Tapsi Kumar

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations

Part of the Bioinformatics Commons, Computational Biology Commons, and the Genomics Commons

Recommended Citation
Kumar, Tapsi and Kumar, Tapsi, "INVESTIGATING THE ECOLOGY AND EVOLUTION OF NORMAL BREAST TISSUES AND BREAST CANCER WITH SINGLE CELL GENOMICS" (2022). The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access). 1196. https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1196

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.
INVESTIGATING THE ECOLOGY AND EVOLUTION OF NORMAL BREAST TISSUES AND BREAST CANCER WITH SINGLE CELL GENOMICS

by

Tapsi Kumar, B.S., M.S.

APPROVED:

Nicholas Navin, Ph.D.
Advisory Professor

Andrew Futreal
Andrew Futreal, Ph.D.
Co-Advisory Professor

Ken Chen, Ph.D.

Bora Lim, M.D.

Clinton Yam, M.D.

APPROVED:

Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences
INVESTIGATING THE ECOLOGY AND EVOLUTION OF NORMAL BREAST TISSUES AND BREAST CANCER WITH SINGLE CELL GENOMICS

A

DISSERTATION

Presented to the Faculty of
The University of Texas MD Anderson Cancer Center
UTHealth Graduate School of Biomedical Sciences

of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by

Tapsi Kumar, B.S., M.S.

Houston, Texas

August 2022
COPYRIGHT

All the work presented in this thesis is my own work earned through blood, sweat and tears and therefore I give permission to reproduce with accreditation.
DEDICATION

To my loving parents, grandparents and my pillar of support, my husband, who encouraged me to pursue a PhD and endured my incessant uncertainties and cribbing and still kept me sane, I dedicate all my work to you.
ACKNOWLEDGMENTS

I whole heartedly thank my mentor Nick Navin, for being my scientific role model, for giving me the best projects, teaching me how to do good science, shaping my scientific curiosity by his unnerving enthusiasm, always having my back and making me the scientist I am today. I am immensely thankful to my co-mentor Andy Futreal for giving me the opportunity to work in such a large-scale project, believing in me for taking it to completion, also having my back and supporting me in front of our collaborators and instilling the confidence in me to present my work. I couldn’t have asked for such amazing mentors.

Throughout my PhD, I have been very lucky to have an extremely supportive advisory committee, Ken Chen, whom I thank for teaching me computational method development and always giving useful inputs to further refine my project, Bora Lim who not only taught me the basic breast biology but also made me think from a clinical lens and always was super supportive and caring during these past years. I sincerely thank Clinton Yam for providing useful inputs in the ARTEMIS project and for agreeing to join my committee to help me with my PhD. I am also thankful to past members of my committee Arvind Rao and Anshumali Shrivastva for always asking very inquisitive questions and helping me think from a purely computer science perspective.

I am immensely grateful to Shanshan Bai and Min Hu with whom I started my journey of the breast cell atlas project, who taught me my first single cell dissociation, spent odd hours doing those terrible time course experiments, helped me with my never ending linux issues and were lovely travel buddies with whom I had a lot of fun in Boston and NY and good times we had that I would rather not mention here – “What happens in NY, stays in NY”.

v
I would like to thank Jake Leighton with whom I joined the lab and who always has been a very good friend as we always ended up getting similar fellowships and talked nonstop about our furry friends.

So many amazing lab members Charissa Kim, Alex Davis, Aislyn Schalek, Naveen Ramesh, Yun Yan, Runmin Wei, Darlan Minussi, Siyuan He with whom I have had morning, afternoon and evening coffee breaks and fun times outside the lab. I thank Emi Sei for all the fun and chirpiness we had in the lab, from filling up Nick’s room with balloons to surprise decorating other’s benches in pink tape and for the confetti I still find in my pockets. All other lab members, past and present and growing, have been like a second family.

To the people with whom I closely worked with in Navin Lab - Yiyun Lin, Yun Yan (Team TNBC), Kaile Wang and Kris Wang (Team DCIS), thank you for being awesome project mates and friends. I would also like to thank Anna Casasent for being a good friend, mentor, a good listener and being so caring.

I am grateful to members of Futreal lab and Zhang lab, Hannah Beird for giving me my first project during my rotation and always patiently answering all my questions, past members Kadir Akdemir (now an amazing PI) for guiding me in choosing my thesis mentors and motivating me to think about cancer research during my time at Rice, Latasha Little for being so efficient in managing all the samples shipped from across the world and helping me learn the experimental details and all other members John, Henry, Mingchu, Per and others for their useful inputs during lab meetings.
I have learnt so much from my collaborators at Houston, I thank Dr. Thompson for always being so supportive and positive, Dr. Kim for explaining me the clinical aspects of PDAC, Dr. Fei Yang and Dr. Edwin-Parra for their unwavering support in the VECTRA projects, Dr. Vinod Ravi for being so supportive and introducing me to the clinical side of Sarcoma and TLS projects, and Dr. Litton for giving me an opportunity to work in the TALA project and external collaborators at UC Irvine, Kai Kessenbrock, Devon Lawson and Kevin Nee for their tireless support in the breast atlas project.

I would like to thank members of Team PRECISION at the Netherlands, UK, and US, Jelle Wesseling, Esther Lips, Lodewyk Wessels, Proteti Bhattacharya, Misha Sheinman, Tycho Bismeijer, Mathilde Almekinders, Elly Sawyer, Maria Roamn-Escorza, Serena Nik-Zainal and Shelley Hwang for being amazing collaborators, mentors and coauthors.

I thank my lab rotation advisors, Traver Hart with whom I started my PhD and who encouraged me to pursue doing a rotation in other labs, Jennifer Wargo and Tatiana Karpinets who gave me an opportunity to be involved in my first paper published in Science, Chris Bristow who mentored me during my internship and taught me to be meticulous, organized and acquire excellent presentation skills.

I am very thankful to Lurtese Gardner and Elva Lopez who have always been so cheerful, wonderful and caring and all the administrative staff at the Department of Genetics, Genomic Medicine, T32 and GSBS. All the members of GSBS staff, Dr. Mattox, Brenda, Amy, Karen, Lily, Joy and Bunny who helped me and always supported me to grow as a student. I’m thankful to Dr. Sanjay Shete and Dr. Prahlad Ram for being amazing QS directors.
When I first started my PhD, I received so many scientific inputs from friends at MD Anderson and Rice - Samir Amin who always explained science with such passion and helped me refine my scientific thought process over several evening tea sessions, Kunal Rai who helped me prepare for interviews, Ayush Raman who helped me with my STATS assignment, Shubham Manchanda who always in his own way thinks of me as a Science Super hero working on some Nobel prize worthy projects, Titiksha Vats, Pushpendra Shishodia, Akshat Jariwala, Aparna Subramaniam, Ajay Shastri, Gautam Sikka and Archana Sharma and little Garvin who have made living in Houston so much fun.

None of this would not have been possible without the love and support of Sahil Seth, my best friend since 2007 and husband, who stood with me, hugged me, listened to my nonstop whining and taught me so much. My parents and grandparents whom I have missed so much these past years, and Sahil’s parents and family and our extended families, aunts and uncles and cousins who supported me in my path to getting a PhD. Last but not the least, my fur babies, Twisty, Satchmore, Snoopy and Patchy who make me smile every day.

Finally, this project would not have been possible without the generosity of our funders, the NCI T32 fellowship, Dr. Ann Killary and Dr. Wistuba, Rosalie B. Hite fellowship, Cancer Research UK and the Chan Zuckerberg foundation.

Thank you everyone!
INVESTIGATING THE ECOLOGY AND EVOLUTION OF NORMAL BREAST TISSUES AND BREAST CANCER WITH SINGLE CELL GENOMICS

Tapsi Kumar, B.S., M.S.

Advisory Professors: Nicholas Navin, Ph.D. and Andrew Futreal, Ph.D.

There is vast cellular heterogeneity in human breast tissues, with different transcriptional programs in the stromal, epithelial, and immune components, however, it remains unclear how their reprogramming and interplay leads to the progression of invasive phenotypes such as Triple-Negative Breast cancer (TNBC). To define the microenvironmental alterations that occur during cancer, we first established a human breast cell atlas, a reference of normal breast cell types from disease free women. We profiled 535,941 cells from 62 women and 124,024 nuclei from 20 women revealing 11 major cell types and 52 cell states that reflect different biological functions that can be organized into 4 major spatial domains (adipose, connective, ducts, lobules). We then compared this atlas against one of the more aggressive subtypes of breast cancer, TNBC in which patients lack estrogen, progesterone, and HER2 growth receptors. Approximately 10-14% of all breast cancer patients are classified as TNBC, with a majority developing resistance to neoadjuvant chemotherapy (NAC) and having a 5-year survival of <30%. Tumor evolution and changes in components of the tumor microenvironment (TME) such as cancer associated fibroblasts (CAFs), tumor associated macrophages (TAMs), and tumor endothelial cells (TECs), have shown to play a role in tumor progression. One knowledge gap is the role macroenvironmental changes confer on chemoresistance in TNBC patients. To order to investigate the transcriptional reprogramming of the TME, we performed single cell RNA sequencing on 100 treatment-naïve TNBC biopsies and compared its ecosystem to the normal breast ecosystem captured earlier. Our data suggest reprogramming and increased proportions of certain cell states such as TECs, CAFs, and effector...
T cells enrichment in TNBC and positive correlation with pathological complete response. Aside from transcriptional changes in normal and late stage invasive breast cancers microenvironments, we also investigated the genomic evolution in early stage breast cancer, Ductal Carcinoma In Situ (DCIS). DCIS is the most common form of pre-invasive breast cancer and despite treatment, a small fraction (5-10%) of DCIS patients present with invasive disease many years later. A fundamental question is whether the invasive disease recurring in the same breast is established by tumor cells in the initial DCIS or represents a new unrelated disease. Whole exome sequencing of 24 longitudinally matched DCIS and recurrent invasive breast cancer revealed clonally related recurrence in ~80% patients whereas ~20% pointed towards independent evolution which was also validated by single cell DNA sequencing in a subset of 4 cases. Overall, we established a breast cell atlas that provides an invaluable reference for the research community to understand how normal cell types are reprogrammed in diseases such as TNBC. We established a secondary cell atlas of TNBC patients and compared the two atlases to characterize the transcriptomic features of progression and predictive of pathological complete response in TNBC samples. Lastly, we showed that not all DCIS may be precursors to invasive cancer which paves a way for accurate risk evaluation of DCIS, treatment de-escalation strategies, and the identification of predictive biomarkers.
Table of Contents

Approval Page i
Title Page ii
Copyright iii
Dedication iv
Acknowledgements v
Abstract ix
List of Illustrations xvi
List of Tables xviii

Chapter 1: Introduction

1.1 – Normal breast tissue biology and cell types 1

1.2 – Triple-Negative Breast Cancer 3
   1.2.1 Resistance to chemotherapy as a major clinical problem 5
   1.2.2 New Standard of care for TNBC 6

1.3 – Ductal Carcinoma in situ 7

1.4 – Landscape of Tumor microenvironment
   1.4.1 Stromal ecosystem 9
   1.4.2 Immune microenvironment 10
   1.4.3 Crosstalk between tumor and its microenvironment 11

1.5 – Evolutionary model of Genome evolution in DCIS 13
   1.5.1 Single cell approaches to measure Tumor Evolution 14

1.6 – Computational tools for processing single cell data
   1.6.1 Single cell RNA methods and benchmarking 15
   1.6.2 Single cell DNA Copy Number Profiling Methods 17
Chapter 2: Materials and Methods

2.1 – Materials

2.1.1 Normal Breast tissue study collection protocol
2.1.2 TNBC Tissue Samples
2.1.3 Pure primary DCIS and subsequent recurrence cancer Samples

2.2 – Application of existing single cell sequencing Methods

2.2.1 Single cell RNA sequencing
2.2.2 Single nuclei RNA sequencing
2.2.3 Single cell DNA sequencing

2.3 – Single cell and nuclei RNA Sequencing Data Analysis

2.3.1 Data preprocessing, QC and filtering
2.3.2 Cellular and Nuclei landscape construction
2.3.3 Assigning Cluster Annotation
2.3.4 Determination of Expression states
2.3.5 Cellcycle analysis, Differential Expression and Pathway Enrichment
2.3.6 Regulatory network analysis
2.3.7 Copy number inference from single cell RNA data
2.3.8 Integration of normal and tumor breast single cell data

2.4 – Single cell DNA sequencing Data Analysis

2.4.1 Data preprocessing, QC and filtering
2.4.2 Data clustering and visualization
2.5 – Application of Existing Bulk Next-Generation Sequencing Methods

2.5.1 Whole-Exome Sequencing

2.6 – Whole-Exome Sequencing Data Analysis

2.6.1 Somatic Variant Calling

2.6.2 Germline Variant Calling

Chapter 3: A spatially resolved single cell genomic atlas of the adult human breast

3.1 – Introduction and Rationale

3.2 – Results

3.2.1 Overview of clinical samples profiled

3.2.2 Major cell types in adult human breast tissues

3.2.3 Epithelial Cells of the Ducts and Lobules

3.2.4 Fibroblasts & Mesenchymal Cell Types

3.2.5 Adipose Tissues of the Breast

3.2.6 Vascular & Lymphatic Endothelial Cells

3.2.7 Pericytes of the Breast

3.2.8 Immune Ecosystem of the breast

3.3 – Discussion

Chapter 4: Chemotherapy Response in Triple-negative breast cancer defined by single cell sequencing

4.1 – Introduction and Rationale

4.2 – Results

4.2.1 Overview of the clinical trial study cohort
4.2.2 TNBC single cell atlas 60
4.2.3 Comparison of stromal microenvironment in Normal and TNBC breast tissues 61
4.2.4 Cell of origin of TNBC tumor cells 64
4.3 – Discussion 65

Chapter 5: Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer
5.1 – Introduction and Rationale 66
5.2 – Results
   5.2.1 Overview of clinical samples profiled 68
   5.2.2 Whole exome sequencing to assess clonal relatedness 69
   5.2.3 Single cell genome sequencing to reconstruct clonal lineages 72
   5.2.4 Genomic aberrations in recurrent invasive disease 75
5.3 – Discussion 75

Chapter 6: Discussion and Future Directions
6.1 – Conclusions 78
6.2 – Future Directions
   6.2.1 Integration of single cell atlas 79
   6.2.2 Multiomics view of the disease to infer mechanisms of chemoresponse 80
   6.2.3 Longitudinal profiling of TNBC to study evolution of chemoresistance 80
   6.2.4 Tumor initiation and invasion in DCIS progression 81
   6.2.5 Evolutionary models in DCIS-IDC using single cell sequencing 81
6.2.6 Spatial assays to capture heterogeneity 82

6.2.7 Applications of single cell sequencing to clinical practice 83

6.3 – Limitations 83

Bibliography 87

Vita 103
LIST OF ILLUSTRATIONS

Figure 1 - Anatomy and pathology of the major cell types in normal human breast tissues.

Figure 2 - Illustration showing invasive TNBC microenvironment

Figure 3 - Illustration of a breast duct with DCIS tumor cells

Figure 4 - Evolutionary models of primary DCIS to recurrent disease progression

Figure 5 - Overview and workflow of different breast tissue samples used in generating the atlas and experimental workflow for single cell sequencing.

Figure 6 - Overview and workflow of the TNBC tissues obtained from the pre-treatment biopsies

Figure 7 - Experimental workflow of longitudinally matched DCIS and recurrent DCIS/IDC samples

Figure 8 - Celltypes in an adult human breast

Figure 9 - UMAP clustering of 535,941 single cell RNA data across 100 tissues

Figure 10 - Major cell type proportions of matched left and right breasts from 22 women and averages across all left and all right breasts

Figure 11 - Consensus heatmap of the top 7 genes expressed in each cell type

Figure 12 - UMAP clustering of 124,024 single cell RNA data across 20 tissues

Figure 13 - Consensus heatmap of the top 7 genes expressed in each nuclei type

Figure 14 - H&E image showing the bilayered duct

Figure 15 - UMAP clustering of epithelial cells and nuclei

Figure 16 - Dotplot showing the top keratins expressed in epithelial cell types

Figure 17 - Expression of hormone genes in in epithelial nuclei

Figure 18 – Cellstates of epithelial cells and defining marker genes

Figure 19 – Cellstates of fibroblasts and defining marker genes

Figure 20 - Violin plot showing canonical genes expressed in beige and white adipocytes
Figure 21 - Cellstates of lymphatic endothelial cells and defining marker genes
Figure 22 - Cellstates of vascular endothelial cells and defining marker genes
Figure 23 - Cellstates of pericytes and defining marker genes
Figure 24 - Chemokine signature across pericytes cellstates and top GO gene signatures
Figure 25 - Cellstates of T/NK cells and defining marker genes
Figure 26 - Cellstates of B cells and defining marker genes
Figure 27 - Cellstates of myeloid cells and defining marker genes
Figure 28 - Clinical endpoint of TNBC tissues obtained from the pre-treatment biopsies
Figure 29 - UMAP and top gene heatmap of ~500,000 cells from 101 TNBC tissues
Figure 30 - Heatmap showing top transcription factors in each cell type
Figure 31 - Comparison of fibroblasts in normal and TNBC tissues
Figure 32 - Comparison of pericytes in normal and TNBC tissues
Figure 33 - Comparison of endothelial cells in normal and TNBC tissues
Figure 34 - Ternary plot showing the distribution of tumor cells from TNBC tissues based on the expression of top marker genes. Heatmap shows the expression of the genes across each sample
Figure 35 - Clonality assessment using whole exome sequencing
Figure 36 - Distribution of variant allele frequency of mutations
Figure 37 - Mutational landscape of primary DCIS samples
Figure 38 - Lineage tracing for 4 patients
Figure 39 - Clonal lineage reconstruction by single cell genome sequencing of unrelated pairs
Figure 40 - Mueller plots showing clonal frequencies
Figure 41 - Phylogeny reconstruction by single cell genome sequencing
Figure 42 - Clonal lineage reconstruction by single cell genome sequencing of related pairs
LIST OF TABLES

Table 1 - Summary of popular single cell RNA computational tools

Table 2 - Clinical Meta data of all normal breast samples

Table 3 - Clinical Meta data of all TNBC breast samples

Table 4 - Clinical characteristics of the DCIS samples
CHAPTER 1: INTRODUCTION

1.1 Normal Breast biology and cell types

The human breast is a glandular organ that consists of a very intricate network of ducts and lobules that play a physiological role in producing and transporting milk to offspring. Further during the lifetime of a woman, it undergoes physiological changes typically controlled by hormonal regulation. The mammary epithelial system is embedded into a fat rich tissue supported by a dense layer of vasculature and lymphatic vessels that provides nutrition and protection (Figure 1).

