Single-cell Assessment of Breast Cells from Genetically Diverse Individuals: A Model for Precision Toxicology to Evaluate Racial Disparities in Breast Cancer

by

Tsu Wei Tasha Gloria Thong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Toxicology)
in the University of Michigan
2022

Dissertation Committee
Associate Professor Justin A. Colacino, Chair
Associate Professor Michele L. Cote, Wayne State University
Professor Jun Li
Research Assistant Professor Laurie Svoboda

Tsu Wei Tasha Gloria Thong

tashagth@umich.edu

ORCID: 0000-0001-5086-4358

Dedication

To every little kid (or adult) out there who loves to ask questions and wants to know how the world works. May you never lose your curiosity and wonder.

Acknowledgements

First, I would like to thank my advisor Dr. Justin Colacino for his unwavering support and belief in me. His mentorship struck the perfect balance of challenging me to think independently and problem solve while being reliably available to help whenever I needed it. I truly believe that this backdrop allowed me to dream up crazier experiments than I could have imagined at the start of my PhD and be courageous enough to try them. I will be forever grateful that he took a chance on me and let me work in the lab when I had never cultured a cell or written a single line of code before. I am thankful for his continued investment in my academic growth and achievement, and dare I say I even like coding now? Lastly, I want to thank him for creating a collaborative and fun lab environment where I was comfortable to be myself around my colleagues, as well as encouraging us to have lives beyond the lab. I look forward to being collaborators and friends in the future.

I also want to thank my other committee members Dr. Jun Li, Dr. Michele Cote, and Dr. Laurie Svoboda for their continued guidance, time, and expertise over the course of my academic journey. Their feedback and support have been integral to my success and completion of my dissertation. I would also like to give a huge thank you to the members of the Colacino lab, past and present. Their infectious laughter and conversation made long lab days pass quickly. I especially want to thank lab members Katelyn Polemi and Lauren Middleton for all their help establishing the Komen samples and keeping me company on some of those especially late nights. Thank you to Dr. Michael Brooks in the Wicha lab for helping me run my single cell experiments. I also want to thank my MIDAS buddies Dr. Yutong Wang, Dr. Alex Vargo, and Dr. Umang Varma for encouraging me, helping me debug code, and answering my endless questions during the beginning of my computational journey. Thank you to Dr. Kelly Bakulski for answering questions and providing code to guide me through my DNA methylation analyses. Thank you to all of my fellow doctoral students across the school of public health and beyond who have commiserated as well as celebrated with me during the ups and downs of this journey. Thank you to all of my dear friends across the globe and my housemates who have facilitated the life aspect of work-life balance.

And finally, a special thank you to my mom Aileen Thong, dad Chew Thong, and sister Tammie Thong for their unconditional love and support spanning continents. Although we have been living in different time zones for 10 years now, they have always made time to support me from afar and listened to me cry over countless video calls. Their belief in me through all of my doubts and struggles in the program and moments when I did not believe in myself allowed me to persevere and cross the finish line. I am especially thankful to my mom who had me making chromosomes out of play doh at age six and instilled in me a love of science for as long as I have known. Thank you for letting me crawl around in the dirt catching earthworms and grasshoppers and allowing me the curiosity and freedom to wonder about the world around me.

Table of Contents

List of Figures

List of Tables

Abstract

Background: Breast cancer is the most diagnosed cancer and the leading cause of cancer death among women worldwide. In the US, women of African ancestry have significantly worse cancer outcomes at every age group, yet the biological basis underlying this disparity is unknown. African American communities in the US are disproportionately exposed to environmental toxicants. Of these, Bisphenol-A (BPA) and its analogues are of interest due to their ability to alter mammary gland morphogenesis and stemness *in vivo* and *in vitro.* Together, these disparities highlight the need for better methods of precision toxicology to model environmental carcinogenesis of mammary cells from genetically diverse women.

Specific Aims: The goal of this dissertation was to characterize differences in normal breast stem cell biology between women of African (AA) and European (EA) ancestry and assess the effect of environmental stressors on normal breast stem cells to provide insight into mechanisms driving racial disparities in breast cancer through the following aims 1) Establish normal mammary cell cultures in the stem cell enriching conditional reprogramming (CR) culture and characterize its effects using single cell RNA-sequencing (scRNA-seq) 2) Quantify transcriptomic, epigenomic, and genotypic differences between normal mammary stem cells from AA and EA women grown in CR culture using scRNA-seq, DNA methylation analysis, and SNP genotyping 3) Quantify transcriptomic effects of bisphenol toxicant exposure on normal mammary stem cells from genetically diverse women using scRNA-seq.

Results: Results from Aim 1 indicate that the CR culture retains the epithelial cell lineages of the breast, luminal and myoepithelial, as well as inter-individual heterogeneity in gene expression. Additionally, CR mammary cells differentially express breast cancer and stem cell associated genes, exhibit a more developmentally immature transcriptomic phenotype, and promote the emergence of a unique hybrid stem cell population following reprogramming. In Aim 2 we identified differentially expressed genes (DEGs) between luminal (n= 639 genes) and myoepithelial (n = 483 genes) conditionally reprogrammed AA and EA cells. Of these DEGs, 8 genes in the luminal subset and 13 genes in the myoepithelial subset significantly overlapped

with genes which have been previously identified as differentially expressed in tumors from The Cancer Genome Atlas. By integrating gene expression data with DNA methylation results, we identified differentially methylated CpG sites on DEGs in luminal (n=221 CpGs) and myoepithelial (n=55) cells. Integrating genotyping data from our samples and GTEx ALFA alleles, we identified ancestry associated eQTLs for a number of differentially expressed genes. Analyses from Aim 3 show that Bisphenols (-A, -S, and -F) elicit gene expression changes on normal mammary cells across a range of human relevant doses, especially at 25 μM and elicit their own distinct gene expression signatures. Genetically diverse individuals display distinct gene expression differences in response to bisphenols, potentially indicating inter-individual susceptibility.

Conclusions: Our establishment of samples from diverse donors in stem cell enriching conditions provides a novel, *ex vivo* system for modeling environmental carcinogenesis and capturing inter-individual heterogeneity. The integrated use of single-cell transcriptomic analyses, epigenomic profiling, and genotyping, provide a systems approach to gain precise insights into the molecular biology of normal mammary stem cells and the influences of gene-environment interactions on stem cell biology. We hope the mechanistic insights gained from this study will help to achieve the long-term goals of reducing breast cancer incidence and disparities and aid in the development of better targeted therapies.

Chapter 1

Introduction

Breast Cancer as a Global Public Health Issue

Breast cancer is the most commonly diagnosed cancer in the world, with an estimated 2.3 million new cases accounting for 11.7% of all global cancer incidence in 2020.[1] Global rates of incidence have continuously increased in all age groups, and similarly, rates of mortality have increased in most age groups and regions.[2] These staggering numbers highlight the pressing need for public health action now more than ever to alleviate the global burden of breast cancer and alter the course of the upward trends in incidence and mortality.

Breast Cancer Subtypes

Breast cancer is a vastly heterogenous disease with prevalence of the four major molecular subtypes heterogeneously distributed across the globe. Developed originally from microarray gene expression data and clinically identified using immunohistochemistry, the four major subtypes are now primarily determined by the 50 gene PAM50 classifier and include Luminal A (Lum A), Luminal B (Lum B), HER2 positive, and basal/ triple negative breast cancer (TNBC).[3,4] The luminal subtypes are characterized by hormone positive expression, specifically for estrogen receptor (ER) and progesterone (PR). Lum A accounts for the largest proportion of all breast cancer (40%) compared to Lum B (20%), with the major differences being higher expression of proliferation and cell cycle genes, lower PR expression, and higher tumor grade in Lum B. HER2+ cancers account for approximately 15-20% of breast cancer cases and are characterized by expression of the human epidermal growth factor receptor-2 (HER2) protein, a proliferative phenotype, and poor prognosis.[5,6] Lastly, while not perfectly synonymous, basal and TNBC are estimated to account for 15% of breast cancer cases, and are characterized as highly aggressive and proliferative tumors with a high rate of metastasis.[5] Triple negative cancers refer to those classified as lacking ER, PR, and HER2 protein expression through immunohistochemical classification.

While these classifications have transformed the way breast cancers are diagnosed and treated, advances in gene expression profiling and a growing increase in genomic breast cancer datasets point to vast heterogeneity even within subtypes, with the TNBC subtype exhibiting the greatest genetic diversity and additional subgroups.[6] This heterogeneity in gene expression across breast tumors highlights the need for continued characterization and profiling of breast tumors for better diagnostic and therapeutic gene targets.

Breast Cancer Disparities

Disparities in breast cancer have been widely characterized and are attributed to a multifactorial mix of biological, socio-economic, lifestyle, environmental, geographic, and access factors.[7,8] Global trends show a rising increase in breast cancer incidence especially in low and middle income countries (LMICs), and mortality rates are also higher in these LMICs.[9] Within the US, African American (AA), Hispanic American, and American Indian women have a higher likelihood of being diagnosed with advanced stage breast cancer and have worse survival outcomes than European American (EA) women.[10,11] TNBC, the most aggressive subtype, is a striking example of this. Women of African ancestry are 2-3 times more likely to develop TNBC than women of European ancestry and have worse clinical outcomes, yet the biological basis underlying this disparity is still unknown.[7,12]

While the biological basis for these disparities is elusive, there are well characterized non-biological factors which contribute to these outcomes. Together, the intersection of these social determinants of health –race , poverty, socioeconomic, and structural factors – compound to significantly inflate disparities. The 40% higher mortality rate observed in AA compared to EA women is largely attributed to differences in access to treatment and high quality care linked to poverty and socioeconomic status.[13] Lack of insurance, transportation, health literacy, and ability to take time away from work are just a handful of barriers which influence access to treatment and likely cause delays in diagnosis and more advanced stage distribution which negatively impact survival outcomes.[13,14] Lifestyle factors such as obesity and diet are directly tied to poverty and socioeconomic status which have been linked to the rising incidence rates in AA women. While breast cancer outcomes are multifactorial and include a mix of non-biological and biological differences such as subtype, biological differences likely reflect a combination of

genetic influences, lifestyle, and environmental factors, highlighting the need for a more integrated and holistic view of breast cancer disparities.[8]

An important distinction to note is the use of self-reported race versus genetic ancestry in the majority of racial disparity studies in breast cancer. While self-reported AA identity correlates closely to African ancestry, this socially constructed category encompasses a vast array of genetically diverse individuals with both African ancestry and in some cases substantial non-African ancestry.[15] Additionally, women with West African ancestry have significantly higher prevalence of ER negative and TNBC compared to women with East African ancestry.[16,17] Due to this, it is likely that using self-reported AA and EA race as monolithic comparison groups are not precise enough to capture nuanced differences in ancestry which may account for functional differences in breast cancer disparities.

Stem Cells and Carcinogenesis

In 1838, German physiologist Johannes Muller described tumors as abnormal continuations of embryonic development based on morphological similarities.[18] Centuries later, an extensive body of experimental evidence has solidified this proposed link between cancer and embryonic development. Stem cells are defined by two key characteristics, *self-renewal*, the ability to produce offspring cells with identical potency to themselves, and *differentiation potential*, the ability divide and produce mature cells through differentiation.[19] These capabilities along with rapid proliferation, invasion, migration, and angiogenesis are crucial in transforming a single fertilized embryo into a mature organism composed of numerous cell types and complex tissues.[20] Cancer cells have been observed to exhibit these same behaviors during progression and metastasis, and even more strikingly, oncogenesis and embryonic development have been shown to share key regulatory pathways Wnt, Notch, and Hedgehog.[19–21] Because of these similarities in behavior and regulation, cancer is considered a disease of "dysregulated development", where cancer stem cells (CSCs), a subset of cancer cells with stem like characteristics, have been proposed to drive tumor initiation, progression, and long term growth.[22,23] Outside of development and cancer, stem cells play a key role in adult tissues, differentiating and replacing cells in response to "wear and tear" and tissue injury.[24,25] Adult tissue stem cells have been characterized in numerous cell types including the breast, and are proposed targets of carcinogenic transformation due to their slow dividing and long-lived nature.

Physiology of the Human Breast and Mammary Stem Cells

The mammary gland is a unique and dynamic organ which fully matures after birth and undergoes structural changes over the life course during *in utero* development, puberty, pregnancy, lactation, and involution.[26] A tubular, bi-layered structure, the mammary gland is composed primarily of two cell types distinct in morphology and function— luminal and basal/myoepithelial cells.[27] The luminal epithelium is composed of cells that form ducts and secretory alveoli and is encased by the myoepithelium which contracts during lactation, squeezing milk from alveolar cells.[27] Due to the fact that the mammary gland undergoes major structural remodeling during distinct developmental time points, these periods are often considered "windows of susceptibility", where exposure to environmental stressors may prove to be especially deleterious.[28]

Mammary stem cells (MSCs) play a crucial role in the functional remodeling during these developmental windows due to their plasticity and differentiation potential.[29] Seminal rodent studies demonstrated the differential potential of MSCs, showing that an entire mammary gland can be reconstituted *in vivo* from mammary epithelial fragments and even from a single stem cell.[30,31] Additionally, serial transplantation of clonal outgrowths in these studies confirmed the self-renewal capabilities of these MSCs. The discovery of MSC markers has been crucial to their isolation and characterization. In studies of bipotent MSCs and progenitors, these populations have been characterized as expressing CD44+/CD24- and CD49f+/EpCAM+ surface marker phenotypes and studies of normal MSCs and progenitors have found high expression of enzymatic marker aldehyde dehydrogenase 1 (ALDH1).[32–34]

Recent studies have found compelling ancestry-specific differences in gene expression and behavior of normal and cancerous mammary cells, strengthening the possibility that there is an inherited genetic component to breast cancer disparities between AA and EA women.[35] In addition to being a putative marker of both normal and malignant stem cells, ALDH1 expression has been associated with aggressive, high grade tumors and poor clinical outcomes.[36] In a study of benign and cancerous breast tissue from Ghanaian women, immunohistochemical analysis revealed that a significantly higher proportion of individuals with benign breast conditions (n=40, 58%) had tissues expressing stromal ALDH1 expression relative to individuals with stromal ALDH1 expressing cancerous tissue (n=44, 42.3%; p=0.043).[35]Additionally, when

malignant tumors were grouped by subtype, expression of ALDH1 was highest in TNBCs. Considering the high incidence of aggressive TNBCs among western sub-Saharan African women and the known association of ALDH1 with tumor aggressiveness and future breast cancer risk, stromal ALDH1 expression of benign tissue may be predictive of cells most at risk for oncogenic transformation to TNBCs.

Breast Cancer, Mammary Stem Cells, and the Environment

While approximately 10% of breast cancer diagnoses are attributable to known hereditable risk factors, such as BRCA mutations, the vast majority (70-90%) of cases occur in women with no family history, suggesting that extrinsic factors are primarily responsible for lifetime cancer risk.[37–39] The implications of this for public health intervention are hopeful—environmental factors are characterizable and modifiable, and thus present tangible targets for breast cancer prevention. African American communities in the US are disproportionately exposed to toxic environmental chemicals and focusing prevention efforts on environmental factors has potential to reduce racial disparities in breast cancer. Understanding the relationship between environmental factors, genetics, and breast cancer risk has proven extremely complex. There are a vast number of exposures and mixtures in the environment and heterogeneity in these exposures by individual, however, there is a lack of human relevant model systems that capture inter-individual heterogeneity available to assess these exposures. Mammary stem cells present a promising means to study gene-environment interactions due to their implicated role in carcinogenesis, the ability to identify and experimentally culture them, and emerging evidence that shows stem cell proportions in normal tissue to be highly variable by individual.[40,41]

Epigenetics, defined as heritable changes in gene function that do not entail a change in DNA sequence, is a key regulatory mechanism involved in cancer and development and a promising means for characterizing gene-environment interactions in mammary stem cells.[42] During prenatal development, puberty, pregnancy, lactation, and menopause, mammary stem cells sense environmental signals which trigger epigenetic alterations that drive the structural and functional remodeling of the breast.[43,44] Due to the rapid proliferation, major restructuring, and hormone signaling during these windows, it is hypothesized that mammary stem cells are particularly susceptible to environmental insult during these periods. Therefore, it is crucial to characterize if and how environmental exposures alter the epigenome of mammary stem cells.

Although we are exposed to an overwhelming number of environmental exposures daily, population level studies such as the National Health and Nutrition Examination Survey (NHANES) of the US population are a valuable resource for prioritizing exposures for experimental assessment. NHANES biomonitoring data has measured endocrine disrupting chemicals (EDCs), defined as chemicals that interfere with hormone function, ubiquitously across the US population and at disproportionately high levels in Non-Hispanic black and low income individuals.[45,46] Of these, bisphenol-A (BPA) and its analogues, are of particular interest in relation to breast cancer due to their classification as xenoestrogens and ability to alter mammary gland morphogenesis and stemness *in vivo* and *in vitro.*

An *in vitro* study of MCF-7 ER-positive cancer cells found that compared to the vehicle control, cells exposed to 10nM BPA showed increased mRNA and protein expression of ALDH1, and SOX2, known to be involved in pluripotency and self-renewal.[47] Additionally, the authors exposed mammospheres derived from MCF-7 cells, patient-derived xenografts, and normal mammary mouse cells to 10nM BPA and found an increase in mammosphere size in all three cell types. The increased expression of ALDH1 and SOX2 and the increase in mammosphere size due to BPA exposure led the authors to conclude that 10nM BPA enhances stem cell populations *in vitro*. In an *in vivo* study, the offspring of pregnant CD-1 mice injected with 25μg/kg BPA during E8.5-E18.5 were found to have significant mammary defects compared to those dosed with a sesame oil vehicle control.[48] Furthermore, the most severe defects were observed during E12.5-16.5, the critical time in development when the ERα negative mammary bud is completely surround by the ERα positive stroma. These observations indicate that timing of exposure and stromal interactions influence severity of BPA induced mammary defects. A study of second trimester pregnant women in northern and central California measured urinary levels of BPA and its analytes and found total BPA measures ranging from 0.37μg/g to 1348 μg/g.[49] This range encompasses the dose of 25μg/kg BPA found to have toxic effects in CD-1 mice, implicating that human exposure to BPA may be significant enough to cause in utero alterations in mammary gland morphogenesis. In utero exposure of CD-1 mice to bisphenol analogues bisphenol-AF(BPAF) or bisphenol-S (BPS) during gestation days 10-17 resulted in significant alterations mammary gland development at doses 0.05,0.5, and 5mg/kg for both chemicals.[50] Epigenetic alterations to bisphenol exposure have also been characterized, as observed in MCF-7 breast cancer cells, where BPS was found to alter DNA

methylation in the promoter region of breast cancer related genes *CDH1*, *SFN*, and *TNFRSF10C*, as well as in transposons.[51] Additionally, exposure to low dose BPA (4nM) was found to hypermethylated the *LAMP3* gene in both MCF-7 cells and primary tumor cells.[52]

The ability of BPA and its analogues to alter mammary gland morphogenesis and stemness *in vivo* and *in vitro* makes them promising candidates for experimentation to assess the relationship between environmental exposures and mammary stem cells.

Models of Precision Toxicology for Assessment of Environmental Exposures

Understanding the relationship between environmental exposures, genetics, and normal mammary stem cell biology requires a relevant model system, and the conditional reprogramming (CR) culture developed by Liu and colleagues is a promising method for this.[53] While animal and cell line models are widely used to characterize effects of environmental toxicants, they are unable to capture the genetic diversity and inter-individual heterogeneity of primary tissue. Primary tissue is challenging to culture due to the fact that normal tissue samples obtained from healthy donors, such as from core biopsies, are usually limited both in size and expansion capacity in culture. The CR culture is a co-culture of human epithelial cells and irradiated mouse fibroblasts in the presence of the Rho kinase inhibitor (Y-27632) resulting in rapid expansion and de-differentiation of primary cells. CR culture enriches for a stem-like phenotype, maintains the cellular heterogeneity of the parent tissue, and retains the ability for CR cells to differentiate into the tissue of origin once removed from CR conditions.[53] For these reasons, and the ability to indefinitely expand normal mammary epithelial populations, CR culture is a promising model for precision toxicology and an advantageous system for characterizing gene-environment influences on mammary stem cells from diverse individuals.

Leveraging Multi-omics, Big Data, and Cutting Edge Technology to Address Breast Cancer Disparities

While novel *in vitro* models such as CR culture show immense promise, the era of big data and -omics is upon us, and as these cutting edge computational and experimental techniques evolve in parallel, the integration of the two provides even more immense potential to inform our understanding of mammary stem cells and the environment.

Advances in single-cell RNA-sequencing (scRNA-seq) approaches have made it an increasingly versatile and informative tool. A variety of studies thus far have demonstrated the utility of unbiased single-cell transcriptome profiling for the discovery of new cell states, profiling heterogenous cell and tissue mixtures, regulatory relationships between cells, mapping developmental trajectories, and more.[54–56] A study performing scRNA-seq of normal mammary cells identified clusters with expression profiles consistent with epithelial and mesenchymal populations found in the normal breast as well as an interesting hybrid epithelial/mesenchymal population with high expression of genes overexpressed in TNBC.[41] The growing number of publicly available single-cell datasets also provides great opportunity for dataset integration and comparison which is especially useful in cases where physically generating an appropriate dataset is challenging. Giraddi and colleagues have generated a transcriptome atlas of the mouse mammary gland sampled at varying stages of development spanning embryonic day 16 to adult.[54] Although an equivalent dataset does not exist for the human mammary gland, computational alignment of normal mammary cells to the mouse transcriptome atlas can provide insight into the developmental maturity of normal mammary cells. For these reasons, single-cell transcriptomic profiling provides a versatile and promising way to analyze how the molecular profiles of human mammary cells differ by ancestry and provide insight into biological factors underlying cancer disparities.

While transcriptomic profiling may reveal differences in gene expression between AA and EA women, epigenetic profiling offers insight into the mechanisms responsible for these transcriptomic differences. Of the epigenetic regulators, DNA methylation has been the most utilized to understand the relationship between environmental exposures and epigenetic reprogramming. Numerous studies using a wide range of common environmental toxicants such as air pollution, heavy metals, polycyclic aromatic hydrocarbons, BPA, and more, have shown that environmental exposures heavily influence epigenetic reprogramming.[57] Of these, BPA has been shown to epigenetically reprogram MCF-7 breast cancer cells, repressing the *LAMP3* locus through an estrogenic pathway.[52] Epigenetic reprogramming plays a key role in mammary gland development, extensively regulating gene expression of stem cells in order for functional differentiation to occur, after which differing mammary epithelial cell types exhibit distinct DNA methylation landscapes.[58] Similarly, genome wide hypomethylation has been commonly observed in breast cancers as well as hypermethylation in over 100 promoter genes, indicating

that epigenetic reprogramming is a highly involved mechanism in carcinogenesis. Thus, characterizing epigenetic patterns of mammary stem cells may be a key link in understanding mechanisms through which environmental exposures influence breast cancer disparities

Knowledge Gap and Rationale of Thesis

The overall goal of this dissertation is to characterize differences in normal breast stem cell biology between women of African and European ancestry and assess the effect of environmental stressors on normal breast stem cells from these women to provide insight into biological mechanisms driving racial disparities in breast cancer (**Figure 1.1**). This dissertation integrates a novel *ex vivo* culture of normal mammary cells, cutting edge multi-omic techniques, and an example of this model system for precision toxicology to address this knowledge gap. To date, studies using normal breast tissue, especially from primary core biopsies are limited, thus our proposed culture of normal breast tissue from epidemiologically well characterized AA and EA women collected from the Komen Normal Tissue Bank in stem cell enriching conditions is novel. Together, these advanced methods create a promising interdisciplinary strategy to tackle the complex relationship between gene-environment interactions and their influence on mammary stem cell biology and breast cancer outcomes.

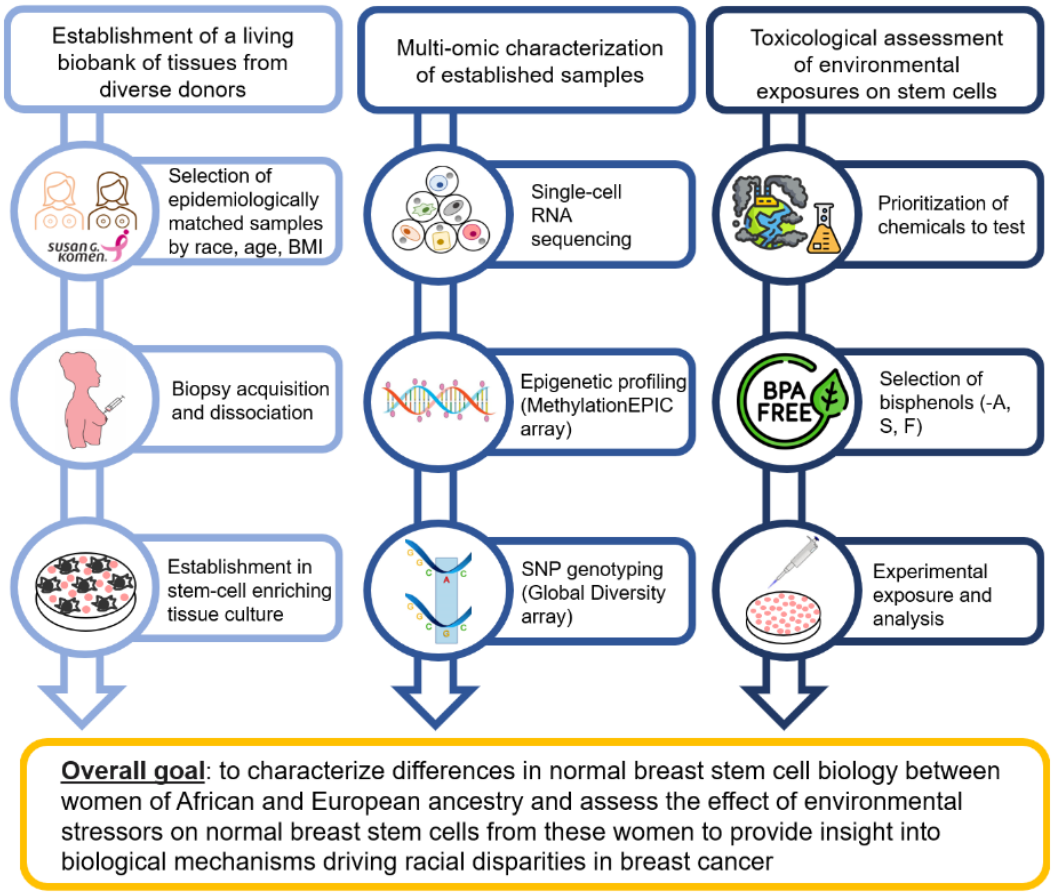


Figure 1.1 Graphical abstract of study overview and motivation.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/CAAC.21660

2. Zeinomar N, Oskar S, Kehm RD, Sahebzeda S, Terry MB. Environmental exposures and breast cancer risk in the context of underlying susceptibility: A systematic review of the epidemiological literature. *Environ Res*. 2020;187:109346. doi:10.1016/j.envres.2020.109346

3. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370

4. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

5. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol*. 2014;5(3):412. doi:10.5306/WJCO.V5.I3.412

6. Johnson KS, Conant EF, Soo MS. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *J Breast Imaging*. 2021;3(1):12-24. doi:10.1093/JBI/WBAA110

7. Dietze EC, Sistrunk C, Miranda-Carboni G, O'Regan R, Seewaldt VL. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat Rev Cancer*. 2015;15(4):248-254. doi:10.1038/nrc3896

8. Carey LA, Perou CM, Livasy CA, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *JAMA*. 2006;295(21):2492-2502. doi:10.1001/JAMA.295.21.2492

9. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev*. 2017;26(4):444-457. doi:10.1158/1055-9965.EPI-16-0858

10. Newman LA. Breast Cancer Disparities: Socioeconomic Factors versus Biology. *Ann Surg Oncol*. 2017;24(10):2869-2875. doi:10.1245/S10434-017-5977-1/TABLES/2

11. Ooi SL, Martinez ME, Li CI. Disparities in breast cancer characteristics and outcomes by race/ethnicity. *Breast Cancer Res Treat*. 2011;127(3):729-738. doi:10.1007/S10549-010-1191-6/TABLES/6

12. Newman LA, Kaljee LM. Health Disparities and Triple-Negative Breast Cancer in African American Women. *JAMA Surg*. 2017;152(5):485. doi:10.1001/jamasurg.2017.0005

13. Stringer-Reasor EM, Elkhanany A, Khoury K, Simon MA, Newman LA. Disparities in Breast Cancer Associated With African American Identity. *Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet*. 2021;41(41):e29-e46. doi:10.1200/EDBK_319929

14. Yedjou CG, Sims JN, Miele L, et al. Health and Racial Disparity in Breast Cancer. *Adv*

Exp Med Biol. 2019;1152:31. doi:10.1007/978-3-030-20301-6_3

15. Newman LA, Carpten J. Integrating the Genetics of Race and Ethnicity Into Cancer Research: Trailing Jane and John Q. Public. *JAMA Surg*. 2018;153(4):299. doi:10.1001/JAMASURG.2017.5080

16. Sung H, DeSantis CE, Fedewa SA, Kantelhardt EJ, Jemal A. Breast cancer subtypes among Eastern-African-born black women and other black women in the United States. *Cancer*. 2019;125(19):3401-3411. doi:10.1002/CNCR.32293

17. Jiagge E, Jibril AS, Chitale D, et al. Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer. *Ann Surg Oncol*. 2016;23(12):3843-3849. doi:10.1245/S10434-016-5420-Z/TABLES/5

18. Capp J-P. Cancer Stem Cells: From Historical Roots to a New Perspective. *J Oncol*. 2019;2019:1-10. doi:10.1155/2019/5189232

19. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. doi:10.1038/35102167

20. Ma Y, Zhang P, Wang F, Yang J, Yang Z, Qin H. The relationship between early embryo development and tumourigenesis. *J Cell Mol Med*. 2010;14(12):2697-2701. doi:10.1111/j.1582-4934.2010.01191.x

21. Reya T, Clevers H. Wnt signalling in stem cells and cancer. *Nature*. 2005;434(7035):843-850. doi:10.1038/nature03319

22. Papaccio F, Paino F, Regad T, Papaccio G, Desiderio V, Tirino V. Concise Review: Cancer Cells, Cancer Stem Cells, and Mesenchymal Stem Cells: Influence in Cancer Development. *Stem Cells Transl Med*. 2017;6(12):2115-2125. doi:10.1002/sctm.17-0138

23. Clevers H. The cancer stem cell: premises, promises and challenges. *Nat Med*. 2011;17(3):313-319. doi:10.1038/nm.2304

24. Goichberg P. Current Understanding of the Pathways Involved in Adult Stem and Progenitor Cell Migration for Tissue Homeostasis and Repair. *Stem Cell Rev Reports*. 2016;12(4):421-437. doi:10.1007/s12015-016-9663-7

25. Shaw TJ, Martin P. Wound repair: a showcase for cell plasticity and migration. *Curr Opin Cell Biol*. 2016;42:29-37. doi:10.1016/J.CEB.2016.04.001

26. Hennighausen L, Robinson GW. Signaling Pathways in Mammary Gland Development. *Dev Cell*. 2001;1(4):467-475. doi:10.1016/S1534-5807(01)00064-8

27. Macias H, Hinck L. Mammary Gland Development. *Wiley Interdiscip Rev Dev Biol*. 2012;1(4):533. doi:10.1002/WDEV.35

28. Russo J. The Windows of Susceptibility to Breast Cancer. *Pathobiol Breast Cancer*. Published online 2016:1-20. doi:10.1007/978-3-319-40815-6_1

29. Mammary Stem Cells: Premise, Properties, and Perspectives. *Trends Cell Biol*. 2017;27(8):556-567. doi:10.1016/J.TCB.2017.04.001

30. Shackleton M, Vaillant F, Simpson KJ, et al. Generation of a functional mammary gland from a single stem cell. *Nature*. 2006;439(7072):84-88. doi:10.1038/nature04372

31. Daniel CW, Deome KB, Young JT, Blair PB, Faulkin LJ. The in vivo life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *J Mammary Gland Biol Neoplasia*. 2009;14(3):355-362. doi:10.1007/s10911-009-9139-3

32. Stingl J, Eaves CJ, Zandieh I, Emerman JT. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat*. 2001;67(2):93-109. doi:10.1023/A:1010615124301

33. Ghebeh H, Sleiman GM, Manogaran PS, et al. Profiling of normal and malignant breast tissue show CD44high/CD24low phenotype as a predominant stem/progenitor marker when used in combination with Ep-CAM/CD49f markers. *BMC Cancer*. 2013;13:289. doi:10.1186/1471-2407-13-289

34. Ginestier C, Hur MH, Charafe-Jauffret E, et al. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell*. 2007;1(5):555-567. doi:10.1016/j.stem.2007.08.014

35. Schwartz T, Stark A, Pang J, et al. Expression of aldehyde dehydrogenase 1 as a marker of mammary stem cells in benign and malignant breast lesions of Ghanaian women. *Cancer*. 2013;119(3):488-494. doi:10.1002/cncr.27737

36. Jiagge E, Chitale D, Newman LA, Newman A. Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *J Pathol*. 2018;188(2). doi:10.1016/j.ajpath.2017.06.020

37. Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature*. 2016;529(7584):43-47. doi:10.1038/nature16166

38. Gray J, Evans N, Taylor B, Rizzo J, Walker M. State of the Evidence: The Connection Between Breast Cancer and the Environment. *Int J Occup Environ Health*. 2009;15(1):43-78. doi:10.1179/107735209799449761

39. Forman MR, Winn DM, Collman GW, Rizzo J, Birnbaum LS. Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod Toxicol*. 2015;54:6-10. doi:10.1016/j.reprotox.2014.10.019

40. Nakshatri H, Anjanappa M, Bhat-Nakshatri P. Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. *Sci Rep*. 2015;5:13526. doi:10.1038/srep13526

41. Colacino JA, Azizi E, Brooks MD, et al. Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*. Published online 2018. doi:10.1016/j.stemcr.2018.03.001

42. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med*. 2009;27(5):351-357. doi:10.1055/s-0029-1237423

43. Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol Rev*. 2020;100(2):489-523. doi:10.1152/PHYSREV.00040.2018/ASSET/IMAGES/LARGE/Z9J0012029270010.JPE

G

44. Terry MB, Michels KB, Brody JG, et al. Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research. *Breast Cancer Res 2019 211*. 2019;21(1):1-16. doi:10.1186/S13058-019-1168-2

45. Attina TM, Malits J, Naidu M, Trasande L. Racial/ethnic disparities in disease burden and costs related to exposure to endocrine-disrupting chemicals in the United States: an exploratory analysis. *J Clin Epidemiol*. 2019;108:34-43. doi:10.1016/j.jclinepi.2018.11.024

46. Ruiz D, Becerra M, Jagai JS, Ard K, Sargis RM. Disparities in Environmental Exposures to Endocrine-Disrupting Chemicals and Diabetes Risk in Vulnerable Populations. *Diabetes Care*. 2018;41(1):193-205. doi:10.2337/dc16-2765

47. Lillo MA, Nichols C, Seagroves TN, Miranda-Carboni GA, Krum SA. Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Horm Cancer*. 2017;8(2):90-99. doi:10.1007/s12672-017-0286-5

48. Hindman AR, Mo XM, Helber HL, et al. Varying Susceptibility of the Female Mammary Gland to In Utero Windows of BPA Exposure. *Endocrinology*. 2017;158(10):3435-3447. doi:10.1210/en.2017-00116

49. Gerona RR, Pan J, Zota AR, et al. Direct measurement of Bisphenol A (BPA), BPA glucuronide and BPA sulfate in a diverse and low-income population of pregnant women reveals high exposure, with potential implications for previous exposure estimates: a cross-sectional study. *Environ Health*. 2016;15:50. doi:10.1186/s12940-016-0131-2

50. Tucker DK, Hayes Bouknight S, Brar SS, Kissling GE, Fenton SE. Evaluation of Prenatal Exposure to Bisphenol Analogues on Development and Long-Term Health of the Mammary Gland in Female Mice. *Environ Health Perspect*. 2018;126(8):087003. doi:10.1289/EHP3189

51. Huang W, Zhao C, Zhong H, Zhang S, Xia Y, Cai Z. Bisphenol S induced epigenetic and transcriptional changes in human breast cancer cell line MCF-7. *Environ Pollut*. 2019;246:697-703. doi:10.1016/J.ENVPOL.2018.12.084

52. Weng Y-I, Hsu P-Y, Liyanarachchi S, et al. Epigenetic influences of low-dose bisphenol A in primary human breast epithelial cells. *Toxicol Appl Pharmacol*. 2010;248(2):111. doi:10.1016/J.TAAP.2010.07.014

53. Liu X, Krawczyk E, Suprynowicz FA, et al. Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc*. 2017;12(2):439-451. doi:10.1038/nprot.2016.174

54. Giraddi RR, Chung C-Y, Heinz RE, et al. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep*. 2018;24(6):1653-1666.e7. doi:10.1016/j.celrep.2018.07.025

55. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18(1):35-45. doi:10.1038/nri.2017.76

56. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257-272. doi:10.1038/s41576-019-0093-7

57. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health*. 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629

58. Holliday H, Baker LA, Junankar SR, Clark SJ, Swarbrick A. Epigenomics of mammary gland development. *Breast Cancer Res*. 2018;20(1):100. doi:10.1186/s13058-018-1031-x

59. Thong T, Forté CA, Hill EM, Colacino JA. Environmental exposures, stem cells, and cancer. *Pharmacol Ther*. Published online 2019. doi:10.1016/j.pharmthera.2019.107398

60. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest*. 2009;119(6):1420-1428. doi:10.1172/JCI39104

61. Merrell AJ, Stanger BZ. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style. *Nat Rev Mol Cell Biol*. 2016;17(7):413-425. doi:10.1038/nrm.2016.24

62. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med*. 2013;19(11):1438-1449. doi:10.1038/nm.3336

63. Jolly MK, Boareto M, Huang B, et al. Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol*. 2015;5:155. doi:10.3389/fonc.2015.00155

64. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

65. Jolly MK, Tripathi SC, Jia D, et al. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*. 2016;7(19):27067-27084. doi:10.18632/oncotarget.8166

66. Dontu G, Abdallah WM, Foley JM, et al. In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev*. 2003;17(10):1253-1270. doi:10.1101/gad.1061803

67. Colacino JA, McDermott SP, Sartor MA, Wicha MS, Rozek LS. Transcriptomic profiling of curcumin-treated human breast stem cells identifies a role for stearoyl-coa desaturase in breast cancer prevention. *Breast Cancer Res Treat*. 2016;158(1):29-41. doi:10.1007/s10549-016-3854-4

68. Liu X, Ory V, Chapman S, et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am J Pathol*. 2012;180(2):599-607. doi:10.1016/j.ajpath.2011.10.036

69. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/J.CELL.2015.05.002

70. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33(5):495-502. doi:10.1038/nbt.3192

71. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Birol I, ed. *Bioinformatics*. 2019;35(3):487-496. doi:10.1093/bioinformatics/bty644

72. Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, Chang HY. Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells. *Cell Stem Cell*. 2008;2(4):333-344. doi:10.1016/j.stem.2008.02.009

73. Kuleshov M V., Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

74. Lim E, Vaillant F, Wu D, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*. 2009;15(8):907-913. doi:10.1038/nm.2000

75. Bach K, Pensa S, Grzelak M, et al. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun*. 2017;8(1):2128. doi:10.1038/s41467-017-02001-5

76. Nguyen QH, Pervolarakis N, Blake K, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-04334-1

77. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71-e71. doi:10.1093/nar/gkv1507

78. Wang Y, Thong T, Saligrama V, Colacino J, Balzano L, Scott C. A gene filter for comparative analysis of single-cell RNA-sequencing trajectory datasets. *bioRxiv*. Published online May 17, 2019:637488. doi:10.1101/637488

79. Abd El-Rehim DM, Pinder SE, Paish CE, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol*. 2004;203(2):661-671. doi:10.1002/path.1559

80. O'Hare MJ, Ormerod MG, Monaghan P, Lane EB, Gusterson BA. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. *Differentiation*. 1991;46(3):209-221. doi:10.1111/j.1432-0436.1991.tb00883.x

81. Marcato P, Dean CA, Pan D, et al. Aldehyde Dehydrogenase Activity of Breast Cancer Stem Cells Is Primarily Due To Isoform ALDH1A3 and Its Expression Is Predictive of Metastasis. *Stem Cells*. 2011;29(1):32-45. doi:10.1002/stem.563

82. Trzpis M, McLaughlin PMJ, De Leij LMFH, Harmsen MC. Epithelial cell adhesion molecule: More than a carcinoma marker and adhesion molecule. *Am J Pathol*. 2007;171(2):386-395. doi:10.2353/ajpath.2007.070152

83. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A*. 2003;100(7):3983-3988. doi:10.1073/pnas.0530291100

84. Gerdur Ísberg Ó, Kim J, Fridriksdottir AJ, et al. A CD146 FACS Protocol Enriches for Luminal Keratin 14/19 Double Positive Human Breast Progenitors. *Sci Rep*. 2019;9(1):14843. doi:10.1038/s41598-019-50903-9

85. Jolly MK, Somarelli JA, Sheth M, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Ther*. 2019;194:161-184. doi:10.1016/j.pharmthera.2018.09.007

86. Casey T, Bond J, Tighe S, et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res Treat*. 2009;114(1):47-62. doi:10.1007/s10549-008-9982-8

87. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173(2):338-354.e15. doi:10.1016/j.cell.2018.03.034

88. Feng Y, Sun B, Li X, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. *Breast Cancer Res Treat*. 2007;103(3):319-329. doi:10.1007/s10549-006-9385-7

89. Folgueira MAAK, Maistro S, Katayama MLH, et al. Markers of breast cancer stromal fibroblasts in the primary tumour site associated with lymph node metastasis: A systematic review including our case series. *Biosci Rep*. 2013;33(6). doi:10.1042/BSR20130060

90. Li J-M, Tseng C-W, Lin C-C, et al. Upregulation of LGALS1 is associated with oral cancer metastasis. *Ther Adv Med Oncol*. 2018;10:175883591879462. doi:10.1177/1758835918794622

91. Falato C, Lorent J, Tani E, et al. Ki67 measured in metastatic tissue and prognosis in patients with advanced breast cancer. *Breast Cancer Res Treat*. 2014;147(2):407-414. doi:10.1007/s10549-014-3096-2

92. Montes de Oca R, Gurard-Levin ZA, Berger F, et al. The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol Oncol*. 2015;9(3):657-674. doi:10.1016/j.molonc.2014.11.002

93. Ding K, Li W, Zou Z, Zou X, Wang C. CCNB1 is a prognostic biomarker for ER+ breast cancer. *Med Hypotheses*. 2014;83(3):359-364. doi:10.1016/j.mehy.2014.06.013

94. Li Y, Chen Y-L, Xie Y-T, et al. Association Study of Germline Variants in CCNB1 and CDK1 with Breast Cancer Susceptibility, Progression, and Survival among Chinese Han Women. Miao X, ed. *PLoS One*. 2013;8(12):e84489. doi:10.1371/journal.pone.0084489

95. Wang Y, Zhang C, Mai L, Niu Y, Wang Y, Bu Y. PRR11 and SKA2 gene pair is overexpressed and regulated by p53 in breast cancer. *BMB Rep*. 2019;52(2):157-162. doi:10.5483/BMBRep.2019.52.2.207

96. Ren Z, Yang T, Zhang P, Liu K, Liu W, Wang P. SKA2 mediates invasion and metastasis in human breast cancer via EMT. *Mol Med Rep*. 2019;19(1):515-523. doi:10.3892/mmr.2018.9623

97. Han JY, Han YK, Park GY, et al. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines. *Sci Rep*. 2015;5. doi:10.1038/srep15993

98. Hamy AS, Bieche I, Lehmann-Che J, et al. BIRC5 (survivin): a pejorative prognostic marker in stage II/III breast cancer with no response to neoadjuvant chemotherapy. *Breast Cancer Res Treat*. 2016;159(3):499-511. doi:10.1007/s10549-016-3961-2

99. Ghaffari K, Hashemi M, Ebrahimi E, Shirkoohi R. BIRC5 genomic copy number variation in early-onset breast cancer. *Iran Biomed J*. 2016;20(4):241-245. doi:10.7508/ibj.2016.04.009

100. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

101. Yu M, Bardia A, Wittner BS, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (80-)*. 2013;339(6119):580-584. doi:10.1126/science.1228522

102. Sun P, Yuan Y, Li A, Li B, Dai X. Cytokeratin expression during mouse embryonic and early postnatal mammary gland development. *Histochem Cell Biol*. 2010;133(2):213-221. doi:10.1007/s00418-009-0662-5

103. Ruiz R, Herrero C, Strasser-Weippl K, et al. Epidemiology and pathophysiology of pregnancy-associated breast cancer: A review. *Breast*. 2017;35:136-141. doi:10.1016/j.breast.2017.07.008

104. Johansson ALV, Andersson TML, Hsieh CC, et al. Tumor characteristics and prognosis in women with pregnancy-associated breast cancer. *Int J Cancer*. 2018;142(7):1343-1354. doi:10.1002/ijc.31174

105. Degnim AC, Visscher DW, Hoskin TL, et al. Histologic findings in normal breast tissues: comparison to reduction mammaplasty and benign breast disease tissues. *Breast Cancer Res Treat*. 2012;133(1):169-177. doi:10.1007/s10549-011-1746-1

106. Rheinwatd JG, Green H. Seria cultivation of strains of human epidemal keratinocytes: the formation keratinizin colonies from single cell is. *Cell*. 1975;6(3):331-343. doi:10.1016/S0092-8674(75)80001-8

107. Domchek SM, Yao S, Chen F, et al. Comparison of the Prevalence of Pathogenic Variants in Cancer Susceptibility Genes in Black Women and Non-Hispanic White Women With Breast Cancer in the United States. *JAMA Oncol*. 2021;7(7):1. doi:10.1001/JAMAONCOL.2021.1492

108. Shimelis H, LaDuca H, Hu C, et al. Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst*. 2018;110(8):855-862. doi:10.1093/jnci/djy106

109. Roelands J, Mall R, Almeer H, et al. Ancestry-associated transcriptomic profiles of breast cancer in patients of African, Arab, and European ancestry. *NPJ Breast Cancer*. 2021;7(1). doi:10.1038/S41523-021-00215-X

110. Nievergelt CM, Maihofer AX, Shekhtman T, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*. 2013;4(1):13. doi:10.1186/2041-2223-4-13

111. Thong T, Wang Y, Brooks MD, et al. Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Front Cell Dev Biol*. 2020;8:288. doi:10.3389/FCELL.2020.00288/BIBTEX

112. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol 2018 365*. 2018;36(5):411-420. doi:10.1038/nbt.4096

113. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048/ATTACHMENT/1990279C-028D-47D4-A5D3-C89B04E14795/MMC3.XLSX

114. Chapter 4 DE analyses between conditions | Multi-Sample Single-Cell Analyses with Bioconductor. Accessed May 25, 2022. http://bioconductor.org/books/3.14/OSCA.multisample/multi-sample-comparisons.html

115. Gulati GS, Sikandar SS, Wesche DJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*. 2020;367(6476):405-411. doi:10.1126/SCIENCE.AAX0249

116. Using escape to perform ssGSEA analyses on single-cell RNA-seq data. Accessed May 25, 2022. http://www.bioconductor.org/packages/release/bioc/vignettes/escape/inst/doc/vignette.html

117. GSEA | MSigDB | Browse Gene Sets. Accessed May 25, 2022. https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

118. Cell-Cycle Scoring and Regression • Seurat. Accessed May 19, 2022. https://satijalab.org/seurat/articles/cell_cycle_vignette.html

119. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/BIOINFORMATICS/BTS480

120. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363. doi:10.1093/BIOINFORMATICS/BTU049

121. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res*. 2016;44(3):e20. doi:10.1093/NAR/GKV907

122. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics and Chromatin*. 2015;8(1):1-16. doi:10.1186/1756-8935-8-6/FIGURES/6

123. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208. doi:10.1186/s13059-016-1066-1

124. Fortin J-P, Triche TJ, Hansen KD, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2017;33(4):558-560. doi:10.1093/bioinformatics/btw691

125. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287. doi:10.1038/NG.3656

126. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443. doi:10.1038/NG.3679

127. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(10):1654. doi:10.1001/JAMAONCOL.2017.0595

128. Wu H, Wang C, Wu Z. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*. 2015;31(2):233-241. doi:10.1093/BIOINFORMATICS/BTU640

129. Yap SQ, Mathavarajah S, Huber RJ. The converging roles of Batten disease proteins in neurodegeneration and cancer. *iScience*. 2021;24(4):102337. doi:10.1016/J.ISCI.2021.102337

130. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet (London, England)*. 2005;365(9460):671-679. doi:10.1016/S0140-6736(05)17947-1

131. Pani T, Rajput K, Kar A, et al. Alternative splicing of ceramide synthase 2 alters levels of specific ceramides and modulates cancer cell proliferation and migration in Luminal B breast cancer subtype. *Cell Death Dis 2021 122*. 2021;12(2):1-22. doi:10.1038/s41419-021-03436-x

132. D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, Troester MA. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res Treat*. 2015;152(2):437-448. doi:10.1007/S10549-015-3474-4/TABLES/6

133. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(12):1654-1662. doi:10.1001/JAMAONCOL.2017.0595

134. Barrow MA, Martin ME, Coffey A, et al. A functional role for the cancer disparity-linked genes, CRYβB2 and CRYβB2P1, in the promotion of breast cancer. *Breast Cancer Res*. 2019;21(1):105. doi:10.1186/s13058-019-1191-3

135. Field LA, Love B, Deyarmin B, Hooke JA, Shriver CD, Ellsworth RE. Identification of differentially expressed genes in breast tumors from African American compared with Caucasian women. *Cancer*. 2012;118(5):1334-1344. doi:10.1002/CNCR.26405

136. Li H, Cheng C, You W, et al. TMEM100 Modulates TGF-β Signaling Pathway to Inhibit

Colorectal Cancer Progression. *Gastroenterol Res Pract*. 2021;2021. doi:10.1155/2021/5552324

137. Ma J, Yan T, Bai Y, et al. TMEM100 negatively regulated by microRNA-106b facilitates cellular apoptosis by suppressing survivin expression in NSCLC. *Oncol Rep*. 2021;46(2):1-14. doi:10.3892/OR.2021.8136

138. Zhuang J, Huang Y, Zheng W, et al. TMEm100 expression suppresses metastasis and enhances sensitivity to chemotherapy in gastric cancer. *Biol Chem*. 2020;401(2):285-296. doi:10.1515/HSZ-2019-0161/MACHINEREADABLECITATION/RIS

139. Donald CD, Sun CQ, Lim SD, et al. Cancer-Specific Loss of β-Defensin 1 in Renal and Prostatic Carcinomas. *Lab Investig 2003 834*. 2003;83(4):501-505. doi:10.1097/01.lab.0000063929.61760.f6

140. Lee J, Han JH, Jang A, Kim JW, Hong SA, Myung SC. DNA Methylation-Mediated Downregulation of DEFB1 in Prostate Cancer Cells. *PLoS One*. 2016;11(11). doi:10.1371/JOURNAL.PONE.0166664

141. Lo JH, Battaglin F, Baca Y, et al. DEFB1 gene expression and the molecular landscape of colorectal cancer (CRC). *https://doi.org/101200/JCO20224016_suppl3523*. 2022;40(16_suppl):3523-3523. doi:10.1200/JCO.2022.40.16_SUPPL.3523

142. Li YH, Hu YQ, Wang SC, Li Y, Chen DM. LncRNA SNHG5: A new budding star in human cancers. *Gene*. 2020;749:144724. doi:10.1016/J.GENE.2020.144724

143. Chi JR, Yu ZH, Liu BW, et al. SNHG5 Promotes Breast Cancer Proliferation by Sponging the miR-154-5p/PCNA Axis. *Mol Ther Nucleic Acids*. 2019;17:138. doi:10.1016/J.OMTN.2019.05.013

144. Damas ND, Marcatti M, Côme C, et al. SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nat Commun 2016 71*. 2016;7(1):1-14. doi:10.1038/ncomms13875

145. Shen M, Jiang YZ, Wei Y, et al. Tinagl1 Suppresses Triple-Negative Breast Cancer Progression and Metastasis by Simultaneously Inhibiting Integrin/FAK and EGFR Signaling. *Cancer Cell*. 2019;35(1):64-80.e7. doi:10.1016/J.CCELL.2018.11.016/ATTACHMENT/1E19EDC0-441D-4164-93EE-F3A5B9EC190A/MMC1.PDF

146. Musetti SN, Huang L. Tinagl1 Gene Therapy Suppresses Growth and Remodels the Microenvironment of Triple Negative Breast Cancer. *Mol Pharm*. 2021;18(5):2032-2038. doi:10.1021/ACS.MOLPHARMACEUT.1C00008

147. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob Heal*. 2020;8(8):e1027-e1037. doi:10.1016/S2214-109X(20)30215-1

148. Zota AR, Shamasunder B. The environmental injustice of beauty: framing chemical exposures from beauty products as a health disparities concern. *Am J Obstet Gynecol*. 2017;217(4):418.e1-418.e6. doi:10.1016/J.AJOG.2017.07.020

149. Ash M, Boyce JK. Racial disparities in pollution exposure and employment at US industrial facilities. *Proc Natl Acad Sci U S A*. 2018;115(42):10636-10641. doi:10.1073/PNAS.1721640115/SUPPL_FILE/PNAS.1721640115.SAPP.PDF

150. Polemi KM, Nguyen VK, Heidt J, Kahana A, Jolliet O, Colacino JA. Identifying the link between chemical exposures and breast cancer in African American women via integrated in vitro and exposure biomarker data. *Toxicology*. 2021;463. doi:10.1016/J.TOX.2021.152964

151. Nguyen VK, Kahana A, Heidt J, et al. A comprehensive analysis of racial disparities in chemical biomarker concentrations in United States women, 1999–2014. *Environ Int*. 2020;137:105496. doi:10.1016/J.ENVINT.2020.105496

152. Atlas E, Dimitrova V. Bisphenol S and Bisphenol A disrupt morphogenesis of MCF-12A human mammary epithelial cells. *Sci Rep*. 2019;9(1):1-10. doi:10.1038/s41598-019-52505-x

153. LaPlante CD, Catanese MC, Bansal R, Vandenberg LN. Bisphenol S Alters the Lactating Mammary Gland and Nursing Behaviors in Mice Exposed During Pregnancy and Lactation. *Endocrinology*. 2017;158(10):3448-3461. doi:10.1210/EN.2017-00437

154. Winkler J, Liu P, Phong K, et al. Bisphenol A replacement chemicals, BPF and BPS, induce protumorigenic changes in human mammary gland organoid morphology and proteome. *Proc Natl Acad Sci U S A*. 2022;119(11). doi:10.1073/PNAS.2115308119/-/DCSUPPLEMENTAL

155. Erler C, Novak J. Bisphenol A Exposure: Human Risk and Health Policy. *J Pediatr Nurs*. 2010;25(5):400-407. doi:10.1016/J.PEDN.2009.05.006

156. Konieczna A, Rutkowska A, Rachoń D. Health risk of exposure to Bisphenol A (BPA). *Rocz Panstw Zakl Hig*. 2015;66(1):5-11. Accessed August 5, 2022. https://pubmed.ncbi.nlm.nih.gov/25813067/

157. Zimmers SM, Browne EP, O'Keefe PW, et al. Determination of free Bisphenol A (BPA) concentrations in breast milk of U.S. women using a sensitive LC/MS/MS method. *Chemosphere*. 2014;104:237-243. doi:10.1016/j.chemosphere.2013.12.085

158. Lehmler H-J, Liu B, Gadogbe M, Bao W. Exposure to Bisphenol A, Bisphenol F, and Bisphenol S in U.S. Adults and Children: The National Health and Nutrition Examination Survey 2013-2014. *ACS omega*. 2018;3(6):6523-6532. doi:10.1021/acsomega.8b00824

159. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol 2017 361*. 2017;36(1):89-94. doi:10.1038/nbt.4042

160. Cheng M, Michalski S, Kommagani R. Role for Growth Regulation by Estrogen in Breast Cancer 1 (GREB1) in Hormone-Dependent Cancers. *Int J Mol Sci*. 2018;19(9). doi:10.3390/IJMS19092543

161. Hodgkinson K, Forrest LA, Vuong N, Garson K, Djordjevic B, Vanderhyden BC. GREB1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene 2018 3744*. 2018;37(44):5873-5886. doi:10.1038/s41388-018-0377-y

162. Haines CN, Braunreiter KM, Mo XM, Burd CJ. GREB1 isoforms regulate proliferation independent of ERα co-regulator activities in breast cancer. *Endocr Relat Cancer*. 2018;25(7):735-746. doi:10.1530/ERC-17-0496

163. Jiang S, Katayama H, Wang J, et al. Estrogen-Induced Aurora Kinase-A (AURKA) Gene Expression is Activated by GATA-3 in Estrogen Receptor-Positive Breast Cancer Cells. *Horm Cancer*. 2010;1(1):11. doi:10.1007/S12672-010-0006-X

164. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Mol Cancer 2021 201*. 2021;20(1):1-27. doi:10.1186/S12943-020-01305-3

165. Ma XF, Zhang J, Shuai HL, Guan BZ, Luo X, Yan RL. IKKβ/NF-κB mediated the low doses of bisphenol A induced migration of cervical cancer cells. *Arch Biochem Biophys*. 2015;573:52-58. doi:10.1016/J.ABB.2015.03.010

166. Sauer SJ, Tarpley M, Shah I, et al. Bisphenol A activates EGFR and ERK promoting proliferation, tumor spheroidformation and resistance to EGFR pathway inhibition in estrogen receptor-negativeinflammatory breast cancer cells. *Carcinogenesis*. 2017;38(3):252-260. doi:10.1093/CARCIN/BGX003

167. Li S, Wang B, Tang Q, Liu J, Yang X. Bisphenol A triggers proliferation and migration of laryngeal squamous cell carcinoma via GPER mediated upregulation of IL-6. *Cell Biochem Funct*. 2017;35(4):209-216. doi:10.1002/CBF.3265

168. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia (United States)*. 2017;19(8):649-658. doi:10.1016/j.neo.2017.05.002

169. Yang N, Wang C, Wang Z, et al. FOXM1 recruits nuclear Aurora kinase A to participate in a positive feedback loop essential for the self-renewal of breast cancer stem cells. *Oncogene*. 2017;36(24):3428-3440. doi:10.1038/ONC.2016.490

170. Eterno V, Zambelli A, Villani L, et al. AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. *Sci Reports 2016 61*. 2016;6(1):1-13. doi:10.1038/srep28436

171. Tan Y, Wang Q, Xie Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. *Int J Oncol*. 2019;54(1):87-97. doi:10.3892/ijo.2018.4598

172. Jian W, Deng XC, Munankarmy A, et al. KIF23 promotes triple negative breast cancer through activating epithelial-mesenchymal transition. *Gland Surg*. 2021;10(6):1941. doi:10.21037/GS-21-19

173. Li Z, Yang HY, Zhang XL, et al. Kinesin family member 23, regulated by FOXM1, promotes triple negative breast cancer progression via activating Wnt/β-catenin pathway. *J Exp Clin Cancer Res*. 2022;41(1):1-15. doi:10.1186/S13046-022-02373-7/FIGURES/9

Chapter 2

Characterizing the Transcriptomic Effects of Conditional Reprogramming Culture on Normal Mammary Cells

Introduction

As the field of cancer biology has evolved, a growing body of work has reinforced the critical role of stem-like cells in cancer. Due to long-observed similarities between embryonic development and oncogenesis, cancer is often considered a disease of "dysregulated development".[19,20] A characteristic shared by stem cells and cancer cells is cellular plasticity – the ability to transition and adopt alternative cell fates in response to environmental signals and stressors.[59] Plasticity is crucial for stem cells during embryonic development, for example during gastrulation when epiblast cells undergo the epithelial to mesenchymal transition (EMT) to form mesoderm which gives rise to the mesenchyme.[60] In adult stem cells, plasticity plays an important role in homeostasis and wound repair. This is demonstrated by adult tissue stem cells in the liver and intestinal epithelium which have been shown to de-differentiate or even trans-differentiate into cell types of a different lineage in order to replace damaged cells.[61] For cancer cells in tumors of epithelial origin, EMT plasticity and its reverse MET, are crucial for primary tumors to be able to adopt mesenchymal characteristics in order to disseminate, metastasize, and re-epithelialize at the metastatic site.[62]

Emerging evidence now suggests that these transitions occur along a continuum rather than as discrete switches in cell state. Transitioning hybrid cells exhibiting phenotypic markers of multiple cell states (epithelial/mesenchymal and luminal/basal) have been identified by us and others in both normal and carcinogenic breast tissue.[41,63,64] These cellular states are defined by the co-expression of known marker genes, for example the epithelial marker *EPCAM* and the mesenchymal marker *VIM* or the luminal marker *KRT18* and the basal/myoepithelial marker *KRT14*. Additionally, recent evidence shows that these hybrid cells exist in metastable states, not just as transient hybrids.[65] These hybrid populations are of particular interest due to their implicated role in promotion of tumorigenesis, metastasis, and aggressiveness of breast cancer.[63]

There are a limited number of studies which have observed these hybrid populations in the normal breast, and of these, the low proportions of hybrid cells identified have made them challenging to characterize.

