Developing a computational pipeline for high throughput quantitative phenotyping

By

Mark Dane

Presented to the
Department of Medical Informatics and Clinical Epidemiology
and the
Oregon Health and Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Master of Science

June 2014

School of Medicine

Oregon Health and Science University

CERTIFICATE OF APPROVAL

This is certify that the Master's thesis of

Mark Dane

has been approved

____

Shannon McWeeney, Ph.D.

____

Laura M. Heiser, Ph.D.

____

Zhi Hu, Ph.D.

Table of Contents

Acknowledgements

My mid-life career change has relied on a few incredible people to teach me the ways of computational biology and support me through this transition. Dr. Shannon McWeeney created my academic path and matched her high expectations with deep knowledge of the field and how to bring me into it. I'm grateful for her energies, guidance and persistence. My abilities stem from her teachings and my shortcomings are of my own doing.

Drs. Laura Heiser, Juha Rantala and Nick Wang have generously taught me how to use my knowledge in a working lab. They have trusted me as a partner in their work, focused my efforts on appropriate topics and been patient as I stumbled and slipped along the way.

Dr. Joe Gray created the opportunity for me to become part of his lab and has provided steadfast backing of my analysis. Drs. Zhi Hu, Jim Korkola and Amanda Esch have helped me understand the biology of their experiments and the results of our analysis. Spencer Watson and Dr. Sophia Bornstein have kindly allowed me to use the data from their experiments in my thesis.

The stalwart love and support of my wife Marti Dane has carried me past feelings of overwhelming stress and difficulty. She's learned all of the topics as I learned them, reframed our challenges to make them achievable and parented our children during my transformation. I am completely and forever (and happily) in her debt.

Abstract

Cellular microarrays are a combination of well plate and microarray technologies that create an efficient high-throughput method to quantitate the phenotypes of cells in response to perturbations such as microenvironments, RNAi and drugs. This thesis covers the development and use of an extensible computational pipeline to process High Content Screening data from cellular microarrays. Extensive Exploratory Data Analysis is performed to identify technical variations and inform normalizations that filter for biological variations. Robust rank product scoring combines replicate and channel data to prioritize hits for different biological questions.

The pipeline has been applied to Cell Spot Microarrays and MEArray datasets, uncovering experiment design issues and generating hit lists for validation. A method to generate simulated datasets with variations that mimic those in actual microarrays has been developed and used to optimize the pipeline and predict the results for different experiment designs.

Introduction

Cell-Based Experiments

Human cell lines are generally accepted as effective models for understanding biological mechanisms and identifying therapeutic drugs.(1) Since the mid-1980s with the establishment of the US National Cancer Institute 60 cell lines(2) and later as 100's of additional cell lines have been immortalized and quantified (3,4), drug discovery often begins with cell line experiments. An important aspect of cell lines is their ability to recapitulate the phenotypes of their source tumors(5). This enables discoveries made in cell lines to be transferable to humans.

Many cell line experiments are performed in well plates where isolated cell populations grow exposed to controlled perturbations. New technologies are merging well plate and microarray methods to enable high throughput cellular microarray experiments.(6–8) Immunofluorescent stains, automated microscopes, machine vision and statistical analysis form the core of these high-throughput experiments that perturb cell populations in known ways and measure the resulting phenotypes. Initially performed in well plates these cell-based experiments are moving into microarrays, increasing the density of experiments by several magnitudes. While initial well plate experiments reported a single measure of cell population viability(9,10), Immunofluorescent stains and machine vision quantitate and locate up to four specific types of molecules along with cell morphology and texture.(11)

Initial applications of cellular microarray experiments performed in Oregon Health and Science University's Center for Spatial Systems Biomedicine have focused on cancer research using established cell lines. These experiments create perturbations through either short interrupting RNA (siRNA) knockdown of genes or through the manipulation of the extracellular matrix proteins and growth factors. The siRNA experiments follow the protocol of Rantala et al.

to create Cell Spot Microarrays (CSMAs). (6,12) The Microenvironment microarrays (MEArrays) follow a protocol under development led by LaBarge et al.(13,8,14) While the spotted perturbations differ, the downstream imaging and data analysis of CSMA and MEArrays are similar.

Cellular well plate experiments and expression microarray experiments have mature workflows with computational pipelines that prepare their raw data for analysis. Cellular microarrays are an emerging technology that blends these two methods, creating a need for a new pipeline. A cellular microarray pipeline can be assembled from established well plate and microarray methods in a modular fashion. The modules can be selected based on robustness, interpretability, ease of implementation, computational speed and supportability.

The cellular microarray pipeline discussed here includes extensive Exploratory Data Analysis (EDA) to detect batch effects, outliers and signal variances. The EDA also guides selection of normalization approaches that remove technical sources of variation. Scoring will prioritize the normalized data within the context of the experiment's biological questions.

Microarray experiments in this context are typically positioned as screens at the beginning of a series of experiments(15) with the role of evaluating a large number of perturbations and determining which warrant further investigation. Imaging the microarrays with a high-speed, low-resolution scanning system that gathers population-level data is an efficient way to investigate each microarray population.(7) This High Throughput Screening (HTS) approach uses fluorescent intensity data at the population level. HTS uses simple machine vision methods to quantitate protein and DNA levels but does not determine any cell-level data such as morphology, texture or location. Statistical analysis of HTS data identifies which populations differ from the majority, often accepting a high false positive rate in order to capture more true positives(16).

High Content Screening (HCS) is an example of a moderately resource intensive next step after HTS. HCS uses 10-20x magnification and machine vision to capture individual cell data.

This can be aggregated to create population-level data or analyzed at the cell level. Using HTS to define 10-20% of microarray populations as hits and then applying HCS to these populations provides a method that captures many high-quality hits at reasonable resource levels. The data generated from HCS can be combined with other omics data and/or used to define hits for new downstream experiments.

In-vitro Cell Line Experiments

Since Robert Hooke first viewed cells through a microscope in 1655(17) scientist have been learning about biological processes by direct imaging. Today, immunofluorescent staining can locate and quantitate specific molecules associated with molecular processes such as proliferation, apoptosis and differentiation.(12,18) Automated microscopes with motorized XY stages, autofocus, multiple wavelength light sources, color filters and high-resolution cameras routinely gather the images of thousands of cell populations.(19) Machine vision algorithms transform these images to numeric measurements, enabling computational biologists to apply statistical methods to determine the state of the cells.(20)

In vitro biological experiments that use living cells from established cell lines started with HeLa cells from Henrietta Lacks' ovarian cancer tumors.(21) Cell-based experiments rely on the immortalized cells being similar to in vivo cells,(5) those in a living organism. This similarity enables cell line-based experiments to efficiently yield important contributions to biological understanding. The cell lines used in this thesis all stem from cancer cells. They emulate human in vivo cells and can be ethically subjected to lethal doses of drugs, small molecules or radiation.

Researchers are still learning about the limitations of in vitro cell-based experiments. The limitations that come from using a simplified model are important when designing a pipeline of experiments that includes drugs for human diseases. Each experiment in the pipeline typically uses a model that is more human-like and has increasing costs and complexity. Typically lacking in microarray assays are physical perturbations such as stretch and pressure changes(22) and cell-to cell signaling and interactions that come from 3-dimensional cell-based contact(23).

Well Plates

Many cell-based assays have been carried out in well plates with 96, 384 or 1536 clear-bottomed wells per plate(9,10). Well plates enable populations of 100's to 1000's of cells to grow in the same macro-environment while each well is isolated, exposing the populations to different treatments. Often, CellTiterGlo is used to measure the amount of ATP in a well population and interpreted as a biomarker for cell count. This is a mature technology that has been scaled up for large experiments. The process is tightly controlled and the analysis is well defined(24–27). Experiments that span 100's of plates are often run in automated core facilities where automated equipment and specialized staff (28) generate reproducible results that contain minor technical variations.

The downsides to well plate-based experiments include the cost of the reagents used in the wells, the need for specialized handling equipment and the corresponding lab space needs. Equipment needs extend to dedicated pipetting equipment, well plate storage systems, well plate handling robots and material handling compatible measuring systems. Another downside is there are known spatial affects around the perimeter of a well plate due to a gradient of evaporation. This is typically mitigated by leaving these wells empty or using them for controls.

Well plate treatments used in screening assays can include gene silencing with RNA interference (RNAi)(29,30), exposure to different microenvironments, exposure to drugs, and combinations of two or three of these perturbations.

siRNAs are RNAi molecules with 21 nucleotide sequences that become part of an RNA-induced silencing complex (RISC). RISCs are endonucleases that use siRNAs to base pair match to mature RNA transcripts, and then degrade them through a process known as gene silencing.

This has proven to be an effective tool to discover a gene's impact on a cell population's phenotype. While highly selective, the 21 base sequences in siRNAs do have differing on-target and off-target rates(31) that vary across cell lines. The variation in these rates is a technical artifact that can be mitigated through the use of pooled siRNAs. The target gene for the pool will be the same while the off-target genes are typically different. This reduces the phenotype response from the off-target genes and concentrates the response due to silencing the on-target gene, even in situations where only two of the siRNAs in the pool have high on-target rates.

Microenvironments include the extracellular matrix proteins, growth factors structural elements and scaffolding that affect cells through signaling and physical interactions.(32) Microenvironments are implicated in several processes important to cancer research such as metastasis and drug resistance. Specific microenvironments can be mimicked and controlled in well plates by letting the cells grow in contact with any combination of proteins or growth factors.(13,33,34)

Drugs are a common perturbation in well plate assays. However, analysis of drug perturbations is not directly addressed in this thesis.

Manufacturing Cellular Microarrays

Cell based microarrays miniaturize the screening experiments that are typically run in well plates. Experiments with spot sizes that enable approximately 100 cells to adhere can go from fifteen 384 well plates to a single microarray that is 1/8 the size of a well plate. Four microarrays printed on a convenient divided plate show a 60:1 reduction in footprint over 384 well plates. Descriptions of the two types of cell based microarrays addressed in this thesis follow.

Cell Spot Microarrays

There have been several methods published for spotting siRNAs onto flat plates to form microarrays.(6,7,35) As described in Rantala *et al*., Cell Spot Microarrays (CSMAs) are created by contact printing siRNAs, transfection agents and extracellular matrix gel onto flat plastic plates. The microarray manufacturing process creates pin spotted local regions referred to as pin grids. Ranging from 7x7 to 11x11 rows and columns, all spots within a pin grid are created by the same pin. Each pin grid defines a subgroup that may differ from the rest due to the physical characteristics of the pins and their holders. Depending on the spot size and spacing, a CSMA can have 2000-6000 spots. After printing, a CSMA is stable and can be stored for months.

The protocol for CSMA experiments is to flood the microarray with a suspension of cells for a short period of time, enabling cells to adhere to the spots. After a light rinsing, cells release from in between the spots and are cultured as separate populations exposed only to the reagents of their spot. Typically after 24-72 hours, the cells are fixed and stained for analysis.

Microenvironment Microarrays

Microenvironment microarrays (MEArrays) share CSMA manufacture and experiment protocols with the exception of the printed reagents. Instead of spots of siRNAs, MEArrays have extracellular matrix (ECM) proteins, adhesion molecules, growth factors and cytokines. Typical ECM proteins used in MEArrays are from the collagen family, fibronectin, and lamilin among others. Typical growth factors and cytokines include the WNT family, interferon, angiopoietins and more. MEArrays are also used to test drug interactions by exposing all cell populations on a microarray to the same drug concentration.

Imaging and Machine Vision

The phenotypes of cells in well plate experiments are quantitated by machine vision collecting one or more measurements from immunofluorescently stained cells. Measurements can be made at the individual cell or population level. The level of the measurements determines the analysis methods and the potential for results(36). The simpler level is measuring the total intensity of one fluorescent biomarker such as CellTiterGlo (CTG) that is proportional to the number of cells in a population. Multiple lasers emitting at different frequencies or wide spectrum light sources that are filtered to desired frequencies can provide four distinct excitation colors. Antibodies labeled with different colored fluorophores can then report on four different biological endpoints. With one endpoint reporting cell count, the others can be normalized to the cell count to determine per cell activity of the other endpoints.

The next step up in complexity is to increase the magnification to 10x or higher and use machine vision to directly measure individual cell values. Machine vision at the cell level first segments the image to identify cells, then measures the per cell biomarker intensities and morphological features such as size, circularity and texture. Derived features such as the ratio of nucleus area to overall cell area or cell cycle state based on several features can also be used in an analysis. A basic feature vector from machine vision based on individual cells might have only the cell counts and the biomarker intensities while a complex feature vector adds forty morphology features of the nucleus and cytosolic areas.(35)

Impacts of High Throughput Methods

Both CSMAs and MEArrays enable scaling up experiments at lower costs and with minimal specialized equipment. In a study of microarrays similar to CSMAs, Wood *et al*. found a 15-30-fold decrease in costs of experiments of microarrays vs. well plates. Lower costs and handling constraints make it feasible for researchers to expand experiments to cover multiple cell lines, thousands of perturbations and tens of end points.

The transition from well plates to spotted microarrays results in two potential sources of compromises to the data. As the cell populations reduce from 1000’s of cells in each well to 50-200 cells at each spot, there is a greater possibility that measurements are not statistically robust. That is, measurements can be skewed by a small number of outlier cells. Robustness can be improved by increasing the number of replicate spots and plates while still maintaining a large cost and complexity advantage over well plates. Analysis from lower cell populations can use robust methods and stick to conservative conclusions to overcome potential issues. This type of analysis remains compatible with using the microarrays as screens to determine hits for downstream processing.

The second potential source of compromises to the data comes from the cell seeding and staining methods that flood the well that the microarray sits in. All spots are exposed to the same solution but gradients in the number of cells and the amount of stain taken up by the cells can appear in the data.

Additional challenges exist on the computational side of microarray experiments due to the vast amount of data they generate. The big picture analysis plan is to combine data generated across microarrays and identify phenotypes of interest. Combining the data is preceded by an extensive Exploratory Data Analysis (EDA) with a goal of ensuring the dataset is free of unexpected technical artifacts.(37) EDA also informs the normalizations that will be used to remove the technical variations and combine the biological results. Finally, scoring will look across the endpoints individually or in combinations to determine which hits match the biological questions. Throughout the analysis, results must be visualized to aid in understanding. This computation analysis and visualization requires flexible and robust tools such as those provided in the language R(38) and associated Bioconductor packages. (39)

Exploratory Data Analysis

Microarrays are used in large experiments that run over weeks and months leading to differences in reagents, machine performance, operator performance, environmental conditions and other technical variations. In addition to informing the normalization methods and parameters designed to minimize these technical variations, EDA will determine if any microarrays must be excluded from an analysis. Using primarily visual techniques, EDA identifies microarrays with abnormal variations, excessive outliers and potential confounders.

EDA Plots and Summaries

Since cellular microarrays share many characteristics with expression microarrays and pin-spotted cDNA microarrays, the EDA techniques developed for those platforms are appropriate to be used on cell based microarrays(40). Table 1 highlights some common microarray EDA methods and their areas of focus.

Table 1 Focus of visualization for cellular microarray EDA

Plot Type	Data to be plotted	Focus of visualization
Pseudo Images	Channel values across plates	Spatial variations
Spatial Variation Index	Channel values across plates	Spatial variations
Scatter plots	Channel-to-channel	Correlations between channels
	Replicate-to-replicate	Correlations between replicates
Boxplots	Channel values across pin grids	Pin-to-pin variations
	Coefficients of Variation	Replicate analysis
	Channel values across plates	Plate to plate variations
Histograms	Channel intensities within a plate	Shape of distribution and location of controls
QQ Plots	Channel values across plates	Normality of distribution

Normalization

The data from cellular microarrays must be normalized before it can be analyzed. The goal for normalization is to reduce systemic variations so that biological variations can be compared across channels, microarrays and cell lines. CSMAs and MEArrays are hybrids of spotted slide microarrays and well plates. Both spotted slide microarrays(40–46) and well plate experiments(46,47) have extensive normalization methods developed and characterized and some of these are appropriate for CSMAs and MEArrays.

Normalizations can begin at the per channel pin level, proceed to the per channel microarray level, and then be combined across channels and microarrays. Since normalization can undesirably remove biological signal, it is prudent to do the least amount of normalization that allows combining data within the experiment. It is also important to match the normalization to the source of technical variation. For instance, if the source is due to pins performing differently, a pin-based normalization will be effective. If instead, the variations are solely due to global variations such as staining or cell adhesion gradients, then pin grid normalization may be skipped and microarray-level normalizations can be performed. Some data collected at the population level will need to be normalized to data that is proportional to the cell count of each spot.

After normalization, control treatments should be analyzed for separation of the positive controls from the negative controls to provide a QA check on each microarray. The separations can be quantified using well plate methods based on calculating the Z’-factor.

Scoring

Scoring is the process that prioritizes what should be included in the downstream analysis based on the biological questions addressed by the experiment. Multi-channel experiments that track multiple phenotypes provide multiple reasons for selecting hits(48). This ability to answer a broader range of biological questions increases the complexity of the scoring and requires it to be tailored to the experiment.

