Investigating the Role of Environmental Exposures and piRNA Expression in Breast Cancer Initiation and Progression

by

Katelyn Marie Polemi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Toxicology)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Justin A. Colacino, Co-Chair
Professor Dana C. Dolinoy, Co-Chair
Assistant Professor Monika Burness
Professor Maureen A. Sartor
Assistant Professor Laurie K. Svoboda

Katelyn M. Polemi

kmpolemi@umich.edu

ORCID iD: 0000-0002-1093-9631

Dedication

This dissertation is dedicated to my parents Dean and Rosemarie Polemi. Thank you for everything. Your endless love, patience, and belief in me have been the foundation of my strength and motivation.

Acknowledgements

I would like to thank my primary research mentors Dr. Dana Dolinoy and Dr. Justin Colacino, for their support, guidance, and encouragement throughout this journey. Their expertise, insightful feedback, and dedication have been invaluable, shaping my understanding and inspiring me to achieve more than I thought possible. I am so grateful for their patience and excitement as they shared their knowledge with me. Thank you for being exceptional role models and for your immeasurable contributions to my growth and success. I would also like to thank my committee members, Dr. Laurie Svoboda, Dr. Maureen Sartor and Dr. Monika Burness, all of whom shared their time and expertise with me throughout this process. This dissertation was also made possible by support from Dr. Kai Wang and Anagha Tapaswi, both of whom helped ensure these projects were completed.

I would like to extend my deepest gratitude to my family and friends who have supported me during this experience. Thank you to my parents, my brother Andrew Polemi and sister-in-law Danae Polemi, my Aunt Donna Lee Gates, and my Fiancé Jacob Byrd. Thank you all for always being there to celebrate my successes and lift me up during challenging times. Your sacrifices, guidance, and wisdom have been invaluable, and I am so grateful for everything you have done to help me reach this point. To my friends Rachel Morgan, Tomoko Ishikawa, Chanese Forte, Becca Ingwers, Samantha Herrin, Sabrina Stegman, and Mackenzie McCoy, your companionship, understanding, and constant cheerleading have been a source of strength and

inspiration. Thanks for always being there to listen, laugh, cry, and celebrate. To Rachel Morgan, thank you for being a sounding board for my ideas, for your patience in explaining complex concepts, and for all of your support. Finally, a HUGE thank you to my fiancé Jake Byrd. These last couple of years have not been easy, and you have stood by my side through my best days and all of my worst days.

Table of Contents

List of Tables

List of Figures

Abstract

Among women, breast cancer is the most prevalent form of cancer worldwide and has the second highest mortality rate of any cancer in the United States. Genetic predispositions are thought to account for 15-20% of all cases; therefore, 80 - 85% of cases occur in women with no family history of the disease. Consequently, the mechanisms of development of many breast cancers remain unknown. Breast tumors are heterogeneous, resulting from acquisition of morphological alterations and cancer hallmarks including stemness and cellular plasticity.

Epigenetics is defined as mitotically heritable changes in gene function that do not alter the underlying DNA sequence, and epigenetic mechanisms regulate cellular plasticity and breast carcinogenesis. Epigenetic mechanisms such as DNA methylation and histone modifications have been extensively studied in cancer; however, small non-coding RNA and their role in cancer progression remain unclear. PIWI-interacting RNA (piRNA) are a class of small, non-coding RNAs which regulate transposons and repression of transposition. A growing number of studies show expression of PIWILs in breast tumors; however, the role of piRNAs in breast cancer development remains unclear.

In the US, there are over 85,000 chemicals in use and only 3% of those are fully tested for human safety. These chemicals interact with our epigenome and may play an important role in carcinogenesis. Cadmium (Cd) is a naturally occurring heavy metal

and a known lung carcinogen, however, its role in breast cancer remains controversial. *In vitro* studies show that breast cells exposed to Cd are malignantly transformed through estrogen receptor independent mechanisms.

The overall goal of this dissertation is to examine the epigenomic, transcriptomic, and morphological changes linked to long term Cd exposure in breast cells and investigate how piRNAs are involved in breast carcinogenesis. In Aim 1, I define key players in piRNA biogenesis and machinery and use the Comparative Toxicogenomics Database (CTD) to assess what chemical exposures are associated with changes in piRNA-associated machinery expression as well as disease states linked to such changes in expression. My results indicate that aldehydes, metals, personal care products, pesticides, and polybrominated diphenyl ethers (PBDEs) impact expression of piRNA-associated genes. In Aim 2, I perform the first baseline characterization of the piRNA system in breast cells using two different cell lines (non-tumorigenic MCF10A and cancerous MCF7) in two culture conditions (2D-monolayer and 3D-mammospheres), with sodium periodate to identify piRNA transcripts. My results show distinct piRNA profiles in the two cell lines, as well as distinct piRNA expression patterns for 2D vs 3D culture conditions. In Aim 3, I investigate the role of long term (40-week), low dose cadmium chloride exposure (0.25μM and 2.5μM) on cancer-associated morphological alterations and cellular plasticity. My results show that the luminal marker, Keratin 8, decreases over time in both the control and treated groups, while the myoepithelial marker, Keratin 14, increases over time in the controls but decreases in the treated groups, with a divergence from the controls observed at week 30 and 40. RNA sequencing data indicated activation of the MYC oncogene, suggesting a potential

shift in cellular behavior and increased proliferation associated with long-term, low dose cadmium exposure. Taken together, these assessments contribute to our understanding of the piRNA profile of breast cancer and highlight novel new mechanisms by which cadmium may promote breast cancer progression.

Chapter 1

Introduction

Breast Cancer Background

Breast cancer is the most common cancer diagnosed in women and is the second leading cause of cancer-related mortality among women in the world (Menon et al. 2024). Treatment and survival rates vary among the different subtypes of breast cancer. There are four main molecular subtypes of breast cancer largely defined by gene expression profiling of three tumor markers, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (HER2). The subtypes are Luminal A ($ER^+/PR^+/HER2^-$), Luminal B ($ER^+/PR^+/HER2^{+/-}$), HER2-enriched ($ER^-/PR^-/HER2^+$), and Triple Negative Breast Cancer ($ER^-/PR^-/HER2^-$) (Prat *et al.*, 2015). Survival, treatment, and metastasis are all specific to the clinical subtype. The underlying mechanisms and initiating factors for each molecular subtype remain unclear. Of the four main subtypes, Triple Negative Breast Cancer (TNBC) tumors are the most aggressive. The aggressive nature of TNBC including its invasive and migrating properties are characteristic of cancer stem cells (Neophytou *et al.*, 2017). Metastasis is the leading cause of breast cancer associated deaths and most commonly occurs in the brain, bone, lungs, and liver (Jin and Mu, 2015).

Cancer Stem Cell Hypothesis

The cancer stem cell hypothesis states that tumors originate from tissue stem cells or their dysregulated progeny and that tumor growth is driven by a unique subgroup of cells which contain stem cell-like properties (Wicha *et al.*, 2006). Hallmark traits of stem cells include self-renewal and differentiation (Pattabiraman and Weinberg, 2014). Cancer stem cells (CSCs) are difficult to target and resistant to therapeutics, leading to poor clinical outcomes in tumors where they are present. Mammary stem cells (MaSCs) are multipotent stem cells which drive the development of the mammary gland with characteristics such as self-renewal and differentiation into all specialized mammary epithelial cells (Woodward, 2005). Oncogenic activity can occur at any point during the pathway from mammary stem cell to differentiation leading to transformation into breast cancer stem cells (BCSCs). BCSCs produce neoplastic progeny with unique traits such as dedifferentiation and tumor-initiation potential (Jiagge *et al.*, 2018). The majority of the tumor is comprised of non- BCSCs, lacking the self-renewal and dedifferentiating properties of BCSCs, however, the small subpopulation of BCSCs account for metastatic virulence and chemotherapeutic resistance. Non- targeted chemotherapeutics primarily focus on the bulk population of non-BCSCs forming the tumor, permitting the survival of the BCSC population and providing little disruption of their metastatic progression and recurrence (Jiagge *et al.*, 2018).

It has been proposed that the different molecular subtypes of breast cancer arise from cells at different positions within the differentiation pathway leading to significant cellular heterogeneity within each tumor (Brooks *et al.*, 2015). Black et al. 2010 described an embryonic stem cell-like (ES) gene expression signature in aggressive human tumors. They reported that expression of genes enriched in embryonic stem

cells were overexpressed in poorly differentiated tumors. NANOG is a transcription factor which suppresses tissue-specific gene expression, and therefore maintains pluripotency in embryonic stem cells (Zhao *et al.*, 2017). Stem cell pluripotency factors NANOG, Oct4, Sox2, and c-Myc have a higher frequency of expression in poorly differentiated tumors compared to well-differentiated tumors (Black *et al.*, 2010). This ES like signature was indicated to be associated with high-grade estrogen (ER) negative tumors, often of the basal-like subtype, and with poor clinical outcome (Black *et al.*, 2010). CSCs are comprised of at least two different phenotypic states, a proliferative epithelial- like state and a more invasive mesenchymal-like state, however, cells transitioning between these conditions can also result in a hybrid state (Brooks *et al.*, 2015). BCSCs maintain the plasticity to transition between these two states by mediation from epigenetic alterations (Pal *et al.*, 2015).

Cell of Origin Hypothesis

CSCs are known to maintain tumor propagation with self-renewal and multipotency attributes, conversely, the cell of origin (or tumor-initiating cell) is an independent concept for cancer initiation. The cell of origin refers to a normal cell which undergoes oncogenic transformation as a result of an oncogenic hit, resulting in the tumor initiating cell. This initial transformed cell then passes on its genetic mutations to other tumor cells (Gupta et al., 2005). Recently, Hanahan presented phenotypic plasticity as a new hallmark of cancer, acknowledging that plasticity can function in a variety of ways, including dedifferentiation, where a cell reverses back to a progenitor-like state, bypassing of a progenitor cell’s terminal differentiation resulting in expansion of cancer cells in a partially differentiated progenitor-like state, and transdifferentiation

where committed cells switch to cell lineages (Hanahan, 2022). Phenotypic plasticity plays a role in the cellular origins of cancer, tumor heterogeneity, and metastasis (Gupta et al., 2019, Burkhardt et al., 2022). Studies aiming to identify the cell of origin in breast cancer found that aggressive basal-like cancers exhibit greater similarity to normal luminal progenitor cells(Lim et al., 2009). This suggests that basal-like breast cancers may originate from dysregulated luminal progenitor cells rather than basal stem cells(Keller et al., 2012). Further, these cells acquire phenotypic plasticity and shift from luminal-like to basal-like characteristics(Chiche et al., 2019).

Luminal-to-basal Transition

The luminal to basal transition in breast cancer refers to when cells shift from a luminal phenotype to a more aggressive basal-like phenotype (Koren et al. 2015). Key features of this process include changes in cellular morphology and gene expression, resulting in more aggressive tumor properties (Grosse-Wilde et al. 2015, Jolly et al. 2015) . Luminal cells are generally associated with a more differentiated state, responds to therapies, and has a better prognosis (Orrantia-Borunda et al. 2022), whereas basal-like cells are associated with a more stem-like state, a worse prognosis and resistance to therapies (Botti et al. 2019). Luminal cells are characterized by the expression of markers such as keratin 8 (KRT8) and keratin 18 (KRT18) while basal cells are characterized by the expression of markers keratin 5 (KRT5) and keratin 14 (KRT14) (Thong et al. 2020). Additionally, studies indicate that breast stem cells display phenotypic plasticity allowing them to transition between mesenchymal and epithelial states or exist in a hybrid luminal/basal state (Thong et al. 2020, Pasani et al. 2020). This hybrid state is associated with increased metastatic potential and stemness

characteristics and are identified by their co-expression of KRT8, a luminal marker, and KRT 14, a basal marker (Jolly et al. 2015, Grosse-Wilde et al. 2015, Thong et al. 2020). Hybrid basal/luminal cells are a subset of alveolar luminal cells in normal breast tissue and accumulate in ducts as people age (Gray et al. 2022). These cells express both basal/myoepithelial and hormone-sensing markers, leading to reduced commitment to a specific cell type, a characteristic which resembles the behavior seen in basal-like breast cancers such as triple negative breast cancer (TNBC) (Gray et al. 2022).

Several mechanisms known to regulate luminal to basal plasticity include genetic and epigenetic regulation, microenvironmental factors, microRNAs, and cell-cell interactions. Key transcription factors such as SLUG, ZEB1, and SOX9 have been shown to drive the luminal to basal transition through activation of NOTCH, WNT and PI3K/AKT signaling pathways. Additionally, changes in the expression levels of genes such as KRT5, KRT14, and VIM are associated with basal like phenotypes and can be regulated by the transcription factors above. Epigenetic regulation such as changes in DNA methylation and histone modifications have been shown to play a role in the luminal to basal transition. Hypermethylation of a luminal marker GATA3 has been shown to contribute to the loss of luminal characteristics. Studies suggest that histone deacetylases (HDACs) can repress luminal genes and promote basal-like gene expression.

Microenvironmental factors such as extracellular matrix components (ECM), hypoxia, and cytokines/growth factors can also promote basal-like traits and enhancement of cellular plasticity through changes in ECM composition, induced hypoxia inducible factors (HIFs), and cytokines such as TGF-B and IL-6. Finally, non-

coding RNAs such as microRNAs and long non-coding RNAs (lncRNAs) have also been implicated as mechanisms initiating the luminal to basal transition by regulation of gene expression post-transcriptionally and modulating chromatin structure.

Epithelial to Mesenchymal Transition

The epithelial to mesenchymal transition (EMT) is an established example of phenotypic plasticity and plays an essential role in tumor development and progression. Transitioning from an epithelial state to a mesenchymal state alters the adhesion molecules expressed by the cell therefore allowing the cell to assume the migratory and invasive behaviors of stem cells (Nieto *et al.*, 2016). EMT is known to be involved in tissue differentiation and wound healing through which cells develop stem cell-like characteristics such as proliferation, loss of apico-basolateral polarity, and dispersing of cell-cell junctions (Cheung *et al.*, 2015). Transcription factors known to regulate EMT including Slug, Snail, and Twist promote cell migration and invasion observed in TNBC metastasis (Neophytou *et al.*, 2018). EMT is characterized by the reduction of epithelial markers E-cadherin and claudin, increase of mesenchymal markers N-cadherin and vimentin, the secretion of matrix metalloproteinases MMPs, and cytoskeleton reorganization (Nieto *et al.*, 2016). EMT is thought to play a role in cancer progression linked to the ability of cells to migrate and invade. Clinical trials have indicated that after chemotherapy treatment, the surviving tumor cells display stem cell-like and EMT-like properties and gene expression profiles (Creighton *et al.*, 2009).

Epigenetic Gene Regulation

Epigenetics is broadly defined as mitotically heritable modifications to the genome that do not alter the underlying DNA sequence itself, and these mechanisms

play important roles in the regulation of gene expression. Generally, epigenetic mechanisms are classified under three categories: DNA methylation, histone modifications, and small, non-coding RNA (ncRNA) (Skvortsova et al. 2018). DNA methylation, the most studied epigenetic mechanism, is known to regulate gene expression by either recruiting proteins involved in gene repression or by inhibiting binding of transcription factors to DNA (Moore et al. 2013). Histone modifications control transcription through the regulation of chromatin conformation; loose structure of chromosomes allows transcription of a specific region, whereas a tight chromosome structure prevents transcription of a given region (Morgan and Shilatifard et al. 2020). Several classes of small ncRNAs regulate gene expression without modifying the underlying DNA sequence including micro-RNA (miRNA), small interfering RNA (siRNA), and PIWI-interacting RNA (piRNA). Of these small ncRNAs, miRNAs and siRNAs are the most well characterized in the roles of regulating gene expression and RNA interference, respectively (Zhang et al. 2019). piRNA are a class of ncRNAs that form a complex with PIWIL proteins to regulate gene expression through the regulation of transposable elements (TEs) (Aravin and Bourc'his, 2008). piRNA has been extensively characterized in the germline, however, recent evidence has suggested that piRNA are present in a tissue specific manner in somatic tissue (Perera et al. 2019).

The Role of PIWILs and piRNA in Cancers and Breast Cancer

Epigenetic regulation is critical in normal breast growth and development and maintains the transcriptional potential of genes. Traditional epigenetic mechanisms including DNA methylation and histone modification have both been extensively studied in breast cancer biology (Jovanovic *et al.*, 2010). However, the role of piRNA in breast

cancer development and progression remains unclear. PIWILs play a role in both piRNA biogenesis and function; therefore, understanding their role in breast cancer can help elucidate the function of piRNA. Many studies have shown the expression of PIWILs in tumors; however, different PIWILs may play unique roles depending on the tissue or tumor type (Erber *et al.*, 2020; Lu *et al.*, 2012; Meseure *et al.*, 2020). Recently, reactivation of PIWIL 1 and 2 has been identified in various types of tumors (Erber *et al.*, 2020). PIWILs have been shown to play significant roles in somatic tissues, specifically in stem cells (Zhang *et al.*, 2013a; Wu *et al.*, 2010). Several hypotheses indicate that the PIWI-piRNA complex may contribute to cancer development and progression by promoting CSCs capable of important properties such as EMT. Studies have found an association between PIWIL activity and EMT (Zhang *et al.*, 2013a). In vitro studies in in prostate cancer cell lines demonstrated that silencing the expression of PIWIL2 significantly decreased cell invasion and migration (Yang *et al.*, 2015). There is an abundance of conflicting data around PIWILs expression in breast tumors. Several studies have indicated PIWIL2 expression in breast cancer and implicated its role as a stem cell protein (Erber *et al.*, 2020; Zhang *et al.*, 2013b; Lu *et al.*, 2012). Additionally, one study showed PIWIL2 expression is increased in breast cancer and is associated with increased expression of the estrogen receptor (Lin Heng *et al.*, 2018). Another study showed PIWIL 2 and 4 were overexpressed in various breast cancer cell lines compared to normal mammary whereas PIWILs 1 and 3 were undetectable (Meseure *et al.*, 2020). Conflicting data showed increased expression of PIWIL1 in breast cancer and that it was associated with an advanced histological tumor grade and poorer clinical

outcome for patients (Cao *et al.*, 2016). The expression and roles of PIWILs in breast cancer are still unclear and therefore require further investigation.

Many studies have shown that piRNAs act as epigenetic regulators involved in carcinogenic processes including angiogenesis, invasiveness, growth, and metastasis of tumors (Dana et al. 2020). Recent studies have implicated specific piRNAs in the metastasis and progression of breast cancer (Ding *et al.*, 2021; Zhang *et al.*, 2013a; Huang *et al.*, 2013; Chalbatani *et al.*, 2019). Although these studies have implicated piRNA expression in breast cancer, none of the studies performed sodium periodate treatment to validate the transcript as an actual piRNA. Most studies validate the piRNA transcript they are interested in using by referencing the piRNABank or piRNABase, however, these web resources on classified and clustered piRNAs have not been validated using a reliable method such as sodium periodate treatment. Isolated smRNA is treated with sodium periodate, which elicits a Beta-elimination reaction that piRNA containing a 2'Omethylation are resistant to, therefore enriching piRNA within the sample prior to sequencing (Ohara *et al.*, 2007).

Several studies have shown not only expression of piRNA in breast cancer but have determined their regulatory function and mechanism. One study observed that piR-021285 is involved in breast tumorigenesis by promoting invasiveness through DNA methylation (Fu *et al.*, 2015). A recent review by Qian et al. 2021 lists piRNA transcripts involved in breast cancer including whether it is upregulated/downregulated, has a regulatory function and has a mechanism (Qian *et al.*, 2021). However, the mechanism behind the up-regulation or deregulation of the piRNAs involved in breast cancer is still undetermined.

Overview of Chemical Exposure and Its Effect on Breast Cancer Risk

Gray et al. 2017 introduced a series of concepts to support a link between environmental toxicants and their influence on increased risk of breast cancer. The list of concepts included low-dose and non-monotonic responses, interactions between environmental toxicants, gene-environment interactions and epigenetic changes, cell-cell interactions and the timing of exposures. Studies have been conducted and indicate that there are disparities in our cellular response to low doses of chemicals compared to high doses of chemicals (Vandenberg, 2014). For example, carcinogenic effects of bisphenol A (BPA) on breast cancer have been indicated at lower concentrations in addition to higher concentrations of BPA (Wang *et al.*, 2017). Therefore, although the effects are dose dependent, they are not linear, and thus a range of concentrations of such chemicals from low to high need to be further studied.

In addition to each individual chemical having a unique mechanism of action that may drive carcinogenesis, it is important to recognize that people are exposed to a multitude of chemicals (Jiang *et al.*, 2018). The genetics of an individual can contribute to the increased risk of developing breast cancer, but there is still much unknown about gene-environment interactions and their consequence of increasing susceptibility to cellular changes. Importantly, besides more direct mechanisms including DNA damage and DNA adduct formation, chemicals contribute to the initiation of cancer by altering the epigenetic regulation of genes involved in cellular processes such as cell proliferation and signaling pathways altering gene expression (Gray et al. 2017).

Two additional concepts outlined by Gray et al. 2017 include the importance of cell-cell interactions and the consequence of disruption and susceptibility during stages

of development. The Tissue Organization Field Theory (TOFT) acknowledges that cell proliferation is the default state for cells and chemical signals regulate cell interactions with neighboring cells in an organ (Soto and Sonnenschein, 2011). Interruptions of cell-cell interactions alter gene expression and influence the development of diseases such as breast cancer. Finally, susceptibility is dependent on stages of development and the duration of exposure. The timing of exposure and duration of the exposures influence the magnitude of toxic effects.

Chemical Exposure: Cadmium

Cadmium (Cd) is a naturally occurring toxic metal found in small amounts in the soil, water, air, and food. Exposure to cadmium occurs through ingestion of contaminated food and water and through inhalation of cigarette smoke and industrial pollution (Genchi et al. 2020). Cd absorption takes place mainly in the respiratory tract and to a lesser extent via the gastrointestinal tract; it then enters the blood stream and accumulates in the kidneys, liver, and gut (Satarug, 2018; Tinkov et al. 2018). Cd is excreted slowly from the body through urine, saliva, and milk during lactation (Godt et al. 2006; Shawahna et al. 2023). Cadmium has been suggested to contribute to the development of lung, bladder, prostate and pancreas cancers (Mezynsk and Brzosk, 2018).

Cadmium is a well-established human health risk, however its role in breast cancer remains controversial. Although it and has been implicated in breast cancer initiation and promotion by numerous mechanistic studies (Tarhonska et al. 2023), multiple epidemiological studies have found null relationships (Filippini et al. 2020, Julin et al. 2012, Adams et al. 2012). This is combated with case- control studies that have

found higher concentrations of cadmium in urine of breast cancer cases compared to controls (Gallagher et al. 2010, Strumylaite et al. 2014). Additionally, exposure to cadmium in utero has been shown to alter mammary gland development and gene expression in mice (Parodi *et al.*, 2017). A recent study found that breast cancer cells exhibited distinct epigenetic profiles after exposure to varying doses (1μM -60μM) of cadmium (Liang *et al.*, 2020). Several studies have shown that long-term exposure to low doses of cadmium increases EMT. Ponce et al. 2015 treated breast cancer epithelial cells from the MCF-7 cell line, with cadmium for 6 months resulting in decreased expression of E-cadherin, a characteristic of EMT. Another study by (Benbrahim-Tallaa *et al.*, 2009) treated the non-malignant breast epithelial cell line MCF10A with cadmium for 40 weeks resulting in cellular transformation to a more basal-like cancer phenotype. The mechanism underlying these long-term cadmium exposures and increased EMT remains unclear. Wei et al. 2017 conducted a 4-week long experiment using MCF10A cells to show that cadmium promotes EMT through modulation of Snail. However, the mechanism of cadmium induced Snail is still unknown. Another short-term study using normal breast epithelial cells showed that cadmium at doses relevant to human exposure induced alterations in breast stem cell proliferation and differentiation by inhibiting HIF-1a (Rocco *et al.*, 2018). Taken together, these data indicate a gap in knowledge of the mechanism underlying cadmium exposure and breast cancer progression.

Experimental Design

The research in this dissertation utilizes *in vitro* models to determine how piRNA expression changes relative to the differentiation state of breast cancer and elucidate

how long-term, low-dose exposure of cadmium affects differentiation state, acquisition of stem cell-like properties and transcriptional properties of MCF10A cells. Additionally, this dissertation focuses on a secondary analysis of publicly available data from the Comparative Toxicogenomics Database to characterize piRNA processing and PIWIL family member expression. An overview of these experimental aims can be found in **Figure 1.1**.

Aim 1

The first aim of this dissertation characterizes piRNA processing and PIWIL family member expression using the Comparative Toxicogenomics Database (CTD). The process of piRNA transcript biogenesis and piRNA/PIWIL function in breast cancer is still unclear and current research remains contradictory; therefore, we hypothesized that providing a detailed review of the current available data on these processes would provide better characterized targets for future studies. Additionally, the effect of environmental exposures, such as chemicals, on piRNA/PIWIL expression and function are poorly understood. This secondary analysis of data from CTD will allow us to better prioritize chemicals of interest to investigate piRNA dysregulation by toxicants in numerous diseases.

A comprehensive review of literature was performed using PubMed on genes with a known association with piRNA biogenesis. Key phrases including, but not limited to, “piRNA biogenesis”, “piRNA machinery”, “PIWIL machinery”, were used to select our key list of genes. This list of genes and their functions can be found in Chapter 2, **Table 2.1**. Diseases highlighted in the literature search included breast cancer, numerous brain diseases, and cardiovascular health.

To generate the spreadsheets used in the analysis, the piRNA-associated gene list was input into CTD Batch Query. After inputting our gene list, we selected the data to download in CSV format including chemical-gene interactions and gene disease interactions. The resulting spreadsheets contained data entailing detailed gene-chemical-disease interactions collected in the published literature.

The spreadsheets provided by CTD allowed us to perform a chemical prioritization based on the known publications associated with the genes in our piRNA list. The top 50 chemicals were then used to determine the top 50 diseases that have an association with our piRNA gene list. Here we were able to determine the state of piRNA research in the field of environmental health and provide new hypothesis driven methodology.

Our hypothesis is that data from CTD will highlight what toxicants and disease outcomes are most commonly associated with our curated list of piRNA-related genes. This new hypothesis generating method will allow future studies to target genes and chemicals for investigation into mechanisms of action.

Aim 2

Aim 2 of this dissertation determines baseline piRNA expression in non-tumorigenic MCF10A and cancerous MCF7 cell lines grown in monolayer and identifies differences in piRNA expression by 2D and 3D states using mammospheres of MCF10A and MCF7 cells. The investigation into the role of piRNA in cancer is an increasingly popular topic; however, current methods often use inappropriate identification of piRNAs. For example, recent work has explored which piRNA transcripts are implicated in breast cancer and their function using piRNABANK as a

validator of piRNA (Krishnan *et al.*, 2016; Hashim *et al.*, 2014; Wang *et al.*, 2016). However, since piRNABANK does not validate using sodium periodate treatment, there is a risk of false positives, meaning that potential piRNA targets identified may not actually be piRNA transcripts. Sodium periodate treatment selects for small RNAs containing the 2'-O-methylation signature or piRNA (Ohara *et al.*, 2007). In addition, contradicting evidence on the expression of PIWIL proteins and piRNA in both normal tissue and cancer makes investigation into the role they play in cancer progression and metastasis extremely difficult. This aim will result in the most accurate piRNAome of two cell lines, MCF10A and MCF7, in two different stemness states with sodium periodate treatment for piRNA validation.

This aim seeks to address several gaps in knowledge: 1) a validated piRNAome of a human non-tumorigenic cell line, MCF10A, and a human cancer cell line, MCF-7 2) a validated piRNAome of these two cell lines in two different differentiation states, monolayer and mammospheres.

To isolate and culture mammary stem cells, Dontu et al., 2003 developed the mammosphere assay, where cells that are capable of forming spheres represent the mammary stem cells and can undergo limited self-renewal. This assay allows us to examine the breast stem cell proliferation capacity due to each mammosphere containing a single sphere forming stem cell. In this aim, we use the mammosphere assay to investigate the differential expression pattern of piRNAs in 2D conventional cell culture and this stem-like 3D mammosphere condition.

Mature piRNAs are denoted by a 2'-O-methylation at their 3' end placed by methyltransferase HEN1 during biogenesis (Kirino and Mourelatos et al. 2007). Sodium

periodate treatment enriches actual piRNAs in the sample by eliciting a Beta-elimination reaction which piRNAs are resistant to due to their 2'-O-methylation (Ohara et al. 2007). This treatment eliminates other transcripts that may be other forms of smRNA and not confirmed piRNA. This approach creates a comprehensive baseline profile for both 2D (monolayer) and 3D (mammosphere) states for MCF10A and MCF-7 cells which will allow future studies to 1) identify possible piRNA targets implicated in breast cancer progression and 2) compare the differences of piRNA expression when these cell lines are exposed to environmental factors. We hypothesized that non-tumorigenic ER- MCF10A cells and cancerous ER+ MCF-7 cells will exhibit different piRNA expression profiles, and 2D and 3D states will also exhibit distinct piRNA expression profiles.

Aim 3

The third aim of this work examines the phenotypic and morphological effects of chronic, low dose cadmium exposure on normal breast epithelial cell line MCF10A. MCF10A cells were kept in culture for 40-weeks under three conditions: Control – 0 μM, low dose - 0.25 μM, and high dose- 2.5 μM. The cells were split between ~70-90% confluency (ensuring no overgrowth of the cells). Three batches were performed to produce three biological replicates. Analysis for each batch was performed after the 40-weeks had been completed. In this aim, we profile key features of these cells after long-term, low dose exposure to cadmium including their differentiation state, their acquisition of cancer stem cell-like properties, and their transcriptional profiles.

The mammary gland is organized into a tree-like structure composed of hollow branches with an inner layer of luminal epithelial cells that face the lumen and are surrounded by an outer layer of myoepithelial cells (**Figure 4.3**). Both ductal and

alveolar luminal cells express keratin 8 and 18 (KRT8/18) genes. KRT8 dimerizes with KRT18 to form an intermediate filament in the cytoplasm of epithelial cells and plays an important role in the structural integrity of the cell and cellular differentiation (NIHa, 2024). The outer myoepithelial/basal layer expresses keratin 5 and 14 (KRT5/14). KRT14 is usually found as a heterotetramer with two KRT5 molecules and for the cytoskeleton of epithelial cells (NIHb,2024). A recent study by Thong et al. 2020 identified a hybrid population of cells which co-expressed both the luminal marker KRT8 and the basal marker K14. This evidence suggests that these cells may have a high plasticity and are transitioning between epithelial and mesenchymal cellular states, or that these cells may be suspended in a hybrid epithelial/mesenchymal state (Thong et al., 2020). Additionally in mammary glands, there are two breast stem cell populations including ALDH1A3+ luminal stem cells and CD44+/CD24- basal stem cells (Visvader and Stingl, 2014 and Van Keymeulen et al. 2011). Recent studies suggest that there is an additional population of cells that express both ALDH1A3+ and CD44+/CD24- and are more likely to form mammospheres than the ALDH+ cells alone (Colacino et al. 2018). In this aim, we investigate the role of long term (40-week) low dose Cd (0.25 μM and 2.5 μM) exposure on cancer stem cell markers and cellular plasticity in non-tumorigenic MCF10A cells.

We developed two high content image-based immunocytochemistry assays to measure the impact of the long-term cadmium exposure of the cells in an unbiased manner every 10 weeks starting at week 0 through week 40. Quantification of KRT8 and KRT14 (markers of luminal and basal cells, respectively) were used to test cell plasticity. Quantification of CD24-/CD44+ and ALDH1A3 expression (markers of cancer

stem cells in breast cancer) were used to test stemness. RNA sequencing and differential gene expression analysis were also performed from samples collected every 10 weeks and gene expression patterns analyzed via gene set enrichment and clustering analysis. Our hypothesis was that long-term, low dose cadmium exposure would result in a phenotypic shift and induce stem cell-like properties.

References

Alluri, P., & Newman, L. (2014). Basal-like and Triple Negative Breast Cancers: Searching For Positives Among Many Negatives. *Surgical Oncology Clinics of North America*, *23*(3), 567–577.

Baccarelli, A., & Bollati, V. (2009). Epigenetics and environmental chemicals. *Current Opinion in Pediatrics*, *21*(2), 243–251.

Benbrahim-Tallaa, L., Tokar, E. J., Diwan, B. A., Dill, A. L., Coppin, J.-F., & Waalkes, M. P. (2009). Cadmium Malignantly Transforms Normal Human Breast Epithelial Cells into a Basal-like Phenotype. *Environmental Health Perspectives*, *117*(12), 1847–1852.

Bessette, D. C., Tilch, E., Seidens, T., Quinn, M. C. J., Wiegmans, A. P., Shi, W., Cocciardi, S., McCart-Reed, A., Saunus, J. M., Simpson, P. T., Grimmond, S. M., Lakhani, S. R., Khanna, K. K., Waddell, N., Al-Ejeh, F., & Chenevix-Trench, G. (2015). Using the MCF10A/MCF10CA1a Breast Cancer Progression Cell Line Model to Investigate the Effect of Active, Mutant Forms of EGFR in Breast Cancer Development and Treatment Using Gefitinib. *PLoS ONE*, *10*(5), e0125232.

Botti, G., Cantile, M., Collina, F., Cerrone, M., Sarno, S., Anniciello, A., & Di Bonito, M. (2019). Morphological and pathological features of basal-like breast cancer. *Translational Cancer Research*, *8*(Suppl 5), S503–S509. https://doi.org/10.21037/tcr.2019.06.50

Breast Cancer Facts & Figures. (n.d.). Retrieved May 11, 2024, from

Brooks, M. D., Burness, M. L., & Wicha, M. S. (2015). Therapeutic Implications of Cellular Heterogeneity and Plasticity in Breast Cancer. *Cell Stem Cell*, *17*(3), 260–271.

Cao, J., Xu, G., Lan, J., Huang, Q., Tang, Z., & Tian, L. (2016). High expression of piwi-like RNA-mediated gene silencing 1 is associated with poor prognosis via regulating transforming growth factor-β receptors and cyclin-dependent kinases in breast cancer. *Molecular Medicine Reports*, *13*(3), 2829–2835.

Castillo Sanchez, R., Gomez, R., & Perez Salazar, E. (2016). Bisphenol A Induces Migration through a GPER-, FAK-, Src-, and ERK2-Dependent Pathway in MDA-MB-231 Breast Cancer Cells. *Chemical Research in Toxicology*, *29*(3), 285–295.

Chalbatani, G. M., Dana, H., Memari, F., Gharagozlou, E., Ashjaei, S., Kheirandish, P., Marmari, V., Mahmoudzadeh, H., Mozayani, F., Maleki, A. R., Sadeghian, E., Nia, E. Z., Miri, S. R., Nia, N. zainali, Rezaeian, O., Eskandary, A., Razavi, N., Shirkhoda, M., & Rouzbahani, F. N. (2018). Biological function and molecular mechanism of piRNA in cancer. *Practical Laboratory Medicine*, *13*, e00113.

Chen, S., Ben, S., Xin, J., Li, S., Zheng, R., Wang, H., Fan, L., Du, M., Zhang, Z., & Wang, M. (2021). The biogenesis and biological function of PIWI-interacting RNA in cancer. *Journal of Hematology & Oncology*, *14*(1), 93.

Chiche, A., Di-Cicco, A., Sesma-Sanz, L., Bresson, L., de la Grange, P., Glukhova, M. A., Faraldo, M. M., & Deugnier, M.-A. (2019). P53 controls the plasticity of mammary luminal progenitor cells downstream of Met signaling. *Breast Cancer Research*, *21*(1), 13.

Colacino, J. A., Azizi, E., Brooks, M. D., Harouaka, R., Fouladdel, S., McDermott, S. P., Lee, M., Hill, D., Madden, J., Boerner, J., Cote, M. L., Sartor, M. A., Rozek, L. S., & Wicha, M. S. (2018). Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*, *10*(5), 1596–1609.

Collins, A., & Politopoulos, I. (2011). The genetics of breast cancer: Risk factors for disease. *The Application of Clinical Genetics*, *4*, 11–19.

Creighton, C. J., Li, X., Landis, M., Dixon, J. M., Neumeister, V. M., Sjolund, A., Rimm, D. L., Wong, H., Rodriguez, A., Herschkowitz, J. I., Fan, C., Zhang, X., He, X., Pavlick, A., Gutierrez, M. C., Renshaw, L., Larionov, A. A., Faratian, D., Hilsenbeck, S. G., … Chang, J. C. (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13820–13825.

Dandawate, P. R., Subramaniam, D., Jensen, R. A., & Anant, S. (2016). Targeting cancer stem cells and signaling pathways by phytochemicals: Novel approach for breast cancer therapy. *Seminars in Cancer Biology*, *40–41*, 192–208.

Ding, X., Li, Y., Lü, J., Zhao, Q., Guo, Y., Lu, Z., Ma, W., Liu, P., Pestell, R. G., Liang, C., & Yu, Z. (2021). piRNA-823 Is Involved in Cancer Stem Cell Regulation Through Altering DNA Methylation in Association With Luminal Breast Cancer. *Frontiers in Cell and Developmental Biology*, *9*, 641052.

Doherty, L. F., Bromer, J. G., Zhou, Y., Aldad, T. S., & Taylor, H. S. (2010). In Utero Exposure to Diethylstilbestrol (DES) or Bisphenol-A (BPA) Increases EZH2

Expression in the Mammary Gland: An Epigenetic Mechanism Linking Endocrine Disruptors to Breast Cancer. *Hormones & Cancer*, *1*(3), 146–155.

Dong, P., Xiong, Y., Konno, Y., Ihira, K., Xu, D., Kobayashi, N., Yue, J., & Watari, H. (2021). Critical Roles of PIWIL1 in Human Tumors: Expression, Functions, Mechanisms, and Potential Clinical Implications. *Frontiers in Cell and Developmental Biology*, *9*, 656993.

Dontu, G., Abdallah, W. M., Foley, J. M., Jackson, K. W., Clarke, M. F., Kawamura, M. J., & Wicha, M. S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes & Development*, *17*(10), 1253–1270.

Erber, R., Meyer, J., Taubert, H., Fasching, P. A., Wach, S., Häberle, L., Gaß, P., Schulz-Wendtland, R., Landgraf, L., Olbricht, S., Jung, R., Beckmann, M. W., Hartmann, A., & Ruebner, M. (2020). PIWI-Like 1 and PIWI-Like 2 Expression in Breast Cancer. *Cancers*, *12*(10), 2742.

Ercan, C., van Diest, P. J., & Vooijs, M. (2011). Mammary Development and Breast Cancer: The Role of Stem Cells. *Current Molecular Medicine*, *11*(4), 270–285.

Fanelli, G. N., Naccarato, A. G., & Scatena, C. (2020). Recent Advances in Cancer Plasticity: Cellular Mechanisms, Surveillance Strategies, and Therapeutic Optimization. *Frontiers in Oncology*, *10*, 569.

Filippini, T., Torres, D., Lopes, C., Carvalho, C., Moreira, P., Naska, A., Kasdagli, M.-I., Malavolti, M., Orsini, N., & Vinceti, M. (2020). Cadmium exposure and risk of breast cancer: A dose-response meta-analysis of cohort studies. *Environment International*, *142*, 105879.

Fite, K. (2017). Dysregulation of Phospholipase D (PLD) isoforms increases breast cancer cell invasion. *Browse All Theses and Dissertations*.

Fu, A., Jacobs, D. I., Hoffman, A. E., Zheng, T., & Zhu, Y. (2015a). PIWI-interacting RNA 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. *Carcinogenesis*, *36*(10), 1094–1102.

Fu, A., Jacobs, D. I., Hoffman, A. E., Zheng, T., & Zhu, Y. (2015b). PIWI-interacting RNA 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. *Carcinogenesis*, *36*(10), 1094–1102.

Gallagher, C. M., Chen, J. J., & Kovach, J. S. (2010). Environmental cadmium and breast cancer risk. *Aging (Albany NY)*, *2*(11), 804–814.

Grosse-Wilde, A., Fouquier d'Hérouël, A., McIntosh, E., Ertaylan, G., Skupin, A., Kuestner, R. E., del Sol, A., Walters, K.-A., & Huang, S. (2015). Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. *PLoS ONE*, *10*(5), e0126522. https://doi.org/10.1371/journal.pone.0126522

Gray, G. K., Li, C. M.-C., Rosenbluth, J. M., Selfors, L. M., Girnius, N., Lin, J.-R., Schackmann, R. C. J., Goh, W. L., Moore, K., Shapiro, H. K., Mei, S., D'Andrea, K., Nathanson, K. L., Sorger, P. K., Santagata, S., Regev, A., Garber, J. E., Dillon, D. A., & Brugge, J. S. (2022). A Human Breast Atlas Integrating Single-Cell Proteomics and Transcriptomics. *Developmental Cell*, *57*(11), 1400-1420.e7. https://doi.org/10.1016/j.devcel.2022.05.003

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46.

Hashim, A., Rizzo, F., Marchese, G., Ravo, M., Tarallo, R., Nassa, G., Giurato, G., Santamaria, G., Cordella, A., Cantarella, C., & Weisz, A. (2014). RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. *Oncotarget*, *5*(20), 9901–9910.

Heng, Z. S. L., Lee, J. Y., Subhramanyam, C. S., Wang, C., Thanga, L. Z., & Hu, Q. (2018). The role of 17β-estradiol-induced upregulation of Piwi-like 4 in modulating gene expression and motility in breast cancer cells. *Oncology Reports*, *40*(5), 2525–2535.

Herceg, Z., & Vaissière, T. (2011). Epigenetic mechanisms and cancer: An interface between the environment and the genome. *Epigenetics*, *6*(7), 804–819.

Holoch, D., & Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nature Reviews. Genetics*, *16*(2), 71–84.

Huang, G., Hu, H., Xue, X., Shen, S., Gao, E., Guo, G., Shen, X., & Zhang, X. (2013a). Altered expression of piRNAs and their relation with clinicopathologic features of breast cancer. *Clinical and Translational Oncology*, *15*(7), 563–568.

Huang, G., Hu, H., Xue, X., Shen, S., Gao, E., Guo, G., Shen, X., & Zhang, X. (2013b). Altered expression of piRNAs and their relation with clinicopathologic features of breast cancer. *Clinical and Translational Oncology*, *15*(7), 563–568.

Jiagge, E., Chitale, D., & Newman, L. A. (2018). Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *The American Journal of Pathology*, *188*(2), 271–279.

Jin, X., & Mu, P. (2015). Targeting Breast Cancer Metastasis. *Breast Cancer : Basic and Clinical Research*, *9*(Suppl 1), 23–34.

Jolly, M. K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J. N., & Levine, H. (2015). Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Frontiers in Oncology*, *5*, 155. https://doi.org/10.3389/fonc.2015.00155

Jolly, M. K., Tripathi, S. C., Jia, D., Mooney, S. M., Celiktas, M., Hanash, S. M., Mani, S. A., Pienta, K. J., Ben-Jacob, E., & Levine, H. (2016). Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*, *7*(19), 27067–27084. https://doi.org/10.18632/oncotarget.8166

Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, *3*(6), 415–428.

Jovanovic, J., Rønneberg, J. A., Tost, J., & Kristensen, V. (2010). The epigenetics of breast cancer. *Molecular Oncology*, *4*(3), 242–254.

Kallunki, T., Barisic, M., Jäättelä, M., & Liu, B. (2019). How to Choose the Right Inducible Gene Expression System for Mammalian Studies? *Cells*, *8*(8), 796.

Khan, S., Suryavanshi, M., Kaur, J., Nayak, D., Khurana, A., Manchanda, R. K., Tandon, C., & Tandon, S. (2021). Stem cell therapy: A paradigm shift in breast cancer treatment. *World Journal of Stem Cells*, *13*(7), 841–860.

Kim, J.-Y., Choi, H.-G., Lee, H.-M., Lee, G.-A., Hwang, K.-A., & Choi, K.-C. (2017). Effects of bisphenol compounds on the growth and epithelial mesenchymal transition of MCF-7 CV human breast cancer cells. *Journal of Biomedical Research*, *31*(4), 358–369.

Kirino, Y., & Mourelatos, Z. (2007). Mouse Piwi-interacting RNAs are 2’-O-methylated at their 3’ termini. *Nature Structural & Molecular Biology*, *14*(4), 347–348.

Koren, S., & Bentires-Alj, M. (2015). Breast Tumor Heterogeneity: Source of Fitness, Hurdle for Therapy. *Molecular Cell*, *60*(4), 537–546. https://doi.org/10.1016/j.molcel.2015.10.031

KRT14 keratin 14 [Homo sapiens (human)]—Gene—NCBI. (n.d.). Retrieved May 7, 2024, from

Lee, J. H., Jung, C., Javadian-Elyaderani, P., Schweyer, S., Schütte, D., Shoukier, M., Karimi-Busheri, F., Weinfeld, M., Rasouli-Nia, A., Hengstler, J. G., Mantilla, A., Soleimanpour-Lichaei, H. R., Engel, W., Robson, C. N., & Nayernia, K. (2010a). Pathways of Proliferation and Antiapoptosis Driven in Breast Cancer Stem Cells by Stem Cell Protein Piwil2. *Cancer Research*, *70*(11), 4569–4579.

Lee, J. H., Jung, C., Javadian-Elyaderani, P., Schweyer, S., Schütte, D., Shoukier, M., Karimi-Busheri, F., Weinfeld, M., Rasouli-Nia, A., Hengstler, J. G., Mantilla, A., Soleimanpour-Lichaei, H. R., Engel, W., Robson, C. N., & Nayernia, K. (2010b). Pathways of Proliferation and Antiapoptosis Driven in Breast Cancer Stem Cells by Stem Cell Protein Piwil2. *Cancer Research*, *70*(11), 4569–4579.

Li, W., Martinez-Useros, J., Garcia-Carbonero, N., Fernandez-Aceñero, M. J., Orta, A., Ortega-Medina, L., Garcia-Botella, S., Perez-Aguirre, E., Diez-Valladares, L., Celdran, A., & García-Foncillas, J. (2020). The Clinical Significance of PIWIL3 and PIWIL4 Expression in Pancreatic Cancer. *Journal of Clinical Medicine*, *9*(5), 1252.

Liang, Z.-Z., Zhu, R.-M., Li, Y.-L., Jiang, H.-M., Li, R.-B., Tang, L.-Y., Wang, Q., & Ren, Z.-F. (2020). Differential epigenetic and transcriptional profile in MCF-7 breast cancer cells exposed to cadmium. *Chemosphere*, *261*, 128148.

Liao, T.-T., & Yang, M.-H. (2020). Hybrid Epithelial/Mesenchymal State in Cancer Metastasis: Clinical Significance and Regulatory Mechanisms. *Cells*, *9*(3), 623.

Lillo, M. A., Nichols, C., Seagroves, T. N., Miranda-Carboni, G. A., & Krum, S. A. (2017). Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Hormones & Cancer*, *8*(2), 90–99.

Liu, J. J., Shen, R., Chen, L., Ye, Y., He, G., Hua, K., Jarjoura, D., Nakano, T., Ramesh, G. K., Shapiro, C. L., Barsky, S. H., & Gao, J.-X. (2010). Piwil2 is expressed in various stages of breast cancers and has the potential to be used as a novel biomarker. *International Journal of Clinical and Experimental Pathology*, *3*(4), 328–337.

Liu, J., Zhang, S., & Cheng, B. (2018). Epigenetic roles of PIWI-interacting RNAs (piRNAs) in cancer metastasis (Review). *Oncology Reports*, *40*(5), 2423–2434.

Lu, Y., Zhang, K., Li, C., Yao, Y., Tao, D., Liu, Y., Zhang, S., & Ma, Y. (2012). Piwil2 Suppresses P53 by Inducing Phosphorylation of Signal Transducer and Activator of Transcription 3 in Tumor Cells. *PLoS ONE*, *7*(1), e30999.

Maleki Dana, P., Mansournia, M. A., & Mirhashemi, S. M. (2020). PIWI-interacting RNAs: New biomarkers for diagnosis and treatment of breast cancer. *Cell & Bioscience*, *10*(1), 44.

Menon, G., Alkabban, F. M., & Ferguson, T. (2024). Breast Cancer. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK482286/

Meseure, D., Vacher, S., Boudjemaa, S., Laé, M., Nicolas, A., Leclere, R., Chemlali, W., Champenois, G., Schnitzler, A., Lesage, L., Dubois, T., & Bieche, I. (2020).

Biopathological Significance of PIWI–piRNA Pathway Deregulation in Invasive Breast Carcinomas. *Cancers*, *12*(10), 2833.

Muñoz-de-Toro, M., Markey, C. M., Wadia, P. R., Luque, E. H., Rubin, B. S., Sonnenschein, C., & Soto, A. M. (2005). Perinatal Exposure to Bisphenol-A Alters Peripubertal Mammary Gland Development in Mice. *Endocrinology*, *146*(9), 4138–4147.

Murray, T. J., Maffini, M. V., Ucci, A. A., Sonnenschein, C., & Soto, A. M. (2007). Induction of mammary gland ductal hyperplasias and carcinoma in situ following fetal bisphenol A exposure. *Reproductive Toxicology (Elmsford, N.Y.)*, *23*(3), 383–390.

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., & Thiery, J. P. (2016). EMT: 2016. *Cell*, *166*(1), 21–45.

Ohara, T., Sakaguchi, Y., Suzuki, T., Ueda, H., Miyauchi, K., & Suzuki, T. (2007). The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated. *Nature Structural & Molecular Biology*, *14*(4), 349–350.

Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., & Ramírez-Valdespino, C. A. (2022). Subtypes of Breast Cancer. In H. N. Mayrovitz (Ed.), *Breast Cancer*. Exon Publications.