![Figure 1: Anatomy and pathology of the major cell types in normal human breast tissues.](image)

The mammary epithelium consists of two distinct cell types that comprise the epithelial bilayer, an inner layer of luminal cuboidal epithelial line the lactiferous ducts and an outer layer of myoepithelial cells in direct contact with the basement membrane. These ducts expand from the nipple and dilate into the lactiferous sinuses just beneath the areola. Lactiferous ducts further subdivides into lobes and lobules, of which the adult female mammary gland has 15 to 20 and 20
to 40, respectively. Each lobule ends in small bulb-like glands known as terminal ductal lobular units, wherein milk gets produced in response to prolactin. The myoepithelial cells are spindle-shaped and have smooth muscle cell properties and participate in the process of milk ejection during lactation. Together, the epithelial components of the mammary gland constitute 10 to 15% of its overall volume (Duivenvoorden et al., 2017).

Apart from epithelial cells, other cells such as adipocytes are a major component of the breast. The volume of adipose tissue leads to variation in the breast size amongst women rather than the epithelial component itself. Adipocytes are found abundant in the interlobular areas, while its presence in the intralobular stroma is scarce. Adipocytes play a role in regulating mammary gland homeostasis, as it is involved in epithelial growth, intercellular communication, angiogenesis, and milk production. It also acts as a reservoir of interstitial fluids such as prolactin etc (Pandya and Moore, 2011).

Fibroblasts are also abundantly found in the breast stroma. They play a role in growth factor synthesis, metalloproteinase (MMP) production, and extracellular matrix (ECM) deposition (Inman et al., 2015). The ECM consists of different types of collagens, fibronectin, proteoglycans, hyaluronic acid and tenascins that supports epithelial survival by suppressing apoptosis, with the help of MMP-2. The ECM also helps with tumor suppression because of the intact basement membrane that are thin, 100-nm sheets of glycoproteins and proteoglycans, laminins, and collagen IV fibrils (Muschler and Streuli, 2010).

The breast is a highly vascular organ comprising of vascular and lymphatic cells that work in close association with the mammary epithelial structures. They play a role in angiogenesis and supplying nutrients to the breast (Betterman et al., 2012). Immune cells play a role in immune-surveillance and ductal morphogenesis, especially macrophages, mast cells, and eosinophils. Mast cells also serve during mammary involution.
Delineating the cellular composition of breast has been challenging since it required reproducible protocols. In the past, histopathology has identified the major cell types in the breast, but a deeper understanding of different cellular expression states has not been studied well. First, the time to dissociation of the breast tissue into single cells needs to be fairly minimum so as to minimize perturbation of cellular features. Second, to find rare subpopulations present at the 1% frequency needs 1000’s of cell to be sequenced. The characterization of the breast tissue is even more challenging as it undergoes changes due to life events such as menopause, postnatal development etc. Recent studies have used single cell sequencing to study the epithelial cells in the breast in both mice and humans but have not extensively looked at other cell types (Nguyen et al., 2018).

My thesis is overcoming this challenge and aimed at creating the first and largest human breast cell atlas at a single cell resolution in order to capture all the cells and expression cell states present in an adult human breast.

1.2 Triple Negative Breast Cancer

Triple Negative Breast Cancer (TNBC) is a highly aggressive form of invasive breast cancer that accounts for 20% of all newly diagnosed cases. It is defined by the lack of estrogen, progesterone and HER2 receptors (Hudis and Gianni, 2011). As compared to DCIS, the tumor cells in TNBC acquire a migratory and invasive phenotype via epithelial-to-mesenchymal transition (EMT) and break the basement membrane and invade the whole breast (Figure 2) (Ahmed and Betz, 2015).
TNBC microenvironment (TME) refers to the ecosystem in which the tumor forms, grows, and metastasizes along with its internal tumor cell. The TME consists of immune cells such as T, B and myeloid cells that have intrinsic immunogenicity. Complex mechanisms of disease progression and tumorigenesis have been studied by focusing on the TME. TME plays an important role in all stages of tumor progression, including tumor growth, immune modulation and chemo resistance (Janiszewska et al., 2019). Intratumor heterogeneity is defined by the phenotypic and molecular diversity within the tumor which can be due to the differentiation state of the cancer cell of origin, mechanisms of cell plasticity, dynamic genomic aberrations, and genomic evolution and positive selection (Keren et al., 2018). Further, it also guides the identification of new biomarkers in stromal components in order to predict clinical outcome and
guide therapy in TNBCs. Apart from tumor cells, the TME consists of extracellular matrix (ECM) and other stromal cell types, including endothelial cells, immune cells, fibroblasts, and adipocytes.

It has been known that the interactions between cancer cells and their surrounding ecosystem plays a role in tumor progression and conferring resistance. Therefore, understanding the unique transcriptional profiles of breast cells, the differential expression of genes, and the role of surrounding stroma and tumor cells that distinguish TNBC from normal breast, will help understand how the tumor microenvironment may contribute to tumor progression and poorer prognosis of TNBC and even enable treatments to be more effectively tailored to patients.

1.2.1 Resistance to chemotherapy as a major clinical problem

TNBC with therapeutic resistance is very aggressive, has only 50% pCR compared with other breast cancer subtypes (Bianchini et al., 2016), and is the major challenge of TNBC treatment. The standard of care for most TNBC patients is neoadjuvant chemotherapy (NAC), a mixture of taxanes (mitotic inhibitors) and anthracyclines (DNA intercalators). Around 30%–50% patients following NAC evolve resistance, leading to poor overall survival (Foulkes et al., 2010, Liedtke et al., 2008). The underlying basis for chemoresistance in TNBC patients remains poorly understood, because of lack of high-resolution single cell methods that can resolve intratumor heterogeneity and identify rare resistant cell populations. A major study resolved this ITH by subtyping TNBC into seven groups based on their transcriptional profiles, namely basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), luminal androgen receptor (LAR), and unstable (UNS), each of which have different clinical response (Lehmann et al., 2011). A study from our lab used single-cell DNA, whole-exome sequencing and RNA-seq to in 20 primary TNBC patients to study genome evolution following NAC. This work focused on 3 groups of patients - clonal extinction (those who achieved
pCR) vs. clonal persistence (non responders) patients defined by whether somatic mutations (nonsynonymous mutations and indels) were detectable following NAC. This data suggested that at the genomic level, chemoresistance evolution occurs through the selection of pre-existing point mutations and copy number changes suggesting adaptive resistance, while chemoresistance evolution at the transcriptomic level occurs through transcriptional reprogramming in signaling pathways associated with chemoresistance, suggesting acquired resistance (Kim et al., 2018).

1.2.2 The New Standard of care for TNBC

TNBC is chemotherapy sensitive and the standard of care for a new diagnoses include NAC with anthracycline (e.g., DNA intercalating agent), alkylating agents (e.g., cyclophosphamide), an anti-microtubule agent taxane, and an anti-metabolite fluorouracil (5-FU), followed by surgery. However, for relapsed/refractory TNBC patients, there is no standard chemotherapy regimen. For late stage TNBC available therapies include anti-metabolites capecitabine and gemcitabine, non-taxane microtubule inhibitor eribulin, and platinum based DNA cross-linkers. For chemo insensitive patients, available targeted therapies include angiogenesis inhibitors (bevacizumab), PARP1 inhibitors, EGFR inhibitors, tyrosine kinase and ERK inhibitors, and mTOR inhibitors. When surgery cannot be done, such patients have newer treatment options available.

TNBC has high levels of TILs in its TME, but also expresses high level of programmed cell death-ligand 1 (PD-L1)[]. Emerging studies targeting immune microenvironment, specifically the PD-1/PD-L1 ligand-receptor pathway have been explored and atezolizumab (anti-PD-L1 antibody) in combination with nanoparticle albumin-bound (nab)-paclitaxel was approved as a first-line therapy (standard of care, SOC) by the US Food and Drug Administration (FDA) based on the IMpassion130 trial (NCT02425891) in 2019. This immunochemotherapy became SOC for patients with PD-L1+, unresectable, locally advanced or metastatic TNBC. However,
Pembrolizumab, PD-1 inhibitor was approved by FDA for patients with high-risk, early-stage, TNBC as SOC in combination with NAC, and then continued as a single agent as adjuvant treatment after surgery [FDA approves pembrolizumab for high-risk early-stage triple-negative breast cancer. FDA. Updated July 27, 2021. Accessed November 5, 2021. bit.ly/3mQntFK]. The approval was based on findings from the phase 3 KEYNOTE-522 trial (NCT03036488), which showed that patients treated with pembrolizumab plus chemotherapy (n = 784) experienced pCR rate of 63% vs 56% in patients receiving placebo plus chemotherapy (n = 390) [Keytruda. Prescribing information. Merck; 2021. Accessed November 5, 2021. bit.ly/3bIYDRB].

1.3 Ductal Carcinoma in situ

Ductal Carcinoma in situ is an early stage or stage 0 breast cancer in which the tumor cells are only confined inside the mammary gland ducts (Figure 3). With the advent of highly sensitive mammography screening techniques, it now represents for ~20% of all newly diagnosed breast cancer cases across the world, with over 60,000 cases in the US alone, ~7000 in the UK and ~2500 in the Netherlands (Shee et al., 2019).

Figure 3: Illustration of a breast duct with DCIS tumor cells
The epidemiological (Kerlikowske, 2010) and genetic predisposition to developing DCIS are similar to the ones for invasive breast cancer such as age, pregnancy and live birth status, family history, high breast density, BRCA1/2 mutation (Mazzola et al., 2013).

The treatment for DCIS varies across countries with surgeons in the US preferring to suggest the patients to undergo a combination of wide local excision (WLE) and radiation (47%) or a mastectomy (28%) or a WLE only (22%) (Worni et al., 2015), whereas in the Netherlands 48% of women with mastectomy, 26% with WLE only, and 26% with WLE + RT (Elshof et al., 2018). However, it has been shown that aggressive treatment regimens may not be needed since the overall survival of low grade DCIS patients with or without breast conserving surgery is the same (Narod et al., 2015).

Hormonal therapy, such as Tamoxifen (Estrogen receptor inhibitor) is rarely given to post menopausal women with DCIS in many countries due to reports of a number of side effects and non-significant reduction in the overall survival (Maxwell et al., 2018). The current dilemma in the clinical management of low grade DCIS is whether the different types of treatments and surgery actually reduces the incidence of recurrence of invasive breast cancer and impact overall survival or if active surveillance should be used to monitor the DCIS lesions (Francis et al., 2015; Hwang et al., 2019).

A fundamental biologic question is whether the invasive disease arises from tumor cells in the initial DCIS or represents new unrelated disease. This has been addressed by several studies using pathological and immunohistochemical assays by showing concordance based on nuclear grade (Bijker et al., 2001), p53 status and ER/HER2 staining (Karlsson et al., 2014). However, these assays are limited by batch affects in assessing the features, inconsistent grading due inter-observer reproducibility and the fact that the above-mentioned features are quite common in most DCIS tumors. Limited genomic studies have been done to assess clonality between longitudinally
matched DCIS and recurrent disease. There is significant gap in knowledge in understanding if the initial DCIS and subsequent ‘recurrence’ share a common genetic lineage, or alternatively represent independent diseases that emerge from different initiating cells in the same breast. To remedy this problem and establish the genetic lineage, I performed extensive genomic characterization of pure DCIS with matched recurrent tumors and validated the results using single cell DNA sequencing.

1.4 Landscape of Tumor microenvironment

1.4.1 Stromal ecosystem

The tumor is very complex tissue comprising of a variety of cells such as mesenchymal cells like fibroblasts, cells of the vasculature like endothelial cells, pericytes and adipocytes. Past studies have shown that these cells are recruited by the tumor cells to build the tumor microenvironment to help them proliferate, invade and even shape its response to treatment (Langerud et al., 2018).

Fibroblasts are an important component of the tumor stroma in TNBC that supports the production of ECM. In early stages of tumor progression activated fibroblasts acts as a barrier to inhibit the progression via gap junctions, however these later get primed by tumor cells and transform into cancer associated fibroblasts (CAFs) that express high levels of fibroblast activating protein (FAP), smooth muscle actin (ACTA2), vimentin (VIM) (Leonardi et al., 2012). The origin of CAFs is still not understood very well but increasing literature evidence supports its lineage from resident fibroblasts, mesenchymal stem cells, epithelial cells via EMT and even endothelial cells via endothelial-mesenchymal transition (Shiga et al., 2015). They play a role in tumor initiation, progression, and local invasion (Mao et al., 2013) via angiogenesis, remodeling of the ECM, cancer-associated inflammation and the regulation of epithelial cell differentiation (Luo et
al., 2015). In TNBC progression, tumor cells activate normal resident fibroblasts by secreting transforming growth factor (TGF-beta) and platelet derived growth factor (PDGF) and converting them into CAFs (Yu and Di, 2017). Moreover, another study reported the activation of CAFs to further recruit macrophages via chemokine CXCL16 expressed by macrophages in TNBC (Allaoui et al., 2016). Podoplanin, found in lymphatic endothelial cells also has high expression in CAFs and was associated with higher grade and TNBC (Niemiec et al., 2014). These cells also recruit endothelial progenitor cells into the tumor core to stimulate angiogenesis via secretion of stromal cell derived factor (SDF-1) (Orimo and Weinberg, 2006). The role of CAFs in drug resistance possibly by enhancing the physical barrier and limiting the drug delivery to tumor core has been reported.

Endothelial cells provide nutritional and metabolic support for the growth and expansion of tumors. Recent studies report the role of surface protein Vascular endothelial growth factor (VEGF), i.e highly dysregulated in TNBC (Bender and Mac Gabhann, 2013), on endothelial cells to affect tumor cells by eliciting a pro-angiogenic response (Yoshiji et al., 1996). Further, increased endothelial cell retraction leads to increased adhesion of tumor cells further leading to local invasion into the endothelial monolayer, and tumor cells can also further cause contraction of endothelial cells (Lee et al., 2003). In TNBCs, endothelial cells via VEGF and PDGF can lead to formation of vessels around the tumor (Plantamura et al., 2014). Expression of VEGF is also reported to be higher in TNBC than in other invasive tumors (Linderholm et al., 2009).

1.4.2 Immune microenvironment

The tumor immune microenvironment (TiME) consists of immune cells such as T, natural killer (NK), B and myeloid cell types that play a role in immune surveillance and immune escape leading to tumor progression, invasion and metastasis of TNBC. The increase in tumor antigen
expression leads to immune system recognition and tumor infiltrating lymphocytes play a role in elimination of tumor cells and removal, however, high expression of immune checkpoints also further inhibits immune cell function and leads to tumor immune escape (Santoni et al., 2018).

Tumor infiltrating lymphocytes (TILs) have been widely studied in TNBC due to their pronounced infiltration in the invasive tumor margin vs the tumor core which has been attributed to high rate of somatic mutations leading to increased neoantigens and heightened immune response (Stanton and Disis, 2016). TILs invading the tumor regions mainly consists of CD4 T helper cells, CD4 T regulatory cells and cytotoxic and effector CD8 T cells. CD4 T regulatory cells play a role in sustaining immunological tolerance and immune suppression. They act by inhibiting the activation and differentiation of CD4+ Th cells and CD8+ cytotoxic T cells (Shang et al., 2015). TNBCs usually have higher average number of CD4 T regulatory cells than other breast cancer subtypes (Stanton et al., 2016). CD8+ cytotoxic T cells are the key effector cells that mediate effective anti-tumor immunity, leading to better prognosis (Huang et al., 2015). In TNBC, following the adjuvant chemotherapy, higher levels of TILs have been associated with better clinical outcomes (Tian et al., 2016).

Tumor associated macrophages (TAMs) play a dual role in host defense system depending on their polarized phenotype. Macrophages are recruited by tumor cells via secretion of cytokines and chemokines and they migrate to the tumor region to form a tumor microenvironment. The classical macrophage (M1) release proinflammatory cytokines to induce a Th1 immune response to promote an inflammatory response. The anti-inflammatory macrophages (M2) are associated with tumor progression, and actively respond to stimuli such as interleukin10 (IL-10) and TGF-β to further inhibit the Th1 immune response leading to tumor invasion. Furthermore, M2 macrophages also can promote angiogenesis by secreting VEGF. TNBC cells secrete cytokines such as IL-2, IL-10, TGF-β and granulocyte colony stimulating factor (G-CSF) to promote the
transformation of M1 into M2. High levels infiltrating TAMs, based on CD68 as the identifiable marker, correlate with high levels of IL-6 and CCL5 and are associated with bad clinical outcomes in TNBC (Yuan et al., 2014).

1.4.3 Crosstalk between tumor and its microenvironment

These close interactions between intrinsic and extrinsic components of the microenvironment is characteristic of the general microenvironment of TNBC and contributes to progression and chemoresponse. Interactions between stromal cells such as fibroblasts and tumor cells are associated with bad prognosis in TNBC patients. CAFs are associated with immune cell infiltration either directly via secreted chemokines, cytokines and surface proteins or indirectly via ECM deposition and matrix remodeling which affects the migration of immune cells (Yamauchi et al., 2020). One study reported the different CAF phenotypes and its association with signatures predictive of tumor immune evasion, with a low inflammatory CAF signature in TNBC samples patients having high cytotoxic CD8 T cell infiltration having better prognosis (Wu et al., 2020). However, CAF heterogeneity and its phenotypes, possible functions, and relationships with the immune infiltrate in TNBC is still limited by studies with small sample sizes.

