

Predicting Gene Function from Images of Cells

by

Thouis Raymond Jones

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

© Massachusetts Institute of Technology 2007. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
May 1, 2007

Certified by.....
Polina Golland
Assistant Professor
Thesis Supervisor

Accepted by.....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Predicting Gene Function from Images of Cells

by

Thouis Raymond Jones

Submitted to the Department of Electrical Engineering and Computer Science
on May 1, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Science

Abstract

This dissertation shows that biologically meaningful predictions can be made by analyzing images of cells. In particular, groups of related genes and their biological functions can be predicted using images from large gene-knockdown experiments. Our analysis methods focus on measuring individual cells in images from large gene-knockdown screens, using these measurements to classify cells according to phenotype, and scoring each gene according to how reduction in its expression affects phenotypes.

To enable this approach, we introduce methods for correcting biases in cell images, segmenting individual cells in images, modeling the distribution of cells showing a phenotype of interest within a screen, scoring gene knockdowns according to their effect on a phenotype, and using existing biological knowledge to predict the underlying biological meaning of a phenotype and, by extension, the function of the genes that most strongly affect that phenotype. We repeat this analysis for multiple phenotypes, extracting for each a set of genes related through that phenotype, along with predictions for the biology of each phenotype.

We apply our methods to a large gene-knockdown screen in human cells, validating it on known phenotypes as well as identifying and characterizing several new cellular phenotypes that have not been previously studied.

Thesis Supervisor: Polina Golland

Title: Assistant Professor

Acknowledgments

Thanks to Anne for starting and shepherding this work.

Thanks to Polina and David for their advice, guidance, and teaching. Especially, I thank Polina for teaching me to think critically and carefully about this work, and David for helping me to understand biological research and how I could contribute to it.

Thanks to all the members of the Sabatini Lab for welcoming me into their group, for teaching me, and for sharing their data and advice. In particular, I am indebted to Jason, Doug, Rob, Jan, Mike, Adam, Steve, Shomit, Xana, and Peggy. Thanks also to Dave, Serena, Greg, Noa, and Piyush at the Broad Institute, for their helpful discussions.

Thanks to the vision group, especially Thomas and Wanmei. Thanks also to the graphics group. I especially thank Frédo, for his support as I sought a new path.

Thanks to my committee for their suggestions and help in improving this work.

Thanks to Ron and Sarah at MERL.

Thanks to my parents and my wife's parents, for their love and support. Thanks also to my son Nathan, for helping me keep perspective, and his impending brother, for impetus.

Most importantly, thank you to my wife Ann, for everything.

Contents

1	Introduction	13
1.1	Making and Interpreting Predictions from Gene-Knockdown Screens	14
1.2	Related Work	17
1.3	Contributions	19
2	Biological Background	21
2.1	Gene-Knockdown Screens	21
3	From Images to Cytological Profiles	27
3.1	Illumination and Staining Variation	28
3.1.1	Correcting Illumination and Staining	30
3.2	Identifying Nuclei	39
3.3	Identifying Cells	41
3.3.1	Validation	44
3.3.2	Connection to Geodesic Active Contours	47
3.4	Measuring Cells	49
3.5	Summary	51
4	From Cytological Profiles to Phenotypes	53
4.1	From Measurements to Phenotypes	54
4.1.1	Simple Phenotypes	54
4.1.2	Complex Phenotypes	57

4.2	GentleBoosting for Cell Classification	58
4.3	Summary	60
5	From Phenotypes to Phenotype Profiles and Related Genes	63
5.1	Scoring Genes	64
5.1.1	Scoring Genes from Gene Knockdowns	69
5.1.2	Phenotype Profiles and Predicting Related Genes	71
5.2	Summary	73
6	From Phenotype Profiles to Biological Function	75
6.1	Methods for Connecting to Existing Biological Knowledge	76
6.2	An Interpretation of Phenotype Profiles	80
6.3	Summary	81
7	Results	83
7.1	Cell Cycle Control Genes	83
7.2	Novel Phenotypes	87
7.3	Interactions Between Phenotypes	97
7.4	Summary	101
8	Discussion	103
8.1	Future Directions	104

List of Figures

1-1	An overview of our methods.	14
1-2	Examples of cellular phenotypes.	16
2-1	Gene transcription and translation, and gene silencing by siRNA or shRNA.	23
2-2	High-throughput screening formats.	24
2-3	A typical image from a screen.	25
3-1	Steps in the image processing pipeline.	27
3-2	Mean intensity of the DNA-stained channel versus actual cell density, for a screen of 5600 gene-knockdowns.	29
3-3	Median nuclear DNA staining intensity, plotted in a grid corresponding to the physical location on the slide where the image was captured.	30
3-4	Histogram of cellular DNA content, before and after correction illumination and staining variation.	30
3-5	Distribution of pixel values for two staining channels, actin and DNA.	38
3-6	Human cells stained for DNA.	40
3-7	<i>Drosophila melanogaster</i> Kc cells stained for DNA, and their segmentation.	41
3-8	Segmenting Cells.	43
3-9	Our segmentation algorithm applied to synthetic images.	45
3-10	Our segmentation algorithm applied to noisy synthetic images.	45

3-11	Histogram of signed distances and cumulative distribution of absolute distances between automatic and manual cell boundary segmentations in the validation set.	47
3-12	Comparison of our cellular segmentation to manual segmentation boundaries.	48
3-13	Cytological Profiles.	51
4-1	A diagram of the cell cycle.	54
4-2	Identifying the phase of the cell cycle of individual cells from measurements of DNA and phospho-histone H3, a marker for mitosis.	55
4-3	Interface for building an automatic classifier.	61
5-1	Histograms and Binomial model fit.	65
5-2	A graphical representation of the Beta-Binomial model.	66
5-3	Histograms and Beta-Binomial model fit.	68
5-4	Phenotype profile for 4N phospho-histone H3 negative (G2 phase) cells.	72
6-1	GSEA scoring cell-cycle genes against phenotype profile for 4N phospho-histone H3 negative (G2 phase) cells.	78
6-2	An example gene network regulating a phenotype.. . . .	80
7-1	Many phenotypes interact with one another.	99
7-2	Some phenotypes perturb the cell-cycle distribution.	100
7-3	Limiting phenotypes to a single phase in the cell cycle does not always remove their dependence.	100

List of Tables

7.1	Gene sets from GSEA that correlate with the G2/4N phenotype profile.	84
7.2	Gene sets from GSEA that correlate with the anaphase vs. metaphase phenotype profile.	85
7.3	Compounds that correlate with the G2/4N phenotype profile in the Connectivity Map.	87
7.4	Top scoring genes for the Actin Dots phenotype.	88
7.5	Gene sets from GSEA that correlate with the Actin Dots phenotype profile.	91
7.6	Gene sets from GSEA that correlate with the Actin Ring phenotype profile.	92
7.7	Gene sets from GSEA that correlate with the Crescent Nuclei phenotype profile.	94
7.8	Gene sets from GSEA that correlate with the Peas-in-a-pod phenotype profile.	96
7.9	Gene sets from GSEA that correlate with the Crescent Nuclei phenotype profile when limited to 4N cells.	98

Chapter 1

Introduction

This dissertation shows how biologically meaningful predictions can be made by analyzing images of cells. Specifically, the predictions are sets of genes with similar biological function. In most cases, we also make a prediction for the function, method of action, or some other common element between those genes, connecting to existing biological knowledge.

The data for generating these hypotheses come from image-based screens, in which groups of cells are grown under a wide variety of conditions, then stained and imaged. Each image is of a group of cells under one of the conditions. In this work, the images are taken from gene-knockdown screens, in which each condition corresponds to a particular gene's expression being targeted for suppression. The size of the experiments discussed here are on the order of a few thousand gene-knockdowns.

Our approach to forming predictions of related genes and their biological function is outlined in figure 1-1. Within images, we identify individual cells and measure a large number of features for each cell. These measurements make up the cell's *cytological profile*. Using the original images, human guidance, and the cytological profiles, we build a classifier for each phenotype of interest. For each phenotype, the corresponding classifier is used to label every cell in the screen according to whether that cell shows the corresponding phenotype or not. Each gene is scored according to how reducing its expression affects the relative number of cells showing the phenotype. The genes and their scores form a *phenotype profile*. We predict that genes with the

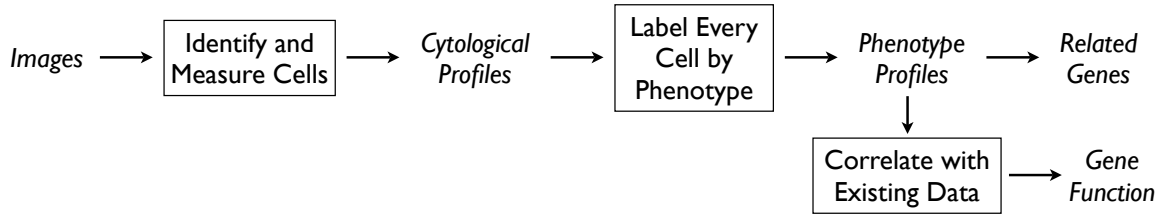


Figure 1-1: An overview of our methods. Our goal is to use images of cells to predict related genes, and their biological function. In order to do this, we first identify and measure cells in the image, forming a cytological profile for each cell. Using human guidance, the original images, and the cytological profiles, we build an automatic classifier for a phenotype of interest. We use this classifier to label every cell according to whether it shows that phenotype. We then score each gene according to how reducing its expression affects the fraction of cells with the phenotype. The gene scores form a phenotype profile. Those genes scoring most strongly are predicted to be related. We correlate the phenotype profile with existing biological data in order to predict the biological mechanisms of the phenotype and the genes that affect it. Italics indicate different forms of data.

strongest effect on the phenotype are related. To predict the underlying biology of the phenotype, and by extension the function of the genes that affect it, we look for correlations between the phenotype profile and existing biological data, such as annotated sets of genes or expression profiles.

Our goal in this work is to expand and improve the tools for analyzing large, image-based screens, in particular gene-knockdown screens. Although sequencing projects have made it possible to identify almost all of the genes in several organisms, little is known about the function of most of these genes. Our methods allow us to predict which genes are related within the cell, from their common effect on cell appearance, as well as function, by comparison to existing biological knowledge.

1.1 Making and Interpreting Predictions from Gene-Knockdown Screens

This work relies on the following fundamental assumption:

If two genes produce a similar appearance in cells when knocked down, that is evidence that they have a similar biological func-

tion.

This is perhaps obvious, given that the appearance of cells is a result of the expression of their genes in combination with their environment, but it is worth stating outright.

We will quantify similarity of appearance by analyzing specific phenotypes. Thus, our predictions about genes' similarity and functions are based on which knockdowns cause cells to exhibit a particular phenotype more or less often than others. In this work, we define a phenotype as a binary trait, which a cell does or does not possess. It may be visually apparent (qualitative) or defined by cytological measurements (quantitative). This approach is slightly artificial, as there exist continuous-valued phenotypes, such as size of a cell or the concentration of a particular protein within the cell. We do not deal with such phenotypes here, though some of our methods could be adapted to these phenotypes. Prior approaches to analyzing images of cells have generally scored images as a whole [52], or measured individual cells but taken the mean of these values to score images [31]. In contrast, we follow a cell-centric approach to improve our ability to detect changes that might otherwise be lost because the phenotype is present in only a fraction of cells due to natural cell-to-cell variability or low penetrance of the phenotype [60].

Some phenotypic differences may only manifest themselves under specific cellular conditions. Any two genes with different sequence have some difference in function; however, we may not be able to detect this difference without very targeted experimentation. For the same reason, two genes with many functions in common may still cause very different phenotypes in an experiment, due to whatever slight differences there are between them. These caveats indicate that we should be aware that any predictions are only hypotheses that must be evaluated via further experimentation.

To label cells by phenotype, we first process the images in order to identify and measure individual cells. The combined measurements for a cell form its *cytological profile*. From cytological profiles, cells can be classified according to whether they present a particular phenotype or not. In some cases, the phenotype can be identified by thresholding a single or small number of measurements in the cytological profile.

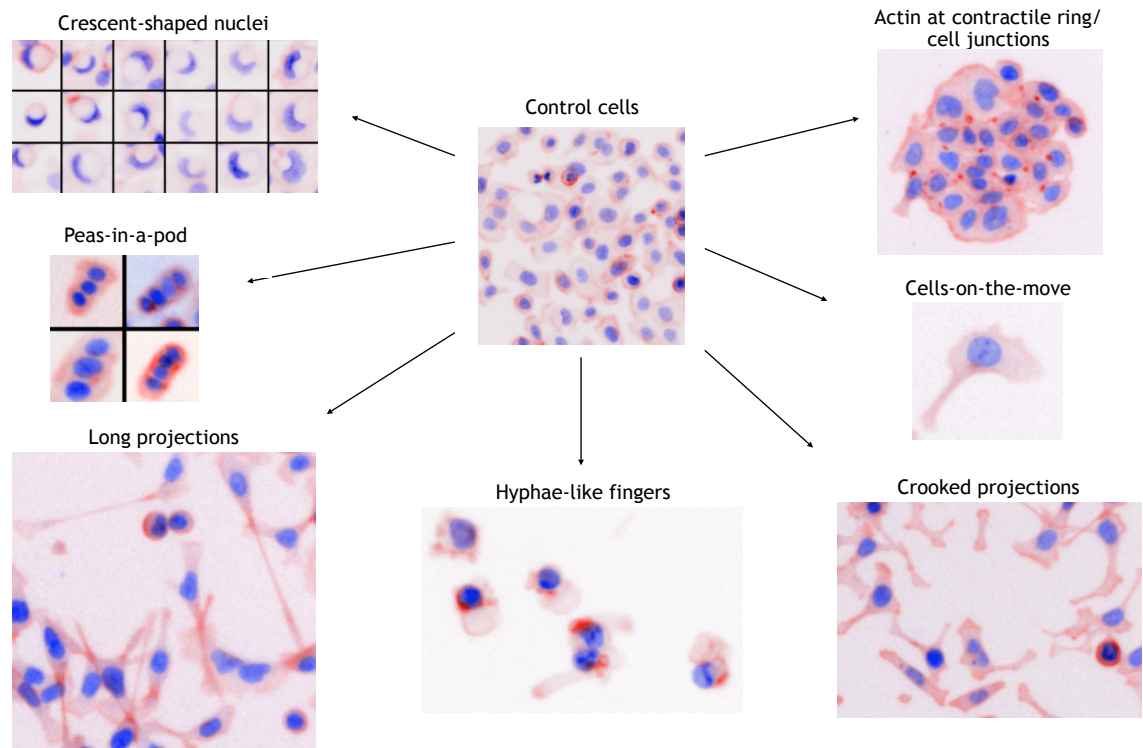


Figure 1-2: Examples of cellular phenotypes. These phenotypes are easily identified visually, but not easily measured via image processing. (The cell images have been color-reversed for printing. Compare to figure 2-3)

For example, a cell's total amount of DNA, an indicator of its progress through the cell cycle, can be measured by totalling the intensity of the DNA stain in the nucleus. In other cases, the cellular phenotype may be a morphological change, such as a crescent-shaped nucleus or cells with long projections, as illustrated in Figure 1-2. In these cases, a single measurement is unlikely to capture the difference between cells with and without the phenotype. To handle these phenotypes, we build automatic classifiers that distinguish between phenotype-positive and phenotype-negative cells. In either case, thresholding a few measurements or using an automatically trained classifier, we label every cell in the screen as having the phenotype or not.

Once individual cells have been classified as having a phenotype or not, we identify which gene knockdowns enhance or suppress that phenotype, by identifying genes that cause the fraction of cells showing the phenotype under that treatment to be significantly higher or lower than expected. We fit a model to the distribution of

phenotype-positive and phenotype-negative cells. Those genes that deviate significantly from the model have the largest effect on the phenotype of interest, and are putatively related through their common (or opposite) effects on cells.

However, if we concentrate on only the genes with the most significant effect on a phenotype, we neglect much of the information available from the screen. The continuum of scores, one for each gene targeted in the screen, has information about how each gene affects the phenotype, even those that do so less strongly. Such genes may be only peripherally involved in the cellular process producing the phenotype, but the full set of screen-wide scores and their relative effects can provide insights into what cellular processes are involved in producing the phenotype of interest.

In order to interpret the full set of scores from a screen, we treat them as a *phenotype profile*, analogous to an expression profile. Expression profiles are large-scale measurements of gene expression, captured via microarrays [85], and quantifying for each gene on the microarray how its expression (i.e., activity) changes between two conditions (e.g., control and treated) or classes of cells (e.g., malignant versus benign tumors). Phenotype profiles operate in the opposite direction: for each gene, we measure how the phenotype of a cell changes when that gene's expression is perturbed, rather than how that gene is perturbed in cells of a particular phenotype. As is done with expression profiles, we look for correlations between the phenotype profiles and biological data using previously developed tools [89, 55].

1.2 Related Work

The application of automatic image analysis to cell images, and large screens in particular, is relatively new. Until recently, large screens were analyzed by researchers examining and scoring individual images by eye [27, 52]. Some groups have developed automatic image analysis methods; though these have been designed to score single phenotypes rather than for general applicability [62, 78, 101], they do demonstrate that image analysis can be a powerful tool for biological discovery. There are several vendors of commercial software for cell image analysis [35], usually bundled with

automated microscopes, but these methods are proprietary and are seldom described technically, are difficult to extend with new algorithms, and do not provide access to the raw data they collect, limiting the analysis methods available to the user.

The fundamental operations in our cell image processing are illumination correction, cell identification, and measurement of individual cells, each applied in a modular fashion. The illumination correction methods we have developed are related to methods from the MRI literature for bias field estimation, in particular those by Wells *et al.* [98] and van Leemput *et al.* [58]. We rely on automatic thresholding methods for initial nuclear foreground identification, primarily the histogram-based method introduced by Otsu [77]. To separate individual nuclei within the foreground, we borrow freely from others' work on identifying, segmenting, and separating clumped nuclei in images, especially from the work of Wahlby *et al.* [93, 94].

We introduce a new method for segmenting cells in images, in which we compute Voronoi regions of nuclei in a manifold with a metric guided by image features. Our method is a generalization of fixed-offset methods, such as those used by Perlman *et al.* [78], which do not take into account image features in locating boundaries, and watershed-based methods [70], which rely solely on image features. Our algorithm shows some similarities with Geodesic Active Contours *et al.* [19]. After cells have been segmented, we measure them in a variety of ways, to form a cytological profile for each cell. The features we have chosen have been guided by prior work on cell image analysis [13, 81].

Processing the images from a screen produces a large database of measurements for every cell in the screen, with each cell associated with one of the gene-knockdowns in the screen (in a many-to-one relationship). To make predictions about related genes and genetic function, we identify which cells have a particular phenotype, for which we sometimes make use of automatic classifiers. Our method of training these classifiers is similar to Example-Based Image Retrieval [91], implemented with GentleBoosting [33].

Automatic classification has been applied to images of cells, previously. Boland & Murphy use neural networks to classify staining of subcellular compartments in

individual cells [13]. Danckaert *et al.* [23] perform a similar task using a modular neural network and multiple images from confocal microscopy. Both of these methods have yet to be applied to questions of gene function. Support vector machines have been used by Harder *et al.* to classify fluorescently-tagged nuclei according to phase in the cell cycle in gene knockdown experiments [40, 41, 53], as part of the MitoCheck project [3]. Support vector machines with a linear decision boundary were used by Looet *al.* [64] to classify drug-treated cells versus controls; they then extract the normal to the decision boundary as a signature for the drug’s phenotypic effect, for use in characterization of drugs via similarity of effect on phenotype.

Our representation of the output of a screen as phenotype profiles, and the use of these profiles to find correlation with existing biological data to predict gene function, is unique to the best of our knowledge. Others have examined screens’ output as a distribution across gene knockdowns of phenotype-based scores [32], but not used the continuous nature of the scores directly. Rather, they interpret the scores as evidence that signalling networks within the cell are more complex and graded than usually presented.

1.3 Contributions

The primary contributions of this work are new image analysis techniques for cell images, and methods for analyzing and understanding data from large, image-based cell screens.

Accurate per-cell measurements are the foundation of our approach to predicting gene function. To improve the quality of our measurements, we must correct for biases introduced by experimental conditions and imperfections in instrumentation. We adapt techniques for bias image correction from the field of MRI processing [58, 98] to handle correction of illumination and staining variation in large screens. To improve segmentation and measurements of individual cells in images, we introduce a new method for identifying cell boundaries. This method works well even when cells are crowded, based on computing the Voronoi diagram of the cells’ nuclei under a met-

ric controlled by image features. Our illumination correction and cell segmentation methods are discussed in chapter 3, in conjunction with the full image-processing system.

We make several contributions in the area of analyzing data from large screens. In chapter 4, we apply automatic classifiers to label cells according to whether they have a particular phenotype, based on example-based image retrieval. In chapter 5, we introduce a model for the distribution of phenotype-positive and phenotype-negative cells, and use this model to score genes according to how they affect the relative number of phenotype-positive cells. We show in chapter 6 how we can use the continuum of scores to predict the underlying biology of the phenotype, and by extension, the function of the genes that most strongly affect that phenotype. We apply our methods to images from a large gene-knockdown screen in human cells [71], validating it on known phenotypes as well as identifying and characterizing several new cellular phenotypes that have not been previously studied.

We would like to emphasize that the work described in this dissertation was performed as part of the CellProfiler project [17]. CellProfiler is an open-source software platform for analyzing large collections of cell images, designed in particular for handling the output of large gene-knockdown screens. It has hundreds of users worldwide [1], and is being continually improved. The image processing methods described here have all been incorporated into this system. The phenotype-based analysis techniques we describe here are being incorporated into a companion system, CellVisualizer [2], which will be released in the coming months.

Chapter 2

Biological Background

In order to establish context and motivate this work, it is useful to provide a brief overview of gene-knockdown screens. However, a full understanding of the biological basis for such screens is not necessary to comprehend the goals and methods in this dissertation. For a more detailed discussion of gene knockdown and other types of screens, the interested reader is directed to the review by Carpenter and Sabatini [18].

2.1 Gene-Knockdown Screens

Genome sequencing projects provide the information necessary to identify almost all of the genes in several organisms, including humans. However, little is known about most of these genes and their purpose within the cell. There are several methods by which gene function can be predicted, such as looking for genes with similar sequence within and across species [14], or using similarities in the patterns of gene expression as an indicator of similarity of function [90, 95]. However, these methods are often quite noisy and are significantly removed from the true situation of interest, *in vivo* biological function. In particular, expression profiling measures the average response of cells, rather than their individual behavior [60]. Knockout organisms or cell lines, in which a gene has been excised or otherwise inactivated by direct modification of the DNA coding that gene, can be created in the lab, but only via expensive and low-throughput methods.

A powerful and high-throughput method to explore gene function is to make use of the existing machinery for gene regulation in the cell. RNA interference (RNAi), discovered in *Caenorhabditis elegans* worms in 1998 by Fire *et al.* [29], allows the suppression of a gene's activity by introduction of long double-stranded RNA (dsRNA) into an organism or cells in culture. This leads to cleavage and degradation of complementary messenger RNA before it can be translated from mRNA to its corresponding protein, silencing the expression of the targeted gene. This allows biologists to easily silence a single gene, and study the effect of its removal.

In *Drosophila melanogaster*, (fruit flies, a common model organism for genetics experiments), long dsRNA (>150-800 nucleotides) can be injected into embryos or introduced into some cell lines in culture. In mammalian cells, long dsRNA in cells induces nonspecific cell death. To work around the cells' natural defenses, dsRNAs around only 20-25 nucleotides long, called small interfering RNAs (siRNAs), are used, though in general, siRNAs are less effective than long dsRNAs. To silence a particular gene, a subsequence from that gene is chosen as the target for the siRNA. The process of RNAi gene knockdown is illustrated in figure 2-1.

A given siRNA's effectiveness varies depending on the particular genomic subsequence it targets, in ways that are not well understood. For this reason, it is common to produce several siRNA constructs targeted to the same gene, with the hope that one will be effective at reducing the expression of that gene. Further complicating matters, RNA silencing does not require a perfect match between nucleotides, so off-target hits are an issue even when a siRNA's sequence is chosen to be unique to the targeted gene. Targeting multiple, nonoverlapping subsequences of the gene also helps control for off-target effects, in which a siRNA matches to more than one gene's sequence.

An effective method for delivering siRNA into cells is via a virus engineered with DNA encoding a short hairpin RNA (shRNA), essentially an siRNA with the two complementary strands joined at one end. The virus injects its payload into the cell, which then integrates randomly into the host cell's DNA. The cell then expresses the shRNA, knocking down the targeted gene via RNA interference, and passes on the

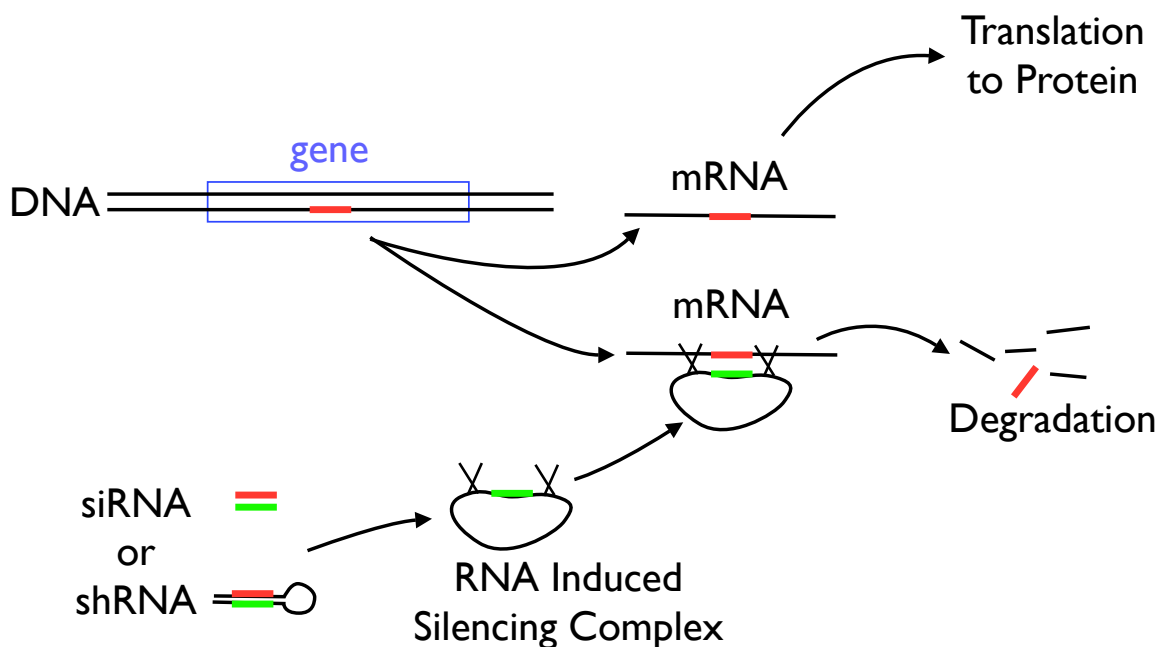


Figure 2-1: Gene transcription and translation, and gene silencing by short interfering RNA (siRNA) or short hairpin RNA (shRNA). In the normal course of events (the upper path), a gene is transcribed from the DNA into messenger RNA (mRNA), which is then translated into a protein within the cell. When siRNA or shRNA are present (the lower path), they combine with the RNA induced silencing complex. The complex cleaves mRNA with the complementary subsequence, preventing its translation to protein, and silencing the gene's expression. In the illustration above, red indicates the gene's targeted subsequence, and green the complementary subsequence.