Pal, B., Chen, Y., Bert, A., Hu, Y., Sheridan, J. M., Beck, T., Shi, W., Satterley, K., Jamieson, P., Goodall, G. J., Lindeman, G. J., Smyth, G. K., & Visvader, J. E. (2015). Integration of microRNA signatures of distinct mammary epithelial cell types with their gene expression and epigenetic portraits. *Breast Cancer Research : BCR*, *17*(1), 85.

Parodi, D. A., Greenfield, M., Evans, C., Chichura, A., Alpaugh, A., Williams, J., Cyrus, K. C., & Martin, M. B. (2017). Alteration of Mammary Gland Development and Gene Expression by In Utero Exposure to Cadmium. *International Journal of Molecular Sciences*, *18*(9), 1939.

Pasani, S., Sahoo, S., & Jolly, M. K. (2020). Hybrid E/M Phenotype(s) and Stemness: A Mechanistic Connection Embedded in Network Topology. *Journal of Clinical Medicine*, *10*(1), 60. https://doi.org/10.3390/jcm10010060

Pasculli, B., Barbano, R., & Parrella, P. (2018). Epigenetics of breast cancer: Biology and clinical implication in the era of precision medicine. *Seminars in Cancer Biology*, *51*, 22–35.

Pattabiraman, D. R., & Weinberg, R. A. (2014). Tackling the cancer stem cells—What challenges do they pose? *Nature Reviews Drug Discovery*, *13*(7), 497–512. https://doi.org/10.1038/nrd4253

Perera, B. P. U., Tsai, Z. T.-Y., Colwell, M. L., Jones, T. R., Goodrich, J. M., Wang, K., Sartor, M. A., Faulk, C., & Dolinoy, D. C. (2019). Somatic expression of piRNA and associated machinery in the mouse identifies short, tissue-specific piRNA. *Epigenetics*, *14*(5), 504–521.

Polemi, K. M., Nguyen, V. K., Heidt, J., Kahana, A., Jolliet, O., & Colacino, J. A. (2021). Identifying the Link Between Chemical Exposures and Breast Cancer in African American Women via Integrated in Vitro and Exposure Biomarker Data. *Toxicology*, *463*, 152964.

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., Díez, M., Viladot, M., Arance, A., & Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, *24*, S26–S35. 8

PubChem. (n.d.). *KRT8—Keratin 8 (human)*. Retrieved May 7, 2024, from

Qian, L., Xie, H., Zhang, L., Zhao, Q., Lü, J., & Yu, Z. (2021). Piwi-Interacting RNAs: A New Class of Regulator in Human Breast Cancer. *Frontiers in Oncology*, *11*, 695077.

Redig, A. J., & McAllister, S. S. (2013). Breast cancer as a systemic disease: A view of metastasis. *Journal of Internal Medicine*, *274*(2), 113–126.

Ross, R. J., Weiner, M. M., & Lin, H. (2014). PIWI proteins and PIWI–interacting RNAs in the soma. *Nature*, *505*(7483), 353–359.

Sala-Hamrick, K. E., Tapaswi, A., Polemi, K. M., Nguyen, V. K., & Colacino, J. A. (2024). High-Throughput Transcriptomics of Nontumorigenic Breast Cells Exposed to Environmentally Relevant Chemicals. *Environmental Health Perspectives*, *132*(4), 047002.

Tarhonska, K., Lesicka, M., Janasik, B., Roszak, J., Reszka, E., Braun, M., Kołacińska-Wow, A., & Jabłońska, E. (2022). Cadmium and breast cancer – Current state and research gaps in the underlying mechanisms. *Toxicology Letters*, *361*, 29–42. https://doi.org/10.1016/j.toxlet.2022.03.003

Tam, W. L., & Weinberg, R. A. (2013). The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature Medicine*, *19*(11), 1438–1449.

Tan, L., Mai, D., Zhang, B., Jiang, X., Zhang, J., Bai, R., Ye, Y., Li, M., Pan, L., Su, J., Zheng, Y., Liu, Z., Zuo, Z., Zhao, Q., Li, X., Huang, X., Yang, J., Tan, W., Zheng,

J., & Lin, D. (2019). PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. *Molecular Cancer*, *18*(1), 9.

Tharmapalan, P., Mahendralingam, M., Berman, H. K., & Khokha, R. (2019). Mammary stem cells and progenitors: Targeting the roots of breast cancer for prevention. *The EMBO Journal*, *38*(14), e100852.

Thong, T., Wang, Y., Brooks, M. D., Lee, C. T., Scott, C., Balzano, L., Wicha, M. S., & Colacino, J. A. (2020). Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Frontiers in Cell and Developmental Biology*, *8*.

Tong, M., Deng, Z., Yang, M., Xu, C., Zhang, X., Zhang, Q., Liao, Y., Deng, X., Lv, D., Zhang, X., Zhang, Y., Li, P., Song, L., Wang, B., Al-Dherasi, A., Li, Z., & Liu, Q. (2018). Transcriptomic but not genomic variability confers phenotype of breast cancer stem cells. *Cancer Communications*, *38*(1), 56.

Van Keymeulen, A., Rocha, A. S., Ousset, M., Beck, B., Bouvencourt, G., Rock, J., Sharma, N., Dekoninck, S., & Blanpain, C. (2011). Distinct stem cells contribute to mammary gland development and maintenance. *Nature*, *479*(7372), 189–193.

Vandenberg, L. N., Maffini, M. V., Wadia, P. R., Sonnenschein, C., Rubin, B. S., & Soto, A. M. (2007). Exposure to Environmentally Relevant Doses of the Xenoestrogen Bisphenol-A Alters Development of the Fetal Mouse Mammary Gland. *Endocrinology*, *148*(1), 116.

Visvader, J. E., & Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes & Development*, *28*(11), 1143–1158.

Wang, K., Perera, B. P. U., Morgan, R. K., Sala-Hamrick, K., Geron, V., Svoboda, L. K., Faulk, C., Dolinoy, D. C., & Sartor, M. A. (2024). piOxi database: A web resource of germline and somatic tissue piRNAs identified by chemical oxidation. *Database*, *2024*, baad096.

Wang, Y., Shi, L., Li, J., Li, L., Wang, H., & Yang, H. (2019). Long-term cadmium exposure promoted breast cancer cell migration and invasion by up-regulating TGIF. *Ecotoxicology and Environmental Safety*, *175*, 110–117.

Wang, Z., Liu, N., Shi, S., Liu, S., & Lin, H. (2016). The Role of PIWIL4, an Argonaute Family Protein, in Breast Cancer. *The Journal of Biological Chemistry*, *291*(20), 10646–10658.

Wicha, M. S., Liu, S., & Dontu, G. (2006). Cancer Stem Cells: An Old Idea—A Paradigm Shift. *Cancer Research*, *66*(4), 1883–1890. https://doi.org/10.1158/0008-5472.CAN-05-3153

YANG, Y., ZHANG, X., SONG, D., & WEI, J. (2015). Piwil2 modulates the invasion and metastasis of prostate cancer by regulating the expression of matrix metalloproteinase-9 and epithelial-mesenchymal transitions. *Oncology Letters*, *10*(3), 1735–1740.

Zhang, H., Ren, Y., Xu, H., Pang, D., Duan, C., & Liu, C. (2013). The expression of stem cell protein Piwil2 and piR-932 in breast cancer. *Surgical Oncology*, *22*(4), 217–223.

Figures and Tables

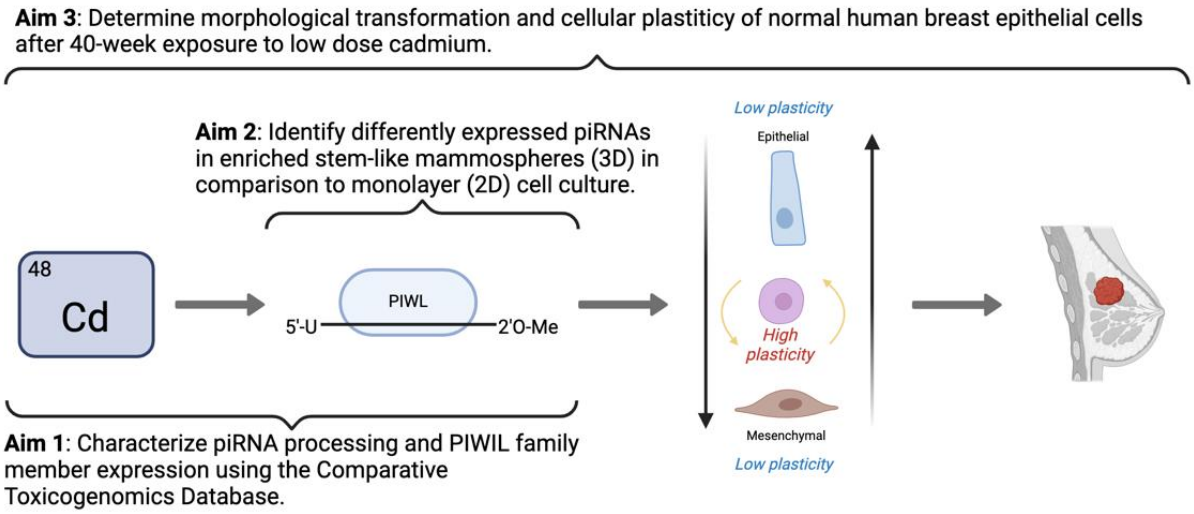


Figure 1.1: Schematic of Dissertation Aims.

Chapter 2

Assessment of Associations Among Environmental Toxicants and piRNA Epigenetic Machinery Using the Comparative Toxicogenomics Database

Abstract

In the US, there are over 85,000 chemicals in use and only 3% of those are fully tested for human safety. On average, people are exposed to thousands of chemicals daily and how these chemicals interact with our epigenome may play an important role in disease development and progression. Epigenetic mechanisms such as DNA methylation and histone modifications have been extensively studied in the context of environmental exposures and disease; however, small, non-coding RNAs have, by comparison, been relatively understudied. PIWI-interacting RNA (piRNA) are one such class known for their regulation of transposable elements (TEs). In humans, piRNA transcripts associate with 4 PIWIL (P-element induced wimpy testis like) proteins (PIWIL1-4) to direct the silencing of TEs via DNA methylation. Improper regulation of TEs have been implicated in the origin of numerous adverse health outcomes including

metabolic syndrome, neurological disorders, and cancer. piRNA are also differentially expressed in cancer, neurodegenerative diseases, and cardiovascular disease. Using the robust publicly available Comparative Toxicogenomic Database (CTD), we performed a comprehensive review of piRNA biogenesis, activities, and pathways associated with environmental toxicants. CTD identifies relationships between genes and chemicals, genes and diseases, and chemicals and disease in the published literature to provide a resource to investigate known interactions, and using an inference network, it predicts interactions that are likely to occur. We found that aldehydes, metals, personal care and consumer product compounds (PCCPCs), pesticides, and polybrominated diphenyl ethers (PBDEs) were linked to expression of piRNA-associated genes. Notably, bisphenol A (BPA) accounted for many of the PCCPC-associated changes, including changes in gene expression and alterations in DNA methylation of genes encoding for Tudor domain containing proteins (TDRDs) as well as PIWILs, both responsible for the biogenesis and processing of piRNAs. Additionally, the insecticide, parathion, and volatile organic compound (VOC), trichloroethylene, were associated with increases in DNA methylation at several piRNA-associated genes. This analysis will allow us to better prioritize chemicals of interest to investigate piRNA dysregulation by toxicants in the risk and outcomes of numerous diseases.

Introduction

An extensive body of evidence links chemical exposures to changes in epigenetic programming across multiple organisms and various life stages (Cantone and Fisher, 2013; Reik et al. 2001). Epigenetics is broadly defined as mitotically heritable changes in gene function that do not alter the underlying DNA sequence.

Epigenetic mechanisms play a key role in the maintenance of cell type-specific gene expression and are an interface between the environment and the genome. Moreover, widespread epigenetic reprogramming is evident in carcinogenesis, cancer metastasis, neurodegenerative diseases, and cardiovascular diseases (Zhang et al. 2020). Multiple epigenetic regulatory mechanisms are dysregulated by toxicant exposures, including DNA methylation and histone modifications. Epigenetic dysregulation may represent a common mechanism by which toxicants may impact the development of chronic diseases.

An emerging class of epigenetic regulators which have been understudied with respect to environmental exposures are the small non-coding RNAs, piRNA. piRNA work in concert with DNA methylation machinery to protect the integrity of the genome, primarily through the regulation of transposable elements (TEs) (Aravin and Bourc'his, 2008). piRNA are extremely important in the germline as they are responsible for repressing transposons and maintaining genomic integrity. Replication and reinsertion of TEs can introduce genetic mutations that result in dysregulated developmental processes and detrimental cellular changes (Garcia-Perez et al. 2016). While piRNA has been extensively studied in the germline, our group is interested in studying the role of this class of ncRNA in the soma and with regard to developmental processes and chronic diseases, such as cancer, cardiac disease, and neurodegenerative disease. Numerous studies have demonstrated piRNA dysregulation in diseases including breast cancer (Krishnan et al. 2016), cardiac disease (Rayford et al. 2021), neurological diseases (Sato et al. 2023), and metabolic diseases (Jacovetti et al. 2021). Even more specifically, studies have shown that piRNA act as epigenetic regulators involved in

carcinogenic processes including angiogenesis, invasiveness, growth, and metastasis of tumors (Dana et al. 2020), including the metastasis and progression of breast cancer (Ding *et al.*, 2021; Zhang *et al.*, 2013a; Huang *et al.*, 2013; Chalbatani *et al.*, 2019)

Developing *in silico* methods to predict which toxicant exposures are most likely to impact piRNA machinery would provide a new strategy to prioritize resources for experimental and epidemiological investigation. The need for prioritization strategies is particularly salient in light of the many chemicals to which individuals may be exposed, with nearly 85,000 chemicals currently in use in the US and about 1,000 new chemicals added annually (ATSDR, 2023). These chemicals are found in personal care products, food and food packaging, household cleaning products, and much more. The Comparative Toxicogenomics Database (CTD) provides an outstanding resource to develop these new strategies. The CTD is a comprehensive repository of scientific evidence pertaining to chemical exposures, their effects on the genome and epigenome, as well as their associations with various disease outcomes. The CTD curates data from epidemiological and toxicological investigations of genomic, transcriptomic, epigenomic, and proteomic effects of chemical exposures and ranks chemical-disease inferences based on local network topology of chemicals, genes and diseases. Thus, the CTD is a powerful tool when assessing gene-toxicant-disease relationships and in designing toxicological experiments. Here, we used the CTD to explore what toxicants and disease outcomes are most commonly associated with a curated list of piRNA-related genes, prioritizing a set of exposures, molecular mechanisms, and disease outcomes for further analysis. This work highlights the merit

in using CTD to assess what associations are well established, which require additional evidence, and where gaps in knowledge remain.

Methods

Review piRNA Associated Gene List

A comprehensive review of literature was performed using PubMed to identify genes with known associations with piRNA biogenesis. Key phrases including, but not limited to, “piRNA biogenesis”, “piRNA machinery”, “PIWIL machinery”, were used to select our key list of genes. This list of genes and their functions can be found in **Table 2.1**. Diseases highlighted in the literature search included breast cancer, numerous neurological diseases, and cardiovascular health.

To generate the spreadsheets used in the analysis, the piRNA-associated gene list was inputted into the CTD “Batch Query” tool on May 16th, 2024. After inputting our gene list, we selected the data to download in CSV format including chemical-gene interactions, chemical associations, and disease associations. The resulting spreadsheets contained data entailing detailed gene-chemical-disease interactions collected in the published literature. The spreadsheets produced by CTD include CTD_Genes_Chemicals spreadsheet (**Table 2.2**) and the CTD_Genes_Diseases spreadsheet (**Table 2.3**).

The CTD_Genes_Chemicals spreadsheet includes columns entitled: “Input”, “ChemicalName”, “ChemicalID”, “CasRN”, “GeneSymbol”, “GeneID”, “Organisms”, “OrganismsID”, “Interaction”, “InteractionActions”, and “PubMedIDs”. The Input column refers to the list of genes that we input into CTD. ChemicalName column provides the known chemical name for each chemical. The GeneSymbol column provides the gene

symbol for genes. The CasRN column provides the CAS registry number, a unique identifier to distinguish chemical substances or molecular structures when there are many other possible names (CAS Registry, American Chemical Society). The Interaction column summarizes the interaction between the chemical and the gene. The InteractionActions column summarizes the effect (decrease, increase or affects) on the chemical modification (expression or methylation). Finally, the PubMedIDs include the PubMed IDs for the citation used to generate the data given that specific row in the table.

The CTD_Genes_Diseases spreadsheet generated from CTD was Genes_Diseases (Table 2.3). The Genes_Diseases spreadsheet includes “Input”, “DiseaseNames”, “DiseaseID”, “GeneSymbols”, “GeneID”, “DiseaseCategories”, “DirectEvidence”, “InferenceChemicalName”, “InferenceScore”, “OmimIDs”, and “PubMedIDs”. The Input column refers to the list of genes that we input into CTD. The DiseaseNames column provides names of diseases in CTD. The DiseaseID column provides unique identifiers assigned to the disease by the U.S. National Library of Medicine’s Medical Subject Headings (MeSH). GeneSymbols column provided the gene symbol for genes. The GeneID column provides the gene ID for the NCBI gene database. DiseaseCategory column provides the CTD’s “Merged Disease Vocabulary” (MEDIC) which is a modified subset of descriptors from the “Diseases” branch of MeSH. The DirectEvidence column indicates if there is direct evidence provided for the gene-disease association by either a marker/mechanism (a gene shown to be a biomarker of the disease or that plays a known role in the etiology of the disease) or a therapeutic (a gene this is or may be a therapeutic target in the treatment of the disease). The

InferenceChemicalName column provides the name of a chemical that CTD has inferred a relationship with between the disease and the chemical. These inferred relationships are established by CTD-curated chemical-gene and gene-disease interactions, where a gene is associated with a disease (inferred relationship) because that gene has a curated interaction with a chemical, and that chemical has a curated association with a disease. The InferenceScore column provides the score of the inference based on CTD's chemical-gene-disease networks. The inference score considers all interactions associated within each network and results in either a high number, indicating there is a high degree of similarity between the network and similar scale free network, or a low number, indicating the opposite. For example, for a single chemical, one chemical-disease relationship has a score of 15 and another relationship with the same chemical and different disease might have a score of 2. The chemical-disease relationship with the higher score could reflect that more genes may be involved in the network, or there may be less "hub genes" which are genes known to interact with a lot of chemicals and a lot of diseases. Therefore, the higher inference score of 15 signifies the inference network is less likely to be due to a random network (https://ctdbase.org/help/diseaseGeneDetailHelp.jsp). The OmimIDs are unique identifiers assigned to the disease by the Online Mendelian Inheritance in Man (OMIM). Finally, the PubMedIDs columns include the PubMed IDs for the citation used to generate the data given that specific row in the table.

Chemical Prioritization

A flow chart of the methods can be found in **figure 2.1**. Prioritization of the top 50 chemicals associated with piRNA was conducted by using the number of publications

with a direct association with our piRNA genes. Using the CTD_Genes_Chemicals spread sheet, we selected for only human, mouse, and rat organisms. Next, we filtered the chemicals to only include individual chemicals that are considered environmental exposures. Large mixtures of chemicals, such as JP8 aviation fuel, were removed in order to investigate independent chemicals. Additionally, pharmaceuticals, such as Cyclosporine, were also removed due to the exposure of the compound only being found in those who take the medication. As exceptions, acetaminophen, Valproic acid, and Jinfukang were kept due to their accessibility and common use. Due to the number of chemicals with the same CasRN and different ChemicalName, chemical names were then collapsed to create consistent names. For example, chemicals “cadmium” and “cadmium chloride” were given the same name “cadmium”. The number of publications for each chemical/gene interaction were then tallied to create the list of the top 50 chemicals. The top chemicals can be found in **table 2.4** and a description of each chemical and can be found in **supplementary table 2.1**.

Chemical X Gene Expression and Methylation

The “Interaction” and “InteractionActions” columns from the environmental chemicals prioritized in methods section 1.2 were used to look at the effect of the chemicals on piRNA-related gene expression and DNA methylation. These columns describe the interacting chemical for a gene and then a brief description of the interaction. Expression and DNA methylation InteractionActions were separated into new data frames. A new column was created to add action values where “increases expression/methylation” = +1, “decreases expression/methylation” = -1, and “no affects” or any N/A’s = 0. Affects expression is used when the reference does not describe a

specific directionality for the interaction. Expression data were also separated into mRNA and protein expression. Action values for expression and DNA methylation were summed across chemical name and gene symbol and used to create an average. Therefore, if a gene x chemical interaction has 3 publications resulting in increases expression, 1 publication for affects expression, and 1 publication for decreases expression, then the action value would be ~ 0.4. A matrix was formatted out of the produced data then the package “pheatmap” was used to make figure 2.3.

Disease Association

To investigate the overlay between diseases linked with the top 50 chemicals and our list of genes, the CTD_Genes_Diseases spread sheet was used. The spread sheet provided gene disease relationships and inferred chemicals that were found in the gene’s network. After collapsing the inference chemical name (as done above in chemical prioritization), the top 50 chemicals were used to identify the top diseases from the “CTD_Genes_Diseases” spreadsheet. The number of publications for each disease/gene/inference chemical were then tallied to create the top 50 diseases.

Results

Chemical Prioritization

To visualize the data provided by CTD_Genes_Chemicals, we created a waterfall plot of the top 50 chemicals with the number of reported associated with the piRNA-related genes, sorted by the number of publications (**Figure 2.2a**). Here we see that bisphenol A contains the highest number of reported associations with our piRNA-related genes (66). Valproic acid, benzo(a)pyrene, and Dioxins follow bisphenol A with 49, 48, and 40 publications, respectively. We see a drop in the number of publications

beginning with Arsenic (24), followed by aflatoxin B1 (17), acetaminophen (12), and bisphenol S (12). The remaining chemicals were identified to have interactions with piRNA-related genes in fewer than 10 publications. We then wanted to visualize the number of publications in the top 50 chemicals for each gene (**2.2b**). SND1 and PRMT5 have the most publications out of all the genes with 113 and 97, respectively. The rest of the following genes were reported in fewer than 60 publications. The gene with the least number of publications is PIWIL3 with 5 publications.

Chemical x Gene Interactions

Next, we aimed to determine the effects of the top 50 chemicals on our piRNA-related genes (**Figure 2.3)**. Here, the scale represents the range of averages of the publications for each gene and chemical interaction. Therefore, dark red indicates all publications concur that the gene is increasing in expression, orange indicates some publications saw an increase in expression while others observed either a decrease in expression or did not specify, yellow indicates the gene expression was either unspecified in the publication or publications saw conflicting directionality, light blue indicates most the genes decreased in some publications and were either unspecified or contradictory in others, and dark blue indicates all publications concur that the gene is decreasing in expression.

Unsurprisingly, the chemicals and genes with the most publications, as shown in figure 2.2, are those that had the majority of the reported expression data including the chemicals bisphenol A and valproic acid as well as the genes PRMT5 and SND1. To find the total number of genes that report expression data for our top chemicals, we counted all dark red interactions as increased, all dark blue as decreased, and all

orange and light blue as contradictory evidence. In total, figure 2.3a shows that piRNA-related gene expression (mRNA) is shown to increase in 99 instances, decreases in 101 instances and have unspecified effects or contradictory evidence in 23 instances. The five most published chemicals from figure 2.2b include: bisphenol A, valproic acid, benzo(a)pyrene, dioxins, and arsenic. Figure 2.3a shows that most of the piRNA-related genes, bisphenol A exposure either decreases expression or has an unspecified effect. Two genes, TDRD9 and SND1 are a dark orange, indicating that some publications did see an increase in expression for these genes related to bisphenol A exposure. Valproic acid, on the other hand, shows that most of the genes increased in expression relative to chemical exposure. PLD6, TDRD6, and TDRD12 are all dark red, indicating the evidence behind the increase in expression is strong. Arsenic had 6 genes with strong evidence of decreased expression, 4 genes with strong evidence of increased expression, and 2 genes with unspecified effects. Dioxins show strong evidence for increased expression of PIWIL2, FKBP6, and TDRD1. Additionally, dioxins show strong evidence of decreased expression for PIWIL4, GPAT2, and TDRD9. However, 8 of the genes that interacted with dioxins show a combo of increased expression and unspecified effects and TDRKH shows a combination of decreased expression and unspecified effects. SND1 had altered expression data for 23 of the chemicals. Eight of those chemicals had strong evidence for increasing expression, 9 of those chemicals showed strong evidence in decreasing expression, and 6 of those chemicals had unspecified data. PLD6 had altered expression data for 21 of the chemicals. 9 had strong evidence for increasing expression, 8 had strong evidence of decreasing in

expression and 4 had unspecified data. PRMT5, SPOCD1 and PIWIL4 followed to make up the top 5 genes with mRNA expression data in our list.

Figure 2.3b shows that piRNA-related protein expression increases in 8 instances, decreases in 6 instances and have unspecified effects in 1 instance. SND1, PRMT5, TDRKH, MAEL, and TDRD6 were the only genes with protein data. PRMT5 showed strong evidence of increased expression for 5 chemicals and strong evidence of decreased expression for one chemical, bisphenol A. SND1 shows strong evidence for increased expression in 3 chemicals and strong evidence of decreased expression in 3 chemicals. Chemicals bisphenol F, bisphenol AF, and thapsigargin clustered together while aflatoxin B1, benzo(a)pyrene, and endosulfan clustered together.

In addition to expression, CTD interactions include DNA methylation data. Figure 2.3c shows the effects of the top 50 chemicals on piRNA-related gene DNA methylation. Strikingly, valproic acid shows strong evidence for increase in methylation for 12 genes, with no evidence of unspecified or decreased methylation. Aflatoxin B1 also shows strong evidence of increase in methylation for 7 genes also shows a decrease in methylation for PRMT5 and PIWIL1. Interestingly, benzo(a)pyrene indicates a range of increased methylation data with 3 genes, PLD6, SPOCD1, and RNF17, showing strong evidence and 11 genes showing a range of increased and unspecified methylation. Bisphenol A shows the widest variety of methylation effects with 5 genes having strong evidence of increased methylation, 3 genes with unspecified, and 4 genes showing strong evidence of a decrease in methylation. In addition, dioxins show 9 genes with unspecified effects.

Chemical x Gene x Disease: Inferenced Associations

We then aimed to understand what relationships were built between piRNA-related genes, chemicals, and disease outcomes. **Figure 2.4** shows the associations between piRNA-related genes and disease by inference from the top 50 chemicals (as shown in figure 2.2). The piRNA-related gene list is on the y-axis and the top 50 disease names across the x-axis. Only gene-disease relationship data were used for inference for the chemicals that matched our top 50 environmental chemical list; therefore, the inference score represents the chemical X gene X disease network created in CTD for our specified environmental chemical and piRNA-related gene list. As expected, the genes with the most data from our list, PRMT5 and SND1, show the most associations with the top disease and the top inferenced chemicals. In particular, PRMT5 and SND1 both indicate very high inference scores for diseases, including weight loss, necrosis, and chemical and drug induced liver injury. The following top disease names with the highest associations with these genes include prenatal exposure delayed effects, hyperplasia, inflammation, hepatomegaly, and kidney disease. PLD6 and SPOCD1 also show some of the highest inference scores among the inference chemical and disease names associations. Specifically, SPOCD1, along with SND1 and PRMT5, see higher inference scored associations with birth weight, cognition disorders, and cell transformation, neoplastic disease names. Additionally, we observe that the gene with the least amount of association with our top diseases and top inference chemicals is PIWIL3.

Discussion

CTD is an accessible research tool which provides manually curated information from published scientific literature about chemical/disease/gene relationships. CTD

integrates chemical, gene, phenotype, anatomy, disease, taxa, and exposure subject matter from published data to generate inferences to assist and guide environmental health research. CTD reports the impressive inclusion of 17,100 chemicals, 54,300 genes, and 7,270 diseases, demonstrating the wide range of the repository (Davis et al. 2023). Researchers can use this database to detect knowledge gaps and develop unique hypothesis through CTD's inference network. In this paper we aimed to explore what toxicants and disease outcomes are most commonly associated with our curated list of piRNA-related genes and provide an in-depth review on the usability of this database.

Figure 2.2a allowed us to visualize the number of publications available in CTD's repository showing interactions between our top 50 chemicals and piRNA-related gene list. This provided an overall look into the chemicals most frequently investigated when looking at these genes. Additionally, this figure indicates those chemicals which indeed have an interaction with our genes and require further investigation. We observe that many endocrine disrupting chemicals including bisphenols, dioxins, and phthalates, are associated with piRNA-related genes. Flame retardants, metals such as arsenic, lead copper, and cadmium and pesticides including chlorpyrifos and parathion are also among the top 50 chemicals. **Supplementary table 2.1** provides short descriptions of each of the chemicals and links to their PubChem pages. After understanding the amount of chemical data associated with our gene list in CTD, we wanted to visualize the number of publications which report associations between chemical exposures and piRNA-related genes to determine what genes were most frequently observed in publications included by CTD. **Figure 2b** shows a breakdown of the number of

publications which report associations with the top 50 chemical exposures for each of our genes. SND1 and PRMT5 have the most publications, while PIWIL3 has the fewest number of publications. This was very surprising at first, we were expecting PIWILs to have much more data due to their overwhelming presence in the piRNA literature, however, because this data is focused on interactions between these genes and chemicals, it may show that the current literature on environmental exposures and piRNA-related genes is lacking. Alternatively, the lack of data in our results for PIWIL3 could potentially indicate that PIWIL3 might not change with chemical exposure, however we cannot necessarily tell the difference between lacking data and non-affected data here. It is notable that with increased research in 'omics, we would expect to see more data on these genes, however, due to the manual input of data into CTD, the full amount of research may not be able to be investigated quite yet. Taken together, figures 2a and 2b create a visual representation of the state of research for piRNA-related genes and environmental chemicals accessible in CTD.

Next, we investigated the published interactions between the chemicals and genes provided by CTD. The "Interactions" and "InteractionActions" columns assigned to each chemical and gene pair described the observed interaction from the publication or publications the data was curated from. **Figure 2.3** examined the mRNA expression differences (**figure 2.3a**) protein expression differences (**figure 2.3b**) and methylation differences (**figure 2.3c**) after chemical exposure. These heat maps are another visualization of our data; however, it is also difficult here to see how much data is missing vs data that have no relationships. Interestingly, some of the interactions were supported by additional publications while other interactions were contradicted by

additional evidence. For example, bisphenol A is observed to increase expression of PRMT5 mRNA in one publication, however, a second publication observed a decrease in PRMT5 mRNA. In order to account for both observations, we chose to create a scale of evidence strength by using action values, therefore, the stronger the color the more supporting evidence. A large portion of the data is reported as “affects expression”, which indicates that the reference does not specify a more specific degree. This may be due to the gene being sequenced, however was not specifically reported on in the publication. Therefore, the ”affects” data could be investigated further to determine direction in a given organism, tissue, developmental stage, or disease state. Certain chemicals maintained nearly consistent patterns, such as valproic acid, which had either strong evidence or contradictory evidence favoring increasing expression for 7 out of the 12 genes data was present for. Another chemical was a nearly consistent pattern is bisphenol A, where 12 out of 14 of the chemicals showed either had strong evidence or favored decreased expression. These patterns may represent important mechanisms of action to further investigate in disease outcomes. Using figure 2.3a and 2.3b together, we can observe separate effects on protein vs mRNA expression. Figure 2.3b indicates bisphenol A results in a decrease in protein expression of PRMT5, however, figure 2.3a shows non-specified/no results for mRNA expression. When looking into the data further, there are three publications contributing to this result, one which shows an increase in PRMT5 mRNA expression, one which shows a decrease, and one which indicates “affects expression”. These results were then averaged, and the final observation was that there was contradictory evidence of directionality. This contradictory evidence indicates further investigation into this research is required to

determine what might cause conflicting evidence, or to determine stronger evidence in either direction. Because the publications are linked to each interaction, we were able to investigate what may have caused this contradictory result for BPA and PRMT5. The publication which reported “no affect” (Kim et al. 2019) performed an RNA seq, however no data on PRMT5 was reported in the paper. The publication which reported “Increased expression” (Ali et al. 2014) investigated reproductive toxicity. Finally, the publications which reported “decreased expression” (Thongknor et al. 2019 & Sukjamnong) investigated neuronal activity. Therefore, the contradictory evidence we may observe could be due to how the chemical affects different tissues/systems.

Further, we examined the DNA methylation data within our CTD gene chemical interactions in figure 2.3c. Similarly to the expression data, patterns arise in the DNA methylation data with chemicals including valproic acid, aflatoxin B1, and benzo(a)pyrene. All three of these chemicals report strong evidence showing increased methylation of their interacting piRNA-related genes. Figure 2.3c continues to show the missing and/or non-specific data for chemical-gene methylation and therefore emphasizes the need to investigate these chemicals further.

Finally, we were able to use the inference networks developed in CTD to determine if our list of piRNA-related genes act as intermediates between chemical and disease relationships. As introduced in the methods, CTD inference scores reflect the similarity between chemical-gene-disease networks. Inference networks are derived from two conditions, the direct relationship between gene and disease, and the direct relationship between gene and chemical. With these two conditions meet, CTD can then “infer” the chemical to disease relationship through the shared gene intermediate.

Figure 2.4, uses direct relationships between genes (G) and chemicals (C), prioritized and investigated in figures 2.2 and 2.3, and the direct relationship between genes (G) and diseases (D), obtained in our Genes_Diseases spreadsheet, to show the inferred relationship between our top 50 environmental chemicals and diseases via our piRNA-related gene intermediates. Here, we can determine that PRMT5 and SND1 have the highest inferenced chemical scores for diseases including weight loss, and chemical and drug induced liver injury. **Table 2.3**, Genes_Diseases, reports the disease names, genes, inferenced chemicals, and inference score, which allows us to find the exact chemicals responsible for these high inference scores. PRMT5 is directly associated with chemical and drug induced liver injury, as well as directly associated with bisphenol A resulting in an inference score of 170.51. A high score such as this is interpreted as the more likely this inference network has atypical connectivity, or it is not due to random network. From this interpretation, we can conclude that there is a high chance that bisphenol A exposure will result in chemical and drug induced liver injury, such as digestive system disease, through PRMT5 gene interactions. Upon investigation to validate this score, we found numerous publications indicating bisphenol A damages the intestinal barrier function in the gastrointestinal tract and promotes inflammatory processes in the stomach and intestine (Ambreen et al. 2019, Zhao et al. 2019).

Although it is apparent that the gene-chemicals and gene-disease relationships with the most publications are those that have the strongest inference networks, this encourages us to begin to look at those relationships with less evidence. For example, breast neoplasms show relatively higher inference scores for PRMT5, SND1, TDRD6,

and SPOCD1. This observation inspires us to investigate the gene-chemical-disease inference score to further understand how the interactions may result in that outcome.

In this study, our results show an effective way to use multiple facets from CTD to evaluate the current state of research for environmental exposures and piRNA-related genes and prioritize future studies to investigate the effects of the environment on piRNA. This study also integrated the utility of CTD in conducting focused hypothesis driven research. Although CTD proved to be helpful, there are some limitations that need to be addressed. First of all, when aiming to prioritize the chemicals in figure 2.2, we ran into issues with the consistency of a few chemical names, specifically metals and dioxins. For example, chemical names for metals, such as cadmium, were recorded as “cadmium” however, upon further investigation into the publications used to create the interaction, the publications used cadmium chloride. In order to obtain the most out of the data as possible, we collapsed the chemicals into an overall chemical name, “cadmium”. Other metals including arsenic, copper, and lead were also collapsed due to similar situations. Similarly, dioxins were collapsed into the same grouping to prevent biased data. We included this as a limitation due to not knowing the full extent of what chemical names are included as individuals and which are reported correctly in the supporting literature. Additionally, as we completed each step of the methods, we would do random spot checks to make sure everything was working as planned. There were several instances where the gene and disease interaction reported seemed to have no relevance in the reference used to support the interaction. This is a limitation due to producing doubt in the relationship reported. We assume this may be due to some sequencing that took place outside of the paper, or possibly a reference within the

publication that saw the interaction, however, to do a deep dive into where that data may have come from would take a tremendous amount of time. Therefore, trust in the accuracy of the curation of these relationships is required to feel confident in the conclusions made.

Finally, our last limitation interrogates the use of the term “affects” in the gene-chemical interaction data. CTD qualifies interactions using the degrees “increases”, “decreases”, “affects”, and “does not affect”. The “affects” degree is used when the reference does not describe a more specific degree (CTD- Advanced Chemical-Gene Interaction Query, 2024). Additionally, any interaction having the “does not affect” degree is excluded from the public data. Figure 2.3 indicates expression and methylation data for the gene-chemical interactions; however, a significant amount of the results indicates the gene-chemical interaction “affects” or has unspecific degree. Further investigation into these publications which report “affects” shows us that the gene of interest may be grouped into a large list of genes, where it is not individually reported, or may be highlighted in an enrichment pathway, or may even just be acknowledged in the supplementary tables. For small scale investigations, it would be easier to dive into these papers to find the effect of a single gene, however with larger scale investigations, looking into each publication for the exact interaction, if any are given at all, would be very time consuming. Moreover, excluding interactions given “does not affect” degree adds another level of complexity for investigations such as this one because one cannot tell if an interaction has been studied and has had no effect, or whether the interaction hasn’t been studied at all. This can complicate hypothesis

generating research due to it its misleading representation of completed investigations vs those that have yet to be researched.

Conclusion

In conclusion, our analysis demonstrates that our piRNA-related gene list resulted in the prioritization of the top 50 environmental chemicals and top 50 diseases to further investigate the role of piRNA-related effects. Our results also show that the field of toxicogenomics still has many unexplored avenues, however, studies such as this one can help prioritize hypothesis on a larger scale. Future studies can incorporate more aspects from CTD including pathway and gene ontology categories. CTD is a powerful and effective database that provides accessible gene-chemical, gene-disease, and inference networks to advance how environmental exposures affect human health. The database will continue to progress developing and furthering its innovative resource promoting researchers in generating testable hypothesis about environmental health.

References

Ali, S., Steinmetz, G., Montillet, G., Perrard, M.-H., Loundou, A., Durand, P., Guichaoua, M.-R., & Prat, O. (2014). Exposure to Low-Dose Bisphenol A Impairs Meiosis in the Rat Seminiferous Tubule Culture Model: A Physiotoxicogenomic Approach. *PLoS ONE*, *9*(9), e106245. https://doi.org/10.1371/journal.pone.0106245

Ambreen, S., Akhtar, T., Hameed, N., Ashfaq, I., & Sheikh, N. (2019). In Vivo Evaluation of Histopathological Alterations and Trace Metals Estimation of the Small Intestine in Bisphenol A-Intoxicated Rats. *Canadian Journal of Gastroenterology & Hepatology*, *2019*, 9292316. https://doi.org/10.1155/2019/9292316

Aravin, A. A., & Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. *Genes & Development*, *22*(8), 970–975. https://doi.org/10.1101/gad.1669408

Cantone, I., & Fisher, A. G. (2013). Epigenetic programming and reprogramming during development. *Nature Structural & Molecular Biology*, *20*(3), 282–289. https://doi.org/10.1038/nsmb.2489

Chalbatani, G. M., Dana, H., Memari, F., Gharagozlou, E., Ashjaei, S., Kheirandish, P., Marmari, V., Mahmoudzadeh, H., Mozayani, F., Maleki, A. R., Sadeghian, E., Nia, E. Z., Miri, S. R., Nia, N. zainali, Rezaeian, O., Eskandary, A., Razavi, N., Shirkhoda, M., & Rouzbahani, F. N. (2018). Biological function and molecular mechanism of piRNA in cancer. *Practical Laboratory Medicine*, *13*, e00113. https://doi.org/10.1016/j.plabm.2018.e00113

Davis, A. P., Wiegers, T. C., Johnson, R. J., Sciaky, D., Wiegers, J., & Mattingly, C. J. (2022). Comparative Toxicogenomics Database (CTD): Update 2023. *Nucleic Acids Research*, *51*(D1), D1257–D1262. https://doi.org/10.1093/nar/gkac833

Ding, X., Li, Y., Lü, J., Zhao, Q., Guo, Y., Lu, Z., Ma, W., Liu, P., Pestell, R. G., Liang, C., & Yu, Z. (2021). piRNA-823 Is Involved in Cancer Stem Cell Regulation Through Altering DNA Methylation in Association With Luminal Breast Cancer. *Frontiers in Cell and Developmental Biology*, *9*, 641052. https://doi.org/10.3389/fcell.2021.641052

Du, W. W., Yang, W., Xuan, J., Gupta, S., Krylov, S. N., Ma, X., Yang, Q., & Yang, B. B. (2016). Reciprocal regulation of miRNAs and piRNAs in embryonic development. *Cell Death & Differentiation*, *23*(9), 1458–1470. https://doi.org/10.1038/cdd.2016.27

Estrogenic Endocrine-Disrupting Chemicals: Molecular Mechanisms of Actions on Putative Human Diseases: Journal of Toxicology and Environmental Health, Part B: Vol 17 , No 3—Get Access. (n.d.). Retrieved June 2, 2024, from https://www.tandfonline.com/doi/full/10.1080/10937404.2014.882194

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development (Cambridge, England)*, *143*(22), 4101–4114. https://doi.org/10.1242/dev.132639

Huang, G., Hu, H., Xue, X., Shen, S., Gao, E., Guo, G., Shen, X., & Zhang, X. (2013). Altered expression of piRNAs and their relation with clinicopathologic features of breast cancer. *Clinical and Translational Oncology*, *15*(7), 563–568. https://doi.org/10.1007/s12094-012-0966-0

Jacovetti, C., Bayazit, M. B., & Regazzi, R. (2021). Emerging Classes of Small Non-Coding RNAs With Potential Implications in Diabetes and Associated Metabolic Disorders. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.670719

Kim, B.-Y., Kim, M., Jeong, J. S., Jee, S.-H., Park, I.-H., Lee, B.-C., Chung, S.-K., Lim, K.-M., & Lee, Y.-S. (2019). Comprehensive analysis of transcriptomic changes induced by low and high doses of bisphenol A in HepG2 spheroids *in vitro* and rat liver *in vivo*. *Environmental Research*, *173*, 124–134. https://doi.org/10.1016/j.envres.2019.03.035

Krishnan, P., Ghosh, S., Graham, K., Mackey, J. R., Kovalchuk, O., & Damaraju, S. (2016). Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. *Oncotarget*, *7*(25), 37944–37956. https://doi.org/10.18632/oncotarget.9272

Liu, Y., Zhang, J., Li, A., Zhang, Y., Li, Y., Yuan, X., He, Z., Liu, Z., & Tuo, S. (2019). Identification of PIWI-interacting RNA modules by weighted correlation network analysis. *Cluster Computing*, *22*(1), 707–717. https://doi.org/10.1007/s10586-017-1194-8

Maleki Dana, P., Mansournia, M. A., & Mirhashemi, S. M. (2020). PIWI-interacting RNAs: New biomarkers for diagnosis and treatment of breast cancer. *Cell & Bioscience*, *10*, 44. https://doi.org/10.1186/s13578-020-00403-5

Rayford, K. J., Cooley, A., Rumph, J. T., Arun, A., Rachakonda, G., Villalta, F., Lima, M. F., Pratap, S., Misra, S., & Nde, P. N. (2021). piRNAs as Modulators of Disease

Pathogenesis. *International Journal of Molecular Sciences*, *22*(5), 2373. https://doi.org/10.3390/ijms22052373

Reik, W., Dean, W., & Walter, J. (2001). Epigenetic Reprogramming in Mammalian Development. *Science*, *293*(5532), 1089–1093. https://doi.org/10.1126/science.1063443

Sato, K., Takayama, K., & Inoue, S. (2023). Role of piRNA biogenesis and its neuronal function in the development of neurodegenerative diseases. *Frontiers in Aging Neuroscience*, *15*, 1157818. https://doi.org/10.3389/fnagi.2023.1157818

Sukjamnong, S., Thongkorn, S., Kanlayaprasit, S., Saeliw, T., Hussem, K., Warayanon, W., Hu, V. W., Tencomnao, T., & Sarachana, T. (2020). Prenatal exposure to bisphenol A alters the transcriptome-interactome profiles of genes associated with Alzheimer's disease in the offspring hippocampus. *Scientific Reports*, *10*(1), 9487. https://doi.org/10.1038/s41598-020-65229-0

Thongkorn, S., Kanlayaprasit, S., Jindatip, D., Tencomnao, T., Hu, V. W., & Sarachana, T. (2019). Sex Differences in the Effects of Prenatal Bisphenol A Exposure on Genes Associated with Autism Spectrum Disorder in the Hippocampus. *Scientific Reports*, *9*(1), 3038. https://doi.org/10.1038/s41598-019-39386-w

Tóth, K. F., Pezic, D., Stuwe, E., & Webster, A. (2016). The piRNA Pathway Guards the Germline Genome Against Transposable Elements. *Advances in Experimental Medicine and Biology*, *886*, 51–77. https://doi.org/10.1007/978-94-017-7417-8_4

Zhang, H., Ren, Y., Xu, H., Pang, D., Duan, C., & Liu, C. (2013). The expression of stem cell protein Piwil2 and piR-932 in breast cancer. *Surgical Oncology*, *22*(4), 217–223. https://doi.org/10.1016/j.suronc.2013.07.001

Zhang, L., Lu, Q., & Chang, C. (2020). Epigenetics in Health and Disease. In C. Chang & Q. Lu (Eds.), *Epigenetics in Allergy and Autoimmunity* (pp. 3–55). Springer. https://doi.org/10.1007/978-981-15-3449-2_1

Zhang, T., & Wong, G. (2022). Dysregulation of Human Somatic piRNA Expression in Parkinson's Disease Subtypes and Stages. *International Journal of Molecular Sciences*, *23*(5), 2469. https://doi.org/10.3390/ijms23052469

Zhao, Z., Qu, W., Wang, K., Chen, S., Zhang, L., Wu, D., & Chen, Z. (2019). Bisphenol A inhibits mucin 2 secretion in intestinal goblet cells through mitochondrial dysfunction and oxidative stress. *Biomedicine & Pharmacotherapy*, *111*, 901–908. https://doi.org/10.1016/j.biopha.2019.01.007

Figures and Tables

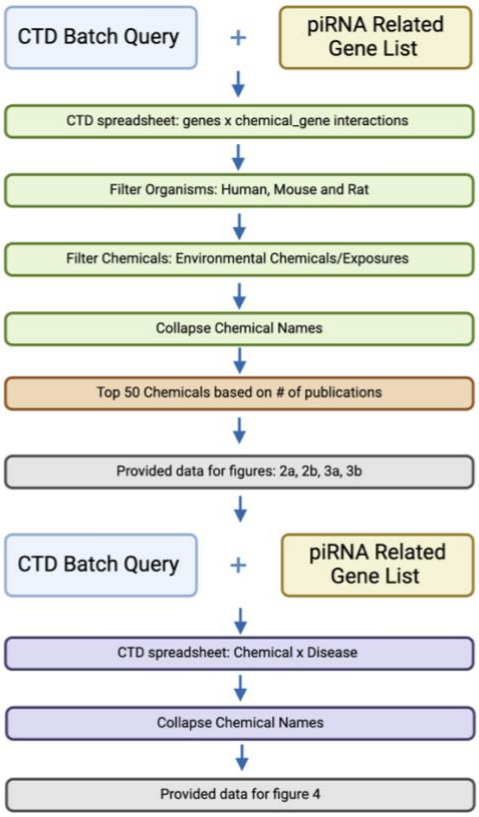


Figure 2.1: Flow chart of Aim 1 methods. After an in-depth literature search for piRNA-related genes, the list was input into the CTD batch query to acquire gene-chemical interactions. This data was then filtered to only include data from human, mouse, and rats. Next, only individual environmental chemicals were included while all other chemical groupings and pharmaceuticals were filtered out. The chemical names were then collapsed to reflect the appropriate names. Finally, a prioritized list of the top 50 chemicals was produced and the data from those chemicals were used in figures 2 and 3. The list of piRNA related genes were then re-input into CTD to acquire gene-disease associations. This data was then filtered to only show data for the prioritized chemicals from the previous step. The chemical names were again collapsed, and the data was used to create figure 4.

