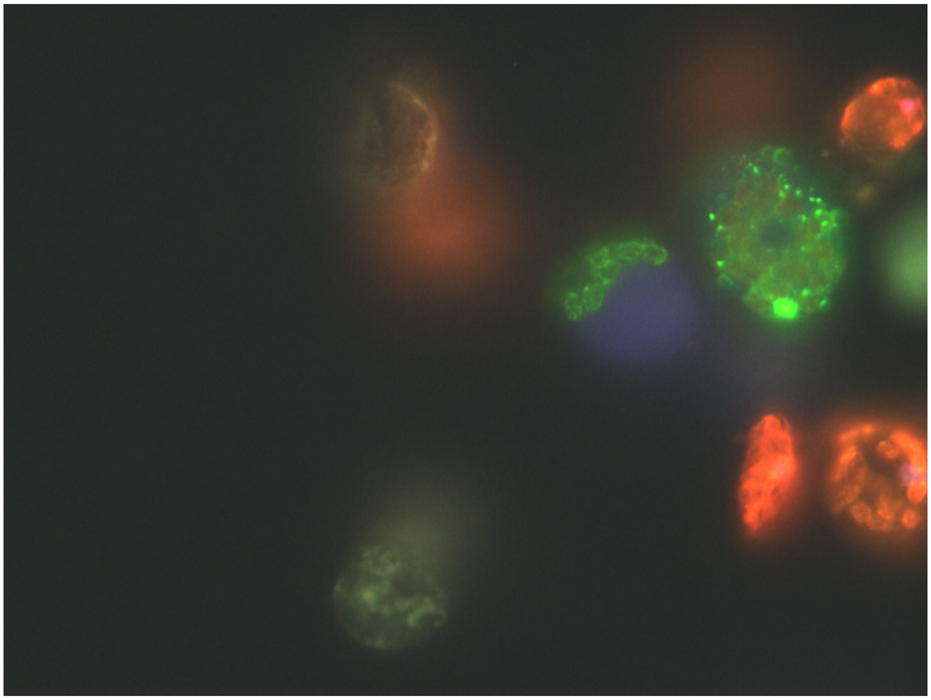


Figure 32: This false color image shows a sample image from the set of protoplast images collected. Rab-F2a (green), DAPI (blue) and chloroplast autofluorescence (red) are shown. Rab-F2a is involved in protein transport and is normally found in the membrane or surface of the vacuole.

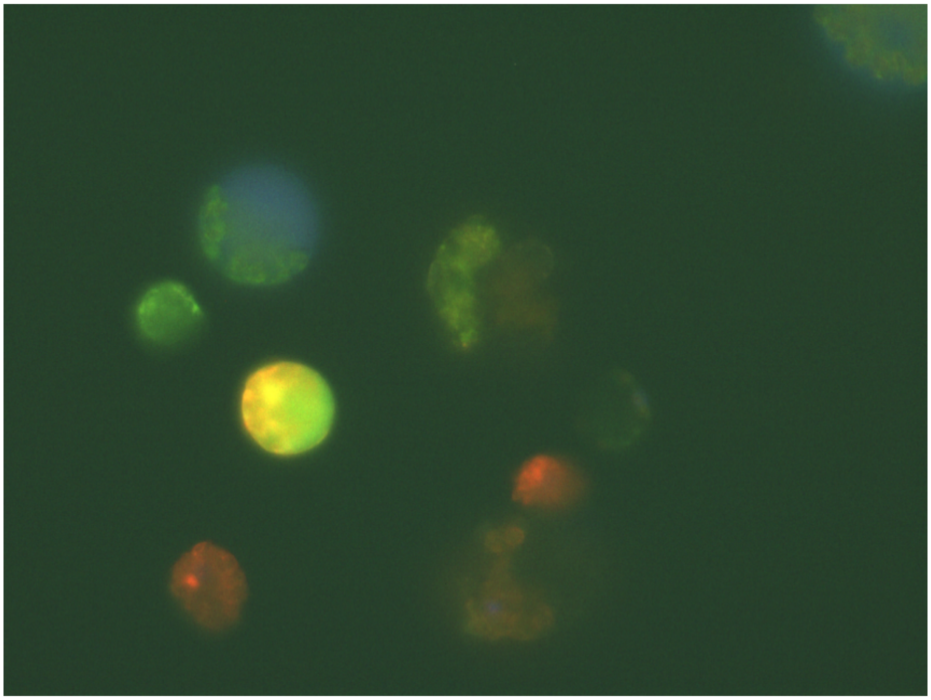


Figure 33: This image was selected from a well that contained cells tagged for Rab-F2a and treated with Tyrphostin. An image of untreated cells taken from the same plate is shown in Figure 8.

Hough Detected Protoplasts

It was hypothesized that the lack of focus in the entire field contributed to the difficulty in identifying cells very well using the Hough transform. The features calculated for these images were heavily dependent on the spatial distribution of the detected fluorescence in each image. Because of the lack of focus for some fields that resulted from the highly variable size of protopalsts, these sorts of features some information may have been lost.

One reason for choosing the mixture model based approach was to have the ability to apply a biological interpretation to the results if necessary. By visualizing the clusters of cells for the K and circle confidence chosen, an expert could rather easily interpret the effects of these drugs. However, using our model selection criteria, a K of 3 was selected. With a K of 3, each cluster appeared to be composed of cells with diverse visual appearances. An alternative model selection method could improve this issue. The underlying assumption in model selection that all lines were different from one another may have been too strong an assumption to use with this particular method. Using minimal expert intervention, one could select a model which separates the controls into the three visually similar groups: talin; ER-GK, Rab-F2a; M4, 313 YFP, AHA-ATPase.

The final conclusions resulting from the mixture model based approach did not indicate that any differences were statistically significant. It is possible that this could have been remedied by taking more images, but the fact that some effects were very close to being statistically significant using the classifier error-based assessment, which used the same protoplast data would indicate that the problem actually lied in the model selection.

PhenoRipper-Based Mixture Models

The results from the PhenoRipper based methods indicated that there were significant effects detected. There is a significant difference between the Hough transform based feature calculation and those of the PhenoRipper method. This is that the spatial arrangement of the colors within a protoplast makes a significant difference in the feature calculation, whereas in the PhenoRipper calculations, the spatial distribution of colors within a patch does not make a difference in the calculation. Because of the variation in the size of cells and the effect on focus within images, this may have made the PhenoRipper method more suitable for the task of recognizing drug effects as the results were likely much less affected by images being out of focus.

Confirmatory Efforts

As a result of these efforts further imaging experiments have been executed to confirm these results. These were executed using confocal microscopy, likely allowing for a more thorough assessment of the location patterns using feature calculations based on the spatial distribution of the tagged protein within the cell relative to the reference channels. It would also be interesting to try to induce a phenotype change in talin using a larger dose of Oryzalin or possibly nocodazole.

Chapter 3: Active Learning for High Content Screening Using **Arabidopsis thaliana** *Protoplasts*

One fundamental weakness in much active learning work has been the demonstration of utility in the real world. Many publications have relied on simulations to demonstrate the success of active learning processes. In addition, many active learning simulation papers have utilized experimental data that is already in the final form that will be utilized by the active learner. The primary difference between this work and that of standard active learning papers is that when the "oracle" is queried for answers in standard active learning, answers are given. In this case, the "oracle" is queried and data is given implying that the act of processing that data is built into the active learning process in some capacity. In the case of high-content screening processes, it is possible that novel phenotypes may be discovered and the presence of these phenotypes may force changes in phenotype assessment results. In this chapter, the results from a partially completed active learning-driven HCS campaign are described as well as simulations demonstrating performance for completion of the remainder of the available experimental space.

Protoplast Preparation

Protoplasts were prepared as described in Chapter 2. Protoplasts were generated from six cell lines expressing different proteins tagged with GFP (AHA-ATPase, ER-GK, Talin, 313-YFP, Rab-F2A, M4). Three concentrations of nine drugs (benzylphosphonic acid, Brefeldin-A, Damnacanthal, endothall, N9-isopropyl olomoucine, oryzalin, tyrphostin, ZM-449829, media + vehicle) were used. This resulted in a true experimental space of (27 drugs + 1 control) × 6 lines or 168 experiments. It is hypothesized that across all proteins there will be some proteins that show correlated effects from the treatment with the same drugs and some drugs that will have correlated effects when treating the same proteins. If this is not the case, it would be impossible to understand the entire experimental space without testing exhaustively. Because we wanted to demonstrate the potential of an active learning process without testing a very large number of targets and drugs, we chose to duplicate the experimental space along the drug and tagged protein axes. This resulted in an experimental space of 12 proteins × (54 drugs + 1 control) or 660 total experiments.

Two active learning threads were initialized. For each round, a set of experiments was selected. On the day prior to the selection of experiments and plating of cells, protoplasts were manually prepared from fresh *Arabidopsis* plants. Cells were allowed to recover overnight. On the first day of the round, cells were plated by hand into 60 wells of a 96 well plate. Of these 60 wells, 12 were untreated controls and the remaining 48 were duplicated experimental conditions. Because protoplasts are very fragile, robotic plating of the cells was not an option. After cell plating, drugs were robotically added to the wells in the order in which they were to be imaged. The timing of the addition of the drugs was matched to the timing of the imaging so that total exposure time would be synchronized across all experiments. The two active learning threads were executed until they were approximately 12% and 25% complete. At this point, experiment execution timing became a significant issue and data was more rapidly gathered using a random selection method to fill out the remainder of the unduplicated experimental space with at least two wells of every condition.

Active Learning Method

Active learning was performed using an active learning method developed by Armaghan Naik. This method utilizes discrete phenotype information to describe the experimental space. This method detects correlations between proteins across different conditions and correlations between conditions across different proteins. Utilizing this correlation information, predictions can be made using imputation. An example of such imputation would include the situation in which Target A under Condition D is observed to be the same as Target B under Condition D. If we then observed Target A under condition C, we could predict that Target B under Condition C would have the same phenotype. This active learning method attempts to build a model around clusters of targets and conditions. It then chooses experiments that are expected to improve this model.

Image Analysis

For this active learning method, a discrete label was required for each experimental condition. The first step of the image analysis process was identical to that used in Specific Aim 1. A Hough transform was used to identify protoplasts. Features were then calculated for these protoplasts, which were then filtered and clustered using K-means and a threshold on the circle confidence from the Hough transform. A mixture model was formed at each round to describe all experimental conditions observed up to that point in the active learning process.

Phenotype Clustering

We utilized a hierarchical clustering method to determine the phenotype labels. In each round of clustering an entire hierarchical clustering tree was formed. At each point in which two cluster labels were merged, the product of the distance between the two clusters and the total number of members in each cluster was calculated. The clustering was stopped at the step prior to the step that had the highest product avoiding combining the observations that are the most distant and the most numerous.

Active Learning Simulations

All images from both active learning threads and the additional thread used to fill out the experimental space were divided into wells. The wells were then divided as evenly as possible into experimental conditions (i.e., protein-drug pairs). For some unduplicated conditions, only two wells worth of images were available. Because the experimental space was duplicated along both the proteins and drugs, there were some duplicated conditions for which there were no experimental wells (114 duplicated conditions missing resulting in 83% coverage of the duplicated space). These experimental conditions were left empty for the duration of the simulations. During the actual active learning threads, a batch size of 24 was utilized and this was maintained for all simulations.

In order to assess our active learning process, the final model and its associated predictions were considered to be the ground truth. However, as the rounds progressed, the phenotype labels for a single experimental condition were likely to change from round to round as new conditions were added to the pool of data used in clustering. As a result a comparison that ignored the value of the resulting labels was needed. In order to calculate the equivalence of a clustering, all inter-observation relationships were considered meaning that if they have the same label in one clustering, an equivalent clustering would also have them in the same cluster but it may not have the same label. This is illustrated in Figure 34 with an example clustering compared to two alternative clusterings. One of these clusterings is a perfect match in terms of recovering label information while having different labels. The other is an imperfect match and the agreement calculation is shown for both. The agreement calculation is the mean of the proportion of must link relationships recovered and the proportion of cannot link relationships recovered.

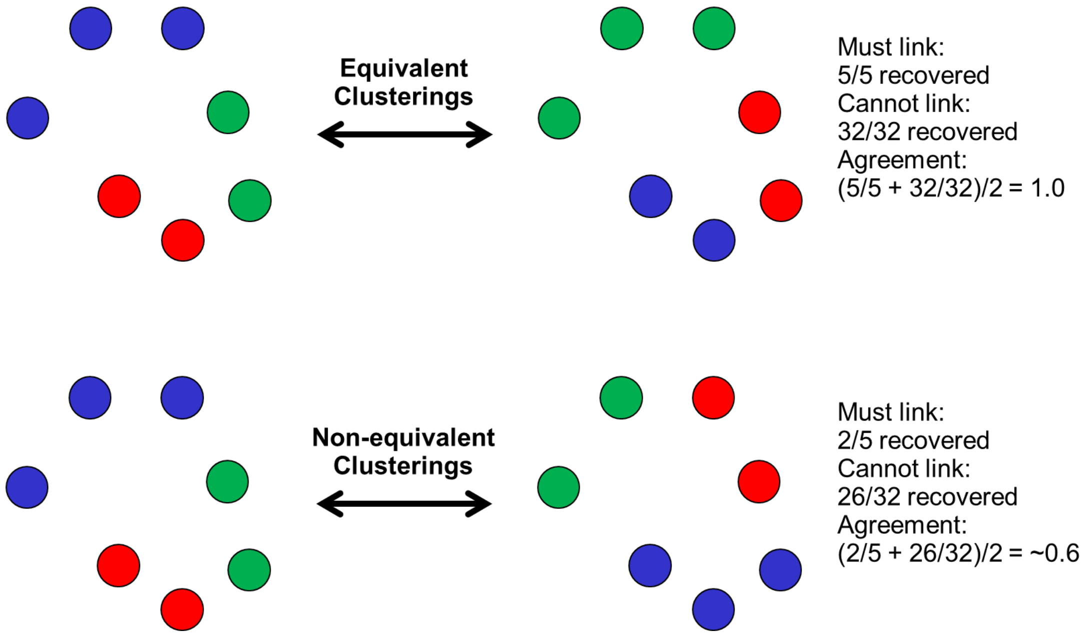


Figure 34: One clustering is compared to two possible clusterings. One of which is a perfect recapitulation of the label information but with different labels. The other is imperfect and the agreement calculation is shown for both.

Because we artificially duplicated the experimental space, we also calculated how well the hidden duplications were recovered. In a round, for each pair of duplicated drugs, all pairs of predictions were compared for equivalence. The fraction of matching pairs was reported as the duplication recovery.

In order to test the utility of the active learning method, multiple simulations (>10) were run in which the initial starting set included only the control wells. From this initial set, a model was learned using hierarchical clustering to determine the final phenotypes. Regardless of the model used to generate the phenotypes, all simulations used the predictions resulting from the active learning model. At each step, the current model predictions were compared to the final model. The results are shown in Figure 35. For both traces, the phenotypes were determined and the predictions were made using identical methods, but the difference between these two traces lies in the selection method used. There was no statistical difference between the resulting traces.

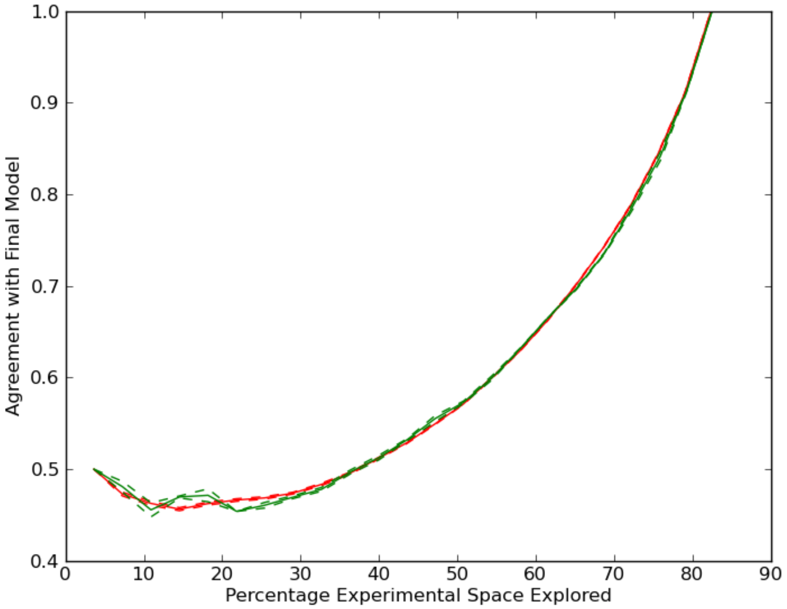


Figure 35: The mean and standard error are shown for the resulting agreement between simulated active learning runs using hierarchical clustering to determine the final phenotype labels. The green line indicates the performance using random selection and the red line indicates performance using active learning selection. There was no significant improvement using active learning.

The ability for the model to recover the line and drug duplications at each round was measured and the results are shown in Figure 36 and Figure 37. In both cases, using active learning to select the experiments resulted in an improvement in the duplication recovery rate.

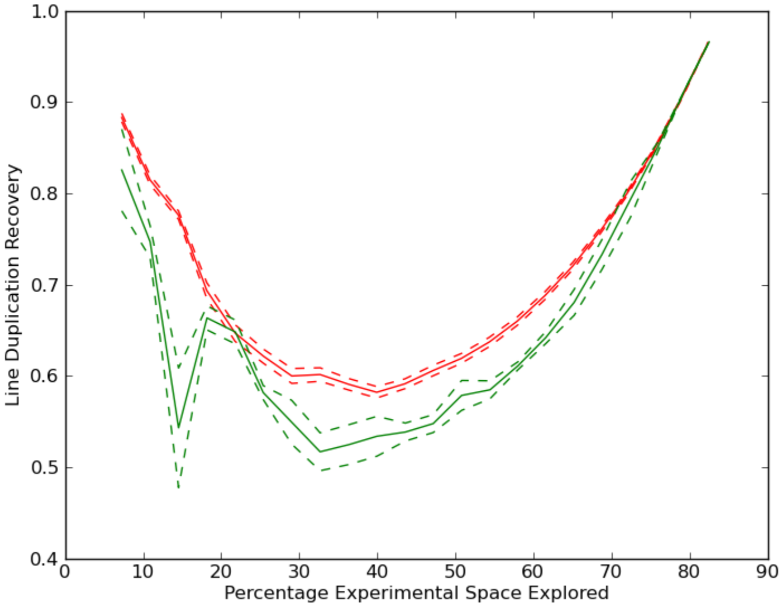


Figure 36: The mean and standard error of the line duplication recovery is shown as a function of experimental space exploration is shown. The red line indicates the recovery while selecting experiments using active learning and the green line represents recovery using random selection.