In this study, we integrate multiple single-cell RNA sequencing datasets from the human and mouse in order to characterize the cell state distributions of the normal mammary (NM) gland throughout the life course, as well as after being perturbed into an enriched stem cell state following *in vitro* culture using the conditional reprogramming (CR) method.[53] Through an integrated analysis of single cell RNA-seq data with bulk human breast cancer transcriptomics from the Cancer Genome Atlas (TCGA), we investigate mammary stem cell populations and hybrid cell states, elucidating roles for these cells in mammary gland development and cancer.

Materials and Methods

Human tissue procurement

Tissue procurement was approved by the University of Michigan Institutional Review Board (HUM00042409). Normal mammary (NM) tissue was obtained from voluntary reduction mammoplasties performed at the University of Michigan hospital. Samples were processed following the protocol of Dontu et al. by enzymatic and mechanical digestion into single cell suspensions, as previously described. [66,67]

Conditional reprogramming

NM cells isolated from mammoplasty dissociation were co-cultured with irradiated 3T3 J2 mouse fibroblasts (Kerafast) using F-media in adherent conditions according to the protocol of Liu et al.[53,68] In order to establish an effective feeder layer, irradiated J2 fibroblasts were plated at a density of 12,000 cells per cm^2.[53] Once plated, the co-cultured cells were incubated in a humidified incubator at 37˚C/5% CO_2. Conditionally reprogrammed (CR) cells were allowed to grow up to 80% confluence and 0.05% trypsin/EDTA (Gibco, cat. no. 25300054), was used to differentially trypsinize cells from the adherent culture dishes. Differential trypsinization detaches irradiated J2s first, leaving behind an enriched population of CR mammary cells to be used for experimentation or cryopreservation.

To culture and irradiate J2 fibroblasts, cells were plated with J2 media in T-150 flasks (250,000-500,000 cells) and allowed to grow up to 80% confluence. J2s were carefully cultured to not exceed 90% confluence. Once confluent, J2s were trypsinized using 0.05% trypsin/EDTA, resuspended in J2 media, and placed on ice to be transported for irradiation. J2s were irradiated at 30 greys for 6 minutes, viability was assessed via acridine orange/propoidium iodide staining, and cryopreserved at 300,000-500,000 cells per vial in recovery cell culture freezing medium (Gibco, cat. no. 12648010). Following irradiation, 100,000 irradiated J2s and non-irradiated controls were plated for comparisons in order to ensure success of irradiation.

J2 media was prepared by combining 500 mL DMEM (Gibco, cat. no. 11965-092), 50 mL bovine calf serum (ATCC cat. no. 30-2030), 5.5 mL 200mM L-glutamine (Gibco, cat. no. 25030081), and 5.5 mL 100X Pen-Strep (Gibco, cat. no. 15140122).F-media was made by combining 623.83 µL of 12mM Rho-associated kinase (ROCK) inhibitor Y-27632 (Cayman Chemical, cat. no. 10005583), 194.48 µL of 96µg/mL hydrocortisone (STEMCELL Technologies, cat. no. 07925), 8.98 µL of 10 µg/mL epidermal growth factor (STEMCELL Technologies, cat. no. 78006.1), 935 µL of 4mg/mL insulin (Invitrogen-LifeTechnologies, cat. no. 12585014), and 62.83 µL of 1.2 µM cholera toxin (Sigma-Aldrich, cat. no. C8052) and 561 mL of complete DMEM (500 mL DMEM (Gibco, cat. no. 11965-092), 50 mL heat inactivated fetal bovine serum (Sigma-Aldrich cat. No F4135), 5.5 mL 200mM L-glutamine (Gibco, cat. no. 25030081), 5.5 mL 100X Pen-Strep (Gibco, cat. no. 15140122)).

Single cell RNA-sequencing

NM cells and their CR counterparts (n=3 pairs) were removed from liquid nitrogen storage and individually thawed, centrifuged, and counted. Cell mixtures were diluted with 0.01%FBS+PBS solution to achieve a final concentration of 100 cells/uL for each 5 mL sample (500,000 cells/sample). Samples were placed on ice and transferred for drop-seq analysis according to the protocol of Macosko et al.[69]

The drop-seq microfluidic device was assembled and calibrated to dispense oil droplets (Bio-Rad cat # 186-4006), cells, and Barcoded Bead SeqB (Chemgenes) beads at optimal velocity. NM and CR samples were loaded into the apparatus and the cell and microbead containing droplets were collected in 50 mL conical tubes (Falcon). Following droplet collection,

a series of wash, transfer, and centrifuge steps were performed in order to prepare the microbeads for sequencing.

After bead purification the following workflow was performed in order to generate DNA for sequencing. To generate cDNA strands from RNA hybridized to bead primers, RT mix was added to the microbeads and incubated. Following incubation, microbeads were rinsed and resuspended in exonuclease mix to remove excess bead primers that did not capture any RNA, rinsed, and then prepped for PCR. A 13 cycle PCR program was run to amplify cDNA and the generated cDNA library was then purified and analyzed on a BioAnalyzer High Sensitivity Chip. The purified cDNA was then tagmented using Nextera XT, PCR amplified, analyzed again using the BioAnalyzer. Following these steps, the library was ready for sequencing on the NextSeq 500.

Single-cell data analyses

Raw data processing: The "Drop-seqAlignmentCookbookv1.2Jan2016" software was used to transform raw sequencing data into gene expression measurements for each individual cell. The paired end reads were aligned to a mixed human (hg19) and mouse (mm10) reference genome and then grouped by cell according to the cell bar code. Next, a digital expression matrix was generated from the unique molecular identifier (UMI) counts for each gene in each cell. We performed quality control (QC) filtering on the raw data, filtering out cells with greater than 5% of mitochondrial genes and fewer than 200 total genes. Following QC filtration, we performed global log-normalization, scaling by percent mitochondrial genes, detection of highly variable genes, and principal component analysis dimension reduction. QC and downstream data processing were performed using the Seurat R package v3 unless specified otherwise.[70]

Unbiased clustering and cell type identification: Graph based unbiased clustering and PCA based tSNE dimension reduction were performed on NM, CR, and pooled NM and CR samples at a resolution of 0.5. Cell type identification was performed by identifying marker genes for individual clusters (Supplemental Table 1 and 2) as well as assessing expression of a pre-selected a panel of cell type marker genes. Cluster markers were identified using the FindMarkers function in Seurat and marker gene expression was assessed using the FeaturePlot and VlnPlot functions. Due to the differences in numbers of cells captured and analyzed between the NM and CR samples, we performed a normalized analysis by randomly down-sampling each individual

sample to 200 cells and performing the same clustering and marker gene assessment as performed on the full dataset.

Differential gene expression analysis: In order to isolate epithelial subsets of NM and CR cells for direct comparison, we filtered out stromal and immune cells from NM samples, and mouse cells were filtered out from CR samples. We performed differential expression analysis between pooled epithelial CR and NM samples with PQLseq, which uses a penalized quasilikelihood and a heredity correlation matrix.[71] The hereditary matrix was designed with the hierarchical data structure in mind in order to account for random effects of individual samples. Doing so prevents any one individual with a large number of cells relative to any other to dominate the analysis, and provides a more powerful analysis compared to a naïve approach. Differential gene expression between the NM and CR cells within each individual was performed using the FindMarkers function in Seurat. DEGs between NM and CR for each individual were plotted by average log2FC and correlation coefficients were calculated for each comparison.

Embryonic stem cell score: To estimate the similarities of the gene expression pattern of each cell to an embryonic stem cell, we calculated an "embryonic stem cell gene expression score." The proportion of total reads which belong to genes in the Embryonic Stem Cell Core set from the Gene Set Enrichment Analysis (GSEA) were calculated on a per cell basis.[72] A higher stem cell score can be interpreted as a greater proportion of reads for a given cell being derived from embryonic stem cell-associated genes.

Transcription factor and enrichment analyses: The top 1000 DEGs in CR compared to NM by log2FC were uploaded to the Enrichr web server to identify ENCODE and ChEA transcription factors enrichment.[73] To characterize the enrichment of the mammary stem cell and luminal progenitor gene sets reported by Lim et. al (2009)[74] and ROCK pathway gene signatures, each of these gene sets was overlapped with CR DEGs. The overlapped genes were then plotted by CR vs NM log2FC to visualize gene signature enrichment.

Identification of hybrid populations: Hybrid populations were identified in using expression of *KRT14*, *KRT18*, *VIM*, and *EPCAM*. Cells expressing marker genes at the 50th percentile or greater were deemed "high expressors." High expressors for the *KRT14/KRT18* or *EPCAM/VIM* marker combinations were identified as "double positive" hybrids, and high expressors for all four marker genes were identified as "quadruple positive" hybrids. Differential gene expression

analysis between quadruple hybrids and all other NM and CR epithelial cells was performed using FindMarkers in Seurat. Differentially expressed genes upregulated in quadruple hybrid cells were intersected with the MSigDB Hallmark Epithelial Mesenchymal Transition gene set (n=200) in order identify EMT related genes expressed in quadruple hybrids.

Integration and alignment of NM, CR, Bach, Nguyen, and the Cancer Genome Atlas (TCGA) breast cancer RNA-seq samples to the Giraddi mouse mammary transcriptome atlas

Dataset Descriptions: To contextualize our findings in NM and CR cells, we also performed an integrated analysis with three other single cell RNA-seq mammary gland datasets generated from mice and humans, as well as a comparative analysis using bulk breast cancer RNA-seq data from TCGA. The "Bach" dataset derives from single cell RNA-seq profiling of mouse mammary gland from four developmental stages: nulliparous, mid gestation, lactation, and post involution.[75] The "Nguyen" dataset is comprising of single cell RNA-seq profiling of human mammary gland generated from adult voluntary reduction mammoplasty patients.[76] The "Giraddi" dataset is comprised of single cell RNA-seq data from multiple timepoints during the lifecourse: embryonic day 16, embryonic day 18, postnatal day 0, postnatal day 4, and adult.[54] Bulk RNA-seq counts of TCGA breast tumors were obtained from the National Cancer Institute's genomic data commons portal using the TCGAbiolinks R package.[77]

Data pre-processing: The raw counts data of NM, CR, Bach, and Nguyen cells were normalized using either the "multiBatchNorm" or the "normalize" function in the R package scran. For gene filtering, a modified version of the CORGI algorithm was used on the Nguyen and Giraddi datasets, hereinafter referred to as "CORGI genes."[54,76] The CORGI gene filtering algorithm works by randomly sampling subsets of genes and scoring the subsets based on the structuredness of the data.[78] Genes that lead to more structured data are encouraged and vice versa. The TCGA dataset was pre-processed in the same way as the single-cell samples.

Down-sampling and cell selection: To generate the Giraddi reference dataset used for alignment, the full mouse mammary dataset was randomly down-sampled to 1000 cells spanning the 4 developmental stages (embryonic day 16, embryonic day 18, gestational day 4, and adult). Proportions of cells in the generated reference dataset reflect the proportions of cells from each developmental stage in the original dataset. The Bach mouse dataset was down-sampled by randomly selecting 250 cells from each of the 4 adult developmental stages (nulliparous, mid-

gestation, lactation, and post-involution) for a total of 1000 cells. NM, CR, and Nguyen datasets were also randomly down-sampled to 1000 cells each.

Batch correction: For batch correction, the "mnnCorrect" function in scran was used with default parameters on the logcounts on CORGI genes. The Giraddi dataset was input into the mnnCorrect as the first argument, i.e., as the reference atlas. Subsequently, the NM, CR, Bach, Nguyen, and TCGA samples were then projected onto the Giraddi developmental trajectory for comparative analysis.

Pseudotime analysis: In order to place the various datasets onto a developmental timeline, we leveraged the Giraddi mouse atlas as a reference. Pseudotime is computed directly onto the two-dimensional PCA plots by taking the dot product with an "arrow-of-time" vector that differentiates between the adult and embryonic cell populations in the Giraddi dataset. The same arrow-of-time vector was then applied to the NM, CR, Bach, Nguyen, and TCGA samples. A generalized linear model was used to determine significantly different pseudotime means between TCGA subtypes.

Dataset availability: The drop-seq data for the NM and CR samples are available on the Gene Expression Omnibus (GSE146792).

Results

Normal Mammary Cells Contain a Mixture of Stromal and Epithelial Cells and Cluster By Subtype

As a first step towards characterizing the distribution of phenotypic states of epithelial cells in the human mammary gland, we performed unbiased clustering of NM scRNA-seq data to determine the cell types and proportions present in the samples. Samples were analyzed from three individuals, here termed "NM11", "NM15", and "NM23". tSNE visualization revealed that the majority of clusters contained cells from each of the individuals (**Figure 2.A1**). To determine the identity of the 6 clusters (**Figure 2.1B**), a panel of known cell type and stem cell marker genes (**Figure 2.1C**) along with the top marker genes for each cluster identified by Seurat were used to characterize the clusters. The two major epithelial subtypes of the breast were identified by *KRT18* (luminal) and *KRT14* (myoepithelial) expression (**Figure 2.1C**).[79,80] Clusters 0 and 2 (**Figure 2.1B**) represent two distinct luminal populations which both highly

express epithelial marker *EPCAM* but differentially express stem cell marker *ALDH1A3*, which is preferentially expressed in cluster 0.[81,82] Mammary stem cell markers *ITGA6* and *CD44* also exhibited varying expression by cluster, with *ITGA6* showing low expression in clusters 0-3 and *CD44* exhibiting moderate to high expression across all clusters.[83] The myoepithelial Cluster 3 was almost entirely composed of cells from one individual (NM15), indicating variation in cell type proportions by individual. We identified cluster 4 identified as fibroblasts (*DCN*), cluster 1 as endothelial cells (*SERPINE1, AKAP12*), and cluster 5 as a small population of immune cells (*PTPRC*). Thus, prior to CR, normal mammary cells are composed of a mixture of stromal, immune, and epithelial cells and cluster primarily by cell type.

Conditionally reprogrammed mammary cells cluster by CR status and by individual

Marker analysis of the CR samples revealed that samples were depleted of fibroblasts, endothelial cells, and immune cells, but retained luminal and myoepithelial populations (**Figure 2.A1.A and B**). We identified two clusters (7 and 8) of mouse fibroblasts, using the mouse gene *Gapdh* as a marker, which we excluded in downstream analyses (**Figure 2.A1B**). To characterize CR alterations specifically in epithelial cells, we grouped NM and CR epithelial cells together for analysis. Unbiased clustering of the NM and CR cells revealed that NM samples remained relatively well mixed amongst each other, whereas CR samples distinctly clustered by individual (**Figure 2.2A**). While CR11 and CR15 exhibited some overlap in clustering, CR23 remained distinct from the other samples. Samples clustered by CR status along tSNE_1 and both NM and CR samples clustered as myoepithelial and luminal cells (**Figure 2.2B and C**). *KRT14* was selectively expressed in NM and CR myoepithelial populations, however, *KRT18* expressing CR cells also co-expressed moderate levels of *KRT14.* To determine if this clustering behavior was representative of CR gene expression alterations or due to the greater proportion of CR to NM cells, the same clustering and marker gene identification was performed on a randomly down-sampled subset comprised of 200 cells from each NM and CR sample. This subset of cells displayed the same clustering patterns and marker gene expression as the full dataset (**Figure 2.A1 C-F**). The co-expression (*KRT18*/*KRT14*) of luminal and myoepithelial markers was the first indication that the CR process could induce a hybrid state phenotype worthy of further investigation.

Conditionally Reprogrammed Mammary Cells Differentially Express Breast Cancer and Stem Cell Associated Genes

To gain mechanistic insight into the effects of the CR process, we compared gene expression patterns between NM and CR cells with differential gene expression (DGE) analysis. DGE between the NM and CR epithelial cells resulted in 3177 genes differentially expressed between the two cell populations (FDR<0.05) (**Figure 2.2D**). DGE was also conducted between the NM and CR cells of each individual and the overlap of differentially expressed genes (DEGs) was compared between individuals (Figure A2.2A). DEGs by individual were consistent with those found in combined NM and CR analysis, with both analyses identifying *LGALSI* as one of the most differentially upregulated genes in CR. Comparing DEGs between NM and CR by individual also revealed that the DEGs between samples 11 and 15 are highly correlated with each other (r=0.896) whereas DEGs between samples 11 and 23 (0.713) and between samples 15 and 23 (0.79) are less well correlated. Because the CR process requires the ROCK pathway small molecule inhibitor Y-27632, we assessed DEGs overlapping with ROCK associated pathway genes (**Figure 2.A2E**). Unsurprisingly, *ROCK2* was the most significantly downregulated gene in this pathway (log2FC=-1.25) in CR cells. We used the top 1000 DEGs in CR and inputted them to the Enrichr web server to identify transcription factors likely driving this process. Gene targets of known stem cell associated transcription factors *E2F4, FOXM1, BRCA1, SOX2, KLF4,* and *MYC* were all identified as enriched in CR upregulated genes (**Figure 2.A2F**).

To further investigate whether NM and CR cells exhibit differences in expression of stem cell associated genes, we performed analyses using overall gene expression as well as NM and CR DEGs. We estimated how "embryonic stem cell-like" each cell was by calculating the proportion of total transcripts annotated to embryonic stem cell (ESC) associated genes expressed in each NM and CR sample.[72] CR samples had higher ESC scores than their NM counterparts (**Figure 2.2E**), providing further evidence that CR cells express a more developmentally immature phenotype. To further characterize this phenotype in comparison to stem and progenitor cells in the normal breast, we overlapped NM and CR DEGs with mammary stem cell (MaSC) and luminal progenitor gene expression signatures reported by Lim et. al (2009). [74] Of the MaSC associated DEGs, 211/282 of the genes were upregulated in CR (**Figure 2.2F**), whereas only 68/144 luminal progenitor associated DEGs were upregulated in CR (**Figure**

2.A2D). Together, these analyses suggest that the CR process enriches for a stem cell-like state, and that the CR transcriptomic signature is reminiscent of ESCs and MaSCs.

Conditionally reprogrammed cells reflect a more developmentally immature phenotype

Due to the enrichment of stem cell associated genes in CR cells, we chose to further investigate this link in the context of mammary gland development. We integrated our data with the mouse mammary single-cell transcriptome atlas generated by Giraddi et. al (2018) which spans mouse mammary gland development from embryonic day 16 to adulthood (**Figure 2.3A**).[54] We calculated pseudotime estimates for each cell across the mouse developmental trajectory. Pseudotime estimates correlate to the developmental timepoint during which each cell was isolated, the more negative the pseudotime estimate the more embryonic-like the cell (**Figure 2.3B**). Using the CORGI alignment algorithm, we used the Giraddi data as a reference to map our NM and CR samples onto the mammary gland developmental trajectory. The majority of NM cells aligned to the adult mouse cells, whereas CR cells spanned the trajectory with a distinct population aligning to the embryonic mouse cells (**Figure 3.3C**). When CR cells were labeled by individual, CR15 and CR23 had cells spanning the whole trajectory, whereas CR11 mapped mostly to mouse mammary gland at post-natal day 4 and adulthood (**Figure 3.3D**).

Hybrid Stem Cell Populations Emerge Following Conditional Reprogramming

A growing number of studies have characterized hybrid stem cell populations in the normal and cancerous breast and have linked these epithelial/mesenchymal (E/M) or luminal/basal (L/B) hybrid phenotypes to aggressiveness of cancer.[41,84,85] Additionally, emerging evidence shows that stem cells can stably exist in hybrid states and that these hybrid phenotypes may be metastable.[65] To investigate the presence of hybrid populations in normal mammary cells, we assessed the co-expression of the luminal and basal (here used interchangeably with myoepithelial) markers *KRT18*/*KRT14* (L/B) and the epithelial and mesenchymal markers *EPCAM*/*VIM* (E/M) to identify "double positive" hybrid cells. Overlap of *EPCAM/VIM* and *CDH1/VIM* double positive populations indicate that *EPCAM* and *CDH1* are both effective epithelial markers, however, *EPCAM* was ultimately chosen as the epithelial marker for hybrid identification due to its overall higher expression in NM and CR cells (**Figure A2.1 G,H**). Co-expression of all four markers *KRT18*/*KRT14/ EPCAM*/*VIM* identified "quadruple positive" hybrid cells. "Triple positive" hybrid combinations were also assessed, however, we found these

redundant and less informative than the double positive and quadruple positive marker combinations (Figure A2.4). We identified L/B, E/M, and quadruple positive hybrid populations in the Giraddi dataset, NM, and CR cells, with CR cells expressing the highest proportions of both double positive hybrids and quadruple hybrids (Table A11). CR15 expressed the highest proportion of hybrid cells among the individuals.

In the Giraddi dataset, E/M hybrids spanned both basal and luminal branches of the trajectory, L/B hybrids mostly mapped to adult luminal and post-natal day 4, and quadruple hybrids mapped along the luminal branch around embryonic day 18 and post-natal day 4 (**Figure 2.3E**). E/M hybrids were the only cells to map to the basal adult cells and the embryonic cells. To investigate the developmental maturity of hybrid CR cells, we mapped the E/M, L/B, and quadruple hybrids to the mouse developmental trajectory (**Figure 2.3F**). Almost all of the hybrid CR cells mapped to mouse cells spanning embryonic day 16 through post-natal day 4, with a few mapping to the adult populations. Interestingly, the hybrid E/M cells map along both the luminal and basal trajectories, however, the L/B hybrids almost exclusively map along the luminal trajectory. The majority of the quadruple hybrids also mapped along the luminal trajectory.

To further characterize the different cell types, pseudotime analysis was performed on the CR, NM, and hybrid populations. Pseudotime estimates for CR cells were more negative than NM cells, indicating a more developmentally immature phenotype (**Figure 2.3G**). Pseudotime analysis of the Giraddi mouse hybrid populations revealed that hybrid E/M cells are the most developmentally immature, followed by the quadruple hybrids, and then hybrid L/B cells (**Figure 2.3H**). CR hybrids exhibited a similar pattern to the mouse hybrids, where hybrid E/M cells were the most developmentally immature, quadruple hybrids were intermediate, and L/B hybrids were the most mature (**Figure 2.3I**). Pseudotime differences between hybrid populations in the CR cells were less pronounced than in the mouse hybrid cells. We also calculated the embryonic stem cell score for the NM and CR hybrid cell populations and found that E/M and quadruple hybrids expressed a higher embryonic stem cell score, whereas L/B hybrids were less distinct (**Figure 2.A3 A-C**). From this we concluded that the CR process causes an enrichment of hybrid cells and that these hybrid populations are transcriptionally similar to mammary cells in early development. Finally, E/M and L/B hybrids appear to represent distinct cellular populations with quadruple positive hybrid cells falling somewhere in between.

Differential gene expression analysis between quadruple positive hybrids and all other epithelial NM and CR cells identified 4052 genes upregulated and 2660 genes downregulated in quadruple hybrids (**Figure 2.A2 B**). The most significant DEG upregulated in the quadruple positive hybrids was extracellular matrix gene *COL14A1* which has been found to be upregulated in cancerous breast stroma compared to normal breast stroma.[86] We further investigated the DEGs from the quadruple hybrids by calculating the overlap of these genes with the MSIGDB EMT hallmark gene set (Table S13). We found that 82 out of the 200 (41%) genes differentially expressed in the quadruple hybrids were EMT related genes (**Figure 2.A2 C**). Together, these data provide compelling evidence for the presence of plastic hybrid cells in the normal and developing breast, specifically early in development.

Hybrid Stem Cell Populations are Enriched During Gestation and Lactation

The enrichment of these hybrid populations early in breast development aligns with the current understanding of the highly dynamic nature of mammary gland morphogenesis. This led us to investigate another highly dynamic and proliferative developmental stage of the breast: gestation and lactation. We incorporated the adult mouse mammary developmental dataset generated by Bach et. al (2017), which spans the nulliparous, mid-gestation, lactation, and post-involution time points.[75] Alignment of the Bach dataset to the Giraddi developmental trajectory revealed a striking chronological pseudotime arc (**Figure 2.4A**). Beginning at the nulliparous stage, mammary cells exhibit a developmentally mature pseudotime, reflected by alignment to Giraddi mouse adult cells. Mammary cells during the gestation stage exhibit a more developmentally immature phenotype, indicated by a drop in pseudotime. Through the lactation and post-involution stages, pseudotime of mammary cells sequentially increases to re-stablize at a pseudotime similar to the developmental maturity of the nulliparous stage. Mapping of these cells to the Giraddi trajectory demonstrated that the nulliparous and post-involution stages mapped most closely to the luminal and basal adult cells, the lactation stage mapped most closely to adult basal cells, and the gestation stage mapped most closely to the embryonic day 18 cells (**Figure 2.4B**). L/B, E/M, and quadruple hybrids were also identified in the Bach dataset and mapped to the Giraddi trajectory (**Figure 2.4C-E**). Proportions of hybrid cells were calculated for each stage (**Figure 2.4F**). The highest proportion of E/M hybrids were found in the gestation stage which also expressed the highest proportion of L/B hybrids, followed closely by the

lactation stage. Interestingly, the lactation stage expressed the highest proportion of quadruple hybrids, followed by the gestation stage. Although the pseudotime estimates for the nulliparous and post-involution stages were similar, the post-involution stage had an approximately 5-fold lower proportion of hybrid cells. The enrichment of hybrid populations during the gestation and lactation stages suggests the importance of these cells during pregnancy-associated mammary gland morphogenesis.

To further extend and validate these findings in human patient samples, we also explored the distribution of hybrid cells in the Nguyen dataset, which is generated largely from nulliparous patients. We aligned the Nguyen data to the Giraddi developmental trajectory and found that cells largely clustered with the mouse adult luminal and basal cells (**Figure 2.A5B**). In the Nguyen dataset, there were approximately 11 and 12% of cells classified as E/M and L/B hybrids (S5J), respectively, which is comparable to the proportion of these hybrid cells in the nulliparous mice from the Bach dataset (10 and 16%). The proportion of these cells in the post-involution mouse cells from the Bach dataset were 3% and 2%, respectively.

Basal Breast Cancers are the most Transcriptionally Distinct and Developmentally Immature of Breast Cancer Subtypes

All our prior findings about hybrid cell states and developmental phenotypes were characterized in normal human and mouse mammary cells. Our next step was to leverage this data to inform our understanding of breast cancer subtype biology. To do this, we assessed gene expression of breast tumors from the Cancer Genome Atlas (TCGA). Principal component analysis of the TCGA tumors without any alignment showed that basal tumors clustered as the most distinct from the other subtypes, with luminal A and luminal B overlapping, and the other subtypes grouping between the luminal and basal subtypes (**Figure 2.5A**). We mapped the bulk TCGA tumor RNA expression data onto the Giraddi mouse developmental trajectory and found that normal, luminal A, and luminal B tumors mapped most closely to the adult cells, HER2 tumors mapped to slightly more immature cells, and basal tumors spanned pseudotime along the basal trajectory (**Figure 2.5B**). Pseudotime estimates by subtype revealed that the luminal A subtype exhibits a significantly more developmentally mature phenotype than the luminal B ($p=4.97E-07$), Her2 ($p=0.0495$), and basal ($p<2E-16$) subtypes, with the basal subtype exhibiting the most immature pseudotime estimate (**Figure 2.5C**). As a next step, we assessed the link

between the pseudotime estimates of gene expression and breast cancer outcomes. Of the top 10 annotated genes with the most negative pseudotime estimates, 5 were significantly associated with poor prognosis in breast cancer patients (**Figure 2.5D**). Our results suggest that "phenotypic developmental maturity" of cancer cells, particularly at timepoints strongly associated with the hybrid E/M state may be a distinguishing factor of the subtypes and that pseudotime-associated genes have prognostic implications for breast cancer patients.

Discussion

Through our integrated analysis of normal human and mouse mammary data and TCGA tumor data, we witness an overarching theme – "developmentally immature" pseudotime is linked to the likelihood of hybrid cells which express a stem-like gene expression signature. Here, we demonstrate that hybrid cells are more developmentally immature and embryonic stem cell-like. Further, we identify an increased proportion of hybrid cells at particular important timepoints during development: in particular the *in utero* period, gestation, and lactation. Others have found associations between an "embryonic stem-cell like" gene expression signature and aggressiveness of cancers.[87] Hybrid E/M cells present a particularly interesting population to further explore in the context of aggressive cancers due to their expression during the lowest pseudotime estimates and their mapping along the basal mouse trajectory. Together, this suggests that hybrid cells/states and their stem-like plasticity are important mediators in development and cancer and that this intersection is a promising future direction to explore.

The precision of single-cell RNA-seq allowed us to characterize NM tissue as comprised of stromal, immune, and epithelial cells. When we perturbed NM cells *in vitro* with the conditional reprogramming method, we identified that CR cells only contained luminal and myoepithelial populations, with a small subpopulation of mouse fibroblasts, which were used as a feeder layer to support the growth of the CR cells. The CR process appears to enhance inter-individual heterogeneity, where post-CR samples cluster much more distinctly by individual. Given that the CR samples and NM counterparts were derived from the same individual, the preferential clustering by CR status is indicative that the CR process likely induces major transcriptomic alterations as well as depletion of immune and stromal cells.

DGE between NM and CR samples allowed us to identify a number of significant genes. Understanding their molecular functions may provide crucial mechanistic insight into the CR process, the enrichment of the embryonic stem cell phenotype we observed, and the connection between stemness and cancer. Of these genes, *LGALS1* stands out due to its significant upregulation overall in CR cells as well as in comparisons of DEGs by sample. A member of the galectin family of proteins which modulate proliferation and cell-cell/cell-matrix interactions, upregulation of *LGALS1* expression in breast cancer adjacent fibroblasts has been linked to metastasis and has also been found to be altered in lymph node metastases compared to primary breast tumors.[88,89] Outside of the breast, *LGALS1* has also been linked to invasiveness and metastasis in oral cancer.[90] Amongst the other highly significant upregulated CR genes by p-value and log2FC, *SKA2*, *MKI67*, *HJURP*, *BIRC5*, and *CCNB1* have all been found to be upregulated in breast cancer tissues and all five except for *SKA2* have been identified as prognostic markers for breast cancer.[91–96] Additionally, *BUB1* and *BIRC5* have been linked to stemness, where depletion of *BUB1* reduced cancer stem cell potential in MDA-MB-231 and MCF-7 breast cancer cell lines and *BIRC5* is commonly expressed in embryonic tissues and cancer but not in adult tissues.[97–99] Experimental evidence continues to support the link between stemness and cancer, and our results showing enrichment of a stem-like phenotype and breast cancer related genes in CR cells adds to this body of work. It is striking that just the induction of stem-like proliferation and de-differentiation of normal mammary adult cells by ROCK inhibitor Y-27632 upregulates numerous genes which overlap with breast cancer and metastasis, providing further experimental evidence that cancers are potentially hijacking normal stem cell mechanisms.

Another key finding of our study was the emergence of hybrid cell populations post-CR. Our characterization of these populations is consistent with what has been previously reported and provides additional insight into the "developmental maturity" of these hybrid states. Hybrid E/M cells have been found in human primary tumors and lymph nodes where they have been shown to exhibit enhanced tumor initiation and metastatic potential and have been implicated in contributing to therapy resistance and poor survival. [85,100,101] Similarly, L/B hybrids have been characterized in both normal and cancer tissue from humans and are believed to be derived from luminal progenitors. [79,84] This hypothesis of L/B hybrids being luminal in origin is consistent with what we observed, where L/B hybrids in both human and mouse map only along the

luminal trajectory of the mouse mammary gland developmental atlas, whereas E/M hybrids map to both luminal and basal trajectories. Sun et al. have shown that in the developing mouse mammary gland, *KRT5*/*KRT14* (L/B) hybrids are observed beginning from embryonic day 15.5 up until adulthood (8-12 weeks).[102] While these populations dramatically decreased after 3 weeks, it is important to note that they were still present in the normal adult mouse mammary gland. Additionally, they also identified a distinct population of cells expressing *KRT6*, a multipotent mammary epithelial progenitor marker, which emerged at embryonic day 16.5 and was found localized to the nipple sheath. Expression of *KRT6* was also correlated to the boundary of the mammary mesenchyme, separate from luminal and basal localization. Considering the proximity to the mammary mesenchyme and its distinctness from luminal and basal progenitors, the *KRT6* population in the mouse mammary gland may be analogous to the E/M hybrids we identified in the CR population. The embryonic origin of hybrid populations in the developing mouse mammary gland and their persistence through adulthood suggests that hybrid populations in the human mammary gland also arise during embryogenesis and are maintained through adulthood.

Pseuodotime analysis of mouse, NM, CR, and breast tumor samples suggest that the "developmental maturity" state of a cell or tumor plays a direct role in its biological behavior. Of the hybrid populations in both mouse and CR cells, the E/M hybrids exhibited the lowest estimated pseudotime. Based on prior knowledge implicating E/M hybrids in tumorigenesis and metastasis, this population may be of particular interest in the future to target for cancer prevention and therapy. To understand the impact of variation in pseudotime on our understanding of breast subtype biology, we calculated pseudotime estimates of bulk tumor RNA-seq data from TCGA samples. On average, none of the TCGA tumor subtypes exhibited pseudotime scores corresponding to adult mouse cells. Instead, average subtype scores corresponded to post-natal day 4 and earlier in development. While these are bulk samples being aligned to single-cell mouse samples, this suggests that regardless of cell type, a more developmentally immature phenotype is characteristic of cancers. Among the subtypes, basal cancers preferentially map to the most developmentally immature cells in the mammary gland and express the lowest pseudotime scores. This difference in "developmental maturity" may be a key distinction between basal cancers and other subtypes and may play a major role in the aggressiveness and low survival outcomes observed clinically and epidemiologically.

One of our most exciting findings was the characterization of hybrid cells in the adult mouse mammary gland during pregnancy. The enrichment of hybrid populations during gestation and lactation and the loss of them in the subsequent post-involution stage suggests that these hybrid states are inducible and transient. This transiency provides compelling evidence that these hybrid populations are instrumental to the dynamic modifications in breast morphogenesis which occur during pregnancy and lactation. This arc of mouse hybrid enrichment and stabilization parallels the transient increase in breast cancer risk during and immediately following pregnancy, which decreases over time. The time period during which hybrid populations are most prevalent in the mouse breast overlaps with pregnancy associated breast cancer (PABC) risk in humans, diagnosed between pregnancy and one year following birth.[103] This overlap in time period, as well as the parallel transiency of mouse hybrid populations and PABC risk, supports the presence of these hybrid populations in the human breast during pregnancy and implicates their involvement in PABC. The pathophysiology of PABC is characterized by metastatic, high grade tumors, and survival is inversely correlated with time since birth.[103] Consistent with this is the finding that ER-/ PR-/HER2+, and triple-negative tumors were more common in women diagnosed with PABC compared to nulliparous women.[104] Based on our other findings that basal breast tumors exhibit the most "developmentally immature" pseudotime estimates and the link between hybrid cells and aggressive cancers, characterizing hybrid populations and "developmental maturity" of PABCs could inform prognostic and therapeutic treatment.

Our study had a number of limitations. One was the source and sample size of mammary tissue. Mammoplasty tissue has been critiqued as not being fully representative of the "normal" breast, and due to the de-identification of the samples we also lack demographic data on the women from whom they were obtained for our study, although we were able to supplement our findings with additional human data from the Nguyen study. [105] Moreover, the conditional reprogramming methodology only supports the outgrowth of epithelial cells from samples, a phenomenon which has been linked to the J2 fibroblast co-cultures since the 1970s.[106] A better understanding of stromal/epithelial interactions in regulating these hybrid stem cell states is an important future direction of research. These future experiments could, for example, assess the impact of adult fibroblasts or cancer associated fibroblasts on the reprogramming process. Future complementary analyses of conditional reprogramming using breast cancer samples could also

provide important insights into the impacts of enhanced stemness and developmental immaturity on tumor characteristics. While single-cell technology is rapidly evolving and improving, we acknowledge that in this study we are only capturing expression of a subset of the genes expressed in each individual cell. Another limitation is the potential for unanticipated bias from using the subset of CORGI selected genes for alignment with the human mammary cells and TCGA tumor samples to the mouse developmental trajectory.

Overall, we showcase a computational analysis which leverages publicly available data to gain insight into the relationship between hybrid cell populations, stemness, and cancer. We and others have identified significant inter-individual heterogeneity in proportions of stem cells in mammary tissue.[40,41] Our ongoing work is utilizing single-cell RNA-seq of normal mammary tissue from epidemiologically well characterized women to understand how known epidemiological risk factors for cancer influence the "stemness" of breast epithelial cells. Quantification of reprogramming efficiency during conditional reprogramming across samples from diverse women could provide a functional readout of "stemness" or reprogramming capacity and their relations to known cancer risk factors, such as age, ethnicity, or genetic predisposition to cancer. Future work can focus on identifying the localization of these hybrid states in the adult mammary gland using advanced techniques, such as spatial transcriptomics. Overall, these results provide further evidence to support investigating the role of stem cells, and particularly hybrid E/M cells, in normal development and characterizing how this biology is hijacked during tumorigenesis. Understanding the biology of these cells will likely provide novel targets for the prevention and therapy of breast cancers, including aggressive subtypes with fewer therapeutic options.

Tables and Figures

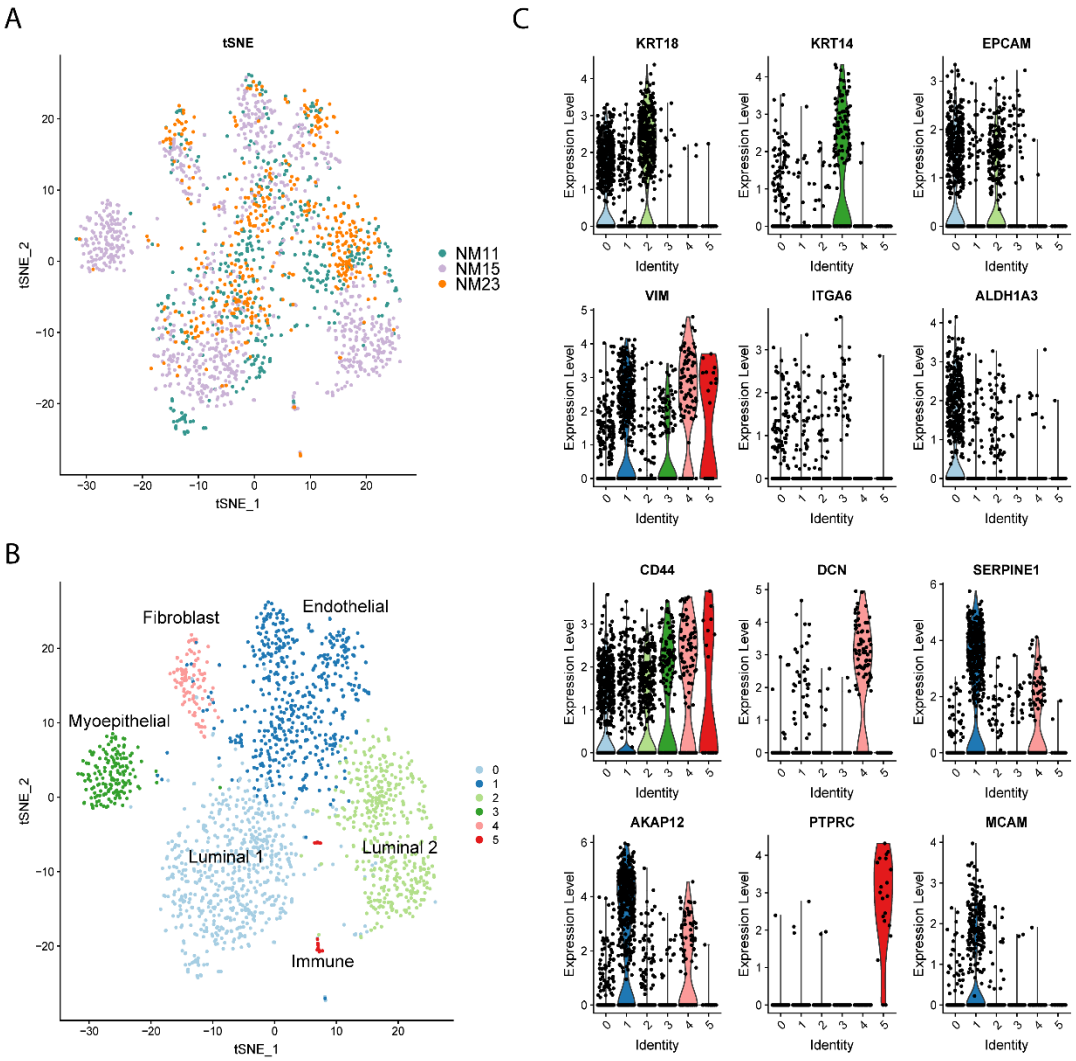


Figure 2.1: Unbiased clustering and cell type identification of NM cells. (A) tSNE dimension reduction of NM samples colored by individual (B) Unbiased clustering of NM samples colored by cell cluster (C) Expression of known cell type marker genes by cluster across all NM samples

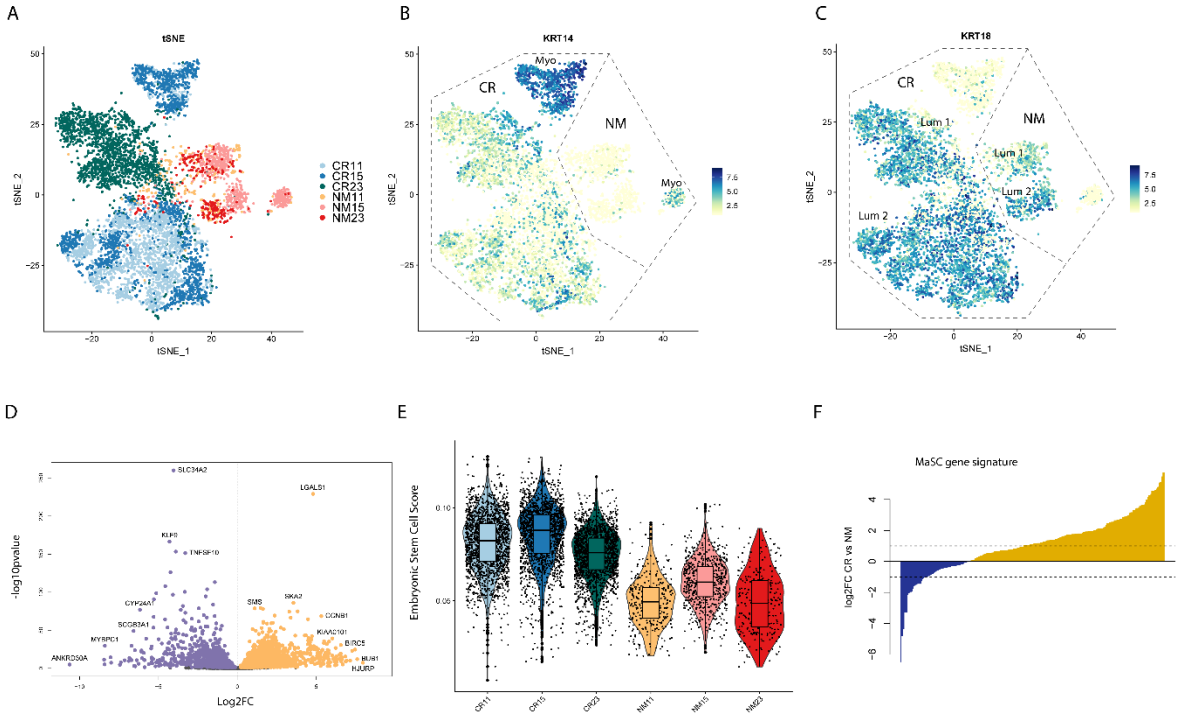


Figure 2.2: Unbiased clustering and differential gene expression between NM and CR. (A) tSNE dimension reduction of NM and CR samples by individual (B) FeaturePlots of myoepithelial marker gene (*KRT14*) and (C) luminal marker gene (*KRT18*) expression (D) Differentially expressed genes between NM and CR epithelial cells. Significantly upregulated genes in CR (FDR<0.05) are colored in orange. Significantly upregulated genes in NM are colored in purple (E) Distribution of cells from NM and CR samples scored by embryonic stem cell gene expression (F) Comparison of overlap between NM and CR differentially expressed genes and the mammary stem cell (MaSC) gene expression signature reported in Lim et al. (2009). Yellow genes indicate MaSC genes more highly expressed in CR vs NM.

