April 2019

Contribution of Retrotransposons to Breast Cancer Malignancy

Isaac D. Raplee
University of South Florida, iraplee@health.usf.edu

Follow this and additional works at: https://scholarcommons.usf.edu/etd

Part of the Bioinformatics Commons, and the Oncology Commons

Scholar Commons Citation
https://scholarcommons.usf.edu/etd/7900

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Contribution of Retrotransposons to Breast Cancer Malignancy

by

Isaac D. Raplee

A dissertation submitted in partial fulfillment
Of the requirements for the degree of
Doctor of Philosophy Medical Sciences
Department of Molecular Medicine
Morsani College of Medicine
University of South Florida

Major Professor: Caralina Marin de Evsikova, Ph.D.
Andriy Marusyk, Ph.D.
Denise Cooper, Ph.D.
Paula Bickford, Ph.D.
Peter Medveczky, M.D.

Date of Approval:
March 20, 2019

Keywords: RNA-seq, bioinformatics, breast cancer, transposable elements

Copyright © 2019, Isaac D. Raplee
# TABLE OF CONTENTS

List of Tables .................................................................................................................. vii

List of Figures .................................................................................................................. viii

Abstract ............................................................................................................................ xi

Chapter One: Breast Cancer, Transposons, and Transcriptomics ............................... 1

1.1 Breast Cancer ........................................................................................................... 1

1.1.1 Statistics ................................................................................................................ 1

1.1.2 Risk Factors .......................................................................................................... 2

1.1.2a Nongenetic, Non-modifiable Risk Factors ......................................................... 3

1.1.2b Nongenetic, Modifiable Risk Factors ................................................................. 4

1.1.2c Genetic Risk Factors .......................................................................................... 5

1.1.3 Types and Subtypes .............................................................................................. 6

1.1.4 Models of Breast Cancer ...................................................................................... 8

1.1.4a In Vitro ................................................................................................................ 8

1.1.4b In Vivo ................................................................................................................ 9

1.1.4c Characteristics During Transformation .............................................................. 11

1.1.4d MCF10A and Derived Cell Lines ........................................................................ 13

1.2 Transposable Elements (TEs) .................................................................................. 14

1.2.1 Classification of Transposable Elements in Humans ........................................... 15

1.2.2 Transposition Properties and Consequences of Transposition ....................... 17

1.2.2a Insertion Mutagenesis and Chromosomal Rearrangement .............................. 17

1.2.2b Insertional Site Preference and Selection ......................................................... 18
Chapter Two: Bioinformatics Optimization of Clinical Samples ........................................... 63

2.1 Introduction .................................................................................................................. 63

2.1.1 Research Purpose and Approach ........................................................................... 64

2.2 Methods ...................................................................................................................... 66

2.2.1 Breast Cancer Samples ......................................................................................... 66

2.2.2 RNAseq Reads Alignment ................................................................................... 67

2.2.2a STAR ..................................................................................................................... 67

2.2.2b HISAT2 ............................................................................................................... 68

2.2.3 Gene Expression Counts ....................................................................................... 68

2.2.4 Data Normalization and Quality Control ............................................................... 69

2.2.4a Normalization ...................................................................................................... 69

2.2.4b Quality Control ................................................................................................... 70

2.2.5 Differential Gene Expression Analysis .................................................................. 70

2.2.5a DESeq2 .............................................................................................................. 71

2.2.5b edgeR ................................................................................................................ 72

2.2.6 Gene Enrichment Analysis Using VLAD .............................................................. 72

2.3 Results ....................................................................................................................... 73

2.3.1 Output of Aligners ............................................................................................... 73

2.3.2 Quality Control .................................................................................................... 73

2.3.3 Gene Expression Profiling ................................................................................... 74

2.3.3a Highly Expressed Genes ................................................................................... 74

2.3.3b Alignment to Pseudogenes ................................................................................. 75

2.3.4 Differential Gene Expression Analysis ................................................................. 76

2.3.4a edgeR of HISAT2 and STAR bams ................................................................. 78

2.3.4b DESeq2 of HISAT2 and STAR bams ............................................................... 78

2.3.4c Comparison of DESeq2 and edgeR Results ..................................................... 79

2.4 Discussion .................................................................................................................. 80
Chapter Three: TE and Transcript Analysis of Clinical Samples

3.1 Introduction ................................................................. 84

3.1.1 Determining Transposable Element Expression During Breast Cancer Progression ............................................ 86