DNA producing the shRNA to daughter cells during replication.

The process to create a virus carrying the DNA for a particular shRNA can go astray in a variety of ways, and the eventual product may not function as intended. To help account for the variability in effectiveness of viral infection and delivery of the DNA, the construct includes an antibiotic-resistance gene, which is integrated into the cell in the same way as the DNA coding the shRNA. This confers resistance on cells in which the infection by the virus and integration of the DNA are successful, allowing selection for successfully infected cells by treating them with the antibiotic.

Large libraries of viral constructs targeting every gene in the human genome, with several constructs per gene, are being produced. Much of the data used in this dissertation comes from one of the first screens performed using such a library [71],

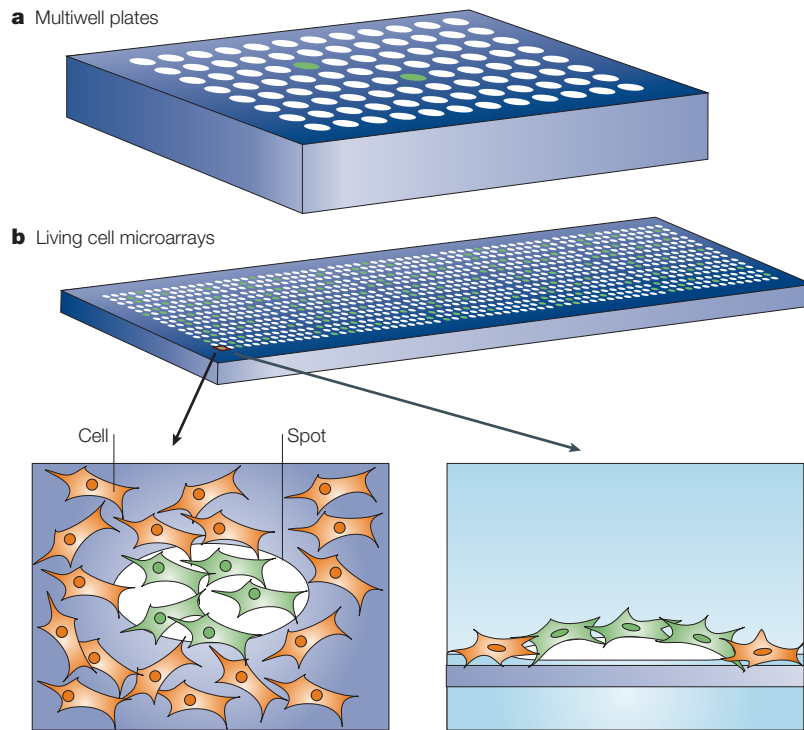


Figure 2-2: High-throughput screening formats. (a) multiwell plates. Each well contains a gene-knockdown reagent, and cells are added to the wells. The bottoms of the well are transparent, allowing cells to be imaged after fixing and staining. (b) Living cell microarrays. Reagents are printed as spots on a glass slide, then cells are grown on the surface of the slide. Cells that grow on a spot take up the reagent at that spot. Cells are fixed, stained, and imaged on the slide. (Figure from Carpenter and Sabatini [18]).

with each shRNA applied to cells in a different well of a multiwell plate, as in figure 2-2(a). We also use data from living cell microarrays, in which spots of dsRNAs are printed on glass slides, and *Drosophila* cells are grown directly on the slides, as in figure 2-2(b). At the printed spots, cells take up the dsRNA and the targeted genes are knocked down.

In both cases, after cells have grown for some time under gene knockdown conditions they are fixed and stained, then imaged with a robotic microscope. A typical image from a screen is shown in figure 2-3. Each image from this step corresponds to cells grown under a different condition, i.e., with a different knockdown. A single screen might target from a few thousand to a few tens of thousands of genes. Screens with long dsRNAs in *Drosophila* cells typically have one long dsRNA per gene, with

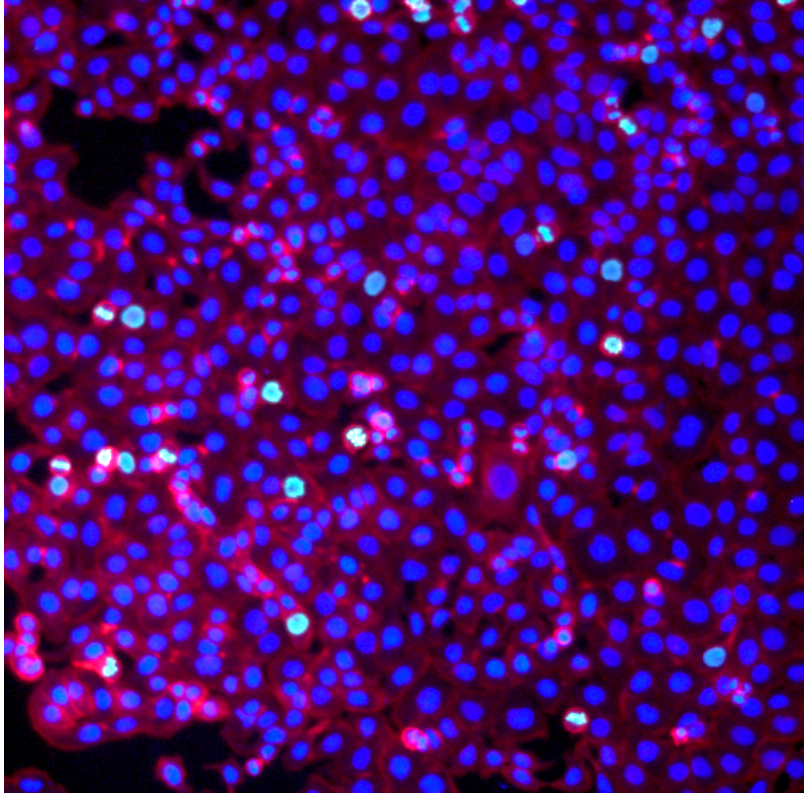


Figure 2-3: A typical image from a gene-knockdown screen. The blue channel is a stain for DNA, showing the nuclei, the red channel shows actin, a cytoskeletal protein, providing cell outlines, and the green channel is a stain for phospho-histone H3, a nuclear marker for division in the cell cycle (appearing as teal due to overlapping the nuclei). This image is of HT29 cells, a human colon cancer line, grown under a control condition (i.e., no gene is actually knocked down in these cells).

multiple replicates in each screen for quality control. In human cell screens, which use the less consistent shRNAs, each gene in the screen might be targeted by 5 or more hairpins, again with replicates to improve data quality. In virus-based screens, at least one replicate will be treated with the antibiotic to which infected cells are resistant. Comparison of the number of cells in antibiotic treated versus untreated conditions gives an estimate of infection efficiency of each virus. This allows us to distinguish knockdown of a gene critical to cell survival, in which case both conditions show few cells, from one in which the viral construct failed in some way, in which only cells treated with the antibiotic die.

It is important to note that the particular stains chosen to prepare the cells for imaging affect which visual features we have available to measure. We can probe

only a small subset of the cells' appearance in any one experiment, primarily due to the small number of stains that can be applied to cells and separated during imaging (usually three or four). For data to be collected for individual cells with the methods in the next chapter, it is necessary to at least label the nucleus of the cell, usually by staining the DNA with a fluorescent dye. If information about cell shape and size is desired, some cytoskeletal protein such as actin must be stained, as well, leaving room for only one or two more stains (with current technology). Therefore, it is important to choose which proteins should be stained carefully, based on the goals of the screen. For instance, a screen exploring genetic regulators of the cell cycle benefit most from stains for one or more proteins known to vary in concentration during the cell cycle.

Lab automation and these screens' natural parallelism makes it feasible to conduct a large screen in about a week. The next step is analyzing the images from the screen in order to extract biologically relevant information. Some large screens have been conducted using visual inspection at this step (e.g., [52]), but experiments based on this approach are prone to human error and subjective bias, are difficult to replicate, and tedious. Automatic image analysis offers a much more effective manner of analyzing these screens.

Chapter 3

From Images to Cytological Profiles

This chapter describes the image processing techniques that allow us to identify and measure individual cells in images. The steps in this process, outlined in figure 3-1, are correcting illumination and staining variation, identification of individual nuclei, from which we identify individual cells, and measurement of the cells. We demonstrate and validate a new method for correcting illumination and staining variation in large screens, and a new method for cell segmentation.

When we process images from a screen, the goal is to identify individual cells in the set of images and to measure the properties of each cell. Illumination and staining variation corrupt these measurements by biasing image intensities. To correct for illumination and staining variation, we must model and correct their effects on image formation. As most measurements depend on a cell's shape, area, or masked pixels, it is similarly necessary to find the boundaries of cells with a good level of accuracy.

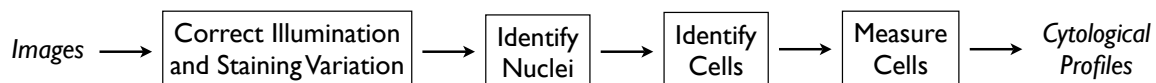


Figure 3-1: Steps in the image processing pipeline. Images are first corrected for illumination and staining variation, then nuclei and cells are identified, and cells are measured. The results are cytological profiles, one for every cell in the set of images.

With these goals, the image processing proceeds as follows. First, we separate the foreground and background of the image for nuclei via automatic thresholding. Then, overlapping and clumped nuclei are separated using a combination of morphological and intensity-based heuristics. Algorithms for nuclear segmentation and declumping have been developed previously by several researchers; we borrow freely from their methods and adapt them as needed. After nuclei are identified, we identify the cellular foreground using the same methods as for nuclei. Then, the cell body corresponding to each nucleus is identified by assigning each pixel in the cellular foreground to one nucleus, via an algorithm based on the Voronoi diagram of the nuclei in a metric space defined by image differences in the cell-body stained image. Finally, each nucleus and its cell are measured in a variety of ways, based on their shape and staining pattern. These methods are implemented as part of the CellProfiler image analysis software [17].

3.1 Illumination and Staining Variation

All physical experiments are subject to noise and biases; large screens are no exception. There are many sources of error, but the two most dominant are illumination variation, due to imperfections in the microscope's optical path or imager, and staining variation, from different concentrations of stain applied to different locations on a slide or plate.

The effect of illumination variation is apparent in the images in figure 3-2(a), in which a full set of 5600 DNA-stained images from a large screen are averaged together (see figure 2-3 for an example of a single DNA and actin stained image). Between the boundary and the middle of the image, there is a relative difference of 1.5 in mean intensity. The density of cells is nearly uniform, however, as shown in figure 3-2(b), as determined by processing images to identify and locate nuclei. Note that a flat cell distribution is not the case in all experiments, particularly well-based screens. The physical dimension of each image is small enough ($\sim 100 \mu\text{m}$) that variation of staining concentration across the image is not a reasonable explanation. Although it

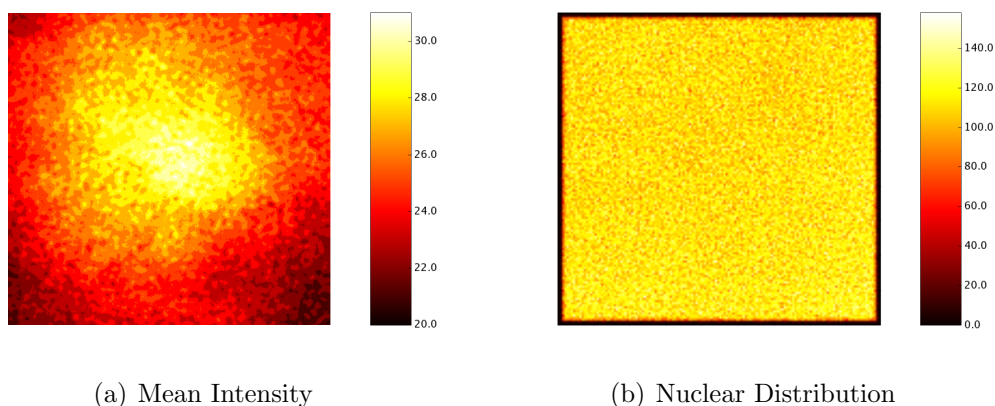


Figure 3-2: Mean intensity of the DNA-stained channel versus actual cell density, for a screen of 5600 gene-knockdowns, on the left. The mean staining image shows a variation of a factor of roughly 1.5 from the edge to the middle of the image. However, the distribution of nuclei, on the right, is almost flat, indicating that the intensity variation is due to illumination and optical path imperfections in the microscope. NB: the distribution of cells is not always uniform, as in this case.

is possible that some biological change in the cells has resulted in a different amount of stain entering the cell, or changing the affinity of stain and DNA, the far simpler explanation is that the illuminant, optical path, and image sensor of the microscope have introduced a bias in intensity.

Staining variation causes biases in images at larger scales. For example, figure 3-3 shows the median DNA staining for cells, computed separately for each image in a screen. Each imagewise median is plotted according to the physical location on the glass slide from where that image was taken. As the gene knockdowns in this experiment were randomly ordered, the large-scale variation in median DNA staining intensity is almost surely due to varying concentration of stain applied to the slide.

If illumination and staining variation are not corrected before segmenting and measuring cells, the resulting measurements are significantly degraded. In figure 3-4, histograms of DNA content for each cell in a screen (the same as produced the data in figures 3-2 and 3-3) are shown before and after corrections for illumination and staining are applied. Prior to correction, it is difficult to separate the subpopulations of cells in $2N$ (2 copies of each chromosome) and $4N$ (4 copies, prior to cellular division). After correction, these subpopulations are easily distinguished. We use

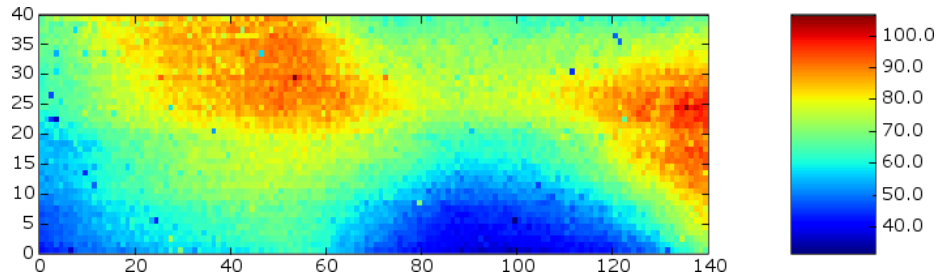


Figure 3-3: Median nuclear DNA staining intensity, calculated for each image, and plotted in a grid corresponding to the physical location on the slide where the image was captured. The large-scale variation in median nuclear DNA stain is probably due to varying DNA stain concentration.

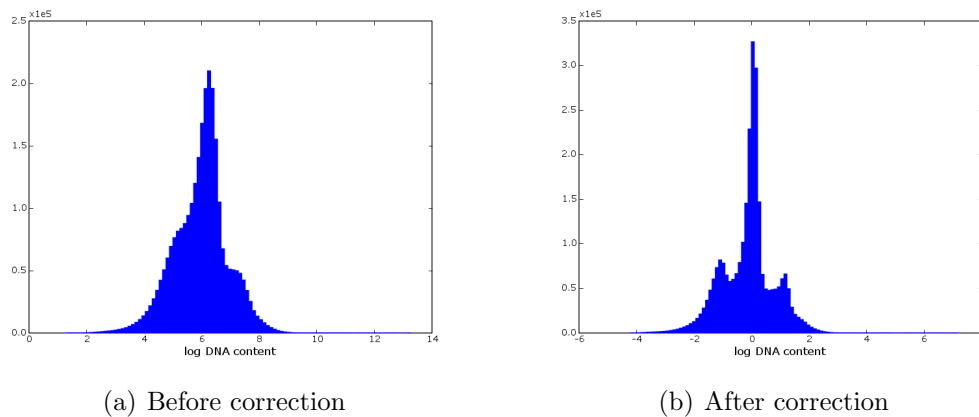


Figure 3-4: (a) Histogram of cellular DNA content (plotted on a log scale), for all cells in a screen ($\sim 14,000,000$ cells), measured without correcting illumination and staining variation. (b) The same data, after correcting illumination and staining variation.

integrated DNA staining intensity, computed within the nucleus, as a proxy for DNA content.

3.1.1 Correcting Illumination and Staining

The Physical Model of Image Formation

It is useful to consider a simplified physical model of how images of cells are formed in the experiments we are considering, in order to understand how to correct biases introduced by illumination and staining variation.

The value of a pixel is essentially a product of several variables: illumination in-

tensity, stain concentration within the image, and the stained protein’s concentration within a cell at that location. We include optical path transparency and camera sensitivity into the illumination intensity, since they act in concert and in the same manner. Our goal is to measure the amount of protein within each pixel (in arbitrary units), without the biases of the other two terms. We assume that the optical path and illuminant are fixed from image to image. We also assume that staining concentration does not vary across the image, and that in slide-based experiments, the stain concentration varies slowly between adjacent images (see chapter 2 for a discussion of the two screening formats).

In mathematical terms, we have

$$\text{image}_{x,y,i,j} \propto c_{x,y} * l_{x,y} * s_{i,j}, \quad (3.1)$$

for a pixel at x, y in an image located at physical position i, j within the experiment, with illumination function l and protein concentration at c varying with x, y , and staining concentration s varying with i, j . We wish to control for the variation in l and s , in order to measure c . Note that c and s are essentially images in this context; below we will simplify their form by modeling s via a basis and representing c via a mixture model.

This model assumes that the concentrations of proteins (and their affinity for staining) do not vary depending on image position. This is probably not strictly correct, especially since we know the density of cells varies based on image position in some cases, and this implies a biological cause that presumably could affect protein levels, as well. However, we expect this effect to be significantly smaller, in general, than that due to illumination. This is an area where further work would be of value.

Most existing approaches to correcting illumination for image-based cell screens operate on single image at a time and assume a flat distribution of cells [26, 40], usually by fitting a smooth function to the intensity image, and using that function as the illumination correction l . Unfortunately, this conflates cell density in the image with illumination intensity. Especially in well-based screens, cell density can vary

significantly within an image. Furthermore, these approaches require post-processing to adjust for staining variation. We avoid both of these difficulties with our approach.

Modeling and Correcting Illumination and Staining Variation

In order to correct illumination and staining variation, it is necessary to estimate how they vary from pixel to pixel or location to location on the slide, respectively.

We do this by randomly sampling pixels from each image, recording their log-transformed intensities in the multiple staining channels as well as their location in the image and physical layout within the experiment. The log-transformation converts the multiplicative model from equation 3.1 to an additive one. If we assume that the noise from each log-transformed term is roughly Gaussian (observed experimentally, see figure 3-5), then each pixel is drawn from a distribution of the form

$$\log \mathbf{p}_{x,y,i,j} \sim \sum_c a_c \mathcal{N}(\mu_c + \mathbf{L}_{x,y} + \mathbf{S}_{i,j}, \Sigma_c), \quad (3.2)$$

where $\mathbf{p}_{x,y,i,j,c}$ is the pixel at location x, y in image coordinates and i, j in the slide (or physical layout) coordinates. The illumination function \mathbf{L} varies with x, y (see figure 3-2). The staining \mathbf{S} varies with i, j (see figure 3-3). The pixel class c depends on where within (or without) a cell the pixel was sampled from, and could be any of the subcellular compartments (nucleus, cytoplasm, some other stained organelle) or background. The distribution of classes is modeled by the mixing coefficients a_c , and the different compartment-dependent correlations between stains by their mean intensities μ_c and covariance matrices Σ_c . Note that we treat pixels as multidimensional vectors, and will correct all staining channels simultaneously. We will set c to the number of visually distinguishable staining compartments, usually three for background, cell body, and nucleus. To make the model identifiable, we constrain \mathbf{L} and \mathbf{S} to have zero mean; this does not reduce the model’s power, as we are operating in the log-intensity domain, where linear shifts correspond to scaling the images by a positive constant.

Expectation-Maximization for Modeling Variation

Our goal is to estimate the illumination $\mathbf{L}_{x,y}$ and staining variation $\mathbf{S}_{i,j}$ in equation 3.2. We represent these as linear combinations of some smooth basis set \mathbf{B} , separating the basis elements of \mathbf{L} and \mathbf{S} for simplicity. As discussed above, each element of \mathbf{B} has zero mean, to ensure the model is identifiable. The bases are combined according to the coefficients \mathbf{f} . Rewriting equation 3.2,

$$\log \mathbf{p}_{x,y,i,j} \sim \sum_c a_c \mathcal{N}(\mu_c + \mathbf{B}_{x,y,i,j} \mathbf{f}, \Sigma_c), \quad (3.3)$$

we can then derive the equations necessary to optimize this model's parameters $\theta = \{a, \mu, \Sigma, \mathbf{f}\}$, via Expectation-Maximization [4], the standard approach with mixture models where we do not know class labels *a priori*.

For simplicity of notation, we replace $p_{x,y,i,j}$ with \mathbf{p}_n and $\mathbf{B}_{x,y,i,j}$ with \mathbf{B}_n . We write the log-likelihood of the model as

$$\ell(\theta) = \sum_n \log p(\mathbf{p}_n; \theta) \quad (3.4)$$

$$= \sum_n \log \sum_{z_n} p(\mathbf{p}_n, z_n; \theta), \quad (3.5)$$

where z_n are the hidden class labels (i.e., background plus subcellular compartment types) of the pixels, and the interior summation is over all possible labelings z_n for pixel \mathbf{p}_n . For any distribution Q_n over the z_n , we have

$$\ell(\theta) = \sum_n \log \sum_{z_n} p(\mathbf{p}_n, z_n; \theta) \quad (3.6)$$

$$= \sum_n \log \sum_{z_n} Q_n(z_n) \frac{p(\mathbf{p}_n, z_n; \theta)}{Q_n(z_n)} \quad (3.7)$$

$$\geq \sum_n \sum_{z_n} Q_n(z_n) \log \frac{p(\mathbf{p}_n, z_n; \theta)}{Q_n(z_n)}, \quad (3.8)$$

where we have applied Jensen's inequality in the last step. If we choose $Q_n(z_n)$ to be the posterior distribution of the z_n given the current estimate for the parameters

$\hat{\theta}$, then the bound is tight at $\theta = \hat{\theta}$ [79]. In the EM algorithm, we will repeatedly maximize the lower bound of equation 3.8 to find a new $\hat{\theta}_{t+1}$ based on the current $\hat{\theta}_t$.

In the E-step of the algorithm, we find $Q_n(z_n) = p(z_n|\mathbf{p}_n; \hat{\theta}_t)$, which we will write as $z_n^{(c)}$ with c the probability for each possible label.

In the M-step, we maximize the lower bound of equation 3.8 to find $\hat{\theta}_{t+1}$,

$$\theta_{t+1} = \operatorname{argmax}_{\theta} \sum_n \sum_{z_n} Q_n(z_n) \log \frac{p(\mathbf{p}_n, z_n; \theta_t)}{Q_n(z_n)} \quad (3.9)$$

$$= \operatorname{argmax}_{\theta} \sum_n \sum_c z_n^{(c)} \log \frac{p(\mathbf{p}_n, z_n = c; \theta_t)}{z_n^{(c)}} \quad (3.10)$$

$$= \operatorname{argmax}_{\theta} \sum_n \sum_c z_n^{(c)} \log \frac{p(z_n = c; \theta_t) p(\mathbf{p}_n | z_n = c; \theta_t)}{z_n^{(c)}} \quad (3.11)$$

$$= \operatorname{argmax}_{\{a, \mu, \Sigma, \mathbf{f}\}} \sum_n \sum_c z_n^{(c)} \log \frac{a_c e^{(-\frac{1}{2}(\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f})^T \Sigma_c^{-1} (\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f}))}}{z_n^{(c)} (2\pi)^{d/2} |\Sigma_c|^{1/2}}, \quad (3.12)$$

where d is the dimension of the pixel data, in this case the number of channels. To maximize this expression, we take the partial derivatives w.r.t. the elements of θ (except for a_c , as it is a special case which we will deal with separately),

$$\nabla_{\mu_c}(\dots) = \sum_n z_n^{(c)} (\Sigma_c^{-1} (\mathbf{p}_n - \mathbf{B}_n \mathbf{f}) - \Sigma_c^{-1} \mu_c) \quad (3.13)$$

$$\nabla_{\Sigma_c}(\dots) = -\frac{1}{2} \sum_n z_n^{(c)} (\Sigma_c^{-1} + \Sigma_c^{-1} (\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f})(\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f})^T \Sigma_c^{-1}) \quad (3.14)$$

$$\nabla_{\mathbf{f}}(\dots) = \sum_n \sum_c z_n^{(c)} \mathbf{B}_n^T \Sigma_c^{-1} (\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f}), \quad (3.15)$$

where we have written partials as gradients to deal with the multiplicity of dimensions and parameters for each mixture c .

The standard approach at this point would be to set these each to zero, and solve simultaneously for the update of $\hat{\theta}_{n+1}$. However, this is difficult because of the nonlinear dependence between the solutions for μ, Σ , and \mathbf{f} . To make this tractable, we will split the EM algorithm into two phases: an EM update of $\{a, \mu, \Sigma\}$ followed by an EM update of $\{a, \mathbf{f}\}$, during each of which the other variables are held constant. This is essentially coordinate ascent, with a single EM optimization along the active

coordinates. Note that we are still guaranteed to improve the log-likelihood in each pair of steps, as

$$(\nabla_{\mu,\Sigma}(\dots) = \mathbf{0} \text{ and } \nabla_{\mathbf{f}}(\dots) = \mathbf{0}) \iff \nabla_{\mu,\Sigma,\mathbf{f}}(\dots) = \mathbf{0}. \quad (3.16)$$

Before detailing the full algorithm, we return to the case of a_c . We have the constraint that $\sum_c a_c = 1$, which we deal with by adding a Lagrange multiplier term $\lambda(1 - \sum_c a_c)$ to equation 3.12. Taking the partial derivative,

$$\nabla_{a_c}(\dots) = \sum_n \frac{z_n^{(c)}}{a_c} - \lambda, \quad (3.17)$$

which we rearrange to get

$$a_c = \sum_n \frac{z_n^{(c)}}{\lambda}. \quad (3.18)$$

Summing over c , and taking into account the constraint and the definition of $z_n^{(c)}$, allows us to solve for λ

$$\sum_c a_c = \sum_c \sum_n \frac{z_n^{(c)}}{\lambda} \quad (3.19)$$

$$1 = \sum_c \sum_n \frac{z_n^{(c)}}{\lambda} \quad (3.20)$$

$$\lambda = \sum_c \sum_n z_n^{(c)} \quad (3.21)$$

$$\lambda = \sum_n \sum_c z_n^{(c)} \quad (3.22)$$

$$\lambda = \sum_n 1 = N, \quad (3.23)$$

where N is the number of pixels, which we can now substitute back into equation 3.18 to give the update rule for a ,

$$a_c = \sum_n \frac{z_n^{(c)}}{N}. \quad (3.24)$$

We now detail the full algorithm.