Colony-stimulating factors (CSFs) are important regulators of myeloid cell lineage affecting their proliferation, activation and differentiation. Tumor cells can upregulate the expression of these CSFs via the mTOR pathway further promoting the migration and infiltration of immunosuppressive myeloid derived suppressor cells (MDSCs) creating an immunosuppressive TME (Welte et al., 2016). Tumor cells express PD-L1, whereas T cells express their receptors Programmed Death 1 (PD-1) and the Cytotoxic T lymphocyte antigen 4 (CTLA-4) are able to evade immune destruction is mediated by cell-cell contact. In normal conditions, PD-1 is expressed on T and B lymphocytes, providing peripheral tolerance and protection against
autoimmunity, while its ligand PD-L1 is mainly expressed on the surface of antigen-presenting cells. In cancers, the cells can acquire the capability to overexpress PD-L1 and PD-L2. Although the mechanism is not completely understood, the PD-1/PD-L1/PD-L2 axis is able to induce anergy and/or apoptosis of PD-1+ T cells, attenuating the anti-tumor immune response and promoting Treg immunosuppressive activity (Barriga et al., 2019).

At present, research on the TNBC immune microenvironment is mainly limited to TILs, TAMS, and CAFs, while few small-scale studies are looking into tumor associated adipocytes, NK cells, tumor associated neutrophils and their functional characteristics such as the distribution, colocalization, and interaction of immune cells and cytokines in the TNBC microenvironment.

1.5 – Evolutionary models of genome evolution in DCIS

Three models of tumor progression have been proposed namely, independent evolution, evolutionary bottleneck and multiclonal invasion. The independent evolution model suggests that two different initiating cells in normal breast tissue give rise separately to recurrent tumor subpopulations. The evolutionary bottleneck posits that subclones diverge and expand in ducts after which a single clone

**Figure 4:** Evolutionary models of primary DCIS to recurrent disease progression.
is selected and remains dormant until it expands to form the recurrent disease many years to decades late. Additionally, the multi-clonal invasion, suggests that multiple subclones escape the ducts and establish the invasive recurrent disease.

To date, these models have mainly been informed by bulk genomic studies of single time point samples using synchronous DCIS-IDC tissues, in which invasion has already occurred or unmatched DCIS-IDC pairs (Bergholtz et al., 2020), where interpatient heterogeneity can confound genomic comparisons. In one previous study, it was reported that multi-clonal invasion was common in synchronous DCIS-IDC [PMID: 29307488], but it remains unclear if this model is also common in pure DCIS cases that have recurrent IDC cancer many years later. A few other studies have shown evidence that most recurrences have a clonal genetic relationship with their matched primary DCIS pairs (Trinh et al., 2021). In one of the largest bulk genomic studies using a cohort of matched DCIS and recurrent invasive cancers, it was estimated that 75% of patients share a direct genetic lineage (Lips et al., 2021) but this study did not distinguish between the two evolutionary models of progression. Moreover, many of these previous studies were limited to ‘bulk’ genomic methods that could not accurately resolve intratumor heterogeneity (ITH) in DCIS and infer models of genomic evolution in matched cases of pure DCIS with recurrent DCIS/IDC cancers.

1.5.1 Single cell approaches to measure genome evolution

Over the last decade, the field of single cell genomics has undergone rapid progress since the development of the first single cell DNA sequencing (scDNA-seq) methods to profile copy number in human tissues (Navin et al., 2011). While initial methods were based on whole-genome-amplification (WGA) chemistries and were limited to profiling a few cells at a time (Zong et al., 2012), the development of microdroplet, nanowell and combinatorial indexing methods have
vastly increased cell throughput to thousands of cells, while greatly reducing costs (Minussi et al., 2021). However, these methods still require fresh tissue samples or snap-frozen samples, which prevents their application for analyzing archival FFPE tissue samples. This issue represents a major technical barrier, since most clinical tissues with long-term patient outcome data have been stored as FFPE blocks by pathologists at hospitals and universities around the world. The problem is that FFPE preservation leads to formalin-induced DNA-protein crosslinking, as well as double-stranded breaks leading to small fragmented DNA molecules (Hoffman et al., 2015), that are difficult to amplify with existing WGA based methods. While one previous study using a first-generation DOP-PCR based chemistry has demonstrated initial feasibility in performing scDNA-seq from FFPE tissues, this method was limited to profiling a very small number of cells and required lengthy experimental procedures (Martelotto et al., 2017).

1.6 Computational tools for processing single cell datasets

1.6.1 Single cell RNA methods and benchmarking

Experimental techniques for scRNA-seq have bloomed and increased over time, leading to an increasing number of computational methods being developed across different programming languages and platform. A list of popular and most used tools used for processing single cell RNA sequencing data have been summarized in the table below.
<table>
<thead>
<tr>
<th>Type of Analysis</th>
<th>Name/Method</th>
<th>Language</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall analysis</td>
<td>Seurat</td>
<td>R</td>
<td>A comprehensive toolkit with functionalities to analyze scRNA, sRNA and scATAC data.</td>
</tr>
<tr>
<td></td>
<td>Scanpy</td>
<td>Python</td>
<td>A scalable toolkit for analyzing single-cell gene expression data</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>PCA</td>
<td>R</td>
<td>Uses a linear combination of genes that best capture the variance in the data</td>
</tr>
<tr>
<td>Reduction</td>
<td>NMF</td>
<td>R</td>
<td>Non-negative matrix factorization, usually gives very granular clusters, generates more interpretable dimensions by attempting to find discrete components</td>
</tr>
<tr>
<td></td>
<td>Diffusion Maps - Destiny</td>
<td>R</td>
<td>Generates a smooth dimensional reduction under the assumption that the cells follow a continuous path</td>
</tr>
<tr>
<td></td>
<td>UMAP</td>
<td>R</td>
<td>Similar to t-SNE but also preserves the global structure; Computationally more efficient but may sometimes result in spurious signals in smaller datasets</td>
</tr>
<tr>
<td>Cell Annotation</td>
<td>SingleR</td>
<td>R</td>
<td>SingleR leverages reference transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells.</td>
</tr>
<tr>
<td></td>
<td>scMAP</td>
<td>R</td>
<td>A tool for unsupervised projection of cells from a scRNA-seq dataset on to the cell-types identified in a new dataset.</td>
</tr>
<tr>
<td></td>
<td>cellassign</td>
<td>R</td>
<td>Automated, probabilistic assignment of scRNA-seq to known types. Required a priori input of cell type markers and then assigns a probability value to each cell type.</td>
</tr>
<tr>
<td></td>
<td>Garnett</td>
<td>R</td>
<td>Supervised regression based classifier that requires cell type marker file as an input to train and predict new datasets.</td>
</tr>
<tr>
<td></td>
<td>scANVI</td>
<td>Python</td>
<td>Single-cell ANnotation using Variational Inference (scANVI) is a semi-supervised variant of scVI designed to leverage annotations from single dataset on a short, uses deep generative models</td>
</tr>
<tr>
<td></td>
<td>Distance-based - PCAKmeans, SCUBA</td>
<td>R</td>
<td>and data points are assigned to the cluster around the nearest centroids. During cluster allocation of data points, the in-cluster sum of squares is reduced, and the position of centroids is iteratively optimized. Fichable and time-efficient. Sensitive to shape and number of data points, need the number of clusters a</td>
</tr>
<tr>
<td></td>
<td>Hierarchial - cellTree</td>
<td>R</td>
<td>Generates clusters into a hierarchical structure, no need of knowing a priori the number of clusters. No assumptions are made about the distribution of data points, and each cluster links to another by branches and is nested like a hierarchical tree in the form of a dendrogram. Computationally intensive.</td>
</tr>
<tr>
<td></td>
<td>Graph based - Seurat, Scanpy, SNN, Cliq</td>
<td>R</td>
<td>Python</td>
</tr>
<tr>
<td></td>
<td>Density based - Monode2</td>
<td>R</td>
<td>Assigns clusters based on distribution of data points, highly sensitive to parameters</td>
</tr>
<tr>
<td></td>
<td>Model based - BISCUIT</td>
<td>Python</td>
<td>Directly modeling gene expression as counts and inferring cluster membership from these models.</td>
</tr>
<tr>
<td>Pseudotime/ Trajectory</td>
<td>Monocle</td>
<td>R</td>
<td>Tree based method, no prior needed, cannot create a cyclic trajectory, can be slow when too many data points</td>
</tr>
<tr>
<td></td>
<td>Slingshot</td>
<td>R</td>
<td>Tree based method, no priors needed, cannot create a cyclic trajectory</td>
</tr>
<tr>
<td></td>
<td>SCORPIUS</td>
<td>R</td>
<td>Linear trajectory method, good at ordering cells</td>
</tr>
<tr>
<td></td>
<td>PAGA</td>
<td>Python</td>
<td>Graph based method. Require prior root information, can create cyclic and linear, bifurcated trajectories</td>
</tr>
<tr>
<td>Differential Expression</td>
<td>MAST</td>
<td>R</td>
<td>Uses a linear hurdle model to account for confounders, and DE is determined using the likelihood ratio test.</td>
</tr>
<tr>
<td></td>
<td>SCDE</td>
<td>R</td>
<td>Uses a Bayesian approach that incorporates an evidence-based approach to evaluate the likelihood of the average level of gene expression for individual cells and measure the fold changes. Highly sensitive.</td>
</tr>
<tr>
<td></td>
<td>DESEQ2</td>
<td>R</td>
<td>Then, it will estimate the gene-wise dispersions and shrink these estimates to generate more accurate estimates of dispersion to model the counts. Finally, DESeq2 will fit the negative binomial model and perform hypothesis testing using the Wald test or Likelihood Ratio Test.</td>
</tr>
<tr>
<td>Doublet Detection</td>
<td>Scublet</td>
<td>R</td>
<td>Simulates doublets from the observed data and uses a k-nearest-neighbor classifier to calculate a continuous doublet score</td>
</tr>
<tr>
<td></td>
<td>doubletCells</td>
<td>R</td>
<td>It generates artificial doublets by adding two randomly selected droplets’ gene expression profiles. For each droplet, it calculates the proportion of artificial doublets, in a neighborhood in the PC space, whose number of dimensions is specified by users.</td>
</tr>
<tr>
<td></td>
<td>DoubleDetection</td>
<td>Python</td>
<td>It generates artificial doublets by adding two randomly selected droplets’ gene expression profiles and pools these artificial doublets with the original droplets. Then it conducts Louvain clustering on the pooled droplets.</td>
</tr>
<tr>
<td></td>
<td>DoubleFinder</td>
<td>R</td>
<td>It generates artificial doublets by averaging two randomly selected droplets’ gene expression profiles. The doublet score of each droplet is defined as the proportion of artificial doublets among its k-nearest neighboring droplets in the PC space, whose number of dimensions is specified by users.</td>
</tr>
</tbody>
</table>

Table 1: Summary of popular single cell RNA computational tools
1.6.2 Single cell DNA Copy Number Profiling Methods

High throughput and cost-effective methods of copy number profiling using single cell DNA sequencing (scDNA-seq) have vastly improved the data quality and cell throughput. New and emerging methods using nanowells, microfluidics, flow sorting etc have enabled sequencing thousands of cells using methods such as Direct Library Preparation (DLP), combinatorial indexing and 10X Chromium Single Cell CNV. scDNA-seq data applications range from studying tumor evolution, ITH and DNA mosaicisms. However, one of the biggest challenges faced in analyzing such datasets has been the limited number of comprehensive and user-friendly computational methods available to find meaningful biology from these sparse datasets. Our lab developed a method called CopyKit a comprehensive software suite written in R that does preprocessing of scDNA data, including binning the sequencing reads in genomic regions, copy number segmentation, filtering of low-quality and diploid cells followed by clustering of cells into subclones and visualization (Navin et al., 2022).

1.7 – Dissertation Summary

Human breast is a highly complex organ with various transcriptional machineries harbored by each cell population, however it’s still unclear on how its reprogramming and interplay leads to progression to invasive phenotypes and therapeutic resistance in Ductal Carcinoma In Situ (DCIS) and Triple-Negative Breast cancer (TNBC). To investigate this problem, I used single cell and whole-exome sequencing to study the normal breast, DCIS and TNBC to understand the role of tumor, stromal and immune cell type reprogramming that either promotes or suppresses tumor progression in DCIS and chemoresistance in TNBC. To address this problem, I used innovative single cell RNA, DNA sequencing and spatial genomic technologies in addition to whole-exome
sequencing to study DCIS and TNBC to understand the role of tumor, stromal and immune cell type reprogramming that either promotes or suppresses tumor progression in DCIS and resistance to chemotherapy in TNBC. To investigate the genome evolution in DCIS, we performed genomic analyses on the initial DCIS lesion and paired invasive recurrent tumors in 34 patients together with single cell DNA sequencing in a subset of 4 cases. Our data showed that in 80% the invasive recurrence was clonally related to the initial DCIS, suggesting that tumor cells were not eliminated during the initial treatment. Surprisingly however, 20% were clonally unrelated to the DCIS, representing new independent lineages. In addition, we established a normal breast cell atlas from tissues collected from 60 disease free women by single cell RNA sequencing of 248,687 cells and 89,301 nuclei from 20 women. Our data identified 10 major cell types and 32 cell states that reflect different biological functions that can be organized into 4 major spatial domains (adipose, connective, ducts, lobules). Unexpectedly, our data reveal an abundance of pericytes and endothelial cells in the breast, as well as a rich ecosystem of immune cells that are located in the ducts and lobules. Collectively, this data provides an unprecedented atlas of the normal cell types and cell states in the human breast, that will serve as an important reference for studying breast development and diseases such as breast cancer. Lastly, we used single cell RNA sequencing to profile 100 treatment naïve TNBC biopsies with pathological complete response (pCR) information and compared the tumor and normal breast ecosystem and found populations like CAFs, tumor endothelial cells and proliferating endothelial cells enriched in TNBC tissues suggesting their potential reprogramming and role in invasive tumor progression. Further we found certain cell states such as cytotoxic CD8 T cells, Matrix metalloproteases (MMP) high macrophages as well as proliferating macrophages associated with pCR response.
CHAPTER 2: MATERIALS AND METHODS

[Note: Part of this chapter is based on the manuscript that is published at Nature Genetics (Lips and Kumar et al. 2022, Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer - doi: https://doi.org/10.1101/2021.03.22.21253209). Nat Genetics articles are published under a CC BY license (Creative Commons Attribution 4.0 International License), which establishes that authors retain copyright in their articles.]

2.1 Materials

2.1.1 Normal Breast tissue study collection protocol

Freshly dissected human breast tissue samples were obtained from the University of Texas MD Anderson Cancer Center and Baylor St. Luke’s Medical Center in collaboration with Dr. Bora
Lim, Dr. Thompson and Dr. Winocour with informed consent and with the material transfer agreement (MTA) between external institutes approved by the Internal Review Board (IRB). Surgical tissue was transported from the pathology to Navin Lab in an ice-cold DMEM medium.

The tissue was collected from women who underwent either a mastectomy or a reduction mammoplasty at the three institutes. Tissue was further used for performing single cell and nuclei RNA sequencing (Figure 5).

2.1.2 TNBC Tissue samples

Fresh triple-negative breast cancer (TNBC) tissue samples were obtained from the ARTEMIS: A Robust TNBC Evaluation fraMework to Improve Survival (ARTEMIS: NCT02276443) trial at the University of Texas MD Anderson Cancer Center led by Dr. Stacy Moulder and Dr. Clinton Yam.

The ARTEMIS trial is a randomized Phase II clinical trial designed to select chemo-insensitive patients based on their genomic signatures and treat them with targeted therapy and evaluate if it leads to better pathologic complete response and better long-term survival compared with standard

![Figure 6](image)
chemotherapy. Patients undergo baseline molecular and IHC evaluation of their tumor biopsy and receive standard anthracycline-based chemotherapy and undergo standard ultrasound at baseline, after 2 courses, and after 4 courses of treatment. Based on the results of the molecular evaluation, they may then choose to continue with standard taxane +/- platinum-based chemotherapy or participate in an experimental clinical trial designed to match their genomic signature and triple-negative subtype. Patients insensitive to standard chemotherapy are advised to participate in a more targeted therapy based clinical trial treating their tumor subtype. We collected fresh core tissue biopsies (size of the core) at the pre-treatment timepoint and was transported from the pathology to Navin Lab in an ice-cold DMEM medium (Figure 6).

2.1.3 Pure primary DCIS and subsequent recurrence cancer samples

Cases of pure primary DCIS that, after treatment, had subsequently developed recurrent disease were identified from:

1. the Sloane project, a national audit of women with non-invasive neoplasia within the United Kingdom National Health Service Breast Screening Programme (Ethical approval REF 08/S0703/147, 19/LO/0648) median follow up 5.3 years.

2. the Dutch DCIS cohort study, a nation-wide, population-based patient cohort derived from the Netherlands Cancer Registry (NCR), in which all women diagnosed with primary DCIS between 1989 and 2004 were included, and has a median follow up time of 12 years 26. This cohort was linked to the nationwide network and registry of histology and cytopathology in the Netherlands (PALGA). The study was approved by the review boards of the NCR (ref. no. 12.281) and PALGA (ref.no. LZV990) and the institutional review board of the Netherlands Cancer Institute under number CFMPB166, CFMPB393 and CFMPB688.
3. the Duke Hospital cohort, a hospital-based study of women (age 40-75 years) diagnosed with DCIS between 1998 and 2016, with a median follow up of 7.9 years (IRB approvals: Pro00054877, Pro00068646).