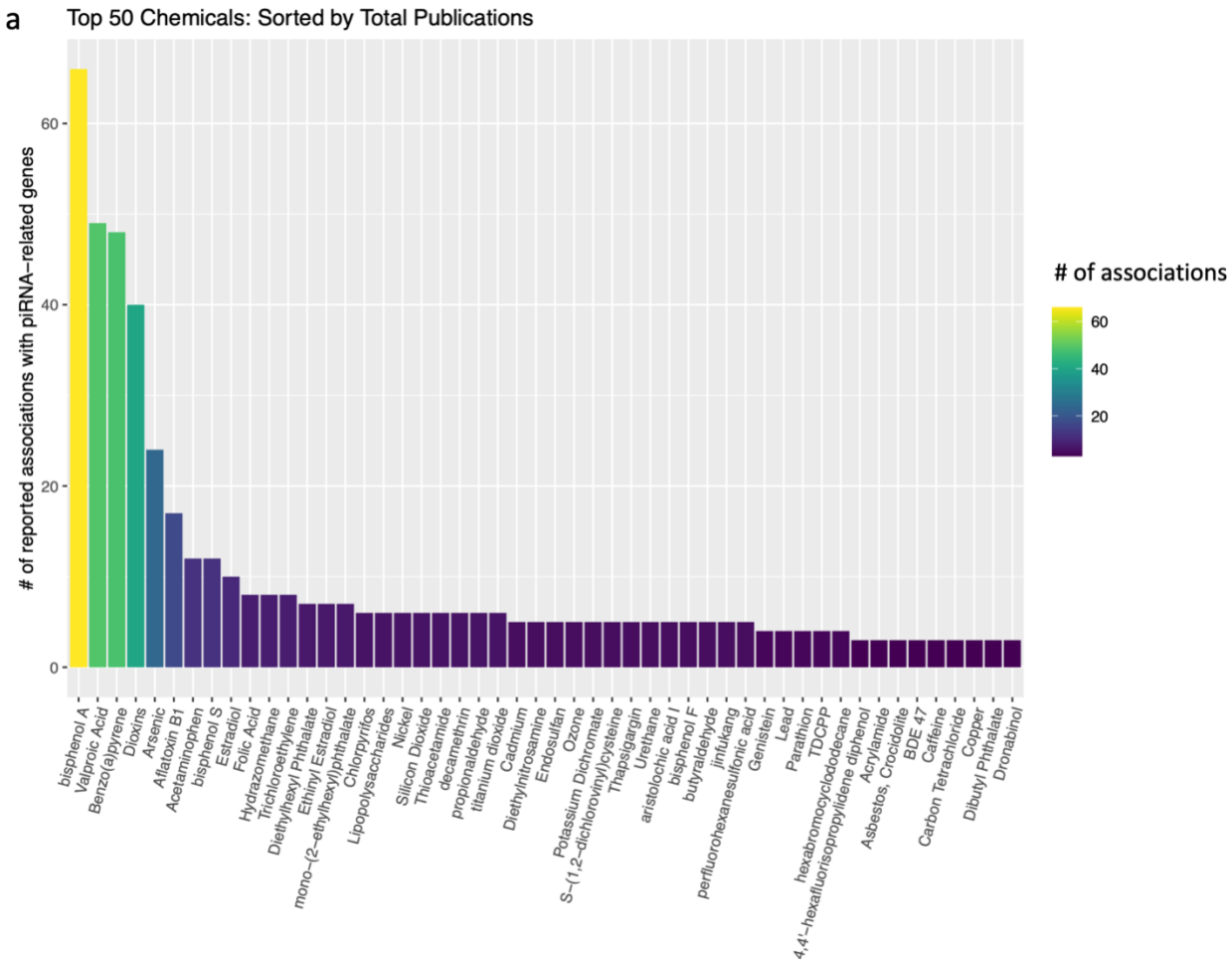


Figure 2.2a: Chemical Prioritization. Visualization of the top 50 environmental chemicals with the highest number of associations with piRNA-related genes.

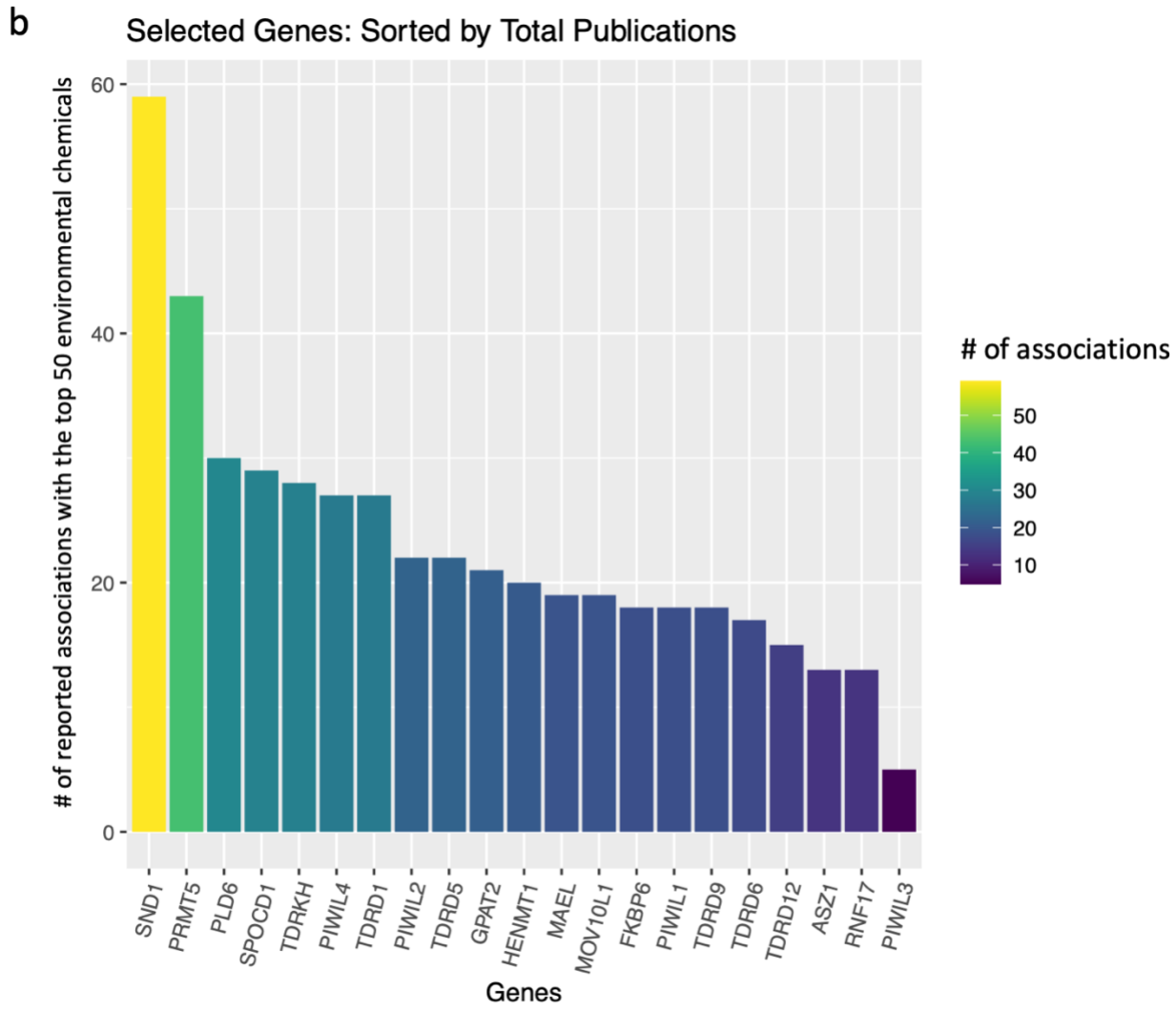


Figure 2.2b: Representation of piRNA-Related Genes. A visualization of the number of genes available in CTD associated with the top-50 chemicals.

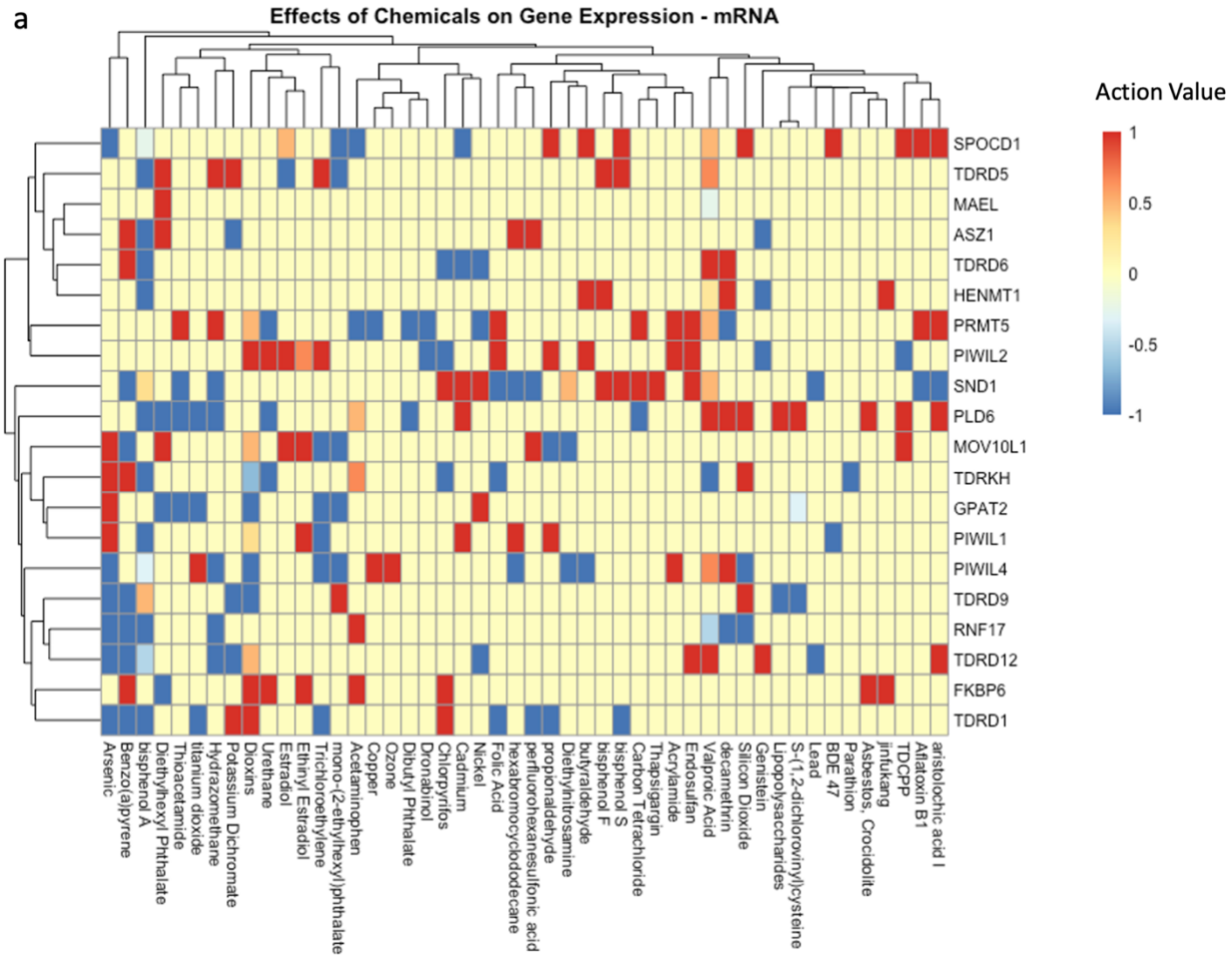


Figure 2.3a: Effects of Chemicals on Gene mRNA Expression. The legend shows the range of action values. Dark red indicates strong evidence for an increase in gene expression, orange indicates contradictory evidence favoring increase in gene expression, yellow indicates “no affects”/contradictory evidence/no data, light blue indicates contradictory evidence favoring decrease in gene expression, and dark blue indicates strong evidence for a decrease in gene expression.

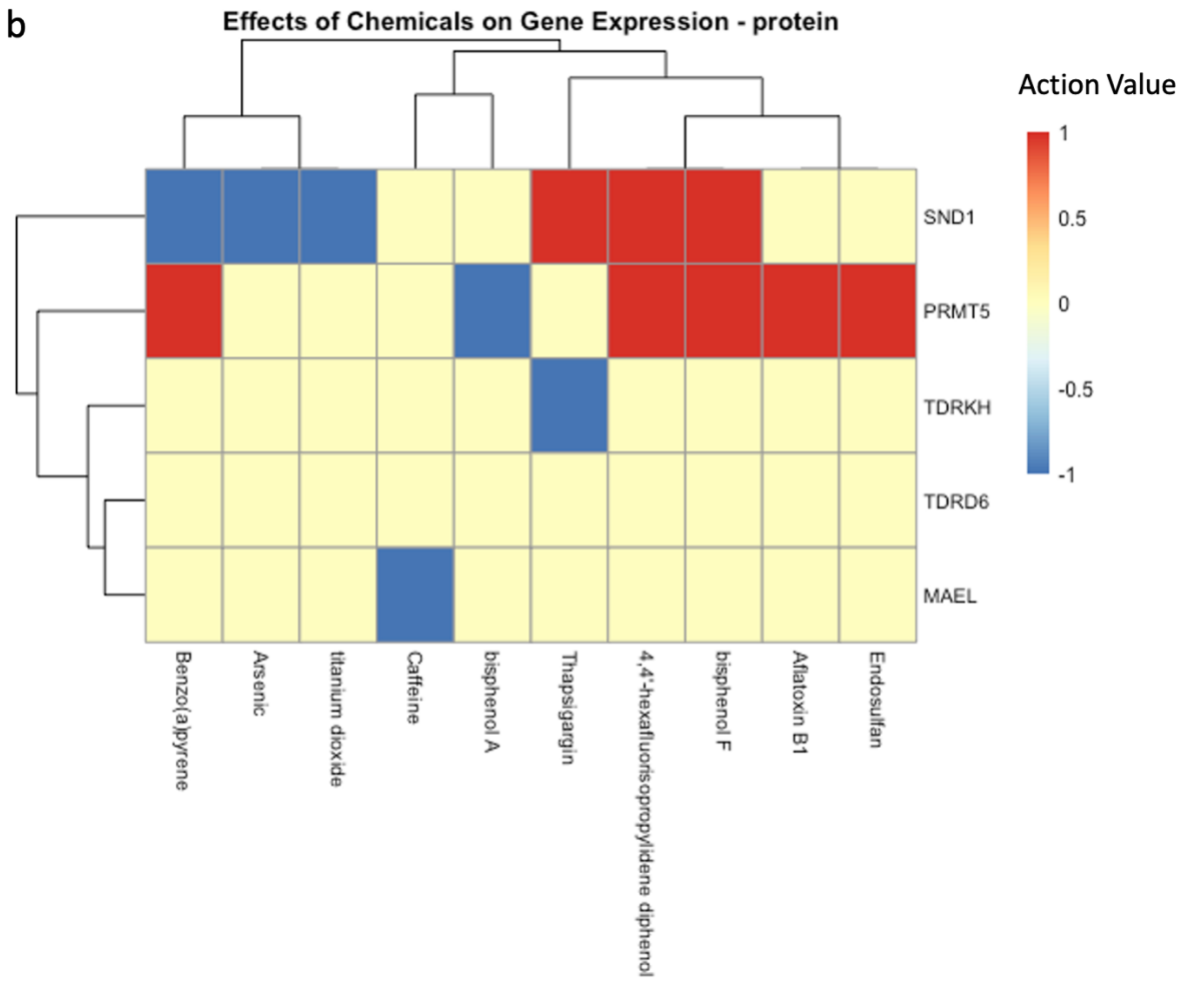


Figure 2.3b: Effects of Chemicals on Protein Expression. The legend shows the range of action values. Dark red indicates strong evidence for an increase in protein expression, orange indicates contradictory evidence favoring increase in protein expression, yellow indicates “no affects”/contradictory evidence/no data, light blue indicates contradictory evidence favoring decrease in protein expression, and dark blue indicates strong evidence for a decrease in protein expression.

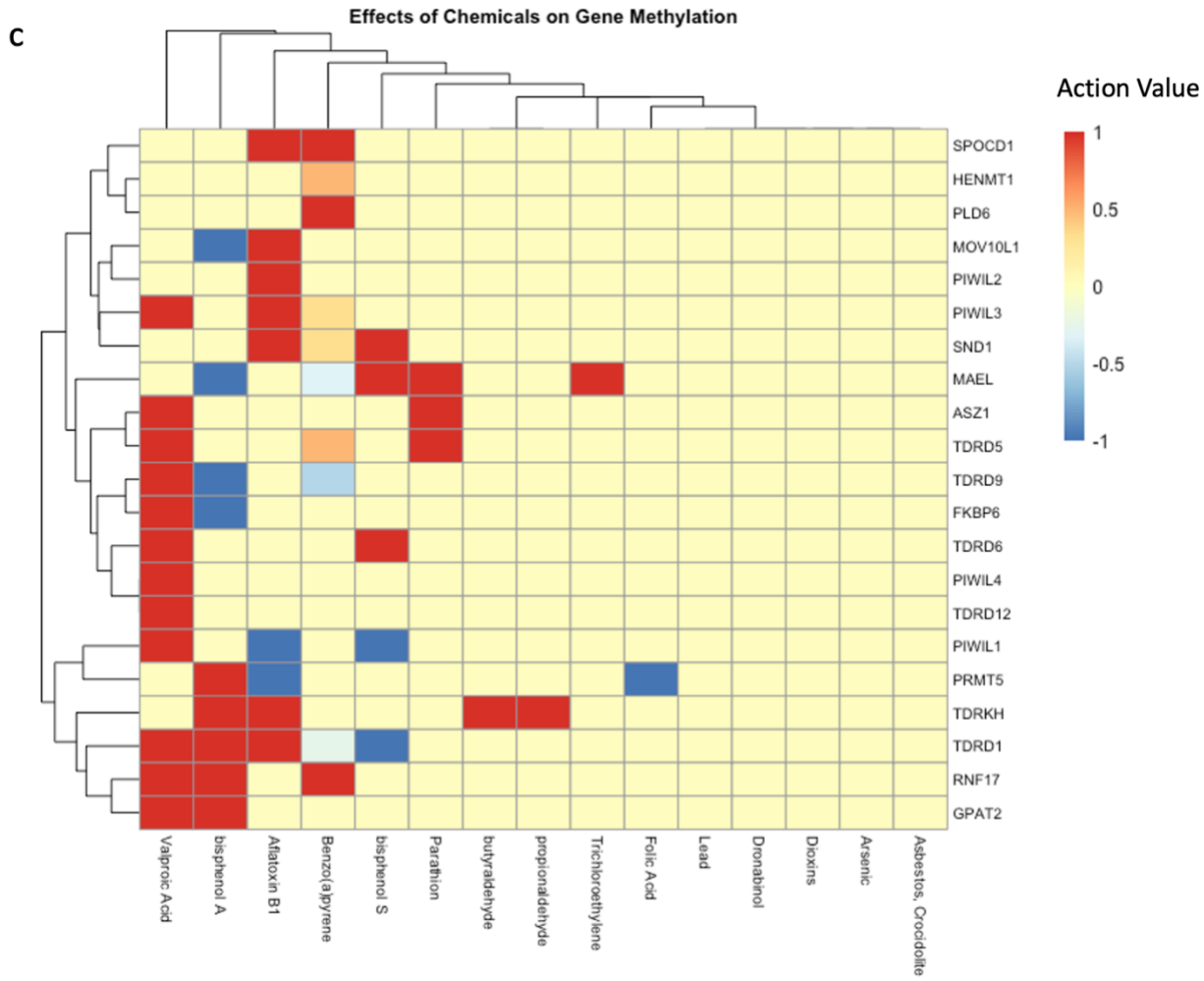


Figure 2.3c: Effects of chemicals on Gene DNA Methylation. The legend shows the range of action values. Dark red indicates strong evidence for an increase in gene methylation, orange indicates contradictory evidence favoring increase in gene methylation, yellow indicates “no affects”/contradictory evidence/no data, light blue indicates contradictory evidence favoring decrease in gene methylation, and dark blue indicates strong evidence for a decrease in gene methylation.

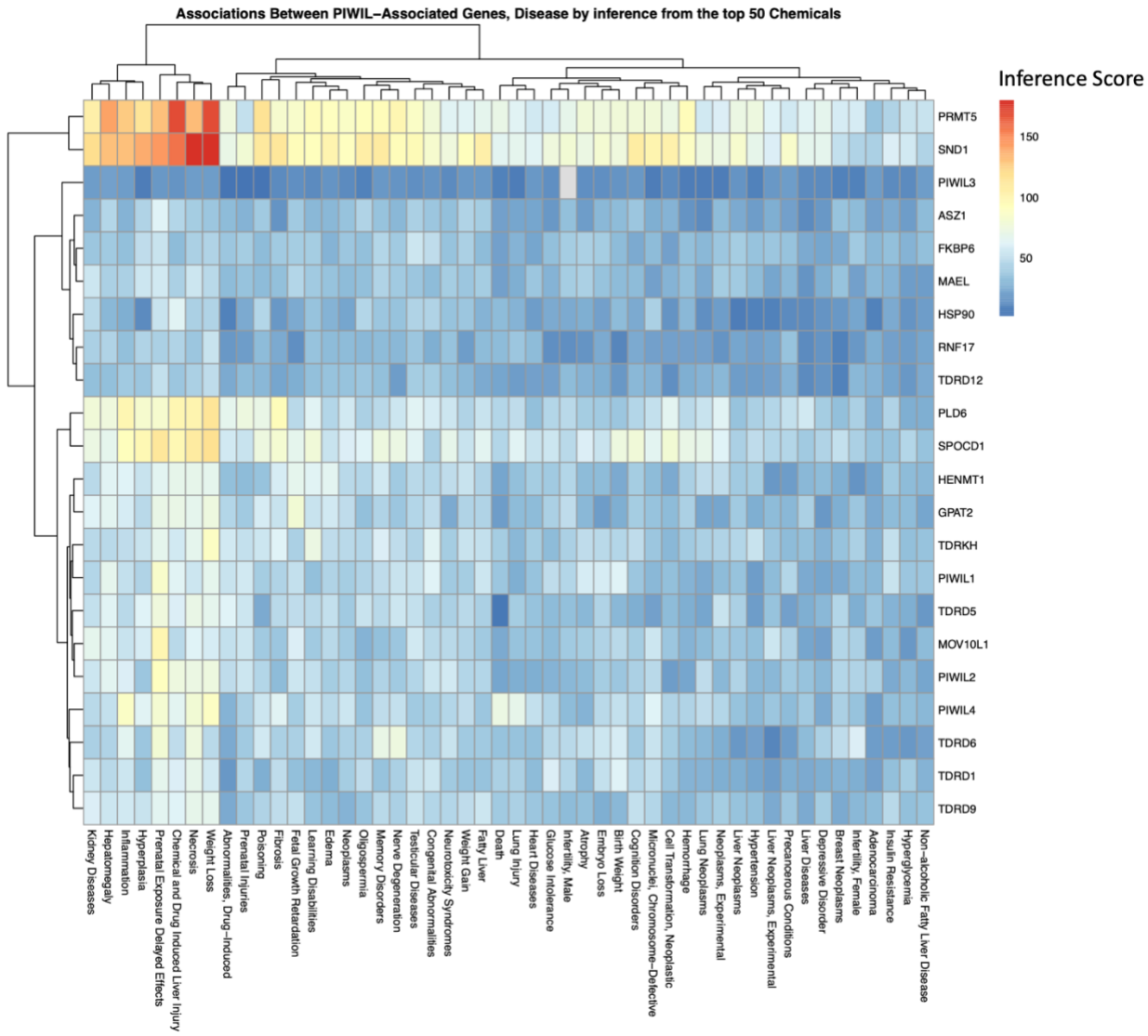


Figure 2.4: Associations Between piRNA-related Genes and Disease by Inference of the Top 50 Environmental Chemicals. Genes and diseases are shown to be associated through inferenced chemicals. Shown are genes on the y-axis and diseases on the x-axis. Here genes (G) have a direct relationship with diseases (D) and a direct relationship with chemicals (C), therefore CTD can then “infer” the C to D relationship via the G intermediate: C → G → D. The inferred chemical scores are used as the scale, red indicating a high score above 150, and blue indicating a lower score below 50.

Table 2.1: piRNA Biogenesis_Machinery. This table can be viewed by following the link.

Table 2.2: CTD_Genes_Chemicals. This table can be viewed by following the link.

Table 2.3: CTD_Genes_Diseases. This table can be viewed by following the link.

Table 2.4: Top 50 Chemicals. This table can be viewed by following the link.

Supplementary Table 2.1: CTD Environmental Chemical List. This table can be viewed by following this link.

Chapter 3

Aim 2: Identify Differentially Expressed piRNAs in Enriched Stem Cell-Like Mammospheres (3D) in Comparison to Monolayer (2D) Cell Culture.

Abstract

Breast cancer is a highly heterogeneous disease and among women is the most prevalent form of cancer worldwide. Epigenetics are thought to play an important role in breast cancer development and progression. Epigenetics is broadly defined as mitotically heritable changes in gene expression that do not alter the underlying DNA sequence. Epigenetic mechanisms play a key role in the maintenance of cell type specific gene expression and are an interface between the environment and the genome. Widespread epigenetic reprogramming is evident in carcinogenesis, progression, and metastasis. Additionally, epigenetic modifications may serve as mechanisms of toxicity and response to certain toxicants. P-element-induced wimpy testis (PIWI)-interacting RNA (piRNA) associate with PIWI proteins to regulate gene expression via DNA methylation. While piRNA were long thought to be exclusively expressed in the germline, work in the mouse has identified piRNA and PIWIL mRNA expression in the soma, and aberrant expression of piRNAs have been detected in human breast cancers. Several piRNAs have been investigated in breast carcinogenesis and metastasis. Preliminary data has implicated piRNAs as potential biomarkers for diagnosis and treatment for cancer. Several studies investigating the

specific function of a single piRNA such as promoting cancer cell proliferation, invasion, and migration. However, many of these studies have bypassed a vital step to investigating these piRNAs. Here, we determine baseline piRNA expression in non-tumorigenic MCF10A and cancerous MCF7 cell lines grown in monolayer and identify differences in piRNA expression by 2D and 3D states using mammospheres of MCF10As and MCF7s. We hypothesize that non-tumorigenic ER- MCF10A cells and cancerous ER+ MCF-7 cells will exhibit different piRNA expression profiles, and 2D and 3D states will also exhibit distinct piRNA expression profiles. The number of piRNA transcripts found in each cell line and morphology are as follows: MCF10A Monolayer (MCF10A-ML) has 3,039; MCF10A Mammospheres (MCF10A-MS) has 10,536; MCF7 Monolayer (MCF7-ML) has 14,225; and MCF7 Mammospheres (MCF7-MS) has 3,525. In conclusion, our results indicate that different cell lines, MCF10A and MCF7, and different cell culture conditions (monolayer vs mammospheres) exhibit different piRNA expression profiles.

Introduction

The breast is a dynamic organ which adapts and responds to the body's hormonal cues and growth factors throughout the lifecycle. The epithelial phenotypes of mammary tissue alter and shift during puberty, menstrual cycle, and pregnancy, fluctuating many times during reproductive periods (Ercan et al, 2011). To perform these dynamic shifts, the breast contains a large number of mammary stem cells (MaSCs). The unique characteristics of these MaSCs such as their ability to repeatedly respond, divide, and differentiate, leaves them vulnerable to acquiring mutations and therefore promoting tumorigenesis (Tharmapalan et al., 2019). Oncogenic activity can

occur at any point during the pathway from mammary stem cell to differentiated adult mammary cell, leading to transformation into breast cancer stem cells (BCSCs) (Jiagge et al. 2018). Additionally, adult mammary cells could undergo dedifferentiation into these more stem-like states (Hanahan, 2022). BCSCs have unlimited renewal capacity and play an important role in tumor heterogeneity, metastasis, recurrence, and chemoresistance (Dandawate et al. 2016). BCSCs are thought to be responsible for aggressive tumors and poor prognosis (Creighton et al. 2009).

There has been a growing emphasis on the role of non-coding RNAs in epigenetic regulation. One such class of small RNA - P-element-induced wimpy testis (PIWI)-interacting RNA (piRNA) - associates with PIWI proteins and binds to target regions where they regulate gene expression by directing DNA methylation machinery. While piRNA were long thought to be exclusively expressed in the germline, work in the mouse has identified piRNA and PIWIL mRNA expression in the soma, and aberrant expression of piRNAs have been detected in human cancers (Ding et al., 2021; Maleki Dana et al., 2020; Perera et al., 2019). Abnormal piRNA expression has been associated with both cancer progression and anti-cancer roles, making piRNA very attractive as targets for prognosis, biomarkers, and treatments. Many studies have shown that piRNA act as epigenetic regulators involved in carcinogenic processes including angiogenesis, invasiveness, growth, and metastasis of tumors (Liu et al. 2018; Holoch and Moazed 2015). Recently studies have implicated specific piRNAs in the metastasis and progression of breast cancer (Chalbatani et al. 2019; Ding et al. 2021; Huang et al. 2013; Zhang et al. 2013). Although these studies have implicated piRNA expression in breast cancer, none of the studies performed sodium periodate treatment

to validate the transcript as an actual piRNA. Most studies identify the piRNA transcript they are interested in using by referencing the piRNABank or piRNABase, however, these web resources on classified and clustered piRNAs have not been validated using a reliable method such as sodium periodate treatment. Sodium periodate treatment selects for small RNAs containing the 2'-O-methylation signature or piRNA (Ohara et al. 2007). A recent web resource generated by our team, piOxi Database, provides a comprehensive analysis of germline and somatic piRNAs identified by chemical oxidation (Wang et al. 2024) to assist broad research applications in the fields of RNA biology, cancer biology, environmental toxicology, and beyond.

Several studies have shown not only expression of piRNA in cancer but have determined their regulatory function and mechanism. One study observed that piR-021285 is involved in breast tumorigenesis by promoting invasiveness through DNA methylation (Fu et al. 2015). A recent review by Qian et al. 2021 lists piRNA transcripts involved in breast cancer including whether the piRNA is upregulated or downregulated, its regulatory function, and its mechanism (Qian et al. 2021). However, the mechanism behind the up regulation or down regulation of the piRNA involved in breast cancer is still undetermined.

The mammosphere assay was established by Dontu et al., 2003 in order to isolate, characterize and culture mammary stem cells. Sphere forming assays are a favorable approach to determine a breast cancer cell's potential to behave like stem cells. Each mammosphere contains about one sphere forming stem cell, thus indicating that the cells capable of sphere formation represent the mammary stem cells that can undergo limited self-renewal (Dontu et al., 2003). Researchers have observed

significant differences in cellular characteristics, gene expression, function when profiling mammospheres (3D culture) vs 2D cultures (Baldasici et al. 2024). Mammospheres exhibit properties of self-renewal, drug resistance and cell proliferation/survival. Additionally, studies have shown an upregulation of genes associated with stemness such as ALDH1, OCT4, SOX2 and NANOG (Rios-Fuller et al 2018). Cells in mammospheres often show increased invasive capabilities and cell-cell interactions indicating their potential for metastasis and interaction with the microenvironment (Bhat et al. 2019). Overall, mammospheres provide a more realistic model of in vivo tumor biology allowing researchers to better study cancer stem cells, drug resistance and mechanisms of metastasis.

The objective of this aim is to determine the baseline piRNA expression in MCF10A and MCF7 cell lines grown in monolayer and identify differences in piRNA expression by 2D and 3D states using mammospheres of MCF10A and MCF7 cells. Our hypothesis is that non-tumorigenic ER- MCF10A cells and cancerous ER+ MCF7 cells will result in different piRNA expression profiles. Additionally, different culture conditions, 2D (monolayer) and 3D (mammosphere), will result in different piRNA expression profiles.

Methods

Cell Culture Conditions

Non-tumorigenic ER- breast epithelial cells, MCF10As, and ER+ mammary gland epithelial cells derived from metastatic breast cancer, MCF7s, are used. The growth media for MCF10A cells include: DMEM/F12 (Thermo, Cat. # 11320-033), HEPES (1M) (Fisher, Cat. # 15630106), Horse Serum (Gibco, Cat. # 16050122), Insulin (4 mg/mL)

(Thermo, Cat. #12585014), Hydrocortisone (96 ug/mL) (StemCell, Cat # 07925), Cholera Toxin (Sigma-Aldrich, Cat. # C8052), Human recombinant epidermal growth factor (EGF) (StemCell, Cat. # 78006.1). The growth medium for MCF7 cells includes: Eagle's Minimum Essential Medium (EMEM) base media (ATCC, Cat. # 30-2003), Heat Inactivated-Fetal Bovine Serum (HI-FBS) (Sigma-Aldrich, Cat. # F4135), Penicillin/Streptomycin 100X (Pen/strep) (Thermo, Cat. # 15140122), Insulin (4 mg/mL) (Thermo, Cat. #12585014). Both MCF10A and MCF7 cells are incubated at 37°C at 5% CO_2. MCF10A cells have about a 24 hour doubling time and are split at 70-90% confluency (Bessette et al. 2015). MCF7 cells have about a 48 hour doubling time and are also split at 70-90% confluency.

Mammosphere Formation

Mammospheres are cultured in MammoCult media containing MammoCult Basal Medium (StemCell, Cat. # 05620), MammoCult Proliferation Supplement (StemCell, Cat. # 05620), Heparin solution (StemCell, Cat. # 07980), Hydrocortisone (96 ug/mL) (StemCell, Cat # 07925), and Penicillin/Streptomycin 100X (Pen/strep)(Thermo, Cat. # 15140122). MCF10A and MCF7 cells are grown up in a monolayer to ~70-80% confluence, cells are washed using PBS (Thermo, Cat. # 10010023) then trypsinized using phenol-red free TrypLE (Thermo, Cat. # 12604013). Once cells have detached, they are collected using their respective media and spun down at 200xg. The pelleted cells then have the remaining media removed before being resuspended in MammoCult. Cells are then counted and plated in Costar™ ultra-low attachment 6-well plates (Corning, Cat. # 07-200-601). Ultra-low attachment wells were seeded with ~300,000 cells/well. Mammospheres had 1 mL MammoCult added after 3 days, then

were collected for extraction after 5 days. Both MCF10A and MCF7 mammospheres are incubated at 37°C at 5% CO_2.

Monolayer and Mammosphere Cell Collection

Cells grown up in monolayer are collected after 2-3 days, or when cells have reached ~70-90% confluence. MCF10A and MCF7 monolayer cells are treated the same during collection. Cells are washed using PBS (Thermo, Cat. # 10010023) then trypsinized using phenol-red free TrypLE (Thermo, Cat. # 12604013). Once cells have detached, they are collected using their respective media and spun down at 200xg. The pelleted cells then have the remaining media removed and are resuspended in 400 uL Trizol (Invitrogen, #15596-026) then kept at -20°C until ready for isolation of RNA and smRNA.

For mammosphere collection, mammospheres from all wells from each 6-well plate are pooled into a 15 ml conical tube (Fisher Scientific, Cat. # 14-959-53A) and spun down at 200xg. The pelleted cells then have the remaining media removed before being resuspended in 750 uL Trizol and stored in -20°C until ready for isolation of RNA and smRNA.

RNA and smRNA Isolation, Sodium Periodate Treatment, and smRNA Sequencing

smRNA is isolated using a combination of Trizol, the RNeasy Mini Kit (Qiagen, Cat. # 74104), and the RNeasy MiniElute Cleanup Kit (Qiagen, Cat. # 74204). Cell lysates are removed from the -20°C freezer and thawed on ice. After lysates are thawed and incubated for ~5 min at room temperature to permit complete dissociation of the nucleoproteins complex, 200 uL of chloroform (Thermo, Cat. # J67241.AP) is added per 1 mL of Trizol and mixed well by pipetting. Left to incubate for 2-3 min. Samples are

then centrifuges for 15 min at 12,000g at 4°C. Here, the mixture will separate into three phases, a lower pink phenol-chloroform (DNA and protein), an interphase, and a colorless aqueous phase containing our RNA and smRNA. Using a 200 uL pipette tip, the aqueous phase is carefully transferred to a new tube without disrupting the interphase to avoid transferring any of the organic phase layer.

RNA is isolated using the RNeasy Mini Kit according to the manufacturer protocols. This kit isolates RNA larger than 200 nts in length. RNA concentration was quantified using the NanoPhotometer system. The RNeasy MinElute Kit is used to isolate smRNA according to the manufacturer protocols. The RNeasy MinElute Kit isolates RNA transcripts shorter than 200 nt in length. Sodium periodate treatment is performed to identify fully mature piRNA transcripts due to the presence of 2'-O-methylation modification on their 3' end. Any transcripts without the 2'-O-methylation modification are degraded, leaving only mature piRNA transcripts. Each smRNA sample is divided into 4 aliquots of 400 ng each, 3 of which undergo sodium periodate treatment leaving 1 to serve as an untreated control. Sodium periodate protocol consists of the following: freshly prepared sodium periodate (Sigma, Cat. #BCBS5360V), 5X borate buffer generated from 150mM borax (Alfa Aesar, Cat. # T29C533), and 150mM boric acid (Fluka Analytics, Cat. # SZBG1280V) adjusted to a pH of 8.6 using sodium hydroxide (Thermo, Cat. #A4782902). A full description of sodium periodate reaction preparation is described in previously published protocols (Perera et al. 2019). The 3 sodium periodate treated technical replicates are recombined following treatment. The final smRNA sample size for the library preparation and sequencing, including treated and untreated, was n = 24.

smRNA libraries are prepared at the University of Michigan Advanced Genomics Core (AGC) using the SMARTer smRNA-Seq Kit (Takara, Cat. # 635031) using Takara smRNA Indexing Primer Set HT. Before sequencing, quality control and library preparation are performed by the AGC using TapeStation (Aligent RNA ScreenTape #5067-5576 and RNA ScreenTape Sample Buffer #5067-5577; analysis software 4.1). The samples are pooled onto one sequencing lane. The kit employs polyadenylation and template switching by extension steps, before adding adapters by PCR. The smRNA library is cleaned and size-selected using AMPure XP Beads (Fisher, Cat. # NC9933872) and sequencing was performed on an Illumina NovaSeq S1 flow cell (200 cycle) ensuring ~38 million reads per sample.

Bioinformatics Identification of piRNA Transcripts

FastQC (v0.11.5) and MultiQC (v1.8) were used to assess the quality of raw sequencing data. Adaptors were trimmed with cutadapt (v4.9) and reads with length ≥10 and ≤ 45 bp were selected for downstream analysis. Bowtie2 (v2.2.9) with 'end-to-end' mode without mismatches was used for alignment in the human genome (hg38) and PePr (v1.1.24) was used for differential peak calling between periodate treated and control groups (35, 36). PePr peak calling was used to identify peaks in treated vs untreated samples with size selection greater than 20bp and less than 45bp. Peak calling was running on a RedHat Linux Server (v7.9) with parameters—shiftsize 0—windowsize 20—threshold 1E-3—peaktype sharp. A False Discovery Rate cutoff of <0.05 was used to select peaks significantly enriched in periodate-treated groups and thus attributable to piRNA rather than other small RNA species. A unique piRNA ID was

assigned to each significant peak based on the genomic coordinates. This resulted in what we call our “piRNA-like” peaks/transcripts. Next, we calculated the expression levels and used in-house python code to find 5’ T and the 10th Adenosine to generate sequence motifs. Finally, we performed a comparison across stem cell states using peak location, sequence identity and expression.

Evaluation of piRNA Sequence Overlap Between MCF10A Monolayer, MCF10A Mammospheres, MCF7 Monolayer and MCF7 Mammospheres.

A visualization of the workflow for the piRNA data analysis is available in **Figure 3.1**. The data from peaks (piRNA-like transcripts) called using PePr are selected based on a length less than 45 bp and input into a data frame named “dat”. To determine whether there were unique or shared piRNAs between the conditions, we used the function bed_intersect between two conditions to identify if the piRNA peak sequences were matched or if they were different. If the sequence was a full match, then we deemed those sequences to be the same piRNA transcripts. Through this process, we obtained two data frames for each condition, one selecting out those piRNA sequences that fully matched between the two conditions being compared marked with “-overlap”, and one selecting for unique piRNAs with sequences not fully matched between the conditions being compared marked with “_no_overlap”. Next, peak annotations are performed using data from either “_overlap” or “_no_overlap” to produce data frames including columns for the annotation symbol and the annotation type. The annotations were built using genome - hg38. This allowed us to identify the specific genes for which these piRNA transcripts were mapped to and type of genomic regions where these

transcripts mapped to. Additionally, we examined the overlaps between genomic regions and repeated annotations. Only overlaps with more than 10 regions were deemed repeats.

Results

Detection of piRNAs in 2D and 3D MCF10A and MCF7 Cell Lines

The number of piRNA transcripts found in each cell line and morphology are as follows: MCF10A Monolayer (MCF10A-ML) has 3,039; MCF10A Mammospheres (MCF10A-MS) has 10,536; MCF7 Monolayer (MCF7-ML) has 14,225; and MCF7 Mammospheres (MCF7-MS) has 3,525 as depicted in **Figure 3.2**.

Comparisons of Number of Unique piRNAs by Cell Line and Cellular State

We investigated the number of piRNAs found in common between the following conditions: between the same cell line but different cellular state, MCF10A-ML vs MCF10A-MS and MCF7-ML vs MCF7-MS; and between the same cellular state and different cell line, MCF10A-ML vs MCF7-ML and MCF10A-MS and MCF7-MS. **Figure 3.3a** indicates that there are 1,628 piRNAs found in common between MCF7-ML and MCF7-MS, while there are 12,597 piRNAs still unique to the MCF7-ML and 1,897 piRNAs still unique to MCF7-MS. Additionally in **Figure 3.3b**, we observe 401 shared piRNAs between MCF10A-ML and MCF10A-MS. MCF10A-MS has 10,135 unique piRNAs while MCF10A-ML has 2,638 unique piRNAs. **Figure 3.4a** shows that the monolayers of both cell lines share 757 piRNAs in common, with 13,468 still unique to the monolayer of MCF7s and 2,282 unique to the monolayer of MCF10As. **Figure 3.4b** shows the mammospheres of both cell lines share 870 piRNAs in common, while MCF7

mammospheres have 2,655 unique piRNAs and MCF10A mammospheres have 9,666 unique piRNAs.

Length of piRNA Transcripts in Each Condition

Next, we analyzed the differences in piRNA transcripts lengths across the different conditions. **Figure 3.5** shows the length distribution of the piRNA transcripts across the different conditions. Peaks 20 bp or smaller and larger than 45 bp are excluded from further analysis. MCF10A-MS has the highest number of piRNA transcripts larger than 40 bp out of the four conditions. All four conditions show the majority of piRNA transcripts are between 30 and 40 bp. Interestingly, we see larger piRNA transcripts (50-60 bp) in the MCF10A mammospheres.

Differentially piRNA Mapped Genes Between Conditions

Figure 3.6a shows the number of unique piRNAs that map to genes in the 2D (monolayer) and 3D (mammosphere), as well as the number of piRNAs shared between the two states of MCF10A cells. MCF10A-MS have 6,465 unique piRNA transcripts mapped to genes compared to the MCF10A-ML, which has 1,150 unique piRNA transcripts. 1,184 unique piRNAs map to genes in both the MCF10A-ML and MCF10A-MS. **Figure 3.6b** shows the number of unique piRNAs that map to genes in the 2D (monolayer) and 3D (mammosphere), as well as the number of piRNAs shared between the two states of MCF7 cells. MCF7-ML has 6,160 unique piRNA transcripts mapped to genes and MCF7-MS has 775 unique piRNA transcripts. The monolayer and mammospheres of the MCF7 cells share 2,020 piRNA transcripts. In order to see the overlap of the piRNAs transcripts found in both cell lines and in both cell states, **Figure 3.7** is a Venn diagram showing all the piRNA transcripts the cell lines and morphological

states have in common. Notably, there are 526 unique piRNA transcripts that map back to genes that all four conditions have in common. Additionally, there are 390 piRNA transcripts shared between both cell line monolayers, and 232 piRNA transcripts shared between both cell line mammospheres.

Differential Expression of Detected piRNAs Between Conditions

We annotated the MCF10 ML and MCF10 MS piRNAs to the human genome hg38 and analyzed the percentage of piRNA mapping to genomic regions in the human genome (hg38). **Figure 3.8a** shows the percentage of piRNAs represented by various genomic regions categorized by piRNAs found only in MCF10A-ML (dark blue), piRNAs found only in MCF10A-MS (light blue), piRNAs which overlapped in both MCF10A-ML and MCF10A-MS (green), and piRNA derived locations expected in a random area of the hg38 genome (yellow). The piRNAs for MCF10A-ML most frequently mapped to genes (~15%), promoters (~10%), and exons (~9%) and a lower proportion of piRNAs mapping to introns (~37%) compared to the expected proportion of exons for a random region of the hg38 genome (~4% exonic regions and ~44% intronic regions). Similarly, for MCF10A-MS, a higher proportion of piRNA map to exons (~30%) promoters (~18%), and genes (~12%) and an extremely lower proportion of piRNAs mapping to introns (~12%) compared to the expected proportion of exons for a random region of the hg38 genome (~4% exonic regions and ~44% intronic regions). Differently, MCF10A-MS piRNAs also mapped in a higher proportion to both 5’UTRs and 3’UTRs compared to the random region. There are distinct differences in the percentage of piRNAs mapping seen between MCF10A-ML and MCF10A-MS. MCF10A-ML have a slightly higher percentage of piRNAs mapping to gene regions compared to MCF10A-MS.

Interestingly, MCF10A-MS has a much higher percentage of piRNAs mapping to 5'UTR, exon, and 3'UTR regions compared to MCF10A-ML. MCF10A-MS also has a slightly higher percentage of piRNAs mapping to promoter regions compared to MCF10A-ML. Notably, MCF10A ML_MS followed more similar patterns to MCF10A-ML than MCF10A-MS.

We then performed analysis of piRNA mapping to repetitive regions in the genome for MCF10A monolayers and mammospheres. **Figure 3.8b** shows that there is a higher proportion of piRNAs from MCF10A-ML mapping to short interspersed nuclear elements (SINEs; ~46%) compared to the proportion of SINEs in random regions of the genome (~31%). Interestingly, SINEs are the only repetitive genomic region that piRNAs from MCF10A-ML mapped to with a greater proportion compared to the random regions. MCF10A-MS piRNAs mapping to repetitive regions in the genome revealed a slightly higher proportion of piRNAs mapping to long terminal repeat (LTR) retrotransposons (~21%) and DNA transposons (~9.5%) compared to the proportion of those repetitive regions in a random region of the genome (~32% mapped to SINEs and ~7% mapped to DNA). Interestingly, piRNA in MCF10A ML_MS overlap had higher proportions of piRNAs mapping to long interspersed nuclear elements (LINEs; ~47%) and LTRs (~23%) compared to those repetitive regions in a random region of the genome (~43% mapped to LINEs and ~18% mapped to LTRs).

We annotated the MCF7 ML and MCF7 MS piRNAs to the human genome hg38 and analyzed the percentage of piRNA mapping to genomic regions in the human genome (hg38). **Figure 3.9a** shows the percentage of piRNAs represented by various genomic regions categorized by piRNAs found only in MCF7-ML (dark blue), piRNAs

found only in MCF7-MS (light blue), piRNAs which overlapped in both MCF7-ML and MCF7-MS (green), and piRNAs expected in a random area of the hg38 genome (yellow). The piRNAs for MCF7-ML most frequently mapped to genes (~13%), promoters (~ 10%), and exons (~ 8%) and a lower proportion of piRNAs mapped to introns (~38%) compared to the expected proportion of exons for a random region of the hg38 genome (~4% exonic regions and ~44% intronic regions). MCF7-MS piRNA mapping show a similar pattern with a higher proportion of piRNAs mapped to genes (~17%), promoters (~14%), and exons (~11%) and a lower proportion of piRNAs mapping to introns (~32%) compared to the expected proportion of exons for a random region of the hg38 genome (~4% exonic regions and ~44% intronic regions). Interestingly, MCF7-MS has a slightly higher proportion of piRNAs mapped to genes, promoters, and exons compared to MCF7-ML. Notably, MCF7 ML_MS followed similar patterns as both MCF7-ML and MCF7-MS.

Lastly, we performed analysis of piRNA mapping to repetitive regions in the genome for MCF7 monolayers and mammospheres. **Figure 3.9b** shows that there is a higher proportion of piRNAs from MCF7-ML mapping to long terminal repeat (LTR) retrotransposons (~28%) compared to the proportion of LTRs in random regions of the genome (~19%). Other than LTRs, the only other repetitive regions MCF7-ML piRNAs mapped in a slightly higher proportion are DNA transposons compared to the randomly generated repetitive regions. Differently, MCF7-MS piRNAs mapping to repetitive regions in the genome revealed a much higher proportion of piRNAs mapping to SINEs (~45%) compared to both the random regions (~31%) and MCF7_ML (~26%). piRNA in MCF10A ML_MS overlap had higher proportions of piRNAs mapping to both LTRs

(~23%) and DNA (~9%) compared to the proportion of those repetitive regions in a random region of the genome.

Discussion

Detection of piRNAs in 2D and 3D MCF10A and MCF7 Cell Lines. Baseline piRNA expression in non-tumorigenic ER- breast epithelial cells, MCF10As, are different from baseline piRNA expression in ER+ mammary gland epithelial cells derived from metastatic breast cancer, MCF7s. Additionally, baseline piRNA expression of each individual cell line in monolayer is different from the baseline expression of each cell line in mammosphere formation. To our knowledge, these findings represent the first baseline characterization of piRNAs in MCF10A and MCF7 cell lines in both monolayer and mammospheres using sodium periodate treatment to select for mature piRNAs. Although piRNA expression has been studied and documented in numerous studies, those studies either focused on a limited number of piRNAs or used resources such as piRNAbank to confirm piRNA expression, which does not use sodium periodate treatment to validate piRNA transcripts (Fu et al., 2015; Hashim et al., 2014). Current research in breast cancer and piRNAs have skipped a vital piece of the puzzle where studies jump straight into the actions of each individual piRNA without understanding how or why piRNAs are being expressed in the first place. By determining the baseline piRNA expression of a sample, we can better investigate how perturbations, such as chemical exposures, effect the expression of piRNAs.

