

The Effect of GATA6 Expression  
and Its Neighborhood Impact Factor  
on Regulating Cell Fate

by

Shaylina Rae Carter

A Thesis Presented in Partial Fulfillment  
of the Requirements for the Degree  
Master of Science

Approved March 2017 by the  
Graduate Supervisory Committee:

Mohammad Reza Ebrahimkhani, Co-Chair  
Samira Kiani, Co-Chair  
Mehdi Nikkhah

ARIZONA STATE UNIVERSITY

May 2017

## ABSTRACT

A genetically engineered line of human induced pluripotent stem cells was used to study the effects of gene expression on cell fate. These cells were designed to activate expression of the gene GATA6 when exposed to the small molecule doxycycline. This gene was chosen because it plays an important role in the developmental biology stages of liver formation. Because of the way the cells were engineered, a given population would have a heterogeneous expression of GATA6 because each cell could have a different copy number of the exogenous gene. This variation allows for the differentiation of multiple cell types, and is used to grow liver organoids. The early liver organoid samples were studied via immunofluorescent staining, imaging, and quantitative image analysis. It was originally hypothesized that absolute gene expression was not the most important factor in determining cell fate, but relative gene expression was. This meant that the spatial location of the cells and their local environment were critical in determining cell fate. In other words, the level of GATA6 of a cell is important, but so is the level of GATA6 in the surrounding cells, or neighborhood, of that cell. This hypothesis was analyzed with the creation of various Neighborhood Impact Factor (NIF) methods. Multiple time points of growth were analyzed to study the temporal effect, in addition to the gene expression and NIF influence on a cell's fate. Direct gene expression level showed correlation with certain cell fate markers. In addition to GATA6 expression levels, NIF results from early and late time point experiments show statistical significance with relatively small neighborhood radii. The NIF analysis was useful for examining the effect of neighboring cells and determining the size of the neighborhood –

how far cells influence one another. While these systems are complex, the NIF analysis provides a way to look at gene expression and its influence in spatial context.

## DEDICATION

To my mother, who has always believed in me. You have always impressed upon me the importance of education, and encouraged me to pursue the highest level of schooling possible.

And to my future, let this degree and experience prepare me for a bright and successful future.

## ACKNOWLEDGMENTS

I would like to give thanks to my committee and lab members – especially Jeremy Velazquez. Without your help this thesis would not have been possible.

I would also like to thank Val and Luke for proofreading my drafts, I don't know why either of you did it, and I'm not sure I would volunteer to do that myself, thank you very much for your comments and feedback.

And Dylan, thank you for always lending a free ear and supporting me throughout this journey.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	vii
LIST OF FIGURES .....	viii
CHAPTER	
1 INTRODUCTION.....	1
2 SCIENTIFIC BACKGROUND.....	5
GATA6 as a Cell Fate Decision Marker .....	5
Early Mesoderm Marker: BRACHYURY .....	6
Early Endoderm Marker: FOXA2 .....	6
Early Ectoderm and Pluripotent Marker: SOX2 .....	7
Endothelial-like Cell Marker: CD34 .....	8
Hepatocyte-like Cell Marker: CEBP $\alpha$ .....	9
3 METHODS .....	10
Genetic Engineering.....	10
Cell Culturing.....	11
Immunofluorescent Staining.....	13
Fluorescence Imaging .....	16
Image Analysis.....	17
Finalized Data Processing Pipeline .....	23
Spatial Visualization and Neighborhood Impact Factor Analysis.....	26
4 RESULTS .....	33
CellProfiler Pipeline Development.....	33

CHAPTER	Page
Image Cropping and Combination .....	35
Spatial Analysis Development.....	37
Early Ectoderm Markers .....	38
Early Time Point Experiments: FOXA2, T, GATA6 .....	40
Early Time Point Experiments: SOX2, T, GATA6 .....	44
Late Time Point Experiment: CEBP $\alpha$ , CD34, HA .....	48
5 DISCUSSION.....	53
Early Time Point Experiments: FOXA2, T, GATA6 .....	53
Early Time Point Experiments: SOX2, T, GATA6 .....	56
Late Time Point Experiment: CEBP $\alpha$ , CD34, HA .....	58
6 CONCLUSION.....	61
7 FUTURE WORK .....	64
Further Experiments.....	64
Software Development.....	66
REFERENCES.....	68
APPENDIX	
A T VS. GATA6 D0, D1, D3, D4.....	72
B FOXA2 VS. GATA6 D0, D1, D3, D4 .....	74
C FOXA2 VS. T D0, D1, D3, D4 .....	76
D DETAILED NIF RESULTS FOR DAY 1: FOXA2, T, GATA6 .....	78
E DETAILED NIF RESULTS FOR DAY 3: FOXA2, T, GATA6 .....	81

APPENDIX	Page
F DETAILED NIF RESULTS FOR DAY 4: FOXA2, T, GATA6 .....	84
G DETAILED NIF RESULTS FOR DAY 1: SOX2, T, GATA6 .....	88
H DETAILED NIF RESULTS FOR DAY 2: SOX2, T, GATA6.....	90
I DETAILED NIF RESULTS FOR DAY 5: SOX2, T, GATA6 .....	93
J DISTRIBUTION OF HA INTENSITY IN CEBP $\alpha$ + AND CD34+ CELLS .....	96
K DISTRIBUTION OF HA INTENSITY IN CEBP $\alpha$ + AND CD34+ CELLS (WITHOUT HA- CELLS) .....	99
BIOGRAPHICAL SKETCH.....	100



## LIST OF TABLES

Table	Page
1 Summary Table for Day 0-6 FOXA2, T, GATA6 Cultures .....	41
2 Summary Table for Days 1, 2, 4, 5, 6 SOX2, T, GATA6 Cultures.....	46
3 Summary of CD34, CEBPa, HA Expression.....	50

## LIST OF FIGURES

Figure	Page
1 Finalized Processing Pipeline in CellProfiler. ....	22
2 Total Expression IF Definition .....	28
3 Local Density IF Definition .....	29
4 Distance Reversed Expression IF Definition .....	29
5 Distance Adjusted Expression IF Definition .....	30
6 Primary Object Identification of AAT .....	33
7 Primary Object Identification of Nuclei .....	34
8 Relate Objects Module.....	34
9 Parent-Child Overlay .....	35
10 Color Composite of 4 Stain Images .....	36
11 Standard Crop Output .....	37
12 Spatial Plot for CEBP $\alpha$ , CD34, HA .....	38
13 Early Ectoderm Marker Test Results .....	39
14 Day 0-6 FOXA2, T, GATA6 Color Composites .....	40
15 Population Gene Expression for FOXA2, T, GATA6.....	42
16 Difference of GATA6 Expression in T+ and FOXA2+ Populations.....	42
17 FOXA2, T, GATA6 NIF Results .....	44
18 Day 1,2,4,5,6 SOX2, T, GATA6 Color Composites .....	45
19 Population Gene Expression .....	47
20 Difference of GATA6 Expression Between T+ and SOX2+ Populations.....	47
21 SOX2, T, GATA6 NIF Results.....	48

Figure	Page
22 Day 14 Color Composite of CEBP $\alpha$ , CD34, HA.....	49
23 Distribution of Cell Types Compared to HA Level.....	51
24 D14: CEBP $\alpha$ , CD34, HA NIF Results .....	52
25 Proposed GATA6 Cell Fate Model.....	62

## CHAPTER 1

### INTRODUCTION

Biological systems are complex and involve many signaling networks, regulatory pathways, environmental factors and cues, and temporal changes and interactions. As of late, there has been significant development and interest in human stem cells because of their ability to proliferate, and their pluripotent state which allows them to differentiate into any cell type within the body. With the discovery of the possibility of turning adult cells back into induced pluripotent stem cells (iPSCs) by Kazutoshi Takahashi and Shinya Yamanaka in 2006, the field of developmental biology with stem cells quickly became one of the most popular and rapidly advancing fields in biology (Takahashi & Yamanaka, 2006). Research with iPSCs has led to an expanding field of “organoid” development. Organoids are “3D tissue structures containing multiple cell types in an organized manner” (Nadkarni et al., 2015). Organoids have been created using iPSCs for many systems within the body: cardiac (Stevens et al., 2009), retinal (Eiraku et al., 2011), pituitary (Suga et al., 2011), small intestine (Spence et al., 2011), thyroid (Antonica et al., 2012), brain (Lancaster et al., 2013), liver (Takebe et al., 2013), stomach (McCracken et al., 2014), lung (Nadkarni et al., 2015), and kidney (Freedman et al., 2015). Most of these organoids were developed by influencing the stem cells’ environments with external growth factors, media, or physical forces or structures to produce a variety of cell types needed to form an organoid. Another common method is to differentiate cells into various cell types separately, and then recombine them to try and replicate the form and function of the desired organ. The approach used in the work presented here was

unique in that the mode of differentiation was from gene activation (internal), rather than from growth factors or supplements (external).

The cell lines used in the experiments conducted for this research had been genetically modified to include a small molecule activation switch for the gene GATA6. This meant that the regulation of differentiation for multiple cell fates was determined from gene expression levels. The internal regulation activated signaling pathways which regulated cell differentiation, more like what happens in biology than the “external” methods mentioned previously (Guye et al., 2016). The pathways that were activated include positive and negative feedback loops which affect how cells communicate with one another.

This concept of cell communication led to the investigation of something known as the “community effect.” The community effect is centered around cell-cell interaction. It has been shown that genetically identical cells can often show gene expression profiles that are significantly different from one another even when cultured in the same physiological condition (Saka et al., 2011). Within these populations, it is also possible to find homogenous groups of cells. One explanation for the formation of these homogenous populations is cell-cell interactions. Theorized definitions of the community effect claim that neighboring cells may override a given cell’s fate, and cause it to conform to the environment if everything around it is part of a different, but consistent cell type. This idea has been shown in muscle precursor cells of *Xenopus* embryos. Ectoderm cells changed their fate to mesodermal cells when placed in an environment that promoted mesoderm signaling. This work led to a theory of the community effect that tried to model this behavior with a mathematical system (Saka et

al., 2011). While this theory of community effect centers around the idea that homogenous environments can overpower a given cell's identity to conform to said environment, it doesn't address the idea of cellular expression in various areas of a cell culture or environment.

The idea of cell-cell communications, coupled with the idea that gene expression guides cell differentiation, led to the one of the founding questions for this thesis: what if not only absolute gene expression is important, but expression compared to its surrounding environment (relative expression) is also important? This work is focused on the concepts of absolute gene expression – specifically GATA6 expression – with regards to cell fate and relative gene expression and its influence on cell fate. Gene expression was quantified using immunofluorescent staining, imaging, and numeric image analysis techniques. Relative gene expression was assessed using the newly defined concept of “Neighborhood Impact Factor” analysis, which quantifies the surrounding genetic expression of a given target cell and compares that with the cell's fate. The main idea behind this analysis was that a cell of medium GATA6 expression might have the possibility to diverge to more than one cell fate, and one of the deciding factors might be whether it is in an environment where it has “high” expression, or in an environment where it is considered to have “low” expression.

Cells were cultured using the liver organoid procedure presented in a paper published in Nature Communications by Dr. Ebrahimkhani and his collaborator Patrick Guye (Guye et al., 2016). GATA6 gene expression and various cell markers were quantified in “early” time point experiments (between 0 to 6 days in age), and late time point experiments (either 10 or 14 days old). The main goal was to compare both direct

gene expression and relative gene expression to cell fate. This work resulted in the development of a prototype of an analytical software tool for assessing Neighborhood Impact Factors.

## CHAPTER 2

### SCIENTIFIC BACKGROUND

#### GATA6 as a Cell Fate Decision Maker

During blastocyst formation, the inner cell mass (ICM) differentiates into two distinct cell types – epiblast (Epi) and primitive endoderm (PrE) cells – which are organized in a salt and pepper pattern. Extensive research has been done to investigate the most important regulating factors during this stage and has shown that in mice *Nanog* or *Gata6* repression can eliminate the formation of either cell type. Without *Nanog* all ICM cells become PrE cells, and without *Gata6* all ICM cells become Epi cells (Bessonard et al., 2014). Similar work has also been performed in embryos and embryonic stem cells and has shown that transcriptional networks based on *NANOG* define the Epi cell fate, and a network based on GATA transcription factors defines the PrE cell fate (Schröter et al., 2015). Research into the GATA family of transcription factors has discovered that the first sub group (GATA1, -2, -3) plays a significant role in hematopoietic cell development, and the second sub group (-4, -5, -6) is predominantly found in the heart, digestive system, and extraembryonic endoderm. It was also found that there is a close link to GATA6 expression and BMP4 expression. It seems that GATA6 plays an important role in promoting complex regulatory signal systems such as that of BMP4 (Nemer & Nemer, 2003), ERK (Bessonard et al., 2014), and FGF/MAPK (Schröter et al., 2015).



### Early Mesoderm Marker: BRACHYURY

Brachyury (also known as T) is an early transcription factor that is characteristic of mesendoderm. Research with human embryonic stem cells (hESCs) has discovered that BMP4 works with FGF2 (via ERK) to promote mesoderm differentiation and prevent endoderm differentiation. This system produced high expression levels of T with co-expression of CDX2; however, T expression preceded and was necessary for the expression of CDX2. Both markers were required for the expression of mesodermal genes. However, BMP4 expression produced mesoderm differentiation and the expression of T, TBX6 and CDX2 (Bernardo et al., 2011). T also interacts with another signaling pathway known as SMAD. It has been demonstrated that by controlling and altering levels of T in cells from different signaling environments, that T is essential for mesoderm but not endoderm formation (Faial et al., 2015).

### Early Endoderm Marker: FOXA2

Once hESCs are expressing mesendodermic markers such as T and other common genes, continued differentiation of these cells will move to definitive endoderm and express SOX17 or FOXA2 (Bernardo et al., 2011). FOXA2 is part of the forkhead box family of liver transcription factors. It has a high level of expression in the liver, and is found at lower levels in other tissues. Studies have shown that knocking out FOXA2 in mice will lead to their death during embryogenesis due to developmental defects, so it is an essential transcription factor for development (Kanaki & Kardassis, 2017). Experiments have shown that Activin treatment of hESCs activated the differentiation to endoderm and promoted the expression of SOX17 and FOXA2 (Bernardo et al., 2011).

Both T and FOXA2 are markers expressed during early development which makes them suitable for image analysis in early time point organoid culture experiments.

#### Early Ectoderm and Pluripotent Marker: SOX2

The Sox gene family was derived from the discovery of the mammalian testis-determining factor, Sry. Sry has a characteristic high-mobility-group (HMG) domain that binds DNA in a sequence-specific way. Proteins that contain an HMG domain with 50% or higher similarity to the domain of Sry are considered Sox proteins (Sry-related HMG box). There are 20 Sox genes present in humans and mice. Sox proteins that share more than 80% of an HMG domain sequence identity are classified into eight groups called A through H. Factors within groups share biochemical properties and have similar biological functions. Factors from different groups fulfil different biological functions even though they recognize the same DNA consensus motif. Sox2 is part of group B, and has been used to study its influence on stem cell decision making and cell fate (Sarkar & Hochedlinger, 2013).

Like FOXA2, Sox2 deletion in zygotes results in early embryonic death. Sox2 is crucial for forming the pluripotent epiblast in blastocyst formation. Sox2 seems to be important for helping stem cells maintain pluripotency, and it appears that it functions in tandem with a few other factors such as Nanog and Oct4. A large portion of genes targeted by Sox2 have composite Oct4/Sox2 consensus binding sites which implies that Sox2 works closely with Oct4 to maintain efficient binding and recruit factors for gene activation. Research that supports this concept has shown that a loss of Sox2 can be partially compensated by the overexpression of Oct4 (Sarkar & Hochedlinger, 2013).

During ESC differentiation Sox2 is expressed in the earliest stages when a cell is becoming neural ectoderm. Sox2 promotes this fate by working directly against T – the early mesendoderm marker – by suppressing it. Sox2 seems to play a significant role in central and peripheral nervous system formation, and is in no way related to endoderm or mesoderm expression (Sarkar & Hochedlinger, 2013).

#### Endothelial-like Cell Marker: CD34

For an organ or organoid to function properly, multiple cell types and germ layers are needed. Vasculature is an obvious necessity for any tissue or organ. Blood vessels are part of the mesoderm, and are a highly integral piece of any working organ. The networks play crucial roles in constructing and reconstructing tissues, promoting wound healing, and delivering metabolic requirements to organs in the body. Being able to replicate these systems in the lab is fundamental for creating applicable human tissues and organoids.

Primitive structures first appear in early development and help guide the differentiation and organization of other cell types. Three-dimensional stem cell culture methods have helped to simulate vasculature development in lab settings. Their significance in research and treatments for diseases from arthritis to cancer, have prompted scientists to work towards a more complete understanding of their development (Lowenthal & Gerecht, 2016).

It is important to have an accurate marker for recognizing vasculature development so that the influence of GATA6 can be properly assessed. Research has shown that CD34 is one of the best-characterized human hematopoietic stem cell-related

antigens (Kanaya et al., 2015). CD34 is a surface marker, which makes it great for visualizing cell cultures during staining, and makes vascular structures easily identifiable.

#### Hepatocyte-like Cell Marker: CEBP $\alpha$

Identifying hepatocyte-like cells is a critical part of monitoring cell fate during the organization of a liver organoid. These cells are derived from endoderm cells. The marker used in the late stage experiments was CCAAT/enhancer binding protein, alpha CEBP $\alpha$ . Studies have shown that differentiated fibroblasts can be reprogrammed into hepatocyte-like cells with the use of biological mixtures which use CEBP $\alpha$  as one of the main factors (Pournasr et al., 2015). CEBP $\alpha$  has been identified as a key hepatocyte transcription factor. It is highly expressed in healthy liver cells (as opposed to some transcription factors such as CEBP $\beta$  that are expressed during liver injury and regeneration). It has also been shown that CEBP $\alpha$  plays a role in guiding cell differentiation into a hepatocyte-like fate (Jakobsen et al., 2013).

## CHAPTER 3

### METHODS

#### Genetic Engineering

Cell lines were specifically engineered to study the influence of GATA6 expression on liver organoid development. This genetic modification was done using lentiviral mediated transgene integration. These cells were first created and used for research that was published in Nature Communications about the self-organization of stem cells for liver organoids using GATA6 (Guye et al., 2016). Cells were lentivirally transduced with a doxycycline inducible-GATA6 cassette and a doxycycline (dox) inducible activator (rttA3). Because of lentiviral transduction, copy numbers of this genetic modification varied in the cell population. This important feature meant that a diverse population of GATA6 expression would be produced even when the entire population was exposed to a consistent level of doxycycline. The exogenous GATA6 would only be activated in the presence of dox, so as soon as doxycycline was removed, these exogenous genes would become inactive. This controllability meant that exogenous GATA6 could be turned on or off at will for varying time points of a cell culture.

Another version of this gene modification was engineered to also add human influenza hemagglutinin (HA) as a tag to GATA6. This cell line was used in a few of the preliminary experiments for measuring the influence of GATA6 on cell fate. This tag could be used to identify which cells had the genetic modification incorporated in their genome, discern between exogenous and endogenous GATA6, and could be used to roughly correlate gene copy number as well. However, this tag only measures the

exogenous gene expression, it does not take into consideration any endogenous GATA6 a cell might express. Staining for GATA6 measures both endogenous and exogenous gene expression, and can be used to identify which cells only express endogenous GATA6 when co-stained and analyzed with HA.

## Cell Culturing

For these experiments cell culturing involved several different cell lines. The most commonly used cell line was PGP1, but C1 cells were also used for additional verification. In addition to the regular PGP1 cell line, some experiments used an HA-tagged PGP1 cell line as well. These cells had HA tagged to the exogenous GATA6 gene that was transfected into them. This cell line was useful for comparing exogenous to endogenous GATA6 gene expression.

All cell cultures used for imaging were cultured in 24-well plates on 12 mm glass coverslips. These cultures were seeded from 6-well plate maintenance wells that had media changes daily, and were passaged every 4 to 5 days as they became confluent. The plates were cultivated in an incubator at 37°C and 5% CO<sub>2</sub>. The maintenance wells were fed with mTeSR-1 to keep them in a pluripotent, non-differentiated state. The passaging processes began by first preparing the wells and coverslips with a layer of Matrigel that the cells would adhere to; this coating was left to sit for one hour at room temperature before cells were added to it. The maintenance wells were then passaged by aspirating the media from the plate and exposing the cells to Accutase for 4-7 minutes in the incubator so they would detach from the surface. During incubation Rho-associated protein kinase (ROCK) inhibitor was added to DMEM F12 media at a 1:3000 ratio

(10 $\mu$ M of ROCK inhibitor). The cells were then lifted from the plate in the Accutase into a new vial containing the ROCK inhibitor-DMEM F12 solution. This mixture was spun down in a centrifuge at 300g for five minutes to make the cells converge at the bottom of the suspension. The media was aspirated and the cells were re-suspended in the ROCK inhibitor-DMEM F12 solution to survive in temporarily. Next, 10  $\mu$ L of this mixture was extracted to use in cell counting. The density of cells in the media was calculated and the cells were then reseeded in a new 6-well plate at a density of 200,000 cells per mL. These cells were then seeded with a 10  $\mu$ M ratio of ROCK inhibitor to mTeSR-1 to reset the confluency for new maintenance wells. If any new experiments were beginning, these plates were also seeded with the same cell density. The following day ROCK inhibitor was removed from the media when feeding the cells. Passaging the maintenance wells took place approximately every four days.

Once cells were seeded for a new experiment, they were dosed with doxycycline (dox) at a 1:1000 ratio with the media for five days following the first day after seeding. The day of seeding was considered day -1, and dox induction began the following day on day 0. During this initial five-day period mTeSR was used to sustain the cells. After day 5 the media for the cells was switched to APEL and the dox was not added to the media. Cells were cultured until as late as day 14 and as early as day 0 before fixing for staining. The fixing process consisted of adding 4% paraformaldehyde to the wells at room temperature for 20 minutes. Next, the wells were washed with PBS for 5 minutes; this wash was repeated 3 times. Finally, the coverslips were removed from the well they were cultured in and added to another well with PBS to sit in the fridge until they were used for staining.

Early stage cell cultures (days 0-6) were induced with dox for the first 5 days. Day 0 stains did not receive dox since they were fixed on the same day that dox was added to all the wells. Late stage cultures (days 14) all had some form of dox redosing so that the endogenous GATA6 would be activated before fixing. The PGP1 CEBP $\alpha$ , CD34, HA stain that was used in early analysis was re-dosed on day 12 with a 48-hour dosing of dox, and fixed after that 48 hours on day 14. In addition to the PGP1 cell line, cell cultures were grown from the C1 cell line. C1 cells were grown and fixed at ages of day 3 and 5.

### Immunofluorescent Staining

All cell cultures used in this experiment were analyzed by immunofluorescent staining. Before the staining process could begin, a plan was necessary to outline the primary and secondary antibodies that would be used. This was critical because each primary antibody is associated with an animal host, and each secondary antibody is associated with a color. The animal hosts must be unique for each antibody so that there is a 1:1 ratio for each imaging band to the antibody. The animal hosts that were used were rat, goat, sheep, and mouse. Unfortunately, goat and sheep could not be used together because it was known from previous experience that these two hosts caused significant cross-binding. These primary antibodies were purchased from either Abcam, Santa Cruz, or R&D Systems. The secondary antibodies had these hosts in each of the three excitation colors: green (488 nm), red (594 nm), and far red (647 nm). The DNA stain (either DAPI or Hoechst) was always imaged in blue (461 nm). Each stain could either be a nuclear, cell surface, or cytoplasmic stain. Sometimes it was possible that



there was bleed over from the red channel to the far-red channel, so if possible it was these channels had different types of stains on them. For example, red would be a surface stain and far red would be nuclear, that way the bleed through wouldn't be an issue when detecting objects in the far red. However, a good portion of the experiments used all nuclear stains because they were the easiest and most reliable to detect with the image processing pipelines. Furthermore, all stains (primary and secondary) had to be tested prior to their use in cell culture experiments, to verify that they were working and accurate.

The staining processes began with a few preparation steps. The 24-well plate containing the coverslips – that were suspended in PBS – was removed from the fridge. An area on a lab bench was cleared and cleaned with ethanol and a strip of parafilm was taped to the bench. Parafilm was used because it was clean and hydrophobic, which helped the solutions stay in droplets for the coverslip to soak in. It also aided with cleanup and prevented contaminants from entering the staining solutions.

The first step of staining was to transfer the coverslips from their face up position in the PBS suspension in the 24-well plate to a face down (side with adhered cells face down) position in a 100  $\mu$ L solution of PBS 0.2% Triton X-100 for 15 minutes. The purpose of this step was to permeabilize the cells so that the antibodies would be able to enter the cell membrane and bind with the appropriate proteins or transcription factors. Next, the coverslips went through a washing stage where they were placed in 200  $\mu$ L of 0.05% PBS Tween-20 for 5 minutes. This wash was repeated three times to remove the Triton from the cells.

After the washes, a blocking step was executed. This consisted of placing the coverslips in 200  $\mu\text{L}$  of 10% donkey serum in 0.05% Tween-20 for 20 minutes. Donkey serum was used because it is the host serum for all the secondary antibodies that were used. This step is necessary to prevent non-specific binding of the secondary antibodies, so that what is detected in the florescent images is only from the binding of the secondary to the primary antibody.

After the blocking step, the primary antibody stage began. The coverslips were placed directly in 30  $\mu\text{L}$  of the primary antibody solution. This solution was created by taking 1  $\mu\text{L}$  of each primary antibody and mixing it with 70  $\mu\text{L}$  of the 5% donkey serum in 0.05% Tween-20 solution. Then the three antibodies were combined so that the resulting solution was a 1:210 ratio of primary antibody to 5% donkey serum Tween-20. From this mixture 30  $\mu\text{L}$  were used for each of the coverslips to be stained with this combination of primary antibodies, and the coverslips stayed in this solution for 60 minutes.