E-step 1

$$z_n^{(c)} = p(z_n | \mathbf{p}_n; \hat{\theta}_t) \quad (3.25)$$

M-step 1: $\{a, \mu, \Sigma\}$

$$a_c = \sum_n \frac{z_n^{(c)}}{N} \quad (3.26)$$

$$\mu_c = \frac{\sum_n z_n^{(c)} (\mathbf{p}_n - \mathbf{B}_n \mathbf{f})}{\sum_n z_n^{(c)}} \quad (3.27)$$

$$\Sigma_c = \frac{\sum_n z_n^{(c)} (\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f}) (\mathbf{p}_n - \mu_c - \mathbf{B}_n \mathbf{f})^T}{\sum_n z_n^{(c)}}, \quad (3.28)$$

where in 3.28 equation μ_c should be the value from 3.27. We then update $\hat{\theta}_t$ with a, μ, Σ from the above to get $\hat{\theta}_{t+\frac{1}{2}}$.

E-step 2

$$z_n^{(c)} = p(z_n | \mathbf{p}_n; \hat{\theta}_{t+\frac{1}{2}}) \quad (3.29)$$

M-step 2: $\{a, \mathbf{f}\}$

We repeat the update for a ,

$$a_c = \sum_n \frac{z_n^{(c)}}{N} \quad (3.30)$$

Then, for \mathbf{f} , we have from equation 3.15,

$$\left(\sum_n \sum_c z_n^{(c)} \mathbf{B}_n^T \Sigma_c^{-1} \mathbf{B} \right) \mathbf{f} = \sum_n \sum_c z_n^{(c)} \mathbf{B}_n^T \Sigma_c^{-1} (\mathbf{p}_n - \mu_c). \quad (3.31)$$

We solve this equation using the pseudoinverse of $\left(\sum_n \sum_c z_n^{(c)} \mathbf{B}_n^T \Sigma_c^{-1} \mathbf{B} \right)$, to handle cases such as an undercomplete basis in \mathbf{B} . We then update $\hat{\theta}_{t+\frac{1}{2}}$ with the new values for \mathbf{f} to get $\hat{\theta}_{t+1}$.

The summations in equation 3.31 can be rewritten in tensor algebra in a straightforward manner, allowing us to represent the set of B_n 's across all pixels as a 3-dimensional tensor ($\#$ channels \times $\#$ pixels \times $\#$ basis elements), and other el-

ements as matrices and vectors. This significantly simplifies and accelerates our implementation.

Note that we could update μ in the second M-step, as well, while holding only Σ constant, as the relation between the optimal values for μ and \mathbf{f} is linear. This would probably result in improved performance. We choose not to make this change to simplify the implementation. Sampling pixels from the thousands of images that make up a screen takes much more time than to compute the illumination and staining correction, and the algorithm has never failed to converge, so there is not much impetus for making this improvement.

This method is easily adapted to both slide-layout and plate-layout experiments. On slides, staining variation is present as a large-scale, smooth function over (i, j) . In plate-based experiments, plate-to-plate variation is modeled with a single offset per plate. If there are known sources of possible plate-pattern biases within the experiment (from moving from 96-well to 384-well plates, for example [83]), these can be incorporated into the bases as well. Illumination variation is modeled with smooth functions, usually one set per channel.

As empirical justification for modeling pixel distributions as a mixture of Gaussians, in figure 3-5 we show the distribution of pixel values (post-correction), for a few million randomly sampled pixels from a screen of 5600 gene knockdowns. Two channels for each pixel were sampled, one of DNA staining, another staining actin, a cytoskeletal protein. Three groups of pixels can be discerned, corresponding to background, cytoplasm (the cell body outside the nucleus), and the nucleus itself. The groups are roughly Gaussian. In practice, our method works well even when the pixel distributions are not well-modeled by a Gaussian.

This approach for correcting illumination and staining variation is similar to some methods of bias field correction in the field of MRI analysis [58, 98]. Our approach differs in that we have a large number of images to correct, which we reduce to a manageable number of pixels via random sampling, and the combination of staining and illumination variation. In many respects, the problems of staining variation and bias field correction can be solved by similar methods. We have explored an entropy-

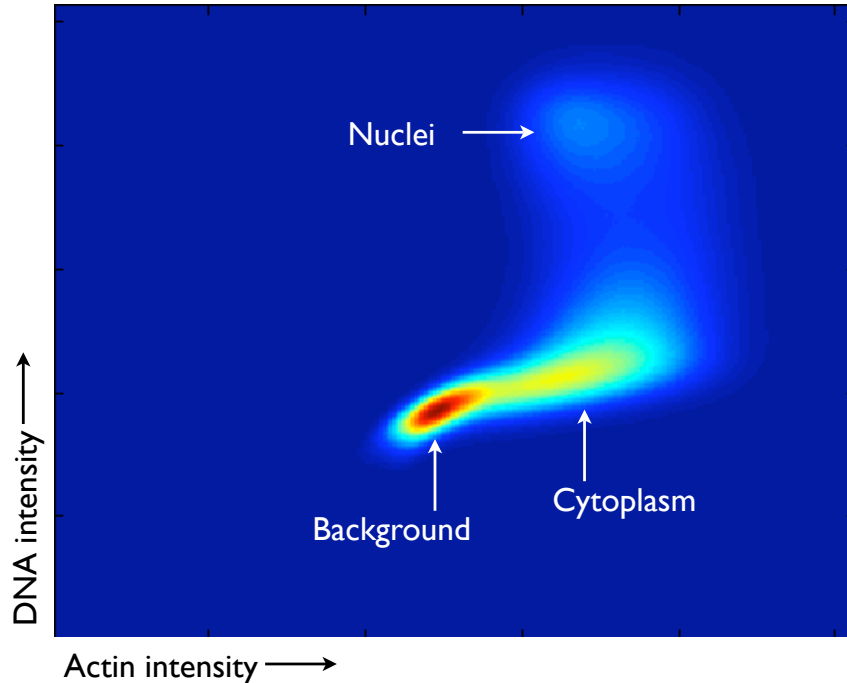


Figure 3-5: Distribution of pixel values for two staining channels, actin and DNA. A 2D histogram of pixel values, randomly sampled from a screen of 5600 gene knock-downs. The horizontal axis corresponds to the log-intensity of the actin-stained channel. The vertical axis is the log-intensity of the DNA-stained channel. Three populations are visible: background pixels (lower left) have low DNA and actin staining intensity. Pixels from the cytoplasm (lower right) have brighter actin staining, but low DNA stain. Pixels taken from nuclei (upper right) have bright DNA and actin staining. The individual groups are roughly Gaussian.

minimization approach, similar to congealing methods for bias field correction [57], but do not make use of it for reasons of efficiency.

We have not explored how well the pixel data from screens actually matches our shifted Gaussian mixture model, nor explored in depth how well we find the correct mixture components in practice. However, it has been our experience that we get nearly identical results even when the fit is incorrect, in the sense of one physical component (from figure 3-5) capturing more than one component in the model (leaving the remaining two to be modeled by a single Gaussian). This is probably because the model’s likelihood is improved if the illumination and staining corrections are more accurate, even when the mixture model is poor. This goes back to our original (and necessary) assumption, that the protein density varies independently

of the illumination and staining variation.

3.2 Identifying Nuclei

The next step in identifying cells is locating and separating individual nuclei. We identify nuclei and cells separately because nuclei are more consistent in shape, are more physically separated than their enclosing cells, and are easily and routinely stained in experiments. Methods to identify and separate individual nuclei have been developed by several researchers in the past [67, 70, 76, 93, 94], and we borrow freely from their work, modifying their approaches to suit our needs. The full explanation of our methods for identifying nuclei is given in the documentation of the CellProfiler project [1]. Here we only summarize the methods which we have used, and do not offer separate validation of these algorithms beyond that already available in the relevant literature.

We first threshold the log-transformed nuclear channel, with an automatic method such as that developed by Otsu [77], or by fitting a mixture-of-Gaussians to pixel values and classifying each pixel according to the resulting model. Note that we do not generally use the mixture model from the illumination correction step at this point, because in many cases gene knockdowns change the distribution of intensities, increasing or decreasing the amount of a particular protein in the cells, or cause nuclei and cells to spread out or shrink, changing the concentration of proteins and causing a mismatch between the screen-wide pixel distribution and that of the individual image. For these reasons, we prefer to calculate the threshold for each image independently, though we have considered using information from the illumination correction step as a prior to guide the choice of a threshold. We include an upper and lower limit on the threshold to catch outlier images, such as those with no cells.

The threshold from the automatic method is often scaled to adjust for cases where the images slightly violate the model implied by the automatic thresholding algorithm. Applying the adjusted threshold gives a foreground mask which we must then separate into individual nuclei.

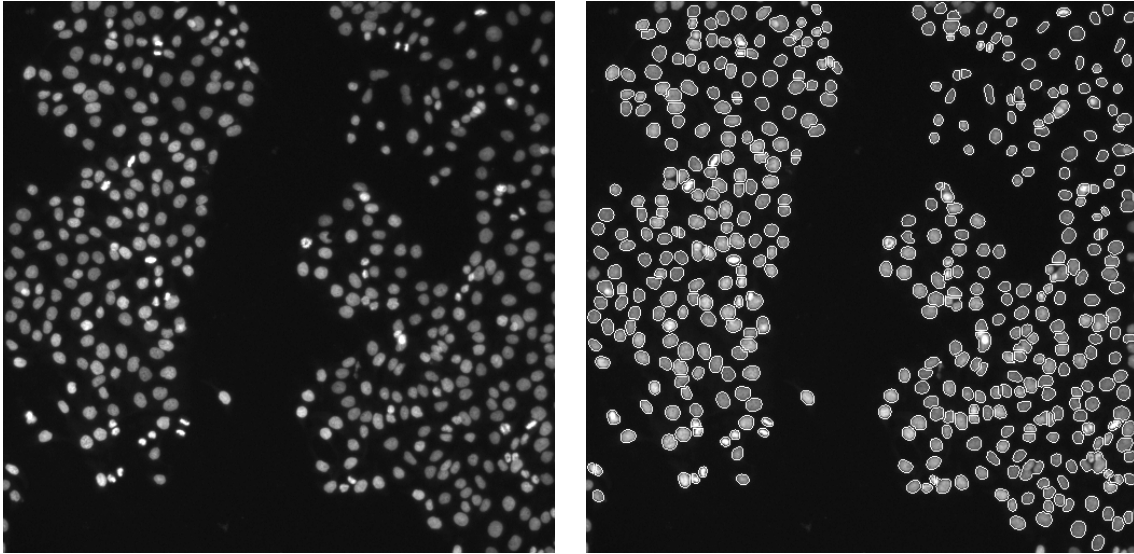


Figure 3-6: Human cells stained for DNA (nucleus), showing the nuclei. When nuclei stain uniformly, and have a smooth and round morphology, we use morphological operations to separate clumped and overlapping nuclei.

Nuclear appearance varies greatly depending on the type of screen, the cell line, the staining protocol, and other experimental factors. We have developed a variety of methods for separating nuclei, in the interest of being flexible enough to handle most cases.

If nuclei are fairly uniform in staining and round in shape, as in figure 3-6, then we usually use a morphological approach to separate clumped nuclei. We apply a distance transform to the binary nuclear mask, replacing each foreground pixel with its distance to the boundary. We then locate maxima in the distance image, and treat these as the nuclear centers. If nuclei are less uniform in shape and stain, as in figure 3-7, we smooth the nuclear stain image and locate maxima in that smoothed image, limited to the foreground pixels.

After identifying maxima, we filter them to remove those that are too close to one another (as defined by the user and adjusted from screen to screen). If two maxima are within the minimum separation, the one with larger magnitude is kept (or one is randomly chosen if they are equal).

To find the nuclear border for each maximum identified as a nuclear center, we use

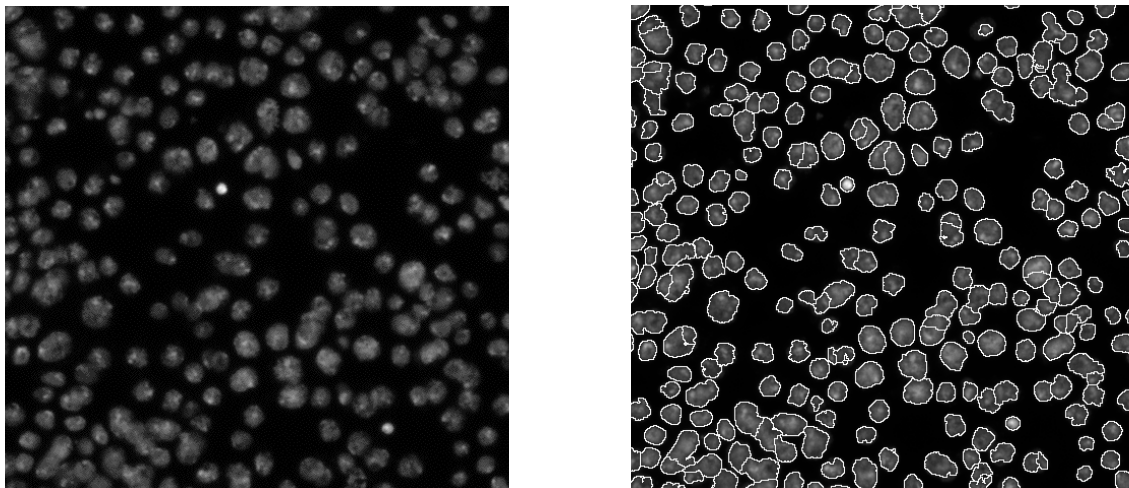


Figure 3-7: *Drosophila melanogaster* Kc cells stained for DNA, and their segmentation. For cell lines and protocols that produce nuclei with less smooth staining and without a smooth shape, we use intensity-based methods to separate overlapping and clumped nuclei. Compare to figure 3-6.

one of two methods. If nuclei show a difference of intensity between their interiors and borders under the staining protocol, we use a watershed transform of the intensity image with the centers as imposed minima. If nuclear shape is consistently round or there is insufficient information in the nuclear channel to identify boundaries, we identify the nuclei as the Voronoi regions of the filtered maxima, confined to the foreground mask.

Finally, nuclei that are above or below user-specified size thresholds are marked for rejection. Large nuclei continue as seeds to identify and separate cells, as discussed in the next section, while small nuclei are treated as noise and removed.

3.3 Identifying Cells

Once nuclei have been identified and separated, we use them as “seed” regions to identify individual cells. We make the simplifying assumption that each cell has only a single nucleus, to speed up image processing, even though there are cases where this is not true. In particular, knockdown of genes that are critical to division in the cell cycle can result in cells with multiple nuclei. Although the one-to-one

nucleus-to-cell assumption results in inaccurate measurements for the rare binucleate and multinucleate cells, this is not much of a concern. Using automatic classifiers (discussed in the next chapter), we can still identify binucleate and multinucleate cells, and can similarly keep them from polluting the quantification of other phenotypes.

The first step in identifying cells is to threshold the cell-staining channel, to get a foreground and background mask, as we did for nuclear identification. As with nuclei, this thresholding is achieved via one of several automatic techniques. Given the mask, identifying individual cells corresponds to assigning each foreground pixel to one of the previously identified nuclei. Our method is based on two heuristics: first, a cellular-foreground pixel is more likely to correspond to a nearby nucleus rather than a far one, and second, a pixel is more likely to correspond to a nucleus with few image boundaries separating them than one with many boundaries between the pixel and the nucleus.

Previous approaches to cell segmentation have been based on one or the other of these heuristics. The watershed transformation [11, 93] is guided by boundaries (peaks or valleys) in the image, but neglects spatial distance. It is therefore sensitive to small gaps in the staining of cellular borders, which are common in our experience, and lead to dramatic missegmentations in many cases. Voronoi regions of nuclei, often implemented as fixed offsets or “donuts” around the nuclei [78], do not take into account image information. Furthermore, fixed offsets inhibit analysis of cell size and morphology changes.

Our approach combines image and spatial information to segment cells. We define each cell as the Voronoi region of its corresponding nucleus in the Voronoi diagram of all of the nuclei, constrained to the cellular foreground mask, as in figure 3-8. However, the Voronoi diagram is computed in the image under a metric that balances image information and spatial distance. The metric is defined in terms of the cell-stained image \mathcal{I} and a regularization parameter λ , as

$$\mathbf{G} = \frac{\nabla \mathbf{g}(\mathcal{I}) \nabla \mathbf{g}^T(\mathcal{I}) + \lambda \mathbf{I}}{1 + \lambda}, \quad (3.32)$$

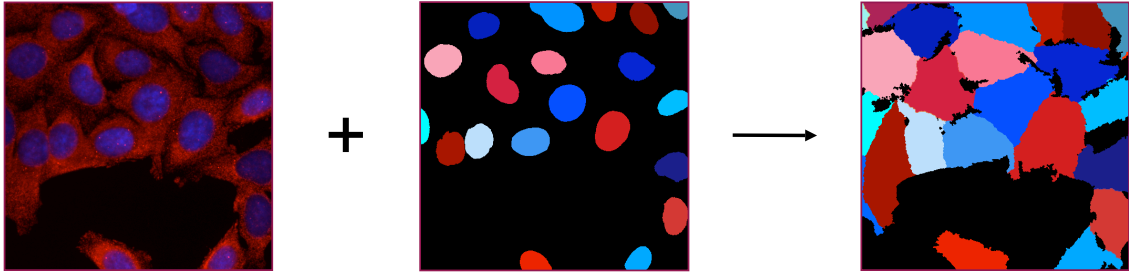


Figure 3-8: Our cell segmentation algorithm takes as input the stained image of the cell (the red channel, in the case above) and a labeling of individual nuclei. The cell-stain channel is thresholded and a single cell per nucleus is identified as the Voronoi diagram of that nucleus, computed under a metric guided by features in the cell-stain image and constrained to the thresholded foreground.

where \mathbf{I} is the 2×2 identity matrix. The function \mathbf{g} maps images to images, and in our application is generally a small-radius blur. Infinitesimal distances under \mathbf{G} are measured by

$$\|d\mathbf{x}\|_{\mathbf{G}}^2 \equiv d\mathbf{x}^T \mathbf{G} d\mathbf{x} = \frac{(d\mathbf{x}^T \nabla \mathbf{g}(\mathcal{I}))^2 + \lambda(d\mathbf{x}^T d\mathbf{x})}{\lambda + 1}. \quad (3.33)$$

The first term in the numerator of this equation, $(d\mathbf{x}^T \nabla \mathbf{g}(\mathcal{I}))^2$, increases distances measured along directions of steep gradients in $\mathbf{g}(\mathcal{I})$. The second term, $\lambda(d\mathbf{x}^T d\mathbf{x})^2$, is a weighted Euclidean distance and serves as a regularizer with strength parameterized by λ . The regularization effect can be seen by

$$\lim_{\lambda \rightarrow \infty} \|d\mathbf{x}\|_{\mathbf{G}}^2 = d\mathbf{x}^T d\mathbf{x} = \|d\mathbf{x}\|_2^2, \quad (3.34)$$

i.e., as λ increases, \mathbf{G} becomes more Euclidean, and the identified cells more like the Voronoi regions without image information.

This metric is similar to that used in Geodesic Active Contours [19], except our goal is to find the nearest nuclei (i.e., shortest paths) without crossing image boundaries, while in Geodesic Active Contours, the goal is to find shortest path following a boundary. The connection is described more fully in section 3.3.2, below.

The behavior of the metric follows our intuitions of cell appearance and shape. In the absence of information in the image, such as edges, pixels are associated with the closest nucleus. If a pixel is roughly equidistant to two nuclei, it is associated with

the nucleus which has the least change in the image between the cell and the nucleus. We demonstrate the behavior of our algorithm on synthetic data in figure 3-9, where the tendency to a more Euclidean behavior can be seen as the regularization term λ increases. The effect of image noise on our algorithm is shown in figure 3-10. Note that in the noisy examples less regularization (i.e., a lower value for λ) is needed to achieve qualitatively similar segmentations. In effect, increasing noise provides a form of regularization by decreasing the relative magnitude of the gradient near edges versus smooth (but noisy) areas of the images.

It is interesting to contrast the details of our algorithm with the probabilistic segmentation algorithm of Ljosa and Singh, which they developed to segment neurites in retinal cells [63]. They model the segmentation problem as a random walk on the pixels, with transitions steps dependent on image intensities. The main differences are that our method computes (implicit) shortest paths from nuclei to pixels, while probabilistic segmentation takes into account all paths, and our method produces a binary segmentation while theirs produces a soft assignment.

Another possible approach, not yet explored to our knowledge, would be to use affinity propagation [30] to identify cells and nuclei by treating cells as individual clusters, with affinity related to image similarity. Cluster centers could be constrained to be within nuclei, and labeling of pixels within a cell would correspond to identifying which pixels belong to a particular cluster (i.e., nucleus).

3.3.1 Validation

To validate our method, we collected a set of manual segmentations by an expert, for 16 images with roughly 80 cells per image. The manual segmentations included nuclei and cells, and were performed in conditions as close to our algorithm above as possible: first the nuclei were outlined, based solely on the nuclear image, then the cells were segmented, using only the cell image and prior nuclear segmentation.

We use these images for validation of our algorithm. For fairness in comparison we use hand-outlined nuclei (rather than automatically identified nuclei), and define the cellular foreground mask as the union of cells identified in the manual segmentation

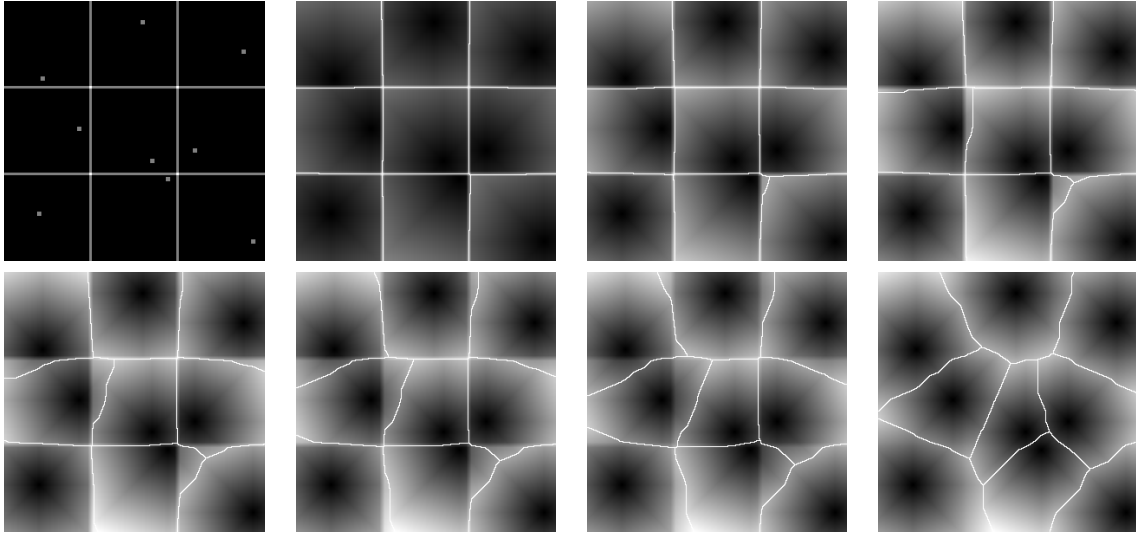


Figure 3-9: Our segmentation algorithm applied to synthetic images. The input image is in the upper left, with seed sites marked with dots. From left to right across the two rows, the resulting distances calculated with our metric are shown, with the resulting segmentation overlaid white lines, for λ equal to 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, and 3.0. The segmentation lines follow the ridges in the distances function. As can be seen, as λ increases, the segmentation approaches the Voronoi diagram of the seed regions.

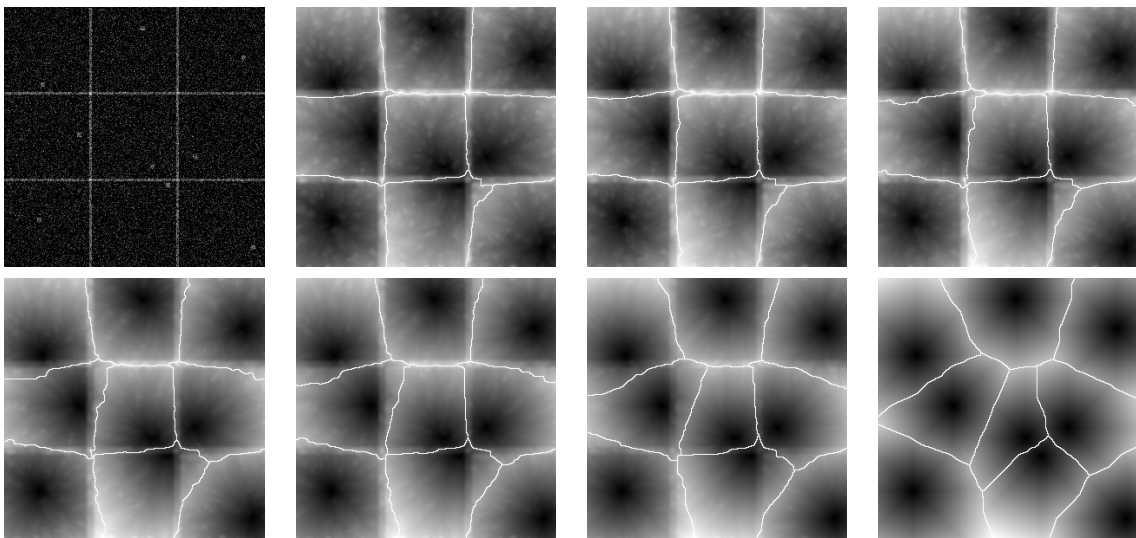


Figure 3-10: Our segmentation algorithm applied to noisy synthetic images. The same input as in Figure 3-9 is used, with zero-mean Gaussian noise with standard deviation 0.5 added to each pixel. Edges were 1.0 and background was 0.0 before noise was added. The layout is the same as in Figure 3-9, but with λ equal to 0.025, 0.05, 0.075, 0.1, 0.125, 0.2, and 0.75.

(slightly expanded), rather than from the automatic segmentation.

These choices allow us to evaluate the cell segmentation algorithm's ability to correctly locate boundaries between cells without propagating errors in nuclear segmentation or thresholding the cellular foreground. For similar reasons, we only evaluate the algorithm on borders between cells that are neighbors in both the manual and automatic segmentations. To evaluate the performance of the algorithm, we quantify the difference between such twice-occurring borders.

When evaluating the algorithm, we use the one-sided signed distance from the automatic segmentation to the manual segmentation (negative inside the lower-indexed manually labeled cell, and zero at the manual border) . We set λ in (3.32) to 0.05 times the distance between the average foreground and background pixel intensities on a per-image basis. This value for λ was found to be close to optimal in our experiments, with fairly stable behavior for a reasonably large range (within a factor of two). Our test set includes a wide variety of cell types, with different sizes and morphologies. In general, most screens would have more homogeneous data, to which λ would be tuned for the entire screen.

Sixteen images made up the test set. Each image was roughly 512x512 pixels on a side, with cells roughly 25 pixels in diameter, and 80 cells per image on average. Across the entire set, there were 21,600 pixels on a cell-cell boundary in the automatic segmentation. A histogram of their signed distances with respect to the manual segmentation is shown in Figure 3-11. Sixty-four percent (14,000) of the boundary pixels in the automatic segmentation are within 2 pixels from the corresponding manual boundary. Ninety-two percent (19,800) of the boundary pixels are within a distance of 5. The accuracy of the hand-labeling is around 3 pixels, based on the width of the marker used to outline the cells.