3.1.2 Research Purpose and Approach .................................... 86

3.2 Methods ........................................................................ 87

3.2.1 Transcript and TESA Pipeline ....................................... 87

3.2.2 Alignment of Clinical RNA-Seq Data Using STAR .............. 88

3.2.3 Quality Control Procedure ............................................ 89

3.2.4 Differential Expression Analysis – DESeq2 ....................... 89

3.2.5 VLAD Gene Ontology Analysis of DE Output .................... 90

3.3 Results ........................................................................... 91

3.3.1 Expression of Transcripts Across Cancer Progression ........ 91

3.3.2 Expression of All TEs Across Cancer Progress .................. 92

3.3.3 VLAD Gene Ontology Output ....................................... 99

3.3.3a Validation of Cancer Pathways .................................... 99

3.3.3b Examples of Novel Pathways ...................................... 103

3.3.3c Conserved Transcripts ............................................. 103

3.4 Discussion ................................................................... 106

3.5 References .................................................................... 108

Chapter Four: Regulation of Retrotransposon Activity in an In Vitro Model ............ 111

4.1 Introduction .................................................................... 111

4.1.1 Testing the Roles of Retrotransposons in the Transformation Pre-Malignant to Malignant Breast Cancer ..................... 112

4.1.1a Pharmacological Approach ........................................ 112

4.1.1b Genetic Engineering Approach .................................... 112

4.1.2 Research Purpose and Approach .................................... 113

4.2 Methods ....................................................................... 114
4.2.1 Cell Culture Conditions of Both Cell Lines ......................................... 114
4.2.2 Drugs ................................................................................................. 114
   4.2.2a 3-Deazaneplanocin A (DZNEP) ...................................................... 114
   4.2.2b 3-Deazaadenosine (DZA) .............................................................. 115
   4.2.2c 5-Azacytidine (AZA) .................................................................. 115
   4.2.2d Raltegravir (RAL) ........................................................................ 115
   4.2.2e Elvitegravir (ELV) ........................................................................ 115
   4.2.2f Azidothymidine (AZT) .................................................................. 115
4.2.3 Cell Viability and Density Dose-Response Curves ............................... 116
4.2.4 Scratch Assay .................................................................................. 117
   4.2.4a Pilot Study ..................................................................................... 117
   4.2.4b WoundMaker™ ........................................................................... 117
4.2.5 Transformation Assay ....................................................................... 118
4.3 Results .................................................................................................. 120
   4.3.1 Cell Viability and Density Dose-Response Curve ............................. 120
   4.3.2 Scratch Assay ................................................................................ 123
      4.3.2a Pilot Study ................................................................................ 123
      4.3.2b WoundMaker™ ....................................................................... 125
   4.3.3 Transformation Assay ................................................................... 130
4.4 Discussion ............................................................................................ 132
4.5 References ........................................................................................... 133

Chapter Five: Perspectives ........................................................................... 135

5.1 Pipeline for the Common Clinical RNA-Seq Dataset ............................. 135
5.2 Novel in silico Method for the Discovery of TE Expression in Breast Cancer Progression .............................................................. 136
5.3 Regulating Retrotransposon Expression and Activity Effects on MCF10A and MCF10DCIS Cell Lines Models ............................................ 138
5.4 Future Directions ................................................................................ 140
   5.4.1 Transposable Element Expression in Breast Cancer Cell Lines ...... 140
   5.4.2 Activity of Transposable Elements ................................................. 140
5.4.3 Patient-Derived Xenograft Models ............................................. 140
5.5 Final Thoughts ............................................................................... 141
5.6 References .................................................................................... 142
LIST OF TABLES

Table 3.1: Differentially expressed TEs for each stage ................................................................. 94

Table 3.2: Top 10 GO terms of DE genes for each stage compared to control and each category ................................................................................................. 104

Table 4.1: Cell viability and density dose-response curve experimental design .................... 116

Table 4.2: Scratch assay pilot study experimental design ............................................................. 117

Table 4.3: WoundMaker™ scratch assay experimental design ...................................................... 118

Table 4.4: Transformation assay experimental design ................................................................. 119

Table 4.5 Results Summary for MCF10A .................................................................................. 131

Table 4.6 Results Summary for MCF10DCIS ............................................................................ 132
LIST OF FIGURES

Figure 1.1: Breast Cancer Progression .................................................................12

Figure 1.2: General Classification of Transposable Elements (TEs) .........................14

Figure 1.3: Classification of Human Transposable Elements ...................................15

