UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
An Imaging-Based Human Embryonic Stem Cell Assay for Teratogenic Activity

Permalink
https://escholarship.org/uc/item/315865xx

Author
Perea, Brian

Publication Date
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

An Imaging-Based Human Embryonic Stem Cell Assay for Teratogenic Activity

by

Brian Perea

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Douglas Clark, Co-chair
Professor David Schaffer, Co-chair
Professor Markita Landry
Professor Sandrine Dudoit

Spring 2017

An Imaging-Based Human Embryonic Stem Cell Assay for Teratogenic Activity

Copyright 2017
by
Brian Perea

Abstract

An Imaging-Based Human Embryonic Stem Cell Assay for Teratogenic Activity

by

Brian Perea

Doctor of Philosophy in Chemical Engineering

University of California, Berkeley

Professor Douglas Clark, Co-chair

Professor David Schaffer, Co-chair

In this work a human pluripotent stem cell (hPSC) assay was proposed as a novel imaging-based screen for identifying substances (teratogens) likely to cause birth defects in humans. The assay was engineered to leverage lessons from the validated mouse Embryonic Stem Cell Test (EST). An automated fluid handler was repurposed to support printing stem cells and biological materials to an innovative two-part microculture platform, the pillar/well chip system. Key milestones were met, demonstrating the applicability of the microculture platform to hPSC cultures over the duration of the proposed assay. Three proof-of-principle experimental data sets were acquired using statistical experimental design to eliminate confounding systematic variation and uncontrolled variables. Changes in cell population and expression of four protein biomarkers were observed with automated fluorescence microscopy and quantified by image analysis. Toxic molecules were found to show similar inhibition of viability to literature values. Further work is needed to optimize quantification and reduce variability in these signals to assess the effects of candidate teratogens on cell differentiation.

To Elisabeth Perea

Contents

Contents ii

List of Figures iv

List of Tables vi

1 A Human Stem Cell Developmental Toxicology Assay 1
- 1.1 The Cost of Failing to Detect Teratogens Early . . . 1
- 1.2 Animal Testing: The Current Gold Standard . . . 2
- 1.3 The Predictive Capacity of Animal Models . . . 3
- 1.4 Data Gaps and Issues of Scale . . . 4
- 1.5 Development of *In Vitro* Teratogenicity Assays . . . 4
- 1.6 Proposed Work . . . 7

2 A Platform for Rapid Microscale Stem Cell Assays 8
- 2.1 A Two-Part System Supports Distinct Microcultures . . . 8
- 2.2 Automated Fluid Handling Enables Controlled Preparation of Experimental Conditions . . . 10
- 2.3 Automated Fluorescence Imaging Supports Acquisition of Rich Biological Data 14
- 2.4 Demonstration of the Platform . . . 16

3 The Differentiation Assay 19
- 3.1 Redesign of the Embryonic Stem Cell Test . . . 19
- 3.2 Experimental Design . . . 22

4 Quantitative Image Analysis 28
- 4.1 The Emerging Role of Imaging in Toxicology . . . 28
- 4.2 Colocalization: Analysis of Multiple Fluorescent Stains . . . 29
- 4.3 Applying Quantitative Colocalization to the Differentiation Assay . . . 31

5 Statistical Analysis of Dose-Response Data 43
- 5.1 Analysis of Experimental Screening Data . . . 43
- 5.2 Experimental Results . . . 46

6 Closing Remarks **55**
6.1 Summary 55
6.2 Outlook 55

Bibliography **57**

A **64**
A.1 Experimental Methods 64
A.2 Preparing the Data Set 65
A.3 Exploratory Analysis of Dose-Response Data 66
A.4 Linear Mixed Effects Model Selection 76
A.5 Comparison of Mixed Effects Model and Nonparametric Results 77
A.6 Cytotoxic Dose-Response Curves 80

List of Figures

2.1 Diagram of the pillar/well chip microculture platform 9
2.2 Reference photo of the automated fluid handler adapted for the pillar/well chip culture platform 11
2.3 Illustration of imaging challenges for stem cell cultures embedded within a hydrogel matrix 12
2.4 Steps in fluid handling for cell culture preparation 13
2.5 A custom Python-based protocol builder simplifies printing in arbitrary positions 14
2.6 With a custom plate definition file, a single well is captured in single field-of-view 15
2.7 Cell seeding is well-controlled 16
2.8 Stem cells remain viable after seeding on chip 17
2.9 Stem cells retain hallmarks of pluripotency on chip 18

3.1 Visual summary of nested experimental design 24

4.1 Sample wells from the the proof-of-principle developmental toxicology data set show various culture morphologies 32
4.2 Single-cell segmentation of Hoechst nuclear stain in a control low-density culture of pluripotent stem cells 33
4.3 Fluorescence image data is a hierarchical data format 34
4.4 Local Binary Pattern algorithm may be used to classify individual pixels based on local intensity differences of surrounding pixels 36
4.5 Texture-based segmentation algorithm robustly segments cell-containing foreground from background 37
4.6 Hoechst nuclear counterstain foreground area correlates linearly with cell count . 38
4.7 Sample simulated images highlighted Manders Overlap M1 and M2 as possible colocalization method candidates 39
4.8 Manders Overlap M1 is a computationally-efficient statistic to approximate colocalization within a cell population 40
4.9 Visual guide to the metrics selected to approximate cell population and biomarker expression 42

5.1 Sample stain montage with definitive endoderm positive control wells highlighted and quantified . . . 47
5.2 Mixed effect analysis enables estimation of cell viability with the Hoechst foreground area approximation . . . 48
5.3 Controls stained for Hoechst and Oct4 . . . 51
5.4 Germ layer stain control plots . . . 52
5.5 Dose-response data for thalidomide . . . 54

A.1 A sample randomized plate map . . . 65
A.2 Cell population trends at endpoint by plate . . . 67
A.3 Cell population within plate trends at endpoint grouped by plate position . . . 68
A.4 Pluripotency within plate trends at endpoint grouped by plate position . . . 69
A.5 Mesodermal marker Brachyury expression within plate trends at endpoint grouped by plate position . . . 70
A.6 Ectodermal marker PAX6 expression within plate trends at endpoint grouped by plate position . . . 71
A.7 Endodermal marker GATA4 expression within plate trends at endpoint grouped by plate position . . . 72
A.8 Example analysis of hierarchical data structure for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area metric . . . 74
A.9 Example analysis of hierarchical data structure for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area metric . . . 76
A.10 Overlay of parametric and nonparametric analyses on sample control and dose-response data . . . 78
A.11 Comparison of parametric and nonparametric analyses on sample control and dose-response data . . . 79
A.12 Viability dose curves for Table 5.1 . . . 80

List of Tables

1.1 Statistics used in analyzing predictive power 3
1.2 ECVAM *in vitro* assay validation statistics calculated from model validation data on 14 test substances (adapted from [27]) 6

3.1 Biomarkers and staining 22
3.2 Overview of experimental replicates 23
3.3 Experimental controls included 25
3.4 Library of compounds screened in proof-of-principle experiments 27

4.1 Overview of quantitative colocalization methods proposed in literature 30
4.2 Pearson R values calculated for subregions highlighted in Figure 4.3 35

5.1 Comparison of estimated IC_{50} values to mouse embryonic stem cell literature values for cytotoxic compounds 49

A.1 Initial cell counts prior to seeding on day 0 67

Chapter 1

A Human Stem Cell Developmental Toxicology Assay

1.1 The Cost of Failing to Detect Teratogens Early

A teratogen is any agent that can disturb the development of an embryo or fetus. Teratogens include maternal infections, radiation, and exposure to chemicals and drugs. Familiar chemical teratogens include alcohol, nicotine, cocaine, and broad-spectrum antibiotics such as doxycycline. One in every 33 babies born in the United States is affected by congenital defects.[1] Approximately 2% to 3% of cases can be attributed to teratogen-induced malformations, and up to 70% of cases have an unknown etiology.[2] Although teratogenic exposures account for a small fraction of total birth defects, defect-associated conditions may persist throughout the life of a child resulting in lifetime medical costs exceeding $500,000 per case.[2] Fetal Alcohol Spectrum Disorders (FASD), for example, lead to facial malformations, growth deficits and neurodevelopmental problems, and represent one of the main causes of mental retardation worldwide. In the United States, 4.0 billion dollars are expended annually to treat and rehabilitate patients with FASD.[3] Teratogens are a costly and preventable source of morbidity and mortality.

Lessons from Thalidomide

Thalidomide is a chemical teratogen marketed in the late 1950's as a safe and non-habit-forming sedative. Widespread use of thalidomide for off-label treatment of morning sickness led to approximately 10,000 children born with thalidomide-attributed defects worldwide.[4] Mortality rates for thalidomide were estimated to be 40% of all cases, therefore the survivors may represent only 60% of the total affected.[5] Thalidomide-induced birth defects included limb reduction anomalies, congenital heart disease, malformations of the inner and outer ear, and ocular abnormalities.[4]

Animal tests conducted on the drug showed no toxic dose in adult animals[6], but the drug was not tested in pregnant animals prior to observations of toxic effects. When the alarming

increase in birth defect rates was connected to thalidomide in 1961, numerous efforts were initiated to validate the teratogenic activity of the drug in animal studies. Thalidomide has since been tested in numerous animal species including mice, rats, rabbits, hamsters, guinea pigs, ferrets, cats, dogs, swine, armadillos, and primates.[7] In the animal models tested, teratogenic effects have been induced only occasionally with no clear consensus emerging among species. The thalidomide tragedy sparked discussions resulting in a new era of drug regulations requiring evidence of safety and efficacy.

1.2 Animal Testing: The Current Gold Standard

Legislation introduced in the United States shortly after the thalidomide tragedy mandated reproductive and developmental studies on animals prior to clinical trials of a drug in addition to more extensive studies before a drug could be approved for marketing. In January 1966 the United States Food and Drug Administration (FDA) issued guidelines for the so-called "3-Segment Study" which divided the reproductive cycle into three segments for evaluation of toxicity: fertility and reproductive performance, *in utero* development and teratology, and perinatal and postnatal effects.[8]

Guidelines for teratology studies recommended testing at least two animal species; mouse, rat, and rabbit being the most common. Treatment was expected to cover the developmental stages of organogenesis (gestation days 6-15 in mice and rats, days 6-18 in rabbits [8]) using the route of administration relevant to the intended application. Drug dosage guidelines specified at least two dose levels, one at the highest maternally tolerated dose level and a second at a dose level some multiple of the intended dose. A concurrent negative control group was mandated, and positive control groups were recommended when available. Each test group was expected to contain at least 20 females for rats or mice, or at least 10 dams for rabbits with untreated males used to produce pregnancies. The guidelines stated that females should be killed and fetuses delivered by cesarean section one to two days prior to the expected birth. Fetus-related endpoints included [8]:

- the number and placement of fetuses
- count of live and dead
- count of early and late resorptions
- individual weight
- external anomalies
- visceral or skeletal anomalies detected by dissection or other methods

The FDA guidelines for teratology form the basis of similar guidelines in the United States for applications in food and pesticides and were adopted more broadly in the European Union, Japan, and Australia in the early 90's after the International Conference on Harmonization of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH) issued guidelines for harmonized international testing requirements.[8]

1.3 The Predictive Capacity of Animal Models

Animal models have contributed greatly to our understanding of biology. [9, 10] Developmental biology in particular has benefited from comparisons between mechanisms conserved across the animal kingdom from *Drosophila* to *Xenopus* and beyond. The use of animal models to explore the origins of biological mechanisms is different altogether from the application of animal models to *predict* human health outcomes. [2, 7, 11, 12] The premise of the former is that homology between growth and differentiation processes governed by genetic mechanisms are conserved across a wide range of animal species—a concept validated by decades of scientific investigation into the processes governing development in numerous animal species. The premise of the latter is that toxicological outcomes in animals are generally predictive of human outcomes.

To be accounted as predictive, toxicology studies using animal models should be scientifically validated using drugs or other agents with known human outcomes. Validation should characterize the predictive power of such studies in terms of sensitivity, specificity, and positive and negative predictive values (Table 1.1). [7, 12] These statistics answer important questions for evaluating the risk associated with a prediction.

- **Positive Predictive Value:** How likely is a positive prediction a true positive?
- **Negative Predictive Value:** How likely is a negative prediction a true negative?
- **Sensitivity or Recall:** How likely is it that a true positive will be detected (predicted as positive)?
- **Specificity:** How likely is it that a true negative will be predicted as negative?

If toxicological animal model studies perform better than random chance, they may be considered predictive to a greater or lesser degree. If they do not, then animal studies contribute no useful information to human health outcomes.

		Human Outcomes	
		+	-
Test	+	TP	FP
Outcomes	-	FN	TN

TP = True Positive
TN = True Negative
FN = False Negative
Sensitivity = TP/(TP + FN)
Specificity = TN/(FP + TN)
Positive Predictive Value = TP/(TP + FP)
Negative Predictive Value = TN/(FN + TN)

Table 1.1: Statistics used in analyzing predictive power

Thalidomide poignantly illustrates flaws in the argument that toxicological animal models are predictive of human developmental outcomes. Even if all the animals mentioned above had been tested for teratogenic activity with thalidomide, the wide variety of outcomes lend little credibility to a consensus-based approach where evidence from different models is weighted with equal importance and even less to the idea that any evidence of risk in any single species is evidence of potential risk in humans. If additional weight were to be given to species genetically closer to humans such as non-human primates, then the correspondence between teratogenic activity in humans and nonhuman primates would be valid for thalidomide only by happenstance—nonhuman primates have no more predictive power than a coin toss when tested with other drugs toxic to humans.[7]

1.4 Data Gaps and Issues of Scale

In addition to concerns about biological relevance, the prohibitively high time and expense of animal toxicology studies precludes such methods from meeting the current testing demand. In 2001, the European Union proposed updating testing requirements for 30,000 chemicals marketed before 1981. If animal-based methods were used, the testing required was estimated to cost 8.68 billion Euros (11 billion US Dollars), 12.8 million animals, and take up to 45 years to complete.[11] In the United States, the Food Additives Amendment of 1958 allows more than 10,000 chemicals to be added directly or indirectly to human food.[13] Of the chemical additives intentionally added to food, 80% lack sufficient testing to estimate minimal consumer safety standards and 93% have no reproductive or developmental toxicity data at all.[13] In clinical pharmacology, more than 60% [14] of pregnant women in the United States take at least one prescription medication during pregnancy. Of the 172 drugs approved by the FDA between 2000 and 2010, the teratogenic risk was undetermined for 168 (98%), and 126 (73%) had no reproductive or developmental toxicology data at all.[14] For the 640 prescription drugs approved by the FDA between 1980 and 2010 teratogenic risk could not be determined for 584 (91%).[14] With such little evidence available to support the FDA-approved safety classifications there is a strong demand for rapid and inexpensive first-pass screens to provide basic developmental toxicology data and to help prioritize compounds for more detailed testing.

1.5 Development of *In Vitro* Teratogenicity Assays

In response to the demand for more rapid and inexpensive developmental toxicology assays, a number of *in vitro* methods have been proposed using animal models commonly studied in developmental biology including fly [15], zebrafish [16], frog [17], chick [18], rat [19], and mouse [20] whole embryo cultures. All methods have shown some success in predicting teratogenic activity for teratogens active in animal-based toxicology studies, but it remains unclear how relevant these tools are to human health outcomes. In practice, whole embryo

culture methods share many of the same limitations in cost and scalability common to animal testing, and even splitting embryos into partial embryo cultures such as in the rat midbrain and limb bud micromass assay [21] resolves little of the highly specialized manual labor and experimental complexity that limit throughput and drive expense.

The Validated Embryonic Stem Cell Test

In the early 90's, a mouse embryonic stem cell test [22] was proposed as a model for teratogenic activity. Embryonic stem cell lines (ESCs) are derived from the inner cell mass of the blastocyst—the earliest stage of embryonic development after fertilization of the egg. ESCs have two highly desirable properties for *in vitro* teratogen assays: (i) they are capable of proliferating indefinitely under appropriate culture conditions, and (ii) they are capable of differentiating into all of the tissues of the body. A stem cell-based developmental toxicology assay eliminates the cost and complexity of maintaining animal facilities since stem cell lines, once established, may be propagated without sacrificing additional animals. In addition, cell-based assays are amenable to scaling using multi-well tissue culture plates and automation of fluid handling and data acquisition. While there are concerns that maternally-associated factors such as pharmacokinetics and metabolism are not accounted for in most *in vitro* assays, the U.S. National Academy of Sciences has proposed a toxicology strategy integrating *in vitro* assays with pharmacokinetic and computational systems biology models to address these limitations. [23] Thus stem cell assays may sidestep the major factors that contribute to differences in toxicity between species by focusing only on disruption of the specific developmental mechanisms, often conserved in evolution, which govern teratogenic activity.

A form of the mouse Embryonic Stem Cell Test (EST) was subjected to formal validation by the European Centre for the Validation of Alternative Methods (ECVAM) [24–26]. Mouse embryonic stem cells (mESCs) may be maintained in the undifferentiated state in the presence of Leukaemia Inhibitor Factor (LIF). Upon withdrawal, mESCs spontaneously differentiate into various cell types. In the validated EST, pluripotent mESCs (cell line D3) were subjected to spontaneous differentiation by LIF withdrawal and exposed to test chemicals over the course of 10 days. Three toxicological endpoints were measured: (i) cytotoxic effects on mESCs, (ii) cytotoxic effects on mouse fibroblasts (cell line 3T3), and (iii) a morphological analysis of beating cardiomyocytes (cardiac muscle cells) arising from differentiating mESCs. Endpoints (i) and (ii) were measured using the MTT colorimetric viability assay and endpoint (iii) was recorded by manually inspecting individual wells and counting the number of beating cell colonies. A prediction model was developed using a three-component Linear Discriminant Analysis [27] to combine and weight the three endpoints such that maximal separation between teratogenic classes was achieved. Manual thresholds were defined for classifying molecules according to the three predefined teratogenic risk levels used in this study: non-teratogenic (Class 1), weakly teratogenic (Class 2), and strongly teratogenic (Class 3). The teratogenic-activity classifications were defined using a consensus approach to animal developmental toxicology studies as follows [28]:

- **strongly teratogenic (Class 3):** strongly embryotoxic and developmentally toxic in all species tested, inducing multiple defects with a high ratio between the adult and developmentally toxic doses
- **weakly teratogenic (Class 2):** chemicals of intermediate activity
- **non-teratogenic (Class 1):** not developmentally toxic at maternally toxic exposures—embryo or fetal toxicity cannot be separated from maternal toxicity

Twenty test substances were selected for the validation study and split into a training and test set composed of 6 and 14 substances, respectively, with a nearly equal distribution of classes in both subsets, and the performance of the EST was compared with results from whole embryo culture (WEC) and limb bud micromass (MM) assays on the same library (Table 1.2).[27] The EST compared favorably with other *in vitro* tests in all performance metrics.

Statistic	**EST**	**MM**	**WEC Model 1**	**WEC Model 2**
Positive Predictive Value (%): non	72	57	56	70
Positive Predictive Value (%): weak	70	71	75	76
Positive Predictive Value (%): strong	100	100	79	100
Negative Predictive Value (%): non	83	85	80	88
Negative Predictive Value (%): weak	89	77	73	82
Negative Predictive Value (%): strong	91	85	96	100
Sensitivity (%): non	70	80	70	80
Sensitivity (%): weak	83	60	45	65
Sensitivity (%): strong	81	69	94	100
Specificity (%): non	52	25	35	36
Specificity (%): weak	33	62	79	64
Specificity (%): strong	100	100	80	100
Accuracy (%):	78	70	68	80

Table 1.2: ECVAM *in vitro* assay validation statistics calculated from model validation data on 14 test substances (adapted from [27])

Although the EST demonstrated promise as a predictive in vitro test for developmental toxicity, a number of concerns have prevented the test from being widely adopted.[23] The test uses mouse cell lines and must extrapolate results to humans. Mouse ESC culture techniques were more advanced when the EST was proposed in the early 90's, but hESC culture is readily accessible today. In addition, the design decision to screen for terminally-differentiated cells in the cardiac lineage does not capture the diversity of mESC fate decisions and requires many days of differentiation to produce cardiomyocytes. [29] The beating colony morphology endpoint itself is qualitative, requires extensive training to discern, is

difficult to standardize between even well-trained practitioners, and fundamentally limits the throughput of the assay. [29, 30]

1.6 Proposed Work

The following chapters will describe a proposed human embryonic stem cell developmental toxicology assay designed to address specific concerns limiting the applicability of the validated Embryonic Stem Cell Test at an industrially-relevant scale. The objective of the assay is to serve as a first-pass screen to provide basic risk data and prioritize substances for more detailed testing. Because maternal factors such as pharmacokinetics and metabolism are not accounted for, the assay results may only be interpreted as indicative of risk to a developing human embryo after a particular substance has penetrated maternal barriers at a set dose level. Additional screens, whether cell-based assays targeting specific mechanisms of toxic activity, *in silico* predictions of bioavailability, or existing common practice animal testing would be required for a reasonable estimation of safety during pregnancy and are beyond the scope of the work considered here.

For the assay to be biologically meaningful and cost effective, experimental methods must be developed to deliver quantitative biological readouts while reducing manual manipulations and evaluation steps. To this end, a proof-of-principle microscale cell culture platform was developed to support facile fluid handling for high density, low volume cell screens. This platform was paired with a high content microscope configured to automatically raster from microculture to microculture and acquire multiple fluorescence images. Methods for quantitative characterization of fluorescence image data were developed to extract measurements of cell population and gene expression levels. These experimental methods were validated individually then used to generate three proof-of-principle developmental toxicology data sets. The statistical analysis techniques needed to analyze the hierarchical data produced in the described experiments are non-trivial, so discussions of the methods explored, relevance of those methods, and outcomes of the experiments are included.

- **Chapter 2** will describe the microscale cell culture platform and associated protocols developed to demonstrate proof-of-principle automation of fluid handling for human pluripotent stem cell screens.
- **Chapter 3** will describe the biological foundations of the modified developmental toxicology assay as well as methodological considerations including the experimental design used to acquire a set of three proof-of-principle data sets.
- **Chapter 4** will describe the methods used to extract quantitative gene expression data from fluorescence images of co-stained cells and how those methods were selected.
- **Chapter 5** will describe the statistical analysis of data tables derived from image analysis of the three experimental data sets.
- **Chapter 6** will summarize the outcomes of the work and evaluate potential future applications.

Chapter 2

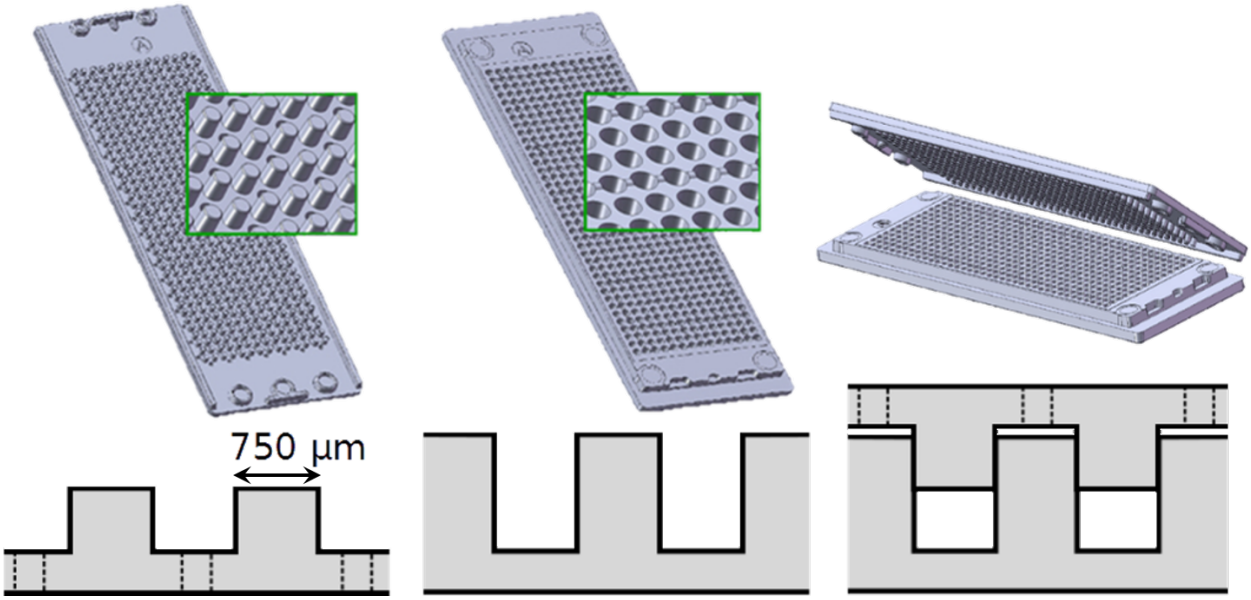
A Platform for Rapid Microscale Stem Cell Assays

2.1 A Two-Part System Supports Distinct Microcultures

A custom microculture chip system and compatible fluid handling apparatus were adapted for high throughput cellular assays using human pluripotent stem cells (hPSCs). The microculture chip system consists of two complementary polystyrene chips, one with 532 microposts, each 750 microns in diameter, and one with 532 complementary wells (Figure 2.1). Pillar chips are perforated with holes situated between pillars to allow for gas permeation. The chips have the footprint of a standard glass coverslip and together enable straightforward liquid handling for high density, low volume microcultures.

The pillar/well chip system has a number of advantages over other microculture platforms and conventional plate formats. Nearly all comparable microculture platforms expose all wells to a uniform media bath, severely limiting the number of soluble factors that may be explored. The pillar/well chip system enables two or three dimensional cell culture while achieving 532 distinct culture environments where individual wells are completely isolated from surrounding wells. The pillar/well chip system supports a culture density similar to conventional 1536-well plates with one tenth of the working volume. The low cell population and reagent requirements combined with complete isolation of individual cultures enables interrogation of arbitrary combinations of signaling cues with low startup costs and a rapid iteration cycle.

Perhaps the greatest advantage of the two-part system is the decoupling of media preparation from ongoing cell cultures. For 1536-well plates reagent removal steps including media changes are impractical and not recommended. This limits the duration of assays in standard plate formats of comparable culture density to the pillar/well chip platform to less than 48 hours with assays less than 24 hours being optimal. Standard 96- and 384-well plates require careful plate washing procedures to avoid detaching adherent cells during media



Complementary polystyrene chips support 532 distinct microcultures. Pillar chip (left) contains an array of 14 x 38 microposts, each 750 microns in diameter, with perforations between pillars to allow for gas permeation. Well chip (center) contains a complementary array of microwells with a working volume of 850 nL per well. A pillar chip may be stamped onto a well chip (right) to form completely isolated cell culture environments.

Figure 2.1: Diagram of the pillar/well chip microculture platform

exchanges. This concern often results in sub-optimal compromises such as exchanging half the media volume to reduce shear stress on the adherent cultures. This process can be time consuming and add additional cell stress as cultures languish at room temperature out of the incubator. In the pillar/well chip system, fresh media chips are prepared at a temperature optimal for reagents without cells present. The fresh media chips are then pre-warmed in humidified chambers in the incubator, and cells are only briefly removed from optimal culture conditions to discard the used media chip and stamp the cell-carrying pillar chip onto the freshly prepared media chip. The two-part pillar/well chip system is an innovative and practical tool for high throughput cell biology.