The MCF10A and MCF7 monolayers act as our baseline for formation of mammospheres using the same cell lines. As stated above, mammosphere formation aids in the isolation, characterization, and culture of mammary stem cells, therefore our

analysis of piRNA detection in mammospheres allows us to identify the characterization of piRNA in stem-like cells. In this study, we compare the stem-like state of cells to the respective monolayer cell line. **Figure 3.2** displays the total number of piRNAs found in each condition. MCF10A monolayer has the fewest number of piRNAs detected with 3,039. MCF7 mammospheres have the second highest with 3,525 and then the largest number of piRNAs detected were in the MCF7 monolayer with 14,225 followed by MCF10A mammospheres with 10,536 total piRNAs. It is unsurprising that the number of MCF10A-ML piRNAs are so much fewer than the MCF10A-MS due to isolation of the stem-like cells in the mammosphere assay. We expected to see a higher number of piRNAs expressed in the mammospheres due to the known characteristics of stem-like cells. The MCF7s however were reversed, there were more detected piRNAs in the monolayer than the mammospheres. We believe this is due to the already transformed characteristics of MCF7 cells. As stated above MCF7s are ER+ mammary gland epithelial cells derived from metastatic breast cancer, therefore these cells express many cancer characteristics which can explain the high number of piRNAs expressed in the monolayer before we even put them into mammospheres. MCF7 mammospheres have a lower number of detected piRNAs than the MCF7 monolayer, showing that these cells may already have more stem-like characteristics in monolayer, therefore there is not as large of a difference after mammosphere formation.

Figure 3.3 showed the number of unique piRNAs found in each condition. Each comparison made in figure 3.3 shows the number of shared piRNAs found in each comparison in addition to the unique number of piRNAs between the two conditions. **Figure 3.3a** and **b** highlight the differences of piRNA number between the cellular state

of each cell line. Here we see there are shared piRNAs between the monolayer and mammosphere of each cell line, these shared piRNAs could play an important role in cell line identity, for example what makes an MCF10A cell an MCF10A cell. **Figure 3.4a** and **3.4b** highlight the differences in piRNA number between cellular states. Here we see that the monolayers of both MCF7 and MCF10As share 779 piRNAs; these shared piRNAs could play a role in maintaining the monolayer. Additionally, we see that mammospheres from both MCF7s and MCF10As share 894 piRNAs; these shared piRNAs could play a role in mammosphere formation.

Length of piRNA Transcripts in Each Condition. In current research, the lengths of piRNA can range from 20-36 nucleotides in length, however, piRNAs can occur in longer transcripts in groups or as preprocessed RNA leaving the exact length of piRNA in somatic tissue to remain unclear (Wang et al., 2024). A study from Perera et al. 2019 demonstrated how adult mouse somatic piRNAs were shorter than germline piRNA, indicating that different lengths in piRNA may result in different functions of the piRNA. Notably, the tissues used in Perera et al. 2019 were normal tissues whereas here we use cell lines of non-tumorigenic and cancerous cells, therefore some of the differences in length could be due to transformation. **Figure 3.5** indicates differential lengths of the piRNA transcripts between different conditions. While MCF10A-ML, MCF7-ML and MCF7-MS showed the majority of the piRNA between lengths 20-45 bp, MCF10A-MS showed a much larger distribution of length of piRNA ranging from 20 to 60 bp. These longer piRNAs could potentially be taking on more unique functions than the shorter piRNAs. Another possible explanation for these longer lengths could be that the piRNA

peak calling is unable to distinguish between multiple transcripts corresponding to the same region in the genome.

Differentially piRNA Mapped Genes between Conditions. As stated previously, breast cancer is a highly heterogeneous disease with different subtypes based on distinct gene expression profiles and clinical behaviors. The underlying molecular mechanisms through which these genes are differentially expressed is still unknown. Current research determines that piRNAs might play a role in the expression of genes involved in breast cancer development (Ross et al., 2014). In this study, we examine the number of piRNA transcripts mapping to genes in each condition. **Figure 3.6a** indicates that 6,465 piRNAs map to unique genes in MCF10A MS compared to MCF10A-ML and 1,150 piRNAs map to unique genes in MCF10A-ML. The genes these piRNAs in MCF10A-MS map to may play a role in the development of a more cancer stem cell-like state. There are 1,184 piRNAs that map to the same genes between both conditions; these genes may be responsible for maintaining the identity of an MCF10A cell. In **figure 3.6b**, MCF7-ML piRNAs map to 6,160 unique genes sharing 2,020 with MCF7-MS. Due to the cancerous characteristics of this cell line, these piRNA may be mapping to genes known to be involved in the luminal subtype of breast cancer.

Differential Location of Detected piRNAs between Conditions. Our analysis reveals that piRNAs are derived from unique locations depending on the cell line and cell culture condition. The monolayers from both cell lines displayed a more similar pattern to the randomly generated regions compared to the mammospheres from both cell lines. This demonstrates that piRNAs involved in stem cell enriched mammospheres may be derived from different genomic regions than those involved in

the maintenance of monolayers. In both the monolayers and mammospheres of MCF10A and MCF7 cells, piRNAs derived from intronic regions were expressed at lower-than-expected levels, while those from exonic regions were expressed higher than expected (Figure 3.8a and 3.9a). Additionally, the proportion of piRNAs mapped to introns and exons in the MCF10A-MS are extremely lower and higher (respectively) compared to the piRNAs mapped in MCF10A-ML. These current results indicate that a higher proportion of piRNAs map to exons rather than intronic regions and reveals a potential new mechanism for both baseline piRNA function in the breast and the role of piRNAs in a stem cell rich environment. Numerous studies, as stated in the introduction, have identified the changing expression of piRNAs in breast cancer, however, the specific mechanisms behind those changes are still unknown. Although more research is needed to determine the exact effect of location specific derived piRNA, it is clear in this experiment that piRNAs map to different regions of the genome depending on cell line and culture condition. We hypothesize that differential expression of piRNAs in stem cell enriched mammospheres and mapping to exonic regions could play a role in the development of breast cancer.

In order to ensure comprehensive and accurate transcriptome analysis, in the future, we would like to analyze the saturate depth of our RNA sequencing. By re-analyzing data with 25, 50, and 75% of the NGS data, we would be able to see the depth of saturation allowing us to determine if further depth is needed in future projects or if we are capturing the most accurate quantification possible.

Further investigation into the actual targets of these piRNA is necessary. Understanding the targets of these piRNAs in each condition could reveal unique roles

for specific piRNAs and uncover potential novel mechanisms in the development of breast cancer. Additional work is also necessary to further characterize piRNA profiles of other subtypes of breast cancer.

Conclusion

To our knowledge, this study is the first of its kind to rigorously evaluate the baseline expression of piRNA in MCF10A and MCF7 monolayer and mammospheres. These results highlight the differential expression of validated piRNAs between cell lines and the impact of the mammosphere assay on piRNA expression. Our results indicate that piRNA expression is dependent on molecular subtype and can change after perturbation such as a mammosphere assay. The importance of creating a validated piRNA baseline of these cell lines is to allow us to examine the effects of perturbations, including chemicals and other environmental exposures, on these piRNAs being able to compare data we obtain in exposure experiments to these baseline experiments.

References

Baldasici, O., Soritau, O., Roman, A., Lisencu, C., Visan, S., Maja, L., Pop, B., Fetica, B., Cismaru, A., Vlase, L., Balacescu, L., Balacescu, O., Russom, A., & Tudoran, O. (2024). The transcriptional landscape of cancer stem-like cell functionality in breast cancer. *Journal of Translational Medicine*, *22*, 530. https://doi.org/10.1186/s12967-024-05281-w

Bhat, V., Allan, A. L., & Raouf, A. (2019). Role of the Microenvironment in Regulating Normal and Cancer Stem Cell Activity: Implications for Breast Cancer Progression and Therapy Response. *Cancers*, *11*(9), 1240. https://doi.org/10.3390/cancers11091240

Chalbatani, G. M., Dana, H., Memari, F., Gharagozlou, E., Ashjaei, S., Kheirandish, P., Marmari, V., Mahmoudzadeh, H., Mozayani, F., Maleki, A. R., Sadeghian, E., Nia, E. Z., Miri, S. R., Nia, N. zainali, Rezaeian, O., Eskandary, A., Razavi, N., Shirkhoda, M., & Rouzbahani, F. N. (2018). Biological function and molecular mechanism of piRNA in cancer. *Practical Laboratory Medicine*, *13*, e00113.

Creighton, C. J., Li, X., Landis, M., Dixon, J. M., Neumeister, V. M., Sjolund, A., Rimm, D. L., Wong, H., Rodriguez, A., Herschkowitz, J. I., Fan, C., Zhang, X., He, X., Pavlick, A., Gutierrez, M. C., Renshaw, L., Larionov, A. A., Faratian, D., Hilsenbeck, S. G., … Chang, J. C. (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13820–13825.

Dandawate, P. R., Subramaniam, D., Jensen, R. A., & Anant, S. (2016). Targeting cancer stem cells and signaling pathways by phytochemicals: Novel approach for breast cancer therapy. *Seminars in Cancer Biology*, *40–41*, 192–208.

Ding, X., Li, Y., Lü, J., Zhao, Q., Guo, Y., Lu, Z., Ma, W., Liu, P., Pestell, R. G., Liang, C., & Yu, Z. (2021). piRNA-823 Is Involved in Cancer Stem Cell Regulation Through Altering DNA Methylation in Association With Luminal Breast Cancer. *Frontiers in Cell and Developmental Biology*, *9*, 641052.

Dontu, G., Abdallah, W. M., Foley, J. M., Jackson, K. W., Clarke, M. F., Kawamura, M. J., & Wicha, M. S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes & Development*, *17*(10), 1253–1270.

Ercan, C., van Diest, P. J., & Vooijs, M. (2011). Mammary Development and Breast Cancer: The Role of Stem Cells. *Current Molecular Medicine*, *11*(4), 270–285.

Fu, A., Jacobs, D. I., Hoffman, A. E., Zheng, T., & Zhu, Y. (2015). PIWI-interacting RNA 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. *Carcinogenesis*, *36*(10), 1094–1102.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46.

Holoch, D., & Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nature Reviews. Genetics*, *16*(2), 71–84.

Huang, G., Hu, H., Xue, X., Shen, S., Gao, E., Guo, G., Shen, X., & Zhang, X. (2013). Altered expression of piRNAs and their relation with clinicopathologic features of breast cancer. *Clinical and Translational Oncology*, *15*(7), 563–568.

Jiagge, E., Chitale, D., & Newman, L. A. (2018). Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *The American Journal of Pathology*, *188*(2), 271–279.

Liu, J., Zhang, S., & Cheng, B. (2018). Epigenetic roles of PIWI-interacting RNAs (piRNAs) in cancer metastasis (Review). *Oncology Reports*, *40*(5), 2423–2434.

Maleki Dana, P., Mansournia, M. A., & Mirhashemi, S. M. (2020). PIWI-interacting RNAs: New biomarkers for diagnosis and treatment of breast cancer. *Cell & Bioscience*, *10*(1), 44.

Ohara, T., Sakaguchi, Y., Suzuki, T., Ueda, H., Miyauchi, K., & Suzuki, T. (2007). The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated. *Nature Structural & Molecular Biology*, *14*(4), 349–350.

Perera, B. P. U., Tsai, Z. T.-Y., Colwell, M. L., Jones, T. R., Goodrich, J. M., Wang, K., Sartor, M. A., Faulk, C., & Dolinoy, D. C. (2019). Somatic expression of piRNA and associated machinery in the mouse identifies short, tissue-specific piRNA. *Epigenetics*, *14*(5), 504–521.

Qian, L., Xie, H., Zhang, L., Zhao, Q., Lü, J., & Yu, Z. (2021). Piwi-Interacting RNAs: A New Class of Regulator in Human Breast Cancer. *Frontiers in Oncology*, *11*, 695077.

Rios-Fuller, T. J., Ortiz-Soto, G., Lacourt-Ventura, M., Maldonado-Martinez, G., Cubano, L. A., Schneider, R. J., & Martinez-Montemayor, M. M. (2018). Ganoderma lucidum extract (GLE) impairs breast cancer stem cells by targeting the STAT3 pathway. *Oncotarget*, *9*(89), 35907–35921. https://doi.org/10.18632/oncotarget.26294

Tharmapalan, P., Mahendralingam, M., Berman, H. K., & Khokha, R. (2019). Mammary stem cells and progenitors: Targeting the roots of breast cancer for prevention. *The EMBO Journal*, *38*(14), e100852.

Wang, K., Perera, B. P. U., Morgan, R. K., Sala-Hamrick, K., Geron, V., Svoboda, L. K., Faulk, C., Dolinoy, D. C., & Sartor, M. A. (2024). piOxi database: A web resource of germline and somatic tissue piRNAs identified by chemical oxidation. *Database*, *2024*, baad096.

Zhang, H., Ren, Y., Xu, H., Pang, D., Duan, C., & Liu, C. (2013). The expression of stem cell protein Piwil2 and piR-932 in breast cancer. *Surgical Oncology*, *22*(4), 217–223.

Figures and Tables

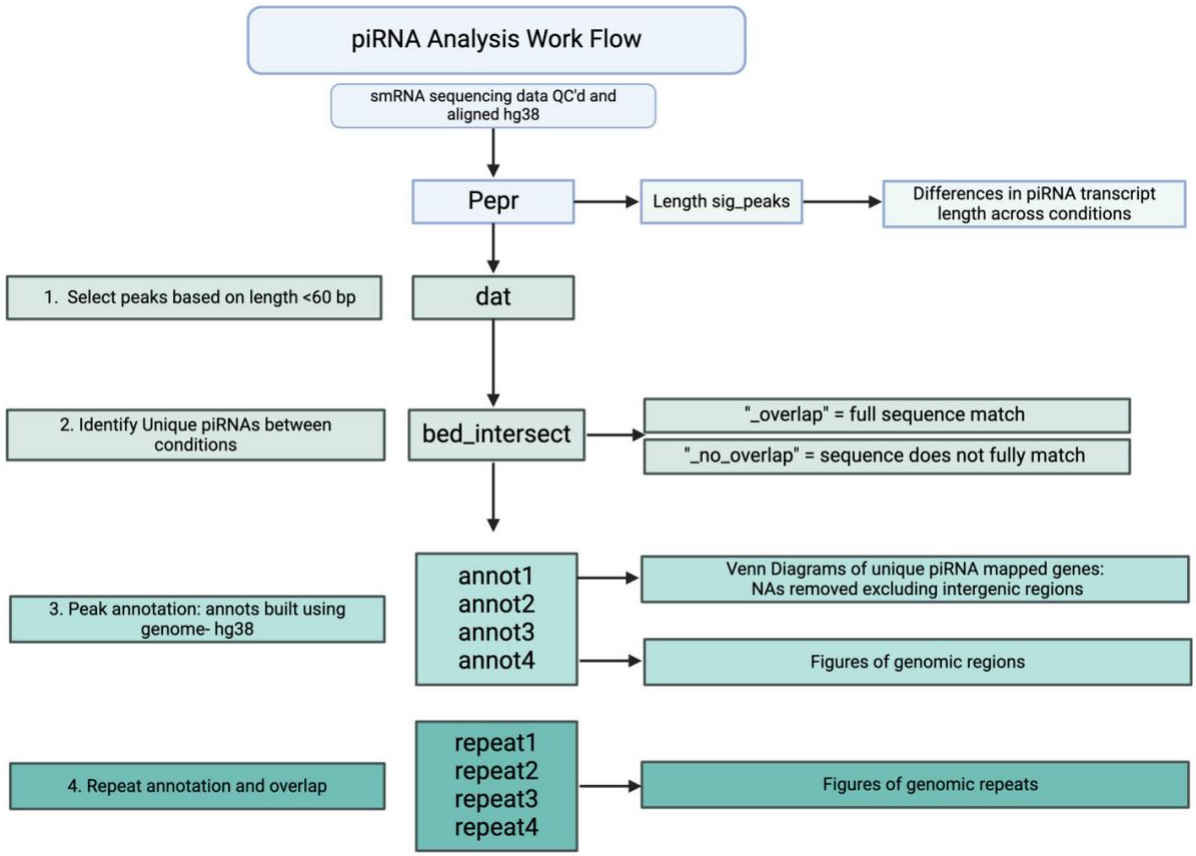


Figure 3.1: piRNA Analysis Workflow. After sequence data returned from the Advanced Genomics Core (ACG) at the University of Michigan, data is run through PePr for size selection and to identify peaks in treated vs untreated samples. The data was then input into R to analyze where we first selected peaks based on length less than 60 bp resulting in data frame “dat”. We then used the function “bed_intersect” in package “valr” allowing us to identify unique piRNAs between conditions. We then annotated these piRNAs using genome hg38. This data is used to create the Venn Diagrams of piRNA mapped genes (NAs were removed) and the figures of genomic regions. Finally, we generated figures for genomic repeats.

Cell line and Morphology	Number of piRNAs
MCF10A – Monolayer	3,039
MCF10A – Mammosphere	10,536
MCF7 – Monolayer	14,225
MCF7 – Mammosphere	3,525

Figure 3.2: Number of piRNAs Found in Each Cell Line (Monolayer vs Mammosphere).

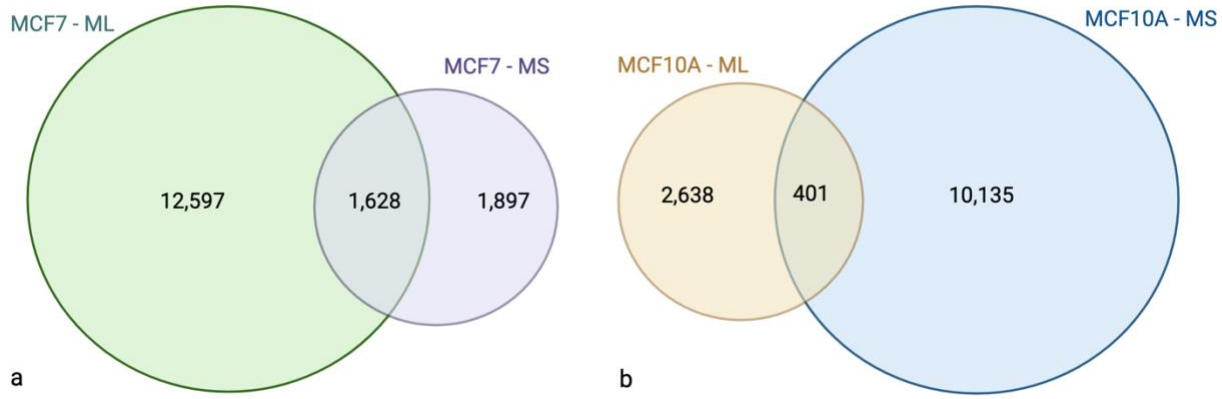


Figure 3.3. Comparisons of the Number of Unique piRNA Transcripts Between Cell Line and Cellular State. a) Venn diagram of piRNAs found in MCF7-ML and MCF7-MS. b) Venn diagram of piRNAs found in MCF10A-ML and MCF10A-MS.

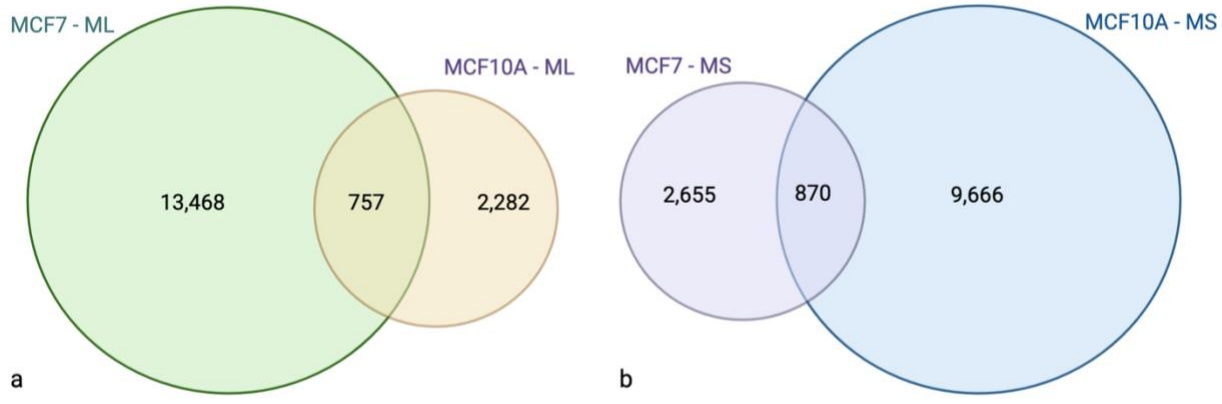


Figure 3.4. Comparisons of the Number of Unique piRNA Transcripts Between Cellular State. a) Venn diagram of piRNA transcripts found in the monolayer of MCF7 and MCF10A cells. b) Venn diagram of piRNA transcripts found in the mammospheres of MCF7 and MCF10A cells.

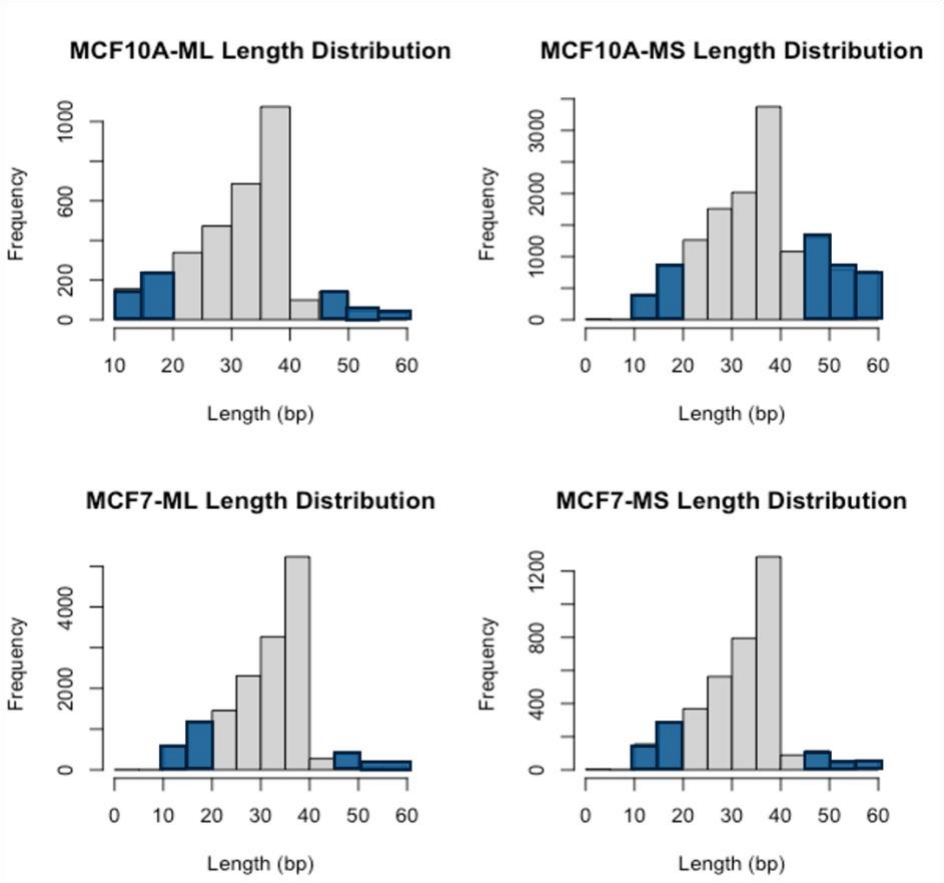


Figure 3.5. The Length Distribution of the piRNA Transcripts Across the Different Conditions. Those peaks under 20bp or larger than 45 bp are excluded from analysis (shown in blue).

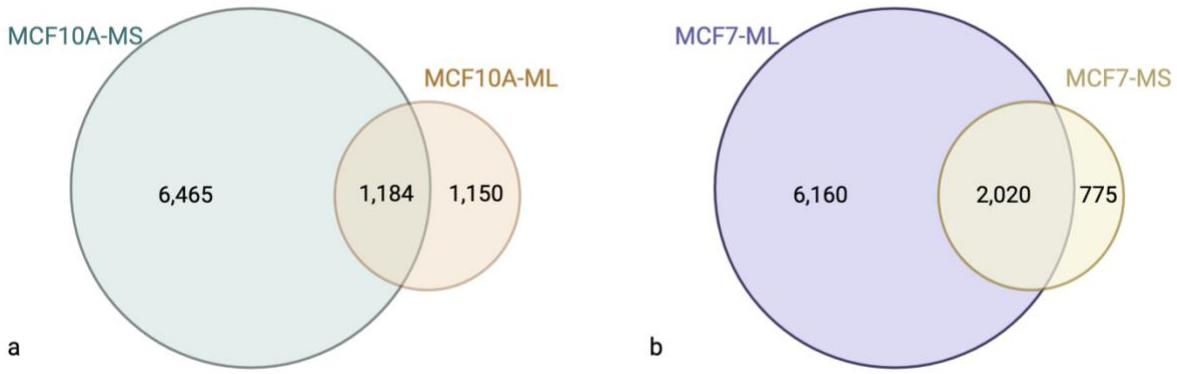


Figure 3.6. Venn diagrams of the Number of piRNA Transcripts Mapping to Genes in Each Condition.

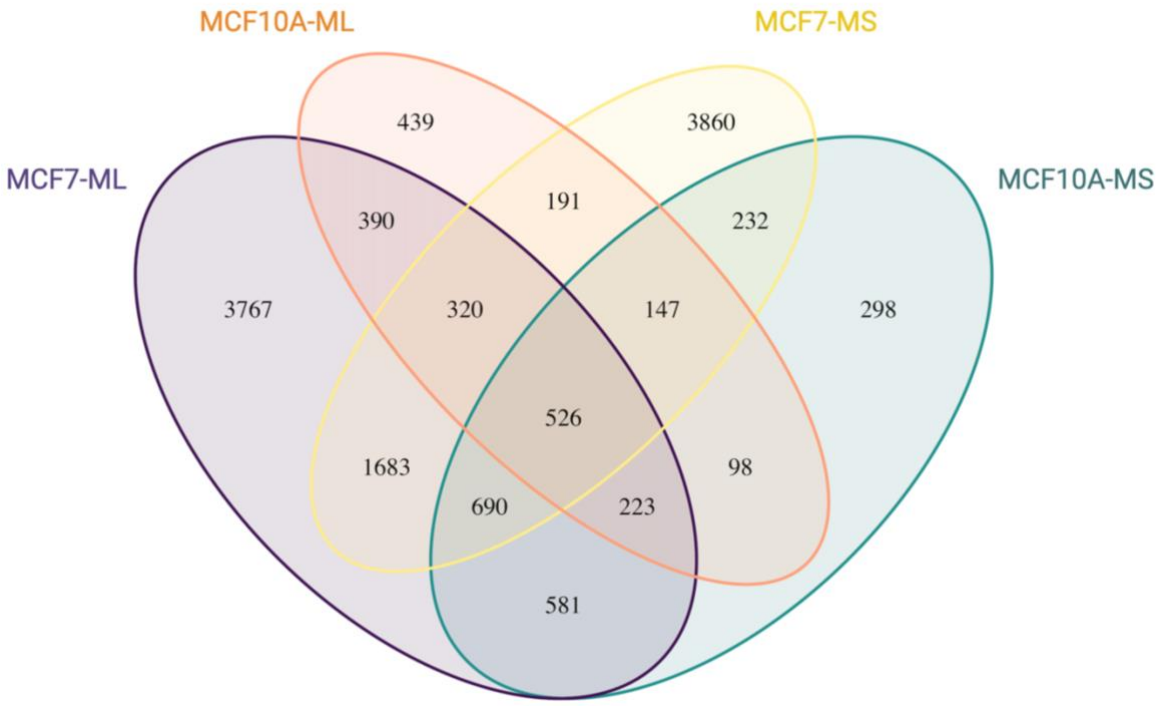
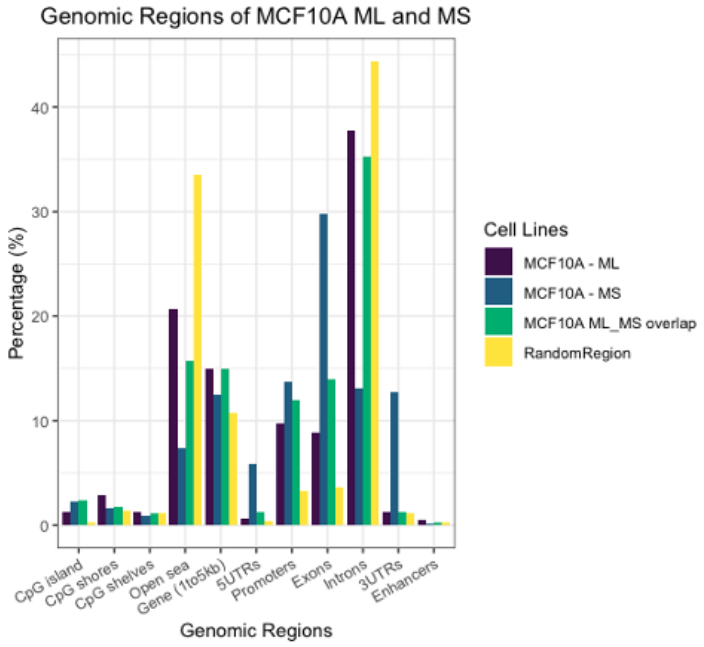


Figure 3.7. Venn Diagram of the Number of piRNA Transcripts Mapping to Genes in Each Condition. (ML- Monolayer and MS- Mammosphere). MCF7-ML is purple, MCF10A-ML is orange, MCF7-MS is yellow, and MCF10A-MS is green.

a)



b)

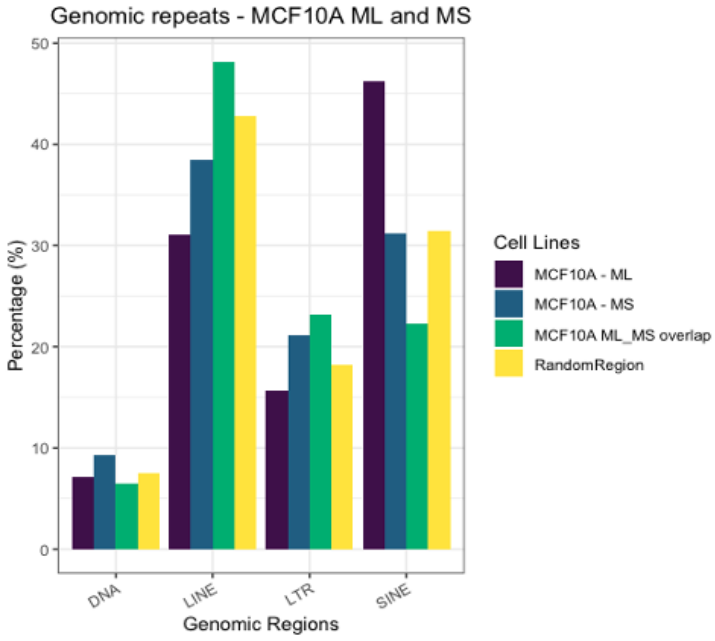
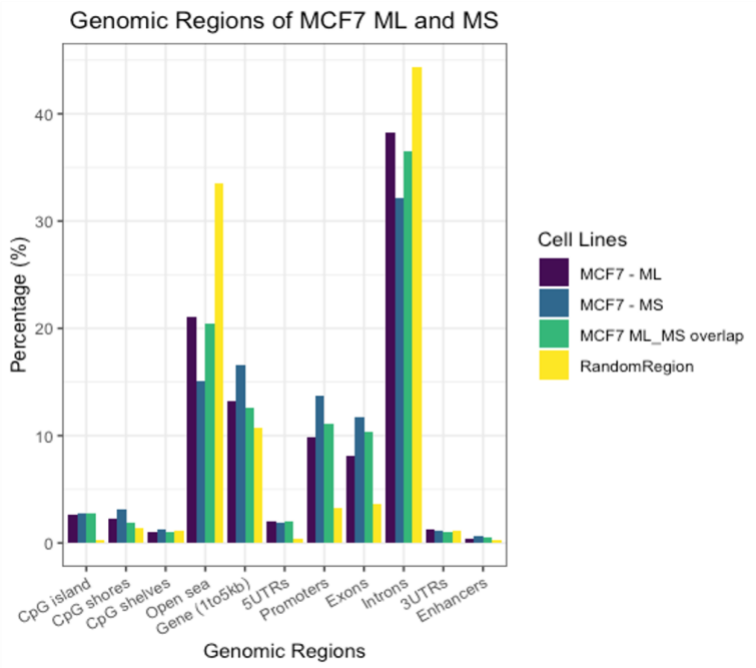


Figure 3.8. Genomic Annotations and Repetitive Regions for MCF10A ML and MS piRNA Expression. a) Genomic annotations were generated for the regions from which piRNA were derived for in MCF10A ML and MS. MCF10A ML (dark blue), MCF10A MS (light blue), and annotated regions found in both MCF10A_ML and MCF10A_MS,

MCF10A ML_MS overlap (green). Random regions were generated to compare our data to (yellow). b) Repetitive annotations were generated for the regions from which piRNA were derived for in MCF10A ML and MCF10A MS. MCF10A ML (dark blue), MCF10A MS (light blue), and repetitive annotations found in both MCF10A_ML and MCF10A_MS, MCF10A ML_MS overlap (green). Random regions were generated to compare our data to (yellow).

a)



b)

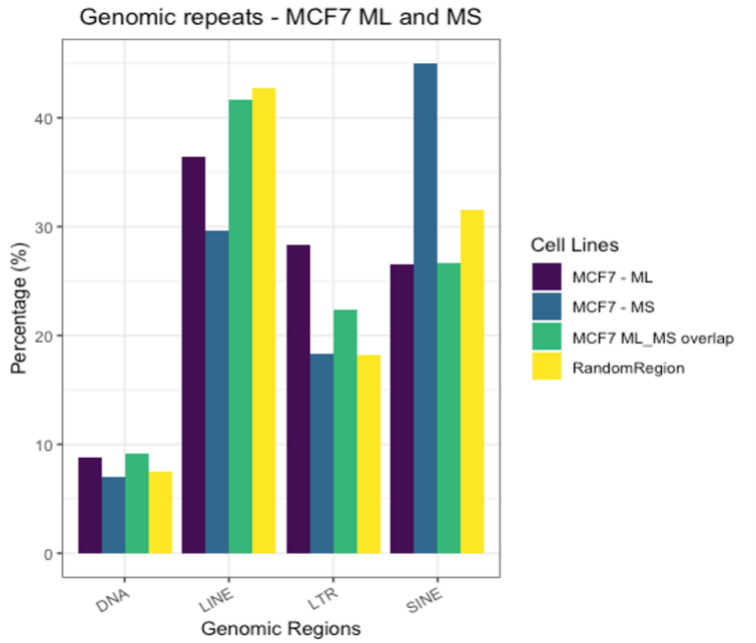


Figure 3.9. Genomic Annotations and Repetitive Rregions for MCF7 ML and MS piRNA expression. a) Genomic annotations were generated for the regions from which piRNA were derived for in MCF7 ML and MS. MCF7 ML (dark blue), MCF7 MS (light blue), and annotated regions found in both MCF7 ML and MCF7 MS, MCF7 ML_MS overlap

(green). Random regions were generated to compare our data to (yellow). b) Repetitive annotations were generated for the regions from which piRNA were derived for in MCF7 ML and MCF7 MS. MCF7 ML (dark blue), MCF7 MS (light blue), and repetitive annotations found in both MCF7 ML and MCF7 MS, MCF7 ML_MS overlap (green). Random regions were generated to compare our data to (yellow).

Chapter 4

Aim 3: Determine Morphological Transformation and Cellular Plasticity of Normal Human Breast Epithelial Cells After 40-Week Exposure to Low Dose Cadmium.

Abstract

As of 2021, breast cancer is the most commonly diagnosed cancer worldwide with the 5th highest mortality rate of any cancer in the US. Breast tumors are highly heterogeneous resulting from the acquisition of morphological alterations and cancer hallmarks including stemness and cellular plasticity. Roughly 80-85% of cases occur in women with no family history of the disease indicating how heavily lifestyle and environmental factors play in disease etiology. Cadmium (Cd) is a naturally occurring toxic heavy metal and a known lung carcinogen, however, its role in breast cancer remains controversial. *In vitro* studies have shown that breast cells exposed to Cd are malignantly transformed through estrogen receptor-independent mechanisms. Here we investigate the role of long term (40-week) low dose Cd (0.25 μM and 2.5 μM) exposure on cancer stem cell markers and cellular plasticity in non-tumorigenic MCF10A cells. We developed two high content image-based immunocytochemistry assays to measure the impact of the long-term cadmium exposure of the cells in an unbiased manner every 10 weeks starting at week 0 through week 40. Quantification of Keratin 8 and Keratin 14 (markers of luminal and basal cells, respectively) was used to test cell plasticity. Quantification of CD24-/CD44+ and ALDH1A3 expression (markers of cancer stem cells

in breast cancer) was used to test stemness. RNA sequencing and differential gene expression analysis were also performed from samples collected every 10 weeks and gene expression patterns analyzed via gene set enrichment and clustering analysis. Our results show that the luminal marker, Keratin 8, decreases over time in both the control and treated groups, while the myoepithelial marker, Keratin 14, increases over time in the controls but decreases in the treated groups, with a divergence from the controls observed at week 30 and 40. We also see an increased population of cells expressing both keratin 8 and keratin 14 markers, indicating hybrid states of these cells and therefore an acquisition of cellular plasticity. Further, our RNA seq data and pathway enrichment analysis reveals genes that change over time in response to low dose cadmium are associated with targets of MYC, a canonical regulator of embryonic stem cells and a strong oncogene implicated in numerous cancers.

Introduction

Breast cancer is a highly heterogeneous disease with four main subtypes based on the expression of hormone receptors estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor (HER2). The four main subtypes are Luminal A ($ER^+/PR^+/HER2^-$), Luminal B ($ER^+/PR^{-/+}/HER2^{+/-}$), HER2 enriched ($ER^-/PR^-/HER2^+$) and triple negative ($ER^-/PR^-/HER2^-$) (Orrantia-Borunda et al. 2022). Triple negative breast cancer (TNBC) is the most aggressive subtype of breast cancer characterized by the lack of expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor (HER2) amplification (Ossovskaya et al. 2011). Patients diagnosed with TNBC more frequently experience metastasis to the lung, liver and brain, contributing to the poor survival observed with this disease (Dietze et al. 2015). In

addition to being a subtype of breast cancer, TNBC itself is extremely complex and diverse. Lehmann et al. 2011 identified six molecular subtypes of TNBC including basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR). Gene ontologies of each subtype reflect unique profiles which identify and define the diversity between the subtypes. A study from Lim et al., 2009 showed that basal-like aggressive breast cancers can derive from dysregulated luminal progenitor cells rather than basal stem cells indicating that these cells have acquired phenotypic plasticity, shifting from luminal to basal-like characteristics (Chiche et al., 2019).

The mammary gland is organized into a tree-like structure composed of hollow branches with an inner layer of luminal epithelial cells that face the lumen and are surrounded by an outer layer of myoepithelial cells (**Figure 4.1a**). Both ductal and alveolar luminal cells express keratin 8 and 18 (KRT8/18) genes. KRT8 dimerizes with KRT18 to form an intermediate filament in the cytoplasm of epithelial cells and plays an important role in the structural integrity of the cell and cellular differentiation (NIHa, 2024). The outer myoepithelial/basal layer expresses keratin 5 and 14 (KRT5/14). KRT14 is usually found as a heterotetramer with two KRT5 molecules and for the cytoskeleton of epithelial cells (NIHb,2024). A recent study by Thong et al. 2020 identifies a hybrid population of cells which co-express both the luminal markers KRT8 and the basal marker K14. This evidence suggests that these cells may have a high plasticity and are transitioning between epithelial and mesenchymal cellular states, or that these cells may be suspended in a hybrid epithelial/mesenchymal state (Thong et al., 2020). Additionally in mammary glands, there are at least two breast stem cell

populations including ALDH1A3+ luminal stem cells with an epithelial phenotype and CD44+/CD24- basal stem cells with a more mesenchymal phenotype (Visvader and Stingl, 2014 and Van Keymeulen et al. 2011). Recent studies suggest that there is an additional population of cells that express both ALDH1A3+ and CD44+/CD24- and are more likely to form mammospheres than the ALDH+ cells alone (**Figure 4.1b**)(Colacino et al. 2018).

The etiological drivers of stemness and cellular plasticity in the normal breast and breast cancer remain poorly understood. Cadmium is a well-established human health risk, however its role in breast cancer remains controversial. Although it and has been implicated in breast cancer initiation and promotion by numerous mechanistic studies, multiple epidemiological studies have found null relationships (Julin et al. 2012, Adams et al. 2012, Florez-Garcia et al. 2023). This is in contrast with case- control studies that have found higher concentrations of cadmium in urine of breast cancer cases compared to controls (Gallagher et al. 2010, Strumylaite et al. 2014). Additionally, exposure to cadmium *in utero* alters mammary gland development and gene expression in mice (Parodi et al., 2017). A recent study determined that exposure of breast cancer cells to a range of cadmium doses (1μM -60μM) resulted in differential epigenetic regulation of important cancer related signaling pathways such as the Wnt (Liang et al., 2020). Several studies have shown that long-term exposure to low doses of cadmium induces phenotypic changes consistent with an epithelial/mesenchymal transition (EMT). Ponce et al. 2015 treated ER+ breast cancer epithelial cells, MCF7, with cadmium for 6 months, resulting in decreased expression of E-cadherin, a characteristic of EMT. Another study by (Benbrahim-Tallaa et al., 2009) treated non-malignant breast epithelial

cells, MCF10As, with cadmium for 40 weeks resulting in cellular transformation to more basal-like cancer phenotype. The mechanism underlying these long-term cadmium exposures and increased EMT remains unclear. Wei et al. 2017 conducted a 4-week long experiment using MCF10A cells to show that cadmium promotes EMT through modulation of SNAIL. However, the mechanism of cadmium induced SNAIL is still unknown. Another short-term study using normal breast epithelial cells showed that cadmium at doses relevant to human exposure induced alterations in breast stem cell proliferation and differentiation by inhibiting HIF-1a (Rocco et al., 2018). Taken together, this data indicated the gap in knowledge of the mechanism underlying cadmium exposure and breast cancer progression.

In this aim, we investigate the role of long term (40-week) low dose Cd (0.25 μM and 2.5 μM) exposure on cancer stem cell markers and cellular plasticity in non-tumorigenic MCF10A cells. We developed two high content image-based immunocytochemistry assays to measure the impact of the long-term cadmium exposure on the cells in an unbiased manner every 10 weeks starting at week 0 through week 40. Quantification of KRT8 and KRT14 (markers of luminal and basal cells, respectively) were used to test cell plasticity. Quantification of CD24-/CD44+ and ALDH1A3 expression (markers of cancer stem cells in breast cancer) were used to test stemness. RNA sequencing and differential gene expression analysis were also performed from samples collected every 10 weeks and gene expression patterns were analyzed via gene set enrichment and clustering analysis. An overview of the aim can be found in **Figure 4.2**.

Methods

MCF10A Cell Culture

Non-tumorigenic ER- breast epithelial cells, MCF10As were used for the 40-week experiment. The growth media for MCF10A cells include: DMEM/F12 (Thermo, Cat. # 11320-033), HEPES (1M) (Fisher, Cat. # 15630106), Horse Serum (Gibco, Cat. # 16050122), Insulin (4 mg/mL) (Thermo, Cat. #12585014), Hydrocortisone (96 ug/mL) (StemCell, Cat # 07925), Cholera Toxin (Sigma-Aldrich, Cat. # C8052), Human recombinant epidermal growth factor (EGF) (StemCell, Cat. # 78006.1). MCF10A cells are incubated at 37°C at 5% CO_2 and have about a 24-hour doubling time and are split at 70-90% confluency (Bessette et al. 2015).

40-Week Low Dose Cadmium Exposure

Figure 4.3 shows the workflow schematic of each biological replicate for the 40-week cadmium exposure. Biological replicate 1 (B1) was initiated 2/14/22 using passage 104 MCF10A cells from Colacino lab cell cryobank. Biological replicate 2 (B2) was initiated 3/1/22 using passage 105 from Colacino lab cell cryobank. Biological replicate 3 (B3) was initiated 7/14/22 using passage 111 from Colacino lab cell cryobank.

MCF1A cells were exposed to two low dose concentrations of Cadmium Chloride (henceforth referred to as $CdCl_2$)(Sigma Aldrich, Cat. # 202908) via cell culture media. Low dose (0.25 μM) $CdCl_2$ and high dose (2.5 μM) $CdCl_2$ exposure continued for the duration of the 40-week exposure, alongside a control with no exposure. All cells were dosed with $CdCl_2$ 24-hours after initial plating then were kept in $CdCl_2$ dosed media for all 40 weeks. All cells were split when cells reached between 70 and 90% confluence (averaging ~3 days between each split). Cells were frozen back and stored in liquid

nitrogen every other split to ensure cells are cryopreserved across the 40-week timeline allowing for analysis of all timepoints simultaneously after completion of exposure.

Immunofluorescence - Keratins and Stemness Assays

After the completion of all 40 weeks of the long-term exposure, cells from week 10, week 20, week 30, week 40 and a plate control were thawed and plated in tissue culture flasks to be grown up. At ~70% confluency, the cells were then split to be plated into a 384 well plate at 750 cells per well (**Figure 4.4**) and incubated for 48 hours before staining. A spreadsheet including the media components, regents and concentration of antibodies used in the staining protocols are included in **supplementary table 4.1**.

Keratin Immunofluorescence Assay (KRT 8/14): After 48 hours of incubation, cells are washed using phosphate buffered saline (PBS) (Fisher, Cat. # 20-012-050), then fixed using 4% paraformaldehyde(Fisher, Cat. # AA433689M) PBS for 10 minutes. Cells were again washed with PBS one time, then permeabilized using 0.1% Triton-X 100 (Sigma Aldrich, Cat. # T8787) for 15 minutes. Cells were washed with PBS two times, then blocked with 1% bovine serum albumin fraction (BSA) (Fisher, Cat. # 50-121-5315) + Glycine (Thermo Scientific, Cat. # AAA1381636) in PBS + Tween (PBST) (Fisher Bioreagents, Cat. # BP337). After removing blocking buffer, primary antibodies, recombinant anti-cytokeratin 8 antibody (EP1628Y) (Abcam, Cat # ab53280) and recombinant anti-cytokeratin 14 antibody (EP1612Y), are added then incubated at 4°C for 8-12 hours. Finally, after incubation, cells were washed three times with PBST, counterstained with the nuclear stain Hoechst 33342 (Fisher Scientific, Cat. # H3570) for 1 hour, then washed another three times using PBST, and imaged using the Cell

Insight CX5 High Content Screening (HCS) Platform (Thermo Fisher Scientific, Cat. # CX51110).

Stemness Immunofluorescence Assay (CD24/44 and ALDH1A3): After 48 hours of incubation, cells are washed using PBS and fixed using 4% paraformaldehyde prepared in PBS for 10 minutes. Cells were again washed with PBS one time, then permeabilized using 0.1% Triton-X 100 for 15 minutes, washed with PBS two times, and then blocked with 1% BSA in PBS. Next, primary antibody solution containing CD24 Antibody (SN3) Alexa Fluor 647 (Santa Cruz, Cat. # sc-19585 AF647), CD44 Antibody Alexa Fluor 488 (Biolegend, Cat. # 103016), and Anti-ALD1A3 Polyclonal Antibody (Abcam, Cat. # ab129815) in PBS are added and incubated at 4°C 8-12 hours. After the incubation, cells are washed using PBST, then a secondary antibody solution containing Goat Anti-Rabbit IgG H&L Alexa Fluor 594 (Abcam, Cat. # ab150080) in PBST is added and incubated for 1 hour. Finally, after secondary antibody incubation, cells are washed three times using PBST then a counterstain containing Hoechst 33342 is added and left to incubate for 30 minutes before washing three more times with PBST then imaging using the Cell Insight CX5 High Content Screening (HCS) Platform.

Cell Profiler and Cell Analyst

Raw image data collected from the CX5 are then input into CellProfiler for image QC, illumination and analysis. CellProfiler and CellProfiler Analyst programs extract quantitative morphometric data from microscopy images of cells to identify biologically relevant morphologic and subtle phenotypic changes among samples. CellProfiler is an open-source software tool used to quantify data from biological images acquired

through high-throughput experiments. CellProfiler measures cellular phenotypes such as size, shape, intensity, and texture of each individual cell from every image allowing for an in depth and robust analysis of high-content screening (Carpenter and Jones, 2008). Pipelines specific to measuring Keratin 8/14, CD24/44 and ALDH1A3 were created and can be found in **Figure 4.5**.

Immunofluorescence Data Analysis

CellProfiler analysis pipeline produced detailed spreadsheets including all standard features in addition to more complex shape and texture features. These spreadsheets were uploaded into R studio and columns including important metadata, image number and intensity measurements for KRT8, KRT14, CD24, CD44, ALDH1A3, and Hoechst, were subset into unique data frames. Average intensity of each marker was collapsed by well for the plate control and weeks 10, 20, 30, and 40 for exposures 0 (control), 0.25 μM, and 2.5 μM. Comparisons between each dose within a week and control was performed using a two-sided t-test in R. Statistical significance was accepted with $p<0.05$ and is designated in figures using *; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$.

In order to investigate the proportion of cells fluorescing either both K8 and K14 or CD24 and CD44, classifications of individual cell intensity were created based on the median intensity of each marker. For keratin staining, cells with higher than the median intensity of KRT8 and KRT 14 were identified as “K8 High” or “K14 High”. Cells with lower than the median intensity were classified as “K8 Low” or “K14 Low”. Classes were then created to identify “Hybrid (K8 high/ K14 High)”, “K8 high/ K14 Low”, “K8 Low/ K14

High", and "K8 Low/ K14 Low". The proportion of each class was determined for each dose of weeks 10, 20, 30 and 40, for each biological replicate, 1, 2, and 3.