We illustrate the behavior of our method on some typical and worst-case cell images in figure 3-12 taken from the validation set.

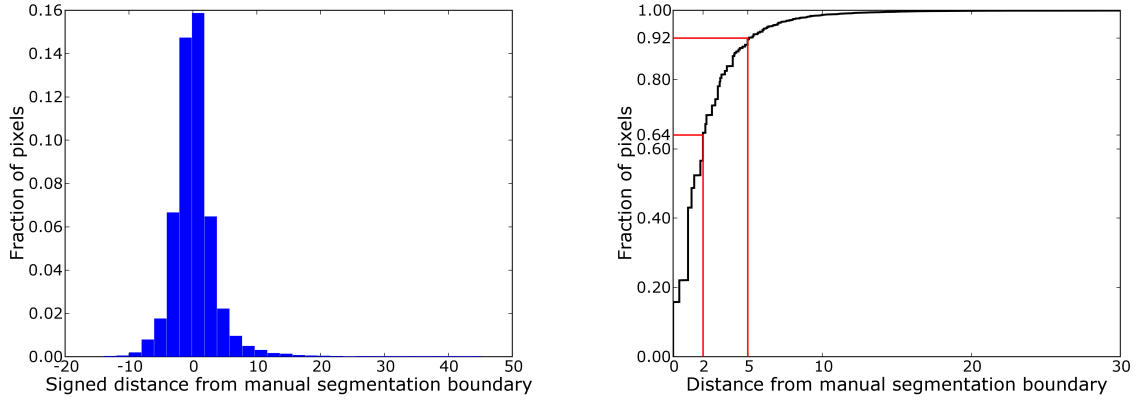


Figure 3-11: Combined histogram for the signed distances and cumulative distribution of absolute distances from automatic segmentation to manual segmentation for all sixteen images in our test set.

3.3.2 Connection to Geodesic Active Contours

Our algorithm is related to Geodesic Active Contours [19]. The full details of their work are not given here, but we discuss the connection briefly.

Active contours can be seen as finding a shortest path in a Riemannian space, where distances between pixels are defined by an edge stopping function $h : \mathbb{R}^+ \rightarrow \mathbb{R}^+$. Examining equation (8) from Caselles *et al.* [19] helps establish the similarity:

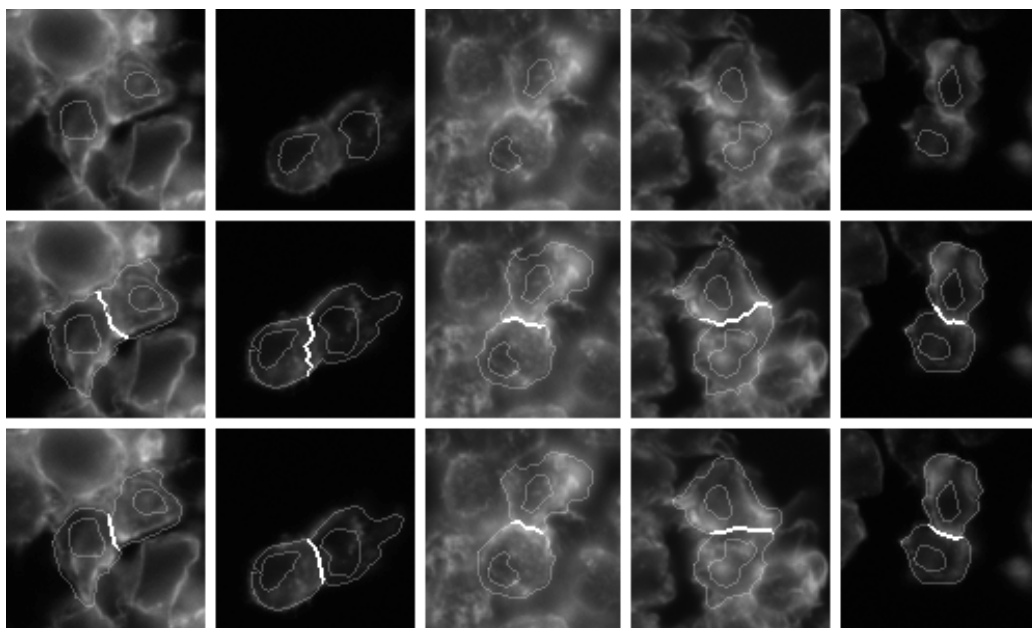
$$\text{Min} \int_0^1 h(|\nabla \mathcal{I}(\mathcal{C}(q))|) |\mathcal{C}'(q)| dq \quad (3.35)$$

where \mathcal{I} is the image, $\mathcal{C}(q)$ is the curve on image that we are minimizing over, and q is the parameter along the curve. The edge stopping function h is strictly decreasing and positive, with $h(\infty) = 0$. The effect of h 's interaction with $\nabla \mathcal{I}$ is such that the minimum curve follows larger gradients in the image.

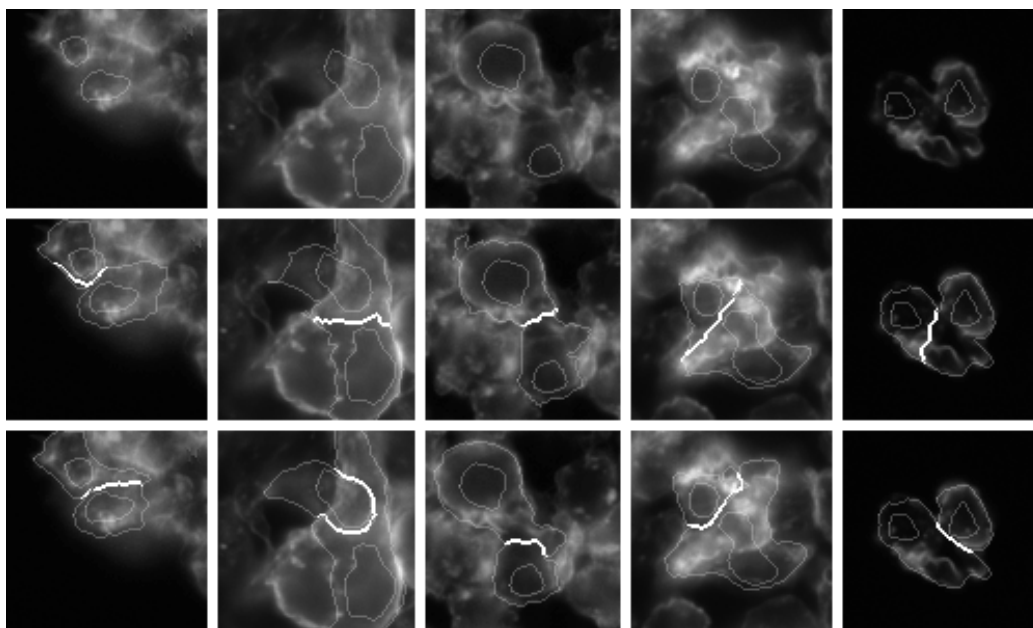
The minimization can be written as (equation (12) of [19])

$$\text{Min} \int_0^{\mathcal{L}(\mathcal{C})} h(|\nabla \mathcal{I}(\mathcal{C}(s))|) ds \quad (3.36)$$

where s is the arclength parameter for \mathcal{C} , and $\mathcal{L}(\mathcal{C})$ is the length of \mathcal{C} . Therefore, active contours can be seen as seeking a minimum length curve where the length



(a) Typical cell segmentation results.



(b) Five worst cases, compared to hand-drawn boundaries.

Figure 3-12: (a) Typical segmentation results, compared to hand-drawn boundaries. (b) Five worst cases, from our hand-drawn boundaries validation set (1280 cells), as measured by maximum distance between hand-drawn and automatically located boundaries. In both (a) and (b), the top rows show the outlined nuclei (manually segmented) overlaid on the cellular stain for two adjacent cells, the second rows the automatic segmentation boundary between the cells from our method, and the bottom row the manual segmentation boundary.

depends on image characteristics [19].

Our goal is different from active contours, since we aim to define boundaries between regions corresponding to nuclei. However, we do seek implicit shortest paths with a distance metric controlled by image characteristics. We do not wish to follow image boundaries, as in active contours, but rather to avoid crossing them. The metrics in the two approaches differ, specifically in their treatment of edges in the image: edges in the active contour setting make points along an edge closer under the Geodesic Active Contours metric, while our metric makes the points across the edge more separated. Moreover, active contours use a directionally uniform metric, while ours is not. Our metric separates points across a boundary, but points along a boundary (perpendicular to the image gradient) are essentially the same distance as equivalently spaced points in a uniform image region.

Overall, the goal is the same in the two approaches, namely, to allow the computation of inter-pixel distances that simplify the problem at hand and map it to a simpler framework. In both cases, the problem reduces to that of finding shortest paths. For active contours, this is often a jumping-off point to more powerful and efficient methods such as level-sets. In contrast, we employ Dijkstra’s algorithm [24] to find the shortest paths.

3.4 Measuring Cells

After cells have been segmented from the background and each other, they are measured along a number of axes. Many of the features we measure have been found useful in other automated cell-image analysis research [13, 81]. The full catalog of features that we measure is detailed the paper describing the CellProfiler project [17].

We measure intensity, texture, and morphology for each of the compartments of the cell: the nucleus, the cytoplasm, and the entire cell. As many biological processes are confined to inter- and intracellular membranes, we treat the nuclear and cellular boundaries as (thin) compartments in the cell, as well. Intensity measures include the total, mean, variance, minimum, and maximum intensity of each staining channel

within a cellular compartment. We measure the correlation between each pair of stains. Texture is measured according to the Haralick and Gabor texture features [34, 39], though generally not in boundaries because they are not wide enough for texture measurements to be meaningful.

Area and shape measures include such measures as total area, perimeter, eccentricity, solidity, and several Zernike moments [51]. These measurements are performed for the binary mask of each of the the cellular compartments.

To capture some information about intercellular interactions, we measure the distance to each cell’s nearest two neighbors, the angle between those neighbors, and the amount of cellular border shared with other cells. These measures are useful for identifying missegmentations from previous steps, particularly binucleate and multinucleate cells that have been incorrectly separated into multiple cells by our mononuclear assumption (cf. section 3.3).

The set of measurements that we capture has been chosen based on previous work, but this set could bias later analysis, or limit what sort of analysis is possible. Our general approach has been to include measurements we know to be useful and meaningful (cell size, total intensity of the various stains, boundary contact information, etc.) along with features previously found to be useful in automatic analysis of cell images (Zernike moments for shape, Gabor texture features for intensity, etc.). In the interest of simplicity, we do not usually adjust this set from screen to screen, nor do we try to trim it down to a smaller set in the interest of capturing more meaningful set of measurements. In general, we expect the measurement set to be “overcomplete,” and have not explored either feature selection or design of more relevant features (though the particular classifier that we use for phenotype identification, discussed in the next chapter, has inherent feature-selection properties.).

We term the combination of measurements made for a single cell its *cytological profile*. A typical screen contains tens of millions of cells, each of which yields a cytological profile. Cytological profiles are grouped by the gene knockdown or other condition that the cells were grown under. A group of cytological profiles, grouped by gene knockdown, are shown in figure 3-13.

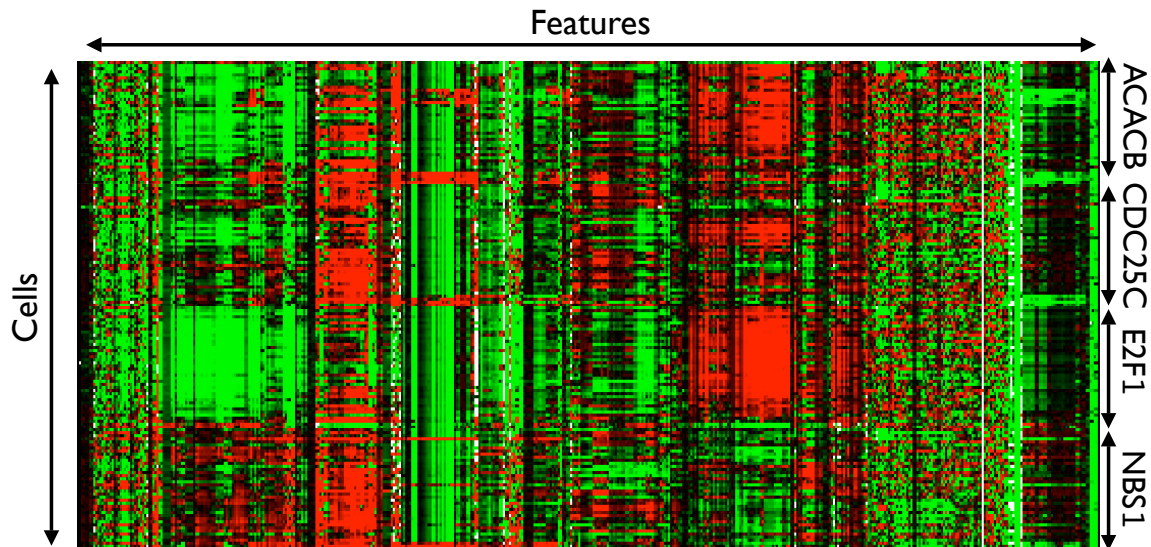


Figure 3-13: Cytological Profiles grouped by which gene knockdown was applied to the cells (captioned at right). Roughly 300 features are measured for each cell, some of which are highly correlated with other features. A typical screen includes a few hundred cells for each gene knockdown, a few thousand knockdowns, and a few million cells overall. Red indicates larger and green lower values, and values have been mean centered and normalized by standard deviation across all cells. (Image from Noa Novershtern.)

In the next chapter, we discuss how once the measurements are collected, we use them, with human guidance, to train automatic classifiers for labeling cells according to phenotype.

3.5 Summary

In this chapter we presented methods for processing a set of images from a screen in order to identify individual cells and measure a large number of features for each one, including cell shape, size, staining texture, and morphology.

As a first step in this measurement process, we introduced a simple model of image formation and presented a method to correct illumination and staining variation in large screens, via this model. We use Expectation-Maximization steps to alternately optimize the parameters of a mixture model for pixel values, in turn with the coefficients of smooth correction functions for illumination and staining variation.

Our method is applicable to very large sets of images, which we reduce to a manageable amount of data via subsampling. Our method avoids some problems with other approaches, such as conflating cell density and illumination intensity.

We introduced a novel segmentation algorithm for separating individual cells within an image. This method is based on a regularized metric that varies according to image features, such that cell boundaries are guided to edges in the image. We have validated this algorithm on a variety of cell types with highly variable morphology and appearance (cf. section 3.3.1).

Chapter 4

From Cytological Profiles to Phenotypes

In the previous chapter, we described how images are processed to identify and measure individual cells, generating cytological profiles for every cell in the images from a screen. In this chapter, we use the cytological profiles to label cells according to whether they have a particular phenotype or not.

The phenotype of interest may be defined by a single or a small set of measurements whose biological meaning is easily interpreted, such as DNA content of the nucleus (an indicator of a cell's phase in the cell cycle). Phenotypes might be visually apparent, such as crescent-shaped nuclei or cells with long projections. These phenotypes can be identified from cellular measurements, but constructing classification rules by hand is difficult. We use machine learning to create an automatic classifier for each phenotype we want to analyze, via a method similar to Example-Based Image Retrieval [91], in which a user hand-labels examples in an iterative fashion to build and improve an automatic classifier.

We can easily repeat this process for several phenotypes, after the images have been processed and cells measured. For each new phenotype, we perform the analyses in the following chapters, as well.

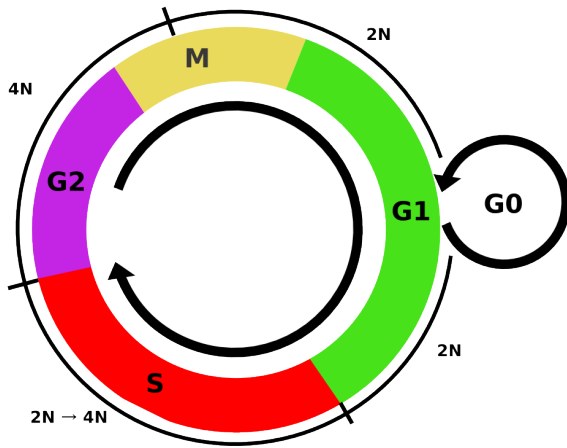


Figure 4-1: A diagram of the cell cycle. Cells in the G1 phase have two copies of each chromosome ($2N$, see text). During the G1 phase cells grow and prepare for DNA duplication and mitosis (cell division). After sufficient growth, cells enter the synthesis (S) phase, and duplicate their DNA to produce four copies of each chromosome ($4N$). They then enter another growth and preparation phase (G2), before beginning mitosis (M), in which the cell separates into two daughter cells (both $2N$). The daughter cells then return to the G1 phase. Cells can enter (temporarily or permanently) a resting phase (G0).

4.1 From Measurements to Phenotypes

Some phenotypes are defined in terms of a few measurements available directly from the image processing step, such as total DNA content or whether some protein is stained in the cell above a threshold. We call these phenotypes simple, since we can usually define rules by hand to label cells according to their phenotype. In other cases, the phenotype is visually apparent, but a connection to the measurements in the cytological profiles is difficult to establish by hand. In this case, we build an automatic classifier to label the cells, guided by human input.

4.1.1 Simple Phenotypes

In some cases, a small number of measurements may define a phenotype that is easily understood, in biological terms. For instance, the total intensity of the DNA stained channel measured within a nucleus is roughly proportional to the total amount of DNA in that nucleus. Given our segmentation of each nucleus as part of the image processing for a screen, this total amount is easily calculated for each cell, as are other easily interpretable measurements such as cellular area or mean intensity of the stained proteins in the nucleus or entire cell.

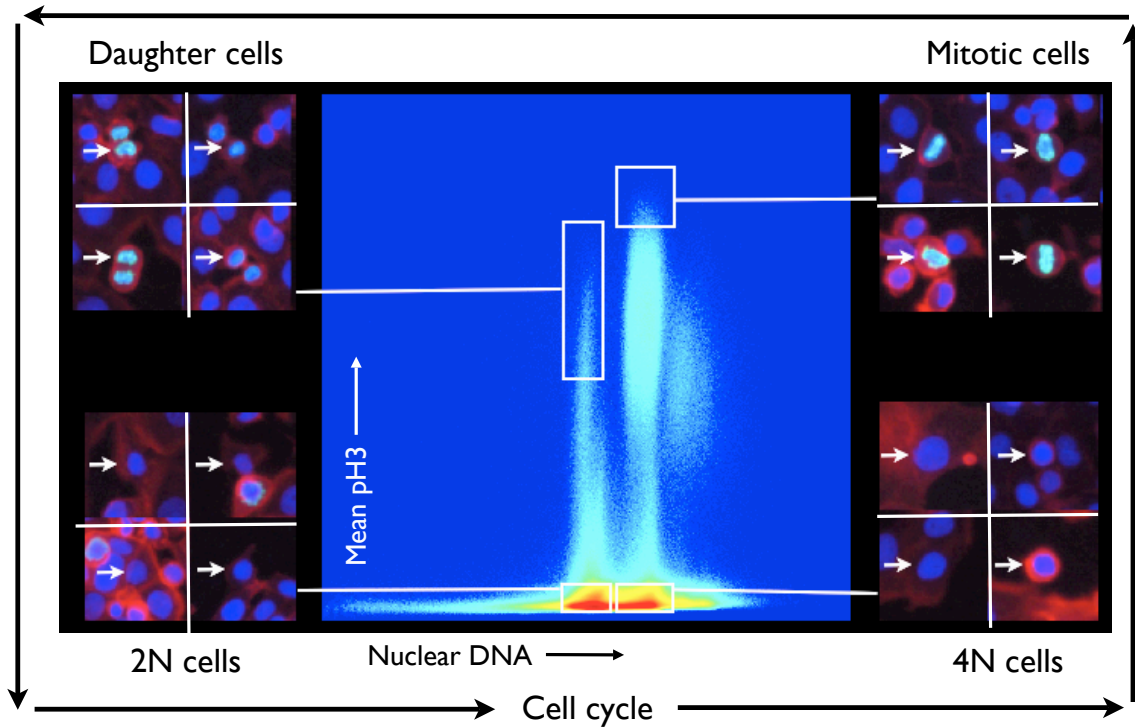


Figure 4-2: Identifying phase of the cell cycle from measurements of DNA and phospho-histone H3, a marker for mitosis. The central plot shows a 2D histogram for all cells in a screen (roughly 14 Mcells). The horizontal axis shows total DNA stain in the cell nucleus, on a log-scale. The vertical axis shows the mean phospho-histone H3 stain in the cell nucleus. Along the bottom of the histogram, two subpopulation are obvious, one corresponding to cells in G1 (lower-left) and one to G2 (lower-right). Above these are subpopulations of cells with the same amount of DNA as those below, but expressing phospho-histone H3. The upper-right subpopulation is of cells entering mitosis, and about to divide, and in the upper-left, cells that have just divided to form two daughter cells. Four randomly selected exemplars from each of the boxed regions are shown to the sides of the histogram, marked by arrows: small, rounded nuclei (G1/2N, lower left), larger nuclei that have duplicated their DNA (G2/4N, lower right), condensed nuclei, in preparation for cell division (early M-phase, 4N, upper right), and just-divided daughter cells (late M-phase, 2N, upper left). (Figure adapted from [46], data from [71])

An example of a phenotype that can be easily extracted from measurements is a cell's approximate phase in the cell cycle. In the cell cycle, shown schematically in figure 4-1, a cell begins in a growing phase (G1), duplicates its DNA during a synthesis phase (S), enters a second growth phase (G2), undergoes mitosis, in which it divides into two daughter cells (M), after which the two daughter cells return to the first growth phase (G1). If cells are stained for both DNA content and one or

more markers for specific phases of the cell cycle, it is possible to use a few simple measurements to label cells according to their phase, as demonstrated in Figure 4-2, where a two-dimensional histogram of measurements are plotted for the full set of cells in a screen. The horizontal axis shows total DNA content of the nucleus. The vertical axis corresponds to mean staining intensity in the nucleus for phospho-histone H3, a marker for mitosis (M phase). At the lower edge of the plot, there are two obvious subpopulations, corresponding to cells in G1 (left), with the normal complement of DNA (two copies of each chromosome, or $2N$), and G2 (right), which have duplicated their DNA (four copies, or $4N$). Above each of these are corresponding subpopulations, of cells in which phospho-histone H3 is being expressed, corresponding to cells that are about to divide (upper right, $4N$ and phospho-histone H3 positive), and cells that have just divided (upper left, $2N$ and phospho-histone H3 positive). During the cell cycle, cells move from subpopulation to subpopulation in the counter-clockwise direction.

Also shown in the figure are randomly selected exemplar cells from each subpopulation, bracketing the histogram. As can be seen, the $2N$ /G1 nuclei are about half the size of the $4N$ /G2 nuclei. The $4N$ cells expressing the highest levels of phospho-histone H3 (upper right) are about to separate into two daughter cells, and show the typical nuclear morphology in which the chromosomes have condensed in preparation for separation. Finally, in the upper left, the cell has divided into two daughter cells, each with the normal complement of DNA ($2N$), but still expressing phospho-histone H3. Soon after this phase, the cells return to G1 and begin the cell cycle again.

In the plot, there are several other subpopulations visible with other amounts of DNA ($8N$, $16N$, as well as $3N$ cells that are phospho-histone H3 positive). The cell line used in this screen is derived from cancer cells, and has known aneuploidy defects (abnormal chromosome number). Cells with non-integral copy numbers are present, though usually only while in the process of duplicating DNA. Some of these subpopulations are caused by these cellular defects; others are due to experimental error (during imaging, some wells were underilluminated), or image processing failures (e.g., very large nuclei are sometimes divided into two or more smaller nuclei).

4.1.2 Complex Phenotypes

In many cases, we are interested in a phenotype that is too complex to be captured in a small number of measurements, but which can be easily recognized by eye. Examples of such are shown in figure 1-2. For such phenotypes, we rely on user guidance to build a training set of positive and negative examples. The user interacts with an automatic classifier in a feedback loop, similar to example-based image retrieval [91], to improve the classifier's performance.

To train a classifier, the user starts with a few images of cells showing the phenotype of interest. They label these cells as phenotype-positive, and a few other cells (usually randomly selected) as phenotype-negative. They can then train a classifier on this limited set, and use the resulting classifier to request more phenotype-positive or -negative cells. Cells are drawn randomly from the entire screen, automatically labelled according to the current iteration of the classifier, and once a sufficient number of the requested type are found, presented to the user. Almost always, there are mislabeled cells in this set in the early stages of training a classifier. The user can correct these errors and iterate, until the classifier is sufficiently accurate.

Once the user is satisfied with the classifier's performance, it can be applied to the entire screen. The total number of phenotype-positive and phenotype-negative cells for each knockdown can be tallied. These counts can then be analyzed using methods described in the following chapters.

There are several benefits to using classifiers to label cells' phenotypes. The classifiers' greater flexibility allows us to screen for phenotypes that can not be analyzed by more traditional methods, such as by total image intensity or translocation assays [27], and pre-screen assay development is simplified, as the screened phenotype can be chosen from a much wider class. For the same reason, phenotypes can be screened with greater specificity, increasing experimenters' confidence that they are measuring what they intend [28]. Since we can train a new classifier for each phenotype of interest, we can mine a single screen multiple times, essentially re-screening for each phenotype. Finally, classifiers help adjust for errors in the image analysis phase,

e.g., missegmented nuclei and cells, provided the errors are consistent and affect the measurements such that they can be identified by the classifier (i.e., image processing errors have an identifiable “phenotype”).

There are some limitations to our approach. The iterative training we use to build the classifier results in a biased sample from the set of positive and negative cells. Therefore, we have no way of knowing that the automatic classifier has not mislabeled significant subpopulations that have not been encountered. Since most of the interesting phenotypes we find tend to be a small fraction of the total cells (in some cases $< 1\%$), we know of no efficient way to correct this. Randomly sampling and labeling the large number of cells required to build sufficient coverage of the phenotype-positive and -negative classes would be prohibitive.

Our approach to building classifiers works particularly well when the biologist performing the screen has positive and negative controls for the phenotype of interest, which is often true. In this case, cells from the positive and negative controls can be used to create the training set directly. In other cases, they may be interested in outliers and interesting phenotypes present but not directly related to the goals of the screen. We have explored some methods for finding novel phenotypes. The space of cellular measurements is of a very high dimension (> 300 features), only a few of which are relevant to any particular phenotype. We discovered the example phenotypes in figure 1-2 through a variety of methods: plotting measurements we expected to produce notable phenotypes, e.g., area, perimeter versus area, size versus intensity, etc. for individual cells as well as averages for each knockdown and looking for outliers; chance encounters, particularly during training classifiers for other phenotypes; and taking some examples from a previous analysis of the screen [71]. In general, we do not have a complete solution to the problem of phenotype discovery.

4.2 GentleBoosting for Cell Classification

GentleBoosting is an ensemble method introduced by Friedman *et al.* [33]. Ensemble classifiers are built by combining simpler classifiers to create a more powerful com-

binned classifier. The overall classifier is built in a series of rounds, during each of which a single weak classifier is added to the ensemble.