2.2 Automated Fluid Handling Enables Controlled Preparation of Experimental Conditions

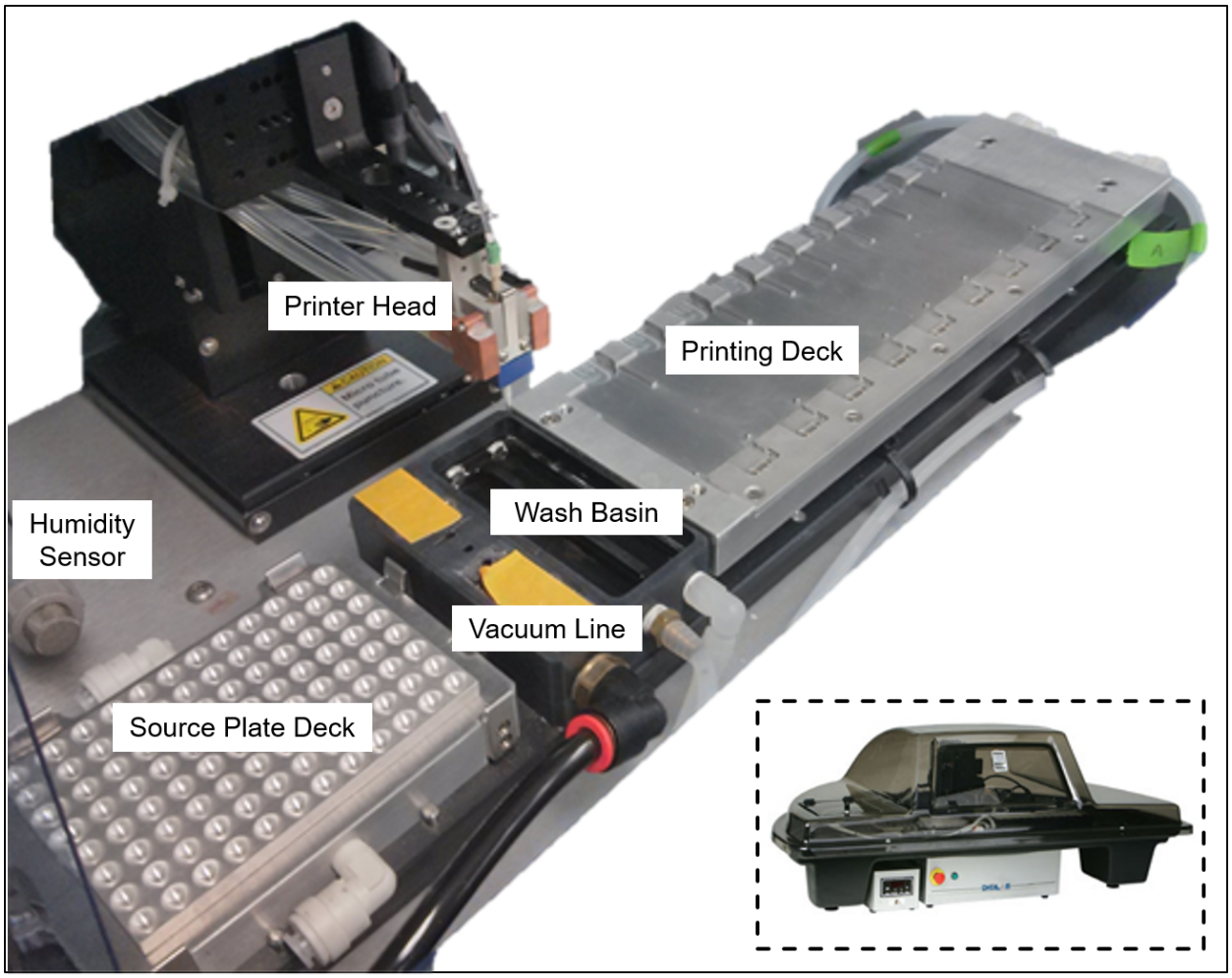
Upgrades to Apparatus

A non-contact Digilab OmniGrid Micro DNA arrayer was repurposed for cellular microarrays. The instrument uses coordinated syringe pump and solenoid valve timing to achieve rapid and precise spotting of uniform droplets. The printing process begins when the printer head aspirates a predefined volume of reagent from a 96-well plate located on the source plate deck (see Figure 2.2). The printer head consists of a ceramic tip connected in series to the solenoid valve and syringe pump (not shown). After aspiration, the instrument balances the pressure in the solenoid valve by dispensing several times, washes the ceramic tip in the wash basin to remove any residual fluid, and dries the tip in the vacuum line before printing.

The microarrayer received several upgrades to enhance compatibility with the pillar/well chip system and with temperature-sensitive biological materials. The original printing deck used vacuum suction to hold glass slides in place during printing. This was incompatible with the perforations in the pillar chips. A chip-compatible aluminum printing deck containing a water channel for temperature control was fabricated. The printing deck was fitted to a water chiller (not shown), enabling controlled cooling or heating of the printed chips. A similar aluminum source plate deck was fabricated and two copper heat exchange fittings were attached to the printer head. A second water chiller was connected to the source plate deck and printer head fittings to achieve orthogonal control over temperature of the experimental reagents before and after printing. Maintaining reagents at 4°C prevents biological hydrogel materials from gelling before printing and preserves reagents during use for more complex protocols.

Culture in Two and Three Dimensions

The pillar/well chip platform has been adapted for stem cell culture in two or three dimensions. Two dimensional culture refers to standard adherent tissue culture practices where stem cells adhere to the flat tissue culture-treated polystyrene pillars. Three dimensional cultures are those in which stem cells are embedded within a droplet of hydrogel scaffold such as Matrigel, alginate, or other biomaterials. Although three dimensional culture conditions are considered to be more representative of *in vivo* tissue conditions, quantitative analysis of three dimensional cultures is complicated by image acquisition and analysis in multiple planes. Figure 2.3 shows six evenly-spaced slices through a 150 micron depth culture of hPSCs embedded in a PEG-PNIPAAM hydrogel, stained with the Hoechst nuclear stain, and imaged using an extra long working objective on a Molecular Devices ImageXpress Micro Confocal microscope. Although the images seem to illustrate a culture with a hollow shell morphology, confocal image slices acquired using a water immersion lens (not shown) found cells distributed evenly throughout. Due to these complications and the lack of a readily



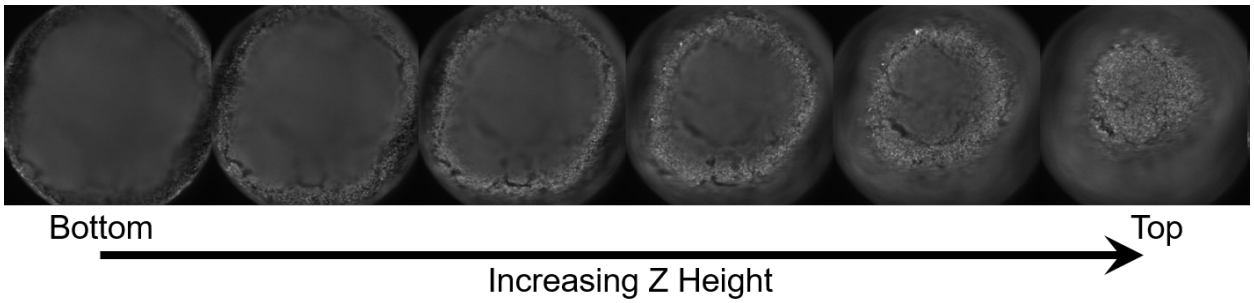
An internal photo of the OmniGrid microarrayer shows upgraded source plate deck, printer head, and printing deck. The instrument is fully enclosed (inset) and may be humidified during operation.

Figure 2.2: Reference photo of the automated fluid handler adapted for the pillar/well chip culture platform

available automated confocal microscope, a two dimensional culture format was selected for development of an *in vitro* teratogen assay.

Culture Preparation

Matrigel-coated wells and the mTeSR1 (Stemcell Technologies Inc.) cell culture medium are the gold standard for robust long term maintenance of hPSCs. Matrigel is a temperature responsive hydrogel composed primarily of common components of extracellular matrices including laminin, collagen, and entactin. The elastic modulus of Matrigel increases as



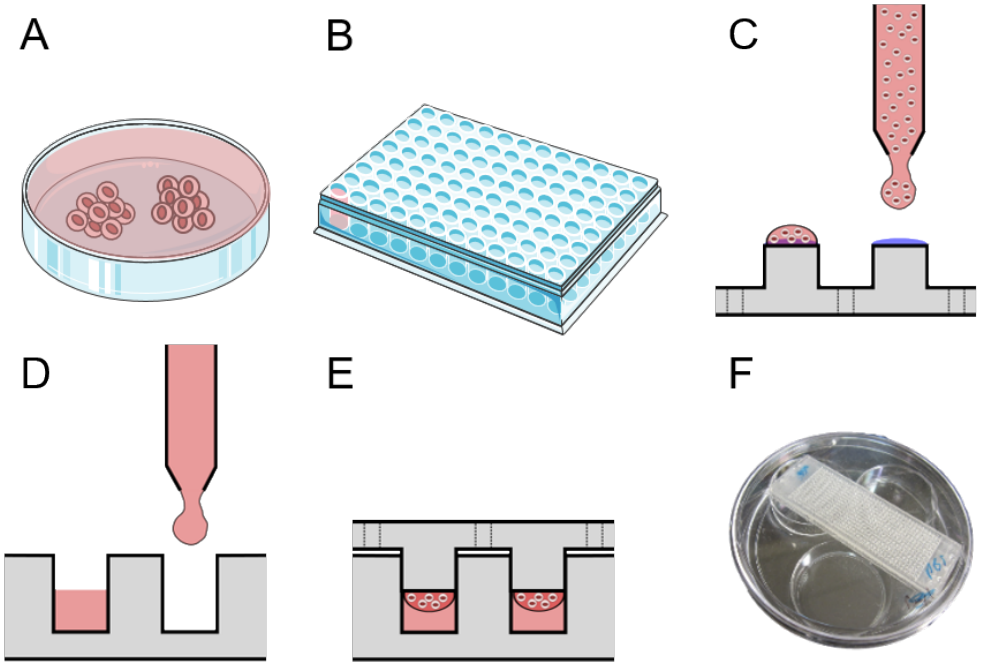
Human pluripotent stem cells cultured for 10 days in a three dimensional PEG-PNIPAAM hydrogel were stained with the Hoechst nuclear stain and imaged at six evenly-spaced Z slices. The apparent lack of cells in the center of the culture is an artifact of imaging.

Figure 2.3: Illustration of imaging challenges for stem cell cultures embedded within a hydrogel matrix

temperature increases, so it is important to maintain the material at low temperature while printing to prevent clogs from developing.

For human stem cell culture, pillar chips are first coated by printing Matrigel solution into a well chip, stamping an uncoated pillar chip onto the Matrigel chip, and incubating in a humidified chamber at 4°C overnight. Human stem cell colonies cultured in conventional 6-well plates (Figure 2.4A) are dissociated into single cells using Accutase, a gentle enzymatic dissociation reagent. Cells are centrifuged and resuspended in mTeSR1 media containing 10 μM Y-27632, a selective and potent inhibitor of Rho-associated, coiled-coil containing protein kinase (ROCK). Resuspended cells are plated in a 96 well plate (Figure 2.4B) and loaded onto the source deck. Cells suspended in media are printed in 60 nL droplets on the microposts and allowed to adhere for 25 minutes in a humidified environment at 37°C and 5% CO_2 (Figure 2.4C). Fresh media containing ROCK inhibitor is printed in 850 nL droplets into the microwells and the fresh media chips are also pre-warmed in a humidified environment in the incubator (Figure 2.4D). The two chips may be stamped together (Figure 2.4E) and placed in a humidified chamber for extended culture (Figure 2.4F). Media may be simultaneously changed for all conditions by printing a fresh well chip, lifting the pillar chip from the used media chip, and stamping the pillar chip onto the fresh media chip. The printer is automatically purged, washed and vacuum dried before and after printing each distinct reagent to avoid cross well contamination.

The microarrayer is housed in an enclosed container. The apparatus is thoroughly sterilized with 70% ethanol internally through an ethanol purge of the syringe pump system and externally by wiping surfaces with ethanol moistened paper towels before and after printing. A humidity sensor is mounted in the container, and humidity may be controlled to prevent evaporation during long printing sessions (greater than 10 minutes).



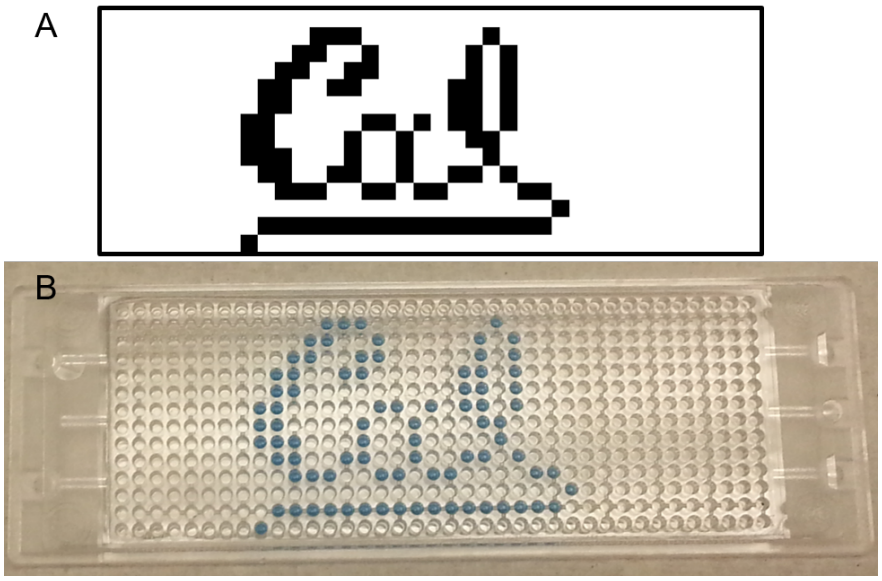
The microarray printer and two-part culture system support easy fluid handling. Human pluripotent stem cells cultured in conventional culture plates (A) are dissociated, resuspended, and added to a single well of a 96-well plate (B) for printing. Stem cells are spotted in 60 nL droplets onto pillar chips pre-coated with Matrigel (C) and allowed to settle for 25 minutes. Wells are printed with 850 nL of media (D) and pre-warmed prior to stamping (E). Cultures are maintained in humidified petri dish chambers (F) in the incubator for extended culture.

Figure 2.4: Steps in fluid handling for cell culture preparation

Arbitrary Dispensation

The microarrayer is programmable and capable of dispensing drops of precise volumes in arbitrary locations for a wide range of biological materials. The software interface (Cartesian AxSys), however, does not support modularization of protocol logic as all variables and functions are declared globally in the program environment. AxSys protocol files are preprocessed text files with specific formatting requirements. To extend the practical range of the instrument and simplify protocol development a set of Python functions and scripts was developed to compile flexibly-defined module blocks into AxSys-compatible text file protocols.

Protocols may be specified using two modes. Figure 2.5 illustrates the first mode in which a 14 by 38 pixel image is colored black in targeted wells (Figure 2.5A) and white elsewhere. Figure 2.5B demonstrates the well chip resulting from the specified positions when spotted with water soluble blue dye. The second mode of protocol specification reads



A printing protocol was generated from a 14 x 38 pixel image (A) to spot 850 nL drops of water-soluble dye in the specified wells of a well chip (B).

Figure 2.5: A custom Python-based protocol builder simplifies printing in arbitrary positions

in an arbitrary experimental design from a spreadsheet as well as printing parameters from a templated print configuration spreadsheet. The experimental design is combined with the relevant settings and encoded into text protocol files. The protocol generator may be configured to position experimental samples and replicates sequentially or fully randomize the location of individual replicates on the chip (see Appendix A.1 for example). Either setting automatically outputs an expanded copy of the experimental design spreadsheet containing the relevant metadata for aligning experimental conditions with plate location information. This allows users to generate modularized, well-validated protocols capable of arbitrarily complex dispensation for an arbitrary number of distinct experimental conditions in a matter of minutes—dramatically decreasing development time and enabling flexible experimental design and randomization.

2.3 Automated Fluorescence Imaging Supports Acquisition of Rich Biological Data

A high throughput cellular assay requires rapid and efficient methods for deriving quantitative metrics from cell cultures. The pillar/well chip platform is well-suited to the emerging paradigm of high content screening (HCS) techniques where an automated fluorescent microscope is configured to acquire fluorescence images across hundreds of unique wells.

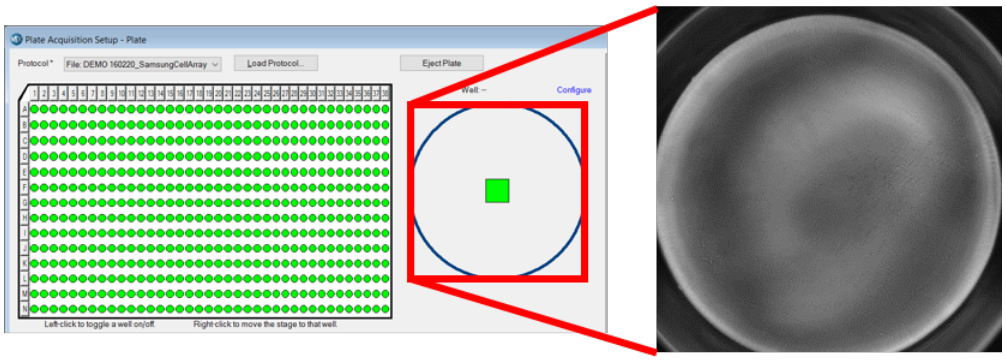


Figure 2.6: With a custom plate definition file, a single well is captured in single field-of-view

Fluorescence image data is a rich source of biological information in which changes in stain intensity and localization as well as cell phenotype, population, and distribution may be quantified by image analysis. An aluminum adapter and custom plate definition file were fabricated to adapt the pillar/well chip system to a Molecular Devices ImageXpress Micro (IXM) automated fluorescence microscope (Figure 2.6). The IXM is fitted with environmental control features and tight excitation and emission filters, enabling automated multi-wavelength imaging of hundreds of microcultures. Due to the small size of the individual wells on the pillar/well chip platform, single wells may be captured in their entirety within a single field-of-view. Image data is exported as Tagged Image Files (TIF), preprocessed, and analyzed using the Python (https://www.python.org/) and R (https://www.r-project.org/) programming languages.

Optimization of Immunocytochemistry

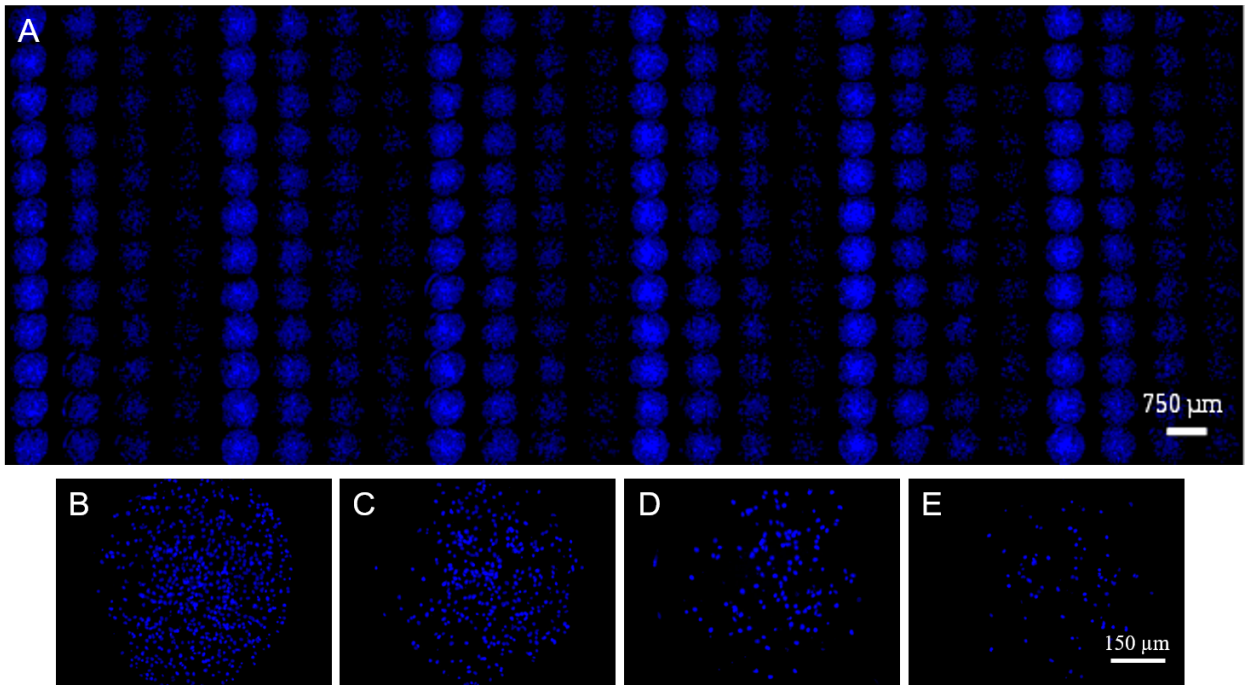
Cell biomarker staining was achieved using multiplexed immunocytochemistry. In brief, cells cultured on the pillar/well chip platform were fixed in 4% paraformaldehyde, permeabolized in 0.25% Triton X-100 to allow for stain permeation, and stained overnight with primary antibodies specific to the desired biomarkers. After washing to remove excess primary antibodies, cells were stained with fluorescently-labeled secondary antibodies specific to the animal antigens for the species in which the primary antibodies were generated. This two-step process achieves amplification of the fluorescence signal via multiple binding events for primary antibodies against a biomarker and for secondary antibodies against a specific animal antigen. A immunocytochemistry protocol was developed and optimized for stain specificity and cell culture preservation on the pillar/well chip platform. Immunocytochemistry treatment steps were applied in bulk, except where noted, by inverting pillar chips into baths of treatment solutions in individual wells of four well plates. For a full protocol, see Appendix A.1.

2.4 Demonstration of the Platform

The following critical milestones for adaptation of the microarray printer and pillar/well chip platform to human pluripotent stem cell culture were met.

Control of Cell Seeding

Initial cell population can have a profound effect on hPSC viability, proliferation, and differentiation outcomes. Thus a platform for quantitative hPSC assays must demonstrate uniform cell seeding and control over seeding density. Figure 2.7A illustrates uniform and tunable control over cell seeding with columns of cultures seeded at various densities.



Montage (A) shows uniformity of cell seeding across columns with 400 (B), 200 (C), 100 (D), and 50 (E) cells seeded per culture.

Figure 2.7: Cell seeding is well-controlled

Maintenance of Cell Viability

Once seeded, cells on the pillar/well chip platform must remain viable. Figure 2.8 shows rat neural progenitor cells and human pluripotent stem cells remain viable after culture in maintenance media for three days on the pillar/well chip platform.

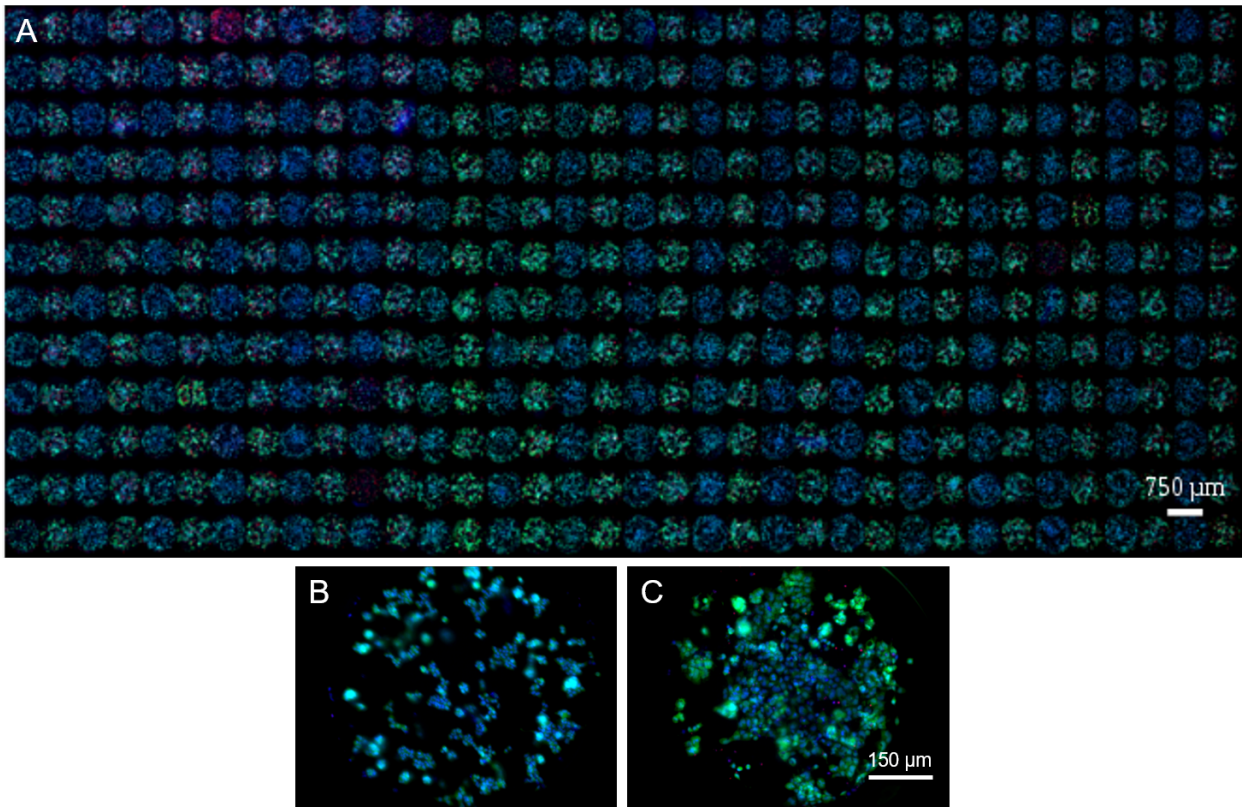
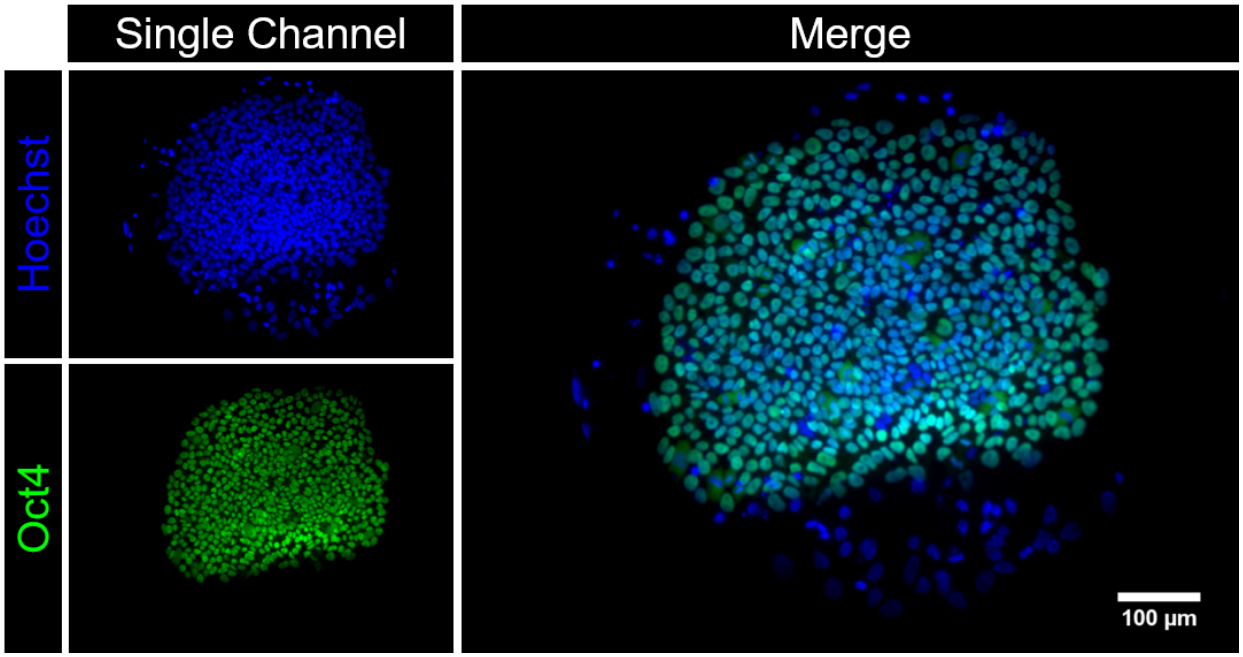


Plate Montage of live/dead assay (A) shows alternating columns of viable rat neural progenitor cells (B) and human pluripotent stem cells (C) after three days in a 3D Matrigel culture. Cells were stained with the live/dead reagents calcein AM (green) and ethidium homodimer (red), respectively, as well as the Hoechst nuclear counterstain (blue).