Reproducibility

Reproducible research is the standard that the data and code used in an analysis is available for others. (49) In addition to confirming initial results, the data and code will enable others to pursue different biological questions from the same data. Cellular microarray experiments generate enormous amounts of data that can be analyzed to answer many different biological questions. Many questions will focus on a subset of the data, perhaps only a type of cell line or particular end points. The type of questions being asked will determine which types of EDA, normalization and scoring are used within the pipeline. Having the code and data available enables other researchers to explore different aspects of the data. To enable reproducibility, the data and all code in the pipeline must be easily accessible as on a public repository or a published GitHub site.

Overview of this Thesis

This thesis concerns developing an extensible computational pipeline for cellular microarrays. Starting with raw imaging data and annotations, the pipeline reproducibly creates a fully annotated dataset of known quality that is ready for downstream analysis. Initially, the cellular pipeline acts at a screen level, producing an unordered list of hits to be validated. This is achieved with modules selected for robustness and ease of integration so that all aspects of the pipeline are functioning. The object-based structure facilitates modularity and extensibility. As different computational techniques are employed, the pipeline will be optimized to prioritize perturbations as opposed to including them in an unordered list.

A screen level pipeline has immediate value for improving cellular microarray technologies. The effects of experiment designs for randomization, control types and replicate numbers can all be quantified. Lab processes such as spotting, cell adherence, staining, imaging and machine vision analysis can be optimized across cell lines.

This thesis develops and evaluates a screen level cellular microarray pipeline and its performance on a simulated dataset and two initial datasets. Implemented around an existing screen-level well plate pipeline, the pipeline also includes microarray EDA, normalization and scoring methods. Initial use of the pipeline has uncovered layout and technical artifacts in preliminary datasets while guiding the design of future experiments and protocols.

Materials and Methods

Dataset Descriptions

To illustrate the pipeline, two experimental data sets and one simulated data set are utilized

Table 2 Characteristics of the datasets

Dataset	Randomization	Replicates	Controls	Treatments
Aberration	Yes	None	Negative controls only	siRNAs targeting 440 genes
MEA	No	6 replicates within each microarray 4 replicate microarrays	None	Combinations of 16 ECM and 24 growth factor molecules
Simulated	Yes	3 replicates within each microarray 4 replicates microarrays	Positive, negative and simulated hits	Controlled perturbations of channel values

Aberration

This dataset is from a siRNA-based CSMA experiment that targets 440 genes (see Appendix A, Table A1) in the HER2-positive breast cancer cell line HCC1954. The channels measure the DNA content, products of apoptosis, proteins associated with DNA damage and the levels of newly synthesized DNA. The image acquisition and machine vision analysis was performed on an Olympus scan^R automated microscope further described in Appendix C. All measurements in this dataset are on a per cell basis. The measurements are the average for each cell in the image, independent of how many cells are being measured.

MEA

The MEA dataset is a subset of the data from an experiment that uses a format based on MEArrays. This dataset has treatments that are all 384 possible pairings of 24 ECM proteins and 16 growth factors (tables A2, A3). The data comes from population level measurements of the

CellMask plasma membrane stain and is proportional to cell count. The image acquisition was performed on a Teacan LS Reloaded (Appendix B) laser scanner using Array-Pro software to run the scanner and measure the intensities.

Simulated

The simulated dataset is created with a randomized layout, positive and negative controls, two replicate microarrays and one channel of values. Technical and biological variations are modeled as factors based on those seen in actual datasets. The factors are multiplied by each other and the base values to determine the sample, hits and positive control values.

There are several known global microarray effects along with unknown but anticipated effects. There are also natural stochastic variations due to using living cells as the basis for study. These variations are modeled by sampling from a gamma distribution. The gamma distribution was chosen as it starts at 0 and can be shaped to emulate the actual intensity values. The sampled base values are shown in Equation 1 where BV_{ij} is the base value in the ith row and jth column.

$$SV_{ijm}= BV_{ij}*PGF_{k}*RF_{ijm}*MSF_{ijm} \qquad \text{(Equation 1)}$$

The simulated dataset reproduces the effects of the pin spotting process by assigning different means and standard deviations to pin grid factors referred to as PGF_k in equation 1 where k goes from 1 to the number of pins in the pin head. Since the pins and holders are manufactured to be similar, these pin means and standard deviations are randomly sampled from a normal distribution.

Replicate microarrays are modeled as variations from a base microarray by multiplying by a matrix of replicate factors RF_{ijm} in Equation 1. The factors are chosen from a normal distribution centered at one and having a selectable standard deviation. The base microarray is not used as one of the replicates.

In addition, some spatial effects are attributed to uneven staining, characteristics of the plastic microarray substrate and physical contamination. These are modeled as ellipsoidal and linear variations with selectable locations and intensities with parameters chosen to appear similar to variations seen in actual datasets. The microarray spatial variation factors, MSF_{ijm} in Equation 1 are the product of the individual spatial variation factors associated with the ellipsoidal and linear variations at each ith row and jth column spot.

SV_{ijm} is the final sample value in the ith row and jth column of the mth replicate.

Controls and Hits

Decreasing the values of specific spots simulates positive controls and hits. Positive controls are blocked throughout the microarray while hits are randomly located. After the sample base value is calculated, the hit or controls factor CF_{ijm} is multiplied to determine the final control value CV_{ijm}.

$$CV_{ijm}= SV_{ijm} * CF_{ijm} \qquad \text{(Equation 2)}$$

$$HV_{ijm}= SV_{ijm} * HF_{ijm} \qquad \text{(Equation 3)}$$

In a similar fashion, hit values HV_{ijm} are created by multiplying the sample values by the hit factor HF_{ijm}.

The ‘strength’ of the positive controls and the hits are determined by sampling from distributions of factors that are centered at numbers between zero and one. For instance, positive control factors centered at 0.4 and hits centered at 0.7 will yield an average control knockdown of 60% and a corresponding hit knockdown of 30%.

Negative controls are at blocked locations that differ from the positive controls and hits. The values associated with the negative controls are determined by the sample and replicate factors so that they provide a subset that represents the non-positive control, non-hit populations. The pseudo images of the intermediate and final simulated microarrays are shown in Figure 1.

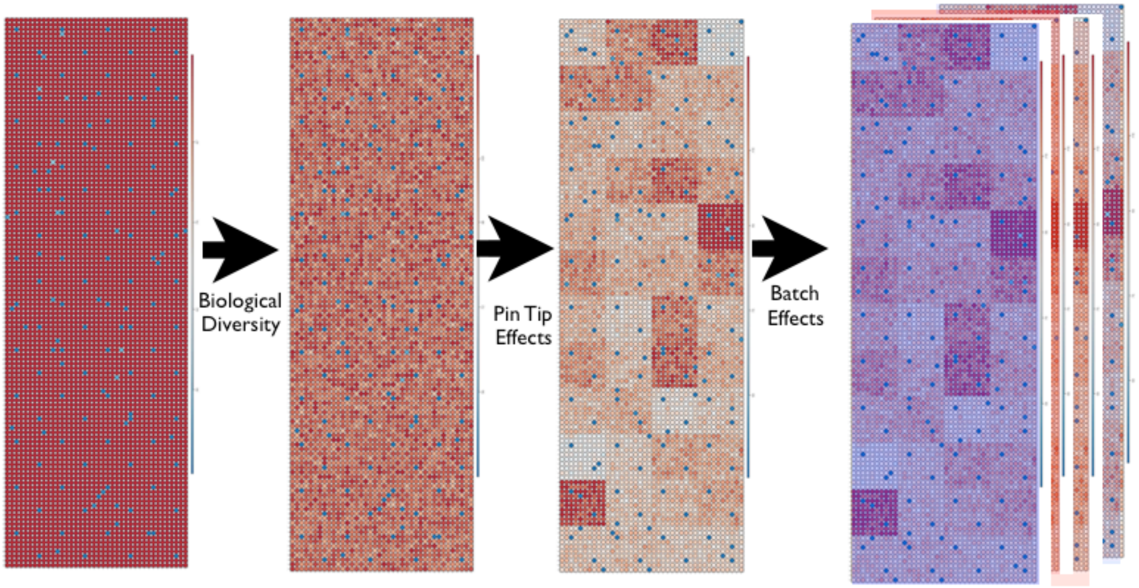


Figure 1 These images show the results from simulated variations in a 4800 spot microarray. The samples start at an intensity of 100, the controls are distributed around an intensity of 30 and hits are distributed around an intensity of 70. The effects of adding each variation are shown in the microarray pseudo-images.

Computational Pipeline

The computational pipeline that prepares the imaging results and annotations for downstream analysis is performed by adding microarray-specific capabilities to the Bioconductor cellHTS2(50) package. cellHTS2 is designed for experiments that are run in 96 and 384 well plates. The pin grids, and global cell seeding and staining attributes of the microarray workflow cause significant differences between well plates and microarrays. These differences define the need for EDA, normalization and visualization extensions that include pin grids and spatial variations. Additionally, cellHTS2 is missing vital support for the analysis of technical replicates within a microarray. Technical replicate EDA, QA and scoring are added to the pipeline as part of this thesis.

There is significant formatting performed to structure the annotations and data for input to the pipeline. A key part of the annotations details the treatments at each spot. These come from the spotting machine in a GAL file format. The imaging data, cell line and staining set information all come from either the Teacan laser scanner or the Olympus scan^R. The annotations and imaging data are merged on the microarray well locations. Additionally meta-data on the experiment such as the experimenter names, dates, and input file names and analysis source code are combined with the results to provide a reproducible framework.

EDA Plots and Evaluation

All EDA plots and statistics are generated using R and Bioconductor packages. Details on each of the EDA components utilized are described below.

Pseudo Images

Plotting an image of a microarray that is colored by intensity values helps to reveal spatial artifacts that might distort the analysis (51,52). An ideal pseudo image of a microarray would show values that are randomly located. Variations due to row, column, perimeter, gradient and localization effects can be visually detected in pseudo images. The magnitude of these spatial effects determines their impact on the dataset. Automatic scaling of the colors to the intensities will maximize the appearance of the variations and so this must be taken into account to avoid a false determination of significant spatial variation.

Spatial Variation Index

A shortcoming of pseudo images becomes apparent in large datasets that contain many channel, drug set and cell line combinations. It is difficult to subjectively identify which of the images should be investigated. Therefore, objective values that quantify the spatial variations are helpful in ranking the images. Existing methods(53, 54) including normalized unscaled standard error (NUSE)(51) and Global NUSE, (GNUSE). (52) “NUSE provides a measure of the precision of its expression estimate on a given array, i, relative to other arrays in the batch ”. (52) GNUSE extends NUSE by replacing the batch with all published microarrays that use the same platform. These methods are not sensitive to whether the high variance spots (or probes) are grouped or spatially distributed. This is significant in cellular microarrays where spatially concentrated staining and time-based imaging variations have been identified. We have created a new value termed Spatial Variation Index (SVI) as part of this thesis. The goal is to quantify the regions of spatial variations on a per microarray channel basis and exclude contributions from isolated spots. An SVI is computed for a microarray channel as follows:

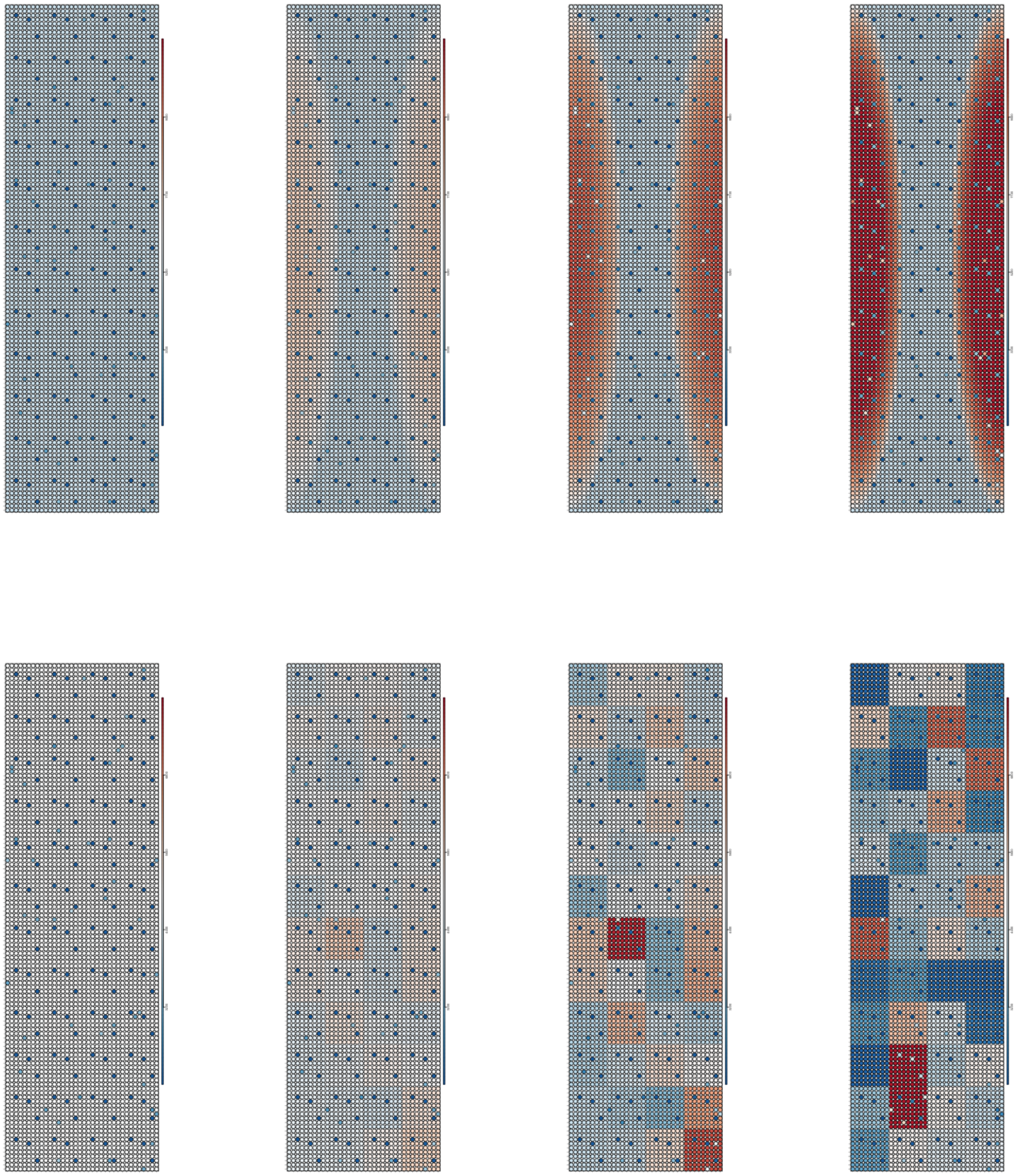
1. Fit a bivariate local regression model to the microarray channel data, defining the rows and columns as the x and y coordinates and the channel values as the z coordinate. Choose parameters that include enough neighboring spots in x and y such that the isolated spots with large z differences from their neighbors have little impact on the model. Initial testing of SVIs on the Simulated dataset used a value of nn=0.01. This results in a 3 dimensional model of the microarray that has been smoothed in the z dimension.

2. Sum the absolute values of the difference of every z value in the model from the model's median z value. This captures the total smoothed variation in the model.

3. Divide the sum of the absolute value difference by the median to normalize to different intensity levels.

4. Divide the median normalized value by the number of spots in the microarray to get the final SVI value. This last step enables comparing SVIs from different sized microarrays.

Figure 2 shows a monotonic relationship between SVI values and spatial variations due to regional artifacts and pin grid differences. Figure 2 also shows that SVI's are not sensitive to perturbations that are at isolated spots. This shows that SVIs can rank microarrays according to their regional and pin grid spatial variations.