We use GentleBoosting in conjunction with decision stumps (decision trees with a single node) whose inputs are the raw measurements from our cell image processing system. We use GentleBoosting because it is easy to implement, robust to noise, and has been shown to outperform other boosting variants in similar image classification tasks in computer vision [61]. Other benefits are that the output of the boosting algorithm can be parsed by a human to a reasonable degree, and can be written into a simple and fast query for application to the full database of cell measurements.

We summarize the GentleBoosting algorithm here, briefly. This explanation is adapted from Torralba *et al.* [92].

At each step of the GentleBoosting algorithm, we seek to choose a single decision stump, or *weak learner*, to add to our ensemble classifier. The choice is made to minimize the weighted squared error,

$$J_{wse} = \sum_{i=1}^N w_i (z_i - h_m(v_i))^2, \quad (4.1)$$

where J_{wse} is the weighted squared error summed over the examples, N the number of training examples, w_i their weights, z_i their labels (-1 or $+1$), h_m is the m th weak learner, and v_i the i th training example. For decision stumps, the weak learners take the form $h_m(v_i) = a\delta(v_i^f > \theta) + b\delta(v_i^f \leq \theta)$, where a and b are the output parameters of h_m , v_i^f is the f th component of v_i , θ a threshold function, and δ the indicator function, 1 when its argument is true and 0 otherwise. The parameters of h_m are chosen by iterating over every measurement f , and every possible threshold θ for that measurement. The outputs a and b can be found via weighted least squares,

$$a = \frac{\sum_i w_i z_i \delta(v_i^f > \theta)}{\sum_i w_i \delta(v_i^f > \theta)} \quad (4.2)$$

$$b = \frac{\sum_i w_i z_i \delta(v_i^f \leq \theta)}{\sum_i w_i \delta(v_i^f \leq \theta)}. \quad (4.3)$$

The weak learner with lowest weighted error according to equation 4.1 is added to the ensemble (via summation of the h_m s). The examples weights are then updated according to

$$w_i \leftarrow w_i e^{-z_i h_m(v_i)}, \quad (4.4)$$

which reweights examples according to how well they are classified by the combined classifier after adding h_m . The complete classifier is given by $H(v_i) = \sum_m h_m(v_i)$. The example weights are initially set to balance the total weights of positive and negative examples, in accordance with the advice in Kinh and Viola [91], since our positive and negative training sets can vary in size significantly.

A system for building automatic classifiers is integrated into CellVisualizer [2], our visualization and analysis system. Users label cells via a drag-and-drop interface, shown in figure 4-3. The user can request more cells with positive or negative labels, sampled from the full screen or a specific control image (useful for seeding the classifier with a few known examples). The user can ask for cells near the decision boundary, in which case several times more cells than they request are randomly sampled, and those with the least decisive labeling are returned to the user. They can check the classifier’s behavior on a test image, seeing every phenotype-positive cell as defined by the current classifier, an especially useful operation on images of positive or negative controls. A training set typically needs to be between 100-1000 cells to train a sufficiently accurate classifier. The drag-and-drop interface makes this process very fast, on the order of a hundred cells per hour.

4.3 Summary

We have demonstrated how cells can be labeled according to phenotype, using their cytological profiles. Some phenotypes are simple, i.e. defined in terms of one or a few measurements, and rules to label cells based on these measurements can be crafted manually. In other cases, the phenotype is easily distinguished visually, but the connection to measurements from image processing is less obvious. In this case, we rely on human guidance, and use example-based image retrieval [91] to quickly

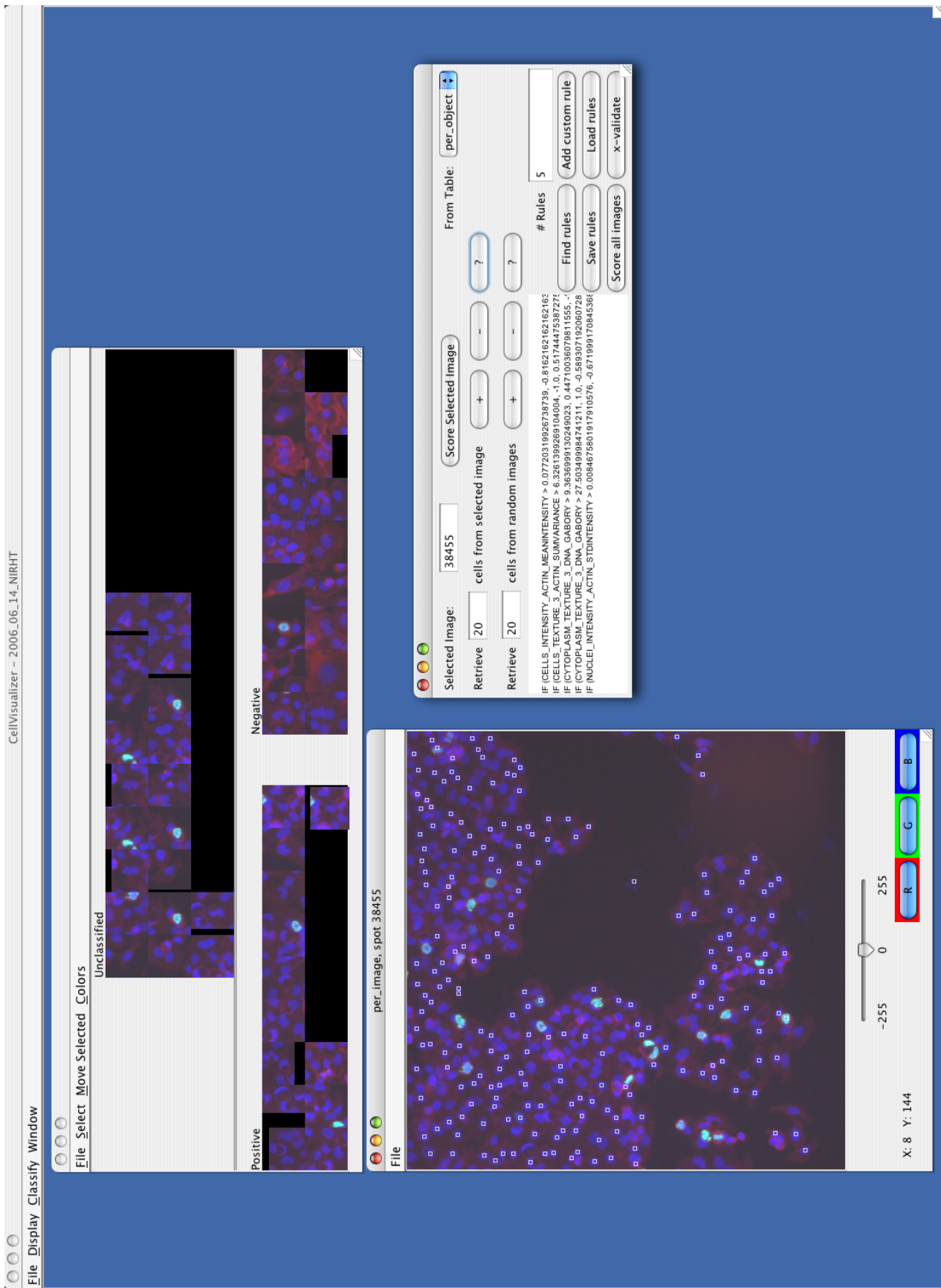


Figure 4-3: The interface for building an automatic classifier. The top window shows the drag-and-drop interface. Cells from the top, unlabelled group can be moved to either the phenotype-positive (left) or phenotype-negative (right) groups. On the lower right, an image from the screen is displayed with cells that are phenotype-positive according to the classifier, allowing the user to quickly evaluate the classifier's performance on a large number of cells. The lower right is the control window, which allows the user to request more cells, train the classifier, score a particular image, and other functions.

build a training set. We can search for multiple phenotypes within a screen, simply by defining a new classifier or set of rules for phenotype.

In our work, we use GentleBoosting and decision stumps to train the automatic classifier [33]. Others have used automatic classifiers to label individual cells according to phenotype. Boland & Murphy use neural networks to classify cells by which subcellular compartment contains a particular stained protein [13]. Harder *et al.* use support vector machines to classify cells according to their phase in the cell cycle, based on nuclear appearance only [40, 41, 53], as part of the MitoCheck project [3]. In a drug-based (rather than genetics-based) screen, Loo *et al.* [64] use linear support vector machines to classify individual cells under treatment with different drugs and chemical compounds, treating every cell under treatment as having the same phenotype. However, they then extract the normal to the decision boundary to use as a signature for each compound, rather than using the classifications directly.

We note that we do not believe that the particular choice of classifier algorithm and training method is vital to our approach; any reasonable method would work. GentleBoosting has some traits that make classification of a large number of cells more tractable, primarily its sparsity and the simplicity (in terms of translation to a database query) of the resulting classifier.

In the next chapter, we show how we can score genes according to how reducing their expression affects the number of cells showing a phenotype, and use this information to predict related genes.

Chapter 5

From Phenotypes to Phenotype Profiles and Related Genes

Once every cell in a screen has been categorized as having a phenotype or not, we calculate, for each knockdown, a score corresponding to how much knocking down that gene enhances or suppresses that phenotype. Those genes whose knockdowns significantly enhance or suppress the same phenotype are predicted to be related.

To quantify the amount a knockdown enhances or suppresses a phenotype, we fit a probabilistic model to the screen-wide set of phenotype-positive and phenotype-negative counts, and score knockdowns according to how they deviate from the model. Some variability is inherent in large screens, from slight changes in experimental conditions between knockdowns, and our model takes this into account. We use methods from Bayesian hypothesis testing, with a scoring function that represents our belief that a knockdown causes an increase or decrease in the number of phenotype-positive cells.

We combine the scores from individual knockdowns targeted to the same gene, giving us a score for each gene probed in the screen. We explicitly adjust for the possibility of off-target effects when combining scores from knockdowns; the adjustment in the scoring function has a regularizing effect that keeps us from giving too much credence to a single knockdown's effect.

The genes that most strongly affect the balance of phenotype-positive and phenotype-

negative cells are predicted to have related function. As discussed in the introduction, we use similarity of appearance as a proxy for similarity of function. Phenotypes are our vehicle for quantifying similarity of appearance.

5.1 Scoring Genes

For every gene knockdown, we have two hypotheses. Either the knockdown enhances the phenotype ($H_{e,kd}$), causing more cells to show that phenotype than expected, or it suppresses the phenotype ($H_{s,kd}$), resulting in fewer than expected cells having the phenotype. For each knockdown, we have two values: the total number of cells found in the corresponding image (t_{kd}), and the number of cells showing the phenotype for that knockdown (n_{kd}). For each knockdown, we wish to compare $P(H_{e,kd}|t_{kd}, n_{kd})$ and $P(H_{s,kd}|t_{kd}, n_{kd})$.

In order to quantitatively evaluate these hypotheses, we must find a model that fits the data in an meaningful and accurate manner.

If each cell were completely independent from all other cells in the screen, and if we assumed that for a particular phenotype, the vast majority of genes have no effect on the the fraction of cells showing the phenotype, then a good model would be to find the screen-wide fraction p of phenotype-positive cells, and treat each cell's appearance with regard to the phenotype as an independent draw from a Bernoulli random distribution with probability of success p . In this case, every knockdown could be scored according to its deviation from the Binomial distribution with parameters t_{kd} and p .

However, this simple model does not accurately represent the observed results from actual screens. In figure 5-1, the fraction of phenotype-positive cells for each knockdown in a screen are presented as histograms, for two different phenotypes. Superimposed on the histograms are analytic distributions using the binomial model above. It is obvious that the binomial model has a much narrower distribution than the actual values from the screens. Other plausible distributions, such as Poisson, produce similar results: a narrow peak near the mean with large number of knockdowns

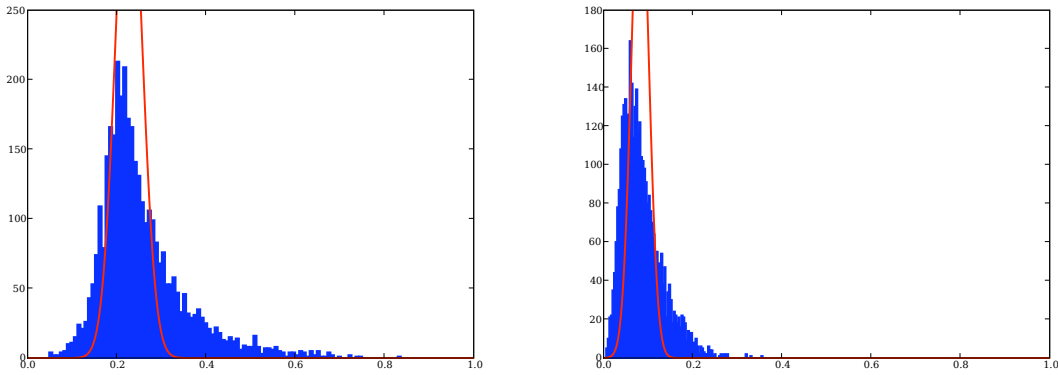


Figure 5-1: Histograms of phenotype-positive fraction of cells across gene knockdowns for a screen and a Binomial model fit to the corresponding data. The median fraction of phenotype-positive cells across knockdowns was used as the probability of success for the binomial. On the left, the phenotype is cells in G2 phase of the cell-cycle 4-1. On the right, cells with crescent-shaped nuclei.

deviating significantly above or below the mean.

The simple binomial model above assumes each cell's phenotype is independently drawn from a single distribution. Given our knowledge of experimental imperfections such as plate effects, staining variation, and other sources of error, we would expect that each group of cells in an well has grown under slightly different conditions, even after taking into account the different gene knockdowns. We believe that screens behave in a more hierarchical manner: each well in a plate or spot on a slide has a slightly different environment, and these environmental differences perturb the balance between phenotype-positive and -negative cell numbers. The single, shared binomial model is not flexible enough to encompass the variability between knockdowns.

To incorporate these effects and model the resulting overdispersion, we move to a slightly more complex model, in which the variability of the probability of a cell having the phenotype is allowed to change from knockdown to knockdown.

The Beta-Binomial Model

Let us consider multiple knockdowns, each with some number of cells that show a phenotype according to some unknown distribution. Based on experimental data, we

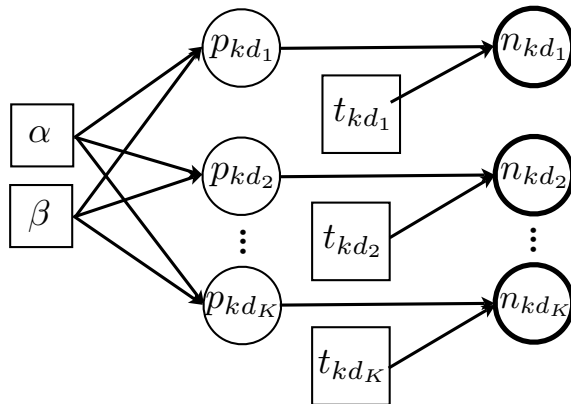


Figure 5-2: The Beta-Binomial model for a phenotypic analysis of a screen. α and β are the shape parameters of the Beta distribution, from which the probability p_{kd} of a cell showing the phenotype is drawn, independently for each gene g_i . The t_{kds} are the total number of cells for each gene, and treated as fixed parameters of the model. The observations are n_{kd} , the number of cells showing the phenotype for each gene. We fit α and β to the data, with the p_{kd} as hidden variables.

believe that the probability of a cell showing the phenotype varies from knockdown to knockdown, even for knockdowns having no effect on the phenotype. A natural way to model that variability is to assume that each cell under a particular knockdown has an identical probability of showing the phenotype, but that this probability is drawn independently for each knockdown from some overarching distribution. The Beta-Binomial is an example of such a model. Under this model, for each knockdown a probability of success p_{kd} is drawn independently from a Beta distribution. Every cell grown in that condition has probability p_{kd} of showing the phenotype, and the total number of phenotype positive cells for that knockdown, n_{kd} , is drawn from a Binomial distribution.

This model, first proposed by Skellam [88], can be written as

$$P_{\text{BetaBinomial}}(n|\alpha, \beta; t) = \int_0^1 P_{\text{Binomial}}(n|p; t) P_{\text{Beta}}(p|\alpha, \beta) dp \quad (5.1)$$

$$= \binom{t}{n} \frac{B(\alpha + n, \beta + t - n)}{B(\alpha, \beta)} \quad (5.2)$$

where n is the number of successes (phenotype positive cells), t the number of trials (total number of cells), and $B(\alpha, \beta)$ is the Beta distribution with shape parameters

α and β . The Beta distribution is flexible enough to exhibit a variety of shapes on $[0, 1]$. The full probability for the data from a screen is just the product of equation 5.2 evaluated for each knockdown,

$$P(\mathbf{n}|\alpha, \beta; \mathbf{t}) = \prod_{i=1}^K P_{\text{BetaBinomial}}(n_{kd_i}|\alpha, \beta; t_{kd_i}), \quad (5.3)$$

where \mathbf{n} and \mathbf{t} are the full set of phenotype-positive and total cells across the knockdowns $kd_1 \dots kd_K$. Note that we are assuming that there are few outliers, i.e., knockdowns that significantly affect the relative number of phenotype-positive and -negative cells, and these outliers do not significantly skew the fit of the model to the data.

Our goal is to find genes that affect the relative number of phenotype-positive cells, and we are less concerned with overall cell viability. For this reason, we do not try to model the total number of cells, and treat the t_{kd} as fixed parameters, rather than random variables. The resulting model is shown in graphical form in figure 5-2.

The integral in equation 5.2 can be computed in closed form, and fitting the model to data with the maximum likelihood estimate is feasible with standard numerical techniques. As in Lowe [65], we reparameterize the Beta function's α and β with

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \nu = \frac{1}{\alpha + \beta}, \quad (5.4)$$

such that μ is the expected probability of the phenotype drawn from the Beta distribution, and ν is a term corresponding to the “width” of the Beta distribution. We form estimates $\hat{\mu}$ and $\hat{\nu}$ by maximizing the full screen's log-likelihood with standard numerical optimization. We use the screen-wide fraction of phenotype-positive cells as the initial guess for μ , and a fixed initial guess of 0.5 for ν .

In figure 5-3, we show the new model fit to the data previously discussed in figure 5-1. The improvement is clearly apparent.

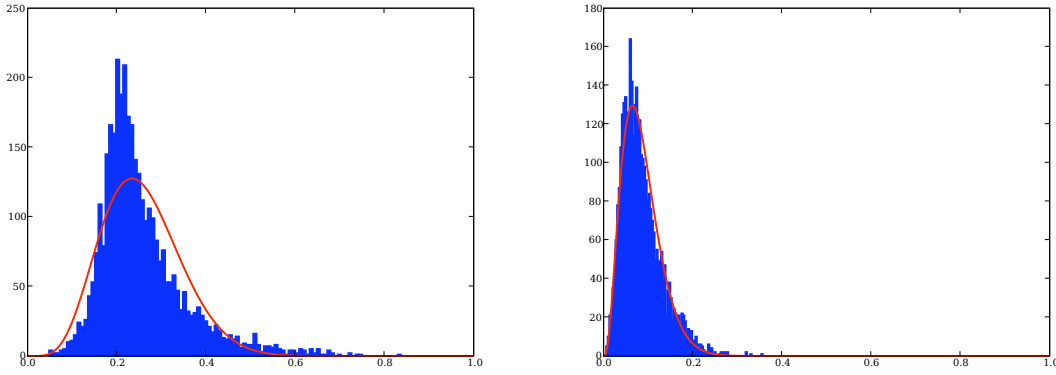


Figure 5-3: Histograms of phenotype-positive fraction of cells across gene knockdowns for a screen and a Beta-Binomial model fit to the corresponding data. On the left, the phenotype is cells in G2. On the right, cells with crescent-shaped nuclei. Compare to figure 5-1. Note that although the fit to the distribution on the left is not necessarily better under any particular measure of goodness-of-fit, it results in far fewer outliers.

Comparing Hypotheses with the Beta-Binomial Model

With the estimates for the parameters of the Beta-Binomial model, we now wish to compare the two hypotheses, $P(H_{e,kd}|n_{kd}; \hat{\mu}, \hat{\nu}, t_{kd})$ and $P(H_{s,kd}|n_{kd}; \hat{\mu}, \hat{\nu}, t_{kd})$ (enhancement and suppression, respectively) for each knockdown. We rewrite these probabilities via Bayes' Rule,

$$P(H_{e,kd}|n_{kd}; \hat{\mu}, \hat{\nu}, t_{kd}) = \frac{P(n_{kd}|H_{e,kd})P(H_{e,kd})}{P(n_{kd})} \quad (5.5)$$

$$P(H_{s,kd}|n_{kd}; \hat{\mu}, \hat{\nu}, t_{kd}) = \frac{P(n_{kd}|H_{s,kd})P(H_{s,kd})}{P(n_{kd})} \quad (5.6)$$

(temporarily dropping the model parameters for clarity).

For reasons that will be explained in the next chapter, we do more than simply decide between the two hypotheses $H_{e,kd}$ and $H_{s,kd}$ for each knockdown. Instead, we calculate the log-likelihood ratio of the hypotheses, and use it as a scoring function for each knockdown. Thus, for each knockdown we compute

$$\log \frac{P(H_{e,kd}|n_{kd})}{P(H_{s,kd}|n_{kd})} = \log \frac{P(n_{kd}|H_{e,kd})}{P(n_{kd}|H_{s,kd})} + C, \quad (5.7)$$

where C is a constant offset equal to $\log \frac{P(H_{e,kd})}{P(H_{s,kd})}$, which we will take to be zero (i.e., enrichment and suppression are equally likely *a priori*), for now.

How are we to interpret the conditional probabilities $P(n_{kd}|H_{e,kd})$ and $P(n_{kd}|H_{s,kd})$? What does it mean for the hypothesis $H_{e,kd}$ to be true? We interpret $H_{e,kd}$ as an event that the hidden parameter p_{kd} , the probability of expressing the phenotype of interest, is from a Beta distribution with larger expected value than $\hat{\mu}$ estimated above (but the same $\hat{\nu}$), and similarly for $H_{s,kd}$. Under this interpretation, the conditional probabilities can be expressed as,

$$P(n_{kd}|H_{e,kd}; \hat{\mu}, \hat{\nu}, t_{kd}) = \frac{1}{Z_e} \int_{\hat{\mu}}^1 P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm \quad (5.8)$$

$$P(n_{kd}|H_{s,kd}; \hat{\mu}, \hat{\nu}, t_{kd}) = \frac{1}{Z_s} \int_0^{\hat{\mu}} P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm, \quad (5.9)$$

where we have assumed that $\hat{\nu}$ is unchanged under the enriched and suppressed hypotheses, we have placed a uniform (uninformative) prior on values for the expectation parameter m of the Beta Binomial under each hypothesis, and Z_e and Z_s are normalizations that constrain $P(n_{kd}|H_{e,kd}, n_{kd} = \hat{\mu}t_{kd}; \dots) = \frac{1}{2}$ (i.e., if n_{kd} is its expected value from the Beta-Binomial model, there is no evidence for or against enhancement or suppression). We combine these equations with equation 5.7, to obtain

$$\log \frac{P(H_{e,kd}|n_{kd})}{P(H_{s,kd}|n_{kd})} = \log \frac{\int_{\hat{\mu}}^1 P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm}{\int_0^{\hat{\mu}} P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm}. \quad (5.10)$$

Using this model and the equations for evaluating relative enrichment and suppression, we can score every knockdown in a screen.

5.1.1 Scoring Genes from Gene Knockdowns

As discussed in section 2.1, there are usually 5-10 unique knockdowns (in the form of short hairpin RNAs) targeted to each gene, across the set of genes screened in our data. In the ideal case, we could combine the evidence from each knockdown targeted

to a gene as the sum of log-likelihood scores for that gene’s knockdowns, as

$$\sum_{kd \in \text{kds}(A)} \log \frac{\int_{\hat{\mu}}^1 P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm}{\int_0^{\hat{\mu}} P_{\text{BetaBinomial}}(n_{kd}|m, \hat{\nu}; t_{kd}) dm}, \quad (5.11)$$

where $\text{kds}(A)$ is the set of knockdowns targeted to a particular gene A , and the individual knockdowns in $\text{kds}(A)$ are assumed to be independent, given the targeted gene. For the data and screen we present here, this is a reasonable assumption, as a design goal when constructing the shRNAs for this screen was that hairpins targeted to the same gene should be well-separated from each other along the gene’s sequence.

However, equation 5.11 neglects an important consideration in knockdown screens: the correspondence between knockdowns and genes is not perfect, nor are knockdowns always effective. In particular, shRNA screens are subject to a large percentage of knockdowns that fail to reduce the targeted gene’s expression level, as well as mismatches resulting in reducing the expression of some other gene, so-called off-target effects. In fact, a single hairpin may suppress expression of multiple genes, each to a different level. In the library of knockdown vectors from which the data presented here was generated, about 40% of hairpins reduce their intended gene’s expression significantly. Even when knockdown from a “good” hairpin produces a distinct phenotype, about one tenth of the time that phenotype is due to an off-target effect, i.e., some other gene’s expression being reduced as well [82]. To account for the possibility of off-target effects, we adjust our scoring function to account for this disconnect between the individual knockdowns and their targeted genes.

From the information above (40% working rate, 10% repeatability of phenotypes in working knockdowns), we can estimate the probability that a gene’s successful knockdown (without any off-target effects) produces the same effect on cells as one of the hairpins targeting it as $(.4)(.9) + (.6)(.5) \approx \frac{2}{3}$, where we have assumed that non-working hairpins enrich or suppress the phenotype of cells randomly and in an unbiased way. We use this value as we combine posteriors from individual knockdowns into a score for the gene they target, while taking into account off-target and non-

working hairpins, as

$$\log \frac{P(H_{e,A}|n_{kd})}{P(H_{s,A}|n_{kd})} = \sum_{kd \in \text{kds}(A)} \log \frac{P(H_{e,A}|H_{e,kd})P(H_{e,kd}|n_{kd}) + P(H_{e,A}|H_{s,kd})P(H_{s,kd}|n_{kd})}{P(H_{s,A}|H_{s,kd})P(H_{s,kd}|n_{kd}) + P(H_{s,A}|H_{e,kd})P(H_{e,kd}|n_{kd})} \quad (5.12)$$

where A is some gene, $H_{e,A}$ the hypothesis that successfully knocking down just A enriches the phenotype, $\text{kds}(A)$ is the set of knockdowns targeted to A , and we have $P(H_{e,A}|H_{e,kd}) = P(H_{e,A}|H_{e,kd}, n_{kd}) = \frac{2}{3}$ and $P(H_{e,A}|H_{s,kd}) = P(H_{e,A}|H_{s,kd}, n_{kd}) = \frac{1}{3}$, and similar values for $P(H_{s,A}|\dots)$.