Figure 1.4: Long Terminal Repeat Retrotransposons Mechanisms of Regulation and Transposition ...................................................................................................................... 19

Figure 1.5: Long Terminal Repeat Retrotransposons Alternative Functions ..............21

Figure 1.6: Long Terminal Repeat Retrotransposon Viral Protein Exaptation ..........22

Figure 2.1: Bioinformatics Pipeline for RNA-Seq Tool Analysis ................................66

Figure 2.2: Performance of HISAT2 and STAR Aligners on the Breast Cancer Series Data .................................................................................................................................... 73

Figure 2.3: PCA Visualization of Gene Expression Data from HISAT2 and STAR Alignments .................................................................................................................................. 73

Figure 2.4: Overlap Between the Highest Expressed Genes in the RNA-seq Breast Cancer Datasets Aligned by HISAT2 or STAR ........................................................................ 74

Figure 2.5: Expression of Retrogenes in HISAT2 and STAR Alignment Data .......... 75
Figure 2.6: MA Plots of Pairwise Comparisons of All Stages ........................................ 77

Figure 2.7: Numbers of Differentially Expressed Genes in Pairwise Comparisons........ 78

Figure 2.8: Overlap of Genes Identified as Differentially Expressed........................... 79

Figure 3.1: Transposon Enrichment Set Analyses (TESA) Pipeline.............................. 87

Figure 3.2: Venn Diagram of Differentially Expressed Transcripts............................ 91

Figure 3.3: Gene Transcriptome and TE Transcriptome .............................................. 92

Figure 3.4: Venn Diagram of Differentially Expressed Transposable Elements ............ 93

Figure 3.5: Differentially Expressed TEs .................................................................... 93

Figure 3.6: Top Ten Overrepresented Gene Ontology Terms of Upregulated
Differentially Expressed Genes in Atypia ................................................................. 99

Figure 3.7: Top Ten Overrepresented Gene Ontology Terms of Upregulated
Differentially Expressed Genes in DCIS ................................................................. 101

Figure 3.8: Top Ten Overrepresented Gene Ontology Terms of Upregulated
Differentially Expressed Genes in IDC ................................................................. 102

Figure 4.1: Alternative Promoter and Nonredundant Genomic Locations of
LTR16B2 .................................................................................................................. 113

Figure 4.2: Initial Dose-Response Curve for SAM Cycle Inhibitors ......................... 121

Figure 4.3: Initial Dose-Response Curve for DNA Methyltransferase Inhibitor ......... 122
ABSTRACT

The components contributing to cancer progression, especially the transition from early to invasive are unknown. Consequently, the biological reasons are unclear as to why some patients diagnosed with atypia and ductal carcinoma in situ (DCIS) never progress into invasive breast cancer. The “one gene at a time” approach does not sufficiently predict progression. To elucidate the early stage progression to invasive ductal cancer, expression signature of transcripts and transposable elements in micropunched samples of formalin-fixed, paraffin embedded (FFPE) tissue was conducted. A bioinformatics pipeline to analyze poor quality, short reads (>36 nts) from RNA-Seq data was created to compare the most common tools for alignment and differential expression. Most samples from patients prepared for RNA-seq analysis are acquired through archived FFPE tissue collections, which have low RNA quality. The pipeline analytics revealed that STAR alignment software outperformed others. Furthermore, our comparison revealed both DESeq2 and edgeR, with the estimateDisp function applied, both perform well when analyzing greater than 12 replicates. Transcriptome analysis revealed progressive diversification into known oncogenic pathways, a few novel biochemical pathways, in addition to antiviral and interferon activation. Furthermore, the transposable element (TE) signature during breast cancer progression at early stages indicated long terminal repeat (LTRs) as the most abundantly differentially expressed TEs. LTRs belong to endogenous retroviruses (ERV), a subclass of TEs. The retroviral and innate immune response activity in DCIS, which indirectly
corroborates the increase in ERV expression in this pre-malignant stage. Finally, to demonstrate the potential role of TEs in the transition from pre-malignant to malignant breast cancer we used pharmacological approaches to alter global TE expression and inhibit retrotransposition activity in control and breast cancer cell lines. It was expected that dysregulation of TEs be associated with increased invasiveness and growth. However, our results indicated that DNA methyltransferase inhibitor 5-Azacytidine (AZA) consistently retarded cell migration and growth. While unexpected, these findings corroborate recent studies that AZA may induce an interferon response in cancer via increased ERV expression. This body of work illustrates the importance of understanding bioinformatics methods used in RNA-seq analysis of common clinical samples. These studies suggest the potential for TEs as biomarkers for disease progression and novel therapeutic approach to investigate in additional model systems.
1.1 Breast Cancer

Breast cancer is a disease of the breast tissue in which cells grow and proliferate aberrantly. Ductal carcinoma is the most common type of breast cancer originating in the ducts that transport breast milk from the lobular glands to the nipples. Some less common breast cancers, e.g., lobular cancers, start in the glands that produce the milk (1) and rare cases start in other tissues of the breast. Ductal carcinomas become invasive upon escaping the basal membrane of the duct and invade the surrounding tissue.