**Figure 7:** Experimental workflow of longitudinally matched DCIS and recurrent DCIS/IDC samples

Formalin-fixed paraffin-embedded (FFPE) tissue specimens of patient-matched DCIS and subsequent recurrence were retrieved and reviewed by specialist breast pathologists to confirm the diagnosis and exclude confounding features (such as microinvasion) (Figure 7).
2.2 Application of existing single cell sequencing methods

2.2.1 Single cell RNA sequencing

The normal breast surgical specimens received in large samples (~200g) as well as tumor breast biopsies were placed in 150 cm sterile tissue culture dishes and are divided into separate 1-2g reactions. Excess adipose tissue was removed prior to dissociation since adipocytes and cells >40um clog 10x Genomics microfluidic chips. Any remaining adipocytes were separated from the cell suspension and removed in the washing steps. Dissociation solution consists of collagenase A (1mg/ml working solution, Sigma #11088793001) dissolved in DMEM F12/HEPES media (Gibco #113300) and BSA fraction V solutions (Gibco# 15260037) mixed at a 3:1 ratio, respectively. For each reaction, a 10cm dish with 2ml dissociation solution was used to mince tissue into a pulp with no visible large pieces. Suspension was transferred into a 50ml conical tube with 40ml of dissociation solution in a rotating hybridization oven for 2 to 6 hours at 37°C until completely digested. Cell suspension was centrifuged at 500g for 5 minutes and supernatant was removed. The pellet was resuspended in 5ml trypsin (Corning #25053CI) at room temperature and incubated in a rotating hybridization oven at 37°C for 5 minutes. Trypsin was neutralized with 10ml DMEM containing 10% fetal bovine serum (FBS) (Sigma #F0926). The solution was mixed by pipetting up and down and filtered through a 70μm strainer (Falcon # 352350). A sterile syringe plunger flange was used to sieve the leftover unfiltered tissue and DMEM was used to wash the remaining single cells off the filter. The flow-through was centrifuged at 500g for 5 minutes and the supernatant was removed. Resulting pellet was nutated at room temperature for 10min in 20ml 1x MACS RBC lysis buffer (MACS #130-094-183) to remove red blood cells (RBCs). To stop RBC lysis, 20ml DMEM was added and then centrifuged at 500g for 5min. The cell pellet was washed in 10mls of cold DMEM and centrifuged at 500g for 5 minutes. Pellet was then resuspended in cold PBS (Sigma #D8537) +0.04%BSA solution (Ambion #AM2616) and
filtered through a 40μm flowmi (Bel-Art #h13680-0040). Trypan blue stained cells were counted in the Countess II FL automated cell counter (Thermo Fisher) and their concentration was adjusted to 700-1200 cells/µl. Single cell capture, barcoding and library preparation were performed by following the 10X Genomics Single Cell Chromium 3’ protocols (V2: CG00052, V3: CG000183, V3.1: CG000204). Detailed description is also publicly available at dx.doi.org/10.17504/protocols.io.t3geqjw.

2.2.2 Single nuclei RNA sequencing

To isolate single nuclei, 0.5-1g fresh breast tissue were placed in a 10cm dish with 2ml lysis buffer. Nuclei lysis buffer consists of NST-DAPI buffer with 0.1U/µl RNase Inhibitor (NEB #M0314L). (NST-DAPI buffer (Leung et al., 2016)) Tissue was minced into a pulp with no visible large pieces. The suspension was filtered through a 40μm cell strainer (Falcon #352340). A sterile syringe plunger flange was used to gently sieve the leftover tissue on the filter and then rinsed with 3ml of lysis buffer. The flow-through was transferred into 5ml DNA LoBind tubes and incubated on ice for 10min. The tube was centrifuged at 500xg for 5min at 4°C. The supernatant was removed and nuclei were washed with 1ml cold lysis buffer and centrifuged again. The nuclei pellet was resuspended in 1% BSA in PBS (Sigma #SRE0036) supplemented with 0.2U/µl RNase Inhibitor. Nuclei were filtered through a 40μm Flowmi cell strainer, counted by hemocytometer under DAPI channel and concentration was adjusted to 700-1200 nuclei/µl. 10X Genomics RNA experiments proceeded immediately to avoid nuclei aggregation. Single cell capture, barcoding, library preparation and sequencing were the same as detailed above. Detailed protocol is also publicly available at dx.doi.org/10.17504/protocols.io.t3aeqje.
2.2.3 Single cell DNA sequencing

FFPE samples were deparaffinized using the FFPE Tissue Dissociation Kit from MACS (Cat#130-118-052). Nuclear suspensions were prepared from the recovered cell suspensions using a DAPI-NST lysis buffer (800 mL of NST (146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM CaCl2, 21 mM MgCl2, 0.05% BSA, 0.2% Nonidet P-40)), 200 mL of 106 mM MgCl2, 10 mg of DAPI). The nuclear suspensions were filtered through a 35 mm mesh and single nuclei were flow sorted (BD FACS Melody) into individual wells of 384-well plates from the aneuploid peak (Supplementary Data File 5). After sorting single nuclei, direct tagmentation chemistry was performed following the Acoustic Cell Tagmentation (ACT) Protocol9. Briefly, nuclei were lysed and tagmentation was performed using TN5 to add dual barcode adapters to the DNA, followed by 12 cycles of PCR. The resulting libraries were QCed for concentration >10ng/ul and pooled for sequencing on the HiSeq4000 (Illumina) instrument at 76 cycles.

2.3 Single cell and nuclei RNA sequencing Data Analysis

2.3.1 Data preprocessing, QC and filtering

Reads from single cells an (Schneider et al., 2017) single nuclei from the 10x chromium were demultiplexed, aligned using the GRCh38.p12 human genome reference (Lander et al., 2001; Schneider et al., 2017) using the default parameters of the Cell Ranger pipeline (version v3.1.0, 10x Genomics). Count matrices were generated for both datasets that were further analyzed using Seurat (Stuart et al., 2019)

Count matrices were processed using the R package Seurat (v. 3.2.3) (Stuart et al., 2019). Cells from each normal breast sample were further filtered for low quality by removing cells with fewer than 500 UMIs or 200 genes detected. Potential doublets and multiplets were classified as
cells that express more than 20,000 UMIS or 5000 genes and were removed. Cells with higher than 10% mitochondrial or 50% ribosomal transcripts were also filtered owing to low quality and dying cells. Similarly, for single cell nuclei, filtering metrics were the same as used for cells, except the min number of genes were 150 as nuclei data is known to express fewer genes.

For the tumor samples, different cutoffs were used to filter the cells such as cells with fewer than 300 UMIs or 150 genes detected, greater than 30% mitochondrial transcripts and 5000 genes were filtered out.

2.3.2 Cellular and Nuclei landscape construction

The major cell type and nuclei clustering for normal breast samples was done by integrating all samples together using Canonical Correlation Analysis (CCA) based integration from the Seurat package. The filtered gene matrices from each sample were normalized using NormalizeData function. To identify highly variable genes, we used the FindVariableFeatures that models the mean-variance relationship of the normalized counts of each gene across cells and identified 5000 genes per sample. We further identified anchors using FindIntegrationAnchors to integrate all patients using following parameters – dims=20, k.filter=30, anchor.features = 3000 and k.score = 30 which were fed into the IntegrateData function with dims=20. The integrated dataset was then used for downstream analysis which included scaling and centering the data using ScaleData, finding the most significant principal components (PC) using RunPCA and utilizing the ElbowPlot to determine the number of PC’s used for clustering. Different resolution parameters for unsupervised clustering were then examined to determine the optimal number of clusters. For the major cell type and nuclei clustering, the first 20 PC’s were used for unsupervised clustering with a resolution = 0.2 yielding a total of 21 cell clusters and for nuclei the resolution = 0.3 yielding 21 nuclei clusters. For visualization, the dimensionality was further reduced using either UMAP

26
methods with Seurat functions RunUMAP. The PC’s used to calculate the UMAP embedding were the same as those used for clustering. Each resulting cluster was further checked for doublets or low quality cells by a 3 step process – 1) Calculated the quality metrics such as nCount\_RNA and mitochondrial content and removed clusters with any outlier values i.e greater or less than 2 sd than the average of all clusters, 2) Checked the top 15 differentially expressed genes of each cluster and removed the clusters where genes were mostly mitochondrial, ribosomal or Hemoglobin, 3) Based on canonical markers, we checked if any cluster had a cells expressing canonical markers from different cell types. Based on the above criteria, we got 10 major and well separated cell types and 12 nuclei clusters on the UMAP space.

Similar steps were followed by breast tumor cells as well, however the parameter selection was different based on the dataset. The breast samples were computationally processed in 2 batches based on the timeline of collection. 1st 60 set of samples were in batch 1 and remaining 40 in batch 2. Both batches were independently filtered, integrated and clustered to identify major cell types using the approach described above. The filtered cells from both batches were then integrated using RunHarmony function instead of CCA based integration due to the large number of cells and computation time. There were 9 major breast cell types identified in the tumor tissues.

2.3.3 Assigning cluster annotation

To annotate the major cell type of each single cell or nuclei, FindAllMarkers was used to find differentially expressed genes in each cluster using the wilcox test statistical framework. The top 12 most significant DEGs were carefully reviewed. Further, we checked each cluster using the known canonical markers such as EPCAM for epithelial cells, PTPRC for immune cells, CD3D/E/G for T cells, CD19/MS4A1/CD79A for B cells, LUM/DCN/COL6A1 for fibroblasts, PECAM1 for endothelial cells and RGS5 for pericytes). We also applied SingleR to
annotate the clusters. The three approaches were combined to infer major cell types for each cell and nuclei cluster according to the resulting annotation given by SingleR, the enrichment of canonical marker genes and top-ranked DEGs in each cell cluster.

2.3.4 Determination of Expression states

Each cell cluster from the normal breast tissues was further extracted and underwent clustering and filtering as described above, however with different parameters. The different parameters used for clustering the expression states of major cells are as follows - B cells (dims = 12; k.param = 20, scaled by nCount_RNA, resolution = 0.3), T cells (dims = 20; k.param = 20, scaled by nCount_RNA, resolution = 0.4), Myeloid cells (dims = 30; k.param = 20, scaled by nCount_RNA, resolution = 0.4), Fibroblasts (), Luminal HR (dims = 35; k.param = 20, resolution = 0.075, scaled by nCount_RNA), Luminal Secretory (dims = 35; k.param = 20, resolution = 0.2 scaled by nCount_RNA), Pericytes (dims = 25; k.param = 20, scaled by nCount_RNA, resolution = 0.4), Lymphatic (dims = 30; resolution = 0.05) and Vascular Endothelial cells (dims = 30; resolution = 0.1). Each round of clustering was followed by filtering for low quality and doublets cells. Expression states were annotated by first investigating the top 20 genes of each cluster and manually curating literature for references followed by identifying the pathways enriched for each state as described in the Pathway Enrichment section.

2.3.5 Cellcycle analysis, Differential Expression and Pathway Enrichment

We utilized the “CellCycleScoring” function from the Seurat package that is based on the cell cycle phase genes from the paper by Tirosh et al (Tirosh et al., 2016). Each cell and nuclei were given a quantitative score for G1, G2/M and S scores. Differentially expressed genes were calculated for each cell cluster relative to other cells within its celltype compartment using the
“FindMarkers” function in Seurat with the wilcoxon test for statistical significance. For Gene set enrichment analysis, ranked genes were selected based on the above test filtered for an adjusted p-value <=0.05 and arranged by average log foldchange values between each cluster and fed into the fgsea package using 1000 permutations. Curated gene sets of KEGG, Biological Processes and Reactome were downloaded from the Molecular Signature Database (MSigDB, http://software.broadinstitute.org/gsea/msigdb/index.jsp) and were used to calculate enrichment. Significantly enriched gene sets were identified with a Benjamini-Hochberg adjusted P value <= 0.05.

2.3.6 Regulatory network Analysis

On the raw count matrix, we employed SCENIC (Aibar et al., 2017) to inference the regulatory networks following the tutorials from https://scenic.aertslab.org/. On the regulon score matrix, we performed differential expression (DE) analysis followed the similar approach in DE gene analysis and identified top regulons for each cell types.

2.3.7 Copy Number inference from single cell RNA data

For all tumor breast sample data, we identified aneuploid cells using the package “CopyKat” (Gao et al., 2021) using the raw count matrix and default parameters.

2.3.8 Integration of normal and tumor breast single cell data

Filtered cells from breast cell atlas at the cell type level were integrated with cells of the TNBC using CCA based integration method in the Seurat package using parameters, k.filter = 50, nfeatures = 2000. We only combined the patients where atleast 50 cells were present. Clustering resolution of integrated_snn_res.0.35 was used to delineate the cell states.
2.4 Single cell DNA sequencing Analysis

2.4.1 Data preprocessing, QC and filtering

To calculate single-cell copy number profiles we demultiplexed sequencing data from each cell into FASTQ files, allowing 1 mismatch of the 8 bp barcode. FASTQ files were aligned to hg19 (NCBI Build 37) using bowtie2 (2.1.0) (Langmead and Salzberg, 2012) and converted from SAM to BAM files with SAMtools (0.1.16) (Li et al., 2009). PCR duplicates were removed based on start and end positions. Copy number profiles were calculated at 220kb resolution using the variable binning method (Baslan et al., 2012). The pre-processing steps to compute DNA copy number profiles have been described in detail previously. Single cells with <10 median reads/bin were excluded for downstream copy number analysis. GC normalized read counts were binned into bins of variable size, averaging 200kb, followed by population segmentation with the multipcf (gamma = 10) method from the R Bioconductor multipcf package (Olshen et al., 2004). The log2 copy number ratio were calculated and used for subsequent analysis. We filtered out noisy single cells with mean 9-nearest neighbor correlation less than 0.85. The mean 9-nearest neighbor correlation is calculated as the average of the Pearson correlation coefficients between any single cell and its 9-nearest neighbors. This step removed single cells with poor whole-genome amplification from the subsequent data analysis.

2.4.2 Data clustering and visualization

Single-cell ratio data was embedded into two dimensions using UMAP, R package ‘uwot’ (v0.1.8, seed = 31, min dist = 0.2, n_neighbors = 30, distance = “manhattan”) (Laks et al., 2019). The resulting embedding was used to create an SNN graph with R Bioconductor package scran.
Subclones were identified with R package ‘dbscan’ (v1.1-5, k_minor = 0.02*#cells) (Traag et al., 2019).

2.4.3 Consensus heatmaps and phylogenetic reconstruction of clonal lineages

Heatmaps were plotted with R package ComplexHeatmap (v2.2.0) (Gu et al., 2016). For each sample, consensus ratio profiles were calculated by taking the median of every segment ratio of all the single cells that were assigned to the same subclone. To reconstruct the clonal lineages, we first calculated pairwise Manhattan distance matrices from consensus profiles of tumor subclones for each sample. Phylogenetic inference for consensus trees were performed using neighbor joining algorithm (Gronau and Moran, 2007) from R package ‘ape’ (v5.5) Diploid cells with segment ratio=1 were used as the root node for each tree. Trees were plotted using R package ‘ggtree’ (v3.2.1) (Xu et al., 2021). The sequence of subclones was defined based on the clonal lineages from the phylogenetic tree calculated from the consensus copy number profiles.

2.5 Application of existing Bulk Next-Generation sequencing Methods

2.5.1 Whole-Exome Sequencing

For DNA isolation, either macrodissection using a light microscope or laser microdissection (LMD) was performed. 8µm sections were stained using nuclear fast red (macrodissection) or toluidine blue (LMD) and DCIS or invasive disease were separated from the normal tissue. Tumor DNA was extracted using the AllPrep DNA/RNA FFPE Kit (Qiagen).

Whole exome sequencing (WES) of the paired DCIS with subsequent recurrence together with matched normal tissue was performed at the Department of Genomic Medicine, MD Anderson Cancer Center. Genomic DNA (18-300 ng) was used to generate sequencing
Libraries were sequenced on NovaSeq 6000 multiplexing 16 tumor samples per lane.

**2.6 Whole-Exome Sequencing Data Analysis**

2.6.1 Somatic Variant Calling and filtering

Data were converted to a FASTQ format and then aligned to the hg19 reference genome using the Burroughs-Wheeler Aligner (BWA). The aligned BAM files were subjected to mark duplication, re-alignment, and re-calibration using Picard v2.21.9 and GATK v4.1.7.0. The BAM files were then analyzed by MuTect and Pindel against the matched normal sample to detect somatic single nucleotide variants (SNV) and insertions/deletions (indels), respectively.

Individuals with normal sample median target coverage (MTC) > 40x and tumor sample MTC > 80x were included for further investigation. Variants were filtered by the following criteria: (1) Log odds score ≥ 10. (2) Exonic variants. (3) Tumor sample coverage at this site ≥ 15. (4) Normal sample coverage at this site ≥ 10. (5) Allele fraction in tumor sample ≥ 0.02. (6) Allele fraction in normal sample < 0.01. (7) Population frequency < 0.01 in ExAC, ESP6500, and 1000g database. (8) Hotspot mutations in *PIK3Ca* and *TP53* were added back to the dataset, if they did not pass these criteria. (9) For non-clonal pairs, private mutations were manually checked in IGV in both the primary and recurrence to ensure that they were indeed private and not filtered out by QC criteria.
2.6.2 Germline Variant Calling

We identified potential sample mismatches using an in house script developed by Dr. John Zhang Lab at the University of Texas at MD Anderson Cancer Center, for computing SNP matching index. Indexed BAM files from both tumor-normal pairs were used as an input to the variant caller Platypus v0.5.2 to identify germline variants. For any pair of Platypus vcf's (two samples), we removed the SNPs from random chromosomes as well as SNPs with coverage < 10, and calculated the number (nAB) of overlapping SNP's (by position), and the number (nGAB) of the same alleles within the overlapping SNPs. The score (match-index %) = nGAB*100/nAB. Using this index, we removed all mismatches with score <90%.
CHAPTER 3: A SPATIALLY RESOLVED SINGLE CELL GENOMIC ATLAS OF THE ADULT HUMAN BREAST

3.1 Introduction

The human breast is an apocrine organ that plays an important physiological role in producing milk to feed a newborn. This glandular function is mediated by an epithelial ecosystem consisting of highly branched lobular units that produce and transport milk via an intricate network of duct units towards the nipple. The mammary epithelial system is embedded into an adipose-rich tissue and surrounded by a dense web of vasculature and lymphatic vessels directly connected to the larger axillary lymph nodes (Figure 8).