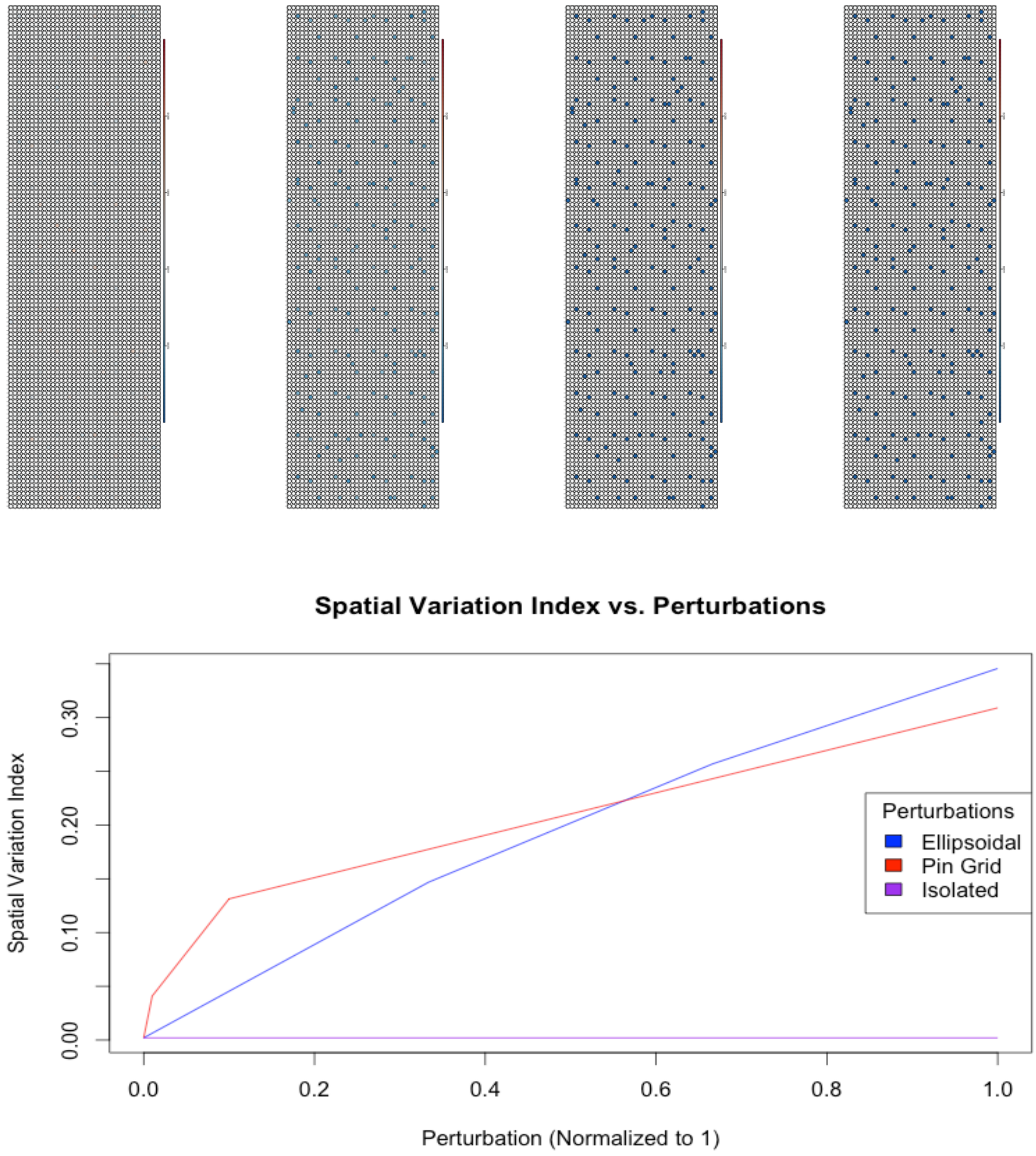


Figure 2 Pseudo Images and Spatial Variation Indices – The four images in the top row show pseudo images with increasing ellipsoidal perturbations, the second row shows increased variation in the pin grids while the third row shows increasing variations in the isolated control and hit locations. The blue and red lines on the bottom graph show the SVIs for the ellipsoidal and pin grid variations increase with increasing perturbation. The purple line shows that SVIs do not change as the isolated perturbations increase. This graph shows SVIs are sensitive to regional variations while being insensitive to isolated variations.

Pseudo Images for Layout

Pseudo images are also useful to visually check the layout of treatments in a microarray. This process begins with assigning contrasting colors to each subgroup of treatments. The definition of a treatment subgroup is treatments with biological reasons to cause similar reactions. Examples include a family of growth factors or siRNAs that target genes whose products from a complex. Creating a pseudo image colored by the treatment subgroups will show localizations of treatments as areas with the same color. These areas may violate assumptions of a random layout and compromise the downstream analysis.

The layout of perturbations on a microarray must be randomized so that the biological effects are not confounded with location. A simple example of confounding would be to place all replicates of a perturbation in a contiguous region. It would not be possible to separate the variations due to the perturbation from variations due to the location. MEArrays are combinations from relatively small sets of compounds and preliminary datasets have had structured layouts. It is important to ensure that the compounds are randomly located throughout the microarray. Pseudo images that assign contrasting colors to each compound can reveal structure in a layout as described in the MEA dataset.

Pseudo Images to locate actual Images

Laser scanners gather low-resolution images of every spot on a microarray resulting in a single image with thousands of cell populations. It is a useful quality check to look at the spot images associated with specific treatments to check for the presence of cells, focus issues and spot misalignments. A variation of the pseudo image for layout helps locate the desired cells by assigning the target subgroup a color of transparent. When this image is overlaid on the actual image of the microarray, only the target subgroup will show through the transparent window, quickly showing the images of those cell populations.

Scatter Plots, Boxplots and Histograms

Scatter plots effectively display how microarray data is correlated across pin grids, channels and treatments. Replicate microarrays are checked for high levels of correlations using scatter plots and by calculating Spearman rank correlation coefficients. Applying unique colors to the controls and hits is an effective way to show their distributions in scatter plots.

Boxplots and histograms show subgroups of the dataset intensities and statistics. Subgroups include plates, ECM and Growth Factor molecules and replicates.

Normalization

The goal for normalization is to remove systemic variations so that biological variations can be analyzed across pin grids, subarrays, channels and cell lines. Cellular microarrays are a hybrid of spotted slides and well plates and it is proposed to combine normalization methods used in both of these domains. HTS analysis of cellular microarrays is based on a single intensity value per channel at each spot. The channels will be normalized independently and sometimes to a cell count channel. If needed, normalizing across cell lines is left to the downstream meta-analysis.

Pin Grid Normalization Methods

Combining the raw pin grid values of a subarray has the undesired effect of ordering the intensity values about their pin means instead of their biological variation. This will lose the biological signal in the technical noise of the experiment. The choice of which pin grid normalization to use is based on the variations in the data. Pin grid variations are not seen in the Aberration and MEA datasets so pin grid normalizations are not performed.

It is appropriate to first perform a log transformation of the microarray raw intensity values to make their distribution more Gaussian. The median normalization process is to subtract each channel's logged pin grid median from the logged channel values of that grid. This centers all of the intensity values within a pin grid at 0 but does not normalize their variances.

The bivariate local regression method as implemented in the R package locfit(55) develops a smoothed model of the data using a proportion of the nearest neighbors. The goal is to match the size of the neighborhood used in the model to the size of the spatial variations on the microarray. The residuals from the model represent the difference between each data point and its local neighborhood. The residuals are used as the normalized data in the analysis.

A key question is how to determine an appropriate level of normalization in a real dataset where the hits are unknown. Most of the information to answer this questions lies in the EDA of the normalized data. The pseudo images of the normalized data should be free of regions of low or high intensity. The pin grid box plots should be centered about the plate median and have similar inter-quartile ranges (the height of the boxes which contain the 1st to 3rd quartiles). If there are positive controls, a histogram should show them as a cluster in one of the tails of the distribution.

Microarray Normalization

Microarray normalization is similar mathematically to pin grid normalization but it its performed at the microarray level. Choosing between median normalization and spatial normalization of a microarray depends on the variations in the dataset and on the layout of the treatments. The simplest case is when there are differences in the mean intensity levels between microarrays but no spatial variations. Normalizing each microarray by dividing all of its values by the median (or subtracting the log of the median if the intensities have been logged) will allow the values across all microarrays to be combined or compared.

If there are spatial variations within a microarray such as gradients or perimeter effects, good normalizations can be achieved by applying bivariate local regression at the microarray level.

However, this assumes the layout of treatments is random. If the treatments form patterns or if there are too many empty locations, bivariate local regression will include biological variation in its model and this will be undesirably removed from the normalized data. In these cases, median normalization will be performed.

Coefficient of Variation

Calculating the Coefficient of Variation (CV) among replicates is appealing as they provide quality values that are independent of positive and negative controls(56). However, the threshold for acceptable CV values is dependent on the data and the number of data points(44,57). In most dataset analyses there are minimum and maximum thresholds for determining valid values. This filtering eliminates spots from inclusion in the technical replicate calculations and reduces the reliability of associated CV values.

Figure 3 shows the 5 percent quantiles of CVs calculated on randomly selected sets of spots and the variations due to difference in the standard deviation of the underlying. The pipeline in this thesis includes the count of values in the CV calculations as a check on validity.

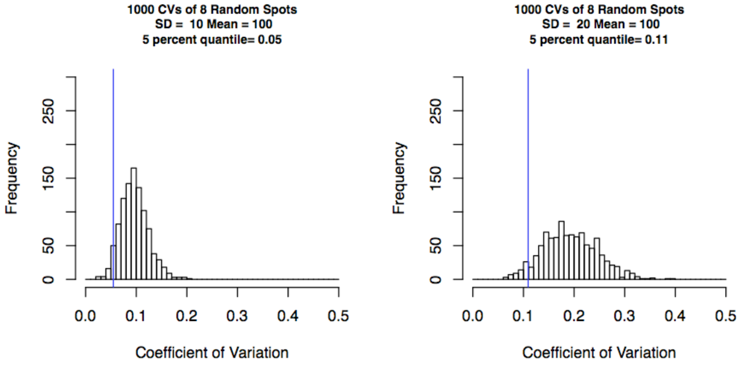


Figure 3 The blue vertical line represents the 5 percent quantile value in the distribution of 1000 CVS of randomly chosen and therefore uncorrelated spots. As the underlying standard deviation increases form 10 to 20, the 5 percent quantile moves from 0.05 to 0.11 making it difficult to determine a valid threshold.

Evaluation of Normalizations

Normalization is first evaluated with a repeat of the EDA methods used on the raw data. Comparisons of the raw and normalized pseudo images, distributions, pin grid boxplots and scatter plots will show if there are still technical variations. The shape of the distributions and the locations of the controls can inform if more normalizations are needed. The pin grid boxplots should have similar medians and interquartile ranges. The analysis can proceed with one or more boxplots different from the rest but conclusions based on that data should be conservative.

The effectiveness of the normalization is evaluated using the area under the curve (AUC) of a Receiver Operating Characteristic (ROC)(58,59) curve and the Positive Predictive Value for identifying the positive controls. As AUC values approach 1, the ability to discern between actual positives and false positives increases. In the simulated dataset, ROC analysis focuses on the created "hits" in addition to the positive. The normalization evaluation will be limited to the simulated dataset as the other datasets do not have positive controls or known hits.

Using a simulated dataset to evaluate the normalization requires setting values for the effect of the positive controls and hits on the phenotypes. These effect values are relative to the biological variations of the samples, which are the perturbations that do not change the phenotypes. When the signal of the positive controls and hits is large relative to the noise of the samples the normalization will be evaluated to perform better. Therefore, it is important that the effect values and biological variations in the simulated data are representative of the actual data and they are not changed during the evaluation.

Scoring

Well plate scoring methods are based on the performance of control treatments or on assumptions of near-normality in the responses of the samples. The sample-based methods have proven more robust and are the focus of this pipeline. Well plate HTS experiments typically assume a Gaussian distribution of the assay-wide intensities and then select the treatments based on their z-scores(9). An alternative to z-scores is to use rank products(60). This method uses the product of the rank of the normalized conditions across replicates, channels and even cell lines. The gamma distribution models the distribution of rank products as previously described. (61) This makes it computationally trivial to estimate the probabilities of a given rank product. Comparisons favor rank products over z-scores(62) in datasets with small numbers of replicates and non-Gaussian distributions. Since cellular microarrays typically exhibit distributions that are not normal and have few replicates, rank product analysis is an appropriate scoring method. The general nature of rank product analysis makes it useful when combining results across channels, drugs and cell lines.

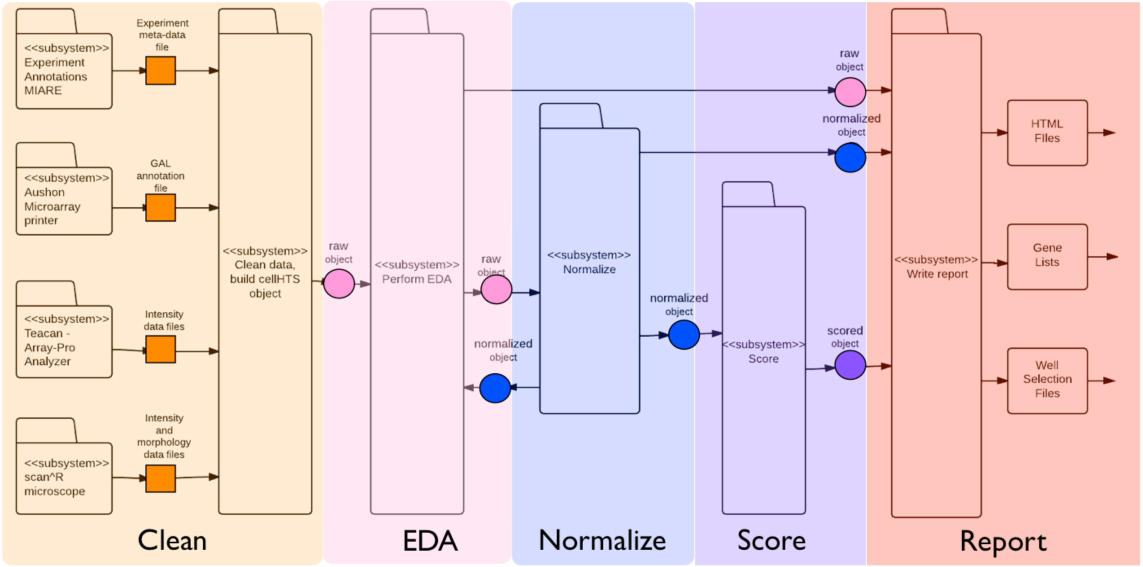
Screen-level scoring results in one or more hit lists of perturbations that match biological questions such as which conditions drive cells into apoptosis or which perturbations cause cells to differentiate. Microarray experiments with multiple biological end points have evidence for multiple hit lists. After normalization and scoring, hits can be called by taking a percentage of the top scoring perturbations(7) or establishing a threshold for hits based on the values of the statistically-derived scores.

Reproducibility

To meet the goals of reproducible research, this pipeline stores the results, the analysis script that created them and a Minimum Information About a Microarray Experiment (MIAME) file(63) that describes the experiment. In order to have the broadest impact, the pipeline developed in this thesis and the Aberration and MEA data published at GitHub repository markdane/CMA2 with DOI: 10.5281/zenodo.10145.

Results

In order to focus on the computational pipeline shown in Figure 4 instead of the datasets, each module of the pipeline will be applied first to the simulated dataset and then to the Aberration and MEA datasets.



<<subsystem>> Experiment Annotations MIARE
Experiment meta-data file
<<subsystem>> Aushon Microarray printer
GAL annotation file
<<subsystem>> Teacan - Array-Pro Analyzer
Intensity data files
<<subsystem>> scan^R microscope
Intensity and morphology data files
<<subsystem>> Clean data, build cellHTS object
raw object
<<subsystem>> Perform EDA
raw object
normalized object
<<subsystem>> Normalize
normalized object
<<subsystem>> Score
scored object
raw object
normalized object
<<subsystem>> Write report
HTML FIles
Gene Lists
Well Selection Files
Clean
EDA
Normalize
Score
Report

Figure 4 Computational pipeline as developed for this thesis. Data in files are solid squares. Data in software objects are in circles. The Clean section gathers the data and metadata from the printer, imaging system and the experimenter and combines them into a raw software object. The EDA displays pseudo images, box plots, scatterplots, histograms and qq-plots of the raw or normalized data. Normalization converts objects of raw data into objects with normalized data. Scoring operates on normalized data to create an object with scores. The report section takes all three types of objects to create HTML, files, plots and tab delimited files.

EDA

The first step in the pipeline is to assemble the data, annotations and the experiment's meta-data into a cellHTS2 object.(20) This is an S4 object in the programming language R(38). The object will be the software organizing entity for the rest of the pipeline, ensuring that data, annotations and processing states are correctly aligned and updated. As soon as the object is built, we can begin the EDA on the raw data. Much of this is performed by the cellHTS2 writeReport function generating html pages viewable with a web browser. writeReport also creates pdf, png and text files that can be accessed directly.

Pseudo Images of Values - Simulated Raw

Figure 5 shows pseudo images of the raw values in the Simulated dataset. The colors in the pseudo images are auto-scaled to the 5 and 95 percentile of the spot values. The color-to-value levels are set for each microarray so that the location and relative amounts of variation can be seen within each microarray. This auto-scaling can be misleading as it does not show the magnitudes of the variation, however these can be seen by cross-referencing the histograms and pin grid boxplots.

The two microarrays in this simulated data set are replicates made by adding random noise to a base microarray of values. This mimics biological and technical variations in the replicates. Visible in the images are the 48 pin grids each comprised of a 9 x10 array of spots. Replicate 1 also has two ellipsoidal and two linear perturbations that are visible in the pseudo images. The center perturbation doubles the intensities at its center with an ellipsoidal roll off towards its edges. The lower left perturbation causes values in the corner to be near 0 and then increase ellipsoidally to normal. The two linear perturbations both double the base intensities. These variations mimic staining and manufacturing artifacts.

Randomly distributed about the plate are hits that reduce the values by 40%. These represent the targets for the screen.

A last item to note in these images is the repeated positions of the three positive controls in each pin grid. The positive controls lower the channel intensities by an average of 70% as shown by the darker blue circles.

The histograms in the middle of Figure 5 show the distributions of all values. The rug below the histogram is colored red for positive controls, blue for negative controls, purple for hits and black for the rest of the samples. The Spatial Variation Index for replicate 1 is 0.227 and .093 for replicate 2.