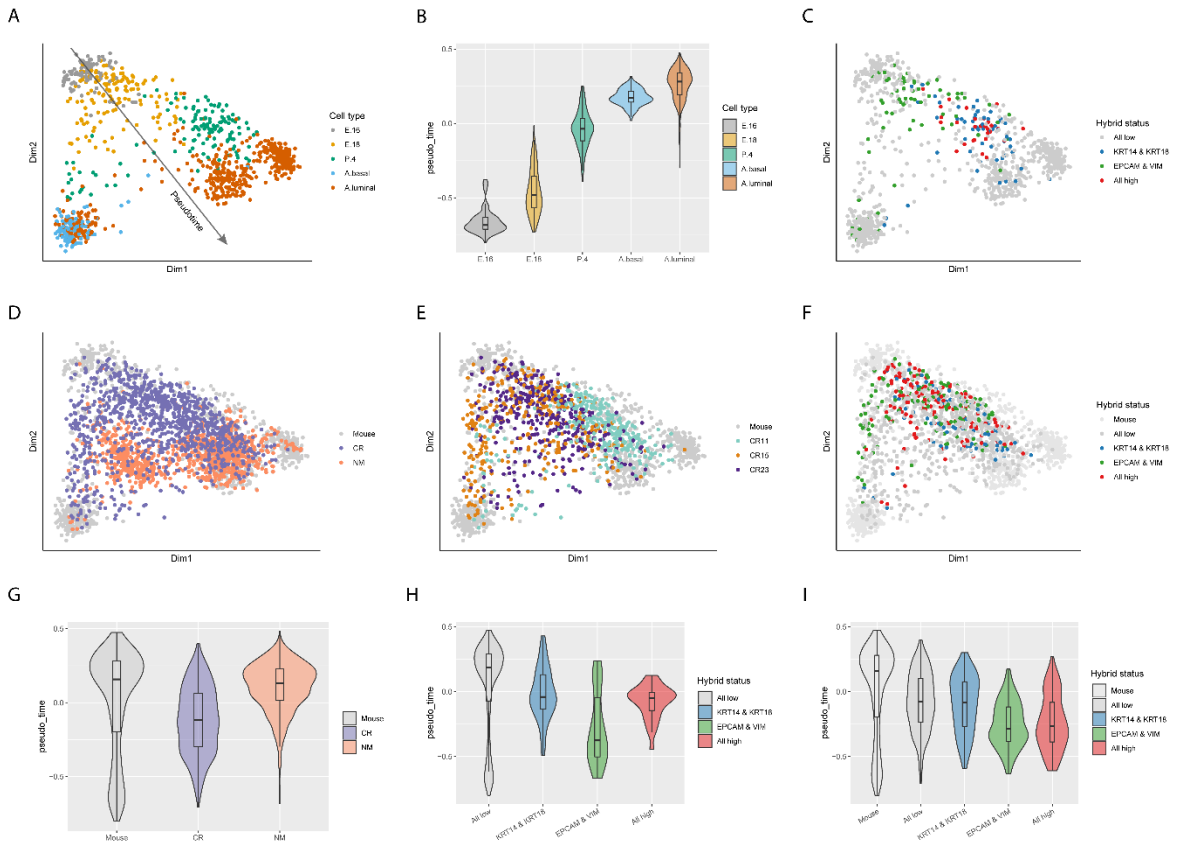
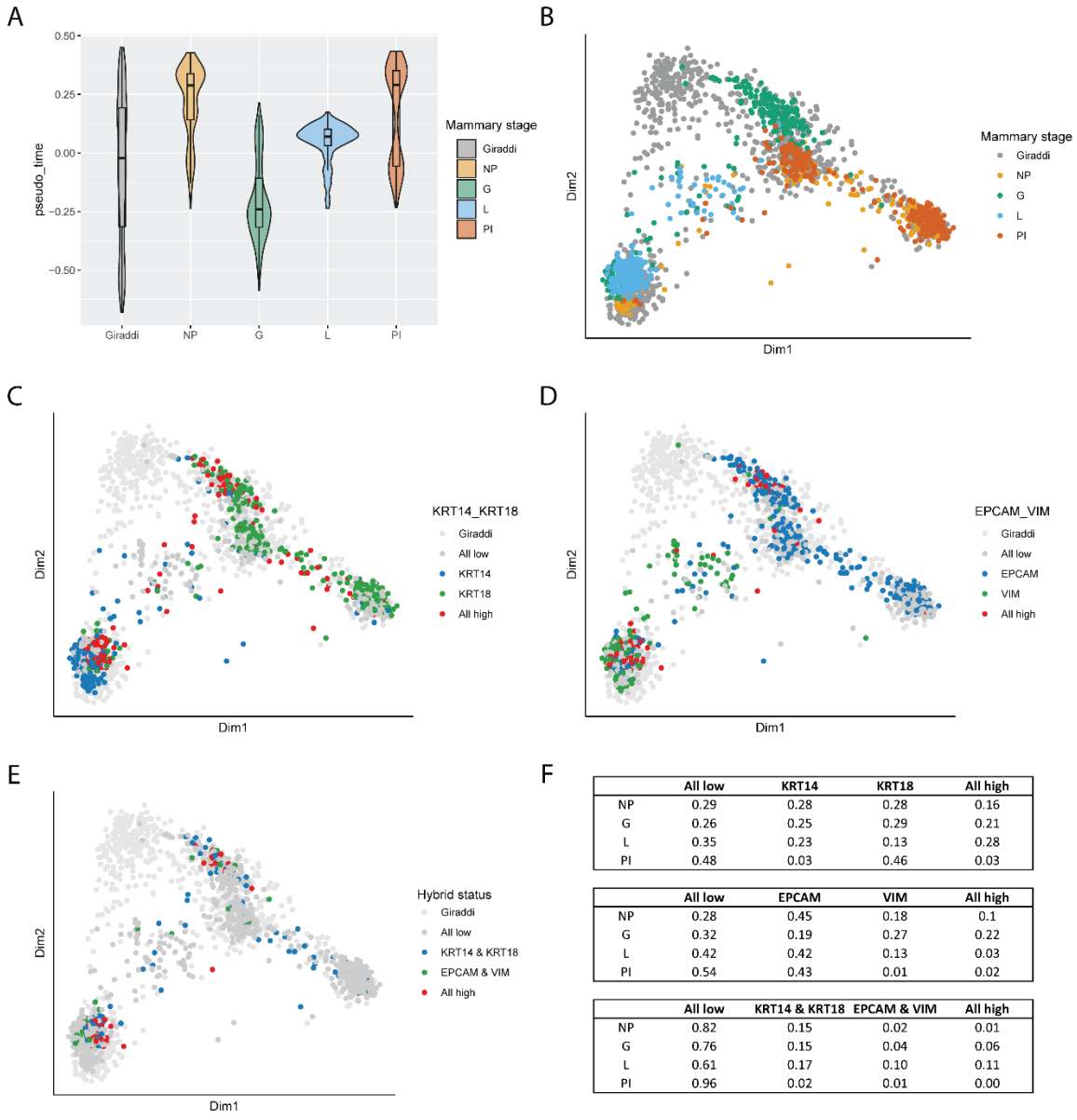


Figure 2.3: Alignment of NM and CR cells to mouse mammary developmental trajectory and characterization of hybrid cells. (A) Principal component analysis plot of single cell RNA-seq data of mouse mammary gland at embryonic day 16 (E.16), embryonic day 18 (E. 18), post-natal day 4 (P.4), and adult basal (A. basal) and adult luminal (A.luminal) cells as reported in Giraddi et al. (2018) (B) Pseudotime estimates of mouse mammary developmental stages (C) NM and CR cells mapped to the developmental trajectory with CoRGI (D) CR samples mapped to the mouse mammary developmental trajectory labeled by individual (E) Hybrid cell identification of mouse mammary cells along the developmental trajectory. Luminal/basal hybrids were identified by concurrent high *KRT/14/KRT18* expression. Epithelial/mesenchymal hybrids were identified by concurrent high *EPCAM/VIM* expression. Quadruple positive hybrid cells were identified by high expression of all four marker genes *KRT14/ KRT18,/EPCAM/VIM* (F) CR cells mapped to mouse developmental trajectory and labeled by hybrid status (G) Pseudotime estimates of NM and CR cells relative to the mouse mammary developmental trajectory cells(H) Pseudotime estimates of mouse hybrid cells (I) Pseudotime estimates of CR hybrid cells.



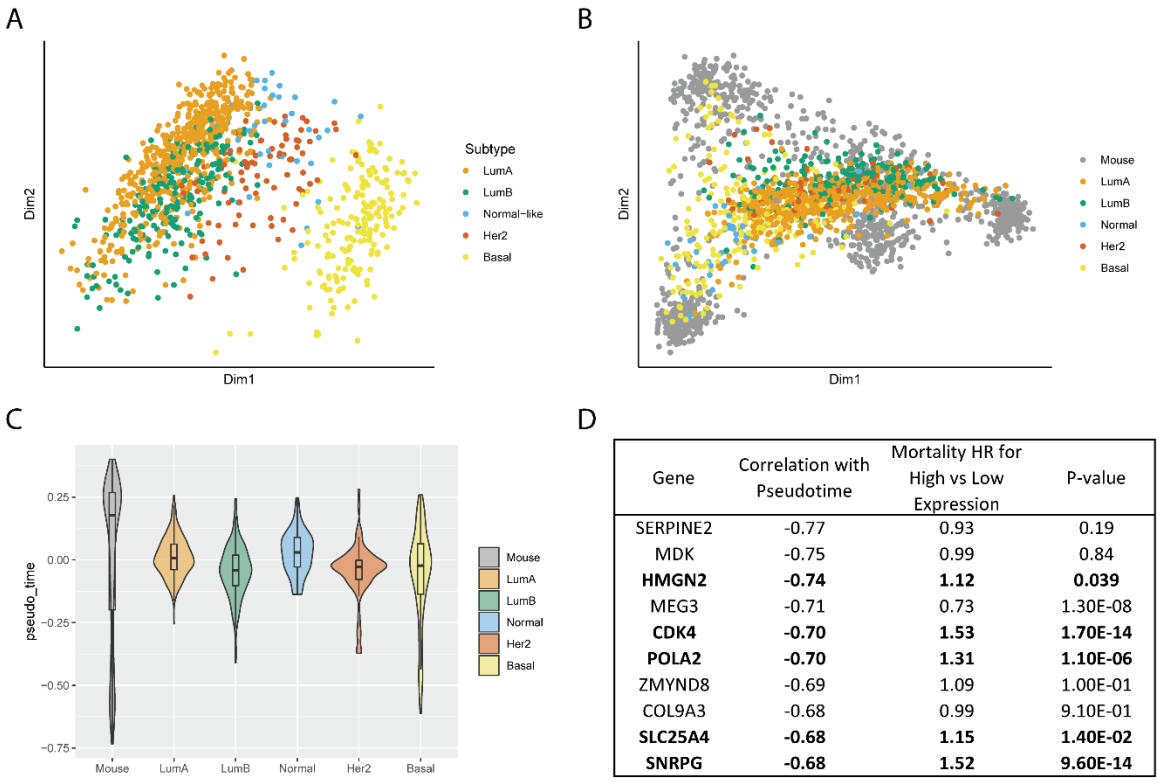
F

	All low	KRT14	KRT18	All high
NP	0.29	0.28	0.28	0.16
G	0.26	0.25	0.29	0.21
L	0.35	0.23	0.13	0.28
PI	0.48	0.03	0.46	0.03

	All low	EPCAM	VIM	All high
NP	0.28	0.45	0.18	0.1
G	0.32	0.19	0.27	0.22
L	0.42	0.42	0.13	0.03
PI	0.54	0.43	0.01	0.02

	All low	KRT14 & KRT18	EPCAM & VIM	All high
NP	0.82	0.15	0.02	0.01
G	0.76	0.15	0.04	0.06
L	0.61	0.17	0.10	0.11
PI	0.96	0.02	0.01	0.00

Figure 2.4: Alignment of Bach mouse mammary developmental dataset to the Giraddi mammary trajectory. (A) Pseudotime estimates of Bach mammary developmental stages: nulliparous (NP), mid-gestation (G), lactation (L), and post-involution (PI) (B) Bach mammary cells mapped to the Giraddi trajectory with CoRGI (C) Bach luminal/basal hybrid cells mapped to Giraddi trajectory (D) Bach epithelial/mesenchymal hybrid cells mapped to Giraddi trajectory (E) Bach quadruple positive hybrid cells mapped to Giraddi trajectory (F) Proportions of Bach hybrid cells by developmental stage.



D

Gene	Correlation with Pseudotime	Mortality HR for High vs Low Expression	P-value
SERPINE2	-0.77	0.93	0.19
MDK	-0.75	0.99	0.84
HMGN2	**-0.74**	**1.12**	**0.039**
MEG3	-0.71	0.73	1.30E-08
CDK4	**-0.70**	**1.53**	**1.70E-14**
POLA2	**-0.70**	**1.31**	**1.10E-06**
ZMYND8	-0.69	1.09	1.00E-01
COL9A3	-0.68	0.99	9.10E-01
SLC25A4	**-0.68**	**1.15**	**1.40E-02**
SNRPG	**-0.68**	**1.52**	**9.60E-14**

Figure 2.5: Alignment of TCGA tumors to Giraddi mammary trajectory (A) Principal component analysis of TCGA bulk breast tumor RNA-seq labeled by subtype (B) Alignment of TCGA tumors to Giraddi developmental trajectory (C) Pseudotime estimates of TCGA tumor subtypes (D) Mortality hazard ratio estimates relative to expression of the top 10 genes most negatively correlated with mouse pseudotime. The more negative the pseudotime estimate, the more highly expressed the gene is in the earliest developmental timepoint.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/CAAC.21660

2. Zeinomar N, Oskar S, Kehm RD, Sahebzeda S, Terry MB. Environmental exposures and breast cancer risk in the context of underlying susceptibility: A systematic review of the epidemiological literature. *Environ Res*. 2020;187:109346. doi:10.1016/j.envres.2020.109346

3. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370

4. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

5. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol*. 2014;5(3):412. doi:10.5306/WJCO.V5.I3.412

6. Johnson KS, Conant EF, Soo MS. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *J Breast Imaging*. 2021;3(1):12-24. doi:10.1093/JBI/WBAA110

7. Dietze EC, Sistrunk C, Miranda-Carboni G, O'Regan R, Seewaldt VL. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat Rev Cancer*. 2015;15(4):248-254. doi:10.1038/nrc3896

8. Carey LA, Perou CM, Livasy CA, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *JAMA*. 2006;295(21):2492-2502. doi:10.1001/JAMA.295.21.2492

9. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev*. 2017;26(4):444-457. doi:10.1158/1055-9965.EPI-16-0858

10. Newman LA. Breast Cancer Disparities: Socioeconomic Factors versus Biology. *Ann Surg Oncol*. 2017;24(10):2869-2875. doi:10.1245/S10434-017-5977-1/TABLES/2

11. Ooi SL, Martinez ME, Li CI. Disparities in breast cancer characteristics and outcomes by race/ethnicity. *Breast Cancer Res Treat*. 2011;127(3):729-738. doi:10.1007/S10549-010-1191-6/TABLES/6

12. Newman LA, Kaljee LM. Health Disparities and Triple-Negative Breast Cancer in African American Women. *JAMA Surg*. 2017;152(5):485. doi:10.1001/jamasurg.2017.0005

13. Stringer-Reasor EM, Elkhanany A, Khoury K, Simon MA, Newman LA. Disparities in Breast Cancer Associated With African American Identity. *Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet*. 2021;41(41):e29-e46. doi:10.1200/EDBK_319929

14. Yedjou CG, Sims JN, Miele L, et al. Health and Racial Disparity in Breast Cancer. *Adv*

Exp Med Biol. 2019;1152:31. doi:10.1007/978-3-030-20301-6_3

15. Newman LA, Carpten J. Integrating the Genetics of Race and Ethnicity Into Cancer Research: Trailing Jane and John Q. Public. *JAMA Surg*. 2018;153(4):299. doi:10.1001/JAMASURG.2017.5080

16. Sung H, DeSantis CE, Fedewa SA, Kantelhardt EJ, Jemal A. Breast cancer subtypes among Eastern-African-born black women and other black women in the United States. *Cancer*. 2019;125(19):3401-3411. doi:10.1002/CNCR.32293

17. Jiagge E, Jibril AS, Chitale D, et al. Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer. *Ann Surg Oncol*. 2016;23(12):3843-3849. doi:10.1245/S10434-016-5420-Z/TABLES/5

18. Capp J-P. Cancer Stem Cells: From Historical Roots to a New Perspective. *J Oncol*. 2019;2019:1-10. doi:10.1155/2019/5189232

19. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. doi:10.1038/35102167

20. Ma Y, Zhang P, Wang F, Yang J, Yang Z, Qin H. The relationship between early embryo development and tumourigenesis. *J Cell Mol Med*. 2010;14(12):2697-2701. doi:10.1111/j.1582-4934.2010.01191.x

21. Reya T, Clevers H. Wnt signalling in stem cells and cancer. *Nature*. 2005;434(7035):843-850. doi:10.1038/nature03319

22. Papaccio F, Paino F, Regad T, Papaccio G, Desiderio V, Tirino V. Concise Review: Cancer Cells, Cancer Stem Cells, and Mesenchymal Stem Cells: Influence in Cancer Development. *Stem Cells Transl Med*. 2017;6(12):2115-2125. doi:10.1002/sctm.17-0138

23. Clevers H. The cancer stem cell: premises, promises and challenges. *Nat Med*. 2011;17(3):313-319. doi:10.1038/nm.2304

24. Goichberg P. Current Understanding of the Pathways Involved in Adult Stem and Progenitor Cell Migration for Tissue Homeostasis and Repair. *Stem Cell Rev Reports*. 2016;12(4):421-437. doi:10.1007/s12015-016-9663-7

25. Shaw TJ, Martin P. Wound repair: a showcase for cell plasticity and migration. *Curr Opin Cell Biol*. 2016;42:29-37. doi:10.1016/J.CEB.2016.04.001

26. Hennighausen L, Robinson GW. Signaling Pathways in Mammary Gland Development. *Dev Cell*. 2001;1(4):467-475. doi:10.1016/S1534-5807(01)00064-8

27. Macias H, Hinck L. Mammary Gland Development. *Wiley Interdiscip Rev Dev Biol*. 2012;1(4):533. doi:10.1002/WDEV.35

28. Russo J. The Windows of Susceptibility to Breast Cancer. *Pathobiol Breast Cancer*. Published online 2016:1-20. doi:10.1007/978-3-319-40815-6_1

29. Mammary Stem Cells: Premise, Properties, and Perspectives. *Trends Cell Biol*. 2017;27(8):556-567. doi:10.1016/J.TCB.2017.04.001

30. Shackleton M, Vaillant F, Simpson KJ, et al. Generation of a functional mammary gland from a single stem cell. *Nature*. 2006;439(7072):84-88. doi:10.1038/nature04372

31. Daniel CW, Deome KB, Young JT, Blair PB, Faulkin LJ. The in vivo life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *J Mammary Gland Biol Neoplasia*. 2009;14(3):355-362. doi:10.1007/s10911-009-9139-3

32. Stingl J, Eaves CJ, Zandieh I, Emerman JT. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat*. 2001;67(2):93-109. doi:10.1023/A:1010615124301

33. Ghebeh H, Sleiman GM, Manogaran PS, et al. Profiling of normal and malignant breast tissue show CD44high/CD24low phenotype as a predominant stem/progenitor marker when used in combination with Ep-CAM/CD49f markers. *BMC Cancer*. 2013;13:289. doi:10.1186/1471-2407-13-289

34. Ginestier C, Hur MH, Charafe-Jauffret E, et al. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell*. 2007;1(5):555-567. doi:10.1016/j.stem.2007.08.014

35. Schwartz T, Stark A, Pang J, et al. Expression of aldehyde dehydrogenase 1 as a marker of mammary stem cells in benign and malignant breast lesions of Ghanaian women. *Cancer*. 2013;119(3):488-494. doi:10.1002/cncr.27737

36. Jiagge E, Chitale D, Newman LA, Newman A. Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *J Pathol*. 2018;188(2). doi:10.1016/j.ajpath.2017.06.020

37. Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature*. 2016;529(7584):43-47. doi:10.1038/nature16166

38. Gray J, Evans N, Taylor B, Rizzo J, Walker M. State of the Evidence: The Connection Between Breast Cancer and the Environment. *Int J Occup Environ Health*. 2009;15(1):43-78. doi:10.1179/107735209799449761

39. Forman MR, Winn DM, Collman GW, Rizzo J, Birnbaum LS. Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod Toxicol*. 2015;54:6-10. doi:10.1016/j.reprotox.2014.10.019

40. Nakshatri H, Anjanappa M, Bhat-Nakshatri P. Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. *Sci Rep*. 2015;5:13526. doi:10.1038/srep13526

41. Colacino JA, Azizi E, Brooks MD, et al. Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*. Published online 2018. doi:10.1016/j.stemcr.2018.03.001

42. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med.* 2009;27(5):351-357. doi:10.1055/s-0029-1237423

43. Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol Rev*. 2020;100(2):489-523. doi:10.1152/PHYSREV.00040.2018/ASSET/IMAGES/LARGE/Z9J0012029270010.JPG

44. Terry MB, Michels KB, Brody JG, et al. Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research. *Breast Cancer Res 2019 211*. 2019;21(1):1-16. doi:10.1186/S13058-019-1168-2

45. Attina TM, Malits J, Naidu M, Trasande L. Racial/ethnic disparities in disease burden and costs related to exposure to endocrine-disrupting chemicals in the United States: an exploratory analysis. *J Clin Epidemiol*. 2019;108:34-43. doi:10.1016/j.jclinepi.2018.11.024

46. Ruiz D, Becerra M, Jagai JS, Ard K, Sargis RM. Disparities in Environmental Exposures to Endocrine-Disrupting Chemicals and Diabetes Risk in Vulnerable Populations. *Diabetes Care*. 2018;41(1):193-205. doi:10.2337/dc16-2765

47. Lillo MA, Nichols C, Seagroves TN, Miranda-Carboni GA, Krum SA. Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Horm Cancer*. 2017;8(2):90-99. doi:10.1007/s12672-017-0286-5

48. Hindman AR, Mo XM, Helber HL, et al. Varying Susceptibility of the Female Mammary Gland to In Utero Windows of BPA Exposure. *Endocrinology*. 2017;158(10):3435-3447. doi:10.1210/en.2017-00116

49. Gerona RR, Pan J, Zota AR, et al. Direct measurement of Bisphenol A (BPA), BPA glucuronide and BPA sulfate in a diverse and low-income population of pregnant women reveals high exposure, with potential implications for previous exposure estimates: a cross-sectional study. *Environ Health*. 2016;15:50. doi:10.1186/s12940-016-0131-2

50. Tucker DK, Hayes Bouknight S, Brar SS, Kissling GE, Fenton SE. Evaluation of Prenatal Exposure to Bisphenol Analogues on Development and Long-Term Health of the Mammary Gland in Female Mice. *Environ Health Perspect*. 2018;126(8):087003. doi:10.1289/EHP3189

51. Huang W, Zhao C, Zhong H, Zhang S, Xia Y, Cai Z. Bisphenol S induced epigenetic and transcriptional changes in human breast cancer cell line MCF-7. *Environ Pollut*. 2019;246:697-703. doi:10.1016/J.ENVPOL.2018.12.084

52. Weng Y-I, Hsu P-Y, Liyanarachchi S, et al. Epigenetic influences of low-dose bisphenol A in primary human breast epithelial cells. *Toxicol Appl Pharmacol*. 2010;248(2):111. doi:10.1016/J.TAAP.2010.07.014

53. Liu X, Krawczyk E, Suprynowicz FA, et al. Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc*. 2017;12(2):439-451. doi:10.1038/nprot.2016.174

54. Giraddi RR, Chung C-Y, Heinz RE, et al. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep*. 2018;24(6):1653-1666.e7. doi:10.1016/j.celrep.2018.07.025

55. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18(1):35-45. doi:10.1038/nri.2017.76

56. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257-272. doi:10.1038/s41576-019-0093-7

57. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health*. 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629

58. Holliday H, Baker LA, Junankar SR, Clark SJ, Swarbrick A. Epigenomics of mammary gland development. *Breast Cancer Res*. 2018;20(1):100. doi:10.1186/s13058-018-1031-x

59. Thong T, Forté CA, Hill EM, Colacino JA. Environmental exposures, stem cells, and cancer. *Pharmacol Ther*. Published online 2019. doi:10.1016/j.pharmthera.2019.107398

60. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest*. 2009;119(6):1420-1428. doi:10.1172/JCI39104

61. Merrell AJ, Stanger BZ. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style. *Nat Rev Mol Cell Biol*. 2016;17(7):413-425. doi:10.1038/nrm.2016.24

62. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med*. 2013;19(11):1438-1449. doi:10.1038/nm.3336

63. Jolly MK, Boareto M, Huang B, et al. Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol*. 2015;5:155. doi:10.3389/fonc.2015.00155

64. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

65. Jolly MK, Tripathi SC, Jia D, et al. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*. 2016;7(19):27067-27084. doi:10.18632/oncotarget.8166

66. Dontu G, Abdallah WM, Foley JM, et al. In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev*. 2003;17(10):1253-1270. doi:10.1101/gad.1061803

67. Colacino JA, McDermott SP, Sartor MA, Wicha MS, Rozek LS. Transcriptomic profiling of curcumin-treated human breast stem cells identifies a role for stearoyl-coa desaturase in breast cancer prevention. *Breast Cancer Res Treat*. 2016;158(1):29-41. doi:10.1007/s10549-016-3854-4

68. Liu X, Ory V, Chapman S, et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am J Pathol*. 2012;180(2):599-607. doi:10.1016/j.ajpath.2011.10.036

69. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/J.CELL.2015.05.002

70. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33(5):495-502. doi:10.1038/nbt.3192

71. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic

sequencing studies. Birol I, ed. *Bioinformatics*. 2019;35(3):487-496. doi:10.1093/bioinformatics/bty644

72. Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, Chang HY. Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells. *Cell Stem Cell*. 2008;2(4):333-344. doi:10.1016/j.stem.2008.02.009

73. Kuleshov M V., Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

74. Lim E, Vaillant F, Wu D, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*. 2009;15(8):907-913. doi:10.1038/nm.2000

75. Bach K, Pensa S, Grzelak M, et al. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun*. 2017;8(1):2128. doi:10.1038/s41467-017-02001-5

76. Nguyen QH, Pervolarakis N, Blake K, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-04334-1

77. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71-e71. doi:10.1093/nar/gkv1507

78. Wang Y, Thong T, Saligrama V, Colacino J, Balzano L, Scott C. A gene filter for comparative analysis of single-cell RNA-sequencing trajectory datasets. *bioRxiv*. Published online May 17, 2019:637488. doi:10.1101/637488

79. Abd El-Rehim DM, Pinder SE, Paish CE, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol*. 2004;203(2):661-671. doi:10.1002/path.1559

80. O'Hare MJ, Ormerod MG, Monaghan P, Lane EB, Gusterson BA. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. *Differentiation*. 1991;46(3):209-221. doi:10.1111/j.1432-0436.1991.tb00883.x

81. Marcato P, Dean CA, Pan D, et al. Aldehyde Dehydrogenase Activity of Breast Cancer Stem Cells Is Primarily Due To Isoform ALDH1A3 and Its Expression Is Predictive of Metastasis. *Stem Cells*. 2011;29(1):32-45. doi:10.1002/stem.563

82. Trzpis M, McLaughlin PMJ, De Leij LMFH, Harmsen MC. Epithelial cell adhesion molecule: More than a carcinoma marker and adhesion molecule. *Am J Pathol*. 2007;171(2):386-395. doi:10.2353/ajpath.2007.070152

83. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A*. 2003;100(7):3983-3988. doi:10.1073/pnas.0530291100

84. Gerdur Ísberg Ó, Kim J, Fridriksdottir AJ, et al. A CD146 FACS Protocol Enriches for

Luminal Keratin 14/19 Double Positive Human Breast Progenitors. *Sci Rep*. 2019;9(1):14843. doi:10.1038/s41598-019-50903-9

85. Jolly MK, Somarelli JA, Sheth M, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Ther*. 2019;194:161-184. doi:10.1016/j.pharmthera.2018.09.007

86. Casey T, Bond J, Tighe S, et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res Treat*. 2009;114(1):47-62. doi:10.1007/s10549-008-9982-8

87. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173(2):338-354.e15. doi:10.1016/j.cell.2018.03.034

88. Feng Y, Sun B, Li X, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. *Breast Cancer Res Treat*. 2007;103(3):319-329. doi:10.1007/s10549-006-9385-7

89. Folgueira MAAK, Maistro S, Katayama MLH, et al. Markers of breast cancer stromal fibroblasts in the primary tumour site associated with lymph node metastasis: A systematic review including our case series. *Biosci Rep*. 2013;33(6). doi:10.1042/BSR20130060

90. Li J-M, Tseng C-W, Lin C-C, et al. Upregulation of LGALS1 is associated with oral cancer metastasis. *Ther Adv Med Oncol*. 2018;10:175883591879462. doi:10.1177/1758835918794622

91. Falato C, Lorent J, Tani E, et al. Ki67 measured in metastatic tissue and prognosis in patients with advanced breast cancer. *Breast Cancer Res Treat*. 2014;147(2):407-414. doi:10.1007/s10549-014-3096-2

92. Montes de Oca R, Gurard-Levin ZA, Berger F, et al. The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol Oncol*. 2015;9(3):657-674. doi:10.1016/j.molonc.2014.11.002

93. Ding K, Li W, Zou Z, Zou X, Wang C. CCNB1 is a prognostic biomarker for ER+ breast cancer. *Med Hypotheses*. 2014;83(3):359-364. doi:10.1016/j.mehy.2014.06.013

94. Li Y, Chen Y-L, Xie Y-T, et al. Association Study of Germline Variants in CCNB1 and CDK1 with Breast Cancer Susceptibility, Progression, and Survival among Chinese Han Women. Miao X, ed. *PLoS One*. 2013;8(12):e84489. doi:10.1371/journal.pone.0084489

95. Wang Y, Zhang C, Mai L, Niu Y, Wang Y, Bu Y. PRR11 and SKA2 gene pair is overexpressed and regulated by p53 in breast cancer. *BMB Rep*. 2019;52(2):157-162. doi:10.5483/BMBRep.2019.52.2.207

96. Ren Z, Yang T, Zhang P, Liu K, Liu W, Wang P. SKA2 mediates invasion and metastasis in human breast cancer via EMT. *Mol Med Rep*. 2019;19(1):515-523. doi:10.3892/mmr.2018.9623

97. Han JY, Han YK, Park GY, et al. Bub1 is required for maintaining cancer stem cells in

breast cancer cell lines. *Sci Rep*. 2015;5. doi:10.1038/srep15993

98. Hamy AS, Bieche I, Lehmann-Che J, et al. BIRC5 (survivin): a pejorative prognostic marker in stage II/III breast cancer with no response to neoadjuvant chemotherapy. *Breast Cancer Res Treat*. 2016;159(3):499-511. doi:10.1007/s10549-016-3961-2

99. Ghaffari K, Hashemi M, Ebrahimi E, Shirkoohi R. BIRC5 genomic copy number variation in early-onset breast cancer. *Iran Biomed J*. 2016;20(4):241-245. doi:10.7508/ibj.2016.04.009

100. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

101. Yu M, Bardia A, Wittner BS, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (80-)*. 2013;339(6119):580-584. doi:10.1126/science.1228522

102. Sun P, Yuan Y, Li A, Li B, Dai X. Cytokeratin expression during mouse embryonic and early postnatal mammary gland development. *Histochem Cell Biol*. 2010;133(2):213-221. doi:10.1007/s00418-009-0662-5

103. Ruiz R, Herrero C, Strasser-Weippl K, et al. Epidemiology and pathophysiology of pregnancy-associated breast cancer: A review. *Breast*. 2017;35:136-141. doi:10.1016/j.breast.2017.07.008

104. Johansson ALV, Andersson TML, Hsieh CC, et al. Tumor characteristics and prognosis in women with pregnancy-associated breast cancer. *Int J Cancer*. 2018;142(7):1343-1354. doi:10.1002/ijc.31174

105. Degnim AC, Visscher DW, Hoskin TL, et al. Histologic findings in normal breast tissues: comparison to reduction mammaplasty and benign breast disease tissues. *Breast Cancer Res Treat*. 2012;133(1):169-177. doi:10.1007/s10549-011-1746-1

106. Rheinwatd JG, Green H. Seria cultivation of strains of human epidemal keratinocytes: the formation keratinizin colonies from single cell is. *Cell*. 1975;6(3):331-343. doi:10.1016/S0092-8674(75)80001-8

107. Domchek SM, Yao S, Chen F, et al. Comparison of the Prevalence of Pathogenic Variants in Cancer Susceptibility Genes in Black Women and Non-Hispanic White Women With Breast Cancer in the United States. *JAMA Oncol*. 2021;7(7):1. doi:10.1001/JAMAONCOL.2021.1492

108. Shimelis H, LaDuca H, Hu C, et al. Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst*. 2018;110(8):855-862. doi:10.1093/jnci/djy106

109. Roelands J, Mall R, Almeer H, et al. Ancestry-associated transcriptomic profiles of breast cancer in patients of African, Arab, and European ancestry. *NPJ Breast Cancer*. 2021;7(1). doi:10.1038/S41523-021-00215-X

110. Nievergelt CM, Maihofer AX, Shekhtman T, et al. Inference of human continental origin

and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*. 2013;4(1):13. doi:10.1186/2041-2223-4-13

111. Thong T, Wang Y, Brooks MD, et al. Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Front Cell Dev Biol*. 2020;8:288. doi:10.3389/FCELL.2020.00288/BIBTEX

112. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol 2018 365*. 2018;36(5):411-420. doi:10.1038/nbt.4096

113. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048/ATTACHMENT/1990279C-028D-47D4-A5D3-C89B04E14795/MMC3.XLSX

114. Chapter 4 DE analyses between conditions | Multi-Sample Single-Cell Analyses with Bioconductor. Accessed May 25, 2022. http://bioconductor.org/books/3.14/OSCA.multisample/multi-sample-comparisons.html

115. Gulati GS, Sikandar SS, Wesche DJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*. 2020;367(6476):405-411. doi:10.1126/SCIENCE.AAX0249

116. Using escape to perform ssGSEA analyses on single-cell RNA-seq data. Accessed May 25, 2022. http://www.bioconductor.org/packages/release/bioc/vignettes/escape/inst/doc/vignette.html

117. GSEA | MSigDB | Browse Gene Sets. Accessed May 25, 2022. https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

118. Cell-Cycle Scoring and Regression • Seurat. Accessed May 19, 2022. https://satijalab.org/seurat/articles/cell_cycle_vignette.html

119. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/BIOINFORMATICS/BTS480

120. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363. doi:10.1093/BIOINFORMATICS/BTU049

121. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res*. 2016;44(3):e20. doi:10.1093/NAR/GKV907

122. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics and Chromatin*. 2015;8(1):1-16. doi:10.1186/1756-8935-8-6/FIGURES/6

123. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina

MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208. doi:10.1186/s13059-016-1066-1

124. Fortin J-P, Triche TJ, Hansen KD, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2017;33(4):558-560. doi:10.1093/bioinformatics/btw691

125. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287. doi:10.1038/NG.3656

126. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443. doi:10.1038/NG.3679

127. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(10):1654. doi:10.1001/JAMAONCOL.2017.0595

128. Wu H, Wang C, Wu Z. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*. 2015;31(2):233-241. doi:10.1093/BIOINFORMATICS/BTU640

129. Yap SQ, Mathavarajah S, Huber RJ. The converging roles of Batten disease proteins in neurodegeneration and cancer. *iScience*. 2021;24(4):102337. doi:10.1016/J.ISCI.2021.102337

130. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet (London, England)*. 2005;365(9460):671-679. doi:10.1016/S0140-6736(05)17947-1

131. Pani T, Rajput K, Kar A, et al. Alternative splicing of ceramide synthase 2 alters levels of specific ceramides and modulates cancer cell proliferation and migration in Luminal B breast cancer subtype. *Cell Death Dis 2021 122*. 2021;12(2):1-22. doi:10.1038/s41419-021-03436-x

132. D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, Troester MA. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res Treat*. 2015;152(2):437-448. doi:10.1007/S10549-015-3474-4/TABLES/6

133. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(12):1654-1662. doi:10.1001/JAMAONCOL.2017.0595

134. Barrow MA, Martin ME, Coffey A, et al. A functional role for the cancer disparity-linked genes, CRYβB2 and CRYβB2P1, in the promotion of breast cancer. *Breast Cancer Res*. 2019;21(1):105. doi:10.1186/s13058-019-1191-3

135. Field LA, Love B, Deyarmin B, Hooke JA, Shriver CD, Ellsworth RE. Identification of differentially expressed genes in breast tumors from African American compared with Caucasian women. *Cancer*. 2012;118(5):1334-1344. doi:10.1002/CNCR.26405

136. Li H, Cheng C, You W, et al. TMEM100 Modulates TGF-β Signaling Pathway to Inhibit Colorectal Cancer Progression. *Gastroenterol Res Pract*. 2021;2021.

doi:10.1155/2021/5552324

137. Ma J, Yan T, Bai Y, et al. TMEM100 negatively regulated by microRNA-106b facilitates cellular apoptosis by suppressing survivin expression in NSCLC. *Oncol Rep*. 2021;46(2):1-14. doi:10.3892/OR.2021.8136

138. Zhuang J, Huang Y, Zheng W, et al. TMEm100 expression suppresses metastasis and enhances sensitivity to chemotherapy in gastric cancer. *Biol Chem*. 2020;401(2):285-296. doi:10.1515/HSZ-2019-0161/MACHINEREADABLECITATION/RIS

139. Donald CD, Sun CQ, Lim SD, et al. Cancer-Specific Loss of β-Defensin 1 in Renal and Prostatic Carcinomas. *Lab Investig 2003 834*. 2003;83(4):501-505. doi:10.1097/01.lab.0000063929.61760.f6

140. Lee J, Han JH, Jang A, Kim JW, Hong SA, Myung SC. DNA Methylation-Mediated Downregulation of DEFB1 in Prostate Cancer Cells. *PLoS One*. 2016;11(11). doi:10.1371/JOURNAL.PONE.0166664

141. Lo JH, Battaglin F, Baca Y, et al. DEFB1 gene expression and the molecular landscape of colorectal cancer (CRC). *https://doi.org/101200/JCO20224016_suppl3523*. 2022;40(16_suppl):3523-3523. doi:10.1200/JCO.2022.40.16_SUPPL.3523

142. Li YH, Hu YQ, Wang SC, Li Y, Chen DM. LncRNA SNHG5: A new budding star in human cancers. *Gene*. 2020;749:144724. doi:10.1016/J.GENE.2020.144724

143. Chi JR, Yu ZH, Liu BW, et al. SNHG5 Promotes Breast Cancer Proliferation by Sponging the miR-154-5p/PCNA Axis. *Mol Ther Nucleic Acids*. 2019;17:138. doi:10.1016/J.OMTN.2019.05.013

144. Damas ND, Marcatti M, Côme C, et al. SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nat Commun 2016 71*. 2016;7(1):1-14. doi:10.1038/ncomms13875

145. Shen M, Jiang YZ, Wei Y, et al. Tinagl1 Suppresses Triple-Negative Breast Cancer Progression and Metastasis by Simultaneously Inhibiting Integrin/FAK and EGFR Signaling. *Cancer Cell*. 2019;35(1):64-80.e7. doi:10.1016/J.CCELL.2018.11.016/ATTACHMENT/1E19EDC0-441D-4164-93EE-F3A5B9EC190A/MMC1.PDF

146. Musetti SN, Huang L. Tinagl1 Gene Therapy Suppresses Growth and Remodels the Microenvironment of Triple Negative Breast Cancer. *Mol Pharm*. 2021;18(5):2032-2038. doi:10.1021/ACS.MOLPHARMACEUT.1C00008

147. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob Heal*. 2020;8(8):e1027-e1037. doi:10.1016/S2214-109X(20)30215-1

148. Zota AR, Shamasunder B. The environmental injustice of beauty: framing chemical exposures from beauty products as a health disparities concern. *Am J Obstet Gynecol*. 2017;217(4):418.e1-418.e6. doi:10.1016/J.AJOG.2017.07.020

149. Ash M, Boyce JK. Racial disparities in pollution exposure and employment at US industrial facilities. *Proc Natl Acad Sci U S A*. 2018;115(42):10636-10641. doi:10.1073/PNAS.1721640115/SUPPL_FILE/PNAS.1721640115.SAPP.PDF

150. Polemi KM, Nguyen VK, Heidt J, Kahana A, Jolliet O, Colacino JA. Identifying the link between chemical exposures and breast cancer in African American women via integrated in vitro and exposure biomarker data. *Toxicology*. 2021;463. doi:10.1016/J.TOX.2021.152964

151. Nguyen VK, Kahana A, Heidt J, et al. A comprehensive analysis of racial disparities in chemical biomarker concentrations in United States women, 1999–2014. *Environ Int*. 2020;137:105496. doi:10.1016/J.ENVINT.2020.105496

152. Atlas E, Dimitrova V. Bisphenol S and Bisphenol A disrupt morphogenesis of MCF-12A human mammary epithelial cells. *Sci Rep*. 2019;9(1):1-10. doi:10.1038/s41598-019-52505-x

153. LaPlante CD, Catanese MC, Bansal R, Vandenberg LN. Bisphenol S Alters the Lactating Mammary Gland and Nursing Behaviors in Mice Exposed During Pregnancy and Lactation. *Endocrinology*. 2017;158(10):3448-3461. doi:10.1210/EN.2017-00437

154. Winkler J, Liu P, Phong K, et al. Bisphenol A replacement chemicals, BPF and BPS, induce protumorigenic changes in human mammary gland organoid morphology and proteome. *Proc Natl Acad Sci U S A*. 2022;119(11). doi:10.1073/PNAS.2115308119/-/DCSUPPLEMENTAL

155. Erler C, Novak J. Bisphenol A Exposure: Human Risk and Health Policy. *J Pediatr Nurs*. 2010;25(5):400-407. doi:10.1016/J.PEDN.2009.05.006

156. Konieczna A, Rutkowska A, Rachoń D. Health risk of exposure to Bisphenol A (BPA). *Rocz Panstw Zakl Hig*. 2015;66(1):5-11. Accessed August 5, 2022. https://pubmed.ncbi.nlm.nih.gov/25813067/

157. Zimmers SM, Browne EP, O'Keefe PW, et al. Determination of free Bisphenol A (BPA) concentrations in breast milk of U.S. women using a sensitive LC/MS/MS method. *Chemosphere*. 2014;104:237-243. doi:10.1016/j.chemosphere.2013.12.085

158. Lehmler H-J, Liu B, Gadogbe M, Bao W. Exposure to Bisphenol A, Bisphenol F, and Bisphenol S in U.S. Adults and Children: The National Health and Nutrition Examination Survey 2013-2014. *ACS omega*. 2018;3(6):6523-6532. doi:10.1021/acsomega.8b00824

159. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol 2017 361*. 2017;36(1):89-94. doi:10.1038/nbt.4042

160. Cheng M, Michalski S, Kommagani R. Role for Growth Regulation by Estrogen in Breast Cancer 1 (GREB1) in Hormone-Dependent Cancers. *Int J Mol Sci*. 2018;19(9). doi:10.3390/IJMS19092543

161. Hodgkinson K, Forrest LA, Vuong N, Garson K, Djordjevic B, Vanderhyden BC. GREB1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene 2018 3744*. 2018;37(44):5873-5886. doi:10.1038/s41388-018-0377-y

162. Haines CN, Braunreiter KM, Mo XM, Burd CJ. GREB1 isoforms regulate proliferation independent of ERα co-regulator activities in breast cancer. *Endocr Relat Cancer*. 2018;25(7):735-746. doi:10.1530/ERC-17-0496

163. Jiang S, Katayama H, Wang J, et al. Estrogen-Induced Aurora Kinase-A (AURKA) Gene Expression is Activated by GATA-3 in Estrogen Receptor-Positive Breast Cancer Cells. *Horm Cancer*. 2010;1(1):11. doi:10.1007/S12672-010-0006-X

164. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Mol Cancer 2021 201*. 2021;20(1):1-27. doi:10.1186/S12943-020-01305-3

165. Ma XF, Zhang J, Shuai HL, Guan BZ, Luo X, Yan RL. IKKβ/NF-κB mediated the low doses of bisphenol A induced migration of cervical cancer cells. *Arch Biochem Biophys*. 2015;573:52-58. doi:10.1016/J.ABB.2015.03.010

166. Sauer SJ, Tarpley M, Shah I, et al. Bisphenol A activates EGFR and ERK promoting proliferation, tumor spheroidformation and resistance to EGFR pathway inhibition in estrogen receptor-negativeinflammatory breast cancer cells. *Carcinogenesis*. 2017;38(3):252-260. doi:10.1093/CARCIN/BGX003

167. Li S, Wang B, Tang Q, Liu J, Yang X. Bisphenol A triggers proliferation and migration of laryngeal squamous cell carcinoma via GPER mediated upregulation of IL-6. *Cell Biochem Funct*. 2017;35(4):209-216. doi:10.1002/CBF.3265

168. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia (United States)*. 2017;19(8):649-658. doi:10.1016/j.neo.2017.05.002

169. Yang N, Wang C, Wang Z, et al. FOXM1 recruits nuclear Aurora kinase A to participate in a positive feedback loop essential for the self-renewal of breast cancer stem cells. *Oncogene*. 2017;36(24):3428-3440. doi:10.1038/ONC.2016.490

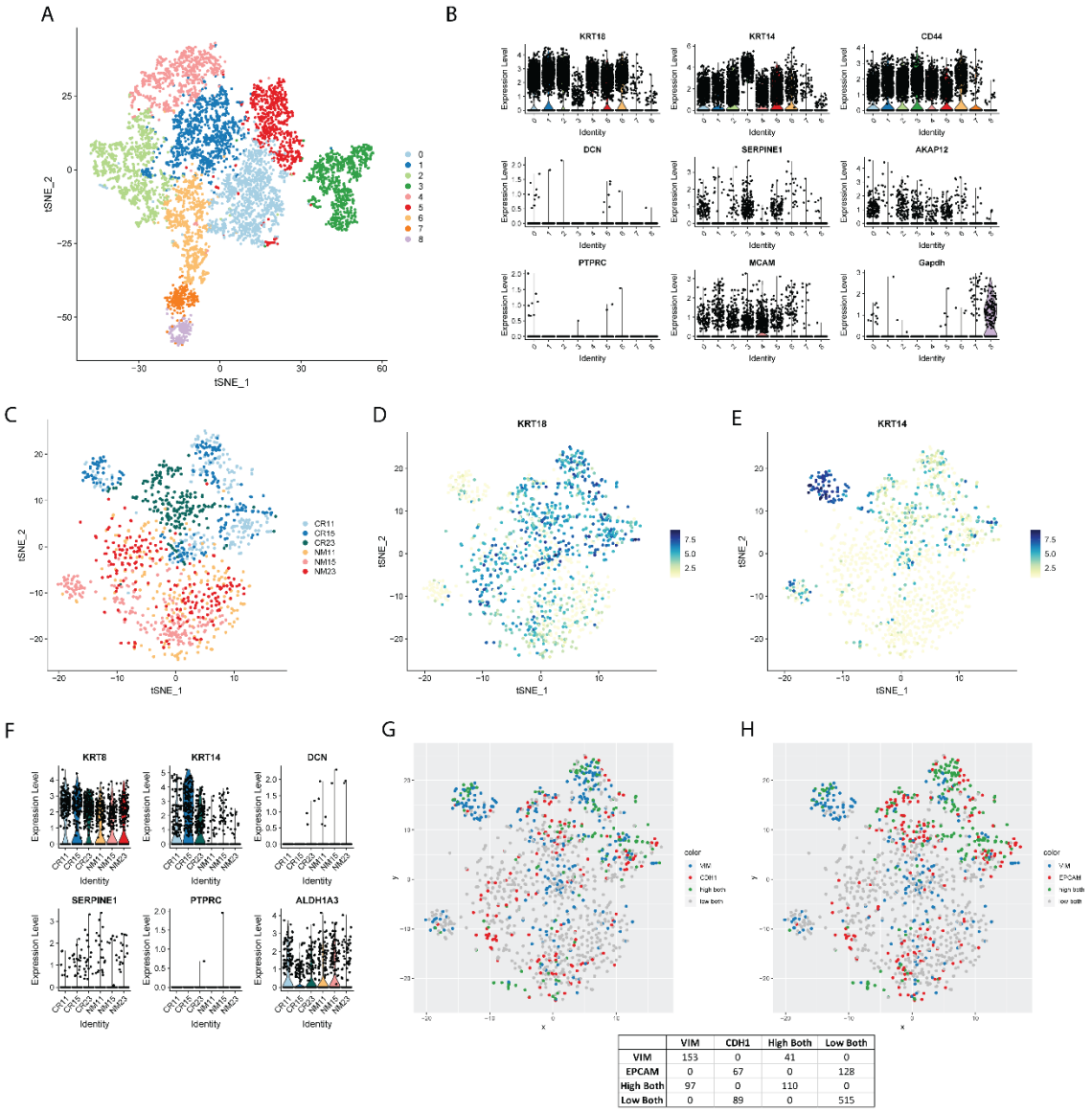
170. Eterno V, Zambelli A, Villani L, et al. AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. *Sci Reports 2016 61*. 2016;6(1):1-13. doi:10.1038/srep28436

171. Tan Y, Wang Q, Xie Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. *Int J Oncol*. 2019;54(1):87-97. doi:10.3892/ijo.2018.4598

172. Jian W, Deng XC, Munankarmy A, et al. KIF23 promotes triple negative breast cancer through activating epithelial-mesenchymal transition. *Gland Surg*. 2021;10(6):1941. doi:10.21037/GS-21-19

173. Li Z, Yang HY, Zhang XL, et al. Kinesin family member 23, regulated by FOXM1, promotes triple negative breast cancer progression via activating Wnt/β-catenin pathway. *J Exp Clin Cancer Res*. 2022;41(1):1-15. doi:10.1186/S13046-022-02373-7/FIGURES/9

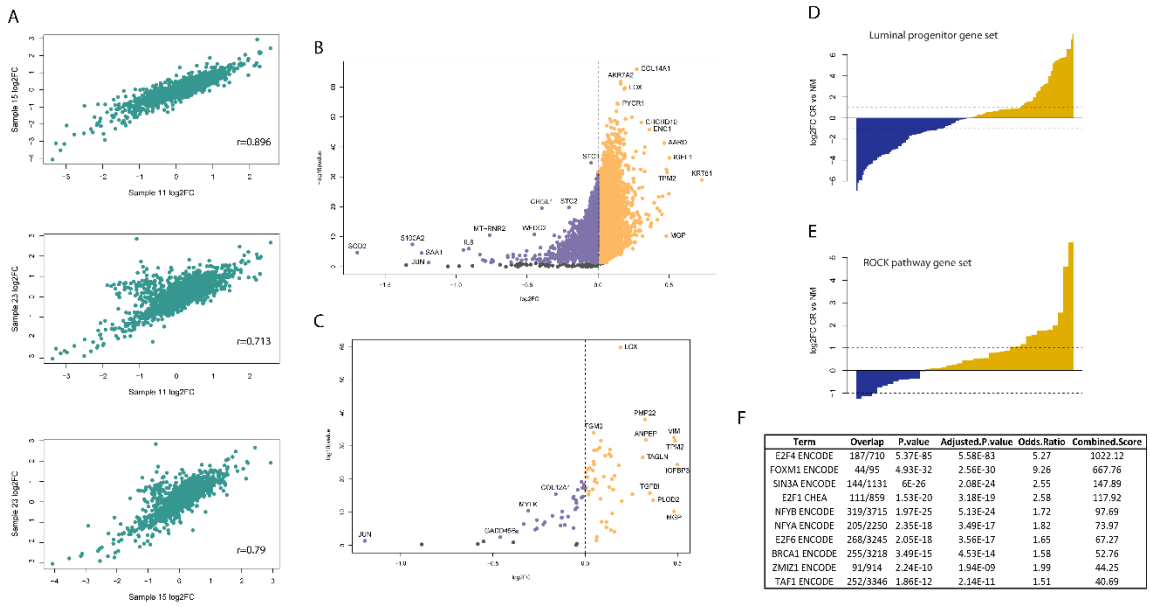
Appendix



	VIM	CDH1	High Both	Low Both
VIM	153	0	41	0
EPCAM	0	67	0	128
High Both	97	0	110	0
Low Both	0	89	0	515

Figure 2.A1: Post-CR single cell unbiased clustering and gene expression (A) tSNE dimension reduction of CR cells colored by cell cluster, identified by unbiased clustering (B) Expression of known cell type marker genes by cluster. Cluster 7 and 8 identified as mouse cells (C) tSNE dimension reduction of NM and CR samples by individual. Each individual sample was down-sampled to 200 cells. (D) NM and CR FeaturePlots of myoepithelial marker gene (*KRT14*) and (E) luminal marker gene (*KRT18*) expression (F) Expression of known cell type marker genes by down-sampled NM and CR individuals (G) Identification of NM and CR

CDH1/VIM double positive cells and (H) *EPCAM/VIM* double positive cells. Table compares the overlap between the *CDH1/VIM* and *EPCAM/VIM* classifications.



Term	Overlap	P.value	Adjusted.P.value	Odds.Ratio	Combined.Score
E2F4 ENCODE	187/710	5.37E-85	5.58E-83	5.27	1022.12
FOXM1 ENCODE	44/95	4.93E-32	2.56E-30	9.26	667.76
SIN3A ENCODE	144/1131	6E-26	2.08E-24	2.55	147.89
E2F1 CHEA	111/859	1.53E-20	3.18E-19	2.58	117.92
NFYB ENCODE	319/3715	1.97E-25	5.13E-24	1.72	97.69
NFYA ENCODE	205/2250	2.35E-18	3.49E-17	1.82	73.97
E2F6 ENCODE	268/3245	2.05E-18	3.56E-17	1.65	67.27
BRCA1 ENCODE	255/3218	3.49E-15	4.53E-14	1.58	52.76
ZMIZ1 ENCODE	91/914	2.24E-10	1.94E-09	1.99	44.25
TAF1 ENCODE	252/3346	1.86E-12	2.14E-11	1.51	40.69

Figure 2.A2: Post-CR differential gene expression and pathway analysis (A) Comparison of differentially expressed genes between NM and CR cells of individual samples. DEGs are plotted by average log2FC. Positive values represent genes upregulated in CR and negative values represent genes downregulated in CR (B) Differential gene expression of quadruple hybrids vs all other NM and CR cells. Significantly upregulated genes in quadruple positive hybrids (FDR<0.05) are colored in orange. Significantly downregulated genes in quadruple hybrids are colored in purple (C) Overlap between quadruple hybrid upregulated genes and EMT related genes. Upregulated EMT genes in quadruple hybrids in orange, and downregulated EMT genes in purple (D) Comparison of overlap between NM and CR differentially expressed genes and the luminal progenitor gene expression signature reported in Lim et al. (2009). Yellow genes indicate luminal progenitor genes more highly expressed in CR vs NM (E) Comparison of overlap between NM and CR differentially expressed genes and the ROCK pathway gene set (F) Top 10 transcription factors associated with top 1000 genes overexpressed in CR cells.

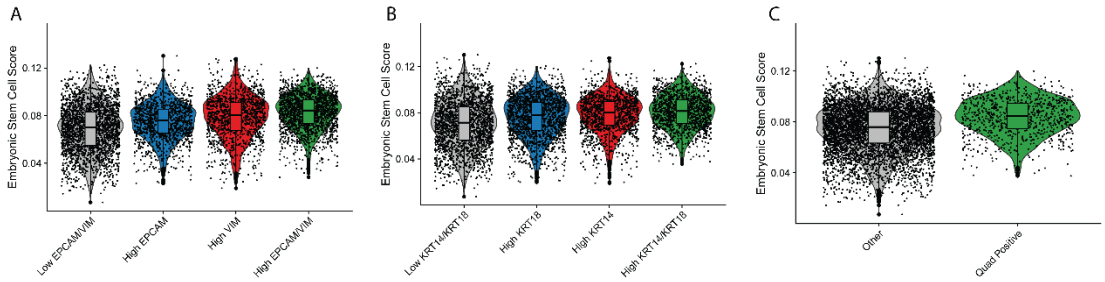


Figure 2.A3: Embryonic stem cell gene signature of NM and CR hybrid cells (A) ESC score of NM and CR cells labeled by *EPCAM/VIM* hybrids status, (B) *KRT14/KRT18* hybrid status, and (C) quadruple positive hybrid status.