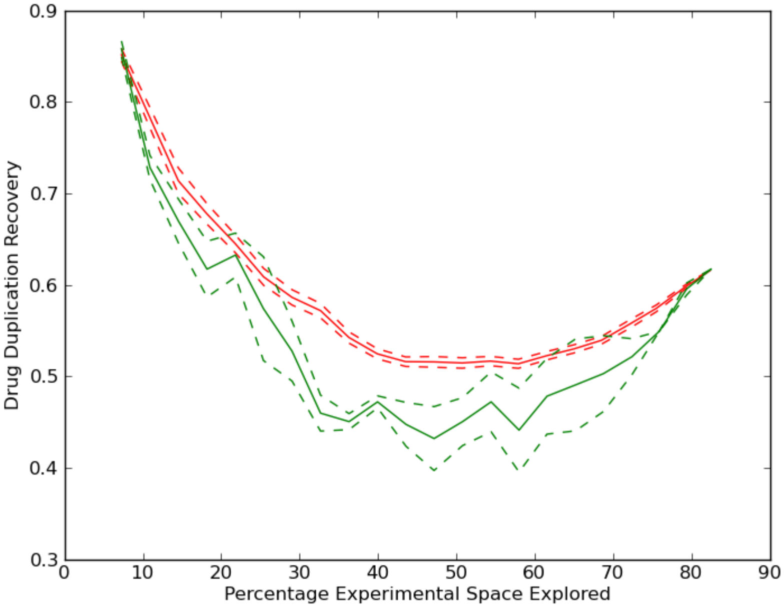


Figure 37: The mean and standard error of the drug duplication recovery is shown as a function of experimental space exploration is shown. The red line indicates the recovery while selecting experiments using active learning and the green line represents recovery using random selection.

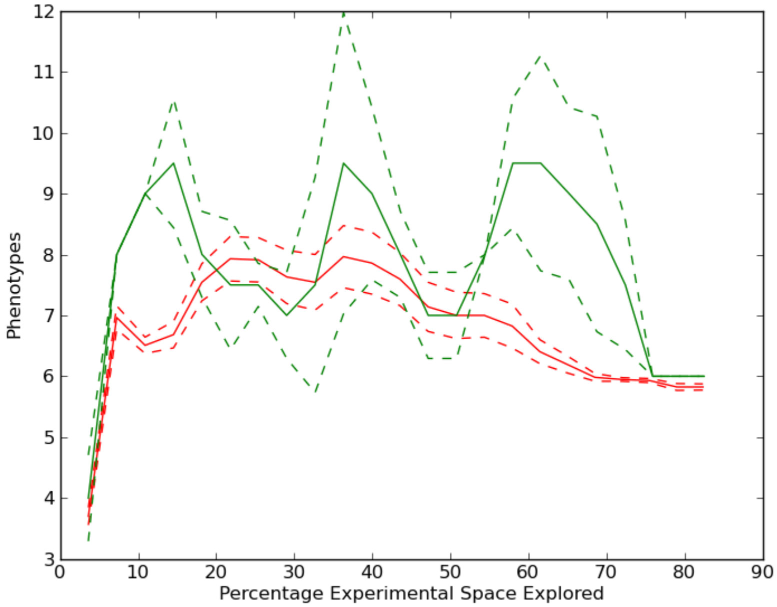


Figure 38: The mean and standard error of the number of phenotypes detected as a function of experimental space exploration is shown. Active learning selection is shown with the red line and random selection is shown with the green line.

Discussion

There was no improvement in the agreement with the final model that resulted from using active learning to select experiments. There are a number of possible reasons for this.

Phenotype Detection Model

It is important to fully understand what the duplication recovery metrics signify. A method which said that every single phenotype for all experimental conditions was identical would have perfect recovery so these measures alone alone cannot tell the whole story.

The mixture model appeared to have sub-optimal performance. One would have expected that if the model was working well, it would have easily identified the line duplications in the first round as all lines were observed in their control conditions. A perfect model would have continued to recognize these duplications as the simulation progressed and the duplication recovery would never have changed. In this case, the line duplication recovery dropped dramatically indicating that it was detecting differences between lines when no such differences should have been detected based on the fact that the data used by the model to make these assessments came from experiments with identical experimental conditions. At the end of both traces, the line recovery rate was at nearly 100% meaning that most line duplications were recovered with complete data which is a positive sign. Additionally, the number of phenotypes at the end of each trace (Figure 38) was approximately six indicating the recovery of line duplications was not attributed to just a single phenotype being detected.

A similar argument can be made for the drug duplication recovery. Because there was no requirement that every drug be tested at least once for a single line, it could take significantly more experiments to guarantee that all duplications would be detected for a perfect model. In this case, even with complete data, the drug duplication recovery was poor at the end of each trace.

Active Learning Method

A discrete experimental space composed of lines and drugs such as this has a set of parameters that can be used to describe it: drug uniqueness, line uniqueness and responsiveness. Line uniqueness refers to how similar the lines are to one another in terms of their affectability as a result of drug exposure. Drug uniqueness refers to how similar the drugs are to one another in terms of what targets lines they can affect. The less unique a system is, the easier it is to learn an accurate predictive model. The more unique a system is the more difficult it is to learn an accurate predictive model. This simulation was utilizing an experimental space which was less than or equal to 50% unique because of the duplications. There is a range of uniqueness in which the active learner used can improve the prediction accuracy relative to that achieved by the same model trained with randomly selected experiments. Responsiveness is the frequency with which a line is affected by a drug. This active learner has been shown to improve prediction accuracy for low ranges of uniqueness and relatively high ranges of responsiveness. There is a risk in the area of drug discovery that the responsiveness may be far too low to see a significant difference between active learning based selection and random selection.

An additional issue with this active learner is that it does have good performance characteristics when the phenotypes appear to be very noisy in nature. Due to the nature of the prediction method used by the active learner, many simulated threads were terminated after exploring only 20-30% of the experimental space. This correlates with a significant drop in the drug duplication recovery rate. The results shown are for the simulated traces that were able to complete (most likely the simulations with the best model accuracy for both random selection and active learning based selection).

Potential Improvements

One possible reason for the problems with the model predictions could be a result of having to split the data up to fulfill the duplicated conditions. It is conceivable that more fields would have helped to give more stable phenotype assessments. Additionally, the phenotype assessment model could have been improved by using the mixtures from a PhenoRipper-based method. With an iterative approach to high-content screening, one must decide whether to analyze new images once or reanalyze the images every round. Using the PhenoRipper method in the first round of a simulation generates a set of superblock definitions that can be used to rapidly process any new images to determine their superblock mixtures. The risk of this approach is that if a new phenotype is observed, but it is measured using old superblock definitions, the new phenotype may not be detected as such. For that reason, it could be advantageous to rerun the entire PhenoRipper pipeline to analyze every image for every round. For these simulations, per round PhenoRipper assessments would have taken a substantial amount of time. For that reason, the mixture model was chosen for use in these simulations.

Chapter 4: Prediction of Biological Responses Using Protein and Compound Features and their Discovery using Active Learning

Introduction

Drug development is a lengthy process that begins with the identification of potential drug targets and ends after testing in clinical trials. The targets are generally identified through basic science studies as being critical components of processes believed to be affected in a disease. Once a target has been identified, the goal is to identify drug-like compounds that either increase or decrease the activity of a target protein. High throughput screening (HTS) is a common way to ascertain the effects of many compounds on a single protein.

The first step in an HTS campaign is to develop an assay that detects the effects of compounds on the target protein. Multiple types of experimental assays can be used for HTS. These screens are commonly performed by adding compounds to a protein in a microtiter plate or to a well of live cells. A plate reader can be used to measure absorbance, fluorescence or luminescence as a reflection of target activity (Inglese, et al., 2007). Alternatively, flow cytometry can be used to measure changes in cellular populations that result from the addition of compounds (Sklar, et al., 2007). Lastly, fluorescence microscopy can be used to detect changes in target localization (Trask, et al., 2009) in a process sometimes referred to as high-content screening. The development of liquid handling robots has made it possible to test millions of compounds in a single pass.

Even with automation, exhaustive high throughput screening can be prohibitively expensive. One approach to reducing the need for experimentation is to generate a model for compound effects *in silico*, a process referred to as virtual screening. There are two common methods (Patani & LaVoie, 1996). During a quantitative structure activity relationship (QSAR) analysis, molecules are checked for the presence or absence of specific structural elements. The vectors describing the molecules are referred to as a "fingerprint." QSAR methods have been used to make predictions about the activity of compounds on target proteins (Kearsley, et al., 1996) (Sheridan, et al., 1996). Molecular docking is an alternative method that requires knowledge of the structure of both target and compound (Lengauer & Rarey, 1996) (Huang & Zou, 2010). Computer simulations are run in which the target and compound are forced into contact and the interaction energy between the target and compound molecule are then estimated. These methods take into consideration features of the target protein and potential drugs. Beyond virtual screening, efforts have also been made to apply machine learning techniques to the wealth of information available in the PubChem database, paying particular attention to the gross imbalance of active to inactive compounds (Han, et al., 2008) (Li, et al., 2009).

Once a compound has been identified as having an effect on a target, the compound may be advanced further along the drug development pipeline. It is not uncommon for previously unknown effects to be discovered after significant investment in a potential drug. These are not discovered earlier because HTS processes are primarily designed to detect the effects of compounds on a single target protein without considering whether or not it affects other targets. One solution to this problem would be to test every compound ($\sim 10^6$) against every possible protein target ($\sim 10^4$). This would require an estimated 10^{10} experiments, the cost of which would likely be prohibitive. An obvious alternative is to use some method to choose a subset of possible experiments that is expected to provide sufficient information to make decisions. This is the province of machine learning approaches termed active learning, which have been specifically developed to permit efficient exploration of large experimental spaces. While active learning is widely used in some fields, its application has been limited in biological problems (Tong & Koller, 2001) (Pournara & Wernisch, 2004) (Liu, 2004) (Stegle, et al., 2009) (Danziger, et al., 2009) (Mohamed, et al., 2010).

Active learning consists of three phases performed in a loop (as illustrated for the work described here in Figure 39). A campaign of experiments can be initialized either using prior results from literature or databases or by randomly selecting a batch of experiments from an experimental space. **(1)** A model is generated to represent the currently available data. **(2)** From that model, experiments are selected for execution that are expected to improve the model. **(3)** The set of experiments is executed and the resulting data are combined with previously collected experimental data. The loop then continues from Step 1 until either a desired accuracy of predictions is achieved or a specified budget has been exhausted. There have been limited previous applications of active learning to the drug discovery process. In these efforts, compound activity was considered to be binary (active or inactive) and effort was focused on only a single target (Warmuth, et al., 2003) (Fujiwara, et al., 2008).

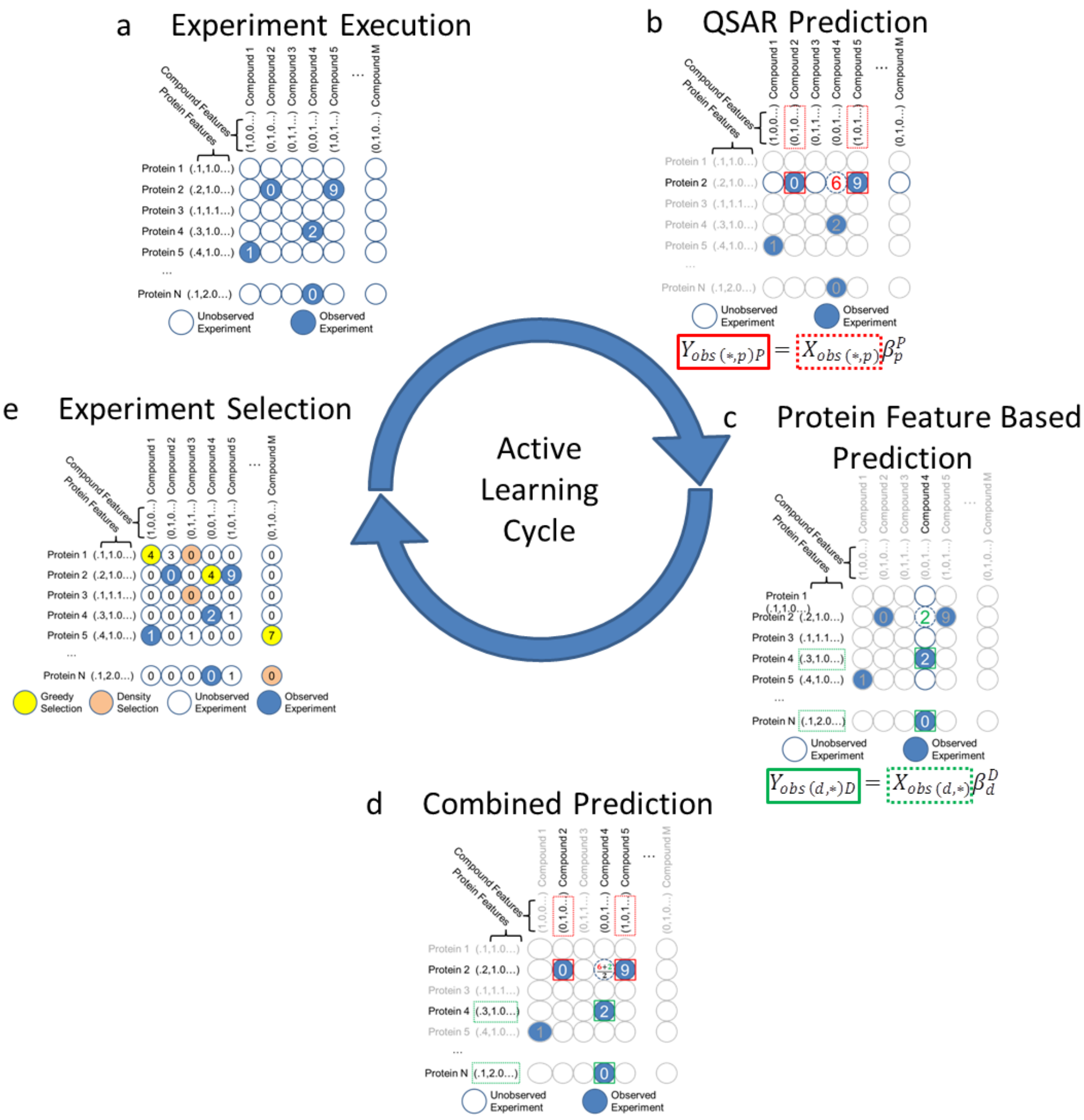


Figure 39: An active learning pipeline is shown for an experimental space with N proteins and M compounds. (a) For each round of active learning, the results from a new set of experiments are observed and added to the current set of data. (b) A model is constructed using the compound features to make predictions for experimental results. (c) A model is constructed using target features to make predictions for experimental results. (d) For the dual regression approach, two predictions are usually made and combined (as shown for Protein 2, Compound 4). (e) Observed experiment values, values predicted from the model, and experiments that would be chosen for the next round of acquisition by different methods are shown.

Here, we designed two models to make predictions about activities for large numbers of combinations of compounds and targets. Our model uses features developed for virtual screening to describe compounds, and features from sequence analysis to describe target proteins. Most importantly, we investigated the utility of applying active learning in combination with these models in order to efficiently discover active compound-target pairs. In tests using data from the PubChem database, we found that active compound-target pairs could be discovered as much twenty-four times faster using active learning than by random selection of experiments.

Results

Dataset

To evaluate our proposed approaches, we chose to use existing experimental results for assays on many targets and many compounds. We therefore began by assembling a large set of compound effect scores from PubChem (http://pubchem.ncbi.nlm.nih.gov). In total, compound activity scores for 177 assays were assembled. Of these assays, 108 were from *in vitro* assays and 69 were from *in vivo* assays. During collection, assay scores were adjusted so that they ranked from -100 (strong inhibitory effect) to 100 (strong activation effect) with scores of zero implying no effect. Of the 600,000 compounds in PubChem across the 177 assays, an average of 30% had a reported activity score for a given assay. (We do not know but assume that the missing values are approximately missing at random.) Of these, we created a dataset of all assay data for 20,000 randomly-chosen compounds, resulting in a system with 3.5 million possible experiments. All combinations of target and compound with scores above 80 or below -80 were marked as hits. (Note that each PubChem assay includes its own rank score cutoff above which a chemical is considered to be "active". Our cutoff of 80 is more stringent than that used for most assays.) Because it is difficult to know what is most relevant to measure about a compound or a target, we calculated many features for each. For each compound, 1559 fingerprint features were calculated. For each assay, 388 features were calculated from the amino acid sequence of the associated protein. These features did not have to be perfect descriptors of molecular properties, but only to reflect aspects of similarity among compounds or proteins.

Model Definition

As an initial approach to constructing a predictive model, we explored using linear combinations of features. Given the large numbers of features involved, lasso regression (Tibshirani, 1996) was used because it allows for efficient feature selection for linear regression models. We note that while the assay scores may be non-linearly related to true activity, and while estimates of true activity may be obtained by further manipulation or testing, we expect them to be good approximate predictors of which combinations of compounds and targets will show high activity.

Three approaches to prediction of the assay scores were used. The first approach used all compound features to predict the activity of each compound in a given assay (analogously to QSAR). Using lasso regression, compound features were selected that were strongly indicative of the activity of a compound on a *single* target. A regression model was learned for each individual target allowing for the selection of compound features unique to a target (Figure 39b). The second approach used all target features to predict the effect on each target of a given compound. When considering all experiments that involved a single compound, lasso regression allowed us to select features of the target protein that were indicative of the likelihood for a target to be affected by that single compound (Figure 39c). The third approach made a combined prediction by averaging the two predictions for each compound-target combination (Figure 39d). We refer to this approach as the *dual regression model*.