Immediately following the primary antibody stage, the coverslips were put through another three washes in 200  $\mu\text{L}$  of PBS 0.05% Tween-20 for 5 minutes each. The purpose of these washes was to remove any non-binded primary antibodies that might still be in the cells or on the coverslip. The next step was to place the coverslips in the secondary antibody solution. Starting with this stage and lasting until the end of the process, all lights were turned off, and the coverslips were covered with a lid during the soaking so that the light did not degrade the secondary antibodies. This solution was created by mixing the secondary antibodies at a 1:400 ratio with the 5% donkey serum-Tween solution. Each of the secondary antibodies (one for each color green, red, and far

red) was mixed at this ratio and then combined. The combined product was used in 30  $\mu$ L droplets for each coverslip. The coverslips stayed in this solution for 60 minutes.

The first wash after staining with the secondary antibodies was in 200  $\mu$ L of a PBS 0.05% Tween-20 solution with DAPI or Hoechst added at a 1:1000 ratio. This soak lasted 10 minutes. The purpose of DAPI and Hoechst is to stain for DNA, meaning that every cell with an intact nucleus would be visible when imaging for this channel. After the DNA stage was finished, two washes of 200  $\mu$ L of Tween were performed for 5 minutes each.

Finally, each coverslip was added to part of glass slide using a 10  $\mu$ L drop of Diamond Antifade reagent. The Diamond Antifade was kept in storage at  $-20^{\circ}\text{C}$  when not in use. The glass slide was then left to dry on a level surface overnight at room temperature before the coverslip was sealed around the edges with nail polish. The purpose of this last step was to keep the coverslip in an airtight state so that the primary and secondary antibodies don't degrade or diffuse throughout time. These glass slides were labeled with numbers that corresponded to a key describing what cell line was used, how old the cells were, and what primary and secondary antibodies were used.

### Fluorescence Imaging

The images were collected using a Leica DMI6000 microscope. The 12mm coverslips were centered and images were collected at an exposure of 627 ms for each channel. Images were collected in frame by frame shots at 10x magnification. The frames were then processed and stitched together by the microscope to produce full coverslip images at 10x resolution. This process was done for each of the four emission

bands. The microscope imaging procedure took anywhere from 15 to 30 minutes per coverslip sample.

## Image Analysis

The main analytical tools used for image analysis in this project were CellProfiler version 2.2.0 (CP), CellProfiler Analyst version 2.2.1 (CPA), SQLite, ImageJ version 1.50i, Java version 1.8 (the computer programming language), and R version 3.32 (a statistical analysis language). CellProfiler and CellProfiler Analyst are free, open-source software programs developed and maintained by the Carpenter Lab at the Broad Institute of Harvard and MIT. They are designed to quantitatively measure and analyze phenotypes from large image data sets. SQLite is a database format that can be produced as an output from a CellProfiler pipeline and is needed as an input for CellProfiler Analyst. These files were instrumental to the numeric analysis of the images (both within CPA and from custom designed Java programs). ImageJ is a free image editing software that allows the images to be manipulated in ways that makes them easier to visualize – for example, adjusting the exposure or contrast, merging multiple bands to create color composites of various stain combinations, and manually cropping or rescaling images. Lastly, Java was used with the integrated development environment (IDE) Eclipse to write programs to automate cropping and file generation of the sub images for analysis, produce summary files of image statistics, and to quantify, develop, and enhance the Neighborhood Impact Factor analysis.

The initial iterations of cellular image analysis consisted of manually selecting and cropping sections of images of arbitrary size. The first images to be analyzed were a

set of stains produced from a culture stained for Alpha-1-Antitrypsin (AAT) and DAPI (DNA stain). This image set was used to develop the first CP pipeline which identified, measured, and exported information about objects within the images.

One of the most important aspects of the CP pipelines that would be used in the later experiments was the identifying of “primary” objects using the CP module named “IdentifyPrimaryObjects”. CellProfiler defines a primary object as something that can be identified first, and then possibly related to other objects to form various relationships such as parent/child, secondary objects, and tertiary objects. In this image set, AAT is a cytoplasmic stain, so the objects looked like approximately round shapes with holes in them where the nuclei were located. The IdentifyPrimaryObjects module developed in this pipeline for AAT was useful for examining other cytoplasm stains used in the future experiments. This cytoplasm module was also particularly useful because it was more complex than the nuclei stain. It was more difficult to create and refine, therefore it provided a better learning dataset for how to use CellProfiler. The IdentifyPrimaryObjects module used to examine nuclei was very helpful in all future experiments, because all experiments either used DAPI or Hoechst to identify and measure all cells present in the well.

In addition to identifying the primary objects in all the channels (in this first case just AAT and DAPI were used), the objects were then always related to their parent nuclei. This was important because stains for multiple markers were used in the same samples, resulting in the possibility that either one, two, or none of the markers could be present in each identified nuclei (parent). This relationship guaranteed that every object was related to a nucleus and therefore belonged to an actual cell, and that all objects

(from every channel) related back to a corresponding nucleus. This was important when trying to compare the expression level of multiple markers, because they may or may not be expressing in the same cells, and knowing that distinction is valuable.

The skills gained from developing this pipeline were quickly transferred to more complex image sets. The next step of development came from analyzing two separate image sets. One was stained with CEBP $\alpha$  (a hepatocyte marker), HA (a tag for exogenous GATA6), and DAPI. The other was stained with CD34 (an endothelial marker), HA, and DAPI. In these image sets DAPI, HA and CEBP $\alpha$  were all nuclei stains, and CD34 was a surface stain. Primary object identification modules were developed for HA, CEBP $\alpha$ , and CD34. These pipelines also included the next step, which was to extract measurements from these objects. Because of these measurements, it was possible to produce graphs comparing the HA intensity values and either CD34 or CEBP $\alpha$ . The data produced from the analysis also allowed the comparison of the size and percent of the populations that were either CD34 or CEBP $\alpha$  positive, and their corresponding HA levels.

The modules developed from this pipeline were immediately applicable to the next experiment which compared CD34 and CEBP $\alpha$  in the same cell culture. This image set contained all four stains on individual channels: CD34, CEBP $\alpha$ , HA, DAPI. ImageJ was used to combine the four bands into a useful color composites for viewing the stains in spatial context with one another. In this CP pipeline, an extra normalization step was applied to the CD34 and CEBP $\alpha$  objects. Once they were related to their parent nuclei, their intensity was normalized by the nuclei intensity. This was to help account for any z-depth intensity loss, or any other abnormality that would cause a section or sections of

the coverslip to appear dim. This pipeline also produced a database file which CPA could access and analyze to produce scatter plots, density plots, and text based statistics. The pipeline also produced CSV data files that could be opened in Excel to produce similar or additional products. This experiment led to the “standard” pipeline that was used in all further experiments for this project.

The final template for the CellProfiler pipeline used to analyze all the images for this project also contained a beginning step for purifying the DAPI or Hoechst (another nuclei stain that was used in some of the experiments) channel. This step was called “despeckling”, and its purpose was to remove small bright spots that were not caused by the stain. To do this I first created a module (called “EnhanceOrSuppressFeatures” in CP) which identified the brightest and smallest spots in the DAPI images. Then I used this output image to subtract from the original DAPI image to produce a “despeckled” DAPI image. This product is what was then used for the rest of the CP pipeline.

The “IdentifyPrimaryObjects” module for the DAPI stain used an adaptive strategy for thresholding (identifying the objects). An adaptive method was chosen for identifying nuclei instead of a global method, because different parts of the image may have different lighting or staining conditions (parts of the image might be lighter overall, compared to other parts of the image). So, an adaptive strategy helps account for that difficulty by breaking up the whole image into smaller pieces – in this case 50x50 pixel pieces – and applying the thresholding method to each piece. The thresholding method that was chosen for identifying DAPI nuclei stains was called Otsu’s method, named after Nobuyuki Otsu. The method calculates the threshold by dividing

into two classes – background and foreground. It does this by minimizing the variance within each class. CellProfiler has an additional option for dividing the pixels into three populations, and then deciding whether the middle population belongs to the foreground or the background. For the DAPI modules I only needed to use the two-population classification because the stain is clear and bright. The other defining factor for the DAPI module was to set the diameter range of the objects from 8 to 30 pixels, which is 1.6 to 24 microns. This was the standard module used for all experiments on the DAPI and Hoechst channels. For identifying other nuclear stains after this (T, FOXA2, ERG, CEBPA), instead of an adaptive method, I used a “per object” method. This method used the areas already defined by the nuclei stain previously to start the search for new objects in its respective channel. This was more accurate than applying an adaptive method, because these stains are all nuclear so they must correlate directly with a nucleus. Also, the objects get related to their nuclei parent later in the pipeline, so this helped make the pipeline more efficient and accurate.

All image sets would go through a pipeline which identified the primary objects in all channels, related those objects to their parent nuclei, measured the intensity of all objects, normalized those values based on the DAPI channel, and produced the necessary output files (SQLite database file and CSV files). This data set was also useful in developing another plot from the spatial information of the cells. The plot depicted individual points for every cell in the image, that was colored according to which population it fell in, and sized (radius of the point) based on the HA level of that cell. This product was the first step at trying to visualize and analyze the spatial importance of neighboring cells on a given cell’s end fate.



The pipeline used in analysis for all the experiments used for neighborhood

analysis consisted of the following modules in the following order:

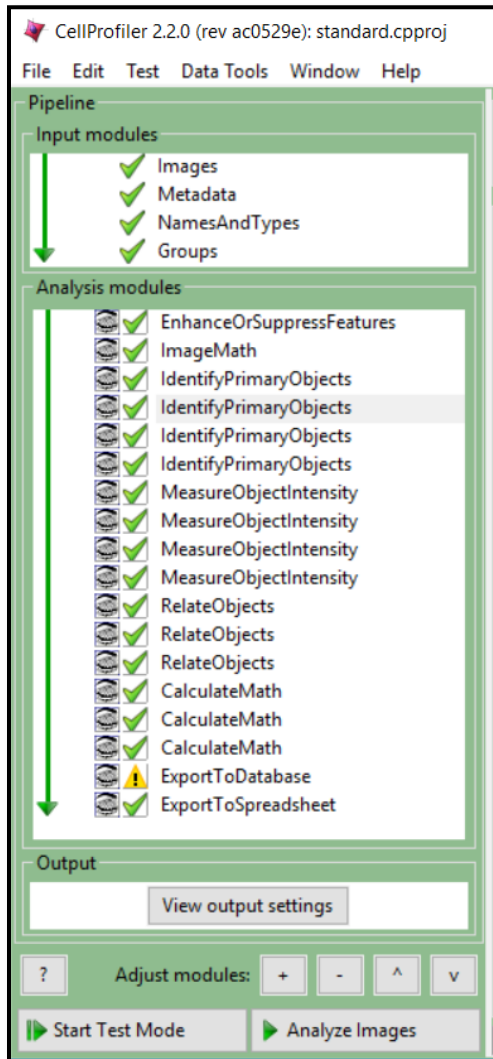


Figure 1 Finalized Processing Pipeline in CellProfiler.

“EnhanceOrSuppressFeatures”, “ImageMath”, “IdentifyPrimaryObjects” x4, “MeasureObjectIntensity” x4, “RelateObjects” x3, “CalculateMath” x3, “ExportToDatabase”, and “ExportToSpreadsheet”. The purpose of the first two modules was to complete the despeckling of the DAPI or Hoechst channel as described above. The “IdentifyPrimaryObjects” were used for each of the four bands (DAPI and three antibodies). The “MeasureObjectIntensity” modules were used for the primary objects on each of their corresponding channels, and the “RelateObjects” modules were used on the three antibody object classes to relate them to

the nuclei stain. The “CalculateMath” modules were used to normalize each of the three marker stains against the DAPI stain. The last two modules exported the data in formats that subsequent programs and scripts could work with to continue the analysis.

## Finalized Data Processing Pipeline

The first formal, standardized portion of the analysis was developed when a Java program was written to extract corresponding portions of each channel from an image set for analysis. This program enforced a consistent and convenient way to create manageable images for the CellProfiler pipeline to process. The cropping program was altered a few times and finalized with the design of taking 9 sub images from each image channel. The cropped images are taken in an evenly spaced 3 x 3 grid from within the cover slip. The size of the cropped images was 2000 x 2000 pixels, which at this magnification is 1.6 x 1.6 mm of surface area. This size was chosen for the stains which were done in 24 well plates and imaged at 10x magnification because the total area analyzed was then approximately 20 percent (see Figure 11).

In addition to the individual channels (typically four, one for each marker), there was an additional image that was cropped: the color composite image. This color composite was made by assigning each channel a unique color (from red, blue, green, yellow, cyan, magenta, gray) and then overlaying and combining all the channels together. This image was also cropped to produce a color composite of each of the nine analyzed sections. These images were useful for visualization to screen for debris or artifacts that may disrupt the CP pipeline. If any of the images had an obvious contaminant in them (a hair or some other outside component), the cropped section was moved so that it only contained valid data in it. A final context image was also created; this image was the color composite of the entire coverslip that had the nine cropped sections outlined in their location on the coverslip. This provided the full context for where each of the cropped sections was coming from, and was useful to verify all of the

cropped sections contained real data, and didn't border on the edge of the coverslip for any reason.

This cropping script produced the output files in a consistent location, with standardized names for each of the images. Having a consistent name and location was important for input into the CellProfiler pipelines. The names of the files were used to identify and analyze the objects from each stain.

The standardized modules for identifying nuclear, surface, and cytoplasmic stains were created in CellProfiler. A new pipeline was created for each image set which used the appropriate combination of the primary identification modules. The pipeline was then run in a testing mode against the image set and visually verified for accurate identification. If the module was not accurate enough, or misidentifying objects, then its settings were adjusted until its results were satisfactory. Then the pipeline was run against the whole image set (nine images for each of the bands) and it produced several output files that were used for further analysis – multiple CSV files and a SQLite database file.

The CSV files were useful in Excel, because the data could be used to produce scatterplots, and histograms. The histograms were made to see the distribution of GATA6 or HA expression in comparison to the cell populations.

The SQLite database file contained information for each object identified in all the images used during the CP run. This meant that all the information gathered across the image set was in one conglomerate database, which made it easy to query statistics on the populations. Two Java programs were written to access this file and create output products, and CellProfiler Analyst depended on the existence of this file to run as well.

CellProfiler Analyst was useful for creating scatter and density plots based off the measurements from the image set. CellProfiler Analyst also had the capability of gating certain populations from a plot and viewing them individually in the color composite image to check for abnormalities or other factors that might be easier to spot visually.

The first Java program that was written used the database file to gather statistics about the stained populations and compare those populations with one another. These statistics were produced in a text file that could be easily compared between various experiments. This Java program also connected with the R statistical language to run a non-paired t-test on the means of two of the populations to check for a statistically significant difference. In addition to this script, if any comparison was desired that wasn't specified in the script, the query could be performed manually in CellProfiler Analyst with its table query interface. This was useful for experiments that didn't depend on GATA6 or HA, as the script was designed to compare populations against a marker such as those.

The second, and more complex, Java script to work with the db file was designed to analyze the "Neighborhood Impact Factor" of GATA6. It did so by creating a "Cell" object. This object consisted of the cell's x and y pixel values, the values for the three biological stain expressions, and the image number the cell was from. The program could then be run to produce a detailed output for the NIF values of a given radius (set before it ran), or it could compute the average NIF values for each of a range of radii to determine the optimal radius. The output files were CSVs that had information for the radius, average IF values of the two marker populations, the percent differences between these IFs and the p-value associated with that difference.

These pieces of software – CellProfiler, CellProfiler Analyst, ImageJ, Java, SQLite, and R – along with Microsoft Excel, were all the analysis tools used for this work. These tools produced all the statistics, plots, and graphs from the cellular stain imaging.

### Spatial Visualization and Neighborhood Impact Factor Analysis

The premise behind the “Neighborhood Impact Factor” (NIF) is that perhaps the absolute expression of any given genetic marker isn’t necessarily the defining factor for driving the cell’s fate. One alternative is that the cell fate could be determined by “relative” genetic expression, meaning that any given cell’s expression is important within the context of the cells around it. So, if there was a given cell with a relatively high expression of GATA6 in a “neighborhood” with many other high expressing GATA6 cells, then maybe it will become CD34 positive and become an endothelial cell, because it has less GATA6 than its neighbors. However, if there was a cell with that same GATA6 expression in a neighborhood with low expressing GATA6 then perhaps it becomes CEBP $\alpha$  positive and becomes a hepatocyte.

This idea was first explored by creating a new data visualization technique. A plot was produced from output of the CellProfiler pipeline. The spatial information (center x and y pixel values), and biomarker intensities were used to produce a new plot. Every cell was plotted using its x and y values, and a circle was placed at the center of those values to represent the cell. The circle was then colored one of four colors to represent the four cell populations, based on the two biomarkers. For example, when analyzing the HA, CEBP $\alpha$ , CD34 data, the four populations were CEBP $\alpha$ + / CD34-,

CEBP $\alpha$ -/CD34+, CEBP $\alpha$ -/CD34-, and CEBP $\alpha$ + /CD34+. The final step of the plot creation was to change the radii of the circles based on the third biomarker (in this case HA). The larger the circle, the higher the intensity of that marker for the given cell. These spatial plots can give a good overview for visualizing the GATA6 or HA distribution in contrast with the different cell fates. This was a good intermediary step for analyzing gene expression impact and visualizing the data on a larger scale, but a more numerical in-depth analysis was still needed. The Neighborhood Impact Factor could go down to a cellular level and analyze each individual cell of interest, within the entire analyzed population. This was important because these spatial plots showed anywhere from 6,000-10,000 cells at once. Viewing that many data points at once is very difficult to try and absorb or draw conclusions from.

The first step was defining how to quantify the Neighborhood Impact. Four different methods were created to quantify the effect of neighboring cells on any given cells. The four methods were called: “Total Expression”, “Local Density”, “Distance Reversed Expression”, and “Distance Adjusted Expression”.

The impact factor (IF) of the Total Expression method is defined as:  $NIF = \sum GATA6 \text{ cell expression}$ . This method simply calculates the sum of the expression of all GATA6 within the neighborhood by finding all the cells within the neighborhood radius and adding up their GATA6 expression levels which are quantified by intensity (see Figure 2).

The Local Density method defines impact factor as:  $NIF =$

$$\frac{\sum(GATA6 \text{ cell expression})}{\text{Number of cells in neighborhood}}$$

This method takes the sum of the expression of the

GATA6 in the neighborhood divided by all the cells in the neighborhood, including those that were GATA6 negative. This method is to check if the density, not just expression, is important for a given cell (see Figure 3).

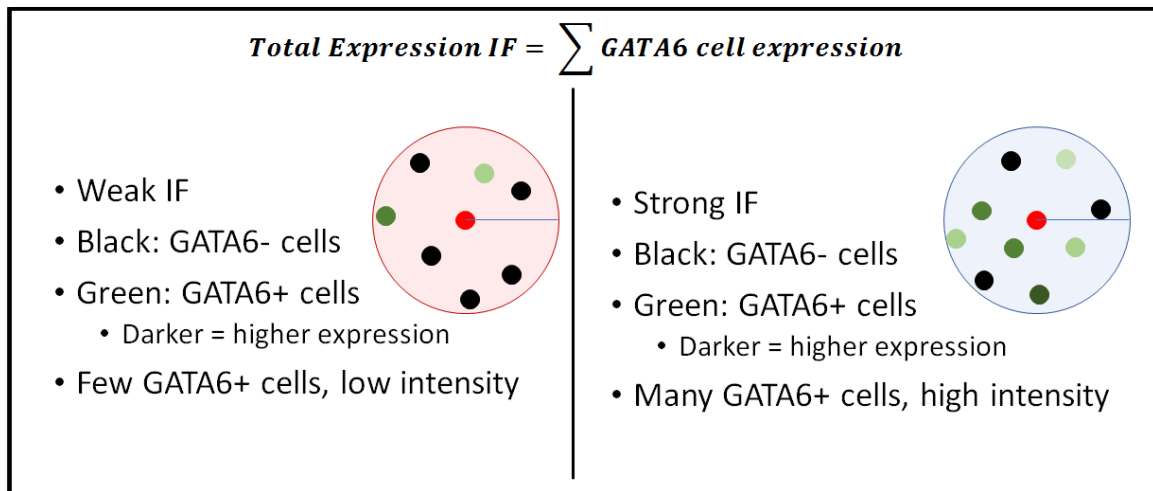
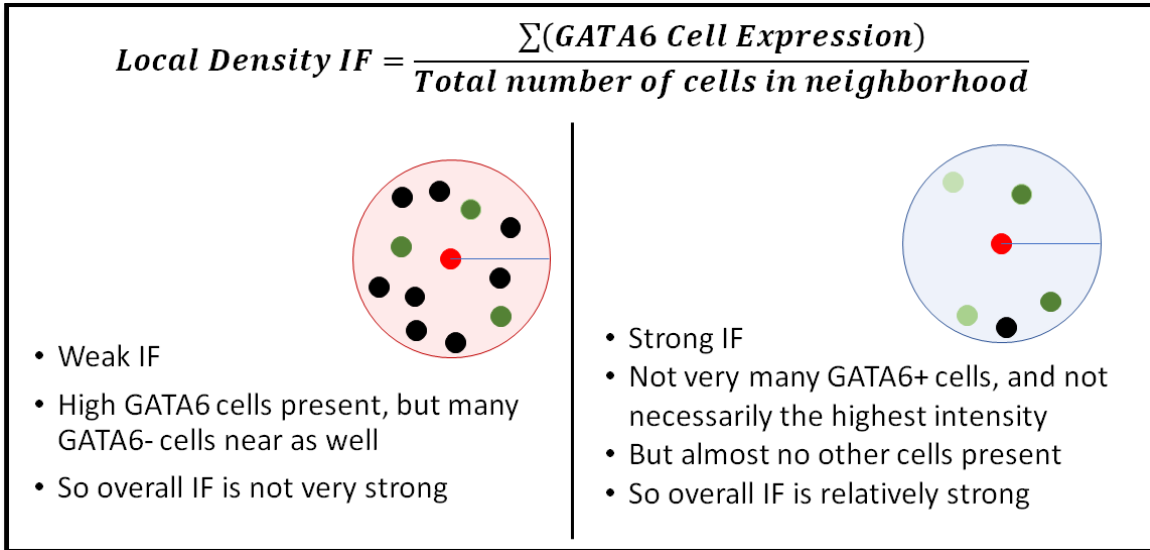


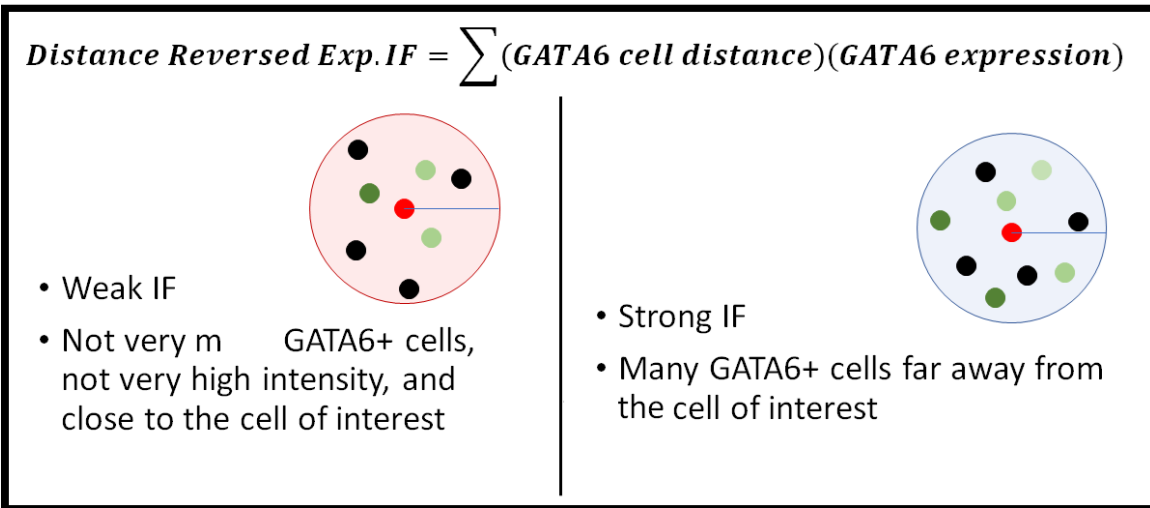
Figure 2 Total Expression IF Definition

The larger red and blue circles represent the radius from the cell of interest. Every cell is depicted as a circle. The red cell is the cell of interest, black cells are GATA6 negative cells, and varying strengths of green are GATA6 positive cells, with darker green meaning higher expression.

An example of a weak Total Expression Impact Factor (IF) is shown on the left, with an example of a strong IF shown on the right.



*Figure 3 Local Density IF Definition*  
 The larger red and blue circles represent the radius from the cell of interest. Every cell is depicted as a circle. The red cell is the cell of interest, black cells are GATA6 negative cells, and varying strengths of green are GATA6 positive cells, with darker green meaning higher expression.  
 An example of a weak Local Density Impact Factor (IF) is shown on the left, with an example of a strong IF shown on the right.