Figure 2.8: Stem cells remain viable after seeding on chip

Maintenance of Pluripotency on Chip

Human pluripotent stem cells seeded on the pillar/well chip platform must retain pluripotency initially and over time and should exhibit the hallmark characteristics of a healthy pluripotent stem cell colony. Figure 2.9 illustrates a representative example of hPSCs stained for the pluripotentcy marker Oct4 after four days of culture on the pillar/well chip platform. The colony exhibits the expected morphology of closely-packed Oct4-positive pluripotent stem cells commonly observed in conventional 6-well plate cultures. Note that Oct4 is localized to the nucleus and stains only the cells in the central colony.



Human pluripotent stem cells retain expression of the nuclearly localized pluripotency marker Oct4 after four days in a 2D culture on chip. Colonies exhibit the expected densely-packed morphology typical of hPSC culture on polystyrene surfaces.

Figure 2.9: Stem cells retain hallmarks of pluripotency on chip

Chapter 3

The Differentiation Assay

3.1 Redesign of the Embryonic Stem Cell Test

The Principles of Teratology

Teratogenic toxicity is characterized by the "Principles of Teratology" [2] which may be summarized as follows:

1. Developmental stages vary in sensitivity to adverse environmental stimuli.
2. Maternal factors including route and degree of exposure, rate of systemic absorption, metabolic activity, and rate of placental transfer govern access of environmental stimuli to embryonic tissues.
3. Teratogens are characteristically manifest in disruptions of development including death, malformation, growth retardation, and functional defects.
4. Teratogenic activity is dose-dependent and increases in frequency as dosage increases from the No Observable Effect Level (NOEL) up to a dose producing 100% lethality.
5. Teratogens act in specific ways on developing cells and tissues to disrupt developmental events.

Design Objectives of a Developmental Toxicology Assay

Taking into consideration the Principles of Teratology, a cell-based developmental toxicology assay should meet the following design objectives:

1. The developmental stages of maximum sensitivity to adverse environmental stimuli should be represented.
2. Maternal factors including route and degree of exposure, rate of systemic absorption, metabolic activity, and rate of placental transfer should be eliminated as confounding factors if they cannot be modeled accurately.
3. Endpoints should capture cell fate decisions leading to disruptions in development including death, malformation, growth retardation, and functional defects.

4. Dose-dependent activity should be assessed and used as a basis of comparison between substances.
5. The assay should be founded on a central biological hypothesis which reasonably explains the principles that govern developmental defects.

Lessons from the Validated Embryonic Stem Cell Test

The validated Embryonic Stem Cell Test (EST) developed a differentiation endpoint suitable for proof-of-principle assessment when the test was proposed in the 90's due to the ease of counting beating cardiomyocyte-containing colonies on a standard transmitted light microscope. Although the existing test and improvements based on molecular endpoints [31] for cardiac fate have shown promising results, the diversity of possible cell fate outcomes is not well represented by heart tissue. Alternative tests targeting other cell fates including neurons, endothelial cells, chondrocytes and osteoblasts [32] have been developed [29] and multiple endpoint tests based on these fates [33] and other differentiation-related biomarkers [34, 35] have been proposed. Preliminary evidence from these studies suggests predictive value in a wide range of differentiation-related biomarkers for classifying the teratogenic risk of control molecules.

Unifying Variations of the Embryonic Stem Cell Test

Results from variations of the EST suggest that many biomarkers related to differentiation may be used to test for teratogenic activity. To consider whether this outcome may be reasonably expected, it may be useful to review the underlying process of fate specification during development. Fate specification during embryogenesis is canonically described as a restriction of potency (or potential cell fates) from the *totipotent* fertilized egg capable of forming all tissues including extra-embryonic tissues to the *pluripotent* embryonic stem cells to *multipotent* progenitors of the three germ layer lineages: ectoderm (e.g. neural tissue), mesoderm (e.g. musculature), and endoderm (e.g. gastrointenstinal tract). The endpoints assessed in previously described variants of the EST largely include markers of terminally differentiated cell types. From teratology we know that sensitivity of the embryo to environmental insults decreases as the fetus reaches full term.[2]

Human Embryonic and Induced Pluripotent Stem Cells

Since the first report of embryonic stem cells (ESCs) derived from humans in 1998 [36], ESCs have held tremendous promise for applications developmental and cell biology, drug discovery, and medicine. In 2006, Kazutoshi Takahashi and Shinya Yamanaka developed a new technology [37] to convert mouse fibroblasts and other somatic cells into induced pluripotent stem cells (iPSCs)—embryonic-like cells exhibiting the hallmark properties and gene expression profiles characteristic of embryonic stem cells. The technology was rapidly adapted to human cell lines in 2007. [38–40] Human iPSCs may be derived from individuals

of virtually any genotype, expanded indefinitely in culture, and differentiated into all the tissues of the body. As a result of these desirable properties, a multitude of patient and disease-specific hiPS cell lines have been derived and adopted into common practice alongside hESCs. [41] These cell lines may be collectively referred to as human pluripotent stem cells (hPSCs).

Hypothesis and Proposed Assay

The following hypothesis is proposed:

A human pluripotent stem cell-based differentiation assay which assesses the earliest developmentally sensitive fate decisions will be highly predictive of the teratogenic risk to diverse tissues in the human embryo should a given substance be presented after maternal effects are accounted for.

To test this hypothesis, the following assay was proposed:

1. Developmental effects of a substance on all tissues will be assessed by measuring changes in cell population and biomarker expression during the critical developmental stage of germ layer specification.
2. Maternal factors will not be assessed and are beyond the scope of the assay.
3. Disruptions in development will be assessed as differences in biomarker expression from an untreated control.
4. Dose-dependent activity will be assessed at five dose levels and used as a basis of comparison between substances.

In addition, the assay will estimate sensitivity differences for different cell lines using one human embryonic stem cell line (H1) and one human induced pluripotent stem cell line (TCTF).

Outcomes Assessed

The proposed assay used differential gene expression of early markers of differentiation to assess alterations in human induced pluripotent stem cell fate outcomes and classify teratogenic risk. By assessing markers associated with pluripotency and each of the three germ layers, the proposed test aimed to capture the full diversity of cell fate outcomes. For these outcomes to be interpreted in terms of teratogenic risk to humans, it is essential that these experiments be conducted in human cell lines to eliminate confounding effects due to species differences.

Experimental endpoints for the proposed test include a general nuclear counterstain (Hoechst 33342) and immunofluorescence stains for expression of key protein biomarkers indicative of the four expected differentiation states: undifferentiated pluripotent stem cells,

ectoderm lineage progenitors, mesoderm lineage progenitors, and definitive endoderm lineage progenitors.

Differentiation responses were assessed using multiplexed antibody staining. All conditions were stained for the pluripotent cell marker octamer-binding transcription factor 4 (Oct4). Germ layer markers paired box 6 (PAX6), brachyury, and GATA-binding protein 4 (GATA4) for ectoderm [42], mesoderm [43], and definitive endoderm [44], respectively, were each represented in one third of the replicates for each experimental condition. Multiplexed antibody staining was enabled by selection of two primary antibodies developed in different species such that two secondary antibodies could be used to fluorescently label distinct species-specific antigens on the primary antibodies.

Stain	**Fate Assessed**	**Fluorescence Channel**	**Total Well Count**		
			TCTF Day 5	**TCTF Day 10**	**H1 Day 10**
Hoechst	Nuclear stain	DAPI	9044	6384	6384
Oct4	Pluripotency	FITC	9044	6384	6384
PAX6	Early ectoderm	TxRed	2660	2128	2128
Brachyury	Early mesoderm	TxRed	3192	2128	2128
GATA4	Definitive endoderm	TxRed	3192	2128	2128

Table 3.1: Biomarkers and staining

Assay Duration

As pluripotent stem cells differentiate, expression of the pluripotency marker Oct4 will be lost and early biomarkers of germ layer specification should begin to be expressed. The specific timing of initiation and duration of biomarker expression varies, thus differentiation was assessed at two timepoints expected to fall within the window of gene expression. A day 5 endpoint was selected for onset of mesoderm lineage biomarker Brachyury, and a day 10 endpoint was selected for onset of ectoderm biomarker PAX6 and definitive endoderm biomarker GATA4.

3.2 Experimental Design

Replication Strategy

The pillar/well chip culture system supports up to 532 distinct cell cultures per plate. A nested experimental design was selected [45, 46] in which all experimental conditions could be included within a single plate and replicated at least three times at the following levels: technical replicates within plates, whole plate replicates, and replicates spaced in time. [47]

Table 3.2 quantitatively summarizes the preliminary data sets acquired using this design, and Figure 3.1 provides a visual overview.

Experimental Design Terminology:

Batches: Complete sets of experimental replicates separated in time and considered biologically independent.

Plates: A single complete set of experimental samples stained with one out of three germ layer stains.

Wells or Technical Replicates: The smallest observed experimental unit stained with Hoechst nuclear stain, Oct4, and one of the three germ layer stains.

Cell Line	Endpoint Day	Total Batches	Total Plates	Stain-Specific Plates			Total Wells
				PAX6	Brachyury	GATA4	
TCTF hiPSC	5	6	17	5	6	6	9044
TCTF hiPSC	10	3	12	4	4	4	6384
H1 hESC	10	3	12	4	4	4	6384

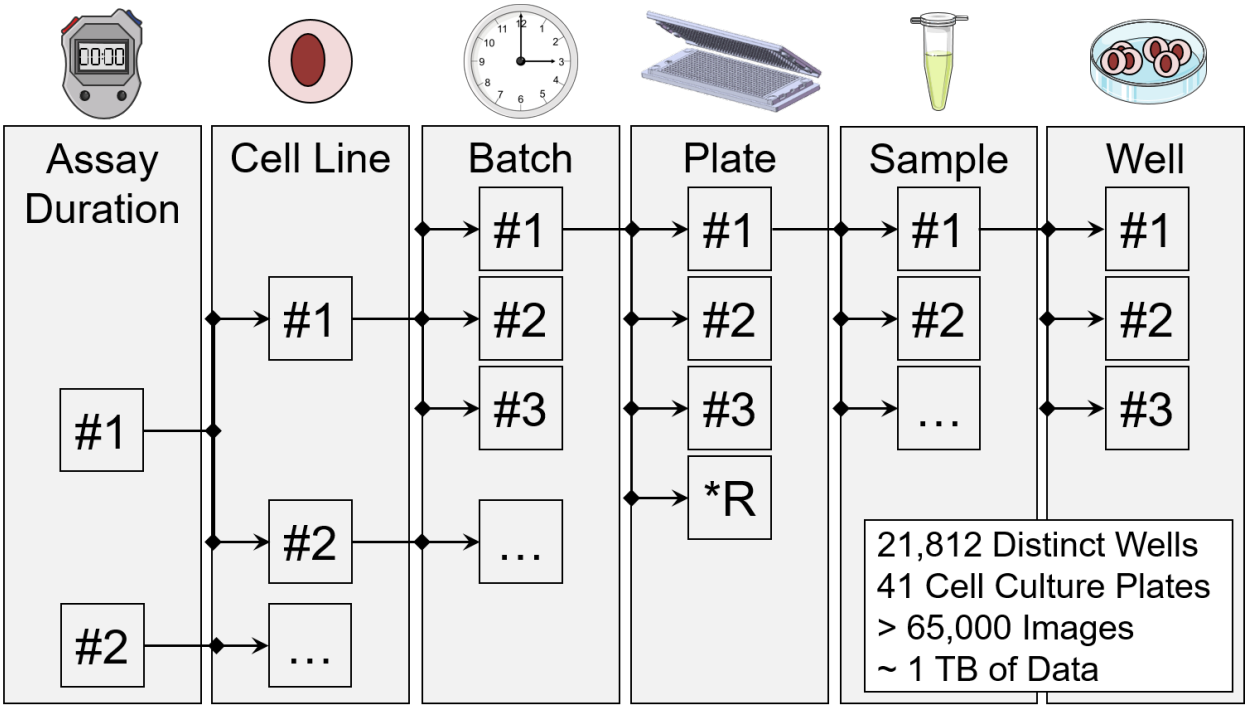
Day 5 endpoint TCTF data was collected from experiments run in collaboration with Dr. Sean McFarland.

Table 3.2: Overview of experimental replicates

Reduction of Confounding Variables

A set of proof-of-principle experiments were designed to test the hypothesis above using the proposed assay while controlling for major known sources of variation and reducing systematic trends due to uncontrolled experimental processes. This systematic approach to planning the experimental structure of replicates is commonly referred to as "blocking" and is an important part of controlling for variation in statistical analyses. [48–50] Each plate contained a full set of experimental treatments and controls with three technical replicates for each unique condition. The position of individual technical replicates within a given plate was randomized to eliminate systematic effects due to plate position. Randomized plate layouts were generated for each batch, but all plates within a given batch were prepared using an identical layout to simplify preparation of fresh media chips. Each batch contained

three to four plates prepared with the same reagents and cells, thus there were a total of 9 to 12 technical replicates for each unique condition per batch. Fresh media chips were printed daily, and systematic effects due to printing order were eliminated by randomly assigning pillar chips to fresh media chips each day. All plates were stained with the nuclear counterstain Hoechst and antibodies against the pluripotency marker Oct4 allowing for direct comparison between plates in terms of cell seeding and maintenance of pluripotency. Each plate was stained for only one of the three differentiation biomarkers, and, where available, a fourth chip was used to completely replicate one of the three differentiation endpoints such that plate-to-plate variation could be assessed for all biomarkers.



This overview shows the highly nested structure of the experimental design. The full compound library (Table 3.4) was replicated on each plate. The symbol (*R) indicates an optional duplicate of one germ layer stain plate (plates 1-3). This duplicate was included in the day 10 endpoint experiments to allow for estimation of plate-to-plate variation of germ layer stains.

Figure 3.1: Visual summary of nested experimental design

Selection of Controls

Selection of appropriate controls (see Table 3.3) was a critical step in assay development since all measured endpoints of the proposed assay were defined relative to the untreated control. The validated EST relied on withdrawal of leukemia inhibitory factor (LIF), an important growth factor for maintaining pluripotency in mouse embryonic stem cells, to

initiate spontaneous differentiation. For the proposed assay, a similar method was used in which fibroblast growth factor 2 (FGF-2), the relevant growth factor for maintenance of pluripotency in human embryonic stem cells, and other supplemental growth factors were withdrawn. A commercially available growth-factor free media, STEMdiff APEL (Stemcell Technologies, Inc.), was selected as the basal medium and untreated control. This medium minimally supports cell viability without providing specific developmental cues and has been used as a basis for several commercially available media formulations for differentiation to the three germ layer endpoints of the proposed assay.

Table 3.3 lists the full set of controls for the proof-of-principle screens. In addition to the untreated APEL control, solvent controls for DMSO and ethanol were included to control for preparation of substances tested. An unstained, untreated control was included for a negative (no signal) control, and a no primary antibody control was included for verifying specificity of the secondary antibody stains. Positive controls were included for each measured endpoint. For pluripotency (Oct4) and cell population (Hoechst) the standard mTeSR1 stem cell maintenance media was included. The ectoderm lineage positive control was based on an efficient differentiation protocol [51], and the mesoderm and endoderm lineage positive controls were commercially available APEL-based differentiation media. Finally, a single well per plate was designated as an image registration control in which multicolored fluorescent beads were printed in a buffer solution. The beads fluoresce in multiple channels and may be useful to confirm the alignment of image frames between exposures.

Control	**Media**	**TCTF Day 5**	**TCTF Day 10**	**H1 Day 10**
Untreated	APEL	yes	yes	yes
Solvent 1	DMSO in APEL	yes	yes	yes
Solvent 2	Ethanol in APEL	yes	yes	yes
Negative, no signal	APEL, no stain	—	yes	yes
Stain specificity	APEL, no primary antibody	—	yes	yes
Positive, pluripotent + cell viability	mTeSR1	yes	yes	yes
Positive, ectoderm	Dual smad inhibitors in APEL	yes	yes	yes
Positive, mesoderm	Mesoderm Induction Media	yes	yes	yes
Positive, endoderm	Definitive Endoderm Induction Media	yes	yes	yes
Image registration	Multicolored beads in PBS	—	yes	yes

Table 3.3: Experimental controls included

Library of Compounds

A library of 31 drugs, chemicals, and nutrients was selected. Molecules in the library exhibit a variety of teratogenic activities in humans and in animals where human data is lacking. Since definitive human data on teratogens is limited, molecules in the library also exhibit various levels of evidence to support their teratogenic risk in humans.

Label Key for Library of Compounds (Table 3.4):

Embryonic Stem Cell Test (EST) Labels: [28]
non: Non-teratogenic in humans or animals
weak: Weakly teratogenic in humans or animals
strong: Strongly teratogenic in humans or animals

Human Evidence Levels: [52, 53]
well-tolerated: Well-tolerated in human pregnancies
risk indicated: Evidence suggests risk in humans or animals
contraindicated: Contraindicated in human pregnancies

Molecule	Description	EST Label	Human Evidence
5-Bromo-2'-deoxyuridine	thymidine analog	strong	risk suggested
6-Aminonicotinamide	vitamin B3 inhibitor	strong	risk suggested
Ascorbic Acid	vitamin C	non	well-tolerated
Asparagine	amino acid	—	—
Aspartic Acid	amino acid	—	—
Boric Acid	antiseptic	weak	risk suggested
Cadmium Sulfate	fungicide	—	risk suggested
Copper (II) Sulfate	potent emetic	—	risk suggested
D-Camphor	antiseptic	non	well-tolerated
Dimethadione	anticonvulsant	weak	risk suggested
Dimethyl Phthalate	insect repellent	—	—
Glutamine	amino acid	—	well-tolerated
Glycine	amino acid	—	well-tolerated
Histidine	amino acid	—	—
Hydroxyurea	antineoplastic	strong	risk suggested
Iron (III) Chloride	solids separation agent	—	—
Isoproterenol	analog of epinephrine	—	risk suggested
Leucine	amino acid	—	—
Lithium Chloride	mood stabilizer	weak	risk suggested
Methotrexate	antineoplastic	strong	contraindicated
Methoxyacetic Acid	weak acid	weak	risk indicated
Penicillin G, sodium salt	antibiotic	non	well-tolerated
Riboflavin	vitamin B2	—	well-tolerated
All-trans-Retinoic Acid	vitamin A	strong	contraindicated
Saccharin sodium hydrate	non-nutritive sweetener	non	well-tolerated
Salicylic Acid sodium salt	anti-inflammatory agent	weak	risk indicated
Sodium Chloride	table salt	non	well-tolerated
Thalidomide	mild sedative	strong	contraindicated
Tyrosine	amino acid	—	risk indicated
Valproic Acid	anticonvulsant	weak	risk indicated
Zinc Sulfate	astringent	—	risk indicated

Table 3.4: Library of compounds screened in proof-of-principle experiments

Chapter 4

Quantitative Image Analysis

4.1 The Emerging Role of Imaging in Toxicology

High content screening (HCS) is an emerging paradigm in which advances in automated microscopy and supporting software enable acquisition, storage, and analysis of images for multiple fluorescent readouts across hundreds of unique cell cultures. This technique is a type of phenotypic screening in which spatially and temporally defined changes in cell morphology or composition (e.g. protein expression) may be observed and quantified via image analysis. Although HCS methods have been largely relegated to a supporting role in favor of fluorescence-activated cell sorting (FACS) in high throughput multi-well plate screens, HCS techniques are gaining increasing popularity in drug discovery and toxicology.

In 2007 the U.S. National Academy of Sciences released *Toxicity Testing in the 21st Century: A Vision and a Strategy*, a report calling for increased efficiency and decreased animal usage in toxicology by transitioning to highly automated *in vitro* toxicology assays based on human cell lines. [23] This report represents a major shift in the way toxicity testing is conducted. Animal studies reveal overt signs of adverse health effects, but extrapolation of the observed outcomes is controversial and little knowledge about the specific biological changes leading to adverse effects is gained. The proposed alternative couples cell-based assays for specific mechanisms of toxicological activity with chemical and bioinformatic predictions based on human health data. The Tox21 pilot program, conceived in 2008, is currently testing 10,000 chemicals in about 50 quantitative high-throughput screening assays using a robotic screening system, 1536-well plates, and high content imaging. [54] A priority of the Tox21 Toxicology Roadmap is to incorporate advances in image analysis and informatics into human cell line-based toxicology assays. The proposed human pluripotent stem cell teratogen assay was engineered to integrate biological advances in developmental toxicology with the emerging technologies prioritized by the Toxicology Roadmap.

Objectives of Image Analysis for the Proposed Assay

The proposed differentiation assay was designed to use nuclearly localized biomarkers (see Table 3.1). When a protein biomarker is expressed the fluorescent probes for that biomarker will localize to the Hoechst nuclear counterstain. Figure 2.9 illustrates localization of the pluripotency marker Oct4 to the nucleus and shows Oct4-positive pluripotent stem cells in the central colony as well as Oct4-negative differentiating progenitor cells in the periphery. As pluripotent stem cells begin to differentiate, Oct4 expression will be lost and expression of one of the three germ layer markers will be gained. In order to use changes in protein expression to predict teratogenic risk, images must be processed to yield quantitative metrics of changes in protein expression. Proliferation and differentiation are tightly interlinked in stem cell populations, therefore quantification of changes in cell population are desired in addition to population-independent metrics of protein expression.

4.2 Colocalization: Analysis of Multiple Fluorescent Stains

The term "colocalization" refers to the observation of spatial overlap or correlation between stains in fluorescence image data. High content analysis is generally concerned with quantifying the colocalization of fluorescently-labeled markers such that conclusions may be drawn from changes in cell morphology, composition, or subcellular localization of fluorescent probes. Although colocalization analyses are among the most common methods applied to fluorescence image data, no consensus approach has emerged from more than three decades of advances.

Cell Counting Methods

Cell counting methods are among the most popular colocalization techniques due to their intuitive interpretation. For small scale studies, fluorescent stains are often overlaid in false color images as shown in Figure 2.9 and the relationship between stains is quantified by manually counting the number of cells stained in both channels and normalizing by the total cell count. This analysis yields a total cell count and a percentage of positively stained cells where each cell is assigned a binary response of either fully co-stained or not. Automated image analysis techniques often mimic the manual approach by segmenting single cell objects within images and applying some filter to return a percentage of positively stained cell objects. The binary response of cell counting methods is unsatisfactory for studies in which stains within a single cell or single region of interest (ROI) must be characterized or for studies in which partial co-staining may be expected. As a result of these and other concerns biologists have invested considerable effort into defining quantitative colocalization statistics.

First Author	Year	Quantitative Methods Introduced	Reference
Manders	1992	Pearson R	[55]
Manders	1993	Manders Overlap Coefficient, Overlap Coefficients k1 & k2, Overlap Coefficients M1 & M2	[56]
Barbarese	1995	Correlation Index, Correlation Distance	[57]
Jiang	1998	Fractional Area Overlap	[58]
Silver	2000	Colocalization Index	[59]
Kreft	2004	Fractional Area Overlap Ratios	[60]
Li	2004	Intensity Correlation Quotient	[61]
Costes	2004	Iterative Correlation Threshold, Modified Manders M1 & M2, Single Channel Block Shuffling Significance Test	[62]
Jaskolski	2005	Modified Li Intensity Correlation Quotient	[63]
Adler	2008	Replicate Based Noise Corrected Correlation	[64]
Wu	2010	Protein-Protein Proximity Index	[65]
Dunn	2011	Dual Channel Block Shuffling Significance Test	[66]
Zinchuk	2011	Modified Wu Protein-Protein Proximity Index	[67]
Singan	2011	Rank-Based Intensity Weighted Manders M1 & M2	[68]
Herce	2013	Herce H Coefficient	[69]
Adler	2014	Adler H Coefficient	[70]
Moser	2016	Normalization by Percentage of Colocalized Objects	[71]

Table 4.1: Overview of quantitative colocalization methods proposed in literature

Quantitative Colocalization Statistics

Quantitative colocalization statistics were introduced by Manders *et al.* in 1992 [55] with the first application of the Pearson correlation coefficient (Pearson R) to measure the linear correlation between stain intensities in pairs of pixels from two fluorescence images. In subsequent years numerous alternative or supporting methods have been proposed (see Table

4.1) to address perceived limitations of previously defined methods. While most methods have been compared to the Pearson correlation coefficient on limited simulated or control biological images, no comprehensive comparison of the available methods exists to guide the selection of appropriate statistics.

4.3 Applying Quantitative Colocalization to the Differentiation Assay

Challenges in Colocalization of Stains in Human Pluripotent Stem Cells

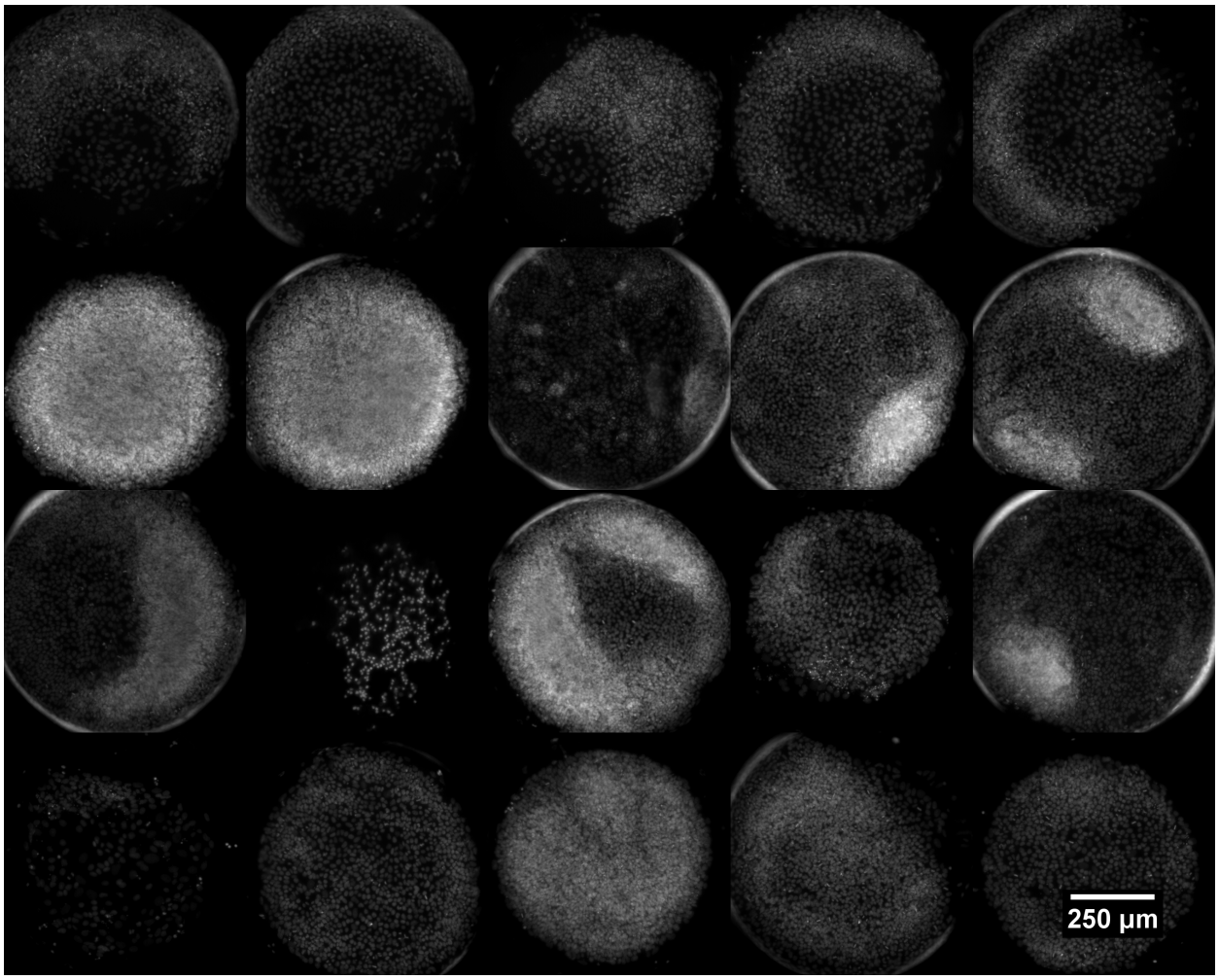
Automated cell counting methods deteriorate when the segmentation of single cell objects becomes unreliable. It is particularly challenging to segment individual nuclei for differentiating human pluripotent stem cell cultures for the following reasons:

- Pluripotent cell colonies are densely packed with little or no space between nuclei to support edge detection.
- Size and shape of nuclei may vary dramatically during differentiation ranging from small and compact in pluripotent stem cells to large and flattened in progenitors making size-based heuristic methods prone to over or under counting of potentially interesting subpopulations.
- Biomarker stain intensity is nonuniform for the heterogeneous cell populations arising during differentiation.
- The intensity of stains is a function of the size and shape of the nucleus.