1.1.1 Statistics

Breast cancer is one of the most common types of cancer in American women. Currently, 1 in 8 women and 1 in 833 men will be diagnosed with breast cancer sometime in their life. Approximately 20% of all breast cancers are diagnosed at pre-malignant stages (1). Noteworthy, only ~one-third of patients diagnosed with a pre-malignant breast cancer will progress into an invasive form (2-4). The American Cancer Society estimates that 266,120 women and 2,550 men will be diagnosed with invasive breast cancer in 2018 (5). The rate of incidence generally climbs with age. According to the CDC, in 2015 woman between the ages of 40 and 44 had an incident rate of 126 per 100,000 women. This rate increased and peaked in women between the ages of 70 and 74 at 462 per 100,000
women. While age is the most prominent risk factor for cancer, ethnicity, race, and sex also play a role in susceptibility. For example, Caucasian and African-American women have the highest rate of new cancers at 126 and 123 per 100,000 respectively, whereas Asian/Pacific Islander, Hispanic, and Native American women had rates at 94, 93, and 71 per 100,000 respectively [6].

Breast cancer is the second leading cause of death in women, behind lung cancer. The chance a woman will die from breast cancer is 1 in 38. The earlier breast cancer is diagnosed, the greater chance of surviving 5 years. If the cancer is at stage 0 or I the chance of surviving 5 years is 100%. Stage II and III have a 5-year survival rate of 93% and 72%, respectively. Stage IV breast cancer has a dramatic decrease in survival rate with only 22% of patients surviving the next 5 years. The average survival rate for the first 5 years after diagnosis is 90% and 83% for 10 years after diagnosis [7].

1.1.2 Risk factors

A clear majority of breast cancers develop sporadically clouting its etiology. However, there are well established risk factors associated with increasing your chances of developing breast cancer. The most common non-genetic, non-modifiable risk factors are age, race, early menarche, late menopause, and history of breast cancer. Non-genetic modifiable risk factors include income, education, insurance status, reproductive patterns, menopausal hormone use, tobacco use, alcohol use, fitness and nutrition. Genetic risks factors include mutations in one of the two BReast CAncer (BRCA) susceptibility genes.
1.1.2a Non-genetic, Non-modifiable Risk Factors

As mentioned previously there is an increase in breast cancer incidence and mortality proportional to increasing age. Disease incidence begins to have a sharp incline starting at age 40 and peaking at age 60 throughout the world. The median age at death due to breast cancer is 68 years. African American women have a median age of death at 62 years while non-Hispanic Caucasian women’s median age of death is 69 [8], which may suggest either a genetic predisposition or health disparity by ethnicity.

Besides the significant disparity in age of death, there additional health disparities between non-Hispanic Caucasian women and African-American women for breast cancer statistics. In addition, African-American women have the highest 5-year breast cancer mortality rate irrespective at stage of diagnosis, indicating that early detection is not yielding better health outcomes. In 2012, the breast cancer death rate was 42% higher in African-American women than in Caucasian women [9]. Even after controlling for stage of disease, tumor characteristics, follow-up, uniform treatment, and other breast cancer risk factors African-American women were more likely to die from breast cancer [10-12]. Furthermore, African-American women are more likely to develop a genetically more aggressive type of cancer, having the largest proportion of HR-/HER2 breast cancer compared to other ethnicities [9, 10]. The HR-/HER2 subtype of cancer is associated with a poor prognosis [13, 14]. Conversely, the lowest incidence and mortality rates in the United States of America belong to Asian/Pacific Islander women followed closely by Alaskan natives, American Indians, and Hispanic women [9, 15]. The reason for lower breast cancer incidence in these ethnic groups is speculated to be associated with different reproductive patterns. Specificity, increased duration of fertility increases the risk of developing breast cancer [16]. For example, girls who start menstruating before the age
of 11, have a 20% higher breast cancer risk than girls who begin at age 13. Furthermore, women who began menopause at 55 years old had a 12% higher chance of breast cancer compared to women who began menopause at 50-54 years old [17].