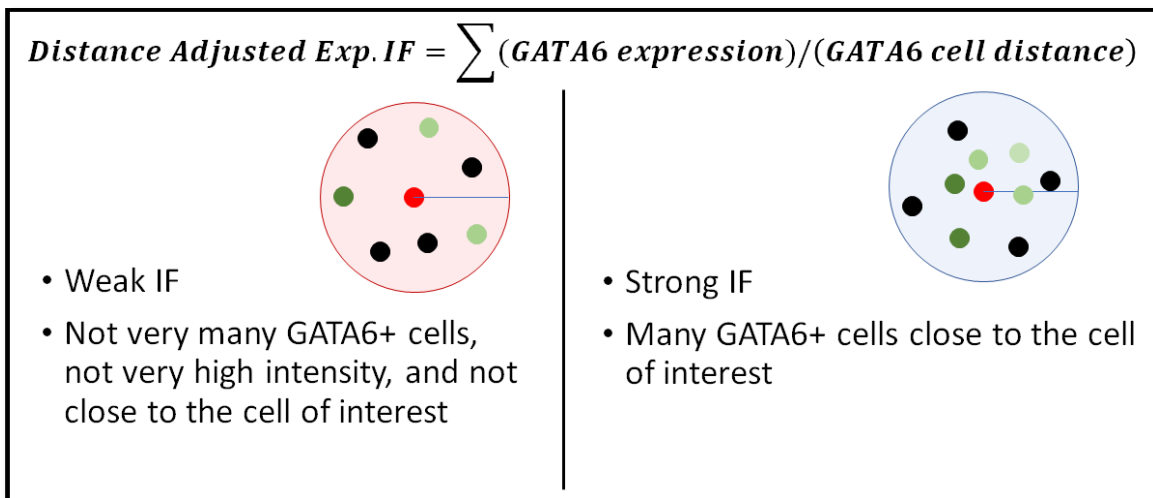
*Figure 4 Distance Reversed Expression IF Definition*  
 The larger red and blue circles represent the radius from the cell of interest. Every cell is depicted as a circle. The red cell is the cell of interest, black cells are GATA6 negative cells, and varying strengths of green are GATA6 positive cells, with darker green meaning higher expression.  
 An example of a weak Distance Reversed Expression Impact Factor (IF) is shown on the left, with an example of a strong IF shown on the right.

The third Neighborhood Impact Factor method, Distance Reversed Expression, defines impact factor as:  $NIF = \sum(\text{GATA6 cell distance})(\text{GATA6 cell expression})$ .



This method is designed in case there are both positive and negative feedback loops in effect. The idea behind it is that there are negative feedback loops affecting the cell of interest, but they only travel a short distance. So, as the GATA6 neighboring cells get farther away their positive GATA6 loop becomes stronger because the negative loop fades (see Figure 4).

The last method, Distance Adjusted Expression, defines impact factor as:  $NIF = \sum(GATA6 \text{ expression}) / (GATA6 \text{ cell distance})$ . This method quantifies the neighborhood with the assumption that as the GATA6 neighboring cells are farther away from the cell of interest, their influence is reduced. Therefore, the distance is inversely proportional to the NIF; the smaller the distance, the larger the impact (see Figure 5).



*Figure 5 Distance Adjusted Expression IF Definition*  
 The larger red and blue circles represent the radius from the cell of interest. Every cell is depicted as a circle. The red cell is the cell of interest, black cells are GATA6 negative cells, and varying strengths of green are GATA6 positive cells, with darker green meaning higher expression.  
 An example of a weak Distance Adjusted Expression Impact Factor (IF) is shown on the left, with an example of a strong IF shown on the right.

The application for the neighborhood analysis is on the vulnerable cell population. This is the population which has the potential to become either of the cell fates of interest. Typically, given a higher GATA6 level, a cell became a hepatocyte, and with

lower or non-existent GATA6 levels, the cell would become an endothelial cell. Because of this, the median GATA6 level cells were chosen as the vulnerable population.

Depending on the distribution of GATA6 intensity, a buffer was chosen which bracketed the mean GATA6 level so that the population of qualifying cells could be analyzed. For the early timepoint experiments with days 1-6, the buffer used was 0.5% of the max GATA6 expression. For day 0, since there were so many fewer cells, this buffer was increased to 5% of the max expression. This population was usually on the range of ~10% of the GATA6+ population. So, out of the total population of analyzed cells in a given dataset (~100,000), usually about 50,000-60,000 or so were GATA6+. Out of these GATA6+ cells, roughly 3,000-5,000 or so fell in the vulnerable mean population. Each one of these cells was examined as a target cell, could have somewhere between 2 to 20 cells within its neighborhood. All of these cells were used to calculate the given IF for that given target cell, and then the average IF from all target cells was used to describe that population.

This population was isolated and analyzed using a Java program that I wrote, which interfaced with the database file produced from the CellProfiler pipeline. When this population was isolated, it was also determined whether each cell expressed either of the markers of interest – for late time points these markers were CEBP $\alpha$  and CD34, for early time points the markers were T and FOXA2. Next, the software was written to test each of the four neighborhood methods on the mean GATA6 population. It did this by looking at each cell within the population of interest and then finding all the neighboring cells within the defined neighborhood radius. Then it used either the GATA6 or HA measurements from those cells to calculate the appropriate neighborhood method. The

outputs of the program contained a CSV file for each of the methods, listing the GATA6 or HA value for each cell, and its Impact Factor. In these files it also separated the cells into Marker 1+/Marker 2-, Marker 1-/Marker 2+, Marker 1+/Marker 2+, and Marker 1-/Marker 2- populations. Finally, the average NIF values for each the purely Marker 1+ population and purely Marker 2+ population were calculated and used to find the percent difference between the two. This was the true purpose of the neighborhood analysis, if the two populations had reasonably different average Impact Factors, then it seemed like the GATA6 neighborhood had an actual influence on the outcome of the cell's fate.

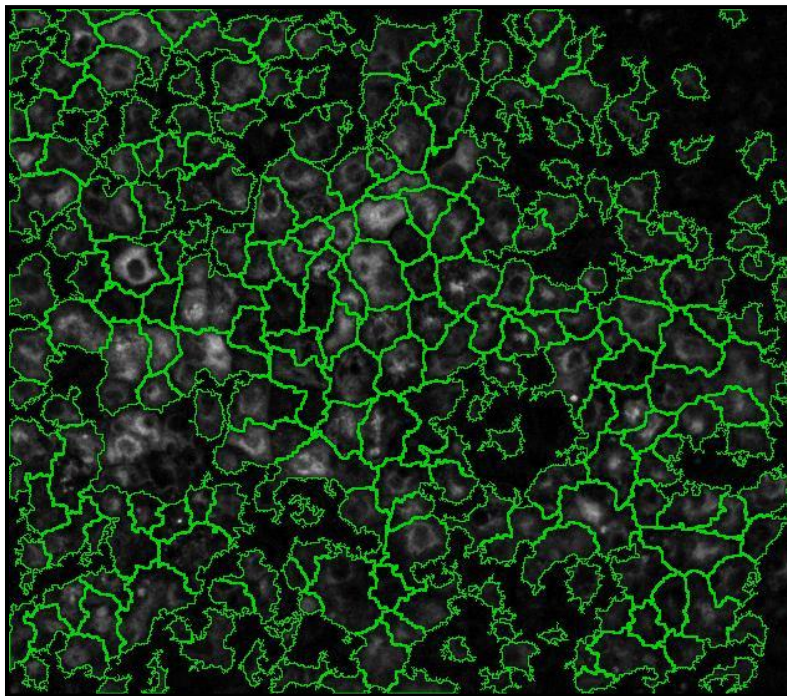
In addition to producing detailed output for each of the neighborhood methods for one set radius another product was made for each method. This product was to try and test the extent or boundary of the neighborhood. In essence, this reran the entire neighborhood analysis calculation 280 times (once for each radius sized 10 – 2800 with 10 pixel increments). The output was a CSV containing the average IF value for each marker population, the percent difference between this average, and the p-value for that difference. This was useful because it could be used to help determine what the proper radius should be when running the analysis, and it was used to see if the neighborhood's influence could change over time. A final product was produced which stepped through the GATA6 expression level in increments of 5% and looked at how many pure marker cells there were at each population slice. This was done to help verify that analyzing the mean GATA6 level was the most appropriate place to define the vulnerable population.

## CHAPTER 4

### RESULTS

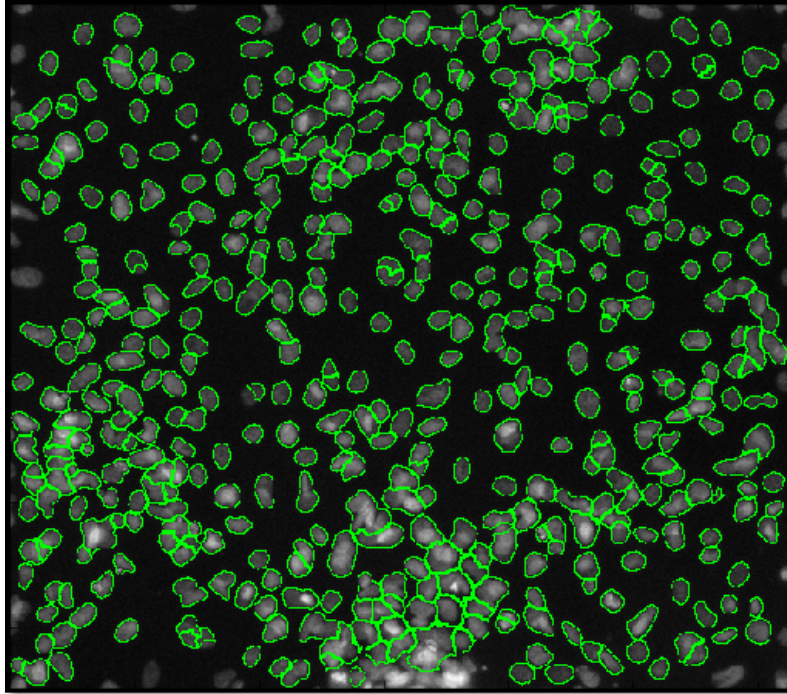
#### CellProfiler Pipeline Development

Identifying primary objects within CellProfiler is one of the most crucial aspects for the image analysis. This function was refined by several iterations and refinements of the IdentifyPrimaryObjects module for each of the individual stain types used in staining. The two main stain types were surface and nuclear. The output from the module for a surface stain can be seen in Figure 6 which was done on the surface stain AAT.



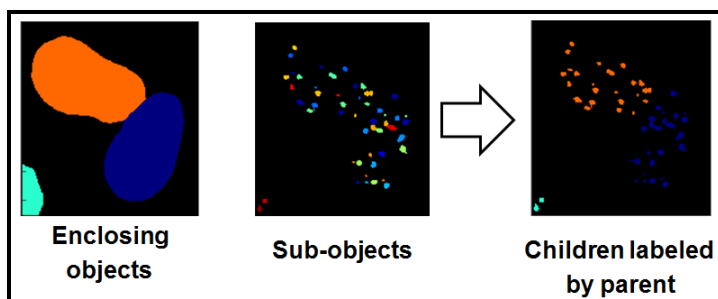
*Figure 6 Primary Object Identification of AAT  
AAT objects have been identified in the green outlines. The original image is behind that, with stain intensity shown with brighter white.*

Nuclear stain object identification was slightly easier to refine because the objects were more detected more easily in CellProfiler, requiring less iteration and modification of the module parameters. An example of the output for a nuclear object identification can be seen in Figure 7.



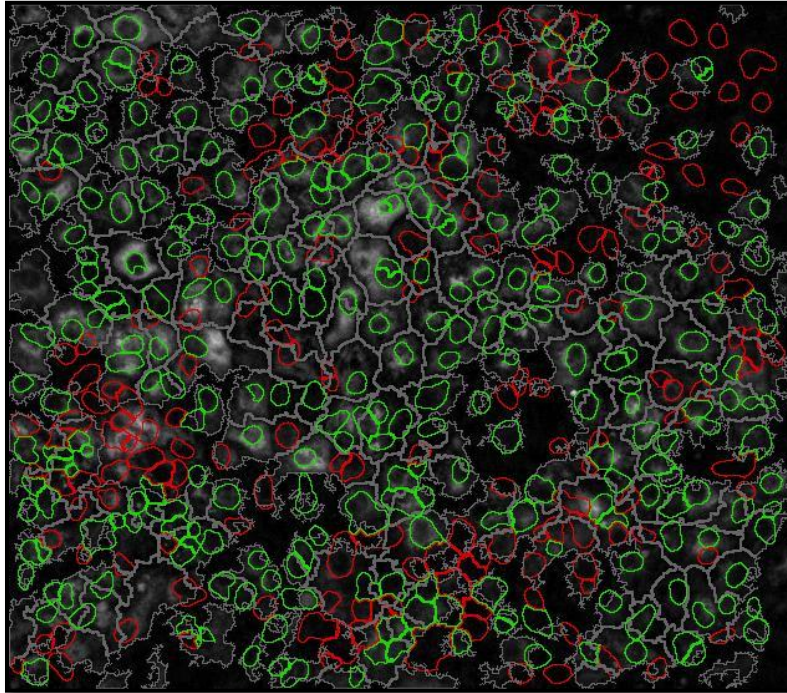
*Figure 7 Primary Object Identification of Nuclei  
Nuclei have been identified in green. This is the from the corresponding DAPI  
image for the AAT stain.*

The next critical part of developing the pipeline was to verify that relating the child objects (the biological marker stains) to the nuclei (DAPI or Hoechst stain) was accurate and accomplishing the desired result. This was verified by producing an overlay of the child objects with the nuclei objects, and then coloring the nuclei objects if they were identified as parents. The premise behind the “Relate Objects” module is that any child object found in the same overlapping space as the parent will then get associated with the parent (see Figure 8).



*Figure 8 Relate Objects Module  
The use of this module is to relate all  
stains (whether nuclear or surface) back  
to a parent nuclei, so that all stains  
could be measured within context of the  
same nuclei object.*

To verify this was working correctly and overlay and filter modules were used to produce the example seen in Figure 9.

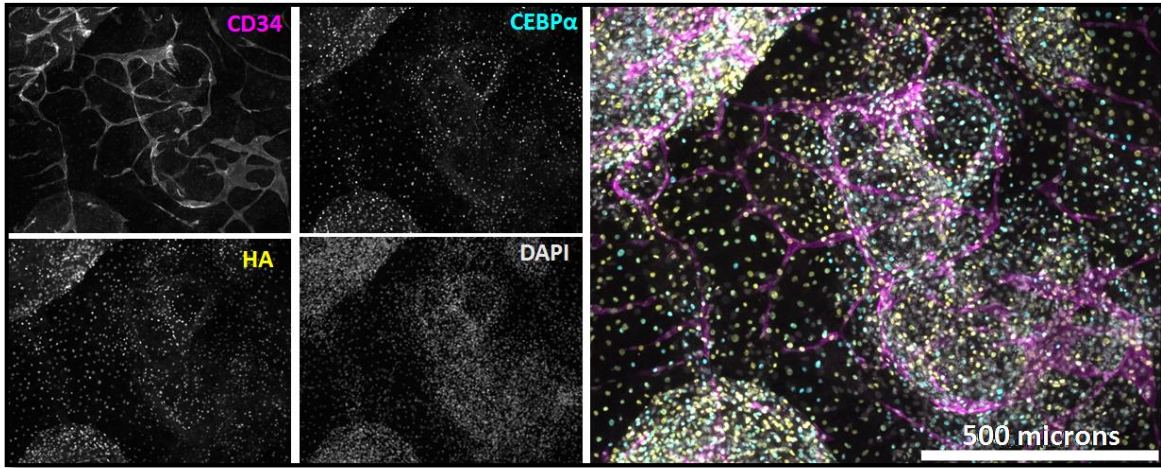


*Figure 9 Parent-Child Overlay*  
The result from a “RelateObjects” module is shown here. Previously identified AAT objects have outlines drawn in gray, and nuclei object outlines are drawn in green and red. The green outlines represent the objects that have been identified as “parents” to the “children” AAT objects, while the red objects have been filtered out as unrelated to an AAT object.

### Image Cropping and Combination

The raw images for each experiment were collected as gray scale TIFFS with one image from each of the four bands. These images needed to first be combined to produce a color composite overlay so all 4 bands could be visualized at one time. Typically, because the DAPI and Hoechst channels were orders of magnitude brighter than the other stains, an exposure increase was applied to the other three channels before combining for the color composite product. This process was done using the ImageJ software package. An example of this process and product is shown in Figure 10 with a day 14 age old

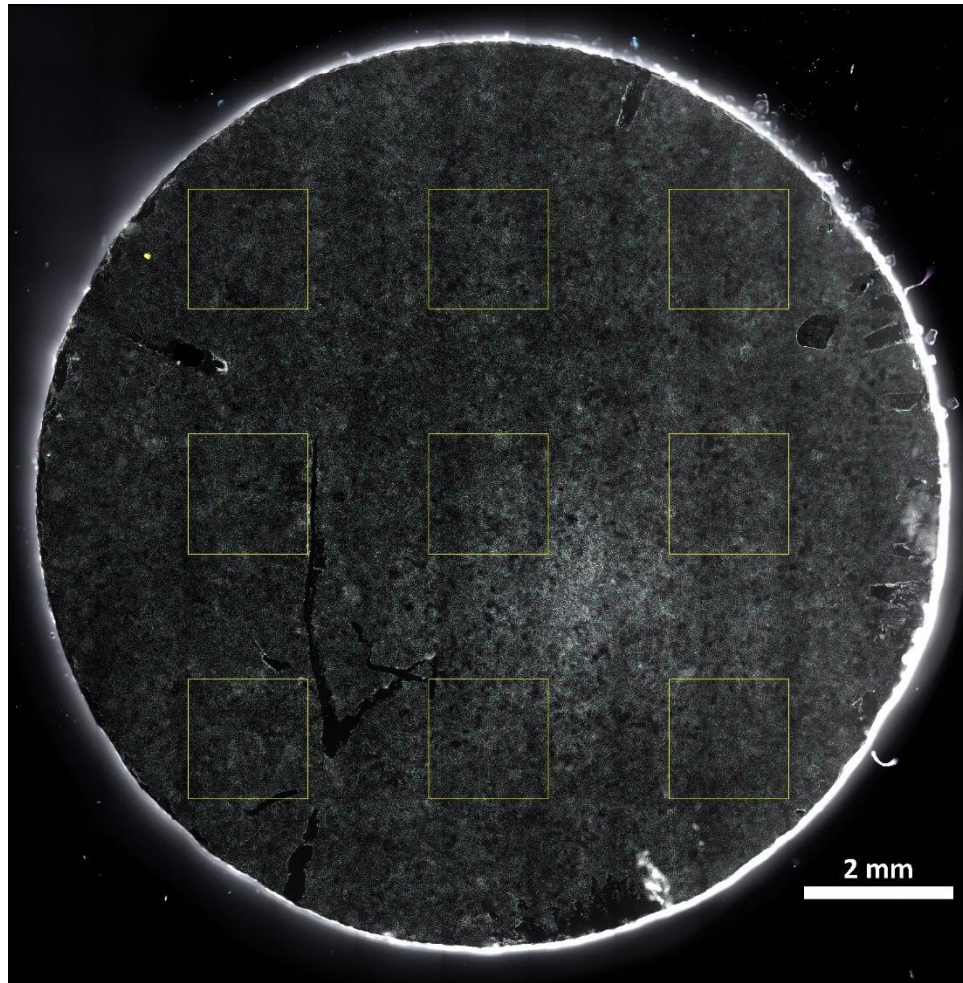
sample that was stained with CD34, CEBP $\alpha$ , HA (exogenous GATA6 marker), and DAPI.



*Figure 10 Color Composite of 4 Stain Images*

*This is an example of a cropped section of a 4-channel stain. The four individual channels are shown to the left with CD34, CEBP $\alpha$ , HA, and DAPI. The right shows the merged color composite of the four channels with magenta being CD34, cyan being CEBP $\alpha$ , yellow being HA, and grey being DAPI.*

These color composites were made using images of the entire coverslip. These images were too large to run through CellProfiler, and they contained the edge of the coverslip within the image which would also cause object identification issues. This led to the Standard Crop Java program to be written and developed. This program took the 5 images as input (4 TIFFs, one for each band, and the color composite JPEG). It then defined nine 1200x1200 regions spaced evenly within the image. It produced an outline image like the one seen in Figure 11, and separate TIFFs in each band for each of the sub cropped outlined images.



*Figure 11 Standard Crop Output*

*This is an example of the cropped regions produced from the Standard Crop Java program. The images used in the CellProfiler pipeline are outlined in yellow. The underlying image in this case is the merged color composite of a PGPI day 3 T, FOXA2, GATA6, Hoechst stain, where T is cyan, FOXA2 is magenta, GATA6 is yellow, and Hoechst is gray.*

## Spatial Analysis Development

As a first step to analyzing the spatial impact of genetic expression, a spatial plot was created using the marker expression levels and cell location as underlying information. The center x and y pixel values for each cell were used as x and y values in a new plot. A circle was placed at every cell location, and it was colorized based on the two cell fate markers. The result was four populations: +/+, +/-, -/+, and -/-. In addition to the colorization, each cell also had a radius that was determined by the GATA6 or HA



expression level. The higher the expression, the larger the radius. An example of these spatial plots can be seen in Figure 12. This first level analysis allowed for visualizing all the expression data in spatial context. This led to the need for quantifying the results and so instigated the creation of the Neighborhood Impact Factor analysis.

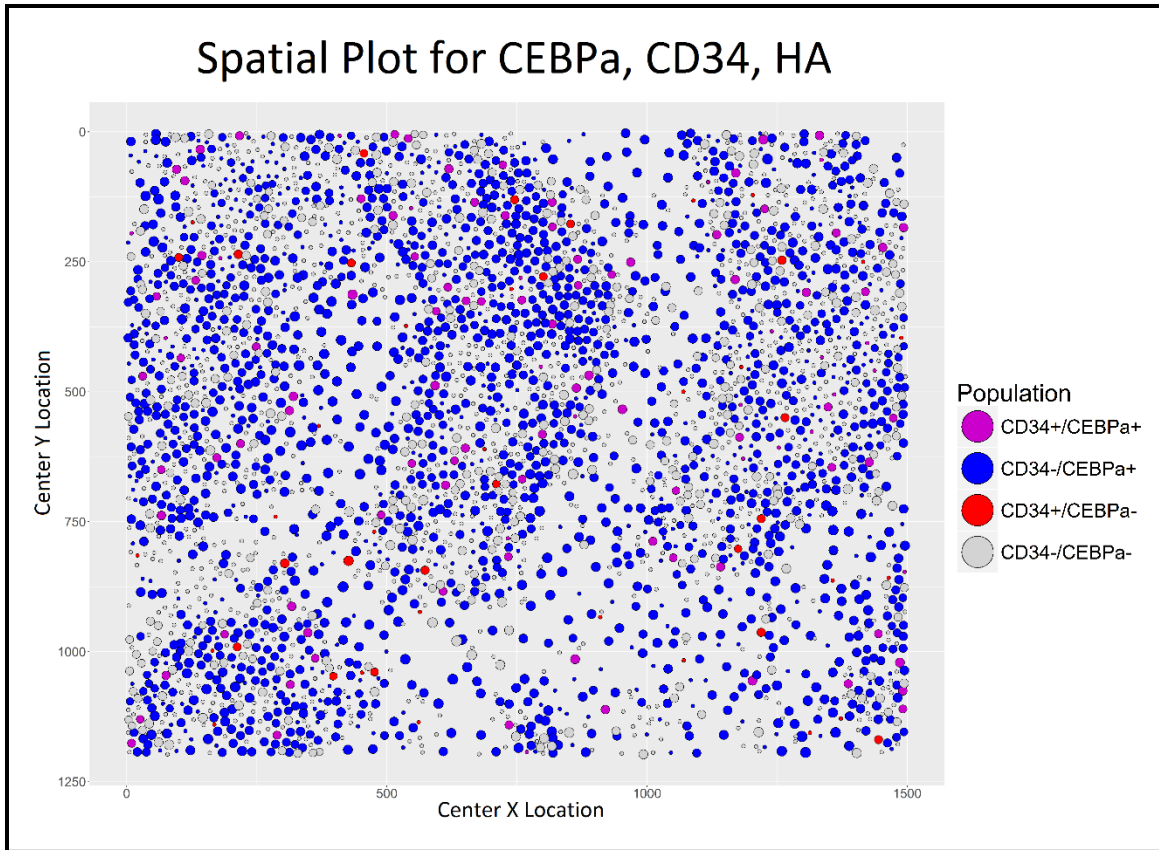


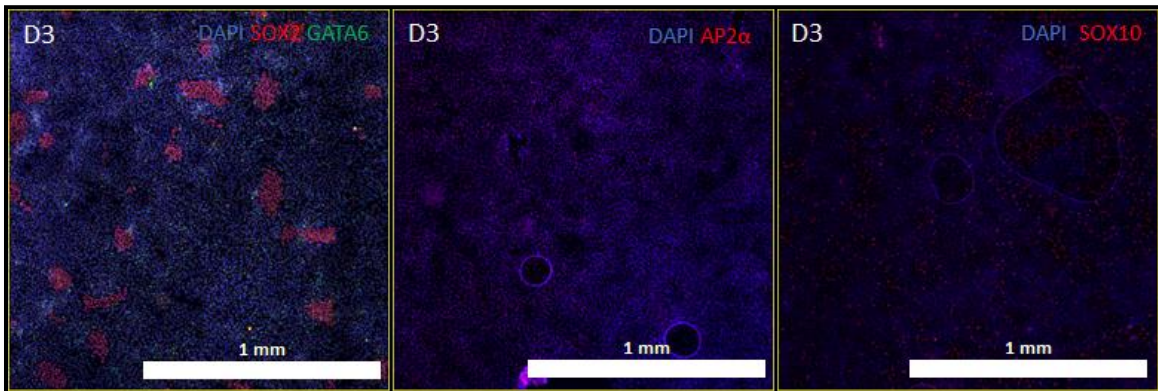
Figure 12 Spatial Plot for CEBPa, CD34, HA  
The spatial plot was derived from information from a sub-cropped section of the day 14 CEBPa, CD34, HA staining. The x and y axis represent pixel values within the cropped image. Each circle represents a detected cell, and it is colored based on the CD34 and CEBPa levels. The radius of the circle for each image is determined by the measured HA amount. The higher the HA expression, the larger the circle representing the cell.