Evaluating Model Performance

We first sought to determine how accurately these models could predict target-compound hits as a function of how much training data was available. To do this, we randomly sampled a sequence of experiments one at a time until 3% of the experimental space had been sampled (note that each combination of assay and compound was considered independently when selecting random experiments). As each experiment was sampled, we combined it with all previous experiments from that sequence to train a model and evaluated its ability to predict hits for all remaining data.

A receiver-operator characteristic (ROC) curve was calculated for each of these models by varying the classification threshold to predict a hit (note that only the prediction threshold was varied; the definition of an actual hit as having an absolute value above 80 was unchanged). Finally, the area under the ROC curve was calculated for each sequence. This process was repeated ten times for each of the three prediction approaches described above. The means and standard errors of the area under the ROC curve for the ten trials for each prediction approach are shown in Figure 39. Two methods were used to generate random predictions for comparison. In the first method, scores were randomly chosen from the set of all scores for all assays. Because these were globally random, the area under the ROC curve was expected to be 0.5. Predictions using all methods performed better than this expectation. For the second method, scores were randomly chosen from those for all compounds for a given target. The predictions from this sort of random predictor were expected to be more accurate than randomizing across all observations (since targets with a lot of hits will be randomly predicted to have a lot of hits), thus it is a more stringent standard for comparison. Predictions using drug features alone or using dual regression performed better than random by this standard. This is despite the fact that less than 0.1% of the combinations were active according to our definition.

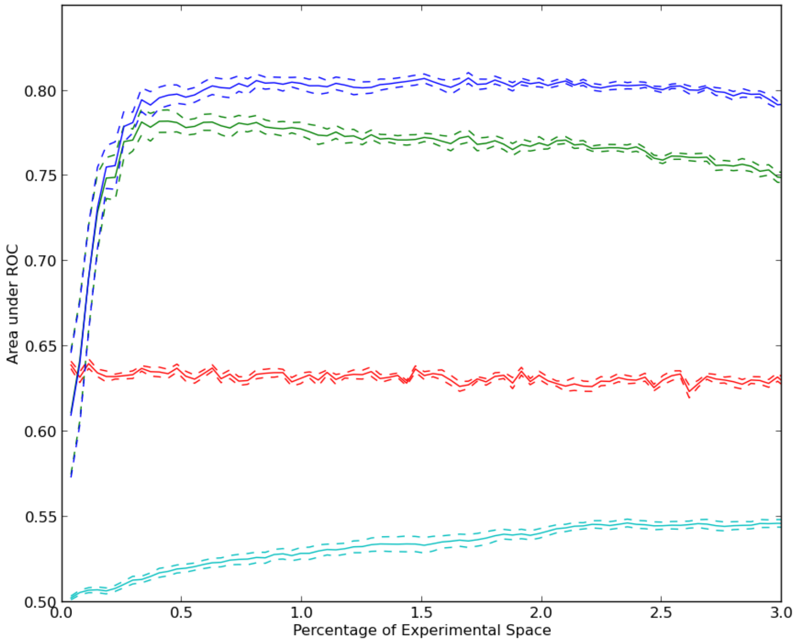


Figure 40: Area under ROC curve for prediction of positive experiments with increasing amounts of random training data. Ten random sequences of experiments were used to select data used for training regression models. After each experiment was chosen, a ROC curve was constructed by gradually raising the threshold on the predicted assay score at which an experiment was considered to be positive. The mean and standard error of the area under the ROC curve after each experiment is plotted for each regression method. The prediction methods shown are within target random prediction (red), regression using target features only (magenta), regression using compound features only (teal) and dual regression (green).

We also considered which features were more informative than others. To make a single set of predictions across the entire space of 20,000 compounds and 177 targets requires the training of 20,177 lasso regression models. The final models trained at 3% of the experimental space (from Figure 40) were analyzed and the proportion of models where the coefficient for each feature was non-zero was calculated. To determine the magnitude of the effect of a feature on prediction, the mean absolute coefficient for each feature (only when it was selected) was calculated. For targets, the most frequently selected features (and those with the largest coefficients) were the amino acid compositions. For compounds, the most frequently selected feature was “Group IIa (Alkaline earth)” and the feature with the largest absolute coefficient was “4M Ring”. Further details about features that were most frequently selected are found in Table 4 and Table 5.

Table 4: The five most frequently selected compound features are shown with their source, description, selection frequency and mean Beta when selected using lasso regression.

Source	Content	Selection Frequency	Mean Absolute Beta
MACCS	4M Ring	0.49	0.70
MACCS	Group IIa (Alkaline earth)	0.58	0.64
MACCS	CSN	0.48	0.62
MACCS	QAAA@1	0.17	0.61
MACCS	OACH2A	0.31	0.58

Table 5: The five most frequently selected target features are shown with their source, description, selection frequency and mean Beta when selected using lasso regression.

Source	Description	Selection Frequency	Mean Absolute Beta
ProtParam	AAComp-(A)	0.62	3.31
ProtParam	AAComp-(E)	0.60	1.23
ProtParam	AAComp-(D)	0.55	1.92
ProtParam	AAComp-(H)	0.54	0.76
ProtParam	AAComp-(C)	0.53	2.85

We also were interested in how applicable a trained model would be to a new target or a new compound. We utilized the same random sampling approach described above. However, for each of the ten trials, the experimental results were held out for a unique 10% of all targets or compounds and a ROC curve was calculated for only the held out experiments. The results (Figure 41) show that when holding out entire compounds or targets, relatively accurate predictions can be made about activities from the remainder. As expected from the results in Figure 40, the predicted activities for held out compounds are more accurate than those for held out targets. Both, however, perform better than random prediction (AUC = 0.5). The results confirm that the regression approach can capture important information about compound effects, even when no information about a compound is provided during training. The fact that scores could be predicted better for new compounds than for new targets may be due to the fact that data was available for many more compounds than targets (and thus there is a higher chance that the model has already seen a similar compound versus a similar target).

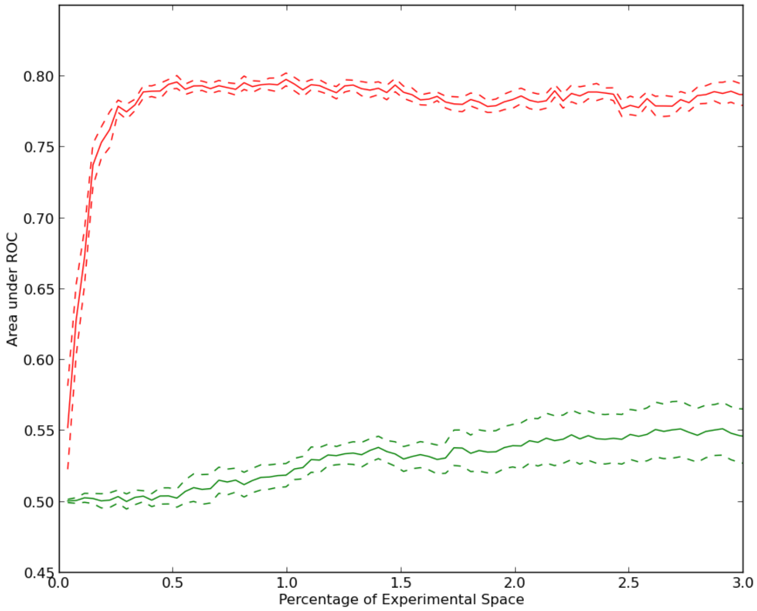


Figure 41: Area under ROC curve for prediction of positive experiments in held out targets or compounds with increasing amounts of random training data. The same approach as in Figure 40 was used, except that only predictions for held out compounds (red) or targets (green) were considered and only dual regression was used. The mean and standard error of the area under the ROC curve after each experiment is plotted.

Active Learning Simulation

Given that our modeling approach performed better than random at predicting relative activity scores, we next determined whether it could be used to successfully drive an active learning process (i.e., to find hits faster than expected at random or to rapidly maximize predictive accuracy). For this, simulations were run for an experimental space of all 177 assays

(129 unique gene targets) and all 20,000 compounds. For this experimental space, rank scores were available in PubChem for 1,043,300 experiments out of 3,540,000 possible experiments. Simulated experiments were restricted to those for which results were available; requests from an active learner for other experiments were skipped.

To initialize a simulation, all experimental results were hidden from the active learner and 384 experiments were selected randomly for "execution." During the execution phase (Figure 39a), results from selected experiments were "revealed" and used for training of a predictive model (Figure 39b-d). A new batch of experiments was then selected using one of a number of active learning methods (Figure 39e). Finally, the data for the selected experiments were added to the pool of previously selected data and the loop continued until 3% of the possible experimental space was explored. Each round consisted of the selection of 384 experiments. Ten separate simulations were run for each method, each starting out with a different set of initial experiments. At each round, the discoveries (combinations whose absolute activity score was greater than or equal to 80) were counted, and the mean count and associated standard error were recorded as a function of the fraction of experimental space so far explored.

We first considered a greedy active learning approach in which unobserved experiments that had the greatest predicted effect (inhibition *or* activation) were selected for measurement in the next round. This greedy approach was used in combination with dual regression, single regression with predictions from compound features for each protein target and single regression with predictions from protein target features for each compound. For comparison, a random selection method was also included. As shown in Figure 42, the greedy dual regression method performed best. After exploration of 3% of the experimental space, an average of approximately 38% of possible discoveries were made. Results for the single regression approaches are also shown. As might be expected from the results in Figure 40, results for prediction from target features only are nearly the same as for random selection. Results using compound features only are much better, but not as good as for dual regression.

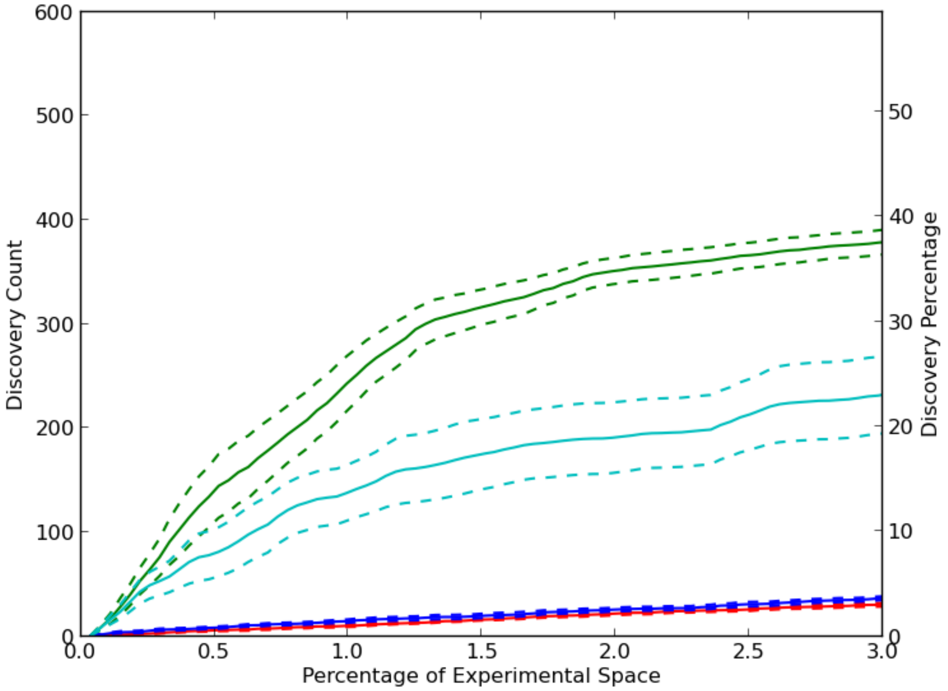


Figure 42: Active learning to discover compound-target hits. The average number of discoveries and standard error for 10 separate trials are shown. The methods were random choice (red), dual regression with greedy selection (green), single regression using only compound features for prediction (blue) and single regression using only target features for prediction (cyan).

The rate of discovery for the greedy method using dual regression decreased as the simulations progressed. Exploration of the experimental space with the greedy algorithm was limited to regions of the feature space that were predicted to have large activities. We considered the possibility that this limited the system's ability to learn a better model, and that this could be overcome by acquiring data in regions where few observations have been made or where the model predictions are uncertain. Therefore, a "density-based" approach was also tested that selected experiments so as to explore the experimental space efficiently without regard to predicted values or experimental results. In this approach experiments were tested that were most similar to unobserved experiments and least similar to observed experiments (Fujii, et al., 1998). A variation on this idea, diversity sampling, was also tested, along with uncertainty sampling in which experiments with the highest uncertainty of their prediction are selected. Results for these approaches are shown in Figure 43. The uncertainty-based selection method performed much better than random but not as well as dual regression with greedy sampling. Density-based and diversity-based sampling performed similarly to random selection. These three classical active learning methods are generally designed to select experiments for execution that will yield the most accurate model, while the results in Figure 43 are for finding hits. We therefore considered the accuracies of the models for each method by calculating the area under the ROC curve (as previously described for Figure 40). As shown in Figure 44, all selection methods, except for uncertainty sampling, resulted in an initial peak followed by a slight, gradual reduction in the accuracy of the models. The better performance of uncertainty sampling compared to dual regression with greedy sampling is consistent with the opposite result in Figure 43. This is because uncertainty sampling does not prefer finding hits over non-hits.

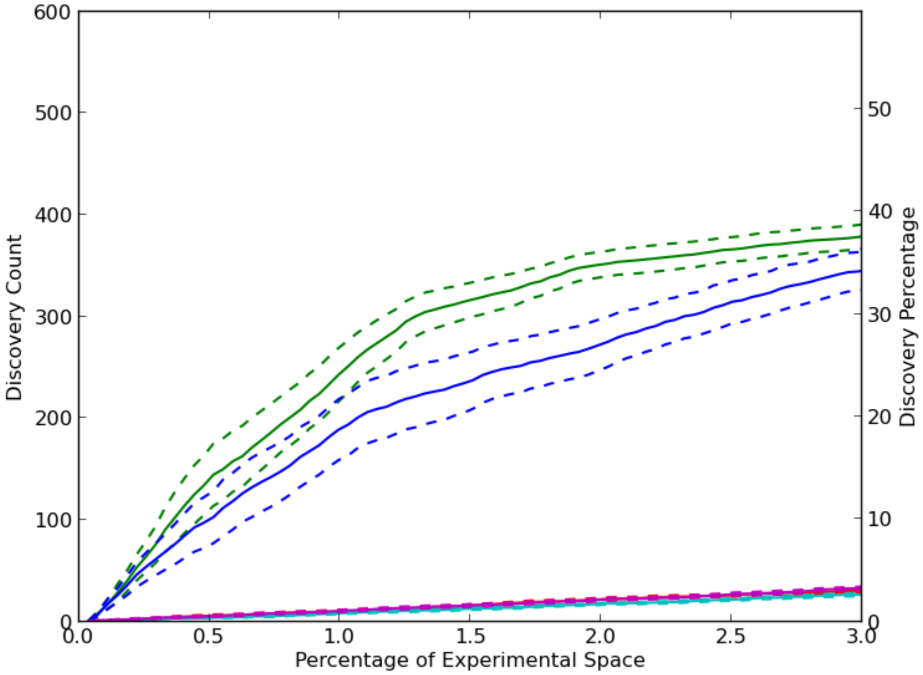


Figure 43: Evaluation of compound-target hit discovery for different active learning methods. The average number of discoveries and standard error for 10 separate trials are shown. The methods were random choice (red), dual regression with greedy selection (green), uncertainty sampling (blue), density-based sampling (teal) and diversity selection (magenta).

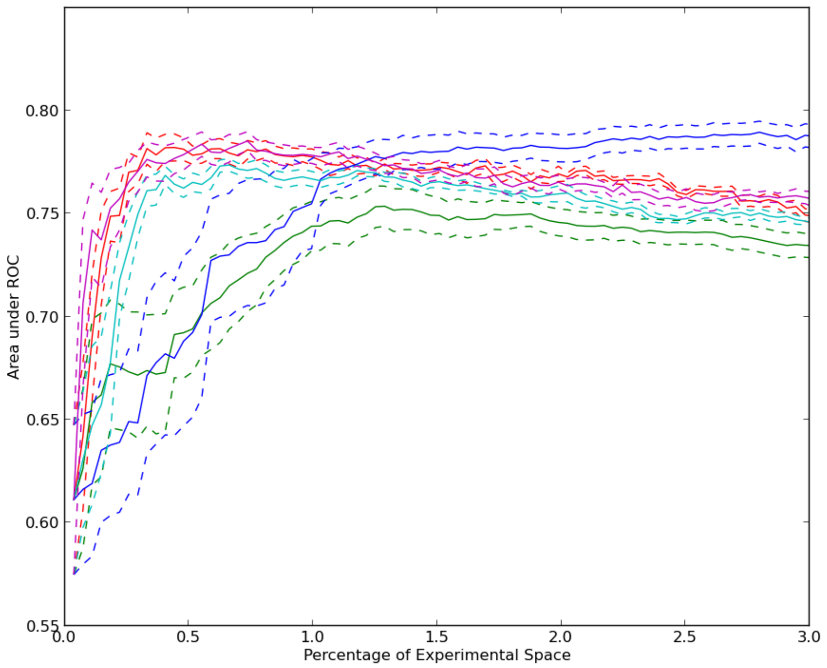


Figure 44: Evaluation of model accuracy for unobserved experiments for different active learning approaches. A dual regression model was trained using all experiments selected with each selection method. The mean and standard error of the area under ROC curve for each regression method is plotted. The selection methods were random choice (red), dual regression with greedy selection (green), uncertainty sampling (blue), density-based sampling (teal) and diversity selection (magenta).

Because uncertainty, diversity and density-based selection methods were designed to yield an accurate model we also tested hybrids of greedy dual regression with each of these methods. The hybrids with density and diversity performed worse than greedy dual regression by itself (not shown) but the hybrid with uncertainty performed slightly better (Figure 45).