![Cell Types of the Human Breast](image)

**Figure 8** – Celltypes in an adult human breast

Based on histology, the human breast can be divided into 4 major regions, 1) lobular regions consisting of densely packed units of branched epithelial lobules, 2) ductal tree like regions that contains bi-layered tubules lined with 2 layers of epithelial cells, 3) adipose regions that is the bulk of the organ, and 4) connective tissue rich in extracellular matrix. Each region is distinct in
the type of cells present and the functional role it has in maintaining the organ and other necessary functions under hormonal control.

Previous studies of human breast tissues have focused on characterizing the epithelial cell types, i.e. a layer of luminal cells that lines the lumen and an outer layer of basal/myoepithelial cells around the basement membrane of the ducts and lobules. These cells have been believed to be the precursors for initiating breast cancer (Fitzgibbons et al., 1998), undergo cyclic variations during life events such as menstruation, pregnancy, lactation (Ewald et al., 2008) and menopause. Recent studies of breast epithelial cells using single cell RNA sequencing have characterized 3 types of epithelial cells, basal/myoepithelial cells, Luminal Secretory cells and Luminal hormone response cells (Nguyen et al., 2018). Other studies have also shown association of epithelial cells states to pregnancy and menopause [gartner]. However, other cell cell types such as stromal, vascular and immune are not studied very well in the human breast tissues thus leaving a gap in knowledge. There is increasing evidence of cross-talk between the epithelial structures and the surrounding stroma to maintain normal mammary homeostasis (Pullan et al., 1996).

The goal of this chapter is to generate a large-scale comprehensive human breast cell atlas (HBCA) using unbiased single cell and nuclei RNA sequencing.

3.2 Results

3.2.1 Overview of the clinical samples profiled

We collected 100 breast tissue samples from 62 women who underwent either a reduction mammoplasty (disease-free), prophylactic mastectomy (cancer-free but may have germline BRCA mutation) and mastectomies (contralateral breast from cancer history in the other breast) (Figure 5, Methods Section 2.1.1). The samples were selected in an unbiased manner collected from women with age 20-65.
<table>
<thead>
<tr>
<th>SampleID</th>
<th>Ethnicity</th>
<th>BMI-group</th>
<th>Age</th>
<th>Surgery</th>
<th>Cancer Type</th>
<th>Pre/Post</th>
<th>Reduction</th>
<th>Tumor Size</th>
<th>Positivity</th>
<th>P53</th>
<th>ER</th>
<th>PR</th>
<th>Estrogen</th>
<th>Progesterone</th>
<th>HER2</th>
<th>N-A</th>
<th>Density</th>
<th>Reduction</th>
<th>BMI</th>
<th>Tumor Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>pt62</td>
<td>black</td>
<td>unknown</td>
<td>42</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>n-a</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt61</td>
<td>white</td>
<td>overweight</td>
<td>40</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt58</td>
<td>white</td>
<td>unknown</td>
<td>39</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt57</td>
<td>white</td>
<td>unknown</td>
<td>39</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt51</td>
<td>white</td>
<td>unknown</td>
<td>38</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt50</td>
<td>white</td>
<td>unknown</td>
<td>38</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt43</td>
<td>white</td>
<td>unknown</td>
<td>37</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt42</td>
<td>white</td>
<td>unknown</td>
<td>37</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt33</td>
<td>white</td>
<td>unknown</td>
<td>36</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt37</td>
<td>white</td>
<td>unknown</td>
<td>36</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt28</td>
<td>white</td>
<td>unknown</td>
<td>35</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt25</td>
<td>white</td>
<td>unknown</td>
<td>35</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt22</td>
<td>white</td>
<td>unknown</td>
<td>34</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt16</td>
<td>white</td>
<td>unknown</td>
<td>34</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt13</td>
<td>white</td>
<td>unknown</td>
<td>33</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt12</td>
<td>white</td>
<td>unknown</td>
<td>33</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt11</td>
<td>white</td>
<td>unknown</td>
<td>33</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt10</td>
<td>white</td>
<td>unknown</td>
<td>33</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>pt02</td>
<td>white</td>
<td>unknown</td>
<td>32</td>
<td>unknown</td>
<td>Cancer Mastectomy</td>
<td>post</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td>normal</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Clinical Meta data of all normal breast samples

We also collected associated clinical meta for all the samples (Table 2). Single cell RNA sequencing was performed on all 100 samples. In addition, we performed single nuclei RNA
sequencing on a subset of 20 samples collected from 10 women. All the samples collected from the sites at MD Anderson Cancer Center and Baylor College of Medicine were freshly obtained post-surgery using a rapid tissue dissociation protocol i.e within a 1-2-hour timeframe wherein the samples were picked up by Navin lab and immediately processed into single cell suspensions from large tissue specimens (50-100gm) for single cell sequencing.

3.2.2 Major cell types in adult human breast tissues

We performed single cell RNA sequencing on 535,941 cells from 100 tissue samples and identified 10 major cell types, including 3 epithelial cell types (basal/myoepithelial, luminal secretory and luminal hormone response), 2 cell types from the vasculature (lymphatic and vascular endothelial cells), 4 immune cell types (T, B, NK and myeloid) and 2 stromal/mesenchymal (fibroblasts and pericytes) (Figure 9).

**Figure 9:** UMAP clustering of 535,941 single cell RNA data across 100 tissues

From 10 patients, we collected both the left and right breast and compared the presence of cell type and found that all the cell types even though had proportional differences, most cell
types were present in all women irrespective of the tissue source or dissociation protocol (Figure 10).

![Figure 10](image_url)

**Figure 10**: Major cell type proportions of matched left and right breasts from 22 women and averages across all left and all right breasts

Most cell types that we found are also reported by previous studies (Ren et al., 2021), but we detected a high number of pericytes and immune cells that have not been reported before in normal breast. Additionally, an abundant population of endothelial cells such as lymphatic and vascular cells were also found suggesting that the breast is a highly vascular organ. We further identified genes expressed in each cell type cluster, some of which were already reported while others are novel cell type markers (Figure 11).
However, we did not detect adipocytes, which by far is known to be the most abundant cell type in the breast, in the scRNA-seq data. This was because the adipocytes are large cells (>100 microns) that are unable to get encapsulated in the GEM bead on the microdroplet platform. We additionally processed snRNA-seq data from 124,024 nuclei from 20 tissues and performed unbiased clustering to get the major cell types. Most cell types, except B cells, were identified from scRNA-seq and also in snRNA-seq, along with adipocytes and an abundant population of mast cells (Figure 12).

Figure 11: Consensus heatmap of the top 7 genes expressed in each cell type

Figure 12: UMAP clustering of 124,024 single cell RNA data across 20 tissues
However, when we compared the top gene markers across each cell type, there was substantial differences possibly owning to the difference in the nuclear vs cytoplasmic RNA pool (Figure 13).

![Consensus heatmap of the top 7 genes expressed in each nuclei type](image)

**Figure 13:** Consensus heatmap of the top 7 genes expressed in each nuclei type

3.2.3 Epithelial cells of the ducts and lobules

The mammary epithelium is involved in the production and secretion of milk into the lumen of the epithelial ducts, which channel towards the nipple during lactation. The breast epithelium is organized into distinct functional structures of 1) terminal ductal lobular units (TDLUs) consisting of highly branched epithelial buds; and 2) an intricate system of epithelial ducts that connect the TDLUs to the nipple of the breast (Figure 1). Histologically, two types of epithelial cells have been classified within the mostly bi-layered epithelium: an inner layer of luminal cells and an outer layer of basal cells (Figure 14), however, some recent studies have shown additional epithelial cell types, such as a basal subset of mammary stem cells and various luminal
epithelial progenitor and mature cell types (Dawson and Visvader, 2021).

**Figure 14:** H&E image showing the bilayered duct

We extracted EPCAM expressing cells from our scRNA-seq and snRNA-seq data and reclustered them into different expression cell states. We identified epithelial cells and epithelial nuclei clustered into 3 major cell types namely basal, luminal secretory and luminal hormone responsive (Figure 15).

**Figure 15:** UMAP clustering of epithelial cells and nuclei
We checked the expression of known keratin genes and saw differences based on the epithelial cell type (Figure 16).

Figure 16: Dotplot showing the top keratins expressed in epithelial cell types

Further, the snRNA-seq data also expressed some key hormone genes such as estrogen (ESR1), progesterone (PGR) and androgen receptor (AR), that were not well detected in scRNA-seq data (Figure 17).
We further reclustered each epithelial cell type to identify the level of heterogeneity and distinct expression cell states (Methods section, 2.3.4). Interestingly, this revealed that while basal cells show little to no distinct heterogeneity (Figure 18a), several distinct and previously unrealized cell states emerged with both luminal cell types with the highest level of heterogeneity detected in LumSec (Figure 18c). LumHR cells displayed three distinct cell states, one large central cluster representing one large, canonical hormone-responsive cell state (Figure 18b), and two smaller clusters marked by genes encoding secretory factors (MUC1, SCGB class genes, etc), and a third cluster marked by genes involved in active hormone signaling such as FASN (Menendez and Lupu, 2017). The lumHR-fasn cluster also showed expression of genes that may relay pro-proliferative signals upon hormone-sensing such as EREG (encoding epiregulin), a function previously associated with lumHR cells (Murrow et al., 2018).

By far the highest degree of heterogeneity among epithelial cell types was observed in LumSec, which harbored a total of five distinct cell states including a large canonical lumSec cluster composed of two cell states marked by KRT23 and HES1, respectively (Figure 18d). Additionally, lumSec cells contained three distinct smaller clusters, namely a cluster of cells expressing both MHC-I and -II genes.
Figure 18: Cell states of epithelial cells and defining marker genes. a) UMAP showing basal/myoepithelial cells, b) UMAP showing expression states of Luminal HR cells, c) UMAP showing expression states of Luminal Secretory cells, d) Heatmap showing the canonical marker genes along with top 7 genes of each state, e) Dotplots showing the expression of SCGB genes, MHC Class I, II genes across epithelial states, f) Violin plot showing expression of G2M score and j) Lactation score in epithelial cell states.

(Figure 18e), a cluster of proliferating cells showing elevated G2M scores (Figure 18f), and a cluster expressing lactation-associated genes (Figure 18j).

3.2.4 Fibroblasts & Mesenchymal Cell types

Mesenchymal cell such as fibroblasts are play a role in the production and deposition of extracellular matrix (ECM) that supports the basement membrane of the epithelial structures. While many past studies have looked into fibroblast heterogeneity in tumor tissues, a comprehensive study at a single cell resolution has not been done in normal tissues. We reclustered the single cell data and investigated the expression cell states in normal fibroblasts and found 3
states: matrix fibroblasts (fibro-matrix), prematrix fibroblasts (fibro-PCOLCE2) and immune fibroblasts (fibro-immuno) (Figure 19a).

**Figure 19**: Cellstates of fibroblasts and defining marker genes. a) UMAP showing the three states of fibroblasts and H&E image showing fibroblasts in the breast tissue, b) Heatmap showing the top 7 genes of each cellstate.

The matrix fibroblasts expressed collagen markers such as COL1A1, COL1A2, COL3A1, COL5A1 and gene signatures involved in ECM and integrin binding. The prematrix fibroblasts expressed precollagen gene PCOLCE2, genes of fatty acid metabolism (FABP4, CD36, PPARG) and also showed gene signatures related to ECM, integrin binding, similar to the matrix fibroblasts. The immune fibroblasts were high in chemokines such as CXCL1, CXCL3, IL6, CXCL8 and
involved in chemokine and cytokine functions, suggesting they are involved in immune cell signaling (Figure 19b).

3.2.5 Adipose Tissues of the breast

Adipocytes are a major component of the breast tissue and is also the most abundant cell type. Interestingly in our scRNA-seq data of over 60 women we did not identify any adipocytes, despite histopathological confirmation. This is due to the technical limitations of the droplet based platforms and large cell size of these cells, it’s very difficult to assay adipocytes. To overcome that, we performed snRNA-seq to measure the transcriptomic profiles of adipocytes. These data identified ADH1B, ADIPOQ, CD36, FHL1, LPL, PLIN1 and PLIN4 as top marker genes of breast adipocytes. (Figure 20).

Figure 20: Violin plot showing canonical genes expressed in beige and white adipocytes

The snRNA-seq data did not indicate that adipocytes had distinct cell states, however we investigated whether known markers of white and beige adipocytes showed heterogeneity, since previous reports reported beige adipocytes in mouse mammary glands. Based on the expression of the genes canonical for white adipocytes, we found that most breast adipocytes expressed mainly white markers, possibly suggesting that the beige adipocytes were not present in the human
breast tissues (Figure 20). These data provide a genomic reference of adipocytes in the breast, which are relevant to breast cancer, development and obesity.

3.2.6 Vascular & Lymphatic endothelial cells

There is a dense network of arteries and veins running through the breast tissue that are primarily involved in providing nutrition and protection (Sagsoz and Ketani, 2010). These structures can also be seen via histopathology. scRNA-seq data showed the 2 major types of endothelial cell clusters - vascular expressing the canonical markers PECAM1 and VWF and lymphatic, expressing known markers such as LYVE1, PROX1 and PDPN (Figure 11). While they are known to exist in the breast, their transcriptional programs and cellular heterogeneity are not well defined.

Reclustering of the lymphatic endothelial cells identified 3 major cell states that corresponded to the immune-lymphatic cells (Lymph-immune), the lymphatic valve cells (Lymph-valve) and a major lymphatic cell state (Lymph-major) (Figure 21). However, these states are not well defined in the literature. The major lymphatic cell state expressed genes including FABP4 and FABP5 and represented the most abundant component of the lymphatic ducts. The lymph-immuno cell state expressed ACKR4 and NTS which resembles cells on the ceiling of subcapsular sinus in human lymph nodes (Takeda et al., 2019), as well as several chemokine ligands (CXCL2, CXCL1, CXCL3, CXCL8) suggesting that they are involved in immune system interactions via chemokine receptors.
Figure 21: Cell states of Lymphatic endothelial cells, a) H&E showing lymphatic ducts in a breast tissue, b) UMAP showing the three cell states of lymphatic cells, c) Heatmap of known canonical genes and top 7 genes of each cell state.

The third cell states are the lymph-valve cells and expressed the known marker CLDN11 (Takeda et al., 2019), which play an important role in preventing lymphatic fluid backflow, since the lymph nodes are a passive system.

Figure 22: Cell states of Vascular endothelial cells, a) H&E showing arteries and veins in a breast tissue, b) UMAP showing the three cell states of vascular cells, c) Heatmap of known canonical genes and top 7 genes of each cell state.

In contrast to the lymphatic structures, the cell states of the vascular structures are histopathologically defined in the breast tissue. Reclustering of the vascular cell scRNA-seq data identified three major cell states, including the arterial endothelial cells (SOX17, GJA4), the venous endothelial cells (ACKR1, SELP) and the capillary endothelial cells (RGCC, CA4) based on known markers of these cell states (Figure 22). These data revealed three vascular cell states that corresponded well to the major vascular structures (veins, arteries, capillaries) and further provided a molecular definition of their expression programs and new top gene markers.

3.2.7 Pericytes of the breast

Pericytes have been mostly reported in the brain tissues, but we found an abundant population of breast pericytes in our scRNA-seq data. It’s a mesenchymal cell type, expressing the
canonical gene RGS5 and play a role in vaso-constriction and immune cell entry, particularly in the capillary beds (Kim, 2019). To delineate their cellular heterogeneity, we reclustered the scRNA-seq pericyte data which identified 3 cell states: myopericytes (peri-myo), immunopericytes (peri-imm) and CREM-pericytes (peri-crem) (Figure 23).

**Figure 23:** Cellstates of pericytes, a) UMAP showing the three cellstates of pericytes, c) Heatmap of top 7 genes of each cellstate
The myopericytes were consistent with their classical role in constricting blood vessels to regulate blood flow (Banerjee et al., 2006) and expressed myo markers such as MYL9, ACTA2, TAGLN, PLCG2 and gene signatures involved in ECM, actin and integrin binding (Figure 23). The immunopericytes expressed CXCL3, C5CL8, CCL20 and showed gene signatures involved in chemokine and cytokine functions, suggesting they are involved in immune cell signaling and mediating the passage of immune cells from vascular structures into the tissue regions (ref). Finally, the peri-CREM cells expressed genes such as CREM, CMSS2 and GCNA, and showed gene signatures associated with hormone and glucocorticoid receptor binding.

Figure 24: Chemokine signature across pericytes cell states and top GO gene signatures

3.2.8 Immune ecosystem of the breast

The normal human breast has a rich ecosystem of different kinds of resident immune cells involved in both innate and adaptive immunity, believed to protect against pathogens and provide
a defense mechanism for babies during lactation and involution. We identified T/NK expressing CD3D, CD3G, FCER1G and KLRF1, B expressing CD79A, MS4A1, MZB1 and myeloid cells expressing CD68, CD14, PTPRC (Figure 11).