Figure 4.1 shows example images from one plate of TCTF day 10 which highlight some of the challenges associated with image analysis. Although a variety of software packages and segmentation algorithms were tested on image data from the proof-of-principle experiments, none were found to produce reliable segmentation of single cell nuclei or even consistent over- or under-segmentation. To pursue quantification of cell population and colocalization, simple approximations of these metrics were sought.

Defining Standards of Comparison for Selection of Colocalization Statistics

Four fluorescence image data sets were previously acquired to compare growth and maintenance of the Oct4 pluripotency marker in four human pluripotent stem cell lines after culture on the pillar/well chip platform for several days. Images from these control data sets were randomly sampled and used to develop a heuristic-based segmentation algorithm capable of reliably segmenting single cell nuclei in control, low density cultures. The algorithm consisted of the following logical steps:



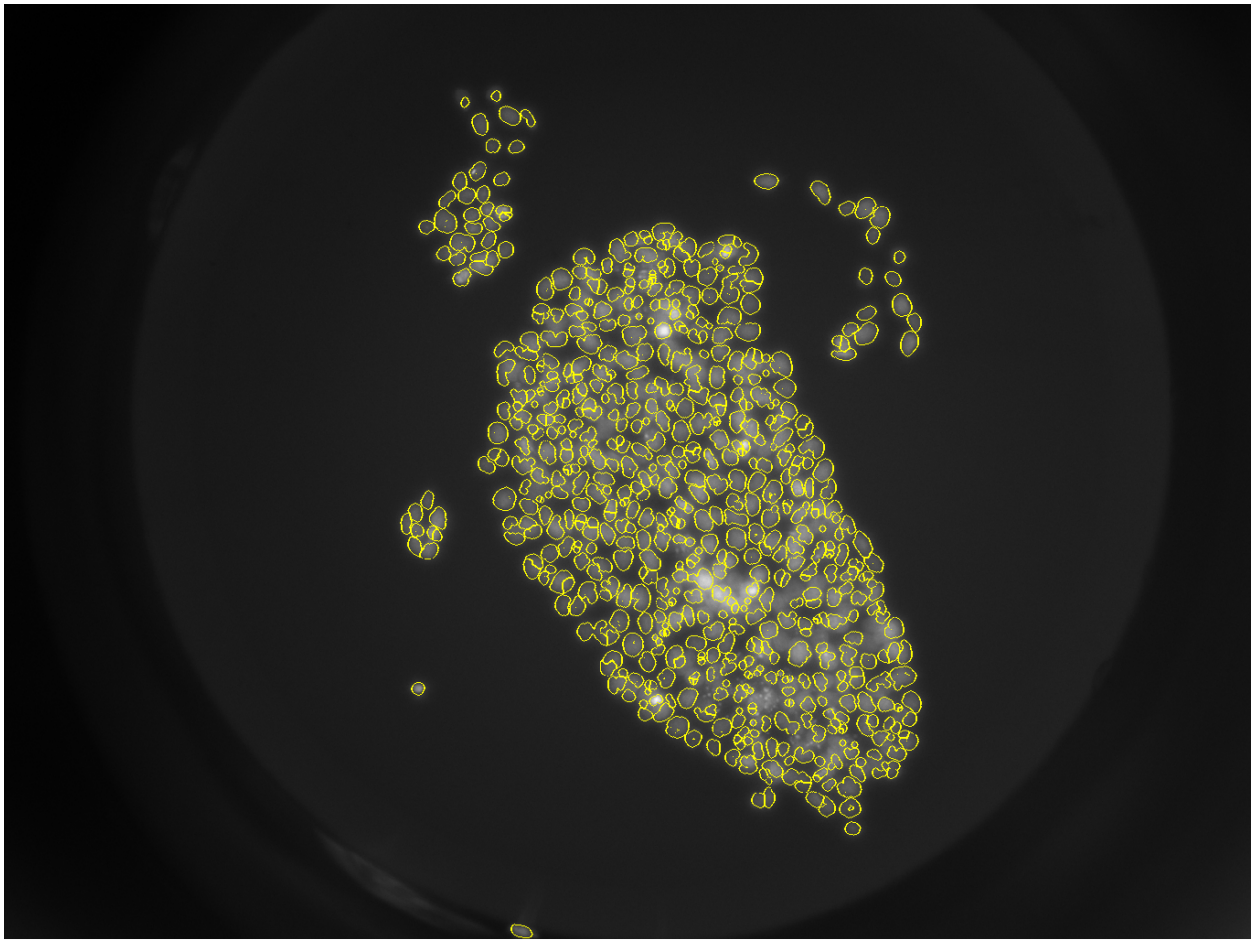
Sample Hoechst nuclear stain grayscale images from wells on a single plate in the TCTF day 10 endpoint data set illustrate some of the challenges faced in image analysis. Size, shape, and stain intensity of single nuclei varies considerably based on local cell density. Single cell segmentation is not feasible in many cases and manual cell counting is challenging. Intensity-based foreground segmentation often preferentially segments bright subpopulations or punctile spots of cellular debris leading to undesirable biases in analysis.

Figure 4.1: Sample wells from the the proof-of-principle developmental toxicology data set show various culture morphologies

1. Illumination correction to flatten out background intensity across the image
2. Adaptive contrast stretching
3. Otsu threshold to separate foreground from background
4. Adaptive threshold to separate clusters of cell objects from within foreground
5. Watershed labeling to separate individual cells from within connected clusters

6. Region Adjacency Graph merging to correct for over segmentation (fracturing of single nuclei) due to watershed algorithm

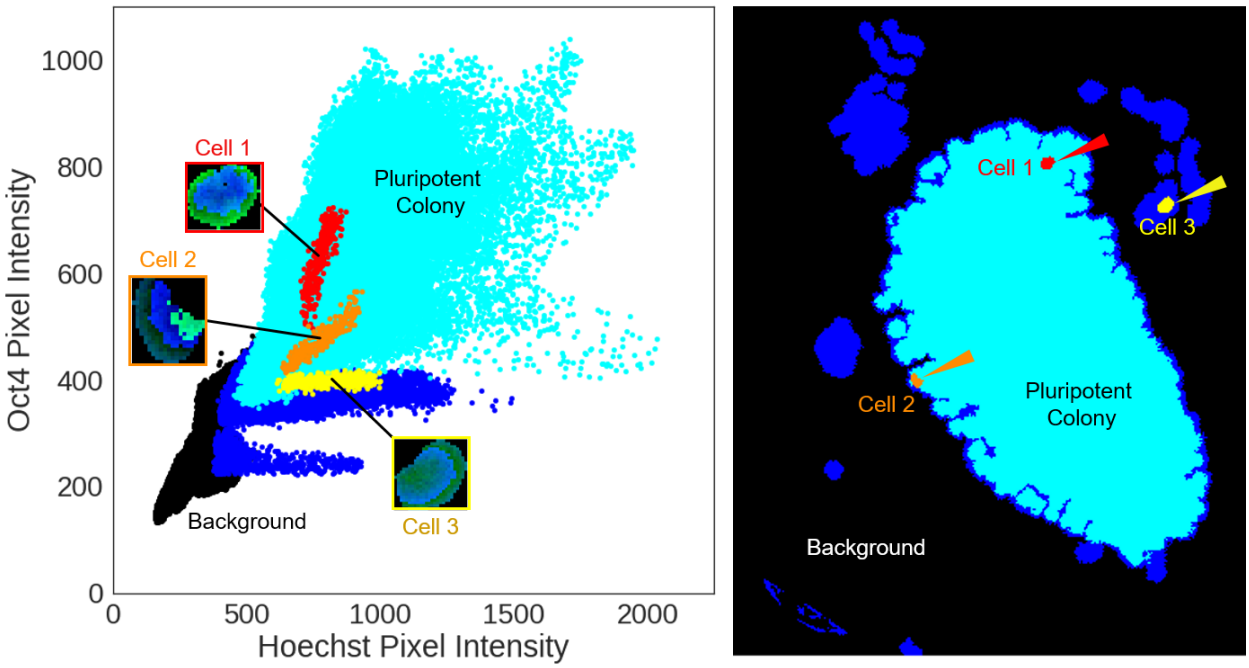
Figure 4.2 illustrates segmentation using this algorithm on an example image.



A single-cell segmentation algorithm was developed using the Python programming language.

Figure 4.2: Single-cell segmentation of Hoechst nuclear stain in a control low-density culture of pluripotent stem cells

Although the algorithm was consistent in segmenting low density control cultures of pluripotent stem cells, the method was complex and poorly suited to the proof-of-principle teratogen screen data. Therefore it was used as a basis of comparison for selecting more robust and efficient image processing methods.



Individual scatter points represent a single pixel position plotted as a function of stain intensity for two compared images. Points in the scattergram (left) are color-coded and matched to the colored ROIs in the segmentation mask (right). The colors are as follows: Black is background pixels. Blue is for pixels in which only the Hoechst nuclear counterstain was observed. Cyan is for the foreground area in which both the counterstain and Oct4 were observed. Red, orange, and yellow represent single cell objects segmented as illustrated in Figure 4.2. The inset images correspond to the false-color merge images of the three single cells shown. As is typical of hPSC cultures, cells exhibit different levels of Oct4 expression depending on their position relative to the densely packed colonies. Cells within the colony (cell 1, red) stain more strongly for Oct4 than cells on the edge (cell 2, orange), and cells in the periphery (cell 3, yellow) rapidly lose Oct4 expression and are primed for differentiation.

Figure 4.3: Fluorescence image data is a hierarchical data format

Lessons from Single Cell Colocalization Results on Control Data

Regions of interest within fluorescence image data may be defined at multiple "levels" of the image hierarchy from whole images, to foreground only, to single cell objects or subcellular compartments. A commonly recommended best practice for qualitative analysis of colocalization in fluorescence image data is plotting a scatter plot (called a "scattergram" or "fluorogram" [72]) in which pixels are plotted as a function of image intensity for a pair of fluorescence images. Figure 4.3 illustrates the scattergram for the nuclear and Oct4 stains of the image shown in Figure 4.2.

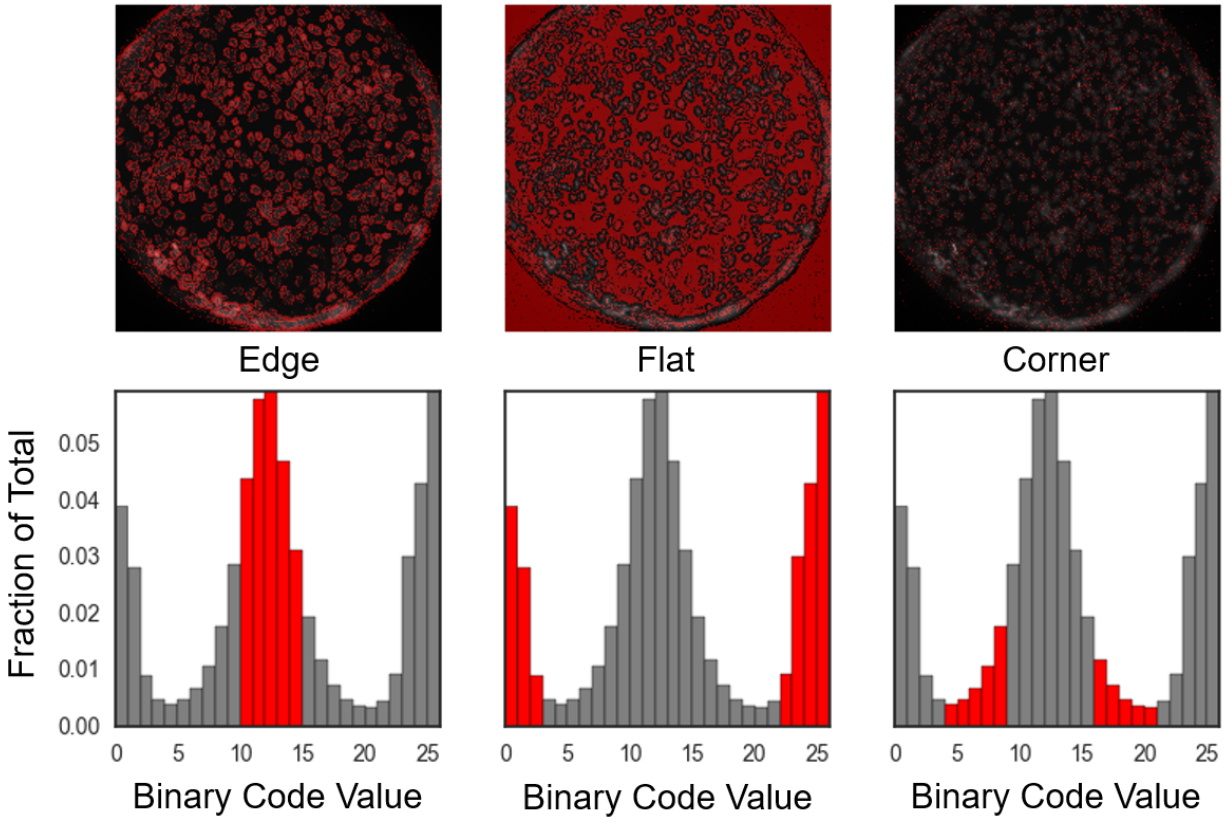
To illustrate the effect that the ROI has on the Pearson correlation statistic, Pearson R

values were calculated for the various subregions highlighted in Figure 4.3.

Region	Color	Pearson R
Whole Image	all	0.925
Background	black	0.918
Foreground (Background Pixels Set to 0)	all	0.985
Foreground Pixels Only	blue + cyan	0.793
Peripheral Cells	blue	0.190
Pluripotent Colony	cyan	0.745
Cell 1	red	0.836
Cell 2	orange	0.931
Cell 3	yellow	0.263
Population Mean of Single Cell Values	—	**0.667**

Table 4.2: Pearson R values calculated for subregions highlighted in Figure 4.3

Application of the analysis shown in Table 4.2 to multiple images revealed that Pearson R values for regions containing multiple cells are insensitive to changes in expression levels and sensitivity is recovered by considering individual cells separately (data not shown). These results were corroborated most recently in a study by Moser *et al.* [71] in which the percentage of positively-stained cells was multiplied by the whole image Pearson R as a correction factor to address the same limitation. While the Moser correction method may show an increased dynamic range, no algorithmic simplification is achieved and the biological interpretation is unclear. Instead, the ideal analysis is one in which a continuous quantitative colocalization statistic may be calculated for individual cells and summarized at the population level by, for example, calculating the arithmetic mean. Such an analysis combines the intuitive interpretation of cell counting methods with the dynamic range of quantitative colocalization statistics. Thus the gold standard metrics for measuring cell population and stain colocalization were defined as the total cell count and the cell population mean Pearson R.



A Local Binary Pattern algorithm with a radius of three pixels and 24 comparison points was used to calculate binary texture features.

Figure 4.4: Local Binary Pattern algorithm may be used to classify individual pixels based on local intensity differences of surrounding pixels

Approximating Cell Population with Cell Foreground Area

If single-cell segmentation is challenging, a logical first-approximation of cell population is the foreground area in which cells are the primary objects. To evaluate whether this approximation was adequate for the intended analysis, a number of image processing techniques were explored to segment the cell-containing foreground from the image background. Image foreground areas from each method were compared to total cell counts over a range of images for both the control low-density culture data sets and the proof-of-principle teratogen screen data sets.

The optimal technique identified was a texture-based segmentation method using the Local Binary Pattern (LBP) algorithm from the Scikit-Image [73] Python package. For each pixel in an image, the LBP algorithm compares the intensity of that pixel with the intensities of the surrounding pixels. If a neighboring pixel is greater in intensity, the algorithm assigns

it a zero value, otherwise it assigns a one. The ones and zeros are then interpreted as a binary number and encoded as a numeric value. As shown in Figure 4.4, a histogram of texture codes may be generated used to classify pixels.

The cell objects in the images considered here are small, rounded blobs frequently broken by edges whereas the background is flat and continuous. To segment the foreground area containing cells, manual histogram thresholds were set to define sharp edges and any continuous flat surfaces touching the edge of the image were discarded. The edges, corners, and small flat areas were combined to define a foreground segmentation mask. Figure 4.5 compares the texture-based segmentation to the single-cell segmentation method on a randomly sampled image from the control image set.

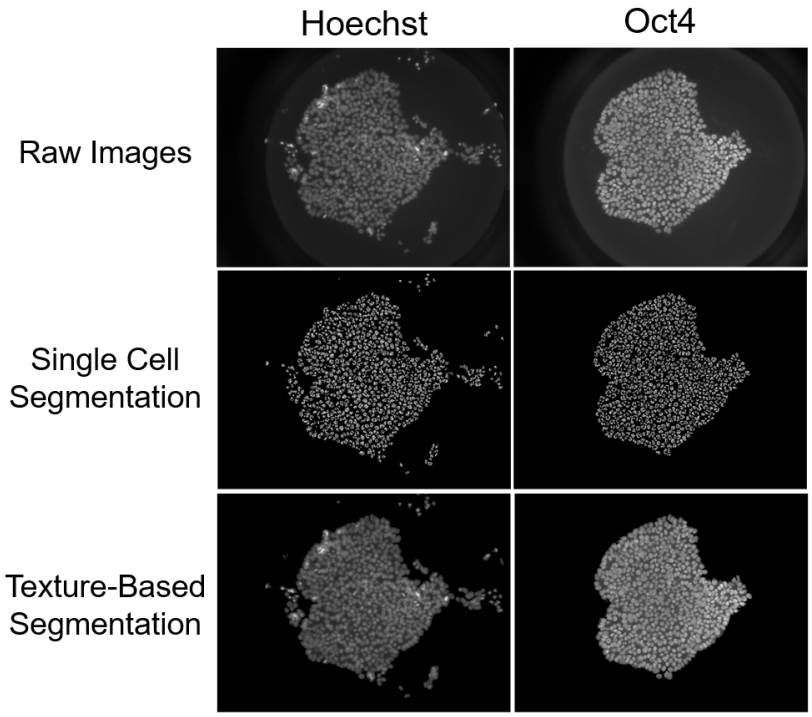
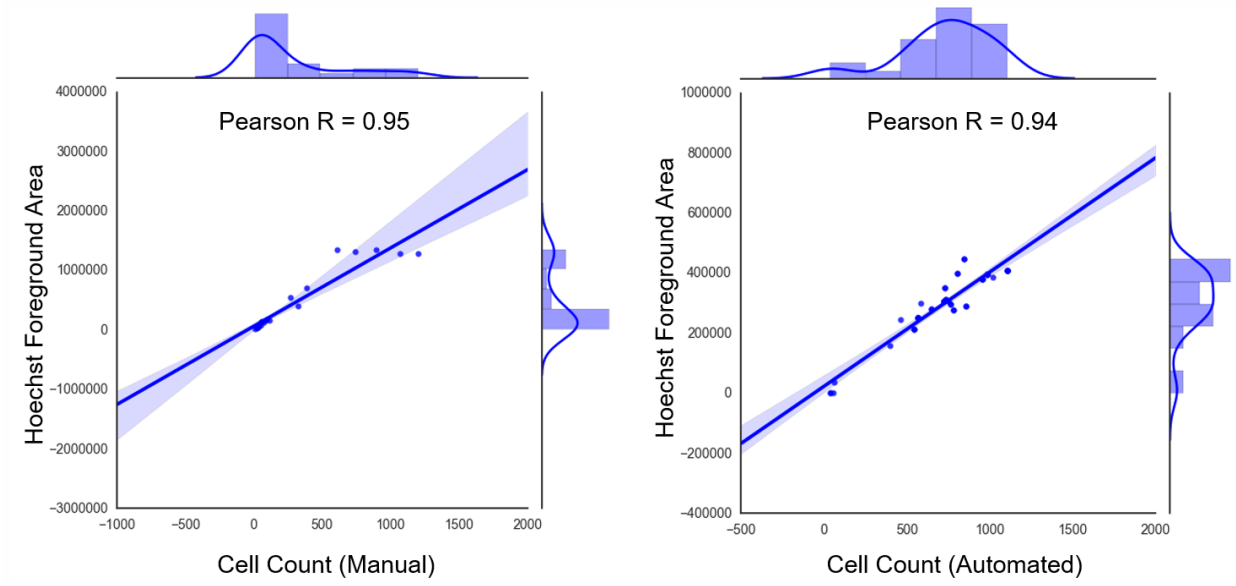


Figure 4.5: Texture-based segmentation algorithm robustly segments cell-containing foreground from background

The texture-based segmentation method was used to generate foreground masks for two sets of images. The first was a small subset of images from the proof-of-principle teratogen screen data with cells counted by hand. The second consisted of randomly sampled images from the control data sets counted automatically using the single cell segmentation method. The results are shown in Figure 4.6.



Hoechst nuclear counterstain foreground area correlates linearly with cell count for hand-counted cultures from the proof-of-principle data sets (left) and for randomly-sampled control cultures counted using the single cell segmentation algorithm (right). The foreground area is the total pixel count within the segmentation mask generated using the Local Binary Pattern segmentation method.

Figure 4.6: Hoechst nuclear counterstain foreground area correlates linearly with cell count

Based on the above analysis, the LBP foreground area (foreground pixel count) was accepted as an adequate approximation of cell population.

Selection of Approximate Methods for Measuring Protein Expression

For characterization of protein expression in the proof-of-principle teratogen screen data sets an approximation of the cell population mean Pearson R was desired from the segmented foreground image. Although quantitative colocalization statistics have been characterized on simulated images containing multiple objects [56, 69] no systematic characterization of methods has compared performance at the single object level with performance across multiple ROIs.

The following colocalization statistics were implemented in the Python programming language, validated on control data from the respective papers (where available), and compared on simulated images of single and multiple cell objects.

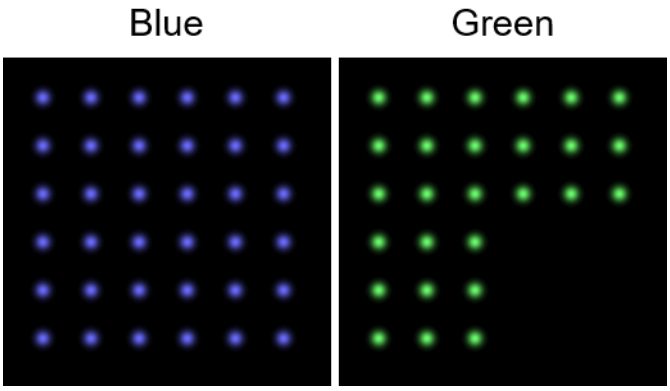
- Adler H_{coef}
- Fractional Area Overlap A1 & A2

- Herce H
- Li Intensity Correlation Analysis Quotient ICQ
- Manders R
- Manders Fractional Overlap M1 & M2
- Manders Intensity Correlation Coefficients k1 & k2

In addition, the following existing methods from Scientific Python (https://www.scipy.org/) were characterized:

- Pearson R
- Spearman Rho
- Kendall Tau

Characterization of these statistics on simulated images highlighted the Manders Fractional Overlap Coefficient M1 as a possible candidate approximation of colocalization based on foreground area.

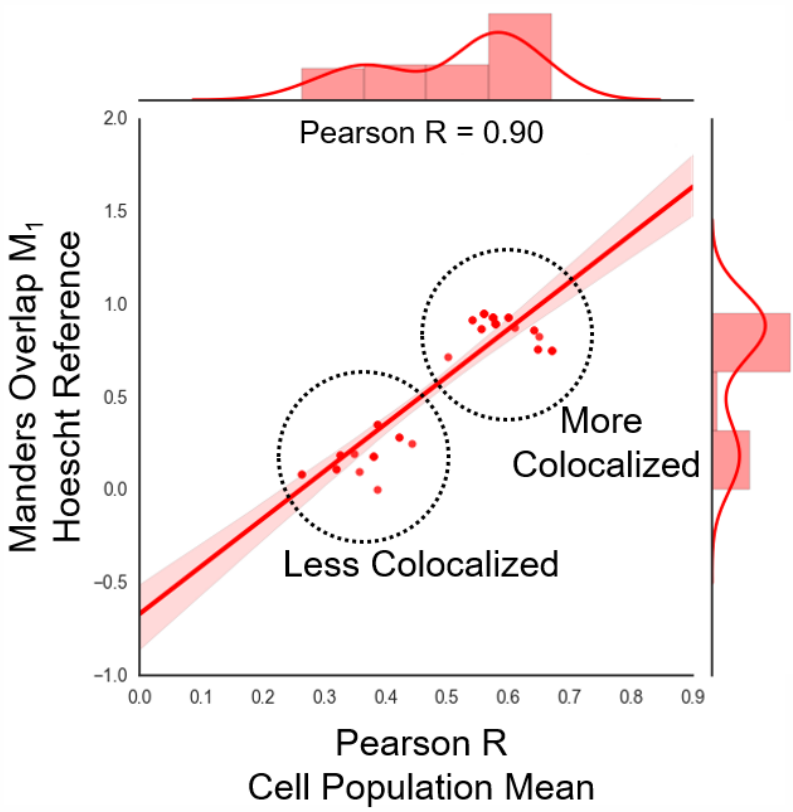


This pair of simulated images contains 36 objects in the blue (left) channel and 27 objects in the green (right) channel. Since the objects are perfectly colocalized, a quantitative colocalization analysis should show that 75% of the blue objects and 100% of the green objects are colocalized. Calculating Manders Overlap M1 (blue objects colocalized with green) yields 0.75 and Manders Overlap M2 (green objects colocalized with blue) returns a value of 1. This matches the expected results.

Figure 4.7: Sample simulated images highlighted Manders Overlap M1 and M2 as possible colocalization method candidates

To evaluate whether the Manders factional overlap coefficients M1 and M2 could approximate the ideal analysis on biological data, the coefficients were compared to single cell results on the control low-density stem cell data. Since all cells are expected to be stained with the Hoechst nuclear stain and Oct4 staining will only occur in a subpopulation of the total Hoechst objects, Manders M1 was defined with the Hoechst-stained total cell population

as the reference and M2 was defined with the Oct4-positive subpopulation as the reference. Figure 4.8 illustrates that Manders M1 correlates with the population mean Pearson R and correctly distinguishes highly co-stained populations from populations in which co-staining is low for all test image pairs.