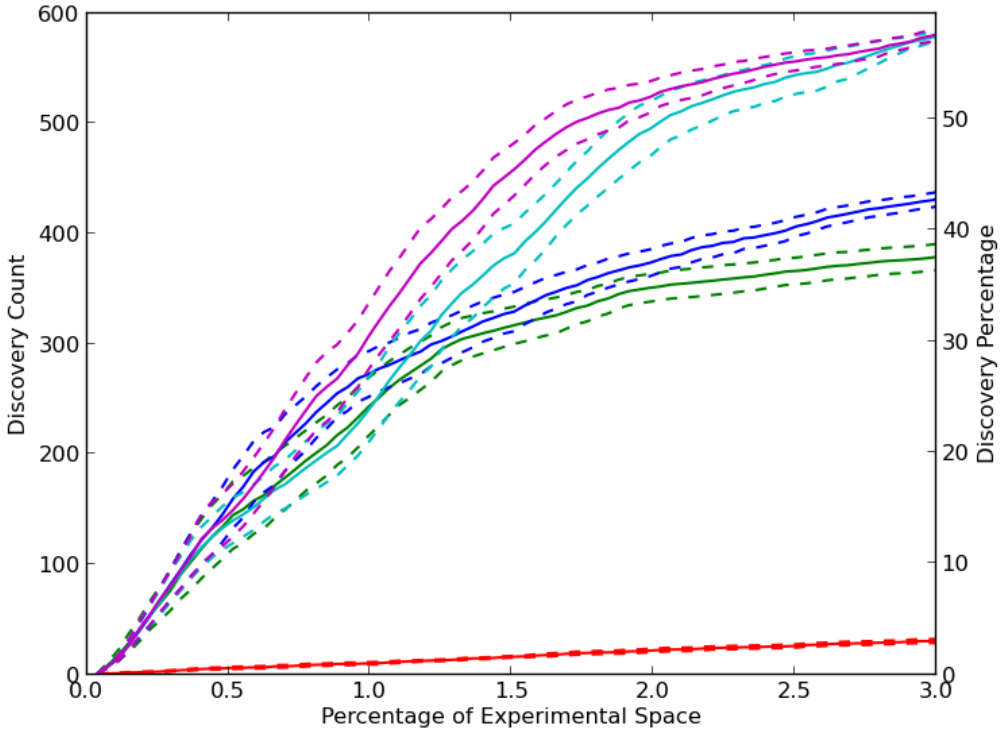


Figure 45: Improved active learning approaches' discovery rates. The average number of discoveries and standard error for 10 separate trials are shown. The methods were random choice (red), greedy dual regression (green), greedy dual regression-uncertainty hybrid (blue), and greedy selection-uncertainty hybrid using memory limits of five (cyan) and ten rounds (magenta).

We also considered the possibility that the decrease in rate of learning for greedy dual regression was due to excessive testing of a given target for new discoveries after all of them have already been made. To address this possibility, we developed a modified approach (which we termed "limited memory") in which only information from a given number of previous rounds was used in the model generation and active learning process. Any requests from the active learner for experiments previously selected and subsequently hidden were skipped. As shown in Figure 45, limiting memory to only the previous 5 or 10 rounds significantly improved the discovery rate. Almost 60% of discoveries were made after only 3% of the experimental space was explored. We also found that limiting memory in the context of hybrid uncertainty methods also improved the quality of the predictive model as measured by the area under the ROC curve in Figure 46.

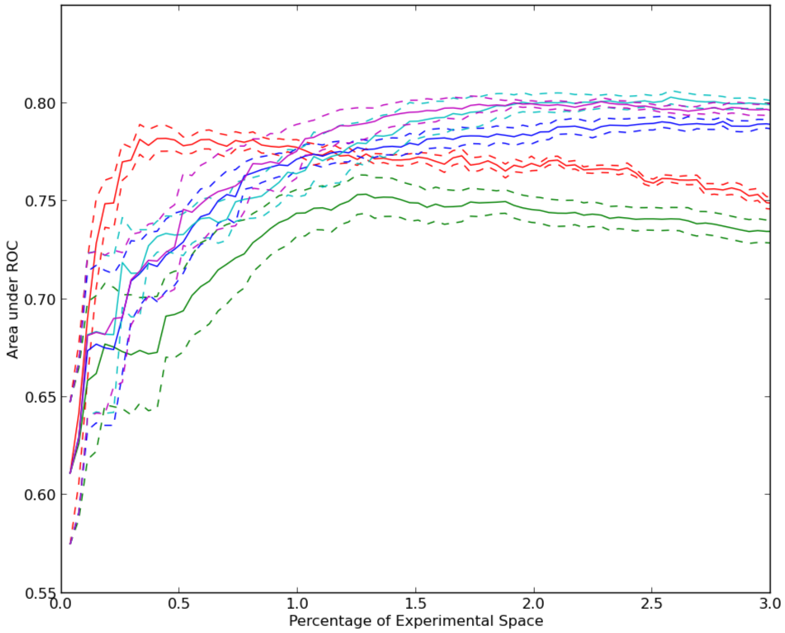


Figure 46: Improved active learning approaches' accuracy. A dual regression model was trained using all experiments selected with each selection method. The mean and standard error of the area under ROC curve for each regression method is plotted. The methods were random choice (red), greedy dual regression (green), greedy dual regression-uncertainty hybrid (blue), and greedy selection-uncertainty hybrid using memory limits of five (cyan) and ten rounds (magenta).

For reasons of computational time, we restricted our analysis to 20,000 compounds. It was therefore of interest to estimate how performance might change if more compounds were included. As a preliminary indication of this, we performed simulations for *smaller* sets of compounds. The results (Figure 47) show that the learning rate is significantly worse for 5,000 compounds than for 20,000, but that it is not much different for 10,000 than for 20,000. This suggests performance for larger sets might be similar.

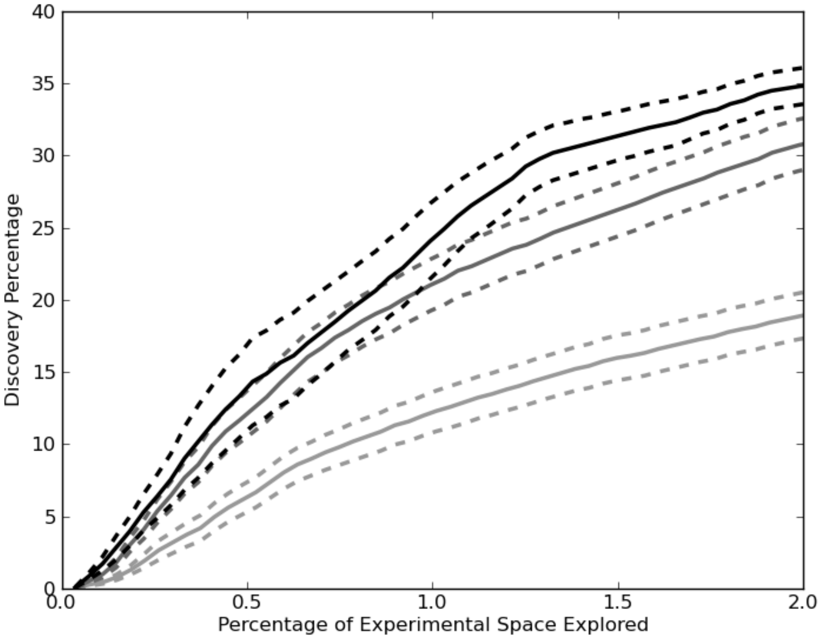


Figure 47: Discovery rates for different numbers of compounds. Simulations were run for multiple active learning methods for the exploration of 2% of the experimental space with various subsets of the compounds. For all runs, dual regression with greedy sampling was used. The average number of discoveries and standard error are shown (n = 10). The drug selection subsets were of the following sizes: 20,000 (black), 10,000 (dark gray) and 5,000 (light gray).

Discussion

We have described a pipeline for executing experiments driven by an active learning system and demonstrated that it can produce the rapid discovery of compounds that affect target proteins using a set of heterogeneous assays. We found that the selection of experiments based only on predictions calculated using compound features (predicting the effect of a compound on a single target) performed significantly better than the selection of experiments based only on predictions from target features (predicting the sensitivity of a target protein to a single compound). Decent performance of the prediction models using compound features is to be expected given past results with QSAR approaches to modeling compound activity on a given target. The comparatively poor performance of the protein models could be a result of multiple issues: poor features, limited data, and heterogeneous data sources. The system included only features that could be calculated from sequence information, and it is likely that this feature set could be improved by the inclusion of features calculated from protein structural information. Importantly, the addition of memory limitations to these models further improves the discovery rate. In this experiment, only information from 177 assays was used. As information from more assays becomes available, predictive models are expected to improve.

Our system could have two closely related purposes. First, this system could be used to predict and test for previously uncharacterized side-effects on other important target proteins. Compared to current high throughput screening methods, the rate of discovery of such effects would be expected to be greatly improved. Alternatively, a set of many target proteins could be constructed (e.g., for a number of diseases), and compounds that affect one or all of these targets could be rapidly discovered concurrently. The analysis of many targets and many compounds in a coordinated fashion increases the average number of experiments performed to find a compound that affects a single target, but decreases the average number per target.

The selection of an appropriate batch size is an important consideration for the utilization of an active learning system. Some experiments might be well suited for dynamic experimentation because a batch of experiments requires relatively little overhead cost and a short time is required to execute the experiments relative to computational time. In this case, a smaller batch size may be chosen.

It is important to note the differences between the approaches presented here and those described previously. In particular, they are distinguished from QSAR and virtual screening approaches by simultaneous consideration of many targets and many compounds, and from chemogenomics approaches by utilizing ligand similarity (Keiser, et al., 2009). However, the most important difference is the emphasis on active rather than passive learning.   We believe active learning will be particularly important as drug development efforts increasingly consider variation among cell types and among individuals. The size of this experimental space clearly precludes exhaustive experimentation.

Many variations on the approaches described here can be considered, including different feature sets and different active learning algorithms (such as information-theoretic scoring (MacKay, 1992) (Settles & Craven, 2008). An exhaustive evaluation of these variations is beyond the scope of this paper, but we have firmly established that significant improvement in learning rates can be achieved. The results also suggest that the paradigm of exploring combinatorial experimental spaces through active learning may be widely applicable in biomedical research beyond drug discovery. This includes any study that seeks to determine the effects of large numbers of perturbagens (such as compounds, siRNAs, or induced mutations) on large numbers of molecular, cellular or histological behaviors (such as enzyme activities, cell shapes or motility, protein expression or localization). As the size of the experimental space grows larger, the more impractical exhaustive experimentation becomes but the more improvement may be expected from active learning. These methods are particularly expected to be valuable for high content screening and analysis, such as for determining where all proteins are located in all cell or tissue types under many conditions.

Methods

Data Preparation

Each assay from the PubChem database (Bolton, et al., 2008) contains gene target information, chemical identifier information and activity scores for all compounds tested in the assay. Various features describing the primary structure of the target protein were calculated using ProtParam (Wilkins, et al., 1998), Protein Recon (http://reccr.chem.rpi.edu/Software/Protein-Recon/Protein-Recon-index.html) (Sukumar, et al., n.d.) and Prosite (de Castro, et al., 2006). In total the assays were described by 388 features. All non-binary features were were z-scored. The compounds in the assays were described with 1559 binary features calculated using OpenBabel (http://openbabel.org) (Guha, et al., 2006). Assays from PubChem targeting human proteins with more than 15,000 entries were manually annotated. For each assay, it was determined what type of effect was being detected for the target (inhibition, excitation, etc.) and the nature of the activity scores reported. Only assays whose activity scores were scaled with a measured effect from the compound were kept for simulation. In general, activity scores were scaled from 0 to 100. When scores were found above 100 in an assay, all scores in that assay were reduced by a constant factor such that the maximum score in the assay was 100. For all assays testing for inhibition, scores were made negative. From the ~600,000 possible compounds, 20,000 were selected randomly for use in simulations of the active learning processes.

Lasso Regression

Linear regression models were trained with the following equations where $Y_{obs(*,p)}$ and $X_{obs(*,p)}$ are the matrices of scores and compound features respectively from all executed experiments with protein p. The regression coefficients learned using lasso regression on the compound features to predict activity across target p are found in β_p^P. Additionally, $Y_{obs(d,*)}$ and $X_{obs(d,*)}$ are the matrices of scores and protein features respectively from all executed experiments with compound d. The regression coefficients learned using lasso regression on the target features to predict activity across compound d are found in β_d^D.

$$Y_{obs(*,p)P} = X_{obs(*,p)}\beta_p^P \tag{1}$$

$$Y_{obs(d,*)D} = X_{obs(d,*)}\beta_d^D \tag{2}$$

Lasso selects a set of features that gives a fit where $|\beta| < s$. The penalty s was selected using cross validation for each linear regression model. Once a model has been trained, predictions about single experiments was made with the following equations:

$$Y_{(d,p)P} = X_p\beta_p^P \tag{3}$$

$$Y_{(d,p)D} = X_d\beta_d^D \tag{4}$$

A combined prediction for $Y_{(d,p)}$ was calculated by taking the mean of the predictions from Equations 3 and 4.

$$Y_{(d,p)} = (Y_{(d,p)P} + Y_{(d,p)D})/2 \tag{5}$$

All regression models were trained using the Least Angle Regression method (Efron, et al., 2004) implemented in SciKits (http://scikits.appspot.com). Penalties were tested between 10^{-4} and 10^{4}. Penalties were selected which minimized the mean squared error of five-fold cross validation within the training data.

Greedy Selection Algorithm

Experiments were selected which had the greatest absolute value of predicted rank score. In some cases, no information was available to make a prediction for an experiment. If no prediction could be made from available data for an experiment, that experiment was predicted to have a rank score of zero. All experiments with equivalent values were treated in random order.

Density-based Selection Algorithm

Each experiment (target, compound) was represented by a single feature vector formed by concatenating the target features and the compound features for that experiment. For computational efficiency, a maximum of 2000 observed and 2000 unobserved experiments were used. Among the two thousand unobserved experiments, selections were made using a density-based sampling method (Fujii, et al., 1998) which attempted to choose experiments which were most distant from already observed experiments.

Uncertainty Sampling Selection Algorithm

For each unobserved experiment, predictions were made using five subsampled training sets for each model. Twenty-five 25 predictions were calculated for each experiment by calculating the mean of each compound prediction with each protein prediction. If a model was impossible to calculate because of a lack of common observations, only five predictions were used. Experiments were selected which had the largest standard deviation of predictions.

Diversity Selection Algorithm

Each experiment was represented by a single vector formed by concatenating the target features and the compound features for that experiment. A random set of 4000 experiments was clustered using the *k*-means algorithm (with *k* being the size of the batch desired, in our case 384). The experiment nearest to each centroid was selected for execution.

Hybrid Selection Algorithms

For each round, half of the experiments were selected using one method and half were selected using another method.

Chapter 5: Automated Generation and Identification of Randomly Tagged Cell Lines for use in an Active Learning Pipeline

RandTag Project

In order to begin to understand the function of a protein within the cell, it is useful to determine where that protein can be found within the cell. To investigate the location patterns of many proteins, the RandTag project (Garcia-Osuna, et al., 2007) was initiated to randomly tag as many proteins as possible using CD-tagging (Jarvik, et al., 1996). A guest exon coding for GFP was inserted into the genome of NIH 3T3 cells using a retrovirus. This allowed for the endogenous expression of stable GFP-tagged proteins. An extensive pipeline was developed to culture these tagged cells for imaging, sequencing and further experimentation. Automation was utilized to rapidly and reliably perform many of the steps in this pipeline. All of this effort resulted in a set of cell lines with tagged proteins that is being utilized in a large active learning process to test for the effects of drugs on protein location pattern.

Manual Tissue Culture

During the infection process, a retrovirus was used to insert a guest exon into the genome of NIH 3T3 cells. If the guest exon is not inserted into an intron, it is highly unlikely that GFP will be expressed. For this reason, this process has a high failure rate. One consequence of this high failure rate is that it was highly unlikely that more than one protein would be tagged in any single cell. Additionally, the result of an infection process was that most cells were not affected. After infection, the cells were allowed to recover which gave the cells time to begin expressing the GFP tagged-protein. After sufficient time (~2 days), the cells were sorted using a flow cytometer. Each tagged cell expressing GFP was sorted into a single well in a 96-well plate. Any cells not expressing GFP (or not at a sufficiently high level) were discarded. The sorted cells were allowed to grow up for 4 – 5 days. In spite of the parent cells being genetically identical to one another, the addition of the GFP tag sometimes significantly affected the viability of the cells. Some cells died in the 96-well plate, some grew slowly and others flourished. After sufficient time, the "fast growing" cells (defined qualitatively) were plated into a 96 well imaging plate along with imaging control lines (mitochondria, nuclear, plasma membrane and endoplasmic reticulum tagged lines in addition to untagged parental lines). The newly infected lines were concurrently plated into a separate 96-well plate for freezing prior to sequencing. Prior to imaging, the media was replaced with optical media and Hoechst was added to the wells. These steps were manually executed at the outset of the RandTag project.

Automation of Tissue Culture

The first major improvement to the RandTag pipeline was the implementation of automated tissue culture methods using an Eppendorf epMotion Liquid Handling Robot. This robot is computer controlled, allowing users to develop protocols using a graphical interface supplied by Eppendorf. Using this interface, one could design protocols by first setting up the robot deck with the necessary equipment and then by adding action commands to the protocol. For each command there were numerous parameters to set such as the aspiration and dispense speed as well as the pattern to be used for motions involving equipment utilizing multiple wells (i.e. multi-well plates, multi-channel reservoirs, tube racks). Seven tools were available for manipulating experimental materials: 50 uL, 300 uL and 1000 uL single and eight-channel pipettors as well as a gripper arm. The gripper could be used to move plates around the robot deck as needed. For example plates that needed to be warmed could be moved to a block referred to as the Thermomixer, which could warm, cool and mix objects placed on that location.

One of the first tasks was to use the robot to change media prior to imaging. After significant testing, optimal parameters for the removal and addition of media to 96 well plates were determined such that cells were not disturbed in the media change process, but the protocol execution itself was sufficiently fast to make the automation worthwhile.

The second major task was to learn how to pass cells using the robot. These simple tasks by themselves did not prove to be the most useful tasks for the RandTag project but they taught us the appropriate parameters to use when executing these steps as part of more substantial protocols.

Automated Protocol Script Generation for EPMotion Robot

In order to design and execute a protocol using the EPMotion, we needed to use a GUI and manually program each individual step of our protocol. In the software supplied with the robot, there was some functionality that allowed the user to define patterns based on a number of samples giving more flexibility to the user for large protocols with many sequential steps between the same starting plate and ending plate. Once a protocol was designed, it was saved to the hard drive for later use. When the protocol was to be run, it was simply loaded and executed. During the actual execution of the robotic protocol, there is no opportunity for dynamic human involvement making it impossible to make changes to a protocol in the middle of execution.