Further clustering of T/NK cells revealed 24 states: CD-Th1 (CCL20, CCR6), CD4-Th1like (PGAP1), CD4-naïve (CCR7, SELL), CD4-Tem, CD4-Treg (FOXP3, CTLA4), CD4-PLCG2 (PLCG2), CD8-ZNF683 (ZNF683), CD8-Trm (ITGA1, KLRC1), CD8-Tem (GZMK, CRTAM), NKT (TRGC2, KLRG1), NK (FCERG1, KLRF1), NK-ILCs (TRDC, TNFRSF18), GD T (GPCPD1, CCDC57), proliferating clusters (ZWINT, STMN1) (Figure 25).
Another immune cell found in the breast was B cell. We reclustered B cells and identified 5 states: B-naïve (CD52, LTB), B-memory-unswitched (IGHD, MYC, EGR3), B-memory-switched (CD27), plasma-b-IgG (IGHG1, IGHG2) and plasma-b-IgA (IGHA1, IGHA2) (Figure 26). Plasma-b-IgA are believed to protect the newborn from pathogens by transferring immunoglobulin IgA either via bloodstream or during lactation (Macpherson et al., 2008).

Figure 26: Cellstates of B cells and defining marker genes, a) UMAP showing states of B cells, b) heatmap showing top 7 genes of each cellstate.

Lastly, we found different subsets of cells from the myeloid lineage: Dendritic cells such as conventional dendritic cells: cDC1 (CLECL9A) and cDC2 (CLECL10A), plasmocytic dendritic cells: pDC (TPM2, TSPAN13), monocyitic dendritic cells: mDC (LAMP3) and Mast cells (TPSAB1). Further there was a huge heterogeneity in monocytes and macrophages. We found 2 subsets of monocytes: Naïve monocytes (mono_naive) expressing the gene CD52 and Activated
monocytes (mono_active) expressing the gene VCAN, THBS1 and S100A8. There were 4 states of macrophages in our data namely, Lipid-associated macrophages (macro_lipo) expressing TREM2, LPL and APOC1, macro_m1 (C3, APOE), macro_m2 (MRC2, LYVE1) and macro_ifN (IFIT1, IFIT2, STAT1). Additionally, a cluster named, mye_prol, expressing all the proliferation

![Cellstates of myeloid cells and defining marker genes](image)

**Figure 27:** Cellstates of myeloid cells and defining marker genes, a) UMAP showing states of myeloid cells, b) heatmap showing top 7 genes of each cellstate.

genes such as MKI67, TOP2A and STMN1 along with canonical myeloid genes like CD68 was also found. M1 and M2 macrophages, characterized by their potential role in wound healing, where M1 are the antigen presenting cells that provide the first layer of defense against pathogens by
producing cytokines further leading to activation of T cells, and M2 facilitating in tissue repair by angiogenesis (Serhan and Savill, 2005) (Figure 27).

3.3 Discussion
Here, we report an unbiased genomic atlas of the adult human breast tissues that comprises 11 major cell types and 52 unique cell states that are organized into 4 major spatial domains in the tissue. While most previous studies (Bhat-Nakshatri et al., 2021) have focused largely on studying the epithelial cells, here we report on many understudied cell types in the adult human breast, including fibroblasts, adipocytes, pericytes, vascular endothelial cells, lymphatic endothelial cells and three major classes of immune cells (T-cells, B-Cells, myeloid cells).

Within the epithelial cells, we define three major cell types, including two luminal cell types (secretory, hormone-responsive) and one basal cell type (myoepithelial). These cells form the bi-epithelial layers of the ducts and lobules. Previously data was unclear about whether the two luminal epithelial cell types were intermixed in the luminal layer of the ducts and lobules, however our spatial data indicates that secretory luminal cells are mainly located in the ducts, while the hormone-responsive epithelial cells are mainly located in the lobules. Interestingly, our data shows that the basal epithelial cell types (which are located in both the ducts and lobules) comprise only a single cell state, acting as both a cell type that secretes the basement membrane of the ducts and lobules and functioning in myo-contractile functions during lactation. In contrast the two luminal epithelial cell types comprised 7 distinct cell states, with diverse biological functions including lactation, HLA expression, proliferation, hormone signaling, secretoglobins and others. Our study provides a comprehensive map of all the cell states present in a normal human breast and serves as a good reference to study diseases related to the breast.
4.1 Introduction

Triple negative breast cancer is an aggressive subtype, with an occurrence of ~20% of all breast cancers and is characterized by lack of estrogen (ER), progesterone (PR) and HER2 receptor (Hudis and Gianni, 2011). The majority of TNBC patients present with high tumor mutation burden with frequent p53 mutations and a high proliferative index (Shah et al., 2012). Past studies using deep sequencing multi-region profiling and single cell sequencing (Kim et al., 2018) have shown that TNBC is a highly heterogeneous disease (Yates et al., 2015) and likely represents multiple clinically and biologically diverse molecular subtypes that have not been defined well at the single cell resolution.

Due to the absence of ER, PR and HER2 receptors, TNBC is not usually treated using targeted therapies and therefore follows standard line of neoadjuvant chemotherapy (NAC), which includes a combination of taxanes (mitotic inhibitor) and anthracyclines (DNA/RNA intercalators). The heterogeneous nature of TNBC is likely the reason for chemoresistance, high metastatic potential and poor survival outcomes (Koren and Bentires-Alj, 2015). Pathological complete response following NAC is achieved in ~50% of early stage TNBC patients (Schmid et al., 2020), and those who do not respond to NAC die within 3 years (Liedtke et al., 2008). However, recent studies have led to the development of new therapeutic agents to treat TNBC such as checkpoint-mediated immune modulatory drugs, PARP inhibitors, antibody-drug conjugates (ADCs), and combinations of chemotherapy and targeted agents.

Previous studies using bulk transcriptomic profiling (Lehmann et al., 2011) have characterized TNBC into different molecular subtypes, namely two basal-like-related subgroups
(basal-like 1 (BL1) and 2 (BL2)), two mesenchymal-related subgroups (mesenchymal (M) and mesenchymal stem-like (MSL)), one immunomodulatory subgroup (IM) and one luminal androgen receptor group (LAR), each with varying degree of survival rates (Masuda et al., 2013). However, based on PAM50 intrinsic subtype classification (Basal-like, HER2-enriched, Luminal A, Luminal B, Normal-like, Claudin-low), majority of TNBC present a basal like signature (Bertucci et al., 2008). Since most of these studies used bulk sequencing modalities, the comprehensive understanding of tumor microenvironment (TME) is still lacking.

Besides tumor cells, TNBC microenvironment also consists of the stromal and immune components. The stromal TME consists of the extra cellular matrix with cells such as fibroblasts, vasculature such as endothelial cells and pericytes. The immune ecosystem in TNBC is composed of cells such as T and B lymphocytes and tissue resident myeloid cells. Previous work of studying TME has shown considerable contribution to tumor growth, chemoresistance, immune evasion and metastatic potential of TNBC (Quail and Joyce, 2013). The phenotypic reprogramming of the TME compared to that in pathologically normal breast tissues have been studied with very limited sample sizes. Previous studies investigating the role of TME in 10 TNBC samples have shown different cellular ecotypes based on the proportion of immune and stromal cells, whereas others have reported extensive heterogeneity in immune cells in 6 TNBC samples.

In this study, we hypothesize that the stromal and immune cells undergo reprogramming in TNBC breast as compared to the normal breast leading to highly invasive potential and resistance to therapies. We performed single cell RNA sequencing on both normal breast tissues and pretreatment TNBC tissue biopsies and compared its cellular ecosystem. We further characterized the transcriptomic features predictive of pathological complete response in TNBC samples. This is the first study to my knowledge to create a TNBC cell atlas with 100 patients.
4.2 Results

4.2.1 Overview of the clinical trial study cohort

We collected 100 treatment naïve core TNBC biopsies from patients enrolled in the clinical trial ARTEMIS. All patients were classified as TNBC based on the absence of estrogen receptor staining, progesterone receptor staining and HER2 copy number by FISH. These samples were biased towards achieving a pathological complete response owning to the design of the clinical trial. The PCR status was made available after 12 weeks of neoadjuvant chemotherapy treatment followed by surgery. We also collected associated clinical meta for all the samples (Table 2). Single cell RNA sequencing was performed on all 100 samples. All the samples collected from MD Anderson Cancer Center were freshly obtained post-surgery using a rapid tissue dissociation protocol i.e within a 1-2 hour timeframe wherein the samples were picked up by Navin lab and immediately processed into single cell suspensions for single cell sequencing. Single cell RNA

**Figure 28:** Clinical endpoint of TNBC tissues obtained from the pre-treatment biopsies.
sequencing data from unmatched pathologically normal breast tissues were taken from the human breast cell atlas project (Method 2.1.1).
<table>
<thead>
<tr>
<th>Sample Id</th>
<th>Patient ID</th>
<th>RCB Status</th>
<th>PCR status</th>
<th>Sample Id</th>
<th>Patient ID</th>
<th>RCB Status</th>
<th>PCR status</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTC01</td>
<td>P1</td>
<td>2</td>
<td>no</td>
<td>ARTC51</td>
<td>P51</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC02</td>
<td>P2</td>
<td>0</td>
<td>yes</td>
<td>ARTC52</td>
<td>P52</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC03</td>
<td>P3</td>
<td>0</td>
<td>yes</td>
<td>ARTC53</td>
<td>P53</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC04</td>
<td>P4</td>
<td>1</td>
<td>no</td>
<td>ARTC54</td>
<td>P54</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC05</td>
<td>P5</td>
<td>0</td>
<td>yes</td>
<td>ARTC55</td>
<td>P55</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC06</td>
<td>P6</td>
<td>2</td>
<td>no</td>
<td>ARTC56</td>
<td>P56</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC07</td>
<td>P7</td>
<td>1</td>
<td>no</td>
<td>ARTC57</td>
<td>P57</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC08</td>
<td>P8</td>
<td>0</td>
<td>yes</td>
<td>ARTC58</td>
<td>P58</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC09</td>
<td>P9</td>
<td>0</td>
<td>yes</td>
<td>ARTC59</td>
<td>P59</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC10</td>
<td>P10</td>
<td>0</td>
<td>yes</td>
<td>ARTC60</td>
<td>P60</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC11</td>
<td>P11</td>
<td>0</td>
<td>yes</td>
<td>ARTC61</td>
<td>P61</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC12</td>
<td>P12</td>
<td>2</td>
<td>no</td>
<td>ARTC62</td>
<td>P62</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ARTC13</td>
<td>P13</td>
<td>3</td>
<td>no</td>
<td>ARTC63</td>
<td>P63</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC14</td>
<td>P14</td>
<td>3</td>
<td>no</td>
<td>ARTC64</td>
<td>P64</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC15</td>
<td>P15</td>
<td>3</td>
<td>no</td>
<td>ARTC65</td>
<td>P65</td>
<td>0</td>
<td>yes+E68:E75</td>
</tr>
<tr>
<td>ARTC16</td>
<td>P16</td>
<td>0</td>
<td>yes</td>
<td>ARTC66</td>
<td>P66</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC17</td>
<td>P17</td>
<td>1</td>
<td>no</td>
<td>ARTC67</td>
<td>P67</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC18</td>
<td>P18</td>
<td>0</td>
<td>yes</td>
<td>ARTC68</td>
<td>P68</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC19</td>
<td>P19</td>
<td>2</td>
<td>no</td>
<td>ARTC69</td>
<td>P69</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC20</td>
<td>P20</td>
<td>0</td>
<td>yes</td>
<td>ARTC70</td>
<td>P70</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ARTC21</td>
<td>P21</td>
<td>0</td>
<td>yes</td>
<td>ARTC71</td>
<td>P71</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC22</td>
<td>P22</td>
<td>2</td>
<td>no</td>
<td>ARTC72</td>
<td>P72</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC23</td>
<td>P23</td>
<td>1</td>
<td>no</td>
<td>ARTC73</td>
<td>P73</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC24</td>
<td>P24</td>
<td>0</td>
<td>yes</td>
<td>ARTC74</td>
<td>P74</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC25</td>
<td>P25</td>
<td>2</td>
<td>no</td>
<td>ARTC75</td>
<td>P75</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC26</td>
<td>P26</td>
<td>2</td>
<td>no</td>
<td>ARTC76</td>
<td>P76</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC27</td>
<td>P27</td>
<td>0</td>
<td>yes</td>
<td>ARTC77</td>
<td>P77</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC28</td>
<td>P28</td>
<td>1</td>
<td>no</td>
<td>ARTC78</td>
<td>P78</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC29</td>
<td>P29</td>
<td>1</td>
<td>no</td>
<td>ARTC79</td>
<td>P79</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ARTC30</td>
<td>P30</td>
<td>0</td>
<td>yes</td>
<td>ARTC80</td>
<td>P80</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC31</td>
<td>P31</td>
<td>3</td>
<td>no</td>
<td>ARTC81</td>
<td>P81</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC32</td>
<td>P32</td>
<td>0</td>
<td>yes</td>
<td>ARTC82</td>
<td>P82</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC33</td>
<td>P33</td>
<td>0</td>
<td>yes</td>
<td>ARTC83</td>
<td>P83</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC34</td>
<td>P34</td>
<td>1</td>
<td>no</td>
<td>ARTC84</td>
<td>P84</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC35</td>
<td>P35</td>
<td>0</td>
<td>yes</td>
<td>ARTC85</td>
<td>P85</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC36</td>
<td>P36</td>
<td>2</td>
<td>no</td>
<td>ARTC86</td>
<td>P86</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>ARTC37</td>
<td>P37</td>
<td>0</td>
<td>yes</td>
<td>ARTC87</td>
<td>P87</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC38</td>
<td>P38</td>
<td>0</td>
<td>yes</td>
<td>ARTC88</td>
<td>P88</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC39</td>
<td>P39</td>
<td>1</td>
<td>no</td>
<td>ARTC89</td>
<td>P89</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC40</td>
<td>P40</td>
<td>0</td>
<td>yes</td>
<td>ARTC90</td>
<td>P90</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC41</td>
<td>P41</td>
<td>unknown</td>
<td>unknown</td>
<td>ARTC91</td>
<td>P91</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ARTC42</td>
<td>P42</td>
<td>unknown</td>
<td>unknown</td>
<td>ARTC92</td>
<td>P92</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC43</td>
<td>P43</td>
<td>0</td>
<td>yes</td>
<td>ARTC93</td>
<td>P93</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC44</td>
<td>P44</td>
<td>0</td>
<td>yes</td>
<td>ARTC94</td>
<td>P94</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC45</td>
<td>P45</td>
<td>2</td>
<td>no</td>
<td>ARTC95</td>
<td>P95</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC46</td>
<td>P46</td>
<td>unknown</td>
<td>unknown</td>
<td>ARTC96</td>
<td>P96</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>ARTC47</td>
<td>P47</td>
<td>2</td>
<td>no</td>
<td>ARTC97</td>
<td>P97</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC48</td>
<td>P48</td>
<td>1</td>
<td>no</td>
<td>ARTC98</td>
<td>P98</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC49</td>
<td>P49</td>
<td>1</td>
<td>no</td>
<td>ARTC99</td>
<td>P99</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC50</td>
<td>P50</td>
<td>1</td>
<td>no</td>
<td>ARTC100</td>
<td>P100</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>ARTC51</td>
<td>P51</td>
<td>0</td>
<td>yes</td>
<td>ARTC101</td>
<td>P101</td>
<td>2</td>
<td>no</td>
</tr>
</tbody>
</table>

**Table 3:** Clinical Meta data of all TNBC breast samples
4.2.2 TNBC single cell atlas

Based on the data from single cell RNA sequencing of 100 TNBC patient samples, we obtained ~500,000 cells from 10 major cell types, including epithelial tumor/anuploid cells, 2 cell types from the vasculature (lymphatic and vascular endothelial cells), 4 immune cell types (T, B, NK and myeloid) and 2 stromal/mesenchymal (fibroblasts and pericytes) (Figure 29).

**Figure 29:** UMAP and top gene heatmap of ~500,000 cells from 101 TNBC tissues
All the cells were present in varying cellular proportions across all 101 samples. We further identified genes and transcription factors (TF’s) expressed in each cell type cluster, some of which were already reported while others are novel cell type markers. The top 7 genes and TF’s of each cell type is shown in (Figure 30). As expected, we did not detect any normal epithelial cells as all epithelial cells already have undergone reprogramming to an invasive aneuploid phenotype in TNBC.

Figure 30: Heatmap showing top transcription factors in each cell type

4.2.3 Comparison of stromal microenvironment in Normal and TNBC breast tissues

We combined stromal cells such as fibroblasts, pericytes and endothelial cells from normal breast cell atlas samples and TNBC tissues and investigated the differences in cellular proportions in the expression cell states. 42,116 fibroblasts were extracted from TNBC (n=29) and HBCA (n=20) and revealed 4 major states namely (Figure 31a, b), Cancer associated fibroblasts (high in FAP, fibroblast activation protein) (Figure 31d), Immuno-fibroblasts (expressed chemokines), Fibro-major (expressing APOE, IGF1) and pre-matrix fibroblasts (high in PCOLCE2, WISP2). The top 5 genes for each cell state in shown in the dotplot (Figure 31c). Interestingly, CAF cells were only found in tumor samples where they play a role in matrix deposition and remodeling, signaling interactions with cancer cells and crosstalk with infiltrating lymphocytes (Sahai et al., 2020). Other
cell states such as immune-fibroblasts (p value = 0) and pre-matrix fibroblasts (p value = 0.015) were significantly higher in normal breast tissues (Figure 31e, f).

![Figure 31](image)

**Figure 31** – Comparison of fibroblasts in normal and TNBC tissues. a), b) UMAP showing fibroblasts in HBCA and TNBC, colored by cellstate and tissue type, c) Dotplot showing top 5 genes of each cellstate, d) Barplots showing frequency of cell states compared between normal and tumor, e) Boxplots comparing the frequencies, p values shown for Wilcoxon 2 sided test, adjusted by FDR.