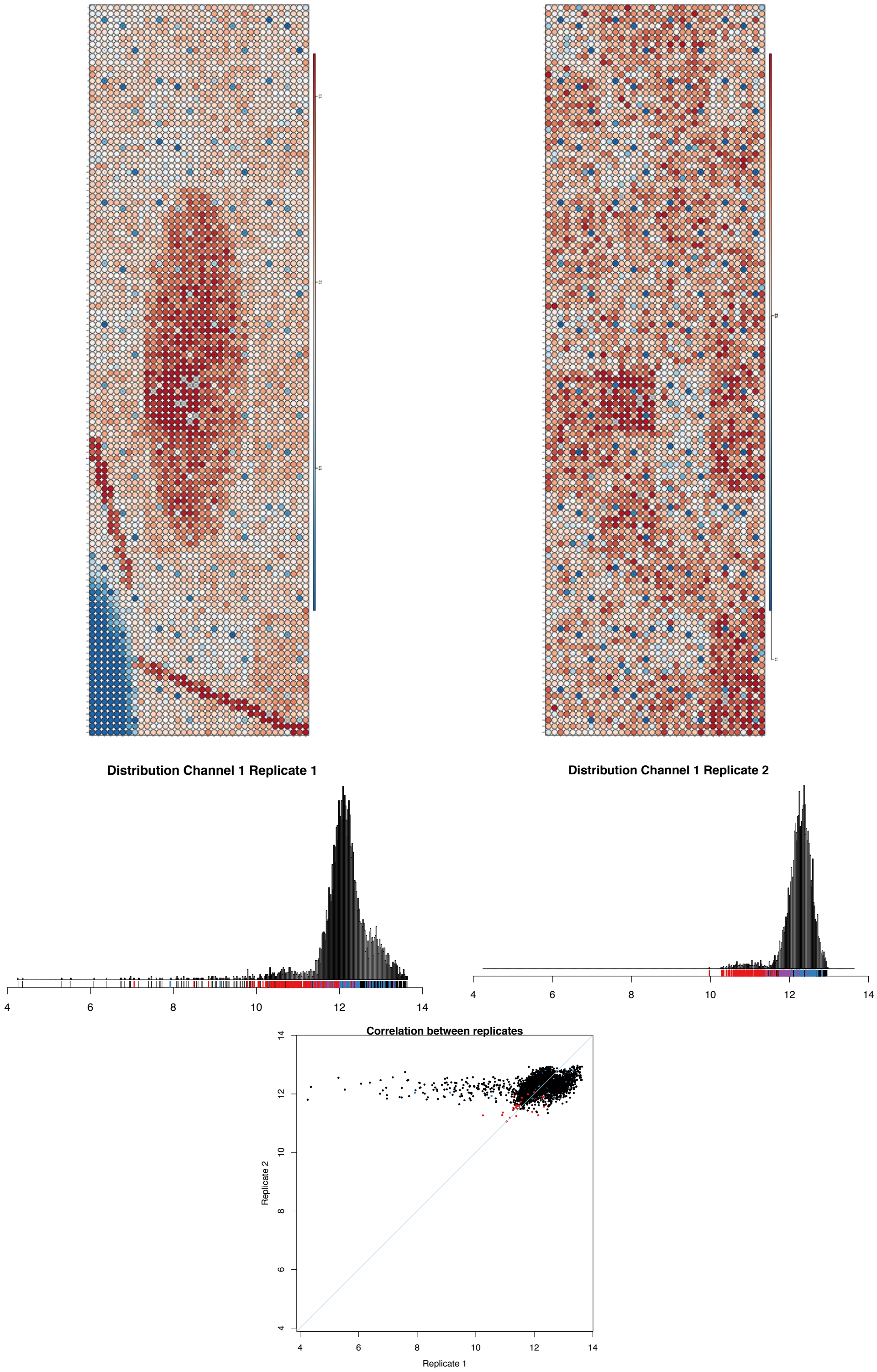


Figure 5 Pseudo images, histograms and a replicate scatter plot of the raw values of replicates 1 and 2 in the Simulated dataset

One of the first steps in the EDA is reviewing the pseudo-images for spatial and pin grid variations while using the histograms to quantify the magnitudes and frequencies. We are looking for issues that can inform the normalization and analysis approaches. Another early step is to locate the controls within the distribution. We expect to see the positive controls within one of the tails and the negative controls spread throughout the body of the distribution. After normalization we will perform a more rigorous controls assessment that includes quantify the microarray quality.

The effects of the spatial perturbations in Replicate 1 can be seen in the right skew of its distribution. Since these perturbations span pin grids, this microarray will benefit from a global, spatially-aware normalization. Visible in replicate 2, there are also obvious pin grid variations with the pin in the 7th row and 2nd column (pin grid 26) having noticeably higher values. These will be quantified using boxplots later in the EDA.

Figure 5, lower panel shows the correlation of all spots in the replicate microarrays. Replicates with high correlation form a linear group with a slope close to one. Perturbations in Replicate 1 cause data points to fall to the left and right of the main linear group and result in a Spearman Rank Correlation value of 0.44. A shortcoming of Spearman Rank Correlation is that it includes all of the data values while the focus of the screen is on the values in the tails of the distribution. This scatterplot confirms the presence of perturbations in one of the replicates that are not in the other but it is not informative about whether they can be removed during normalization.

Pseudo Images and Distributions of Values - Aberration Raw

The Aberration dataset has 4 channels of intensities and the raw pseudo images in the upper panels of Figure 6 show spatial variations in three of them. The Spatial Variations Indices of the channels 1-4 are 0.03, 0.05, 0.06 and 0.02 respectively which correspond to small amounts of spatial variation. The auto-scaling of the pseudo images makes these small amounts of spatial variation visible. This dataset is a good example of the need to use the histograms in the Figure 6 lower panels to determine the magnitude of the variations along with the pseudo images to show their locations. The layout of the siRNAs in this experiment is randomized so these variations are likely due to technical effects in the staining or imaging processes. There are not distinct pin grid differences although these may be masked by the color scaling that is driven by the global variations. The values have been log transformed and the distributions are mostly Gaussian.

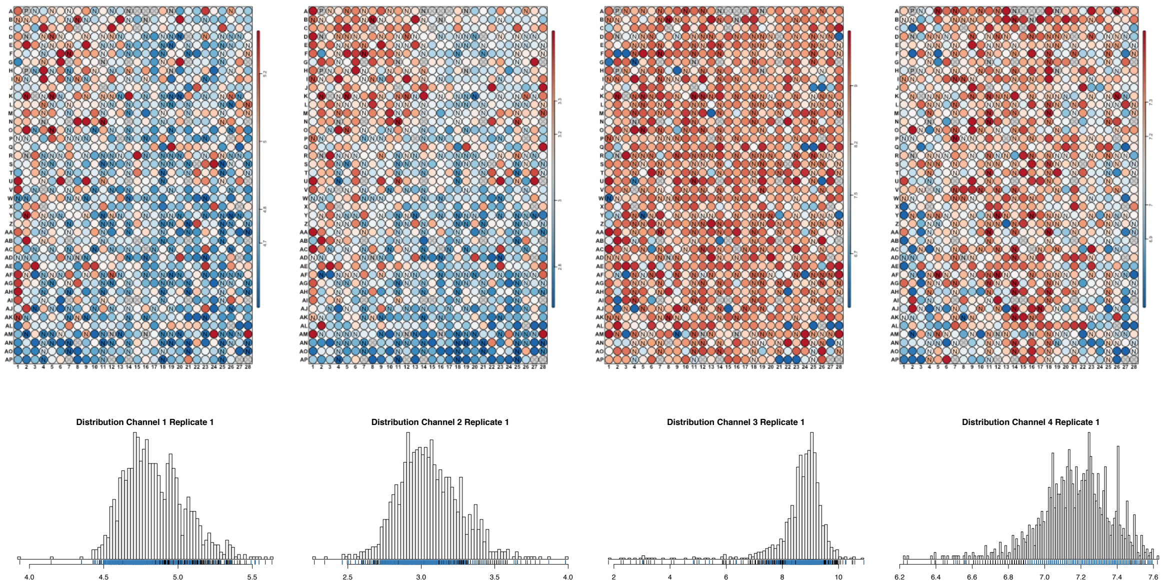


Figure 6 Pseudo images and histograms of the log transformed raw values of the 4 channels in the Aberration dataset

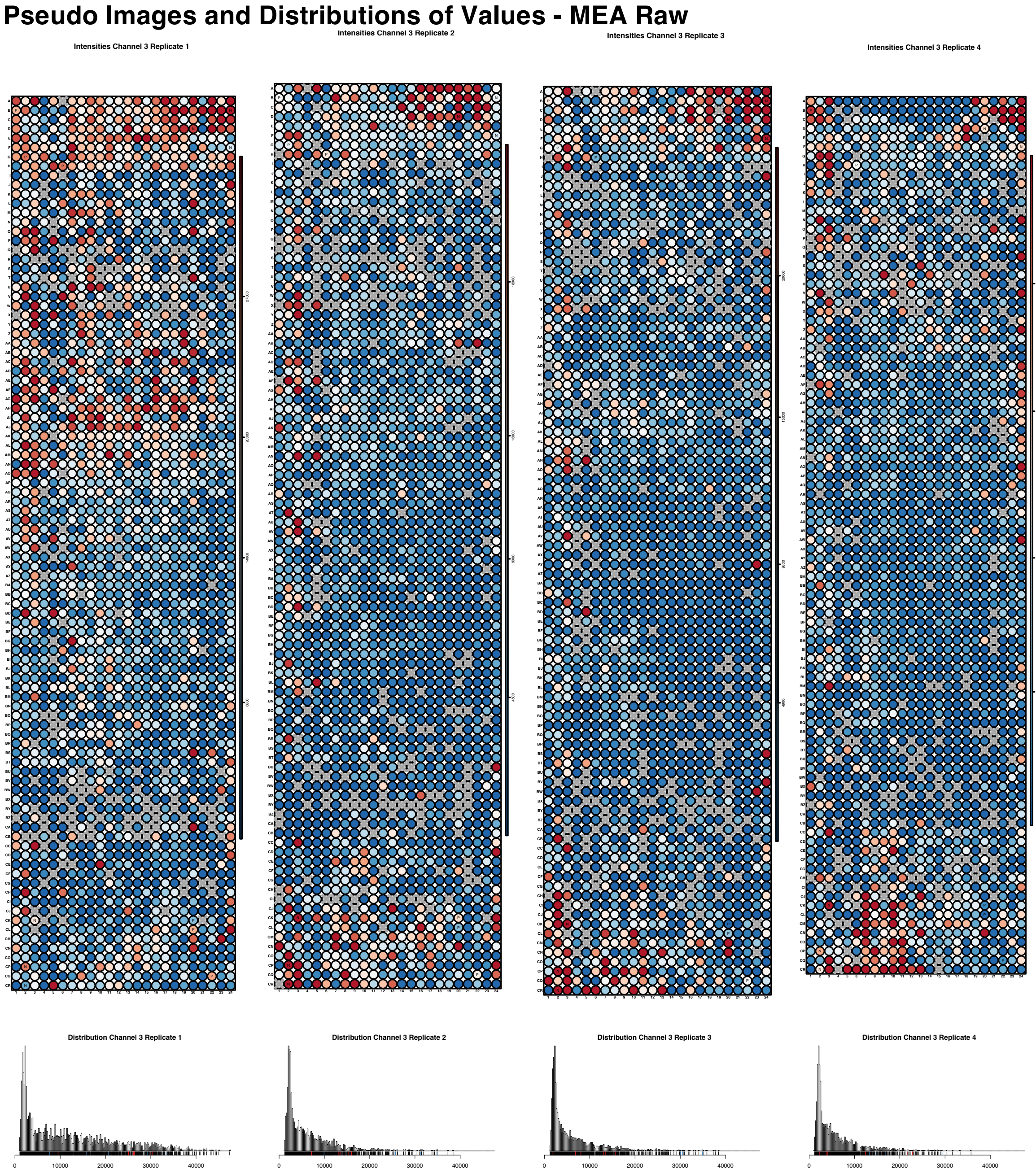


Figure 7 The raw values from the four MEA replicates all show similar skewed right distributions. Log transforming these values will make them easier to interpret.

The MEA dataset shown in Figure 7 has one channel of data from four replicate microarrays. The biomarker in the channel is Cellmask^{TM} from Thermo Fisher Scientific Inc. which is designed to stain intact plasma membranes and provide a measure of cell counts. The gray spots have no treatments and their uneven distribution will have an impact on spatial normalizations. The pseudo images in Figure 7 show spatial variations with similar patterns across the replicates. The Spatial Variation Indices for replicates 1 through 4 are 0.25, 0.14, 0.13 and 0.12 respectively. These variations may be from a combination of the structure in the layout and spatial variations such as staining or imaging.

The distributions in Figure 7 are heavily right skewed starting at 0 and the analysis and interpretation will benefit from being log transformed.

Pseudo Images for Layout - MEA

The perturbations in the MEA dataset are the 384 possible pairings of 16 ECM molecules with 24 growth factors. We can proceed by assigning each ECM molecule one of 16 contrasting colors, then creating an image of the microarray layout. The same process can be applied to the growth factors, creating two different maps of the same microarray. As seen in Figure 8, the perturbation locations are structured with the ECM molecules in columns and the growth factors in rows. There is randomization within the structure, for instance each 8 x 6 pin grid has four randomly located growth factors. However, identical pin grids are created in a symmetrical pattern, confounding the position of the perturbations with their type. This violates an assumption of spatial normalization that could correct for spatial variations within the microarray.

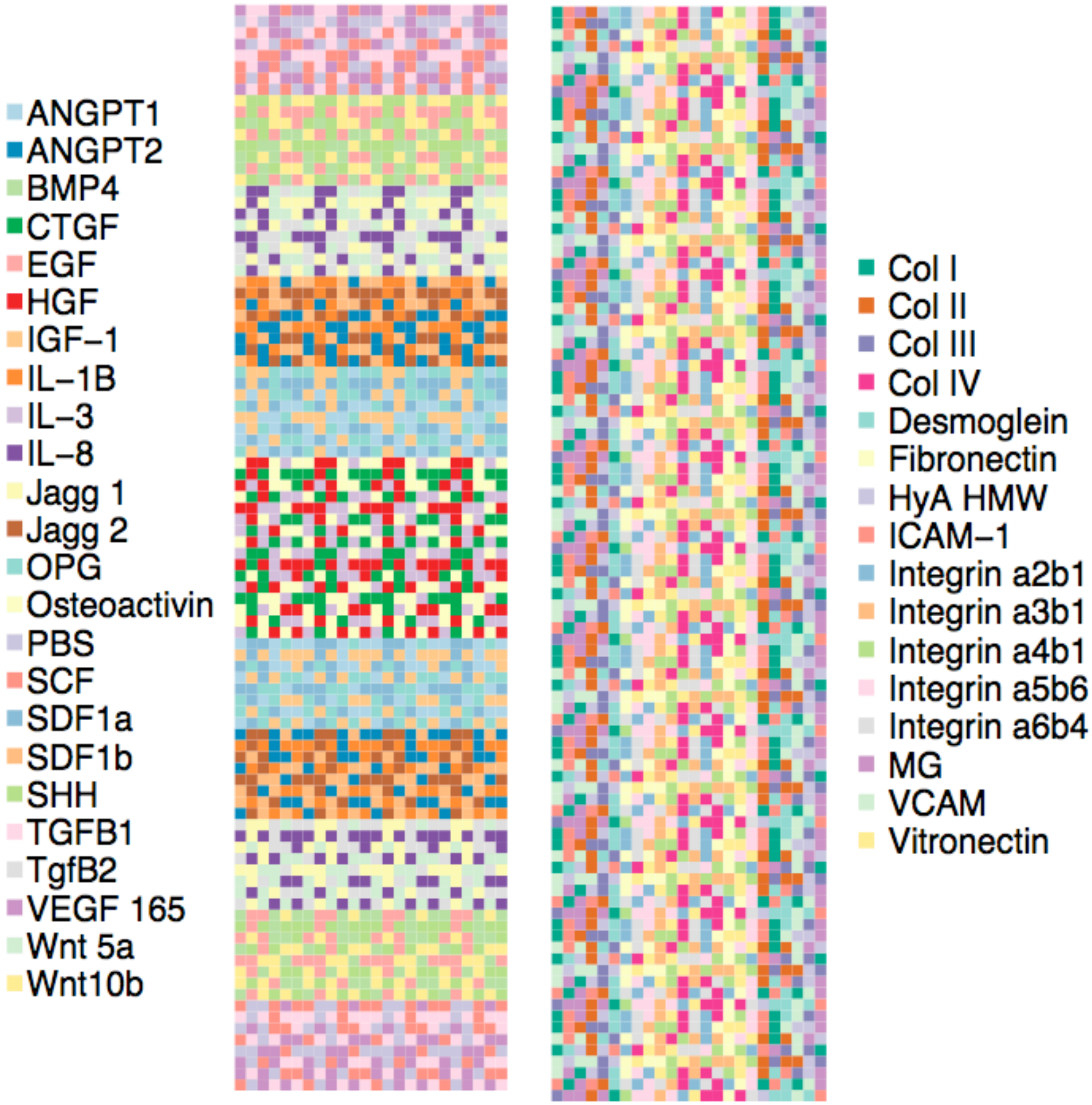


Figure 8 Pseudo Image of MEA layout. Each colored grid represents the growth factors (left) or ECM proteins (right) that are combined at each spot. The symmetrical patterns of the rows and columns show that the layout of treatments is not random.