In the case of the plating of cells for imaging, we wanted to be able to tell the robot which wells to plate for imaging and sequencing. We then desired to plate the cells from each well based on their growth rates into three separate plates: imaging plate, sequencing plate and backup plate. The protocol is illustrated in Figure 48. For every good well of cells, there were at least eight unique steps that would need to be programmed differently for every plate. It became obvious that we needed to work around the user interface.

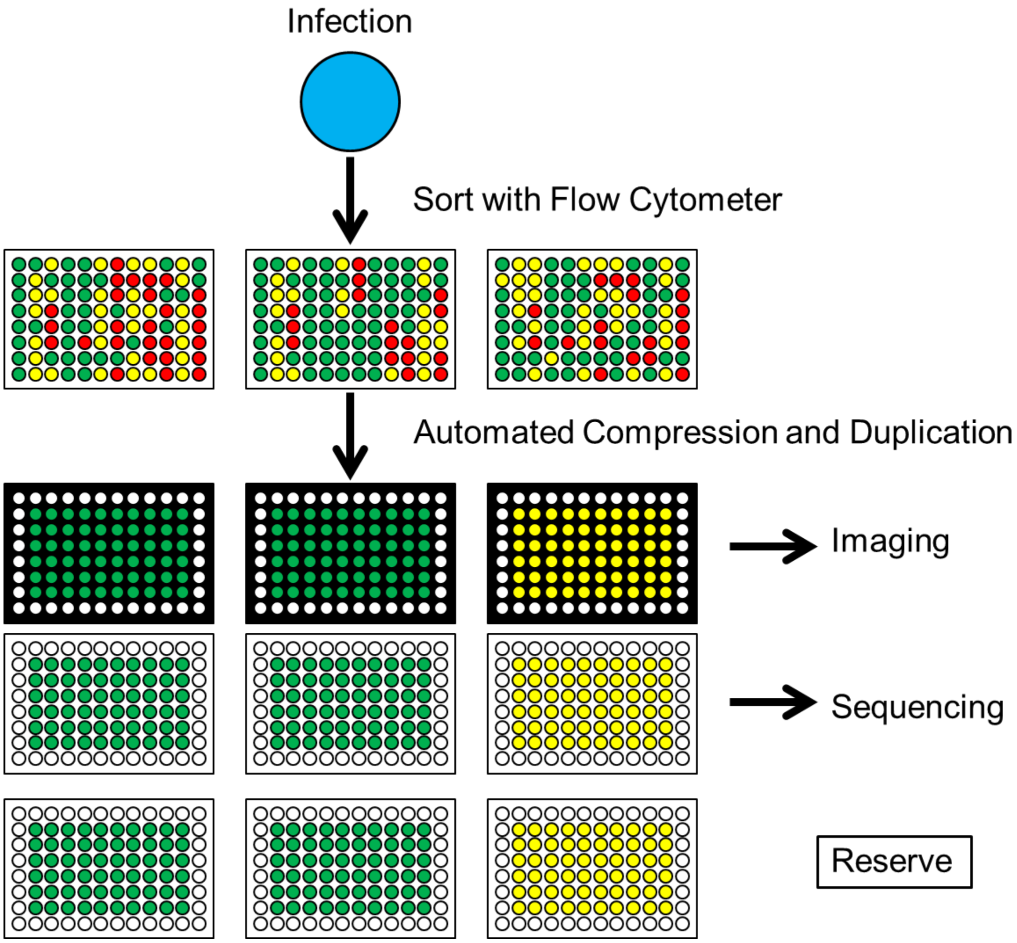


Figure 48: The tissue culture pipeline for the RandTag project is illustrated. After infection, cells were sorted using a flow cytometer. Single cells which expressing GFP were sorted into individual wells in 96-well plates. After a few days of recovery time, different growth rates were observed for the cells. Some cells did not recover at all (red), some grew slowly (yellow) and some grew rapidly (green). The cells from fast and slow growing lines were plated into three identical plates in batches of 60 allowing for the addition of controls to imaging plates. One plate was kept as a reserve in the event that we desired to continue working with a cell line after sequencing or imaging.

In order to implement this procedure using the robot we reverse engineered the protocol files generated by the Eppendorf software. We then designed a robot scripting system that allowed us to programmatically build custom protocols rapidly without using the GUI for the protocol design. We then designed a script such that it would take as input a list of wells and their associated growth rate labels. This script generated protocols that would efficiently plate only the selected cells from the source plate allowing cells in remaining wells to continue growing. The desired cells were plated at identical concentrations in identical positions in the three needed plates: imaging plate, sequencing plate, reserve plate.

Sequencing of RandTag Clones

The RandTag clones were generated through the introduction of a guest exon coding for GFP into an intron of a gene. To determine which protein was tagged, we needed to sequence the genomic DNA around the insertion site. This task was performed using a Splinkerette based method (Devon, et al., 1995) as illustrated in Figure 49. The challenge here was to amplify DNA outside of the region of the insertion site. First, the genomic DNA was digested using DpnII. DpnII was selected because its recognition site is four base pairs; we therefore expect a cut site on average every 256 base pairs, so it is unlikely that the reads will be too long for Sanger sequencing. Also, DpnII can be easily heat inactivated. Splinkerettes were formed by annealing two oligos together. When these two oligos were annealed together, the shorter oligo resulted in the creation of an overlap that matched the cut site for DpnII. On the other end of the oligo, a hairpin loop was formed which was not amplifiable. The longer oligo contained the three primer sites (Splink1, Splink2 and Seq). The splinkerettes were ligated to the ends of the digested DNA. PCR was utilized to amplify between Splink1 and Nest1 within the guest exon to the outermost primer site in the splinkerette. During the first cycle, Splink1 primers cannot bind and thus there is no extension from the splinkerette. Once there is extension from the Nest1 primer, a Splink1 site is created which is then used for extension in the following cycle from Splink1 primers. Next, another PCR reaction was executed such that the amplified region was between Splink2 and Nest2, which were both situated internally to Splink1 and Nest1. Finally, the DNA was sequenced from the Seq primer site in the splinkerette using Sanger sequencing (Sanger, et al., 1977). This sequencing run resulted in a final sequencing read that included the sequence from the splinkerette, followed by a DpnII cut site, followed by the genomic DNA of interest. If the distance between the DpnII cut site in the insertion was not too distant from the next DpnII cut site in the genomic DNA, the LTR could be identified within the read as well. This information was used to confirm the quality of the sequencing read.

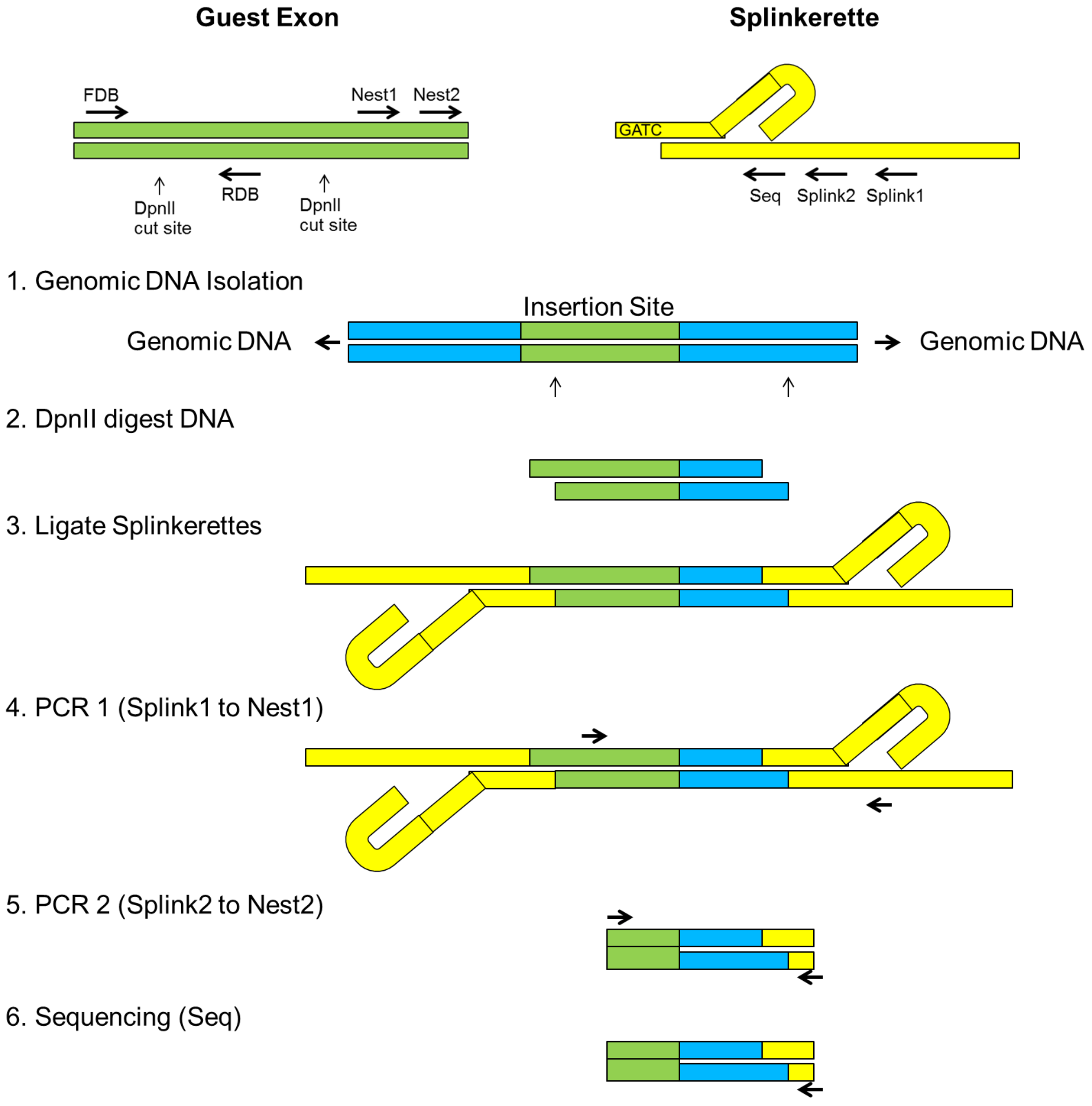


Figure 49: The splinkerette sequencing process is illustrated in order to show how the insertion site is identified. The first step is to isolate the genomic DNA, which is then digested with DpnII. Splinkerettes are ligated to the ends of the digested genomic DNA. DNA is amplified using PCR between a primer site (Splink1) in the splinkerette and a known primer site in the guest exon (Nest1). A second nested PCR amplification is done to further amplify the region of interest between Nest2 and Splink2. The DNA is sequenced from a final primer site (Seq) in the splinkerette for as long as possible.

Sequencing Protocol Optimization

We took advantage of the configuration of the guest exon to test each phase of our preparation for sequencing process. Two primer sites (FDB and RDB) were found within the guest exon; between them was a DpnII cut site.

Genomic DNA Extraction

Two protocols were tested in order to find an appropriate genomic DNA isolation method. We tested two protocols for the quality of the extractions. Because we were trying to execute this process in a high-throughput manner, we wanted the process to be well suited for 96 well plates. We also wanted the extraction process to be as simple as possible. We initially tried a new product called Evogen One (http://www.evogen.com/products/evogen_one.html) as it was claimed that this would be a one-step method for the extraction of PCR ready genomic DNA. We tested the quality of the extraction by assessing our ability to amplify the region of DNA between the FDB and RDB within the insertion site in tagged lines and that we could not amplify that same sequence in DNA extracted from untagged cells. We were unable to amplify the region of interest in the tagged lines, so we investigated other DNA extraction methods. The resulting gel is shown in Figure 50.

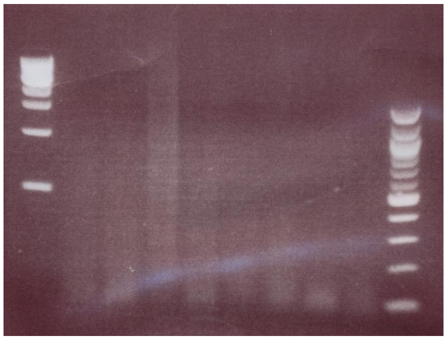


Figure 50: Evogen ONE was used to extract genomic DNA from tagged cells. PCR was run using FDB and RDB primers which should have resulted in a product from DNA from tagged cells. It is clear this failed.

We next tested a proteinase K genomic DNA extraction method. This was a standard proteinase K extraction reaction in which cells were lysed. Proteinase K was used to help to break down proteins during an overnight incubation period. The following day, an ethanol-NaCl DNA precipitation procedure was used. The resulting DNA was spun down and gently washed using 70% ethanol. Water was then added to the DNA in preparation for the initial steps of the sequencing process. Using samples of DNA extracted from tagged and untagged cells using the Proteinase K based extraction method, we ran multiple PCR reactions to amplify the region between the RDB and FDB. As expected, we were able to successfully amplify DNA within in tagged lines, but not in the untagged (Figure 51). We also determined that the product was of the appropriate length. We utilized this proteinase K extraction method for the duration of the project.

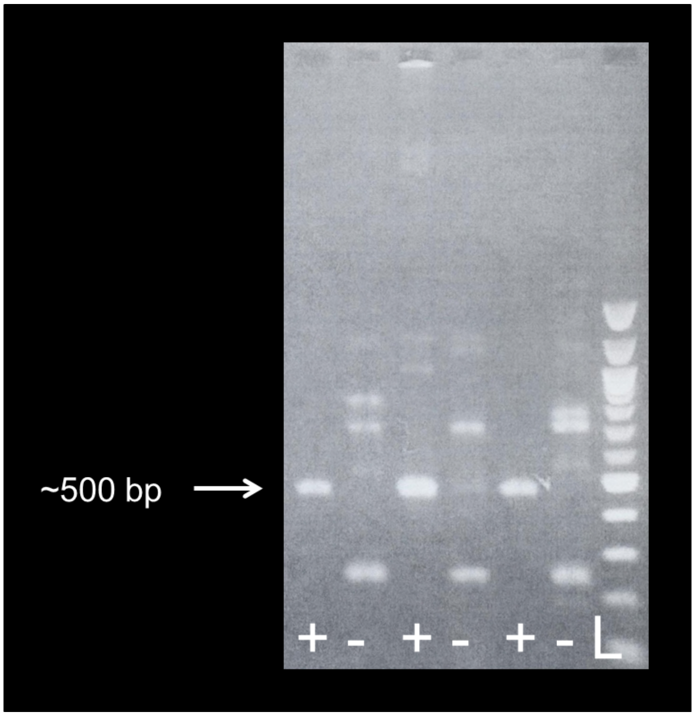


Figure 51: PCR with forward and reverse digest brackets should have resulted in a 512-bp product from genomic DNA from cells tagged with GFP. It appears that is the case; non-specific products were formed from DNA from untagged parental cells.

Digestion with DpnII

In order to test the digestion with DpnII, we extracted DNA and ran a digestion reaction with and without DpnII. We then ran a gel to compare the lengths of the resulting products. With DpnII in the digestion reaction, the lengths were much shorter than for the reactions without DpnII implying that the digestion process was successful.

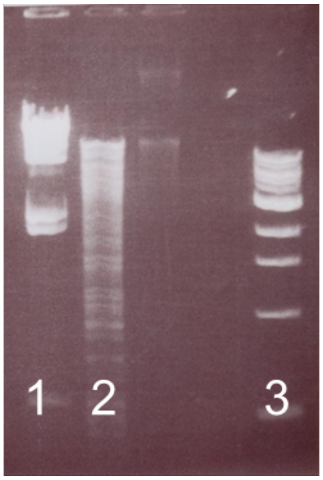


Figure 52: Digestion of genomic DNA with DpnII absent (lane 1) and present (lane 2) yielded smaller products (lane 2) according to the ladder (lane 3).

Splinkerette Ligation with T4 Ligase

In order to test the ligation reactions, we extracted genomic DNA and compared the length of the DNA prior to ligation with the length of the DNA following ligation. The DNA was longer after ligation, implying that the ligation reaction was working as expected (Figure 53).

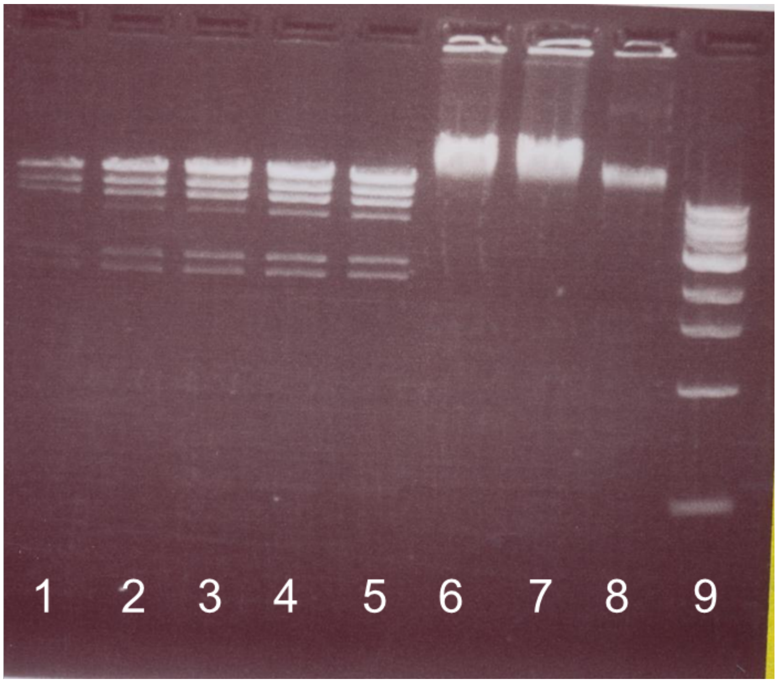


Figure 53: Ligation of genomic DNA (lanes 1- 5) yields larger genetic products (lanes 6-8) according to the ladder (lane 9).