DNA/RNA Extractions

Cells from weeks 10, 20, 30, 40 and a plate control from each of the biological replicates were thawed and plated in T-75 tissue treated flasks. Cells dosed with 2.5 μM Cd took about 24 hours longer to reach confluency compared to controls and 0.25 μM Cd cells. Cells were collected between 70-90% confluency and lysed in 600 μL of 1% β-mercaptoethanol (Thermo, Cat. #21985023) diluted in Buffer RLT solution (Qiagen, Cat. #79216). Lysed samples were then homogenized using QIAshredder columns (Qiagen, Cat. #79656) prior to DNA and RNA extraction using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen, Cat. #80004). DNA and RNA were stored at -80°C until further use.

RNA Quantification

RNA concentration was determined using the Invitrogen Quant IT Ribogreen RNA assay Kit (Thermo Fisher, Waltham, US). Standards ranging from 20 ng/mL to 1ug/mL were prepared using the Ribosomal RNA standard (100 ug/mL) provided in the kit. 2uL of test RNA was diluted in 18uL of 1X TE buffer. 1uL of the prepared RNA sample was diluted to 99uL of 1X TE buffer to get a 1:100 dilution. A 200-fold aqueous Quant-iT Ribogreen reagent was prepared in 1X-TE buffer for a volume of 100uL/mL. The standards and samples were plated in a black bottom clear 96 well plate (Corning, New York, US) using a 125-1250uL multi-channel pipette. The samples were incubated with the Ribogreen reagent for 5 minutes, protected from light. The fluorescence of Quant-iT Ribogreen solution was read on SpectraMax M5e microplate reader

(Molecular Devices, San Jose, CA). The pre-set protocol on SoftMaxPro software version 5.4 for Ribogreen Assay for Nucleic acid was used for analysis. 10ng/uL sample dilution was prepared for each sample, prior to cDNA preparation.

PlexWell cDNA Preparation and Quantification

The manufacturer's protocol for plexWell sequencing was followed to complete cDNA preparation. Briefly, 1uL of 10 ng/uL RNA dilution were added to a 96 well PCR plate (Dot Scientific, Burton, MI, USA) for oligoDT annealing. cDNA was amplified for 12 PCR cycles in the C1000 Thermal Cycler (BioRad, California, USA) and cDNA was purified using an equivalent amount of MAGwise Paramagnetic beads. cDNA was allowed to bind to the beads for 5mins. The plate was placed on a 96 well plate magnet to allow the beads to pellet on one side of the well. The supernatant was removed, and the beads were washed once using 80% ethanol. The cDNA was eluted using 20uL of 10mM Tris solution. Purified cDNA was stored at -20^{o}C for short term storage.

As per plexWell protocol, cDNA concentrations for all samples were determined using the Quant-iT Picogreen dsDNA assay kit (Thermo Fisher, Waltham, US). Standards ranging from 25pg/mL to 25ng/mL were prepared using the Lambda Standard DNA (100 ug/mL). 1uL of prepped cDNA was diluted in 99uL of 1X-TE buffer to get a 1:100 dilution. A 200-fold Quant-iT Picogreen solution was prepared in 1X TE buffer at a volume of 100uL/well. The samples and standards were plated in a flat bottom Corning 96 well plate using a multi-channel pipette. Picogreen solution was added at 100uL/well and samples were incubated at room temperature for 5 minutes, protected from light. The fluorescence was read on the SpectraMax M5e microplate reader (Molecular

Devices, San Jose, CA). The pre-set protocol on SoftMaxPro software version 5.4 for Picogreen assay for Nucleic acid was used for analysis.

The protocol allows a flexibility of using a 5ng to 25ng range of input cDNA for library prep. A manufacturer provided global dilution factor calculator was used to determine the final volume of cDNA to use to ensure a minimum input of 5ng of cDNA. As per plexWell protocol, 6 samples, per cDNA prep, were analyzed on Agilent High Sensitivity DNA Bioanalyzer at the University of Michigan Advanced Genomics Core. The electropherogram for submitted samples detected the summary of fragment sizes. These graphs were compared to example electropherograms provided in the plexWell protocol to check for fragment size discrepancies.

Library Preparation

Post global dilution factor calculation and dilution, 6uL of cDNA at approximately 1.7ng/uL was used for library preparation. The cDNA was added to hard skirted Sample barcode plate provided in the plexWell LP384 Library Preparation Kit (SeqWell, Beverly, MA, USA). Each sample was labeled with a i7 index, also called Sample Barcode, via a tagmentation reaction. Post i7 tagging, the 18ul of each sample were pooled in 2 pools, containing 48 samples each to a final volume of 850uL. An equivalent volume of MAGwise paramagnetic beads was added and cDNA was allowed to bind for 5 minutes. The tubes were placed on a magnetic stand to allow the beads to form a pellet. The pellet was washed two times with 80% ethanol and cDNA was eluted with 40uL of 10mM Tris. Post pooling, Picogreen quantification was completed using the above-mentioned protocol to confirm that the Sample Barcoded eluate was within the protocol recommended ranges. The 2 pools were then labeled with a i5 index, also called Pool

barcode, using a tagmentation reaction. The i5 tagged pools were purified using an equivalent amount of MAGwise paramagnetic beads. The cDNA was allowed to bind for 5 minutes, and the beads formed a pellet when placed on a magnet. The bead pellets were washed with 80% ethanol two times. The cDNA was eluted with 24uL of 10mM Tris. Purified Pool barcoded products were amplified on the C1000 thermal cycler for 8 cycles. After amplification, the products were diluted to 205uL using 10mM Tris. 5uL of unpurified products was stored as control. 200uL of the diluted Pool barcoded product was purified using 0.8 equivalents of the MAGwise paramagnetic beads. cDNA was allowed to bind to the beads for 5 mins. The beads form a pellet when placed on a magnet. The pellet was washed 2 two times with 80% ethanol and the purified multiplexed library was eluted out with 30uL of 10mM Tris. Purified libraries are stored in -20oC for short term storage. Library QC was done on the Agilent Bioanalyzer (High Sensitivity DNA 5000 kit) at the Advanced Genomics core.

SeqWell and RNA Sequencing

Samples were submitted to the University of Michigan Advanced Genomics Core. Libraries were run on Nova seq shared flow cell at 25% of a flow cell (approx 2 billion total reads) at the PE150 on a 300 cycle S4 flow cell.

After RNA seq data was returned, the data was trimmed and aligned to the human genome hg19 build using STAR (https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html) (Dobin et al. 2013). After alignment QC, quantification was performed using FeatureCounts (Liao et al. 2014). This resulted in the count matrix which would be used for differential expression analysis.

Combat - Batch Correction

The Bioconductor package, Combat, was used to adjust for batch effects for B1, B2, and B3 in R Studio (Johnson et al. 2007). Combat returns an expression matrix that has been corrected for batch effects. PCA plots are used to show similarities between the three biological replicates before and after Combat correction (**Figure 4.6**).

Differential Gene Expression and Genes of Interest

The Bioconductor package, EdgeR, was used to examine RNA-seq data to evaluate differential gene expression (DGE) between different conditions. Batch corrected count data from Combat was used to create a list-based object, DGEList, to be further manipulated in R. We then created a design matrix to fit a statistical model in which we use “Week” and “Dose” as the predictors. Lowly expressed genes were filtered out to prepare our data for regression analysis. Next, we calculated the dispersion parameters for each gene to account for the biological variability in the data. Accurate estimation of dispersion is crucial for reliable differential expression analysis; any over- or underestimation of dispersion could lead to false positives or false negatives downstream. This then prepares our RNA-seq data for DGE analysis. Finally, we can subset our count data and metadata for samples with a specific dose to then analyze for differentially expressed genes. First, we create a design matrix for modeling, then ‘fit’ a generalized linear model to the specific RNA-seq count data we are interested in. Last, we perform an F-test to assess the significance of our selected coefficient in the fitted GLM, resulting in test statistics and p-values for each gene indicating the statistical differential expression associated with ‘Week’. We can then identify genes whose expression levels are significantly affected by our time variable, ‘Week’.

To investigate those genes involved in EMT and stemness, we generated a list of 94 genes known to be associated with EMT and stemness markers (Colacino et al. 2018). A line plot indicating the expression over time for weeks 10, 20, 30, 40 and at baseline was created for the 76 of these 94 markers which were present in the expression data.

Gene Clusters and Enrichr Pathway Analysis

Batch corrected counts data were loaded into *Clust*, a method which automatically extracts optimal co-expressed gene clusters (Basel Abu-Jamous and Steven Kelly, 2018). *Clust* was run using the Python package version 1.18.0 (2022). *Clust* generates cluster profiles and a spreadsheet of cluster objects which contains the list of genes which represents each cluster.

Gene lists from each cluster in the cluster objects spreadsheet were then input into a gene enrichment tool, Enrichr (https://maayanlab.cloud/Enrichr/) (Chen et al. 2013; Kuleshov et al. 2016; Xie et al. 2021). Enrichr is a publicly available tool which analyzes gene sets and provides enrichment results. Genes from clusters C0, C1, C2, C3 and C4, were each put into Enrichr gene query. Enrichr compares our gene list against gene libraries under eight categories, “Pathways”, “Ontologies”, “Diseases/Drugs”, “Cell Types”, “Misc”, “Legacy”, and “Crowd”. Each category contains a grid of gene libraries which then display the top 5 enriched terms from each library for our gene list in a bar plot. The bar plots are sorted by significance (p-value) based on color and length, with the longer and lighter the red bar, the more significant the pathway. The Enrichr gene set libraries we investigated included “ENCODE and ChEA

Consensus TFs from ChIP-X", "Reactome 2022", "KEGG 2021 Human", "MSigDB Hallmark 2020", and "GO Biological Processes 2023".

Results

Keratins and Stemness Immunofluorescence

MCF10A cells were treated with two doses, 0.25 μM Cd and 2.5 μM Cd, for 40 weeks to profile the cell's morphological state and acquisition of cancer stem cell-like properties over time. After immunostaining, images obtained by the CX5 microscope were analyzed for luminal (KRT8) and myoepithelial (KRT14) expression. **Figure 4.7** shows the average intensity of Keratin 8 and 14 over 40 weeks by dose for each biological replicate. The baseline represents cells with the same passage as the initial 40-week cells. Figures 4.7a and 4.7d show the luminal marker, KRT8, decreasing over time in the controls, and the basal marker, KRT14, increasing over time in the controls. For B1 and B2, Cells dosed with 0.25 μM Cd and 2.5 μM Cd follow similar trends seen in KRT8 controls, however, for the KRT14 marker, they follow a similar trend as controls for week 10 and 20 then see a dramatic decrease in intensity of KRT14 at week 30 and 40. Although we see a consistent overall trend in controls for B1,B2, and B3, decreasing KRT8 markers and increasing KRT14 markers, we observe differences in the low and high doses of Cd in B3. B3 shows a steady increase of KRT14 for both low and high doses exceeding the intensity of KRT14 in controls (Figure 4.7f).

Figure 4.8 and 4.9 shows the average intensity of stemness markers CD24, CD44 and ALDH1A3. Across the three biological replicates, no specific pattern emerges over the 40 weeks, however, B1 and B3 both show significant decreases in CD24 at

week 30 compared to week 30 control (Figure 4.8a & c). B1 shows a consistent increase in CD44 for controls over time and a significant decrease of CD44 at week 30 and 40 for both the low and high doses of Cd. Interestingly, both B2 and B3 see a decrease in CD44 for controls over time, however, B2 sees an increase in CD44 expression in both the high and low dosed Cd cells while B3 dosed cells follow the controls with CD44 expression decrease overtime. The other stemness marker, ALDH1A3, increases in expression for all controls over time across all replicates (Figure 4.9). Dosed cells from B1 and B2 show an early increase of expression in week 10 and 20 then levels out with the control's expression over time, while B3 shows a significant decrease in ALDH1A3 expression at week 30 and 40.

Keratin Hybrid Cells

Individual cells were categorized by their KRT8 and KRT14 median intensity as KRT8 +, KRT14+, KRT8+/KRT14+ or keratin 8-/14- . **Figure 4.10** breaks down the proportion of keratins by week and dose for each biological replicate. Week 10 controls all show ~12-15% hybrid (KRT8+/KRT14+) cells (shown in red) for all biological replicates. While the largest proportion of cells are indicated to be KRT8+/KRT14- (green) for B2 and B3, B1 shows KRT8-/KT14+ (blue) cells with the largest proportion (**Figure 4.10a**). Interestingly, all three replicates show ~12-15% of the final classification, KRT8-/KRT14- (purple). After 10 weeks of exposure to 0.25 μM Cd, we see an increased percentage of hybrid cells for B1 and B2 in combination with an increase in KRT8-/KRT14+ cells. Surprisingly, we see a very small proportion of hybrids in B3 after 10-week exposure to 0.25 μM Cd, with a very large proportion of KRT8+/KRT14- cells. After 10 weeks of 2.5 μM Cd exposure, we see a much larger

percentage of hybrids across all three biological replicates compared to the controls. In B1 and B2, we observe the other largest proportion of cells to be KRT8-/KRT14+, while B3 contains more KRT8+/KRT14- cells. Interestingly, B3 seems to maintain a significant proportion of KRT8-/KRT14- across doses compared to B1 and B2. **Figure 4.10b** shows keratin hybrid proportions after 20 weeks for each dose and all three biological replicates. Week 20 controls for B1 and B2 indicate that slightly over 25% of the cells are hybrid and more than 50% of the cells are KRT8-/KRT14+. B3 controls show a very large percentage of KRT8-/KRT14- compared to hybrids followed by KRT8-/KRT14+ as the second largest proportion. Week 20 0.25 μM Cd exposure shows an increase in hybrid cells for all three biological replicates, while proportions of KRT8-/KRT14+ remain to be a large population of the cells. B3 indicates an increase of KRT8+/KRT14- compared to controls and B1 shows an increase in KRT8-/KRT14- compared to controls. Week 20 2.5 μM Cd cells maintain a similar pattern to 0.25 μM Cd cells. **Figure 4.10c** shows keratin hybrid proportions after 30 weeks for each dose and all three biological replicates. Week 30 controls indicate that the predominant two populations are the hybrids and KRT8-/KRT14+ cells for B1 and B2. B3 indicates nearly equal proportions of each classification of cell. Week 30 0.25 μM Cd cells show a decrease in the percentage of hybrid cells compared to control for B2 and B3, while B1 shows a slight increase of hybrid cells. Week 30 2.5 μM Cd shows dramatically different proportions of cell classifications compared to both control and 0.25 μM Cd. The B1 hybrid population decreased while the populations of KRT8+/KRT14- and KRT8-/KRT14- increased significantly. Surprisingly, KRT8-/KRT14+ almost completely disappeared in the B3 week 30 2.5 μM Cd cells. While the percentage of hybrid cells

slightly decreased in B2, we were surprised to see the large increase of KRT8-/KRT14-, similar to B1. B3 had the largest population of hybrid cells for week 30 2.5 μM Cd at nearly 50% increasing significantly compared to B3 control cells. Finally, **figure 4.10d** shows keratin hybrid proportions after 40 weeks for each dose and all three biological replicates. The hybrid population represents about 25% of the cells across all three biological replicates. KRT8-/KRT14+ are nearly the only other populations of cells in B1 and B2 week 40 control cells. B3 on the other hand shows almost equal populations of hybrids and KRT8-/KRT14- cells, with the largest percentage of cells being KRT8-/KRT14+. Week 40 0.25 μM Cd cells show a similar pattern to the control cells with exception to a slight increase in hybrid cells for B1 and slight decrease in hybrid cells for B2 and B3. Week 40 2.5 μM Cd cells are dramatically different from the control and 0.25 μM Cd cells. The largest populations of cells in B1 are shown to be KRT8-/KRT14- and KRT8+/KRT14-. Surprisingly, there is almost no population of hybrids in B1 week 40 2.5 μM Cd cells. B2 follows a similar pattern as B1, with an extreme decrease in hybrid cells and extreme increase in both KRT8+/KRT14- and KRT8-/KRT14- cells. B3 cells actually maintain a similar pattern to both control and 0.25 μM Cd cells. Over time, starting at 20 weeks, we observe a consistent pattern of the proportion of hybrid and KRT8+/KRT14- cells in the controls. Our 2.5 μM cells observe a decrease in KRT8+/KRT14+ hybrid cells over time, with an increase in KRT8-/KRT14- cells.

Stemness Populations

40-week immunostained cells were characterized by their median intensity as “CD24/CD44 High” (red), “CD24/CD44 Low” (purple), “CD24 High” (green) and “CD44 High” (blue). Those cells that are “CD44 High” and CD24 Low (blue) are the canonical

breast stem cell population. **Figure 4.11** breaks down the proportion of CD24 and CD44 expressing cells by week and dose for each biological replicate. **Figure 4.11a** shows CD24/CD44 proportions after 10 weeks for each dose and all three biological replicates. B1 and B3 week 10 controls show a similar pattern with “CD24/CD44 High" as the largest proportion of cells, while the largest proportion of cells in B2 are “CD24/CD44 Low”. After 10 weeks of 0.25 μM Cd exposure, “CD24/CD44 High" and “CD24 High” populations increased in B1, while “CD44 High” and increased in both B2 and B3 compared to controls. 10-week 2.5 μM B1 cells show a consistent pattern to controls, while B2 and B3 cells indicate an increase in “CD24 High” cells. Additionally, B2 cells show a significant increase in “CD24/CD44 High” cells and decrease in “CD24/CD44 Low” cells compared to controls. **Figure 4.11b** shows CD24/CD44 proportions after 20 weeks for each dose and all three biological replicates. All biological replicates show similar patterns of CD24/CD44 proportions across controls. Week 20 0.25 μM Cd cells all show an increase in “CD24 High” cells compared to controls. Week 20 2.5 μM Cd cells follow consistent patterns to 0.25 μM cells with an increase in “CD24 High” cells. **Figure 4.11c** shows CD24/CD44 proportions after 30 weeks for each dose and all three biological replicates. Week 30 controls see a drastic change from week 20 controls. “CD24/CD44 High” cells have the highest percentage of cells followed by “CD44 High” cells. Interestingly, we see a consistent pattern of “CD24 High” cells in 0.25 μM Cd compared to control with a decrease in “CD24/CD44 High” cells and an increase in “CD24/CD44 Low” cells. Week 30 2.5 μM Cd cells show and increase in “CD44 High” and “CD24/CD44 Low” cells for B1 and B3, while B2 shows an increase in “CD24/CD44 High” cells. **Figure 4.11d** shows CD24/CD44 proportions after 40 weeks for each dose

and all three biological replicates. Here we see the most variation over the time course, with the highest proportion of cells in B2 controls indicated as “CD24 High”, while the highest proportion of cells in B1 and B3 controls are “CD24/CD44 High”. Week 40 0.25 μM Cd shows a decrease in “CD44 High” expressing cells for B1 and an increase in “CD44 High” expressing cells in B3 compared to controls. Finally, week 40 2.5 μM Cd indicates an increase in “CD44 High” expressing cells for B2, and a decrease in “CD44 High” expressing cells in B1 and B3.

Differential Gene Expression of 40- Week Cadmium Exposure

RNA extractions from each condition for weeks 10, 20, 30, 40, and a plate control for all three biological replicates resulted in RNA sequencing of 39 samples total (13 samples from each biological replicate). **Figure 4.12** shows the number of genes differentially expressed in each direction for time and dose. **Figure 4.12a** shows the number and direction of differentially expressed genes in the controls from each week vs the plate control. **Table 4.1.1** shows the exact number of genes up-regulated and down-regulated for the controls of each week vs the plate control. Overall, week 10 has the smallest amount of differentially up and down regulated genes and week 40 has the largest number of differentially expressed genes. There are more down-regulated genes than up-regulated genes. **Figure 4.12b** shows the number and direction of differentially expressed genes in the 0.25 μM Cd dosed cells from each week vs their respective controls. **Table 4.1.2** shows the exact number of genes up-regulated and down-regulated for the low dose, 0.25 μM Cd, cells from each week vs their respective controls. There are significantly few differentially expressed genes in 0.25 μM Cd dosed cells vs their controls compared to the controls vs the plate controls. Week 20 has the

most down-regulated genes while week 30 and 40 have the most up-regulated genes. 0.25 μM Cd exposure down-regulates more genes than it upregulates. **Figure 4.12c** shows the number and direction of differentially expressed genes in all 2.5 μM Cd dosed cells from each week vs their respective controls. **Table 4.1.3** shows the exact number of genes up-regulated and down-regulated for the low dose, 2.5 μM Cd, cells from each week vs their respective controls. Week 40 shows the most differentially expressed genes for 2.5 μM Cd compared to its control. Interestingly, 2.5 μM Cd has more upregulated genes than down regulated genes, opposite of controls over time and 0.25 μM Cd over time.

Groupings of gene markers for “Stemness”, “EMT”, “Cell Adhesion”, “piRNA Pathway”, and “Inflammatory mediators” were created by performing an in-depth literature search. These lists were then analyzed to see how many, if any, of these genes changed over 40-weeks exposure to low dose cadmium. After 40 weeks of exposure to 2.5 μM Cd, stemness markers WNT5A, ALDH1A3, TGFB2, ABCG2, SNAI1, and SOX4 all showed upregulation compared to Week 40 control cells. Additionally, gene markers of cell adhesion including COL1A1, ICAM1, CDH2, and MMP2 were also upregulated in the 40-week 2.5 μM Cd cells compared to the 40-week control cells. Only one marker of the piRNA pathway, TDRKH, was shown to increase in expression for the 40-week 2.5 μM Cd compared to the 40-week control cells. Surprisingly, there were eight gene markers of EMT, WNT5A, KRT8, CDH2, KRT18, FOXC2, TGFB2, SNAI1, and MMP2, that were shown to increase after 40-weeks of 2.5 μM Cd compared to 40-week control cells. By grouping gene markers into specific

categories, we are able to better analyze specific functions. Gene lists and line plots for additional gene categories are available in. **supplementary figures 4.2-6.**

In order to visualize the expression data of individual genes over time, we highlighted 76 genes of interest (see in methods) which have been shown to play a role in EMT and stemness and created line plots for each individual gene's expression over time. **Figure 4.13** shows seven of these line plots, KRT8, KRT14, CD24, CD44, ALDH1A3, VIM, and CDH1. We looked at these genes to compare the RNA data to the protein expression data seen in figures 7, 8, and 9, and added VIM and CDH1 to analyze mesenchymal and epithelial changes in gene expression, respectively. The rest of the line plots are included in **supplemental Figure 4.1**. In figure 4.13, the y-axis indicates counts per million (CPM) and the x-axis is time. Doses are labeled in the legend, 0 μM Cd is green, 0.25 μM Cd is red and 2.5 μM Cd is blue. Figure 4.13a, we see CPM for KRT8 starts over 650 for the plate control and decreases to 400 CPM for control and 0.25 μM and 300 for 2.5 μM. Control and 0.25 μM follow a similar pattern of decreasing W10 to W20 with a slight increase to W30 and finally a steep drop of expression at W40. Conversely, 2.5 μM slightly decreases W10 to W20 then increases from W20 to W40. Figure 4.13b shows CPM for KRT14 starting low for the plate control at ~900, with an increase to ~3,000 by week 10 for control cells, which continue to increase through week 20 peaking at a little over 8,000 CPM. At W10, both doses are higher than the control cells at ~4,000 CPM, both increase to W20 at slightly under 6,000 CPM. At W20, 0.25 μM and 2.5 μM diverge with 0.25 μM continuing to increase through W40, while 2.5 μM decreases significantly from W20 to W40 finishing at around 2,000 CPM. Figure 4.13c shows CD24 expression data, with plate control starting at 70

CPM decreasing to 40 CPM by W40 showing a slight increase at W20. Both doses show high variability over the 40 weeks, although they follow a similar pattern. Both doses start with a high expression than control at W10, increase in expression to ~80 CPM at W20 before steep decreases to ~55 CPM for 0.25 μM and 35 CPM for 2.5 μM. From W30 to W40, both doses show steep increases of expression with 0.25 μM ending at ~85 CPM and 2.5 μM ending at 70 CPM. Figure 4.13d shows CD44 expression data. Controls start at 750 CPM for plate control, then increase to 1100 CPM for W20 and decreases through W30 to end at ~950 CPM. 0.25 μM starts a little above 950 CPM at W10 then increases at W20 and holds expression around 1050 CPM through to W40. 2.5 μM starts around control and 0.25 μM at W10 then consistently decreases to 850 CPM by W40. Figure 4.13e shows ALDH1A3 expression data. Controls see a slight increase to W10 with a steep decrease to W20 followed by a constant increase to W40 ending around 90 CPM. Both doses start around 50 CPM, hold constant expression through week 30 where they diverge, with 2.5 μM seeing a steep increase to above 100 CPM and 0.25 μM seeing a decrease to ~45 CPM. Figure 4.13f shows expression data for VIM. Controls, 0.25 μM and 2.5 μM follow very similar patterns; VIM starts over 1000 for the plate control and decreases to 800 CPM by week 10.VIM expression drops again for all doses from week 10 to week 20 then slightly increases for week 30, then slightly decreases again for week 40. Overall, we can see that VIM expression decreases over the 40-week time course, however it experiences an increase of expression at week 30. Figure 4.13g shows expression data for CDH1. Controls start around 62 CPM, increase slightly to W10, then drops to 55 at W30 followed by an increase to 75 by W40. 2.5 μM

follows a similar pattern as control, whereas 0.25 μM sees a steep increase to 95 CPM at W20, drops to ~60 CPM at W30 then ends slightly lower than controls at ~70 CPM.

Additional line plots for the groupings of gene markers for “Stemness”, “EMT”, “Cell Adhesion”, “piRNA Pathway” and “Inflammatory Mediators” are included in supplemental figures 4.2-6. In the stemness gene grouping, we observe different patterns of expression data (CPM) for most of the genes over the 40-week time course. Many of the genes including SOX4, TGFB2, SOX9, SOX12, WNT5A, CTNNB1, and CD44, show a similar expression pattern between control and 0.25 μM Cd, where as NOTCH1, SNAI2, WNT3, WNT5A, ABCG2, and CTNNB1 for all three conditions, 0.25 μM Cd, 2.5 μM Cd, and control, while 2.5 μM Cd expression pattern is relatively different after 40-weeks. A few of the genes including ZEB2, NOTCH2, ZEB1, and WNT2B show the final CPM at week 40 to be fairly close between all three conditions. Six stemness genes were significantly different in the 40-week 2.5 μM Cd vs the 40-week control including WNT5A, ALDH1A3, SOX4, TGFB2, ABCG2, and SNAI1.

The cell adhesion gene grouping showed some widely different expression patterns for COL1A1, DSC3, COL4A1, CDH3, CLDN1, ITGB3BP, MMP2, and EPCAM. For most cell adhesion genes, 2.5 μM Cd expression patterns were much different compared to the control and 0.25 μM Cd expression patterns. Expression patterns for genes LAMA3, IGF1R, and ITGA6 followed a close trend between all three conditions, however, ended if slightly different CPM numbers at the end of 40 weeks. Four genes from the cell adhesion grouping were significantly different in the 40-week 2.5 μM Cd vs the 40-week control including COL1A1, ICAM1, CDH2, and MMP2. Of interest from the EMT gene grouping, genes KRT17, KRT14, MET, EPCAM, and CD44 showed widely

different expression patterns of the cadmium dosed cells compared to the control cells. Eight genes involved in EMT were significantly different in the 40-week 2.5 μM Cd vs the 40-week control including WNT5A, KRT8, KRT18, CDH2, FOXC2, TGFB2, SNAI1 and MMP2. In the inflammatory gene group, CCL20, CXCL1, TNFAIP1, TGFB1, and STAT3 all showed very different expression patterns among the three conditions, however there were no significantly different genes in the 40-week 2.5 μM Cd vs the 40-week control.

Our piRNA pathway gene grouping indicated the differential expression patterns of HSP90B1, SND1, TDRKH, PRMT5, AGPAT2, and PLD6. Interestingly, PRMT5 and AGPAT2 show very different expression patterns between the Cd dosed cells and the controls. Both 0.25 μM Cd and 2.5 μM Cd show a higher expression of PRMT5 after 40 weeks compared to the control cells. For AGPAT2, both doses were expressed lower than the controls after 40 weeks. HSP90B1 indicated a higher CPM at W10 than 0.25 μM Cd and the controls, however, expression dropped to under the controls after the 40 weeks. Interestingly, for SND1, the controls and 0.25 μM Cd cells showed a similar expression pattern after the 40 weeks, whereas the 2.5 μM Cd cells zig zagged expression starting high and ultimately ending far below the controls and 0.25 μM Cd cells. One piRNA pathway gene, TDRKH, is differentially expressed between 40-week 2.5 μM Cd vs the 40-week control.

Gene Cluster and Enrichment Analysis

In order to identify co-expressed gene clusters across our three biological replicates, we used *Clust* for cluster extraction of our gene expression data. **Figure 4.14** shows the cluster profile results after running Clust on B1, B2, and B3 raw expression

data. The bottom row shows how each cluster of genes change over time for controls (red), the top row shows how each cluster of genes changes over time for 0.25 μM Cd, and the middle row shows how each cluster of genes changes over time for 2.5 μM Cd. For the 1,381 genes in C0, control and 0.25 μM Cd increase expression at week 10 then steadily hold that increase through the 40 weeks, whereas 2.5 μM Cd increases expression also at week 10, but ends with lower expression at week 40. C1 shows genes in controls and 0.25 μM gradually increases expression over the 40 weeks, with a slight decrease for 0.25 μM Cd at week 30 before ending high again. For 2.5 μM, genes from C1 immediately spike at week 10 then show variability of ending expression over the 40 weeks. C1 genes are more tightly correlated for controls and 0.25 μM Cd than 2.5 μM Cd. Genes in the C2 cluster show a lot of variability. Controls are tightly correlated as they decrease expression by week 10 then jump to higher levels by week 20, drive even higher for week 30 then drop slightly for week 40. C2 0.25 μM Cd clusters show a steady increase over the 40 weeks. C2 2.5 μM Cd clusters increase by week 20 then drop slightly at week 30 and finish with variability at week 40. Genes in C3 controls, 0.25 μM Cd, and 2.5 μM Cd show a similar pattern up to week 30, where controls show a decrease in expression and the dosed cells end the 40 weeks with an increase in those genes' expression. Lastly, those genes in C4 show similarity between the three conditions until week 20, where controls and 0.25 μM Cd continue to decrease ending with low expression, and 2.5 μM Cd shows an increase starting at week 30 with high variability of genes ending with a higher expression at Week 40.

We then input the gene lists per cluster, provided by Clust, into the open resource *Enrichr* to perform gene set enrichment analysis. **Figure 4.15** shows the enrichment

results for genes from cluster C3 (figure 4.15) for the gene sets “ENCODE and ChEA Consensus TFs from ChIP-X”, “GO Biological Process 2023”, and “MSigDB Hallmark 2020”. We used C3 here because the profiles remain very similar through week 30 then drastically change at week 40. The bar plots are sorted by significance (p-value) based on color and length, with the longer and lighter red the bar, the more significant the term or pathway. The most enriched terms for “ENCODE and ChEA Consensus TFs from ChIP-X” include YY1, TAF1, and ATF2. The most enriched terms for “GO Biological Process 2023” include RNA splicing, mRNA Processing, and mRNA Splicing. The most enriched terms for “MSigDB Hallmark 2020” include Myc Targets V1, oxidative phosphorylation, and reactive oxygen species pathway.

Discussion

In this study, we investigate the role of long term (40-week) low dose Cd (0.25 μM and 2.5 μM) exposure on cancer-associated morphological alterations and cellular plasticity. MCF10A cells are exposed to biologically relevant doses, 0.25 μM Cd and 2.5 μM Cd, for 40-weeks and collected at multiple timepoints to profile key features of these cells, including their differentiation state, their acquisition of stem cell-like properties, and their transcriptional profiles. We found that our low dose cadmium (0.25 μM Cd) increased cellular plasticity over 40 weeks defined by the population hybrid KRT8/KRT14 cells. We also found that our high dose (2.5 μM Cd) largely increased cellular plasticity (hybrid KRT8/KRT14 cells) in the first 20 weeks of exposure then indicated a sudden switch to very low populations of hybrid cells with a large population of low expressing KRT8 and low expression KRT14 cells.

Numerous studies have implicated chemical exposures as drivers of breast cancer. Chemicals including bisphenol A (BPA) (Gao et al. 2015), phthalates (Zuccarello et al. 2018), and benzo(a)pyrene (Malik et al. 2018) have each been associated with breast cancer development and progression through mechanisms such as epigenetic alterations and endocrine disruption. Although numerous chemicals have been associated with breast cancer, toxic metals, such as cadmium, have been linked to higher breast cancer risk, but they require further investigation into their mechanisms through which they promote breast cancer. In this study, we perform a long-term low dose cadmium exposure to profile its role in phenotypic plasticity and the development of stemness characteristics.

Keratins and Stemness Immunofluorescence

Using luminal (KRT8) and basal (KRT14) markers, we quantified the effects of prolonged culture, along with the effects of cadmium exposure over time. In Figure 4.7, untreated controls had a steady decrease in KRT8 intensity over time along with a steady increase of K14 intensity over time. This indicates that heterogeneous MCF10A cells express more luminal markers at baseline and over time, with no cadmium exposure, the cells become more basal. With cadmium exposure, we see a slight increase in KRT8 compared to controls, however we see a steep decrease of KRT14 at week 30. This then indicates that long-term, low dose cadmium exposure may be pushing cells into a mesenchymal to epithelial transition (MET) instead of EMT.

Although two of the three biological replicates followed this pattern, the third biological replicate actually exhibited a more basal phenotype, where the cadmium dosed cells decreased in KRT 8 luminal and increased significantly in KRT 14 basal

intensity. This discrepancy between replicates may be due to a number of factors. One might be the higher passage number of B3, which started using p111 cells, whereas B1 and B2 were started with lower passages. Therefore, previous passages of these cells may have undergone some phenotypic changes due to overgrowth or stress.

Next, we quantified CD24/CD44 and ALDH1A3 expression to test stemness. Two non-overlapping populations of cancer stem cells, CD44+/CD24- and ALDH1A3+, have been shown to be highly plastic and may play an important role in metastasis (Liu et al. 2014). In a study by Colacino et al., 2018, normal mammary (mammoplasty tissue) was shown to contain overlapping of these two populations, CD44+/CD24- and ALDH1A3+, and indicated this heterogeneous population had the greatest mammosphere forming potential and expressed higher levels of stemness and EMT-related genes. Our immunofluorescence data from Figure 4.8 showed an overall increase of CD24 expression for controls over time and a non-specific pattern of CD44 alterations. We do see a decrease in CD44 expression for dosed cells over time in B1, however, in B2 we see an increase in CD44. Interestingly, for B1, we don't see a significant decrease in CD44 expression until week 30, whereas for our B2 CD44 expression, we see an immediate significant increase starting at week 10. B3 also sees a decrease in CD44 expression for the dosed cells. Again, we see discrepancies between the batches that makes overall trends difficult to conclude. Figure 4.9 showed an overall trend of increased ALDH1A3 expression in the control cells; however, we observed non-specific expression patterns for ALDH1A3 expression in the dosed cells. B1 and B3 both show ALDH1A3 expression decreasing in 0.25 μM cells, whereas B2 sees an increase. Additionally, we can see a very slight pattern in 2.5 μM cells for B1 and B2 where 2.5

μM shows an overall increase from week 10 to week 20, then a decrease in 30 and 40 weeks.

Keratin Hybrid Cells

We identified phenotypic plasticity through increases in KRT8/KRT14 hybrid populations. Figure 4.10 allowed us to see the proportion of hybrid cells not only over time and from different biological replicates, but also between doses. We can see that week 10 controls contained the fewest hybrids, while the W10 high dose contained the most hybrids compared to the other conditions. This may indicate that the majority of hybrid cells are being produced within the first 10 weeks of the experiment. The proportion of KRT8-/KRT14+ cells increase over time in the control cells and 0.25 μM Cd cells, whereas cells exposed to the high Cd dose contain higher proportions of KRT8+/KRT14- cells. This supports the trend we saw in the immunofluorescence data (Figure 4.7). Additionally, we noticed that B3 starts with a higher proportion of KRT8-/KRT14- cells in both the controls and 0.25 μM Cd exposed cells, however, the proportion of these low/low cells decreases over time in the 2.5 μM Cd cells for B3. Conversely, B1 and B2 start with very little of these low-low cells and increase drastically over time in the 2.5 μM Cd cells. There are a couple factors that may contribute to what these low-low cells are. One may be due to the overall intensity of the individual cells. These cells are designated low-low based on the median intensity of all the cells. If there are populations of cells that are expressing very high intensity vs the other cells, then these low-low cells could just be cells with much lower expression compared to those very high cells. Therefore, if we were to investigate the individual

cells, there is a chance that they do have intensity of KRT8 or 14 just much lower compared to the cells with extremely high intensity. Another possibility is that these cells really don't express (or very little expression) of KRT8 and KRT14. These cells would then have lost their identity, and further investigation into what those cells are would be required.

Stemness Population

We identified stem cell-like cells through populations of high CD44, low CD24 (CD44+/CD24-) cell proportions, which typically identifies stem cells in a mesenchymal phenotype (Colacino et al. 2018; Liu et al. 2014). Figure 4.11 shows the proportion of cells with stem cell like marker expression. We see our highest proportion of CD44+/CD24- (blue) cells in weeks 10 and 20 controls and 0.25 μM then in weeks 30 and 40 2.5 μM Cd. Interestingly, we see a shift of increased proportions of these cells from control and low dose at the beginning of the time course, to the high dose towards the end of the time course. This indicates a possible shift in the cells to a more stemlike state that occurs halfway through the time course. This may also suggest an adaptation of the cells to cadmium over time.

Differential Gene Expression of 40- week Cadmium Exposure

Gene expression data was analyzed after batch correction to allow us to rigorously investigate the effect of cadmium on gene expression over time. Although not shown, differential gene analysis for each control vs the plate control and for each dose vs its own control were performed. In order to get an overall idea of the number of differentially expressed genes, we created up/down plots to see how many genes were down-regulated vs how many were up-regulated, given a specific condition. Figure 4.12

demonstrated the immense effect of long-term cell culture by showing the number of differentially expressed genes in 10 weeks vs the baseline compared to 40 weeks vs the baseline. Additionally, the stark difference in number of differentially expressed genes for 0.25 μM Cd vs 2.5 μM Cd, indicates how much cadmium affects these cells in addition to the time component. Interestingly, the controls and low dose Cd differentially expressed genes seem to be preferentially down-regulated rather than up-regulated, whereas the high dose Cd exposure after 40 weeks shows more up-regulated genes than down regulated genes. This likely indicates different mechanisms, such as gene expression and epigenetic changes, of low dose cadmium exposure vs high dose cadmium exposure.

Due to the large amounts of data, we selected a list of genes of interest to focus on based on stemness, EMT, and luminal to basal transition. Figure 4.13 shows a representative line plots for genes KRT8, KRT14, CD24, CD44, ALDH1A3, VIM, and CDH1, which visualizes the expression pattern for each gene over 40 weeks for all three conditions. The other 75-line plots are available in supplementary figure 4.1. These line plots allow us to track the expression of the gene over time to see when it is most affected. KRT8 expression data over the 40 weeks follows the protein expression data quite closely, whereas the KRT14 expression data sees some major differences for 0.25 μM and controls. Similarly, CD24 and CD44, see some similarities for the doses, however the controls show a different pattern. Because the RNA did undergo batch correction, it makes sense that the RNA expression doesn't follow any of the staining batches perfectly, however, we do see a few overall patterns consistent between the RNA and protein, such as a decrease in KRT8 over time and an overall increase in

KRT14 over time in controls. Additionally, we see an overall slight increase in ALDH1A3 expression in both RNA and protein. Notably, due to the batch correction that the RNA was able to undergo and that the staining data was unable to undergo, we can infer more about what Cd does over time across replicates in the RNA expression data compared to the protein expression data.

Additional line plots for the groupings of gene markers (Supplementary figures 4.2-6) indicated both different and similar gene expression patterns over the 40 weeks. Importantly, the y axis indicated CPM and there are numerous genes with overall lower CPM meaning a lower overall gene expression for that gene. It is important to take that into account when looking at some of the gene patterns in the line plots. These plots allowed us to look at molecular markers more in depth to better understand what may be driving the stemness and proliferative characteristics we see in the 40-week cells. EMT and cell adhesion gene groupings had the most differentially expressed genes indicating their overall role in long term, low dose exposure to cadmium in breast cells.

Interestingly, the gene category, “Inflammatory Mediators”, resulted in no significant changes after 40-week exposure to Cd. This was not as surprising since we are looking at epithelial cells, we would expect more genes changing in stromal cells due to their role as immunoregulators. Stromal cells often interact with immune cells and actively produce chemokines and other inflammatory mediators. Tumor promoting inflammation is a hallmark of cancer and further investigation into how long term, low dose cadmium exposure effects stromal cells would be very interesting.

Gene Cluster and Enrichment Analysis

Clust provides a way to visualize the effect of different gene expression data sets all together by identifying genes which have similar expression patterns over time. Figure 4.14 shows five different gene clusters found across B1, B2 and B3 treatments over weeks. We then put the gene lists provided in each cluster into the publicly available gene set enrichment software Enrichr. Enrichr offers thousands of gene sets, collected in gene set libraries, for gene set enrichment analysis. Here, although we only highlight three libraries, there are numerous gene sets available to investigate comparisons. In Figure 4.15, we use cluster genes from C3 to evaluate targets enriched by our list of genes due to its unique pattern which is closely followed by all three conditions until the last 10 weeks. Our C3 gene list resulted in enrichment of transcription factor YY1 Encode, this means that our gene set highly matched those gene sets known to be regulated by transcription factor YY1. Interestingly, YY1 has been shown to positively regulate transcription in embryonic stem cells and has been implicated in numerous cancers including colon, breast, cervical, bladder and brain (Wang et al. 2018, Hosea et al. 2023). We also see that our gene list is highly similar to those genes involved in RNA and mRNA splicing, indicating that these genes play an important role in the transcription process. mRNA splicing has been shown to be dysregulated in cancer (Bradley and Anczukow, 2023). Finally, our C3 gene list was enriched for the term Myc Targets in gene ontology library “MSigDB Hallmark 2020”. Myc is a very well-studied oncogene known to promote cell proliferation and one of the canonical regulators of embryonic stem cells (Chappell and Dalton 2013). Myc overexpression has been highly reported on, specifically in breast cancer. (Schulze et al. 2020). This indicates that our gene list is very similar to the known subgroup of

genes regulated by Myc. Importantly, Myc is known to sustain/promote proliferative signaling, one of the hallmarks of cancer, indicating that long-term cadmium exposure may be increasing this key characteristic of carcinogenesis.

Conclusion

A previous study by Benbahim-Tallaa et al. 2009, also performed a long-term, 40-week exposure of 2.5 μM cadmium to MCF10A cells. Their results indicated that the chronic exposure to low dose cadmium transformed MCF10A cells to a basal-like phenotype reflected by ER-alpha and HER2 negativity, reduced expression of BRCA1, and increased expression of KRT5 and P63. Interestingly our results indicate that the control cells undergo more of a basal shift compared to the cadmium dosed cells which seem to undergo more of a mesenchymal to epithelial shift (MET). MET has been shown to be important in metastasis and the formation of secondary tumors (Jolly et al. 2016). However, we had two biological replicates that demonstrated for MET while our third replicate did show a bit more of EMT in the immunostaining. Further investigation into the mechanisms behind these different results are required.

Our results show that KRT 8 decreases over time for both controls and low and high Cd doses. KRT 14 increases in controls over time and in both doses decrease diverging from controls at week 30 and week 40. Keratin hybrids emerge around week 10 for both doses of cadmium with the highest percent of hybrids in week 10 highest dose. CD44+/CD24- cells emerge in controls and low dose Cd within 20 weeks, and then in the high dose at week 30 and 40. Gene expression data shows differential expression over time and within doses. The gene set clusters generated from expression data indicate involvement of important transcription factors and pathways

associated with breast cancer. Further, this study demonstrates low-dose cadmium activation of phenotypic plasticity and acquisition of stem cell-like properties, however, additional work is necessary to determine the mechanistic effect of long-term, low dose cadmium exposure and breast cancer.

References

Abu-Jamous, B., & Kelly, S. (2018). Clust: Automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biology*, *19*(1), 172. https://doi.org/10.1186/s13059-018-1536-8

Adams, S. V., Quraishi, S. M., Shafer, M. M., Passarelli, M. N., Freney, E. P., Chlebowski, R. T., Luo, J., Meliker, J. R., Mu, L., Neuhouser, M. L., & Newcomb, P. A. (2014). Dietary Cadmium Exposure and Risk of Breast, Endometrial, and Ovarian Cancer in the Women's Health Initiative. *Environmental Health Perspectives*, *122*(6), 594–600. https://doi.org/10.1289/ehp.1307054

Benbrahim-Tallaa, L., Tokar, E. J., Diwan, B. A., Dill, A. L., Coppin, J.-F., & Waalkes, M. P. (2009). Cadmium Malignantly Transforms Normal Human Breast Epithelial Cells into a Basal-like Phenotype. *Environmental Health Perspectives*, *117*(12), 1847–1852. https://doi.org/10.1289/ehp.0900999

Bradley, R. K., & Anczuków, O. (2023). RNA splicing dysregulation and the hallmarks of cancer. *Nature Reviews Cancer*, *23*(3), 135–155. https://doi.org/10.1038/s41568-022-00541-7

Chappell, J., & Dalton, S. (2013). Roles for MYC in the Establishment and Maintenance of Pluripotency. *Cold Spring Harbor Perspectives in Medicine*, *3*(12). https://doi.org/10.1101/cshperspect.a014381

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128. https://doi.org/10.1186/1471-2105-14-128

Chiche, A., Di-Cicco, A., Sesma-Sanz, L., Bresson, L., de la Grange, P., Glukhova, M. A., Faraldo, M. M., & Deugnier, M.-A. (2019). P53 controls the plasticity of mammary luminal progenitor cells downstream of Met signaling. *Breast Cancer Research*, *21*(1), 13. https://doi.org/10.1186/s13058-019-1101-8

Colacino, J. A., Azizi, E., Brooks, M. D., Harouaka, R., Fouladdel, S., McDermott, S. P., Lee, M., Hill, D., Madden, J., Boerner, J., Cote, M. L., Sartor, M. A., Rozek, L. S., & Wicha, M. S. (2018a). Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*, *10*(5), 1596. https://doi.org/10.1016/j.stemcr.2018.03.001

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15. https://doi.org/10.1093/bioinformatics/bts635

Florez-Garcia, V. A., Guevara-Romero, E. C., Hawkins, M. M., Bautista, L. E., Jenson, T. E., Yu, J., & Kalkbrenner, A. E. (2023). Cadmium exposure and risk of breast cancer: A meta-analysis. *Environmental Research*, *219*, 115109. https://doi.org/10.1016/j.envres.2022.115109

Gallagher, C. M., Chen, J. J., & Kovach, J. S. (2010). Environmental cadmium and breast cancer risk. *Aging (Albany NY)*, *2*(11), 804–814.