### Early Ectoderm Markers

Multiple cultures of days 0-6 cells were created for this project. These cultures were grown using the PGP1 cell line. Cell cultures from a second source did not turn out

and have not been included, and would need to be regrown in future experiments to verify results for that cell line.

The main points of analysis were to determine if the direct expression of GATA6 had an influence on markers representing the three germ lines. Ectoderm, mesoderm, and endoderm were represented by the staining of SOX2, T, and FOXA2, respectively. The first set of cell cultures were stained with GATA6 (which tags both the endogenous and exogenous gene), T, and FOXA2. A second set of cultures was stained with GATA6, SOX2, and FOXA2. The ectoderm marker (SOX2) was chosen after testing a variety of other markers on samples of ages 3 and 5 days: SOX2, FOXG1, AP2 $\alpha$ , OCT4, and SOX10. Some of the stains did not produce measurable results, and SOX2 appeared to have the clearest and cleanest output. Figure 13 shows the results from these ectoderm markers on day 3 samples.

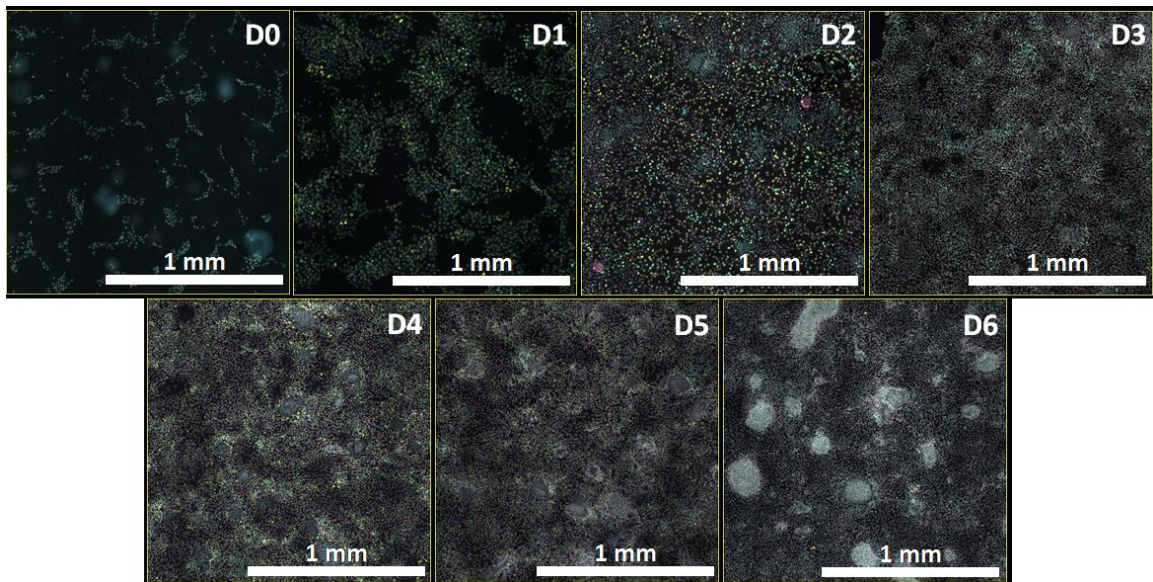


*Figure 13 Early Ectoderm Marker Test Results*

*Shown are day 3 cultures that were each stained with different ectoderm markers. The center and right samples were also stained with FOXG1 and OCT4 respectively, but those bands produced no measurable image so they were not included in the color composites.*

## Early Time Point Experiments: FOXA2, T, GATA6

Figure 14 shows cropped images taken from each of the day 0 through 6 samples that were stained with GATA6, T, and FOXA2. These images were used to produce Table 1 which shows basic statistics about the samples. Cell number and density increased throughout time, as expected. GATA6 expression also increased during the first 3 days before leveling off to a consistent level and then dropping significantly on day 6 when doxycycline was removed from the media.



*Figure 14 Day 0-6 FOXA2, T, GATA6 Color Composites*  
One image from the Standard Crop analysis is shown from each of the cell cultures aged day 0-6. These cultures were stained for Hoechst (nuclei stain), GATA6, T, and FOXA2 with the colors gray, yellow, cyan, and magenta respectively.

It was also shown that a high percent of the population expresses T very early on and tapers off a little as time passes. Concurrently, FOXA2 expression picks up when T expression decreases in days 4 through 6 (see Figure 15). As far as GATA6 expression goes, it seems in general T+ cells have a higher mean GATA6 expression than FOXA2+ cells. The mean GATA6 levels for the T+ and FOXA2+ populations are substantially

lower than the mean GATA6 value for the GATA6+ population but that is to be expected since both the T+ and FOXA2+ populations contain a large amount of GATA6- cells.

The mean GATA6 level for the T+ population is always higher than that in the FOXA2+ population for the PGP1 cells. A statistical t-test was ran on these two populations and was less than 0.05 for each of the experiments, showing that the difference between these two populations is significant.

*Table 1 Summary Table for Day 0-6 FOXA2, T, GATA6 Cultures*

*The columns increase according to the age of the sample (day 0 – day 6). The data represented are the numerical products of the object identification image analysis. The rows in descending order are as follows: total number of cells in the measured sub-images, percent of total population which has GATA6 expression, percent of total population which has FOXA2 expression, percent of total population which has T expression, percent of the FOXA2 population which has above average GATA6 expression, percent of T population which has above average GATA6 expression, percent of FOXA2 population which has below average GATA6 expression (not including no GATA6 expression), percent of T population which has below average GATA6 expression (not including no GATA6 expression), percent of FOXA2 population which has no GATA6 expression, percent of T population which has no GATA6 expression, the mean expression of GATA6 (for GATA6 positive cells only), the mean expression of GATA6 for all FOXA2 positive cells, the mean expression of GATA6 for all T positive cells, and the P-value for the difference between the mean GATA6 expression between the FOXA2 positive and T positive populations.*

Stat	D0	D1	D2	D3	D4	D5	D6
Number of Cells	8349	35991	60831	90999	97583	97295	98976
% GATA6+	23.14%	49.56%	75.45%	80.45%	58.03%	68.12%	22.05%
% FOXA2+	79.91%	55.26%	75.73%	58.02%	62.45%	58.43%	67.47%
% T+	85.47%	61.62%	63.72%	54.28%	47.66%	51.95%	52.10%
% FOXA2+ > mean GATA6	9.73%	15.77%	32.44%	30.61%	23.37%	29.56%	9.48%
% T+ > mean GATA6	9.49%	18.82%	38.32%	39.83%	30.04%	33.97%	11.02%
% FOXA2+ < mean GATA6	10.96%	36.21%	49.56%	53.16%	37.75%	45.73%	17.42%
% T+ < mean GATA6	14.81%	35.46%	45.27%	47.56%	37.79%	43.36%	16.67%
% FOXA2 no GATA6	79.32%	48.02%	17.99%	16.24%	38.88%	24.71%	73.10%
% T+ no GATA6	75.70%	45.72%	16.41%	12.60%	32.17%	22.67%	72.31%
Mean GATA6 expression	0.0083	0.0552	0.1045	0.0458	0.0417	0.0206	0.0080
Mean FOXA2+ GATA6 exp	0.0019	0.0278	0.0883	0.0384	0.0250	0.0167	0.00021
Mean T+ GATA6 exp	0.0020	0.0304	0.1000	0.0447	0.0298	0.0180	0.00023
P-value for mean diff	0.0226	1.53 E-9	4.14 E-70	5.73 E-205	4.50 E-187	1.06 E-55	6.33 E-12

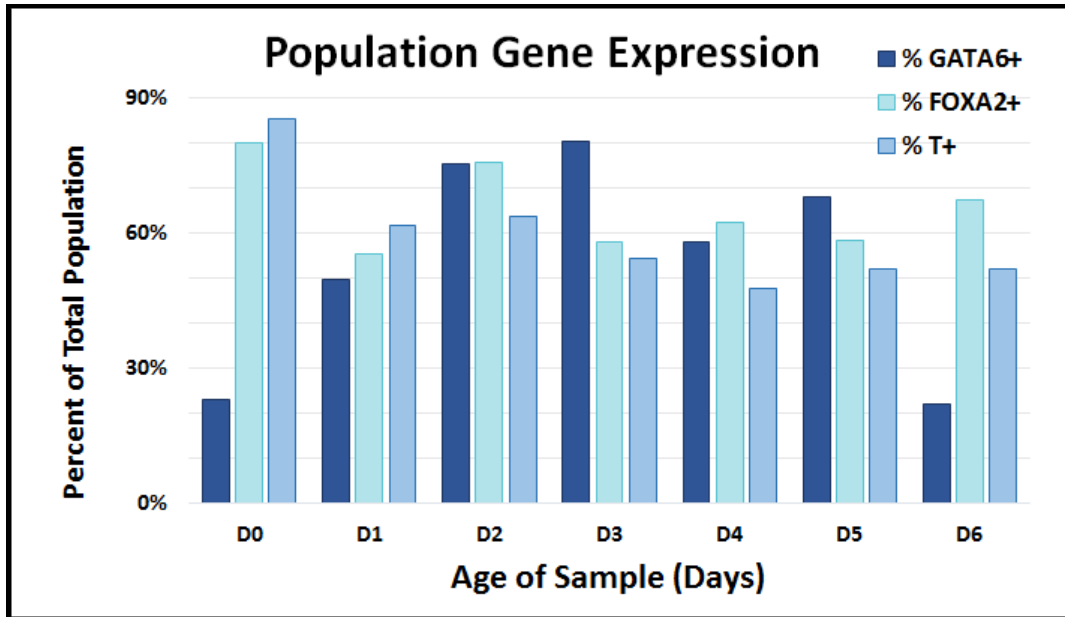


Figure 15 Population Gene Expression for FOXA2, T, GATA6  
 The composition of the overall population is displayed by % positive for each of the three gene markers for every day 0 through 6. Some cells may express more than one marker.

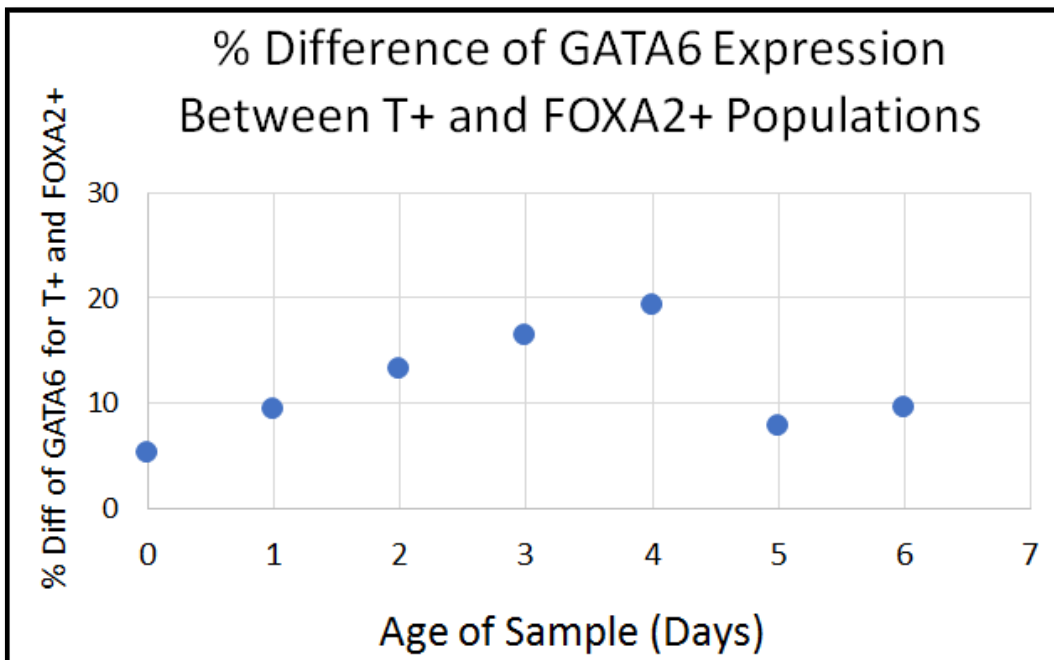


Figure 16 Difference of GATA6 Expression in T+ and FOXA2+ Populations  
 Plotted on the x-axis is the sample (defined by the age it was grown until fixed in days), and the y-axis is the percent difference of the mean GATA6 level between the T+ and FOXA2+ populations. This percent was calculated by subtracting the FOXA2+ mean GATA6 from the T+ mean GATA6 and dividing by the FOXA2+ mean GATA6.

The GATA6 expression levels of the T+ and FOXA2+ populations are significantly different from each other. This difference increases almost linearly from days 0 to 4 (see Figure 16). The percent difference ranges from ~6% to ~19%, with the T+ population always expressing a higher average GATA6 level (see Figure 16).

Plots for days 1, 3, 4, and 5 were created by comparing the following expressions against one another: T vs. GATA6, FOXA2 vs. GATA6, and FOXA2 vs. T. These plots can be seen in Appendices A-C, respectively.

Neighborhood Impact Factor analysis of the early time point stains of FOXA2, T, and GATA6 was summarized using boundary optimization plots. Since the optimal extent of the neighborhood boundary was not known, the NIF analysis for all four methods was run 280 times for each sample set, using a neighborhood radius of 10-2800 pixels to cover the entire range of the image. The boundary optimization program produced data for plotting the average NIF values of both populations vs. the neighborhood radius, as well as the percent difference for these averages, and their respective p-values against the neighborhood radius (see Appendices A-C for more detailed results). That data was used to generate summary plots showing the optimal radii and their corresponding percent difference for each NIF method for days 1, 3, and 4 datasets (see Figure 17).

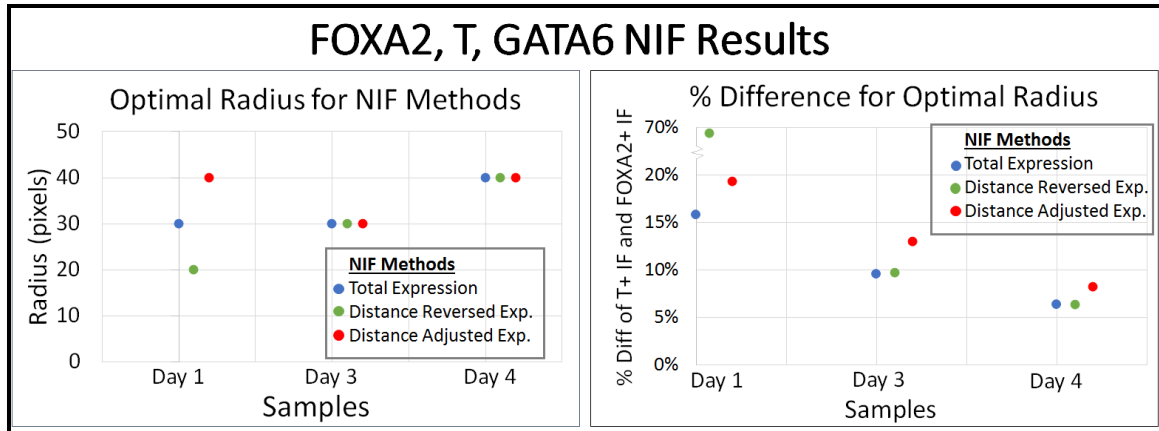


Figure 17 FOXA2, T, GATA6 NIF Results

Shown on the left are the optimal (determined by combination of lowest p-value and largest percent difference) neighborhood radii for each of the three NIF methods for days 1, 3, and 4 samples. On the right are the corresponding percent differences for the two populations' average IF values, with the T+ population having a higher impact factor value than the FOXA2+ population. All results shown had p-values below 0.05, making the results statistically significant.

#### Early Time Point Experiments: SOX2, T, GATA6

Data from the day 1-6 SOX2, T, GATA6 experiments were processed to produce the color composites seen in Figure 18. One of the processed cropped sections from the color composite image is shown for each day of the experiment. The day 0 results have been omitted because the SOX2 channel which was measured in the red emission wavelength seemed to have bled into the far red which was measuring the GATA6 stain, so the image analysis shows unrealistically high levels of GATA6 for a sample that has not been exposed to dox. The day 3 results have been omitted because the stained images did not turn out properly. These experiments seemed to produce cultures which were less differentiated than expected.

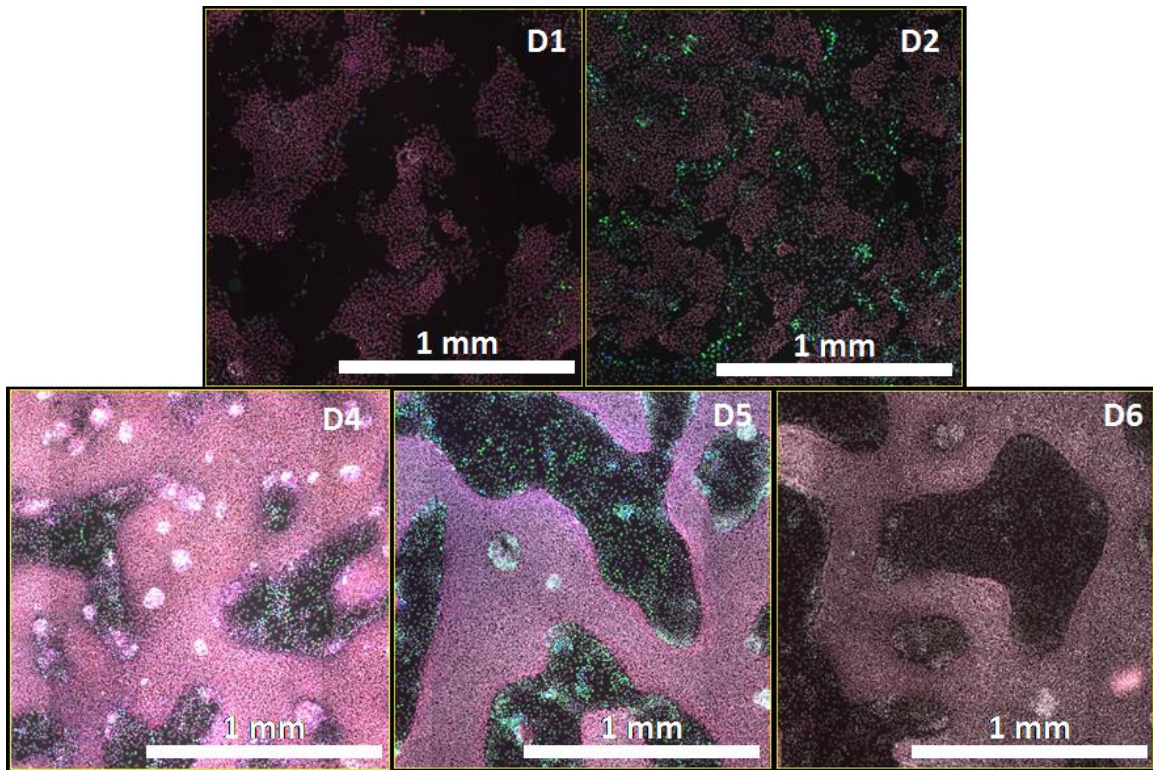


Figure 18 Day 1,2,4,5,6 SOX2, T, GATA6 Color Composites  
 One image from the Standard Crop analysis is shown from each of the cell cultures aged day 1, 2, 4, 5, and 6. These cultures were stained for Hoechst (nuclei stain), GATA6, T, and SOX2 with the colors gray, green, blue, and red respectively.

The single channel grayscale images used to produce the color composites were processed with CellProfiler to produce a database file for each experiment. These files were then analyzed to produce a text file with quantitative statistics for each one, and those stats are summarized in Table 2. The fact that these cultures were less differentiated than previous experiments can be seen from lower T+ percent values than before, and consistently high SOX2+ values (Figure 19). The T+ population was more likely to have GATA6 expression than the SOX2+. It also showed that for the days where the cultures were exposed to dox, the mean GATA6 expression for the T+ population was statistically significantly higher than the GATA6 expression of the SOX2+ population.



Table 2 Summary Table for Days 1, 2, 4, 5, 6 SOX2, T, GATA6 Cultures

The columns increase according to the age of the sample (day 1 – day 6, excluding day 3). The data represented are the numerical products of the object identification image analysis. The rows in descending order are as follows: total number of cells in the measured sub images, percent of total population which has GATA6 expression, percent of total population which has SOX2 expression, percent of total population which has T expression, percent of the SOX2 population which has above average GATA6 expression, percent of T population which has above average GATA6 expression, percent of SOX2 population which has below average GATA6 expression (not including no GATA6 expression), percent of T population which has below average GATA6 expression (not including no GATA6 expression), percent of SOX2 population which has no GATA6 expression, percent of T population which has no GATA6 expression, the mean expression of GATA6 (for GATA6 positive cells only), the mean expression of GATA6 for all SOX2 positive cells, the mean expression of GATA6 for all T positive cells, and the P-value for the difference between the mean GATA6 expression between the SOX2 positive and T positive populations.

Stat	D1	D2	D4	D5	D6
Number of Cells	22428	46683	77927	74623	78949
% GATA6+	53.22%	68.72%	34.49%	41.72%	24.61%
% SOX+	76.66%	77.02%	60.14%	71.97%	70.17%
% T+	53.17%	64.71%	43.73%	57.22%	66.99%
% SOX2+ > mean GATA6	18.51%	13.78%	14.25%	14.08%	10.84%
% T+ > mean GATA6	24.13%	21.24%	16.23%	20.58%	10.50%
% SOX+ < mean GATA6	43.18%	56.79%	27.10%	23.97%	175.61%
% T+ < mean GATA6	41.36%	57.87%	25.37%	30.84%	14.57%
% SOX2 no GATA6	38.32%	29.43%	58.66%	61.96%	73.55%
% T+ no GATA6	34.52%	20.88%	58.40%	48.58%	74.63%
Mean GATA6 expression	0.0502	0.0294	0.0134	0.0343	0.02777
Mean SOX2+ GATA6 exp	0.0391	0.0192	0.0056	0.0127	0.00732
Mean T+ GATA6 exp	0.0449	0.0248	0.0061	0.0182	0.00707
P-value for mean diff	1.69 E-23	2.56 E-130	6.54 E-14	1.74 E-261	0.00573

The data shown in the last three rows of Table 2 was used to create a percent difference plot to highlight the difference in GATA6 expression for the T+ and SOX2+ populations. The differences between the two populations is significant and ranges from ~10% to ~45% (see Figure 20).

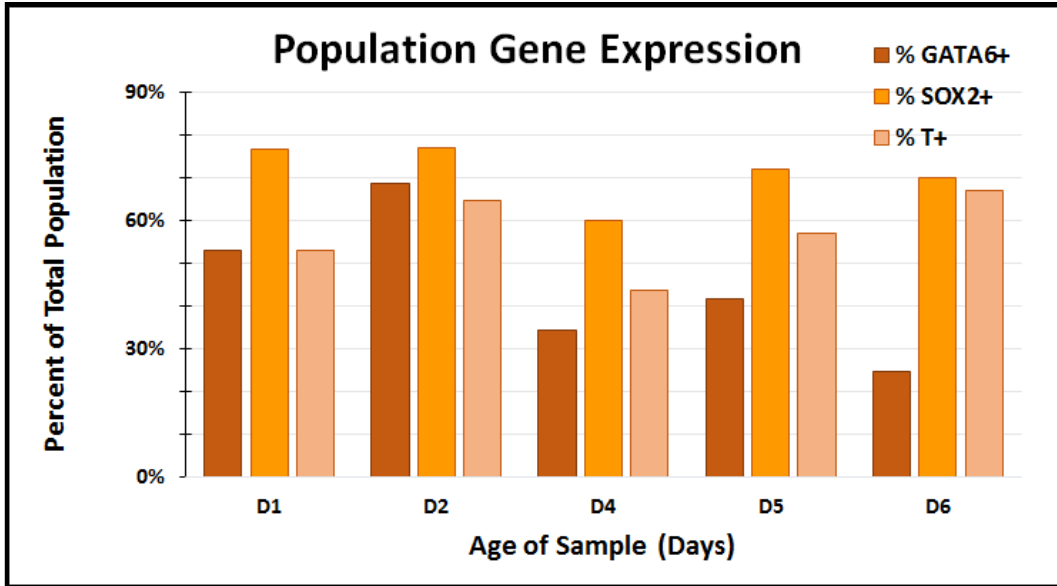


Figure 20 Population Gene Expression  
 The composition of the overall population is displayed by % positive for each of the three gene markers for every day 0 through 6. Some cells may express more than one marker.

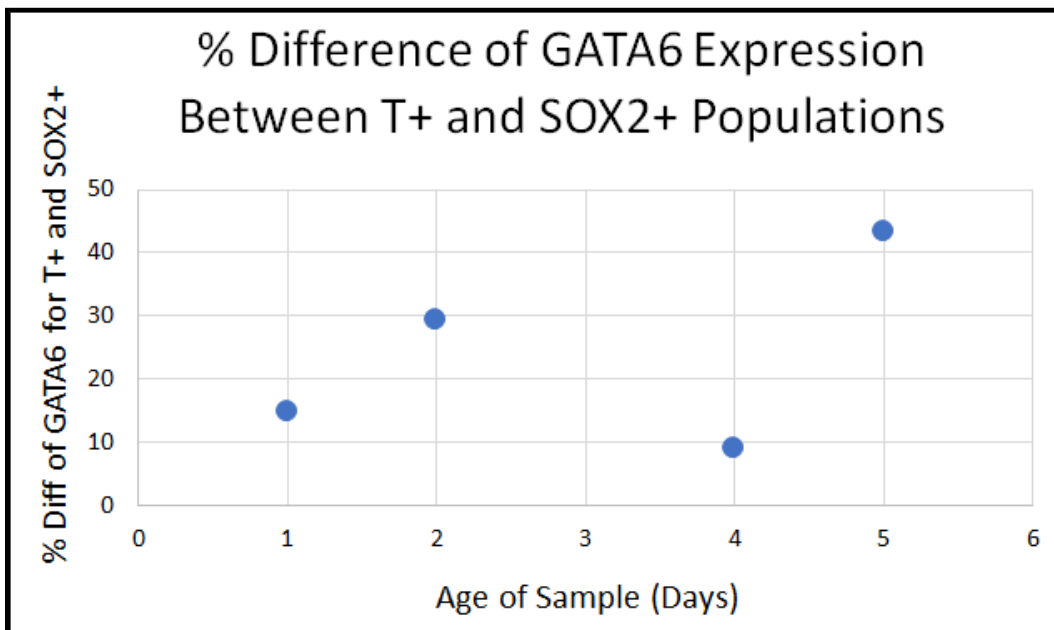


Figure 19 Difference of GATA6 Expression Between T+ and SOX2+ Populations  
 Plotted on the x-axis is the sample (defined by the age it was cultured till in days), and the y-axis is the percent difference of the mean GATA6 level between the T+ and SOX2+ populations. This percent was calculated by subtracting the SOX2+ mean GATA6 from the T+ mean GATA6 and dividing by the SOX2+ mean GATA6.