Nested PCR Reaction Tuning

Prior tests (not shown) revealed that the PCR reaction protocol we were using seemed to be rather sensitive. We wanted to optimize two parameters of the protocol: $MgCl_2$ concentration and annealing temperature. In order to do this, we took advantage of the liquid handling robot and a PCR machine with a heating block capable of generating a temperature gradient. For both nested PCR steps, five concentrations of $MgCl_2$ were tested: 0.0 mM, 0.5 mM, 1.0 mM, 3 mM, 5 mM. We tested twelve annealing temperatures ranging between 53 ˚C and 65 ˚C. It appeared that the first nested PCR reaction was not sensitive to temperature, but was very sensitive to $MgCl_2$ concentration. The best concentration was 5mM $MgCl_z$. These results are shown in Figure 54.

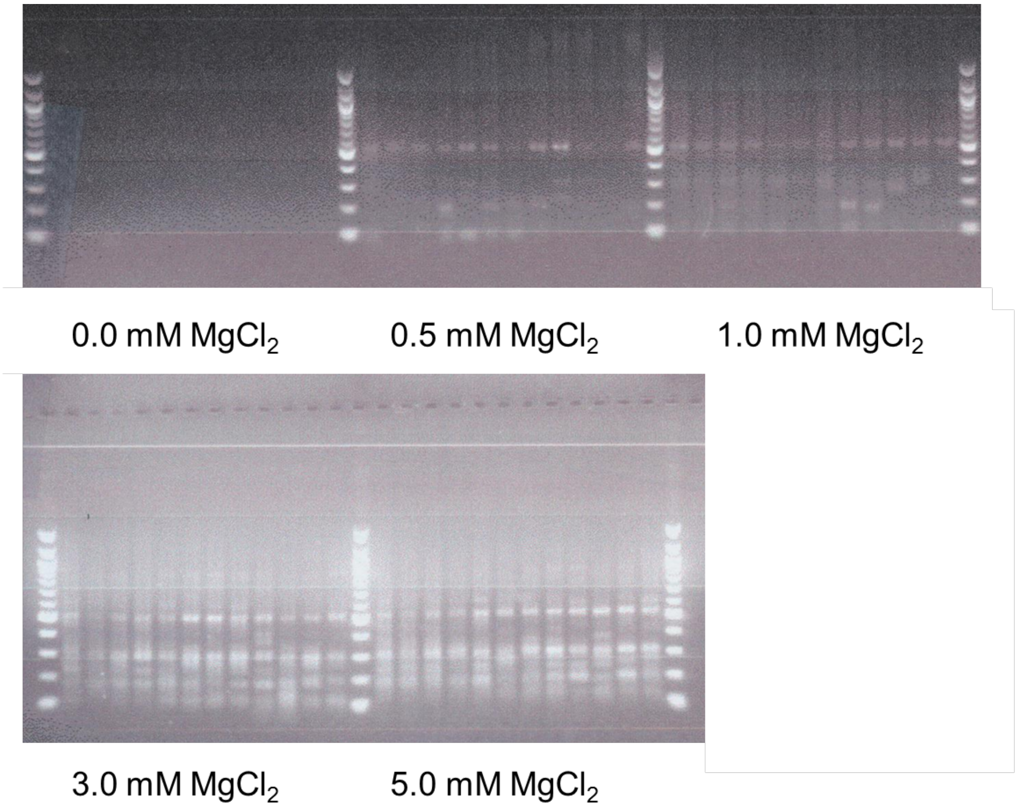


Figure 54: Results from tuning the first nested PCR step are shown for five different concentrations and twelve temperatures ranging from 53 ˚C to 65 ˚C from left to right in each block of lanes between the ladders. These results showed that the first nested PCR step is very sensitive to the $MgCl_2$ concentration.

Using the product from the reaction using 5mM $MgCl_2$ and 59 ˚C annealing temperature, a round of optimization for the second nested PCR step was tested for the same $MgCl_2$ concentrations and temperatures. In this case, there appeared to be a dependence on annealing temperature at lower concentrations of $MgCl_2$. In the second reaction it was found that the best concentration was 3.0 mM for $MgCl_2$ and there did not appear to be any significant dependence on the annealing temperature at that concentration so 59 ˚C was selected again. The resulting gel is shown in Figure 55.

Samples were cleaned up in preparation for sequencing by using EXO-SAP-IT to eliminate any remaining oligos. Finally, samples were transferred into 96-well plates with the appropriate oligo for sequencing from the final primer site (Seq) in the splinkerette. Samples were sent to GeneWiz (www.genewiz.com) for sequencing.

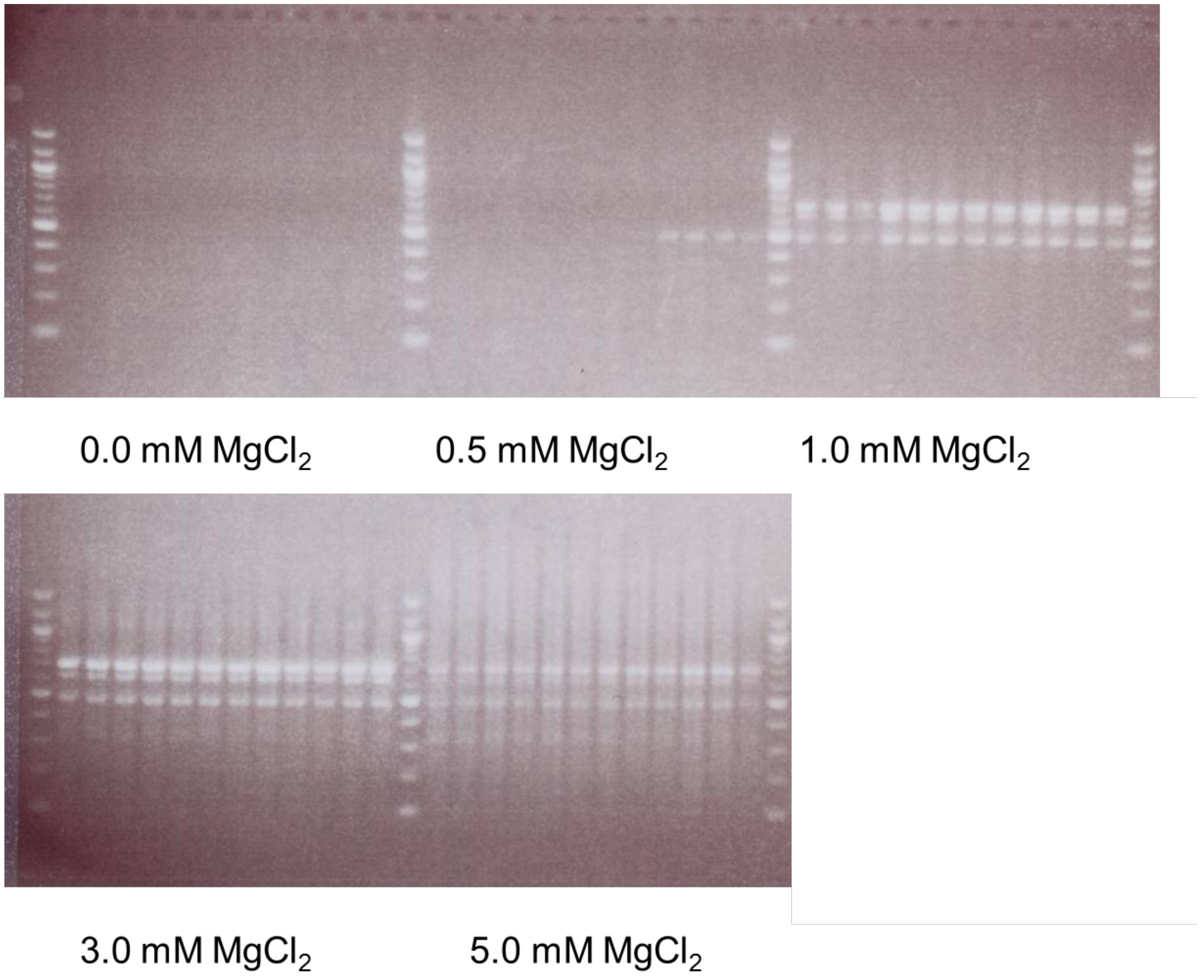


Figure 55: Results from tuning the second nested PCR step are shown for five different concentrations and twelve temperatures ranging from 53 ˚C to 65 ˚C from left to right in each block of lanes between the ladders. These results showed that the second nested PCR was more sensitive to annealing temperature at low concentrations of $MgCl_2$.

Sequencing Read Assessment

In order to maximize our ability to determine what proteins were tagged in each of the clones, we duplicated the ligation protocol for every plate. The pipeline was then completed for each ligation separately. The result was that every single well had two reads from the sequencing facility that were each assessed independently. Using BLAT and the associated tools from UCSC (Kent, 2002) (Karolchik, et al., 2004), the mouse genome was queried for the entirety of the sequencing read. For each hit, that query returned, a start point within the read and a start point in the genome for the beginning of the alignment. It also included a stop point in both the read and the hit. Using the information from the genomic alignment start and stop points, we were able to determine what genes were found in that area.

According to the protocol used to generate the DNA for sequencing, a perfect sequencing read should have started with the ending segment of the long arm of the splinkerette. At the end of that segment should be genomic DNA. The hit from BLAT should begin before or immediately after the end of the splinkerette sequence. The hit may begin before the end of the splinkerette sequence as the end of the splinkerette may match the genomic DNA which was replaced by the splinkerette after digestion and ligation. After the genomic DNA, the LTR from the guest exon should be found. Based on these constraints, the quality of the read was determined by five characteristics of the read that were measured:

Was the LTR present in the read?

Was the splinkerette present in the read?

Was the hit bounded at the beginning by the splinkerette?

Was the hit bounded at the end by the LTR?

What was the hit score above 90?

A sequence was considered present in the read if the p-value for the local alignment of the objective sequence and the read was < 0.01. A hit was considered bounded by a sequence if the hit and the element considered (splinkerette end or LTR beginning) began and ended within seven base pairs of one another. In order to assess how well these measurements could predict the quality of a read, we analyzed all reads. A read for which the splinkerette was present followed by the hit which was followed by the LTR was considered to be a "perfect" hit. All pairs of reads in which at least one of them was perfect were used for assessment. A class system was developed to describe each kind of read. This is described in Table 6. The presence of the splinkerette was a strong indication of the accuracy of the read. Of more than 3,000 sequencing reads, there were more than 400 reads that were considered high accuracy meaning their class accuracy was greater than 95%. More than 370 wells were represented in this list of high quality reads. Amongst these 370 wells, 166 unique accession sets were discovered based on the hits.

Table 6: All reads with a BLAT score greater than 90 are classified into one of these nine classes based on the presence and location relative to the hit of the splinkerette and LTR. The comparable read column shows how many reads of this class were found that had a duplicate read from the same well that was a "perfect" or Class 0 read. The agreement column shows how often the Class 0 agreed with the class in the first column. The final column shows how many total reads of that class were found during all sequencing efforts to date.

Class	LTR Present	Splinkerette Present	LTR Bounded	Splinkerette Bounded	Comparable Reads	Agreement	Total Reads
0	TRUE	TRUE	TRUE	TRUE	106	98%	340
1	TRUE	TRUE	TRUE	FALSE	5	80%	38
2	TRUE	TRUE	FALSE	TRUE	0	NA	0
3	TRUE	TRUE	FALSE	FALSE	0	NA	0
4	TRUE	FALSE	TRUE	FALSE	6	50%	164
5	TRUE	FALSE	FALSE	FALSE	0	NA	0
6	FALSE	TRUE	FALSE	TRUE	24	100%	80
7	FALSE	TRUE	FALSE	FALSE	1	100%	14
8	FALSE	FALSE	FALSE	FALSE	47	79%	387

Imaging for RandTag

When plating cells, only the inner 60 wells were used for new RandTag clones. The remainder of the wells were reserved for imaging controls which consisted of an untagged parental line, a mitochondrial tagged line, an endoplasmic reticulum tagged line, a nuclear tagged line and a plasma membrane tagged line. These wells were imaged using the IC100 which is an automated widefield microscope. All cells were imaged with 2 µM Hoechst because the DNA channel has proven useful for classification and the bright stain of Hoechst was used by the microscope to automatically focus on the cells. Tweny-five fields were imaged for each well.

All images taken from the RandTag project and the associated sequence information were made publicly available (http://pslid.org). The results for the high-quality reads that are published are shown in Table 7. Table 6 includes the gene name, the number of times a clone was sequenced with a given gene tagged, the number of infection runs which resulted in that gene being tagged, subcellular locations from other databases (i.e. Uniprot) and the classification for the gene based on features calculated from images taken of the cells with the tagged protein. In considering the number of unique clones sequenced and the number of unique infections, if there is only a single infection, it is most likely that the unique clones sequenced are sister cells from the same infection. There does seem to be a strong preference toward tagging some genes. For example, the gene "complement component 1, q subcomponent binding protein" was sequenced four times from four different infections. The implication is that either that region of the genome is prone to infection or something about that region of the genome makes it easier to successfully sequence.

Table 7: A list of all tagged targets which were sequenced with high confidence is shown. The target name, number of times a clone with that target tagged was sequenced, the number of unique infections after which that target was found to be tagged, and the documented subcellular location of that target.

Target	Clones Sequenced	Unique Infections	Locations from Other Databases
NADH	1	1	mitochondrial inner

dehydrogenase (ubiquinone) 1 alpha subcomplex, 7 (B14.5a)			membrane,mitochondrion,respiratory chain,mitochondrial respiratory chain complex i,membrane; extracellular region,cytoplasm,mitochondrion,nucleus,endoplasmic reticulum
RAB11a, member RAS oncogene family	1	1	mitochondrion,recycling endosome membrane,plasma membrane,cleavage furrow;endosome,cleavage furrow,trans-golgi network,membrane,mitochondrion,recycling endosome,transport vesicle,plasma membrane; golgi apparatus,cytoplasm,nucleus,peroxisome,mitochondrion,cytoskeleton,plasma membrane
RAN, member RAS oncogene family	1	1	nucleus,melanosome; nucleoplasm,cytosol,protein complex,nucleus,cytoplasm;mitochondrion,cytoplasm,nucleus,golgi apparatus,peroxisome
UTP14, U3 small nucleolar ribonucleoprotein, homolog A (yeast)	1	1	nucleolus,small-subunit processome; cellular_component,nucleus,small-subunit processome;extracellular region,cytoplasm,nucleus,endoplasmic reticulum
actin related	2	1	cell leading edge,focal adhesion,cell

protein 2/3 complex, subunit 2			projection,arp2/3 protein complex; golgi apparatus,cytoplasm,cell leading edge,focal adhesion,arp2/3 protein complex,cytoskeleton,cell projection
collagen, type IV, alpha 2	1	1	collagen type iv; extracellular region,basement membrane,proteinaceous extracellular matrix,collagen,collagen type iv; lysosome,extracellular region,cytoplasm,endoplasmic reticulum,nucleus,mitochondrion
collagen, type V, alpha 1	1	1	basement membrane; extracellular matrix,extracellular region,collagen,collagen type v,basement membrane,proteinaceous extracellular matrix; lysosome,extracellular region,cytoplasm,endoplasmic reticulum,nucleus,mitochondrion
collagen, type XXVIII, alpha 1	1	1	collagen,basement membrane; extracellular region,basement membrane,proteinaceous extracellular matrix,collagen
complement component 1, q subcomponent	4	4	mitochondrion,extracellular space,membrane,nucleus,cytoplasm; extracellular

binding protein			region,cytoplasm,mitochondrion,endoplasmic reticulum
filamin C, gamma	1	1	
guanine nucleotide binding protein (G protein), alpha inhibiting 2	1	1	midbody,cytosol,extrinsic to internal side of plasma membrane,membrane fraction,membrane,centrosome,nucleus,cytoplasm,heterotrimeric g-protein complex,cytoskeleton,membrane raft,plasma membrane; extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,mitochondrion
heat shock protein 90, beta (Grp94), member 1	1	1	plasma membrane part,endoplasmic reticulum lumen,melanosome; endoplasmic reticulum membrane,cytosol,microsome,endoplasmic reticulum,perinuclear region of cytoplasm,plasma membrane part,endoplasmic reticulum lumen,midbody; extracellular region,cytoplasm,nucleus,endoplasmic reticulum
myosin, heavy	1	1	

polypeptide 9, non-muscle			
parvin, alpha	1	1	cytoplasm,focal adhesion,nucleus,lamellipodium; protein complex,cytosol,cytoplasm,focal adhesion,nucleus,membrane,lamellipodium,cytoskeleton,actin cytoskeleton,cell junction,plasma membrane; extracellular region,cytoplasm,mitochondrion,nucleus,peroxisome
phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoribosylaminoimidazole, succinocarboxamide synthetase	1	1	cellular_component; mitochondrion,cytoplasm,endoplasmic reticulum,peroxisome
prolactin family 2, subfamily c, member 3	1	1	extracellular space; extracellular region,extracellular space; extracellular region,cytoplasm,endoplasmic reticulum,lysosome
proteasome	1	1	

(prosome, macropain) subunit, alpha type 5			
protein kinase, cAMP dependent regulatory, type II beta	2	1	camp-dependent protein kinase complex,soluble fraction,cell fraction,mitochondrial inner membrane,cytoplasm,insoluble fraction,centrosome,cytosol,perinuclear region of cytoplasm,membrane raft,plasma membrane; mitochondrion,cytoplasm,nucleus, endoplasmic reticulum,peroxisome
protein phosphatase 2A, regulatory subunit B (PR 53)	2	1	cytoplasm,nucleus; cytoplasm,calcium channel complex,nucleus; extracellular region,cytoplasm,mitochondrion,endoplasmic reticulum,peroxisome
ribosomal protein, large, P1	1	1	ribosome; cytosol,cytoplasm,ribosome,intracellular,ribonucleoprotein complex
signal sequence receptor, delta	2	1	endoplasmic reticulum membrane,integral to membrane; sec61 translocon complex,endoplasmic reticulum,integral to membrane,membrane; lysosome,extracellular region,cytoplasm,endoplasmic reticulum,peroxisome,plasma membrane

solute carrier family 1 (neutral amino acid transporter), member 5	1	1	extracellular region; mitochondrion,cytoplasm,plasma membrane,endoplasmic reticulum,peroxisome
tubulin, alpha 1C	1	1	cytoplasmic microtubule; microtubule,cytoplasm,cytoskeleton,cytoplasmic microtubule,protein complex; lysosome,extracellular region,cytoplasm,nucleus,peroxisome,cytoskeleton
tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide	1	1	mitochondrion,melanosome; kinesin complex,cytoplasm,mitochondrion,axon part,cytosol;extracellular region,cytoplasm,nucleus,endoplasmic reticulum,peroxisome
BCL2-associated athanogene 3	1	1	cytosol,synaptosome; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
DEAD (Asp-Glu-Ala-Asp) box polypeptide 21	1	1	nucleolus,nucleus; cytoplasm,nucleus,plasma membrane,peroxisome