Gao, H., Yang, B.-J., Li, N., Feng, L.-M., Shi, X.-Y., Zhao, W.-H., & Liu, S.-J. (2015). Bisphenol A and Hormone-Associated Cancers: Current Progress and Perspectives. *Medicine*, *94*(1). https://doi.org/10.1097/MD.0000000000000211

Hosea, R., Hillary, S., Wu, S., & Kasim, V. (2023). Targeting Transcription Factor YY1 for Cancer Treatment: Current Strategies and Future Directions. *Cancers*, *15*(13), Article 13. https://doi.org/10.3390/cancers15133506

Hsieh, T.-H., Tsai, C.-F., Hsu, C.-Y., Kuo, P.-L., Lee, J.-N., Chai, C.-Y., Wang, S.-C., & Tsai, E.-M. (2012). Phthalates induce proliferation and invasiveness of estrogen receptor-negative breast cancer through the AhR/HDAC6/c-Myc signaling pathway. *The FASEB Journal*, *26*(2), 778–787. https://doi.org/10.1096/fj.11-191742

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jolly, M. K., Tripathi, S. C., Jia, D., Mooney, S. M., Celiktas, M., Hanash, S. M., Mani, S. A., Pienta, K. J., Ben-Jacob, E., & Levine, H. (2016). Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*, *7*(19), 27067–27084. https://doi.org/10.18632/oncotarget.8166

Julin, B., Wolk, A., Bergkvist, L., Bottai, M., & Åkesson, A. (2012). Dietary Cadmium Exposure and Risk of Postmenopausal Breast Cancer: A Population-Based Prospective Cohort Study. *Cancer Research*, *72*(6), 1459–1466. https://doi.org/10.1158/0008-5472.CAN-11-0735

Koual, M., Tomkiewicz, C., Cano-Sancho, G., Antignac, J.-P., Bats, A.-S., & Coumoul, X. (2020). Environmental chemicals, breast cancer progression and drug resistance. *Environmental Health*, *19*(1), 117. https://doi.org/10.1186/s12940-020-00670-2

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(Web Server issue), W90. https://doi.org/10.1093/nar/gkw377

Liang, Z.-Z., Zhu, R.-M., Li, Y.-L., Jiang, H.-M., Li, R.-B., Tang, L.-Y., Wang, Q., & Ren, Z.-F. (2020). Differential epigenetic and transcriptional profile in MCF-7 breast cancer cells exposed to cadmium. *Chemosphere*, *261*, 128148. https://doi.org/10.1016/j.chemosphere.2020.128148

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., Asselin-Labat, M.-L., Gyorki, D. E., Ward, T., Partanen, A., Feleppa, F., Huschtscha, L. I., Thorne, H. J., Fox, S. B., Yan, M., French, J. D., Brown, M. A., Smyth, G. K., Visvader, J. E., & Lindeman, G. J. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature Medicine*, *15*(8), 907–913. https://doi.org/10.1038/nm.2000

Liu, S., Cong, Y., Wang, D., Sun, Y., Deng, L., Liu, Y., Martin-Trevino, R., Shang, L., McDermott, S. P., Landis, M. D., Hong, S., Adams, A., D'Angelo, R., Ginestier, C., Charafe-Jauffret, E., Clouthier, S. G., Birnbaum, D., Wong, S. T., Zhan, M., … Wicha, M. S. (2014). Breast Cancer Stem Cells Transition between Epithelial and

Mesenchymal States Reflective of their Normal Counterparts. *Stem Cell Reports*, *2*(1), 78. https://doi.org/10.1016/j.stemcr.2013.11.009

Malik, D., David, R. M., & Gooderham, N. J. (2018). Mechanistic evidence that benzo[a]pyrene promotes an inflammatory microenvironment that drives the metastatic potential of human mammary cells. *Archives of Toxicology*, *92*(10), 3223. https://doi.org/10.1007/s00204-018-2291-z

Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., & Ramírez-Valdespino, C. A. (2022). Subtypes of Breast Cancer. In H. N. Mayrovitz (Ed.), *Breast Cancer*. Exon Publications. http://www.ncbi.nlm.nih.gov/books/NBK583808/

Ossovskaya, V., Wang, Y., Budoff, A., Xu, Q., Lituev, A., Potapova, O., Vansant, G., Monforte, J., & Daraselia, N. (2011). Exploring Molecular Pathways of Triple-Negative Breast Cancer. *Genes & Cancer*, *2*(9), 870–879. https://doi.org/10.1177/1947601911432496

Parodi, D. A., Greenfield, M., Evans, C., Chichura, A., Alpaugh, A., Williams, J., Cyrus, K. C., & Martin, M. B. (2017). Alteration of Mammary Gland Development and Gene Expression by In Utero Exposure to Cadmium. *International Journal of Molecular Sciences*, *18*(9), 1939. https://doi.org/10.3390/ijms18091939

Piper, M. M., Bob Freeman, Mary. (2017, June 7). *Alignment with STAR*. Introduction to RNA-Seq Using High-Performance Computing - ARCHIVED. https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Rocco, S. A., Koneva, L., Middleton, L. Y. M., Thong, T., Solanki, S., Karram, S., Nambunmee, K., Harris, C., Rozek, L. S., Sartor, M. A., Shah, Y. M., & Colacino,

J. A. (2018). Cadmium Exposure Inhibits Branching Morphogenesis and Causes Alterations Consistent With HIF-1α Inhibition in Human Primary Breast Organoids. *Toxicological Sciences*, *164*(2), 592–602. https://doi.org/10.1093/toxsci/kfy112

Schulze, A., Oshi, M., Endo, I., & Takabe, K. (2020). MYC Targets Scores Are Associated with Cancer Aggressiveness and Poor Survival in ER-Positive Primary and Metastatic Breast Cancer. *International Journal of Molecular Sciences*, *21*(21), 8127. https://doi.org/10.3390/ijms21218127

Strumylaite, L., Kregzdyte, R., Bogusevicius, A., Poskiene, L., Baranauskiene, D., & Pranys, D. (2014). Association between cadmium and breast cancer risk according to estrogen receptor and human epidermal growth factor receptor 2: Epidemiological evidence. *Breast Cancer Research and Treatment*, *145*(1), 225–232. https://doi.org/10.1007/s10549-014-2918-6

Thong, T., Wang, Y., Brooks, M. D., Lee, C. T., Scott, C., Balzano, L., Wicha, M. S., & Colacino, J. A. (2020). Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.00288

Van Keymeulen, A., Rocha, A. S., Ousset, M., Beck, B., Bouvencourt, G., Rock, J., Sharma, N., Dekoninck, S., & Blanpain, C. (2011). Distinct stem cells contribute to mammary gland development and maintenance. *Nature*, *479*(7372), 189–193. https://doi.org/10.1038/nature10573

Visvader, J. E., & Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes & Development*, *28*(11), 1143–1158. https://doi.org/10.1101/gad.242511.114

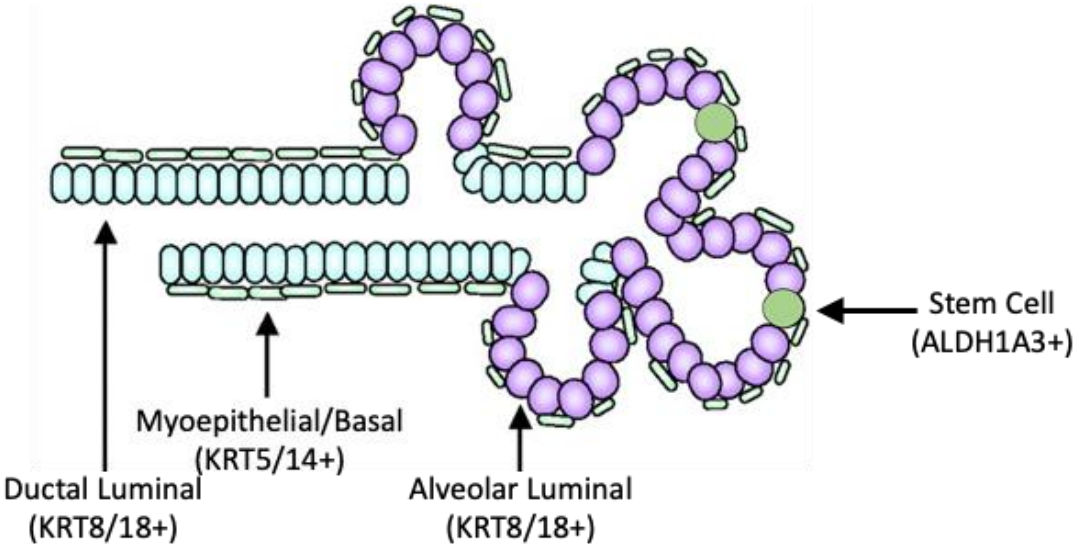
Wang, J., Wu, X., Wei, C., Huang, X., Ma, Q., Huang, X., Faiola, F., Guallar, D., Fidalgo, M., Huang, T., Peng, D., Chen, L., Yu, H., Li, X., Sun, J., Liu, X., Cai, X., Chen, X., Wang, L., … Ding, J. (2018). YY1 Positively Regulates Transcription by Targeting Promoters and Super-Enhancers through the BAF Complex in Embryonic Stem Cells. *Stem Cell Reports*, *10*(4), 1324–1339. https://doi.org/10.1016/j.stemcr.2018.02.004

Wang, Z., & Yang, C. (2019). Metal carcinogen exposure induces cancer stem cell-like property through epigenetic reprograming: A novel mechanism of metal carcinogenesis. *Seminars in Cancer Biology*, *57*, 95–104. https://doi.org/10.1016/j.semcancer.2019.01.002

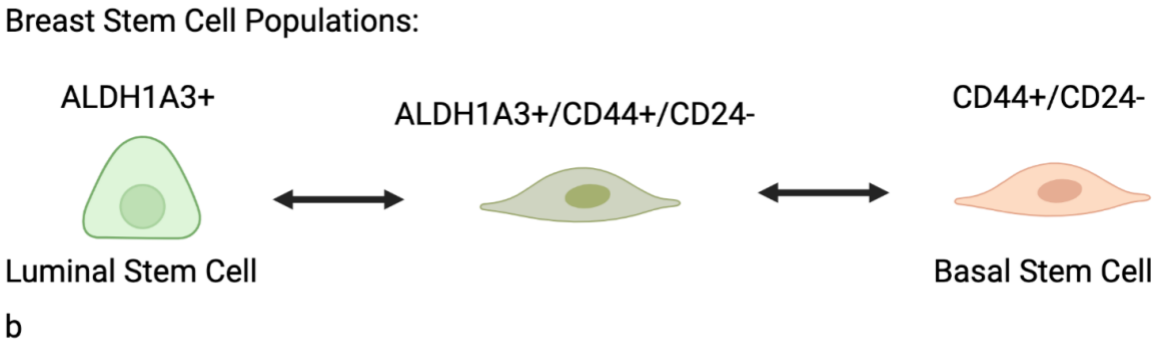
Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3), e90. https://doi.org/10.1002/cpz1.90

Zuccarello, P., Oliveri Conti, G., Cavallaro, F., Copat, C., Cristaldi, A., Fiore, M., & Ferrante, M. (2018). Implication of dietary phthalates in breast cancer. A systematic review. *Food and Chemical Toxicology*, *118*, 667–674. https://doi.org/10.1016/j.fct.2018.06.011

Figures and tables



a



b

Figure 4.1: a) Mammary gland structure depicting differentiated cell types. (Adapted from Woodward et al. 2005) b) schematic of breast stem cell populations.

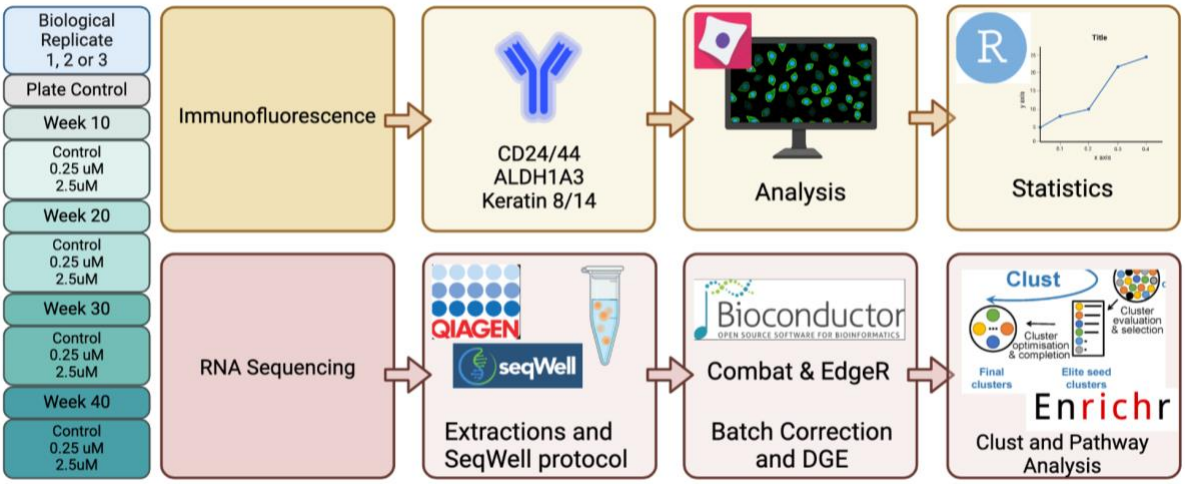


Figure 4.2: Aim 3 Overall Schematic. This figure represents the workflow for aim 3. After each 40-week biological replicate was finished, the cells were prepared for immunofluorescence staining and RNA sequencing. For immunofluorescence, cells from each condition (plate control, Week 10, Week 20, Week 30, and Week 40) were plated in a 384 well plate with 8 replicates per treatment. Next, after 48 hours, cells were fixed and stained for antibodies specific to keratin 8, keratin 14, CD24, CD44, and ALDH1A3. Analysis was then performed by inputting immunofluorescence data into Cell Profiler to quantify intensity and expression. QC and illumination correction were used to flag out blurry images and saturated images prior to unbiased quantification. R studio was used to perform T-tests to compare differences of mean values between controls and treated samples. Significant data in results is annotated with *. For RNA sequencing, cells from each week, condition, and biological replicate were grown up and collected for DNA, RNA, and smRNA extractions. The PlexWell protocol was used to clean up the RNA to convert it to cDNA. Samples were then barcoded, pooled and amplified. Bioconductor packages, Combat and EdgeR, were used to remove batch effects from RNA seq samples and for differential gene analysis for comparison between baseline cells, control cells, and dosed cells, respectively. Clust is an automatic extraction of optimal co-expressed clusters from gene expression data. Gene set enrichment analysis, Enrichr, was used for pathway analysis.

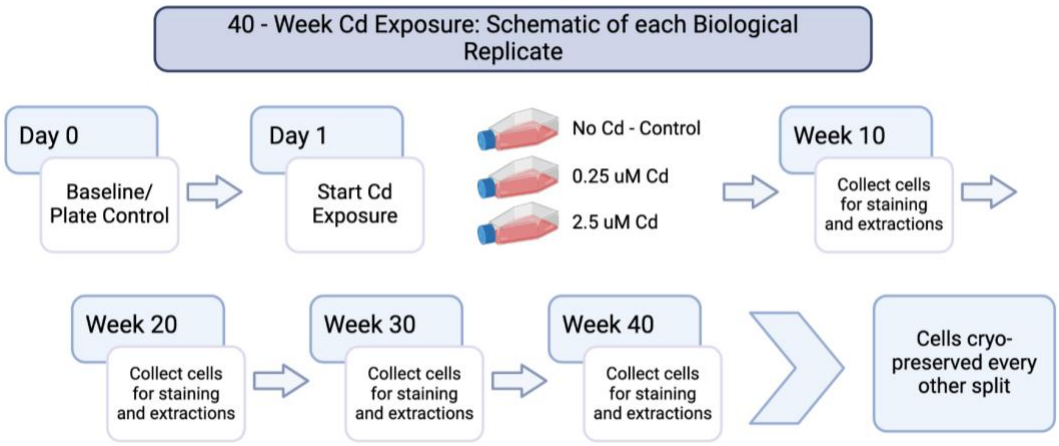


Figure 4.3: 40-week culture schematic for each biological replicate.

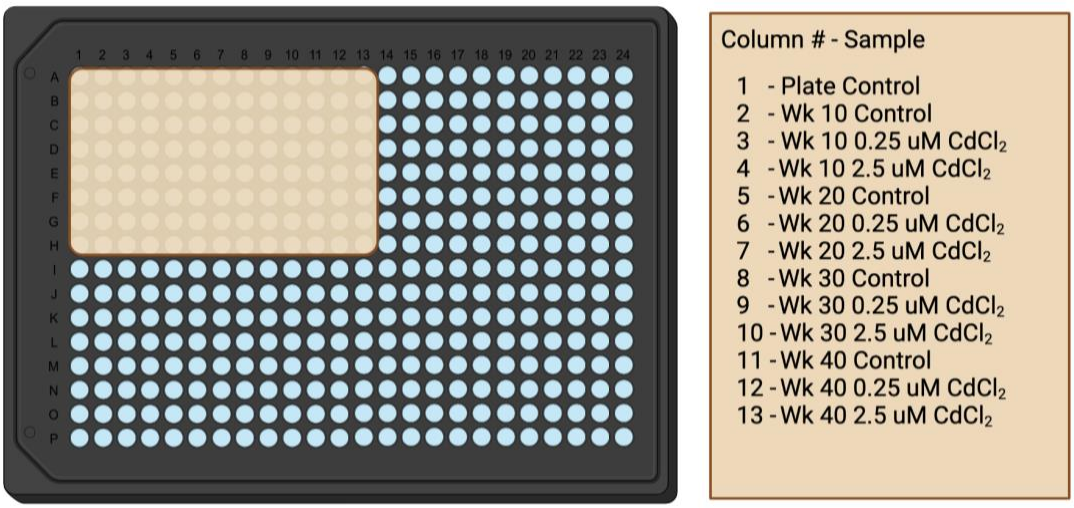


Figure 4.4: Plate layout for 384-well immunostaining.

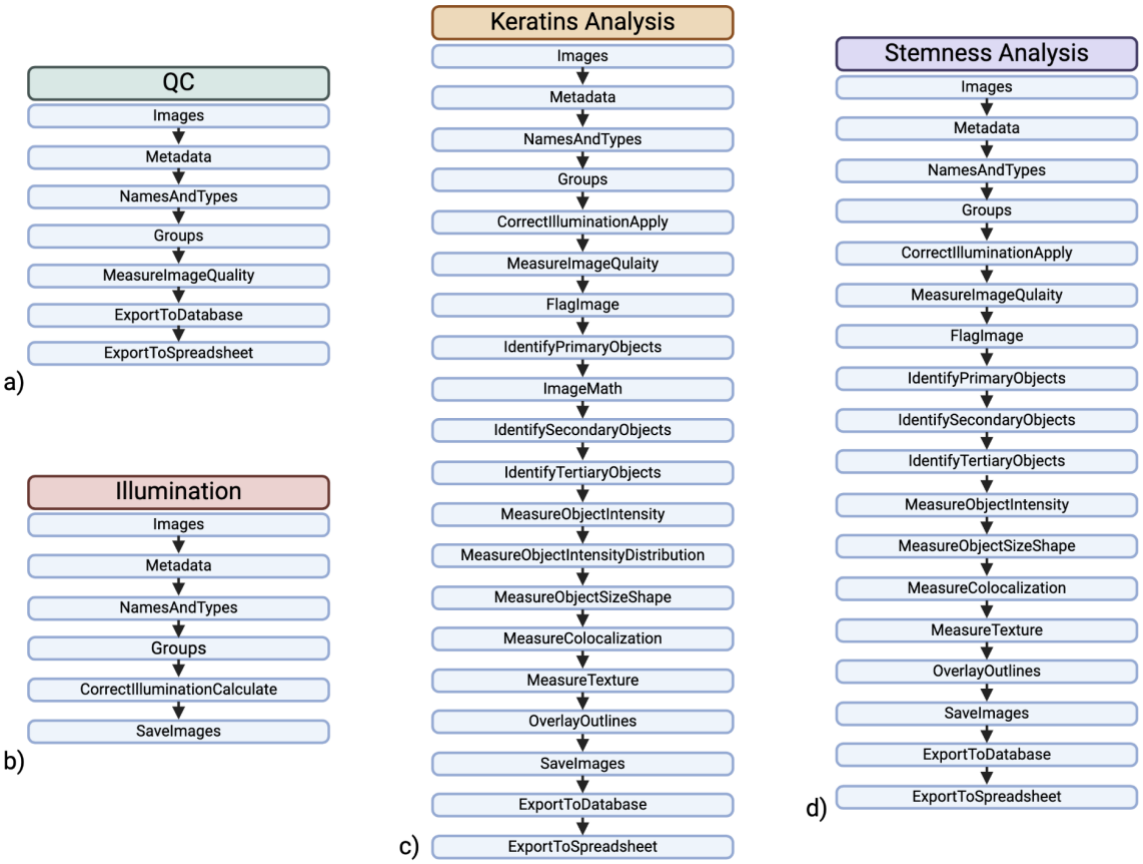
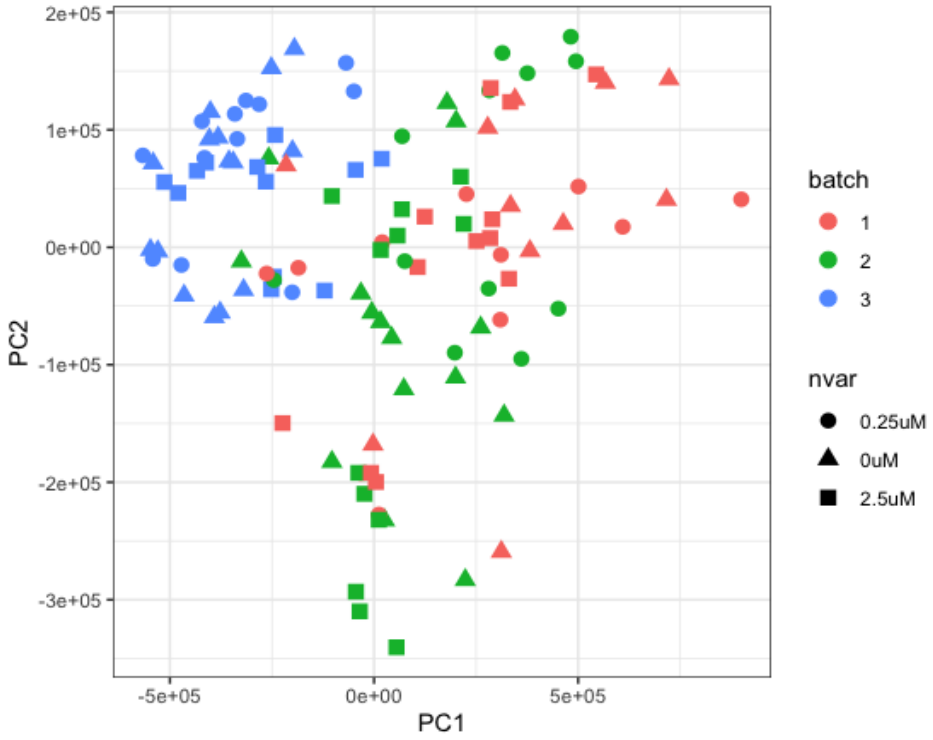


Figure 4.5: CellProfiler Pipelines. a) QC pipeline. b) Illumination pipeline. c) Keratins analysis pipeline. d) Stemness analysis pipeline.

Before Combat (a):



After Combat (b):

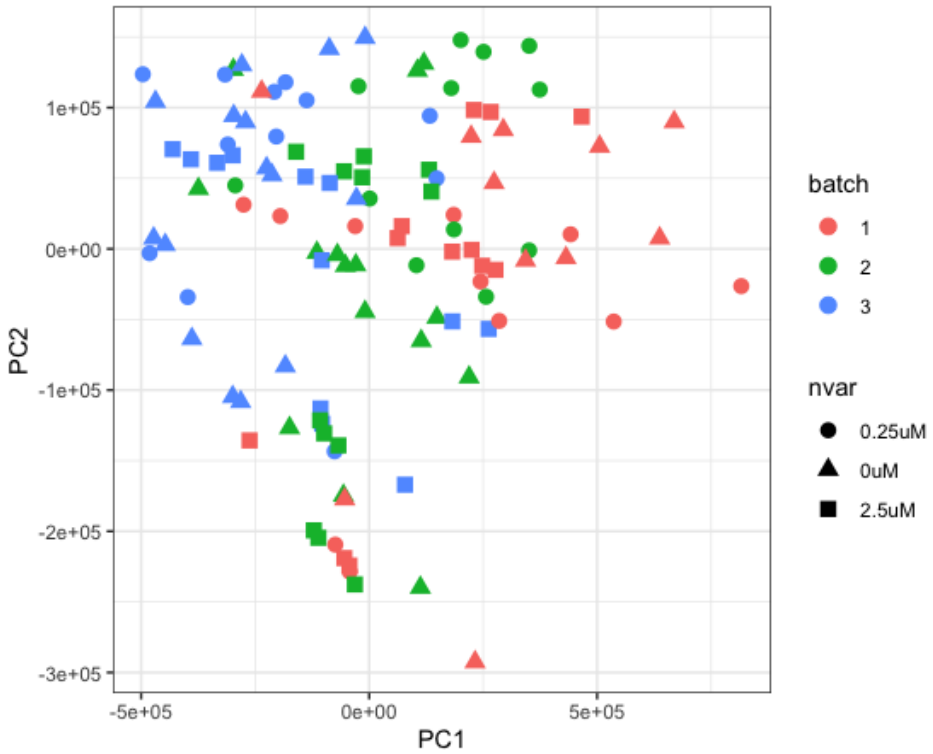


Figure 4.6: Batch Correction using Bioconductor Package, Combat. Here, “Batch” represents biological replicate. Batch 1 is red, batch 2 is green, and batch 3 is blue. a) Before Combat correction. b) After Combat correction.

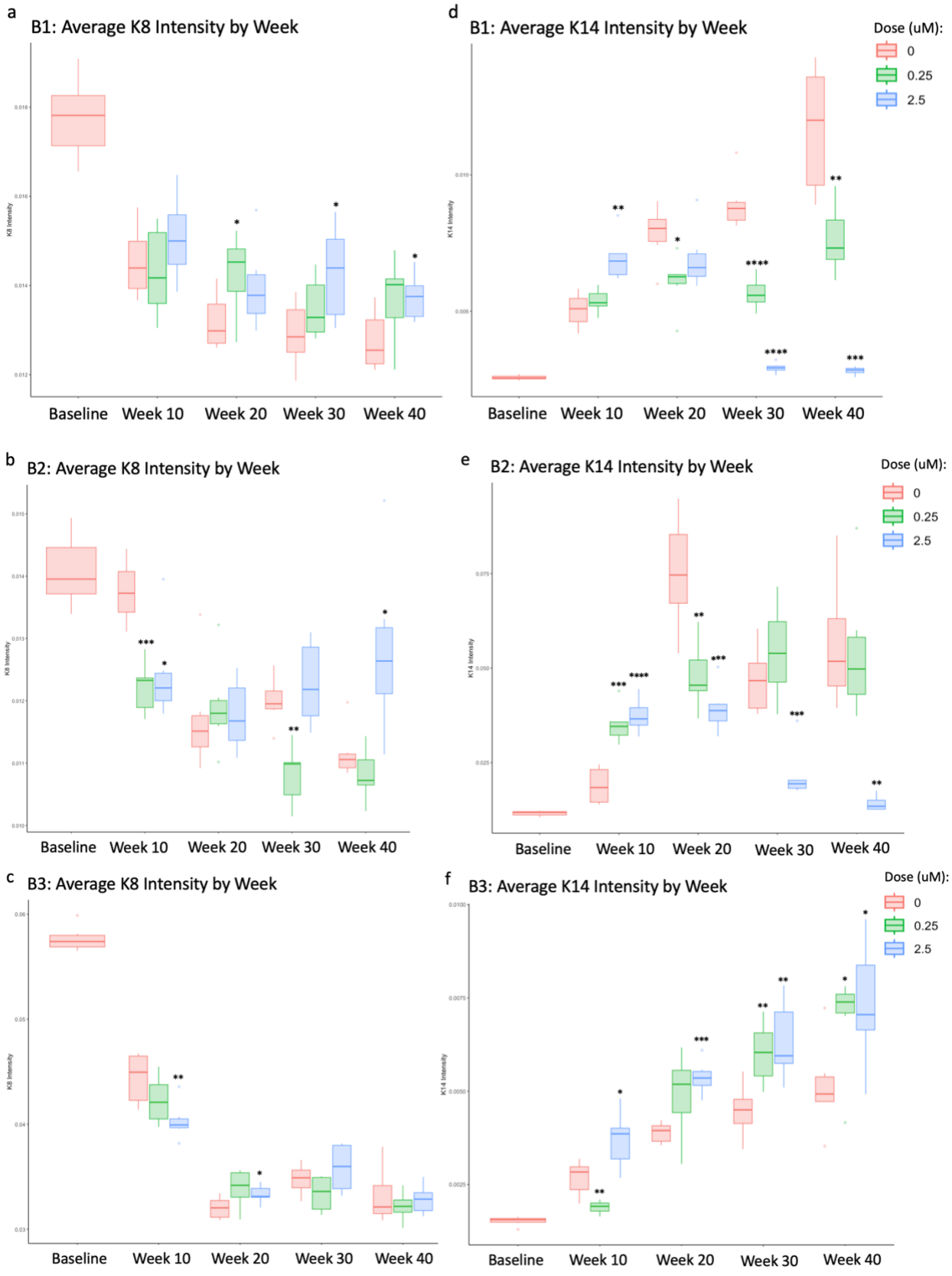

a
B1: Average K8 Intensity by Week
K8 Intensity
Baseline
Week 10
Week 20
Week 30
Week 40
d
B1: Average K14 Intensity by Week
K14 Intensity
Dose (uM):
0
0.25
2.5
b
B2: Average K8 Intensity by Week
e
B2: Average K14 Intensity by Week
c
B3: Average K8 Intensity by Week
f
B3: Average K14 Intensity by Week

Figure 4.7: Average intensity of K8 and K14 by week. B1 represents biological replicate 1, B2 represents biological replicate 2, and B3 represents biological replicate 3. Red indicates control (0 μM), green indicates the low dose (0.25 μM), and blue indicates the high dose (2.5 μM). a, b, c) The Average intensity of K8 by week. d,e, f) The average intensity of K14 by week. Significant data is annotated with *. Statistical significance was accepted with $p<0.05$; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$.

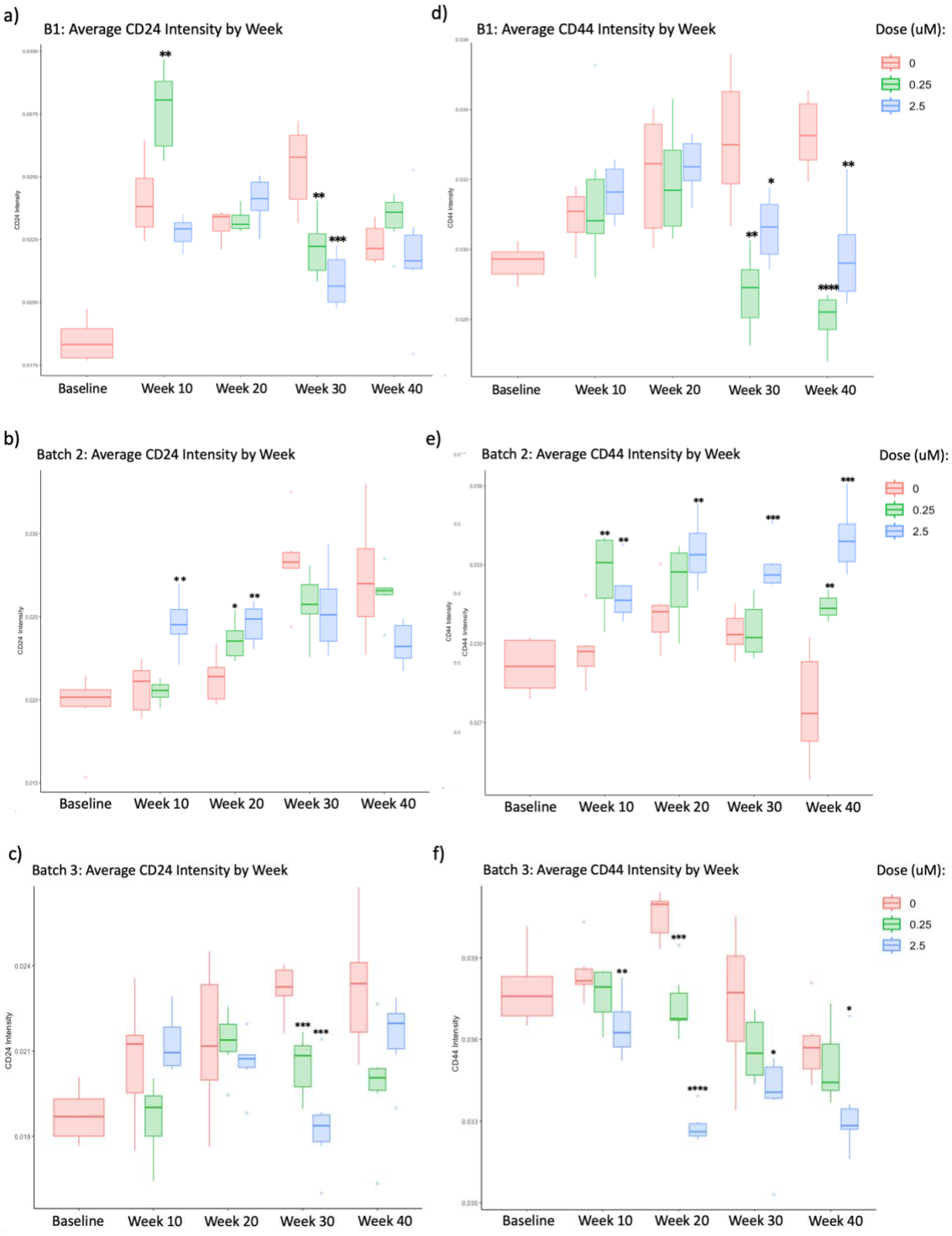
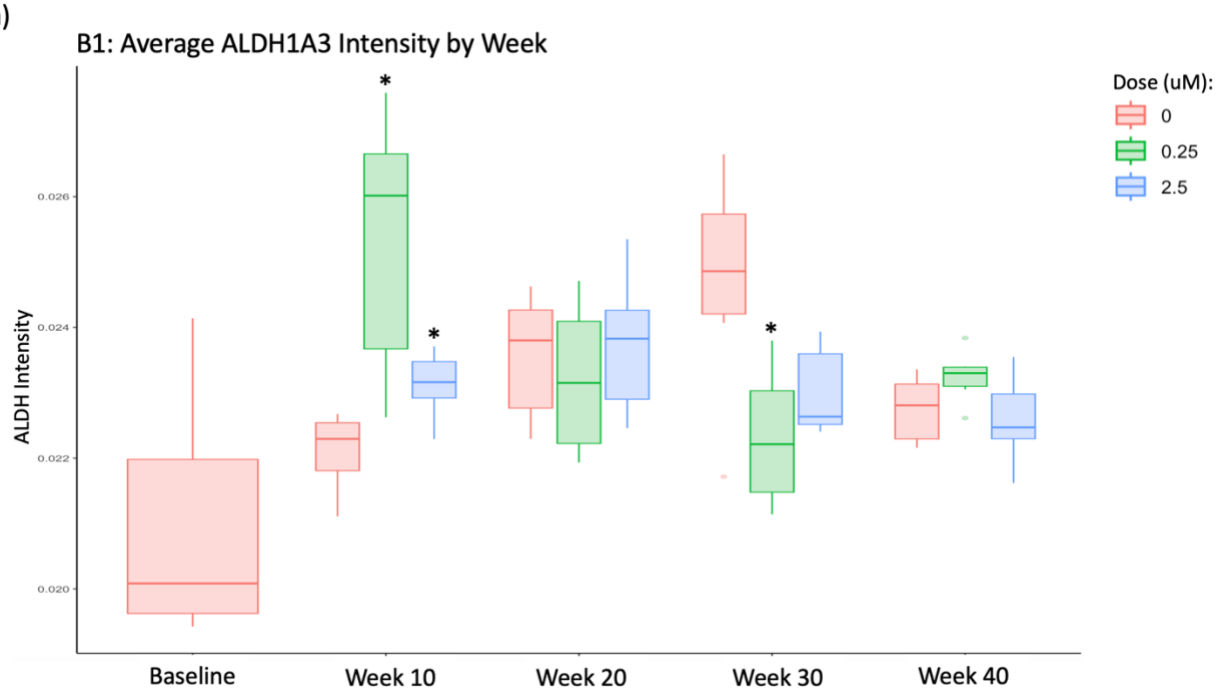


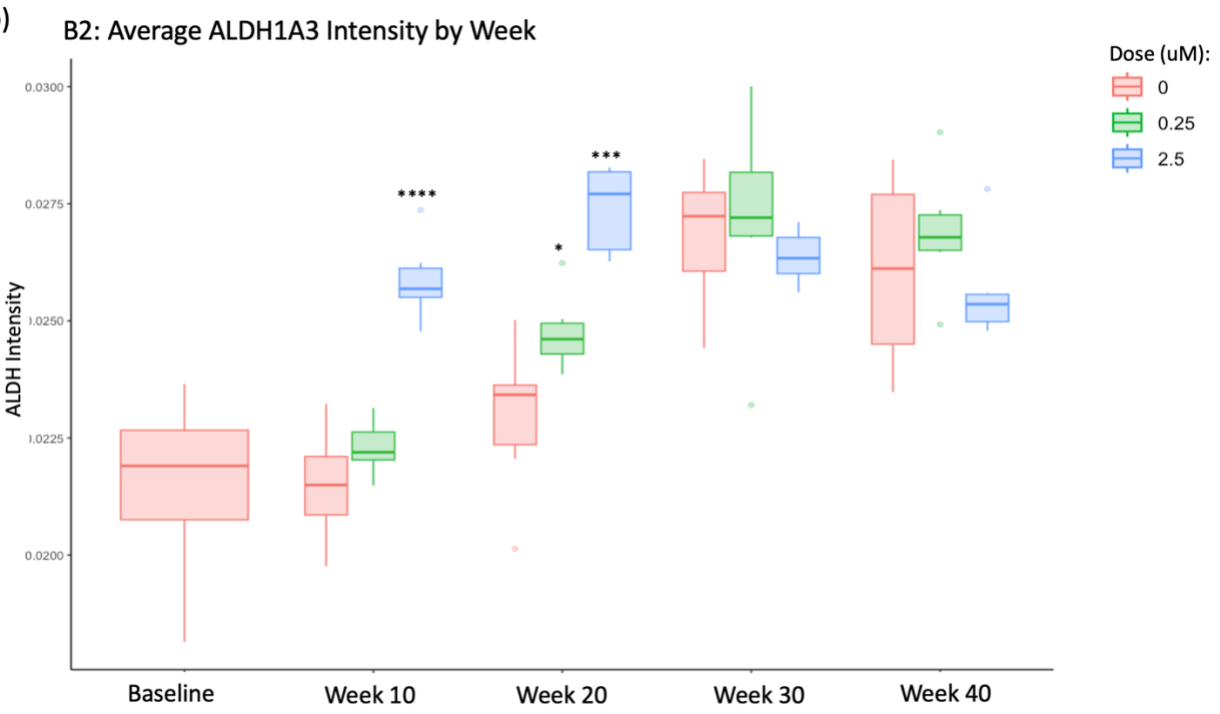
Figure 4.8: Average intensity of CD24 and CD44 by week. B1 represents biological replicate 1, B2 represents biological replicate 2, and B3 represents biological replicate

3. Red indicates control (0 μM), green indicates the low dose (0.25 μM), and blue indicates the high dose (2.5 μM). a, b, c) The Average intensity of CD24 by week. d, e, f) The average intensity of CD44 by week. Significant data is annotated with *. Statistical significance was accepted with $p<0.05$; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$.

a)


B1: Average ALDH1A3 Intensity by Week
ALDH Intensity
0.026
0.024
0.022
0.020
Baseline
Week 10
Week 20
Week 30
Week 40
Dose (uM):
0
0.25
2.5

b)


B2: Average ALDH1A3 Intensity by Week
ALDH Intensity
0.0300
0.0275
0.0250
0.0225
0.0200
Baseline
Week 10
Week 20
Week 30
Week 40
Dose (uM):
0
0.25
2.5

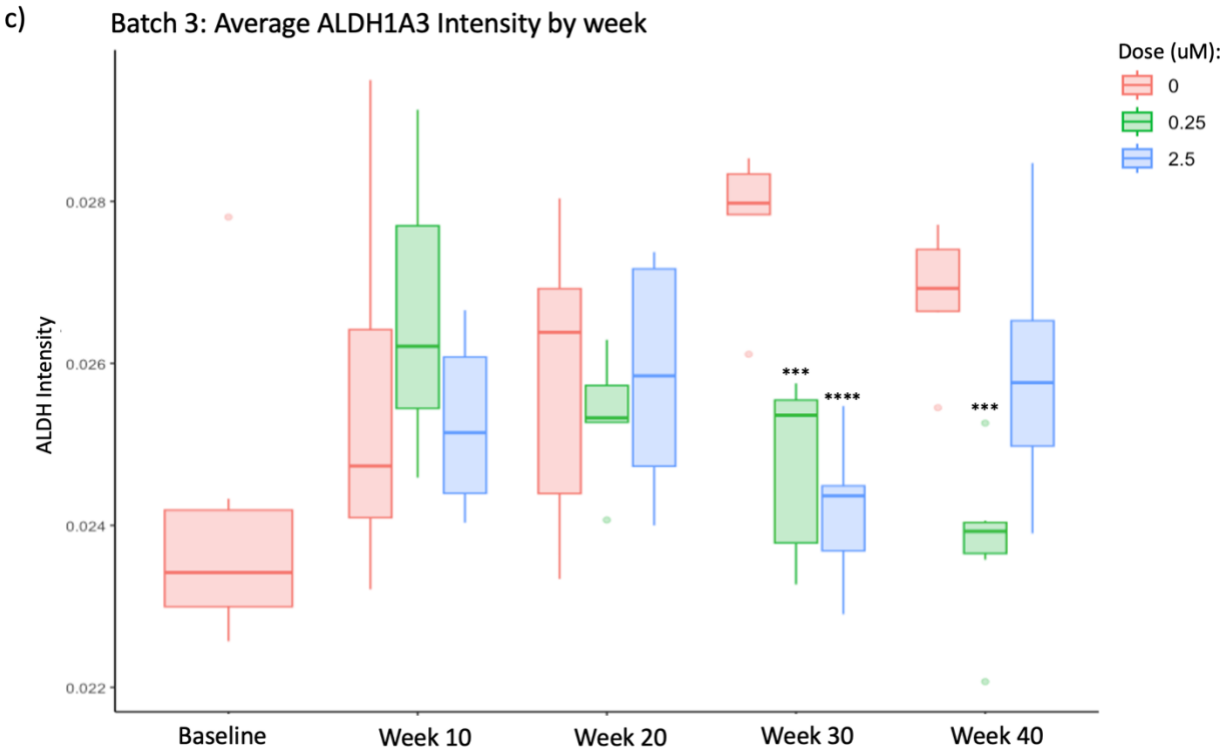
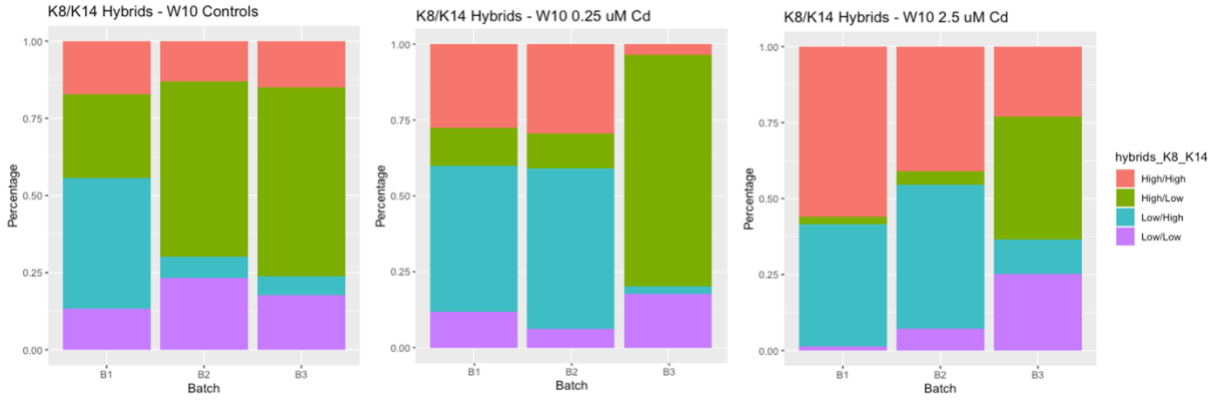
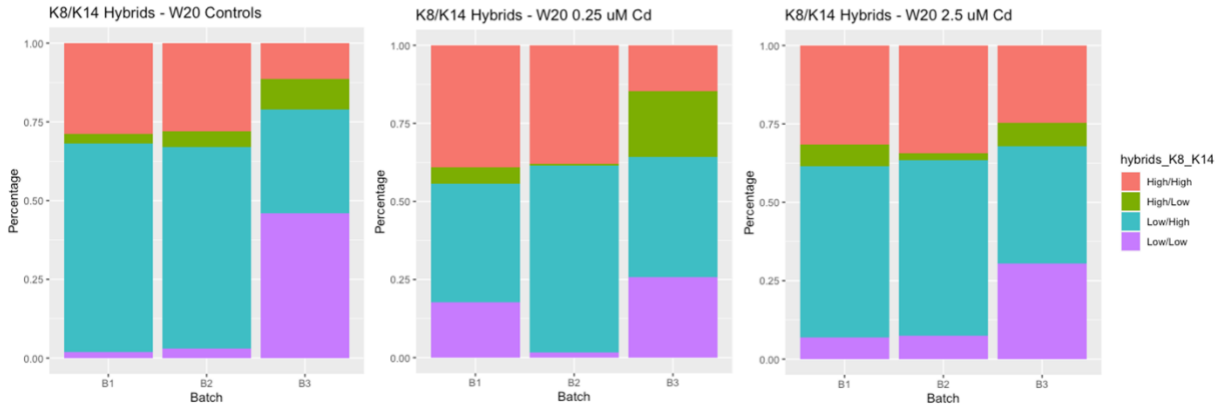


Figure 4.9: Average intensity of ALDH1A3 by week. B1 represents biological replicate 1, B2 represents biological replicate 2, and B3 represents biological replicate 3. Red indicates control (0 μM), green indicates the low dose (0.25 μM), and blue indicates the high dose (2.5 μM). a) The Average intensity of ALDH1A3 by week for B1. b) The average intensity of ALDH1A3 by week for B2. c) The average intensity of ALDH1A3 by week for B3. Significant data is annotated with *. Statistical significance was accepted with $p<0.05$; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$.