We found 4 expression states of pericytes (n = 11,290) present in both TNBC (n=30) and normal breast (n = 22) (Figure 32a, b), namely peri-crem (expressing CREM gene), immune-pericytes (high in chemokines), matrix-pericytes (high in POSTN, CD36) and myo-pericytes (expressing PLN, ACTA2, MYL9). The top 5 genes of each state are shown in the dotplot (Figure 32c). Peri-crem state (p value = 0) was mostly found in normal breast whereas myo-pericytes was high in TNBC tissues (p value = 0.01) (Figure 32e, f).
Breast vasculature contains both lymphatic and vascular endothelial cells. From the 26,739 endothelial cells from both HBCA (n = 23) and TNBC (n = 35) samples, we got 7 expression states (Figure 33a, b, c). Vascular cells comprised of arterial (GJA4), veinous (ACKR1) and capillary (CA4) representing the veins and arteries. Both capillaries (p value = 0.013) and venous cells (p value = 0) were found in higher proportion in normal breast than TNBC (Figure 33d, e). We also found lymphatic cells (expressing LYVE1), valve cells (PROX1), tip cells (INSR) and proliferating endothelial cells (MKI67). Interestingly both tip cells (p value = 0) and proliferating cells (p value = 0) were higher in TNBC samples suggesting their potential role in forming new vessels and transfer of nutrients into the growing tumor tissue (Dudley, 2012) (Figure 33d, e).
Figure 33 – Comparison of endothelial cells in normal and TNBC tissues. a), b) UMAP showing pericytes in HBCA and TNBC, colored by cellstate and tissue type, c) Dotplot showing top 5 genes of each cellstate, d) Barplots showing frequency of cell states compared between normal and tumor, e) Boxplots comparing the frequencies, p values shown for wilcox 2 sided test, adjusted by FDR.

4.2.4 Cell of origin of TNBC tumor cells

We overlayed the top gene markers of the three epithelial cell types onto the tumor cells extracted from TNBC tissues (Figure 34), heatmap shows the patients with high expression of unique Basal, Luminal HR and Luminal Secretory cell lineages. We observed that most patients have a luminal epithelial
origin, with a mixed phenotype coming from luminal Secretory and Luminal HR populations, and a small set of patients with basal cell of origin.

4.3 Discussion

We performed single cell RNA sequencing on 100 treatment naïve TNBC biopsies and compared its ecosystem to the normal breast ecosystem captured earlier. We generated an atlas of all the different cell types present in a TNBC tumor microenvironment including tumor, stromal and immune cells. For each cell state, we found novel gene markers specific to each state that provides a useful reference to be used for deconvolution of bulk datasets. Our data further suggests reprogramming and increased proportions of certain cell states such as TECs, CAFs in the TNBC tissues. While inferring the possible cell of origin of tumor cells in TNBC, we found that most tumor cells are transcriptionally similar to luminal epithelial cells. Further subtyping of TNBC tumor cells is required to delineate the association of per response rate with molecular subtypes.
CHAPTER 5: GENOMIC ANALYSIS DEFINES CLONAL RELATIONSHIP OF DUCTAL CARCINOMA IN SITU AND RECURRENT INVASIVE BREAST CANCER

[Note: This chapter is based on the manuscript that is published at Nature Genetics (Lips, E and Kumar, T. et al., Genomic profiling defines variable clonal relatedness between invasive breast cancer and primary ductal carcinoma in situ - doi: https://doi.org/10.1101/2021.03.22.21253209). Nature Portfolio Journals are published under a CC BY license (Creative Commons Attribution 4.0 International License) which gives authors the copyright in their articles.]

5.1 Introduction and Rationale

Ductal Carcinoma In Situ (DCIS) is the most common form of pre-invasive breast cancer and is often detected during routine mammography, representing 20% of all screened breast cancers. There is evidence that many DCIS lesions, particularly low-grade lesions, will never progress to invasive disease (Maxwell et al., 2018; Ryser et al., 2019). However almost all patients with DCIS are still treated with surgery with or without radiotherapy to prevent progression to invasive disease that can occur years or even decades after the initial DCIS. Despite treatment, there remains a subgroup of patients who still develop recurrent disease as demonstrated in a study of 7934 patients treated by lumpectomy with or without radiotherapy, where 5.3% developed an invasive recurrence and 3.8% a DCIS recurrence with a median follow-up of 9.4 years (Shaaban et al., 2021).
As many DCIS lesions will never progress to invasive disease, some women may receive intensive treatment without any clinical benefit (Maxwell et al., 2018; Ryser et al., 2019). Therefore, there is a great unmet clinical need to develop treatment strategies that avoid overtreatment. Clinical trials of treatment de-escalation are currently underway, for example by leaving out radiotherapy or even by refraining from surgery in the absence of invasion (Elshof et al., 2015; Hwang et al., 2019). However, there is limited knowledge of whether invasive recurrences are clonally related to the initial DCIS disease making accurate evaluation of the risk of progression and the assessment of the prognostic value challenging. Detailed analysis of this question in a large cohort of patients with matched recurrent tissues samples and long-term clinical outcome data are currently lacking.

Thus, the central question of this study is whether the initial DCIS and subsequent ‘recurrence’ share a common genetic lineage, or alternatively represent independent diseases that emerge from different initiating cells in the same breast (ipsilateral). This question has been difficult to address, in part due to the logistical challenges in collecting matched longitudinal samples that are years to decades apart, and the technical challenges in performing genomic assays on archival formalin fixed paraffin embedded (FFPE) materials of this age. Consequently, most studies have focused on studying synchronous DCIS-invasive ductal carcinoma, which are single time point samples that have areas of DCIS and regions of invasive cancer cells co-occurring in the same tissue section. These studies represent more advanced cases, in which invasion has already occurred, and as expected most data have shown the cancer cells from the in situ and invasive areas are clonally related in their genetic profiles (Casasent et al., 2018; Hernandez et al., 2012; Kim et al., 2015; Minussi et al., 2021).
In contrast the genomic data on ‘pure DCIS’ with matched recurrent DCIS or invasive tumors from many years later is far more limited. One small study that used genomic data to assess clonal relatedness showed that in two of eight pairs of primary DCIS and subsequent invasive disease copy number aberrations were not concordant (Ostrovna et al., 2010), suggesting that perhaps not all cases are clonally related to the initial DCIS. Thus, while the subsequent cancers following DCIS are often termed ‘recurrences’ as a clinical definition, the genetic studies supporting their clonal relationship is still lacking.

Here, to investigate the genomic concordance and clonal relationship of pure DCIS and subsequent ipsilateral recurrent DCIS or invasive cancers, we pooled samples from three countries resulting in the largest cohort to date of DCIS cases with 5-17 years of clinical data follow up. We used whole exome sequencing to investigate genomic concordance of pure DCIS and recurrent invasive tumors. We further applied single cell DNA-seq methods to validate these results in a subset of cases (N=4).

5.2 Results

5.2.1 Overview of clinical samples profiled

In total, 24 primary DCIS and their matched recurrences were analyzed (Figure 7). All recurrences were ipsilateral, meaning that they occurred in the same breast as the initial primary DCIS lesion. The median age at diagnosis of the primary DCIS was 57 years (range 34-87 years) and median time to the recurrence was 4 years (0.4-17.5 years). Of the primary DCIS samples, 52% were high grade, 67% were ER positive (ER+), 29% were Her2 positive (Table 4). Only 13% (12/95) of the primary DCIS that developed an invasive recurrence received radiotherapy as part
of their primary treatment, in contrast to 53% (18/34) of the primary DCIS that recurred as pure DCIS.

**Table 4:** Clinical characteristics of the DCIS samples

5.2.2 Whole exome sequencing to assess clonal relatedness

We employed whole exome sequencing (WES) to survey somatic mutations in 24 DCIS-invasive (INV) recurrence pairs. The numbers of shared and private mutations were highly variable for the different tumor pairs, ranging from 0 to 112 shared mutations and 4-646 private mutations (Figure 35). Shared mutations had significantly higher (P<0.001, Wilcoxon Rank Test) allele frequencies compared to private mutations (Figure 36), consistent with early clonal selection, with the most common shared mutations occurring in *TP53* and *PIK3CA* (Figure 38).

**Figure 35:** Clonality assessment using whole exome sequencing. The total number of mutations, followed by the mutations type (primary private, recurrence private and shared) and the final clonality conclusion plotted for each DCIS-invasive recurrence pair.
Invasive recurrences had a higher number of private mutations than their matched primary DCIS, \( P=0.039 \), Wilcoxon Rank Test, two-sided, (Figure 37c).

**Figure 36:** Distribution of variant allele frequency of mutations. a) Scatter plots showing the variant allele frequency of mutations in 3 clonal related pairs. b) Similar as a, for 3 clonal unrelated pairs, c) Boxplots comparing variant allele frequency primary private \((n=702)\), recurrence private \((n=1257)\) and shared mutations \((n=433)\). Minima and maxima are present in the lower and upper bounds of the boxplot, respectively. Adjusted \( p \)-values for Holm–Bonferroni method \( p = 5.96 \times 10^{-84} \) and \( p = 3.84 \times 10^{-39} \), were calculated with two-sided Wilcoxon test. For the shared mutations, both the variant allele frequency in the DCIS tissue (Primary shared) and the invasive recurrence (Recurrence shared) are shown.

**Figure 37:** Mutational landscape of primary DCIS samples. Distribution of mutations in subsequent samples with DCIS and invasive recurrences based on WES data \((n=24)\).

Of the twenty-four DCIS-invasive cancer pairs, 83\% \((20/24)\) showed clear evidence of clonal relatedness, including three cases of primary DCIS that developed an invasive recurrence.
despite having undergone a mastectomy. The remaining four pairs (17%) did not harbor any shared mutations that could be detected in our analyses, with 11-70 mutations being detected in the primary DCIS and 36-329 in the invasive recurrence (Figure 35). Lineage inference of clonally unrelated pairs showed clearly different subclones and drivers in the primary and recurrence lesions (Figure 38). Analysis of the clonally related pairs revealed that the primary DCIS consisted of multiple subclones some of which had expanded in the invasive recurrence, dominant subclones remained at high frequencies in both the primary and recurrent samples (Figure 38).

**Figure 38:** Lineage tracing for 4 patients, a) 2 patients with clonal unrelated tumors, b) Similar as a, for 2 clonal related tumor pairs.

Our results based on WES show that most DCIS-invasive recurrence pairs are clonally related, and further show that in some cases the genomic profiles are highly similar. Notably however, others had diverged genomically, acquiring many additional events but are still related to the DCIS through a common ancestor. These data also revealed a small number of cases (N=4) that did not share any genomic alterations between primary DCIS and invasive
recurrences, suggesting that an independent tumor representing a second cancer, emerged in the same breast.

5.2.3 Single cell genome sequencing to reconstruct clonal lineages

To validate the bulk genomic profiling classifications, we applied single cell DNA sequencing (scDNA-seq) to profile genomic copy number in 2294 cells from primary and recurrent disease from FFPE tissue collected from 4 DCIS patients (Methods 2.2.3). In the two clonally unrelated sample pairs (P122 recurred as invasive disease, P110 as pure DCIS), unbiased clustering identified three major subclones in P122 and seven subclones in P110, in which each of the individual clones was specific to either the primary DCIS or recurrent tumors (Figure 39).
**Figure 39:** Clonal lineage reconstruction by single cell genome sequencing. a, UMAP plots of single cell copy number profiles from FFPE tissue showing clusters of subclones at primary time point or recurrence for two DCIS patients with independent lineages. b, Clustered heatmaps of single cell copy number profiles for two DCIS cases where the recurrence event represents an independent lineage, with selected breast cancer genes annotated below the heatmap.

**Figure 40:** Mueller plots showing clonal frequencies. Lineages are reconstructed from neighbour-joining trees using timescape, with selected breast cancer genes annotated, and chromosomal gains and losses indicated by plus and minus signs.
In P122 the clustered heatmaps showed that subclone 1 (c1) was specific to the primary DCIS and had a number of copy number aberration (CNA) events but showed no common CNA events or Clonal

**Figure 41:** Phylogeny reconstruction by single cell genome sequencing. a), Consensus copy number heatmaps of subclones calculated from clusters of single cell copy number profiles from clonally-related (b,) and clonally unrelated (c,) patients. d, Neighbor-joining trees of clonal lineages constructed from consensus subclones from clonally related and unrelated patients rooted by a diploid node, where the recurrence event represents an independent lineage, with selected breast cancer genes annotated below the heatmap.

breakpoints with the recurrent subclones (c2-3) (Figure 39b). Similarly, in the non-clonal P110 patient, common chromosomal losses on 16q and 17p (TP53) were detected in all of the subclones (c1-5) from the primary sample but did not share any CNA events with the subclones (c6-c7) in
the invasive disease. We computed consensus subclone CNA profiles from the single cells (Figure 41c) and reconstructed clonal lineages (Figure 41c) and Muller plots of subclonal frequencies, which confirmed the independent lineages in both DCIS tumors (Figure 40).

We further investigated the clonal substructure of two clonally related patients (P082, P042) classified by bulk DNA-seq. Unbiased clustering identified eight subclones in P082 and six subclones in P042 (Figure 42a).

**Figure 42:** Clonal lineage reconstruction by single cell genome sequencing of related pairs a) UMAP plots of single cell copy number profiles from FFPE tissue for two clonally related pairs showing subclones at the primary DCIS and at the recurrence time points. b) Mueller plots of the same two clonally related pairs showing clonal frequencies and lineages reconstructed from neighbour-joining trees using timescape, with selected breast cancer genes annotated, again with gains and losses annotated with plus and minus signs, respectively.

In contrast to the two clonally unrelated pairs, these tumors shared a large number of CNA events between the primary and invasive tumors. In P082 chromosomal gains in 8q (MYC), 17q (ERBB2) and 20 (AURK4) and losses in 11q (PGR), 16q, 17p (TP53) were shared among all of the eight subclones, while in P042 chromosomal gains in 1q, 8q (MYC) and 17q (ERBB2) and losses in 8p, 11q (PGR) 16q, 17p (TP53) were present in all six subclones. Furthermore, in P082 multiple subclones (c1, c2, c3, c7, c8) with the same genotypes were detected in both the primary and invasive disease. Consensus subclone CNA profiles were computed from the single cells (Figure 41a) and used to reconstruct clonal lineages (Figure 41c), which identified subclones that expanded
in the invasive disease and harbored CNA events associated with recurrence, including subelones c4-c7 in P082 and c4-c6 in P042 (Figure 42b). Collectively, the single cell data validated the clonal classifications estimated by bulk DNA-seq and further resolved direct and independent clonal lineages, revealing chromosomal events and genes associated with recurrence.

5.2.4 Genomic aberrations in recurrent invasive disease

In the patients classified as clonally related we compared mutations between the primary DCIS and invasive recurrences to identify genomic events that occurred at a higher frequency in the invasive disease and were thus associated with invasive recurrence. Strikingly, our data show that most mutations detected in the matched invasive breast cancer were already established in the primary DCIS and there were no clear genomic markers of invasive progression. The most frequently mutated genes were PIK3CA (24% DCIS, 27% INV) and TP53 (24% DCIS, 27% INV).

5.3 Discussion

In this large dataset of matched DCIS-INV recurrence pairs, we have confirmed that primary DCIS can be a precursor to subsequent invasive cancer. However notably, our data show that not all ipsilateral invasive breast cancers are clonally related to the prior DCIS, but rather 18% represent de novo primary cancers in the same breast. As such our results confirm on a much larger and detailed scale, the small study by Gorringe et al showing unrelated recurrences (Ostrovnaya et al., 2010). In patients with non-clonally related recurrences, the major question is what factors constituted the basis of risk in these women who developed two clonally different cancers in the same breast over time. We can speculate that such cases may be due to genetic predisposition as there is clear evidence that the known invasive breast cancer predisposition genes and polymorphisms also predispose to DCIS, particularly ER+ DCIS (Petridis et al., 2019, 2016).
However, the limited data on rare germline variants in this study does not support this theory. Another possibility is that a cancer field effect in the breast has greatly increased the probability of developing new cancers compared to the general population. The “sick lobe” theory and field cancerization concepts may explain why in an affected breast prone to tumorigenesis despite wide local excision of DCIS with histologically clear margins and radiotherapy, a secondary tumor can emerge that is unrelated to the initial DCIS (Curtius et al., 2018; Heaphy et al., 2009). Studies on tumour adjacent normal tissue raise the possibility that non-malignant pre-cancerous cells contribute to recurrences. (Danforth, 2016; Heaphy et al., 2006).

Our data also show that at least 75% of the invasive recurrences are elonally related to the initial DCIS diagnosis, sharing a common genomic lineage that was established from the same ancestral cell in the breast. The genomic data show a very high concordance in both the driver mutations and chromosomal amplifications that were detected in both the DCIS and invasive disease in these patients. These findings support that genomic evolution occurs at the earliest stages of breast cancer progression (within the ducts) in which driver events including TP53, PIK3CA mutations and Her2 amplifications are present at the DCIS stage, prior to breaking through the basement membrane of the ducts to establish the invasive disease. These findings suggest that specific genomic mutations per se do not drive invasion but perhaps a critical combination of mutations and CNAs is required, or alternatively, that characteristics of the tumor and surrounding microenvironment are present at the earliest stages of progression to permit invasion in the later stages of the disease (Casasent et al., 2018, 2017). Further characterization of the DCIS microenvironment may reveal key stromal and immune factors that may create conditions that are permissive for invasion.
The finding that one in five ipsilateral invasive cancers following DCIS are not clonally related has fundamental biologic implications: first, DCIS can no longer be considered solely as a precursor lesion, but rather also a risk lesion for development of further invasive disease. This is similar to the role that has been ascribed to lobular carcinoma in situ (LCIS) where there is both an increased risk of subsequent ipsilateral and contralateral invasive disease (King et al., 2015). Second, the true risk of recurrence from the same population of preinvasive tumor cells has likely been overestimated, thereby confounding the potential benefit of radiotherapy, as radiation likely prevents clonal progression rather than preventing initiation of a new neoplastic process. Third, these data have important implications for accurate identification of predictive biomarkers for invasive progression, since in clonally unrelated DCIS, the notion of biomarkers predictive of invasion is irrelevant. These data may explain why it has been so challenging to identify predictive biomarkers of progression to invasive disease to further underscoring the need to characterize DCIS more comprehensively in the context of the stroma in future studies.