Neighborhood Impact Factor analysis of the SOX2, T, GATA6 early time point experiments found that the optimal radius of each experiment was between 24 to 48 microns – 24 and 48 microns (approximately the width of 10 to 20 nuclei). With these radii used for the NIF analysis, the results showed a significant difference of ~10% to 22% between the average IF value for the T+ population compared to the average IF value for the SOX2+ population (see Figure 21).

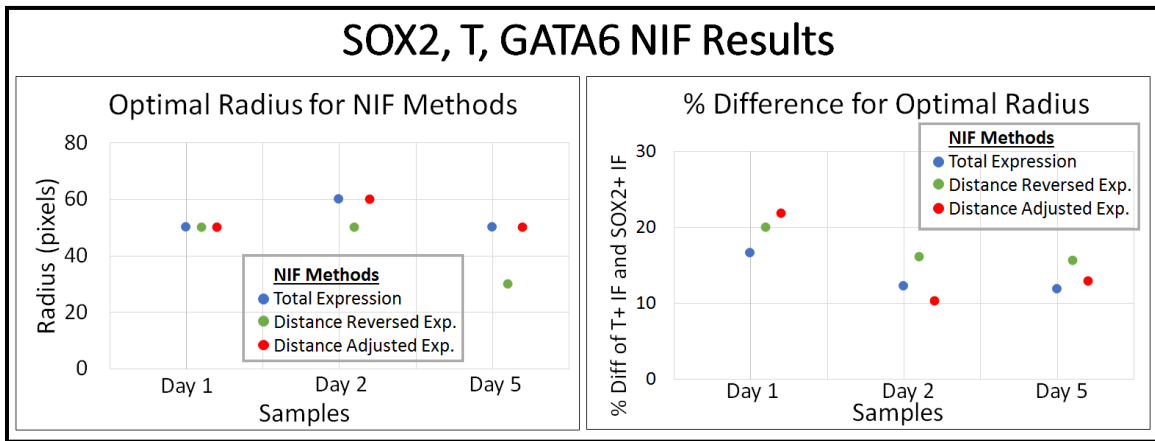
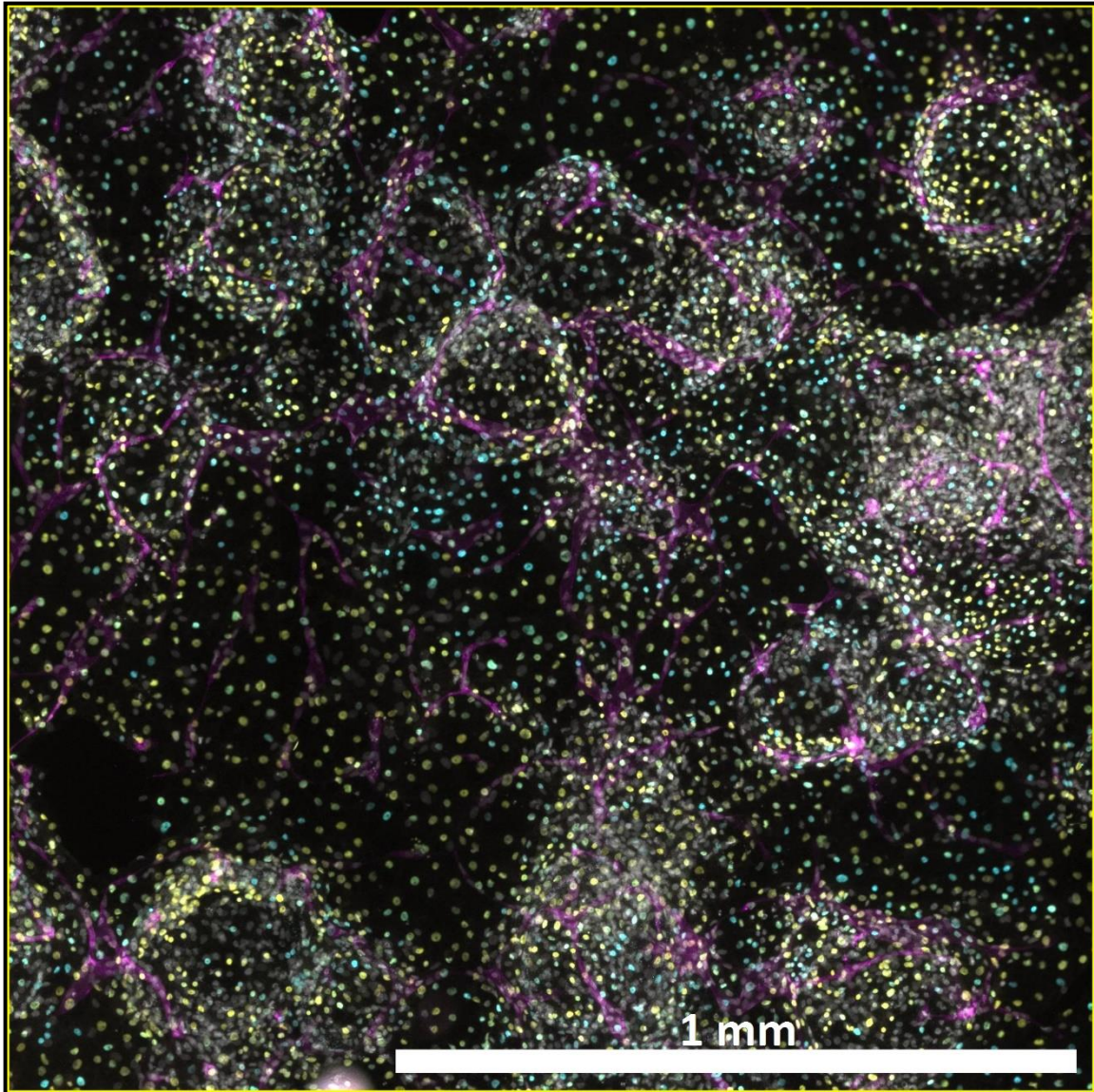


Figure 21 SOX2, T, GATA6 NIF Results  
 Shown on the left are the optimal (determined by combination of lowest p-value and largest percent difference) neighborhood radii for each of the three NIF methods for days 1, 2, and 5 samples. On the right are the corresponding percent differences for the two populations' average IF values, with the T+ population having a higher average Impact Factor than the SOX2+ population. All results shown had p-values below 0.05, making the results statistically significant.

#### Late Time Point Experiment: CEBP $\alpha$ , CD34, HA

Cells cultured to day 14 first went through 5 days with doxycycline (during days 1-5), and then another 48 hours of dox exposure starting on day 12. Cropped images used for analysis can be seen in Figure 22. As can be seen in the figure CD34 is a surface stain for mesoderm and highlights the vascular structures which have formed in the organoid. The other three stains, DAPI, CEBP $\alpha$ , and HA are nuclear stains, so overlap can be clearly seen in the produced “green” nuclei in the image.



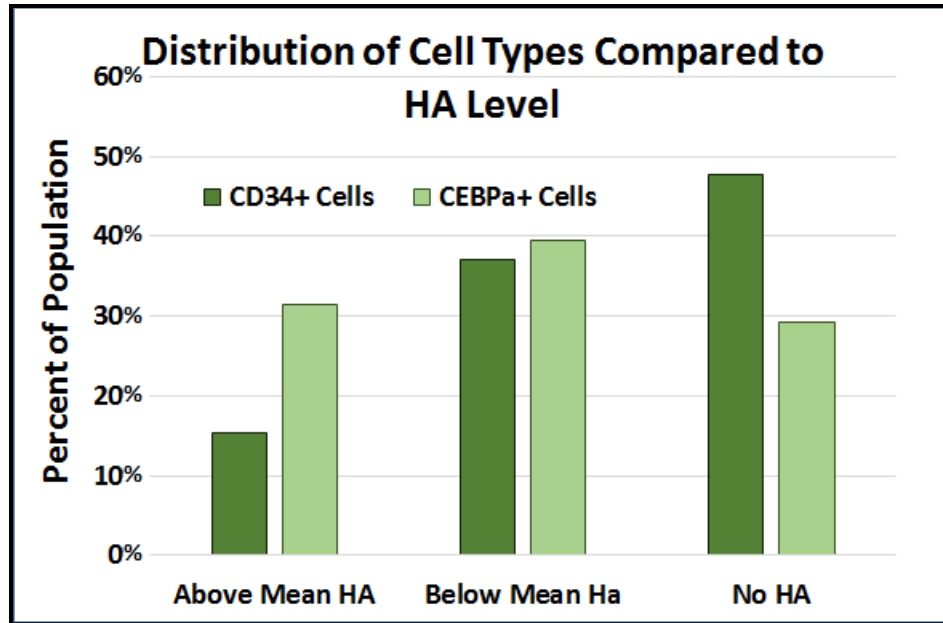
*Figure 22 Day 14 Color Composite of CEBP $\alpha$ , CD34, HA  
Shown is one of the sub cropped images of the day 14 CEBP $\alpha$ , CD34, HA stain. The composite was made with CD34 as magenta, CEBP $\alpha$  as cyan, HA as yellow, and DAPI as gray.*

A summary of the marker expressions was produced using the output from CP and can be seen in Table 3. The middle few rows detail the breakdown of the CEBP $\alpha$ + and CD34+ populations with respect to HA levels. These numbers are more clearly summarized in Figure 23, where it is shown that the CEBP $\alpha$ + has higher levels of GATA6,

but also that the CD34+ population still has ~15% of its cells with HA levels above the mean HA value. The last few rows of Table 3 show the difference in GATA6 expression for the CD34+ and CEBP $\alpha$ + populations. The percent difference between the GATA6 in the two populations is 62.6%, with the CEBP $\alpha$ + population having the higher expression level of GATA6.

Stat	Value
Number of Cells	71414
% HA+	45.84%
% CD34+	4.47%
% CEBP $\alpha$ +	49.59%
% CD34+ > mean HA	15.35%
% CEBP $\alpha$ + > mean HA	31.42%
% CD34+ < mean HA	37.00%
% CEBP $\alpha$ + < mean HA	39.43%
% CD34+ no HA	47.65%
% CEBP $\alpha$ + no HA	29.16%
Mean HA expression	0.31338
Mean CD34+ HA exp	0.14102
Mean CEBP $\alpha$ + HA exp	0.22924
P-value for mean diff	4.40 E-161

*Table 3 Summary of CD34, CEBP $\alpha$ , HA Expression*  
*Shown to the right are the measured statistics for the day 14 experiment. The rows from top to bottom are as follows: total population size, percent of total population that expressed HA, percent of total population that expressed CD34, percent of total population that expressed CEBP $\alpha$ , percent of the CD34 population that had above average HA expression, percent of the CEBP $\alpha$  population that had above average HA expression, percent of the CD34 population that had below average HA expression (not including cells with no HA expression), percent of CEBP $\alpha$  population that had below average HA expression (not including cells with no HA expression), percent of CD34 population that had no HA expression, percent of CEBP $\alpha$  population that had no HA expression, average HA expression for the entire CD34 positive population, average HA expression for the entire CEBP $\alpha$  positive population, P-value for the difference between those two means.*



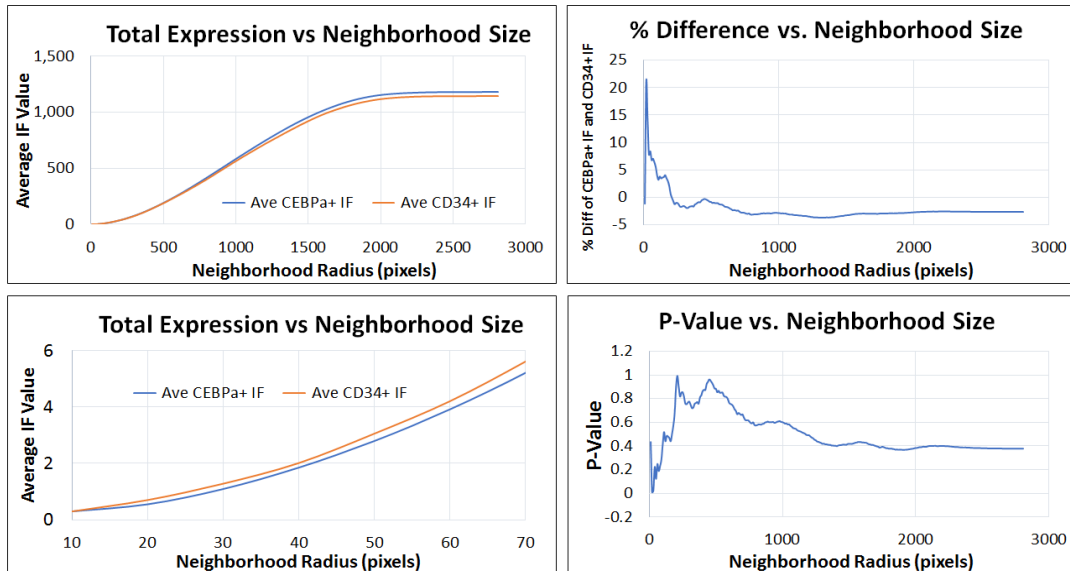
*Figure 23 Distribution of Cell Types Compared to HA Level*  
Each of the two populations (CEBPα+ and CD34+) are divided into three categories: those with HA values above the mean HA level, those below the mean HA level, and those with no HA.

To further visualize the difference in HA levels for the CD34 positive and CEBPα positive populations, histograms were produced from preliminary image analysis of the CD34 and CEBPα images (see Appendices J and K). These figures were created before the standard crop procedure and used four smaller (1200x1200) sections of the original image to represent the images.

Neighborhood Impact Analysis on the data produced from the last stage sample resulted in statistically significant findings for the Total Expression and Distance Adjusted Expression IF methods (see Figure 24).

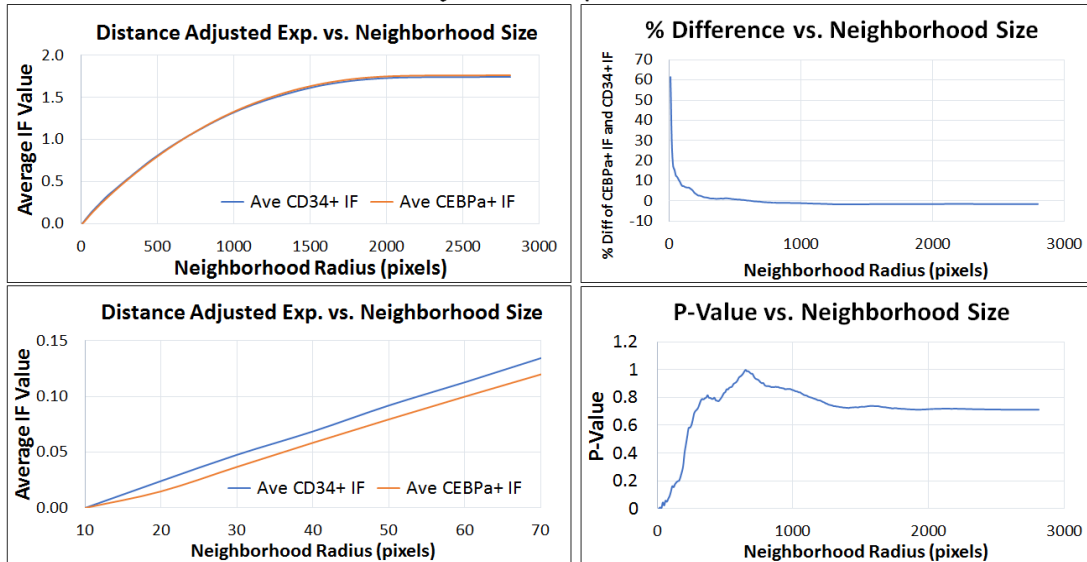
## D14: CEBP $\alpha$ , CD34, HA NIF Results

### Total Expression Results



Optimal Radius: 30 pixels, % Diff: 20.9%, P-value: 0.013

### Distance Adjusted Expression Results



Optimal Radius: 30 pixels, % Diff: 29.3%, P-value: 0.006

Figure 24 D14: CEBP $\alpha$ , CD34, HA NIF Results

Shown in the upper half of the figure are 4 plots detailing the Total Expression NIF method and in the lower half are the 4 plots for the Distance Adjusted Expression IF method. The upper left plot of each section shows the average IF values for the two populations (CEBP $\alpha$ + and CD34) against the neighborhood radius, with the lower left showing a zoomed in region of that same plot. The upper right plot for each section has the corresponding percent difference for the two populations vs the neighborhood radius, and the lower right shows the p-values for those differences. The bottom line of each section shows the optimal radius (in pixels), its percent difference, and its p-value.

## CHAPTER 5

### DISCUSSION

#### Early Time Point Experiments: FOXA2, T, GATA6

One thing that is clearly apparent, and expected, is the expression of GATA6 grows during the first few days (days 0-3), levels off and then declines on day 6, as soon as the doxycycline is removed from the media. This is confirmed by the fact that the percent of the total population of cells increases and then declines on day 6 (see Tables 1 and 2). These results show that the genetically modified gene circuit is working and actively triggering the expression of GATA6.

The expression of GATA6 seems to be driving cell differentiation towards mesendoderm fate. T seems to be expressed in the largest population – as high or higher than FOXA2 – early on (days 0-3). This could be because the fact that T is a mesendoderm marker, so this implies its expression comes first because it still has the potential to become endoderm and gain FOXA2 expression. In days 4 through 6, FOXA2 is expressed at a higher level, and this could be because some of the T+ population has switched to gain FOXA2 expression (see Table 1 and Figure 15). Also, it seems like once the initial cell fate decision has been made between ectoderm and mesendoderm (high T expression), then a second decision comes up which is between mesoderm and endoderm and that is why the increase of the FOXA2 population occurs at slightly later time points. The data from this table defines the T+ and FOXA2+ populations as a cell which expresses any of that marker. That means that some of the T+ cells are also FOXA2+. So, it is possible that as FOXA2 expression becomes stronger,



that the cell loses its T expression, and that is why the T population is shrinking a little as the FOXA2 population is growing.

This point is further supported by mean GATA6 levels for both the T+ and FOXA2+ populations. The T+ population has higher GATA6 expression than the FOXA2 population for all days 0 through 6. However, the first four days this difference seems to increase almost linearly, and then has a sharp decrease on days 5 and 6 (see Figure 16). This shows that high GATA6 expression is initially pushing cells to a mesendoderm fate – making the first decision of the three germ layers – and then as time progresses, some of those cells make the decision of endoderm over mesoderm, and the difference in GATA6 between those populations lowers quite a bit. It's possible that if measurements were taken for the next few days it might be shown that the mean GATA6 levels between the two populations equalizes or even perhaps, the FOXA2 population might surpass the T population.

In addition to what is seen from the summary of the two populations (from Table 1 and Figure 15 and 16), the data from the individual cells also tells a similar story. Examining T expression against GATA6 expression for several of the time points has a similar trend as well (see Appendix A). On day 0 the level of GATA6 and T expression is very low, as to be expected because the cells have not been introduced to dox and should be in a pluripotent state. On day 1 however, a dramatic increase in the expression of T and GATA6 is seen. This expression appears to increase on day 3 and then taper off on day 4. What is especially interesting to note on day 4 is the highest level of GATA6 is quite a bit lower in these T+ cells than in day 1 or day 3. At the same time, FOXA2 expression is doing the opposite with regards to GATA6. Initially, there is very low

expression of FOXA2 and GATA6 on day 0, as is seen with T expression. On day 1 there is a dramatic increase of FOXA2, and on day 3 an increase of both FOXA2 and GATA6 is seen in the FOXA2 population (see Appendix B). Then, on day 4, unlike what happens to GATA6 expression for the T population, GATA6 expression for the FOXA2 population actually increases. This could be when some cells that were initially part of the T population have started to differentiate towards endoderm and have become FOXA2 positive.

When assessing the direct expression of GATA6 it appears that first T is most heavily promoted, and then FOXA2. This is seen when looking at the T+ and FOXA2+ populations as a whole by analyzing their size relative to the entire population, and comparing their mean GATA6 expression. This explanation seems to also hold true when examining the individual cells of each population. The T population seems to lose some of its highest expressing GATA6 cells, right at the same time the FOXA2 population is gaining its highest expressing GATA6 cells.

The NIF results for these experiments seemed to show consistent results. The Local Density method results have not been shown for any of the experiments because that method did not seem to produce significant and meaningful results. So, it appears that the density of local neighborhood for a cell does not play a significant role in the decision making for cell fate – at least not in the experiments conducted for this research, and in the context of GATA6 gene expression. The remaining three methods: Total Expression, Distance Reversed Expression, and Distance Adjusted Expression were all used for the FOXA2, T, GATA6 days 1, 3, and 5 experiments. All the methods produced an optimal radius on the range of 16 to 32 microns (see Figure 17). For these

experiments, radii of these sizes contained approximately 2-14 neighboring cells. The results showed that the T+ population usually had a neighborhood with 15-20% higher Impact Factor values (see Figure 17). This leads to two possible conclusions: cells begin expressing T when they are surrounded by high levels of GATA6, or that cells become T positive with slightly lower relative levels of GATA6 compared to their neighborhoods. Because the NIF analysis holds the target cell GATA6 level to a very small population (usually 0.05% around the mean GATA6 expression), the difference of cell fate between T and FOXA2 cannot be determined from that (direct GATA6 expression). And since T+ cells have a higher NIF value than FOXA2, it could mean that the FOXA2 cells really are the highest expressing GATA6 cells in their neighborhood, and GATA6 ultimately does promote the endoderm cell fate with highest expression. This also seems to be entirely possible, because the extent of the optimal neighborhoods (~24 microns in radius) is fairly short. Meaning that the T+ cells really are “low” in GATA6 in respect to their respective neighborhood, and FOXA2+ cells are “high” expressing compared to their neighbors.

#### Early Time Point Experiments: SOX2, T, GATA6

The cell cultures used in the SOX2 experiments seemed to be less differentiated than usual cultures from these time points. Regardless, they still showed significant difference between the GATA6 in SOX2+ cells and T+ cells. The GATA6 expression in these cultures was less than in the FOXA2, T, GATA6 experiments, but still shows a consistent expression during days 1 to 5, and has a decline on day 6 when the dox is removed. The SOX2+ population stays at a consistent level, with approximately 70% of

the total population expressing this gene for every time point (see Table 2 and Figure 19). The last three rows of the table highlight the difference in GATA6 expression for the two populations, T+ and SOX2+. As to be expected, the mean GATA6 expression for the T+ population was consistently higher than that of the SOX2+ population – ranging from at least 10% to over 40% higher (see Figure 20).

The results from this experiment also helped confirm the findings and theory presented by Schode et al. that NANOG and GATA6 have inhibitory effects on one another. The SOX2 population can serve as an analogous population to a NANOG positive population, and it shows low levels of GATA6. Whereas, the T population acts as the mesendoderm population which is promoted by GATA6 and shows much higher levels of that gene. These results support the theory that GATA6 promotes endoderm, and the lack of GATA6 leaves cells in a pluripotent state (Schode et al., 2014).

The NIF results for these experiments was very similar to the results from the FOX2, T experiments in that the optimal radius was in the range of 24-48 microns, and the percent difference was between 10-25% (see Figure 21). These results help confirm that the range of cell-cell communication – in terms of the NIF effect – is short range and on the order of 3-4 cells away (48 microns) at most. These neighborhoods contained anywhere from 9 to 18 cells per target cell. This shows that the T+ cells were in a consistently higher neighborhood than their counterpart SOX2+ cells, which could mean that high presence of GATA6 promotes T expression and not SOX2 expression. It also could point to the fact that the cells migrate, and T+ and SOX2+ cells could migrate away from each other and towards their more similar populations which would also influence the GATA6 expression in each of these neighborhoods. If the SOX2+ cells move to seek

out other SOX2+ cells, they are bringing their own low GATA6 expression to a low GATA6 neighborhood.

Late Time Point Experiment: CEBP $\alpha$ , CD34, HA

The late time point experiment produced a sample with endothelial like cells which were stained and visible with CD34 (see Figure 22). CD34 was expressed in about 4.5% of the total population of cells, with about half the population expressing CEBP $\alpha$  (see Table 3). These cells were grown with the cell line that had a genetically engineered HA tag on the exogenous GATA6 genes, and this experiment was stained for HA, not GATA6. So, the HA expression represents the exogenous GATA6 that was in each cell, and also represents the portion of GATA6 expression which was directly triggered by the presence of dox. The CD34+ cells also dramatically less HA than the CEBP $\alpha$ + population. When comparing marker populations, it is shown that over 47% of CD34+ cells had no HA, with that number only reaching 29% in CEBP $\alpha$ + cells. Also, the mean HA expression for CEBP $\alpha$ + population was almost double that of the CD34+ population with normalized intensities of 0.23 and 0.14, respectively. These results certainly support the idea that high expressing GATA6 cells (since HA is a marker for GATA6) are pushed towards an endodermic – hepatic – cell fate.