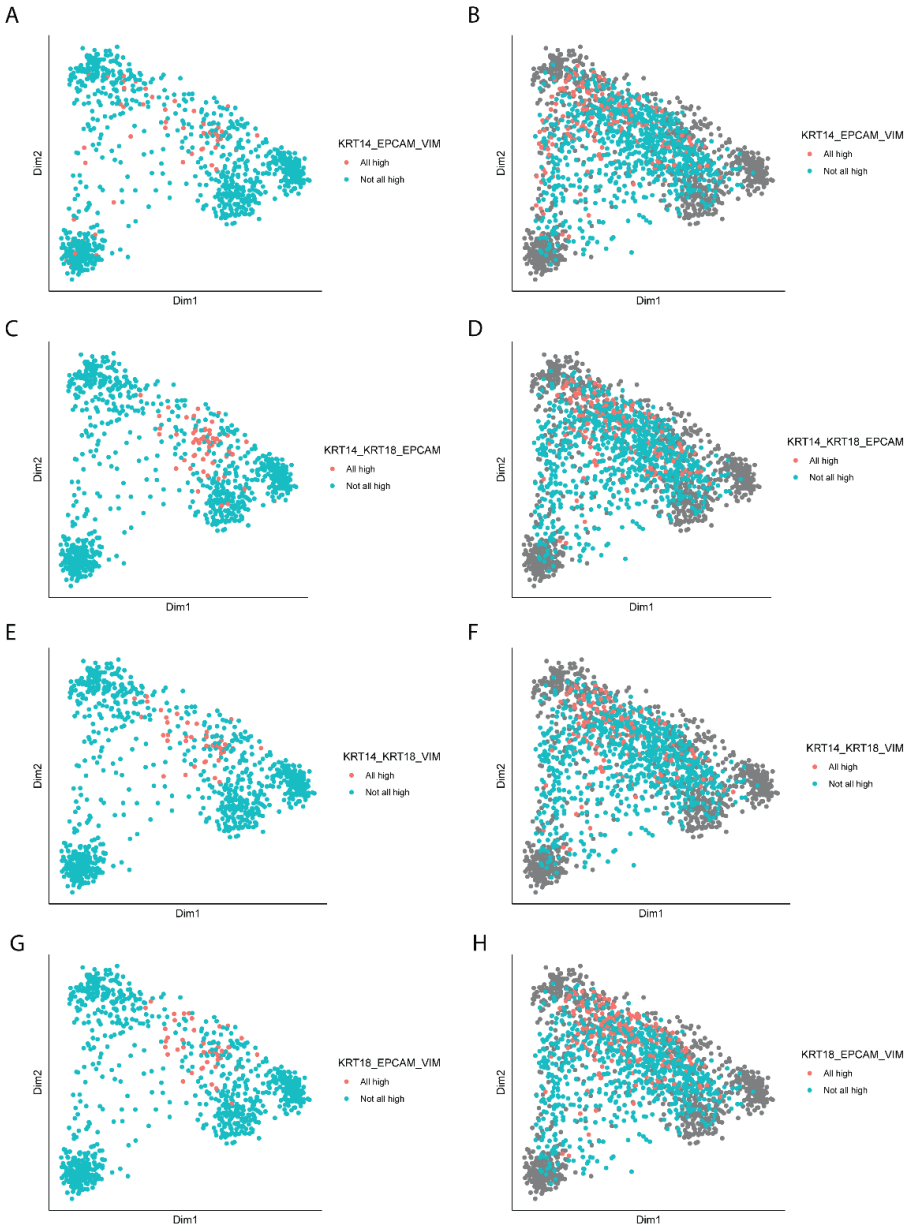
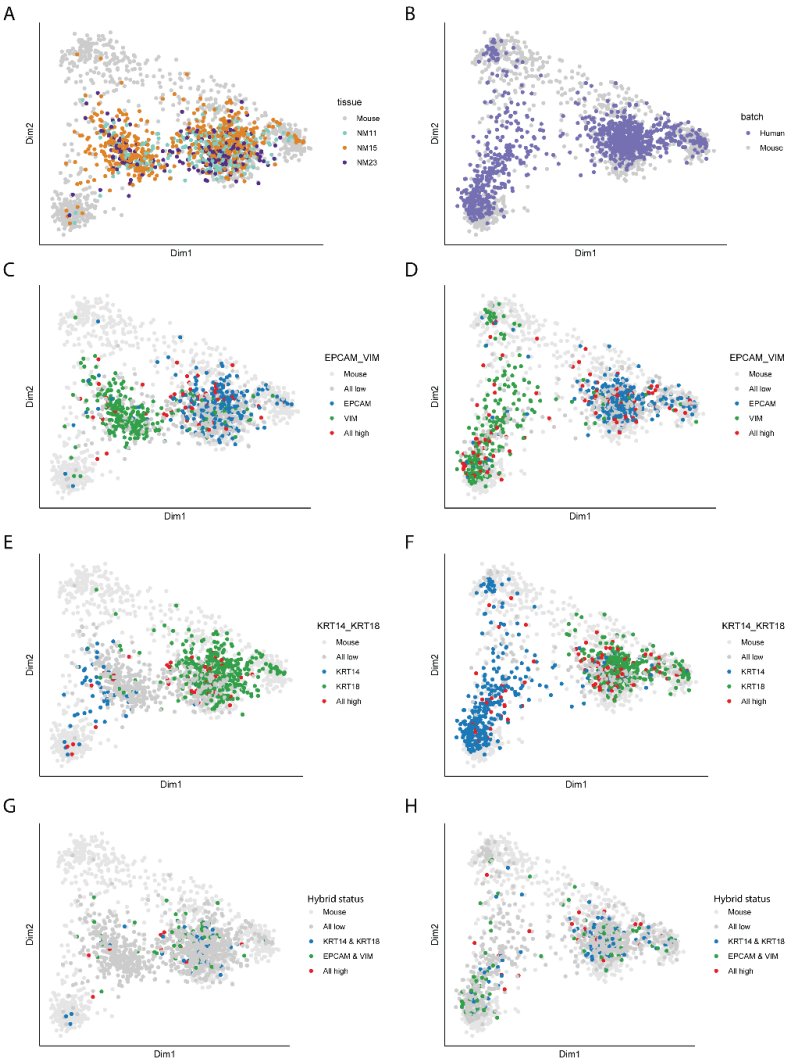


Figure 2.A4: Comparison of triple positive Giraddi mammary cells and aligned CR cells

Localization of (A) mouse *KRT14/EPCAM/VIM* triple positive cells, (B) CR *KRT14/EPCAM/VIM* triple positive cells, (C) mouse *KRT14/KRT18/EPCAM* triple positive cells, (D) CR *KRT14/KRT18/EPCAM* triple positive cells, (E) mouse *KRT14/KRT18/VIM* triple positive cells, (F) CR *KRT14/KRT18/VIM* triple positive cells, (G) mouse *KRT14/EPCAM/VIM* triple positive cells, and (H) CR *KRT14/EPCAM/VIM* triple positive cells when aligned to the mouse mammary developmental trajectory with CoRGI.



I

All low	EPCAM	VIM	All high
0.447	0.216	0.276	0.061

All low	KRT14	KRT18	All high
0.523	0.053	0.38	0.044

J

All low	EPCAM	VIM	All high
0.364	0.285	0.237	0.114

All low	KRT14	KRT18	All high
0.227	0.332	0.323	0.118

Figure 2.A5: Comparison of hybrid NM cells and Nguyen human mammary cells (A) Alignment of NM cells to Giraddi trajectory by individual (B) Alignment of Nguyen mammary cells to Giraddi trajectory (C) NM *EPCAM/VIM* hybrids (D) Nguyen *EPCAM/VIM* hybrids (E) NM KRT14/KRT18 hybrids (F) Nguyen KRT14/KRT18 hybrids (G) NM quadruple positive hybrids (H) Nguyen quadruple positive hybrids (I) Proportions of NM cells by hybrid status (J) Proportions of Nguyen cells by hybrid status

Chapter 3

Establishing a Biobank of Normal Mammary Tissue from Genetically Diverse Individuals and Quantifying Differences Using a Multi-omic Approach

Introduction

In the US, women of African ancestry have significantly worse cancer outcomes at every age group, yet the biological basis underlying this disparity is still unknown.[10,11] Triple negative breast cancer (TNBC), the most aggressive subtype, is a striking example of this disparity, with African American (AA) women 2-3 times more likely to develop it than European American (EA) women.[7,12] Although approximately 10% of breast cancer diagnoses are attributable to hereditable risk factors, such as BRCA mutations, the vast majority (70-90%) of cases occur in women with no family history, suggesting that extrinsic factors are primarily responsible for lifetime cancer risk.[37]

Vast amounts of resources and time have been poured into characterizing race associated differences in gene expression and single nucleotide polymorphisms (SNPs) that might account for differences in breast cancer risk and outcomes. However, population-based studies conducted to characterize hereditary ancestry differences in known breast cancer risk associated genes such as *BRCA1, BRCA2, PALB2, and BARD1*, have not identified a "smoking gun". In a study conducted by Domchek and colleagues comparing 3946 Black women and 25,287 non-Hispanic white women with breast cancer in the US, they report no difference in prevalence of germline pathogenic variants in the 12 evaluated breast cancer susceptibility genes.[107] Shimelis et al performed multigene hereditary cancer panel testing in order to identify women at elevated risk of TNBC and reported similar trends between EA and AA women in pathogenic variants associated with TNBC risk. [108] Results from population-based studies such as these indicate that germline pathogenic variants alone are not the primary cause of race associated differences observed in breast cancer incidence and survival and that other factors, which could interact with these variants need to be better characterized. Genotypic analysis revealed that variations in somatic mutation as well as copy number variations in breast tumors were more specific to

subtype rather than ancestral subgroup.[109] Similarly, while DEGs have been identified between tumors from AA and EA individuals, these differences have not been linked to significant differences in survival.[109] Additionally, once DEGs are identified, it is challenging to determine if differences in expression are due to ancestry related differences or functional differences in tumor subtype.

A likely contributor to the challenges in identifying race associated DEGs and SNPs is the issue of sample size, specifically a lack of AA samples in these large sequencing and profiling projects.[12] Additionally, individuals of African ancestry in the United States represent a vast range of genetic diversity, representative of highly variable, cross continental diasporic spread. This highlights the pressing need for increased sampling of AA individuals, especially because of the vast genetic diversity they represent.

The majority of studies characterizing race-associated differences have been conducted using tumor tissues from diverse individuals. Tumors functionally exhibit high levels of intratumor genetic heterogeneity due to their carcinogenic transformation, thus comparisons of tumor tissue between AA and EA are limited in distinguishing true race-associated differences versus those that arise due to carcinogenesis. There is an ever-growing need for biologically relevant models of the breast, and normal mammary tissue presents a promising source of characterization for race associated differences. However, it is both challenging to acquire these samples as well as culture them in the lab. The conditional reprogramming (CR) culture developed by Schlegel et al provides a promising method to establish and expand normal mammary samples from diverse individuals.[53]

The era of big data and multi-omics technology provide immense potential for characterizing the biological basis of racial disparities in breast cancer. With high throughput RNA sequencing, epigenetic profiling, and genotyping technology rapidly improving in profiling capabilities and decreasing in cost, integrating these pieces of data can provide key insight into gene expression differences as well as potential mechanisms driving them.

In this study, we aimed to address the knowledge gap of uncovering the biological basis for racial disparities in breast cancer by characterizing transcriptomic differences between CR cultured normal mammary cells from AA and EA individuals. We established a biobank of normal mammary punch biopsy samples from diverse individuals in CR culture and profiled

these using single cell RNA-sequencing (scRNA-seq), DNA methylation bisulfite sequencing, and SNP genotyping. By performing a high throughput multi-omic analysis of CR normal mammary cells integrated with publicly available data and gene sets, we provide a transdisciplinary approach for better characterization of race associated differences in gene expression and their mechanistic drivers. We also present an *ex vivo* model of the normal breast for precision toxicology. To our knowledge, this is the first study integrating single cell RNA-seq, DNA methylation, and SNP genotyping of normal mammary cells from the same experimental dataset.

Materials and Methods

Komen Sample Acquisition

Cryopreserved normal mammary tissue samples were acquired from the Susan G. Komen normal tissue bank and immediately stored in liquid nitrogen upon arrival. All samples were nulliparous and either self-identified African American (n=17) or European American (n=32), matched on age, BMI, and days since last period. Samples were acquired with institutional review board approval (University of Michigan IRB approval number HUM00094966).

Ancestry Analysis: Ancestry data for each sample was obtained from the Komen Normal Tissue Bank who performed the genotyping and ancestry analysis. Genotyping was performed using the Illumina HumanHap650 Y Beadchip and ancestry proportions were quantified using the Ancestry Informative Marker Panel of SNPs developed by Nievergelt et al.[110]

Tissue dissociation and establishment in conditional reprogramming (CR) culture

Samples were processed by enzymatic and mechanical digestion following an adapted version of the protocol by Nakshatri et al (2015)[40]. For dissociation, samples were thawed (37°C, 5% CO_2) and poured over a 70 µm filter into a 15mL conical tube (FALCON) containing enzyme mixture (300 µl 10x Collagenase/hyaluronidase (STEMCELL technologies cat. no.07912) in 2.7mL F-media (Thong et al 2020)).[111] The tissue was gently scraped off the filter using a sterile spatula into a new conical tube, and the enzyme mixture remaining in the original tube was used to rinse the filter. The conical tube containing the sample in enzyme mixture was placed in a pre-heated shaking incubator (37°C, 135 RPM) for 1-2 hours depending on appearance of tissue sample (size, color). Samples were removed from the shaking incubator

and pipetted up and down 20x with a 1mL or 5mL pipet depending on size. This combination of incubation and pipetting was repeated every hour until the tissue chunk dissociated into single cells or stopped reducing in size. For samples that did not fully dissociate from incubation and pipetting, tissue chunks were chopped on a petri dish using two scalpels, pipetted 20x with a filtered glass pipet, and filtered again through a 70 µm filter. Once dissociated, samples were checked for single cell suspensions, cell counted and assessed for viability (LUNA-FL Dual Fluorescence cell counter), and centrifuged at 500g for 5 minutes. Samples were resuspended in 5mL of F-media, counted again, and plated in adherent 6 well plates containing pre-plated irradiated J2 fibroblasts. Samples which successfully grew multiple colonies were maintained and expanded according to the CR protocol outline in Thong et al (2020). [111]

Conditional reprogramming culture and sample collection for single cell RNA-sequencing

All of the samples which were successfully established in CR were grown simultaneously in culture to account for batch effects and split into two sample collection days and single cell RNA-seq (drop-seq) runs. Cryopreserved samples were thawed, counted, and 500,000 cells from each passage 1 sample were added to pre-plated irradiated J2s in a T-75 flask with F-media. The first day of plating included samples (KCR8519, KCR7889, KCR8565, KCR8483, KCR8523, KCR8580, KCR7554, KCR79530) and the remaining samples (KCR8195, KCR7518, KCR7195, KCR8474, KCR8617,KCR 8302, KCR8451, KCR8084) were plated the following day. Samples were monitored every day for confluence and media was changed every 3 days. Due to differences in growth rate, a number of samples (KCR8519, KCR8483, KCR7953, KCR7554, KCR8580, KCR7889, KCR8195) required passaging to prevent over confluence. In total, samples were in culture for 4-7 days depending on if they were collected in the first batch (KCR7518 (EA), KCR8302(AA), KCR7195(AA), KCR8519(AA), KCR8523 (EA), KCR8084 (EA), KCR7554 (EA), KCR7953 (EA)) or the second batch (KCR8195 (AA), KCR8483 (AA), KCR8580 (AA), KCR7889 (EA), KCR8474 (EA), KCR8451(AA), KCR8617(AA), KCR8565 (EA)) 3 days later.

For sample collection, differential trypsinization according to the CR protocol was performed to remove the J2 fibroblasts and 500,000 cells at a concentration of 100 cells/ µL from each sample were collected for drop-seq.

Single cell RNA-sequencing

Samples were transported on ice and processed for drop-seq analysis according to the protocol of Macosko et al (2015)[69] and Thong et al (2020).[111]

DNA extraction

For the remainder of cells not used for drop-seq, J2 fibroblasts were removed using the Mouse Cell Depletion Kit (Miltenyi cat. no. 130-104-694) and DNA was extracted using the DNeasy blood and tissue kit and protocol (QIAGEN cat. no. 69504). DNA quality was assessed by NanoDrop nucleic acid quantification (ThermoFisher) and frozen at -80°C for further experimentation

DNA sequencing

DNA samples stored at -80°C were thawed and NanoDrop quantification was used to dilute DNA to a concentration of 100ng/µL. 25µL from each sample was sent to the University of Michigan Advanced Genomics Core for genotyping. PicoGreen/Qubit analysis and normalization were performed to normalize samples and genotyping was performed using the Infinium Global Diversity Array-8 according to the manufacturer's instructions (Illumina) at the University of Michigan Advanced Genomics Core.

Epigenetic profiling

DNA samples stored at -80°C were thawed and sent to the University of Michigan Epigenomics Core for analysis using the EPIC Methylation Array (Illumina). DNA was quantified using the Qubit High Sensitivity dsDNA assay and quality was assessed using the TapeStation genomic DNA kit. 250ng of each sample were bisulfite converted using Zymo's EZ DNA Methylation kit according to the manufacturer's (Illumina) incubation parameters. Cleaned up samples were then sent to the UM Advanced Genomics Core for hybridization to the Infinium MethylationEPIC BeadChip array, washing, and scanning, according to the manufacturer's instructions (Illumina EPIC Datasheet).

Single cell data analysis

Raw data processing : Raw data was processed according to the protocol of Thong et al (2020).[111]

QC and high quality cell selection: The R package Seurat[112,113] was used for all of the initial QC and cell selection. Cells were filtered and selected if they met the QC criteria of less than 20% mitochondrial genes, gene counts between 200 and 6000, and RNA counts of less than 50,000. One of the samples KCR7889 was excluded from downstream analyses due to a sequencing issue where the Read1 of the Illumina sequencer read off the 5' mRNA sequence instead of the cell barcode sequence in the library. After QC and filtering, a total of 8 AA samples (n=8240 cells) and 7 EA samples (n=4577 cells) were available for further analysis. Once high quality cells were selected, they were normalized using the "LogNormalize" method in Seurat and highly variable features were identified. Data were then scaled by regressing on percent mitochondrial genes. Linear dimension reduction was run on the scaled data using variable genes to identify the top principle components (PCs).

Unbiased clustering and cell type identification :Unbiased clustering was performed using the top 10 PCs at a resolution of 0.05. Non-linear dimension reduction (UMAP) was also performed using the top 10 PCs. Cell type markers were used to identify myoepithelial (*KRT5/KRT14*) and luminal (*KRT8/KRT18*) cells which separated into two clear clusters at a resolution of 0.05.

Differential gene expression analysis: Cells were split by type (luminal and myoepithelial) and differential gene expression (DGE) analysis was performed using the Pseudobulk method according to the pipeline outlined by Amexquita et al.[114] For each cell type, pseudobulk samples were generated by aggregating across the Sample variable. Pseudobulk sample libraries with low counts (<100 cells) and genes with low counts were filtered out. Composition biases were corrected for by computing normalization factors with the trimmed mean of M-value. Multi-dimensional scaling, negative binomial dispersions, and quasi-likelihood dispersions were performed to calculate mean and per-gene variance. Generalized linear models (GLM) were fit to the counts for each gene and differential gene expression testing was performed between AA and EA cells for each cell type.

CytoTRACE: The Cellular (Cyto) Trajectory Reconstruction Analysis using gene Counts and Expression (CytoTRACE) package in R was used to predict the differentiation state of cells.[115] CytoTRACE leverages the number of detectably expressed genes per cell as a read out of developmental potential. Each individual cell was assigned a CytoTRACE score where cells

with a higher number of genes per cell were considered more transcriptionally diverse, which correlates with a more "stem cell" like gene expression profile.

Hybrid Analysis: Luminal/myoepithelial hybrid cells were identified by pulling the gene counts for marker genes *KRT5*/*KRT8* and *EPCAM/VIM* Cells in the upper 50% quantile for each gene were considered "highly expressed" and cells which highly co-expressed both *KRT5* and *KRT8* or *EPCAM* and *VIM* were labeled as hybrid cells.[111] The "FindMarkers" function in Seurat was used to perform differential expression testing using the Wilcoxon rank sum test to identify marker genes distinguishing the hybrid cells from all other cells in order to identify a hybrid cell signature.

Single cell Gene Set Enrichment Analysis: The escape package in R was used to perform single cell gene set enrichment analysis (GSEA),[116] assigning each individual cell an enrichment score for relevant gene sets. Gene sets were pulled from the C2 and Hallmark libraries from the Molecular Signatures Database (MsigDB)[117] and enrichment was performed using escape.

Cell Cycle Scoring and Regression : Cell cycle phase scores were generated for each cell using the Seurat cell cycle scoring pipeline.[118] Scores were then used to assign phases S or G2/M to each cell and stored in the metadata.

Epigenetics Data Analysis

QC and Filtering : Initial QC and filtering was performed by the University of Michigan Epigenomics Core using the Snakemake bioinformatics workflow.[119] Raw IDAT files were read into R using the minfi Bioconductor package[120] and the Enmix Bioconductor package[121] was used to perform initial qc based on detection p-values and signal intensity. Probes were removed if the detection p-value was <0.05 in more than 5% of samples (11841 probes), if a sample had more than 5% of probes with detection p-value <0.05 (0 samples), if a probe was within 2bp of a SNP (58337 probes),[122,123] or was located on the X or Y chromosome (19681 probes). After filtering a total of 776,977 probes were used for downstream analysis and were corrected, normalized, and had intensities background and dye-corrected using the NOOB background correction.[124]

Differential methylation was calculated by constructing a regression model using ebayes (limma) by race. The differentially methylated probes were annotated using the EPIC annotation

and merged with the UCSC reference genome to assign them to differentially expressed genes from the single cell analysis.

DNA Data Analysis

Imputation of SNPs via TopMed: We imputed polymorphisms genome-wide for each sample based on the GlobalDiversity SNP array results generated by the Advanced Genomics Core using the TOPMed Imputation Server.[125] Briefly the TOPMed Imputation Server performs an established quality control pipeline to validate variants, phases haplotypes via eagle2[126] , and imputes genotypes against the TOPMed reference panel (consisting of data from 97,256 individuals) using minimac4.[125]

Assigning genotypes to individuals for race associated SNPs of interest: For genes of interest identified through differential gene expression, we looked at single tissue specific eQTLs in GTEx to identify race-associated SNPs we could extract genotypes in our samples with. Using the dbSNP feature, we looked at the SNP report ALFA allele frequency and calculated the difference in allelic frequency between the European and African American populations. The SNP was selected if the difference in allelic frequency was > 10%. Using the SNP report we also noted the chromosome location. After loading in the TOPMed imputed tabix file for the relevant chromosome the SNP was located on, we used the SNP location from GTEx to identify the hard calls for each individual. Based on the reference and alternate allele in GTEx we assigned genotypes for each of our samples for the SNP. R2 >0.3 was used as the cutoff.

Results

Establishment of Komen Biopsy samples in Conditional Reprogramming Culture and Multi-omic Characterization

Cryopreserved punch biopsy samples from AA (n=17) and EA (n=32) women epidemiologically matched on age, BMI, and days since last period were dissociated and cultured in CR conditions. Of these, 8 AA and 8 EA samples were successfully established and expanded with enough cells and genetic material to use for single cell RNA-seq, DNA methylation, and SNP genotyping analysis (**Figure 3.1**). AA samples had a higher establishment rate (53.3%) compared to EA samples (26.7%) in CR culture conditions. SNP ancestry analysis from all acquired samples exhibit a dynamic range of ancestry percentages across both self-

identified race groups, indicating a diverse range of genetic phenotypes even within race groups (**Appendix Figure 3.A1).** H&E histology staining of available samples strongly suggest that establishment success in CR culture is highly dependent on epithelial and cellular composition of the original cryopreserved biopsy tissue, with samples containing a higher proportion of epithelial material more likely to be established versus those with a higher adipocyte content (**Appendix Figure 3.A2**).

Normal Mammary Cells From AA and EA Women Cluster by Cell Type, Cell Cycle Phase, and are Highly Variable by Individual

After undergoing QC and filtering, a total of 12,817 cells (n=8240 AA, n = 4577 EA) and 33,794 genes passed the selection criteria and were used for downstream single cell analysis (**Appendix Figure 3.A3**). Unbiased clustering at a resolution of 0.075 and analysis of mammary cell type markers revealed that cells primarily cluster by luminal (*KRT8*) and myoepithelial (*KRT5*) cell types and expression of epithelial/mesenchymal (*EPCAM/VIM*) and stem cell (*ITGA6/ALDH1A3*) markers was consistent with our findings in Chapter 2 **(Figure 3.2).** More specifically, the luminal cluster showed local sub-clustering into two smaller clusters with differences in *ALDH1A3* stem cell marker expression. Labeling by cell type, race, and individual revealed that cells from each race group were distributed throughout both luminal and myoepithelial clusters, with local sub-clustering highly influenced by gene expression of individual samples **(Fig 3.3 A-C)**. In addition to clustering by cell type, an analysis of gene expression by cell cycle phase revealed that cells cluster by phase in a dimensional direction consistent with the biological process progression –G1, S, to G2M **(Figure 3.3 D)**. Further analysis of cell type distributions across race groups and samples revealed that cell type distribution is highly variable by individual. The AA group was evenly split with 4 samples composed of a majority luminal cells and 4 samples primarily myoepithelial, whereas the EA group had 5 samples primarily luminal and 2 samples primarily myoepithelial in cell type distribution **(Figure 3.3 E-F)**.

Single Cell Gene Set Enrichment (ssGSEA) of Breast Cancer and Stem Cell Associated Gene Sets Reveals Highly Variable Inter-individual Differences Within AA and EA Race Comparisons

To investigate global gene signature differences between AA and EA mammary cells, enrichment scores were calculated for each individual cell using breast cancer and stem cell associated gene sets curated from MsigDB. Heatmap clustering revealed that enrichment scores were highly variable by individual, even within race groups, and that gene sets also grouped by biological function **(Appendix Figure 3.4A)**. Unsurprisingly, the luminal A and luminal B breast cancer gene sets (SMID_BREAST_CANCER_LUMINAL_A_UP and SMID_BREAST_CANCER_LUMINAL_B_UP) grouped with the luminal progenitor (LIM_MAMMARY_LUMINAL_PROGENITOR_UP) gene set. The basal breast cancer gene set (SMID_BREAST_CANCER_BASAL_UP) grouped with the embryonic stem cell (WONG_EMBRYONIC_STEM_CELL_CORE) and mammary stem cell (LIM_MAMMARY_STEM_CELL_UP) gene sets, and the ERBB2 breast cancer gene set (SMID_BREAST_CANCER_ERBB2_UP) grouped most closely with the mammary luminal mature (LIM_MAMMARY_LUMINAL_MATURE_DN) gene set. Enrichment scores across gene sets were also variable by luminal and myoepithelial cell types (**Appendix Figure 3.A4)**. Linear mixed effects modeling revealed no significant enrichment differences by race for any of the gene sets after accounting for inter-individual differences **(Appendix Table 3.A1)**.

Differential Gene Expression Reveals Transcriptomic Differences Between AA and EA Normal Mammary Cells, Including DEG in The Cancer Genome Atlas (TCGA) Tumors

To characterize transcriptomic differences between AA and EA mammary cells and identify specific genes driving them, differential gene expression was performed between AA and EA luminal and myoepithelial populations. After controlling for inter-individual differences, we identified 589 genes upregulated and 483 genes downregulated in AA vs EA ($p<0.05$) myoepithelial cells (**Figure 3.5 A**). In luminal cells, we identified 297 genes upregulated and 342 genes downregulated ($p<0.05$) in AA cells compared to EA (**Figure 3.5 B**). After adjusting for multiple comparisons, only 2 genes in the myoepithelial subset were identified as significantly (FDR<0.05) differentially expressed– *MTND4P24* (log2FC =6.14, p= 2.99E-09) and *RP11-673C5.1* (log2FC = 2.68, p = 7.29E-08). No genes in the luminal subset passed the FDR cutoff of 0.05 (**Appendix Table 3.A2**). Luminal and myoepithelial DEGs were compared with race associated DEGs identified as differentially expressed in TCGA tumors by Huo et al.[127] Of these, 8 genes in luminal cells and 13 genes in myoepithelial cells were differentially expressed

between AA and EA in both our normal mammary samples and TCGA tumors (**Table 3.1**). *CRYBB2* stood out as a particular gene of interest due to its differential expression in both the luminal and myoepithelial subtypes as well as being differentially upregulated in AA TCGA tumors for every breast cancer subtype.[127]

CytoTRACE Prediction of Cellular Differentiation Potential Reveals Differences By Cell Type

Differentiation potential is a key feature of stem cells and cancer, therefore we assessed this in AA and EA cells across luminal and myoepithelial subtypes using the Cellular (Cyto) Trajectory Reconstruction Analysis using gene Counts and Expression (CytoTRACE) computational method.[115] CytoTRACE predicts the differentiation state of cells by leveraging transcriptional diversity, or the number of genes per cell, as a measure of developmental potential. Average gene counts per cell showed a range across individual samples as well as within cells from the same individual (**Appendix Figure 3.A5**). AA samples showed on average slightly higher gene counts compared to their EA counterparts (**Appendix Figure 3.A5**). CytoTRACE scores across clusters showed the highest CytoTRACE scores in clusters which corresponded to the G1 and G2 phases identified in the previous cell cycle analysis (**Figure 3.6A, Figure 3.3D**). Higher CytoTRACE scores indicate higher transcriptional activity and developmental potential. Comparisons of CytoTRACE scores between luminal and myoepithelial cells from all individuals showed that myoepithelial cells on average express a higher CytoTRACE score than luminal cells (Wilcoxon $p<0.05$) (**Figure 3.6B)**. CytoTRACE scores vary highly by individual in both luminal an myoepithelial cell types, and after adjusting for individual differences there was no significant difference in CytoTRACE score by race (**Figure 3.6C-D)**.

Hybrid Stem Cell Populations Exhibit a Stem-like Developmentally Immature Phenotype and are Associated with Poor Cancer Outcomes

Chapter 2 showed the emergence of hybrid stem cell populations following conditional reprogramming, and we aimed to characterize these populations in this dataset across more genetically diverse individuals. Here we used the marker gene combinations of *KRT8/ KRT5* to identify luminal/basal (used interchangeably here with myoepithelial)(L/B) and EPCAM/VIM to identify epithelial/mesenchymal (E/M) hybrids. We found that L/B hybrids were mostly

localized to the luminal cluster (**Figure 3.7A)** and that E/M hybrids were present in both luminal and myoepithelial clusters (**Figure 3.7B**). This localization of hybrids was consistent with our findings from Chapter 2. Hybrid L/B cells were identified in both AA and EA groups with proportions varying greatly by individual and numerous individuals in both groups exhibiting very few of the L/B hybrids (**Figure 3.7C**). Hybrid E/M cells were also identified in all individuals, with higher overall proportions in both AA and EA groups (**Figure 3.7D**). Proportions of hybrid E/M were variable by individual but with less pronounced differences than for L/B hybrids. CytoTRACE scoring of both L/B and E/M hybrids revealed that hybrid cells exhibit the most developmentally immature phenotype compared to their counterparts (**Figure 3.7E-F**).

The developmentally immature phenotype of hybrid cells has previously been linked to cancer stem cells and aggressiveness of tumors, therefore we used ssGSEA of stem cell and metastasis associated gene sets to further characterize this. L/B and E/M hybrid cells exhibit distinct enrichment signatures from each other, but showed similarities overall (**Figure 3.8A-B**) . In particular, both L/B and E/M hybrid cells exhibit enrichment for the ALONSO_METASTASIS_EMT_UP gene set and a small subset of E/M hybrids exhibit a particularly strong signal for the VANTVEER_BREAST_CANCER_METASTASIS_UP gene set.

Differential Gene Expression of Hybrid Stem Cells Reveals Transcriptomic Differences between AA and EA Mammary Cells, Including DEGs in TCGA tumors

To quantify and further characterize gene expression signatures of L/B and E/M hybrid cells and identify race-associated differences, we performed differential gene expression analysis between AA and EA hybrid populations. We were able to identify both significantly upregulated (n=40, adj. $p<0.05$) and downregulated (n=64, $p<0.05$) DEGs between AA and EA L/B hybrids (**Figure 3.9A**). E/M hybrids exhibited greater transcriptomic differences by race, characterized by a larger number of DEGs upregulated (n=337, adj. $p<0.05$) and downregulated (n=494, adj. $p<0.05$) in AA hybrids compared to EA (**Figure 3.9B**). Race associated DEGs of hybrid cells were overlapped with DEGs from TCGA tumors, resulting in 5 overlapping genes shared by L/B hybrids and TCGA tumors and 11 overlapping genes for E/M hybrids. Of these, *CRYBB2* stands out due to its upregulated in AA cells for both hybrid types and our previous identification of

upregulation in both luminal and myoepithelial AA cells. Other genes of interest include *TINGAGL1*, *DEFB1,* and *SNHG5* which are downregulated in both hybrid cell types as well as in TCGA tumors

Differentially Methylated CpG Sites Between AA and EA Mammary Cells Correspond to DEGs in Both Luminal and Myoepithelial Cell Types

DNA methylation was analyzed for AA and EA samples in order to gain mechanistic insight into potential upstream causes of differential gene expression characterized earlier in the study. A total of 776,187 CpG sites were successfully annotated to the EPIC reference and fit to a regression model using ebayes (limma). Of these, 20,694 CpGs were differentially methylated ($p<0.05$), however, none were significant after adjusting for multiple comparisons (FDR<0.05). The CpG annotation was used to identify the genes they were located on and after filtering by p value (<0.05) and percent methylation change (>10%), 88 CpG sites were identified to be differentially methylated between AA and EA luminal cells on luminal DEGs (**Figure 3.11A**), and 50 CpGs on myoepithelial DEGs (**Figure 3.11B**). When assessing DNA methylation alongside RNA expression, *PRDM16* stood out as a DEG of interest (log2FC= -3.1, p =0.03) in the luminal population due to it having 31 differentially methylated CpGs ($p<0.05$) in AA compared to EA samples. CpGs for *PRDM16* ranged from -10 to -22.6% difference in methylation and 8 were located on islands, 15 on shores, and 8 on shelves. (**Appendix Table 3.A3**). *ANKR7L* was a DEG of interest in the myoepithelial population (log2FC=2.73, p=0.01) containing 6 CpGs ranging from 15.8-29.9% differential methylation and all of which were located on an island. Lastly, *GSTM1* stood out as a DEG of interest due to its upregulation in myoepithelial (log2FC =3.58, p=0.014) and luminal (log2FC=2.75, p=0.17) cells as well as TCGA tumors. *GSTM1* contained one significant CpG with -10.51% differential methylation in myoepithelial cells which was located on an island.

Identification of Race Associated eQTLs on DEGs Between AA and EA Mammary Cells Show Genotypic Differences by Race Consistent with Gene Expression in Normal Mammary Tissue (GTEx) and TCGA Tumors

A number of the most significant DEGs between AA and EA showed no significant differences in CpG methylation, which led us to investigate whether these expression differences could be accounted for by differences in eQTLs in our samples with imputed genotypes.

CRYBB2 was the first gene we investigated due to its upregulation in both luminal (log2FC = 5.58, p=0.013) and myoepithelial (log2FC=3.94, p=0.036) AA cells as well as in TCGA tumors (log2FC=1.23, p= 3.02E-32). We selected two SNPs and identified genotypes of our samples at these loci due to their differences in allelic frequency by race (rs5996939: 44%, rs6519611: 29%)(**Figure 3.12A**). For rs5996939, the A allele was associated with higher expression in GTEx mammary tissue, and 6/7 AA samples had 1 or more A allele at this locus compared to 1/6 for EA samples (**Figure 3.12B-C).** Similarly, the G allele for rs6519611 was associated with higher expression, and was present in 6/7 AA and 1/6 EA samples. In TCGA tumors, CRYBB2 is significantly upregulated in AA tumors as well as in the TNBC tumor subtype (**Figure 3.12D-E**).

We also investigated eQTLs on *SNHG5,* which was significantly downregulated in AA cells for both hybrid cell types (L/B: log2FC= -0.9, p = 9.06E-22, E/M: log2FC =-0.96, p = 2.94E-47) as well as TCGA tumors (log2FC=-0.62, p=3.32E-7). We looked at the SNPs rs1059307 and rs9450311 which had allelic frequency differences of 28% and 31%, respectively (**Figure 3.13A**). The G allele was associated with lower expression in GTEx mammary tissue and 7/7 AA samples had a G allele, with 5/7 being GG for this loci (Figure 3.13B-C). Only 1 EA sample expressed the GG genotype for this eQTL and 4/6 were GT. In TCGA, AA tumors significantly expressed lower levels of SNHG5 compared to EA counterparts, and TNBC tumors also significantly expressed lower levels compared to normal (Figure 3.13 D-E).

Lastly, we also looked at two SNPs in *CLN8,* a gene downregulated in AA luminal cells (log2FC= -0.54, p=3.41E-.2) and TCGA tumors (log2FC=-0.36, p=6.21E-7) with allelic frequency difference of 41% for rs6558535 and 13% for rs6558532 (**Figure 3.14A**). For rs6558535 the T allele associated with lower in expression in GTEx mammary tissue was present in 5/7 AA samples and 3/6 EA samples. The C allele for rs6558532 was present in all AA and EA samples, with 4/7AA samples and 1/6 EA samples exhibiting the CC genotype.

Overall, we found this integration of DNA SNPs, DEGs from RNA-seq, and TCGA tumor data informative and promising to identify race associated eQTLs.

Discussion

In this chapter we leverage a multi-omic systems biology approach to characterize transcriptomic differences and potential mechanisms driving these differences between mammary cells from genetically diverse AA and EA donors. The integration of high throughput and computationally rigorous tools allowed us to characterize transcriptomic signatures of luminal, myoepithelial, and hybrid cell types at single cell resolution, gain insight into DNA methylation and/or genotype differences potentially contributing to differences in gene expression, and compare normal mammary stem cells with TCGA tumor expression signatures. From this, we gleaned two major conclusions 1) there are no strikingly clear race associated differences in transcriptomic expression and DNA methylation between AA and EA mammary cells that can be deconvoluted from inter-individual and cell type heterogeneity 2) while DNA methylation and genotyping analysis can give us potential mechanistic insight into a handful of genes of interest we still have a limited understanding of the underlying causes of the differential expression we characterized.

While the inter-individual heterogeneity of our samples made it challenging to distinguish differences in gene expression and DNA methylation by race, what it does confirm is that the CR culture is a viable model system to study mammary cells from genetically diverse individuals as it retains inter-individual differences even after reprogramming. The ancestry data (**Appendix 3.A1)** provided to us by the Komen normal tissue bank showcases the genetic diversity of all our individual donors, even within self-identified AA and EA race groups. This genetic diversity across our samples was likely the reason that differentially expressed genes and CpG sites were not significant after adjusting for multiple comparisons due to our sample size of 8 AA and 7 EA individuals being underpowered to detect differences. Future studies may need to include greater sample size to account for this diversity or a more nuanced integration of ancestry/genotype information rather than the broad AA and EA race groups. Using the R package PROPER (PROspective Power Evaluation for RNAseq)[128], we ran a power simulation analysis to estimate the minimum sample size needed in order to detect significant differences using the number of genes and transcript counts observed in our dataset as the simulation inputs. Based on the simulation, the minimum number of samples required in each group is estimated to be 20-25 to achieve a target power of 0.8 and FDR <0.05 (**Appendix 3.A6**). Due to this, our

sample size of 8 AA and 7 EA was likely underpowered (~0.6) to detect differences between such diverse individuals.

Other factors contributing to the inter-individual heterogeneity of our samples include the sample collection process as well as the expansion in CR conditions. Samples from the Komen Normal Tissue Bank are collected at different sites across the country and differences in sample collection and personnel may introduce heterogeneity in the collected samples. Additionally, the punch biopsy samples are so small that even samples taken from different locations on the same breast may capture different distributions of cells and cell types. Thus, the cellular composition of the original sample highly influences the success of establishment and expansion in CR and may introduce heterogeneity in cell type distribution if cells take hold in CR conditions differently. Additionally, as mentioned in Chapter 2, the CR process appears to enhance inter-individual heterogeneity post-CR compared to NM cells.

One strength of this study was the way single cell RNA-seq was leveraged sequentially, beginning with cell type identification upstream and stratified differential gene expression leading to a more targeted downstream analysis of DNA methylation and eQTL assessment. The upstream identification of luminal and myoepithelial cells was crucial to the stratified differential gene expression analysis we performed which allowed us to identify distinct cell type specific transcriptomic differences between AA and EA mammary cells.

While there were some overlapping genes differentially expressed in both cell types such as *CRYBB2* and *GSTM1*, stratified analysis by cell type allowed us to identify genes like *CLN8,* which was significantly downregulated in luminal but not myoepithelial AA cells as well as in TCGA tumors. *CLN8* has been most characterized as a mutated gene in neuronal ceroid lipofuscinoses (NCLs), a rare family of neurodegenerative disorders. Emerging studies have identified genetic alterations in NCL genes across multiple cancer types.[129] In brain cancers, *CLN8* has been showed to act as a tumor suppressor, with lower expression linked to worse cancer outcomes.[129] More specifically in the breast, *CLN8* has also been included as part of a gene set used as a strong prognostic factor for the development of breast cancer metastasis and has also been identified to be located on a domain of *CERS2,* known to be alternatively spliced in Luminal B cancers.[130,131] Kaplan-Meier survival analysis of the breast also shows low expression of *CLN8* associated with lower chance of survival. Together, alongside our identification of race

associated eQTLs consistent with the downregulation in expression, *CLN8* emerges as a prime candidate for further characterization in normal tissue and breast tumors.

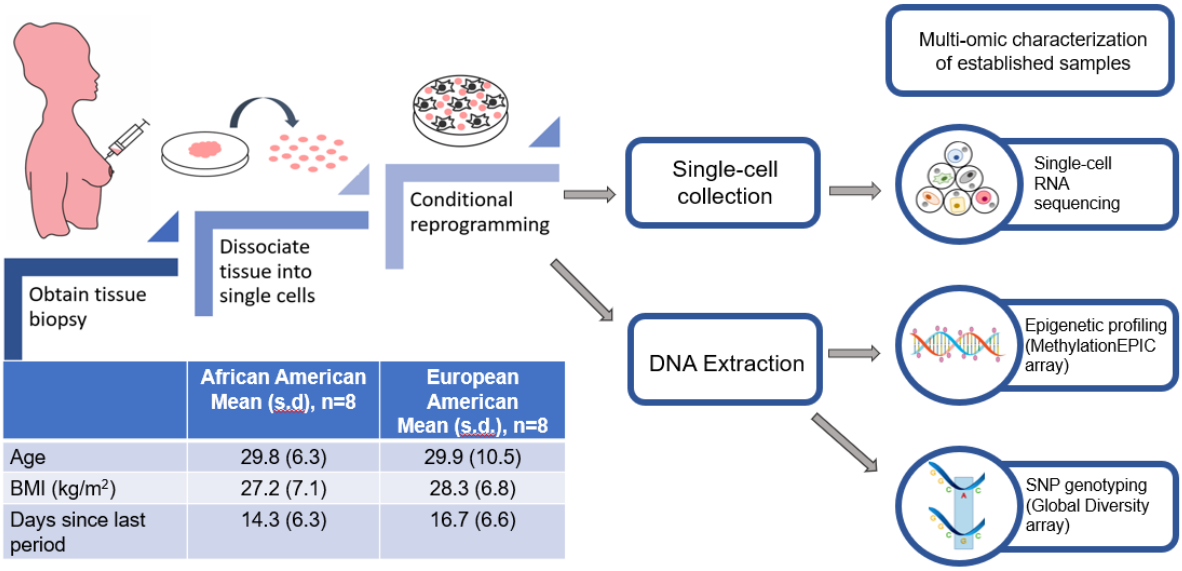
Conversely, *CRYBB2* has been consistently identified as differentially upregulated in AA vs EA normal mammary tissue and tumors, although its mechanism of action and identification as a race specific gene are still being elucidated.[132–135] In addition to *CRYBB2*, its pseudogene *CRYBB2P1* is also garnering increasing interest for its potential function as a regulatory non-coding RNA of *CRYBB2*. Barrow et. al report that *CRYBB2* and *CRYBB2P1* independently promote tumor growth likely through distinct mechanisms.[134] We report significant overexpression of *CRYBB2* and *CRYBB2P1* in both luminal, myoepithelial cell types as well as identify race specific *CRYBB2* eQTLs reflected in the corresponding genotypes of our samples. While there were no significant differentially methylated CpGs for *CRYBB2*, *CRYBB2P1* had one significant CpG (cg04572826, % methylation difference = 20%, p=0.006) which has been identified to be located on a promoter regulatory element. Together, these findings add to the existing body of evidence that *CRYBB2* and *CRYBB2P1* are race specific genes linked to breast cancer and breast cancer outcomes. To our knowledge, this is the first parallel integration of gene expression, eQTL genotype imputation, and DNA methylation analysis from the same experiment and individuals providing mechanistic evidence to account for *CRYBB2* and *CRYBB2P1* gene expression differences by race.

As seen in Chapter 2, hybrid stem cells are of particular interest in relation to stemness and breast cancer, and we were able to further characterize L/B and E/M hybrid population across more individuals in this chapter. Hybrid L/B and E/M exhibit distinct transcriptomic signatures from each other while simultaneously exhibiting an enriched stem-like, developmentally immature phenotype compared to their non-hybrid counterparts. GSEA allowed us to characterize these global signatures as enriched in gene sets related to stem cells, metastasis, EMT, and poor prognosis. Differential gene expression between AA and EA hybrid cells also highlighted a few additional genes interest, *TINAGL1*, *DEFB1*, *SNHG5*, and *TMEM 100*—all of which were significantly downregulated in both L/B and E/M hybrids as well as TCGA tumors. *TMEM100* expression has been linked to inhibit and suppress gastric, lung, and colorectal cancer.[136–138] Similarly, DEFB1 downregulation has been identified in renal, prostate, and colorectal cancers.[139–141]. *SNHG5* has been previously identified as a promising diagnostic

and therapeutic target for multiple cancer types, however most of these findings report upregulation of *SNHG5* promoting cancer survival and progression.[142–144] This provides the opportunity for future work to characterize the functional implications of downregulation of *SNHG5*. Lastly, *TINAGL1* presents an especially interesting finding due to its differential expression in AA vs EA hybrid cells, but not overall in luminal and myoepithelial cells. More specifically, *TINAGL1* has been found to suppress triple negative breast cancer progression and metastasis *in vitro* and *in vivo* and has also been successfully targeted for gene therapy to suppress TNBC growth.[145,146] Due to the higher prevalence of TNBC among AA individuals, *TINAGL1* is a promising target to study further in the characterization of breast cancer disparities as well as overall breast cancer outcomes.

Overall, this multi-omics systems biology approach to characterizing transcriptomic differences and mechanisms in AA and EA mammary cells showcases an integrated and targeted pipeline for identifying differentially expressed genes and potential mechanisms with functional implications for breast cancer and breast cancer outcomes. These results also suggest that genetic ancestry analysis rather than self-identified race may be a better predictor of breast carcinogenesis and outcomes. The highly variable inter-individual differences in gene expression, DNA methylation, and eQTL genotypes we report highlight the pressing need to use and develop better model systems which are able to capture the genetic variation of diverse individuals.

Tables and Figures



	African American Mean (s.d), n=8	European American Mean (s.d.), n=8
Age	29.8 (6.3)	29.9 (10.5)
BMI (kg/m^2)	27.2 (7.1)	28.3 (6.8)
Days since last period	14.3 (6.3)	16.7 (6.6)

Figure 3.1. Experimental Design and Sample Demographics. Conditional reprogramming of AA and EA matched normal mammary samples and multi-omic characterization of differences.

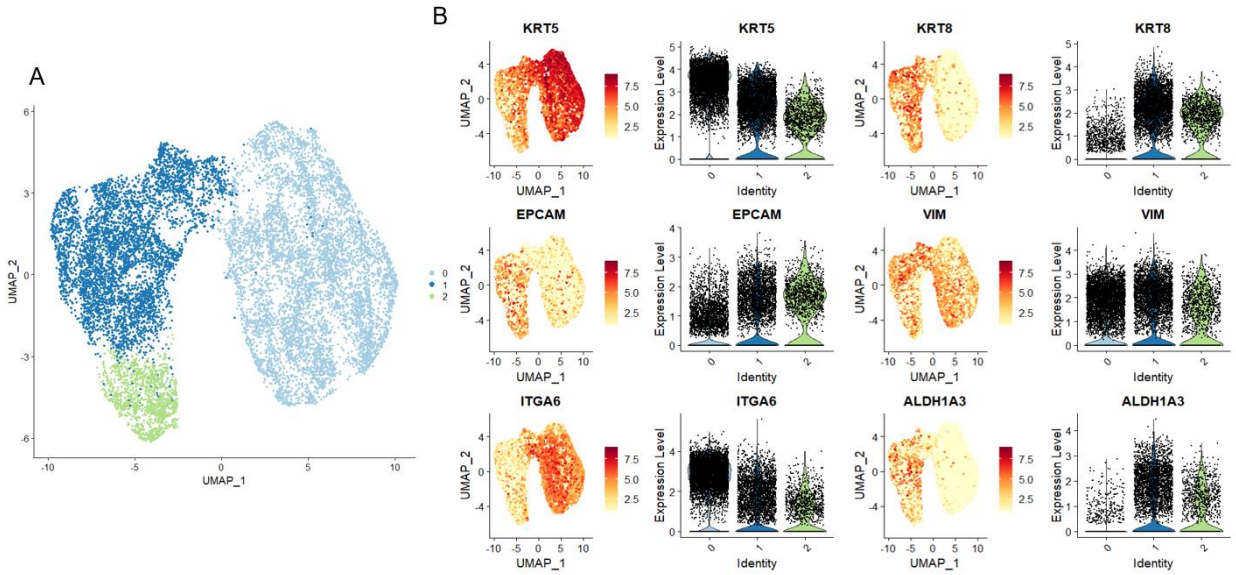
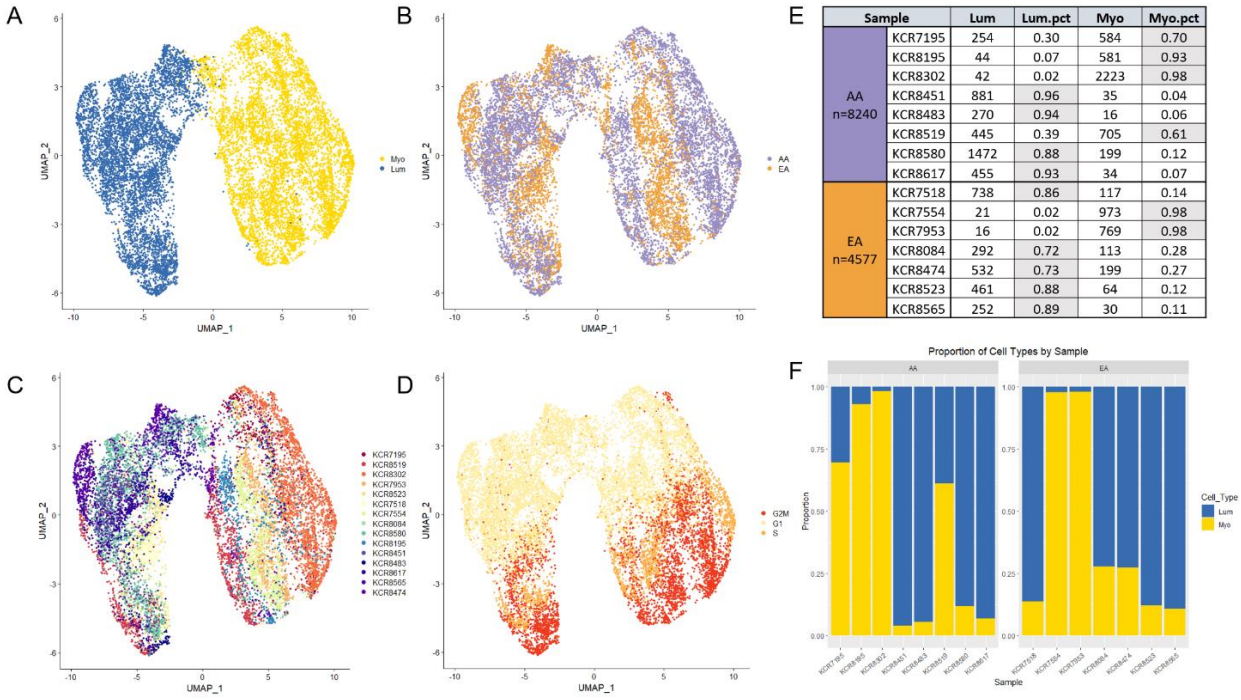


Figure 3.2. Unbiased clustering and marker gene expression A) UMAP clustering of cells at 0.075 resolution and B) Cell marker expression across clusters.



Sample		Lum	Lum.pct	Myo	Myo.pct
AA n=8240	KCR7195	254	0.30	584	0.70
	KCR8195	44	0.07	581	0.93
	KCR8302	42	0.02	2223	0.98
	KCR8451	881	0.96	35	0.04
	KCR8483	270	0.94	16	0.06
	KCR8519	445	0.39	705	0.61
	KCR8580	1472	0.88	199	0.12
	KCR8617	455	0.93	34	0.07
EA n=4577	KCR7518	738	0.86	117	0.14
	KCR7554	21	0.02	973	0.98
	KCR7953	16	0.02	769	0.98
	KCR8084	292	0.72	113	0.28
	KCR8474	532	0.73	199	0.27
	KCR8523	461	0.88	64	0.12
	KCR8565	252	0.89	30	0.11

Figure 3.3. Clustering and cell type distributions. A) Labeled by cell type from *KRT5/KRT8* gene expression B) Labeled by race C) Labeled by sample ID D) Labeled by cell cycle phase E) Cell counts stratified by sample and cell type. Grey cells indicate the cell type with the majority proportion. F) Cell type proportions by individual

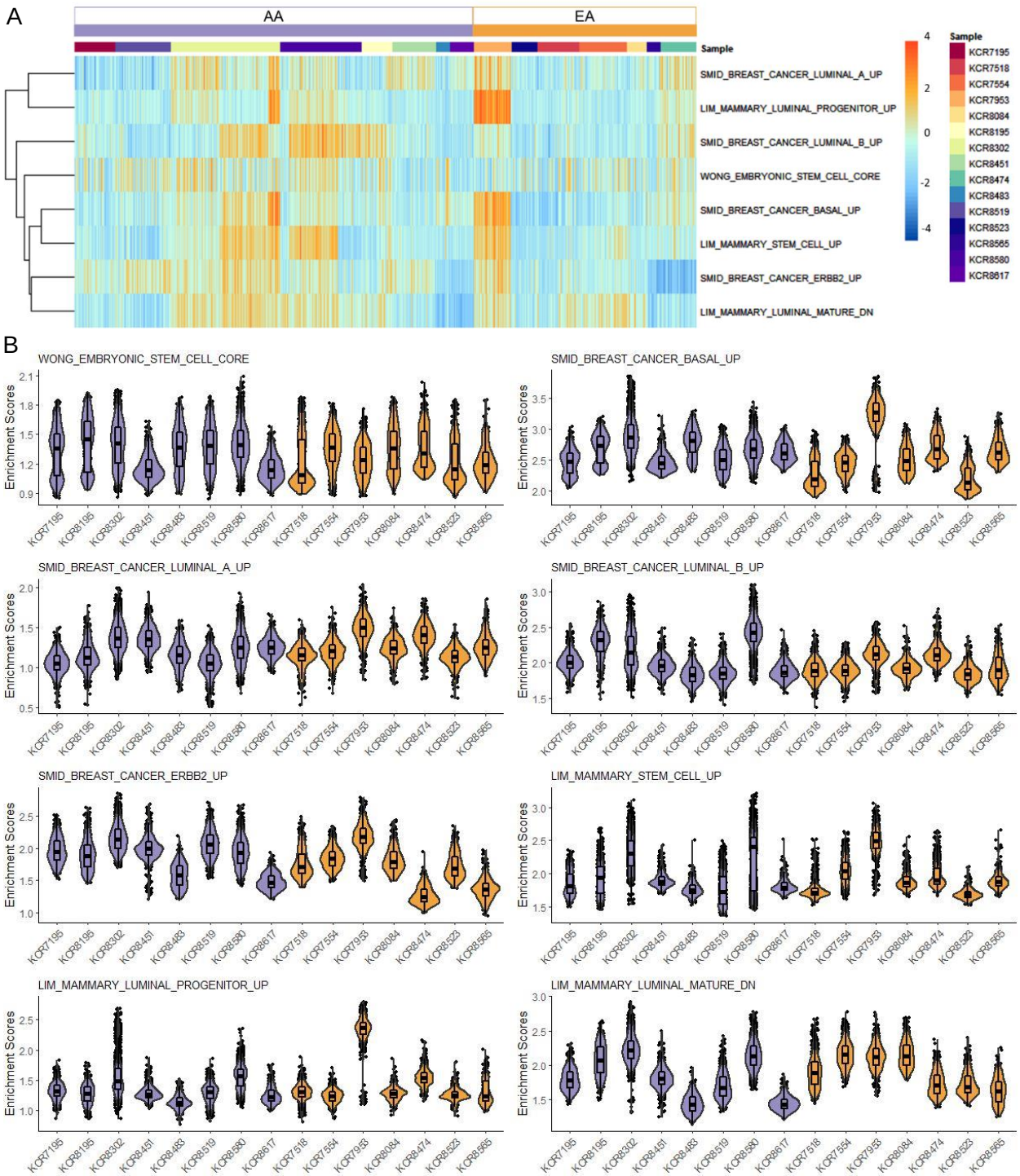


Figure 3.4. ssGSEA of breast cancer and stem cell associated gene sets A) Heatmap of enrichment scores by sample stratified by race where each vertical column represents an individual cell. B) Violin plots of enrichment scores for individual gene sets across samples. Purple indicates AA individuals and orange indicates EA individuals.