Pseudo Images to locate Actual Images

It is often informative to view the images underlying the quantitated data. While the image of an entire scanned microarray can be seen with a simple image viewer, we often want to see specific subgroups of spots. This can be done by making a pseudo image of the layout that has the subgroup transparent and then overlaying the layout on top of the whole microarray image.

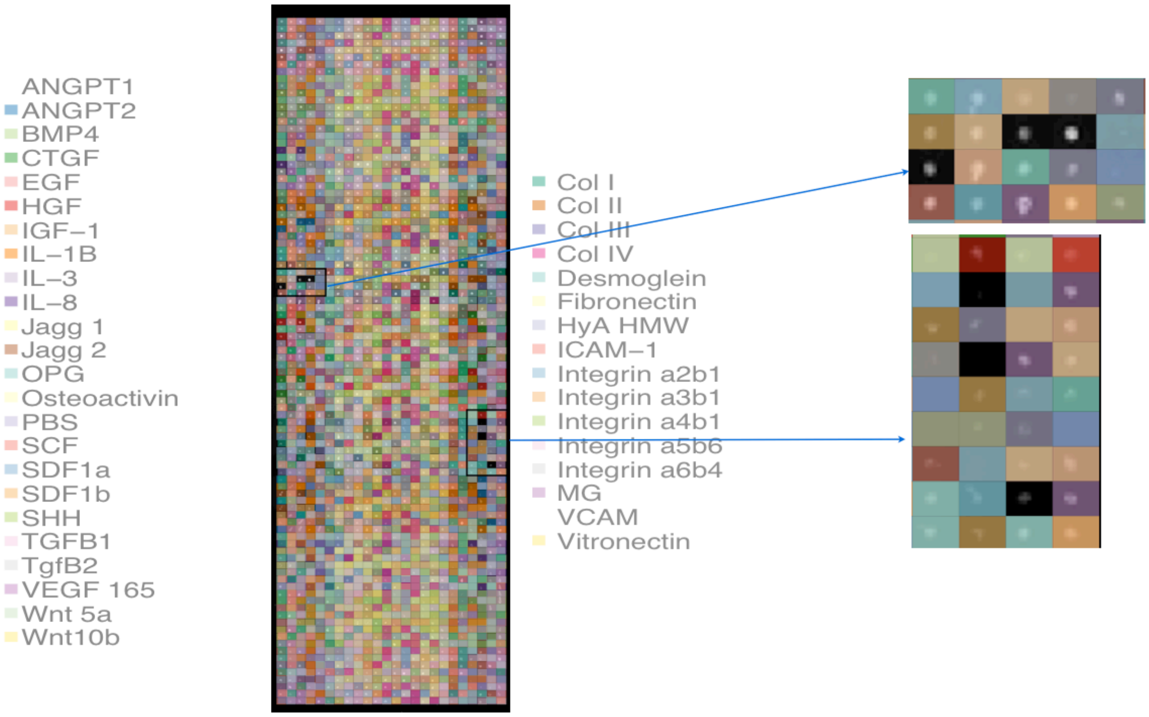


Figure 9 Pseudo images to locate vision images – The combined ECM and growth factor colored grids are overlaid on the tiff image of the microarray. The machine vision images of selected ECMs and growth factors are readily located through transparent windows.

During the EDA of the MEA dataset the colors for VCAM and ANGPT1 were made transparent in order to locate the images behind data that had high CV values. The two right-side panels in Figure 9 show the six VCAM+ANGPT1 images. The upper panel shows a neighborhood of bright spots which result in high signals while the lower panel shows a neighborhood of dim spots which result in low signals. These images provide evidence that locations on the microarray are driving the signal in addition to the biology of the spots.

Pin Grid Level EDA – Simulated Raw

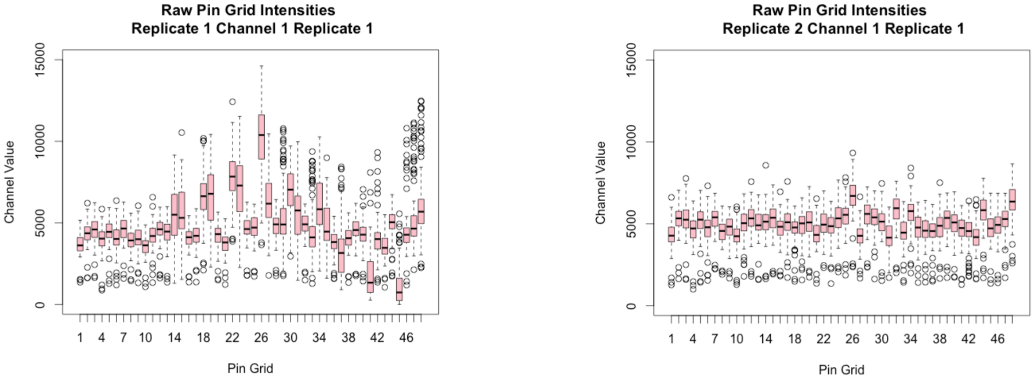


Figure 10 Boxplots of raw values from the Simulated data set.

The boxplots of the Simulated dataset raw values grouped by pin grids in Figure 10 show both pin grid and spatial variations. The pin grid variations show up as singular boxplots that differ from their direct neighbors such as pin grid 26 in Replicate 1. Global variations affect contiguous regions and can be seen in sequential boxplots or in effects that repeat with an index of four. The index of four stems from the four columns of pins in the spotting pinhead. Numerous examples of spatial variations in Replicate 1 can be seen in the boxplots that are contiguous and those that are indexed by four.

Much of the downstream processing yields better results and is easier to visualize when the data is log transformed. As an example, Figure 11 shows log transforms of the Simulated raw data. The relative magnitudes of the spatial variations in the lower left corner of Replicate 1 are now more apparent.

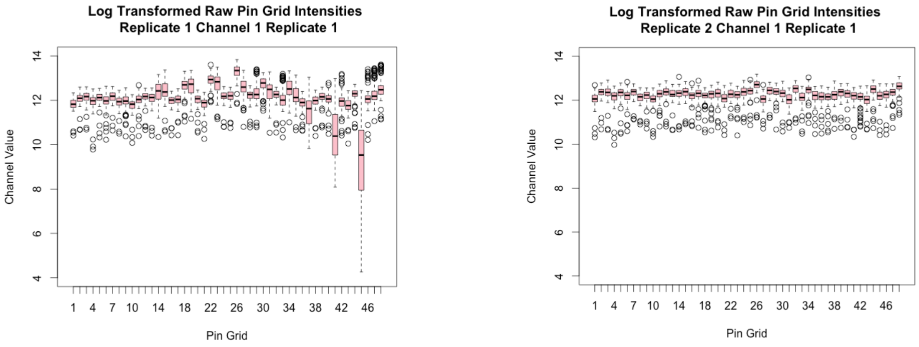


Figure 11 Box plots of the log transformed raw values from the Simulated data set.

Pin Grid Level EDA – Aberration Raw

Figure 12 shows boxplots of the raw channel values of the Aberration dataset. Inspecting these plots for microarray variations shows decreasing values in channels 1 and 2 from the top of the microarray to the bottom. A possible source for this is uneven staining. It could also be a time-based effect due to lengthy measuring times. Global spatial effects are apparent in the up and down wave pattern seen in all channels. These are from values in the center columns being different than those on the edges. These boxplots show the location and magnitude of the variations but do not directly determine its source. Pin grids 17 and 24 are consistent examples of higher variance across all channels.

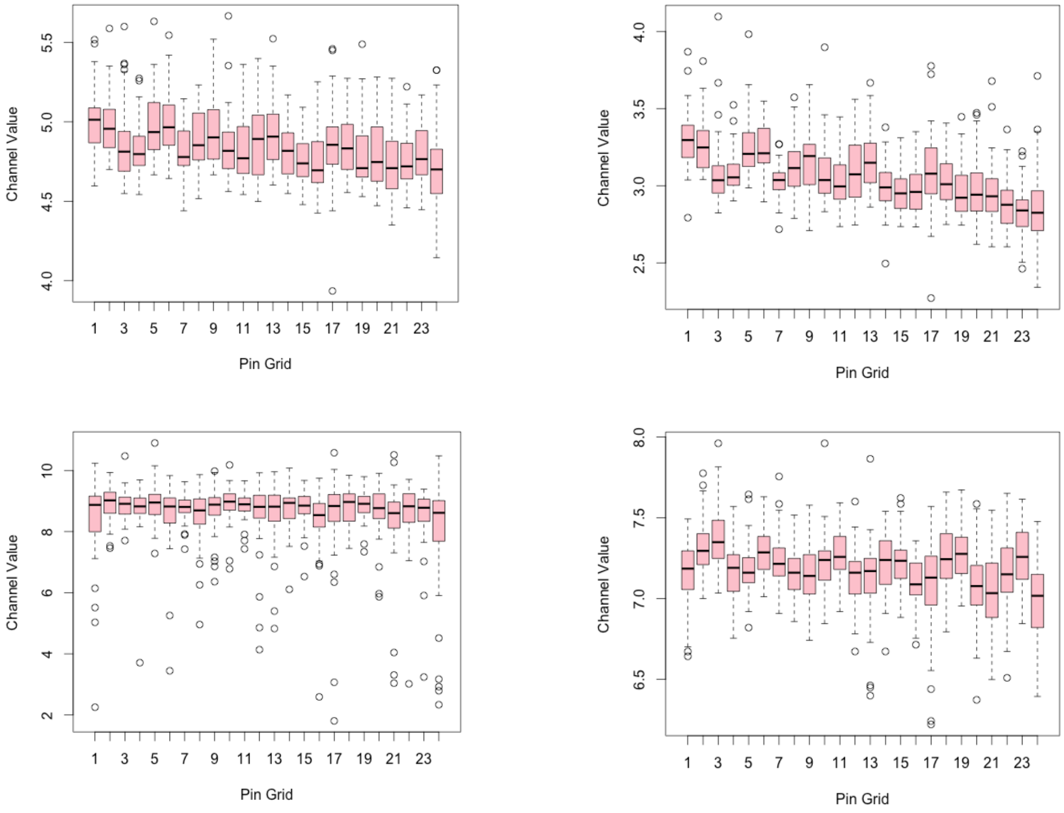


Figure 12 Log transformed raw values of the four channels of Aberration. These boxplots show global variations in the data such as the general decrease in values of channel 1, upper left and channel 2, upper right. Also seen is the pattern of the second and third boxplots being the highest in a row of four in channel 4, lower right. None of the pin grids appear as outliers.

Pin Grid Level EDA – MEA

The boxplots in Figure 13 show the high amount of variance in the four MEA replicate microarrays. There are patterns common across the replicates such as the high values in the first grids, a general trend down and then an upswing in the last grids. There are variations correlated across the rows that cause similarities in groups of four grids. There are also singular grids that differ from their neighbors and from the other replicates.

The EDA of the MEA data is showing significant spatial variations. Unfortunately, the layout of the biological treatments is not random and the variations can be due to the biological treatments, their locations on the microarray or both factors.

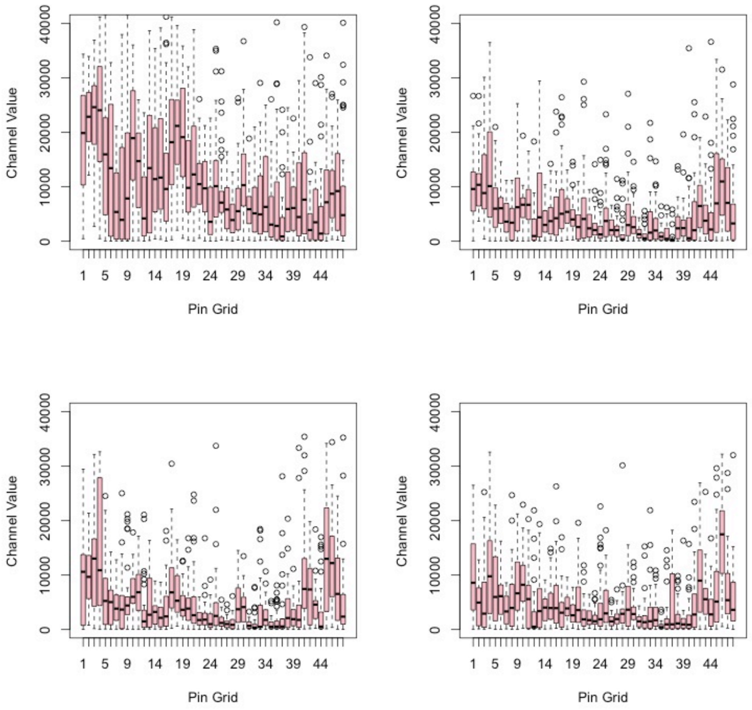


Figure 13 Boxplots of MEA pin grids. These boxplots show global and singular pin grid variations in the raw data of the MEA replicates. The along the microarray row variations are seen in the repetitive patterns of groups of four pin grids. High variation is seen in the first 21 pin grids of replicate 1 (upper left) and in pin grid 4 of replicate 3 (bottom left).

Microarray Normalization and EDA - Simulated

The next process in the pipeline is to normalize the data to remove technical artifacts and enable combining data from pin grids, microarrays, channels and experiments. The EDA of the simulated data showed both pin grid and global spatial effects. Since normalizations can be performed at either or both levels, we will start at the microarray level using a spatially aware normalization. Figure 14 shows the pseudo images and distributions after normalizing with a bivariate local regression. The ellipsoidal spatial variations are gone and only a remnant of the linear perturbation in the lower part of the microarray is visible.

The distributions in Figure 14 show a promising scenario with the positive controls far to the left and most of the hits between the controls and the majority of samples. The box plots in Figure 15 show similar medians and variations in all pin grids even though there was no pin grid specific normalization.

The scatterplot in Figure 16 still shows variation between the replicates as seen by the spread of the data points orthogonal to the diagonal. A Spearman rank correlation value of 0.48 has been calculated for these replicates. Scatterplots and correlation values are not very informative of the quality of screens(64) as they are driven by the large majority of genes that do not cause significant phenotypic differences. We are interested in the small number of perturbations found in the tails of the distributions that cause significant phenotypic differences.

At this point in the pipeline we can evaluate the controls by inspecting their distributions and calculating the Z’-factor as shown in the lower panels in Figure 16. Replicate 1 has a Z’-factor of -0.15 and Replicate 2 is -0.08 Well plate experiments look to have Z’-factors greater than 0. There is not an established rating for Z’-factors for cellular microarrays.

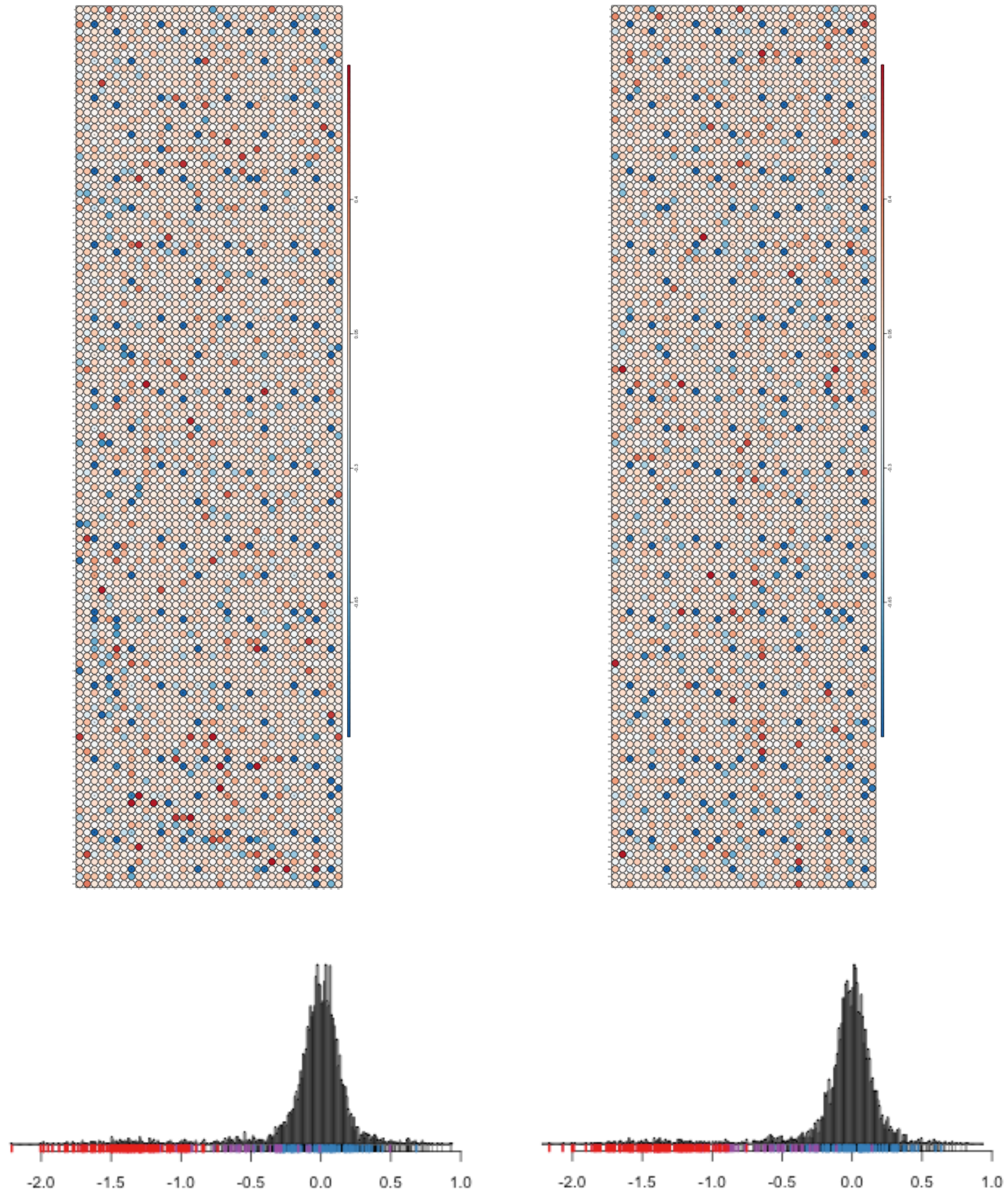


Figure 14 After spatial normalization, the log transformed Simulated values are distributed with most of the positive controls (red) at the lower end and most of the hits (purple) between the positive controls and the samples.