Manders Overlap M1 classifies positively and negatively stained cell populations without computationally-expensive single-cell segmentation. Individual scatter points represent single cultures characterized by both methods. Cell population mean Pearson R (x-axis) is the average linear correlation between stains for single cells. Manders overlap M1 (y-axis) is the fraction of Hoechst signal colocalized with Oct4 stain. The linear fit and 0.9 Pearson R value (center top) illustrate that estimated values for the two methods of measuring colocalization are linearly related. The scatter about the linear fit shows some intrinsic error in the linear relationship, but both metrics achieve a clear separation of more colocalized and less colocalized cell populations. By averaging multiple replicates to reduce error in the linear relationship the Manders M1 metric may be used as a continuous approximation of the cell population mean Pearson R. Alternatively, thresholds could be defined at $x = 0.5$ and $y = 0.5$ to classify to bin single cultures into positively- (> 0.5) or negatively-stained (< 0.5) categories.

Figure 4.8: Manders Overlap M1 is a computationally-efficient statistic to approximate colocalization within a cell population

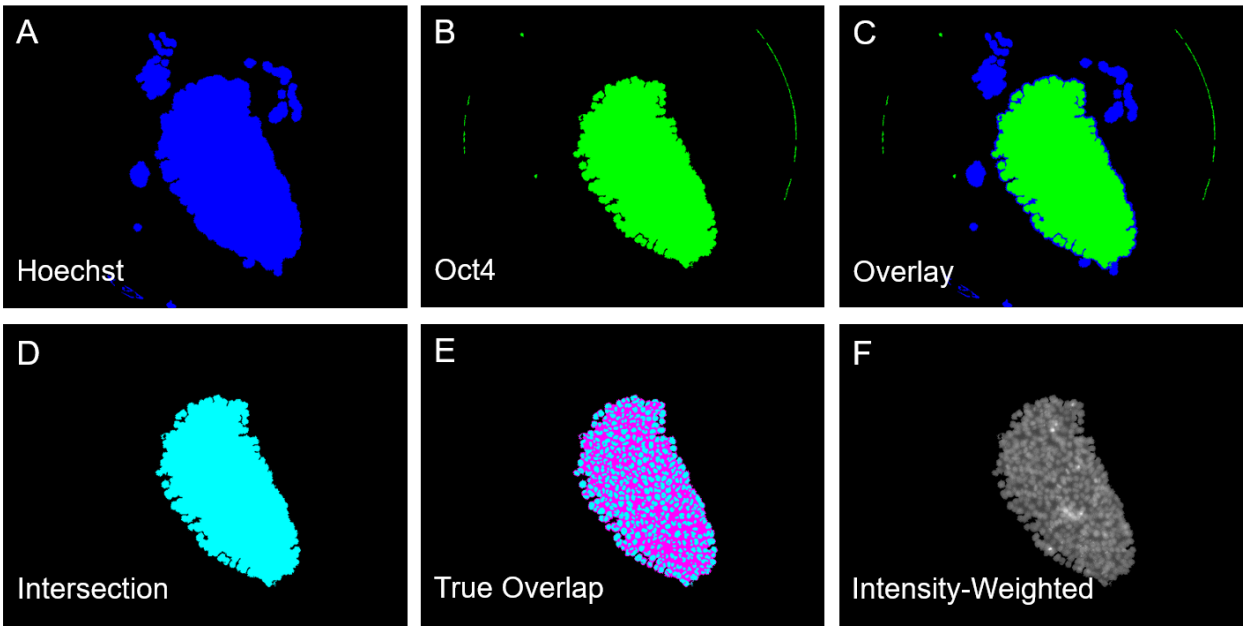
Overview of the Final Image Analysis

The objective of image analysis for the proof-of-principle teratogen screen was the quantification of changes in cell population and protein expression. Due to technical challenges in segmenting single nuclei within dense human pluripotent stem cell colonies and the heterogeneous cell populations arising from differentiation, approximate metrics were identified from a range of published methods using control biological and simulated image data sets. The final metrics selected are the following:

$$Area_{stain} = \frac{number\ of\ pixels\ in\ foreground}{number\ of\ pixels\ in\ image}$$

$$M_1 = \frac{total\ signal\ in\ costained\ region\ of\ interest}{total\ signal\ in\ region\ of\ interest}$$

Figure 4.9 walks through the logical steps of the final image analysis applied to the three proof-of-principle teratogen screen data sets.



Calculating foreground area and overlap coefficients begins with segmenting the foreground area for Hoechst (A) and the stain of interest–Oct4 shown here (B). The segmentation masks are then overlaid (C) and a logical AND operation is applied to find the overlapping intersection (D). Since the intersection consists of true foreground (cells, cyan) and false positive foreground (space between cells, magenta) (E), pixels are weighted by intensity values (F). The Manders overlap coefficient M1 is calculated by computing the sum of intensity values in the overlapping area and normalizing by the sum of the intensity values in the Hoechst foreground image. This yields an intensity-independent fraction of colocalized reference signal to the total reference stain signal with values ranging between 0 and 1.

Figure 4.9: Visual guide to the metrics selected to approximate cell population and biomarker expression

Chapter 5

Statistical Analysis of Dose-Response Data

5.1 Analysis of Experimental Screening Data

Objectives of the Analysis

The objective of experimental data analysis was to calculate summary statistics from dose-response data to use as a basis for comparison of toxicity between molecules in the compound library. These statistics should summarize observed differences in cell population and biomarker expression from the untreated control. Sigmoidal curves are often fit to dose-response data and used to calculate a half maximal effective concentration or EC_{50} as a measure of potency. Due to the large variation observed in exploratory data analysis (see Appendix A.3) for dose-response replicates and the expected correlation between technical replicates (individual cell culture wells) within single experimental batches due to shared reagents, cells, and laboratory conditions, a statistical analysis was desired to estimate the expected values and confidence intervals for each individual experimental treatment (one molecule at one dose level) given the highly nested experimental design prior to calculating summary statistics from dose-response curves.

ANOVA and Nested Data Sets

Analysis of variance (ANOVA) models are commonly used statistical techniques for analyzing high throughput screening data by testing for statistically significant differences among group means. [48] The ANOVA may serve similar analysis goals to Student's t-test when more than two groups are considered. Like the t-test, ANOVA analyses depend on assumptions of normality, statistical independence, and equal variances. For the three data sets considered, all assumptions are violated if individual cell culture wells are considered separately. Replicates of experimental conditions within a single batch are, in fact, repeated measurements using the same cells, reagents, and culture processes at the same time. Since

the data sets are structured hierarchically with wells nested within plates and plates nested within batches, implementing an ANOVA analysis is complicated by specification of the degrees of freedom. A test on technical replicates (individual wells) would artificially inflate the confidence level whereas a test on the replicates aggregated by plate and batch would be overly conservative.

Mixed Effects Modeling

Data sets with repeated (interdependent) measurements are commonly analyzed using a method called "linear mixed effects modeling." This method is a subclass of linear models and is closely related to linear regression and ANOVA-type analyses. Mixed effect modeling derives its name from the specification of so-called "fixed" and "random" effects when fitting a linear model to a data set. Fixed effects are estimated parameters typically specified in the study and of particular interest. The response due to a single experimental treatment (one molecule at one dose level) may be a fixed effect. Standard linear models estimate fixed effects assuming that experimental observations are independent. When experimental observations have some interdependence on another factor, random effects may be introduced to assume a different baseline response for each factor. For example, responses may depend on experimental grouping such as the cell culture plate on which they are located, and different groups may have different nominal responses. In mixed effects models, the nominal responses of random effects are also estimated parameters, and assumptions may be made about the population from which these parameters are drawn. This gives structure to the error term of the model and enables characterization of variation due to some experimental variable. Mixed effect models are seeing increasing application in biological data analysis in recent years due to the fact that biological data often contains interdependent measurements. [46, 74–76]

The key advantages of mixed effect modeling are:

1. No assumptions are made regarding independence. The method is thus especially valuable for analyzing repeated sampling, time series, and hierarchical data sets.
2. Fixed effects are nearly always specified relative to some standard of comparison (the intercept) during model formulation. This makes calculating the difference between a sample and the untreated control (also called the "effect size") relatively easy.
3. A mixed effect model may estimate both the confidence intervals (or statistical significance) and the magnitude of the effect size for comparisons between group means.
4. Sources of variation are quantified. Random effects are estimated, enabling identification of systematic sources of error and their magnitude.
5. Information from the full data set is pooled to improve confidence in individual estimates. For example, wells from all samples on the same chip contribute estimating a parameter for the noise introduced by plate-to-plate variation. This so-called "shrinkage"

can have a profound impact on the final parameter estimates.

The standard form of a linear mixed effects model is,

$$y_{measured} = X\beta_{fixed} + Zb_{random} + \varepsilon_{error} \tag{5.1}$$

where

- y is the n-by-1 response vector, and n is the number of observations,
- X is an n-by-p fixed-effects design matrix, and p is the number of fixed-effects,
- β is a p-by-1 fixed-effects vector,
- Z is an n-by-q random-effects design matrix, and q is the number of random-effects,
- ε is an n-by-1 residual error vector.

The assumptions of a linear mixed effects model are as follows.

- The random-effects vector b and error vector ε are independent of one another.
- Normal prior distributions are assumed for the random-effects vector b and error vector ε:

  $$b \sim N(0, \sigma^2 D(\theta)) \text{ and } \varepsilon \sim N(0, \sigma^2 I)$$

  where D is a symmetric and positive semidefinite matrix parameterized by variance component vector θ, I is the n-by-n identity matrix, and σ^2 is the error variance.

Fitting Mixed-Effects Models in the R Programming Language

A mixed-effects model formula was developed via iterative model building (see Appendix A.4). Fixed effects were specified to enable interpretation of the model intercept as the effect size of the untreated control and the slopes as contrasts for the difference in effect size due to each experimental treatment relative to the untreated control. A random effects structure was developed to account for plate-to-plate and batch-to-batch differences. The mixed-effects models were fit using the `lme4` package [77] in the R programming environment. Although an outlier analysis was conducted (see Appendix A.3) no outliers were excluded from the final analysis. The following illustrates code used to set the untreated control as the intercept and fit a model using the restricted maximum likelihood (REML) estimator:

```
> data_set <- within(data_set, treatment_id <- relevel(
  as.factor(treatment_id), ref = "untreated_control"))
> feature_model <- lmer(feature ~ 1 + treatment_id +
  (1|batch_id\plate_id), data_set, REML=TRUE)
```

For each of the three proof-of-principle data sets considered, a total of nine models was fit corresponding to each of the following nine image measurements or "features":

1. **Area Hoechst:** Hoechst Foreground Area
2. **Area Oct4:** Oct4 Foreground Area
3. **Area PAX6:** PAX6 Foreground Area
4. **Area Brach:** Brachyury Foreground Area
5. **Area GATA4:** GATA4 Foreground Area
6. **Overlap Oct4:** Hoechst/Oct4 Manders Overlap Coefficient M1
7. **Overlap PAX6:** Hoechst/PAX6 Manders Overlap Coefficient M1
8. **Overlap Brachyury:** Hoechst/Brachyury Manders Overlap Coefficient M1
9. **Overlap GATA4:** Hoechst/GATA4 Manders Overlap Coefficient M1

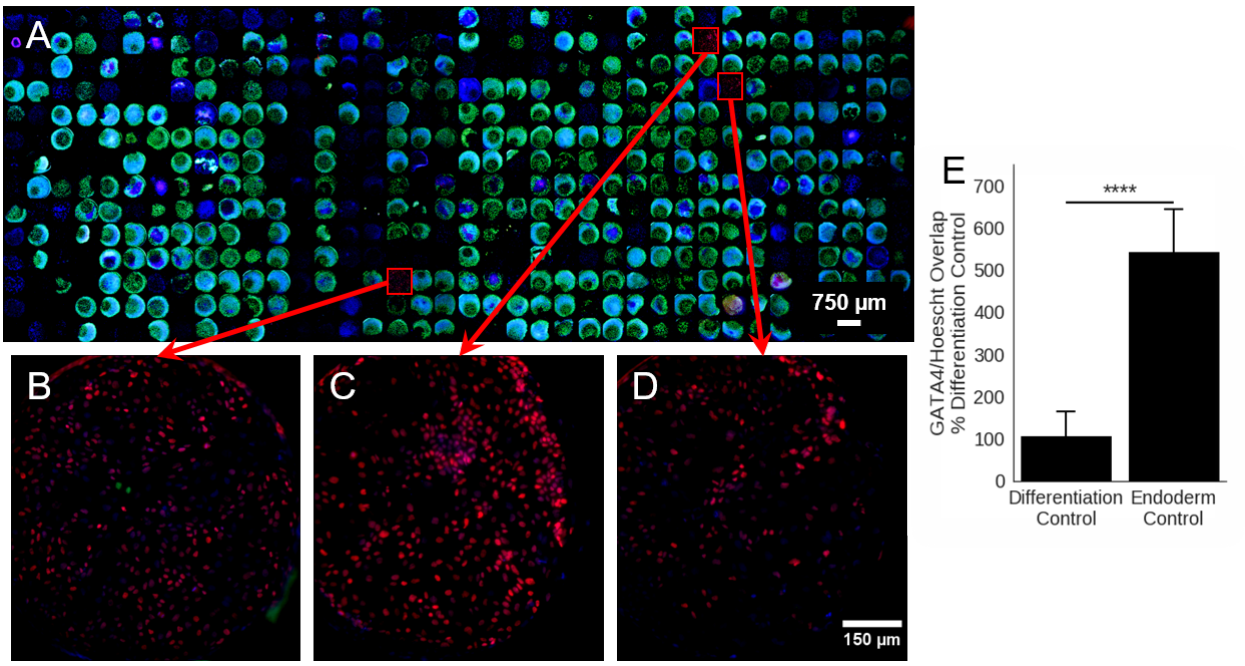
An analysis of the model residuals was conducted to evaluate the model fit for each parameter. While several large residual values were observed, these occurred randomly and no systematic trends were identified in the residuals or random effects. This suggests that the fitted models may adequately satisfy the assumptions described above. As a check to evaluate the validity of the linear mixed effect model results, estimated effect sizes and confidence intervals were compared to nonparametric bootstrap analyses of the same data (see Appendix A.5). Close agreement in effect size estimates were observed and confidence intervals showed general agreement and highlight trade-offs between the parametric and nonparametric analyses.

5.2 Experimental Results

Innovative solutions were developed to overcome practical challenges in image and statistical analysis of the proof-of-principle teratogen screen data sets. To demonstrate the validity of these methods to high content screening data, confirmation that the statistical results agree with the observed and expected biological outcomes was desired.

Quantification of Colocalization using Approximate Metrics and Mixed Effect Modeling

Cells in the definitive endoderm positive control media are expected to lose expression of the pluripotency marker Oct4 and gain expression of the definitive endoderm marker GATA4. Figure 5.1 shows a sample chip image montage and highlights the wells treated with the definitive endoderm positive control medium. The expected changes in staining are observed in all three replicate wells. Quantification of staining for all replicate wells in the TCTF day 10 endpoint data set reveals strong GATA4 marker expression consistent with the observed results.

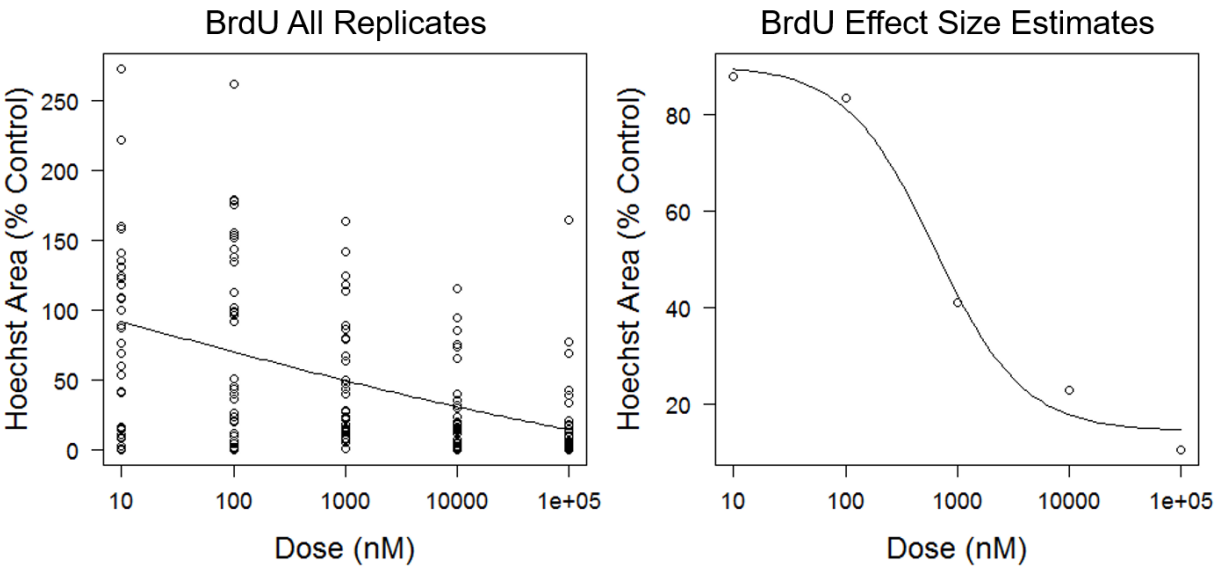


A montage of merged false color fluorescence images (A) gives an overview of data from a sample chip of the TCTF day 10 endpoint experiment. The Hoechst nuclear counterstain is shown in blue, Oct4 pluripotency stain in green, and GATA4 definitive endoderm stain in red. The three technical replicates of the definitive endoderm positive control have been highlighted and are magnified (B-D) to illustrate staining. Quantification of endoderm colocalization (E) for all replicates in the TCTF day 10 data set shows strong expression of GATA4 for the day 10 endpoint TCTF data. Error bars represent 95% confidence intervals, p-value < 0.0001 using a Wald approximation.

Figure 5.1: Sample stain montage with definitive endoderm positive control wells highlighted and quantified

Quantification of Cytotoxicity using Foreground Area and Mixed Effect Modeling

A subset of the compound library consists of molecules known to be generally toxic to cells. When human pluripotent stem cells die they detach from substrate surfaces. Thus the total cell population approximates the total number of viable cells. To evaluate whether the mixed effect analysis and foreground area approximation of cell population could support a viability assay, dose-response curves were generated and compared to IC_{50} (half maximal viability) values from the mouse Embryonic Stem Cell Test literature (see Table 5.1). Although some differences in sensitivity may be expected due to species and cell line differences, in several cases the observed and literature IC_{50} values agree within expected error (order of magnitude). Figure 5.2 shows application of the described analysis to 5-Bromo-2'-deoxyuridine (BrdU), a synthetic nucleoside analog of thymidine which incorporates into



Two dose-response curves were fit with the four-parameter log-logistic function before (left) and after (right) mixed effect model analysis of 5-Bromo-2'-deoxyuridine (BrdU) data from the H1 day 10 data set. On left, single circles represent Hoechst foreground area estimates for single wells (technical replicates). On right, single circles represent estimated values for individual BrdU treatments (individual doses) from the mixed effect model analysis. The mixed effect model pools data from all plates in the H1 day 10 data set, and no outliers were omitted. The IC_{50} values are 26 nM before and 625 nM after mixed effect analysis. The literature IC_{50} for mouse embryonic stem cells is 940 nM.

Figure 5.2: Mixed effect analysis enables estimation of cell viability with the Hoechst foreground area approximation

newly synthesized DNA of replicating cells.

As Figure 5.2 illustrates, individual cell cultures may be highly variable and treating technical replicates as independent observations yields misleading results. Only when the interdependence between replicates is accounted for in the mixed effect model analysis does the trend become clear and a dose-response curve emerge with the expected IC_{50} value from literature. As highlighted by Table 5.1 and Appendix A.6 the mixed effect model analysis does not always result in such clean dose-response data; variation not accounted for by random effects terms remains apparent in the estimated fixed effects.

Molecule	TCTF Day 10	H1 Day 10	TCTF Day 5	Literature
5-Bromo-2'-deoxyuridine	—	625 nM	—	940 nM
Cadmium Sulfate	12 μM	225 μM	—	3.6 μM
Hydroxyurea	91 nM	8.7 μM	—	2 μM
Methotrexate	—	14 nM	—	160 nM

Dashes indicate indeterminate values. See Appendix A.6 for corresponding dose-response curves.

Table 5.1: Comparison of estimated IC_{50} values to mouse embryonic stem cell literature values for cytotoxic compounds

Evaluation of Controls

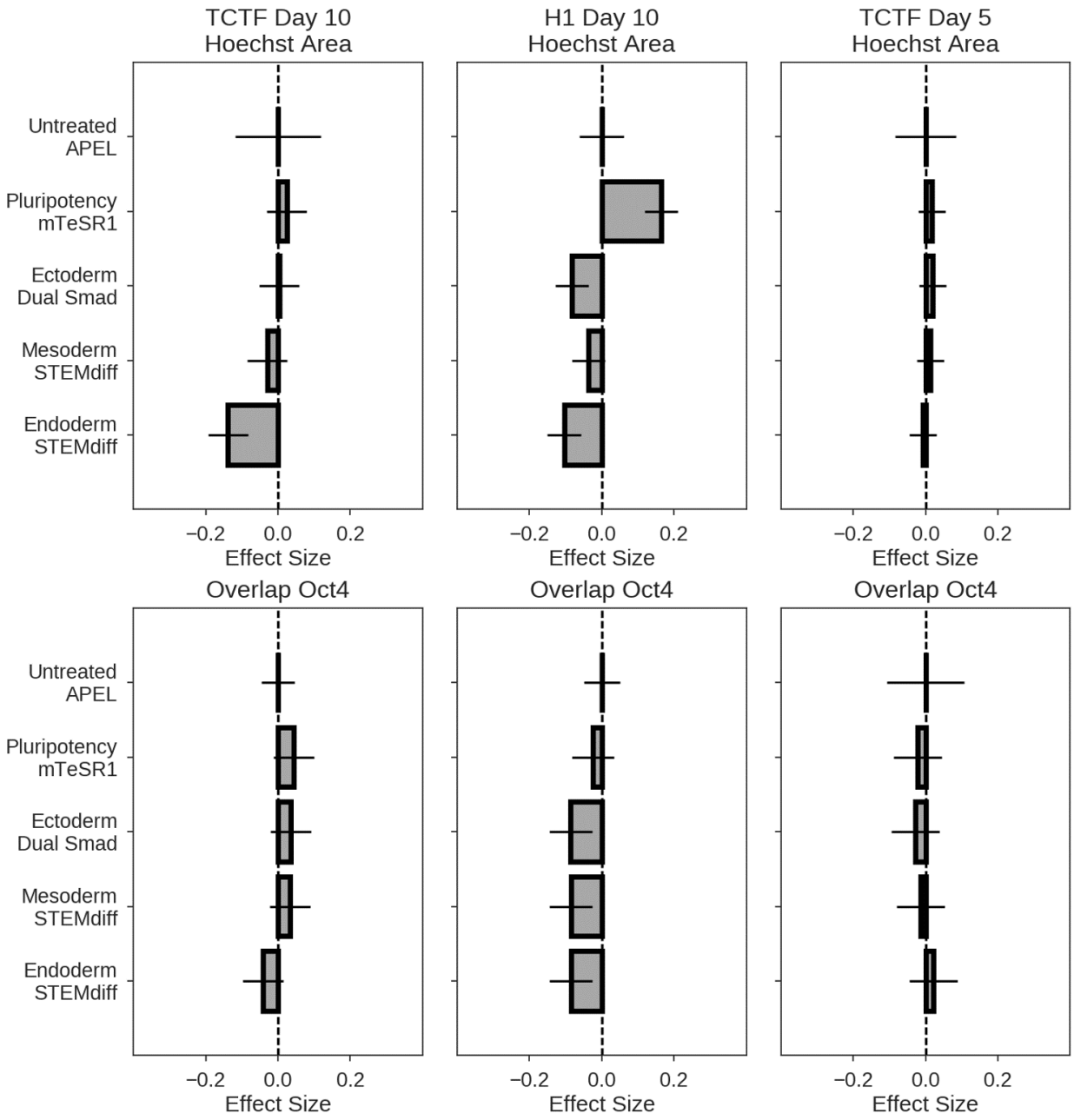
Previously reported mouse Embryonic Stem Cell Test variants assessed endpoints for terminally differentiated cell populations such as cells of the heart, central nervous system, or bone tissue. As cells gain expression of terminal tissue markers they may reasonably be expected to remain positive for those markers for the duration of the assay. While this provides some security in terms of endpoint detection, such tissue markers necessarily limit the scope of the assay to particular lineages and may not represent the full diversity of signaling pathways sensitive to disruption by teratogens.

In contrast, the proposed differentiation assay sought to characterize early markers of tissue progenitor specification. The expected outcome of the differentiation assay was a loss of pluripotency marker Oct4 expression and a gain of one of three germ layer biomarkers: PAX6 (ectoderm), Brachyury (mesoderm), or GATA4 (definitive endoderm). While the proposed endpoints capture developmentally-sensitive progenitor populations representative of a broad spectrum of potential tissue lineages, expression of progenitor biomarkers is expected to be transient with optimal windows of expression between day 5 and 10. Experiments conducted at day 5 and day 10 aimed to capture early and late windows of biomarker expression, respectively. Positive controls were included for all expected cell fate outcomes on all cell culture plates to gauge the degree to which marker expression at the time points examined.