Another biological risk factor is density of breast tissue increases the risk of breast cancer [18]. Women with breast density greater than 50% and 26%-50% were 2.3 and 1.6 times higher risk for breast cancer diagnosis compared to women with breast density between 11%-25% [19, 20]. These observation are further convoluted because denser breasts are harder to diagnose from impaired mammographic detection [21].

1.1.2b Nongenetic, Modifiable Risk Factors

Reproductive patterns. Women who have their first child at a younger age and a greater number of children have a decreased breast cancer risk. Conversely, delaying childbearing, having fewer children, decreased length of time breastfeeding, and hormonal contraceptives are linked to increased breast cancer risk [22, 23]. There is a 50% reduced lifetime risk in women having a first child before the age of 20. There is an association of decreased breast cancer risk when the first child is closer to menarche [5, 24]. Premenopausal breast cancer risk is increased by 5% for each year the first birth is delayed from menarche. Furthermore, each full-term pregnancy was shown to lead to a 12% decrease in postmenopausal breast cancer risk [25]. For every year of breastfeeding there is a 4% reduction in a woman's risk of breast cancer.

Hormone Replacement Therapy. There is a higher risk correlated with prolonged use of post-menopausal hormones use, from starting hormone replacement therapy at the start of menopause, and using combined estrogen and progestin [26]. The risk appears to dissipate back to baseline after 5 years of discontinued use of synthetic hormones [27].
**Tobacco Use.** There is evidence that smoking prior to menopause will increase a woman’s risk of developing breast cancer. Furthermore, women who started smoking before their first pregnancy had a 21% higher risk of developing breast cancer. Conversely, some studies show that women who start smoking after menopause had a decreased risk of developing breast cancer, possibly due to an antiestrogenic effect of tobacco [28, 29].

**Alcohol Use.** Many studies have found a dose-dependent association of alcohol consumption and increased breast cancer risk [30, 31]. Other studies found chronic alcohol consumption associated with breast cancer incidence. Further studies identified binge drinking associates with cancer incidence [32]. While the type of alcohol consumed didn’t affect the incidence rates, the majority of breast cancers associated with alcohol consumption are the estrogen receptor positive subtype [33-35].

**Socioeconomic Status.** Socioeconomic status is more complicated risk factor for breast cancer. For example, women with higher socioeconomic status have significantly higher rates of breast cancer incidence [36-38], whereas, the 5-year survival rates for each stage at diagnosis are lower in patients who reside in lower socioeconomic areas [39]. Reduced access to routine health checkups is linked to poor outcomes for breast cancer. For instance, the group with the highest mortality rate include women without insurance or only Medicaid [40].

**1.1.2c Genetic Risk Factors**

The most common hereditary breast and ovarian cancer is caused by genetic mutations in one of the two BREast CAncer (BRCA) genes, BRCA1 and BRCA2. These mutated genes account for 3% and 10% of all breast cancers and ovarian cancers in women, respectively [41]. Women, without a genetic mutation common in breast cancer, have a 12% chance of being diagnosed with breast cancer sometime in their lifetime.
However, women with a BRCA1/2 mutation have a 45-65% chance of being diagnosed within their lifetime [42-44]. Men who have the BRCA2 gene mutation substantially higher chance of a breast cancer incident. As mentioned previously, less than 1 in 1,000 men will develop breast cancer, however men with a mutation in BRCA2 have a 50-80% in 1,000 chance of developing breast cancer [45-48]. The genetic causes of breast cancer account for 5 to 10% of the overall breast cancer cases in women, whereas in men, up to 40% of cases are caused by a BRCA2 mutation [49]. In addition to mutations in the BRCA genes, there is supportive evidence the following list of genes increases the risk of breast cancer: ATM, BARD1, CDH1, CHEK2, NBN, NF1, PALB2, PTEN, RAD51D, STK11, and TP53.