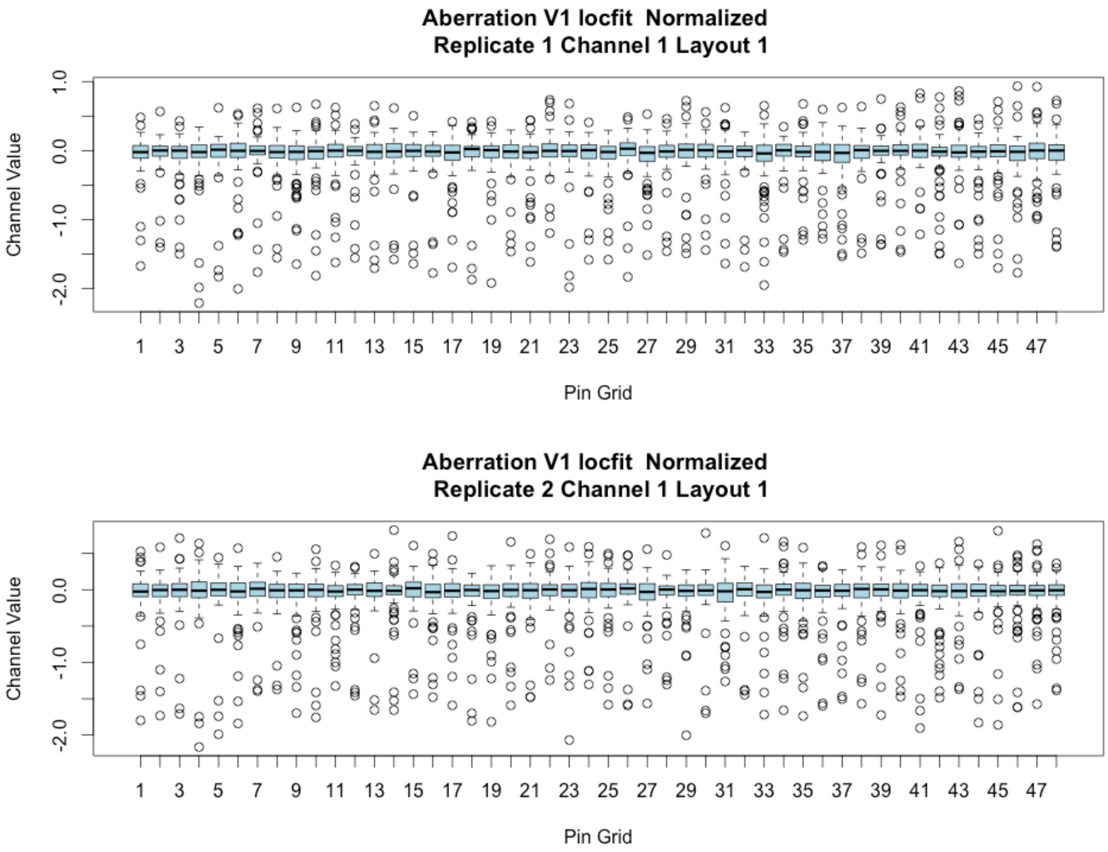
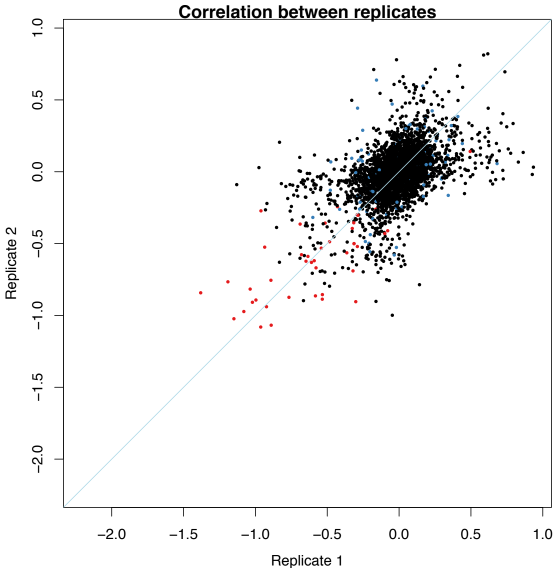


Figure 15 Spatially normalized Simulated data. Using a bivariate local regression to normalize the values centers each pin grid close to 0 and results in similar variations across the microarray.



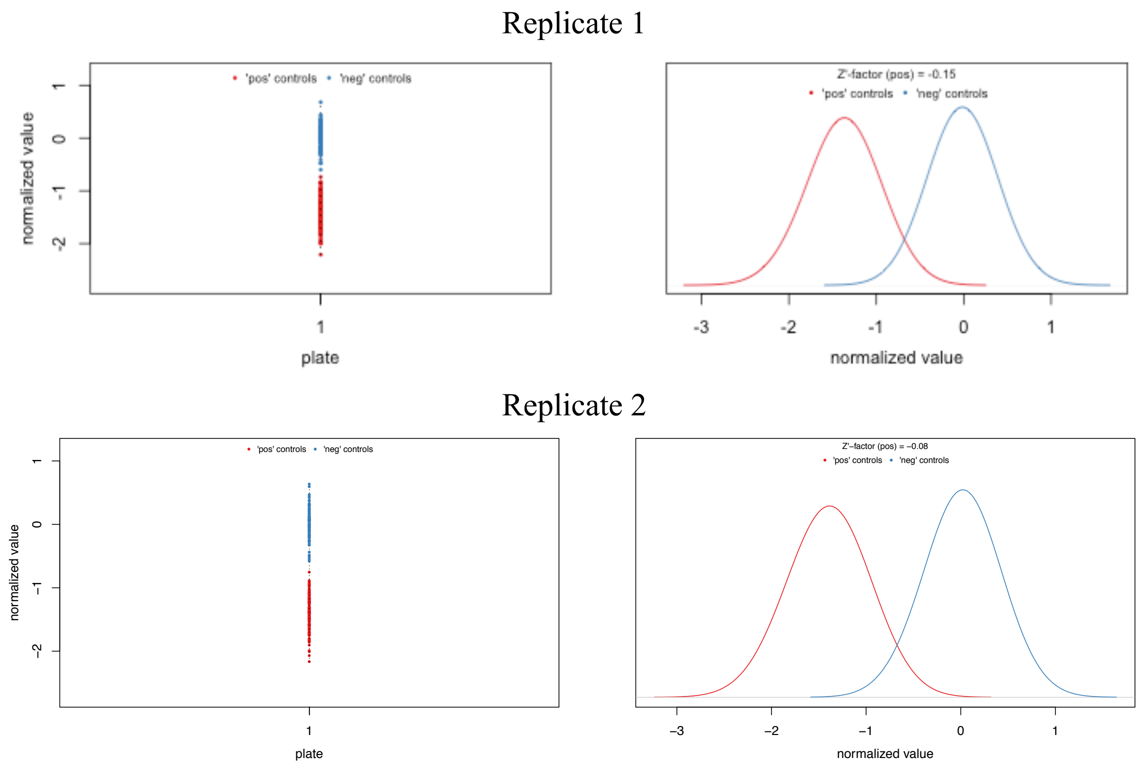


Figure 16 Scatter plot and Controls plots for the two normalized replicate microarrays in the Simulated dataset. After the spatial normalization the scatterplot still shows variation at the tails of the distribution. This is an indicator that there are going to be false positives in the final hit lists. The lower figures analyze the controls, showing there is overlap which results in Z'factors of -0.15 on the highly perturbed replicate 1 and -.08 on replicate 2.

Evaluation of Normalization - Simulated

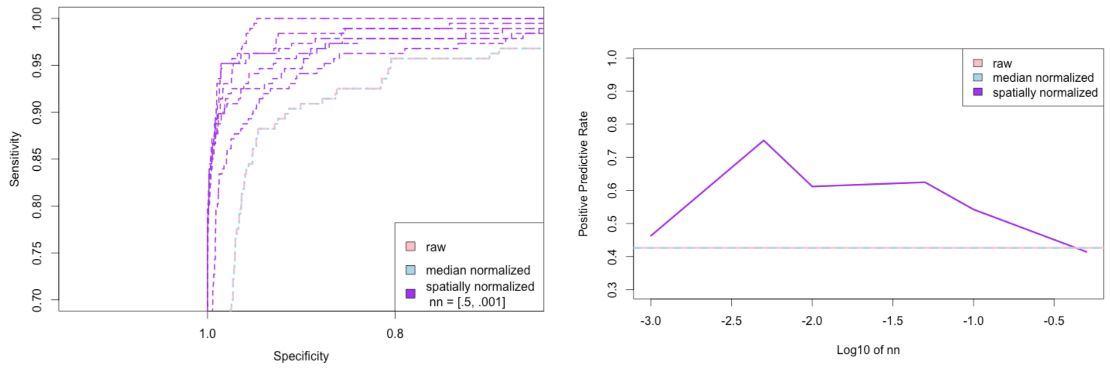


Figure 17 Using Positive Predictive Value (PPV) to evaluate the number of neighbors (nn) parameter in the bivariate local regression. The left panel is partial ROC curves of raw, median normalized and spatially normalized intensities with nn values between 0.001 and 0.5. The right panel shows the highest PPV at an nn of 0.005 [$\log_{10}$(.005)= -2.3].

The spatial normalization fits a bivariate local regression to each spot using its neighborhood of spots. The weights of the spots in the neighborhood are based on a parameter nn, the number of neighbors. Smaller values of nn reduce the size of the neighborhood. Conceptually, we would like a neighborhood that captures the technical variations but has little effect on the isolated spots that represent biological variation. Figure 17 shows the results from using ROC analysis to pick a threshold and then calculating the resulting Positive Predictive Value defined as True Positives/(True Positives + False Positives). The value of nn=0.005 is selected as the value to normalize and score the Simulated dataset. This type of parameter selection requires positive controls so it cannot be performed on either of the experimental datasets.

Microarray Normalization and EDA – Aberration

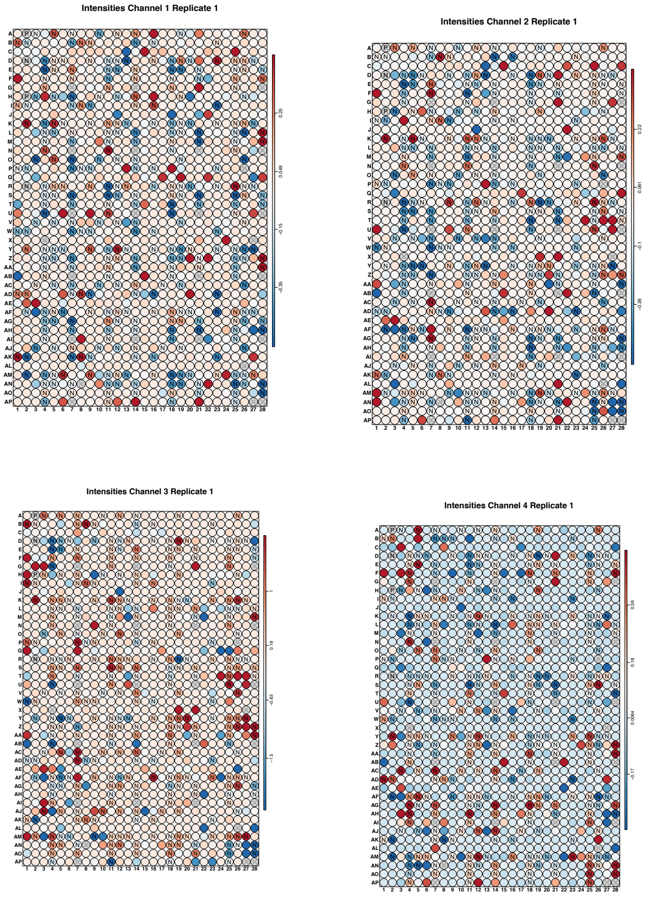


Figure 18 Pseudo images of the spatially normalized Aberration channel data

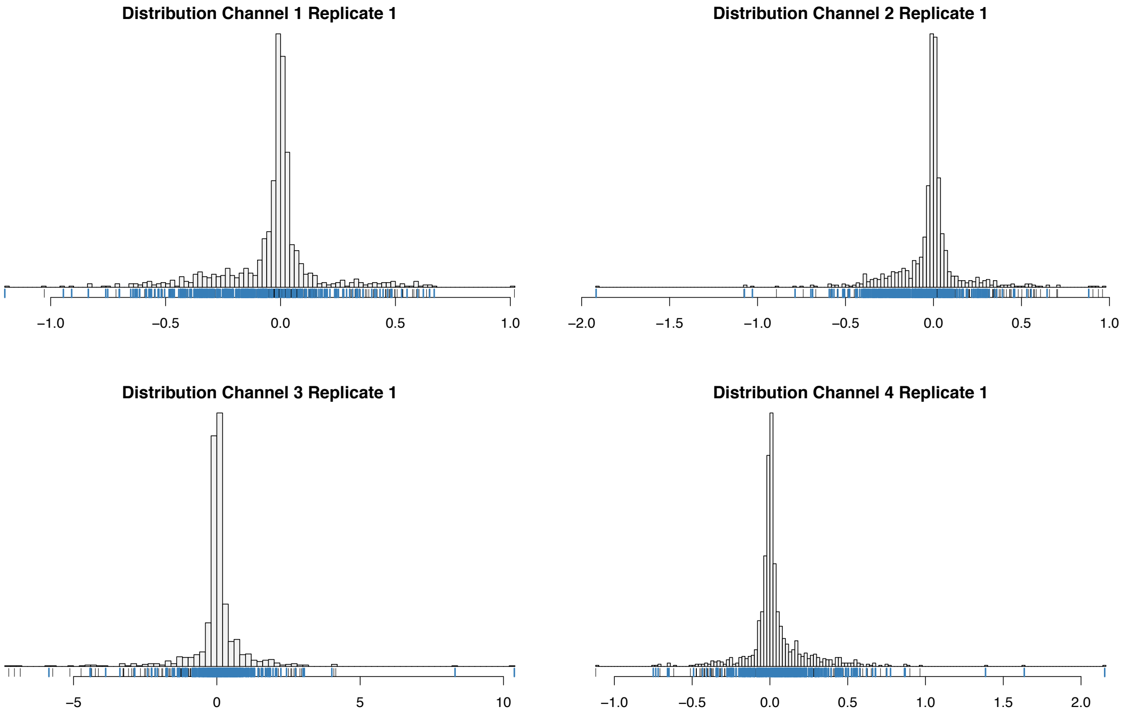


Figure 19 Distributions of the log transformed and spatially normalized Aberration channels

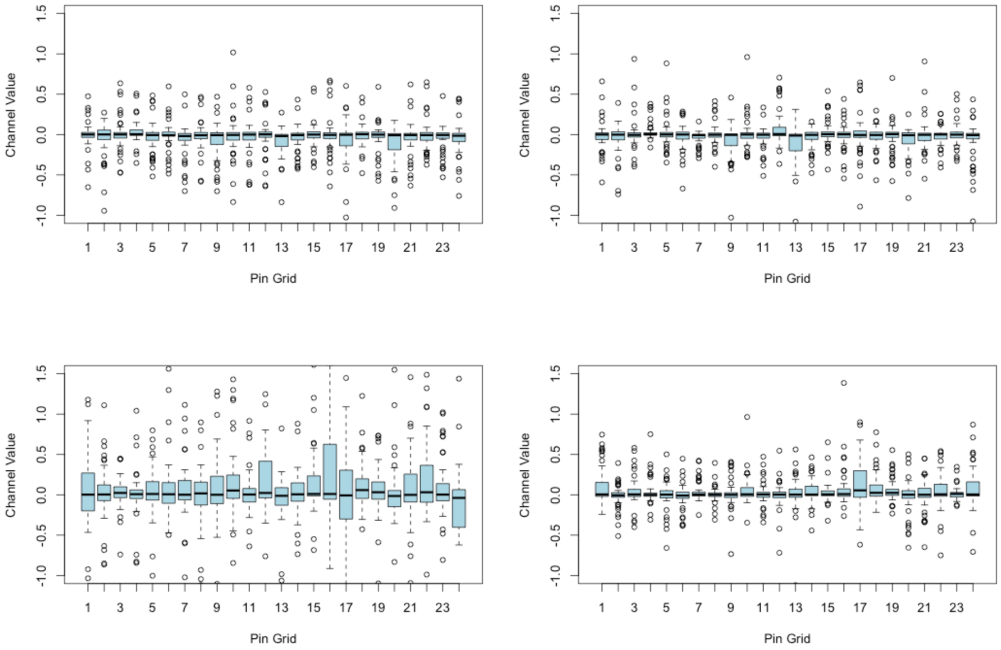


Figure 20 Normalized Aberration dataset pin grid boxplots. All four channels show mostly normalized data with just a few high variance pins such as channel 3, pins 16 and 17 (lower left panel) and channel 4, pin 17 (lower right panel).