The NIF analysis for this sample showed interesting results. Only the Total Expression and Distance Adjusted methods produced statistically significant findings. The optimal neighborhood was very like that of previous experiments, at 24 microns (see Figure 23). This radius incorporated an average of just over 6 cells in every neighborhood, and produced a difference of 21-29% between the average IF values for

the two populations, with the CD34 population having the higher NIF. This result could be used to show support for the idea that relative gene expression is playing an important role in the cell fate decision between endothelial and hepatic cell fates. Cells with a mean level of exogenous GATA6 in high GATA6 neighborhoods became CD34+, while cells with that same GATA6 expression in lower GATA6 neighborhoods became CEBP $\alpha$ +. This means, that the CD34+ cells actually had “low” GATA6 expression, for the neighborhood they were in, and can be an explanation for why they became endothelial cells, even expressing significant amounts of GATA6. Since the NIF optimal radius seems to be similar at early and late time points, and the percent difference of population IFs at the late time point experiment is almost twice that of the early time point FOXA2, T, GATA6 experiments, it could mean that the decision to become endoderm was made at some point between days 6 and 14. The neighborhood size seems to hold constant, but once the differences in neighborhoods become great enough, it can force high GATA6 expressing cells to become mesoderm instead of endoderm, because their expression can be characterized as “low”.

A possible explanation for the effect of the NIF impact could be explained by the influence GATA6 has with the WNT signalling pathway. It has been shown that overexpression of GATA6 can activate the expression of a WNT reporter construct. GATA6 has also been discovered to be bound to the WNT6 promoter and corresponds with upregulation of the WNT6 gene. This pathway has been shown to lead cells to an endodermal fate, which is consistent with what is seen in the results for this research (Hwang & Kelly, 2012). Another reason that points towards the WNT pathway as the method of cell to cell communication is the fact that the neighborhood radius is relatively

small. Studies have shown that it appears that WNT signalling plays a role in highly localized communication for cells immediately surrounding one another, which seems to be the interactions described by the results from the NIF analysis (Farin et al., 2016).

## CHAPTER 6

### CONCLUSION

There seems to be a correlation between GATA6 expression level and cell fate as defined by markers representing the three germ layers – SOX2 for ectoderm, FOXA2 for endoderm, and T for mesoderm (and mesendoderm). When comparing SOX2+ and T+ populations the T+ population had significantly higher GATA6 expression. This was to be expected and confirmed the theory that GATA6 expression promotes mesendoderm over ectoderm.

The findings from the FOXA2 and T comparison are a little more complex. The T+ population consistently had higher GATA6 expression than the FOXA2+ population. A possible explanation for this is that GATA6 promotes both meso- and endoderm early on, and so that is why T is expressed with high GATA6 and is expressed faster (higher percent of the total population on days 1 to 3) than FOXA2. It's also possible that as the cells mature, they make a second decision between meso- and endoderm and gain FOXA2 expression. This is supported by the fact that T+ GATA6 expression seems to decline while FOXA2+ GATA6 expression is increasing.

The outcomes of the Neighborhood Impact Factor analysis seem to point to an influence of short cell to cell communication. The optimal radius for most of the NIF methods for the experiments ranged from 24-48 microns, which was about 3-4 cells across, and these neighborhoods contained an average of 7 to 18 to cells each. This analysis showed that FOXA2+ cells were in higher IF neighborhoods than T+ cells with the same GATA6 expression, and that this difference in average IFs decreased as the cells aged. This could support the idea that some T+ cells were becoming FOXA2+ as



the cultures matured. The NIF results from the day 14 experiment showed that cells in with a CD34+ cells seemed to be found in neighborhoods with a 20% higher IF value than CEBP $\alpha$ + cells with the same exogenous GATA6 expression. This supports the idea that relative gene expression is an influential factor in determining cell fate, and that cells with GATA6 can become mesoderm, even though it is typically associated with promoting endoderm.

These findings of the relation of direct GATA6 levels, and their context in their immediate environment led to the definition of the model proposed in Figure 25. Both GATA6 expression and high GATA6 NIF seem to be related to the initial decision of ectoderm/pluripotent cell vs. mesendoderm. It seems that high levels of GATA6 expression go on to promote hepatocyte-like cell fate, and cells with a high GATA6 NIF and mean GATA6 expression are be prompted to form endothelial-like cells.

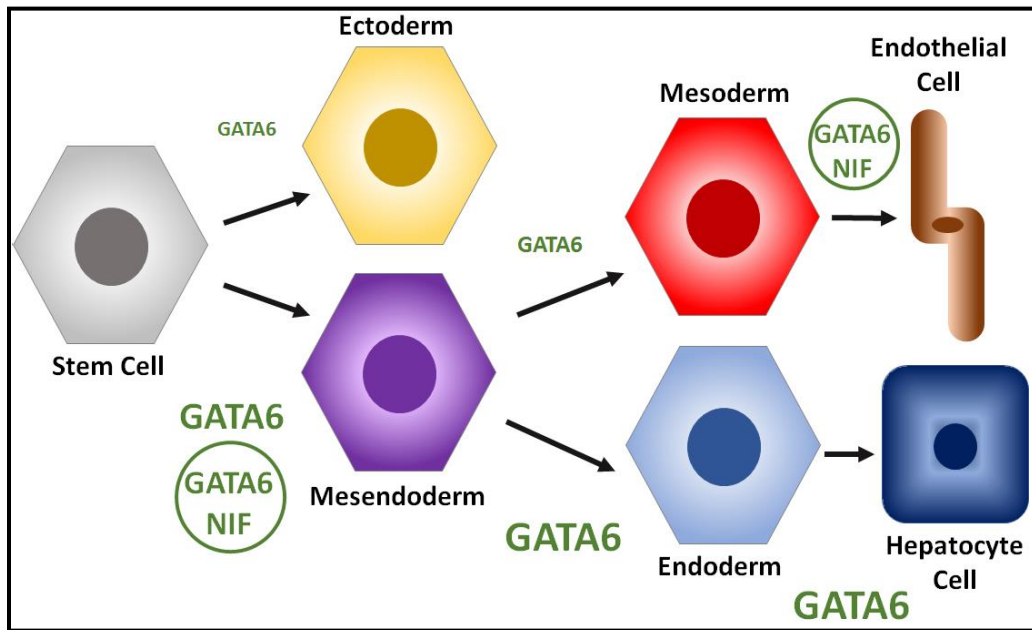


Figure 25 Proposed GATA6 Cell Fate Model  
Proposed model from stem cells to mature fate formation as promoted by GATA6 expression and high NIF values.

Of the four NIF methods, three of them seemed to play possible effects on early time point experiments, and two of them were significant for the day 14 study. The Local Density method did not seem to produce any sort of correlation results and does not appear to be an appropriate way to measure the effect of surrounding cells. The Total Expression and Distance Adjusted Expression methods were the most significant for the late time point, and they may play a critical role in determining cell fate on ages between 6 and 14. The Reversed Expression method showed significant results for the early time point experiments, but it is possible this is only short lived and as the cells mature, this influence disappears. This new form of quantification and analysis of gene expression in a spatial context seems to provide useful information to describing and explaining cell fate when culturing organoids.

## CHAPTER 7

### FUTURE WORK

#### Further Experiments

The most immediate work is to complete further research to prepare this work to be submitted for acceptance in a reputable journal. To have a paper ready for submission, several more experiments and analyses need to be carried out. A weakness of the current data is the way cells were categorized as positive and negative for gene expression. In this research, if a cell had any expression of a gene, it was considered positive. In reality though, the true negative population may possess some expression, but that expression is below a threshold limit for whether or not it is part of the positive population. Therefore, it would be important to reanalyze this data carefully to select for thresholding limits for the different gene populations, and see what the results look like with those thresholds in place.

Additionally, a few follow up studies have been suggested which would bring about a more robust result set for this research. First, it might be helpful to engineer a few cell lines that have inhibitors and activators for the early time point markers of T and FOXA2. These would be important because they will demonstrate the ability to accurately guide cell fate by directly modifying the genes that are important for defining cell fate.

Other follow up studies would include a more defined experiment centered around the importance of the “redose” of dox in the late time point experiments before fixing for staining and imaging. In this research, two dosing strategies were used on day 10 samples, and three dosing strategies were used for day 14 samples. It might be

necessary to investigate the organoid response more thoroughly under multiple dosing strategies. The purpose for the redose is so it is possible to see the endogenous levels of GATA6 that are in the cells, because these levels were present during the first 6 days of cell culture. If no redose was performed, the staining of GATA6 might not show very high expression at all because this late in the culture it is likely that most GATA6 expression has tapered off. The concern that is brought about from this practice is that the redose might cause cells with a defined fate to modify their cell type under the influence of such high GATA6 presence. Therefore, the length of the dose (exposure to doxycycline) and the time following the dose until the fixing of the cultures should be evaluated more accurately before publishing this work to a journal.

Another interesting study would be to purposely sort three populations of cells based on GATA6 expression: a “high expression” group, “medium expression” group, and “low expression” group. Then, cells from the medium group could be planted in each of the remaining groups, tracked, and studied to see what their eventual cell fate becomes. If the Neighborhood Impact Factor hypothesis holds true, it would be expected that a “medium” expression cell in a low expression neighborhood would be treated as a high expression cell (and maybe ultimately end up with a hepatic cell fate). Meanwhile, if that same cell was placed in the high expression neighborhood it might be treated as a low expressing cell and possibly end up with a mesodermal or ectodermal cell fate. Using this population model brings up another possible study related to relative gene expression. It would be interesting to culture the three populations separately and see if they still develop into multi-cell type organoids like those that develop in a heterogeneous GATA6 expression population without sorting. If similar organoids

develop, it might be further support for the idea that when it comes to gene expression, relative expression is more important than absolute expression. Because even though these populations won't have as large a variation in expression levels, a small amount of variation will still exist. And if the organoid develops it means that even the small changes in gene expression are significant, if the population is so homogeneous to begin with.

### Software Development

In addition to future hands-on experiments that should be performed, there is the possibility of continuing the development of the Java programs that were written to perform the analysis for this thesis. The final Java program that was written to assess the Neighborhood Impact Factor analysis could be enhanced to provide a graphical user interface so any user could easily work with the tool and their own data. The tool would have a standard, predefined input type that users could create from CellProfiler Analyst or other common cellular imaging tools. The Java program would allow the user to easily point to their specified input file and then define which biological markers were the determining cell fate markers, and which biological marker was the neighborhood influencing marker. In this work the cell fate markers used were the following combinations T and FOXA2, SOX2 and FOXA2, CD34 and CEBP $\alpha$ , ERG and CEBP $\alpha$ , and the neighborhood influence factor was either GATA6 or HA (which was a marker for exogenous GATA6). The program could be packaged and marketed to academic institutions for general use in organoid and developmental biology research fields. Also,

if this work was completed, it would make a good case for writing a technical paper and publishing that as well.

## REFERENCES

- Antonica, Francesco, Dominika Figini Kasprzyk, Robert Opitz, Michelina Iacovino, Xiaio-Hui Liao, Alexandra Mihaela Dumitrescu, Samuel Refetoff, Kathelijne Peremans, Mario Manto, Michael Kyba, and Sabine Costagliola. "Generation of functional thyroid from embryonic stem cells." *Nature* 491.7422 (2012): 66-71. *Science Direct*. Web.
- Bernardo, Andreia S., Tiago Faial, Lucy Gardner, Kathy K. Niakan, Daniel Ortmann, Claire E. Senner, Elizabeth M. Callery, Matthew W. Trotter, Myriam Hemberger, James C. Smith, Lee Bardwell, Ashley Moffett, and Roger A. Pedersen. "BRACHYURY and CDX2 Mediate BMP-Induced Differentiation of Human and Mouse Pluripotent Stem Cells into Embryonic and Extraembryonic Lineages." *Cell Stem Cell* 9.2 (2011): 144-55. Web.
- Bessonnard, S., L. De Mot, D. Gonze, M. Barriol, C. Dennis, A. Goldbeter, G. Dupont, and C. Chazaud. "Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network." *Development* 141.19 (2014): 3637-648. Web. 12 Jan. 2017.
- Eiraku, Mototsugu, Nozomu Takata, Hiroki Ishibashi, Masako Kawada, Eriko Sakakura, Satoru Okuda, Kiyotoshi Sekiguchi, Taiji Adachi, and Yoshiki Sasai. "Self-organizing optic-cup morphogenesis in three-dimensional culture." *Nature* 472.7341 (2011): 51-56. Web.
- Faial, T., A. S. Bernardo, S. Mendjan, E. Diamanti, D. Ortmann, G. E. Gentsch, V. L. Mascetti, M. W. B. Trotter, J. C. Smith, and R. A. Pedersen. "Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells." *Development* 142.12 (2015): 2121-135. Web.
- Farin, Henner F., Ingrid Jordens, Mohammed H. Mosa, Onur Basak, Jeroen Korving, Daniele V. F. Tauriello, Karin De Punder, Stephane Angers, Peter J. Peters, Madelon M. Maurice, and Hans Clevers. "Visualization of a short-range Wnt gradient in the intestinal stem-cell niche." *Nature* 530.7590 (2016): 340-43. Web.
- Freedman, Benjamin S., Craig R. Brooks, Albert Q. Lam, Hongxia Fu, Ryuji Morizane, Vishesh Agrawal, Abdelaziz F. Saad, Michelle K. Li, Michael R. Hughes, Ryan Vander Werff, Derek T. Peters, Junjie Lu, Anna Baccei, Andrew M. Siedlecki, M. Todd Valerius, Kiran Musunuru, Kelly M. McNagny, Theodore I. Steinman, Jing Zhou, Paul H. Lerou, and Joseph V. Bonventre. "Modelling kidney disease with CRISPR-mutant kidney organoids derived from human pluripotent epiblast spheroids." *Nature Communications* 6 (2015): 8715. Web.
- Guye, Patrick, Mohammad R. Ebrahimkhani, Nathan Kipniss, Jeremy J. Velazquez, Eldi

- Schoenfeld, Samira Kiani, Linda G. Griffith, and Ron Weiss. "Genetically engineering self-organization of human pluripotent stem cells into a liver bud-like tissue using Gata6." *Nature Communications* 7 (2016): 10243. Web. 29 Jan. 2016.
- Hwang, Jason T.k., and Gregory M. Kelly. "GATA6 and FOXA2 Regulate Wnt6 Expression During Extraembryonic Endoderm Formation." *Stem Cells and Development* 21.17 (2012): 3220-232. Web.
- Jakobsen, J. S., J. Waage, N. Rapin, H. C. Bisgaard, F. S. Larsen, and B. T. Porse. "Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries." *Genome Research* 23.4 (2013): 592-603. Web.
- Kanaki, Maria, and Dimitris Kardassis. "Regulation of the human lipoprotein lipase gene by the forkhead box transcription factor FOXA2/HNF-3 $\beta$  in hepatic cells." *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1860.3 (2017): 327-36. Web.
- Kanaya, Kurando, Masaaki, Teiji Okazaki, Toru Nakamura, Miki Horii-Komatsu, Cantas Alev, Hiroshi Akimaru, Atsuhiko Kawamoto, Hidetoshi Akashi, Hiroyuki Tanaka, Michio Asahi, and Takayuki Asahara. "Sonic Hedgehog signaling regulates vascular differentiation and function in human CD34 positive cells." *Stem Cell Research* 14.2 (2015): 165-76. Web.
- Lancaster, Madeline A., Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S. Bicknell, Matthew E. Hurles, Tessa Homfray, Josef M. Penninger, Andrew P. Jackson, and Juergen A. Knoblich. "Cerebral organoids model human brain development and microcephaly." *Nature* 501.7467 (2013): 373-79. Web.
- Lowenthal, Justin, and Sharon Gerecht. "Stem cell-derived vasculature: A potent and multidimensional technology for basic research, disease modeling, and tissue engineering." *Biochemical and Biophysical Research Communications* 473.3 (2016): 733-42. Web.
- McCracken, Kyle W., Emily M. Catá, Calyn M. Crawford, Katie L. Sinagoga, Michael Schumacher, Briana E. Rockich, Yu-Hwai Tsai, Christopher N. Mayhew, Jason R. Spence, Yana Zavros, and James M. Wells. "Modelling human development and disease in pluripotent stem-cell-derived gastric organoids." *Nature* 516.7531 (2014): 400-04. *Science Direct*. Web.
- Nadkarni, Rohan R., Soumeya Abed, and Jonathan S. Draper. "Organoids as a model system for studying human lung development and disease." *Biochemical and Biophysical Research Communications* 473.3 (2016): 675-82. Web.
- Nemer, Georges, and Mona Nemer. "Transcriptional activation of BMP-4 and regulation

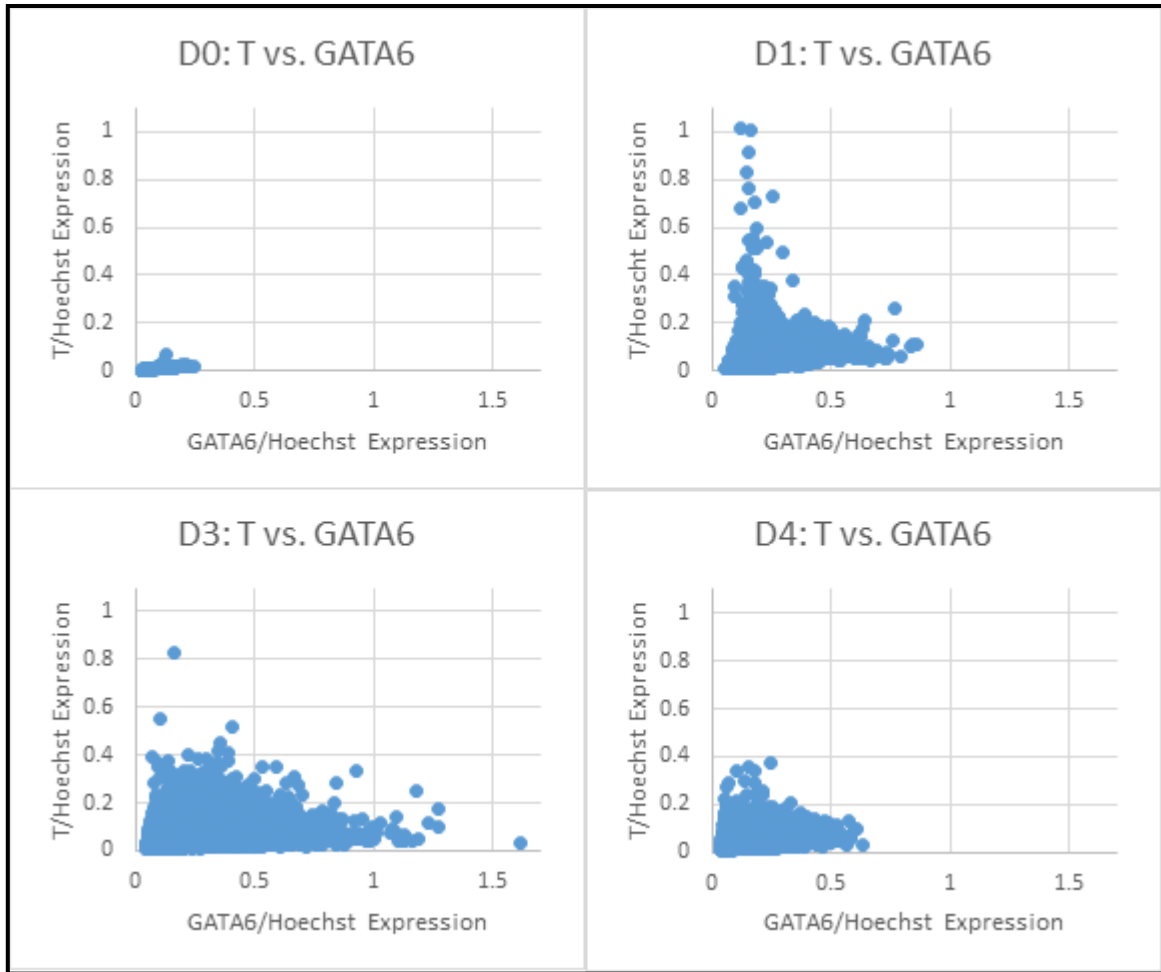


- of mammalian organogenesis by GATA-4 and -6." *Developmental Biology* 254.1 (2003): 131-48. Web. 23 Jan. 2017.
- Pournasr, Behshad, Mohammad Hassan Asghari-Vostikolaee, and Hossein Baharvand. "Transcription factor-mediated reprogramming of fibroblasts to hepatocyte-like cells." *European Journal of Cell Biology* 94.12 (2015): 603-10. Web.
- Saka, Y., Lhoussaine, C., Kuttler, C., Ullner, E., & Thiel, M. (2011). Theoretical basis of the community effect in development. *BMC Systems Biology*, 5(1), 54. doi:10.1186/1752-0509-5-54
- Sarkar, Abby, and Konrad Hochedlinger. "The Sox Family of Transcription Factors: Versatile Regulators of Stem and Progenitor Cell Fate." *Cell Stem Cell* 12.1 (2013): 15-30. Web.
- Schröter, C., P. Rue, J. P. Mackenzie, and A. Martinez Arias. "FGF/MAPK signaling sets the switching threshold of a bistable circuit controlling cell fate decisions in embryonic stem cells." *Development* 142.24 (2015): 4205-216. Web. 12 Jan. 2017.
- Spence, Jason R., Christopher N. Mayhew, Scott A. Rankin, Matthew F. Kuhar, Jefferson E. Vallance, Kathryn Tolle, Elizabeth E. Hoskins, Vladimir V. Kalinichenko, Susanne I. Wells, Aaron M. Zorn, Noah F. Shroyer, and James M. Wells. "Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro." *Nature* 470.7332 (2011): 105-09. *Science Direct*. Web.
- Stevens, Kelly R., Lil Pabon, Veronica Muskheli, and Charles E. Murry. "Scaffold-Free Human Cardiac Tissue Patch Created from Embryonic Stem Cells." *Tissue Engineering Part A* 15.6 (2009): 1211-222. *Science Direct*. Web.
- Suga, Hidetaka, Taisuke Kadoshima, Maki Minaguchi, Masatoshi Ohgushi, Mika Soen, Tokushige Nakano, Nozomu Takata, Takafumi Wataya, Keiko Muguruma, Hiroyuki Miyoshi, Shigenobu Yonemura, Yutaka Oiso, and Yoshiki Sasai. "Self-formation of functional adenohypophysis in three-dimensional culture." *Nature* 480.7375 (2011): 57-62. Web.
- Takahashi, Kazutoshi, and Shinya Yamanaka. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126.4 (2006): 663-76. *Science Direct*. Web.
- Takebe, Takanori, Keisuke Sekine, Masahiro Enomura, Hiroyuki Koike, Masaki Kimura,

Takunori Ogaeri, Ran-Ran Zhang, Yasuharu Ueno, Yun-Wen Zheng, Naoto Koike, Shinsuke Aoyama, Yasuhisa Adachi, and Hideki Taniguchi. "Vascularized and functional human liver from an iPSC-derived organ bud transplant." *Nature* 499.7459 (2013): 481-84. *Science Direct*. Web.

APPENDIX A

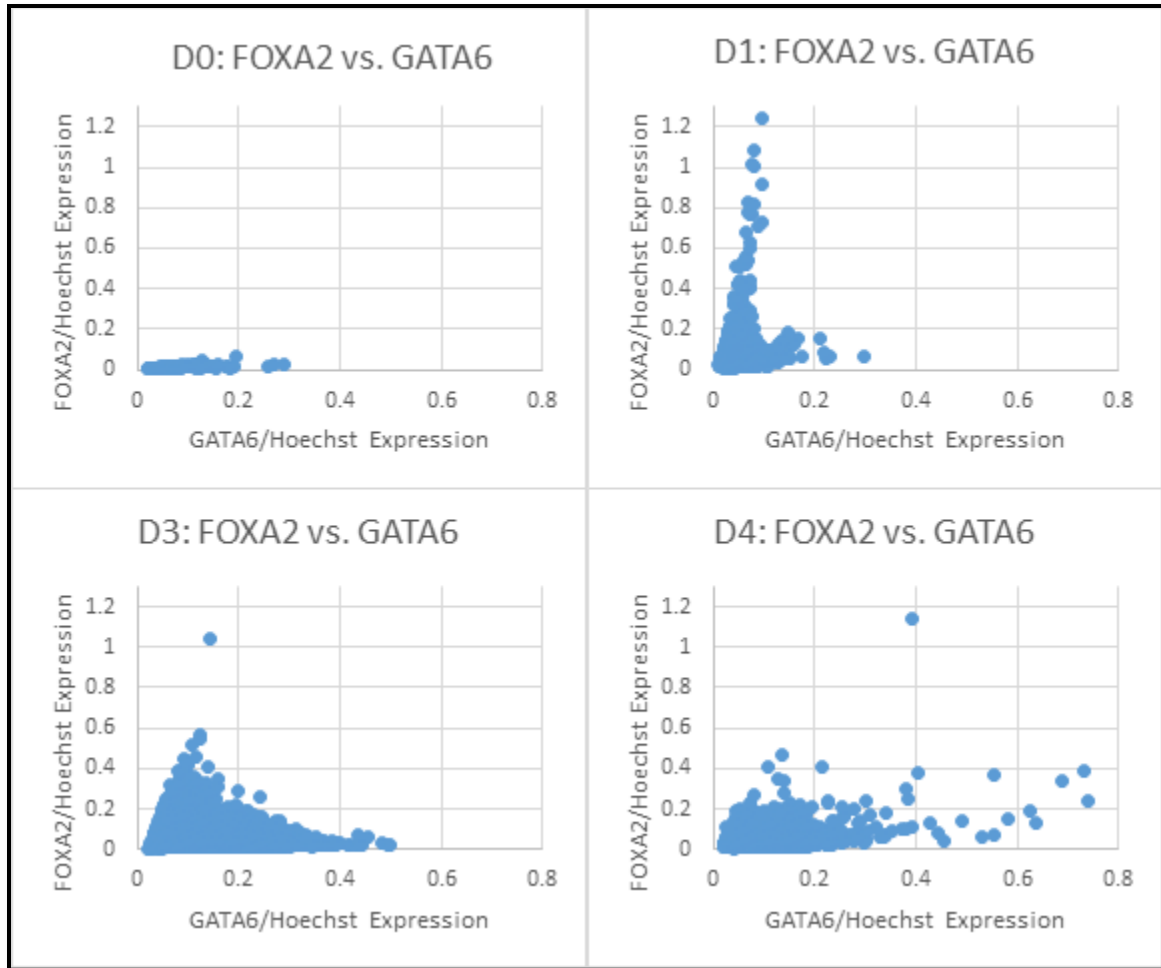
T VS. GATA6 D0, D1, D3, D4



Shown above are plots for day 0, day 1, day 3, and day 4 samples. The x-axis is the normalized GATA6 expression (the measured GATA6 intensity divided by the measured Hoechst intensity). The y-axis is the normalized T expression (the measured T intensity divided by the measured Hoechst intensity).