The additional terms in the numerator and denominator of equation 5.12 act as a regularizer. For a single knockdown, when $P(H_{e,kd}|n_{kd}) \approx 1$, the maximum effect on the targeted gene's score from that knockdown is $\log \frac{P(H_{e,A}|H_{e,kd})}{P(H_{s,A}|H_{e,kd})}$. Similarly, when $P(H_{s,kd}|n_{kd}) \approx 1$, the effect on the score is bounded below by $\log \frac{P(H_{e,A}|H_{s,kd})}{P(H_{s,A}|H_{s,kd})}$. Thus, the probability of a phenotype presenting due to an off-target effect provides an upper and lower bound on the evidence from any single knockdown. When $P(H_{e,kd}|n_{kd}) = P(H_{s,kd}|n_{kd})$, i.e., there is no real evidence for enhancement versus suppression of the phenotype, the hairpin's score does not affect the gene's score, as expected.

It remains to estimate C in equation 5.7, which is straightforward using controls, or from the median score across the knockdowns, under the assumption that most genes have no effect on a particular phenotype. We leave $C = 0$ in this work, as it does not deviate too far from that value in our experiments, regardless of the method used to calculate it.

5.1.2 Phenotype Profiles and Predicting Related Genes

We refer to the set of genes and their scores from equation 5.12 as a phenotype profile. For each phenotype we analyze in a screen, there is a corresponding phenotype profile. The parallel with expression profiles is intentional. In an expression profile, genes are "scored" based on their relative expression levels between two conditions, while in phenotype profiles, we score genes based on how changing their expression levels changes relative number of phenotype positive cells. The connection between

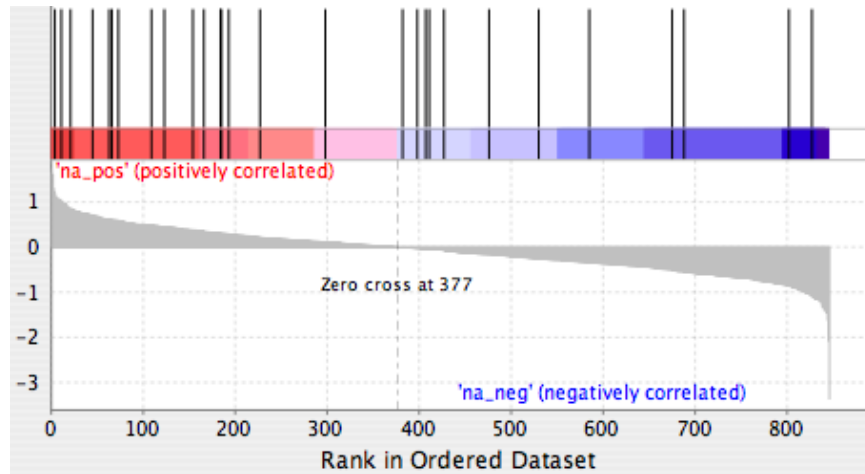


Figure 5-4: Phenotype profile for 4N phospho-histone H3 negative cells. The genes are sorted by decreasing score, left-to-right, i.e., genes that enhance the phenotype are on the left. Marked on the upper view of the profile are genes annotated as being cell-cycle related [15]. This set is significantly biased to the left, i.e., genes that cause G2 phase cells (p-value of 0.018 from permutation testing)

expression and phenotype profiles is explored in the next chapter.

We show an example of a phenotype profile in figure 5-4 for the phenotype of 4N phospho-histone H3-negative (G2 phase) cells, as defined by gating the lower-right subpopulation in figure 4-2. Genes annotated as being involved in the cell cycle are marked in the (sorted) phenotype profile, and these genes are significantly biased towards larger positive scores (p-value of 0.018, as calculated by methods discussed in the next chapter), indicating that such genes cause an increase in cells in the G2 phase, validating the phenotype as cell-cycle related. In the screen, knockdowns for 847 genes pass quality control (e.g., filters for successful infection, cf. chapter 2).

The genes that score most strongly for a particular phenotype are those that we predict to be related through some common biological mechanism or function. Our work does not allow direct predictions of which genes might interact directly due to the gulf between cellular phenotypes and genetic networks. Rather, our methods should be viewed as an attention focusing device and generator of hypotheses for more direct investigation in the laboratory.

5.2 Summary

In this chapter, we have introduced methods for scoring individual knockdowns via a screen-wide probabilistic model of the distribution of phenotype-positive and -negative cells. We choose a Beta-Binomial model, as it allows us to account for the natural variation in the fraction of cells showing a phenotype, separate from the effects of the actual gene knockdowns.

Our scores are represented as the log likelihood ratio of two hypotheses, that a knockdown enriches the phenotype or that it suppresses it. We do not use the scores to choose a particular hypotheses, but rather maintain them as indicators of belief. This allows us to combine scores from knockdowns into scores for genes in a straightforward manner. Our method for combining knockdown scores explicitly includes the possibility of off-target effects. Modeling off-target effects can be seen as acting as a regularizer on knockdown scores as they are combined into gene scores, preventing any single knockdown from unduly affecting its targeted gene's score.

We refer to the set of genes and their scores from a single phenotype as a phenotype profile. This representation is similar to that proposed by Friedman and Perrimon [32]. The intention, there and here, is to represent continuous information about screening data, whether motivated by a view of genetic interaction networks as more graded than simple on/off interactions, or, in our case, by the noisy readouts of the effects of genes as modulated through a particular phenotype.

Our model for scoring allows only two possibilities for each knockdown: enrichment or suppression of the phenotype. We assume that most knockdowns do not actually affect the balance of phenotypes, implying that most log-likelihood ratios will be near zero. This is the case, but exploring a model with an additional null-effect class would be valuable. We also adjust the priors of the two hypotheses such that knockdowns with the expected number of phenotype-positive cells have a zero log-likelihood ratio. Alternatively, these priors and other model parameters could be estimated directly from the data via a mixture model. This would likely be effective for a model that included a null-effect component, as well.

In the next chapter, we use phenotype profiles to make predictions about the biological mechanisms of that phenotype, and by extension, the genes that most strongly affect it.

Chapter 6

From Phenotype Profiles to Biological Function

We have shown how to score genes according to their effect on a phenotype, to form a phenotype profile of the genes. We predict that the highest scoring genes are more likely to be related, in terms of gene function. If the phenotype being scored is biologically interpretable, the set of high-scoring genes may be sufficient to suggest biological experiments to validate the prediction that they are related. However, there are cases when knowledge of the phenotype is quite limited. It may have been encountered during a screen, and while easily identified, nothing more may be known about the phenotype. Alternatively, some of the biological basis of the phenotype might have been established, but further insight is desired.

In this chapter, we show how we can establish connections between phenotype profiles and existing biological data to predict the biological basis of a phenotype, and by association, the biological function of the genes that affect that phenotype.

6.1 Methods for Connecting to Existing Biological Knowledge

Recall from the previous chapter that a phenotype profile is the set of scores,

$$\log \frac{P(\text{gene enhances phenotype})}{P(\text{gene suppresses phenotype})}, \quad (6.1)$$

estimated for each gene in the screen, and computed in terms of models for the distribution of phenotype-positive cells, and the connection between knockdowns' and their targeted genes' effects on the phenotype of interest.

We will treat these scores similarly to expression profiles, which represent relative levels of genes or gene products in cells (e.g., DNA, mRNA, proteins. see chapter 2), comparing two different conditions, often control and perturbed (such as by a drug) states. In our case, we have measured the change in cells due to a change in a gene's expression, rather than the change in a gene's expression due to a change in cells. The interpretation of the scores is the same, however: a large score for a gene indicates a probable connection between the gene and the cellular change.

To biologically interpret phenotype profiles, we will use existing tools for correlating gene scores and existing datasets. We will explore two tools, here, Gene Set Enrichment Analysis [89] and the Connectivity Map [55]. There are several other tools available that could be applied, as well, most geared toward helping to interpret or make predictions from expression profiles.

Gene Set Enrichment Analysis, or GSEA, is based around a large collection of gene sets. Each gene in a set shares a common trait with other genes in that set, such as common biological function, proximity on a chromosome, common regulatory motif, shared membership in a particular genetic pathway, and so on. These sets have been gathered from a variety of sources, including computational analysis of genome and expression data, as well as curation from published papers. At the current time, GSEA includes a few thousand gene sets, with more added regularly.

GSEA's input is an expression profile. For each set S in its database, GSEA

computes an enrichment score ($ES(S)$) relative to that profile, based on a weighted form of the Kolmogorov-Smirnov statistic,

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_r}, \text{ where } N_r = \sum_{g_i \in S} |r_j|^p \quad (6.2)$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{N - N_S} \quad (6.3)$$

$$D(S, i) = P_{\text{hit}}(S, i) - P_{\text{miss}}(S, i) \quad (6.4)$$

$$ES(S) = \operatorname{argmax}_{D(S, i)} |D(S, i)|, \quad (6.5)$$

where g_j and r_j are the j th gene and its score in the profile, respectively, S the gene set being scored in GSEA, N_S the number of genes in the profile that are also in S (note that expression and phenotype profiles are not always over the same set of genes). When $p = 0$, the enrichment score is the standard Kolmogorov-Smirnov statistic. When $p = 1$, the enrichment score weights the genes in S by their score in the profile, normalized by the sum of correlations over all the genes in S , essentially a weighted correlation. As suggested in the original work on GSEA [89], we use $p = 1$. To assess significance, GSEA permutes the gene labels g_j relative to the scores r_j many times, and recomputes $ES(S)$ to estimate the null distribution.

Returning to the 4N phospho-histone H3-negative (G2 phase) phenotype from the previous chapters, we show the GSEA result for a set of genes annotated as cell-cycle related by Brentani *et al.* [15] in figure 6-1 (an expanded form of figure 5-4. The elements of S that are also in the phenotype profile are marked by the vertical black bars overlaid on the color-coded phenotype profile. The enrichment score $ES(S)$ (the maximum deviation of the running enrichment score $D(S)$) is about 0.43, which has a p-value of 0.018 from permutation testing with 10,000 permutations. This gene set is significantly biased towards genes that cause an increase in G2 phase cells.

The Connectivity Map [55] takes an obverse approach to GSEA: it relies on a large collection of expression profiles. Each expression profile compares gene expression levels in cells under control conditions to expression levels when the cells are treated

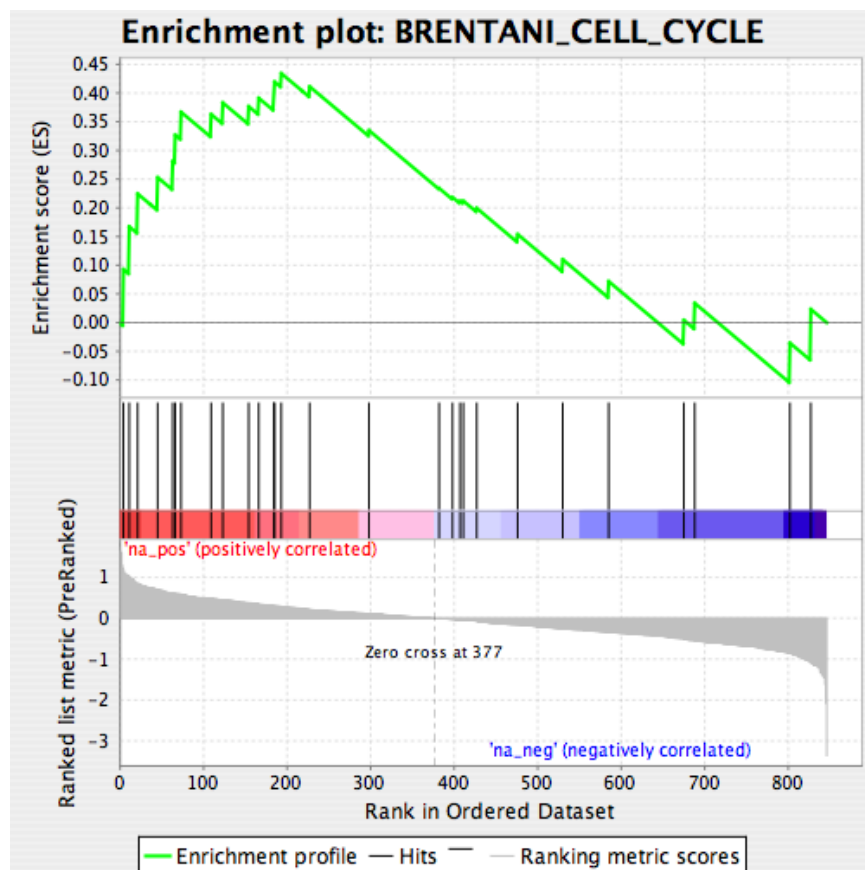


Figure 6-1: GSEA scoring cell-cycle genes against phenotype profile for 4N phospho-histone H3 negative (G2 phase) cells. The upper graph shows the running score $D(S)$, with the enrichment score $ES(S)$ equal to $D(S)$'s maximum deviation from zero. Steps up in $D(S)$ occur when genes in the profile are also in S , also shown as black lines in the lower plot. This set is significantly biased to the left (p-value of 0.018 from permutation testing)

with a drug or chemical. At the time of writing, the Connectivity Map contains 453 mRNA expression profiles for 164 different combinations of compounds and cell lines.

The Connectivity Map's input is two sets of genes taken from experiment: upregulated and downregulated, relative to controls or comparing two tissue types (e.g., cancerous and benign tumors). For each differential expression profile in its database, it computes a similar score to that of GSEA for the up- and down-regulated input sets. For sets S_{up} and S_{down} and a profile R , the Connectivity Map score is computed

as,

$$P_{\text{hit}}(R, S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{1}{N_S}, \quad (6.6)$$

$$P_{\text{miss}}(R, S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{N - N_S} \quad (6.7)$$

$$D(R, S, i) = P_{\text{hit}}(R, S, i) - P_{\text{miss}}(R, S, i) \quad (6.8)$$

$$CS(R) = \operatorname{argmax}_{D(R, S_{\text{up}}, i)} |D(R, S_{\text{up}}, i)| - \operatorname{argmax}_{D(R, S_{\text{down}}, i)} |D(R, S_{\text{down}}, i)|, \quad (6.9)$$

which is essentially the difference of the $ES(S_{\text{up}})$ and $ES(S_{\text{down}})$ scores from GSEA, with $p = 0$, which is also the difference in the Kolmogorov-Smirnov statistics of S_{up} and S_{down} relative to R . The score $CS(R)$ is computed for every profile in the Connectivity Map database. Many of the chemicals and drugs in the the database are represented multiple times, often with varying concentrations. This allows some evaluation of significance from permutation testing, by comparing the average $CS(R)$ for the multiple profiles of one chemical to the distribution of the average score of randomly selected set of the same size.

We apply these tools directly to phenotype profiles, exactly as we would if they were expression profiles. Doing so with GSEA is trivial. To apply the Connectivity Map, it is necessary to choose score thresholds to generate the upregulated and downregulated sets, which in our case are genes that enhance and suppress the phenotype, respectively. In general, we choose a positive and negative threshold at the same absolute log-likelihood score, near the knee in the curve of the phenotype profile scores (see figure 6-1), usually resulting in around fifty genes in each set. We have not investigated a more principled approach to choosing these thresholds. As part of the validation of our methods, we analyzed the G2-phase phenotype with the Connectivity Map. The results of this analysis are presented in the next chapter.

The correlations found by these tools allow us to form predictions of the biological significance of the phenotype of interest. Just as when these tools are used on expression data, they are intended as hypotheses to steer scientists' attention, suggest

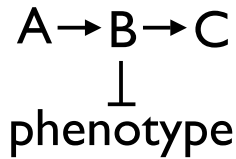


Figure 6-2: An example gene network regulating a phenotype. Expression of gene A positively regulates gene B. Increased expression of B suppresses the phenotype, while increasing expression of C. Knockdown of A or B enhances the phenotype. From the true network, we would expect gene C’s expression to be low in cells showing the phenotype. However, actual knockdown of C would not affect the phenotype, preventing us from correctly estimating its expression in this phenotype’s profile.

new avenues of exploration, and help validate prior results.

6.2 An Interpretation of Phenotype Profiles

Phenotype profiles are estimations of whether a particular gene’s knockdown causes more or fewer cells to take on some phenotype. We treat them analogously to expression profiles, in order to find correlations between genes that cause the phenotype and possible explanations for how they do so. We interpret strong correlations as evidence that some genetic pathway, biological system, or drug’s method of action is related to the phenotype, via a sort of “guilt by association.”

A gene’s score in a phenotype profile is a quantification of the evidence about the effect of reducing that gene’s expression. However, we can interpret phenotype profiles in another way, as noisy estimations of gene expression in cells showing a particular phenotype. This interpretation can be seen as a (loose) application of Bayes’s law:

$$P(\text{phenotype present} \mid \text{gene expression reduced}) \propto P(\text{gene expression reduced} \mid \text{phenotype present})$$

This estimation is noisy, of course, and dramatically simplified from the realities of genetic regulation. Even without considering noise, there is no concept of causality or parallel effects present in this model. As an example, consider the network in figure 6-2. Gene A positive regulates gene B, while B negatively regulates some phenotype and positively regulating gene C. Knocking down the expression of gene

A or B results in an increase in the phenotype, while knocking down C does not significantly affect the phenotype. By our methods, we would estimate that A and B are underexpressed in cells showing the phenotype, but that C's level is unchanged in them. Yet, depending on how tightly coupled gene B's expression and the phenotype are, C would be underexpressed in these cells. It is fairly easy to construct other counterexamples. In general, we can only claim to have estimated the expression levels of genes "upstream" of a phenotype. Our hope is that the knockdowns of unrelated and "downstream" genes do not affect the phenotype too much.

Phenotype profiles should not be taken as *true* expression profiles, by any means, but this approach does provide us with some intuition for interpreting them and the results from tools such as GSEA and the Connectivity Map. Our use of large ensembles of scores when interpreting phenotypes, the full profile in the case of GSEA and sets of high- and low-scoring genes for the Connectivity Map, is in part motivated by the need to incorporate a large amount of data to overcome the noise and failures of estimation above.

6.3 Summary

This chapter introduces methods for connecting phenotype profiles to biological data. We use existing tools which seek correlations between sets of genes and expression profiles, or in our case, phenotype profiles.

A strong correlation between a phenotype profile and some aspect of known biology leads us to predict a relationship between the phenotype generating the phenotype profile and that biological aspect, and by association, to predict a relation between that known biology and the genes that score most strongly in relation to the phenotype. The existing data might be known genetic pathway, a set of genes with a common annotation, genes differentially expressed in one tissue type compared to some other, a drug or chemical's effect on gene expression, or from a number of other sources. We have validated these approaches on a cell-cycle related test case, which we explore further in the next chapter.

We have also explored an interpretation of phenotype profiles as estimates of expression profiles from phenotype information. This interpretation provides us with intuition as to why we expect expression analysis tools to be effective when applied to phenotype profiles, and what the correlations they might find indicate about the phenotypes we apply them to.

In the next and final chapter, we present results from and further validation of the methods and techniques we have introduced in this and prior chapters, applied to predicting related genes and their underlying biological functions.

Chapter 7

Results

We present validation of and novel results from our methods in this chapter. For validation purposes, we look at two cases, cells in the G2 phase of the cell cycle, and cells undergoing cytokinesis.

We present results for several phenotypes that have not been previously studied. For those phenotypes, we present predictions for the biological mechanisms of the phenotype, where possible.

Our image processing and cell measurement methods have been validated separately, as discussed in previous chapters, but also through their application to a variety of screens and other quantitative imaging problems [8, 9, 17, 22, 42, 56, 71, 99].

7.1 Cell Cycle Control Genes

We return to the 4N phospho-histone H3-negative (G2-phase) phenotype, as a positive control for the methods introduced in this dissertation. We generated a phenotype profile from the scores for each gene, and processed it with Gene Set Enrichment Analysis to find which gene sets show a significant correlation with the profile. The highest scoring gene sets are listed in table 7.1, as ordered by p-value from GSEA's permutation testing.

Of those gene sets scoring with a positive enrichment score (i.e., with gene sets more likely to cause the G2 phenotype when knocked down), seven of ten are cell cycle

Gene Set	Overlap	Size	Score (NES)	p-value	MSigDB ID
Protein modification	96	150	-1.53	0.0083	c2:519
Upregulated in VHL-rescued renal carcinoma vs. normal	33	519	-1.62	0.013	c2:1620
B-cell antigen receptor pathway	15	40	1.72	0.015	c2:569
Upregulated by sulindac in SW260 colon carcinoma cells	16	131	-1.62	0.02	c2:1428
Upregulated in VHL-null renal carcinoma vs normal	27	447	-1.56	0.027	c2:1776
Cancer related genes involved in the cell cycle	28	86	1.52	0.033	c2:513
Genes downregulated in response to rapamycin	17	229	1.54	0.041	c2:625
Downregulated in fibroblasts by HMCV infection	44	421	1.42	0.042	c2:1283
G1 pathway	15	28	1.51	0.05	c2:193
Up-regulated in mouse hematopoietic stem cells	19	241	-1.5	0.051	c2:1460
Cell cycle	23	84	1.46	0.053	c2:500
Downregulated by butyrate in SW260 colon carcinoma cells	15	109	1.51	0.055	c2:1390
BCR signaling pathway	22	46	1.46	0.057	c2:560
Genes upregulated by hypoxia or HIF1 activation	19	107	-1.47	0.058	c2:737
Genes involved in mRNA processing	23	47	-1.45	0.061	c2:542
Downregulated in XPC-defective fibroblasts	15	190	1.47	0.064	c2:1245
Genes involved in DNA damage signaling	22	95	1.42	0.068	c2:525
Genes upregulated in well-functioning transplanted kidney biopsies	42	565	-1.38	0.073	c2:836
Up-regulated in mouse hematopoietic stem cells	18	227	-1.41	0.084	c2:1461
Down-regulated following treatment with Et-743	33	269	-1.36	0.089	c2:1338

Table 7.1: Top twenty gene sets from GSEA that correlate with the 4N phenotype profile, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set. Bolded sets are related to the cell cycle.

Description	Overlap	Size	NES	p-value	MSigDB ID
Upregulated in mouse hematopoietic stem cells and progenitors (I)	35	610	-1.76	0.0037	c2:1458
Upregulated in mouse hematopoietic stem cells and progenitors (II)	35	610	-1.75	0.0052	c2:1457
Upregulated in mouse hematopoietic stem cells and progenitors (III)	36	624	-1.7	0.0073	c2:1456
Genes expressed in T-cell acute lymphocytic leukemia	25	282	1.73	0.0074	c2:1030
Genes overexpressed in leukemia cells.	25	259	1.73	0.0079	c2:666
Downregulated in human granulosa cells by luteinizing hormone	21	77	-1.7	0.01	c2:1548
Downregulated in human granulosa cells by follicle stimulation hormone	21	77	-1.7	0.011	c2:1340
Cancer related genes involved in the cell cycle	28	86	1.62	0.016	c2:513
Downregulated genes following Apc loss [38, 50]	17	405	1.68	0.017	c2:1047
Genes involved in mRNA splicing	17	58	-1.67	0.018	c2:543
Downregulated in Wilms' tumor versus fetal kidney	22	162	1.61	0.02	c2:888
Upregulated in Wilms' tumor versus fetal kidney	20	180	-1.63	0.023	c2:889
Downregulated in WS1 skin fibroblasts by high-dose UV-C light	37	297	-1.54	0.026	c2:1733
Signal transduction through calcium, calcineurin, and NF-AT (mouse)	22	100	1.55	0.031	c2:495
Upregulated in plasma cells compared to PPCs	27	310	1.49	0.04	c2:693
Genes related to chemotaxis	17	45	-1.55	0.041	c2:562
Cell Cycle (Kegg pathways)	26	90	1.49	0.043	c2:457
Differentially expressed in developmental of Va14i NKT cells	38	493	-1.46	0.044	c2:1091
Upregulated in VHL-rescued renal carcinoma vs normal	33	519	-1.47	0.046	c2:1620
Downregulated by BAF57-rescue in Bt549 breast cancer [96]	18	335	1.5	0.049	c2:1182

Table 7.2: Top twenty gene sets from GSEA that correlate with the anaphase vs. metaphase profile. Positive scores indicate that the gene set correlates with genes that enrich the anaphase phenotype. Bold entries are cell-cycle related, with citations where this is not apparent from the gene set itself (Ultraviolet (UV) light causes DNA damage leading to cell-cycle arrest).

related. We have listed gene sets that negatively correlate with genes that cause the phenotype, though these are not as directly interpretable as cell-cycle related. We have noticed this to commonly be the case, possible reasons for which we discuss below.

We generated an anaphase vs. metaphase phenotype profile, in which we considered anaphase cells as phenotype-positive and metaphase as negative, but ignored all other cells. This is a useful technique for evaluating putatively related phenotypes where cells not in either phenotype might dominate the computation. In this case, most other cells are in some other phase of the cell cycle, such as G1 or G2.

In this case, because of the lower number of cells showing either phenotype (around 3% of all cells), we expect lower accuracy from any predicted gene function. The results from GSEA are shown in table 7.2. Again, several cell-cycle related sets appear, including one with known specific function in the metaphase/anaphase transition (loss of the gene *Apc* [38, 50]). Note that several related gene sets appear multiple times, including three related to mouse hematopoietic stem cells from three slightly different characterizations of these genes, two related to granulosa cells, and up- and downregulated genes in Wilms' tumor. These repeats are an artifact of their overrepresentation in GSEA's database of gene sets.

We analyzed the G2 phenotype with the Connectivity Map. The top five most significant results for this phenotype are shown in table 7.3. Many drugs affect the cell cycle (particularly those used in the treatment of cancer, many of which are present in the Connectivity Map's database), so it is unsurprising that a link between the drugs listed in table 7.3 and the cell cycle can be established. The table lists a citation deemed most plausible for the screened cell line (HT29) and this phenotype (G2), from the large number available for each compound tying it to the cell cycle.

We analyzed the anaphase vs. metaphase phenotype via the Connectivity Map, with similar results. The two most strongly scoring compounds for this phenotype were LY-294002 (also scoring in the G2 phenotype), which is known to affect the anaphase-metaphase transition [43, 75], and all-trans retinoic acid, known to affect proliferation in the HT29 cell line (the line used in this screen) [73]. We should note

Name	Type	Count	Mean score	p-value	Citation
LY-294002	PI3K inhibitor	17	-0.515	0.0001	[87]
Valproic acid	HDAC1 inhibitor	18	0.101	0.0145	[86]
Rosiglitazone	Ppar γ activator	4	-0.224	0.0208	[44]
Nu-1025	PARP inhibitor	2	-0.644	0.0233	[69]
Prochlorperazine	Calmodulin inhibitor	3	-0.385	0.0457	[47]

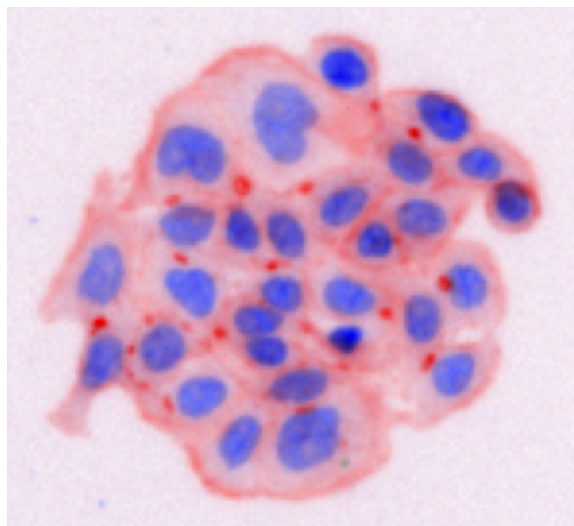
Table 7.3: Compounds that correlate with the G2/4N phenotype profile in the Connectivity Map. Positive scores indicate positive correlation between the up- and downregulated sets of genes. Note that knockdown experiments reduce gene expression, so we expect negative scores to positive correlate with the phenotype-positive cells. Count is the number of times the compound appears in the connectivity map database (as a replicate or under a different concentration), across which we calculate a mean score. P-values are determined by permutation of expression profiles (not of genes in the up- and downregulated sets). Citations listed were selected based on plausibility of explanation for the connection to the cell-cycle.

that the Connectivity Map was generated by profiling expression in several different cell lines, but primarily in MCF7 cells (a breast cancer line). Neither the HT29 cell line, nor any other colon cancer cell lines, were used to create the profiles. As many drugs affect cells in a very tissue-specific manner, the hypotheses generated from the Connectivity Map should be balanced by the similarities in drug-response from cell-line to cell-line.