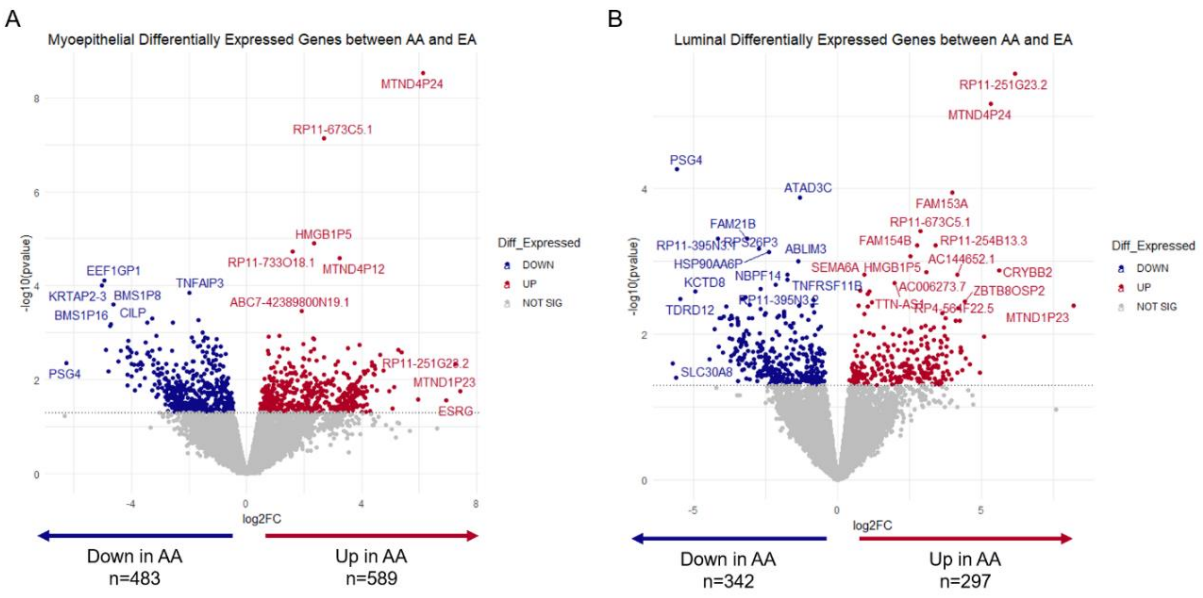


Figure 3.5. Differentially expressed genes between AA and EA mammary cells by cell type A) DEGs in myoepithelial cells. Genes shown in red are significantly upregulated ($p < 0.05$) and genes shown in blue are significantly downregulated ($p < 0.05$) in AA cells compared to EA. B) DEGs in luminal cells.

Table 3.1 Overlapping DEGs between AA and EA normal mammary cells and TCGA tumors stratified by cell type. Genes highlighted in red were upregulated in AA compared to EA and blue were downregulated.

	Myo Komen DEGs			TCGA DEGs			
Gene	Log2FC	PValue	FDR	Log2FC	Pvalue	FDR	Adj Pvalue
LRRC37A	4.10	1.86E-02	8.99E-01	-0.46	1.29E-10	1.28E-07	2.43E-06
GSTM1	3.58	1.38E-02	8.99E-01	2.90	3.80E-10	3.40E-07	7.15E-06
CRYBB2	2.40	3.63E-02	8.99E-01	1.23	3.02E-32	2.84E-28	5.67E-28
MAB21L1	1.54	1.14E-02	8.99E-01	-0.65	1.24E-06	2.09E-04	2.34E-02
FAM3A	0.63	2.34E-02	8.99E-01	0.38	4.20E-11	4.64E-08	7.89E-07
RPL23AP7	-1.10	4.89E-02	8.99E-01	-0.37	1.71E-07	5.03E-05	3.22E-03
LRRC37A2	-1.13	4.23E-02	8.99E-01	-0.96	2.42E-22	1.51E-18	4.54E-18
NBPF15	-1.25	4.68E-03	8.99E-01	-0.71	5.49E-08	2.02E-05	1.03E-03
SNHG5	-1.25	3.76E-02	8.99E-01	-0.62	3.32E-07	7.62E-05	6.25E-03
PTTG3P	-2.15	1.71E-02	8.99E-01	0.41	1.08E-06	1.92E-04	2.04E-02
AMPH	-2.26	4.75E-02	8.99E-01	-0.87	1.62E-06	2.52E-04	3.05E-02
RPS26P11	-2.45	1.55E-02	8.99E-01	0.39	3.38E-09	1.87E-06	6.35E-05
NBPF14	-2.60	2.46E-03	8.99E-01	-0.54	2.23E-06	3.07E-04	4.20E-02

	Lum Komen DEGs			TCGA DEGs			
Gene	Log2FC	PValue	FDR	Log2FC	Pvalue	FDR	Adj Pvalue
CRYBB2	5.58	1.32E-03	1.00E+00	1.23	3.02E-32	2.84E-28	5.67E-28
GSTM1	2.75	1.67E-02	1.00E+00	2.90	3.80E-10	3.40E-07	7.15E-06
LRRC37A	2.59	1.72E-02	1.00E+00	-0.46	1.29E-10	1.28E-07	2.43E-06
POLR2J3	1.93	1.41E-02	1.00E+00	0.50	2.14E-08	9.16E-06	4.03E-04
CLN8	-0.54	3.41E-02	1.00E+00	-0.36	6.21E-07	1.25E-04	1.17E-02
EP400NL	-0.57	5.00E-02	1.00E+00	-0.44	5.33E-10	4.36E-07	1.00E-05
NBPF14	-1.75	1.51E-03	1.00E+00	-0.54	2.23E-06	3.07E-04	4.20E-02
GSTT2	-3.22	3.08E-03	1.00E+00	0.96	2.38E-07	6.15E-05	4.48E-03

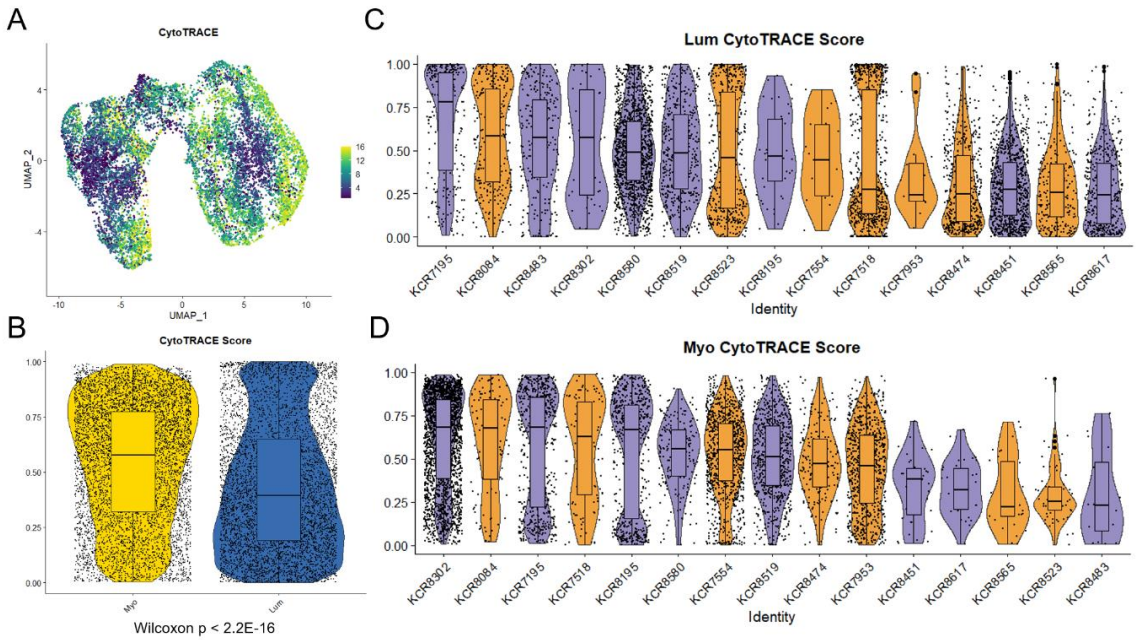


Figure 3.6. CytoTRACE scoring to assess transcriptional diversity across cell types and individuals A) FeaturePlot of CytoTRACE scores across clusters B) Differences in CytoTRACE scores by cell type. The higher the score the more transcriptionally active. C) Luminal CytoTRACE scores by AA and EA individuals D) Luminal CytoTRACE scores by AA and EA individuals

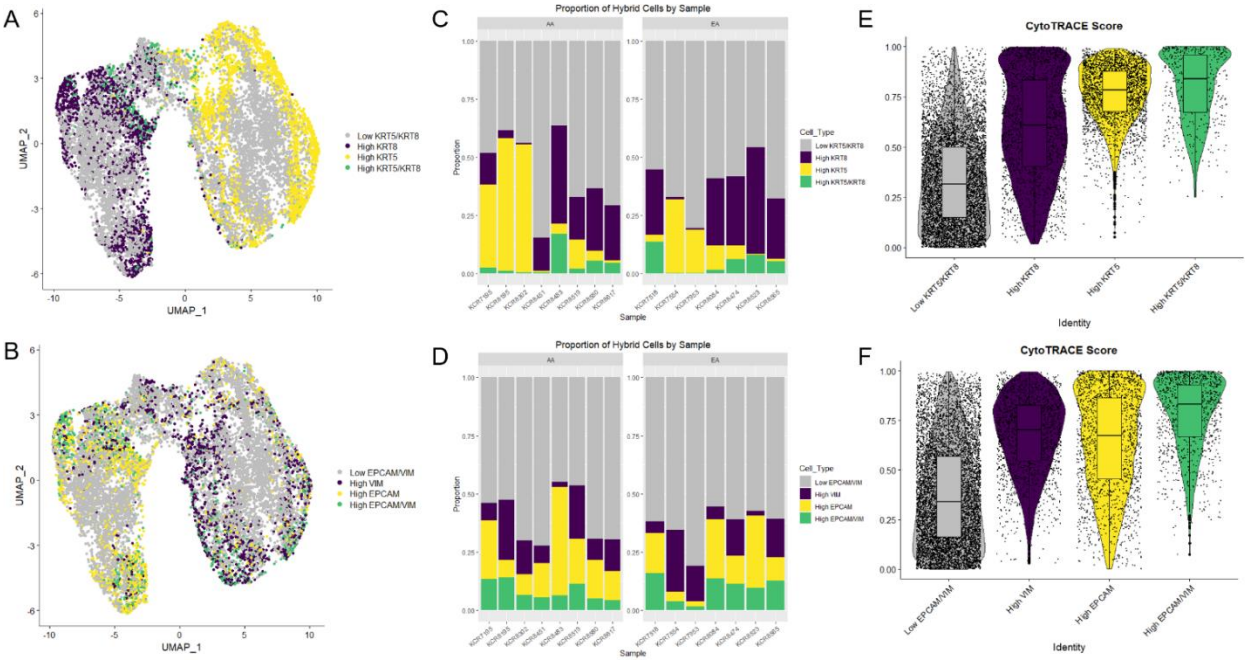


Figure 3.7. Hybrid KRT5/KRT8 and EPCAM/VIM cells A-B) Across clusters C-D) Individual sample proportions of hybrid cells stratified by race E-F) CytoTRACE scoring

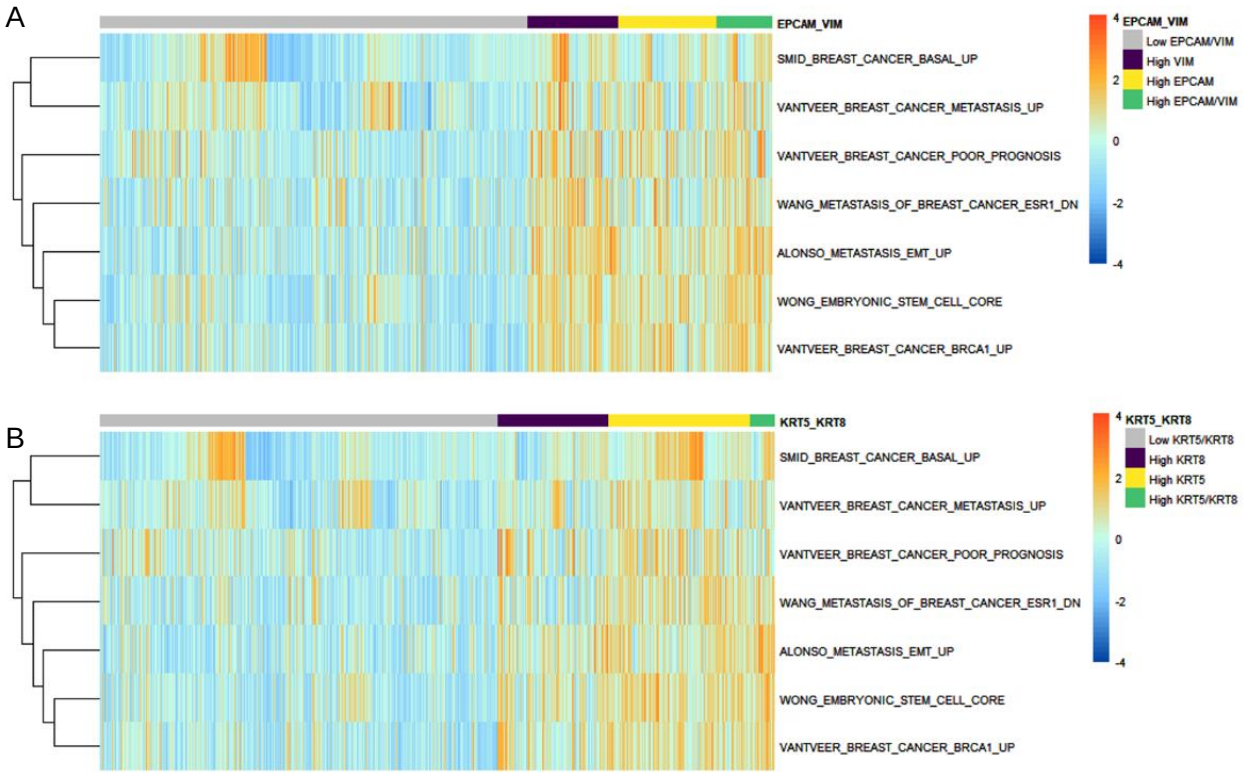


Figure 3.8. ssGSEA of stem cell and metastasis associated gene sets in hybrid cells A) Heatmap of enrichment scores for EPCAM/VIM hybrid cells B) Heatmap of enrichment scores for KRT5/KRT8 hybrid cells

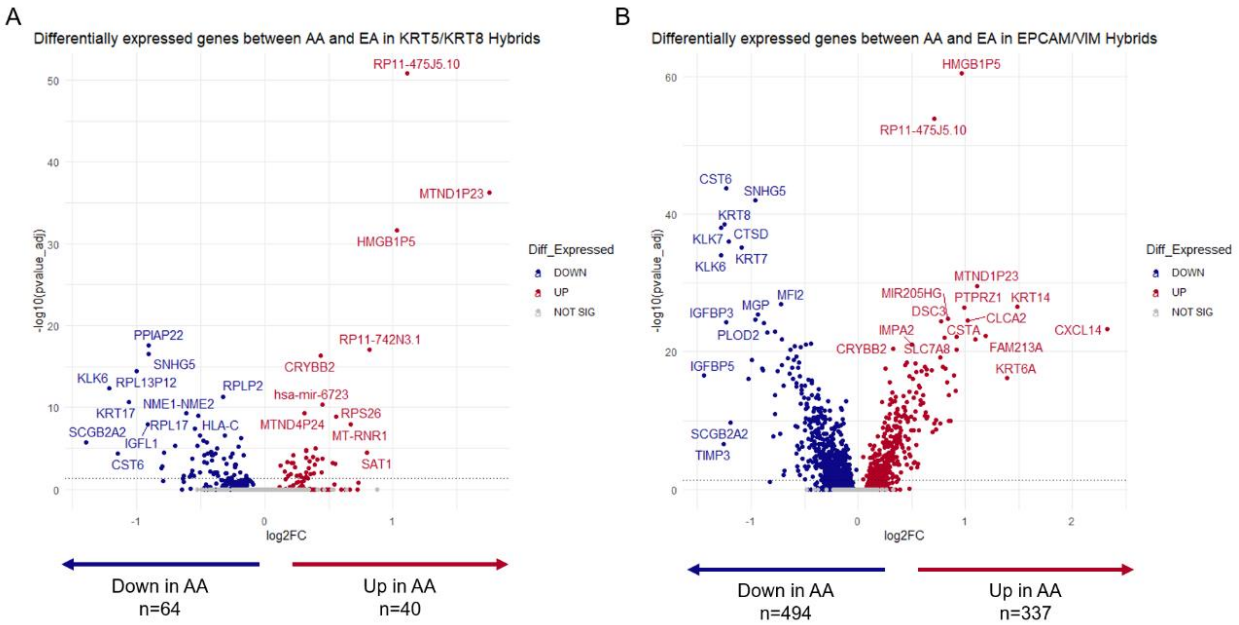


Figure 3.9. Differentially expressed genes between AA and EA hybrid cells A) DEGs in KRT5/KRT8 hybrids. Genes shown in red are significantly upregulated (adj. p <0.05) and genes shown in blue are significantly downregulated (adj. p<0.05) in AA cells compared to EA B) DEGs in EPCAM/VIM hybrids

Table 3.2 Overlapping DEGs between AA and EA hybrid cells and TCGA tumors. Genes highlighted in red were upregulated in AA and blue were downregulated

	Hybrid EPCAM/VIM DEGs			TCGA DEGs			
Gene	Log2FC	Pvalue	Adj. Pvalue	Log2FC	Pvalue	FDR	Adj. Pvalue
RPS26	0.48	1.08E-21	3.65E-17	0.45	2.13E-08	9.16E-06	4.00E-04
CRYBB2	0.33	1.02E-25	3.44E-21	1.23	3.02E-32	2.84E-28	5.67E-28
PLA2G4A	0.23	3.32E-14	1.12E-09	-0.69	4.78E-07	1.02E-04	8.99E-03
ANKRD29	0.14	7.94E-10	2.68E-05	-0.75	3.70E-07	8.28E-05	6.95E-03
DFNA5	0.12	1.72E-08	5.82E-04	-0.43	1.22E-06	2.07E-04	2.30E-02
NBPF15	-0.14	6.63E-07	2.24E-02	-0.71	5.49E-08	2.02E-05	1.03E-03
TMEM100	-0.15	2.34E-12	7.90E-08	-0.87	2.11E-06	2.92E-04	3.97E-02
NBPF14	-0.32	9.45E-17	3.19E-12	-0.54	2.23E-06	3.07E-04	4.20E-02
TINAGL1	-0.35	6.03E-10	2.04E-05	-0.64	6.08E-08	2.16E-05	1.14E-03
DEFB1	-0.48	1.35E-17	4.56E-13	-1.09	1.50E-06	2.46E-04	2.83E-02
SNHG5	-0.96	2.94E-47	9.93E-43	-0.62	3.32E-07	7.62E-05	6.25E-03

	Hybrid KRT5/KRT8 DEGs			TCGA DEGs			
Gene	Log2FC	Pvalue	Adj. Pvalue	Log2FC	Pvalue	FDR	Adj. Pvalue
RPS26	0.56	3.43E-14	1.16E-09	0.45	2.13E-08	9.16E-06	4.00E-04
CRYBB2	0.43	1.23E-21	4.16E-17	1.23	3.02E-32	2.84E-28	5.67E-28
TINAGL1	-0.42	7.86E-10	2.65E-05	-0.64	6.08E-08	2.16E-05	1.14E-03
DEFB1	-0.64	1.28E-06	4.34E-02	-1.09	1.50E-06	2.46E-04	2.83E-02
SNHG5	-0.90	9.06E-22	3.06E-17	-0.62	3.32E-07	7.62E-05	6.25E-03

A

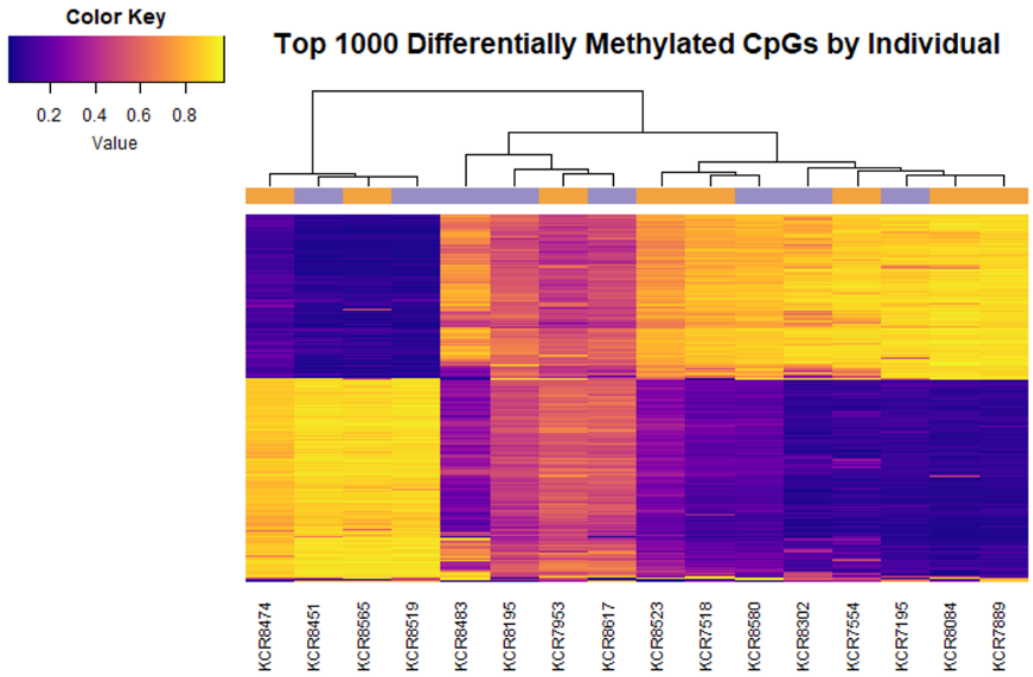


Figure 3.10. Heatmap clustering of top 1000 differentially methylated CpGs by individual and race. A) Each row represents a CpG. Variance was calculated across samples

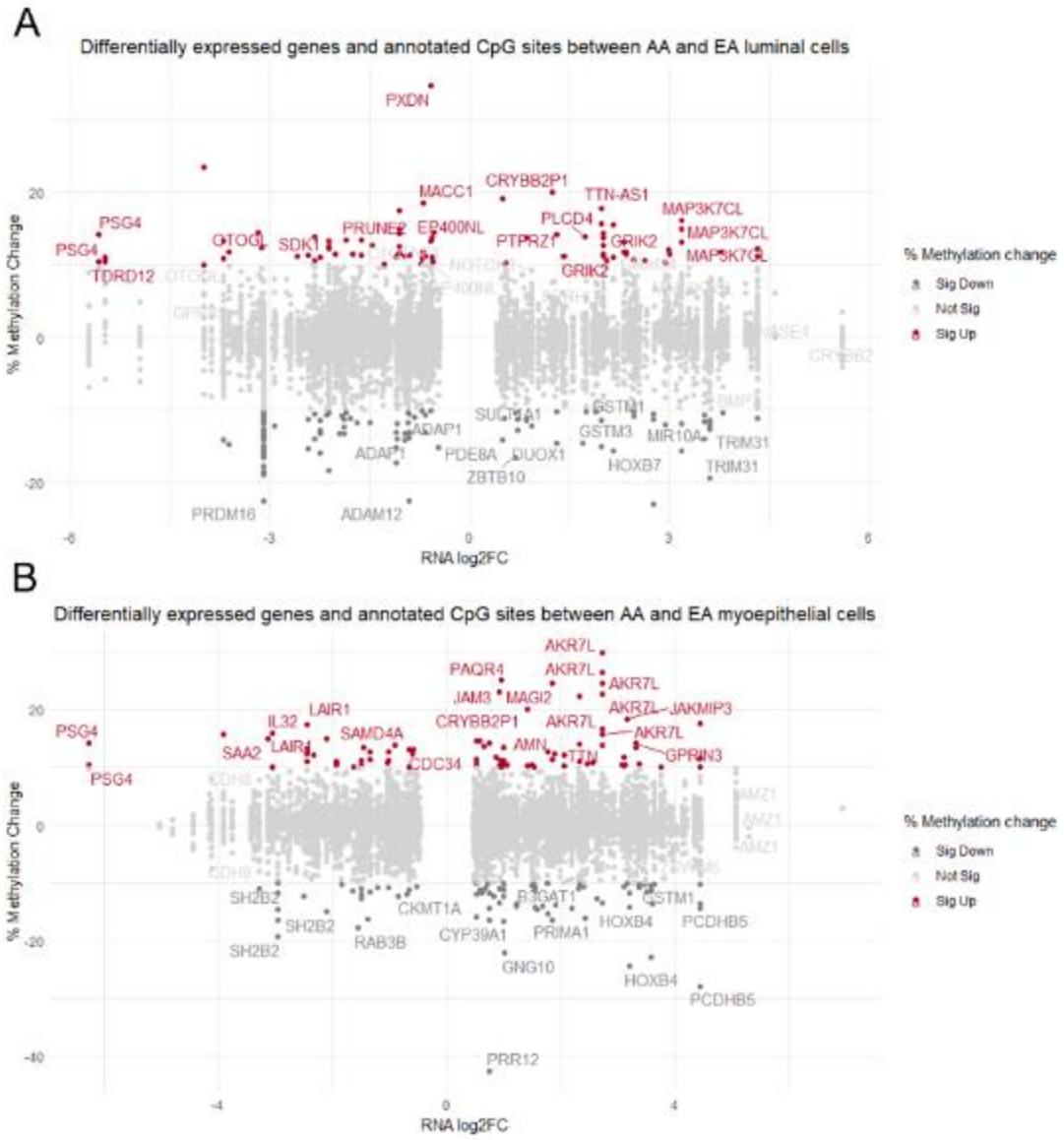


Figure 3.11. Percent methylation change of differentially methylated CpG sites on differentially expressed genes between AA and EA mammary cells A) Luminal DEGs and annotated CpG sites. Differentially methylated CpGs with greater than 10% methylation change are colored in red if up and dark grey if down. B) Myoepithelial DEGs and annotated CpG sites.

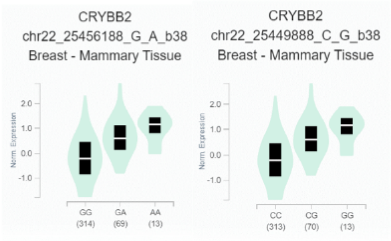
A

Gene	SNP	eQTL p_val	Ref	Alt	EA ref	EA alt	AA ref	AA alt	% Diff	Position	Lum Log2FC	Myo Log2FC
CRYBB2	rs5996939	3.60E-13	G	A	0.92	0.08	0.48	0.52	44%	chr22:25456188	5.58	3.94
	rs6519611	6.20E-13	C	G	0.96	0.04	0.66	0.34	29%	chr22:25449888		

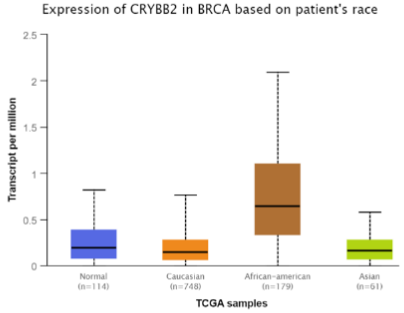
B

Sample		Genotype	
		rs5996939	rs6519611
AA	KCR7195	GA	CG
	KCR8195	GA	CG
	KCR8302	GA	GG
	KCR8451	GG	CC
	KCR8483	GA	CG
	KCR8519	AA	GG
	KCR8580	AA	GG
EA	KCR7518	GG	CC
	KCR7554	GA	CG
	KCR7953	GG	CC
	KCR8084	GG	CC
	KCR8474	GG	CC
	KCR8523	GG	CC

C

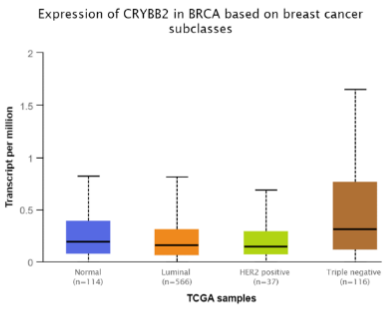


D



Comparison	Pvalue
Normal-vs-Caucasian	6.36E-01
Normal-vs-AfricanAmerican	**<1E-12**
Normal-vs-Asian	4.26E-01
Caucasian-vs-AfricanAmerican	**1.62E-12**
Caucasian-vs-Asian	5.85E-01
AfricanAmerican-vs-Asian	**9.10E-11**

E



Comparison	Pvalue
Normal-vs-Luminal	3.40E-01
Normal-vs-HER2 Positive	3.24E-0.1
Normal-vs-TNBC	**6.21E-04**
Luminal-vs-HER2 Positive	4.60E-01
Luminal-vs-TNBC	**1.50E-03**
HER2 Positive-vs-TNBC	1.26E-01

Figure 3.12. *CRYBB2*: Integration of imputed genotypes with race associated eQTLs and comparisons to TCGA tumor expression A) Aggregate allele frequency of AA and EA from dbGaP and RNA expression by cell type B) Imputed genotypes of AA and EA individuals from TOPMed C) eQTL violin plots of expression by genotype D) Expression of *CRYBB2* gene in TCGA tumors by race from UALCAN E) Expression of *CRYBB2* in TCGA tumors by subtype from UALCAN

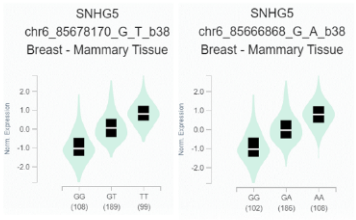
A

Gene	SNP	eQTL p_val	Ref	Alt	EA ref	EA alt	AA ref	AA alt	% Diff	Position	Lum Log2FC	Myo Log2FC
SNHG5	rs1059307	2.40E-137	G	T	0.46	0.54	0.74	0.26	28%	chr6_85678170	-1.09	-1.25
	rs9450311	2.40E-115	G	A	0.45	0.55	0.76	0.24	31%	chr6:85666868		

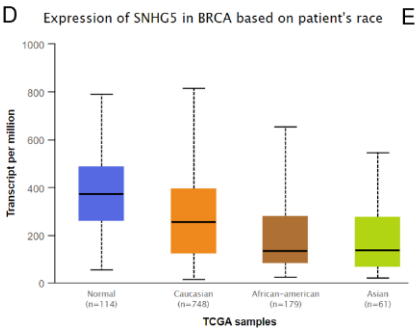
B

Sample		Genotype	
		rs1059307	rs9450311
AA	KCR7195	GT	GA
	KCR8195	GT	GA
	KCR8302	GG	GG
	KCR8451	GG	GG
	KCR8483	GG	GG
	KCR8519	GG	GG
	KCR8580	GG	GG
EA	KCR7518	GT	GA
	KCR7554	GT	GA
	KCR7953	GT	GA
	KCR8084	TT	AA
	KCR8474	GT	GA
	KCR8523	GG	GG

C

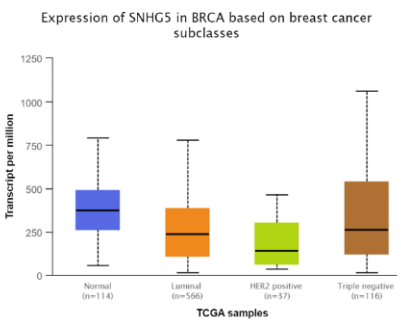


D



Comparison	Pvalue
Normal-vs-Caucasian	**2.73E-06**
Normal-vs-AfricanAmerican	**1.69E-07**
Normal-vs-Asian	**1.74E-09**
Caucasian-vs-AfricanAmerican	**2.44E-02**
Caucasian-vs-Asian	**6.61E-04**
AfricanAmerican-vs-Asian	2.13E-01

E



Comparison	Pvalue
Normal-vs-Luminal	3.40E-01
Normal-vs-HER2 Positive	3.24E-0.1
Normal-vs-TNBC	**6.21E-04**
Luminal-vs-HER2 Positive	4.60E-01
Luminal-vs-TNBC	**1.50E-03**
HER2 Positive-vs-TNBC	1.26E-01

Figure 3.13. *SNHG5*: Integration of imputed genotypes with race associated eQTLs and comparisons to TCGA tumor expression A) Aggregate allele frequency of AA and EA from dbGaP and RNA expression by cell type B) Imputed genotypes of AA and EA individuals from TOPMed C) eQTL violin plots of expression by genotype D) Expression of *SNHG5* gene in TCGA tumors by race from UALCAN E) Expression of *SNHG5* in TCGA tumors by subtype from UALCAN

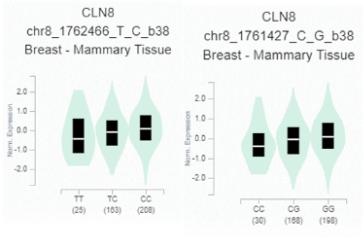
A

Gene	SNP	eQTL p_val	Ref	Alt	EA ref	EA alt	AA ref	AA alt	% Diff	Position	Lum Log2FC
CLN8	rs6558535	6.70E-12	T	C	0.32	0.68	0.73	0.27	41.0%	chr8:1762466	-0.54
	rs6558532	1.60E-11	C	G	0.28	0.72	0.41	0.59	13%	chr8:1761427	

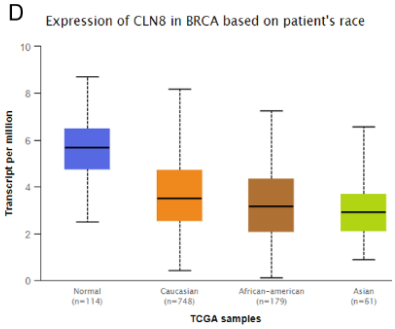
B

Sample		Genotype	
		rs6558535	rs6558532
AA	KCR7195	TC	CC
	KCR8195	CC	CG
	KCR8302	TT	CC
	KCR8451	TT	CC
	KCR8483	TC	CG
	KCR8519	CC	CG
	KCR8580	TC	CC
EA	KCR7518	CC	GG
	KCR7554	TC	CG
	KCR7953	TT	CC
	KCR8084	TC	CG
	KCR8474	CC	GG
	KCR8523	CC	GG

C

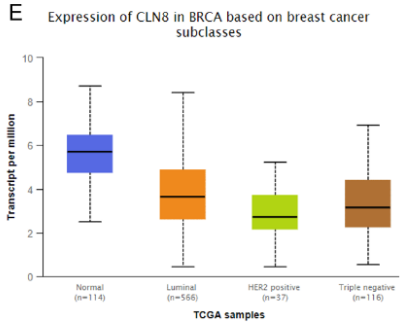


D



Comparison	Pvalue
Normal-vs-Caucasian	**<1E-12**
Normal-vs-AfricanAmerican	**<1E-12**
Normal-vs-Asian	**3.22E-12**
Caucasian-vs-AfricanAmerican	**1.98E-03**
Caucasian-vs-Asian	**3.07E-02**
AfricanAmerican-vs-Asian	8.32E-01

E



Comparison	Pvalue
Normal-vs-Luminal	**1.62E-12**
Normal-vs-HER2 Positive	**1.36E-09**
Normal-vs-TNBC	**2.83E-12**
Luminal-vs-HER2 Positive	**2.62E-02**
Luminal-vs-TNBC	5.04E-01
HER2 Positive-vs-TNBC	1.89E-01

Figure 3.14. *CLN8*: Integration of imputed genotypes with race associated eQTLs and comparisons to TCGA tumor expression A) Aggregate allele frequency of AA and EA from dbGaP and RNA expression by cell type B) Imputed genotypes of AA and EA individuals from TOPMed C) eQTL violin plots of expression by genotype D) Expression of *CLN8* gene in TCGA tumors by race from UALCAN E) Expression of *CLN8* in TCGA tumors by subtype from UALCAN

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/CAAC.21660

2. Zeinomar N, Oskar S, Kehm RD, Sahebzeda S, Terry MB. Environmental exposures and breast cancer risk in the context of underlying susceptibility: A systematic review of the epidemiological literature. *Environ Res*. 2020;187:109346. doi:10.1016/j.envres.2020.109346

3. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370

4. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

5. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol*. 2014;5(3):412. doi:10.5306/WJCO.V5.I3.412

6. Johnson KS, Conant EF, Soo MS. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *J Breast Imaging*. 2021;3(1):12-24. doi:10.1093/JBI/WBAA110

7. Dietze EC, Sistrunk C, Miranda-Carboni G, O'Regan R, Seewaldt VL. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat Rev Cancer*. 2015;15(4):248-254. doi:10.1038/nrc3896

8. Carey LA, Perou CM, Livasy CA, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *JAMA*. 2006;295(21):2492-2502. doi:10.1001/JAMA.295.21.2492

9. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev*. 2017;26(4):444-457. doi:10.1158/1055-9965.EPI-16-0858

10. Newman LA. Breast Cancer Disparities: Socioeconomic Factors versus Biology. *Ann Surg Oncol*. 2017;24(10):2869-2875. doi:10.1245/S10434-017-5977-1/TABLES/2

11. Ooi SL, Martinez ME, Li CI. Disparities in breast cancer characteristics and outcomes by race/ethnicity. *Breast Cancer Res Treat*. 2011;127(3):729-738. doi:10.1007/S10549-010-1191-6/TABLES/6

12. Newman LA, Kaljee LM. Health Disparities and Triple-Negative Breast Cancer in African American Women. *JAMA Surg*. 2017;152(5):485. doi:10.1001/jamasurg.2017.0005

13. Stringer-Reasor EM, Elkhanany A, Khoury K, Simon MA, Newman LA. Disparities in Breast Cancer Associated With African American Identity. *Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet*. 2021;41(41):e29-e46. doi:10.1200/EDBK_319929

14. Yedjou CG, Sims JN, Miele L, et al. Health and Racial Disparity in Breast Cancer. *Adv Exp Med Biol*. 2019;1152:31. doi:10.1007/978-3-030-20301-6_3

15. Newman LA, Carpten J. Integrating the Genetics of Race and Ethnicity Into Cancer Research: Trailing Jane and John Q. Public. *JAMA Surg*. 2018;153(4):299. doi:10.1001/JAMASURG.2017.5080

16. Sung H, DeSantis CE, Fedewa SA, Kantelhardt EJ, Jemal A. Breast cancer subtypes among Eastern-African-born black women and other black women in the United States. *Cancer*. 2019;125(19):3401-3411. doi:10.1002/CNCR.32293

17. Jiagge E, Jibril AS, Chitale D, et al. Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer. *Ann Surg Oncol*. 2016;23(12):3843-3849. doi:10.1245/S10434-016-5420-Z/TABLES/5

18. Capp J-P. Cancer Stem Cells: From Historical Roots to a New Perspective. *J Oncol*. 2019;2019:1-10. doi:10.1155/2019/5189232

19. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. doi:10.1038/35102167

20. Ma Y, Zhang P, Wang F, Yang J, Yang Z, Qin H. The relationship between early embryo development and tumourigenesis. *J Cell Mol Med*. 2010;14(12):2697-2701. doi:10.1111/j.1582-4934.2010.01191.x

21. Reya T, Clevers H. Wnt signalling in stem cells and cancer. *Nature*. 2005;434(7035):843-850. doi:10.1038/nature03319

22. Papaccio F, Paino F, Regad T, Papaccio G, Desiderio V, Tirino V. Concise Review: Cancer Cells, Cancer Stem Cells, and Mesenchymal Stem Cells: Influence in Cancer Development. *Stem Cells Transl Med*. 2017;6(12):2115-2125. doi:10.1002/sctm.17-0138

23. Clevers H. The cancer stem cell: premises, promises and challenges. *Nat Med*. 2011;17(3):313-319. doi:10.1038/nm.2304

24. Goichberg P. Current Understanding of the Pathways Involved in Adult Stem and Progenitor Cell Migration for Tissue Homeostasis and Repair. *Stem Cell Rev Reports*. 2016;12(4):421-437. doi:10.1007/s12015-016-9663-7

25. Shaw TJ, Martin P. Wound repair: a showcase for cell plasticity and migration. *Curr Opin Cell Biol*. 2016;42:29-37. doi:10.1016/J.CEB.2016.04.001

26. Hennighausen L, Robinson GW. Signaling Pathways in Mammary Gland Development. *Dev Cell*. 2001;1(4):467-475. doi:10.1016/S1534-5807(01)00064-8

27. Macias H, Hinck L. Mammary Gland Development. *Wiley Interdiscip Rev Dev Biol*. 2012;1(4):533. doi:10.1002/WDEV.35

28. Russo J. The Windows of Susceptibility to Breast Cancer. *Pathobiol Breast Cancer*. Published online 2016:1-20. doi:10.1007/978-3-319-40815-6_1

29. Mammary Stem Cells: Premise, Properties, and Perspectives. *Trends Cell Biol*. 2017;27(8):556-567. doi:10.1016/J.TCB.2017.04.001

30. Shackleton M, Vaillant F, Simpson KJ, et al. Generation of a functional mammary gland from a single stem cell. *Nature*. 2006;439(7072):84-88. doi:10.1038/nature04372

31. Daniel CW, Deome KB, Young JT, Blair PB, Faulkin LJ. The in vivo life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *J Mammary Gland Biol Neoplasia*. 2009;14(3):355-362. doi:10.1007/s10911-009-9139-3

32. Stingl J, Eaves CJ, Zandieh I, Emerman JT. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat*. 2001;67(2):93-109. doi:10.1023/A:1010615124301

33. Ghebeh H, Sleiman GM, Manogaran PS, et al. Profiling of normal and malignant breast tissue show CD44high/CD24low phenotype as a predominant stem/progenitor marker when used in

combination with Ep-CAM/CD49f markers. *BMC Cancer*. 2013;13:289. doi:10.1186/1471-2407-13-289

34. Ginestier C, Hur MH, Charafe-Jauffret E, et al. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell*. 2007;1(5):555-567. doi:10.1016/j.stem.2007.08.014

35. Schwartz T, Stark A, Pang J, et al. Expression of aldehyde dehydrogenase 1 as a marker of mammary stem cells in benign and malignant breast lesions of Ghanaian women. *Cancer*. 2013;119(3):488-494. doi:10.1002/cncr.27737

36. Jiagge E, Chitale D, Newman LA, Newman A. Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *J Pathol*. 2018;188(2). doi:10.1016/j.ajpath.2017.06.020

37. Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature*. 2016;529(7584):43-47. doi:10.1038/nature16166

38. Gray J, Evans N, Taylor B, Rizzo J, Walker M. State of the Evidence: The Connection Between Breast Cancer and the Environment. *Int J Occup Environ Health*. 2009;15(1):43-78. doi:10.1179/107735209799449761

39. Forman MR, Winn DM, Collman GW, Rizzo J, Birnbaum LS. Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod Toxicol*. 2015;54:6-10. doi:10.1016/j.reprotox.2014.10.019

40. Nakshatri H, Anjanappa M, Bhat-Nakshatri P. Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. *Sci Rep*. 2015;5:13526. doi:10.1038/srep13526

41. Colacino JA, Azizi E, Brooks MD, et al. Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*. Published online 2018. doi:10.1016/j.stemcr.2018.03.001

42. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med*. 2009;27(5):351-357. doi:10.1055/s-0029-1237423

43. Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol Rev*. 2020;100(2):489-523. doi:10.1152/PHYSREV.00040.2018/ASSET/IMAGES/LARGE/Z9J0012029270010.JPEG

44. Terry MB, Michels KB, Brody JG, et al. Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research. *Breast Cancer Res 2019 211*. 2019;21(1):1-16. doi:10.1186/S13058-019-1168-2

45. Attina TM, Malits J, Naidu M, Trasande L. Racial/ethnic disparities in disease burden and costs related to exposure to endocrine-disrupting chemicals in the United States: an exploratory analysis. *J Clin Epidemiol*. 2019;108:34-43. doi:10.1016/j.jclinepi.2018.11.024

46. Ruiz D, Becerra M, Jagai JS, Ard K, Sargis RM. Disparities in Environmental Exposures to Endocrine-Disrupting Chemicals and Diabetes Risk in Vulnerable Populations. *Diabetes Care*. 2018;41(1):193-205. doi:10.2337/dc16-2765

47. Lillo MA, Nichols C, Seagroves TN, Miranda-Carboni GA, Krum SA. Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Horm Cancer*. 2017;8(2):90-99. doi:10.1007/s12672-017-0286-5

48. Hindman AR, Mo XM, Helber HL, et al. Varying Susceptibility of the Female Mammary Gland to In Utero Windows of BPA Exposure. *Endocrinology*. 2017;158(10):3435-3447. doi:10.1210/en.2017-00116

49. Gerona RR, Pan J, Zota AR, et al. Direct measurement of Bisphenol A (BPA), BPA glucuronide and BPA sulfate in a diverse and low-income population of pregnant women reveals high exposure, with potential implications for previous exposure estimates: a cross-sectional study. *Environ Health*. 2016;15:50. doi:10.1186/s12940-016-0131-2

50. Tucker DK, Hayes Bouknight S, Brar SS, Kissling GE, Fenton SE. Evaluation of Prenatal Exposure to Bisphenol Analogues on Development and Long-Term Health of the Mammary Gland in Female Mice. *Environ Health Perspect*. 2018;126(8):087003. doi:10.1289/EHP3189

51. Huang W, Zhao C, Zhong H, Zhang S, Xia Y, Cai Z. Bisphenol S induced epigenetic and transcriptional changes in human breast cancer cell line MCF-7. *Environ Pollut*. 2019;246:697-703. doi:10.1016/J.ENVPOL.2018.12.084

52. Weng Y-I, Hsu P-Y, Liyanarachchi S, et al. Epigenetic influences of low-dose bisphenol A in primary human breast epithelial cells. *Toxicol Appl Pharmacol*. 2010;248(2):111. doi:10.1016/J.TAAP.2010.07.014

53. Liu X, Krawczyk E, Suprynowicz FA, et al. Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc*. 2017;12(2):439-451. doi:10.1038/nprot.2016.174

54. Giraddi RR, Chung C-Y, Heinz RE, et al. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep*. 2018;24(6):1653-1666.e7. doi:10.1016/j.celrep.2018.07.025

55. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18(1):35-45. doi:10.1038/nri.2017.76

56. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257-272. doi:10.1038/s41576-019-0093-7

57. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health*. 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629

58. Holliday H, Baker LA, Junankar SR, Clark SJ, Swarbrick A. Epigenomics of mammary gland development. *Breast Cancer Res*. 2018;20(1):100. doi:10.1186/s13058-018-1031-x

59. Thong T, Forté CA, Hill EM, Colacino JA. Environmental exposures, stem cells, and cancer. *Pharmacol Ther*. Published online 2019. doi:10.1016/j.pharmthera.2019.107398

60. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest*. 2009;119(6):1420-1428. doi:10.1172/JCI39104

61. Merrell AJ, Stanger BZ. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style. *Nat Rev Mol Cell Biol*. 2016;17(7):413-425. doi:10.1038/nrm.2016.24

62. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med*. 2013;19(11):1438-1449. doi:10.1038/nm.3336

63. Jolly MK, Boareto M, Huang B, et al. Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol*. 2015;5:155. doi:10.3389/fonc.2015.00155

64. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

65. Jolly MK, Tripathi SC, Jia D, et al. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*. 2016;7(19):27067-27084. doi:10.18632/oncotarget.8166

66. Dontu G, Abdallah WM, Foley JM, et al. In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev*. 2003;17(10):1253-1270. doi:10.1101/gad.1061803

67. Colacino JA, McDermott SP, Sartor MA, Wicha MS, Rozek LS. Transcriptomic profiling of curcumin-treated human breast stem cells identifies a role for stearoyl-coa desaturase in breast cancer prevention. *Breast Cancer Res Treat*. 2016;158(1):29-41. doi:10.1007/s10549-016-3854-4

68. Liu X, Ory V, Chapman S, et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am J Pathol*. 2012;180(2):599-607. doi:10.1016/j.ajpath.2011.10.036

69. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/J.CELL.2015.05.002

70. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33(5):495-502. doi:10.1038/nbt.3192

71. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Birol I, ed. *Bioinformatics*. 2019;35(3):487-496. doi:10.1093/bioinformatics/bty644

72. Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, Chang HY. Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells. *Cell Stem Cell*. 2008;2(4):333-344. doi:10.1016/j.stem.2008.02.009

73. Kuleshov M V., Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

74. Lim E, Vaillant F, Wu D, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*. 2009;15(8):907-913. doi:10.1038/nm.2000

75. Bach K, Pensa S, Grzelak M, et al. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun*. 2017;8(1):2128. doi:10.1038/s41467-017-02001-5

76. Nguyen QH, Pervolarakis N, Blake K, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-04334-1

77. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71-e71. doi:10.1093/nar/gkv1507

78. Wang Y, Thong T, Saligrama V, Colacino J, Balzano L, Scott C. A gene filter for comparative analysis of single-cell RNA-sequencing trajectory datasets. *bioRxiv*. Published online May 17, 2019:637488. doi:10.1101/637488

79. Abd El-Rehim DM, Pinder SE, Paish CE, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol*. 2004;203(2):661-671. doi:10.1002/path.1559

80. O'Hare MJ, Ormerod MG, Monaghan P, Lane EB, Gusterson BA. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. *Differentiation*. 1991;46(3):209-221. doi:10.1111/j.1432-0436.1991.tb00883.x

81. Marcato P, Dean CA, Pan D, et al. Aldehyde Dehydrogenase Activity of Breast Cancer Stem Cells Is Primarily Due To Isoform ALDH1A3 and Its Expression Is Predictive of Metastasis. *Stem Cells*. 2011;29(1):32-45. doi:10.1002/stem.563

82. Trzpis M, McLaughlin PMJ, De Leij LMFH, Harmsen MC. Epithelial cell adhesion molecule: More than a carcinoma marker and adhesion molecule. *Am J Pathol*. 2007;171(2):386-395. doi:10.2353/ajpath.2007.070152

83. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A*. 2003;100(7):3983-3988. doi:10.1073/pnas.0530291100

84. Gerdur Ísberg Ó, Kim J, Fridriksdottir AJ, et al. A CD146 FACS Protocol Enriches for Luminal Keratin 14/19 Double Positive Human Breast Progenitors. *Sci Rep*. 2019;9(1):14843. doi:10.1038/s41598-019-50903-9

85. Jolly MK, Somarelli JA, Sheth M, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Ther*. 2019;194:161-184. doi:10.1016/j.pharmthera.2018.09.007

86. Casey T, Bond J, Tighe S, et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res Treat*. 2009;114(1):47-62. doi:10.1007/s10549-008-9982-8

87. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173(2):338-354.e15. doi:10.1016/j.cell.2018.03.034

88. Feng Y, Sun B, Li X, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. *Breast Cancer Res Treat*. 2007;103(3):319-329. doi:10.1007/s10549-006-9385-7

89. Folgueira MAAK, Maistro S, Katayama MLH, et al. Markers of breast cancer stromal fibroblasts in the primary tumour site associated with lymph node metastasis: A systematic review including our case series. *Biosci Rep*. 2013;33(6). doi:10.1042/BSR20130060

90. Li J-M, Tseng C-W, Lin C-C, et al. Upregulation of LGALS1 is associated with oral cancer metastasis. *Ther Adv Med Oncol*. 2018;10:175883591879462. doi:10.1177/1758835918794622

91. Falato C, Lorent J, Tani E, et al. Ki67 measured in metastatic tissue and prognosis in patients with advanced breast cancer. *Breast Cancer Res Treat*. 2014;147(2):407-414. doi:10.1007/s10549-014-3096-2

92. Montes de Oca R, Gurard-Levin ZA, Berger F, et al. The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol Oncol*. 2015;9(3):657-674. doi:10.1016/j.molonc.2014.11.002

93. Ding K, Li W, Zou Z, Zou X, Wang C. CCNB1 is a prognostic biomarker for ER+ breast cancer. *Med Hypotheses*. 2014;83(3):359-364. doi:10.1016/j.mehy.2014.06.013