APPENDIX B

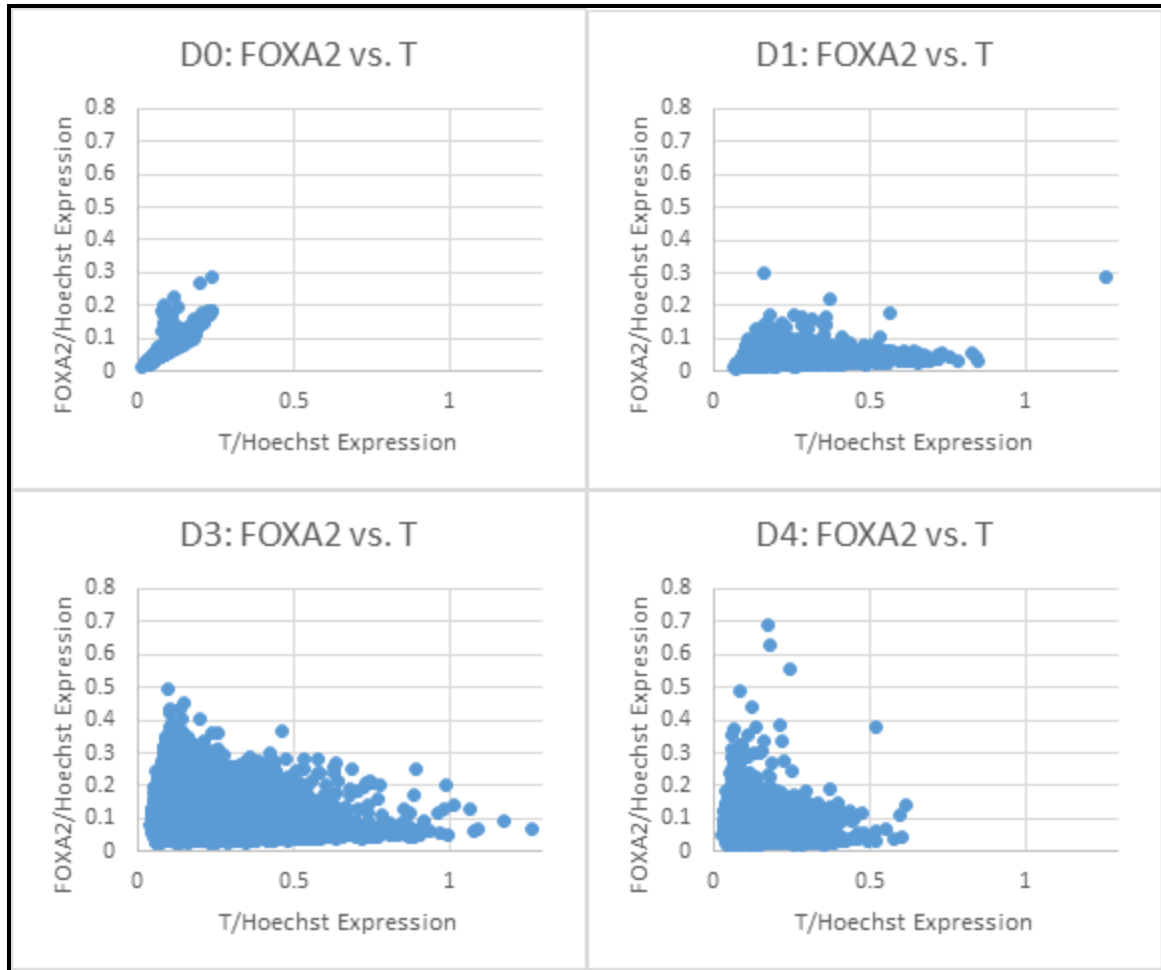
FOXA2 VS. GATA6 D0, D1, D3, D4



Shown above are plots for day 0, day 1, day 3, and day 4 samples. The x-axis is the normalized GATA6 expression (the measured GATA6 intensity divided by the measured Hoechst intensity). The y-axis is the normalized FOXA2 expression (the measured FOXA2 intensity divided by the measured Hoechst intensity).

APPENDIX C

FOXA2 VS. T D0, D1, D3, D4



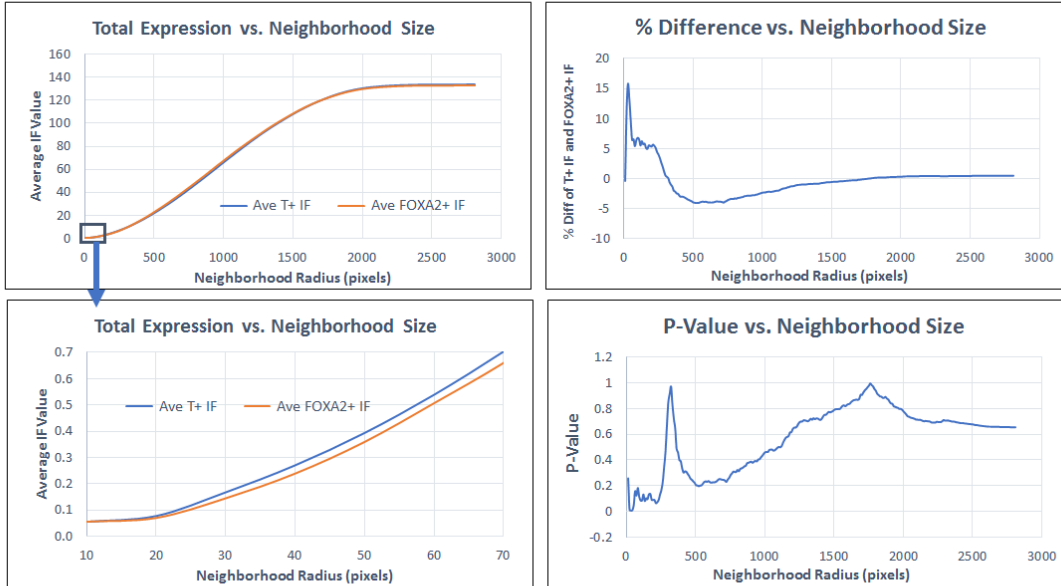
Shown above are plots for day 0, day 1, day 3, and day 4 samples. The x-axis is the normalized GATA6 expression (the measured GATA6 intensity divided by the measured Hoechst intensity). The y-axis is the normalized FOXA2 expression (the measured FOXA2 intensity divided by the measured Hoechst intensity).



APPENDIX D

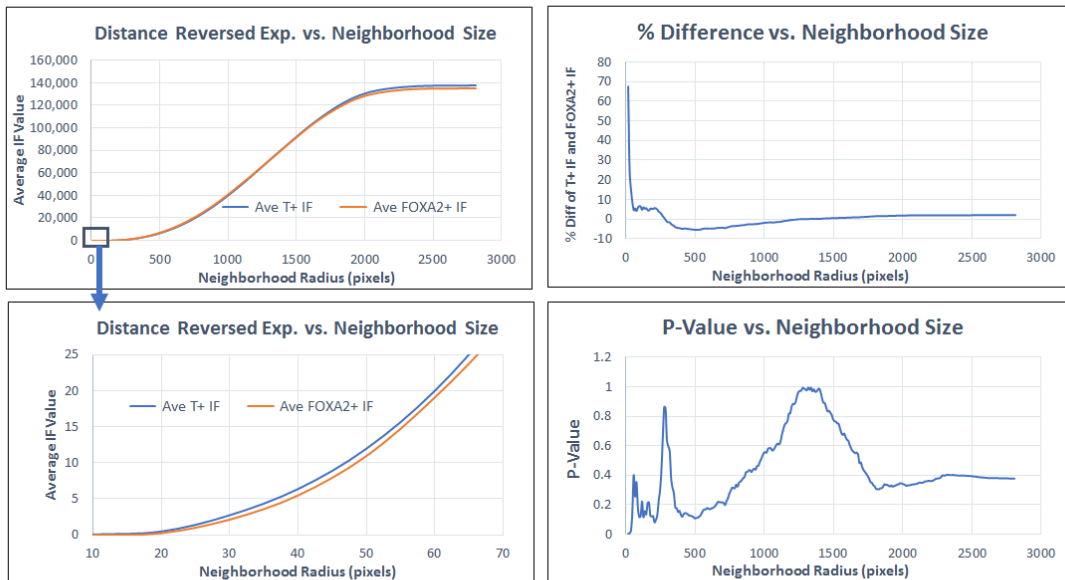
DETAILED NIF RESULTS FOR DAY 1: FOXA2, T, GATA6

## D1: T, FOXA2, GATA6 Total Expression Results



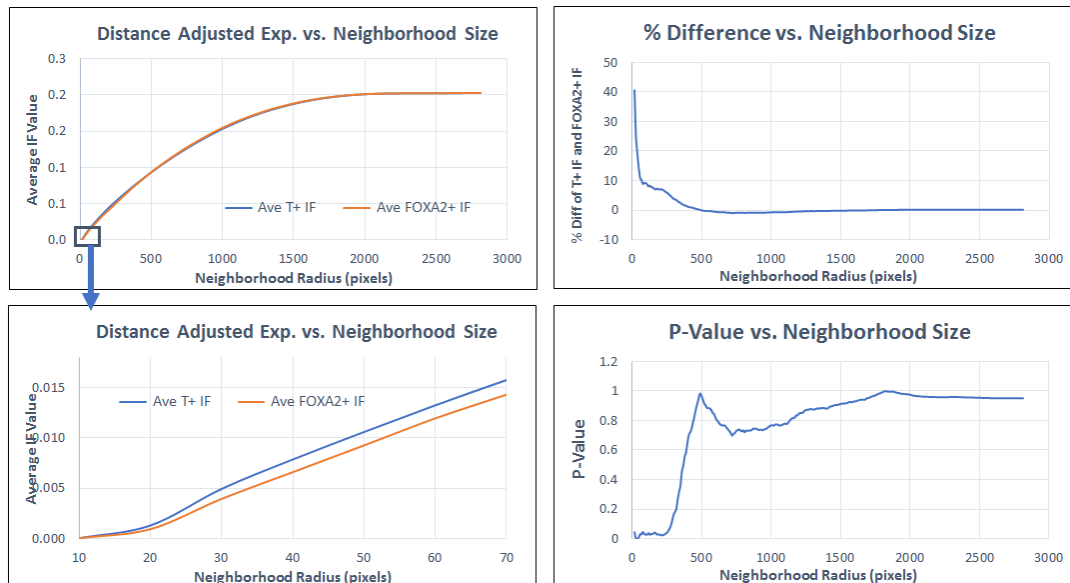
**Optimal Radius: 30 pixels, % Diff: 15.8%, P-value: 0.005**

## D1: T, FOXA2, GATA6 Distance Reversed Expression Results



**Optimal Radius: 20 pixels, % Diff: 67.7%, P-value: 0.002**

## D1: T, FOXA2, GATA6 Distance Adjusted Expression Results



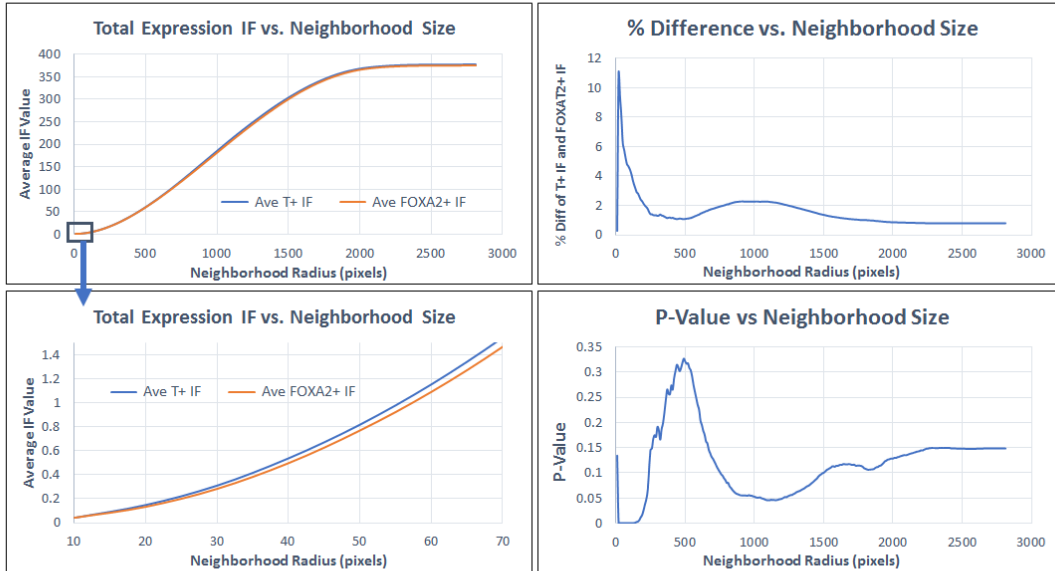
**Optimal Radius: 40, % Diff: 19.3%, P-value: 0.003**

Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 1 – FOXA2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and FOXA2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels.

APPENDIX E

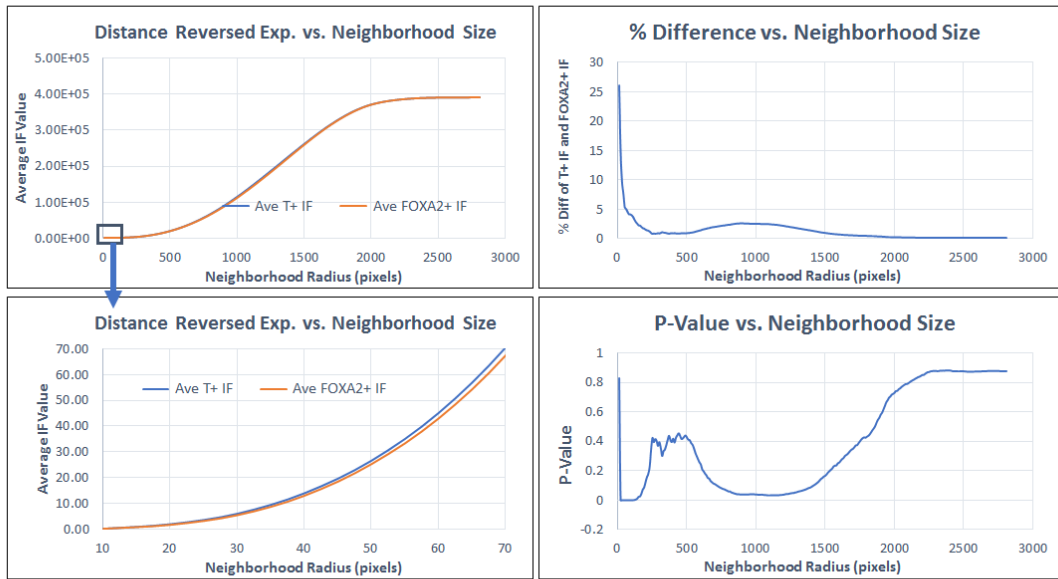
DETAILED NIF RESULTS FOR DAY 3: FOXA2, T, GATA6

### D3: T, FOXA2, GATA6 Total Expression Results



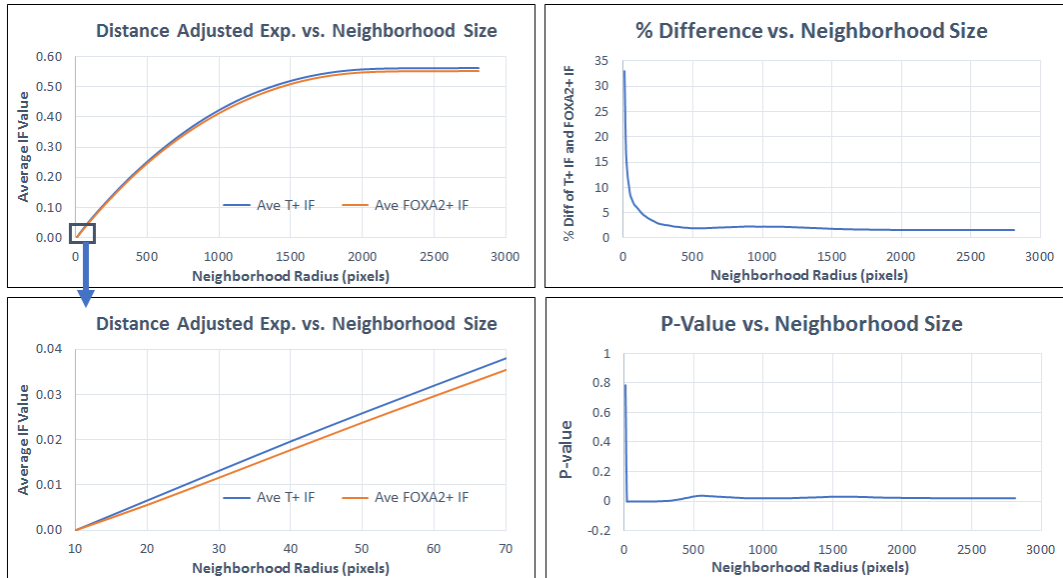
**Optimal Radius: 30, % Diff: 9.5%, P-value: 4.68E-11**

### D3: T, FOXA2, GATA6 Distance Reversed Expression Results



**Optimal Radius: 30 pixels, % Diff: 9.7%, P-value: 2.93E-8**

### D3: T, FOXA2, GATA6 Distance Adjusted Expression Results



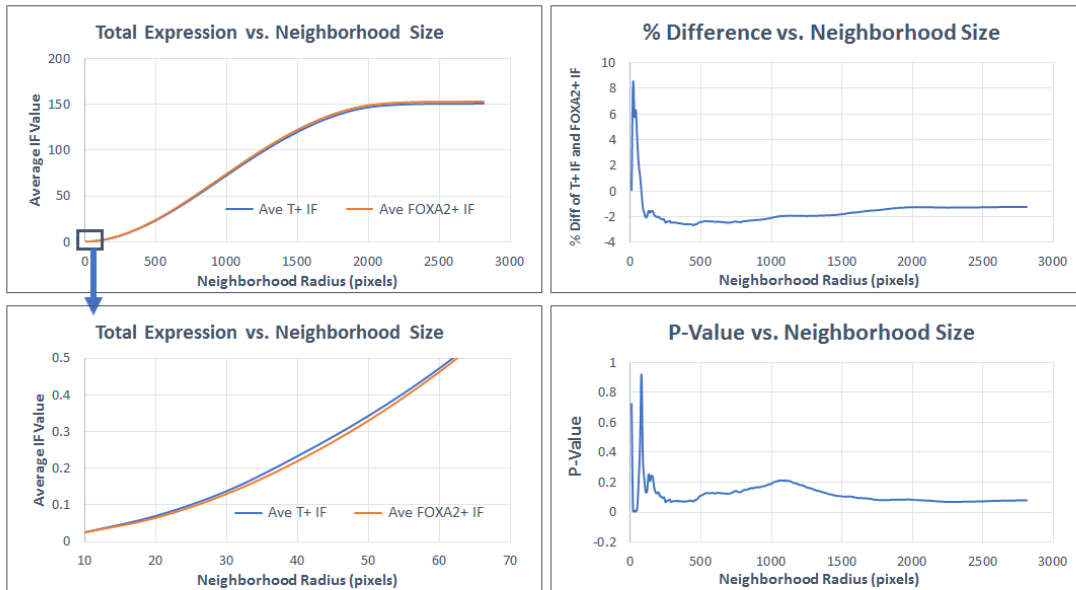
**Optimal radius: 30 pixels, % Diff: 13.0%, P-value: 1.54E-13**

Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 3 – FOXA2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and FOXA2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels.

APPENDIX F

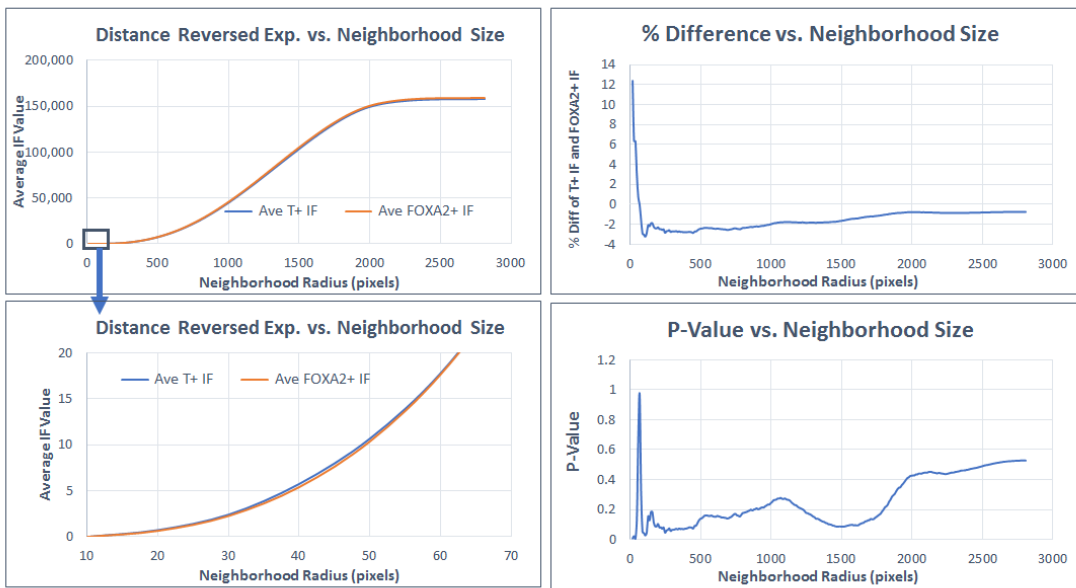
DETAILED NIF RESULTS FOR DAY 4: FOXA2, T, GATA6

## D4: T, FOXA2, GATA6 Total Expression Results



**Optimal Radius: 40 pixels, % Diff: 6.3%, P-value: 0.001**

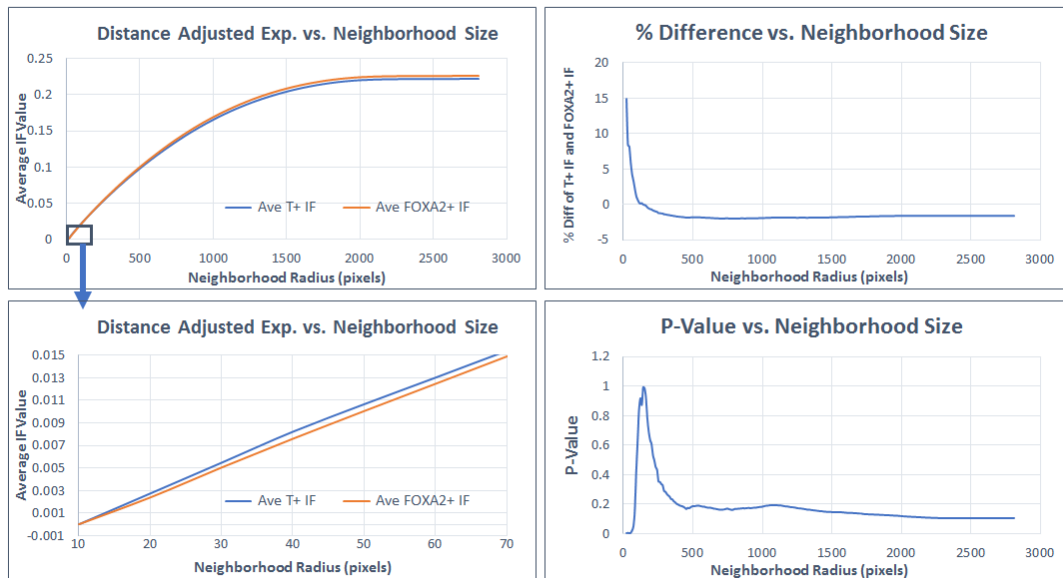
## D4: T, FOXA2, GATA6 Distance Reversed Expression Results



**Optimal Radius: 40 pixels, % Diff: 6.3%, P-value: 0.005**



## D4: T, FOXA2, GATA6 Distance Adjusted Expression Results



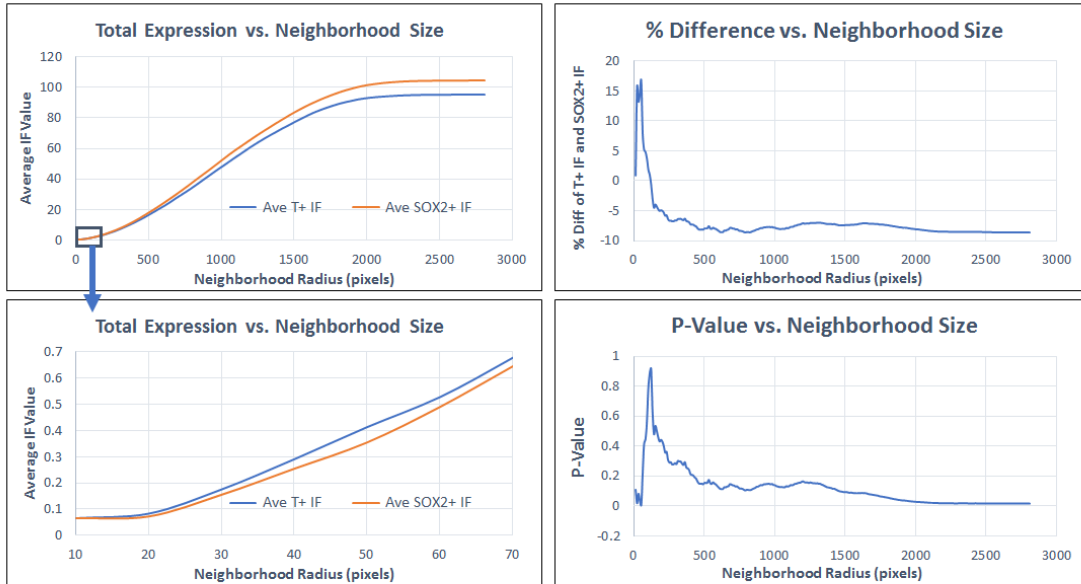
**Optimal Radius: 40 pixels, % Diff: 8.2%, P-value: 3.6E-4**

Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 4 – FOXA2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and FOXA2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels.