Figure 5.3 shows Hoechst and Oct4 stain effect size estimates (difference from untreated control) for the set of positive controls. Controls from the H1 day 10 data set (column middle) show the expected behavior with cell population (Hoechst Area) increasing in supportive pluripotency maintenance media mTeSR1 and generally decreasing in differentiation conditions relative to the minimally supportive basal medium APEL. H1 results for Oct4 expression also show expected decreases in pluripotency as cells differentiate. TCTF day 10 results (column left) show notable differences relative to the H1 cell line. TCTF cell populations appear to be relatively insensitive to proliferation or differentiation conditions with the exception of a marked decrease in population in endodermal differentiation media. Oct4 staining of TCTF day 10 data again suggests that the cell line is particularly sensitive to

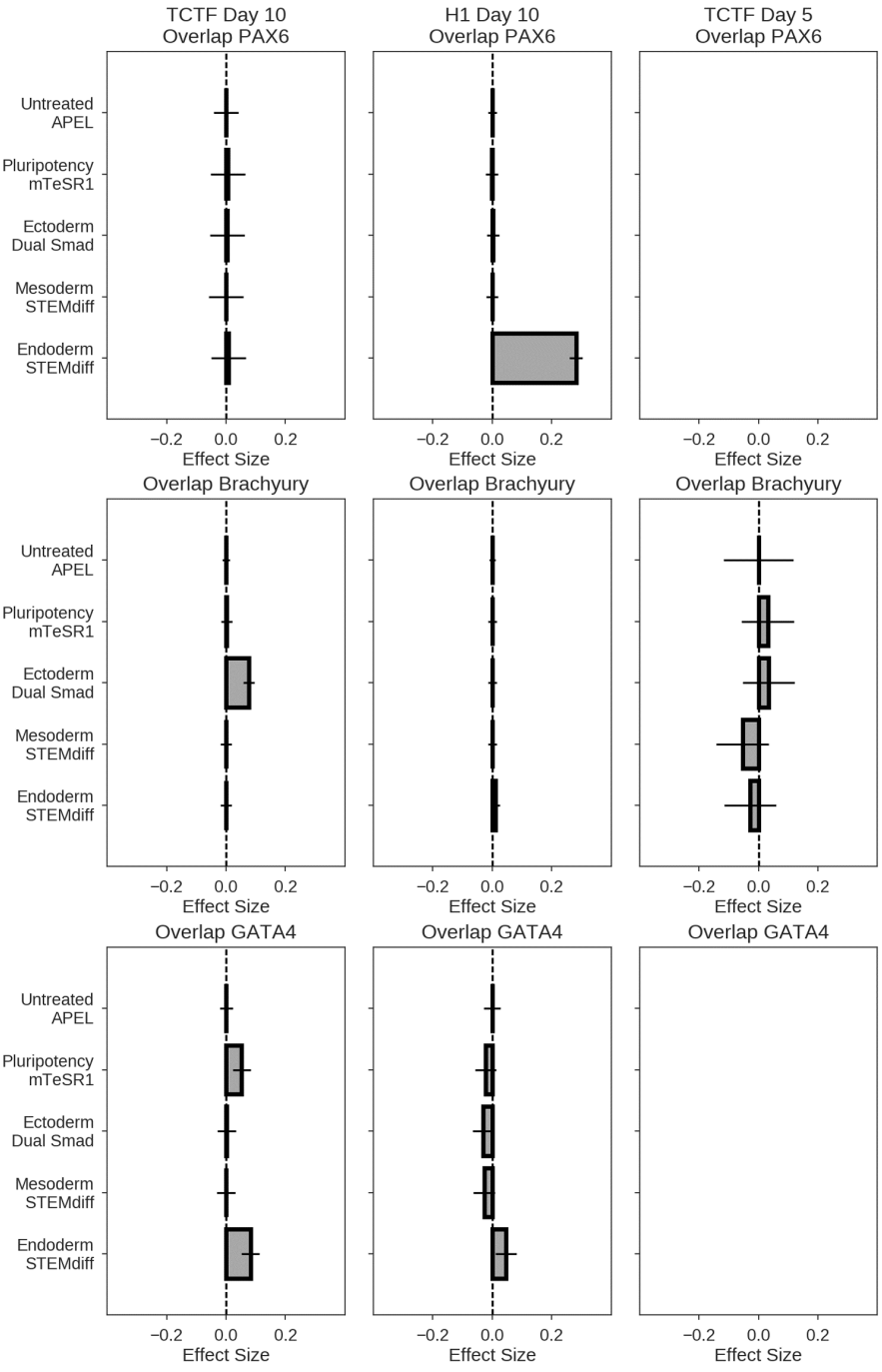
endodermal differentiation as a smaller fraction of the population remains pluripotent in the endoderm differentiation media than in any other control. Controls from the TCTF day 5 (column right) data set showed no substantive difference from the untreated control in terms of cell population or pluripotency. This is generally consistent with the observed insensitivity of the cell line to control conditions at day 10. The noted sensitivity of TCTF to endodermal differentiation at day 10 is not observed at day 5, but this may be time-dependent since endoderm lineage specification is expected to occur late in germ layer specification.

Figure 5.4 shows germ layer stain profiles of the control conditions. Although effect size estimates (difference from untreated control) do not show the magnitude of untreated control, it showed little to no staining for all germ layer stains and data sets. Mixed effects models could not be fit for TCTF day 5 Overlap PAX6 and Overlap GATA4 measurements due to the variation in the data. This is consistent with expectations that PAX6 and GATA4 should be expressed after day 6. PAX6 expression is not observed in the ectodermal positive control at day 5 or day 10, which suggests either a limited window of expression between days 6 and 9 or systematic issues with the dual smad media. Interestingly, the ectoderm marker PAX6 is markedly increased in H1 cells exposed to the endoderm positive control whereas no difference in PAX6 expression is observed in the comparable TCTF day 10 data. Similarly, mesoderm marker Brachyury appears to be upregulated in the TCTF day 10 ectoderm control whereas the comparable H1 data shows no difference. These two cases illustrate the tightly interlinked cell fate decisions during germ layer specification and may be examples of cell line-specific differences in sensitivity to developmental cues. Day 5 should fall within the optimal window of Brachyury expression, but the mesoderm positive control dose not show the expected strong increase in Brachyury expression. Finally, GATA4 expression increased in both day 10 endoderm positive controls as expected.



Effect size estimates of positive controls show the contrasts (difference relative to untreated control) estimated via mixed effect modeling. Error bars are 95% confidence intervals.

Figure 5.3: Controls stained for Hoechst and Oct4



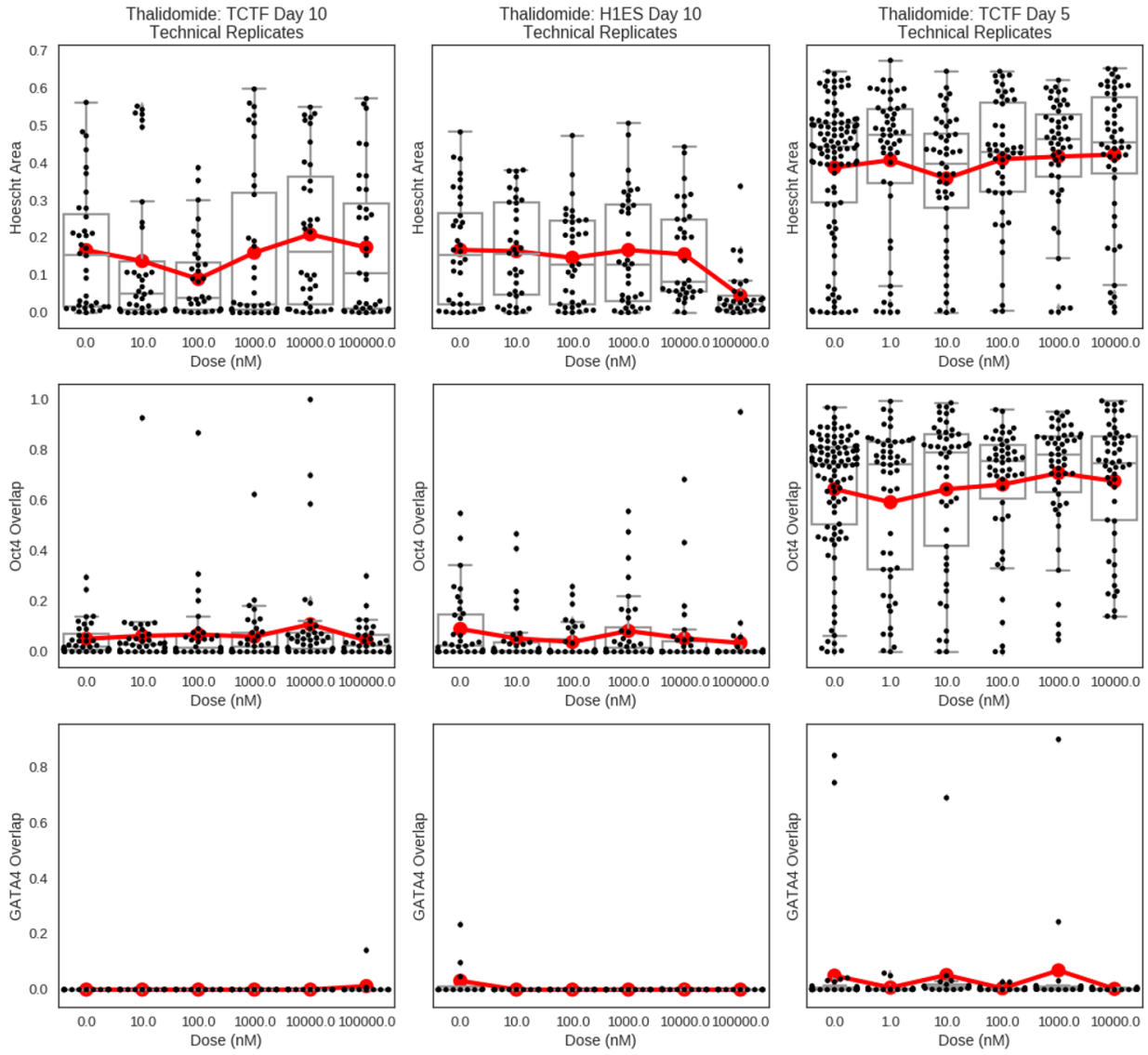
Effect size estimates of positive controls show the contrasts (difference relative to untreated control) estimated via mixed effect modeling. Error bars are 95% confidence intervals.

Figure 5.4: Germ layer stain control plots

Diagnosis of the Data Sets

Taken together, the positive control results show adequate Hoechst, Oct4, and GATA4 stain profiles with observed outcomes matching the expected biological results. Further optimization may be required to identify the optimal timepoints for observing PAX6 and Brachyury expression or eliminate systematic sources of variation leading to no staining in the respective positive controls. The untreated control served as the basal media into which the experimental substances were supplemented. Withdrawal of growth factors maintaining pluripotency via transferring cells to the untreated control media was insufficient to drive differentiation and cells in the untreated control condition largely remained pluripotent.

In terms of the differentiation assay, had molecules of the compound library exhibited sufficient activity to induce differentiation, the low level of spontaneous differentiation in the untreated control would have been desirable for understanding the specific changes in cell fate due to experimental treatment. The available evidence suggests instead that even potent human teratogens such as Thalidomide do not exhibit sufficient activity to induce differentiation (as measured via strong loss of Oct4 or gain of a differentiation biomarker) independent of other differentiation cues (see Figure 5.5). Biological signaling is often regulated by inhibition and disinhibition rather than activation, so if insufficient positive staining is observed in the untreated condition then any inhibition due to experimental treatment may not be distinguishable. A time series experiment in which endpoints are measured daily for positive control media staining should resolve issues of timing in biomarker expression or systematic variation in control media formulation. With optimized timepoints and positive germ layer staining, the positive control media conditions for ectoderm, mesoderm, and endoderm could be used as the basal media conditions and untreated controls for three separate germ layer screens. These two steps would resolve many of these issues discussed here, and a defined rather than spontaneous differentiation may bring sources of biological variation under control.



Dose-response data for thalidomide reveals no strong dose-dependent effects on pluripotency marker Oct4 or definitive endoderm marker GATA4. Single black dots represent observed values for single wells (technical replicates). Red dots track mean values for each dose.

Figure 5.5: Dose-response data for thalidomide

Chapter 6

Closing Remarks

6.1 Summary

In this work a human pluripotent stem cell (hPSC) assay was proposed as a novel imaging-based screen for identifying substances (teratogens) likely to cause birth defects in humans. The assay was engineered to leverage lessons from the validated mouse Embryonic Stem Cell Test (EST). An automated fluid handler was repurposed to support printing stem cells and biological materials to an innovative two-part microculture platform, the pillar/well chip system. Key milestones were met, demonstrating the applicability of the microculture platform to hPSC cultures over the duration of the proposed assay. Three proof-of-principle experimental data sets were acquired using statistical "Design of Experiments" principles to eliminate confounding systematic variation and uncontrolled variables. Changes in cell population and expression of four protein biomarkers were observed with automated fluorescence microscopy and quantified by image analysis. Toxic molecules were found to show similar inhibition of viability to literature values. Further work is needed to optimize quantification and reduce variability in these signals to assess the effects of candidate teratogens on cell differentiation.

6.2 Outlook

One of the primary conclusions of the described work was that none of the three proof-of-principle data sets showed discriminative power in separating molecules of different levels of teratogenic activity. From analysis of the dose-response data it is clear that the biomarkers of germ layer differentiation were expressed at low levels if at all for all experimental treatments except positive control media. The one common variable shared by these treatments is the basal media formulation. Selection of a uniform, growth-factor free basal medium and untreated control was motivated by the following considerations:

- It represents the closest basal medium condition to that used in the the validated EST.

- It minimizes potential confounding effects due to synergistic or antagonistic interactions between the treated molecules and components of the basal medium.
- It allows for clear definition of positive, negative, and untreated controls which encapsulate the full dynamic range of the experimental outcomes.
- It maximizes the number of replicates that can be pooled for statistical analysis since wells for different germ layer endpoints were experimental replicates using the same basal medium formulations and were stained for the nuclear stain and Oct4.
- It vastly simplifies the experimental procedure—unique basal media formulations for each germ layer endpoint would triple the number of unique conditions that needed to be prepared.
- Pooling of technical replicates within a given batch minimizes the number of technical replicates needed within a single plate and thereby maximizes the number of unique experimental conditions (molecules, doses, and controls) that could be included on a single plate.

The commercially available STEMdiff APEL growth factor-free medium was used to minimally support cell viability while isolating cells from confounding signaling cues due proteins or small molecules commonly present in standard maintenance medias. The only cues for fate specification were (1) withdrawal of the growth factor required for maintaining pluripotency (FGF), and (2) supplementation with the specified small molecule at a particular dose. The distinctly low germ layer biomarker expression at day 5 and day 10 endpoints suggests that additional cues may be required to initiate differentiation and observe dose-dependent effects of substances from the compound library. If such is the case, then it may not be feasible to study the dose-dependent effects of a molecule on differentiation in isolation—interactions between multiple fate specification cues may be required.

The above analysis does not provide grounds for evaluating the proposed hypothesis that an assay which screens the earliest developmentally sensitive fate decisions would be maximally predictive of the teratogenic risk to a human embryo should a given substance be presented after maternal effects are accounted for. Thus future studies seeking to develop a human pluripotent stem cell assay for teratogenic activity should do so in the context of combinations of factors known to direct hPSC differentiation to each cell fate of interest. Such an assay would use positive control media as the basal media formulations and untreated controls and would require separate experiments (with distinct basal media formulations) for each endpoint of interest. With an assay modified as described, the proposed hypothesis could be evaluated appropriately.

Bibliography

1. *Update on Overall Prevalence of Major Birth Defects — Atlanta, Georgia, 1978–2005* https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5701a2.htm. (2017).
2. Finnell, R. H. Teratology: General Considerations and Principles. *Journal of Allergy and Clinical Immunology* **103,** S337–S342. ISSN: 0091-6749 (Feb. 1999).
3. Bakoyiannis, I. *et al.* Fetal Alcohol Spectrum Disorders and Cognitive Functions of Young Children. ENG. *Reviews in the Neurosciences.* ISSN: 0334-1763. doi:10.1515/revneuro-2014-0029 (June 2014).
4. Kim, J. H. & Scialli, A. R. Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease. en. *Toxicological Sciences* **122,** 1–6. ISSN: 1096-6080, 1096-0929 (Jan. 2011).
5. Lenz, W. A Short History of Thalidomide Embryopathy. en. *Teratology* **38,** 203–215. ISSN: 1096-9926 (Sept. 1988).
6. Somers, G. F. Pharmacological Properties of Thalidomide (α-Phthalimido Glutarimide), A New Sedative Hypnotic Drug. en. *British Journal of Pharmacology and Chemotherapy* **15,** 111–116. ISSN: 1476-5381 (Mar. 1960).
7. Shanks, N., Greek, R. & Greek, J. Are Animal Models Predictive for Humans? *Philosophy, Ethics, and Humanities in Medicine : PEHM* **4,** 2. ISSN: 1747-5341 (Jan. 2009).
8. Collins, T. F. History and Evolution of Reproductive and Developmental Toxicology Guidelines. *Current Pharmaceutical Design* **12,** 1449–1465 (Apr. 2006).
9. Pijnenborg, R., Robertson, W. B., Brosens, I. & Dixon, G. Review Article: Trophoblast Invasion and the Establishment of Haemochorial Placentation in Man and Laboratory Animals. *Placenta* **2,** 71–91. ISSN: 0143-4004 (Jan. 1981).
10. Alvarado, A. S. & Tsonis, P. A. Bridging the Regeneration Gap: Genetic Insights from Diverse Animal Models. en. *Nature Reviews Genetics* **7,** 873–884. ISSN: 1471-0056 (Nov. 2006).
11. Bailey, J., Knight, A. & Balcombe, J. The Future of Teratology Research Is in Vitro. eng. *Biogenic amines* **19,** 97–145. ISSN: 0168-8561 (2005).

12. Knight, A. Animal Experiments Scrutinised: Systematic Reviews Demonstrate Poor Human Clinical and Toxicological Utility. eng. *ALTEX* **24,** 320–325. ISSN: 1868-596X (2007).

13. Neltner, T. G., Alger, H. M., Leonard, J. E. & Maffini, M. V. Data Gaps in Toxicity Testing of Chemicals Allowed in Food in the United States. *Reproductive Toxicology* **42,** 85–94. ISSN: 0890-6238 (Dec. 2013).

14. Adam, M. P., Polifka, J. E. & Friedman, J. M. Evolving Knowledge of the Teratogenicity of Medications in Human Pregnancy. eng. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics* **157C,** 175–182. ISSN: 1552-4876 (Aug. 2011).

15. Bournias-Vardiabasis, N. & Teplitz, R. L. Use of Drosophila Embryo Cell Cultures as an in Vitro Teratogen Assay. en. *Teratogenesis, Carcinogenesis, and Mutagenesis* **2,** 333–341. ISSN: 1520-6866 (Jan. 1982).

16. Selderslaghs, I. W. T., Van Rompay, A. R., De Coen, W. & Witters, H. E. Development of a Screening Assay to Identify Teratogenic and Embryotoxic Chemicals Using the Zebrafish Embryo. *Reproductive Toxicology* **28,** 308–320. ISSN: 0890-6238 (Nov. 2009).

17. Fort, D. J. *et al.* Comparative Sensitivity of Xenopus Tropicalis and Xenopus Laevis as Test Species for the FETAX Model. eng. *Journal of applied toxicology: JAT* **24,** 443–457. ISSN: 0260-437X (2004 Nov-Dec).

18. Griffin, G. F. & Chu, F. S. Toxicity of the Alternaria Metabolites Alternariol, Alternariol Methyl Ether, Altenuene, and Tenuazonic Acid in the Chicken Embryo Assay. en. *Applied and Environmental Microbiology* **46,** 1420–1422. ISSN: 0099-2240, 1098-5336 (Jan. 1983).

19. Zhang, C. *et al.* Development of a Streamlined Rat Whole Embryo Culture Assay for Classifying Teratogenic Potential of Pharmaceutical Compounds. eng. *Toxicological Sciences: An Official Journal of the Society of Toxicology* **127,** 535–546. ISSN: 1096-0929 (June 2012).

20. Lipinski, R. J. *et al.* Dose- and Route-Dependent Teratogenicity, Toxicity, and Pharmacokinetic Profiles of the Hedgehog Signaling Antagonist Cyclopamine in the Mouse. *Toxicological Sciences* **104,** 189–197. ISSN: 1096-6080 (July 2008).

21. Flint, O. P. & Orton, T. C. An in Vitro Assay for Teratogens with Cultures of Rat Embryo Midbrain and Limb Bud Cells. *Toxicology and Applied Pharmacology* **76,** 383–395. ISSN: 0041-008X (Nov. 1984).

22. Heuer, J., Bremer, S., Pohl, I. & Spielmann, H. Development of an in Vitro Embryotoxicity Test Using Murine Embryonic Stem Cell Cultures. eng. *Toxicology in vitro: an international journal published in association with BIBRA* **7,** 551–556. ISSN: 0887-2333 (July 1993).

23. Krewski, D. *et al.* Toxicity Testing in the 21st Century: A Vision and a Strategy. eng. *Journal of Toxicology and Environmental Health. Part B, Critical Reviews* **13,** 51–138. ISSN: 1521-6950 (Feb. 2010).

24. Scholz, G., Pohl, I., Genschow, E., Klemm, M. & Spielmann, H. Embryotoxicity Screening Using Embryonic Stem Cells in Vitro: Correlation to in Vivo Teratogenicity. en. *Cells Tissues Organs* **165,** 203–211. ISSN: 1422-6421, 1422-6405 (1999).

25. Scholz, G. *et al.* Prevalidation of the Embryonic Stem Cell Test (EST)-A New In Vitro Embryotoxicity Test. eng. *Toxicology in vitro: an international journal published in association with BIBRA* **13,** 675–681. ISSN: 0887-2333 (1999 Aug-Oct).

26. Seiler, A. E. M. & Spielmann, H. The Validated Embryonic Stem Cell Test to Predict Embryotoxicity in Vitro. eng. *Nature protocols* **6,** 961–978. ISSN: 1750-2799 (July 2011).

27. Elke Genschow *et al.* The ECVAM International Validation Study on In VitroEmbryotoxicity Tests: Results of the Definitive Phase andEvaluation of Prediction Models. *ATLA* **30,** 151–176 (2002).

28. Brown, N. A. Selection of Test Chemicals for the ECVAM International Validation Study on in Vitro Embryotoxicity Tests. European Centre for the Validation of Alternative Methods. eng. *Alternatives to laboratory animals: ATLA* **30,** 177–198. ISSN: 0261-1929 (2002 Mar-Apr).

29. Buesen, R. *et al.* Trends in Improving the Embryonic Stem Cell Test (EST): An Overview. eng. *ALTEX* **21,** 15–22. ISSN: 1868-596X (2004).

30. Suzuki, N., Ando, S., Yamashita, N., Horie, N. & Saito, K. Evaluation of Novel High-Throughput Embryonic Stem Cell Tests with New Molecular Markers for Screening Embryotoxic Chemicals In Vitro. *Toxicological Sciences* **124,** 460–471. ISSN: 1096-6080 (Dec. 2011).

31. Seiler, A., Visan, A., Buesen, R., Genschow, E. & Spielmann, H. Improvement of an in Vitro Stem Cell Assay for Developmental Toxicity: The Use of Molecular Endpoints in the Embryonic Stem Cell Test. *Reproductive Toxicology* **18,** 231–240. ISSN: 0890-6238 (Mar. 2004).

32. Chen, X. *et al.* Developing Osteoblasts as an Endpoint for the Mouse Embryonic Stem Cell Test. *Reproductive Toxicology* **53,** 131–140. ISSN: 0890-6238 (June 2015).

33. zur Nieden, N. I., Kempka, G. & Ahr, H. J. Molecular Multiple Endpoint Embryonic Stem Cell Test—a Possible Approach to Test for the Teratogenic Potential of Compounds. *Toxicology and Applied Pharmacology* **194,** 257–269. ISSN: 0041-008X (Feb. 2004).

34. Palmer, J. A. *et al.* Establishment and Assessment of a New Human Embryonic Stem Cell-Based Biomarker Assay for Developmental Toxicity Screening. en. *Birth Defects Research Part B: Developmental and Reproductive Toxicology* **98,** 343–363. ISSN: 1542-9741 (Aug. 2013).

35. Van Dartel, D. A. M. *et al.* Monitoring Developmental Toxicity in the Embryonic Stem Cell Test Using Differential Gene Expression of Differentiation-Related Genes. en. *Toxicological Sciences* **116,** 130–139. ISSN: 1096-6080, 1096-0929 (Jan. 2010).

36. Thomson, J. A. *et al.* Embryonic Stem Cell Lines Derived from Human Blastocysts. en. *Science* **282,** 1145–1147. ISSN: 0036-8075, 1095-9203 (Nov. 1998).

37. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. English. *Cell* **126,** 663–676. ISSN: 0092-8674, 1097-4172 (Aug. 2006).

38. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. eng. *Cell* **131,** 861–872. ISSN: 0092-8674 (Nov. 2007).

39. Yu, J. *et al.* Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. eng. *Science (New York, N.Y.)* **318,** 1917–1920. ISSN: 1095-9203 (Dec. 2007).

40. Park, I.-H. *et al.* Reprogramming of Human Somatic Cells to Pluripotency with Defined Factors. eng. *Nature* **451,** 141–146. ISSN: 1476-4687 (Jan. 2008).

41. Cherry, A. B. C. & Daley, G. Q. Reprogrammed Cells for Disease Modeling and Regenerative Medicine. *Annual Review of Medicine* **64,** 277–290 (2013).

42. Dimanlig, P. V., Faber, S. C., Auerbach, W., Makarenkova, H. P. & Lang, R. A. The Upstream Ectoderm Enhancer in Pax6 Has an Important Role in Lens Induction. en. *Development* **128,** 4415–4424. ISSN: 0950-1991, 1477-9129 (Nov. 2001).

43. Schulte-Merker, S. & Smith, J. C. Mesoderm Formation in Response to Brachyury Requires FGF Signalling. *Current Biology* **5,** 62–67. ISSN: 0960-9822 (Jan. 1995).

44. Holtzinger, A., Rosenfeld, G. E. & Evans, T. Gata4 Directs Development of Cardiac-Inducing Endoderm from ES Cells. *Developmental biology* **337,** 63–73. ISSN: 0012-1606 (Jan. 2010).

45. Krzywinski, M., Altman, N. & Blainey, P. Points of Significance: Nested Designs. en. *Nature Methods* **11,** 977–978. ISSN: 1548-7091 (Oct. 2014).

46. Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V. & van der Sluis, S. A Solution to Dependency: Using Multilevel Analysis to Accommodate Nested Data. en. *Nature Neuroscience* **17,** 491–496. ISSN: 1097-6256 (Apr. 2014).

47. Blainey, P., Krzywinski, M. & Altman, N. Points of Significance: Replication. en. *Nature Methods* **11,** 879–880. ISSN: 1548-7091 (Sept. 2014).

48. Krzywinski, M. & Altman, N. Points of Significance: Analysis of Variance and Blocking. en. *Nature Methods* **11,** 699–700. ISSN: 1548-7091 (July 2014).

49. Altman, N. & Krzywinski, M. Points of Significance: Sources of Variation. en. *Nature Methods* **12,** 5–6. ISSN: 1548-7091 (Jan. 2015).

50. Altman, N. & Krzywinski, M. Points of Significance: Split Plot Design. en. *Nature Methods* **12,** 165–166. ISSN: 1548-7091 (Mar. 2015).

51. Studer, L. Derivation of Dopaminergic Neurons from Pluripotent Stem Cells. eng. *Progress in brain research* **200,** 243–263. ISSN: 1875-7855 (2012).

52. Schardein, J. *Chemically Induced Birth Defects, Third Edition* en. ISBN: 978-0-8247-0265-6 (Taylor & Francis, Apr. 2000).

53. Briggs, G. G., Freeman, R. K. & Yaffe, S. J. *Drugs in Pregnancy and Lactation: A Reference Guide to Fetal and Neonatal Risk* en. Google-Books-ID: OIgTE4aynrMC. ISBN: 978-1-60831-708-0 (Lippincott Williams & Wilkins, 2011).

54. Attene-Ramos, M. S. *et al.* The Tox21 Robotic Platform for the Assessment of Environmental Chemicals – from Vision to Reality. *Drug Discovery Today* **18,** 716–723. ISSN: 1359-6446 (Aug. 2013).

55. Manders, E. M., Stap, J., Brakenhoff, G. J., van Driel, R. & Aten, J. A. Dynamics of Three-Dimensional Replication Patterns during the S-Phase, Analysed by Double Labelling of DNA and Confocal Microscopy. eng. *Journal of Cell Science* **103 (Pt 3),** 857–862. ISSN: 0021-9533 (Nov. 1992).

56. Manders, E. M. M., Verbeek, F. J. & Aten, J. A. Measurement of Co-Localization of Objects in Dual-Colour Confocal Images. en. *Journal of Microscopy* **169,** 375–382. ISSN: 1365-2818 (Mar. 1993).

57. Barbarese, E. *et al.* Protein Translation Components Are Colocalized in Granules in Oligodendrocytes. eng. *Journal of Cell Science* **108 (Pt 8),** 2781–2790. ISSN: 0021-9533 (Aug. 1995).

58. Jiang, L. & Rogers, J. C. Integral Membrane Protein Sorting to Vacuoles in Plant Cells: Evidence for Two Pathways. *The Journal of Cell Biology* **143,** 1183–1199. ISSN: 0021-9525 (Nov. 1998).

59. Silver, M. A. & Stryker, M. P. A Method for Measuring Colocalization of Presynaptic Markers with Anatomically Labeled Axons Using Double Label Immunofluorescence and Confocal Microscopy. eng. *Journal of Neuroscience Methods* **94,** 205–215. ISSN: 0165-0270 (Jan. 2000).