94. Li Y, Chen Y-L, Xie Y-T, et al. Association Study of Germline Variants in CCNB1 and CDK1 with Breast Cancer Susceptibility, Progression, and Survival among Chinese Han Women. Miao X, ed. *PLoS One*. 2013;8(12):e84489. doi:10.1371/journal.pone.0084489

95. Wang Y, Zhang C, Mai L, Niu Y, Wang Y, Bu Y. PRR11 and SKA2 gene pair is overexpressed and regulated by p53 in breast cancer. *BMB Rep*. 2019;52(2):157-162. doi:10.5483/BMBRep.2019.52.2.207

96. Ren Z, Yang T, Zhang P, Liu K, Liu W, Wang P. SKA2 mediates invasion and metastasis in human breast cancer via EMT. *Mol Med Rep*. 2019;19(1):515-523. doi:10.3892/mmr.2018.9623

97. Han JY, Han YK, Park GY, et al. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines. *Sci Rep*. 2015;5. doi:10.1038/srep15993

98. Hamy AS, Bieche I, Lehmann-Che J, et al. BIRC5 (survivin): a pejorative prognostic marker in stage II/III breast cancer with no response to neoadjuvant chemotherapy. *Breast Cancer Res Treat*. 2016;159(3):499-511. doi:10.1007/s10549-016-3961-2

99. Ghaffari K, Hashemi M, Ebrahimi E, Shirkoohi R. BIRC5 genomic copy number variation in early-onset breast cancer. *Iran Biomed J*. 2016;20(4):241-245. doi:10.7508/ibj.2016.04.009

100. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

101. Yu M, Bardia A, Wittner BS, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (80-)*. 2013;339(6119):580-584. doi:10.1126/science.1228522

102. Sun P, Yuan Y, Li A, Li B, Dai X. Cytokeratin expression during mouse embryonic and early postnatal mammary gland development. *Histochem Cell Biol*. 2010;133(2):213-221. doi:10.1007/s00418-009-0662-5

103. Ruiz R, Herrero C, Strasser-Weippl K, et al. Epidemiology and pathophysiology of pregnancy-associated breast cancer: A review. *Breast*. 2017;35:136-141. doi:10.1016/j.breast.2017.07.008

104. Johansson ALV, Andersson TML, Hsieh CC, et al. Tumor characteristics and prognosis in women with pregnancy-associated breast cancer. *Int J Cancer*. 2018;142(7):1343-1354. doi:10.1002/ijc.31174

105. Degnim AC, Visscher DW, Hoskin TL, et al. Histologic findings in normal breast tissues: comparison to reduction mammaplasty and benign breast disease tissues. *Breast Cancer Res Treat*. 2012;133(1):169-177. doi:10.1007/s10549-011-1746-1

106. Rheinwatd JG, Green H. Seria cultivation of strains of human epidemal keratinocytes: the formation keratinizin colonies from single cell is. *Cell*. 1975;6(3):331-343. doi:10.1016/S0092-8674(75)80001-8

107. Domchek SM, Yao S, Chen F, et al. Comparison of the Prevalence of Pathogenic Variants in Cancer Susceptibility Genes in Black Women and Non-Hispanic White Women With Breast Cancer in the United States. *JAMA Oncol*. 2021;7(7):1. doi:10.1001/JAMAONCOL.2021.1492

108. Shimelis H, LaDuca H, Hu C, et al. Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst*. 2018;110(8):855-862. doi:10.1093/jnci/djy106

109. Roelands J, Mall R, Almeer H, et al. Ancestry-associated transcriptomic profiles of breast cancer in patients of African, Arab, and European ancestry. *NPJ Breast Cancer*. 2021;7(1). doi:10.1038/S41523-021-00215-X

110. Nievergelt CM, Maihofer AX, Shekhtman T, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*. 2013;4(1):13. doi:10.1186/2041-2223-4-13

111. Thong T, Wang Y, Brooks MD, et al. Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Front Cell Dev Biol*. 2020;8:288. doi:10.3389/FCELL.2020.00288/BIBTEX

112. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol 2018 365*. 2018;36(5):411-420. doi:10.1038/nbt.4096

113. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048/ATTACHMENT/1990279C-028D-47D4-A5D3-C89B04E14795/MMC3.XLSX

114. Chapter 4 DE analyses between conditions | Multi-Sample Single-Cell Analyses with Bioconductor. Accessed May 25, 2022. http://bioconductor.org/books/3.14/OSCA.multisample/multi-sample-comparisons.html

115. Gulati GS, Sikandar SS, Wesche DJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*. 2020;367(6476):405-411. doi:10.1126/SCIENCE.AAX0249

116. Using escape to perform ssGSEA analyses on single-cell RNA-seq data. Accessed May 25, 2022. http://www.bioconductor.org/packages/release/bioc/vignettes/escape/inst/doc/vignette.html

117. GSEA | MSigDB | Browse Gene Sets. Accessed May 25, 2022. https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

118. Cell-Cycle Scoring and Regression • Seurat. Accessed May 19, 2022. https://satijalab.org/seurat/articles/cell_cycle_vignette.html

119. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/BIOINFORMATICS/BTS480

120. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363. doi:10.1093/BIOINFORMATICS/BTU049

121. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res*. 2016;44(3):e20. doi:10.1093/NAR/GKV907

122. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics and Chromatin*. 2015;8(1):1-16. doi:10.1186/1756-8935-8-6/FIGURES/6

123. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208. doi:10.1186/s13059-016-1066-1

124. Fortin J-P, Triche TJ, Hansen KD, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2017;33(4):558-560.

doi:10.1093/bioinformatics/btw691

125. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287. doi:10.1038/NG.3656

126. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443. doi:10.1038/NG.3679

127. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(10):1654. doi:10.1001/JAMAONCOL.2017.0595

128. Wu H, Wang C, Wu Z. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*. 2015;31(2):233-241. doi:10.1093/BIOINFORMATICS/BTU640

129. Yap SQ, Mathavarajah S, Huber RJ. The converging roles of Batten disease proteins in neurodegeneration and cancer. *iScience*. 2021;24(4):102337. doi:10.1016/J.ISCI.2021.102337

130. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet (London, England)*. 2005;365(9460):671-679. doi:10.1016/S0140-6736(05)17947-1

131. Pani T, Rajput K, Kar A, et al. Alternative splicing of ceramide synthase 2 alters levels of specific ceramides and modulates cancer cell proliferation and migration in Luminal B breast cancer subtype. *Cell Death Dis 2021 122*. 2021;12(2):1-22. doi:10.1038/s41419-021-03436-x

132. D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, Troester MA. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res Treat*. 2015;152(2):437-448. doi:10.1007/S10549-015-3474-4/TABLES/6

133. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(12):1654-1662. doi:10.1001/JAMAONCOL.2017.0595

134. Barrow MA, Martin ME, Coffey A, et al. A functional role for the cancer disparity-linked genes, CRYβB2 and CRYβB2P1, in the promotion of breast cancer. *Breast Cancer Res*. 2019;21(1):105. doi:10.1186/s13058-019-1191-3

135. Field LA, Love B, Deyarmin B, Hooke JA, Shriver CD, Ellsworth RE. Identification of differentially expressed genes in breast tumors from African American compared with Caucasian women. *Cancer*. 2012;118(5):1334-1344. doi:10.1002/CNCR.26405

136. Li H, Cheng C, You W, et al. TMEM100 Modulates TGF-β Signaling Pathway to Inhibit Colorectal Cancer Progression. *Gastroenterol Res Pract*. 2021;2021. doi:10.1155/2021/5552324

137. Ma J, Yan T, Bai Y, et al. TMEM100 negatively regulated by microRNA-106b facilitates cellular apoptosis by suppressing survivin expression in NSCLC. *Oncol Rep*. 2021;46(2):1-14. doi:10.3892/OR.2021.8136

138. Zhuang J, Huang Y, Zheng W, et al. TMEm100 expression suppresses metastasis and enhances sensitivity to chemotherapy in gastric cancer. *Biol Chem*. 2020;401(2):285-296. doi:10.1515/HSZ-2019-0161/MACHINEREADABLECITATION/RIS

139. Donald CD, Sun CQ, Lim SD, et al. Cancer-Specific Loss of β-Defensin 1 in Renal and Prostatic Carcinomas. *Lab Investig 2003 834*. 2003;83(4):501-505. doi:10.1097/01.lab.0000063929.61760.f6

140. Lee J, Han JH, Jang A, Kim JW, Hong SA, Myung SC. DNA Methylation-Mediated Downregulation of DEFB1 in Prostate Cancer Cells. *PLoS One*. 2016;11(11). doi:10.1371/JOURNAL.PONE.0166664

141. Lo JH, Battaglin F, Baca Y, et al. DEFB1 gene expression and the molecular landscape of colorectal cancer (CRC). *https://doi.org/101200/JCO20224016_suppl3523*. 2022;40(16_suppl):3523-3523. doi:10.1200/JCO.2022.40.16_SUPPL.3523

142. Li YH, Hu YQ, Wang SC, Li Y, Chen DM. LncRNA SNHG5: A new budding star in human cancers. *Gene*. 2020;749:144724. doi:10.1016/J.GENE.2020.144724

143. Chi JR, Yu ZH, Liu BW, et al. SNHG5 Promotes Breast Cancer Proliferation by Sponging the miR-154-5p/PCNA Axis. *Mol Ther Nucleic Acids*. 2019;17:138. doi:10.1016/J.OMTN.2019.05.013

144. Damas ND, Marcatti M, Côme C, et al. SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nat Commun 2016 71*. 2016;7(1):1-14. doi:10.1038/ncomms13875

145. Shen M, Jiang YZ, Wei Y, et al. Tinagl1 Suppresses Triple-Negative Breast Cancer Progression and Metastasis by Simultaneously Inhibiting Integrin/FAK and EGFR Signaling. *Cancer Cell*. 2019;35(1):64-80.e7. doi:10.1016/J.CCELL.2018.11.016/ATTACHMENT/1E19EDC0-441D-4164-93EE-F3A5B9EC190A/MMC1.PDF

146. Musetti SN, Huang L. Tinagl1 Gene Therapy Suppresses Growth and Remodels the Microenvironment of Triple Negative Breast Cancer. *Mol Pharm*. 2021;18(5):2032-2038. doi:10.1021/ACS.MOLPHARMACEUT.1C00008

147. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob Heal*. 2020;8(8):e1027-e1037. doi:10.1016/S2214-109X(20)30215-1

148. Zota AR, Shamasunder B. The environmental injustice of beauty: framing chemical exposures from beauty products as a health disparities concern. *Am J Obstet Gynecol*. 2017;217(4):418.e1-418.e6. doi:10.1016/J.AJOG.2017.07.020

149. Ash M, Boyce JK. Racial disparities in pollution exposure and employment at US industrial facilities. *Proc Natl Acad Sci U S A*. 2018;115(42):10636-10641. doi:10.1073/PNAS.1721640115/SUPPL_FILE/PNAS.1721640115.SAPP.PDF

150. Polemi KM, Nguyen VK, Heidt J, Kahana A, Jolliet O, Colacino JA. Identifying the link between chemical exposures and breast cancer in African American women via integrated in vitro and exposure biomarker data. *Toxicology*. 2021;463. doi:10.1016/J.TOX.2021.152964

151. Nguyen VK, Kahana A, Heidt J, et al. A comprehensive analysis of racial disparities in chemical biomarker concentrations in United States women, 1999–2014. *Environ Int*. 2020;137:105496. doi:10.1016/J.ENVINT.2020.105496

152. Atlas E, Dimitrova V. Bisphenol S and Bisphenol A disrupt morphogenesis of MCF-12A human mammary epithelial cells. *Sci Rep*. 2019;9(1):1-10. doi:10.1038/s41598-019-52505-x

153. LaPlante CD, Catanese MC, Bansal R, Vandenberg LN. Bisphenol S Alters the Lactating Mammary Gland and Nursing Behaviors in Mice Exposed During Pregnancy and Lactation. *Endocrinology*. 2017;158(10):3448-3461. doi:10.1210/EN.2017-00437

154. Winkler J, Liu P, Phong K, et al. Bisphenol A replacement chemicals, BPF and BPS, induce

protumorigenic changes in human mammary gland organoid morphology and proteome. *Proc Natl Acad Sci U S A*. 2022;119(11). doi:10.1073/PNAS.2115308119/-/DCSUPPLEMENTAL

155. Erler C, Novak J. Bisphenol A Exposure: Human Risk and Health Policy. *J Pediatr Nurs*. 2010;25(5):400-407. doi:10.1016/J.PEDN.2009.05.006

156. Konieczna A, Rutkowska A, Rachoń D. Health risk of exposure to Bisphenol A (BPA). *Rocz Panstw Zakl Hig*. 2015;66(1):5-11. Accessed August 5, 2022. https://pubmed.ncbi.nlm.nih.gov/25813067/

157. Zimmers SM, Browne EP, O'Keefe PW, et al. Determination of free Bisphenol A (BPA) concentrations in breast milk of U.S. women using a sensitive LC/MS/MS method. *Chemosphere*. 2014;104:237-243. doi:10.1016/j.chemosphere.2013.12.085

158. Lehmler H-J, Liu B, Gadogbe M, Bao W. Exposure to Bisphenol A, Bisphenol F, and Bisphenol S in U.S. Adults and Children: The National Health and Nutrition Examination Survey 2013-2014. *ACS omega*. 2018;3(6):6523-6532. doi:10.1021/acsomega.8b00824

159. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol 2017 361*. 2017;36(1):89-94. doi:10.1038/nbt.4042

160. Cheng M, Michalski S, Kommagani R. Role for Growth Regulation by Estrogen in Breast Cancer 1 (GREB1) in Hormone-Dependent Cancers. *Int J Mol Sci*. 2018;19(9). doi:10.3390/IJMS19092543

161. Hodgkinson K, Forrest LA, Vuong N, Garson K, Djordjevic B, Vanderhyden BC. GREB1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene 2018 3744*. 2018;37(44):5873-5886. doi:10.1038/s41388-018-0377-y

162. Haines CN, Braunreiter KM, Mo XM, Burd CJ. GREB1 isoforms regulate proliferation independent of ERα co-regulator activities in breast cancer. *Endocr Relat Cancer*. 2018;25(7):735-746. doi:10.1530/ERC-17-0496

163. Jiang S, Katayama H, Wang J, et al. Estrogen-Induced Aurora Kinase-A (AURKA) Gene Expression is Activated by GATA-3 in Estrogen Receptor-Positive Breast Cancer Cells. *Horm Cancer*. 2010;1(1):11. doi:10.1007/S12672-010-0006-X

164. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Mol Cancer 2021 201*. 2021;20(1):1-27. doi:10.1186/S12943-020-01305-3

165. Ma XF, Zhang J, Shuai HL, Guan BZ, Luo X, Yan RL. IKKβ/NF-κB mediated the low doses of bisphenol A induced migration of cervical cancer cells. *Arch Biochem Biophys*. 2015;573:52-58. doi:10.1016/J.ABB.2015.03.010

166. Sauer SJ, Tarpley M, Shah I, et al. Bisphenol A activates EGFR and ERK promoting proliferation, tumor spheroidformation and resistance to EGFR pathway inhibition in estrogen receptor-negativeinflammatory breast cancer cells. *Carcinogenesis*. 2017;38(3):252-260. doi:10.1093/CARCIN/BGX003

167. Li S, Wang B, Tang Q, Liu J, Yang X. Bisphenol A triggers proliferation and migration of laryngeal squamous cell carcinoma via GPER mediated upregulation of IL-6. *Cell Biochem Funct*. 2017;35(4):209-216. doi:10.1002/CBF.3265

168. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia (United States)*.

2017;19(8):649-658. doi:10.1016/j.neo.2017.05.002

169. Yang N, Wang C, Wang Z, et al. FOXM1 recruits nuclear Aurora kinase A to participate in a positive feedback loop essential for the self-renewal of breast cancer stem cells. *Oncogene*. 2017;36(24):3428-3440. doi:10.1038/ONC.2016.490

170. Eterno V, Zambelli A, Villani L, et al. AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. *Sci Reports 2016 61*. 2016;6(1):1-13. doi:10.1038/srep28436

171. Tan Y, Wang Q, Xie Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. *Int J Oncol*. 2019;54(1):87-97. doi:10.3892/ijo.2018.4598

172. Jian W, Deng XC, Munankarmy A, et al. KIF23 promotes triple negative breast cancer through activating epithelial-mesenchymal transition. *Gland Surg*. 2021;10(6):1941. doi:10.21037/GS-21-19

173. Li Z, Yang HY, Zhang XL, et al. Kinesin family member 23, regulated by FOXM1, promotes triple negative breast cancer progression via activating Wnt/β-catenin pathway. *J Exp Clin Cancer Res*. 2022;41(1):1-15. doi:10.1186/S13046-022-02373-7/FIGURES/9

Appendix

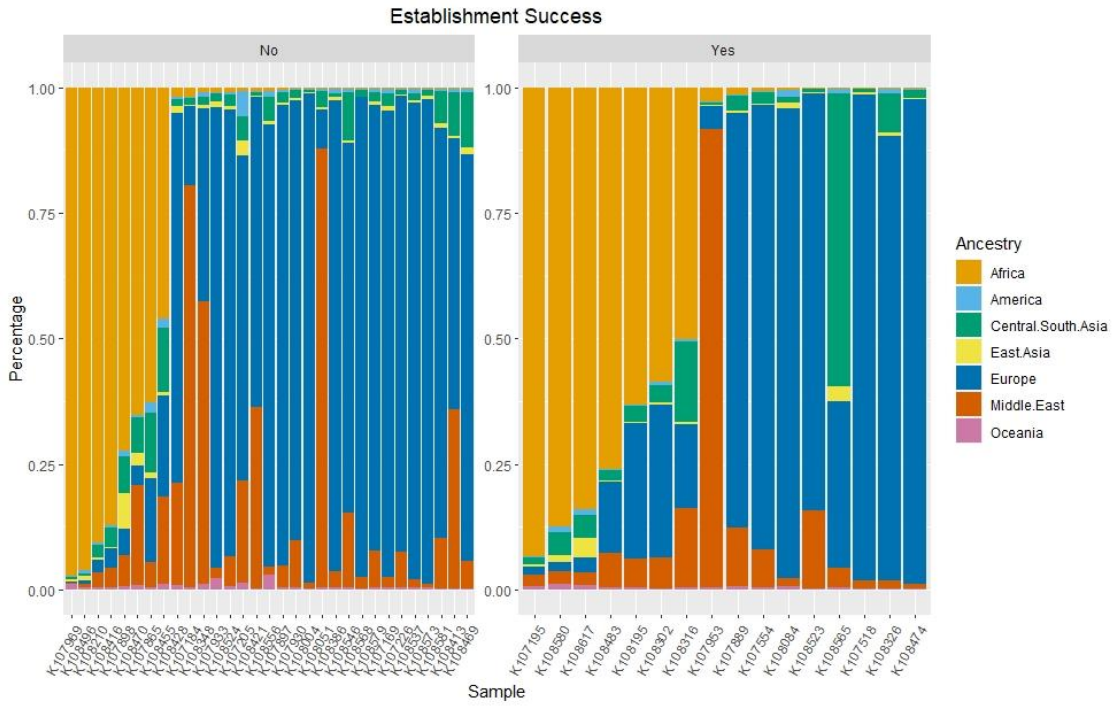


Figure 3.A1. Ancestry analysis of komen tissue samples. Quantification of ancestry percentages from SNP genotyping stratified by establishment success in CR culture.

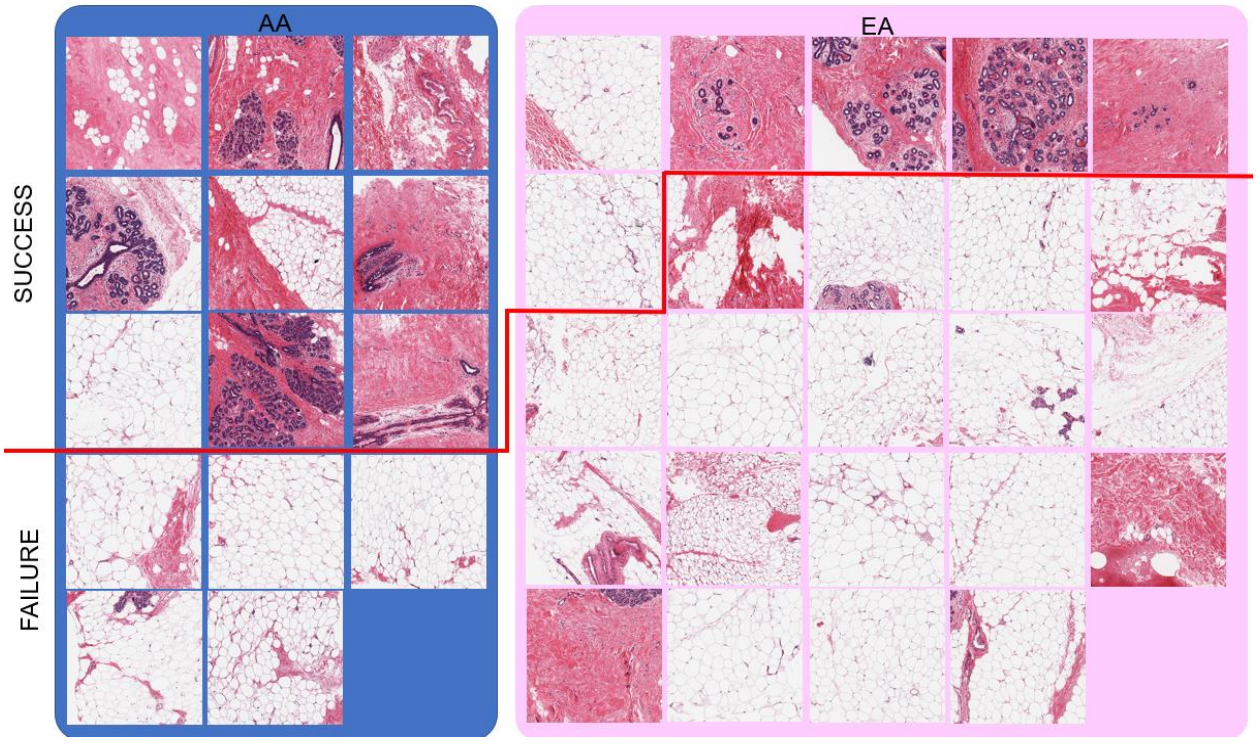
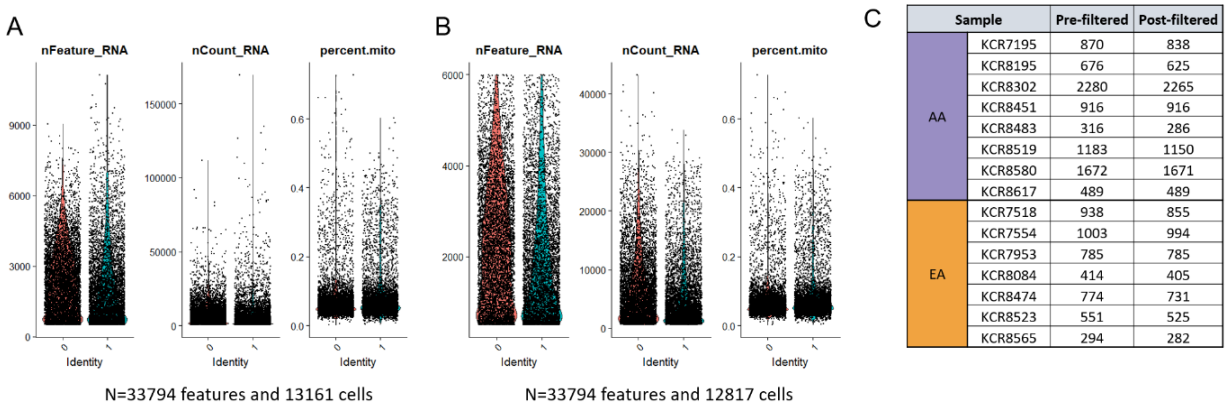


Figure 3.A2. Histological composition of Komen tissue samples. H&E histology depiction of Komen samples stratified by race and establishment success.



C

Sample		Pre-filtered	Post-filtered
AA	KCR7195	870	838
	KCR8195	676	625
	KCR8302	2280	2265
	KCR8451	916	916
	KCR8483	316	286
	KCR8519	1183	1150
	KCR8580	1672	1671
	KCR8617	489	489
EA	KCR7518	938	855
	KCR7554	1003	994
	KCR7953	785	785
	KCR8084	414	405
	KCR8474	774	731
	KCR8523	551	525
	KCR8565	294	282

Figure 3.A3. Quality control metrics for single-cell filtering split by race A) AA and EA cells pre-filtering B) Post-filtering distributions C) Number of high quality cells filtered by individual

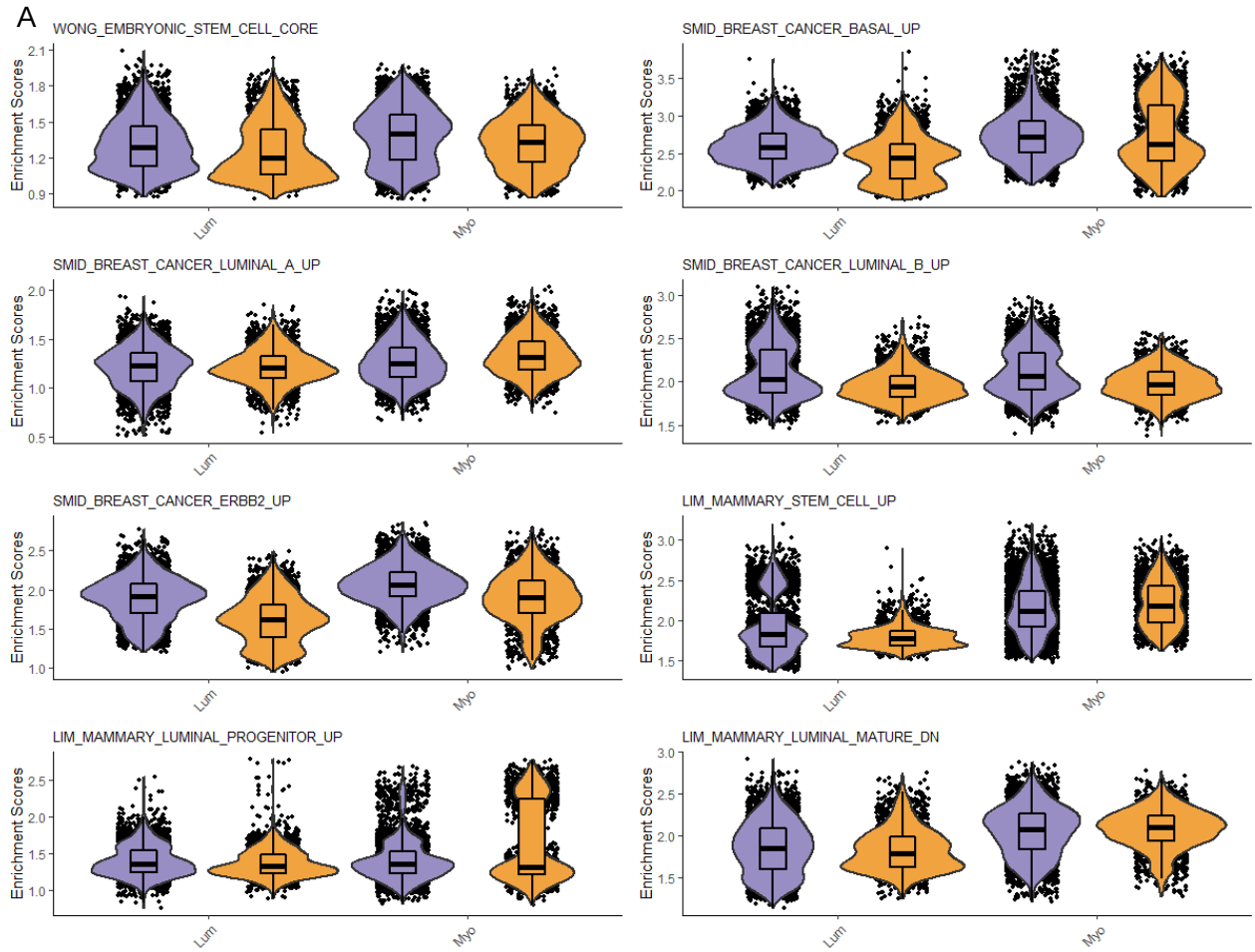


Figure 3.A4. ssGSEA of breast cancer and stem cell associated gene sets A) Stratified by cell type and race

Table 3 A.1 Linear mixed effects modeling results of ssGSEA of breast cancer and stem cell associated gene sets by race. AA race was used as the reference group and coefficients of fixed effects are shown in this table.

Gene Set	Estimate	Std. Error	df	t value	P value
WONG_EMBRYONIC_STEM_CELL_CORE	-4.00E-02	4.70E-02	1.30E+01	-8.51E-01	4.10E-01
SMID_BREAST_CANCER_BASAL_UP	-6.50E-02	1.35E-01	1.30E+01	-4.82E-01	6.37E-01
SMID_BREAST_CANCER_LUMINAL_B_UP	-8.80E-02	9.00E-02	1.20E+01	-9.75E-01	3.47E-01
SMID_BREAST_CANCER_ERBB2_UP	-1.70E-01	1.40E-01	1.30E+01	-1.23E+00	-2.40E-01
SMID_BREAST_CANCER_LUMINAL_A_UP	6.20E-02	6.60E-02	1.20E+01	9.31E-01	3.69E-01
LIM_MAMMARY_STEM_CELL_UP	2.30E-02	1.20E-01	1.30E+01	1.86E-01	8.55E-01
LIM_MAMMARY_LUMINAL_MATURE_DN	8.88E-02	1.40E-01	1.30E+01	6.34E-01	5.37E-01
LIM_MAMMARY_LUMINAL_PROGENITOR_UP	1.30E-01	1.48E-01	1.30E+01	8.79E-01	3.95E-01

Table 3 A.2 Top 100 differentially expressed genes between AA and EA mammary cells

Top 100 Luminal DEGs				**Top 100 Myoepithelial DEGs**			
Gene	**log2FC**	**PValue**	**FDR**	**Gene**	**log2FC**	**PValue**	**FDR**
RP11-251G23.2	6.14	2.62E-06	5.41E-02	MTND4P24	6.14	2.99E-09	6.21E-05
MTND4P24	5.31	6.75E-06	6.97E-02	RP11-673C5.1	2.68	7.29E-08	7.57E-04
PSG4	-5.59	5.41E-05	3.72E-01	HMGB1P5	2.34	1.25E-05	8.69E-02
FAM153A	3.97	1.14E-04	5.40E-01	RP11-733O18.1	1.59	1.89E-05	9.81E-02
ATAD3C	-1.33	1.31E-04	5.40E-01	MTND4P12	3.24	2.55E-05	1.06E-01
RP11-673C5.1	2.87	3.84E-04	1.00E+00	EEF1GP1	-4.96	7.66E-05	2.65E-01
FAM21B	-3.15	4.82E-04	1.00E+00	KRTAP2-3	-5.04	9.78E-05	2.90E-01
RP11-395N3.1	-4.16	4.92E-04	1.00E+00	TNFAIP3	-1.98	1.40E-04	3.64E-01
RP11-254B13.3	3.39	5.95E-04	1.00E+00	BMS1P8	-4.65	2.49E-04	5.75E-01
FAM154B	2.76	5.99E-04	1.00E+00	ABC7-42389800N19.1	1.90	3.47E-04	7.22E-01
RPS26P3	-2.75	6.64E-04	1.00E+00	PNMA2	-3.29	5.06E-04	8.99E-01
HSP90AA6P	-2.40	7.34E-04	1.00E+00	RP11-305B6.1	-1.67	5.40E-04	8.99E-01
HMGB1P5	2.50	8.36E-04	1.00E+00	CILP	-3.45	6.02E-04	8.99E-01

ABLIM3	-1.37	9.98E-04	1.00E+00	RPS3AP49	-2.57	6.16E-04	8.99E-01
CRYBB2	5.58	1.32E-03	1.00E+00	BMS1P16	-4.73	6.65E-04	8.99E-01
AC006273.7	3.08	1.41E-03	1.00E+00	RP11-160C18.2	-4.75	7.14E-04	8.99E-01
NBPF14	-1.75	1.51E-03	1.00E+00	POMZP3	-1.53	9.99E-04	8.99E-01
AC144652.1	4.15	1.53E-03	1.00E+00	IL2RG	-3.13	1.15E-03	8.99E-01
SEMA6A	0.92	1.54E-03	1.00E+00	MTATP6P1	1.15	1.19E-03	8.99E-01
TNFRSF11B	-1.76	1.77E-03	1.00E+00	SLC6A8	0.77	1.20E-03	8.99E-01
TTN-AS1	1.97	1.98E-03	1.00E+00	MT-TH	-1.43	1.26E-03	8.99E-01
RP11-395N3.2	-2.16	2.10E-03	1.00E+00	TMEM229B	-1.92	1.26E-03	8.99E-01
Z73979.1	-2.68	2.37E-03	1.00E+00	MEF2C	2.31	1.33E-03	8.99E-01
HIP1R	0.78	2.53E-03	1.00E+00	CTC-529P8.1	-1.36	1.34E-03	8.99E-01
KCTD8	-4.96	2.57E-03	1.00E+00	D2HGDH	-1.05	1.35E-03	8.99E-01
KIT	1.10	2.60E-03	1.00E+00	GRIA4	2.33	1.38E-03	8.99E-01
SAT1	1.03	2.81E-03	1.00E+00	PLBD1	-4.11	1.47E-03	8.99E-01
GSTT2	-3.22	3.08E-03	1.00E+00	XYLT1	-1.50	1.49E-03	8.99E-01
TDRD12	-5.48	3.24E-03	1.00E+00	ITGB1P1	-1.85	1.58E-03	8.99E-01
RP11-303E16.7	-3.31	3.28E-03	1.00E+00	RP11-269C4.1	-3.52	1.59E-03	8.99E-01
RP11-785H5.1	-0.87	3.36E-03	1.00E+00	DEPTOR	1.97	1.76E-03	8.99E-01
ZBTB8OSP2	4.41	3.56E-03	1.00E+00	CTD-2215L10.1	3.92	1.85E-03	8.99E-01
PELI2	1.18	3.66E-03	1.00E+00	RPL7P23p	-1.62	1.92E-03	8.99E-01
CHCHD2P9	-3.07	3.93E-03	1.00E+00	CELF2	-1.08	1.95E-03	8.99E-01
AFAP1L1	-1.34	4.05E-03	1.00E+00	HMGB1P8	-1.33	1.97E-03	8.99E-01
SMPD1	-0.89	4.07E-03	1.00E+00	MT-TS2	-0.90	1.97E-03	8.99E-01
C11orf54	0.70	4.08E-03	1.00E+00	MIR29B1	-1.71	1.99E-03	8.99E-01
MTND1P23	8.18	4.09E-03	1.00E+00	AC079150.2	-4.14	2.02E-03	8.99E-01

PLD1	1.02	4.11E-03	1.00E+00	AC132192.1	3.52	2.15E-03	8.99E-01
GPR55	-3.88	4.15E-03	1.00E+00	PXK	1.24	2.16E-03	8.99E-01
RP4-564F22.5	4.18	4.35E-03	1.00E+00	CRYBB2P1	1.42	2.20E-03	8.99E-01
KRT121P	-2.68	4.53E-03	1.00E+00	NXPH4	2.36	2.23E-03	8.99E-01
CTB-147C22.8	-1.87	4.93E-03	1.00E+00	BCYRN1	-1.17	2.24E-03	8.99E-01
TFEB	-0.83	5.05E-03	1.00E+00	SLC15A2	-3.03	2.27E-03	8.99E-01
ZGPAT	-2.50	5.08E-03	1.00E+00	SLC7A11-AS1	5.29	2.28E-03	8.99E-01
BMP7	3.61	5.14E-03	1.00E+00	RP11-519G16.2	-4.88	2.30E-03	8.99E-01
DPP4	-1.88	5.31E-03	1.00E+00	CTC-484P3.1	-4.06	2.34E-03	8.99E-01
SULT1A1	0.93	5.33E-03	1.00E+00	RP3-417G15.1	-1.16	2.34E-03	8.99E-01
LUZP2	-2.53	5.49E-03	1.00E+00	MBNL2	0.75	2.40E-03	8.99E-01
TFPI2	-2.05	5.54E-03	1.00E+00	RP11-1036F1.1	-3.28	2.44E-03	8.99E-01
KLK6	-1.27	5.55E-03	1.00E+00	AC007292.1	-4.38	2.45E-03	8.99E-01
GPR83	-3.71	5.57E-03	1.00E+00	NBPF14	-2.60	2.46E-03	8.99E-01
COL22A1	-1.87	5.65E-03	1.00E+00	TRIM22	-1.54	2.49E-03	8.99E-01
CTA-390C10.10	-3.97	5.65E-03	1.00E+00	C17orf97	2.23	2.52E-03	8.99E-01
RP11-326G21.1	-2.10	5.91E-03	1.00E+00	RP11-429J17.2	-1.38	2.59E-03	8.99E-01
RP11-545M17.1	3.73	5.96E-03	1.00E+00	AP000688.14	5.39	2.62E-03	8.99E-01
RP1-102K2.6	-3.08	5.96E-03	1.00E+00	HMGN1P37	-3.42	2.66E-03	8.99E-01
EI24P2	-4.04	5.96E-03	1.00E+00	AC091492.2	-4.01	2.67E-03	8.99E-01
RP11-326A19.5	-3.69	5.99E-03	1.00E+00	RP11-297L17.6	-1.15	2.75E-03	8.99E-01
SRIP3	-4.12	6.01E-03	1.00E+00	DBIP1	-3.82	2.78E-03	8.99E-01
MILR1	-3.51	6.03E-03	1.00E+00	RP11-111F16.2	-3.47	2.80E-03	8.99E-01
AGAP7	-2.98	6.07E-03	1.00E+00	HMGXB3	0.75	2.82E-03	8.99E-01
SNX6P1	-3.76	6.24E-03	1.00E+00	RIN2	-0.72	2.87E-03	8.99E-01

COL6A2	-1.02	6.30E-03	1.00E+00	RP11-554E23.4	4.63	2.95E-03	8.99E-01
SNORD3B-1	4.08	6.46E-03	1.00E+00	GAREML	-1.04	2.99E-03	8.99E-01
PLXDC1	4.22	6.50E-03	1.00E+00	C16orf72	-0.89	3.12E-03	8.99E-01
C2orf71	-3.47	6.79E-03	1.00E+00	RP11-361H10.3	-4.07	3.19E-03	8.99E-01
RNA5SP283	-3.48	6.87E-03	1.00E+00	CDH8	-3.39	3.25E-03	8.99E-01
XXbac-BPG181M17.6	-3.48	6.87E-03	1.00E+00	TBC1D17	1.15	3.26E-03	8.99E-01
RP11-764K9.4	-1.54	7.03E-03	1.00E+00	ADAMTS6	-1.46	3.27E-03	8.99E-01
FAM195B	-1.83	7.09E-03	1.00E+00	CDV3P1	-4.16	3.40E-03	8.99E-01
RNA5SP248	-3.48	7.36E-03	1.00E+00	XXbac-BPG252P9.9	2.67	3.42E-03	8.99E-01
KIRREL3	-3.70	7.41E-03	1.00E+00	BANK1	1.98	3.44E-03	8.99E-01
PRRG3	-1.67	7.65E-03	1.00E+00	RP11-69E11.4	1.99	3.64E-03	8.99E-01
RP5-968J1.1	3.05	7.65E-03	1.00E+00	STARD4-AS1	2.18	3.73E-03	8.99E-01
MET	-0.58	7.96E-03	1.00E+00	RPS7P4	-3.23	3.78E-03	8.99E-01
AC005003.1	1.59	7.98E-03	1.00E+00	RP11-75L1.2	-1.34	3.80E-03	8.99E-01
RP11-181C3.1	2.89	8.02E-03	1.00E+00	SMN2	-1.08	3.85E-03	8.99E-01
POLD4	-1.62	8.03E-03	1.00E+00	TMEM139	-1.17	3.88E-03	8.99E-01
hsa-mir-6723	3.32	8.07E-03	1.00E+00	IFITM1	2.42	3.97E-03	8.99E-01
HIST1H4A	3.67	8.15E-03	1.00E+00	JAKMIP3	3.17	4.04E-03	8.99E-01
RP11-494K3.2	-4.30	8.41E-03	1.00E+00	STAU2-AS1	-4.45	4.06E-03	8.99E-01
RP11-462L8.1	2.31	8.60E-03	1.00E+00	XXyac-YRM2039.2	-1.85	4.11E-03	8.99E-01
MFI2	-0.68	8.61E-03	1.00E+00	RPL35P5	-1.89	4.14E-03	8.99E-01
RP11-537I16.2	3.22	8.85E-03	1.00E+00	FCHO1	4.35	4.21E-03	8.99E-01
DIP2A-IT1	2.50	8.88E-03	1.00E+00	SERBP1P1	-1.52	4.23E-03	8.99E-01
CLIC2	-3.19	9.02E-03	1.00E+00	TMEM97	1.05	4.27E-03	8.99E-01

RP11-449P15.2	-2.13	9.28E-03	1.00E+00	RNA5SP252	-3.69	4.31E-03	8.99E-01
ANKRD36B	0.84	9.28E-03	1.00E+00	C16orf45	1.49	4.41E-03	8.99E-01
SLC6A20	2.54	9.31E-03	1.00E+00	PSG4	-6.27	4.43E-03	8.99E-01
FAM153C	2.47	9.49E-03	1.00E+00	CTB-13H5.1	-1.30	4.49E-03	8.99E-01
S1PR3	-1.49	9.54E-03	1.00E+00	RP11-297C4.1	3.76	4.50E-03	8.99E-01
LRRFIP1P1	1.10	9.61E-03	1.00E+00	RNA5SP248	-3.08	4.52E-03	8.99E-01
RPL35P3	-2.71	9.77E-03	1.00E+00	AC079354.3	-4.01	4.64E-03	8.99E-01
CTC-228N24.1	-3.25	9.86E-03	1.00E+00	NBPF15	-1.25	4.68E-03	8.99E-01
CTD-2555O16.2	-3.25	9.86E-03	1.00E+00	RNA5SP195	-2.78	4.68E-03	8.99E-01
PPARGC1A	1.89	9.93E-03	1.00E+00	NPM1P21	-2.11	4.71E-03	8.99E-01
RP11-390D11.2	-3.26	9.98E-03	1.00E+00	RP11-251G23.2	7.25	4.72E-03	8.99E-01
CASC14	-3.26	9.98E-03	1.00E+00	RNU4-80P	-3.25	4.75E-03	8.99E-01
RP11-347P5.1	-3.09	1.00E-02	1.00E+00	TMEM156	-3.76	4.77E-03	8.99E-01

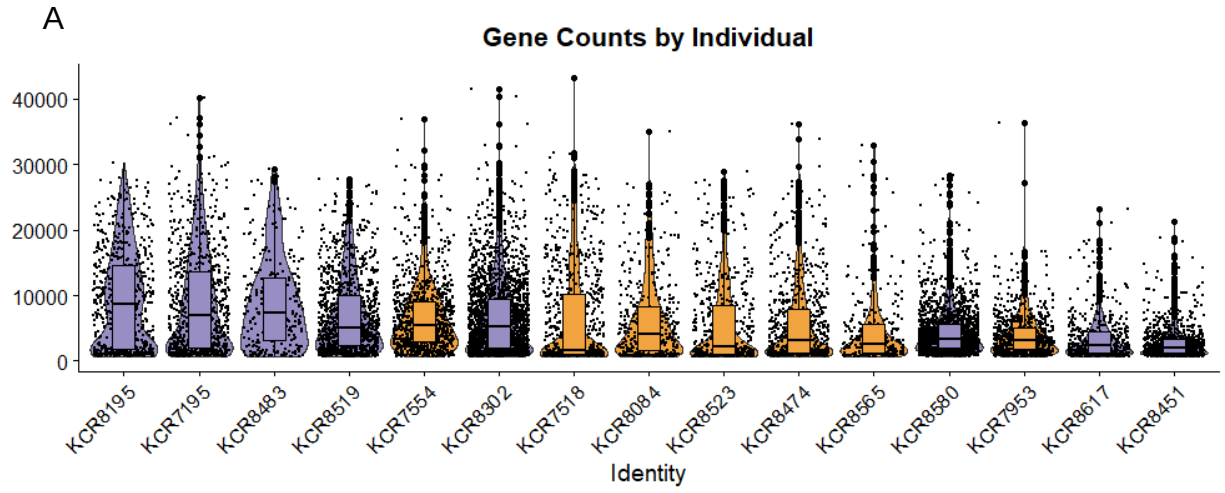


Figure 3.A5. Quantification of gene counts for cells across individual samples. Samples are colored by race and ranked from high to low average gene counts per individual.

Table 3A.3. Differentially methylated CpG sites for selected differentially expressed genes and their annotations

Gene	Name	pct.meth.change	DNAm.pvalue	Islands_Name	Relation_to_Island
AKR7L	cg18202521	24.60	7.15E-03	chr1:19600249-19600942	Island
AKR7L	cg13935437	29.85	1.07E-02	chr1:19600249-19600942	Island
AKR7L	cg11376198	26.55	1.28E-02	chr1:19600249-19600942	Island
AKR7L	cg20677058	16.68	4.52E-02	chr1:19600249-19600942	Island
AKR7L	cg12798157	22.67	1.56E-02	chr1:19600249-19600942	Island
AKR7L	cg09045262	15.75	2.95E-02	chr1:19600249-19600942	Island
GSTM1	cg16180556	-10.51	3.47E-02	chr1:110230238-110230614	Island
PRDM16	cg14761019	-22.59	1.08E-02	chr1:3028179-3028557	Island
PRDM16	cg23022057	-18.89	2.05E-02	chr1:3351365-3351632	S_Shore
PRDM16	cg22730864	-18.69	1.69E-02	chr1:3163969-3164643	N_Shelf
PRDM16	cg01713250	-18.45	2.09E-03		OpenSea
PRDM16	cg04423188	-17.69	9.94E-03	chr1:3327991-3329424	Island
PRDM16	cg02481237	-16.61	6.47E-03	chr1:3071899-3072239	N_Shore
PRDM16	cg26846424	-16.24	8.45E-03	chr1:3310102-3311035	Island
PRDM16	cg12096707	-15.97	1.55E-04	chr1:3071899-3072239	S_Shore
PRDM16	cg18381051	-15.84	3.94E-02	chr1:2983925-2987962	Island
PRDM16	cg06748955	-15.52	1.89E-02		OpenSea
PRDM16	cg26407161	-15.29	1.48E-02	chr1:3028179-3028557	S_Shore
PRDM16	cg06970772	-15.18	4.37E-04	chr1:3275029-3275284	N_Shelf
PRDM16	cg10021614	-15.10	1.41E-02	chr1:3028179-3028557	S_Shore
PRDM16	cg02404410	-15.05	2.74E-03	chr1:3163969-3164643	N_Shelf
PRDM16	cg24514678	-14.83	1.99E-02	chr1:3147845-3148081	N_Shelf
PRDM16	cg06911744	-14.10	4.02E-02	chr1:3275029-3275284	S_Shore
PRDM16	cg22726349	-13.45	4.83E-03	chr1:2990030-2990718	Island
PRDM16	cg12599275	-13.42	2.59E-03		OpenSea
PRDM16	cg19243842	-13.39	1.60E-02	chr1:3059050-3059268	N_Shore
PRDM16	cg27444751	-12.90	1.67E-02	chr1:3028179-3028557	S_Shore
PRDM16	cg12884780	-12.67	6.06E-03	chr1:3080934-3081292	N_Shelf
PRDM16	cg01071314	-11.73	2.47E-02	chr1:3275029-3275284	S_Shelf
PRDM16	cg01602345	-11.69	2.66E-02	chr1:3275029-3275284	S_Shelf
PRDM16	cg01062116	-11.61	1.65E-02	chr1:3351365-3351632	S_Shore
PRDM16	cg01593475	-11.39	3.51E-02	chr1:3203765-3204453	S_Shelf
PRDM16	cg11229543	-11.00	2.30E-02	chr1:2983925-2987962	Island
PRDM16	cg09077530	-10.98	1.25E-02		OpenSea
PRDM16	cg14706825	-10.98	4.44E-02		OpenSea
PRDM16	cg08876274	-10.97	1.89E-02	chr1:3028179-3028557	Island
PRDM16	cg14030836	-10.95	2.15E-04	chr1:2990030-2990718	S_Shore
PRDM16	cg26300135	-10.87	3.08E-02	chr1:3321269-3322310	S_Shore

PRDM16	cg10588310	-10.66	5.66E-03	chr1:3331986-3332227	Island
PRDM16	cg14659814	-10.59	4.14E-02		OpenSea
PRDM16	cg08110058	-10.39	3.88E-02	chr1:3071899-3072239	S_Shore
PRDM16	cg22396632	-10.30	7.29E-03	chr1:3080934-3081292	N_Shore
PRDM16	cg15059176	-10.18	2.73E-02	chr1:3307069-3307285	N_Shore
PRDM16	cg16011907	-10.07	4.21E-02		OpenSea
PRDM16	cg21604357	-10.05	2.66E-02	chr1:3028179-3028557	N_Shore

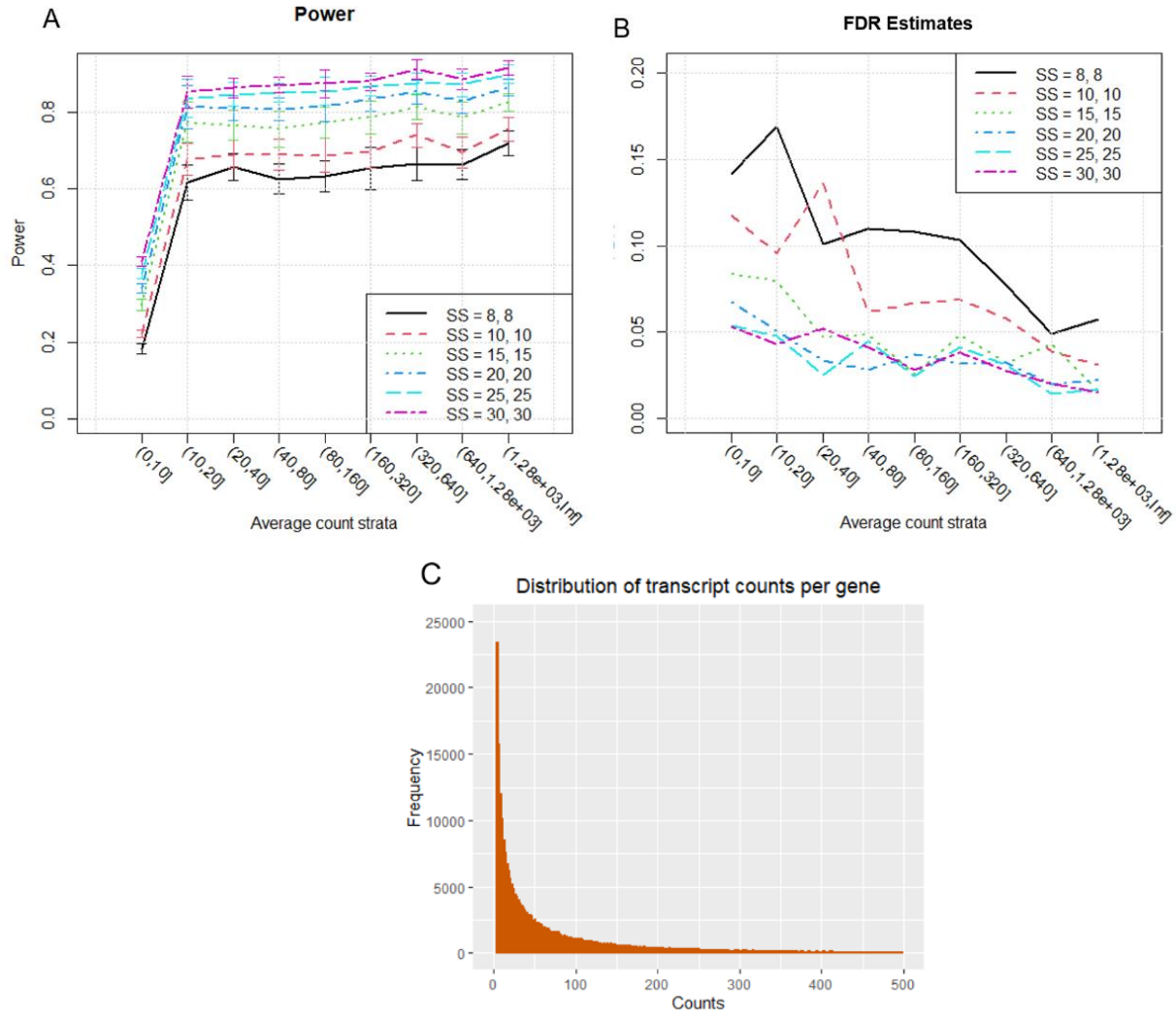


Figure 3.A6. Power simulation for sample size A) Power analysis for sample size groups based on number of counts **B)** FDR estimates for sample size and counts **C)** Distribution of counts per gene

Chapter 4

Characterizing Inter-individual Differences in Transcriptomic Response to Bisphenol Exposure of Normal Mammary Cells

Introduction

Every year, the number of environmental chemicals in production and circulation increases, yet the experimental evaluation of the risk they pose on human health progresses at a much slower rate. A growing body of evidence points to inter-individual differences in response and susceptibility to these toxicants, highlighting the need for experimental models of precision toxicology able to capture gene-environment interactions and their effects on disease outcomes.

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death among women worldwide.[1,147] Although approximately 10% of breast cancer diagnoses are attributable to hereditable risk factors, such as BRCA mutations, the vast majority (70-90%) of cases occur in women with no family history, suggesting that extrinsic factors are primarily responsible for lifetime cancer risk.[37] The implications of this for public health intervention are hopeful—environmental factors are characterizable and modifiable, and thus present tangible targets for breast cancer prevention.