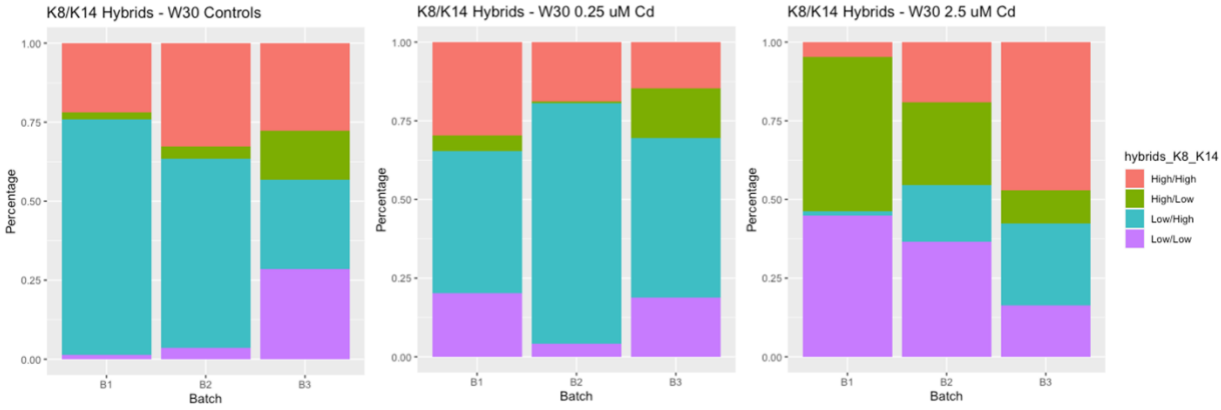
a


K8/K14 Hybrids - W10 Controls
K8/K14 Hybrids - W10 0.25 uM Cd
K8/K14 Hybrids - W10 2.5 uM Cd
Percentage
Batch
B1
B2
B3
hybrids_K8_K14
High/High
High/Low
Low/High
Low/Low

b


K8/K14 Hybrids - W20 Controls
K8/K14 Hybrids - W20 0.25 uM Cd
K8/K14 Hybrids - W20 2.5 uM Cd
Percentage
Batch
B1
B2
B3
hybrids_K8_K14
High/High
High/Low
Low/High
Low/Low

c


K8/K14 Hybrids - W30 Controls
K8/K14 Hybrids - W30 0.25 uM Cd
K8/K14 Hybrids - W30 2.5 uM Cd
Percentage
Batch
B1
B2
B3
hybrids_K8_K14
High/High
High/Low
Low/High
Low/Low

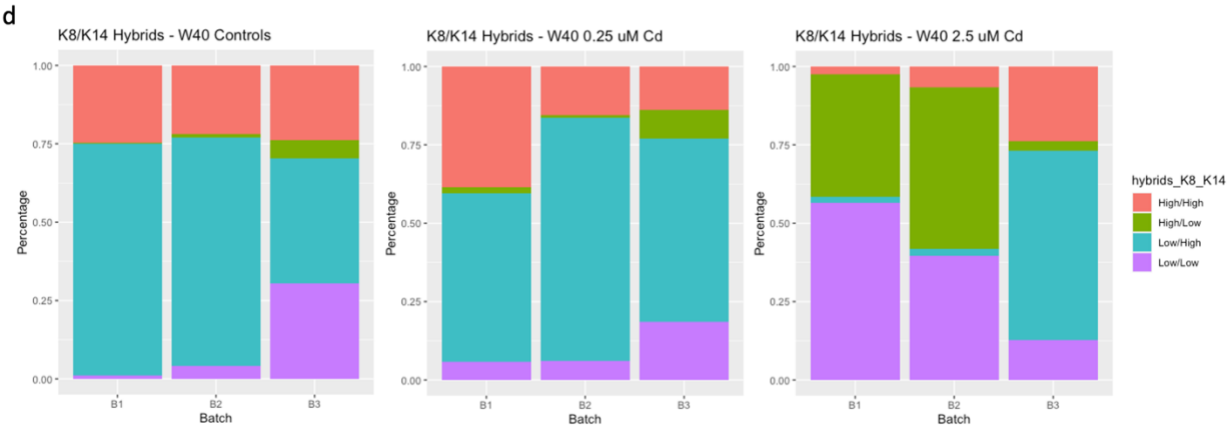
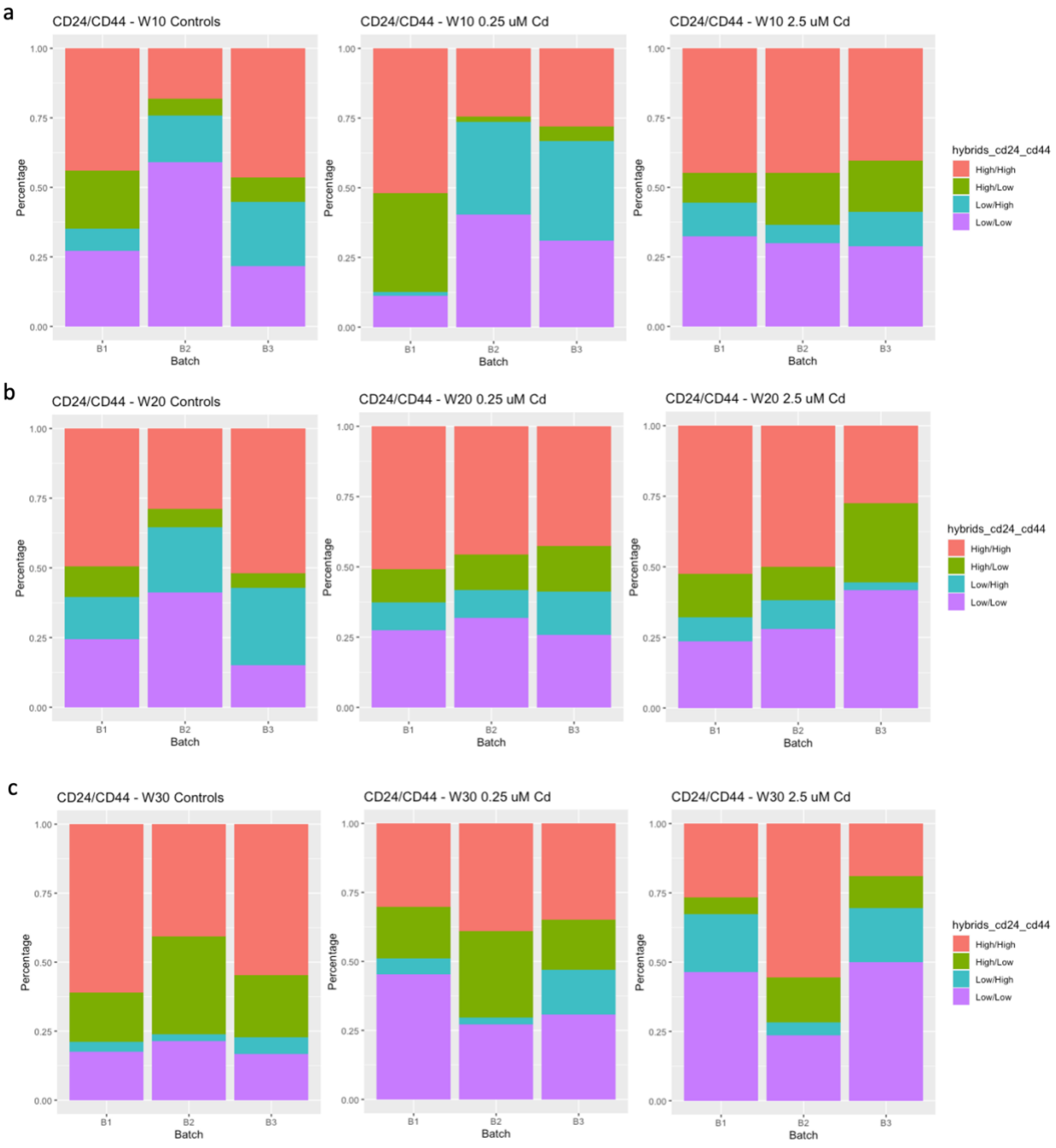


Figure 4.10: Proportion of K8/K14 hybrid cells per condition for all three biological replicates. Red indicates K8+/K14+ cells, green indicates K8+/K14- cells, blue indicates K8-/K14+ cells, and purple indicates K8-/K14- cells. a) Week 10 proportion of K8/K14 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. b) Week 20 proportion of K8/K14 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. c) Week 30 proportion of K8/K14 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. d) Week 40 proportion of K8/K14 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3.


a
CD24/CD44 - W10 Controls
CD24/CD44 - W10 0.25 uM Cd
CD24/CD44 - W10 2.5 uM Cd
b
CD24/CD44 - W20 Controls
CD24/CD44 - W20 0.25 uM Cd
CD24/CD44 - W20 2.5 uM Cd
c
CD24/CD44 - W30 Controls
CD24/CD44 - W30 0.25 uM Cd
CD24/CD44 - W30 2.5 uM Cd
Percentage
Batch
B1
B2
B3
hybrids_cd24_cd44
High/High
High/Low
Low/High
Low/Low

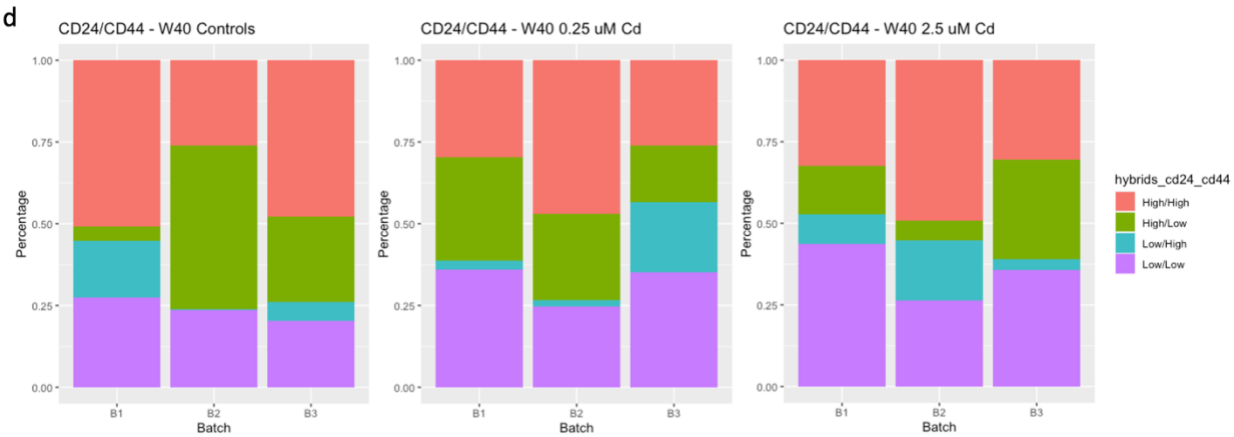


Figure 4.11: Proportion of CD24/CD44 hybrid cells per condition for all three biological replicates. Red indicates CD24+/CD44+ cells, green indicates K8+/CD44- cells, blue indicates CD24-/CD44+ cells, and purple indicates CD24-/K14- cells. a) Week 10 proportion of CD24/CD44 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. b) Week 20 proportion of CD24/CD44 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. c) Week 30 proportion of CD24/CD44 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3. d) Week 40 proportion of CD24/CD44 cells: Controls. 0.25 μM, and 2.5 μM for biological replicate 1, 2, and 3.

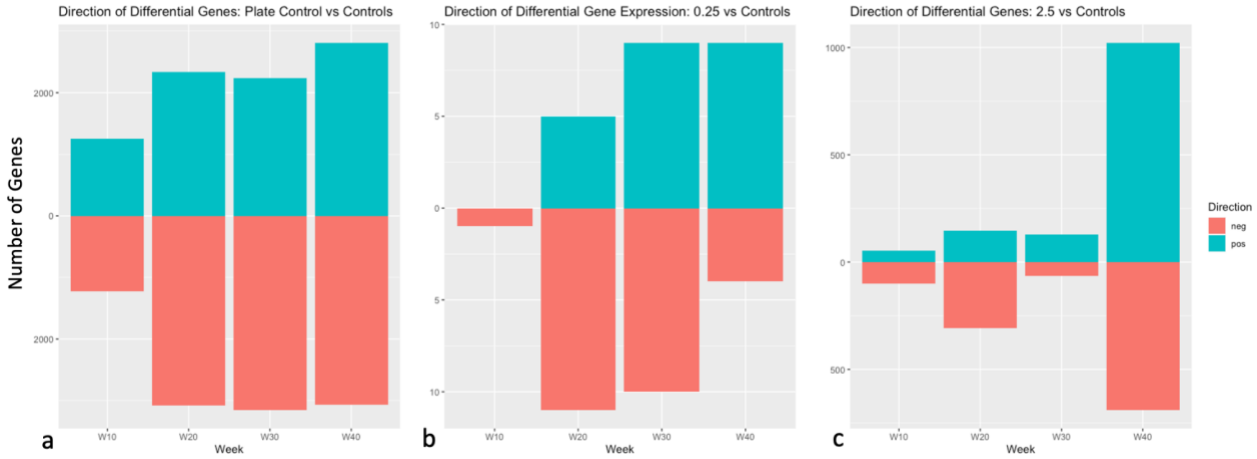
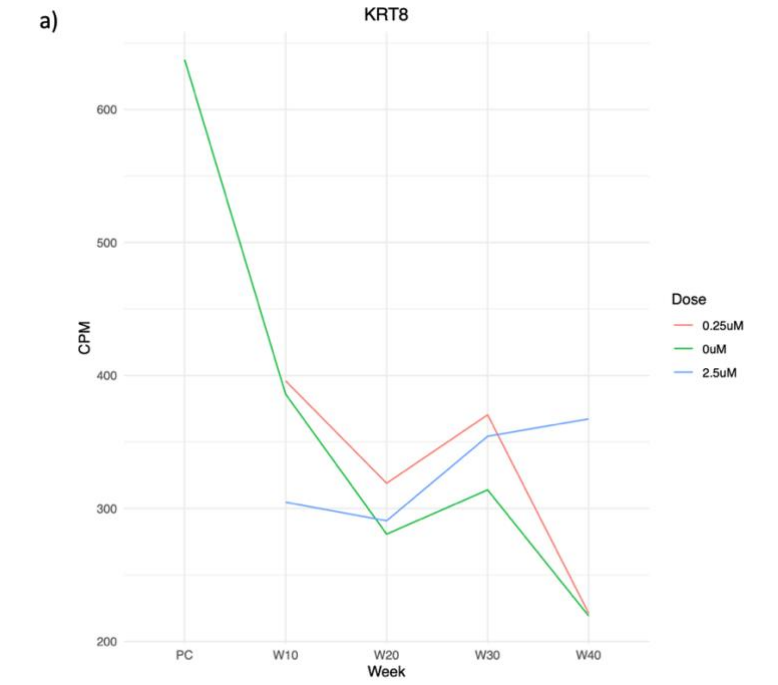
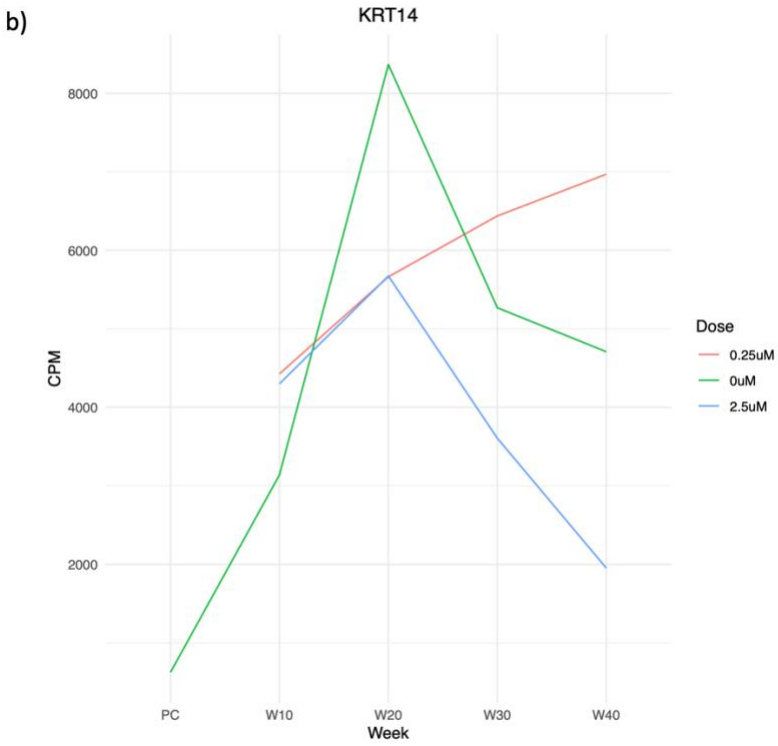


Figure 4.12: Direction of Differential Gene Expression by Treatment. Genes that are down-regulated are represented in red and genes that are up-regulated are represented in blue.

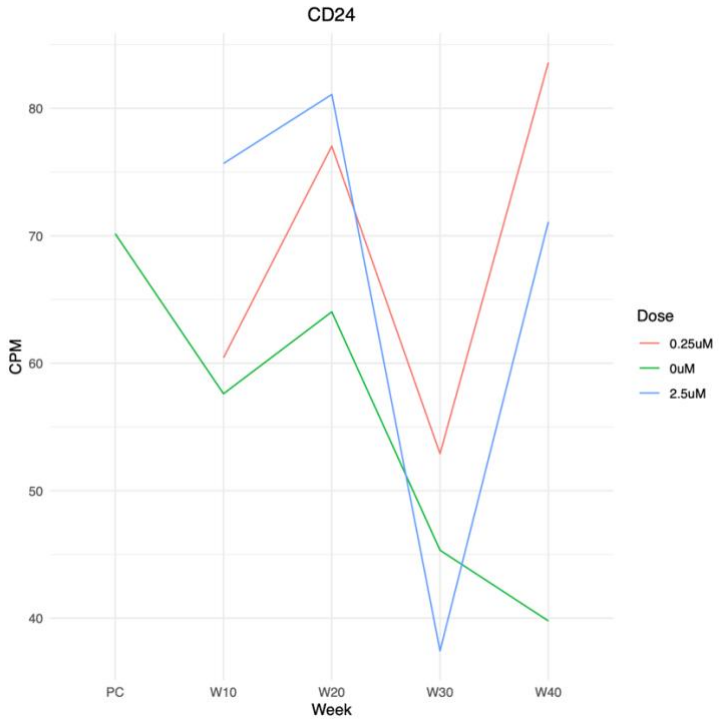
a)


KRT8
600
500
400
300
200
CPM
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

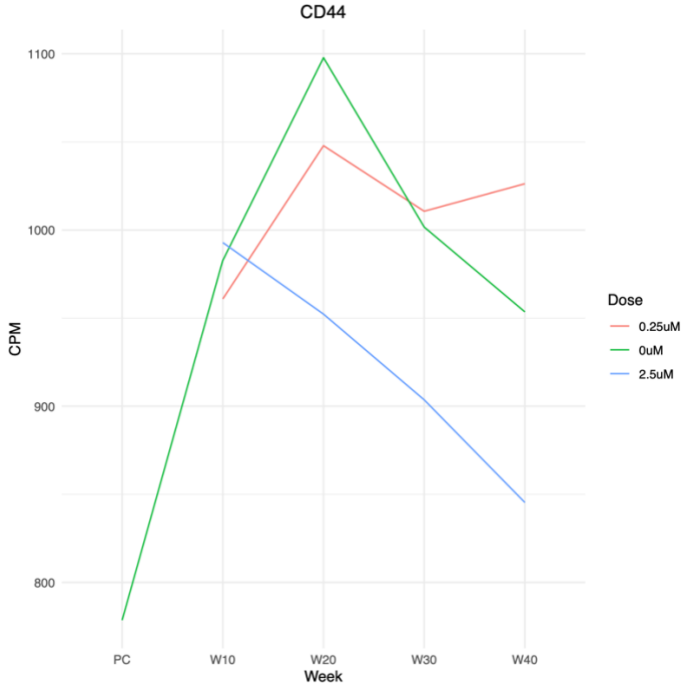
b)


KRT14
8000
6000
4000
2000
CPM
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

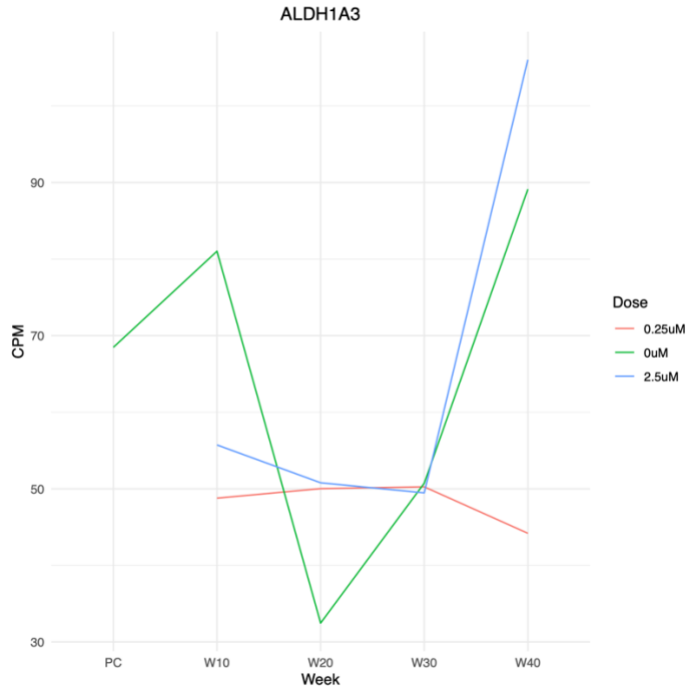
c)


CD24
80
70
60
50
40
CPM
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

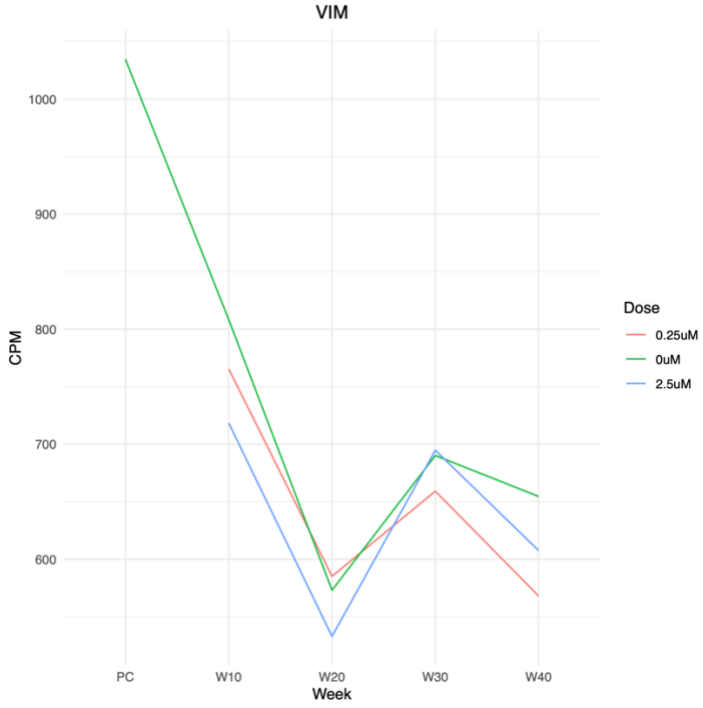
d)


CD44
1100
1000
900
800
CPM
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

e)


ALDH1A3
CPM
90
70
50
30
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

f)


VIM
CPM
1000
900
800
700
600
PC
W10
W20
W30
W40
Week
Dose
0.25uM
0uM
2.5uM

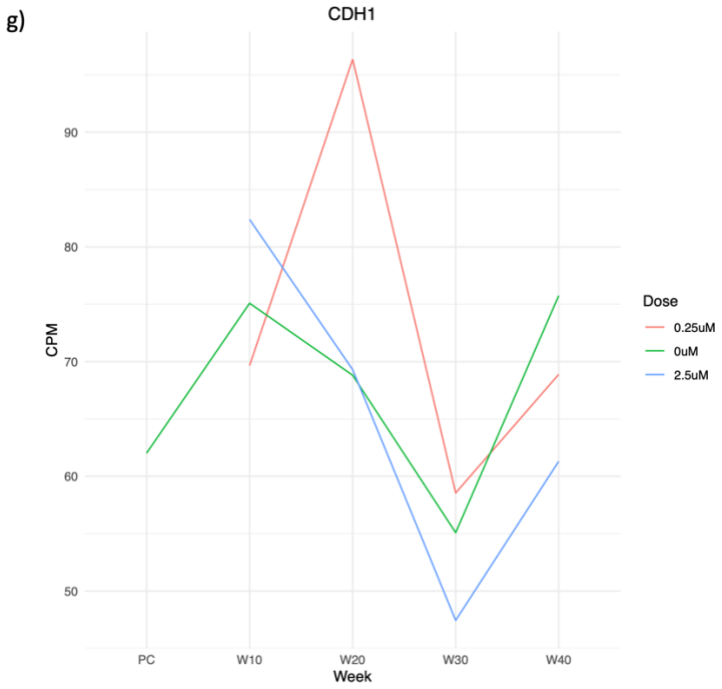


Figure 4.13: Line plot of gene expression for genes: a) KRT 8, b) KRT14, c) CD24, d) CD44, e) ALDH1A3, f) VIM and g) CDH1 over time.

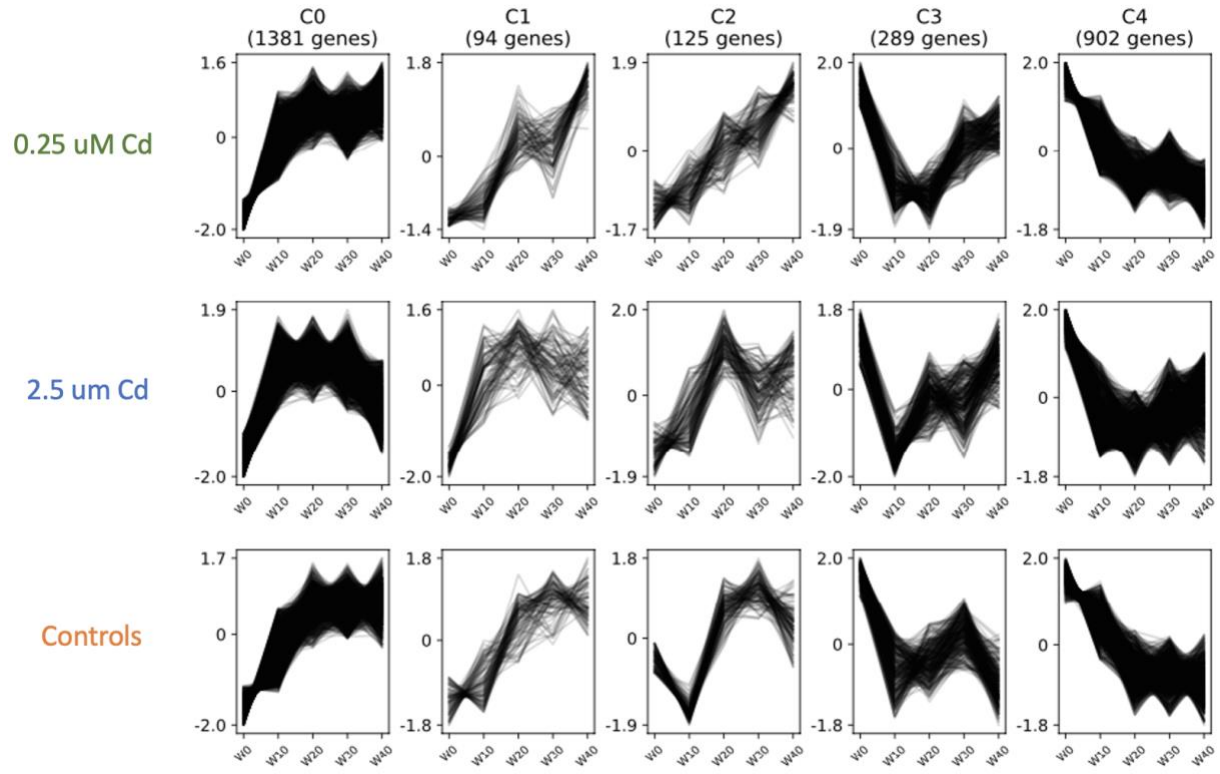


Figure 4.14: Clust profiles results. Controls profiles are shown in the bottom row, 0.25 μM profiles are shown in the top row, and 2.5 μM profiles are shown in the middle row.

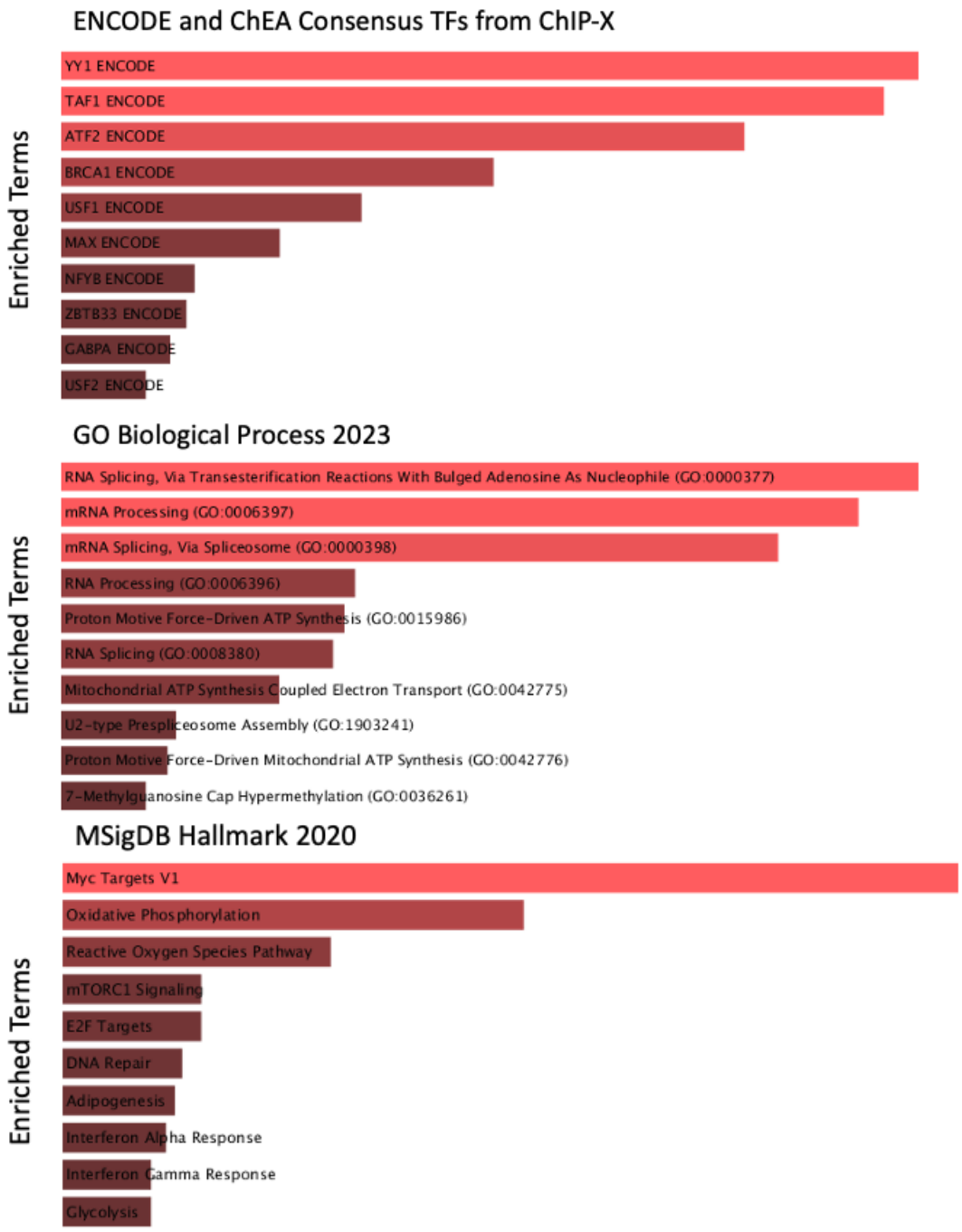
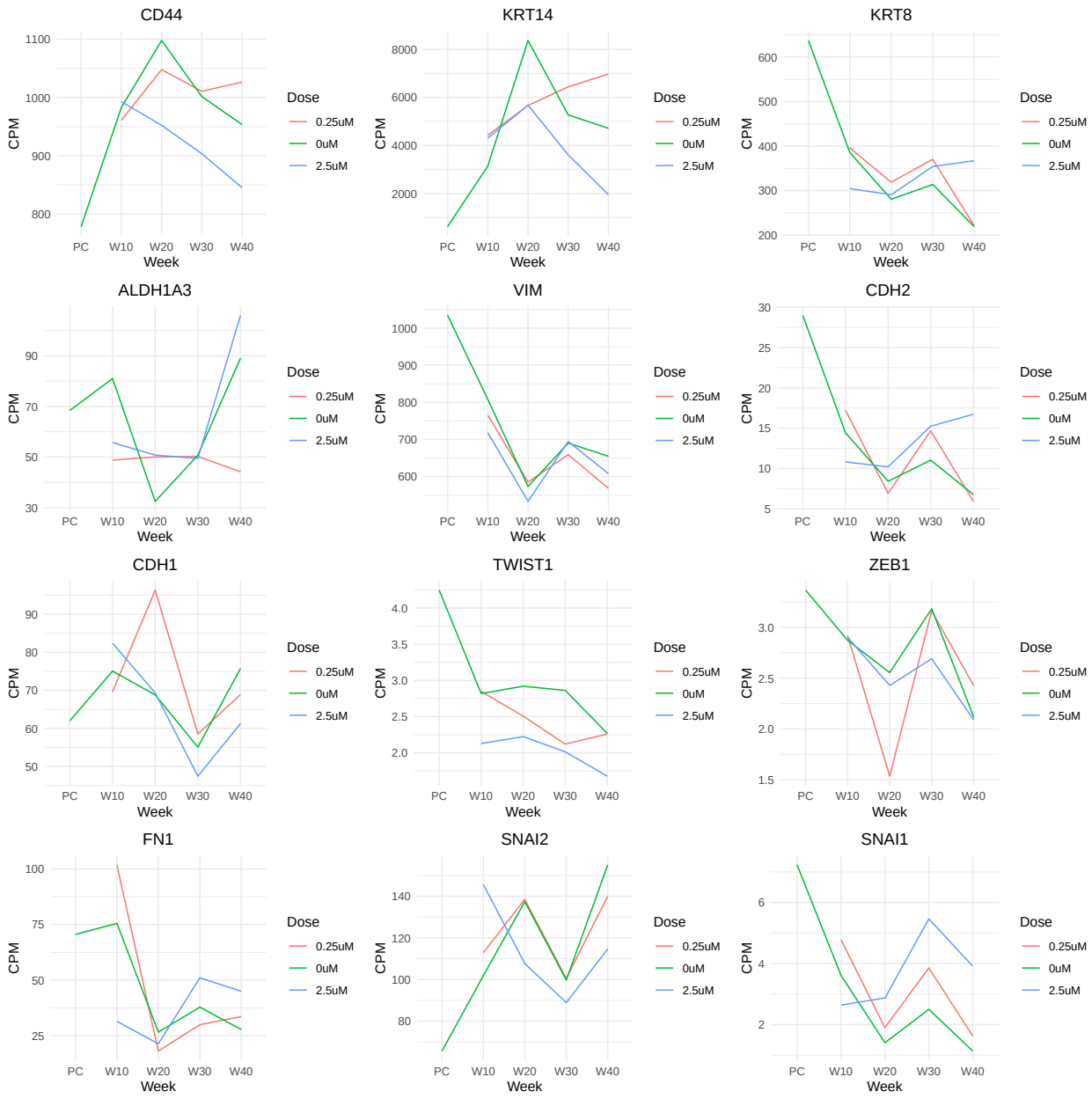


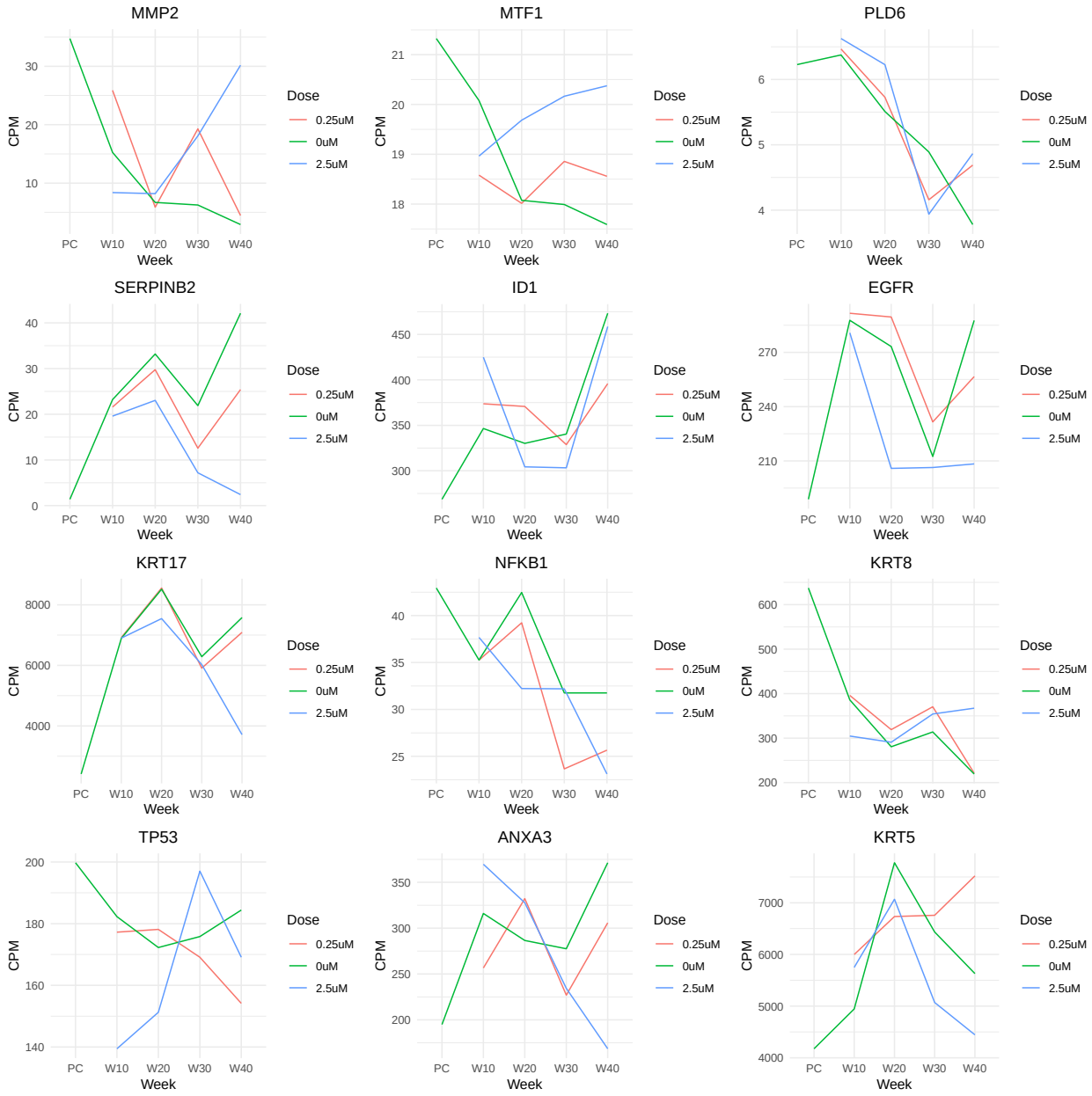
Figure 4.15: Pathway analysis from Enrichr using genes from cluster C3 of the Clust analysis. The bars plots are sorted by significance (p-value) based on color and length, with the longer and lighter red the bar, the more significant the pathway.

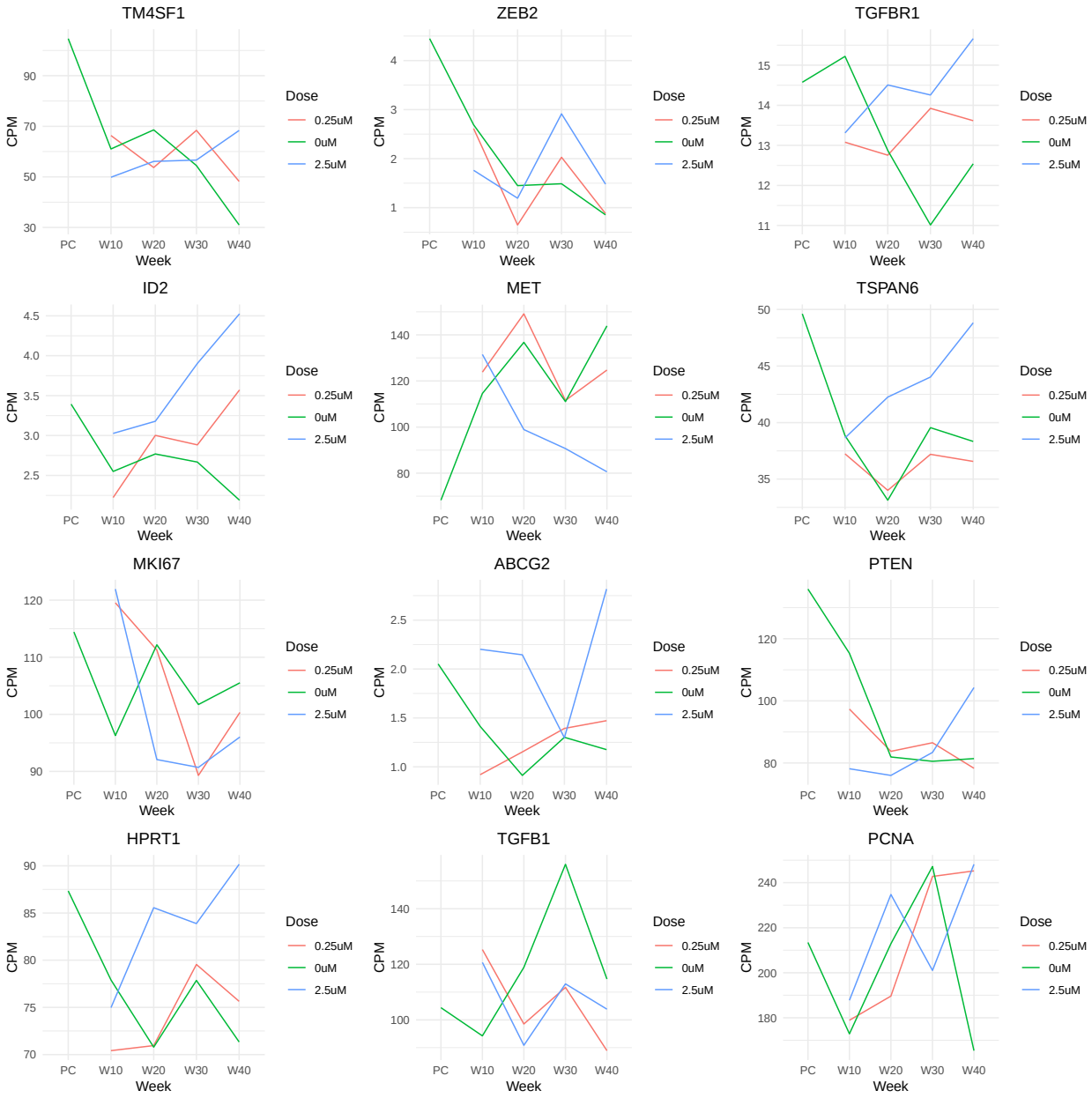
Table 4.1.1: Up_Down Plot – Controls. This table can be viewed by following the link.

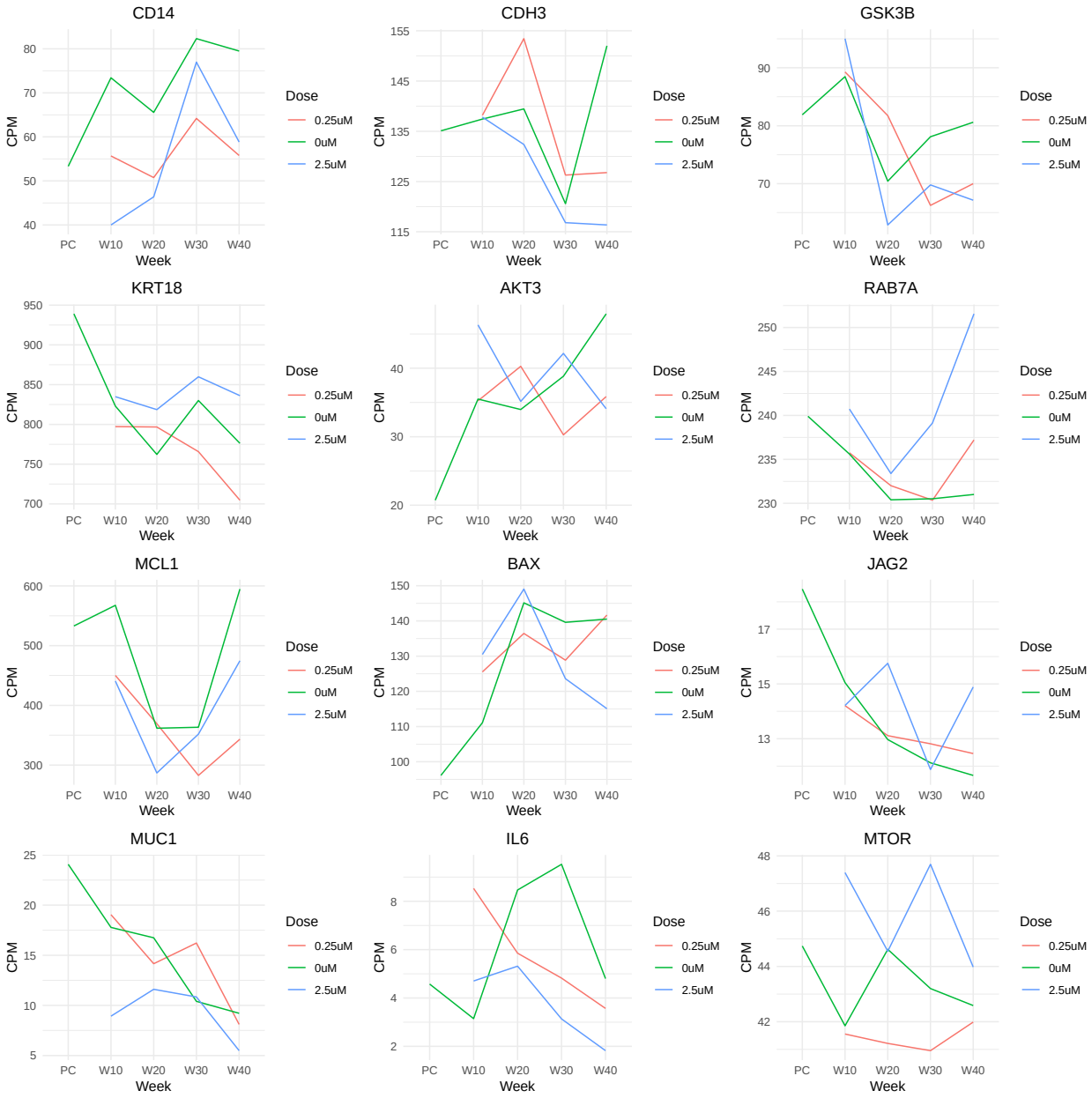
Table 4.1.2: Up_Down Plot – 0.25 μM. This table can be viewed by following the link.

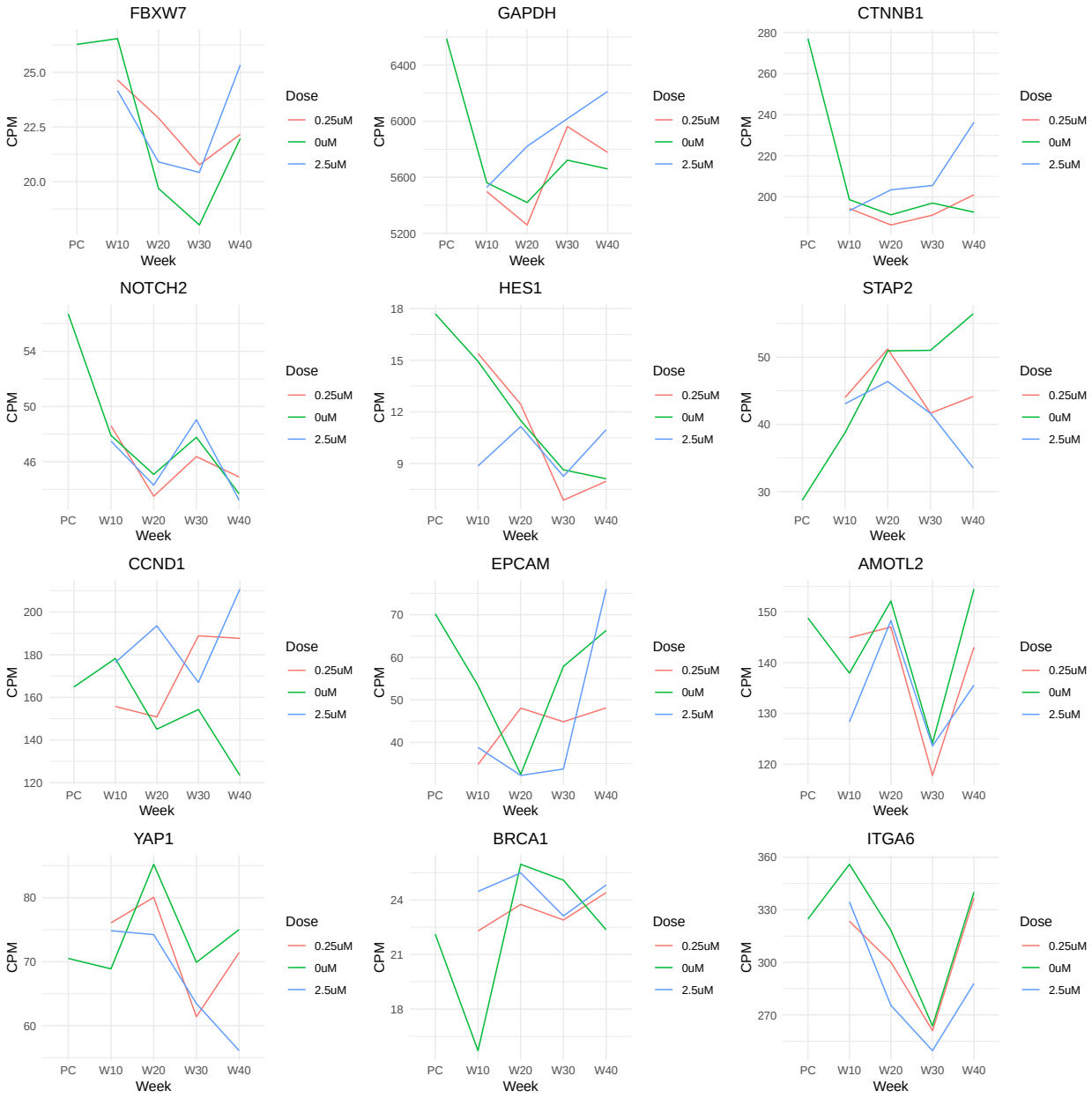
Table 4.1.3: Up_Down Plot – 2.5 μM. This table can be viewed by following the link.