DEAH (Asp-Glu-Ala-Asp/His) box polypeptide 57	1	1	cellular_component; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
E2F transcription factor 3	1	1	transcription factor complex; nucleoplasm,transcription factor complex,nucleus;mitochondrion,cytoplasm,nucleus,plasma membrane,endoplasmic reticulum
Ewing sarcoma breakpoint region 1	1	1	cajal body,nucleolus,cytoplasm,intracellular,nucleus,membrane,plasma membrane;extracellular region,cytoplasm,nucleus,plasma membrane,peroxisome
KDM1 lysine (K)-specific demethylase 6B	1	1	nucleus; nucleus; extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,plasma membrane
LIM domain containing preferred translocation partner in lipoma	1	1	cytoplasm,nucleus; cytoplasm,focal adhesion,nucleus,cell junction; extracellular region,cytoplasm,nucleus,peroxisome

MACRO domain containing 2	2	1	cellular_component; extracellular region,cytoplasm,mitochondrion,nucleus,peroxisome
PDZ and LIM domain 5	1	1	cytosol,postsynaptic density,membrane fraction,actin cytoskeleton,z disc; extracellular region,cytoplasm,mitochondrion,nucleus
RAS-related C3 botulinum substrate 1	1	1	cytosol,lamellipodium,membrane fraction,melanosome,extrinsic to plasma membrane;cytosol,cytoplasm,membrane fraction,intracellular,golgi membrane,cytoplasmic vesicle,cytoplasmic membrane-bounded vesicle,membrane,lamellipodium,cell projection,plasma membrane,extrinsic to plasma membrane; golgi apparatus,extracellular region,cytoplasm,nucleus,peroxisome,mitochondrion,plasma membrane
RNA binding motif protein 3	1	1	cytoplasm,nucleus,dendrite; large ribosomal subunit,cytoplasm,nucleus,cell projection,dendrite
SET nuclear	2	2	cytosol,nucleoplasm,perinuclear region of

oncogene			cytoplasm,endoplasmic reticulum;nucleus,cytoplasm,protein complex,perinuclear region of cytoplasm,endoplasmic reticulum;extracellular region,cytoplasm,nucleus,endoplasmic reticulum,peroxisome
SH3-domain kinase binding protein 1	1	1	synaptosome,cytoplasmic vesicle membrane,focal adhesion,cytoskeleton,synapse;synaptosome,synapse,cytosol,cytoplasm,cytoplasmic vesicle,nucleus,membrane,cytoskeleton,cell junction,plasma membrane;cytoplasm,nucleus,endoplasmic reticulum,peroxisome
Sec61 beta subunit	1	1	
U2 small nuclear ribonucleoprotein auxiliary factor (U2AF) 2	1	1	nuclear speck; ribonucleoprotein complex,nucleus,spliceosomal complex,nuclear speck;mitochondrion,cytoplasm,nucleus,endoplasmic reticulum,peroxisome
WD repeat	1	1	nucleolus; cellular_component,nucleus; c

domain 43			ytoplasm,nucleus,plasma membrane,endoplasmic reticulum,peroxisome
acetyl-Coenzyme A acetyltransferase 1	1	1	mitochondrial inner membrane; mitochondrion,mitochondrial inner membrane,mitochondrial matrix; mitochondrion,cytoplasm,endoplasmic reticulum,peroxisome
actinin alpha 4	1	1	cortical cytoskeleton,ribonucleoprotein complex,stress fiber,pseudopodium; protein complex,nucleolus,ribonucleoprotein complex,pseudopodium,stress fiber,cytoplasm,nucleus,perinuclear region of cytoplasm,cortical cytoskeleton,actin cytoskeleton; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
adenosine kinase	5	2	cytosol,nucleus; cytosol,nucleus; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
annexin A2	1	1	melanosome,basement membrane,sarcolemma,early endosome,schmidt-lanterman incisure,extrinsic to plasma membrane; protein complex,stress fiber,extracellular

			region,cytoplasm,membrane fraction,cell junction,myelin sheath,basement membrane,perinuclear region of cytoplasm,sarcolemma,early endosome,proteinaceous extracellular matrix,schmidt-lanterman incisure; extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,plasma membrane
annexin A5	1	1	cytoplasm,intracellular,sarcolemma,intercalated disc,cell projection,plasma membrane;extracellular region,cytoplasm,nucleus,peroxisome,mitochondrion,plasma membrane
bromodomain containing 2	1	1	cytoplasm,nucleus; cytoplasm,nucleus; extracellular region,cytoplasm,mitochondrion,nucleus
caldesmon 1	9	4	actin filament,membrane fraction,neuronal cell body,dendrite,focal adhesion,actin cap,dendritic spine,postsynaptic density,actin cytoskeleton,plasma membrane; extracellular region,cytoplasm,nucleus,endoplasmic

			reticulum,peroxisome
cytochrome b5 reductase 3	2	1	cytosol,endoplasmic reticulum membrane,mitochondrial inner membrane,mitochondrial outer membrane; endoplasmic reticulum membrane,soluble fraction,mitochondrial inner membrane,cytoplasm,endoplasmic reticulum,membrane,mitochondrion,mitochondrial outer membrane; golgi apparatus,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,mitochondrion
death-associated protein	1	1	cellular_component
epidermal growth factor-containing fibulin-like extracellular matrix protein 1	1	1	extracellular space,proteinaceous extracellular matrix; extracellular region,extracellular space,proteinaceous extracellular matrix; golgi apparatus,extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,plasma membrane
eukaryotic translation	1	1	cytoplasm,mrna cap binding complex; cytoplasm,mrna cap binding

initiation factor 4E member 2			complex;mitochondrion,cytoplasm,nucleus,peroxisome
glutamate dehydrogenase 1	1	1	mitochondrial inner membrane,mitochondrial matrix
hedgehog interacting protein-like 1	1	1	extracellular region,membrane; extracellular region,cellular_component,membrane
hematological and neurological expressed 1-like	1	1	cytoplasm,nucleus; cytoplasm,nucleus
heterogeneous nuclear ribonucleoprotein A3 isoform a	1	1	spliceosomal complex; neuron projection,nucleolus,ribonucleoprotein complex,heterogeneous nuclear ribonucleoprotein complex,cytoplasm,catalytic step 2 spliceosome,spliceosomal complex,nucleus; mitochondrion,cytoplasm,nucleus,peroxisome
leucine rich repeat containing 59	1	1	
minichromoso	2	1	mcm complex; nucleoplasm,mcm

me maintenance deficient 4 homolog (S. cerevisiae)			complex,nucleus;mitochondrion,cytoplasm,nucleus,endoplasmic reticulum
mitogen-activated protein kinase kinase 1	1	1	cytosol; axon part,perikaryon,golgi apparatus,microtubule,cytosol,cytoplasm,cell cortex,dendrite,perinuclear region of cytoplasm,dendrite cytoplasm,soluble fraction,plasma membrane; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
nascent polypeptide-associated complex alpha polypeptide	1	1	cytoplasm,nucleus; cytoplasm,nucleus
nidogen 1	1	1	extracellular region,cytoplasm,proteinaceous extracellular matrix,basement membrane,basal lamina; extracellular region,plasma membrane,endoplasmic reticulum,golgi apparatus,lysosome
nuclear protein 1	1	1	nucleus; nucleus; extracellular region,cytoplasm,mitochondrion,nucleus

paired related homeobox 1	7	3	nucleus; nucleus; mitochondrion,cytoplasm,nucleus,endoplasmic reticulum,peroxisome
phosphoenolpyruvate carboxykinase 2 (mitochondrial)	1	1	mitochondrion; mitochondrion,soluble fraction; golgi apparatus,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,mitochondrion
plectin 1	1	1	insoluble fraction,hemidesmosome,contractile fiber,sarcolemma; insoluble fraction,basal plasma membrane,cytosol,cytoplasm,apical plasma membrane,hemidesmosome,intermediate filament cytoskeleton,focal adhesion,perinuclear region of cytoplasm,sarcolemma,cytoskeleton,plasma membrane,contractile fiber; golgi apparatus,extracellular region,cytoplasm,nucleus,peroxisome,mitochondrion,plasma membrane
polymerase I and transcript release factor	2	1	cytosol,microsome,endoplasmic reticulum,mitochondrion,nucleus,caveola;nucleoplasm,cytoplasm,microsome,endoplasmic reticulum,mitochondrion,nucleus,membrane,caveola,plasma

			membrane;cytoplasm,nucleus,endoplasmic reticulum,peroxisome
protein kinase C and casein kinase substrate in neurons 2	2	2	cytosol,cytoplasmic membrane-bounded vesicle; cytosol,cytoplasm,cytoplasmic vesicle,trans-golgi network; cytoplasm,nucleus,golgi apparatus,peroxisome
recombination signal binding protein for immunoglobulin kappa J region	1	1	cytoplasm,transcription factor complex; cytoplasm,nucleolus,nucleus,transcription factor complex; cytoplasm,nucleus,plasma membrane,endoplasmic reticulum,peroxisome
ribosomal protein L18	1	1	ribosome; cytosol,cytoplasm,ribosome,intracellular,ribonucleoprotein complex
ribosomal protein L23	1	1	ribosome,nucleolus; cytosolic ribosome,nucleolus,ribonucleoprotein complex,cytosol,cytoplasm,ribosome
ribosomal protein L23A	1	1	
ribosomal protein L7	3	3	cytosolic large ribosomal subunit; ribonucleoprotein complex,large ribosomal

			subunit,ribosome,intracellular,cytosol,cytosolic large ribosomal subunit;mitochondrion,cytoplasm,nucleus,peroxisome
ribosomal protein S26	1	1	ribosome
septin 9	6	3	microtubule,cytoplasm,cytoskeleton,perinuclear region of cytoplasm,stress fiber;cytoplasm,nucleus,plasma membrane,peroxisome
small nuclear ribonucleoprotein D1	1	1	
spectrin beta 2	1	1	sarcomere,cuticular plate,spectrin,plasma membrane; protein complex,nucleolus,cytosol,cytoplasm,spectrin,cuticular plate,nucleus,membrane,cytoskeleton,cortical cytoskeleton,plasma membrane;cytoplasm,nucleus,plasma membrane,golgi apparatus,peroxisome
sphingosine-1-phosphate receptor	1	1	integral to plasma membrane; integral to plasma membrane,plasma membrane,integral

3			to membrane,membrane; mitochondrion,nucleus ,plasma membrane,endoplasmic reticulum,peroxisome
staphylococcal nuclease and tudor domain containing 1	1	1	mitochondrion,nucleus,melanosome,rna-induced silencing complex;mitochondrion,cytoplasm,nucleus,rna-induced silencing complex;mitochondrion,cytoplasm,nucleus,endoplasmic reticulum,peroxisome
stearoyl-Coenzyme A desaturase 1	1	1	endoplasmic reticulum membrane,integral to membrane; membrane,endoplasmic reticulum,integral to membrane,microsome; golgi apparatus,lysosome,cytoplasm,endoplasmic reticulum,peroxisome,mitochondrion,plasma membrane
suppression of tumorigenicity 13	1	1	cytoplasm; cytosol,cytoplasm,protein complex; cytoplasm,nucleus,endoplasmic reticulum,peroxisome
syntaxin 3	1	1	apical plasma membrane,integral to membrane; neuron projection,snare complex,azurophil granule,cell-cell

			junction,plasma membrane enriched fraction,apical plasma membrane,specific granule,integral to membrane,membrane,growth cone,plasma membrane; golgi apparatus,extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,plasma membrane
thymopoietin	2	2	chromatin,nuclear envelope,nucleus,nuclear membrane,nuclear inner membrane;mitochondrion,cytoplasm,nucleus, endoplasmic reticulum,golgi apparatus
tissue inhibitor of metalloproteinase 3	1	1	basement membrane; extracellular region,cytoplasm,proteinaceous extracellular matrix,basement membrane; extracellular region,cytoplasm,endoplasmic reticulum,lysosome
tropomyosin 4	1	1	cortical cytoskeleton,podosome; filamentous actin,stress fiber,cytoplasm,cytoskeleton,podosome,cortical cytoskeleton; extracellular region,cytoplasm,endoplasmic

			reticulum,nucleus,peroxisome,cytoskeleton
tubulin cofactor A	1	1	microtubule,cytoplasm; microtubule,cytoplasm,cytoskeleton; extracellular region,cytoplasm,endoplasmic reticulum,nucleus,peroxisome,mitochondrion
vimentin	1	1	
vinculin	1	1	
zinc finger protein 207	1	1	nucleus; cytoplasm,nucleus,golgi apparatus,peroxisome

Implementation and Execution of Active Learning Process Using RandTag Clones

In order to further demonstrate the utility of active learning for high content screening experiments, an experiment was planned such that we would select 47 cell lines from the RandTag project and 47 drugs to treat these cell lines. These cell lines and drugs were selected concurrently so that we would have a set that included some experimental conditions in which we would expect to see location pattern changes when imaging. If we had very few location pattern changes, the learning problem would have been trivial and active learning would not have been worth the effort. We would then duplicate the experimental space as in Chapter 3 for a total of 96 drugs including two vehicles and 96 lines including two untagged parental lines. The fact that these lines and drugs were duplicated was to be hidden from the active learner.

Each round, we would attempt to execute 96 experiments selected using active learning. Each experiment was to be executed in triplicate using 384 well imaging plates.

Clone Selection for Active Learning

The reserve plates were generated during the RandTag project primarily to serve as a source from which to gather cell lines for use in further experimentation. As the RandTag project progressed, the phenotypes of new images were monitored so that we could select 47 tagged lines representing a broad spectrum of phenotypes. We also watched closely as new sequencing results were delivered and assessed; if an interesting protein was tagged, we would collect the cells from the reserve plate for use in the active learning experiment. For cell lines that appeared to grow at a good rate and also had potentially interesting phenotypes, some drugs were tested for effects on the location patterns. The results of these short experiments were assessed only by eye.

Because of the possibility for passage number-dependent phenotypes, we decided to only use cells for experiments that were within 3 passes of one another. In order to synchronize cells at a single passage, cells were grown up in bulk in "tissue culture factories" that provided approximately 2400 square centimeters of growth area. This allowed us to grow up and subsequently freeze enough cells to last the duration of the experiment with some to spare. In total, we selected 47 lines that represented 18 phenotypes which were qualitatively identifiable when untreated.

Drug Selection for Active Learning

An initial large set of drugs (~60) were selected and purchased. If we determined through literature review that a compound was known to affect the location pattern of a protein, we accepted that drug for the active learning project. Other compounds which we hypothesized to have effects were tested. If any effects were observed the drugs were accepted. Nearly half of the initial set of compounds had no observed effects so we adjusted concentrations of some of the accepted drugs in that way creating new drug conditions for the active learning experiment. We desired to avoid repeatedly freezing and thawing of the drugs, so in order to only freeze and thaw once, drugs were solubilized in pure DMSO and a very small volume (~1.3 µL) of each drug in DMSO was added to 1.5 mL tubes. These tubes were then frozen.

Experiment Execution

Experiments were selected using the same active learner as in Chapter 3. A batch of 96 experiments was selected for execution in triplicate each round and an additional 18 control experiments were added. A program was developed to take in a list of experiments. This list of experiments would then be randomly assigned to positions on the 384-well plate such that experiments of like conditions would not likely be found next to one another. Three kinds of protocols were generated for each set of experiments.

In order to plate the cells reliably from round to round, each line was counted by hand using a hemocytometer and diluted such that the final count of cells in each well would be close to 5400 cells in a volume of 80 µL per well. The cell plating protocol was designed to pause after plating every single cell line so that it would not progress beyond our ability to count cells for plating. When a new cell line was counted, we would simply allow the protocol to proceed and the next cell line would be plated where it needed to go on the imaging plate. Because of equipment constraints the program generating these protocols was designed to split any plating process into multiple individual protocols where necessary. For example, only 24 lines could be plated in a single cell plating protocol, so for every round two protocols needed to be made because all rounds had experiment selections calling for more than 24 lines.

The drugs were pre-plated prior to being added to the experimental wells. By pre-plating, we had better control over the timing of the final plating process as well as a reduction in the total number of tips used during the experiments. A set of frozen 1.5 mL tubes were gathered based on the drugs required for the experiments selected. For each tube, 1 mL of Optimem + 2 µM Hoechst was added. The volume and temperature of the media was significantly greater than the tiny frozen droplet of DMSO and drug, so the droplet thawed quickly in the media. It was then vigorously mixed. After mixing, the drug was plated into a separate drug plate.

Just prior to actually running the experiments, a 384-well plate was made that contained only 1x PBS. By making a plate containing PBS, a wash step could be utilized while not contaminating a reservoir of PBS and yet minimizing wasted tips. For each set of experiments, the final robot protocols developed were for the final plating of the drugs. During the development of these protocols it was discovered that the robot software was unable to handle more than ~1100 steps in a protocol, so the final plating protocol needed to be split into a pair of protocols as each well required four steps: aspirate media and move to waste, add PBS from corresponding well in PBS plate to imaging plate, aspirate PBS and move to waste and add drug/imaging media solution from corresponding position in drug plate. These actions were undertaken in the same order in which they would be imaged on the IC100 at approximately the same rate in order to maintain constant exposure time to the drug. On the IC100, 12 fields were taken per well.

Image Processing for RandTag Active Learning

The active learner took discrete class labels as input. In order to determine these labels, field level features were calculated for all fields. Despite of our best efforts to accurately count cells and setup the microscope, some fields were poor either because of a lack of cells or poor focus. In order to prevent these sorts of images from affecting our final phenotype determination, a Support vector machine classifier was trained to recognize good and bad images and used to filter out the bad images. For the remaining images, field level features were calculated using "field+" features using PySLIC which took into consideration the protein channel as well as the reference DNA channel when calculating feature values.