AA individuals experience disproportionate exposure to environmental toxicants such as air pollution, beauty products, and endocrine disrupting chemicals, all of which have been linked to human disease.[46,148,149] Additionally, human biomarker analyses in the National Health and Nurses Examinates Study (NHANES) provide quantitative measures for race associated differences in exposure doses as well as toxicological response to chemicals.[150,151]

Of the chemicals AA women are disproportionately exposed to, Bisphenol-A (BPA) and its analogues are of interest due to their ability to alter mammary gland morphogenesis and stemness *in vivo* and *in vitro.*[47,152–154] Bisphenols are ubiquitously used in the manufacturing of plastics and epoxy resins used in food packaging, kitchenware, coatings of cans and jars, dental

materials, toys, thermal paper, and more.[155,156] Due to this widespread use, common routes of exposure include oral, inhalation, and transdermal.[156] BPA and BPS have been measured in human urine and breast milk at concentrations ranging from 1nM to 6uM and MCF-7 ER positive cancer cells have been observed to exhibit alterations in stem-related genes when exposed to concentrations as low as 10nM of BPA.[47,157,158] Classified as xenoestrogens due to their ability to bind and activate the estrogen receptor (ER), bisphenols have been implicated in carcinogenesis, however their mechanism of action remains unclear. Studies characterizing bisphenol exposure of normal mammary cells are limited, thus whether or not bisphenols contribute to carcinogenic transformation of normal mammary cells remains to be seen. Additionally, animal and cell line models are unable capture inter-individual differences in response to bisphenol exposure. Given the ubiquitous and wide dose range of human exposure to bisphenols, characterizing inter-individual differences in response to bisphenol exposure is imperative to our understanding of their health risks.

In this study we aim to characterize the effects of BPA, BPS, and BPF on normal mammary cells from diverse individuals across a range of human relevant doses. We leverage high throughput single cell RNA-seq to gain insight into transcriptomic alterations elicited by bisphenol, gain insight into their mechanisms of action, and test if this is a viable model of the normal breast for precision toxicology that can be expanded to other chemicals.

Materials and Methods

Sample acquisition and selection

This experiment utilized normal mammary samples acquired from the Susan G. Komen normal tissue bank which had been previously established and expanded by conditional reprogramming (CR) culture in Chapter 3.[111] 6 samples (n=3 AA, n=3 EA) which were closely matched by age (AA= 25.6 (1.2), EA= 25.0 (1.7)) , BMI (AA =27.7 (7.6), EA = 27.1 (6.6)), and days since last period (AA= 8.6 (1.5), EA= 15.3 (4.2)) were selected to be used for this experiment.

Bisphenol dose preparation

Bisphenol-A (Sigma-Aldrich cat.no. 239658, lot no. MKCD7508), Bisphenol-S (Sigma-Aldrich cat. no. 103039, lot. no. MKCL2824) and Bisphenol-F (Alfa Aesar cat.

no.AAA1141703, lot. no. 10211585) 10mM stocks were made by dissolving each chemical in DMSO. The 10mM stocks were frozen at -20°C for storage. Intermediate stocks of 5mM, 2.5mM, 160µM, 80 µM , 5µM, and 2.5 µM were made for each chemical by serial dilution with DMSO and F-media the week of the experiment and frozen at -20°C. The chemical doses used on the cells were made fresh the day of dosing by diluting the intermediate stocks with F-media to final concentrations of 25 µM, 0.8 µM, and 0.025 µM, with a final DMSO content of 0.05%. Controls were made by diluting DMSO with F-media to a final concentration of 0.05% DMSO.

Conditional reprogramming culture and dosing

Conditionally reprogrammed cells from each sample (KCR8195, KCR8580, KCR8519, KCR7889, KCR7953, KCR7518) were thawed from cryopreservation and plated individually in CR culture. All samples were plated at passage 1. Once confluence was reached (3-4 days), cells were differentially trypsinized (Gibco 0.05% Trypsin/EDTA), counted, and viability was assessed using acridine orange/propidium iodide staining (LUNA FL Dual Fluorescence Cell Counter). From the cell counts, 650,000 cells from each individual were pooled and the pooled sample containing a mix of cells from all 6 individuals was split and plated into 12 T-75 flasks of 300,000 cells, one flask for each dose (0 µM, 0.025 µM, 0.8 µM, 25µM) of each chemical (BPA, BPS, BPF). After growing in normal CR conditions for 48 hours, each flask was dosed with its designated F-media including one of the bisphenols or DMSO control and grown for another 48 hours.

Sample collection and multiplexing

On the day of sample collection, wash buffers (PBS + 0.04% BSA and PBS + 0.1% BSA, Gibco) for cell multiplexing were prepared fresh. All 12 flasks were differentially trypsinized (Gibco 0.05% Trypsin/EDTA), centrifuged (500gs for 5 minutes), and resuspended in wash buffer. Following the manufacturer's specifications for CellPlex preparation of cell lines, each of the 12 samples was incubated with a unique feature barcode oligonucleotide for multiplexing. (10x Genomics). Once incubation, washing, and oligo tagging was complete, concentration and viability was quantified for each of the 12 samples. 117,000 cells were taken from each sample and pooled, and concentration and viability of pooled sample was measured. The concentration of the pooled sample was adjusted for a final target concentration of 1,300-1,600 cells/µL for targeted cell recovery of 10,000-30,000 cells per Chromium Next GEM Chip G run. The pooled

sample was transported on ice to the University of Michigan Advanced Genomics Core for sequencing.

Single cell RNA-sequencing and library construction

Oligo tagged multiplexed cells were run across two Chromium Next GEM G Chips core using the manufacturer's protocol and reagents. Briefly, CellPlex labeled cells contained within the 10x Master Mix were loaded onto the Chromium chip along with partitioning oil and gel beads in order to generate Gel Beads-in-Emulsion (GEMs). Each GEM consists of an oil droplet encapsulating a CellPlex labeled cell and a barcoded gel bead. Once GEMS were generated, the encapsulated cells were lysed and gel beads were dissolved, releasing primers. Within the GEM, the poly (dT) primer for cDNA generation and the Capture Sequence 2 primer for multiplexing libraries were simultaneously engaged in separate reactions in order to generate cDNA from mRNA and barcoded DNA from the CMO barcodes. cDNA was PCR amplified and multiplexing were constructed. Sample libraries were run on a NovaSeq 6000 DNA sequencer (Illumina).The 10x genomics CellRanger Multi pipeline was used to analyze the generated single cell gene expression and multiplexing libraries. Briefly, "cellranger mkfastq" was run to generate demultiplexed FASTQ files. "Cellranger count" was run on each GEM well that was demultiplexed to generate feature barcode libraries for CMOs and gene expression libraries for each cell. The output gene expression and barcode matrices were then further analyzed downstream for QC and cell selection.

Single cell data analysis

Sample deconvolution using Demuxlet: To identify which cells from the sequenced cell pools derived from which sample, we used the genetic demultiplexing tool Demuxlet.[159] Demuxlet assigns cells to individual samples based on a genetic reference and polymorphisms called from the single cell RNA-seq data. From the CellRanger Multi output, we analyzed the aligned read BAM files for each CellPlex pool using the TOPMED imputed genome VCF files. We used the settings "r2-info 0.8", "doublet-prior 0.05", "alpha 0", and "alpha 0.5". After running Demuxlet, the "best guess" for the Sample ID and "droplet type" were added to the metadata of the gene expression matrix.

QC and high quality cell selection: The R package Seurat[112,113] was used for all of the initial QC and cell selection. Individual gene expression matrixes from the 12 multiplexed samples were merged into a singular Seurat object. Cells were filtered and selected if they met QC criteria of less than 25% mitochondrial genes and gene counts (nFeature_RNA) between 200 and 7500 genes. Additionally, cells were further filtered based on the droplet type identified by Demuxlet and selected if they were identified as singlets. Once high quality cells were selected, they were normalized using the "LogNormalize" method in Seurat and highly variable features were identified. Data was then scaled by regressing on percent mitochondrial genes. Linear dimension reduction was run on the scaled data using variable genes to identify the top principle components (PCs).

Unbiased clustering and cell type identification: Unbiased clustering was performed using the top 10 PCs at a resolution of 0.075. Non-linear dimension reduction (UMAP) was also performed using the top 10 PCs. Cell type markers were used to identify myoepithelial (*KRT5/KRT14*) and luminal (*KRT8/KRT18*) cells which separated into two clear clusters at a resolution of 0.075.

Differential gene expression analysis: Cells were split by cell type (luminal and myoepithelial) and differential gene expression (DGE) analysis was performed using the Pseudobulk method.[114] For each cell type, pseudobulk samples were generated by aggregating across the variables "Sample", "Treatment", and "Dose". Pseudobulk samples with low counts (<20 cells) and genes with low counts (min.total.count =5) were filtered out. Composition biases were corrected for by computing normalization factors with the trimmed mean of M-value. Multi-dimensional scaling, negative binomial dispersions, and quasi-likelihood dispersions were performed to calculate mean and per-gene variance. Generalized linear models (GLM) were fit to the counts for each gene and differential gene expression testing was performed between controls and each dose (0.025µM, 0.8 µM, 25 µM) for each chemical (BPA, BPS,BPF).

CytoTRACE: The Cellular (Cyto) Trajectory Reconstruction Analysis using gene Counts and Expression (CytoTRACE)[115] package in R was used to predict the differentiation state of cells. CytoTRACE leverages the number of detectably expressed genes per cell as a read out of developmental potential. Each individual cell was assigned a CytoTRACE score where cells with

a higher number of genes per cell were considered more transcriptionally diverse, which correlates with a more "stem cell" like gene expression profile.

Hybrid Analysis: Luminal/myoepithelial hybrid cells were identified by extracting the gene counts for marker genes *KRT5* and *KRT14*. Cells in the upper 50% quantile for each gene were considered "highly expressed" and cells which highly co-expressed both *KRT5* and *KRT14* were labeled as hybrid cells.[111] The "FindMarkers" function in Seurat was used to perform differential expression testing using the Wilcoxan rank sum test to identify marker genes distinguishing the hybrid cells from all other cells in order to identify a hybrid cell signature.

Single cell Gene Set Enrichment Analysis: The escape package in R was used to perform single cell gene set enrichment analysis (GSEA),[116] assigning each individual cell an enrichment score for relevant gene sets. Gene sets were pulled from the Hallmark and C2 libraries from the Molecular Signatures Database (MSigDB)[117] and enrichment was performed using escape.

Cell Cycle Scoring and Regression :Cell cycle phase scores were generated for each cell using the Seurat cell cycle scoring pipeline.[118] Scores were then used to assign phases S or G2/M to each cell and stored in the metadata.

Benchmark Dose Analysis: To calculate dose-response relationships for individual genes and pathways and chemical treatments, we performed benchmark dose (BMD) analysis using BMDExpress2. We performed counts per million normalization from the pseudobulk aggregated counts matrices. Counts per million data were then loaded into BMDExpress2 based on chemical exposure (BPA, BPS, and BPF) across doses by cell type (luminal, myoepithelial). Genes were first filtered using analysis of variance (ANOVA) with an unadjusted p-value cutoff of 0.05 across doses. Benchmark dose analyses were conducted on the ANOVA filtered gene list using the hill, power, linear, poly 2, poly3, exponential 2 ,exponential 3, exponential 4, and exponential 5 methods. The benchmark response was a 1 standard deviation change, and the best model was selected using a nested chi-squared test followed by the lowest Akaike information criteria (AIC). Benchmark doses for pathway alterations were calculated for all Gene Ontology categories and Reactome pathways, with pathway enrichment calculated via a Fisher's Exact test. We additionally calculated enrichment for the MSigDB gene sets as described in the section above.

Results

Multiplexed Bisphenol Exposed Cells from Genetically Diverse Individuals Can Be Successfully Deconvoluted using Genetic Variation

Conditionally reprogrammed mammary cells from 6 diverse individuals (n=3 AA, n=3 EA) were cultured, pooled, and exposed to 0, 0.025, 0.8, or 25 μM of BPA, BPS, or BPF (Figure 4.1). Each of the 12 experimental conditions was tagged with a cell multiplexing oligo (CMO) prior to single cell RNA-sequencing. Following sequencing, the 12 experimental conditions were deconvoluted using Cell Ranger and the individual samples were deconvoluted using the Demuxlet algorithm and DNA extracted from Chapter 3. After sample deconvolution, we determined the cell counts for each individual, which ranged from 912 – 9629 cells per individual (**Table 4.2**). We saw more consistency in cell counts across the experimental conditions, with a total of 4970 vehicle controls and bisphenol treated cells ranging from 1268-1675 cells per experimental condition (**Table 4.3**).

Bisphenol Treated Mammary Cells Cluster by Cell Type and Vary by Individual

Unbiased clustering of all cells which met our filtering criteria showed that there was no distinct clustering by bisphenol treatment and that cells primarily cluster by luminal and myoepithelial cell type (**Figure 4.2 B-D**). Both luminal and myoepithelial clusters contain cells from all 6 individuals, however, the proportions of luminal and myoepithelial cells vary highly by individual (**Table 4.2**). Within the luminal and myoepithelial cell type clusters, there also appears to be sub-clustering by individual **(Figure 4.2A).**

Bisphenol Treated Mammary Cells Include Hybrid KRT5/KRT8 and EPCAM/VIM Hybrids Which Vary by Individual

We explored the presence of the hybrid cell types discussed in Chapters 2 and 3 and found that both *KRT5/KRT8* Luminal/Basal (L/B) and *EPCAM*/*VIM* (E/M) hybrids were present in all of our samples following exposure to bisphenols. Consistent with prior findings, L/B hybrids cluster mostly in a subsection of the luminal cell cluster, whereas E/M hybrids are in both clusters, but mostly diffuse across the luminal cluster **(Figure 4.3 A,C).** Proportions of hybrids varied by individual for both hybrid cell types **(Figure 4.3).**

Differential Gene Expression Reveals Individual Bisphenols Elicit Unique Transcriptomic Alterations Highly Variable by Cell Type and Dose

After differential gene expression was performed controlling for differences by individual, the different bisphenols showed distinct transcriptomic differences in directionality and magnitude, which were variable by cell type and dose. For upregulated genes, BPA showed the highest number of DEGs at 25µM for both cell types, BPS for 0.8µM in myoepithelial and 0.025 µM in luminal, and BPF at 0.8 µM for luminal and 0.025 µM for myoepithelial. For downregulated genes, BPF showed the highest number of DEGs for all luminal doses and 0.025 µM for myoepithelial. BPS had the most DEGs for 0.8 µM and 25 µM for myoepithelial cells. **(Figure 4.4)**.

When looking at overlaps between DEGs stratified by cell type and direction of fold change, only a small fraction of total DEGs overlapped across the conditions. The greatest number of overlaps was across 5 out of the 9 doses in the myoepithelial downregulated genes **(Figure 4.5D).** For overlaps between all three bisphenols at the highest dose 25 µM, there were 47 upregulated and 38 downregulated overlapping DEGs in luminal cells, and 30 upregulated and 45 downregulated overlapping DEGs in myoepithelial cells **(Figure 4.5 A-D).** Overall bisphenol exposure appears to induce a more heavily downregulated transcriptomic signature across doses and chemicals.

We further investigated genes which met the FDR <0.05 cutoff to characterize their magnitude and directionality across 0.8 µM and 25 µM. While transcriptomic differences remain by cell type, bisphenol, and dose, there were a few notably consistent alterations. *MUCL1* was significantly downregulated at both doses of BPA, BPF, and BPS in luminal cells. *NFKBIA* was significantly downregulated for both cell types at 0.8 µM and 25 µM for BPS and BPF, however it was not significantly different for BPA. *GREB1*, an estrogen responsive gene, was significantly upregulated at 25 µM BPF and BPS in luminal cells. *GATA3*, associated with ER positive breast cancers, was significantly upregulated at 25 µM BPA and BPF for both luminal and myoepithelial subtypes, but not for BPS.

Overall, characterizing similarities as well as differences in transcriptomic alterations induced by bisphenols stratified by cell type, provides useful insight into potential mechanisms of action and genes of interest to probe further.

Benchmark Dose Analysis and Single Cell Gene Set Enrichment Reveal BPS Elicits Significant and Distinct Transcriptomic Alterations at Low and High Doses

To extrapolate the accumulated gene alterations of bisphenols for doses spanning our 0.025-25 μM range, we performed benchmark dose analysis using the significant genes identified from the differential gene expression analysis. Accumulation, indicating the number of significantly affected genes at the corresponding dose, showed similar rates of increase for BPA and BPF in both luminal and myoepithelial cell types (**Figure 4.7**). BPS exposed myoepithelial cells exhibited the highest accumulation overall spanning all doses, however, the BPS exposed luminal cells showed the highest levels of accumulation at the lowest tested dose and below.

We also performed single cell gene set enrichment (GSEA) for a number of breast cancer, stem cell, and estrogen associated gene sets. Overall, we did not see any strikingly clear enrichment for any of the selected gene sets across chemicals, however, comparisons of the enrichment signatures across bisphenols reinforces the prior observations that BPA, BPF, and BPS elicit distinct transcriptomic alterations with some overlapping similarities (**Figure 4.8 A**).

Due to the high accumulation results of BPS from the benchmark dose analysis and a faint signal from overall GSEA, we decided to take a closer look at the enrichment of the WONG_EMBRYONIC_STEM_CELL_CORE and SMID_BREAST_CANCER_BASAL_UP gene sets across BPS doses. We see significantly different enrichment for both of these gene sets compared to control for both luminal and myoepithelial cells across all doses except 25 μM SMID in myoepithelial cells (Figure 4.8 B-E). Interestingly, at 0.025 μM BPS cells exhibited enrichment scores greater than controls, however 0.8 μM and 25 μM were significantly lower.

This non-monotonic response of BPS at low vs higher doses was also observed when looking at the overlaps between BPS DEGs at these doses and the genes in the Wong gene set. We overlapped all doses of BPS by luminal and myoepithelial subtype with the 335 genes in the Wong gene set and found the following overlaps – 0.025 μM (n=30 genes), 0.8 μM (n=4 genes), 25 μM (n=30 genes). We observed more overlaps in the myoepithelial subset with 0.025 μM (n=37 genes), 0.8 μM (n=89 genes), 25 μM (n=155 genes). None of the luminal-Wong overlapping genes were shared across all three doses. However, for the myoepithelial- Wong overlaps, 30 of the genes were differentially expressed across all 3 doses ($p<0.05$) (Table 4.4). Interestingly, 29/30 genes for 0.025 μM were upregulated, whereas for the other two doses, the

majority of genes were downregulated. A number of these genes have been linked to cell cycle and proliferation, and a few *AURKA, CDCA3, CDC20, CDCA8* have even been linked to breast cancer and estrogenic pathways.

Individuals Exhibit Distinct and Variable Transcriptomic Responses to Bisphenol Exposure

Finally, we aimed to characterize inter-individual response to bisphenol exposure and if transcriptomic alterations are individual specific. For this we selected the top 20-25 most differentially expressed genes (FDR significance) at 25 µM for each bisphenol and looked at fold change differences in the selected genes between control and 25 µM for each individual. Overall we see the most variation in inter-individual response for BPS in both luminal and myoepithelial cells, with a mixture of up and down regulated genes. Conversely, when stratified by individual, 25 µM BPF exposure exhibits a down regulated, low expressing signature across most samples. Together, this indicates differences in transcriptomic response to bisphenols by chemical and individual, and provides preliminary evidence for inter-individual chemical susceptibility.

Discussion

This study leverages a high resolution single cell dataset of bisphenol exposure multiplexed across 3 chemicals, 12 human relevant doses, and 6 genetically diverse individuals to perform integrated transcriptomic analyses characterizing their effects. From these analyses we come to the following main conclusions in our study 1) BPA, BPF, and BPS induce distinct transcriptomic alterations on mammary cells, which vary by cell type 2) The magnitude of transcriptomic alteration is highly dose and chemical dependent, and in some cases non-monotonic 3) Individuals exhibit differences in transcriptomic response to bisphenol DEGs 4) Response to bisphenol exposure in our study does not appear to strongly induce changes in canonical estrogen pathways. These findings add to the growing body of evidence that bisphenols are relevant environmental chemicals to study in relation to breast cancer risk, that they likely have both estrogenic and non-estrogenic mechanisms of action, and that analogues of BPA, which were introduced as replacements, elicit distinct, potentially worse, transcriptomic alterations on mammary cells.

We were able to characterize significant transcriptomic alterations in mammary cells, however, the majority of these DEGs did not seem to highlight a clear estrogenic signal or pathway activation. One reason for this could be that expression of *ESR1* and *ESR2* is fairly low in our cells, likely an effect of the CR culture conditions. Although we did not detect strong enrichment for estrogenic pathway activation, we did identify a number of genes associated with estrogen signaling and response. *GREB1*, an estrogen responsive gene correlated with expression in hormone responsive cancer, was found to be upregulated 25 µM for BPF and BPS in luminal cells but not myoepithelial.[160,161] Isoforms of *GREB1* have been found regulate proliferation in both ER positive and ER negative breast cancer cell lines, with increased or decreased proliferation dependent on isoform specificity.[162] *AURKA*, which is overexpressed in breast cancer and inducible by estrogen, was found to be upregulated at the lowest dose of BPS but down regulated at the two higher doses.[163,164] While others have characterized these genes as estrogen responsive in other in *vitro* and *in vivo* models, our data does not strongly suggest estrogen mediated alterations.

Non-estrogenic effects of bisphenols have been characterized by others and due to our cells not expressing the estrogen receptor, it is likely that the transcriptomic effects we observe are through non-estrogenic pathways. In cervical cancer cell lines, HeLA, siHa, and C-33A, exposure to 1uM and 1nM BPA promoted *in vitro* migration and activity of *NFkB*.[165] *NFKBIA* was significantly downregulated ($p<0.05$) at 0.8uM and 25uM for both BPS and BPF, thus the *NFkB* pathway is a promising target to explore further in our differential gene expression analysis. In estrogen negative breast cancer cell lines SUM149 and SUM190, BPA exposure activated *EGFR* and *ERK* signaling and enhanced proliferation through EGFR mediation.[166] Additionally, the author's note that *EGFR* activation in response to BPA also correlated with an increase in *GPER*, the alternate ER receptor. BPA exposure also was found to trigger tumorigenesis and progression of laryngeal cancer cells via GPER upregulation of IL-6, thus making it another promising non-estrogenic target to investigate further.[167]

The single cell resolution of this data allowed us to characterize differences by luminal and myoepithelial cell type, which adds significant computational rigor to our analysis. Our differential gene expression results show distinct transcriptomic signatures in response to bisphenols between luminal and myoepithelial cell types, suggesting that bisphenols may have

distinct mechanistic targets based on cell type. For example, genes like *GREB1* and *MUCL1* were found to be differentially expressed exclusively in the luminal cell type. *GREB1* is significantly upregulated in luminal TCGA tumors, but significantly down regulated in the HER2 positive and TNBC subtypes.[168] *MUCL1* was significantly downregulated in BPA, BPF, and BPS treated luminal cells, and luminal TCGA tumors also exhibit significant downregulation of this gene.[168] In bisphenol exposed myoepithelial cells, *AURKA*, *GATA3*, and *KIF23* are differentially expressed, but not in luminal cells. These genes are expressed most highly by triple negative TCGA tumors as well as found to promote EMT, proliferation, and self-renewal and are regulated by the *FOXM1* transcription factor.[169–173] This stratified cell type differential gene expression analysis provides greater precision for identifying and prioritizing genes of interest which in turn sheds insight into potential mechanistic differences in response to bisphenols.

Lastly, the non-monotonic dose response we observe with BPS as well as the inter-individual differences in DEG expression highlight the pressing need for improved models of precision toxicology which can model dose-dependent inter-individual differences in response to chemicals. With BPS, we observe upregulation of cell cycle and proliferation associated genes at 0.025 μM, but downregulation of these same genes at the higher doses. This non-monotonic dose response coupled with the distinct transcriptomic signatures and directionality we observe between the low and high doses suggest differential pathway activation dependent on dose of exposure.

While our experimental design allowed us to test 12 different chemical treatments across 6 individuals, these were all captured at one 48 hour time point. Future experiments could be designed to include a range of time points to gain more insight into how cells respond to bisphenols over time. Additionally, while growing pooled cells from different individuals in the same flask helps to correct for potential batch effects, it is possible that cell to cell communication occurs between cells from different samples and influences their growth. Although the same number of cells were taken from each individual sample to pool, some samples proliferated much faster than others leading to an inadequate amount of cells for analysis from certain individuals downstream. Future work to characterize the effects of growing cells from different individuals in one culture can be performed to understand and standardize this process better.

Overall, this study demonstrates the computational power of single cell technology to characterize differences in bisphenol exposure across multiple chemicals, doses, and individuals, and highlights the importance of stratified analyses to detect these differences by dose and individual. Future work will involve expanding the experimental model to test a greater range of chemicals and doses. Additionally, functional experiments and epigenetic profiling could provide crucial mechanistic insight into the effect of bisphenols on normal mammary cells.

Tables and Figures

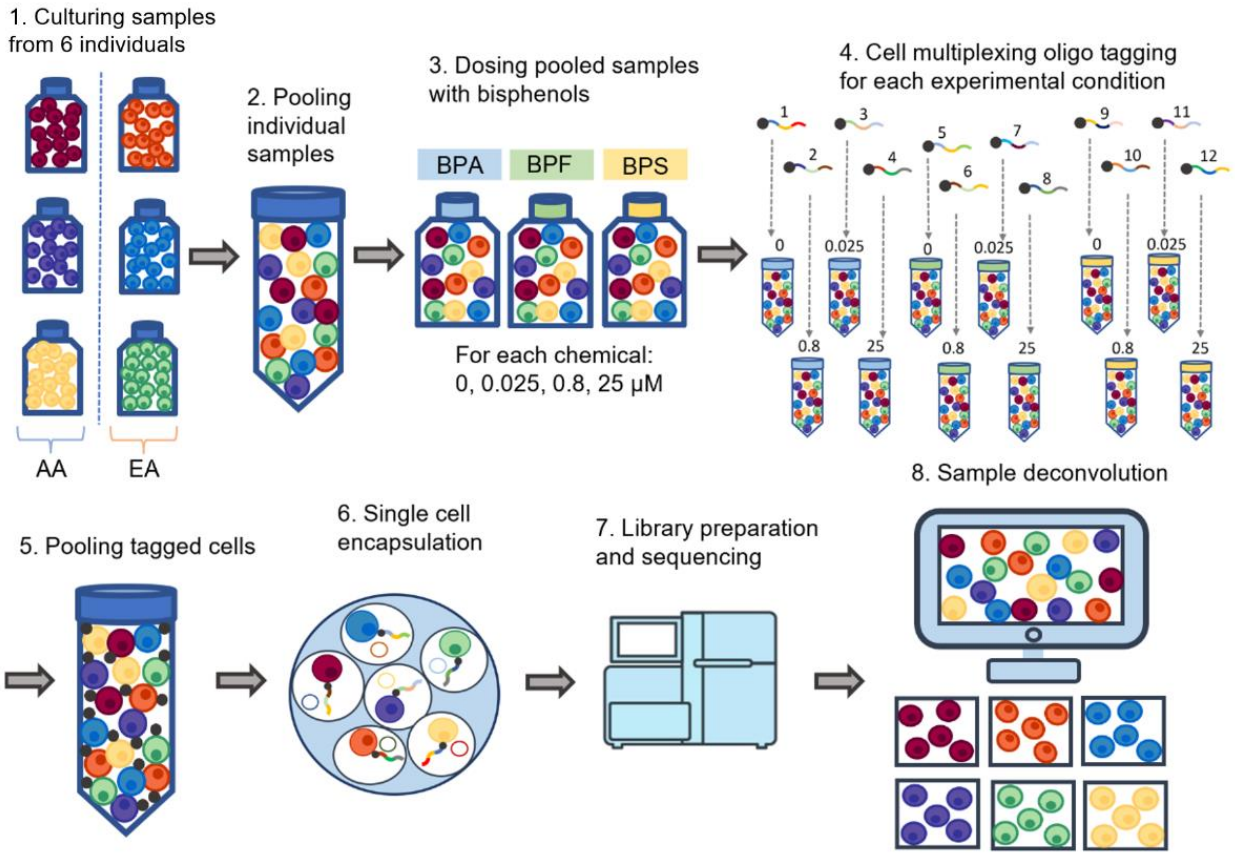


Figure 4.1. Experimental design and overview. Conditional reprogramming of AA and EA matched normal mammary samples, bisphenol exposure and cell multiplexing single cell RNA sequencing.

Table 4.1. Overview of sample demographics.

	African American Mean (s.d.), n=3	European American Mean (s.d.), n=3
Age	25.6 (1.2)	25.0 (1.7)
BMI (kg/m2)	27.7 (7.6)	27.1 (6.6)
Days since last period	8.6 (1.5)	15.3 (4.2)

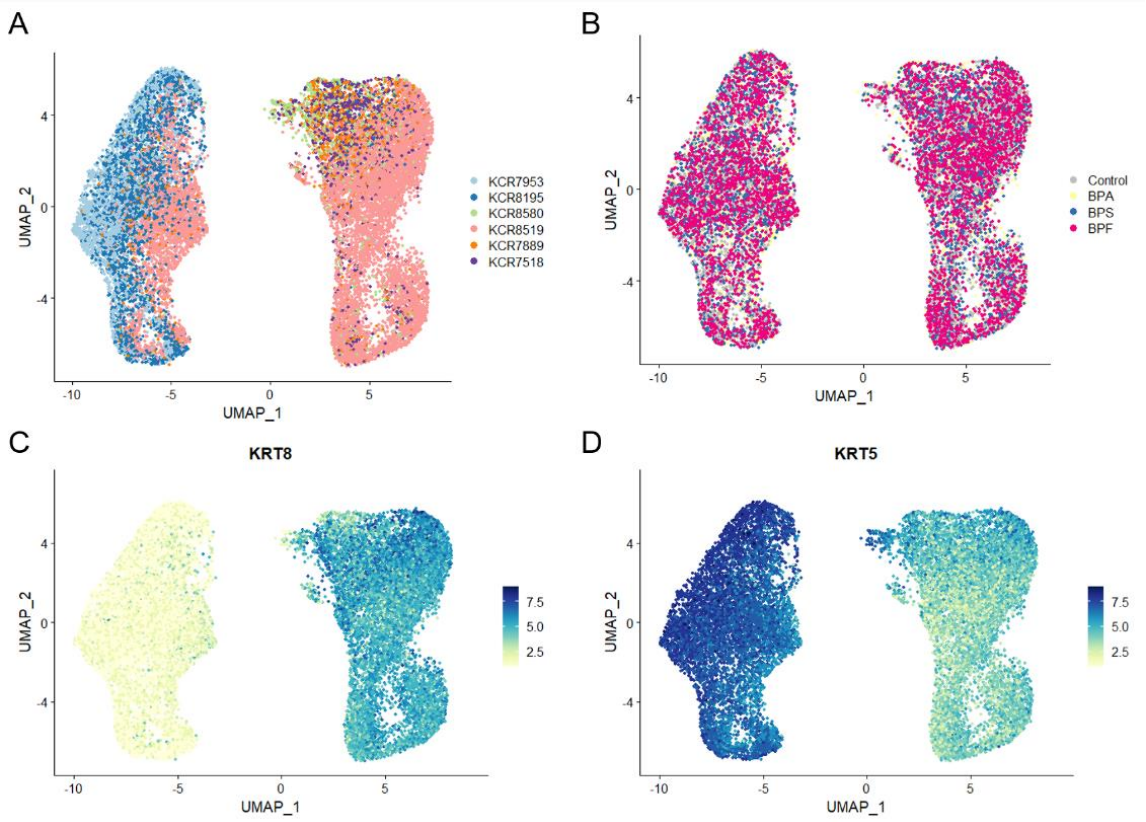


Figure 4.2. Unbiased clustering by UMAP at 0.075 resolution A) Labeled by sample ID B) Labeled by bisphenol treatment C) FeaturePlot of luminal cell type marker *KRT8* D) FeaturePlot of myoepithelial cell type marker *KRT5*

Table 4.2. Cell counts by sample and cell type

	Sample	Lum	Myo	Total Cells
AA	KCR8195	76	2795	2871
	KCR8519	7099	2530	9629
	KCR8580	846	66	912
EA	KCR7518	1102	29	1131
	KCR7889	842	239	1081
	KCR7953	4	2422	2426

Table 4.3. Cell counts by treatment and dose

Treatment	Dose		
	0.025	0.8	25
BPA	1575	1251	1656
BPF	1268	1381	1675
BPS	1296	1457	1521
Control	4970		

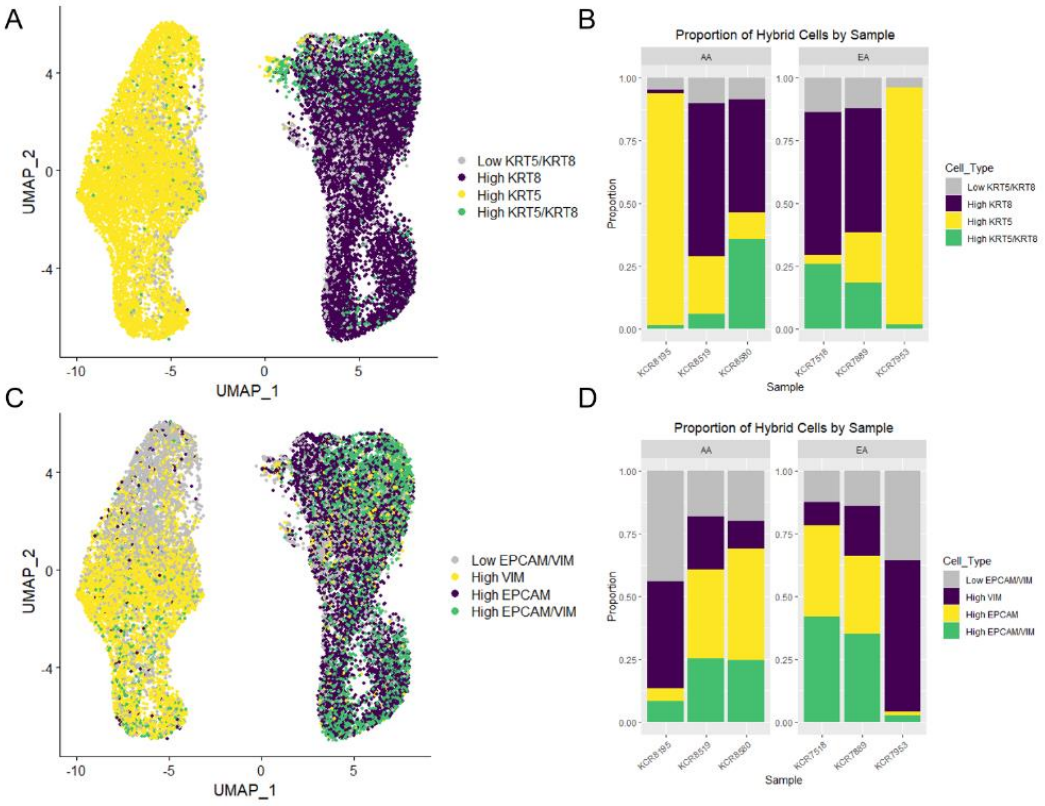


Figure 4.3. Clustering and proportions of hybrid cells A) Clustering of *KRT5/KRT8* hybrid cells ID B) Proportion of *KRT5/KRT8* hybrids by sample C) Clustering of *EPCAM/VIM* hybrid cells D) Proportion *EPCAM/VIM* hybrids by sample

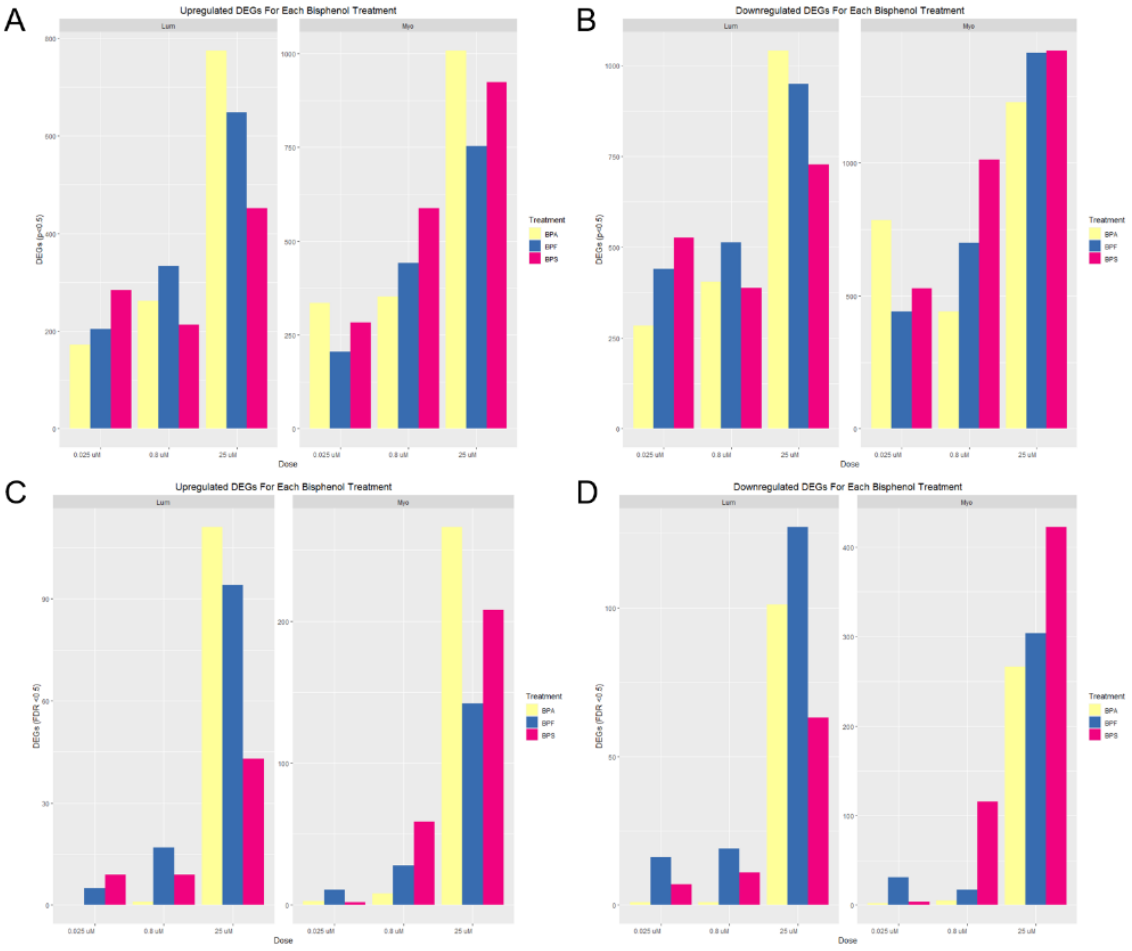
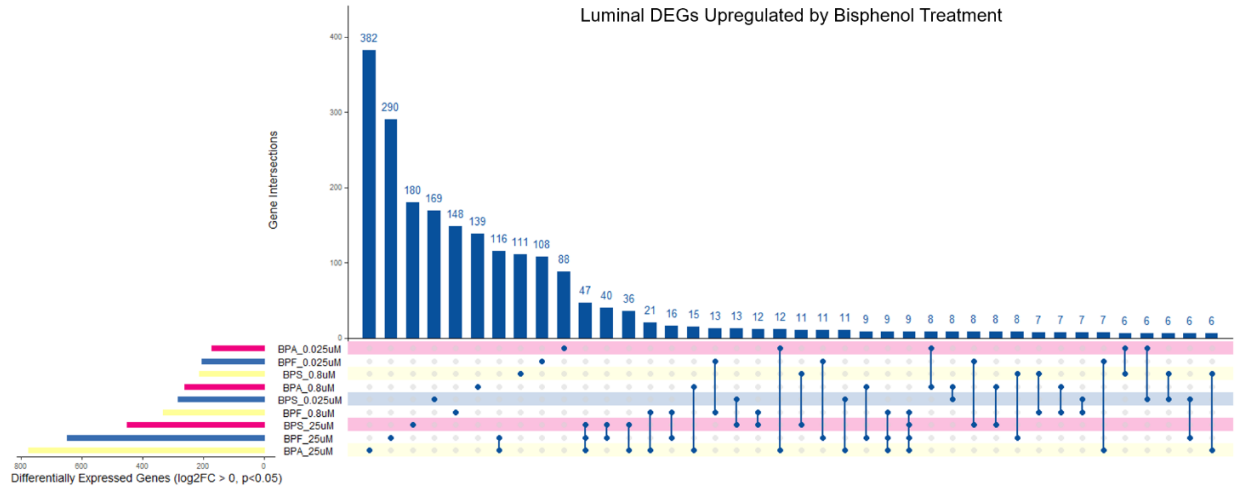
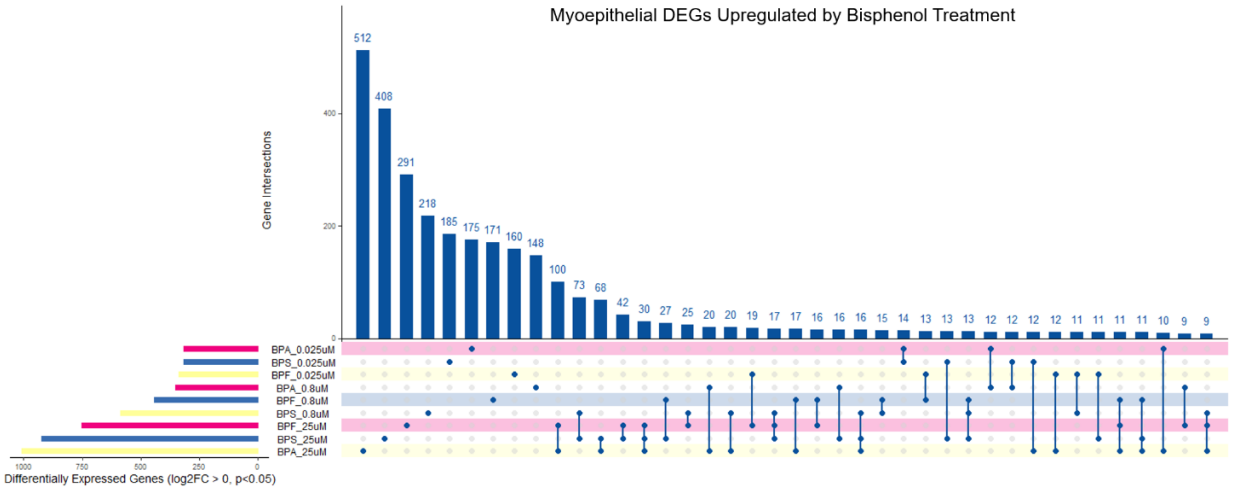


Figure 4.4. Number of differentially expressed genes across bisphenol doses A) Upregulated DEGs split by cell type ($p<0.05$) B) Downregulated DEGs split by cell type ($p<0.05$) C) FDR adjusted upregulated DEGs split by cell type (FDR <0.05) D) FDR adjusted downregulated DEGs split by cell type (FDR <0.05)

A


Luminal DEGs Upregulated by Bisphenol Treatment
Gene Intersections
382
290
180
169
148
139
116
111
108
88
47
40
36
21
16
15
13
13
12
12
11
11
11
9
9
9
8
8
8
8
8
7
7
7
7
6
6
6
6
6
BPA_0.025uM
BPF_0.025uM
BPS_0.8uM
BPA_0.8uM
BPS_0.025uM
BPF_0.8uM
BPS_25uM
BPF_25uM
BPA_25uM
800
600
400
200
0
Differentially Expressed Genes (log2FC > 0, p<0.05)

B


Myoepithelial DEGs Upregulated by Bisphenol Treatment
Gene Intersections
512
408
291
218
185
175
171
160
148
100
73
68
42
30
27
25
20
20
19
17
17
16
16
16
15
14
13
13
13
12
12
12
12
11
11
11
11
10
9
9
BPA_0.025uM
BPS_0.025uM
BPF_0.025uM
BPA_0.8uM
BPF_0.8uM
BPS_0.8uM
BPF_25uM
BPS_25uM
BPA_25uM
1000
750
500
250
0
Differentially Expressed Genes (log2FC > 0, p<0.05)

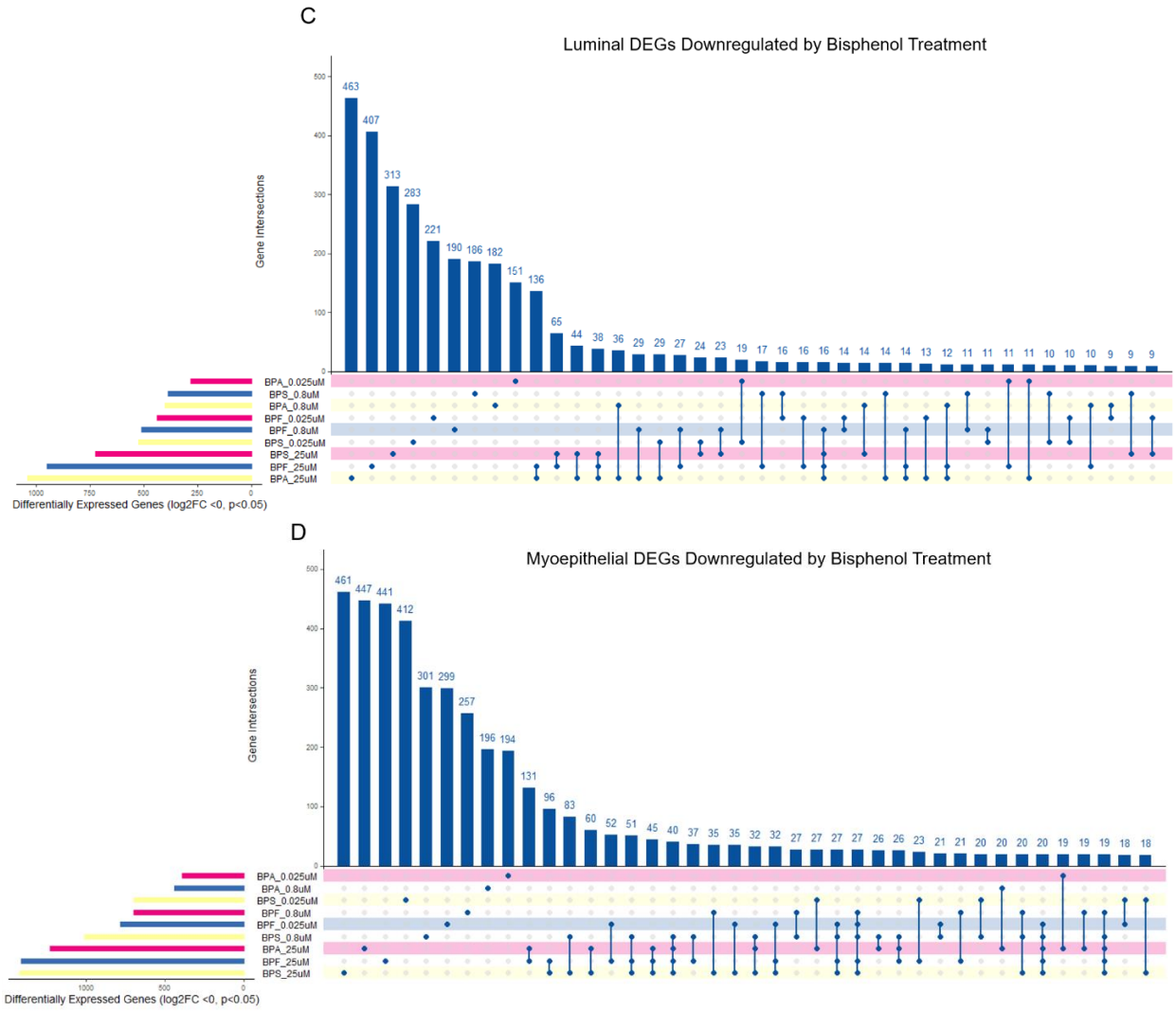


Figure 4.5. Overlapping differentially expressed genes across bisphenol doses A) Upregulated DEGs across luminal cells ($p<0.05$) B) Upregulated DEGs across myoepithelial cells ($p<0.05$) C) Downregulated DEGs across luminal cells ($p<0.05$) D) Downregulated DEGs across myoepithelial cells ($p<0.05$)

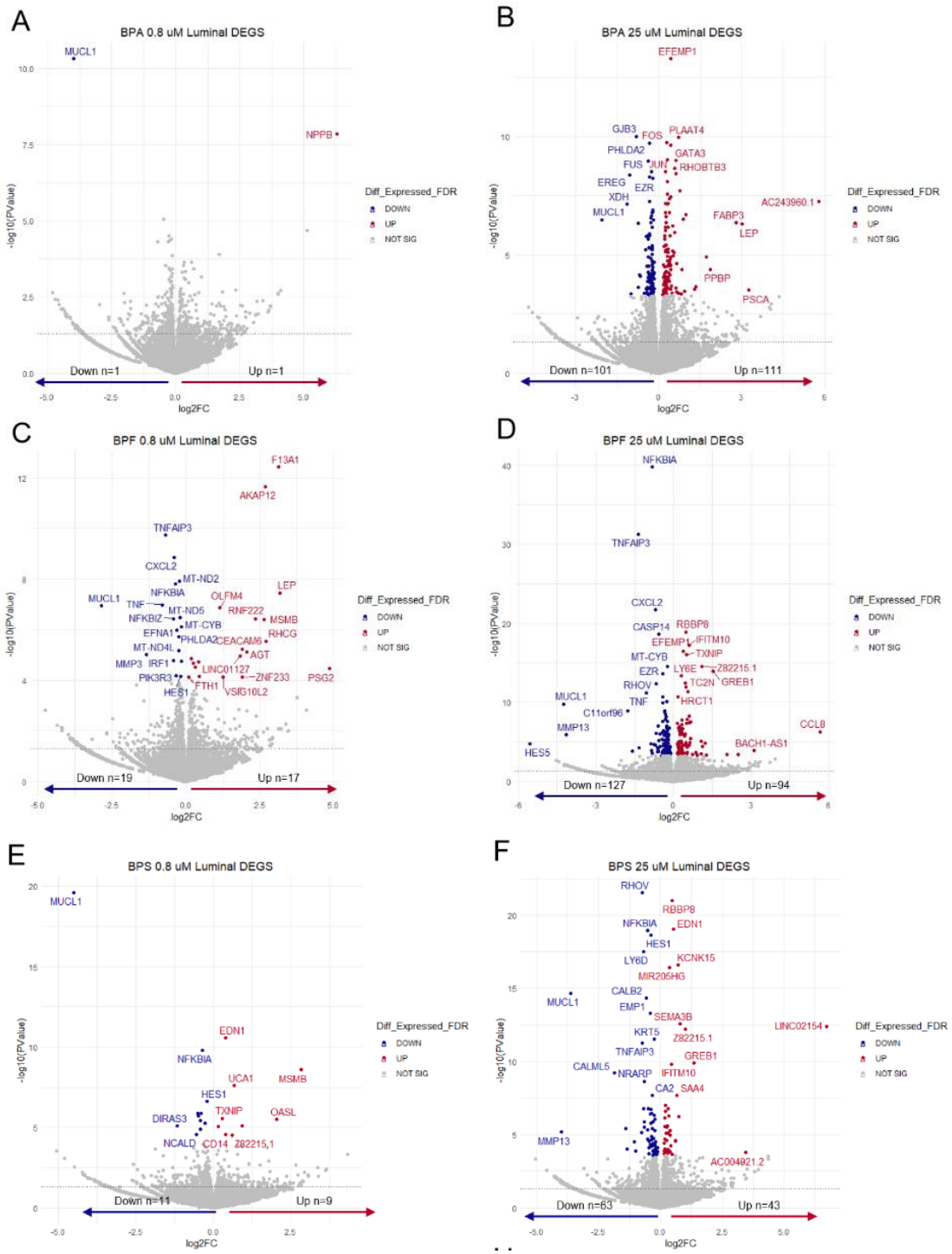


Figure 4.6 (A-F) Differentially expressed genes for 0.8µM and 25µM luminal cells. Genes in color met FDR <0.05 cutoff A) BPA 0.8 µM B) BPA 25 µM C) BPF 0.8 µM D) BPF 25 µM E) BPS 0.8 µM F) BPS 25 µM

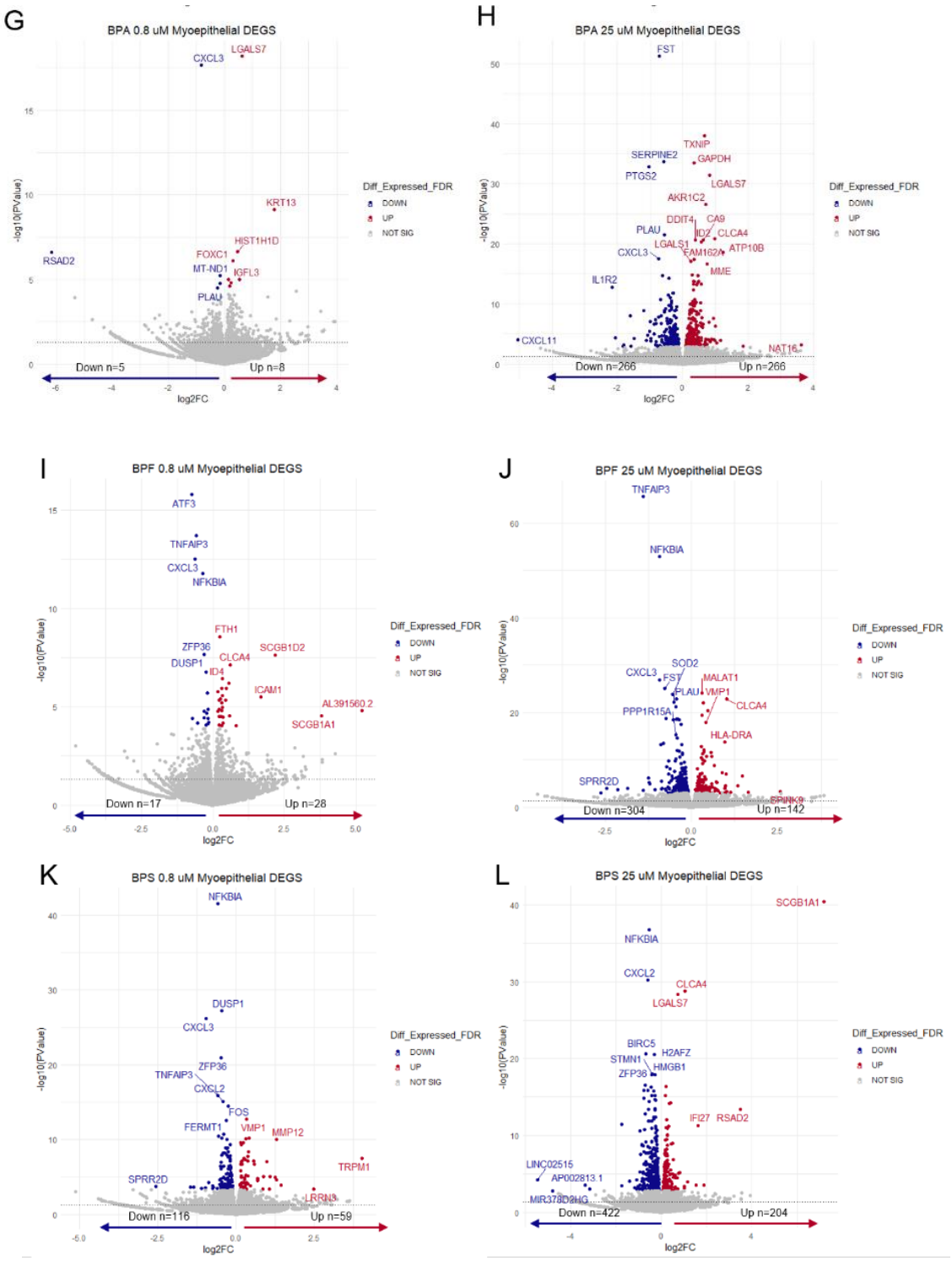


Figure 4.6 (G-L) Differentially expressed genes for 0.8μM and 25μM myoepithelial cells. Genes in color met FDR <0.05 cutoff G) BPA 0.8 μM H) BPA 25 μM I) BPF 0.8 μM J) BPF 25 μM K) BPS 0.8 μM L) BPS 25 μM

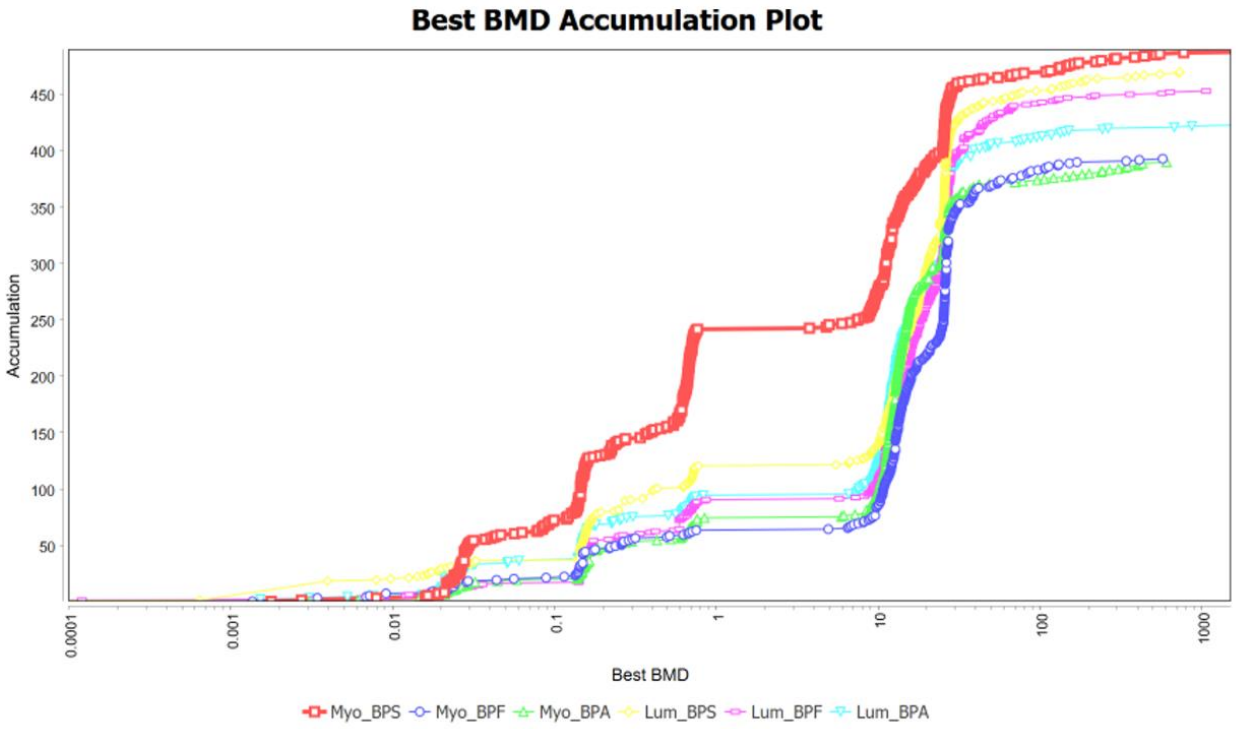


Figure 4.7. Benchmark dose response analysis using DEGs for each bisphenol.
Accumulation indicates the number of significantly affected genes at the corresponding dose.