7.2 Novel Phenotypes

We now consider some of the complex visual phenotypes from figure 1-2, and their possible interpretations from phenotype-profile analysis. Some of these phenotypes were discovered in the initial screen [71], others when following up earlier phenotypes or during analysis of *a priori* interesting cytological measurements, such as cell size or eccentricity.

Actin Dots

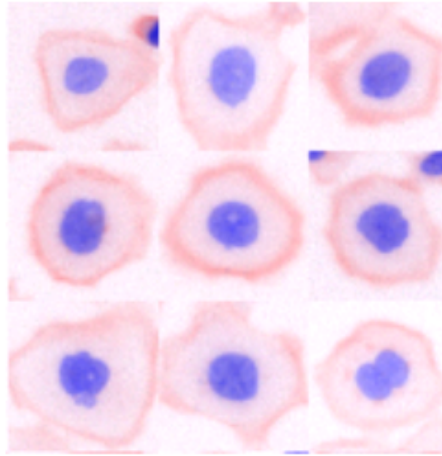


This phenotype, with dots of actin at cell junctions, appears to be related to cell adhesion, based on the most strongly scoring genes and the results from GSEA. The most strongly scoring genes are listed in table 7.4, along with possible links to cellular focal adhesions, the cytoskeletal links from a cell its exterior. The results from GSEA analysis are shown in table 7.5. We discuss this phenotype in conjunction with the Actin Ring phenotype, below, for reasons that will become clear.

Gene	Score	Connection to focal adhesion?
SFRS2	2.25	none known
PKIA	1.28	none known
PTPRC	1.10	focal adhesion GO term [6] from sequence similarity
PTPRB	1.10	yes [5]
TRPM6	1.09	possibly, via association with TRPM7 [21]
...	...	
TEC	-1.36	none known, (same family as TXK)
STK17A	-1.37	none known
PPP1R12A	-1.40	yes, in focal adhesion pathway (Kegg [49])
LCK	-1.54	yes, [37]
TXK	-1.64	none known, (same family as TEC)
PRKCZ	-2.06	maybe, involved in cell-cell (tight) junctions (Kegg [49])
DAPK3	-2.10	maybe, (interacts with PRKCZ <i>in vitro</i> [16])

Table 7.4: Top scoring genes for the Actin Dots phenotype. Several genes are directly or closely related to focal adhesions in cells.

Actin Ring

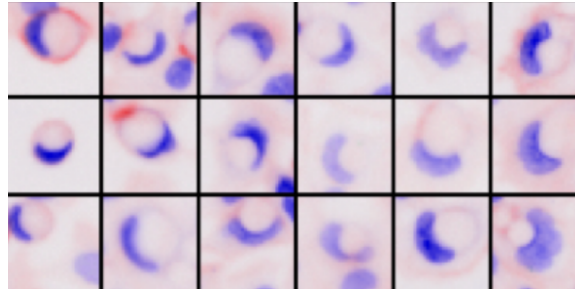


The Actin Ring phenotype shows an increased concentration of cytoskeletal material and a smooth cell border. The appearance of the cells indicates a reorganization of the cytoskeleton, similar to that caused by toxins released by some bacteria, especially *Bacteroides fragilis* [84]. The GSEA analysis of this phenotype, listed in table 7.6, indicate two interesting results. First, this phenotype is anticorrelated with the Actin Dots phenotype, indicating similar underlying mechanisms, operating in an opposite fashion. Second, the connection to human cytomegalovirus (HCMV, member of the herpesvirus family), which is known to cause reorganization of the actin cytoskeleton, and causes a similar phenotype in HT29 cells (cf. figure 9 of [10]). The reorganization of the cytoskeleton may help spread the virus between cells [66]. It is known that HCMV upregulates β_1 -integrin receptors in PC3 cells [12], possibly affecting cell adhesion and cytoskeletal organization, via the integrin-signalling pathway, and also upregulates the PI3K pathway [45]. Both of these pathways appear in the GSEA results for this phenotype.

The opposition of the underlying mechanisms for the Actin Ring and Actin Dots phenotypes opens up the possibility of epistasis studies to determine which genes (of those scoring highly for either phenotype) are up- or downstream relative to the phenotype. Such studies are possible with any two phenotypes which are mutually

exclusive, such as 2N versus 4N DNA content or large versus small cells, but more useful when exploring a more limited portion of the genetic regulatory network.

Crescent Nuclei



One of the most interesting phenotypes we have encountered is one in which the nuclei are crescent shaped and pressed against the cell membrane, usually by what appears to be a vacuole (i.e., a membrane-bound compartment interior to the cell). Often, the cell body appears inflated as well.

One theory being investigated in ongoing experiments is that this phenotype is due to breakdown of ion balance in the cell, in particular, dysregulation of sodium ion channels. This is based on the large structure in each cell that appears to be a vacuole, the lack of staining in the vacuole even with a variety of broadly targeted stains, and a few key genes' knockdowns giving strong phenotypes. In particular, one of the two hairpins against NEDD4L that pass quality control for infection efficiency gives a very strong, positive phenotype. NEDD4L is a known inhibitor of ENaC, a sodium channel protein complex in the cell [48]. Since only two knockdowns of NEDD4L pass quality control, the gene itself does not score strongly.

Another possibility, from examination of the literature (guided by the GSEA results discussed below), is that the vacuole is filled with mucin or another substance. This is based on the phenotype which is very similar to “signet ring cells” [54]. Cancers showing this morphology arise in several different organs [72], and patients with signet ring cell carcinoma (SRCC, defined as more than 50% of tumorous cells showing this phenotype) have a poorer prognosis than non-SRCC carcinoma [74]. The genes that lead to this phenotype are therefore of interest to understand why SRCC

Gene Set	Overlap	Size	Score (NES)	p-value	MSigDB ID
Upregulated in VHL-null renal carcinoma vs. normal	27	447	1.9	0.0016	c2:1776
Upregulated in VHL-rescued renal carcinoma vs. normal	33	519	1.86	0.0022	c2:1620
Differentially expressed cellular adhesion genes (aortic flow)	57	342	-1.67	0.0027	c2:718
Integrin mediated cell adhesion (Kegg pathway)	25	96	-1.62	0.013	c2:469
Genes related to chemotaxis	17	45	-1.62	0.015	c2:562
Genes downregulated in response to glucose starvation	21	157	1.64	0.019	c2:631
Differentially expressed in HGF-activated monocytes	39	661	-1.49	0.03	c2:1097
Genes related to insulin signaling	34	103	-1.5	0.03	c2:538
Genes upregulated in well-functioning transplanted kidney biopsies	42	565	1.46	0.032	c2:836
G1 pathway	15	28	-1.49	0.047	c2:193
G-protein signaling	15	97	-1.49	0.05	c2:374
Downregulated in XPB/TTD fibroblasts by UV-C light	30	165	-1.44	0.053	c2:1754
Downregulated by TPA in HL-60 cells	23	234	-1.44	0.059	c2:1684
Downregulated in human granulosa cells by luteinizing hormone (LH)	21	77	-1.43	0.061	c2:1548
Genes in the BCR signaling pathway	22	46	-1.43	0.062	c2:560
Downregulated in human granulosa cells by follicle stim. hormone (FSH)	21	77	-1.44	0.065	c2:1340
PI3K pathway	19	38	-1.43	0.065	c2:590
Downregulated in fibroblasts by HMCV infection	44	421	-1.38	0.066	c2:1283
Genes in the NFAT pathway	17	53	-1.43	0.066	c2:258
Genes in the MAPK pathway	51	87	-1.36	0.073	c2:241

Table 7.5: Top twenty gene sets from GSEA that correlate with the Actin Dots phenotype profile, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set. Bolded sets are directly related to cytoskeletal organization.

Gene Set	Overlap	Size	Score (NES)	p-value	MSigDB ID
Downregulated in WS1 skin fibroblasts by high-dose UV-C light	37	297	1.59	0.017	c2:1733
Genes related to chemotaxis	17	45	1.6	0.021	c2:562
Genes differentiating BRCA1-linked and BRCA2-linked breast cancers	20	140	1.52	0.038	c2:1218
Integrin signaling pathway (Science stke)	31	82	1.48	0.04	c2:581
Genes in the NFAT pathway	17	53	1.52	0.04	c2:258
Cancer related genes involved in cell signaling	60	198	1.41	0.046	c2:521
PI3K pathway	19	38	1.48	0.052	c2:590
Genes in the MAPK pathway	51	87	1.41	0.052	c2:241
Genes related to IL4 receptor signaling in B lymphocytes	15	27	1.49	0.053	c2:563
Regulated by UV-B light in normal human epidermal keratinocytes	43	397	1.4	0.061	c2:1718
Genes downregulated in response to rapamycin	17	229	1.45	0.066	c2:625
Genes related to the insulin receptor pathway	19	51	1.43	0.07	c2:564
Downregulated in human granulosa cells by luteinizing hormone (LH)	21	77	1.41	0.076	c2:1548
G1 pathway	15	28	1.42	0.079	c2:193
Upregulated in mouse hematopoietic stem cells, versus brain & bone marrow	103	1452	1.3	0.081	c2:1647
Genes related to PIP3 signaling in cardiac myocytes	31	67	1.39	0.081	c2:566
Downregulated in human granulosa cells by follicle stim. hormone (FSH)	21	77	1.4	0.082	c2:1340
Downregulated in fibroblasts following infection with HCMV infection	44	421	1.35	0.084	c2:1283
Calcium regulation in cardiac cells	26	143	-1.34	0.096	c2:356
Enriched in mature T cells	73	1141	1.29	0.098	c2:1053

Table 7.6: Top twenty gene sets from GSEA that correlate with the Actin Ring phenotype profile, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set.

responds more poorly to treatment, as well as for possible drug targets. Signet ring cells are similar in appearance to “goblet cells” [25], which are responsible for mucus secretion in the intestinal and respiratory tracts. The cell line used in this screen can be induced to form a mucin-secreting, goblet-cell enriched line by selection under methotrexate [59] or sodium butyrate [7].

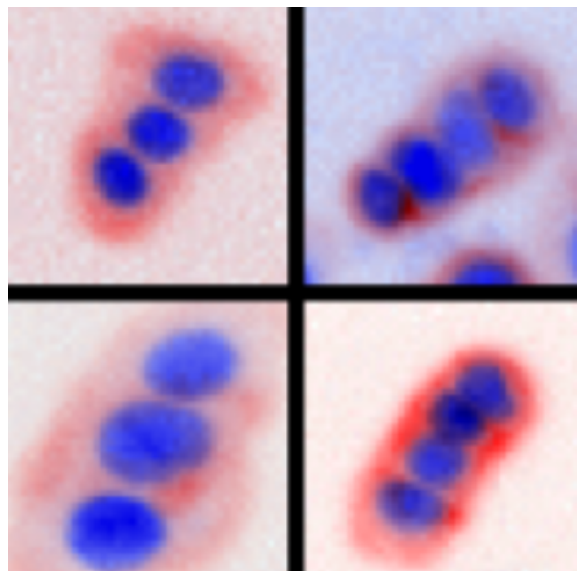
The GSEA results for this phenotype, shown in table 7.7, show evidence in favor of both of the explanations above, though possibly more strongly the second. Two of the top hits are sets of genes downregulated by butyrate in a colon-cancer cell line (SW620) at different time points. Butyrate increases expression of the components of the ENaC complex [100]. However, butyrate also regulates mucin genes [36] in HT29 cells, and as mentioned above, selection under butyrate causes enrichment in cells with this morphology [7]. This phenotype is also seen in pituitary cells when stimulated with luteinizing hormone releasing hormone (which causes release of both luteinizing hormone and follicle stimulating hormone) [97], though the presence of this result in the GSEA list is not necessarily meaningful, given that it also appears for other phenotypes, such as Actin Rings (cf. table 7.6).

The multitude of possibilities, all linked to phenotypes similar in appearance but with quite different underlying biology, leads to another explanation: it is possible that we have detected more than one of the actual mechanisms due to an inability to distinguish their slight differences in the screen at hand. This is a concern in any screen, particularly given the technical limitations in staining protocols and imaging. This situation could be improved by screening replicates with different staining protocols.

Description	Overlap	Size	NES	p-value	MSigDB ID
Genes involved in mRNA splicing	17	58	1.84	0.003	c2:543
Downregulated by butyrate in SW260 colon carcinoma cells (I)	24	248	1.75	0.0069	c2:1397
Cancer related genes involved in the cell cycle	28	86	1.63	0.015	c2:513
Biopetides and GTPase pathway	19	39	1.65	0.017	c2:125
Keratinocyte differentiation pathways.	25	46	1.58	0.022	c2:233
Upregulated in HEK293 cells by infection with reovirus strain T3Abney	20	237	1.61	0.023	c2:1623
Downregulated by butyrate in SW260 colon carcinoma cells (II)	15	109	1.63	0.025	c2:1390
Genes differentiating BRCA1-linked and BRCA2-linked breast cancers	20	140	1.56	0.03	c2:1218
Upregulated in plasma cells compared to PPCs	27	310	1.5	0.037	c2:693
Downregulated in human granulosa cells by follicle stim. hormone (FSH)	21	77	1.52	0.04	c2:1340
Downregulated in human granulosa cells by luteinizing hormone (LH)	21	77	1.52	0.04	c2:1548
Proliferation related genes	54	394	1.41	0.044	c2:646
Genes in the NFAT pathway	17	53	1.52	0.045	c2:258
G1 pathway	15	28	1.51	0.049	c2:193
Cell cycle	23	84	1.48	0.05	c2:500
Upregulated in human granulosa cells by follicle stim. hormone (FSH)	19	78	-1.49	0.054	c2:1341
Genes involved in DNA damage signaling	22	95	1.47	0.055	c2:525
Genes related to IL4 rceptor signaling in B lymphocytes	15	27	1.49	0.056	c2:563
Differentially expressed genes in sickle cell patients	24	372	1.45	0.057	c2:1036
Cancer related genes involved in the cell signaling	60	198	1.36	0.061	c2:521

Table 7.7: Top twenty gene sets from GSEA that correlate with the Crescent Nuclei phenotype profile, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set.

Peas-in-a-pod



The final phenotype we examine in isolation is the “Peas-in-a-pod” phenotype, in which individual cells clump together in short strings, with limited cytoskeletal boundaries between adjacent cells. This phenotype is quite rare, with only a few knockdowns causing it across the entire screen. The results from GSEA analysis of this phenotype are shown in table 7.8.

The relative sparseness of this phenotype within the screen limits our confidence in the hypotheses generated by GSEA. However, the strongest hit, genes downregulated by tumor promoter 12-O-tetradecanoylphorbol-13-acetate (TPA) in HL-60 cells, a human leukemia cell line, led us to seek examples of this chemical’s effect on HT29 cells (the line from this screen). Although there have been no morphology-oriented studies of this cell line under TPA treatment, studies of TPA-induced cell scattering seem to show a similar phenotype in at least three cases: Choi *et al.* [20, figure 2], Martin *et al.* [68, figure 1A(c)], and Rochet-Egly *et al.* [80, figure 2(b,e,h)]. Although not conclusive, these examples strongly hint at a connection to scattering and the underlying biological processes, such as cell-cell contact and extracellular adhesion [68].

Description	Overlap	Size	NES	p-value	MSigDB ID
Downregulated by TPA in HL-60 cells (I)	39	304	1.68	0.0027	c2:1686
Genes downregulated in response to glutamine starvation	32	313	1.62	0.0064	c2:627
Cell Cycle	26	90	1.64	0.0075	c2:457
Up-regulated in mouse mature blood cells (I)	16	257	1.65	0.0075	c2:1473
Up-regulated in mouse mature bod cells (II)	17	331	1.64	0.0085	c2:1472
Genes involved in mRNA splicing	17	58	1.63	0.01	c2:543
Differentially expressed in wild-type versus fetal kidneys (high)	22	162	1.59	0.013	c2:888
Enriched in mouse neural stem cells	139	1838	1.36	0.017	c2:1648
Genes upregulated in control vs. bmyb knockdown in zebra fish	17	208	1.57	0.019	c2:752
Downregulated by TPA in HL-60 cells (II)	35	284	1.47	0.031	c2:1682
Upregulated in VHL-null renal carcinoma vs. normal renal cells	27	447	1.48	0.037	c2:1776
Genes involved in mRNA processing	23	47	1.48	0.038	c2:542
mRNA processing reactome genes	36	121	1.42	0.046	c2:431
Cell-cycle dependent genes regulated by serum in fibroblast cells	15	138	1.47	0.048	c2:1640
Upregulated in the atria of healthy hearts, compared to ventricles	15	198	1.46	0.049	c2:1181
Differentially expressed in wild-type versus fetal kidneys (low)	20	180	1.44	0.054	c2:889
Mad1 affected genes in lymphocytes	17	127	1.44	0.054	c2:944
Upregulated by hypoxia in normal renal cells	18	219	-1.45	0.062	c2:1490
Upregulated by RNAi knockdown of PRMT5 in 3T3 cells	17	187	1.42	0.064	c2:1619
Upregulated in mouse mature blood cells	23	339	1.38	0.076	c2:1471

Table 7.8: Top twenty gene sets from GSEA that correlate with the Peas-in-a-pod phenotype profile, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set.

7.3 Interactions Between Phenotypes

As some of the results above show, many phenotypes interact with one another. This is especially true of the cell-cycle, which in many ways dominates most other phenotypes, as any phenotype that does not allow the cell cycle to progress can be considered an instance of a cell-cycle arrest phenotype.

We visualize interactions between phenotypes by plotting pairwise gene scores for several of the phenotypes we have explored in figure 7-1. As can be seen, many of the phenotypes have a cell-cycle dependency (based on the correlation of scores with the G2 phenotype). The anticorrelation between actin dots and actin rings was explored above. As would be expected, some phenotypes affect cell-cycle distribution, as shown in figure 7-2 for the Crescent Nuclei and Fingers phenotypes.

It would be useful to be able to explore the cell-cycle and morphological effects of these phenotypes separately, but how to do so is an open problem. A simple (but unsuccessful) approach is to only count cells in a particular phase (such as G2) in the phenotype-positive and -negative classes for each knockdown. However, as can be seen in figure 7-3, though this may decrease the dependence between the phenotypes, the effect is slight, and not generally effective.

One case where it does appear to attenuate the (slight) cell-cycle dependency is in the Crescent Nuclei phenotype. We reanalyzed this phenotype limited to 4N cells, with results shown in table 7.9. These results are more difficult to interpret, in relation to the phenotype, and appear more noisy (note, for example, the presence of sets upregulated and downregulated in XPC-defective fibroblasts).

Methods

To analyze phenotype profiles with GSEA, we use version 2.0.1 of the source code and version 2 of the MSigDB (the database of gene sets), limit analysis to gene sets that overlap with at least 15 and no more than 500 genes in the screen, use the weighted statistic described in the previous chapter, and use 20,000 permutations to compute p-values.

Description	Overlap	Size	NES	p-value	MSigDB ID
Cell proliferation genes (from zebra fish)	33	232	-1.67	0.008	c2:757
Cell proliferation genes.	33	232	-1.67	0.0085	c2:508
Down-regulated in glomeruli isolated from Pod1 knockout mice	68	724	-1.55	0.011	c2:1613
Upregulated in XPC-defective fibroblasts	18	151	-1.67	0.018	c2:1246
Cancer related genes involved in protein modification	96	150	-1.45	0.018	c2:519
Up-regulated in mouse hematopoietic stem cells and progenitors from adult bone marrow	36	624	1.55	0.02	c2:1456
Downregulated in XPC-defective fibroblasts (I)	15	190	-1.6	0.024	c2:1245
Mitochondrial genes	17	447	-1.57	0.032	c2:605
Tumor suppressor genes	15	26	-1.54	0.04	c2:599
Upregulated in mouse hematopoietic stem cells from fetal liver	35	610	1.47	0.04	c2:1457
Upregulated in mouse hematopoietic stem cells from adult bone marrow and fetal liver	35	610	1.47	0.04	c2:1458
Genes upregulated in human pulmonary endothelial cells under hypoxic conditions	19	107	-1.51	0.044	c2:737
Genes upregulated in response to glutamine starvation	28	296	-1.48	0.044	c2:628
MYC target genes	22	180	-1.48	0.047	c2:759
p38 MAPk pathway	17	37	1.5	0.051	c2:588
Differentially expressed genes in sickle cell patients	24	372	1.45	0.056	c2:1036
Myb-regulated genes	22	325	-1.45	0.06	c2:1038
Biopeptides and GTPase pathway	19	39	1.45	0.063	c2:125
Downregulated in XPC-defective fibroblasts (II)	62	478	-1.35	0.065	c2:1762
Upregulated in mouse hematopoietic stem cells, compared to brain and bone marrow cells	103	1452	1.27	0.075	c2:1647

Table 7.9: Top twenty gene sets from GSEA that correlate with the Crescent Nuclei phenotype profile when limited to 4N cells, ordered by p-value from permutation testing. We show the overlap between the set and the genes in the screen, the full size of the set, and the normalized enrichment score (NES) and p-value from GSEA. Positive scores indicate that the gene set correlates with genes that enrich the phenotype. The MSigDB ID is the GSEA-specific identifier for the gene set.

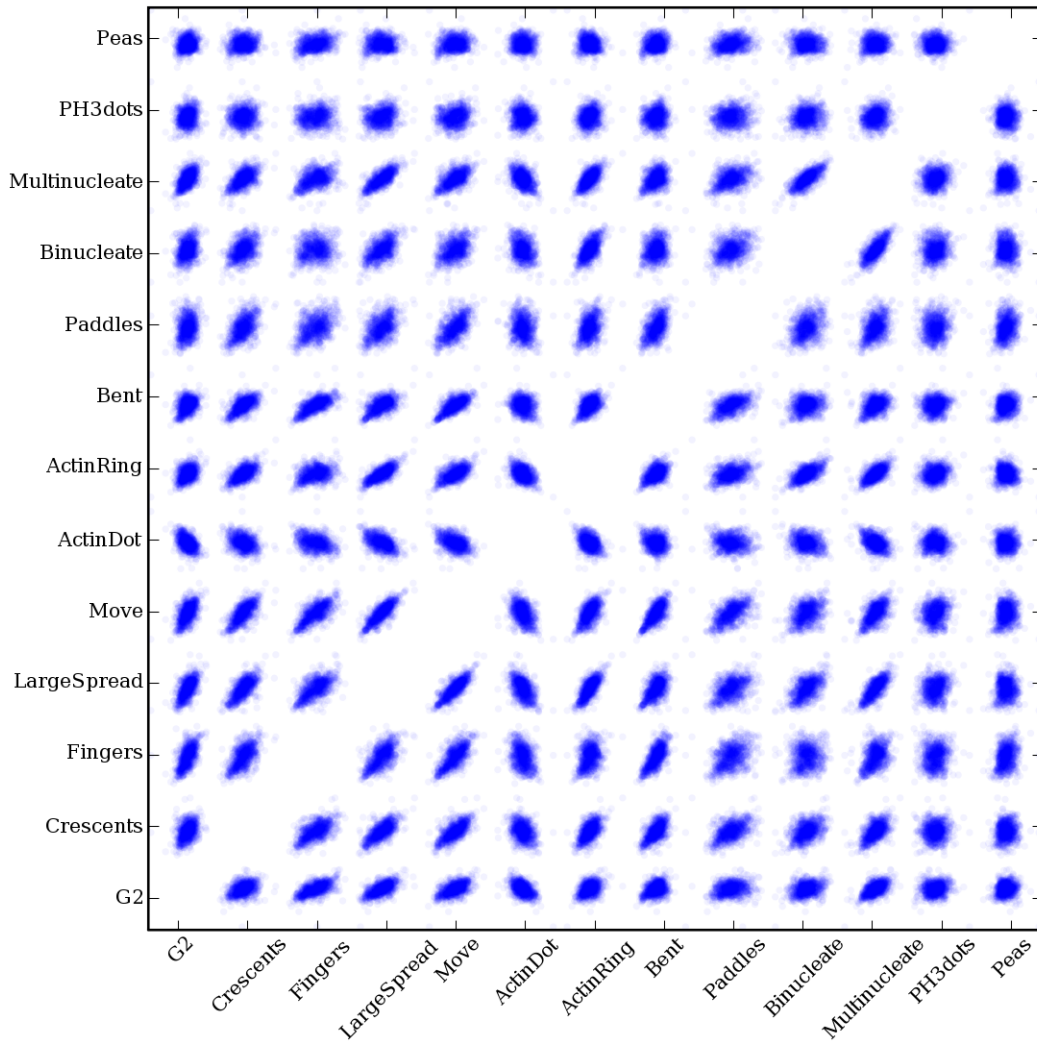


Figure 7-1: Many phenotypes interact with one another. Each small scatterplot shows the paired per-gene scores from the phenotype profile for that column (horizontal axis) and row (vertical axis). Many phenotypes have a cell-cycle interaction or effect, as can be seen in the bottom row of plots. The negative correlation between the Actin Dot and Actin Ring phenotypes is also apparent.

For the Connectivity Map, for uniformity of analysis, we take the top 100 genes by magnitude of their log-likelihood score, and break them into upregulated and downregulated sets by the sign of their score. The Connectivity Map requires input in the form of probe identifiers on a particular expression array (Affymetrix GeneChip Human Genome U133A Array, part number 510681). We translate phenotype profiles

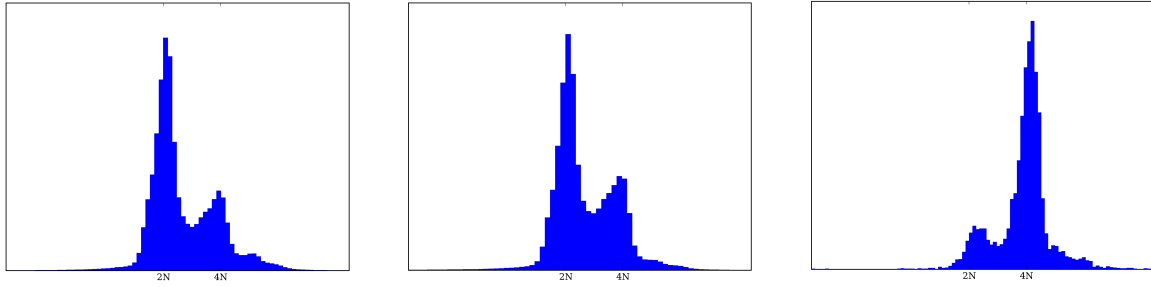


Figure 7-2: Some phenotypes perturb the cell-cycle distribution. From left-to-right, DNA histograms are shown for control cells, cells with the Crescent Nuclei phenotype, and cells with the Fingers phenotype. The latter causes a strong bias toward G2/M (4N) cells.