60. Kreft, M., Milisav, I., Potokar, M. & Zorec, R. Automated High Through-Put Colocalization Analysis of Multichannel Confocal Images. *Computer Methods and Programs in Biomedicine* **74,** 63–67. ISSN: 0169-2607 (Apr. 2004).

61. Li, Q. *et al.* A Syntaxin 1, Galpha(o), and N-Type Calcium Channel Complex at a Presynaptic Nerve Terminal: Analysis by Quantitative Immunocolocalization. eng. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **24,** 4070–4081. ISSN: 1529-2401 (Apr. 2004).

62. Costes, S. V. *et al.* Automatic and Quantitative Measurement of Protein-Protein Colocalization in Live Cells. *Biophysical Journal* **86,** 3993–4003. ISSN: 0006-3495 (June 2004).

63. Jaskolski, F., Mulle, C. & Manzoni, O. J. An Automated Method to Quantify and Visualize Colocalized Fluorescent Signals. eng. *Journal of Neuroscience Methods* **146,** 42–49. ISSN: 0165-0270 (July 2005).

64. Adler, J., Pagakis, S. N. & Parmryd, I. Replicate-Based Noise Corrected Correlation for Accurate Measurements of Colocalization. eng. *Journal of Microscopy* **230,** 121–133. ISSN: 1365-2818 (Apr. 2008).

65. Wu, Y. *et al.* Quantitative Determination of Spatial Protein-Protein Correlations in Fluorescence Confocal Microscopy. eng. *Biophysical Journal* **98,** 493–504. ISSN: 1542-0086 (Feb. 2010).

66. Dunn, K. W., Kamocka, M. M. & McDonald, J. H. A Practical Guide to Evaluating Colocalization in Biological Microscopy. *American Journal of Physiology - Cell Physiology* **300,** C723–C742. ISSN: 0363-6143 (Apr. 2011).

67. Zinchuk, V., Wu, Y., Grossenbacher-Zinchuk, O. & Stefani, E. Quantifying Spatial Correlations of Fluorescent Markers Using Enhanced Background Reduction with Protein Proximity Index and Correlation Coefficient Estimations. eng. *Nature Protocols* **6,** 1554–1567. ISSN: 1750-2799 (Sept. 2011).

68. Singan, V. R., Handzic, K., Curran, K. M. & Simpson, J. C. A Method for Improved Clustering and Classification of Microscopy Images Using Quantitative Co-Localization Coefficients. eng. *BMC research notes* **5,** 281. ISSN: 1756-0500 (2012).

69. Herce, H., Casas-Delucchi, C. & Cardoso, M. New Image Colocalization Coefficient for Fluorescence Microscopy to Quantify (Bio-)Molecular Interactions. *Journal of Microscopy* **249,** 184–194. ISSN: 0022-2720 (Mar. 2013).

70. Adler, J. & Parmryd, I. Quantifying Colocalization: Thresholding, Void Voxels and the Hcoef. *PLoS ONE* **9.** ISSN: 1932-6203. doi:10.1371/journal.pone.0111983. (2016) (Nov. 2014).

71. Moser, B., Hochreiter, B., Herbst, R. & Schmid, J. A. Fluorescence Colocalization Microscopy Analysis Can Be Improved by Combining Object-Recognition with Pixel-Intensity-Correlation. en. *Biotechnology Journal* **12,** n/a–n/a. ISSN: 1860-7314 (Jan. 2017).

72. Demandolx, D. & Davoust, J. Multicolour Analysis and Local Image Correlation in Confocal Microscopy. en. *Journal of Microscopy* **185,** 21–36. ISSN: 1365-2818 (Jan. 1997).

73. van der Walt, S. *et al.* Scikit-Image: Image Processing in Python. eng. *PeerJ* **2,** e453 (2014).

74. Ji, H. & Liu, X. S. Analyzing 'omics Data Using Hierarchical Models. en. *Nature Biotechnology* **28,** 337–340. ISSN: 1087-0156 (Apr. 2010).

75. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal. eng. *Journal of Memory and Language* **68.** ISSN: 0749-596X. doi:10.1016/j.jml.2012.11.001 (Apr. 2013).

76. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and Pitfalls in the Application of Mixed-Model Association Methods. en. *Nature Genetics* **46,** 100–106. ISSN: 1061-4036 (Feb. 2014).

77. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* **67,** 1–48 (2015).

78. Wickham, H. Tidy Data. *Journal of Statistical Software* **59,** 1–23. ISSN: 1548-7660 (2014).

79. Altman, N. & Krzywinski, M. Points of Significance: Analyzing Outliers: Influential or Nuisance? en. *Nature Methods* **13,** 281–282. ISSN: 1548-7091 (Apr. 2016).

80. Nakagawa, S. & Cuthill, I. C. Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists. en. *Biological Reviews* **82,** 591–605. ISSN: 1469-185X (Nov. 2007).

81. Krzywinski, M. & Altman, N. Points of Significance: Error Bars. en. *Nature Methods* **10,** 921–922. ISSN: 1548-7091 (Oct. 2013).

82. Kulesa, A., Krzywinski, M., Blainey, P. & Altman, N. Points of Significance: Sampling Distributions and the Bootstrap. en. *Nature Methods* **12,** 477–478. ISSN: 1548-7091 (June 2015).

Appendix A

A.1 Experimental Methods

Immunocytochemistry

Fixing, Permeabolization, and Primary Antibody Stain

1. Prepare the following solutions in designated volumes for each pillar chip to fix:
    a. 24 mL of PBS at 37°C
    b. 6 mL of 4% paraformaldehyde at 37°C
    c. 6 mL of PBS + 0.25% Triton X100
2. Prepare primary antibody solutions (PBS + primary antibody (or IgG control) + 5% donkey serum)
3. Set up aspirator
4. Print primary antibody well chip
5. Fix cells in 4% PFA at 37°C for 20 min
6. Permeabolize cells for 15 min with PBS + 0.25% Triton X-100 solution at 37°C
7. Wash with PBS at 37°C for 5 min
8. Wash with fresh PBS two more times for 15 min each
9. Dry excess liquid from back of pillar chip using aspirator
10. Stamp fixed pillar chip into the primary antibody chip
11. Incubate in a humidified chamber overnight at 4°C

Secondary Antibody Stain and Hoechst Nuclear Counter Stain

1. Set up aspirator
2. Prepare secondary antibody solutions (PBS + secondary antibodies + Hoechst nuclear stain + 5% donkey serum)
3. Prepare imaging solution of PBS + Hoechst nuclear stain
4. Print secondary antibody well chip
5. Wash with room temperature PBS for 15 min
6. Dry excess liquid from back of pillar chip using aspirator
7. Stamp pillar chip into the secondary antibody chip

8. Incubate for 1 hour in the dark at room temperature
9. While waiting, print imaging solution well chip
10. Wash in PBS for 15 min at room temperature
11. Dry excess liquid from back of pillar chip using aspirator
12. Stamp stained pillar chip into the imaging solution chip

A.2 Preparing the Data Set

In order to analyze data resulting from the proof-of-principle experiments, information from the experimental design, laboratory practices, and image analysis was aligned and compiled into a data table. This process was facilitated by assigning a unique integer-valued "treatment ID" to identify each distinct experimental condition. Treatment IDs were automatically aligned with laboratory protocol metadata including experimental batch, plate ID, and randomized well positions during protocol generation using the previously described custom protocol generator for the microarray printing instrument. A standardized naming prefix containing batch, plate ID, well position, and fluorescence channel information was applied to all image data acquired by the high content microscope. During image analysis, the file names were parsed using regular expressions and used to align image measurements and file paths with experimental design information. Data preparation was done in the Python programming language (https://www.python.org/) using the Jupyter interactive electronic notebook environment (http://jupyter.org/).

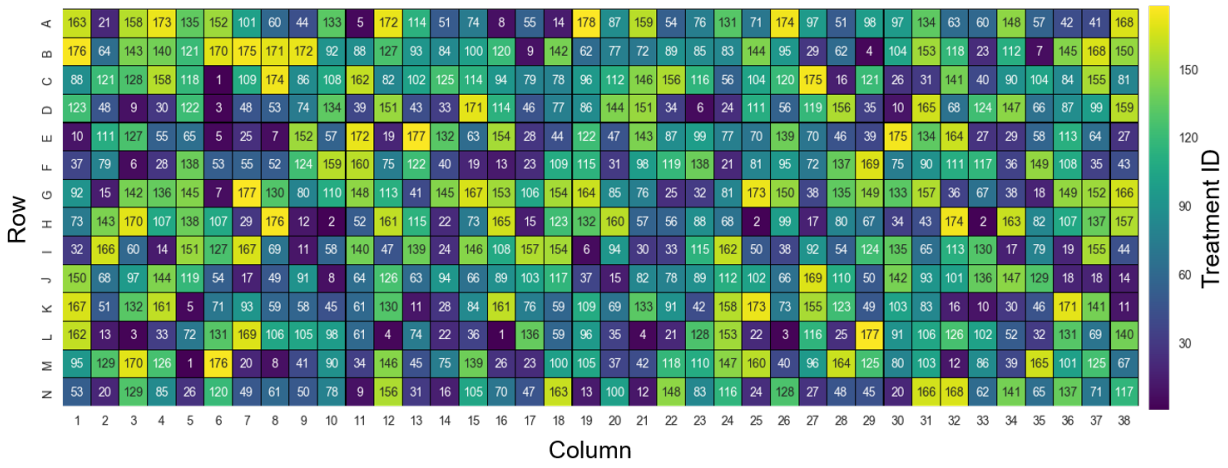


Figure A.1: A sample randomized plate map

Tidying the Data

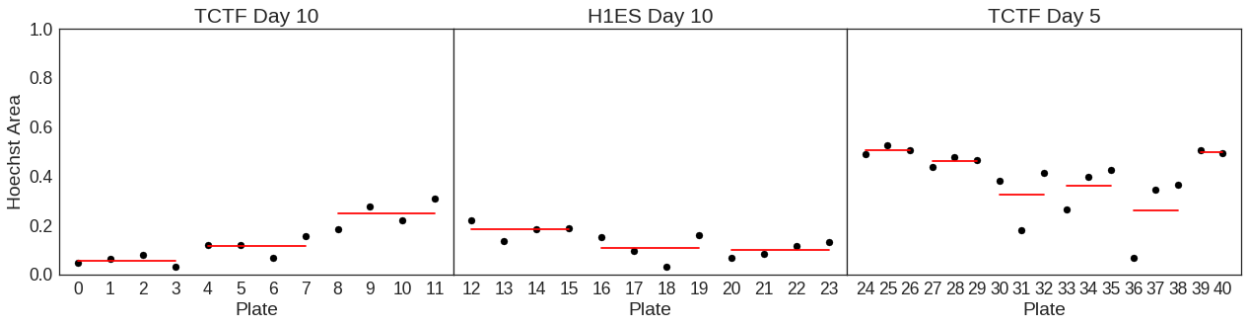
The highly nested structure of the proof-of-principle experiments lends itself to multiple strategies for formatting data tables. Data could be split into separate tables by plate, batch, cell line, experiment duration, or other attributes. The strategy selected for this work was inspired by Hadley Wickham's concept of "tidy data." [78] In brief, statistical data sets are typically compiled in tables consisting of values organized in rows and columns. A tidy dataset is organized such that each row contains all values measured on the same unit and each column contains all values associated with measurements of the same underlying attribute. For the data considered here, each row was defined as a single physical cell culture well. All recorded attributes associated with the experiment were organized into columns with values assigned for each well. Data from all three proof-of-principle data sets were compiled into a single table such that combinations of column values could be used to interactively query the data using the Pandas Python package (http://pandas.pydata.org/).

A.3 Exploratory Analysis of Dose-Response Data

Exploratory data analysis was performed to develop an understanding of the data set and inform model selection.

Systematic Effects

Systematic effects in cell population at the assay endpoint were investigated and evaluated with reference to initial cell seeding (Table A.1). Figure A.2 shows the mean Hoechst area (approximate cell population metric) for all wells within a given plate (black dots) as well as the mean of all plate means within a common batch (red lines). Although differences between plates and between batches are apparent with noted outliers in the TCTF Day 5 endpoint data set, the endpoint Hoechst area does not appear to trend with the initial cell seeding concentration (where available).



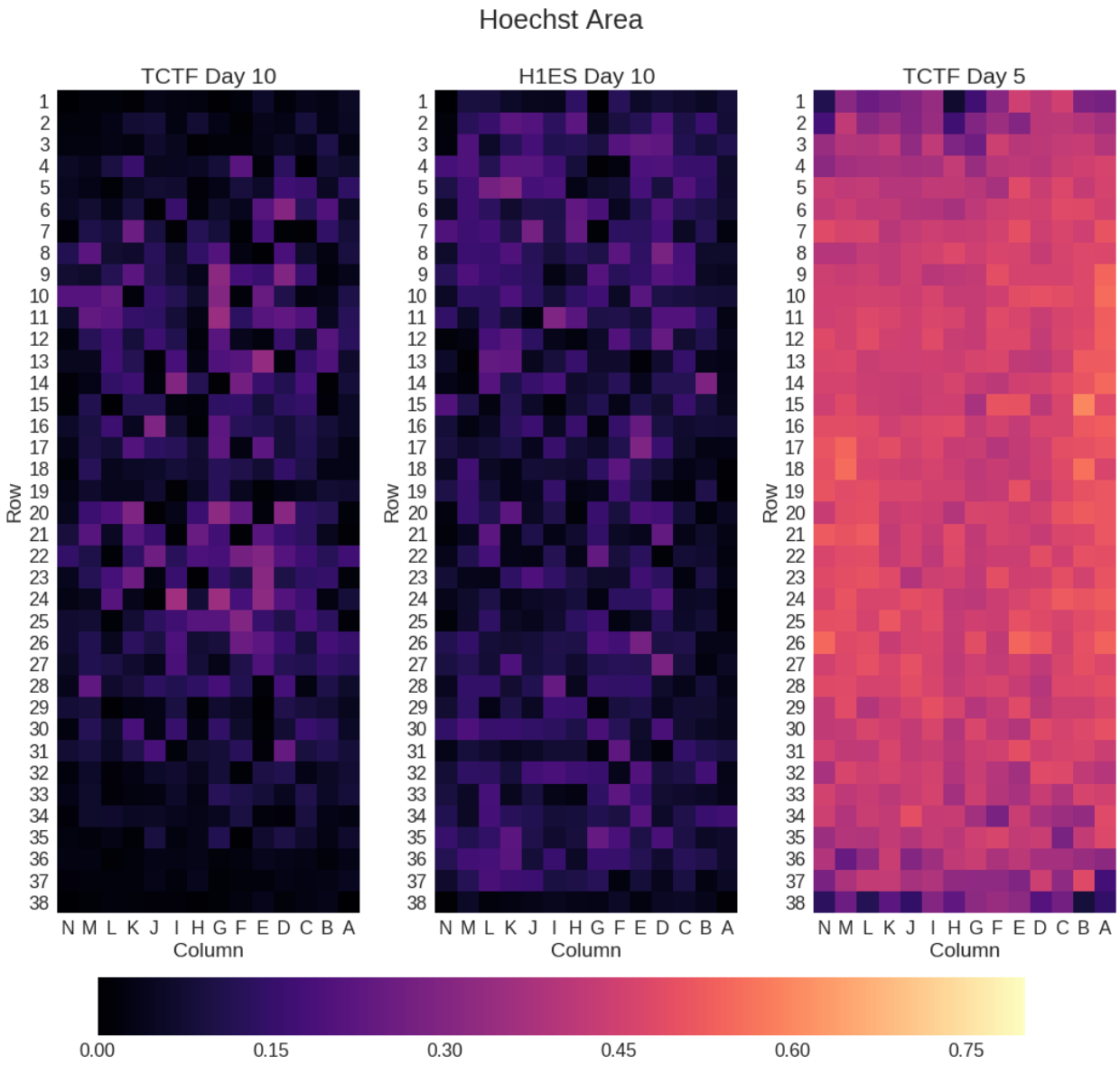
Cell population trends at the assay endpoint as approximated by Hoechst foreground area metric. Length of red lines (x-axis) indicates plates in the same batch seeded at the same time. Height of the red lines (y-axis) indicates mean value of all plates in the batch.

Figure A.2: Cell population trends at endpoint by plate

Data Set	**Batch**	**Plates**	**Cells Seeded (cells/mL)**
TCTF Day 10	0	0-3	$7.61 * 10^6$
	1	4-7	$6.18 * 10^6$
	2	8-11	$5 * 10^6$
H1 Day 10	0	12-15	$9.91 * 10^6$
	1	16-19	$11.2 * 10^6$
	2	20-23	$8 * 10^6$
TCTF Day 5	0	0-3	—
	1	24-26	—
	2	27-29	—
	3	30-32	—
	4	33-35	—
	5	36-38	—
	6	39-40	—

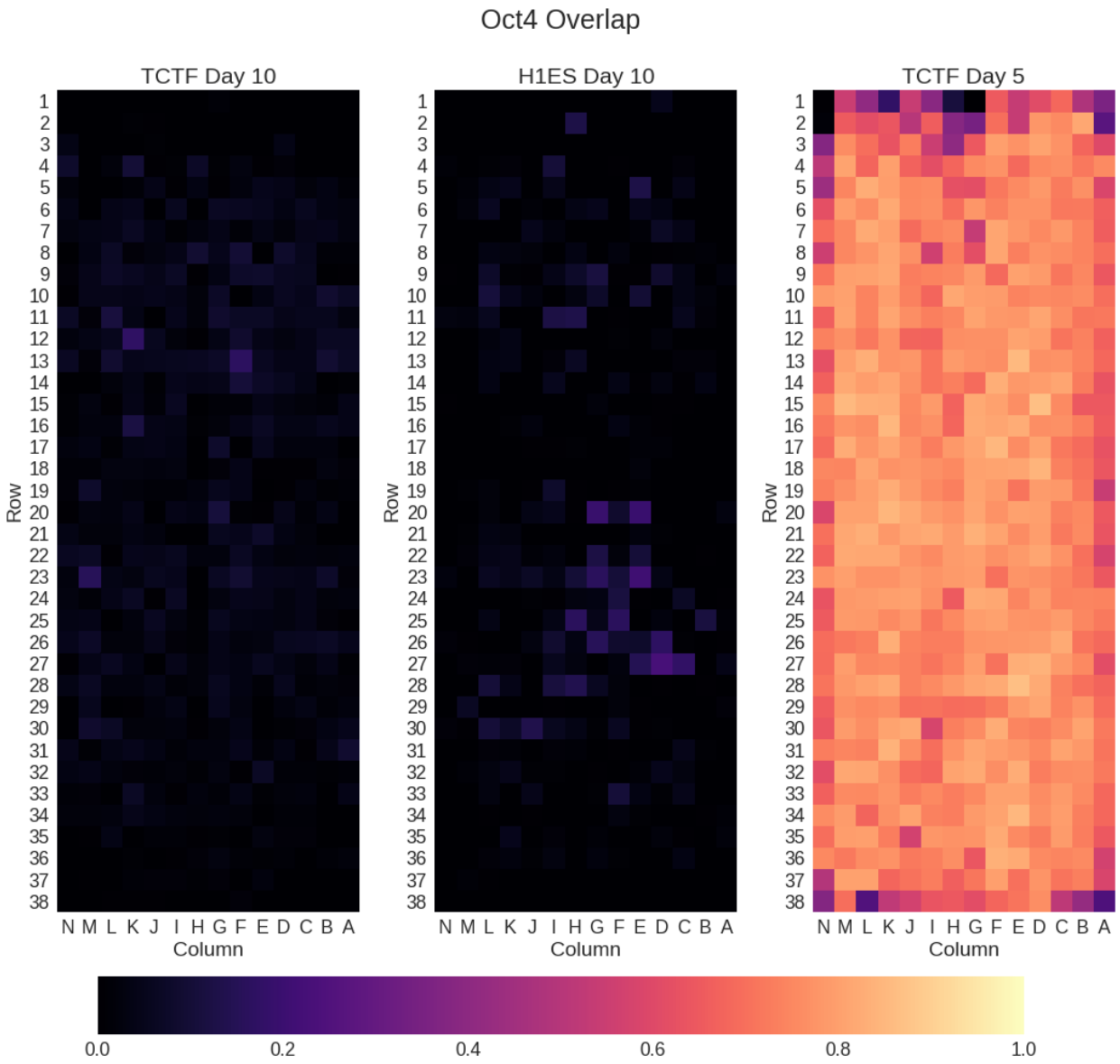
Cells were spotted on Matrigel-pretreated pillar surfaces in 60 nL drops at the specified concentration and allowed to adhere in a humidified environment for 25 minutes prior to stamping into pre-warmed mTeSR1 media containing 10 μM ROCK inhibitor. All plates within a given batch were seeded during the same printing cycle from the same source well.

Table A.1: Initial cell counts prior to seeding on day 0



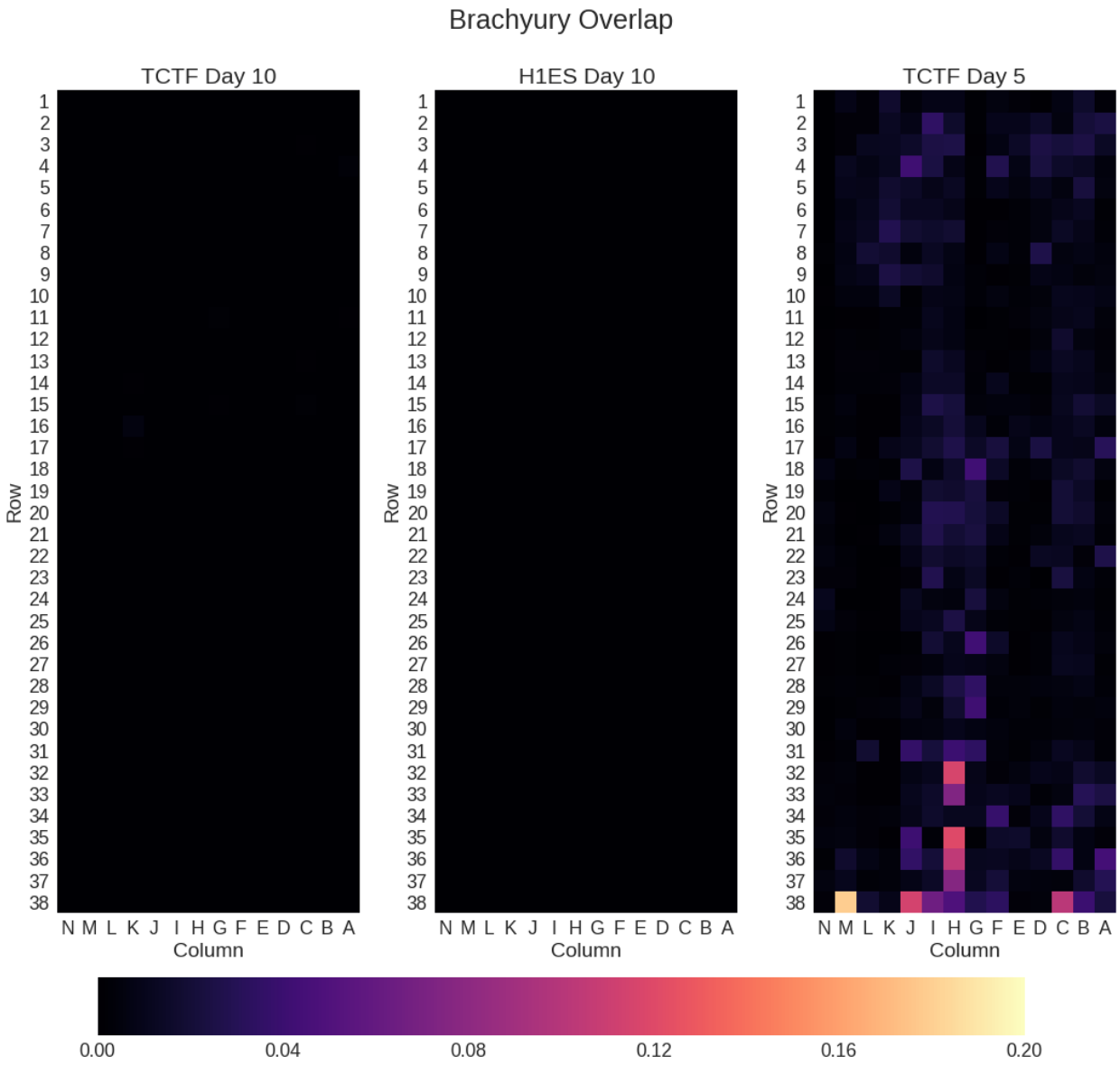
Heatmaps of mean cell population trends at the assay endpoint as approximated by Hoechst foreground area metric grouped by data set and plate position. Cell population position effects were consistently observed near the top and bottom edges of the culture plates.

Figure A.3: Cell population within plate trends at endpoint grouped by plate position



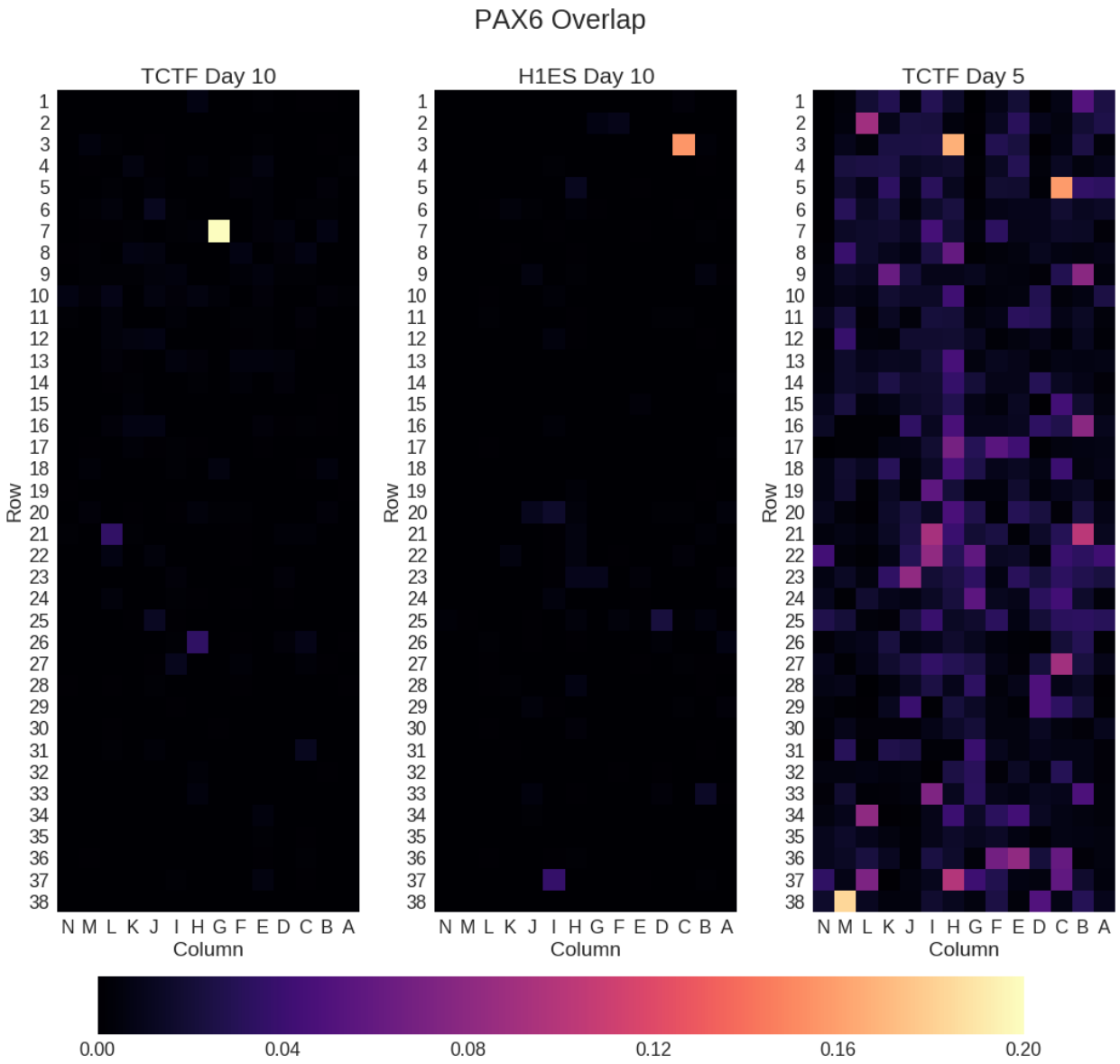
Heatmaps of mean pluripotency marker Oct4 trends at the assay endpoint as approximated by Manders overlap M1 metric grouped by data set and plate position. Both day 10 endpoint data sets (left, center) show markedly lower Oct4 expression relative to the TCTF day 5 (right) data set. The loss of pluripotency over days 6-10 was expected due to withdrawal of growth factors necessary for maintenance of pluripotency.