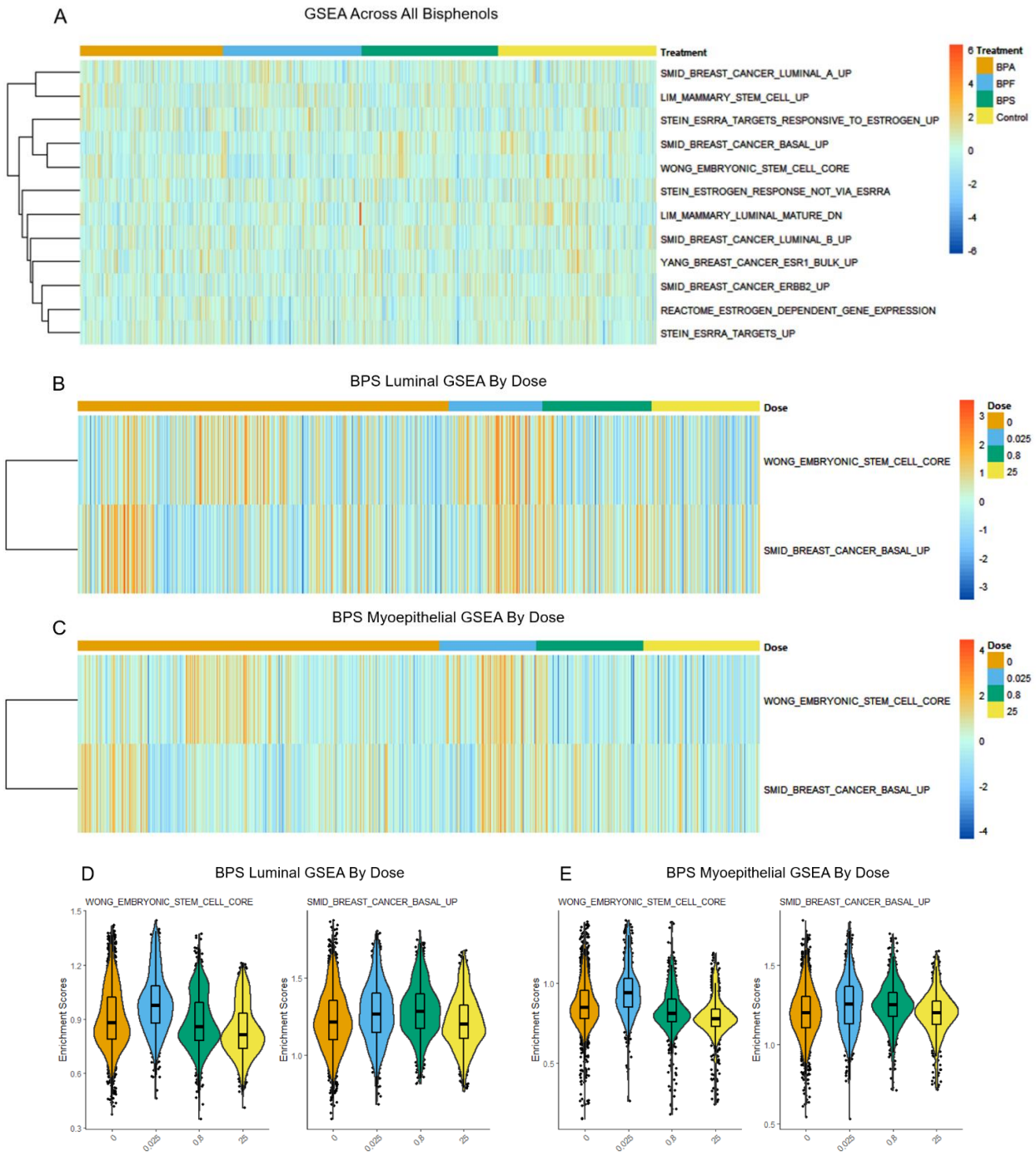


Figure 4.8. Gene set enrichment analysis of breast cancer, stem cell, and estrogen associated gene sets A) Combined heatmap across BPA, BPF, BPS cells and doses B) Heatmap of BPS luminal cells by dose for Wong embryonic stem cell and Smid breast cancer basal gene sets C) Heatmap of BPS myoepithelial cells D) Violinplot of BPS luminal GSEA by dose. All doses were significantly different (linear mixed effects model) from control for both gene sets p(<0.05) E) Violinplot of BPS myoepithelial GSEA by dose. All doses except 25 µM for Smid were significantly different from control p <0.05

Table 4.4 Overlaps between myoepithelial cells and Wong embryonic stem cell gene set

	BPS 0.025 µM			BPS 0.8 µM			BPS 25 µM		
Gene	logFC	Pvalue	FDR	logFC	PValue	FDR	logFC	Pvalue	FDR
AURKA	0.25	5.38E-03	5.66E-01	-0.50	3.56E-07	1.34E-04	-0.59	2.74E-09	7.65E-07
CDC20	0.30	4.02E-03	5.21E-01	-0.40	6.31E-04	6.49E-02	-0.46	1.10E-04	6.55E-03
CDCA3	0.25	2.54E-02	8.49E-01	-0.32	9.23E-03	3.14E-01	-0.55	1.60E-05	1.47E-03
CDCA8	0.27	1.56E-02	7.62E-01	-0.35	5.48E-03	2.41E-01	-0.31	1.05E-02	1.75E-01
CKS1B	0.12	6.09E-03	5.79E-01	-0.24	4.07E-07	1.45E-04	-0.33	7.34E-13	4.34E-10
CKS2	0.11	1.54E-02	7.62E-01	-0.26	2.42E-08	1.43E-05	-0.33	3.74E-13	2.36E-10
DLGAP5	0.24	3.06E-03	4.68E-01	-0.39	1.17E-05	2.94E-03	-0.77	1.30E-16	1.81E-13
GTSE1	0.29	2.27E-03	4.15E-01	-0.40	1.48E-04	2.40E-02	-0.74	1.78E-11	7.58E-09
HMGB2	0.26	3.26E-05	5.31E-02	-0.31	4.36E-06	1.25E-03	-0.57	3.82E-14	2.66E-11
KIF4A	0.39	1.60E-03	3.56E-01	-0.48	8.10E-04	7.76E-02	-0.56	7.03E-05	4.82E-03
KPNA2	0.13	9.51E-03	6.57E-01	-0.17	1.12E-03	9.57E-02	-0.29	2.50E-08	5.40E-06
NDC80	0.26	3.09E-02	8.74E-01	-0.38	4.41E-03	2.16E-01	-0.84	1.98E-09	5.70E-07
NUSAP1	0.23	2.19E-04	1.47E-01	-0.45	6.63E-11	9.97E-08	-0.59	1.43E-16	1.87E-13
PLK1	0.20	4.78E-02	9.58E-01	-0.24	2.64E-02	4.83E-01	-0.64	1.93E-08	4.39E-06
PSMA7	0.07	3.14E-02	8.75E-01	-0.11	4.16E-04	4.83E-02	-0.16	2.00E-07	3.50E-05
SERPINH1	-0.10	3.62E-02	9.03E-01	-0.15	1.30E-03	1.02E-01	-0.16	6.45E-04	2.57E-02
SMC4	0.17	3.85E-03	5.12E-01	-0.25	4.32E-05	8.79E-03	-0.42	1.07E-11	4.76E-09

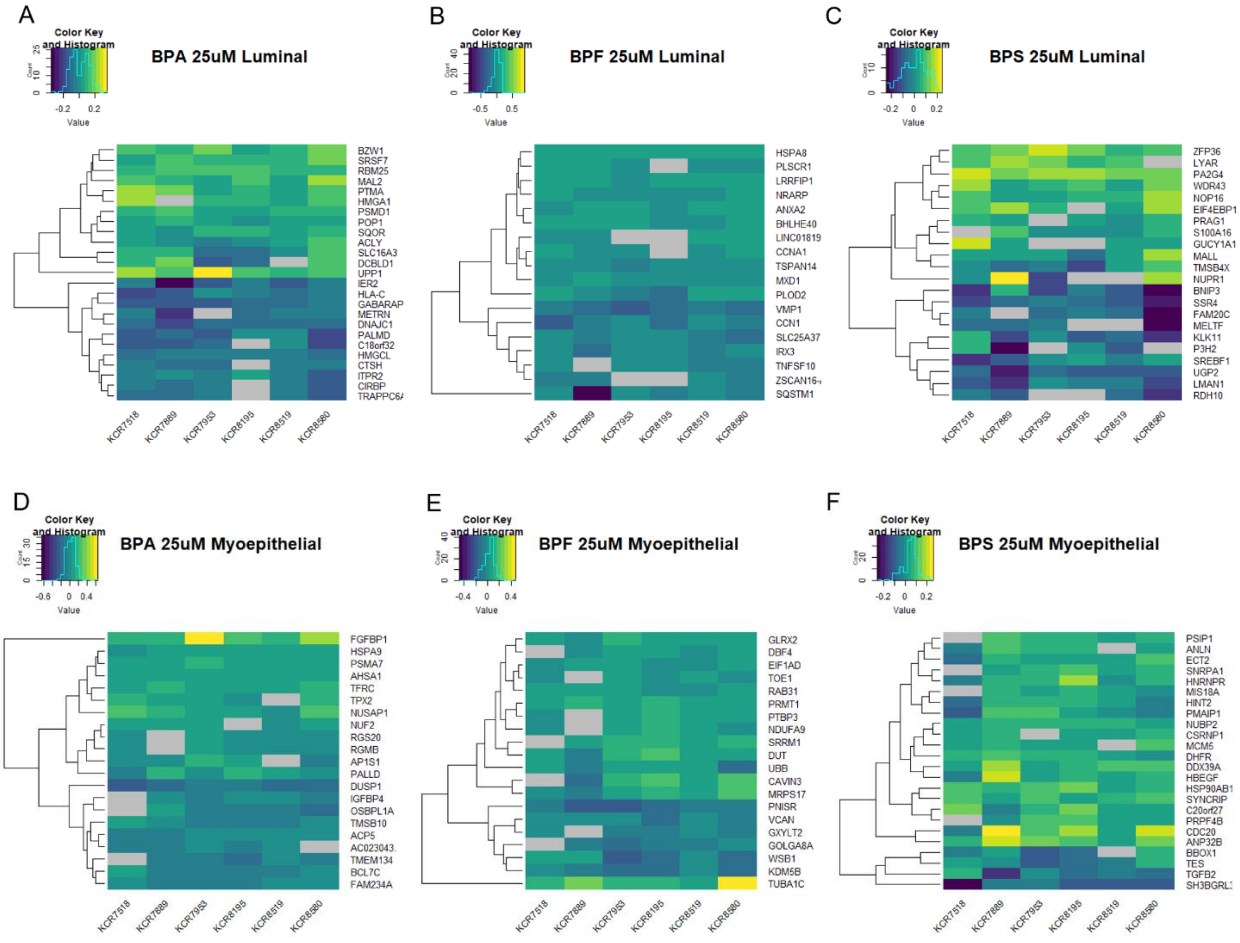


Figure 4.9 Difference in expression of most significant differentially expressed genes by individual. A) BPA 25 µM luminal B) BPF 25 µM luminal C) BPS 25 µM luminal D) BPA 25 µM myoepithelial E) BPF 25 µM myoepithelial F) BPS 25 µM myoepithelial

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/CAAC.21660

2. Zeinomar N, Oskar S, Kehm RD, Sahebzeda S, Terry MB. Environmental exposures and breast cancer risk in the context of underlying susceptibility: A systematic review of the epidemiological literature. *Environ Res*. 2020;187:109346. doi:10.1016/j.envres.2020.109346

3. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370

4. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

5. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol*. 2014;5(3):412. doi:10.5306/WJCO.V5.I3.412

6. Johnson KS, Conant EF, Soo MS. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *J Breast Imaging*. 2021;3(1):12-24. doi:10.1093/JBI/WBAA110

7. Dietze EC, Sistrunk C, Miranda-Carboni G, O'Regan R, Seewaldt VL. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat Rev Cancer*. 2015;15(4):248-254. doi:10.1038/nrc3896

8. Carey LA, Perou CM, Livasy CA, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *JAMA*. 2006;295(21):2492-2502. doi:10.1001/JAMA.295.21.2492

9. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev*. 2017;26(4):444-457. doi:10.1158/1055-9965.EPI-16-0858

10. Newman LA. Breast Cancer Disparities: Socioeconomic Factors versus Biology. *Ann Surg Oncol*. 2017;24(10):2869-2875. doi:10.1245/S10434-017-5977-1/TABLES/2

11. Ooi SL, Martinez ME, Li CI. Disparities in breast cancer characteristics and outcomes by race/ethnicity. *Breast Cancer Res Treat*. 2011;127(3):729-738. doi:10.1007/S10549-010-1191-6/TABLES/6

12. Newman LA, Kaljee LM. Health Disparities and Triple-Negative Breast Cancer in African American Women. *JAMA Surg*. 2017;152(5):485. doi:10.1001/jamasurg.2017.0005

13. Stringer-Reasor EM, Elkhanany A, Khoury K, Simon MA, Newman LA. Disparities in Breast Cancer Associated With African American Identity. *Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet*. 2021;41(41):e29-e46. doi:10.1200/EDBK_319929

14. Yedjou CG, Sims JN, Miele L, et al. Health and Racial Disparity in Breast Cancer. *Adv*

Exp Med Biol. 2019;1152:31. doi:10.1007/978-3-030-20301-6_3

15. Newman LA, Carpten J. Integrating the Genetics of Race and Ethnicity Into Cancer Research: Trailing Jane and John Q. Public. *JAMA Surg*. 2018;153(4):299. doi:10.1001/JAMASURG.2017.5080

16. Sung H, DeSantis CE, Fedewa SA, Kantelhardt EJ, Jemal A. Breast cancer subtypes among Eastern-African-born black women and other black women in the United States. *Cancer*. 2019;125(19):3401-3411. doi:10.1002/CNCR.32293

17. Jiagge E, Jibril AS, Chitale D, et al. Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer. *Ann Surg Oncol*. 2016;23(12):3843-3849. doi:10.1245/S10434-016-5420-Z/TABLES/5

18. Capp J-P. Cancer Stem Cells: From Historical Roots to a New Perspective. *J Oncol*. 2019;2019:1-10. doi:10.1155/2019/5189232

19. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. doi:10.1038/35102167

20. Ma Y, Zhang P, Wang F, Yang J, Yang Z, Qin H. The relationship between early embryo development and tumourigenesis. *J Cell Mol Med*. 2010;14(12):2697-2701. doi:10.1111/j.1582-4934.2010.01191.x

21. Reya T, Clevers H. Wnt signalling in stem cells and cancer. *Nature*. 2005;434(7035):843-850. doi:10.1038/nature03319

22. Papaccio F, Paino F, Regad T, Papaccio G, Desiderio V, Tirino V. Concise Review: Cancer Cells, Cancer Stem Cells, and Mesenchymal Stem Cells: Influence in Cancer Development. *Stem Cells Transl Med*. 2017;6(12):2115-2125. doi:10.1002/sctm.17-0138

23. Clevers H. The cancer stem cell: premises, promises and challenges. *Nat Med*. 2011;17(3):313-319. doi:10.1038/nm.2304

24. Goichberg P. Current Understanding of the Pathways Involved in Adult Stem and Progenitor Cell Migration for Tissue Homeostasis and Repair. *Stem Cell Rev Reports*. 2016;12(4):421-437. doi:10.1007/s12015-016-9663-7

25. Shaw TJ, Martin P. Wound repair: a showcase for cell plasticity and migration. *Curr Opin Cell Biol*. 2016;42:29-37. doi:10.1016/J.CEB.2016.04.001

26. Hennighausen L, Robinson GW. Signaling Pathways in Mammary Gland Development. *Dev Cell*. 2001;1(4):467-475. doi:10.1016/S1534-5807(01)00064-8

27. Macias H, Hinck L. Mammary Gland Development. *Wiley Interdiscip Rev Dev Biol*. 2012;1(4):533. doi:10.1002/WDEV.35

28. Russo J. The Windows of Susceptibility to Breast Cancer. *Pathobiol Breast Cancer*. Published online 2016:1-20. doi:10.1007/978-3-319-40815-6_1

29. Mammary Stem Cells: Premise, Properties, and Perspectives. *Trends Cell Biol*. 2017;27(8):556-567. doi:10.1016/J.TCB.2017.04.001

30. Shackleton M, Vaillant F, Simpson KJ, et al. Generation of a functional mammary gland from a single stem cell. *Nature*. 2006;439(7072):84-88. doi:10.1038/nature04372

31. Daniel CW, Deome KB, Young JT, Blair PB, Faulkin LJ. The in vivo life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *J Mammary Gland Biol Neoplasia*. 2009;14(3):355-362. doi:10.1007/s10911-009-9139-3

32. Stingl J, Eaves CJ, Zandieh I, Emerman JT. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat*. 2001;67(2):93-109. doi:10.1023/A:1010615124301

33. Ghebeh H, Sleiman GM, Manogaran PS, et al. Profiling of normal and malignant breast tissue show CD44high/CD24low phenotype as a predominant stem/progenitor marker when used in combination with Ep-CAM/CD49f markers. *BMC Cancer*. 2013;13:289. doi:10.1186/1471-2407-13-289

34. Ginestier C, Hur MH, Charafe-Jauffret E, et al. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell*. 2007;1(5):555-567. doi:10.1016/j.stem.2007.08.014

35. Schwartz T, Stark A, Pang J, et al. Expression of aldehyde dehydrogenase 1 as a marker of mammary stem cells in benign and malignant breast lesions of Ghanaian women. *Cancer*. 2013;119(3):488-494. doi:10.1002/cncr.27737

36. Jiagge E, Chitale D, Newman LA, Newman A. Triple-Negative Breast Cancer, Stem Cells, and African Ancestry. *J Pathol*. 2018;188(2). doi:10.1016/j.ajpath.2017.06.020

37. Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature*. 2016;529(7584):43-47. doi:10.1038/nature16166

38. Gray J, Evans N, Taylor B, Rizzo J, Walker M. State of the Evidence: The Connection Between Breast Cancer and the Environment. *Int J Occup Environ Health*. 2009;15(1):43-78. doi:10.1179/107735209799449761

39. Forman MR, Winn DM, Collman GW, Rizzo J, Birnbaum LS. Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod Toxicol*. 2015;54:6-10. doi:10.1016/j.reprotox.2014.10.019

40. Nakshatri H, Anjanappa M, Bhat-Nakshatri P. Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. *Sci Rep*. 2015;5:13526. doi:10.1038/srep13526

41. Colacino JA, Azizi E, Brooks MD, et al. Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports*. Published online 2018. doi:10.1016/j.stemcr.2018.03.001

42. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med*. 2009;27(5):351-357. doi:10.1055/s-0029-1237423

43. Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol Rev*. 2020;100(2):489-523. doi:10.1152/PHYSREV.00040.2018/ASSET/IMAGES/LARGE/Z9J0012029270010.JPE

G

44. Terry MB, Michels KB, Brody JG, et al. Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research. *Breast Cancer Res 2019 211*. 2019;21(1):1-16. doi:10.1186/S13058-019-1168-2

45. Attina TM, Malits J, Naidu M, Trasande L. Racial/ethnic disparities in disease burden and costs related to exposure to endocrine-disrupting chemicals in the United States: an exploratory analysis. *J Clin Epidemiol*. 2019;108:34-43. doi:10.1016/j.jclinepi.2018.11.024

46. Ruiz D, Becerra M, Jagai JS, Ard K, Sargis RM. Disparities in Environmental Exposures to Endocrine-Disrupting Chemicals and Diabetes Risk in Vulnerable Populations. *Diabetes Care*. 2018;41(1):193-205. doi:10.2337/dc16-2765

47. Lillo MA, Nichols C, Seagroves TN, Miranda-Carboni GA, Krum SA. Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Horm Cancer*. 2017;8(2):90-99. doi:10.1007/s12672-017-0286-5

48. Hindman AR, Mo XM, Helber HL, et al. Varying Susceptibility of the Female Mammary Gland to In Utero Windows of BPA Exposure. *Endocrinology*. 2017;158(10):3435-3447. doi:10.1210/en.2017-00116

49. Gerona RR, Pan J, Zota AR, et al. Direct measurement of Bisphenol A (BPA), BPA glucuronide and BPA sulfate in a diverse and low-income population of pregnant women reveals high exposure, with potential implications for previous exposure estimates: a cross-sectional study. *Environ Health*. 2016;15:50. doi:10.1186/s12940-016-0131-2

50. Tucker DK, Hayes Bouknight S, Brar SS, Kissling GE, Fenton SE. Evaluation of Prenatal Exposure to Bisphenol Analogues on Development and Long-Term Health of the Mammary Gland in Female Mice. *Environ Health Perspect*. 2018;126(8):087003. doi:10.1289/EHP3189

51. Huang W, Zhao C, Zhong H, Zhang S, Xia Y, Cai Z. Bisphenol S induced epigenetic and transcriptional changes in human breast cancer cell line MCF-7. *Environ Pollut*. 2019;246:697-703. doi:10.1016/J.ENVPOL.2018.12.084

52. Weng Y-I, Hsu P-Y, Liyanarachchi S, et al. Epigenetic influences of low-dose bisphenol A in primary human breast epithelial cells. *Toxicol Appl Pharmacol*. 2010;248(2):111. doi:10.1016/J.TAAP.2010.07.014

53. Liu X, Krawczyk E, Suprynowicz FA, et al. Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc*. 2017;12(2):439-451. doi:10.1038/nprot.2016.174

54. Giraddi RR, Chung C-Y, Heinz RE, et al. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep*. 2018;24(6):1653-1666.e7. doi:10.1016/j.celrep.2018.07.025

55. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18(1):35-45. doi:10.1038/nri.2017.76

56. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20(5):257-272. doi:10.1038/s41576-019-0093-7

57. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health*. 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629

58. Holliday H, Baker LA, Junankar SR, Clark SJ, Swarbrick A. Epigenomics of mammary gland development. *Breast Cancer Res*. 2018;20(1):100. doi:10.1186/s13058-018-1031-x

59. Thong T, Forté CA, Hill EM, Colacino JA. Environmental exposures, stem cells, and cancer. *Pharmacol Ther*. Published online 2019. doi:10.1016/j.pharmthera.2019.107398

60. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest*. 2009;119(6):1420-1428. doi:10.1172/JCI39104

61. Merrell AJ, Stanger BZ. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style. *Nat Rev Mol Cell Biol*. 2016;17(7):413-425. doi:10.1038/nrm.2016.24

62. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med*. 2013;19(11):1438-1449. doi:10.1038/nm.3336

63. Jolly MK, Boareto M, Huang B, et al. Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol*. 2015;5:155. doi:10.3389/fonc.2015.00155

64. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

65. Jolly MK, Tripathi SC, Jia D, et al. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*. 2016;7(19):27067-27084. doi:10.18632/oncotarget.8166

66. Dontu G, Abdallah WM, Foley JM, et al. In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev*. 2003;17(10):1253-1270. doi:10.1101/gad.1061803

67. Colacino JA, McDermott SP, Sartor MA, Wicha MS, Rozek LS. Transcriptomic profiling of curcumin-treated human breast stem cells identifies a role for stearoyl-coa desaturase in breast cancer prevention. *Breast Cancer Res Treat*. 2016;158(1):29-41. doi:10.1007/s10549-016-3854-4

68. Liu X, Ory V, Chapman S, et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *Am J Pathol*. 2012;180(2):599-607. doi:10.1016/j.ajpath.2011.10.036

69. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/J.CELL.2015.05.002

70. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33(5):495-502. doi:10.1038/nbt.3192

71. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Birol I, ed. *Bioinformatics*. 2019;35(3):487-496. doi:10.1093/bioinformatics/bty644

72. Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, Chang HY. Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells. *Cell Stem Cell*. 2008;2(4):333-344. doi:10.1016/j.stem.2008.02.009

73. Kuleshov M V., Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

74. Lim E, Vaillant F, Wu D, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*. 2009;15(8):907-913. doi:10.1038/nm.2000

75. Bach K, Pensa S, Grzelak M, et al. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun*. 2017;8(1):2128. doi:10.1038/s41467-017-02001-5

76. Nguyen QH, Pervolarakis N, Blake K, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-04334-1

77. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71-e71. doi:10.1093/nar/gkv1507

78. Wang Y, Thong T, Saligrama V, Colacino J, Balzano L, Scott C. A gene filter for comparative analysis of single-cell RNA-sequencing trajectory datasets. *bioRxiv*. Published online May 17, 2019:637488. doi:10.1101/637488

79. Abd El-Rehim DM, Pinder SE, Paish CE, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol*. 2004;203(2):661-671. doi:10.1002/path.1559

80. O'Hare MJ, Ormerod MG, Monaghan P, Lane EB, Gusterson BA. Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting. *Differentiation*. 1991;46(3):209-221. doi:10.1111/j.1432-0436.1991.tb00883.x

81. Marcato P, Dean CA, Pan D, et al. Aldehyde Dehydrogenase Activity of Breast Cancer Stem Cells Is Primarily Due To Isoform ALDH1A3 and Its Expression Is Predictive of Metastasis. *Stem Cells*. 2011;29(1):32-45. doi:10.1002/stem.563

82. Trzpis M, McLaughlin PMJ, De Leij LMFH, Harmsen MC. Epithelial cell adhesion molecule: More than a carcinoma marker and adhesion molecule. *Am J Pathol*. 2007;171(2):386-395. doi:10.2353/ajpath.2007.070152

83. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A*. 2003;100(7):3983-3988. doi:10.1073/pnas.0530291100

84. Gerdur Ísberg Ó, Kim J, Fridriksdottir AJ, et al. A CD146 FACS Protocol Enriches for Luminal Keratin 14/19 Double Positive Human Breast Progenitors. *Sci Rep*. 2019;9(1):14843. doi:10.1038/s41598-019-50903-9

85. Jolly MK, Somarelli JA, Sheth M, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Ther*. 2019;194:161-184. doi:10.1016/j.pharmthera.2018.09.007

86. Casey T, Bond J, Tighe S, et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res Treat*. 2009;114(1):47-62. doi:10.1007/s10549-008-9982-8

87. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173(2):338-354.e15. doi:10.1016/j.cell.2018.03.034

88. Feng Y, Sun B, Li X, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. *Breast Cancer Res Treat*. 2007;103(3):319-329. doi:10.1007/s10549-006-9385-7

89. Folgueira MAAK, Maistro S, Katayama MLH, et al. Markers of breast cancer stromal fibroblasts in the primary tumour site associated with lymph node metastasis: A systematic review including our case series. *Biosci Rep*. 2013;33(6). doi:10.1042/BSR20130060

90. Li J-M, Tseng C-W, Lin C-C, et al. Upregulation of LGALS1 is associated with oral cancer metastasis. *Ther Adv Med Oncol*. 2018;10:175883591879462. doi:10.1177/1758835918794622

91. Falato C, Lorent J, Tani E, et al. Ki67 measured in metastatic tissue and prognosis in patients with advanced breast cancer. *Breast Cancer Res Treat*. 2014;147(2):407-414. doi:10.1007/s10549-014-3096-2

92. Montes de Oca R, Gurard-Levin ZA, Berger F, et al. The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol Oncol*. 2015;9(3):657-674. doi:10.1016/j.molonc.2014.11.002

93. Ding K, Li W, Zou Z, Zou X, Wang C. CCNB1 is a prognostic biomarker for ER+ breast cancer. *Med Hypotheses*. 2014;83(3):359-364. doi:10.1016/j.mehy.2014.06.013

94. Li Y, Chen Y-L, Xie Y-T, et al. Association Study of Germline Variants in CCNB1 and CDK1 with Breast Cancer Susceptibility, Progression, and Survival among Chinese Han Women. Miao X, ed. *PLoS One*. 2013;8(12):e84489. doi:10.1371/journal.pone.0084489

95. Wang Y, Zhang C, Mai L, Niu Y, Wang Y, Bu Y. PRR11 and SKA2 gene pair is overexpressed and regulated by p53 in breast cancer. *BMB Rep*. 2019;52(2):157-162. doi:10.5483/BMBRep.2019.52.2.207

96. Ren Z, Yang T, Zhang P, Liu K, Liu W, Wang P. SKA2 mediates invasion and metastasis in human breast cancer via EMT. *Mol Med Rep*. 2019;19(1):515-523. doi:10.3892/mmr.2018.9623

97. Han JY, Han YK, Park GY, et al. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines. *Sci Rep*. 2015;5. doi:10.1038/srep15993

98. Hamy AS, Bieche I, Lehmann-Che J, et al. BIRC5 (survivin): a pejorative prognostic marker in stage II/III breast cancer with no response to neoadjuvant chemotherapy. *Breast Cancer Res Treat*. 2016;159(3):499-511. doi:10.1007/s10549-016-3961-2

99. Ghaffari K, Hashemi M, Ebrahimi E, Shirkoohi R. BIRC5 genomic copy number variation in early-onset breast cancer. *Iran Biomed J*. 2016;20(4):241-245. doi:10.7508/ibj.2016.04.009

100. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. Ben-Jacob E, ed. *PLoS One*. 2015;10(5):e0126522. doi:10.1371/journal.pone.0126522

101. Yu M, Bardia A, Wittner BS, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (80-)*. 2013;339(6119):580-584. doi:10.1126/science.1228522

102. Sun P, Yuan Y, Li A, Li B, Dai X. Cytokeratin expression during mouse embryonic and early postnatal mammary gland development. *Histochem Cell Biol*. 2010;133(2):213-221. doi:10.1007/s00418-009-0662-5

103. Ruiz R, Herrero C, Strasser-Weippl K, et al. Epidemiology and pathophysiology of pregnancy-associated breast cancer: A review. *Breast*. 2017;35:136-141. doi:10.1016/j.breast.2017.07.008

104. Johansson ALV, Andersson TML, Hsieh CC, et al. Tumor characteristics and prognosis in women with pregnancy-associated breast cancer. *Int J Cancer*. 2018;142(7):1343-1354. doi:10.1002/ijc.31174

105. Degnim AC, Visscher DW, Hoskin TL, et al. Histologic findings in normal breast tissues: comparison to reduction mammaplasty and benign breast disease tissues. *Breast Cancer Res Treat*. 2012;133(1):169-177. doi:10.1007/s10549-011-1746-1

106. Rheinwatd JG, Green H. Seria cultivation of strains of human epidemal keratinocytes: the formation keratinizin colonies from single cell is. *Cell*. 1975;6(3):331-343. doi:10.1016/S0092-8674(75)80001-8

107. Domchek SM, Yao S, Chen F, et al. Comparison of the Prevalence of Pathogenic Variants in Cancer Susceptibility Genes in Black Women and Non-Hispanic White Women With Breast Cancer in the United States. *JAMA Oncol*. 2021;7(7):1. doi:10.1001/JAMAONCOL.2021.1492

108. Shimelis H, LaDuca H, Hu C, et al. Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst*. 2018;110(8):855-862. doi:10.1093/jnci/djy106

109. Roelands J, Mall R, Almeer H, et al. Ancestry-associated transcriptomic profiles of breast cancer in patients of African, Arab, and European ancestry. *NPJ Breast Cancer*. 2021;7(1). doi:10.1038/S41523-021-00215-X

110. Nievergelt CM, Maihofer AX, Shekhtman T, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*. 2013;4(1):13. doi:10.1186/2041-2223-4-13

111. Thong T, Wang Y, Brooks MD, et al. Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Front Cell Dev Biol*. 2020;8:288. doi:10.3389/FCELL.2020.00288/BIBTEX

112. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol 2018 365*. 2018;36(5):411-420. doi:10.1038/nbt.4096

113. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048/ATTACHMENT/1990279C-028D-47D4-A5D3-C89B04E14795/MMC3.XLSX

114. Chapter 4 DE analyses between conditions | Multi-Sample Single-Cell Analyses with Bioconductor. Accessed May 25, 2022. http://bioconductor.org/books/3.14/OSCA.multisample/multi-sample-comparisons.html

115. Gulati GS, Sikandar SS, Wesche DJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*. 2020;367(6476):405-411. doi:10.1126/SCIENCE.AAX0249

116. Using escape to perform ssGSEA analyses on single-cell RNA-seq data. Accessed May 25, 2022. http://www.bioconductor.org/packages/release/bioc/vignettes/escape/inst/doc/vignette.html

117. GSEA | MSigDB | Browse Gene Sets. Accessed May 25, 2022. https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

118. Cell-Cycle Scoring and Regression • Seurat. Accessed May 19, 2022. https://satijalab.org/seurat/articles/cell_cycle_vignette.html

119. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/BIOINFORMATICS/BTS480

120. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363. doi:10.1093/BIOINFORMATICS/BTU049

121. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res*. 2016;44(3):e20. doi:10.1093/NAR/GKV907

122. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics and Chromatin*. 2015;8(1):1-16. doi:10.1186/1756-8935-8-6/FIGURES/6

123. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208. doi:10.1186/s13059-016-1066-1

124. Fortin J-P, Triche TJ, Hansen KD, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2017;33(4):558-560. doi:10.1093/bioinformatics/btw691

125. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287. doi:10.1038/NG.3656

126. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443. doi:10.1038/NG.3679

127. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(10):1654. doi:10.1001/JAMAONCOL.2017.0595

128. Wu H, Wang C, Wu Z. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*. 2015;31(2):233-241. doi:10.1093/BIOINFORMATICS/BTU640

129. Yap SQ, Mathavarajah S, Huber RJ. The converging roles of Batten disease proteins in neurodegeneration and cancer. *iScience*. 2021;24(4):102337. doi:10.1016/J.ISCI.2021.102337

130. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet (London, England)*. 2005;365(9460):671-679. doi:10.1016/S0140-6736(05)17947-1

131. Pani T, Rajput K, Kar A, et al. Alternative splicing of ceramide synthase 2 alters levels of specific ceramides and modulates cancer cell proliferation and migration in Luminal B breast cancer subtype. *Cell Death Dis 2021 122*. 2021;12(2):1-22. doi:10.1038/s41419-021-03436-x

132. D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, Troester MA. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res Treat*. 2015;152(2):437-448. doi:10.1007/S10549-015-3474-4/TABLES/6

133. Huo D, Hu H, Rhie SK, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(12):1654-1662. doi:10.1001/JAMAONCOL.2017.0595

134. Barrow MA, Martin ME, Coffey A, et al. A functional role for the cancer disparity-linked genes, CRYβB2 and CRYβB2P1, in the promotion of breast cancer. *Breast Cancer Res*. 2019;21(1):105. doi:10.1186/s13058-019-1191-3

135. Field LA, Love B, Deyarmin B, Hooke JA, Shriver CD, Ellsworth RE. Identification of differentially expressed genes in breast tumors from African American compared with Caucasian women. *Cancer*. 2012;118(5):1334-1344. doi:10.1002/CNCR.26405

136. Li H, Cheng C, You W, et al. TMEM100 Modulates TGF-β Signaling Pathway to Inhibit

Colorectal Cancer Progression. *Gastroenterol Res Pract*. 2021;2021. doi:10.1155/2021/5552324

137. Ma J, Yan T, Bai Y, et al. TMEM100 negatively regulated by microRNA-106b facilitates cellular apoptosis by suppressing survivin expression in NSCLC. *Oncol Rep*. 2021;46(2):1-14. doi:10.3892/OR.2021.8136

138. Zhuang J, Huang Y, Zheng W, et al. TMEm100 expression suppresses metastasis and enhances sensitivity to chemotherapy in gastric cancer. *Biol Chem*. 2020;401(2):285-296. doi:10.1515/HSZ-2019-0161/MACHINEREADABLECITATION/RIS

139. Donald CD, Sun CQ, Lim SD, et al. Cancer-Specific Loss of β-Defensin 1 in Renal and Prostatic Carcinomas. *Lab Investig 2003 834*. 2003;83(4):501-505. doi:10.1097/01.lab.0000063929.61760.f6

140. Lee J, Han JH, Jang A, Kim JW, Hong SA, Myung SC. DNA Methylation-Mediated Downregulation of DEFB1 in Prostate Cancer Cells. *PLoS One*. 2016;11(11). doi:10.1371/JOURNAL.PONE.0166664

141. Lo JH, Battaglin F, Baca Y, et al. DEFB1 gene expression and the molecular landscape of colorectal cancer (CRC). *https://doi.org/101200/JCO20224016_suppl3523*. 2022;40(16_suppl):3523-3523. doi:10.1200/JCO.2022.40.16_SUPPL.3523

142. Li YH, Hu YQ, Wang SC, Li Y, Chen DM. LncRNA SNHG5: A new budding star in human cancers. *Gene*. 2020;749:144724. doi:10.1016/J.GENE.2020.144724

143. Chi JR, Yu ZH, Liu BW, et al. SNHG5 Promotes Breast Cancer Proliferation by Sponging the miR-154-5p/PCNA Axis. *Mol Ther Nucleic Acids*. 2019;17:138. doi:10.1016/J.OMTN.2019.05.013

144. Damas ND, Marcatti M, Côme C, et al. SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nat Commun 2016 71*. 2016;7(1):1-14. doi:10.1038/ncomms13875

145. Shen M, Jiang YZ, Wei Y, et al. Tinagl1 Suppresses Triple-Negative Breast Cancer Progression and Metastasis by Simultaneously Inhibiting Integrin/FAK and EGFR Signaling. *Cancer Cell*. 2019;35(1):64-80.e7. doi:10.1016/J.CCELL.2018.11.016/ATTACHMENT/1E19EDC0-441D-4164-93EE-F3A5B9EC190A/MMC1.PDF

146. Musetti SN, Huang L. Tinagl1 Gene Therapy Suppresses Growth and Remodels the Microenvironment of Triple Negative Breast Cancer. *Mol Pharm*. 2021;18(5):2032-2038. doi:10.1021/ACS.MOLPHARMACEUT.1C00008

147. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob Heal*. 2020;8(8):e1027-e1037. doi:10.1016/S2214-109X(20)30215-1

148. Zota AR, Shamasunder B. The environmental injustice of beauty: framing chemical exposures from beauty products as a health disparities concern. *Am J Obstet Gynecol*. 2017;217(4):418.e1-418.e6. doi:10.1016/J.AJOG.2017.07.020

149. Ash M, Boyce JK. Racial disparities in pollution exposure and employment at US industrial facilities. *Proc Natl Acad Sci U S A*. 2018;115(42):10636-10641. doi:10.1073/PNAS.1721640115/SUPPL_FILE/PNAS.1721640115.SAPP.PDF

150. Polemi KM, Nguyen VK, Heidt J, Kahana A, Jolliet O, Colacino JA. Identifying the link between chemical exposures and breast cancer in African American women via integrated in vitro and exposure biomarker data. *Toxicology*. 2021;463. doi:10.1016/J.TOX.2021.152964

151. Nguyen VK, Kahana A, Heidt J, et al. A comprehensive analysis of racial disparities in chemical biomarker concentrations in United States women, 1999–2014. *Environ Int*. 2020;137:105496. doi:10.1016/J.ENVINT.2020.105496

152. Atlas E, Dimitrova V. Bisphenol S and Bisphenol A disrupt morphogenesis of MCF-12A human mammary epithelial cells. *Sci Rep*. 2019;9(1):1-10. doi:10.1038/s41598-019-52505-x

153. LaPlante CD, Catanese MC, Bansal R, Vandenberg LN. Bisphenol S Alters the Lactating Mammary Gland and Nursing Behaviors in Mice Exposed During Pregnancy and Lactation. *Endocrinology*. 2017;158(10):3448-3461. doi:10.1210/EN.2017-00437

154. Winkler J, Liu P, Phong K, et al. Bisphenol A replacement chemicals, BPF and BPS, induce protumorigenic changes in human mammary gland organoid morphology and proteome. *Proc Natl Acad Sci U S A*. 2022;119(11). doi:10.1073/PNAS.2115308119/-/DCSUPPLEMENTAL

155. Erler C, Novak J. Bisphenol A Exposure: Human Risk and Health Policy. *J Pediatr Nurs*. 2010;25(5):400-407. doi:10.1016/J.PEDN.2009.05.006

156. Konieczna A, Rutkowska A, Rachoń D. Health risk of exposure to Bisphenol A (BPA). *Rocz Panstw Zakl Hig*. 2015;66(1):5-11. Accessed August 5, 2022. https://pubmed.ncbi.nlm.nih.gov/25813067/

157. Zimmers SM, Browne EP, O'Keefe PW, et al. Determination of free Bisphenol A (BPA) concentrations in breast milk of U.S. women using a sensitive LC/MS/MS method. *Chemosphere*. 2014;104:237-243. doi:10.1016/j.chemosphere.2013.12.085

158. Lehmler H-J, Liu B, Gadogbe M, Bao W. Exposure to Bisphenol A, Bisphenol F, and Bisphenol S in U.S. Adults and Children: The National Health and Nutrition Examination Survey 2013-2014. *ACS omega*. 2018;3(6):6523-6532. doi:10.1021/acsomega.8b00824

159. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol 2017 361*. 2017;36(1):89-94. doi:10.1038/nbt.4042

160. Cheng M, Michalski S, Kommagani R. Role for Growth Regulation by Estrogen in Breast Cancer 1 (GREB1) in Hormone-Dependent Cancers. *Int J Mol Sci*. 2018;19(9). doi:10.3390/IJMS19092543

161. Hodgkinson K, Forrest LA, Vuong N, Garson K, Djordjevic B, Vanderhyden BC. GREB1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene 2018 3744*. 2018;37(44):5873-5886. doi:10.1038/s41388-018-0377-y

162. Haines CN, Braunreiter KM, Mo XM, Burd CJ. GREB1 isoforms regulate proliferation independent of ERα co-regulator activities in breast cancer. *Endocr Relat Cancer*. 2018;25(7):735-746. doi:10.1530/ERC-17-0496

163. Jiang S, Katayama H, Wang J, et al. Estrogen-Induced Aurora Kinase-A (AURKA) Gene Expression is Activated by GATA-3 in Estrogen Receptor-Positive Breast Cancer Cells. *Horm Cancer*. 2010;1(1):11. doi:10.1007/S12672-010-0006-X

164. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Mol Cancer 2021 201*. 2021;20(1):1-27. doi:10.1186/S12943-020-01305-3

165. Ma XF, Zhang J, Shuai HL, Guan BZ, Luo X, Yan RL. IKKβ/NF-κB mediated the low doses of bisphenol A induced migration of cervical cancer cells. *Arch Biochem Biophys*. 2015;573:52-58. doi:10.1016/J.ABB.2015.03.010

166. Sauer SJ, Tarpley M, Shah I, et al. Bisphenol A activates EGFR and ERK promoting proliferation, tumor spheroidformation and resistance to EGFR pathway inhibition in estrogen receptor-negativeinflammatory breast cancer cells. *Carcinogenesis*. 2017;38(3):252-260. doi:10.1093/CARCIN/BGX003

167. Li S, Wang B, Tang Q, Liu J, Yang X. Bisphenol A triggers proliferation and migration of laryngeal squamous cell carcinoma via GPER mediated upregulation of IL-6. *Cell Biochem Funct*. 2017;35(4):209-216. doi:10.1002/CBF.3265

168. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia (United States)*. 2017;19(8):649-658. doi:10.1016/j.neo.2017.05.002

169. Yang N, Wang C, Wang Z, et al. FOXM1 recruits nuclear Aurora kinase A to participate in a positive feedback loop essential for the self-renewal of breast cancer stem cells. *Oncogene*. 2017;36(24):3428-3440. doi:10.1038/ONC.2016.490

170. Eterno V, Zambelli A, Villani L, et al. AurkA controls self-renewal of breast cancer-initiating cells promoting wnt3a stabilization through suppression of miR-128. *Sci Reports 2016 61*. 2016;6(1):1-13. doi:10.1038/srep28436

171. Tan Y, Wang Q, Xie Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. *Int J Oncol*. 2019;54(1):87-97. doi:10.3892/ijo.2018.4598

172. Jian W, Deng XC, Munankarmy A, et al. KIF23 promotes triple negative breast cancer through activating epithelial-mesenchymal transition. *Gland Surg*. 2021;10(6):1941. doi:10.21037/GS-21-19

173. Li Z, Yang HY, Zhang XL, et al. Kinesin family member 23, regulated by FOXM1, promotes triple negative breast cancer progression via activating Wnt/β-catenin pathway. *J Exp Clin Cancer Res*. 2022;41(1):1-15. doi:10.1186/S13046-022-02373-7/FIGURES/9

Chapter 5

Conclusions and Future Directions

Summary of Research Findings

In this study, we showcase the utility of CR as a biologically relevant *ex vivo* model of the normal breast and highlight the computational power of multi-omic analyses. Our findings highlight the precision of high resolution single cell transcriptomics and its utility in characterizing distinct differences in cell types (luminal and myoepithelial) and cell states (hybrid cells, cell cycle). Additionally, we characterize inter-individual transcriptomic differences across genetically diverse individuals and in response to bisphenol exposure.

In chapter 2 we identify hybrid cell types in normal mammary tissue, conditionally reprogrammed mammary tissue, and the mouse mammary gland. These hybrid L/B and E/M cells are distinct populations but are both characterized by a stem-like developmentally immature phenotype. The emergence of hybrid populations post-CR and during pregnancy in the mouse mammary gland suggest that these hybrid states are inducible and transient. This transiency provides evidence that hybrid cells may be instrumental in breast remodeling during pregnancy and lactation.

Our integrated multi-omic analysis in chapter 3 revealed that there are no strikingly clear race associated differences in transcriptomic expression and DNA methylation between AA and EA mammary cells that can be deconvoluted from inter-individual and cell type heterogeneity. This suggests that genetic ancestry rather than self-identified race may be a better predictor of breast carcinogenesis and outcomes. We found that we were able to integrate DNA methylation and genotyping data to give us potential mechanistic insight into a handful of genes of interest, although this only accounts for a small fraction of genes we observe to be differentially expressed. The highly variable inter-individual differences in gene expression, DNA methylation, and eQTL genotypes we report highlight the pressing need to use and develop better model systems which are able to capture the genetic variation of diverse individuals.

Finally, our results in chapter 4 show that BPA, BPF, and BPS elicit transcriptomic alterations on mammary cells across a range of human relevant doses. The magnitude and direction of transcriptomic alterations varied by dose, cell type, and by chemical. This is exemplified by the results of overlapping DEGs from TCGA tumors between AA and EA women and the DEGs from 25uM bisphenol exposure. For luminal cells, 25uM BPA was the only chemical which shared a DEG overlap with the TCGA AA tumors in gene *DEFB1*. In myoepithelial cells, 25uM BPA had 7 overlaps (*CLN8, COL14A1, LPCAT2, PLA2G4A, POLR2J3, TINAGL1, TSHZ2*), BPF had 2 overlaps (*COL14A1, LPCAT2*), and BPS had 6 overlaps (*AURKB, CDH1, CDK1, DEFB1, TINAGL1, TSFM*). These differences in number and identity of differentially expressed genes illustrate the varying effects of each bisphenol. Additionally, their overlaps with DEGs from AA tumors provide promising gene targets to explore further. In conclusion, mammary cells from genetically diverse individuals exhibited differences in transcriptomic response to bisphenol exposure, demonstrating the utility of the CR *ex vivo* system as a biologically relevant model of precision toxicology.

Relevance to Human Health

Breast cancer is the most commonly diagnosed in the world and affects individuals at every corner of the globe. These individuals exhibit vast genetic variation and are exposed to distinct combinations of environmental exposures. Our findings highlight that genetic ancestry and inter-individual differences highly influence an individual's transcriptomic response to environmental exposures and may even influence their risk of cancer incidence and survival outcomes.

These findings also add to the growing body of evidence that bisphenols are relevant environmental chemicals to study in relation to breast cancer risk and that analogues of BPA, which were introduced as replacements, elicit distinct, potentially worse, transcriptomic alterations on mammary cells. In particular, BPS may be especially potent at low doses. While traditional toxicology relies heavily on measures such as no observed adverse effect levels (NOAELs) and lowest observed adverse effect levels (LOAELs), given that the low dose of BPS induces a proliferative and cancer stem cell signature, whereas the higher doses appear to downregulate this, reinforces the importance of characterizing dose specific transcriptomic alterations in order to distinguish differences in lowest observed adverse effect levels and the

functional dose of chemicals which actually elicit the lowest adverse effects. Additionally, NOEALs and LOELs are often determined using model systems which are limited in capturing inter-individual differences in toxicant response. Given the vast range and magnitude of human exposure to environmental chemicals and their increased distribution, the pressing need for better of models of precision toxicology has never been more imminent. Expanding the use of precision toxicology models across a range of environmental exposures and characterizing their effects could have major implications for reducing disparities in breast cancer and overall cancer incidence.

Impact and Innovation

To our knowledge we are the first group to perform an integrated multi-omic analysis of normal mammary cells cultured in CR conditions. Additionally, we used advanced computational approaches and integrated publicly available data (TCGA, GTEx, UALCAN) into our analysis pipelines. Our bisphenol experiment also involved a novel multiplexing technique with pooled cells from multiple individuals which we deconvoluted downstream. Our establishment of the CR system for normal mammary cells and successful transcriptomic characterization of genetically diverse individuals will be expanded to test other chemicals.

Limitations

One limitation of this study is that while we are aiming for biologically relevant models of the breast, the CR system is not "true normal". Culturing primary cells in conditions they would not normally want to grow in requires significant alterations to occur in them to be established. Additionally, we observe major differences in cell type proportions by individual and based on the source of our samples we are unable to determine if these proportions reflect the overall composition of an individual's breast or if they represent an expansion of the few cells we acquire after a punch biopsy dissociation. Lastly, we were unable to pick up significant race associated transcriptomic differences. This limitation could be due to our sample size being not high powered enough to detect alterations.

Recommendations for Future Research

The CR model system provides a promising model for precision toxicology and should be expanded for use across more chemicals and mammary cells for individuals. Based on the

inter-individual heterogeneity of our samples, future experiments comparing the AA and EA groups should target a sample size of 20-25 per group in order to be powered enough to detect differences across such diverse individuals. Future work could also be performed to assess the effect of mixtures on normal mammary cells, as these are often the form that environmental exposures take. In order to gain mechanistic insight into the response of mammary cells to bisphenols, DNA methylation could be performed. Finally, techniques such as high content imaging can also be used to further characterize morphometric alterations.