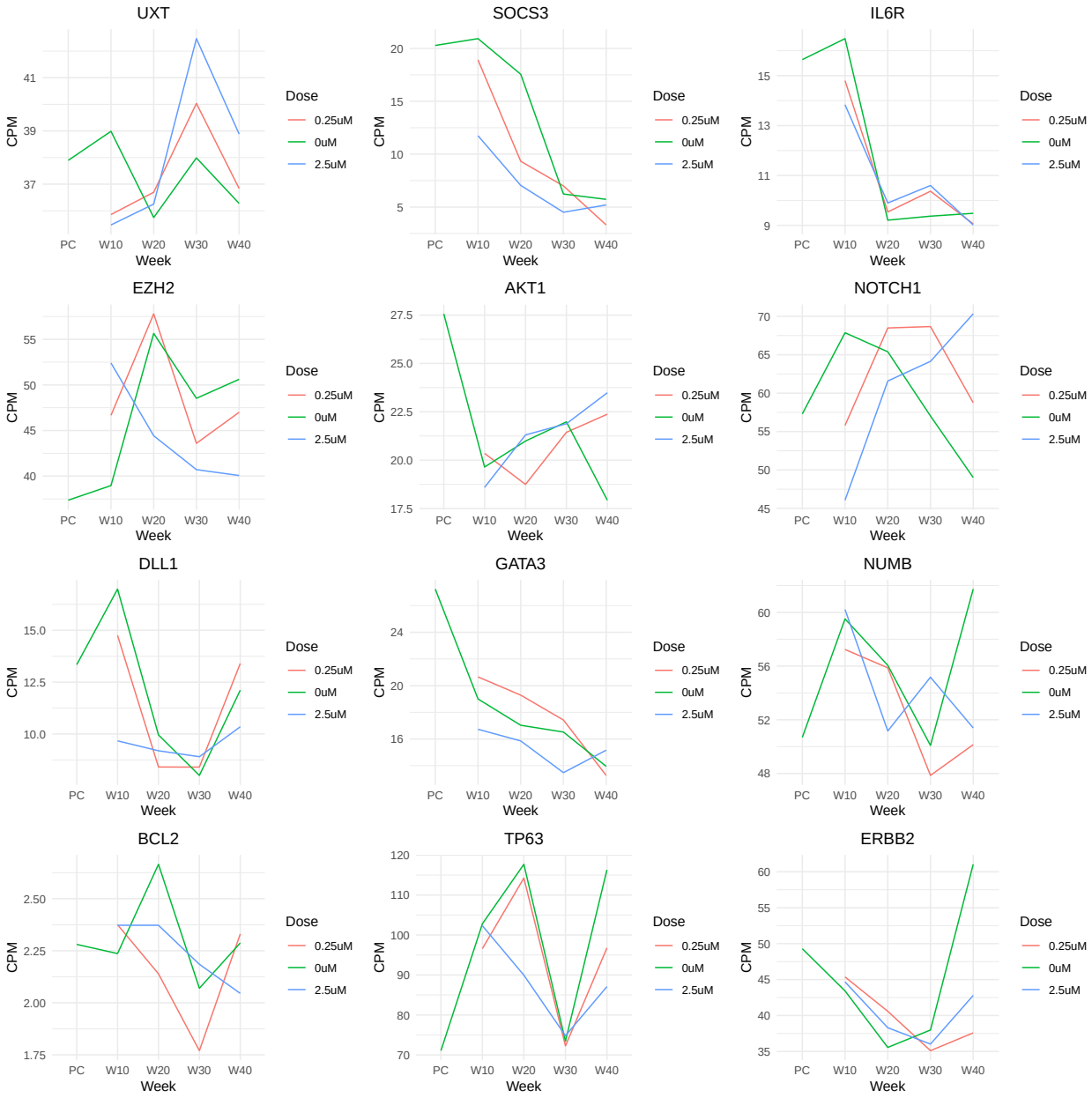


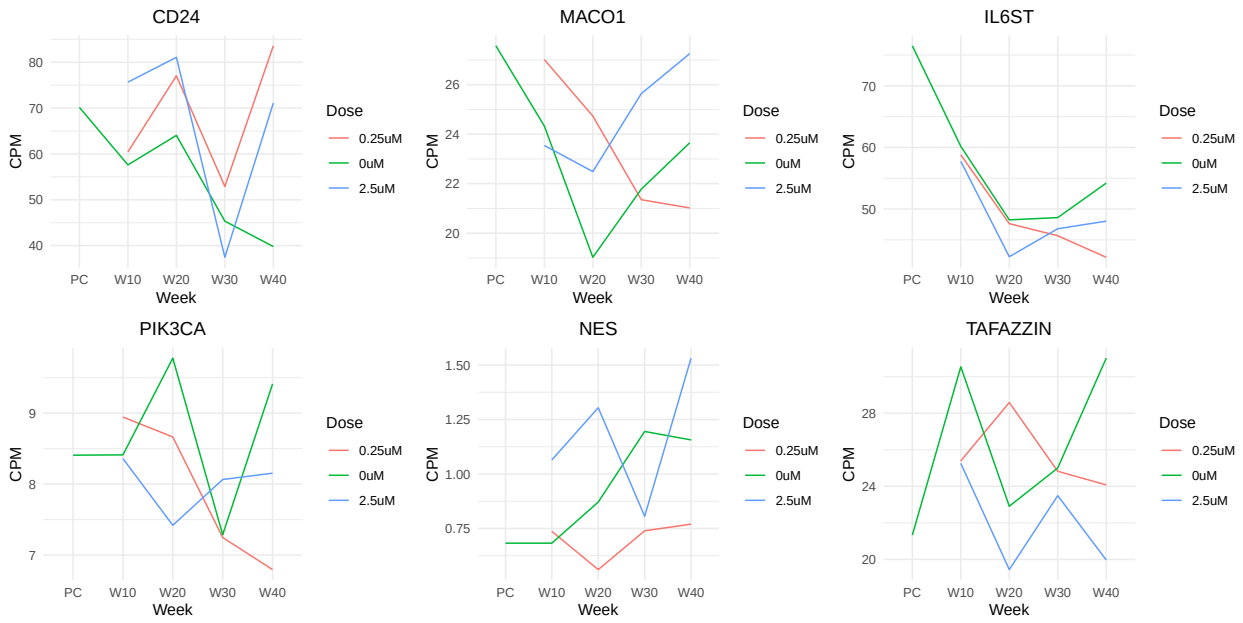

MMP2
MTF1
PLD6
SERPINB2
ID1
EGFR
KRT17
NFKB1
KRT8
TP53
ANXA3
KRT5
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM


TM4SF1
ZEB2
TGFBR1
ID2
MET
TSPAN6
MKI67
ABCG2
PTEN
HPRT1
TGFB1
PCNA
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

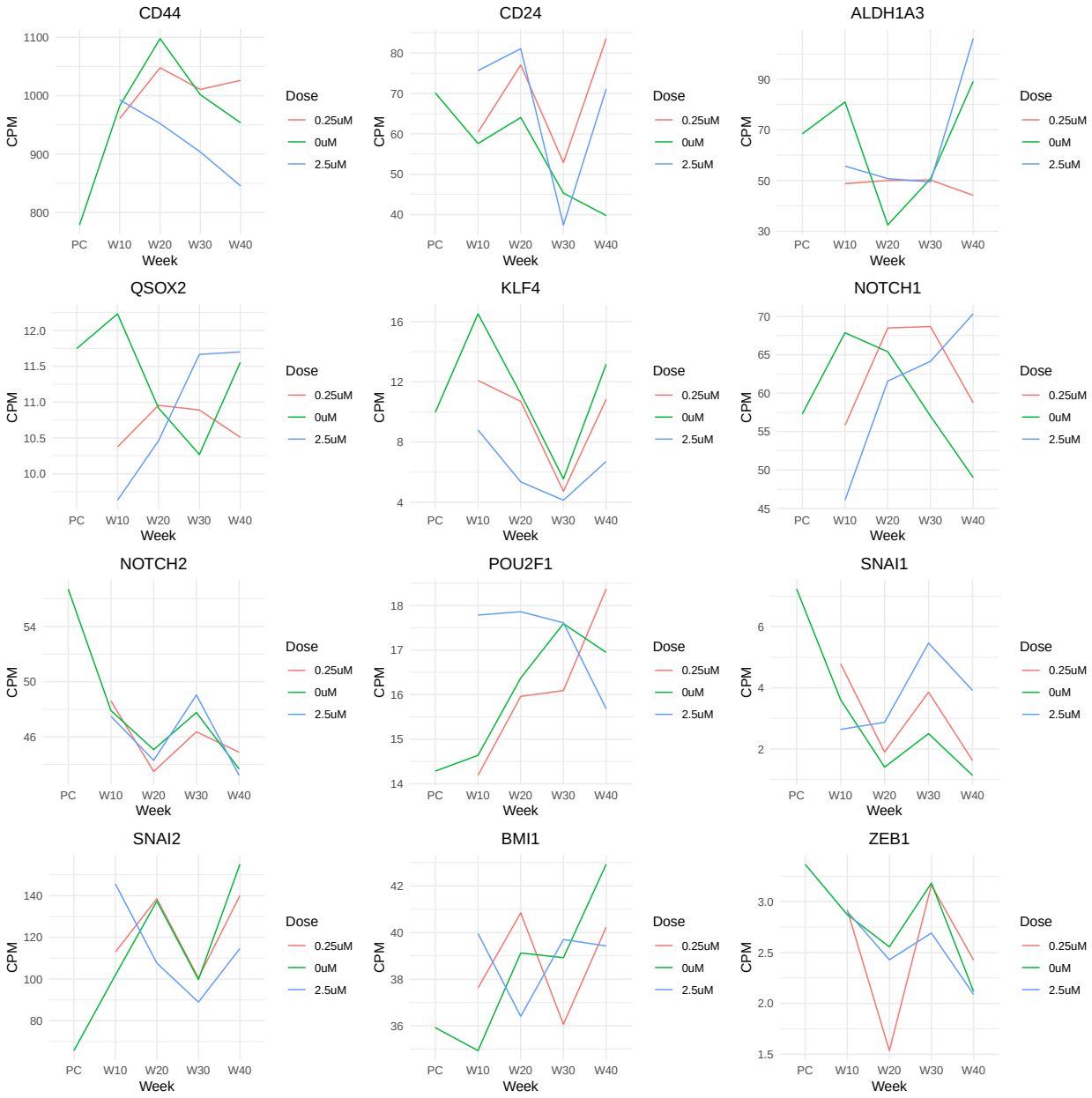

CD14
CDH3
GSK3B
KRT18
AKT3
RAB7A
MCL1
BAX
JAG2
MUC1
IL6
MTOR
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

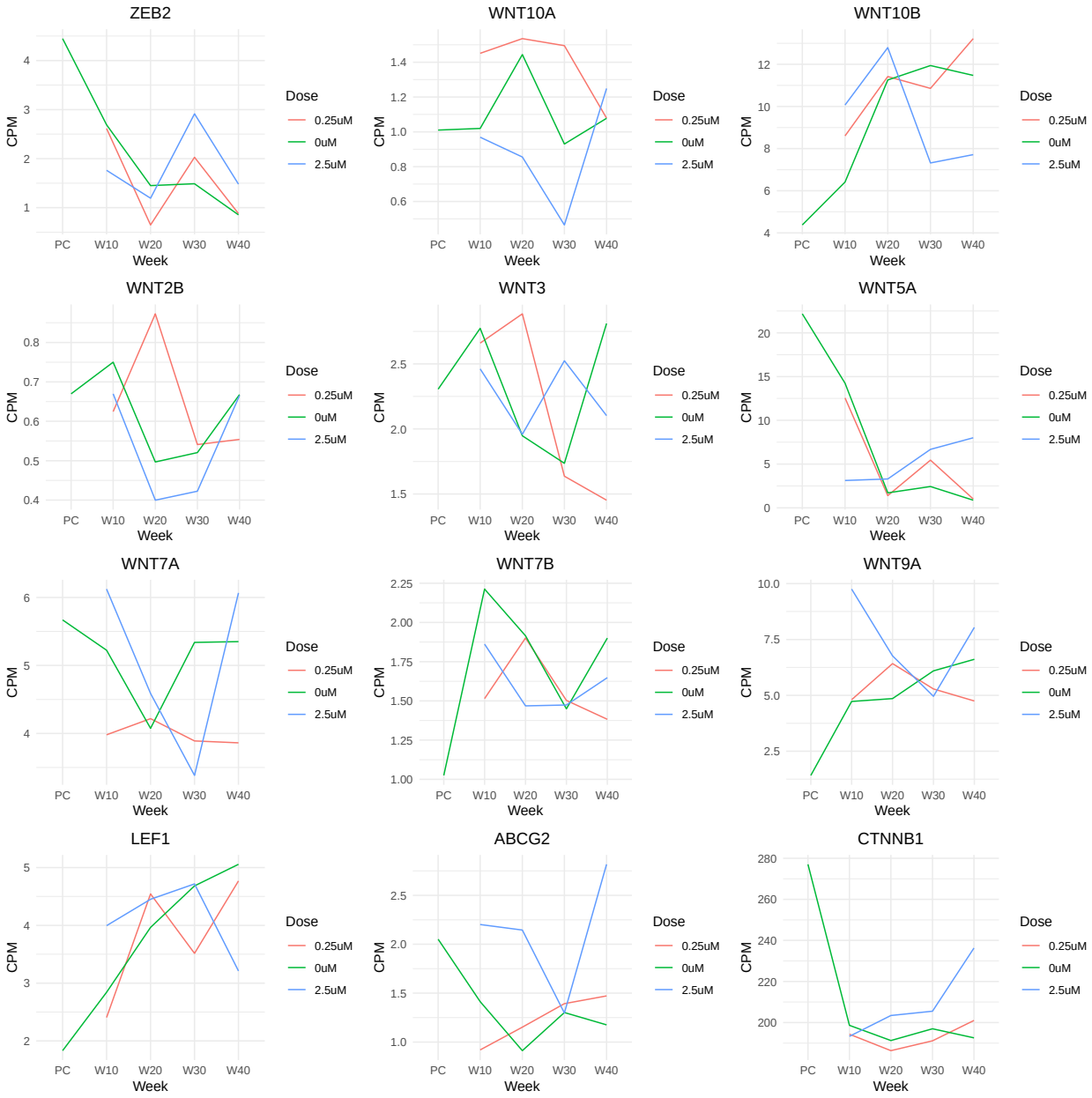

FBXW7
GAPDH
CTNNB1
NOTCH2
HES1
STAP2
CCND1
EPCAM
AMOTL2
YAP1
BRCA1
ITGA6
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

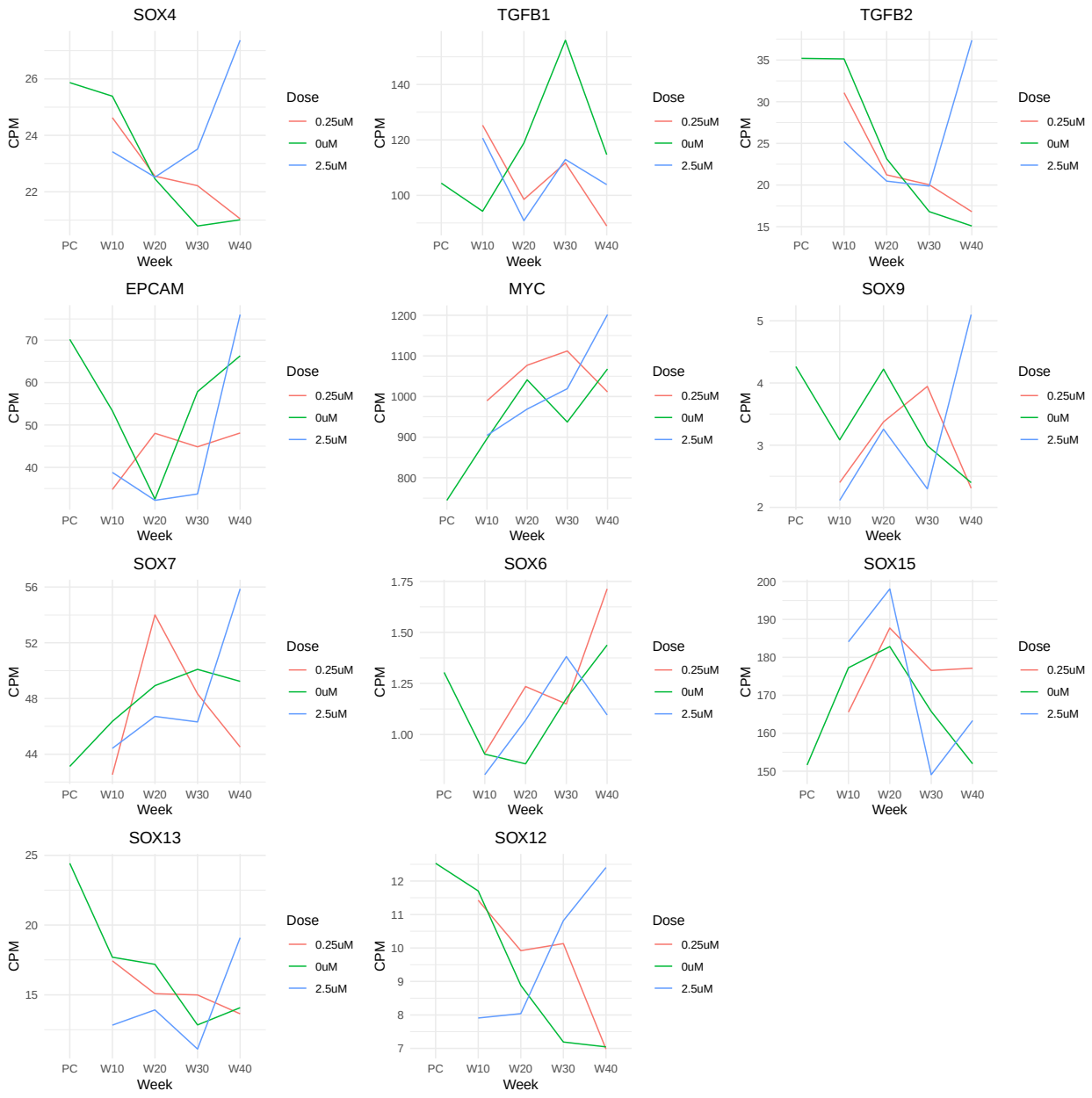

UXT
SOCS3
IL6R
EZH2
AKT1
NOTCH1
DLL1
GATA3
NUMB
BCL2
TP63
ERBB2
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM



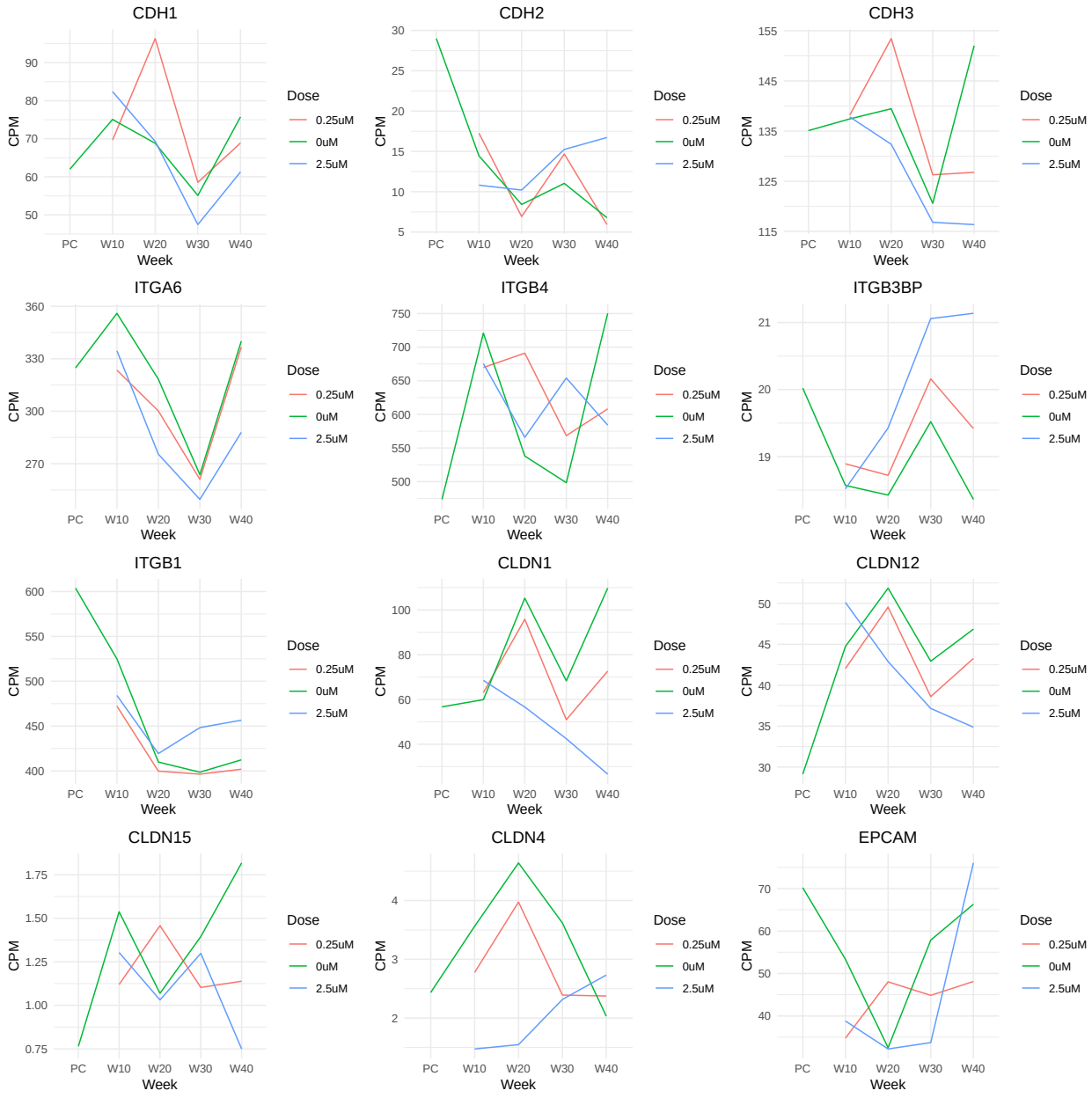
Supplemental Figure 4.1: Visualization of the expression data for all 76 genes of interest in line plots.

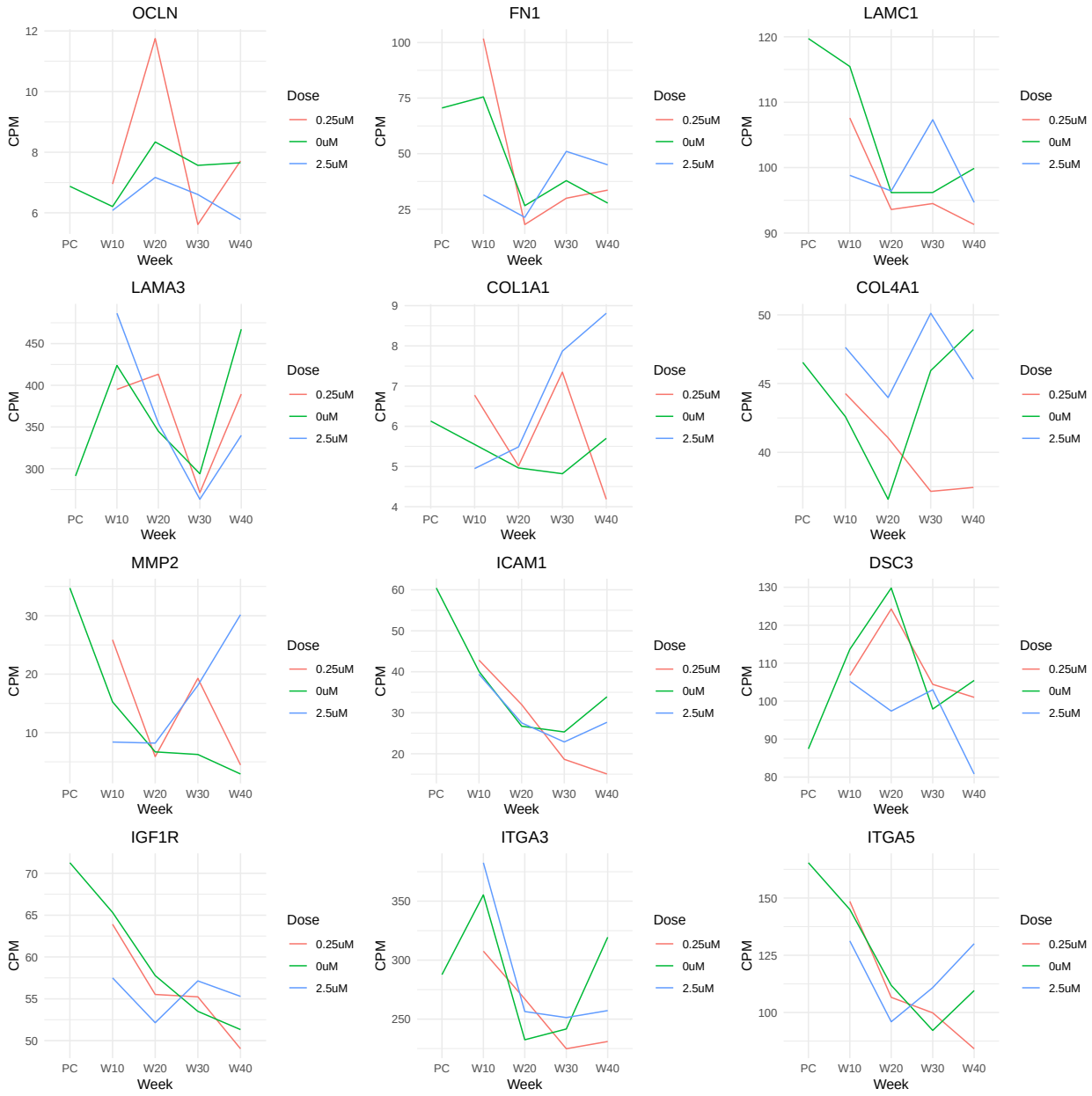

CD44
CD24
ALDH1A3
QSOX2
KLF4
NOTCH1
NOTCH2
POU2F1
SNAI1
SNAI2
BMI1
ZEB1
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM


ZEB2
WNT10A
WNT10B
WNT2B
WNT3
WNT5A
WNT7A
WNT7B
WNT9A
LEF1
ABCG2
CTNNB1
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

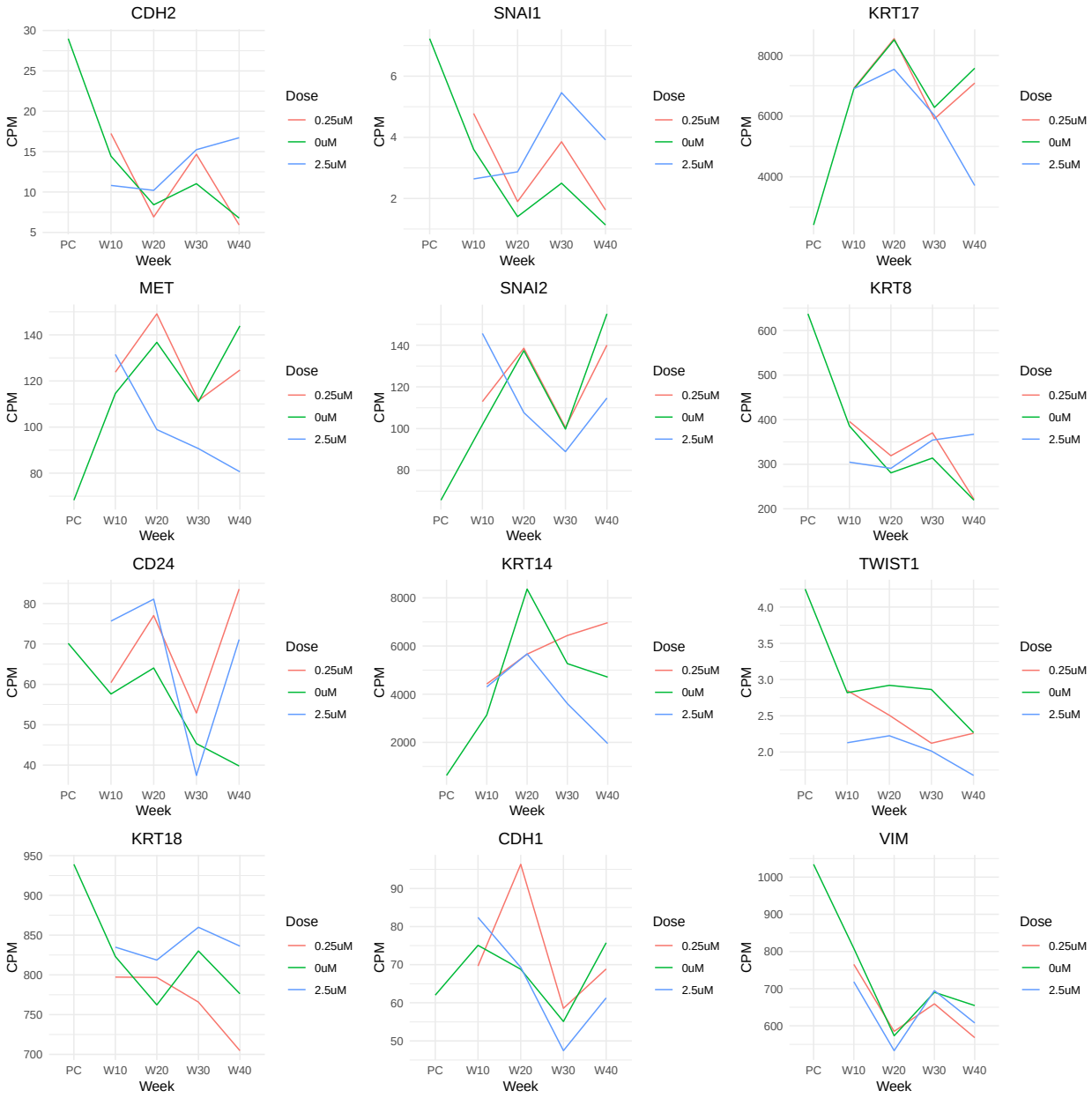


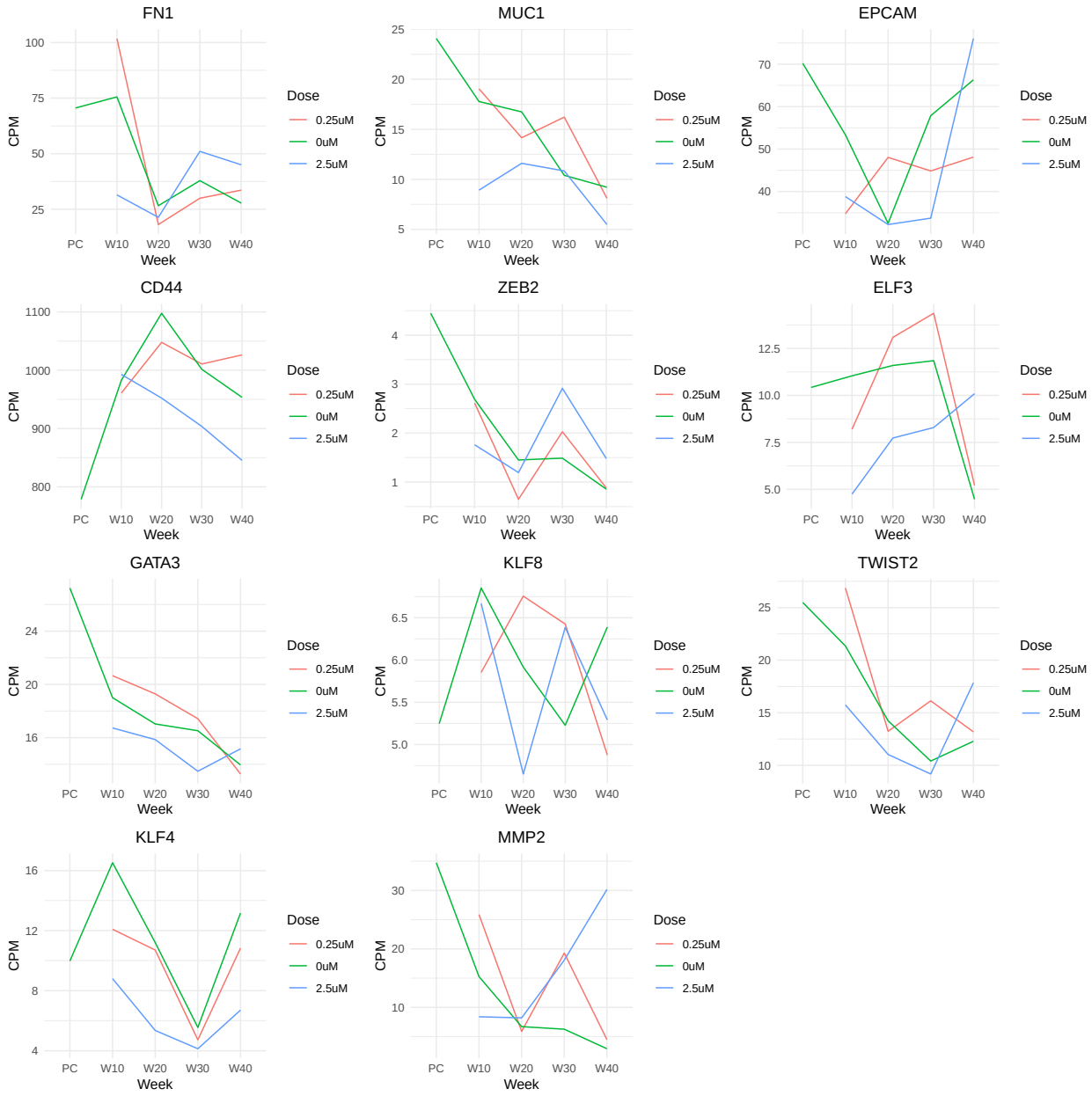
Supplemental Figure 4.2: Line plots for the grouping of gene markers for Stemness.


CDH1
CDH2
CDH3
ITGA6
ITGB4
ITGB3BP
ITGB1
CLDN1
CLDN12
CLDN15
CLDN4
EPCAM
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

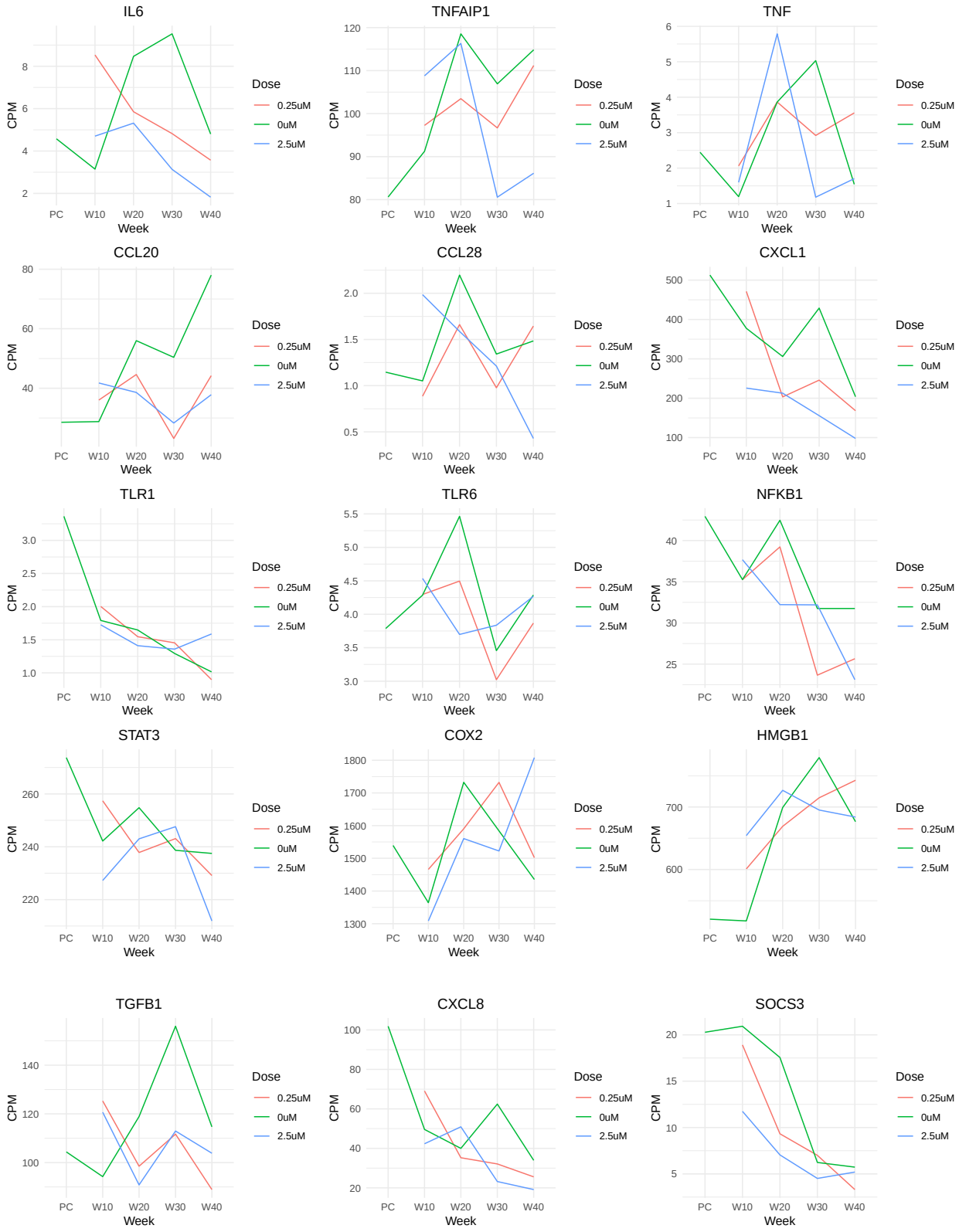


Supplemental Figure 4.3: Line plots for the grouping of gene markers involved in Cell Adhesion.

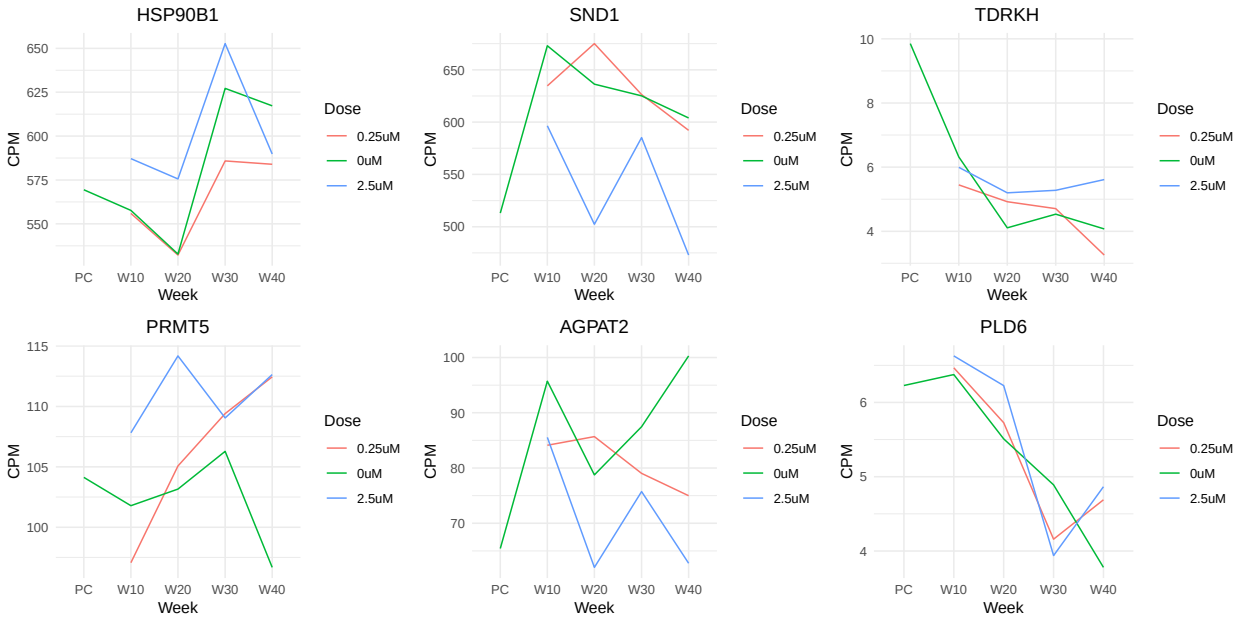

CDH2
SNAI1
KRT17
MET
SNAI2
KRT8
CD24
KRT14
TWIST1
KRT18
CDH1
VIM
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM



Supplemental Figure 4.4: Line plots for the grouping of gene markers involved in EMT.


IL6
TNFAIP1
TNF
CCL20
CCL28
CXCL1
TLR1
TLR6
NFKB1
STAT3
COX2
HMGB1
TGFB1
CXCL8
SOCS3
CPM
Week
PC
W10
W20
W30
W40
Dose
0.25uM
0uM
2.5uM

Supplemental Figure 4.5: Line plots for the grouping of gene markers of Inflammatory Mediators.



Supplemental Figure 4.6: line plots for the grouping of gene markers involved in the piRNA Pathway.

Supplemental Table 4.1: Media Components, Regents, and Concentration of Antibodies Used in Aim 3 Staining Protocols. This table can be viewed by following the link.

Chapter 5

Discussion

Summary and Synthesis of Research Findings

In this dissertation research, we performed three toxicological studies, examining the epigenomic, transcriptomic, and morphological changes linked to long term cadmium exposure in breast cells, with a focus on understanding how a class of small non-coding RNAs, piRNAs, may play a role in the acquisition of disease phenotypes. These studies were intended to explore whether chemical exposures can affect differential piRNA expression resulting in the development and progression of breast cancer.

In Aim 1, we demonstrated that our piRNA-related gene list resulted in the prioritization of the top 50 environmental chemicals and the top 50 diseases to further investigate the role of piRNA-related effects. This allowed us to visualize what data has been collected by other researchers and use it to generate our own hypothesis based on hundreds of literature sources we would not have been able to comb through ourselves. As more and more scientists publish their ‘omics data, this database will likely grow to promote the understanding of mechanisms underlying environmentally influenced diseases. Although we provide a novel way to use the resource, in this aim, we uncovered some of the challenges and limitations to using this database. Information is manually curated into CTD, therefore inconsistencies of factors such as

chemical name, disease name, and disease category can occur. Additionally, in CTD, relationships that show “no affects” and not indicated are included in the database. This makes it difficult to determine whether a chemical or gene has not been studied, or whether there is just no relationship between the two. Finally, while interrogating the publications listed to show interactions between chemicals and genes, we recognized that certain papers would be used as references for interactions, however, the actual gene would not be studied or even referenced in the actual paper and only included in RNA sequencing. Therefore, we believe there may be some publication bias indicating more interactions than there actually are.

In Aim 2, we provided baseline profiles of piRNAs in non-tumorigenic MCF10A cells and cancerous MCF7 cells in both 2D - monolayers and 3D - mammospheres. These profiles, to our knowledge, are the first of their kind to be generated using sodium periodate treatment and investigated in differentiated states. Here, we show not only that piRNAs are present in different breast cell lines, but they change in different cell culture conditions such as monolayer and mammospheres. Importantly, mammospheres are used in research to study mammary stem cell biology through identification of cancer stem cells and can quantify cancer stem cell activity. Therefore, the baseline piRNA profiles of these two different cell lines in mammosphere formation represent differentiation states of breast cancer cells. We also show the distribution of piRNA lengths between the cell lines and culture conditions, which may indicate overlapping piRNA transcripts that are being identified as a single long piRNA transcript. Therefore, our total number of piRNA transcripts may be slightly higher and these overlapping transcripts may annotate back to different regions. These longer piRNAs

are only detected in the MCF10A-MS sample, indicating that the longer piRNAs may play a role in cancer stem cell activity. Further, our results show that where in the genome piRNAs are derived from depends on differentiation state. For example, we see MCF10A MS piRNA are mostly derived from exons, whereas MCF10A ML piRNAs are mostly derived from genes (Figure 3.8). These results lay the foundation for future studies on piRNA exposure and pave the way for further investigation into piRNA profiles in other breast cancer subtypes.

In Aim 3 we examined the phenotypic and morphological effects of chronic, low dose cadmium exposure on normal breast epithelial cell line, MCF10A. Here, we profiled key features of these shifted cells including their differentiation state and their acquisition of cancer stem cell-like properties using typical markers for cellular plasticity and stemness. Our results demonstrated a phenotypic shift of untreated MCF10A cells from highly luminal (high KRT8 expression) to more basal (KRT14 expression) over 40 weeks. Additionally, we identified a population of “hybrid” cells experiencing high cellular plasticity after long term exposure to low dose cadmium using KRT8 (luminal) markers and KRT14 (basal) markers. RNA seq data also allowed us to investigate the pathways being affected by the 40-week cadmium exposure, identifying possible mechanisms underlying cadmium induced breast cancer. This work, to our knowledge, was the first long term *in vitro* cadmium exposure to provide multiple time points. Cells were frozen back at least once a week for 40 weeks. This provides an abundant number of samples at each time point to further investigate different effects of long-term cadmium exposure. Although we identified some unique patterns to the long-term, low dose cadmium exposure, the three biological replicates actually provided multiple conclusions to the

results of chronic cadmium exposure. We know that MCF10A express mostly KRT8, luminal, markers, however the cells are heterogeneous, and they are heavily dependent on their neighbors. Therefore, how the cells are grown up and how they are frozen back could determine how they react to cellular conditions, including exposure to chemicals. This could explain why we saw such different conclusions in the biological replicates.

Our immunostaining and RNA sequencing data highlight several ways that cadmium exposure may impact some of the key characteristics of carcinogenesis. Our immunostaining data indicate a hybrid population of cells expressing both KRT8 (luminal marker) and KRT14 (basal marker), previous studies have associated these hybrid cells with increased metastatic potential and stemness characteristics (Jolly et al. 2015, Grosse-Wilde et al. 2015). Additionally, our cluster results indicate the role of the MYC, an oncogene known to play a significant role in regulating the cell cycle, in our RNA sequencing results. These results taken together implicates one of the key characteristics of carcinogenesis, deregulation of cell cycle control, in long term, low dose cadmium exposure. Our RNA sequencing results show DGE of numerous genes involved in EMT, MET, and luminal to basal shift which implicates another key characteristic of carcinogenesis, invasion and metastasis.

In summary, Aim 1 identifies the top 50 environmental chemicals interacting with piRNA-related genes, and Aim 2 establishes a baseline profile of piRNA in MCF10A and MCF7 cells. Together, these aims allow us to prioritize our investigation of the impact of environmental chemicals on piRNA expression in breast cancer. With the baseline piRNA profiles provided in Aim 2, we can compare any piRNA expression data from either MCF10A cells or MCF7 cells to our baseline to determine differential expression

of piRNAs. Additionally, we can look at what targets these piRNA transcripts have and further investigate DNA methylation of these samples too. The stem-like profiles from the mammospheres also allow us to compare exposure data of these cell lines to investigate whether the chemical of interest pushes the cells to a more stem-like state. Currently, to address a future direction for both Aim 2 and Aim 3, piRNA expression data has been collected and will be analyzed as soon as possible to identify differential piRNA expression after long-term, low dose exposure to cadmium.

Relevance to Human Health

piRNA is a hot field in the cancer biomarker and therapeutic world (Cai et al. 2022; Limanówka et al. 2023; Mai et al. 2020; Tan et al. 2024). However, I believe that basic steps including how piRNAs are generated, and what affects their expression and interrupts their functions have been overlooked by cancer literature. Therefore, this dissertation research took a step back to try to better understand these concepts. With conceptual understanding of where these piRNAs are being derived from and how they are being controlled can provide insights into how they promote disease development and how they might be used or managed to prevent disease progression.

Identification of hybrid populations (i.e. having characteristics of both luminal and basal) developed during the 40-week experiment indicates the role of phenotypic plasticity in long-term low dose cadmium exposure. These hybrid cells are highly vulnerable and can have been deemed a hallmark of cancer since 2022 (Hanahan, 2022). Understanding the effects of chronic cadmium exposure in the development of cancer stem-like properties, differentiation state, and transcriptional profiles will aid us in

further determining mechanistic targets. With mechanistic targets, we could mediate cadmium exposure to lessen these deleterious effects.

Impact and Innovation

Other research findings use unreliable methods to distinguish piRNA transcripts from other miRNAs and short interfering RNAs. This project uses the most rigorous and comprehensive methods currently available to characterize piRNA expression in normal MCF10A and cancerous MCF7 breast cells.

First, the investigation into the role of piRNA in cancer is an increasingly popular topic; however, current methods use inappropriate validation of piRNAs. For example, recent studies have investigated which piRNA transcripts are implicated in breast cancer and their functions, utilizing piRNABank to validate these piRNAs (Krishnan et al., 2016; Hashim et al., 2014; Wang et al., 2016). However, piRNABank does not utilize sodium periodate treatment; consequently, potential piRNA targets identified might not actually be piRNA transcripts. In addition, contradicting evidence on the expression of PIWIL proteins and piRNA in both normal tissue and cancer makes investigation into the role they play in cancer progression and metastasis extremely difficult.

Secondly, considering the vast number of chemicals we are exposed to, there is limited data on the effects of chemical or environmental factors on piRNA expression and function. Further, exploration of how these ncRNAs are affected by chemical stressors is necessary. This work will provide a visualization of the current research state of piRNA and the environment as well as a prioritization of environmental chemicals and diseases linked to piRNA related genes. Additionally, our long-term, low

dose cadmium exposure study will provide valuable insights to further investigate to determine underlying mechanisms of cadmium induced breast cancer.

Recommendation for Future Research

Future work that would further inform the conclusions drawn in Aim 1 would be a deeper analysis into the network that makes up the inference score. Examining the other genes directly involved in piRNA related effects from environmental chemicals could prioritize new mechanisms to further investigate. Additionally, the inferenced relationships prioritized by Aim 1 could be investigated through *in vitro* experiments. Future work that would further inform the conclusions drawn in Aim 2 would be to perform gene set enrichment to determine targets of the piRNA. We could then also investigate if the length of the piRNA matters in targeting genes. Another future direction would be to analyze the baseline piRNA for other subtypes of breast cancer. Finally, future work that would further inform the conclusions drawn in Aim 3 would be to conduct smRNA analysis to investigate piRNA changes in these cells. In fact, the smRNA data has been collected, and we will include its analysis in future publications. We would also like to functionally assess stemness and differentiation capacity in 3D mammospheres and organoids for the different doses across the time course. This mammosphere data has been generated for all three batches and will be analyzed in the future. Finally, to further test whether these cells were truly transformed, we would like to transplant the long-term, low dose cadmium cells into a mouse model to determine if the cells would produce a tumor.

References

Cai, A., Hu, Y., Zhou, Z., Qi, Q., Wu, Y., Dong, P., Chen, L., & Wang, F. (2022). PIWI-Interacting RNAs (piRNAs): Promising Applications as Emerging Biomarkers for Digestive System Cancer. *Frontiers in Molecular Biosciences*, *9*, 848105. https://doi.org/10.3389/fmolb.2022.848105

Grosse-Wilde, A., Fouquier d'Hérouël, A., McIntosh, E., Ertaylan, G., Skupin, A., Kuestner, R. E., del Sol, A., Walters, K.-A., & Huang, S. (2015). Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. *PLoS ONE*, *10*(5), e0126522. https://doi.org/10.1371/journal.pone.0126522

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

Limanówka, P., Ochman, B., & Świętochowska, E. (2023). PiRNA Obtained through Liquid Biopsy as a Possible Cancer Biomarker. *Diagnostics*, *13*(11), Article 11. https://doi.org/10.3390/diagnostics13111895

Mai, D., Zheng, Y., Guo, H., Ding, P., Bai, R., Li, M., Ye, Y., Zhang, J., Huang, X., Liu, D., Sui, Q., Pan, L., Su, J., Deng, J., Wu, G., Li, R., Deng, S., Bai, Y., Ligu, Y., … Lin, D. (2020). Serum piRNA-54265 is a New Biomarker for early detection and clinical surveillance of Human Colorectal Cancer. *Theranostics*, *10*(19), 8468–8478. https://doi.org/10.7150/thno.46241

Tan, L., Mai, D., Zhang, B., Jiang, X., Zhang, J., Bai, R., Ye, Y., Li, M., Pan, L., Su, J., Zheng, Y., Liu, Z., Zuo, Z., Zhao, Q., Li, X., Huang, X., Yang, J., Tan, W., Zheng, J., & Lin, D. (2019). PIWI-interacting RNA-36712 restrains breast cancer progression

and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. *Molecular Cancer*, *18*(1), 9. https://doi.org/10.1186/s12943-019-0940-3

The emerging role of the piRNA/piwi complex in cancer | *Molecular Cancer*. (n.d.). Retrieved June 19, 2024, from https://link.springer.com/article/10.1186/s12943-019-1052-9