Figure A.4: Pluripotency within plate trends at endpoint grouped by plate position



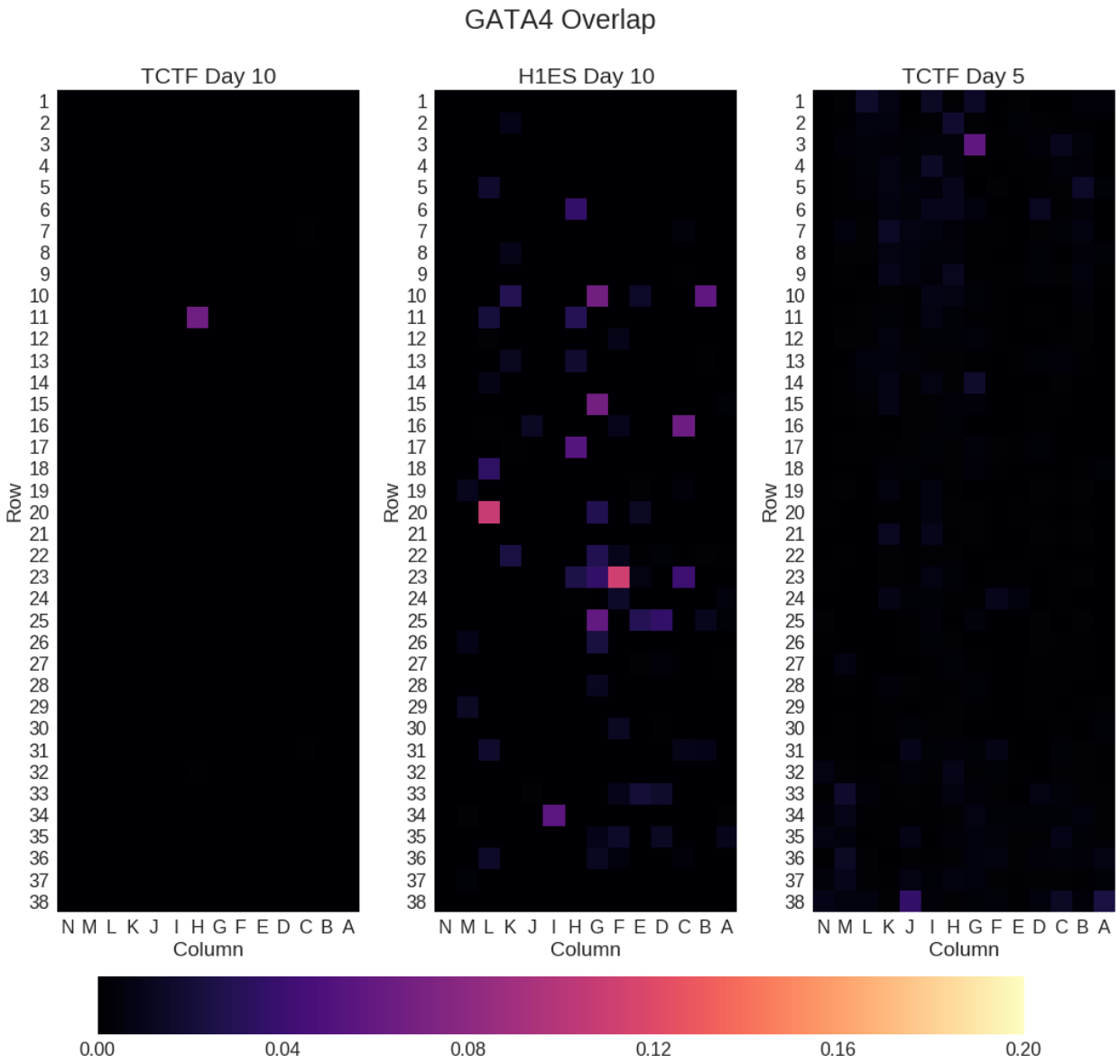
Heatmaps of mean mesoderm lineage marker Brachyury trends at the assay endpoint as approximated by Manders overlap M1 metric grouped by data set and plate position. Brachyury expression was neither expected nor observed in both day 10 endpoint data sets (left, center) due to an expected optimal expression endpoint around day 5, as tested in the TCTF day 5 (right) data set. Brachyury expression in the TCTF day 5 data set showed plate position effects within select plates. These effects are apparent in the streaks of increased expression levels from left to right and increased staining along the bottom edge in the plate summary heatmap.

Figure A.5: Mesodermal marker Brachyury expression within plate trends at endpoint grouped by plate position



Heatmaps of mean ectoderm lineage marker PAX6 trends at the assay endpoint as approximated by Manders overlap M1 metric grouped by data set and plate position. Both day 10 endpoint data sets (left, center) show low levels of PAX6 expression with occasional outlier wells while the TCTF day 5 (right) data set shows notable PAX6 expression with no clear plate position trends. PAX6 expression was expected by day 6.

Figure A.6: Ectodermal marker PAX6 expression within plate trends at endpoint grouped by plate position



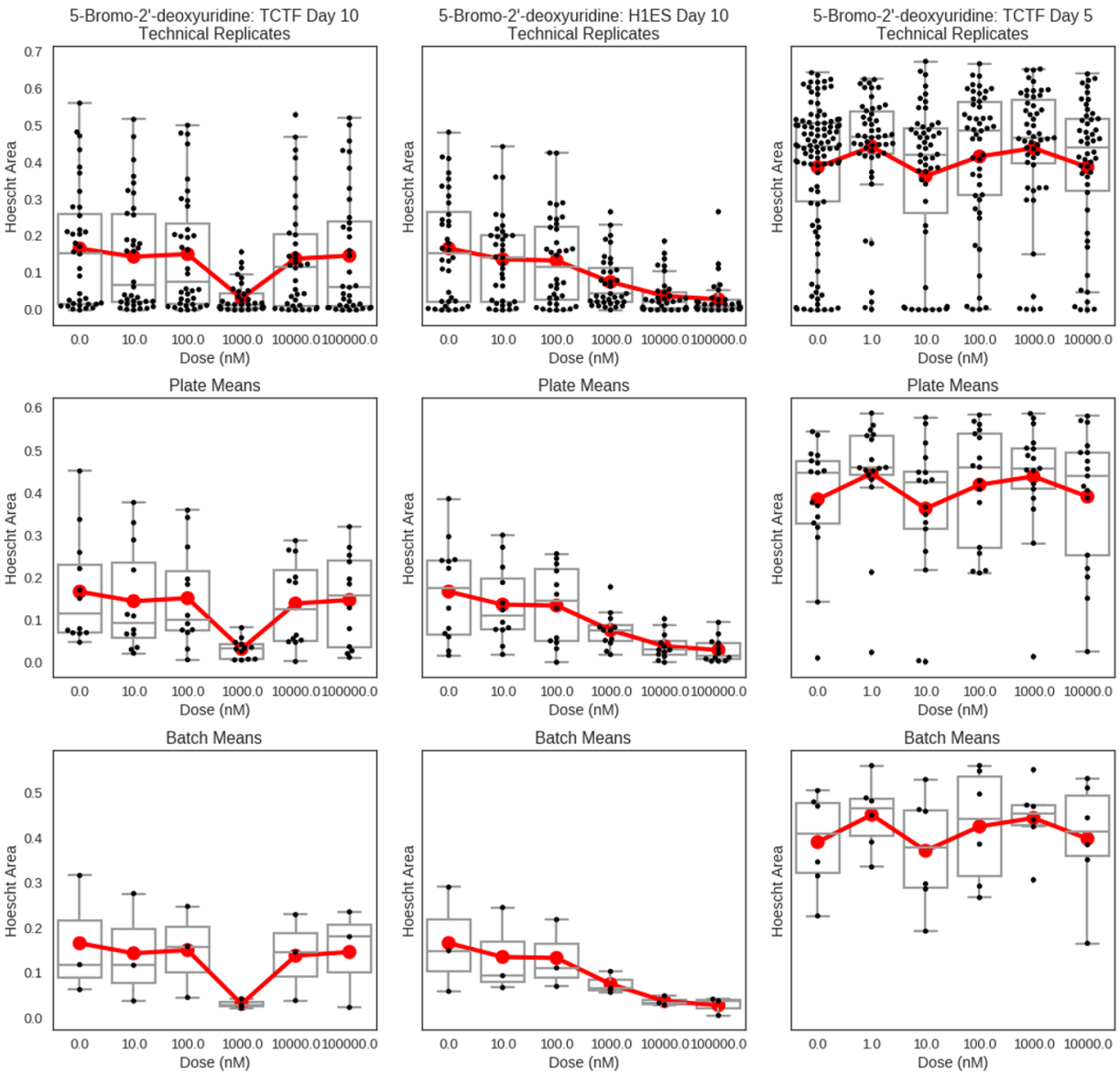
Heatmaps of mean endodermal lineage marker GATA4 trends at the assay endpoint as approximated by Manders overlap M1 metric grouped by data set and plate position. Only the H1 day 10 data set (center) shows notable GATA4 expression levels. GATA4 expression was expected to come up around days 7-9.

Figure A.7: Endodermal marker GATA4 expression within plate trends at endpoint grouped by plate position

APPENDIX A.

Dose-Response Variation and Result of Averaging

As shown in Figure A.8, the impact of the nested experimental structure on dose-responsive trends and variation was investigated in exploratory analysis by averaging results at the technical replicate, plate, and batch levels. Averaging replicates resulted in the same dose-responsive trends and similar estimated values but dramatically impacted the number and distribution of replicates. Fairly good agreement was noted for mean and median values especially in the H1 day 10 data set. As the technical replicate plots illustrate, many wells have few to no cells (Hoechst Area is approximately zero).



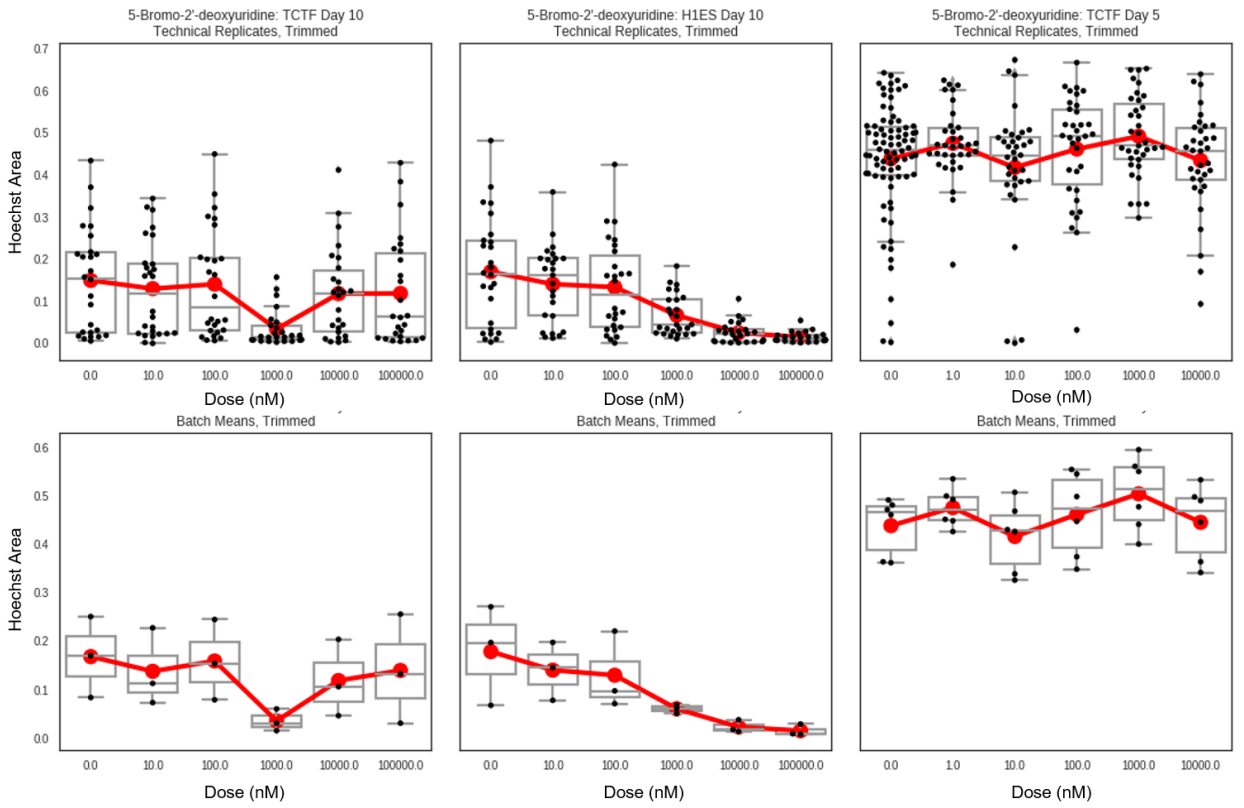
Exploratory analysis of hierarchical data for the BrdU dose-response curve. Black dots represent single technical replicates or wells (top row), means of technical replicates grouped by plate (middle row), or means of plate means grouped by batch (bottom row). Box plots show median and visually summarize dispersion and skewness of data. Red dots are mean values and red lines provide a visual aid to connect mean values between doses.

Figure A.8: Example analysis of hierarchical data structure for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area metric

APPENDIX A.

Outlier Analysis

Evaluating the impact of prospective outlying data is an important part of model selection due to the fact that outliers can have undue leverage on estimated model parameters. [79] Several outlier detection techniques such as the robust Hampel identifier were implemented to detect points of concern within groupings of replicates including experiment, batch, plate, and image measurement as well as within single experimental treatments (one molecule at one dose level). Skewness due to a relatively high frequency of zero values in both cell population (Hoechst area) and biomarker expression (Manders overlap M_1) metrics made the systematic detection of outlying data points unreliable for the methods tested. Therefore a simple trimmed mean approach was taken to evaluate the data before and after dropping the wells containing the most extreme high or low image measurements. Specifically, individual image measurements were assigned rank-order values for each data set. The ranks were then summed across image measurements and the wells with the top five and bottom five most extreme cumulative rank-values were dropped for each experimental treatment (one molecule at one dose level). This resulted in dropping 10 technical replicates (up to one third of replicate wells) for each experimental treatment. As Figure A.8 shows for the 5-bromo-2'-deoxyuridine Hoechst area metric, the estimated values and trends are largely conserved relative to the untrimmed data set (see Figure A.8) and some reduction in variation is achieved. Although a trimmed mean may yield modest improvements, the conventions are subjective and a more conservative approach was ultimately chosen in which all replicates were analyzed with a mixed effect modeling approach.



Analysis of hierarchical data for the BrdU dose-response curve after dropping prospective outlying wells. Black dots represent single technical replicates or wells (top row) or means of plate means grouped by batch (bottom row). Box plots show median and visually summarize dispersion and skewness of data. Red dots are mean values and red lines provide a visual aid to connect mean values between doses.

Figure A.9: Example analysis of hierarchical data structure for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area metric

A.4 Linear Mixed Effects Model Selection

Iterative Model Building

The final mixed effect model formula was selected via iterative model building. Model parameterizations were incrementally modified, and the Bayesian information criterion (BIC) was minimized for comparisons between proposed models. Likelihood-ratio tests were used to verify the significance of fixed effect terms relative to comparable null (random effects-only) model parameterizations. Fixed effects were ultimately specified to enable interpretation of the model intercept as the expected value of the untreated control and the slopes for each

treatment ID as contrasts (or effect sizes) for the difference due to the treatment relative to the untreated control. Since batches represent a progression of time, batches could have been modeled either as fixed or random effects. For simplification, batches were modeled as random effects parameters. This decision assumes that increased culture time would not result in systematic deviations in the observed outcomes —a reasonable assumption for pluripotent stem cell lines over the time frame of these experiments since hPSCs are expected to proliferate indefinitely in culture. If systematic differences in experimental outcomes were expected as a result of increased culture time, then batches would be more appropriately specified as fixed effects parameters. Due to within plate randomization of sample positions, the maximal random effects structure warranted by the experimental design was determined to include batch and plate effects. [75]

A.5 Comparison of Mixed Effects Model and Nonparametric Results

As a check to evaluate the validity of the linear mixed effect model results, estimated effect sizes and confidence intervals [80, 81] were compared to nonparametric bootstrap analyses of the same data. Two nonparametric analysis methods were explored: the bootstrap and stratified bootstrap methods, discussed below.

Bootstrap

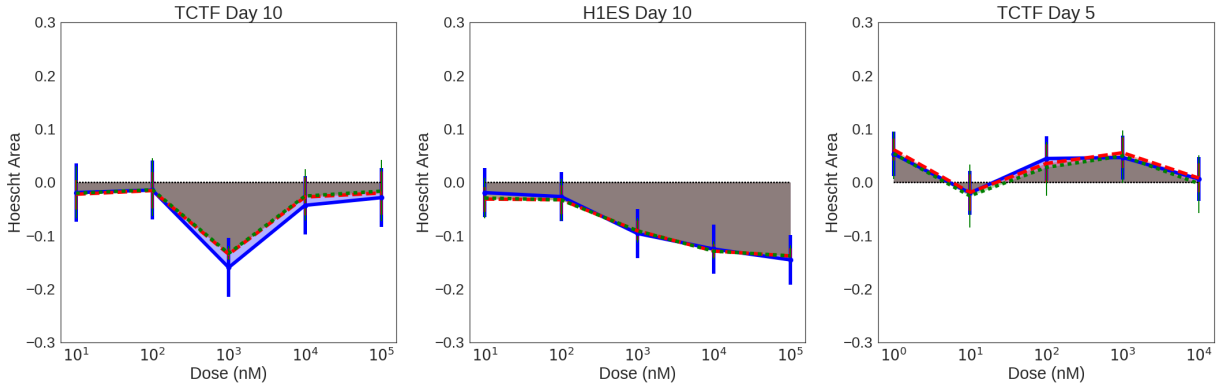
The bootstrap is a nonparametric analysis method based on random resampling of the data with replacement. [82] It does not require assumptions about the distribution shape or explicit assumptions of independence and can be used to calculate arbitrary statistics. In this approach all technical replicates were implicitly treated as independent by iteratively resampling the full data set and calculating the mean of each experimental condition. After ten thousand iterations, the estimated mean, standard error, and 95% confidence intervals of each experimental condition were calculated from the distribution of means calculated from the resampled data sets. The difference relative to the untreated control was calculated from the bootstrap-estimated mean values.

Stratified Bootstrap

In the stratified bootstrap approach, resampled technical replicates were grouped by plate then the overall mean for each experimental condition was calculated by calculating the plate means, batch means, and then overall means for each condition. As with the simple bootstrap, this process was repeated ten thousand times and the estimated overall mean values, standard error, and 95% confidence intervals of each experimental condition were calculated from the distribution of overall means calculated from the resampled data

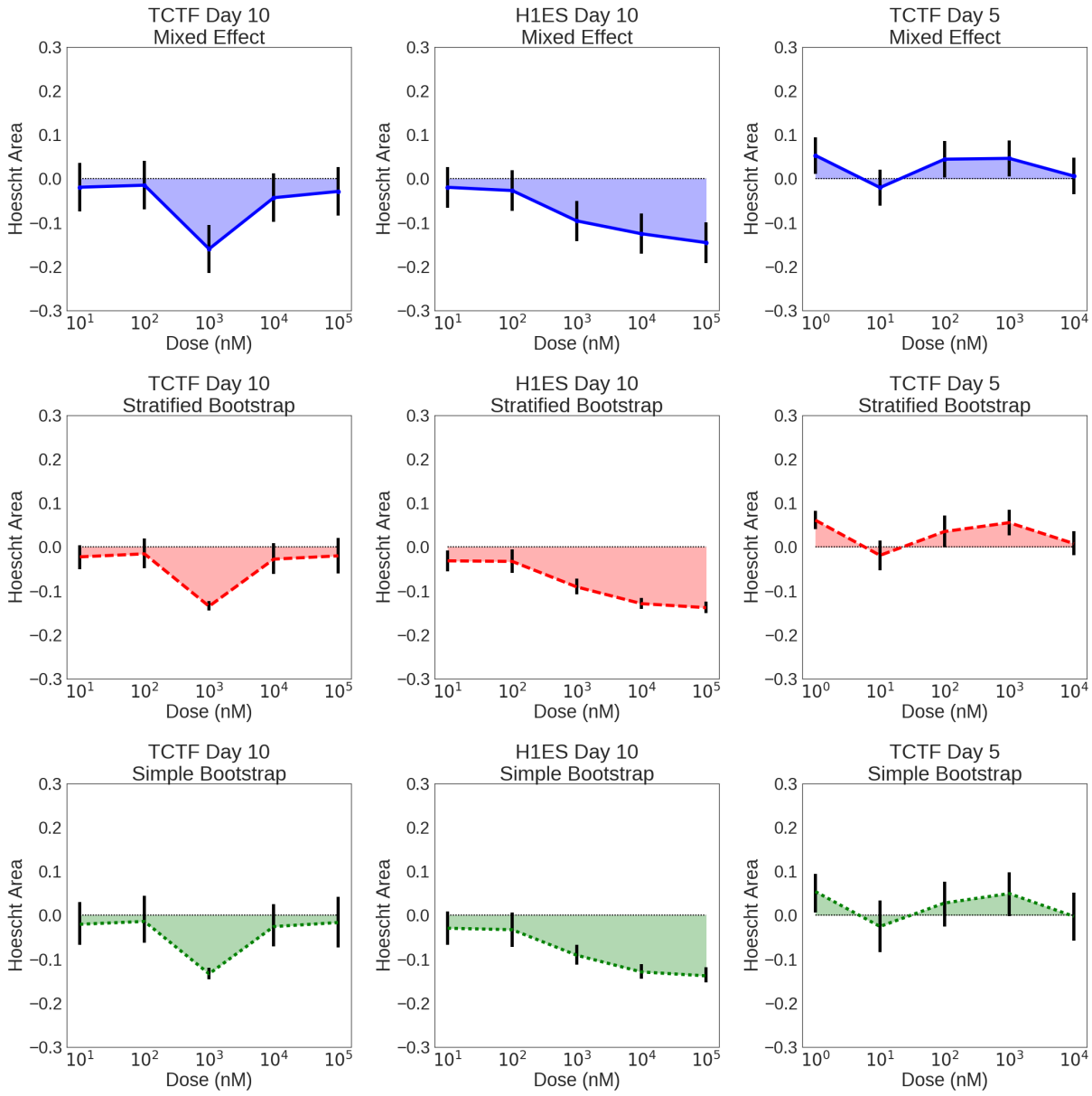
sets. The difference relative to the untreated control was calculated from the stratified bootstrap-estimated mean values.

Results



Overlay of parametric and nonparametric results for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area data shows close agreement between models on estimated effect sizes.

Figure A.10: Overlay of parametric and nonparametric analyses on sample control and dose-response data



Overview of parametric and nonparametric results for 5-bromo-2'-deoxyuridine (BrdU) Hoechst area data. Error bars represent 95% confidence intervals.

Figure A.11: Comparison of parametric and nonparametric analyses on sample control and dose-response data

A.6 Cytotoxic Dose-Response Curves

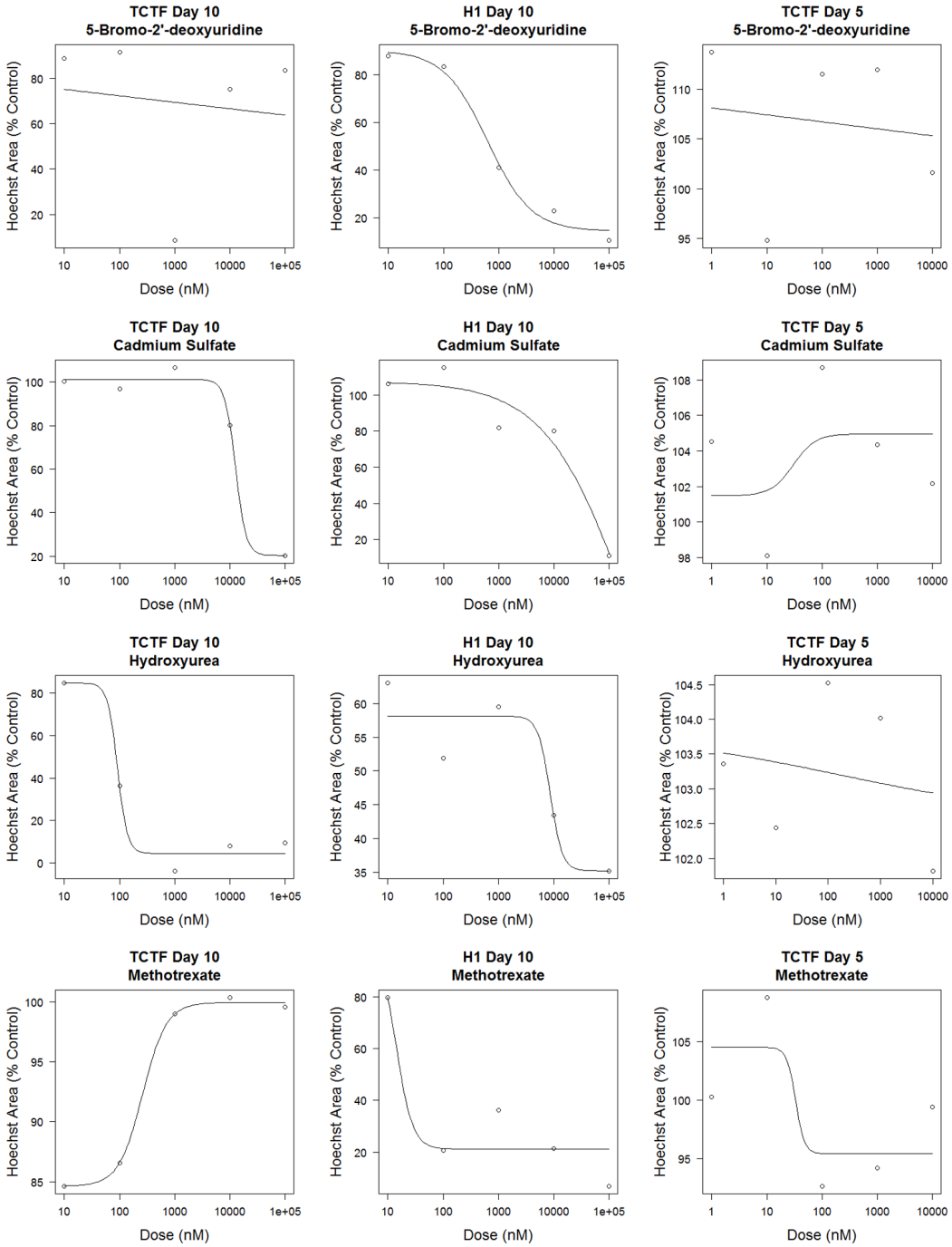


Figure A.12: Viability dose curves for Table 5.1