The spatial normalization of the Aberration dataset has removed most of the spatial variations as seen in Figures 18-20. This is appropriate as the layout is random. If this experiment had replicates, these could be used to filter the biological variations from the technical ones

Microarray Normalization and EDA - MEA

The EDA of the MEA dataset revealed a need for spatial normalization but a layout that has the perturbations confounded with the locations. Applying a spatial normalization will distort the biological variations. Therefore, a median normalization is applied to each microarray so the replicates can be combined. The raw values are also log transformed in order to compress their scale and make the results easier to visualize.

The median normalization results in distributions and boxplots as seen in Figures 21 and 22 but does not change the appearance of the pseudo images shown in Figure 7.

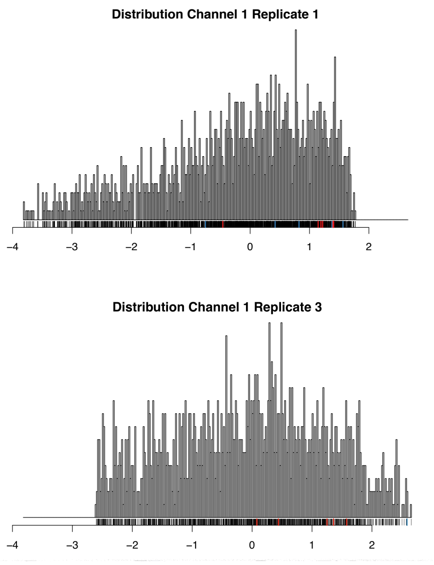


Figure 21 Distributions of the median normalized and log transformed MEA dataset.

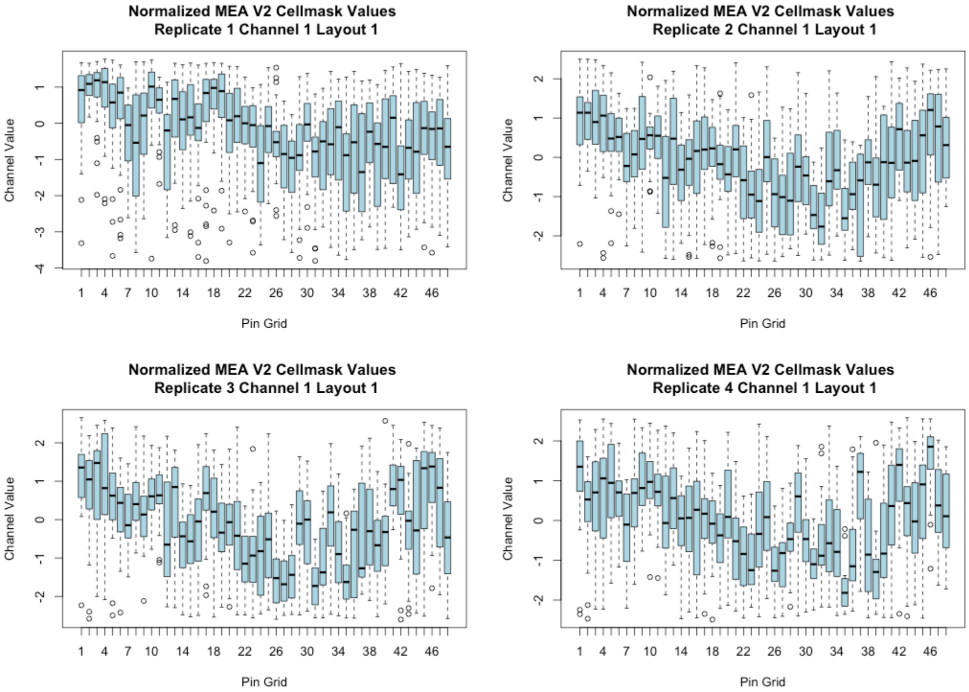


Figure 22 Boxplots of the log transformed and median normalized MEA data have the same pattern as the raw values but are now centered about 0.

Scoring

Z scores and rank product scores for the Simulated dataset are compared using the PPVs. The left panel of Figure 23 shows how rank products are more robust to non-normal distributions. The nn value of 1 does no spatial normalization. With no spatial normalization, rank products return higher PPVs than Z scores until the list size increases to 350. As the spatial normalization is increased by lowering the nn value, the data becomes more normal and the two scoring methods converge. At the theoretical optimal value for normalization of nn=0.005, z scores outperform rank products as seen in the right panel of Figure 23. This may be due to the slight estimation error associated with modeling the rank products with the gamma distribution.

This analysis of the simulated dataset uses the advantage of knowing which samples are hits. Actual datasets will at best have positive controls to evaluate the normalization. This is a reason to favor rank products over z scores enabling the use of higher, more conservative nn values.

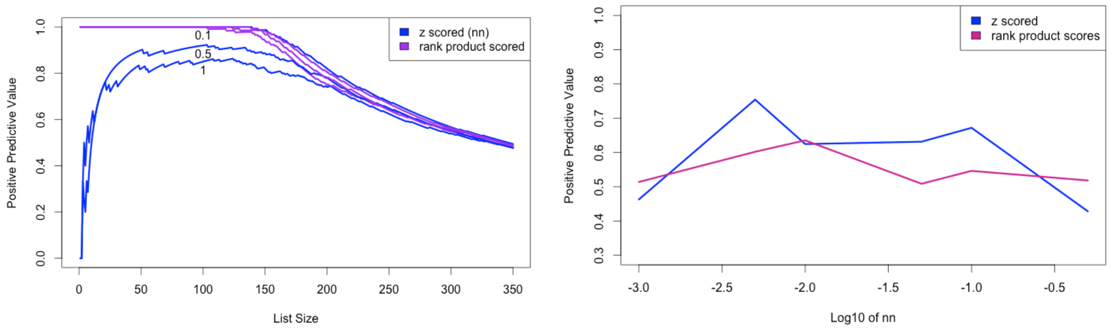


Figure 23 Positive Predictive Values of rank products and Z scores. The left panel shows PPVs form rank products remain stable as the amount of normalization varies. The right panel shows Z scores outperform rank products on the spatially normalized data.

We can score the Aberration dataset by ranking each channel according to its intensities but the experiment does not have positive controls nor replicates. This limits the ability to directly quantify the quality of the hit list. If the dataset is to be used for downstream experiments, the channels can be spatially normalized and a threshold set based on the number of genes that can be efficiently evaluated in the next experiment.

Discussion

As stated in LaBarge *et al*.(8) concerning MEArrays "Currently the major crux of this technology is inadequate methods of analysis." To begin to address this, this pipeline provides a modular solution for data cleaning, EDA, normalization and scoring cellular microarrays.

A simulated dataset, an siRNA dataset and a microenvironment dataset were used to demonstrate the implications of layout randomization, replicates and controls on the ability to normalize and develop hit lists with high positive predictive values. As emerging technologies, these microarrays benefit from extensive EDA as demonstrated with the use of pseudo images, boxplots, histograms and scatterplots before and after normalization. Normalizations that match the technical variations and rank product scoring that is robust to non-Gaussian distributions are key contributors to reducing false positives as shown by ROC curve analysis.

This pipeline has identified problems in layout randomizations, cell adherence, staining, image processing, dye-to-filter mismatches and experiment design. This had led to use of the QA/QC and EDA to inform the ongoing experiment designs to determine the number of replicates and setting expectations of the PPV of the hit lists based on data quality etc.

We note that this pipeline complements existing methods and pipelines. Existing computational methods for the MicroScale microarray from Wood *et al*. do not address layout randomization. They use a form of B score normalization that normalizes to the median of the 6 nearest row and column spots. This type of normalization is appropriate for well plates where there are row and column effects due to multichannel pipettes but does not match the pin grid and global spatial variations found in cell based microarrays. The computational pipeline uses bivariate local regression and pin grid normalizations to match the global and pin tip variations, respectively.

LaBarge *et al*.(8) describes using a Dunnett's t test to compare the ratios of lineage markers in treated spots to control spots and a visualization of the resulting p-values. Mann-Whitney tests are also used to compare the distributions of cell level data. Randomization of the layout, use of replicate microarrays and normalization methods are not considered.

Rantala *et al*.(6) used pin grid normalization on a per channel basis, calculated ratios between channels then assigned z-scores using all samples in the microarray. Rank product analysis was used for comparisons across microarrays and cell lines. Layout randomization and replicate microarrays are not addressed. Negative and positive controls were used to establish effective reverse transfection concentration levels but are not described in the case study assay

The computational pipeline is intentionally modular so that alternative EDA, normalization and scoring solutions can be efficiently test and implemented.

Many biomarkers are must be analyzed on a per cell level. When high speed laser scanning is used to gather population-level data, one of the channels must measure cell count. There would be benefit to developing a method to estimate cell count from the three biomarkers that are measuring differentiation features as opposed to cell count.

For normalization, there is likely benefit from fitting a linear model to the data and its position. The coefficients of the model will show the relative impact of their features. For instance, spatial variations that vary across a microarray will result in larger coefficients for the row or column terms. Principal Component Analysis could be useful as an additional visualization to identify the sources of variations in the raw data and to examine the effect of the normalizations. Open source methods such as ComBat(65) may be used to robustly address known batch effects.

With regard to the scoring, a clustering approach could group perturbations on a portion of the feature vector of their phenotype data. Unsupervised clustering would be useful if the clusters are well separated and they contain perturbations that cause similar phenotypes. Known perturbations could be used to identify the cluster and reveal novel perturbations with similar phenotypes. The extensible nature of this pipeline allows the addition of other modules for more extensive visualization and analysis.

Summary and Conclusions

The developed pipeline performs EDA, normalizes and scores data from cellular microarray experiments and stores the analysis in an interactive and reproducible fashion. The computational methods and code are based on published methods developed for well plate and spotted cDNA microarrays. Software objects that combine the data and metadata are passed between the pipeline modules, making the pipeline extensible and ensuring effective capture of data provenance and lineage.

Several key experiment design considerations have been highlighted through applying this pipeline to experimental and simulated datasets. These include randomization of the perturbations/treatments, using replicates within and across microarrays, including positive controls, matching the dyes to the imaging wavelengths, filtering out low cell count spots, including cell count endpoints and identifying low signal to noise channels.

MEArrays are particularly challenging to fully randomize. They are combinations of less than 100 ECM proteins and growth factors and it requires diligence (and a formal randomization scheme) to ensure that the final layouts do not have organized structure as occurred in the MEA dataset.

Not surprisingly, analysis of the Aberration screen showed how the lack of replicates limits the ability to reduce the technical artifacts and quantify the screen quality.

While negative controls can be replaced with the majority of perturbations that do not cause phenotype responses, positive controls are vital to assessing the quality of an assay and to estimating the PPV of a scored list of hits. Ideally, every biological endpoint would have a perturbation that creates a known effect in all cell lines. This is a challenge in MEArrays due to cell lines having different responses to their microenvironments. Analysis without positive controls further reiterates the importance of EDA in assessing microarray quality.

The ability to create simulated datasets that model variations seen in cellular microarrays has informed the expectations of PPVs in the hit lists. Modeling includes the strength of biological and systemic variations such as the responses to the controls, location and magnitude of global variations, pin grid effects and replicate-to-replicate variations. These tunable parameters inform which normalizations are most effective and aid in assessment of PPVs.

The simulated dataset showed the value of rank product analysis over z scores when combining replicates. This rank product analysis will become more important as additional channels, cell lines and drugs are analyzed in microarrays.

All steps in the analysis are scripted in R and the scripts are included with the results. Combined with the data and a description of the operating environment, this framework was implemented to ensure the pipeline is reproducible, allowing for ease to re-run analyses, facilitate external validation and secondary use of this data.

References

1. Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat Rev Cancer. 2010 Apr;10(4):241–53.

2. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006 Oct;6(10):813–23.

3. McDermott U, Sharma SV, Dowell L, Greninger P, Montagut C, Lamb J, et al. Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. Proc Natl Acad Sci. 2007 Dec 11;104(50):19936–41.

4. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 28;483(7391):603–307.

5. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Acad Sci U S A. 2012 Feb 21;109(8):2724–9.

6. Rantala J, Mäkelä R, Aaltola A-R, Laasola P, Mpindi J-P, Nees M, et al. A cell spot microarray method for production of high density siRNA transfection microarrays. BMC Genomics. 2011;12(1):162.

7. Wood KC, Konieczkowski DJ, Johannessen CM, Boehm JS, Tamayo P, Botvinnik OB, et al. MicroSCALE Screening Reveals Genetic Modifiers of Therapeutic Response in Melanoma. Sci Signal. 2012 May 15;5(224):rs4–rs4.

8. LaBarge MA, Parvin B, Lorens JB. Molecular deconstruction, detection, and computational prediction of microenvironment-modulated cellular responses to cancer therapeutics. Adv Drug Deliv Rev [Internet]. 2014 Feb [cited 2014 Apr 9]; Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169409X14000313

9. Brough R, Frankum JR, Sims D, Mackay A, Mendes-Pereira AM, Bajrami I, et al. Functional Viability Profiles of Breast Cancer. Cancer Discov. 2011 Aug 2;1(3):260–73.

10. Marcotte R, Brown KR, Suarez F, Sayad A, Karamboulas K, Krzyzanowski PM, et al. Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells. Cancer Discov. 2011 Dec 29;2(2):172–89.

11. Laufer C, Fischer B, Billmann M, Huber W, Boutros M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. Nat Methods. 2013 Apr 7;10(5):427–31.

12. Rantala J, Kwon S, Korkola J, Gray J. Expanding the Diversity of Imaging-Based RNAi Screen Applications Using Cell Spot Microarrays. Microarrays. 2013 Apr 11;2(2):97–114.

13. LaBarge MA, Nelson CM, Villadsen R, Fridriksdottir A, Ruth JR, Stampfer MR, et al. Human mammary progenitor cell fate decisions are products of interactions with combinatorial microenvironments. Integr Biol. 2009;1(1):70.

14. Lin J, Bruni FM, Fu Z, Maloney J, Bardina L, Boner AL, et al. A bioinformatics approach to identify patients with symptomatic peanut allergy using peptide microarray immunoassay. J Allergy Clin Immunol. 2012;129(5):1321–8.

15. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&amp;D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov [Internet]. 2010 Nov 19 [cited 2013 Nov 17]; Available from: http://www.nature.com.liboff.ohsu.edu/nrd/journal/v9/n3/fig_tab/nrd3078_F2.html#figure-title

16. Boutros M, Ahringer J. The art and design of genetic screens: RNA interference. Nat Rev Genet. 2008 Jun 3;9(7):554–66.

17. Robert Hooke. Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses. first edition. London: J. Martyn and J. Allestry; 1665.

18. Bridger JM, Volpi EV. Fluorescence in situ Hybridization (FISH). [cited 2014 Feb 23]; Available from: http://link.springer.com/content/pdf/10.1007/978-1-60761-789-1.pdf

19. Zanella F, Lorens JB, Link W. High content screening: seeing is believing. Trends Biotechnol. 2010 May;28(5):237–45.

20. Sklyar O, Huber W. Image analysis for microscopy screens. R News. 2006;6(5):12–6.

21. Skloot R. The Immortal Life of Henrietta Lacks. Reprint edition. Crown; 2010. 402 p.

22. Cheng W-P, Wang B-W, Chen S-C, Chang H, Shyu K-G. Mechanical stretch induces the apoptosis regulator PUMA in vascular smooth muscle cells. Cardiovasc Res. 2012 Jan 1;93(1):181–9.

23. Thoma CR, Stroebel S, Rösch N, Calpe B, Krek W, Kelm JM. A High-Throughput–Compatible 3D Microtissue Co-Culture System for Phenotypic RNAi Screening Applications. J Biomol Screen. 2013 Dec 1;18(10):1330–7.

24. Brideau C, Gunter B, Pikounis B, Liaw A. Improved Statistical Methods for Hit Selection in High-Throughput Screening. J Biomol Screen. 2003 Dec 1;8(6):634–47.

25. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, et al. Statistical Methods for Analysis of High-Throughput RNA Interference Screens. Nat Methods. 2009 Aug;6(8):569–75.

26. Zhang J-H, Chung TDY, Oldenburg KR. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J Biomol Screen. 1999 Apr 1;4(2):67–73.

27. Chung N, Zhang XD, Kreamer A, Locco L, Kuan P-F, Bartz S, et al. Median Absolute Deviation to Improve Hit Selection for Genome-Scale RNAi Screens. J Biomol Screen. 2008 Feb 1;13(2):149–58.

28. General overview of RNAi screening at ICCB-Longwood [Internet]. 2012 [cited 2013 Nov 17]. Available from: http://iccb.med.harvard.edu/wp-content/uploads/2012/08/RNAi_screen_workflow_083112.pdf

29. Willingham AT, Deveraux QL, Hampton GM, Aza-Blanc P. RNAi and HTS: exploring cancer by systematic loss-of-function. Oncogene. 2004;23(51):8392–400.

30. Mohr SE, Perrimon N. RNAi screening: new approaches, understandings, and organisms. Wiley Interdiscip Rev RNA. 2012 Mar;3(2):145–58.

31. Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. 2012 Dec 26;23(4):665–78.

32. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nat Rev Cancer. 2009 Apr;9(4):239–52.

33. Weigelt B, Lo AT, Park CC, Gray JW, Bissell MJ. HER2 signaling pathway activation and response of breast cancer cells to HER2-targeting agents is dependent strongly on the 3D microenvironment. Breast Cancer Res Treat. 2010 Jul 1;122(1):35–43.

34. Brafman DA, Chien S, Willert K. Arrayed cellular microenvironments for identifying culture and differentiation conditions for stem, primary and rare cell populations. Nat Protoc. 2012 Apr;7(4):703–17.

35. Almaça J, Faria D, Sousa M, Uliyakina I, Conrad C, Sirianant L, et al. High-Content siRNA Screen Reveals Global ENaC Regulators and Potential Cystic Fibrosis Therapy Targets. Cell. 2013 Sep;154(6):1390–400.

36. Bauer M, Kim K, Qiu Y, Calpe B, Khademhosseini A, Liao R, et al. Spot identification and quality control in cell-based microarrays. ACS Comb Sci. 2012 Aug 13;14(8):471–7.

37. Tukey JW. Exploratory data analysis. 1977 [cited 2014 Apr 19]; Available from: http://xa.yimg.com/kq/groups/16412409/1159714453/name/exploratorydataanalysis.pdf

38. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2013. Available from: http://www.R-project.org/

39. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

40. Wang X, Hessner MJ, Wu Y, Pati N, Ghosh S. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. Bioinformatics. 2003 Jul 22;19(11):1341–7.

41. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003 Dec;31(4):265–73.

42. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002 Feb 15;30(4):e15.

43. Yang YH, Dudoit S. Normalization: Bioconductor's marray package. 2011 [cited 2014 Jan 17]; Available from: http://stuff.mit.edu/afs/athena.mit.edu/software/r/current/arch/amd64_linux26/lib/R/library/marray/doc/marrayNorm.pdf

44. Tseng GC, Oh M-K, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 2001 Jun 15;29(12):2549–57.

45. Wiles AM, Ravi D, Bhavani S, Bishop AJR. An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme. J Biomol Screen. 2008 Sep;13(8):777–84.

46. Dragiev P, Nadon R, Makarenkov V. Two effective methods for correcting experimental high-throughput screening data. Bioinformatics. 2012 Jul 1;28(13):1775–82.

47. Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R. An efficient method for the detection and elimination of systematic error in high-throughput screening. Bioinformatics. 2007 Apr 26;23(13):1648–57.

48. Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, Steinbrink S, et al. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol [Internet]. 2010 Jun 8 [cited 2013 Oct 27];6. Available from: http://www-ncbi-nlm-nih-gov.liboff.ohsu.edu/pmc/articles/PMC2913390/

49. Peng RD. Reproducible Research in Computational Science. Science. 2011 Dec 2;334(6060):1226–7.

50. Brás L, Boutros M, Huber W. Analysis of multi-channel cell-based screens. 2010 [cited 2013 Apr 23]; Available from: http://bioc.ism.ac.jp/2.11/bioc/vignettes/cellHTS2/inst/doc/twoChannels.pdf

51. Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP. Experimental Design and Low-Level Analysis of Microarray Data. In: Michael F. Miles, editor. International Review of Neurobiology [Internet]. Academic Press; 2004 [cited 2014 Apr 10]. p. 25–58. Available from: http://www.sciencedirect.com/science/article/pii/S007477420460002X

52. McCall MN, Murakami PN, Lukk M, Huber W, Irizarry RA. Assessing affymetrix GeneChip microarray quality. BMC Bioinformatics. 2011 May 7;12(1):137.

53. Kim K, Page GP, Beasley TM, Barnes S, Scheirer KE, Allison DB. A proposed metric for assessing the measurement quality of individual microarrays. BMC Bioinformatics. 2006 Jan 23;7(1):35.

54. Reimers M, Weinstein JN. Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. BMC Bioinformatics. 2005;6(1):166.

55. Catherine Loader. locfit: Local Regression, Likelihood and Density Estimation [Internet]. 2013. Available from: http://CRAN.R-project.org/package=locfit

56. Reed GF, Lynn F, Meade BD. Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. Clin Diagn Lab Immunol. 2002 Nov;9(6):1235–9.

57. Jenssen T-K, Langaas M, Kuo WP, Smith-S?rensen B, Myklebost O, Hovig E. Analysis of repeatability in spotted cDNA microarrays. Nucleic Acids Res. 2002 Jul 15;30(14):3235–44.

58. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77.

59. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006 Jun;27(8):861–74.

60. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004 Aug;573(1-3):83–92.

61. Eisinga R, Breitling R, Heskes T. The exact probability distribution of the rank product statistics for replicated experiments. FEBS Lett. 2013 Mar;587(6):677–82.

62. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008 Jan 18;24(3):374–82.

63. Abeygunawardena N. MIAME - Workgroups - FGED [Internet]. 2007 [cited 2014 Apr 28]. Available from: http://www.mged.org/Workgroups/MIAME/miame.html

64. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods. 2005 May;2(5):345–50.

65. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):e161.

Appendices

Appendix A Dataset Perturbations

Table A1 - Aberration siRNA targets

TMEM190
CD59
ARHGAP5
AKAP6
SSC5D
BEND3
ITGBL1
WNK1
IRS2
VSTM2B
WT1-AS
TBL1XR1
ZNF581
CDH1
PRPF18
PGAP3
TSC2
THRSP
FKSG29
MLH1
PPP1R12B
KCTD14
HSPBP1
MBIP
KLHL12
PDSS2
TMCO3
PPIF
CBLB
KISS1
PLEKHF1
FRMD4A
LOC283050
MRPS25
EPS15
ZNF787
ERBB2
UBL3
CCDC90A
NUMA1
OPTN
SNORA77
NKX2-8
SLC9A3
KDM5B
SLC41A1
DCUN1D2
BEND7
CNTN2
CCDC73
SMARCA4
TMEM183A
UPF2
C19orf51
PEX5L-AS2
LAX1

SFTPA1
C3orf19
ADORA1
NDUFB5
F7
STRN3
NLRP4
ECM1
SYT5
NFASC
HECTD1
TFDP1
KLHDC8A
SEPHS1
CDK18
LOC284578
FAM72A
ARGLU1
EPN1
COCH
RAB20
SLC15A1
NUBPL
RAP2A
NLRP5
NKX2-1
BRIP1
ZC3H11A
EIF3M
PHGDH
ATM
CHEK2
ZNF697
EIF5AL1
PPP1R12C
TNNI3
GOLPH3L
PHKB
GP6
RASA3
RPRD2
FANCD2
NINJ2
HEATR5A
RSF1
C19orf29
WT1
MCM10
LEMD1
TUBD1
GOLT1A
CHAMP1
AP4S1
ZIC5
URI1
MDM4

RDH13
GNB4
ABCC4
VOPP1
FRS2
CD83
NDUFC2
CCDC168
ZNF579
CSTF3
FGF14
TPD52
SLC10A2
LINC00346
LOC389493
SNORA15
RAB11FIP1
MRPS17
MCF2L
KCNMB3
PIK3CA
C19orf12
TCP11L1
C13orf35
CPM
MFSD4
PF4
LAMP1
DCT
GATA3
QSER1
CLDN10
CARD11
SNRPE
SUV420H2
EDN1
MLLT4
SOX13
CCND1
ETV5
SFTA3
STARD3
RANBP9
CDC123
DNAJC3
ECHDC3
B4GALNT3
TMTC4
SBK2
SYT2
G2E3
MCL1
TNNT1
LRRN2
MYO16
CAMK1D

SLC26A9
ABHD13
AFP
NPAS3
BIRC3
MBL1P
ZCCHC24
NUP214
FGFR1
TMEM86B
EPS8L1
ADAMTSL4
HAO2
UPF3A
AVPR1B
HSD3B1
ALG8
CXCL1
ZNF704
RABIF
TP53
ZNF703
LOC148709
MRPL47
LYZ
CPSF6
CLNS1A
TEX30
NLRP2
ZSCAN5B
STMN2
PSPH
GPR124
CCT6A
KCNMB2
ZNF639
ATP11A
C11orf41
FAM71E2
C6orf203
KDELC1
CCNE1
METTL21C
ZNF784
CDC16
TCAP
DOCK9
IGF1R
DZIP1
RBBP5
GDPD4
TUBGCP3
PRRG4
ARHGEF7
LOC650623
RAP1GDS1

PHYH
RNF182
NFKB2
OPTC
NLRP8
POU5F1
SSTR1
PCID2
FMOD
GNA11
TMEM183B
C1orf186
LOC283070
HMGCS2
FAM155A
PLEKHA6
CLEC14A
ANKRD10
NLRP11
ENSA
GRTP1
SCFD1
COX18
PM20D1
ZNF580
PAX9
COL4A1
HS6ST3
U2AF2
PPP1R15B
DSTYK
BLM
FGD5
PDGFRB
ADAM30
NBPF7
SHISA7
TMEM150B
CTSS
RB1
TMEM81
ZFYVE20
RAB7L1
ADIPOR1
PTEN
TNFSF13B
CLYBL
UBE2S
IPO5
FKBP9L
BIVM
ERCC5
TEX29
GBAS
SUMF2
CUL4A

POP4
AQP11
FAM70B
METTL21CP1
TMCC2
CXCL6
ATP4B
ORAO
FOXA1
USP22
PROZ
GPC3
FBXO3
NCOA4
ZNF524
PIK3C2B
FANCG
CHI3L1
LINC00410
LOC127841
ARHGAP5-AS1
CREBBP
ACTL6A
UCMA
CTSE
NKX2-1-AS1
BRSK1
REN
PEX5L
ZSCAN5A
PPP6R1
CHIT1
C14orf126
NUCKS1
MFN1
ANKRD17
USP13
KAT6A
ZMIZ1
COL1A1
KRAS
PTPRH
ZNF628
TARS2
GPC6
NF2
DHTKD1
NUAK2
PAG1
MIPOL1
AFM
ZBTB10
HRAS
BRF2

CYB5R1
IL11
JARID2
NOL7
YEATS4
INTS4
ISOC2
GALP
MRPS28
PROSC
CHCHD2
GPC5
ZMAT3
ADPRHL1
C11orf67
FIZ1
LOC553137
TPP2
A2LD1
NR2C2
PNMT
UGGT2
SMARCB1
SRGAP2
MYC
EFNB2
LINC00303
RPS6KB1
USP6NL
PLAT
SEC61A2
SFTPA2
SPTSSA
DEPDC7
FLJ44054
SIRT5
NUDT5
BTG2
NLRP13
BRAF
GFOD1
CARS2
MYOG
IL6ST
KDM5B-AS1
C10orf47
RAD52
AP3M2
C3orf20
ETNK2
PHACTR1
F10
HORMAD1
SOX1
ALB
CCDC3

ELK4
ERC1
CCDC106
EGLN3
COL4A2
RNF113B
NLRP9
RFPL4A
MYBPH
NTRK1
BRCA2
PER1
NOTCH1
REG4
HSD3B2
NAT14
COX6B2
IKBKB
EP300
RASSF6
ATP2B4
SLC45A3
PRELP
PML
LIG4
PCCA
ERLIN2
OXGR1
LANCL2
GRK1
ZNF444
HIPK3
HEY1
NALCN
SPACA7
DAOA
ZNF713
CARKD
ZNF217
VMP1
ING1
GAS6
PHKG1
PAK1
MAP2K4
EGFR
MAP3K1
BRMS1L
IL8
FGFR2

Table A2- MEA ECM Proteins

Col I	Fibronectin
MG Col II	Integrin a2b1
Col III	Integrin a6b4
Col IV	Integrin a4b1
HyA	Integrin a5b6
HMW	Integrin a3b1
ICAM-1	Vitronectin
Desmoglein	VCAM

Table A3 - MEA Growth Factors

VEGF 165	BMP4	SDF1b	ANGPT1
TGFB1	EGF	IL-1B	SDF1a
SCF	Wnt 5a	ANGPT2	Osteoactivin
PBS	IL-8	Jagg 2	HGF
Wnt10b	TgfB2	OPG	IL-3
SHH	Jagg 1	IGF-1	CTGF

Appendix B: Olympus Scan^R Microscope

Hardware		Olympus IX83 and IX81 microscope
		CCD Hamamatsu: ORCA-ERG, C8484, ORCA 285. Intensified and EMCCD cameras (Hamamatsu) on request
		Motorised Stage Märzhäuser SCAN IM for IX3 and IX2 frames
		Imaging Computer (latest generation PC), 2 Hard-Drives (80 GB and 250 GB), 2GB RAM
Hardware control		MT20
	Short arc burners	150 W Xenon or Mercury-Xenon
	8 Filter positions	Diameter 25 mm
	Filter switch	min. 58 ms (neighbouring positions)
	Attenuation	14 levels, 1% - 100%
	Attenuation switch	<58 ms
	Shutter, on/off time	1 ms
	Operation	all modules in parallel
Image Acquisition		Workflow oriented configuration and user interface
		Variable powerful software auto-focus procedures
		Format Manager with predefined formats (slides, multiwell plates) and editing interface to create and edit customised formats (spotted arrays)
		On-line display
		Fully automated operation
		User interaction: pause, resume, set marker
Image Analysis		On-line & off-line analysis
		Independent software module
		Image processing
		Image analysis & particle detection
		Parameter extraction and calculation
		Cytometric data analysis, gating & classification
		Direkt link between data-points, objects and images
		Data-export
		Complex analysis procedures can be saved as assays
		Predefined assays and advanced scientific assay development functionality
	Fast filter turret	min. switching time 350 ms
		Professional data storage systems
Performance	Image acquisition 2	1 s/position

	colour channels á 200 ms	
	IR-hardware auto-focus	1 - 2 s/position
	software auto-focus	2 - 5 s/position
General Features		Maximum throughput and minimised bleaching/photo-toxicity due to real-time synchronisation
		Homogeneous and high intensity fluorescence illumination by optimised illumination optics
		Thermal insulation of samples and vibration free image acquisition by fiber-optic illumination coupling
		High precision and high endurance components
		Maximum flexibility and modularity by open microscope based system platform
		Fluorescence and transmitted light screening
		"Unlimited" colour channels

Appendix C: Teacan LS Reloaded Laser Scanner

General

- Possible Laser Sources 635 nm, 532 nm, 488 nm, 594 nm
- Detectors 1 or 2 PMTs, optional simultaneous dual color detection
- Emission Filters 1 or 2 filter slides, space for 4 filters each
- 1 Laser System: CY5 (692/45),
- 1-4 Laser System: CY5, ROX (635/35), CY3 (575/50), FITC (535/25) standard filters
- Additional filters can easily be added. Up to 20 filters can be applied per system and registered by LS Software.
- Substrates as microscope slides, microplates or any other substrate up to the size of microplates (15x85x127 mm)
- Scan area slides 22x75 mm one run; 1-96 scan areas over an area up to microplate footprint (automated stitching)
- Autofocus for transparent or not-transparent surfaces with or without segmentations performed in advance to each individual scan
- Depth of Focus 90, 300, or 800 μm reading by 3 software selectable pinhole sizes
- Working Distance 6.5 mm (focus)
- NA 0.6
- Adjustable angle of incidence of laser beam 25-0° for evanescence resonance scanning and three-dimensional structures

Performance

- Signal to electronic noise ratio 5 orders of magnitude
- Sensitivity < 0.1 Fluorophore equivalent/μm^2
- <0.01 Fluorophore equivalent//μm^2 by adjusted laser beam angle to Cy3/Cy5 evanescence resonance slide scanning
- Intrascenic Dynamic Range 4.5 orders of magnitude
- Gain Adjustment 5 orders of magnitude (0.005% to 500%)
- Pixel Resolution 6, 10, 20, 40 μm
- Reading speed 4 minutes for a full slide, 25 minutes a SBS plate (dual color, 10 μm pixel resolution),

Pipeline Software

Pre-release code for the computational pipeline and simulation functions and the thesis datasets are available at GitHub markdane/CMA2 at DOI: 10.5281/zenodo.10145.