CHAPTER 6: DISCUSSION AND FUTURE DIRECTIONS

6.1 Conclusions

In my dissertation, I constructed a comprehensive unbiased human breast cell atlas (HBCA) using single cell technologies from samples collected from 60 women. I also demonstrate the cellular heterogeneity present in a pathologically normal human breast and associate with clinical metadata for the first time in breast single cell analysis. The thesis reported the dynamic
changes in both epithelial and immune cells based on age, menopausal status, as well as breast density/BMI, suggesting the importance of defining normal can vary based on individual person. Further, I identified unique lactation cell status only existing in individuals experienced pregnancies. While these findings were speculated, this is the first study to report from my knowledge. More importantly, we observed similar cell identity and states defining gene markers that were previously reported as disease-related, exist in normal breast cell states.

This study paved the way for comparing the normal breast transcriptomes to the study of invasive and aggressive subtype of breast cancer, TNBC by creating a tumor breast cell atlas using single cell RNA sequencing in 101 TNBC treatment naïve patients. This is the first study to comprehensively profile immune, stromal and tumor components in such a large cohort of TNBC. Comparing the cells in normal to TNBC breast, many transcriptional programs supporting invasive progression were found in the study. Our data suggests reprogramming and increased proportions of certain cell states such as TECs, CAFs and effector T cells enrichment in TNBC and positive correlation with pathological complete response.

Lastly, our study reported the existence of two models of genomic evolution following direct lineage from DCIS to invasive recurrent progression in a cohort of 24 longitudinally matched patients. With the application of whole exome and single cell DNA sequencing, we were able to find a direct clonal genetic lineage in ~80% of the DCIS-invasive recurrences, whereas an independent lineage in ~20% of the cases that was never reported before.

6.2 Future Directions

6.2.1 Integration of single cell atlases

With the single cell breast atlas context in my thesis, I cannot begin to highlight the importance of integrating these atlases. In my thesis, I attempted to integrate normal and tumor
breast cell atlas to inform on the transcriptional reprogramming occurring during invasive disease progression. To comprehensively analyze complex biological processes, a number of approaches have been are being developed to integrate and harmonize the single-cell atlases. Most of these methods are aimed to correct for batch effects that are being generated due to differences in tissue dissociation protocols and operator differences, however they are all based on the same principle of variance stabilization and dimensionality-reduction. Further, they all harmonize to varying degrees of integration that differs based on the tissues and cell type integrated. More flexible but rigorous statistical and computational frameworks are needed to correct for these batch effects and integrate the cell atlases. I anticipate that integrating the cell atlases will inform on key biological questions in cellular taxonomy, developmental hierarchy, discovery of new genes to identify cells, spatially quantify the cellular location to the anatomical structure of the organ, cell to cell communication. Most importantly, atlas should help researchers to compare healthy reference cells to diseased ones in the relevant tissues — and so facilitate the identification of biomarkers of progression and drug response.

6.2.2 Multiomics view of the disease to infer mechanisms of chemoresponse

Multiple types of genomic, proteomic and transcriptomic measurements from a single sample and even a single cell is needed to fully understand the tissue as a whole. These measurements can inform on the genotype-phenotype relationship in a cell. There are few low throughput, labor intensive and expensive technologies looking into such measurements, but further development is needed (Dey et al., 2015; Macaulay et al., 2015) using combinatorial indexing, nanogrids and microdroplets platforms (Yin et al., 2019). TME reprogramming with chemoresistence has been reported in melanoma (Sade-Feldman et al., 2018) using scRNA-seq and in basal cell carcinoma (Satpathy et al., 2019) using scATAC-seq by identification of specific
cell states associated with response. To understand the genomic evolution of chemotherapy resistance, there have been limited studies using scDNA-seq with small sample sizes informing on adaptive vs acquired resistance models, however, these need to be expanded in a much larger cohort with multiple modalities to get a comprehensive view of the chemoresistance mechanism.

6.2.3 Longitudinal profiling of TNBC to study evolution of chemoresistance

In our study, we interrogated transcriptional heterogeneity at a single pretreatment timepoint in 100 patients. Future work will need to be performed in a longitudinal cohort of TNBC patients to inform on the evolutionary model of progression and chemoresistance using both single cell DNA and RNA sequencing. Functional studies will also be needed to validate the cell states conferring chemoresistance or pathological complete response using methods such as IHC, spatial transcriptomics, etc. Other important future directions include an analysis of matched metastatic tumors to understand whether the chemoresistant clones in the primary tumor seed distant metastases and also confer resistance at these organ sites. These studies will become possible with the ability to do multi-modal analysis in the same cell to get a multiomic view of mechanisms of chemoresistance.

6.2.4 Tumor initiation and invasion in DCIS progression

Genomic studies of tumor initiation, invasion and progression in DCIS represents a major gap in knowledge due to difficulties in obtaining pure tumor cells from the breast ducts. Further due to long time to recurrence, ranging from 2-10 years of some of the DCIS cases, studying the progression in the same patient is challenging. Another extension of this study should be the comparison of normal breast microenvironment to DCIS TME in order to capture the transcriptional reprogramming happening from normal > DCIS > invasive progression. The
logistical challenge of obtaining fresh or frozen tissues and the lack of spatial resolution to study genomic, transcriptomic and spatial evolution and transcriptional reprogramming in the tumor cells and the microenvironment during premalignant breast cancer progression needs to be circumvented with more sophisticated technologies.

6.2.5 Evolutionary models in DCIS-IDC using single cell sequencing

There are two evolutionary models that have been proposed to explain the progression of DCIS to IDCs in cases that have direct genomic lineages: 1) an evolutionary bottleneck, in which a single dormant clone is selected and expands to form the invasive disease, and 2) multi-clonal invasion, in which multiple subclones escape the ducts and establish the invasive disease. To date, these models have mainly been informed by bulk genomic studies of single time point samples using synchronous DCIS-IDC tissues, in which invasion has already occurred or unmatched DCIS-IDC pairs, where interpatient heterogeneity can confound genomic comparisons. In one previous study, it was reported that multi-clonal invasion was common in synchronous DCIS-IDC6, but it remains unclear if this model is also common in pure DCIS cases that have recurrent IDC cancer many years later. A few other studies have shown evidence that most recurrences have a clonal genetic relationship with their matched primary DCIS pairs. In one of the largest bulk genomic studies using a cohort of matched DCIS and recurrent invasive cancers, it was estimated that 75% of patients share a direct genetic lineage, but this study did not distinguish between the two evolutionary models of progression. Moreover, many of these previous studies were limited to ‘bulk’ genomic methods that could not accurately resolve intratumor heterogeneity (ITH) in DCIS and infer models of genomic. High throughput whole genome sequencing methods capable of generating high-quality genomic copy number data from FFPE samples are needed to genomic evolution cancer and other human diseases.
6.2.6 Spatial assays to capture heterogeneity

An upcoming method aimed to capturing spatial heterogeneity at the single cell resolution is the emergence of spatially resolved single cell sequencing. However, the method is limited to acquiring cell suspensions as the starting material which inherently loses the spatial context of the cell in the tissue. One such method was developed by our lab, laser capture microdissection (LCM) method that can be combined with scDNA-seq (Casasent et al., 2018) or scRNA-seq (Nichterwitz et al., 2016). However, these are low throughput (<100 cells) and labor intensive. Upcoming methods such as spatial transcriptomic assays (Ståhl et al., 2016) and Slide-Seq (Rodrígues et al., 2019) are high throughput and can interrogate small clusters of cells (e.g., 10–100 cells, depending on the cellular size) across multiple spatial regions, but do not have single-cell resolution, rather multiple cells can be embedded in the wells. In contrast, in situ sequencing technologies, including FISSEQ (Lee et al., 2014) and MERFISH (Chen et al., 2015) can profile cells at a single-cell resolution but are limited in the capture regions and fewer genes. Hence, development of spatial assays is needed to persevere cellular information and spatial context.

6.2.7 Applications of single cell sequencing to clinical practice

With the advent of single cell technologies and decrease in cost in extracting relevant genomic information from these datasets, considerable efforts are being made to apply single cell sequencing not only to cancer research but also clinical medicine. Single cell sequencing has important applications possibly in the field of early detection, diagnostics, molecular subtyping for patient stratification, drug resistance, identification of biomarkers etc.
6.3 Limitations

Single cell technology has become the go to method for resolving cell to cell and organ heterogeneity. Being able to get single cell measurements from both RNA and DNA, can now shed light into the cellular granularity at the finest resolution possible. The information captured from single-cell sequencing is enormous: allowing to be able to apply it for clinical practice. However, due to its potential application, the generation of such datasets needs to be handled properly and robustly. Firstly, the collection and processing of tissues into single cell suspensions is a tricky process and subtle changes such as digestion times, time the tissue was kept in media, RBC lysis, cryopreservation etc, in the protocol can lead to changes in the transcriptome.

Secondly, while single cell DNA and ATAC sequencing methods can work on archival materials (snap-frozen tissues, optimal cutting temperature (OCT) media frozen samples and even formalin-fixed, paraffin-embedded [FFPE]), single cell RNA sequencing required tissue to be acquired freshly post-surgery which poses a major logistically challenge. This has been partly overcome by isolating single nuclei suspensions from snap-frozen tissues, however I observed in my project that nuclear and cytoplasmic RNA differs leading to differences in genes and pathways. Therefore, collection of fresh samples poses a challenge.

Thirdly, while our current breast atlas study significantly advances towards a comprehensive human breast tissue atlas, a drawback has been the lack of ancestral diversity and representation of underrepresented communities in the sample cohort profiles. Thus, our future study should focus on a diversity in genomic research to advance our understanding of disease and improve outcomes for women of all backgrounds.

Fourthly, our study of genomic evolution in DCIS is limited by low sample numbers and only delineates limited evidence of the evolutionary models using scDNA data. A larger cohort is needed to identify common evolutionary models of progression, and to identify specific CNA
events in dormant subclones that were recurrent across all patients which could serve as biomarkers to stratify progression and non-progression patients.

Lastly, our chemoresistance study is limited by the low frequency of stromal cells observed in the TME in a subset of patients making it difficult to posit the differences in a statistically significant way. Additionally, the study captured transcriptomic information from core biopsies tissues that may not reflect the true spatial biases in the tissue. Further, spatial validation is needed to ascertain the expression of MHC genes in tumor vs immune cells.

*Given these innovations and discoveries, I hope that better and high resolution single cell technologies will pave a way for further understanding disease mechanisms and impact clinical management.*


Elshof, L.E., Tryfonidis, K., Slaets, L., van Leeuwen-Stok, A.E., Skinner, V.P., Dif, N.,


Francis, A., Thomas, J., Fallowfield, L., Wallis, M., Bartlett, J.M.S., Brookes, C., Roberts, T.,
Pirrie, S., Gaunt, C., Young, J.,Billingham, L., Dodwell, D., Hanby, A., Pinder, S.E.,
Evans, A., Reed, M., Jenkins, V., Matthews, L., Wilcox, M., Fairbrother, P., Bowden, S.,

Gao, R., Bai, S., Henderson, Y.C., Lin, Y., Schalck, A., Yan, Y., Kumar, T., Hu, M., Sei, E.,


Heaphy, C.M., Bisoffi, M., Fordyce, C.A., Haaland, C.M., Hines, W.C., Joste, N.E., Griffith, J.K.,


Hernandez, L., Wilkerson, P.M., Lambros, M.B., Campion-Flora, A., Rodrigues, D.N., Gauthier,
A., Cabral, C., Pawar, V., Mackay, A., A’Hern, R., Marchiò, C., Palacios, J., Natrajan, R.,


Huang, Y., Ma, C., Zhang, Q., Ye, J., Wang, F., Zhang, Y., Hunborg, P., Varvares, M.A., Hoft,


Ting, J., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko, O., Poon,
J., Chan, T., Vatrt-Watts, S., Walters, P., Abrar, N., Chan, S., Wiens, M., Martin, L.,
Wilder Scott, R., Michael Underhill, T., Chavez, E., Steidl, C., Da Costa, D., Ma, Y.,
Coope, R.J.N., Corbett, R., Pleasance, S., Moore, R., Mungall, A.J., Mar, C., Cafferty, F.,
Gelmon, K., Chia, S., Hannon, G.J., Battistoni, G., Bressan, D., Cannell, I., Casbolt, H.,
Jauset, C., Kovačević, T., Mulvey, C., Nugent, F., Ribes, M.P., Pearsall, I., Qosaj, F.,
Sawicka, K., Wild, S., Williams, E., Aparicio, S., Li, Y., O'Flanagan, C., Smith, A., Ruiz,
T., Balasubramanian, S., Lee, M., Bodenmiller, B., Burger, M., Kuett, L., Tietscher, S.,
Windager, J., Boyden, E., Alon, S., Cui, Y., Emenari, A., Goodwin, D., Karagiannis, E.,
Sinha, A., Wassie, A.T., Caldas, C., Bruna, A., Callari, M., Greenwood, W., Lerda, G.,
Lubling, Y., Marti, A., Rueda, O., Shea, A., Harris, O., Becker, R., Grimaldi, F., Harris,
S., Vogl, S., Joyce, J.A., Hausser, J., Watson, S., Shah, S., Vázquez-García, I., Tavaré, S.,
Dinh, K., Fisher, E., Kunes, R., Walton, N.A., Al Sa’d, M., Chornay, N., Dariush, A.,
Solares, E.G., Gonzalez-Fernandez, C., Yoldas, A.K., Millar, N., Zhuang, X., Fan, J.,
179, 1207-1221.e22.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar,
K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J.,
Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov,
J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan,
A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D.,
Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter,
N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R.,
French, L., Graffham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M.,
Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A.,
Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier,
L.W., McPherson, J.D., Marr, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin,
K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L.,
Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J.,
Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S.,
Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E.,
Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J.,
Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S.,
Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda,
A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig,
R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker,
P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois,
J., Rosenthal, A., Platzter, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J.,
Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W.,
Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M.,
Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki,
K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roel, B.A., Chen, F., Pan,
H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia,
N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravinda, L., Bailey, J.A.,
Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E.,
Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D.,
Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk,
A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D.,
Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J.,
Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld,
A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa,
K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome

Nurs. 19, 366–376.


Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R.,
Jeanty, S.S.F., Li, C., Amamoto, R., Peters, D.T., Turczyk, B.M., Marblestone, A.H.,
1360–1363.


Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakrarathy, A.B., Shyr, Y., Pietenpol,

Leonardi, G.C., Candido, S., Cervello, M., Nicolosi, D., Raiti, F., Travali, S., Spandidos, D.A.,

214–235.


Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N.,
Mumbach, M.R., Pierce, S.E., Corces, M.R., Shah, P., Bell, J.C., Jhutty, D., Nemec, C.M.,
Schmid, P., Cortes, J., Pusztai, L., McArthur, H., Kümmel, S., Bergh, J., Denkert, C., Park, Y.H.,
Hui, R., Harbeck, N., Takahashi, M., Foukakis, T., Fasching, P.A., Cardoso, F., Untch,
M., Jia, L., Karantza, V., Zhao, J., Aktan, G., Dent, R., O'Shaughnessy, J., KEYNOTE-
Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy,
T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., Fulton, R.S., Kremitzki, M.,
Magrini, V., Markovic, C., McGrath, S., Steinberg, K.M., Auger, K., Chow, W., Collins,
J., Harden, G., Hubbard, T., Pelan, S., Simpson, J.T., Threadgold, G., Torrance, J.,
Wood, J.M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy,
Res. 27, 849–864.
Shaaban, A.M., Hilton, B., Clements, K., Provenzano, E., Cheung, S., Wallis, M.G., Sawyer, E.,
Thomas, J.S., Hanby, A.M., Pinder, S.E., Thompson, A.M., Sloane Project Steering
Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K.,
Haffari, G., Bashashati, A., Prentice, I.M., Khattria, J., Burleigh, A., Yap, D., Bernard, V.,
McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J.,
Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S.K., Griffith,
S.-F., Curtis, C., Rueda, O.M., Pharoah, P.D., Damaraju, S., Mackey, J., Hoon, K.,
Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J.F., Meyer,
I.M., Eaves, C.J., Wasserman, W.W., Jones, S., Huntsman, D., Hirst, M., Caldas, C.,


Shee, K., Muller, K.E., Marotti, J., Miller, T.W., Wells, W.A., Tsongalis, G.J., 2019. Am. J.
Pathol. 189, 956–965.

Shiga, K., Hara, M., Nagasaki, T., Sato, T., Takahashi, H., Takeyama, H., 2015. Cancers (Basel) 7,
2443–2458.

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S.,
Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg,
Å., Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén,


Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y.,

Takeda, A., Hollmén, M., Dermadi, D., Pan, J., Brulois, K.F., Kaukonen, R., Lönnberg, T.,
Boström, P., Koskivuo, I., Irjala, H., Miyasaka, M., Salmi, M., Butcher, E.C., Jalkanen, S.,


Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem,
A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-
R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer,


Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T.,
Alexandrov, L.B., Larsimont, D., Davies, H., Li, Y., Ju, Y.S., Ramakrishna, M., Haugland,
H.K., Lilleng, P.K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D.,
Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, I.J., Jiang, B., Vincent, D., Greene-
Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M.R.,
Med. 21, 751–759.

Yin, Y., Jiang, Y., Lam, K.-W.G., Berletch, J.B., Disteche, C.M., Noble, W.S., Steemers, F.J.,


1480.

VITAE

Tapsi Kumar M.S was born in Punjab, India. She graduated from Jaypee Institute of Information technology in 2008 with a Bachelor of Technology in Biotechnology and went on to work in the finance and Information technology industry till 2012. In 2016, she completed her Masters in Statistics from Rice University, Houston, Texas and joined the PhD program at the University of Texas Health Science Center at Houston MD Anderson Cancer Center. She loves animals except reptiles. She has a dog named Satchmore with whom she loves to go on long walks.