Once these features were calculated, we clustered experimental conditions using a hierarchical approach with the distance between two conditions represented by the mean accuracy of a set of balanced nearest neighbor classifiers as was calculated in Chapter 2. A tree was built clustering such that at each level, the pair of clusters with the most overlap (lowest accuracy classifier) was combined to form a new cluster. In order to determine at what threshold to stop clustering, all fields from each experimental condition were randomly divided into two subsets which were then each treated as a new experimental conditions in a new clustering. The clustering of these divided true conditions was continued until the point at which 90% of all pairs of original conditions were clustered into the matching clusters. This process was executed five times and the mean threshold for 90% recovery was used to select the clustering stopping point in the original clustering with undivided experimental conditions. The resulting labels were used as inputs for the active learner in order to select the next batch of experiments to be executed.

Discussion

In this chapter, the implementation of a highly automated active learning pipeline from a largely manual process was described. This required the implementation and testing of a sequencing methodology. The most significant result from this work was the sequencing of 175 clones and their publication.

Chapter 6: Conclusions and Future Work

In this thesis, the primary effort was focused around utilizing active learning in conjunction with high-content screening. This approach was undertaken for two model organisms which required different sorts of experimental efforts, image analysis and machine learning techniques.

In Chapter 2, images gathered from an abbreviated active learning campaign were analyzed using multiple methods to discover drug effects on *Arabidopsis thaliana* protoplasts. These protoplasts were sourced from all above ground tissues. As a result, there was substantial heterogeneity within the populations of protoplasts. In order to address the issue of heterogeneity, methods were utilized whose measurements were based on small regions of the images (circles using Hough transform and small square patches using PhenoRipper). Using these methods, many features were calculated per experimental condition and thus comparison methods were needed that allowed for the comparison of distributions with wildly different sample sizes. We chose to utilize mixture models and measure pairwise classification error between experimental conditions. The results of these analyses seem to indicate that Damnacanthal was likely to affect the distribution of protoplasts and that Tyrphostin may have had an effect as well.

In Chapter 3, active learning simulations were executed to demonstrate the utility of using active learning methods to direct high-content screening campaigns. In this case, protoplast based mixture models were calculated every round. As a result, the inputs to the active learning process were much less stable than those normally utilized for most active learning problems. A comparison between models was utilized to show the benefit of using active learning. For the first half of most simulations, the active learning model had accuracy which was worse than the accuracy of a model trained using randomly selected data. In the second half of the simulations, the active learning method showed higher accuracy than the model trained using randomly selected data. The utility of the active learning method in this case is limited to only situations in which an experimenter is willing to execute at least half of the experiments.

In Chapter 4, methods were described to utilize active learning to dramatically improve the discovery rate and predictive accuracy of models developed to predict the effects of compounds across multiple diverse targets as measured by diverse experimental means. Approaches of this nature could allow for earlier detection of deleterious side effects from new pharmaceuticals saving research efforts as well as improving the safety of new drugs.

In Chapter 5, significant improvements in the RandTag project were implemented including the automation of many crucial steps allowing for an increase in throughput. As a result of the efforts in the RandTag project an active learning experiment was setup using 48 cell lines and 48 drugs. The final experimental space tested was actually 96 cell lines by 96 drugs in size. This required substantial effort in terms of the selection and preparation of the cell lines and drugs to be used. An image analysis method was implemented to detect changes in phenotypes from the treatment with drugs which was based around a hierarchical clustering algorithm using classifier accuracy to determine the difference between two distributions.

Thesis Contributions

1. We executed a high-content screening campaign to detect the effects of drugs on multiple lines of protoplasts with different tagged proteins.

2. We developed methods for the comparison of distributions of observations using nearest neighbor classifiers.
3. We determined that Damnacanthal and Tyrphostin had effects on the protoplasts tested.
4. We demonstrated that in spite of the inability to identify a protein by its location pattern in untreated cells, observing the effects of treatment with a small number of drugs can allow one to accurately identify a tagged protein.
5. We implemented an actual active learning pipeline to test plant protoplasts for effects resulting from drug exposure.
6. We designed a predictive model for the effects of multiple compounds on multiple target proteins utilizing external protein and compound information concurrently.
7. We demonstrated that the predictive model can learn from diverse sources of experimental information.
8. We demonstrated that using that predictive model in conjunction with active learning yields significant rewards in terms of hit discovery rate and accuracy improvement.
9. We substantially improved the throughput of the RandTag project through the addition of automated methods to perform many of the most tedious tasks.
10. We implemented a system for generating customized protocols for the Eppendorf epMotion liquid handling robot using a programmatic interface instead of the graphical interface allowing for significantly more sophisticated protocols.
11. We optimized the sequencing protocol used to determine which proteins were tagged in the RandTag project resulting in the successful sequencing and subsequent publication of 175 unique clones.
12. We sequenced and imaged more than 300 tagged clones for the RandTag project and imaged thousands more.
13. We implemented an active learning pipeline to test for the effects of 48 drug treatments on the location patterns of 48 NIH 3T3 cell lines.

14. We developed a hierarchical clustering method for the clustering of groups of fields with field level features calculated.

Future Work

In the high-content screening analysis of the protoplast experiments, the PhenoRipper methods seemed to show promise in that detected effects had relatively low p-values and seemed to be somewhat consistent across concentrations. These results could be improved by testing alternative parameter sets. The parameters chosen were essentially the recommended parameters used in the original software package. In the current work, the optimal model was selected based on the assumption that duplications of the same experimental condition should have similar measurements. For positive controls, each of the experimental lines was considered to be distinct from all other lines in the control conditions. Some lines do indeed appear to be similar. If more cell lines were to be added, the problem of model selection choosing models that attempted to separate similar phenotypes would be exacerbated. As an alternative, these relationships could be assessed by a protoplast expert. The assessments of these relationships could then be used to drive the model selection process across various sets of parameters using PhenoRipper which could improve the final analysis.

Another potential avenue for improvement would be to increase the accuracy of the protoplast segmentation. In the work presented, protoplasts are found in regions which do contain protoplasts, but these patches were relatively infrequently centered tightly over single protoplasts. Because of the poor segmentation, the best approach was to measure a large number of features for each patch. As a result, we could only assess that populations were different, and we could not assess what systems were being affected by the compounds. With good segmentation, one could utilize approaches from cell-based generative modeling which seek to automatically build compact and statistically accurate models from images which can be used generatively to build new example images of cells based on cell shape, nuclear morphology, object distribution and microtubule patterns (Zhao & Murphy, 2007; Shariff, et al., 2009). By utilizing these methods, one could conceivably infer the effects of compounds on specific systems within the protoplasts.

In order to improve the active learning simulations utilizing the protoplast data, the greatest improvement could be gained by generating a more stable phenotype assessment model than the k-means and hierarchical clustering models that were tested using the protoplast mixture model. This may be as simple as using one of these clustering methods in conjunction with an alternative image analysis method, such as the PhenoRipper system or using the nearest neighbor classifier based approach to assessing differences between protoplasts distributions.

Improvements could be made to the work in Chapter 4 through the development of predictive models that may or may not make use of the features describing each of the targets and compounds. This would allow for the inclusion of diverse sets of treatments not limited to small molecule compounds as well as a diverse set of protein targets. When not using external features, the actual experiments used for the assays could be anything ranging from high-throughput screening results to the results of FDA trials as long as the independent variable across an assay is the treatment. A model for that could be rapidly trained using active learning to make accurate predictions for these sorts of assays would be very useful for drug discovery and development. In order to effectively extend this work, it would be of use to consider the composition of an ideal dataset. First one must consider the source of the data to be used to populate the matrix. Some experimental protocols are very noisy and would be less well suited for this purpose without a large number of experimental replicates per experimental condition. The results of some experiments are categorical in nature, particularly for some high-content screening processes and these could not be used directly in the final matrix without some modification. Additionally, one must consider the experimental protocols in use to generate the data. In the process simulated, a heterogeneous set of experiments were selected for "execution". It might be the case that for some of these experiments it is more practical to run numerous experiments rather than a single experiment. Efficiency improvements in a discovery process using this method would be best realized for experimental processes which have high costs per experiment relative to the costs of executing a single batch of experiments.

The RandTag based active learning project has been implemented and the primary future effort involves the actual execution of the pipeline. To date, 28 out of a total of 96 rounds selected using active learning have been executed. Additional cell lines with new tagged proteins as well as new drugs may be added to the system to demonstrate the ability to make predictions for new proteins and drugs.

Works Cited

Anon., n.d. [Online]

Available at: http://openbabel.sourceforge.net/

[Accessed 10 November 2010].

Balaban, N. Q. et al., 2004. Bacterial Persistence as a Phenotypic Switch. *Science,* 305(5690), pp. 1622-1625.

Barbe, L. et al., 2008. Towards a confocal subcellular atlas of the human proteome. *Molecular Cell Proteomics,* Volume 7, pp. 499-508.

Bay, H., Ess, A., Tuytelaars, T. & Gool, a. L. V., 2008. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding,* pp. 346-359.

Boland, M. V. & Murphy, R. F., 2001. A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells. *Bioinformatics,* Volume 17, pp. 1213-1223.

Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H., 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry,* Volume 4.

Bredel, M. & Jacoby, E., 2004. Chemogenomics: An emerging strategy for rapid target discovery. *Nature Reviews Genetics,* April.pp. 262-275.

Breiman, L., 2001. Random Forests. *Machine Learning.*

Brommage, R. et al., 2008. High-throughput Screening of Mouse Knockout Lines Identifies True Lean and Obese Phenotypes. *Obesity,* pp. 2362-2367.

Burges, C. J., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* pp. 121-167.

Çağatay, T. et al., 2009. Architecture-Dependent Noise Discriminates Functionally Analogous Differentiation Circuits. *Cell,* 139(3), pp. 512-522.

Canny, J., 1986. A Computational Approach to Edge Detection. *IEEE Transactions Pattern Analysis and Machine Intelligence,* 8(6), pp. 679-698.

Chakraborty, S., Balasubramanian, V. & Panchanathan, S., 2010. *Dynamic Batch Size Selection for Batch Mode Active Learning in Biometrics.* s.l., s.n.

Coelho, L. P., Peng, T. & Murphy, R. F., 2010. Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics,* Volume 26, pp. i7-i12.

Costes, S. V. et al., 2004. Automatic and Quantitative Measurement of Protein-Protein Colocalization in Live Cells. *Biophysical Journal,* 86(6), pp. 3993-4003.

Cui, A. & Schneider, J., 2010. *Active Learning for Fast Drug Discovery,* s.l.: s.n.

Danziger, S. A. et al., 2009. Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning. *PLoS Computational Biology,* Volume 5.

de Castro, E. et al., 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research,* pp. 362-365.

Devon, R. S., Porteous, D. J. & Brookes, A. J., 1995. Splinkerettes—improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Research,* pp. 1644-1645.

Duda, R. O. & Hart, P. E., 1972. Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Communications of the ACM,* 15(1).

Efron, B., Hastie, T., Johnstone, I. & Tibrshriani, R., 2004. Least Angle Regression. *The Annals of Statistics,* 32(9).

Faraco, M. et al., 2011. One Protoplast Is Not the Other!. *Plant Physiology,* 156(2), pp. 474-478.

Fujii, A., Inui, K., Tokunaga, T. & Tanaka, H., 1998. Selective Sampling for Example-based Word Sense Disambiguation. *Computational Linguistics,* 24(4), pp. 573-597.

Fujiwara, Y. et al., 2008. Virtual Screening System for Finding Structurally Diverse Hits by Active Learning. *Journal of Chemical Information and Modeling,* Volume 48, pp. 930-940.

Garcia-Osuna, E. et al., 2007. Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. *Annals of Biomedical Engineering,* pp. 1081-1087.

Giuliano, K. A. et al., 2009. Cellular systems biology profiling applied to cellular models of disease. *Combinatorial Chemistry & High Throughput Screening,* Volume 12, pp. 838-848.

Guha, R. et al., 2006. The Blue Obelisk-Interoperability in Chemical Informatics. *Journal of Chemical Information and Modeling,* Volume 46, pp. 991-998.

Han, L., Wang, Y. & Bryant, S. H., 2008. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics,* Volume 9.

Huang, S.-Y. & Zou, X., 2010. Advances and Challenges in Protein-Ligand Docking. *International Journal of Molecular Sciences,* Volume 11, pp. 3016-3034.

Inglese, J. et al., 2007. High-throughput screening assays for the identification of chemical probes. *Nature Chemical Biology,* pp. 466-479.

Jarvik, J. W. et al., 1996. CD-Tagging: A New Approach to Gene and Protein Discovery and Analysis. *BioTechniques,* pp. 896-904.

Karolchik, D. et al., 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research,* Volume 32, pp. D493-D496.

Kearsley, S. K. et al., 1996. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.,* 36(1), pp. 118-127.

Keiser, M. J. et al., 2009. Predicting new molecular targets for known drugs. *Nature,* 1 November.pp. 175-181.

Kent, W. J., 2002. BLAT - The BLAST-Like Alignment Tool. *Genome Research,* Volume 12, pp. 656-664.

Koutsoukas, A. et al., 2011. From in silico target prediction to multi-target drug design: Current databases, methods and applications. *Journal of Proteomics,* 18 May.

Lengauer, T. & Rarey, M., 1996. Computational methods for biomolecular docking. *Current Opinion in Structural Biology,* Volume 6, pp. 402-406.

Li, Q., Wang, Y. & Bryant, S. H., 2009. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics,* Volume 25, pp. 3310-3316.

Liu, Y., 2004. Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. *Journal of Chemical Information and Computer Sciences,* Volume 44, pp. 1936-1941.

MacKay, D. J. C., 1992. Information-Based Objective Functions for Active Data. *Neural Computation,* pp. 590-604.

MacQueen, J., 1967. *Some Methods for Classification and Analysis of Multivariate Observations.* s.l., s.n., pp. 281-297.

Manders, E. M. M., Verbeek, F. J. & Aten, J. A., 1993. Measurement of co-localisation of objects in dual-colour confocal images. *Journal of Microscopy,* Ja, Volume 169, pp. 375-382.

Meijering, E., Smal, I. & Danuser, G., 2006. Tracking in Molecular Bioimaging. *IEEE Signal Processing Magazine,* May.pp. 46-53.

Mohamed, T. P., Carbonell, J. G. & Ganapathiraju, M. K., 2010. Active Learning for human protein-protein interaction prediction. *BMC Bioinformatics,* Volume 11.

Pardo-Martin, C. et al., 2010. High-throughput in vivo vertebrate screening. *Nature Methods,* pp. 634-636.

Patani, G. A. & LaVoie, E. J., 1996. Bioisosterism: a rational approach in drug design. *Chemical Reviews,* pp. 3147-3176.

Paul, S. M. et al., 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery,* Volume 9, pp. 203-214.

Peng, H. & Myers, E. W., 2004. *Comparing In situ mRNA Expression Patterns of Drosophila Embryos.* s.l., s.n., pp. 157-166.

Peng, T. et al., 2010. Automated Unmixing Of Subcellular Patterns: Determining the Distribution of Probes Between Different Subcellular Locations. *Proc. Natl. Acad. Sci. U.S.A.,* Volume 107, pp. 2944-2949.

Perlman, Z. et al., 2004. Multidimensional Drug Profiling By Automated Microscopy. *Science,* Volume 306, pp. 1194-1198.

Pournara, I. & Wernisch, L., 2004. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics,* Volume 20, pp. 2934-2942.

Rajaram, S., Pavie, B., Wu, L. F. & Altschuler, S. J., 2012. PhenoRipper: software for rapidly profiling microscopy images. *Nature Methods,* 9(7), pp. 635-637.

Sanger, F., Nicklen, S. & Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *PNAS,* 74(12), pp. 5463-5467.

Settles, B. & Craven, M., 2008. *An analysis of active learning strategies for sequence labeling tasks.* s.l., s.n., pp. 1070-1079.

Shariff, A., Rohde, G. K. & Murphy, R. F., 2009. *Indirect learning of generative models for microtubule distribution from fluorescence microscope images.* s.l., s.n.

Sheridan, R. P., Miller, M. D. & Underwood, D. J. K. S. K., 1996. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.,* Volume 36, pp. 128-136.

Sklar, L. A., Carter, M. B. & Edwards, B. S., 2007. Flow Cytometry for Drug Discovery, Receptor Pharmacology and High Throughput Screening. *Current Opinions in Pharmacology,* Volume 7, pp. 527-534.

Stegle, O. et al., 2009. Predicting and understanding the stability of G-quadruplexes. *Bioinformatics,* Volume 25, pp. i374-i382.

Sukumar, N., Hepburn, T., Sundling, M. & Breneman, C., n.d. *Protein Recon.* [Online] Available at: http://reccr.chem.rpi.edu/Software/Protein-Recon/Protein-Recon-index.html [Accessed January 2011].

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* Volume 58, pp. 267-288.

Tong, S. & Koller, D., 2001. *Active Learning for Structure in Bayesian Networks.* Seattle, Washington, s.n.

Trask, O. J. et al., 2009. High-Throughput Automated Confocal Microscopy Imaging Screen of a Kinase-Focused Library to Identify p38 Mitogen-Activated Protein Kinase Inhibitors Using the GE InCell 3000 Analyzer. *Methods in Molecular Biology.*

Warmuth, M. K. et al., 2003. Active Learning with Support Vector Machines in the Drug Discovery Process. *Journal of Chemical and Informatio and Computer Sciences,* Volume 43, pp. 667-673.

Wildman, S. A. & Crippen, G. M., 1999. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences,* Volume 35, pp. 868-873.

Wilkins, M. R. et al., 1998. Protein Identification and Analysis Tools in the ExPASy Server. *Methods in Molecular Biology,* Volume 112.

Yang, L. & Carbonell, J., 2009. *Cost Complexity of Proactive Learning via a Reduction to Realizable Active Learning,* Pittsburgh: Carnegie Mellon University.

Zhang, D., Chen, S. & Zhou, Z.-H., 2008. Constraint Score: A new filter method for feature selection with pairwise constraints. *Pattern Recognition.*

Zhao, T. & Murphy, R. F., 2007. Automated Learning of Generative Models for Subcellular Location: Building Blocks for Systems Biology. *Cytometry Part A,* 71A(12), pp. 978-990.