APPENDIX G

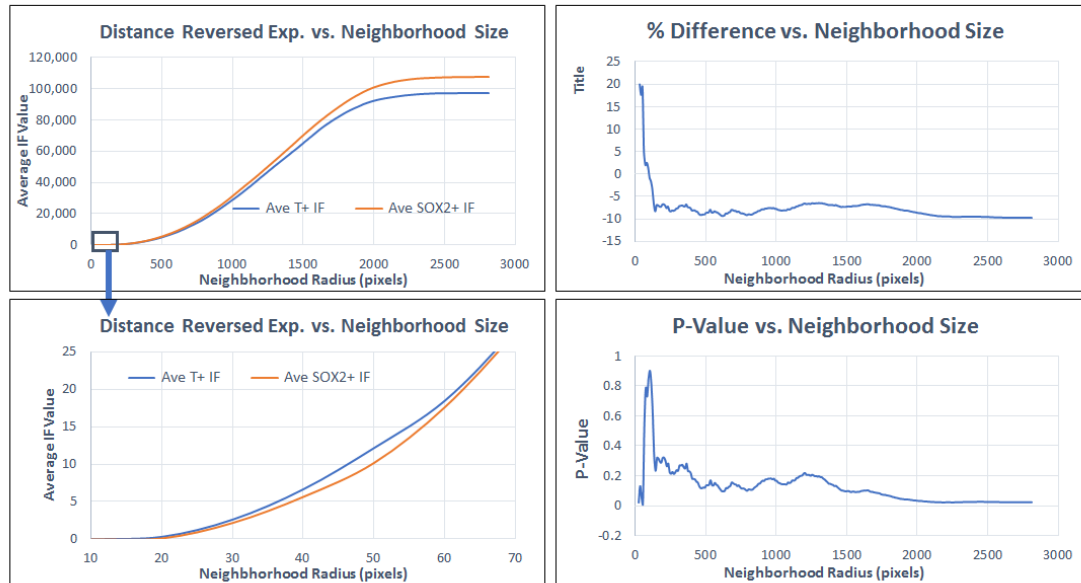
DETAILED NIF RESULTS FOR DAY 1: SOX2, T, GATA6

## D1: T, SOX2, GATA6 Total Expression Results



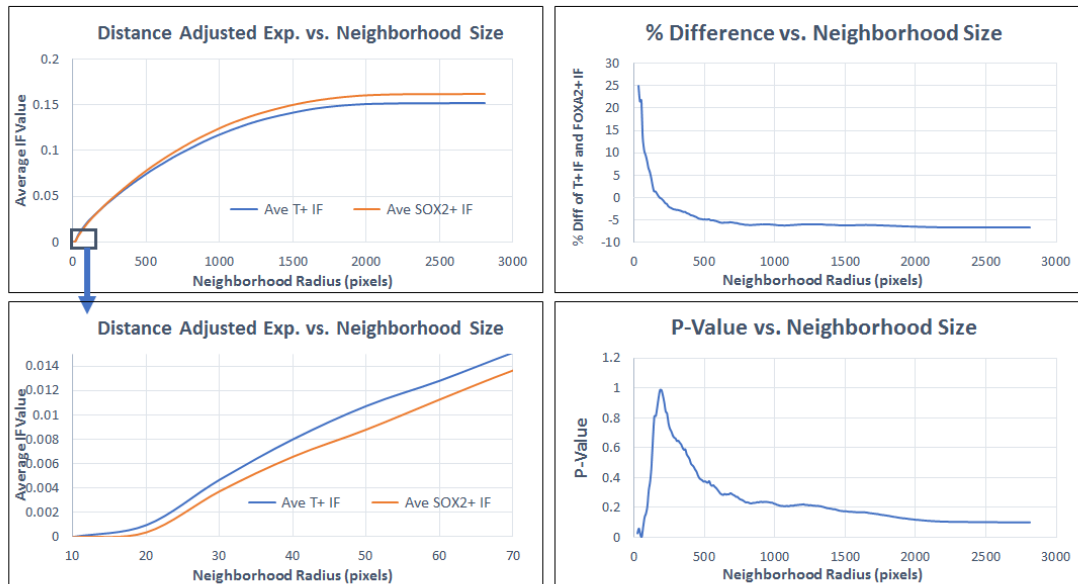
**Optimal Radius: 50 Pixels, % Diff: 16.6%, P-value: 0.007**

## D1: T, SOX2, GATA6 Distance Reversed Expression Results



**Optimal Radius: 50 Pixels, % Diff: 20.0, P-value: 0.013**

## D1: T, SOX2, GATA6 Distance Adjusted Expression Results



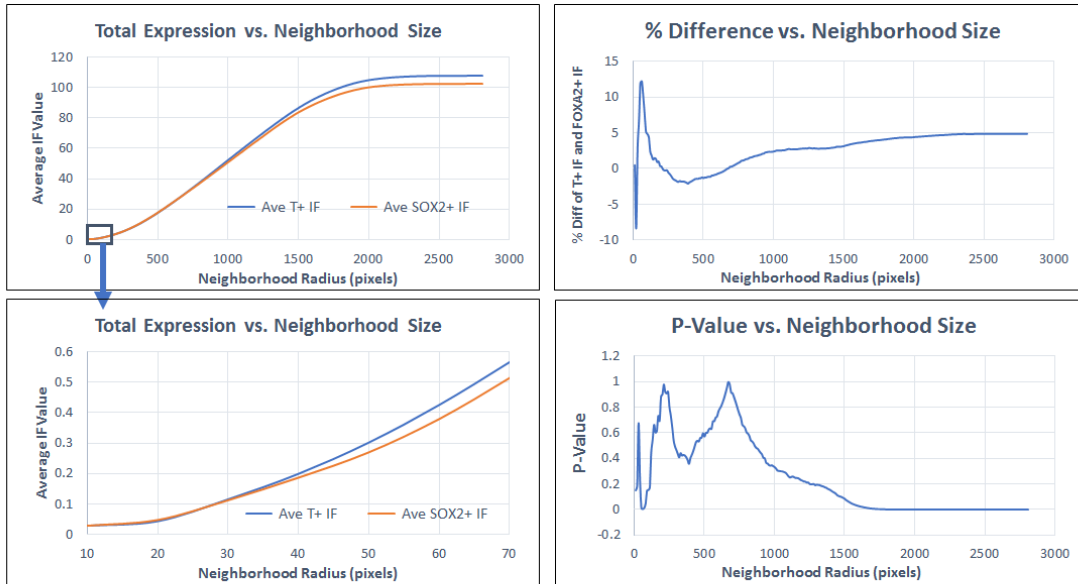
**Optimal Radius: 50 pixels, % Diff: 21.8%, P-value: 0.006**

Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 1 – SOX2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and SOX2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels. Below each set of charts the optimal radius is listed in pixels, along with its corresponding percent difference and p-value.

APPENDIX H

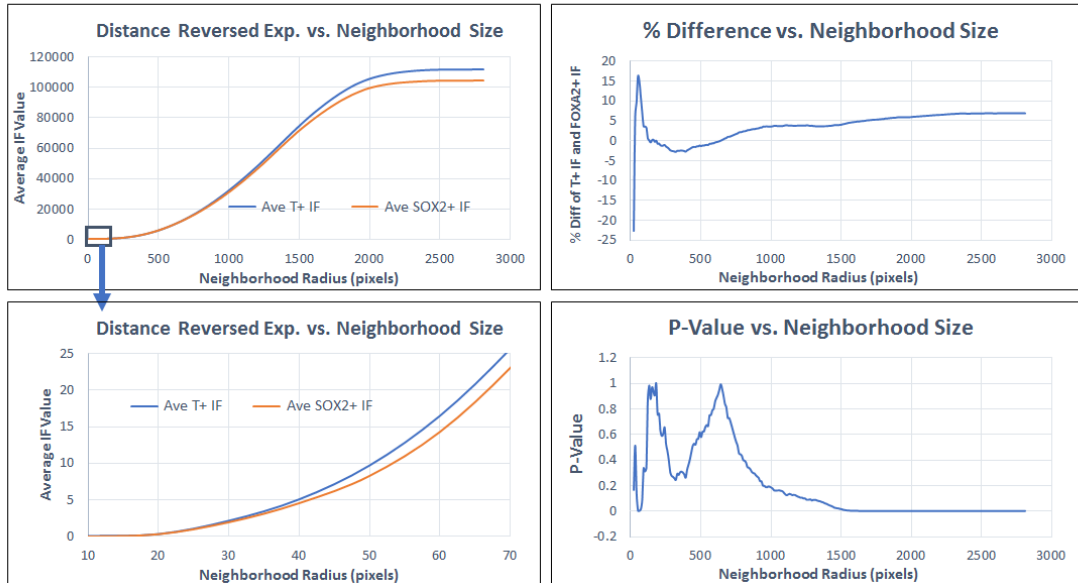
DETAILED NIF RESULTS FOR DAY 2: SOX2, T, GATA6

## D2: T, SOX2, GATA6 Total Expression Results



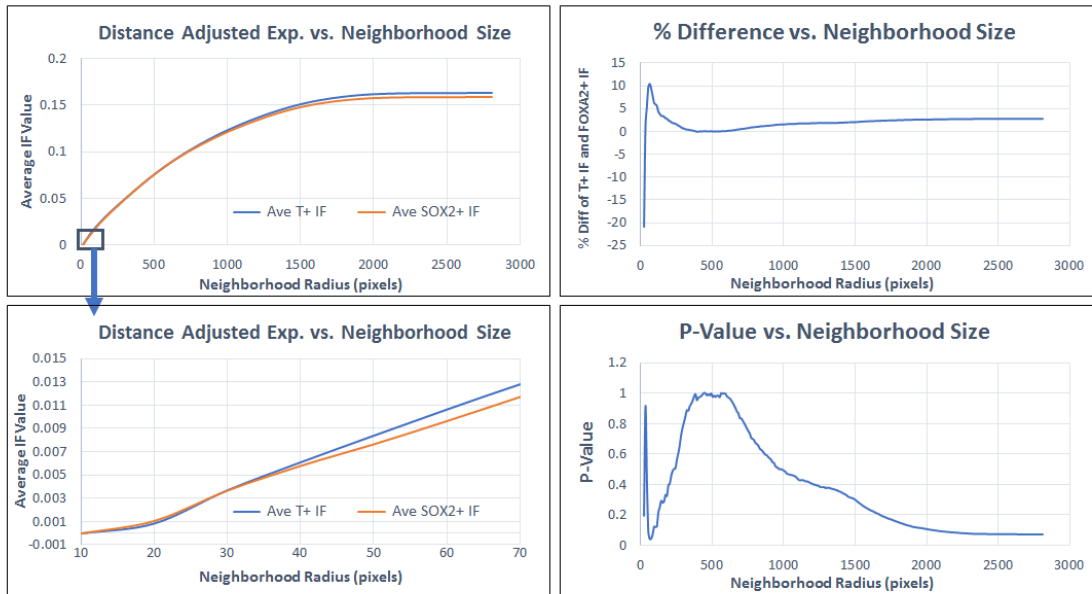
**Optimal Radius: 60 pixels, % Diff: 12.2%, P-value: 0.005**

## D2: T, SOX2, GATA6 Distance Reversed Expression Results



**Optimal radius: 50 pixels, % Diff: 16.1%, P-value: 0.0019**

## D2: T, SOX2, GATA6 Distance Adjusted Expression Results



**Optimal Radius: 60 pixels, % Diff: 10.3%, P-value: 0.043**

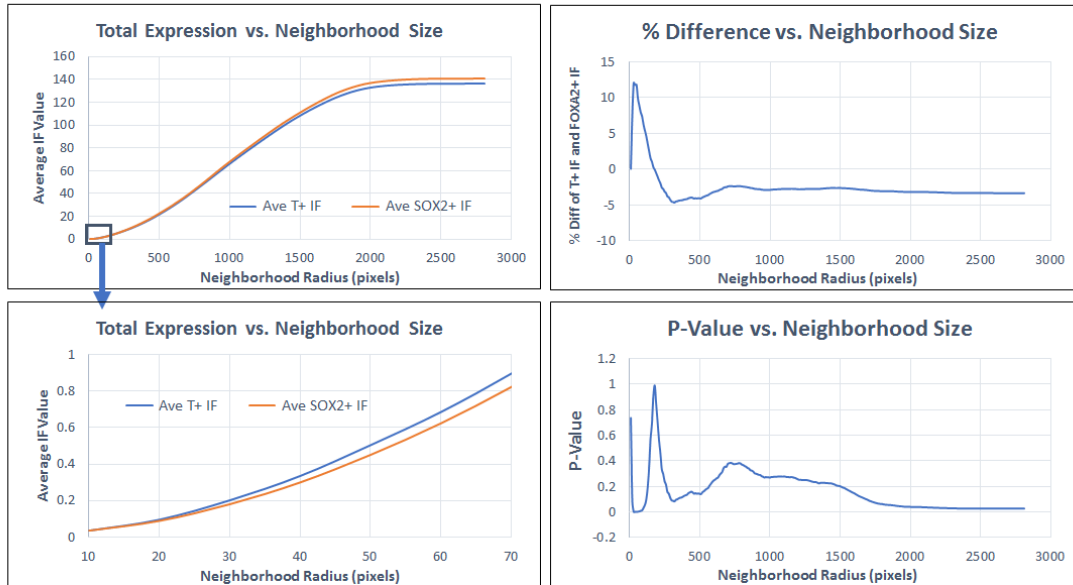
Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 2 – SOX2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and SOX2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels. Below each set of charts the optimal radius is listed in pixels, along with its corresponding percent difference and p-value.

APPENDIX I

DETAILED NIF RESULTS FOR DAY 5: SOX2, T, GATA6

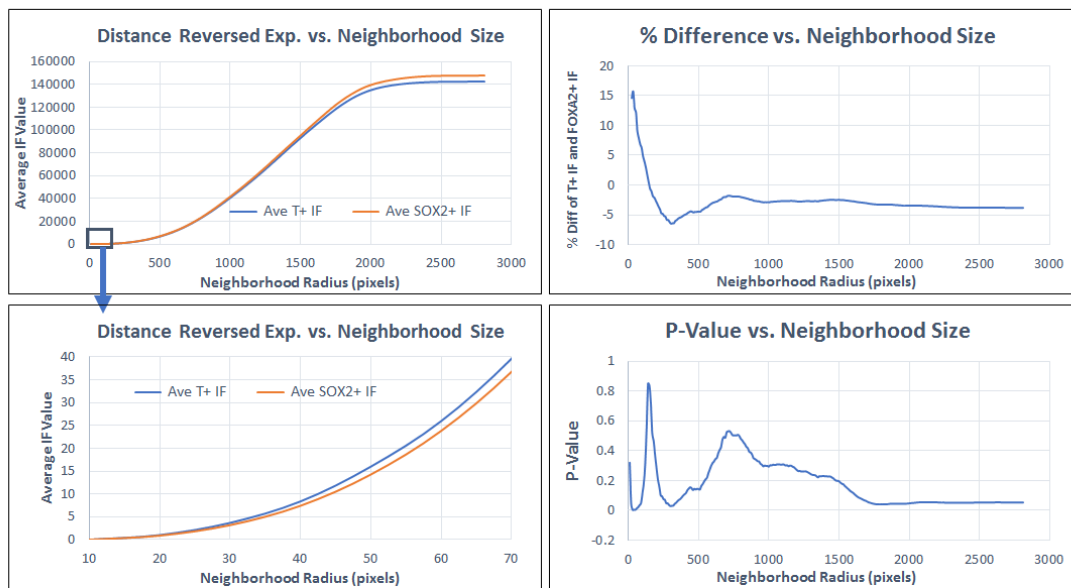


## D5: T, SOX2, GATA6 Total Expression Results



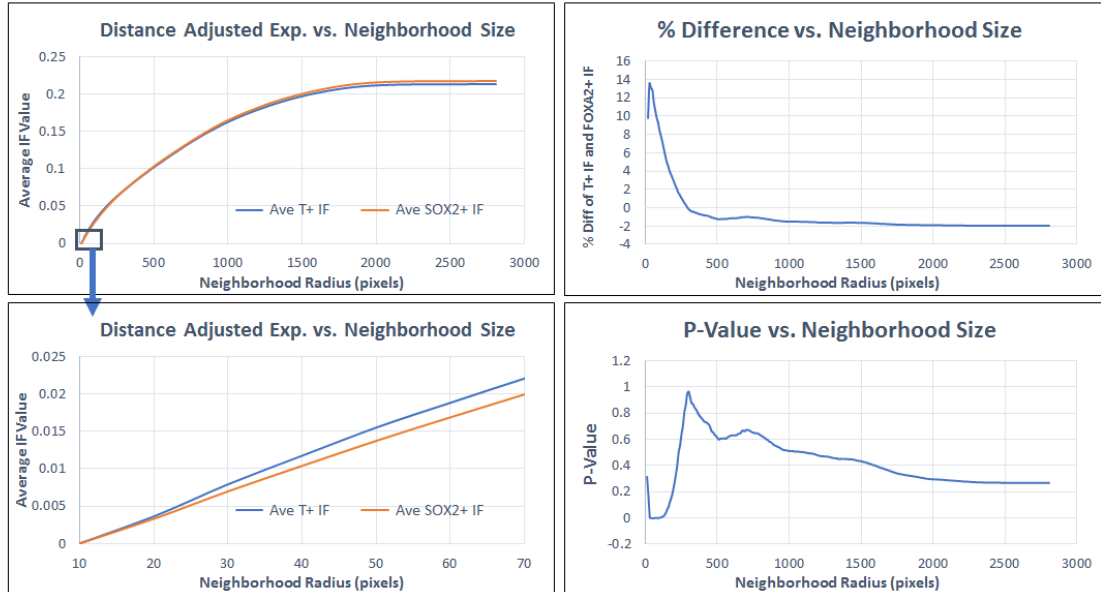
**Optimal Radius: 50 pixels, % Diff: 11.8%, P-value: 0.0008**

## D5: T, SOX2, GATA6 Distance Reversed Expression Results



**Optimal Radius: 30, % Diff: 15.6%, P-value: 0.0014**

## D5: T, SOX2, GATA6 Distance Adjusted Expression Results

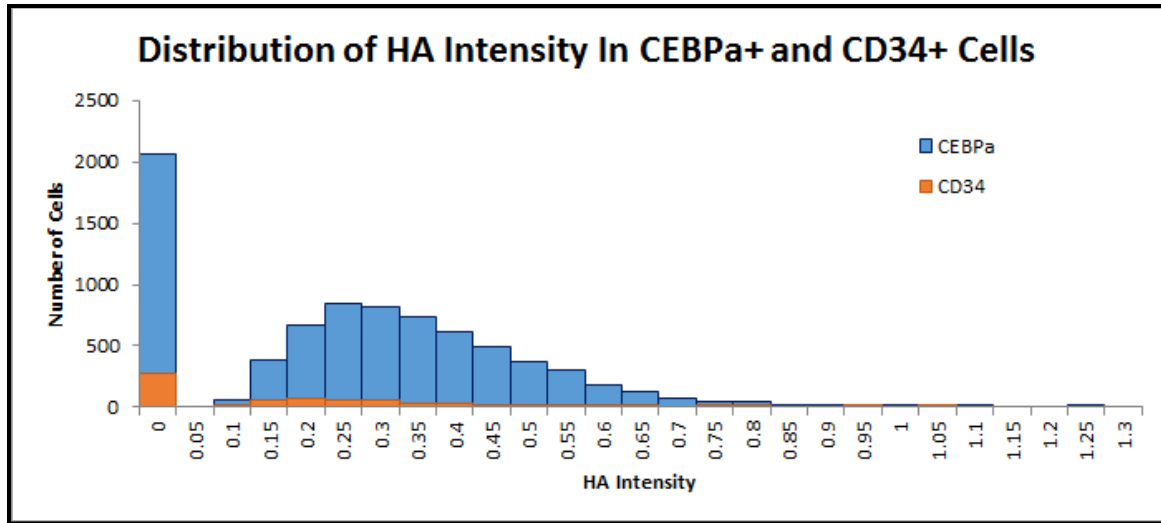


**Optimal Radius: 50 pixels, % Diff: 12.9%, P-value: 0.0009**

Shown above are the detailed results for three of the four designed neighborhood methods for the PGP1 – day 5 – SOX2, T, GATA6 staining. The three methods, in order, are Total Expression, Distance Reversed Expression, and Distance Adjusted Expression (Local Density is not shown because it did not seem to produce results signifying an impact on cell fate). The upper left corner of each segment of the figure shows the average IF of each population (T+ and SOX2+) vs the neighborhood radius from 10 to 2800 pixels (the entire extent of the image). The upper right of each segment shows the corresponding percent difference of the IF averages vs the radius, and the bottom right shows the corresponding P-value for the difference vs the radius. The bottom left corner shows a zoomed in version of the upper left, with the x-axis restricted from 10-70 pixels. Below each set of charts the optimal radius is listed in pixels, along with its corresponding percent difference and p-value.

## APPENDIX J

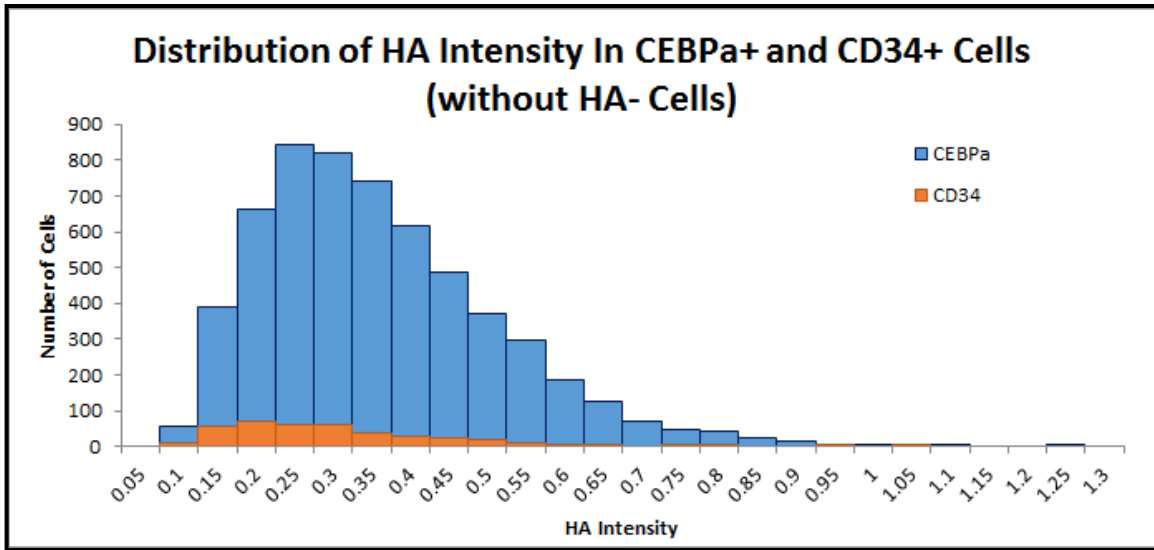
### DISTRIBUTION OF HA INTENSITY IN CEBP $\alpha$ + AND CD34+ CELLS



This plot was created by binning the HA values by every 0.05 intensity increase. The largest bin is at HA=0, and because of this data point, the rest of the histogram is hard to distinguish. It is less visible, but the peak HA value for the CD34 positive population (shown in orange) appears to be somewhere around HA = 0.2, whereas the peak HA value for the CEBP $\alpha$  positive population (shown in blue) appears to be closer to HA = 0.25 or 0.3.

## APPENDIX K

DISTRIBUTION OF HA IN CEBP $\alpha$ + AND CD34+ CELLS (WITHOUT HA- Cells)



This plot was created by binning HA values by every 0.05 intensity increase, and removing the bin of HA=0 so that the rest of the plot can be visualized more easily. Here it is easier to see that the peak HA value appears to be lower for the CD34 positive population (shown in orange) with HA ~ 0.2, than it is for the CEBP $\alpha$  positive population (shown in blue) with HA ~ 0.275.

## BIOGRAPHICAL SKETCH

Shay graduated from ASU with a Bachelor's of Science in Earth and Space Exploration in 2013. Since then she has worked full time as a software developer at the Mars Space Flight Facility here at ASU, and has been pursuing a Master's degree in Biomedical Engineering as a part time student. She has been involved with many sport clubs as a student at ASU and has won a National Championship with the wakeboarding team. Shay joined Dr. Ebrahimkhani's graduate research lab in December of 2015. She has been working on image processing and image analysis for cellular imagery. Her main area of study has been with human induced pluripotent stem cells that are used to develop a liver organoid. Developmental biology and the influence of specific gene expression and how that guides cell fate are the main areas of study in her thesis. She has developed a new way of analysis gene expression, with this concept of "Neighborhood Impact Factor".