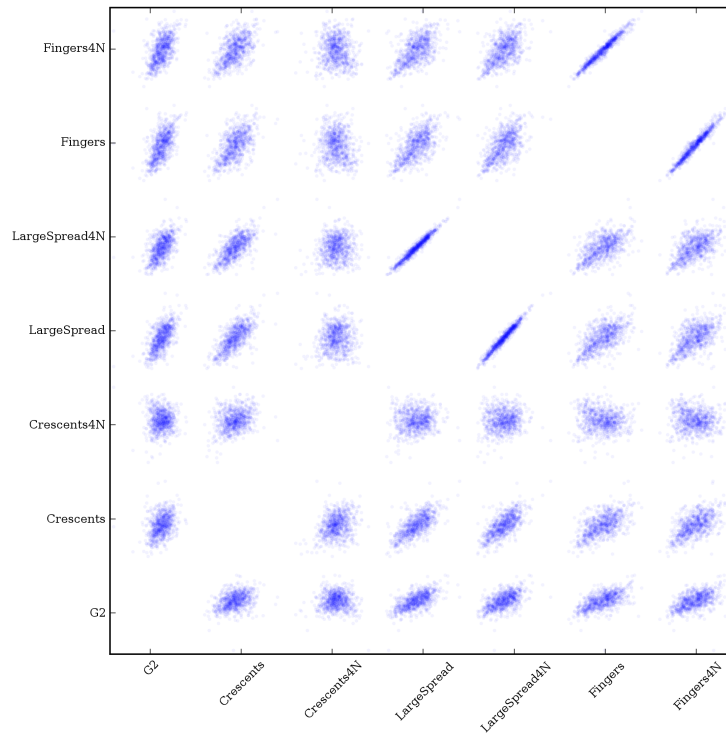


Figure 7-3: Limiting phenotypes to a single phase in the cell cycle does not always remove their dependence. Shown are correlations in scores between three phenotypes and the G2/4N phenotype. Attempting to control for cell-cycle dependence by limiting phenotype-positive and -negative cell populations to 4N cells does not make the scores independent in most cases. An exception is the Crescent Nuclei phenotype, whose slight dependence is further reduced.

(which are listed by gene) to probes via GSEA's Chip2Chip function. There are often multiple probes per gene; we include all probes of each of the 100 genes selected by score, above.

7.4 Summary

In this chapter we have validated our approach to predicting related genes and their function on two cell-cycle related phenotypes, G2-phase cells and cells undergoing cytokinesis. We have also presented results on phenotypes not previously studied, including predictions of genes related to those phenotypes and possible biological meaning of the phenotypes.

Our results for the several of the phenotypes are quite plausible. The Actin Dots and Actin Ring phenotypes, which visually seem to involve cell adhesion and spreading, are correlated with gene sets for cellular adhesion, chemotaxis, and cytoskeletal reorganization by viral infection. Several of the gene sets correlating with the Crescent Nuclei phenotype are known to cause similar phenotypes in the screened cell line, as well as others.

Chapter 8

Discussion

In this work, we have demonstrated that it is possible to extract useful information from images of cells in large gene knockdown screens, and to make meaningful predictions from this information. In our approach, we first correct systematic biases in the images, then identify and measure every cell in each image. The collection of measurements for each cell forms its *cytological profile*. From the measurements of the cells, we identify which cells show a particular phenotype of interest. This phenotype may be defined by particular measurements on the cells, such as DNA content, or may be visually apparent. In the latter case, we use human guidance to train an automatic classifier to identify the cells with the phenotype. We then label every cell in the screen according to whether it shows the phenotype or not, and compute for each knockdown the number of phenotype-positive and phenotype-negative cells. We fit a probabilistic model to these counts, and use that model to score each knockdown according to our belief that it causes or suppresses the phenotype of interest. We combine scores from knockdowns into scores for genes, taking into account possible off-target effects, to form a *phenotype profile* for the screen and that phenotype. We use these scores for our prediction of which genes relate to each other, through the phenotype. To predict the biological meaning of a phenotype, we use existing tools for expression profile analysis to find correlations between phenotype profiles and previously known biological information.

We have validated our methods on several phenotypes, some well understood,

and others not studied previously. For the known phenotypes, we find expected predictions, while for the new phenotypes, we find predictions that are plausible. In both cases, we generate hypotheses that merit further exploration in the laboratory and in future screens.

Our work is one of the first to analyze large screens by measuring and modeling individual cells, and to the best of our knowledge, the first to use screen-wide scores to make predictions about the biological meaning of a phenotype. Our image analysis methods are applicable to a wide variety of cell types and screening and imaging protocols, and our cell-centric, phenotype-based approach to analysing screen data enables a much wider variety of biological questions to be answered via large screens.

8.1 Future Directions

There are several areas for future exploration suggested by our work.

A promising avenue would be the unification of image-based chemical biology and gene knockdown screens. Drug discovery is often a problem of connecting chemicals to the biological processes they affect, the genes they target, or both. Image-based screens of known and potential drug compounds have demonstrated the feasibility of clustering drugs by their effect on cellular phenotypes to predict similarities in biological mechanisms [64]. A screen combining gene knockdowns and chemical treatments would allow several useful biological questions to be answered. For instance, for a given drug, which gene knockdowns cause similar phenotype changes? This can be answered by defining a phenotype from the drug's effect on cells, and applying it to gene-knockdown data. Drug-defined phenotypes would also allow us to predict which biological process is targeted by the drug, via its phenotype profile. Similarly, finding drugs that cause a phenotype similar to a that of a particular gene's knockdown would allow immediate discovery of drugs predicted to act on processes related to that gene.

Our generation and analysis of phenotype profiles have concentrated, for the most part, on single phenotypes in isolation. This does not fully model the case where a gene or genes can affect multiple phenotypes, called *pleiotropy*. We also do not take

into account cases where the presence of one phenotype prevents cells from taking on some other phenotype, similar to the concept of *epistasis* in genetic networks. A more nuanced model of the interaction between phenotypes would improve our predictive ability, by removing confounding factors in the analysis of single phenotypes, and also open up the possibility of characterizing the hierarchical nature of biological processes by examining how they exist in epistatic or pleiotropic relationships as revealed by their effect on phenotypes.

As the number and size of image-based screens grow, data-driven modeling of phenotypes and their variation becomes feasible. There are many opportunities for advanced techniques from machine learning to be applied to such data sets, particularly for automatic phenotype identification. More unsupervised approaches to feature extraction and measurement of cells is an interesting area of future research.

There are also many opportunities for more direct integration with other sources of biological knowledge, such as genomic, proteomic, and expression data. A more accurate picture of the relationship between knockdown target sequence and which genes' expression levels are actually reduced, and by how much, would significantly improve the generation of phenotype profiles from knockdown scores. Integrating expression and protein-protein interaction data would allow us to pick out more complex relationships and strengthen our predictive power. Our methods have been purposefully constructed in a model-based manner, to allow easy integration of multiple types of data and sources of evidence.

To conclude, we have demonstrated that meaningful biological predictions can be made from images of cells, through analysis of the effects of gene knockdowns on cellular phenotypes. We have introduced methods for measuring individual cells in images, for using these measurements to classify cells by phenotype, and for modeling the effects on phenotype for the full set of gene knockdowns in a screen. We use these methods to predict genes related to one another through a number of phenotypes, as well as to connect phenotypes with existing biological knowledge to predict gene function.

Bibliography

- [1] CellProfiler cell image analysis software, <http://cellprofiler.org>.
- [2] CellVisualizer, <http://cellvisualizer.org>.
- [3] Mitochek project, <http://mitochek.org>.
- [4] D. B. R. A. P. Dempster, N. M. Laird. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [5] K. Adamsky, J. Schilling, J. Garwood, A. Faissner, and E. Peles. Glial tumor cell adhesion is mediated by binding of the fniii domain of receptor protein tyrosine phosphatase beta (rptpbeta) to tenascin c. *Oncogene*, 20(5):609–18, 2001.
- [6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.
- [7] C. Augeron and C. L. Laboisie. Emergence of permanently differentiated cell clones in a human colonic cancer cell line in culture after treatment with sodium butyrate. *Cancer Res*, 44(9):3961–9, 1984.

- [8] S. N. Bailey, S. M. Ali, A. E. Carpenter, C. O. Higgins, and D. M. Sabatini. Microarrays of lentiviruses for gene function screens in immortalized and primary cells. *Nat Meth*, 3(2):117–122, 2006.
- [9] A. E. Baltus, D. B. Menke, Y.-C. Hu, M. L. Goodheart, A. E. Carpenter, D. G. de Rooij, and D. C. Page. In germ cells of mouse embryonic ovaries, the decision to enter meiosis precedes premeiotic dna replication. *Nat Genet*, 38(12):1430–4, 2006.
- [10] G. L. Bentz, M. Jarquin-Pardo, G. Chan, M. S. Smith, C. Sinzger, and A. D. Yurochko. Human cytomegalovirus (hcmv) infection of endothelial cells promotes naive monocyte extravasation and transfer of productive virus to enhance hematogenous dissemination of hcmv. *J Virol*, 80(23):11539–55, 2006.
- [11] S. Beucher. The watershed transformation applied to image segmentation. In *Scanning Microscopy International*, volume 6, pages 299–314, 1992.
- [12] R. A. Blaheta, E. Weich, D. Marian, J. Bereiter-Hahn, J. Jones, D. Jonas, M. Michaelis, H. W. Doerr, and J. J. Cinatl. Human cytomegalovirus infection alters pc3 prostate carcinoma cell adhesion to endothelial cells and extracellular matrix. *Neoplasia*, 8(10):807–16, 2006.
- [13] M. V. Boland and R. F. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17(12):1213–23, 2001.
- [14] P. Bork and E. V. Koonin. Predicting functions from protein sequences—where are the bottlenecks? *Nat Genet*, 18(4):313–8, 1998.
- [15] H. Brentani et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc Natl Acad Sci U S A*, 100(23):13418–23, 2003.
- [16] B. Cariou, D. Perdereau, K. Cailliau, E. Browaeys-Poly, V. Bereziat, M. Vasseur-Cognet, J. Girard, and A.-F. Burnol. The adapter protein zip binds

- grb14 and regulates its inhibitory action on insulin signaling by recruiting protein kinase ζ . *Mol Cell Biol*, 22(20):6959–70, 2002.
- [17] A. E. Carpenter, T. R. Jones, M. Lamprecht, D. B. Wheeler, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, 2006.
- [18] A. E. Carpenter and D. M. Sabatini. Systematic genome-wide screens of gene function. *Nat Rev Genet*, 5(1):11–22, 2004.
- [19] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. J. Comput. Vision*, 22:61–79, 1997.
- [20] P. M. Choi, K. M. Tchou-Wong, and I. B. Weinstein. Overexpression of protein kinase c in ht29 colon cancer cells causes growth inhibition and tumor suppression. *Molecular and Cellular Biology*, 10(9):4650–4657, 1990.
- [21] K. Clark, M. Langeslag, B. van Leeuwen, L. Ran, A. G. Ryazanov, C. G. Figdor, W. H. Moolenaar, K. Jalink, and F. N. van Leeuwen. Trpm7, a novel regulator of actomyosin contractility and cell adhesion. *EMBO J*, 25(2):290–301, 2006.
- [22] L. E. Cowen, A. E. Carpenter, O. Matangkasombut, G. R. Fink, and S. Lindquist. Genetic architecture of hsp90-dependent drug resistance. *Eukaryot Cell*, 5(12):2184–8, 2006.
- [23] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes. Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, 3(1):66–73, 2002.
- [24] E. W. Dijkstra. A note on two problems in connection with graphs. *Numerische Math*, 1:269–271, 1959.
- [25] L. G. Dodd, W. F. Moore, and C. R. Eedes. Signet ring adenocarcinoma metastatic to the bronchus and mimicking goblet cell hyperplasia. a case report. *Acta Cytol*, 43(6):1108–12, 1999.

- [26] T. Dorval, A. Ogier, E. Dusch, N. Emans, and A. Genovesio. Bias free features detection for high content screening. In *Proceedings of ISBI*, 2007.
- [27] C. Echeverri and N. Perrimon. High-throughput RNAi screening in cultured cells: a user’s guide. *Nat Rev Genet*, 7(5):373–384, 2006.
- [28] C. J. Echeverri, P. A. Beachy, B. Baum, M. Boutros, F. Buchholz, S. K. Chanda, J. Downward, J. Ellenberg, A. G. Fraser, N. Hacohen, W. C. Hahn, A. L. Jackson, A. Kiger, P. S. Linsley, L. Lum, Y. Ma, B. Mathey-Prevot, D. E. Root, D. M. Sabatini, J. Taipale, N. Perrimon, and R. Bernards. Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods*, 3(10):777–9, 2006.
- [29] A. Fire, S. Xu, M. Montgomery, S. Kostas, S. E. Driver, and C. Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391:806–811, February 1998.
- [30] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.
- [31] A. Friedman and N. Perrimon. A functional RNAi screen for regulators of receptor tyrosine kinase and erk signalling. *Nature*, 444(7116):230–234, 2006.
- [32] A. Friedman and N. Perrimon. Genetic screening for signal transduction in the era of network biology. *Cell*, 128(2):225–31, 2007.
- [33] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, apr 2000.
- [34] D. Gabor. Theory of communication. *Journal of the Institute of Electrical Engineers*, 93:429–441, 1946.
- [35] R. Garippa. A multi-faceted approach to the advancement of cell-based drug discovery. *Drug Discovery World Winter 2004/5*, pages 43–55, 2004.

- [36] E. Gaudier, A. Jarry, H. M. Blottiere, P. de Coppet, M. P. Buisine, J. P. Aubert, C. Laboisse, C. Cherbut, and C. Hoebler. Butyrate specifically modulates muc gene expression in intestinal epithelial goblet cells deprived of glucose. *Am J Physiol Gastrointest Liver Physiol*, 287(6):G1168–74, 2004.
- [37] W. H. Goldmann. p56(lck) controls phosphorylation of filamin (abp-280) and regulates focal adhesion kinase (pp125(fak)). *Cell Biol Int*, 26(6):567–71, 2002.
- [38] F. Gounari, R. Chang, J. Cowan, Z. Guo, M. Dose, E. Gounaris, and K. Khaz-
aie. Loss of adenomatous polyposis coli gene function disrupts thymic develop-
ment. *Nat Immunol*, 6(8):800–9, 2005.
- [39] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image
classification. *IEEE Trans. on Systems, Man, and Cybernetics*, 3(6):610–621,
1973.
- [40] N. Harder, F. Mora-Bermúdez, W. J. Godinez, J. Ellenberg, R. Eils, and
K. Rohr. Automated analysis of the mitotic phases of human cells in 3D flu-
orescence microscopy image sequences. In *Proceedings of MICCAI*, volume 1,
pages 840–848, 2006.
- [41] N. Harder, B. Neumann, M. Held, U. Liebel, H. Erfle, J. Ellenberg, R. Eils, and
K. Rohr. Automated recognition of mitotic patterns in fluorescence microscopy
images of human cells. In J. Kovačević and E. Meijering, editors, *Proc. IEEE In-
ternat. Symposium on Biomedical Imaging: From Nano to Macro (ISBI'2006)*,
pages 1016–1019, Arlington, VA, USA, April 6.-9. 2006.
- [42] K. A. Hartwell, B. Muir, F. Reinhardt, A. E. Carpenter, D. C. SgROI, and R. A.
Weinberg. The spemann organizer gene, goosecoid, promotes tumor metastasis.
Proc Natl Acad Sci U S A, 103(50):18969–74, 2006.
- [43] T. H. Hemstrom, M. Sandstrom, and B. Zhivotovsky. Inhibitors of the pi3-
kinase/akt pathway induce mitotic catastrophe in non-small cell lung cancer
cells. *Int J Cancer*, 119(5):1028–38, 2006.

- [44] A. Jain and D. B. Agus. Ppargamma signaling: one size fits all? *Cell Cycle*, 3(11):1352–4, 2004.
- [45] R. A. Johnson, X. Wang, X. L. Ma, S. M. Huong, and E. S. Huang. Human cytomegalovirus up-regulates the phosphatidylinositol 3-kinase (pi3-k) pathway: inhibition of pi3-k activity inhibits viral replication and virus-induced signaling. *J Virol*, 75(13):6022–32, 2001.
- [46] T. R. Jones, A. E. Carpenter, D. M. Sabatini, and P. Golland. Methods for high-content, high-throughput image-based cell screening. In D. Metaxas, R. Whitaker, J. Rittscher, and T. Sebastian, editors, *Proceedings of 1st Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI, Copenhagen)*, pages 65–72, 2006.
- [47] C. R. Kahl and A. R. Means. Regulation of cell cycle progression by calcium/calmodulin-dependent pathways. *Endocr Rev*, 24(6):719–36, 2003.
- [48] E. Kamynina and O. Staub. Concerted action of enac, nedd4-2, and sgk1 in transepithelial na(+) transport. *Am J Physiol Renal Physiol*, 283(3):F377–87, 2002.
- [49] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The kegg databases at genomenet. *Nucleic Acids Res*, 30(1):42–6, 2002.
- [50] K. B. Kaplan, A. A. Burds, J. R. Swedlow, S. S. Bekir, P. K. Sorger, and I. S. Nathke. A role for the adenomatous polyposis coli protein in chromosome segregation. *Nat Cell Biol*, 3(4):429–432, 2001.
- [51] A. Khotanzad and Y. H. Hong. Invariant image recognition by zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(5):489–497, 1990.
- [52] A. A. Kiger, B. Baum, S. Jones, M. R. Jones, A. Coulson, C. Echeverri, and N. Perrimon. A functional genomic analysis of cell morphology using RNA interference. *J Biol*, 2(1475-4924 (Electronic)):27, 2003.

- [53] V. Kovalev, N. Harder, B. Neumann, M. Held, U. Liebel, H. Erfle, J. Ellenberg, R. Eils, and K. Rohr. Feature selection for evaluating fluorescence microscopy images in genome-wide cell screens. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 276–283, Washington, DC, USA, 2006. IEEE Computer Society.
- [54] N. Kuroda, I. Yamasaki, H. Nakayama, K. Tamura, Y. Yamamoto, E. Miyazaki, K. Naruse, H. Kiyoku, M. Hiroi, and H. Enzan. Prostatic signet-ring cell carcinoma: case report and literature review. *Pathol Int*, 49(5):457–61, 1999.
- [55] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, M. J. W. Irene C Blat, J. Lerner, J.-P. Brunet, K. N. R. Aravind Subramanian, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–1935, 2006.
- [56] M. R. Lamprecht, D. M. Sabatini, and A. E. Carpenter. Cellprofiler: free, versatile software for automated biological image analysis. *Biotechniques*, 42(1):71–5, 2007.
- [57] E. G. Learned-Miller and P. Ahammad. Joint MRI bias removal using entropy minimization across images. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 761–768. MIT Press, Cambridge, MA, 2005.
- [58] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, October 1999.
- [59] T. Lesuffleur, A. Barbat, E. Dussaulx, and A. Zweibaum. Growth adaptation to methotrexate of ht-29 human colon carcinoma cells is associated with their ability to differentiate into columnar absorptive and mucus-secreting cells. *Cancer Res*, 50(19):6334–43, 1990.

- [60] J. M. Levisky and R. H. Singer. Gene expression and the myth of the average cell. *Trends Cell Biol*, 13(0962-8924 (Print)):4–6, 2003.
- [61] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM'03, 25th Pattern Recognition Symposium*, pages 297–304, 2003.
- [62] J. Lindblad, C. Wahlby, E. Bengtsson, and A. Zaltsman. Image analysis for automatic segmentation of cytoplasms and classification of rac1 activation. *Cytometry A*, 57(1):22–33, 2004.
- [63] V. Ljosa and A. K. Singh. Probabilistic segmentation and analysis of horizontal cells. In *Proceedings of the 2006 IEEE International Conference on Data Mining (ICDM)*, pages 980–985, 2006.
- [64] L.-H. Loo, L. F. Wu, and S. J. Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nat Meth*, advanced online publication:–, 2007.
- [65] S. Lowe. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval. In *Proceedings of Eurospeech-99A*, volume 6, pages 2443–2446, 1999.
- [66] E. Maidji, S. Tugizov, T. Jones, Z. Zheng, and L. Pereira. Accessory human cytomegalovirus glycoprotein us9 in the unique short component of the viral genome promotes cell-to-cell transmission of virus in polarized epithelial cells. *The Journal of Virology*, 70(12):8402–8410, 1996.
- [67] N. Malpica, C. O. de Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289–97, 1997.
- [68] M. Martin, P. Simon-Assmann, M. Kedinger, M. Martin, P. Mangeat, F. X. Real, and M. Fabre. Dcc regulates cell adhesion in human colon cancer derived ht-29 cells and associates with ezrin. *Eur J Cell Biol*, 85(8):769–83, 2006.

- [69] M. Masutani, T. Nozaki, K. Wakabayashi, and T. Sugimura. Role of poly(adp-ribose) polymerase in cell-cycle checkpoint mechanisms following gamma-irradiation. *Biochimie*, 77(6):462–5, 1995.
- [70] B. S. Meyer F. Morphological segmentation. *J Visual Communication Image Representation*, 1:21–46, 1990.
- [71] J. Moffat, D. A. Grueneberg, X. Yang, S. Y. Kim, A. M. Kloepper, G. Hinkle, B. Piqani, T. M. Eisenhaure, B. Luo, J. K. Grenier, A. E. Carpenter, S. Y. Foo, S. A. Stewart, B. R. Stockwell, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini, and D. E. Root. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 124(0092-8674 (Print)):1283–98, 2006.
- [72] M. D. Nguyen, B. Plasil, P. Wen, and W. L. Frankel. Mucin profiles in signet-ring cell carcinoma. *Arch Pathol Lab Med*, 130(6):799–804, 2006.
- [73] B. Nicke, A. Kaiser, B. Wiedenmann, E.-O. Riecken, and S. Rosewicz. Retinoic acid receptor [alpha] mediates growth inhibition by retinoids in human colon carcinoma ht29 cells. *Biochemical and Biophysical Research Communications*, 261(3):572–577, 1999.
- [74] A. Nissan, J. G. Guillem, P. B. Paty, W. D. Wong, and A. M. Cohen. Signet-ring cell carcinoma of the colon and rectum: a matched control study. *Dis Colon Rectum*, 42(9):1176–80, 1999.
- [75] A. J. Olaharski, Z. Ji, J.-Y. Woo, S. Lim, A. E. Hubbard, L. Zhang, and M. T. Smith. The histone deacetylase inhibitor trichostatin a has genotoxic effects in human lymphoblasts in vitro. *Toxicol Sci*, 93(2):341–7, 2006.
- [76] C. Ortiz de Solorzano, E. G. Rodriguez, A. Jones, D. Pinkel, J. W. Gray, D. Sudar, and S. J. Lockett. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *J Microsc Oxford*, 193:212–226, 1999.

- [77] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, Mar. 1979.
- [78] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(1095-9203 (Electronic)):1194–8, 2004.
- [79] K. Pohl. *Prior Information for Brain Parcellation*. PhD thesis, MIT, 2005.
- [80] C. Rochette-Egly, M. Kedinger, and K. Haffen. Modulation of ht-29 human colonic cancer cell differentiation with calmidazolium and 12-o-tetradecanoylphorbol-13-acetate. *Cancer Research*, 48(21):6173–6182, 1988.
- [81] K. Rodenacker and E. Bengtsson. A feature set for cytometry on digitized microscopic images. *Anal Cell Pathol*, 25(1):1–36, 2003.
- [82] D. Root. personal communication, 2007.
- [83] D. E. Root, B. P. Kelley, and B. R. Stockwell. Detecting spatial patterns in biological array experiments. *J Biomol Screen*, 8(4):393–8, 2003.
- [84] R. F. Saidi, K. Jaeger, M. H. Montrose, S. Wu, and C. L. Sears. Bacteroides fragilis toxin rearranges the actin cytoskeleton of ht29/c1 cells without direct proteolysis of actin or decrease in f-actin content. *Cell Motil Cytoskeleton*, 37(2):159–65, 1997.
- [85] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.
- [86] C. Schwartz, V. Palissot, N. Aouali, S. Wack, N. H. C. Brons, B. Leners, M. Bosseler, and G. Berchem. Valproic acid induces non-apoptotic cell death mechanisms in multiple myeloma cell lines. *Int J Oncol*, 30(3):573–82, 2007.
- [87] E. Shtivelman, J. Sussman, and D. Stokoe. A role for pi 3-kinase and pkb activity in the g2/m phase of the cell cycle. *Curr Biol*, 12(11):919–24, 2002.

- [88] J. G. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261, 1948.
- [89] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. In *Proc. Natl. Acad. Sci. USA*, volume 102, pages 15545–15550, 2005.
- [90] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12, 1999.
- [91] K. Tieu and P. Viola. Boosting image retrieval. *Int. J. Comput. Vision*, 56(1-2):17–36, 2004.
- [92] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 29(5):In press, May 2007.
- [93] C. Wählby. *Algorithms for Applied Digital Image Cytometry*. PhD thesis, Uppsala University, 2003.
- [94] C. Wahlby, I.-M. Sintorn, F. Erlandsson, G. Borgefors, and E. Bengtsson. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. *J Microsc*, 215(Pt 1):67–76, 2004.
- [95] M. G. Walker, W. Volkmuth, E. Sprinzak, D. Hodgson, and T. Klingler. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res*, 9(12):1198–203, 1999.

- [96] L. Wang, R. A. Baiocchi, S. Pal, G. Mosialos, M. Caligiuri, and S. Sif. The brg1- and hbrm-associated factor baf57 induces apoptosis by stimulating expression of the cylindromatosis tumor suppressor gene. *Mol Cell Biol*, 25(18):7953–65, 2005.
- [97] T. Watanabe, T. Jeziorowski, W. Wuttke, and D. Grube. Secretory granules and granins in hyperstimulated male rat gonadotropes. *J Histochem Cytochem*, 41(12):1801–12, 1993.
- [98] W. Wells, W. Grimson, R. Kikinis, and F. Jolesz. Adaptive segmentation of MRI data. In *IEEE TMI*, volume 15, pages 429–442, 1996.
- [99] D. B. Wheeler, S. N. Bailey, D. A. Guertin, A. E. Carpenter, C. O. Higgins, and D. M. Sabatini. Rnai living-cell microarrays for loss-of-function screens in drosophila melanogaster cells. *Nat Meth*, 1(2):127–132, 2004.
- [100] S. Zeissig, A. Fromm, J. Mankertz, J. Weiske, M. Zeitz, M. Fromm, and J.-D. Schulzke. Butyrate induces intestinal sodium absorption via sp3-mediated transcriptional up-regulation of epithelial sodium channels. *Gastroenterology*, 132(1):236–48, 2007.
- [101] X. Zhou, X. Cao, Z. Perlman, and S. T. C. Wong. A computerized cellular imaging system for high content analysis in monastrol suppressor screens. *J Biomed Inform*, 39(2):115–25, 2006.