1.1.3 Types and Subtypes

Histologically, breast cancer has been categorized, broadly, into carcinoma in situ and invasive. Further classification of carcinoma in situ is distinguished by anatomical location, ductal or lobular, known clinically as ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS), respectively. DCIS has been further histologically subclassified into Comedo, Cribriform, Micropapillary, Papillary, and Solid [50]. While these subclassifications are a valuable histological tool, these subclassifications lack molecular definitions proven to be prognostically significant [51]. Invasive carcinomas have several different histological subtypes, as well. The most common histological subclassification of invasive breast cancer is infiltrating/invasive ductal carcinoma (IDC) accounting for 70% to 80% of all invasive forms [52]. IDC has been further subclassified into well (grade 1), moderately (grade 2), and poorly (grade 3) differentiated based on mitotic index, pleomorphism, glandular formation [53]. The other major subclassifications of invasive
breast cancer are invasive lobular, mucinous, tubular, medullary and papillary carcinomas.

Clinical assessments of IDC utilize molecular markers to distinguish differences in prediction of overall survival. Traditional molecular classification of invasive breast cancer relied on markers for estrogen receptor (ER), progesterone (PR), and human epidermal growth factor receptor (HER2). However, through gene expression and immunohistochemical analysis a spectrum of molecular subtypes suggests prognostic status with five major subtypes. Claudin-low molecular subtype is typically triple negative (ER-, PR-, HER2-), associated with the worst prognosis, and a mesenchymal phenotype [54-56]. The basal subtype usually has a triple negative molecular profile, as well, but mimics basal epithelial cells, similar to normal breast myoepithelial cells, with low to no hormone receptor and HER2 expression and high basal membrane marker expression, such as keratins and EGFR [57, 58]. The basal subtype is associated with a poor prognosis [58, 59]. The middle of the prognostic spectrum has HER2 enriched molecular subtypes [58, 60, 61]. HER2 overexpression tumors are typically ER and PR negative and have enhanced expression of genes in the HER2 amplicon with a high rate of TP53 gene mutation associations [62]. The generally poor prognosis associated with the HER2 overexpression subtype is due to the high risk of early relapse when treatment fails to completely eliminate the cancer [61]. Luminal breast cancers have expression profiles similar to the luminal epithelial component of the breast and are associated with ER expression and downstream activation of ER gene pathways [57]. Luminal B tumors are the triple positive (ER, PR, HER2) subtype when diagnosed through immunohistochemically, however expression profile of luminal B tumors reveal only part of the tumors are HER2+ [63, 64]. Luminal A tumors are the ER+, PR+, and HER2-
subtype and have higher expression of ER related genes but lower expression of proliferative genes than luminal B tumors [60, 65]. In general, the luminal tumor subtypes have a good prognosis and make up the majority of breast cancer tumor diagnoses. However, luminal B tumors have significantly worse outcomes than luminal A tumor subtypes [60].

1.1.4 Models of Breast Cancer

1.1.4a In Vitro

The first breast cancer cell line was derived in the late 1950s from a mammary duct carcinoma [66]. Subsequently, breast cancer cell lines have become the principle models for clarifying apoptosis, genetic contributors, migration, proliferation, and signaling pathways. Currently, the most commonly used breast cancer cell line is MCF-7, which is positive for estrogen hormone receptor expression [67]. In general, genome and RNA sequencing has revealed the genetic and molecular expression profiles for the most common breast cancer cell lines and revealing their shared commonalities within the spectrum of clinical breast cancer subtypes. Briefly, no one cell line models the full spectrum of breast cancer heterogeneity, however panels of cell lines encompass large portions of the molecular subtyping [68-71]. Thus, breast cancer cell lines are an appropriate experimental tool applied to derive information pertaining the specific subtype of breast cancer resembling to its molecular profile [72]. The popularity of breast cancer cell lines as a model of breast cancer is primarily for their ease of use and their being perceived as more “relevant” to human disease when compared to animal models. For example, a study showed that steroid hormone dependence is not well modelled in
mice and the elements that contribute to transformation in mouse and human epithelial cells are debatably different [73-76].

Furthermore, cell lines can be grown as xenografts to provide more clinically relevant information as an in vivo model. Notably, the same cell line may adopt many variants between labs with distinct molecular profiles and phenotypes, which may account for variation when comparing published studies [77]. Another important consideration about breast cancer cell lines is that a majority are grown in 2D. Two-dimensional growth is different than the in vivo breast microenvironment. Recent studies indicate that 3D culture more accurately recapitulates and expresses more pathways found in breast [78-83].

1.1.4b In Vivo

Research using in vivo model systems has provided improved therapeutic options for breast cancer, and observations of disease progression.

Xenografts. To investigate facets of breast cancer biology using cell lines that more closely proximate in vivo, cell lines can be xenografted into immunocompromised animals, such as SCID - “severe combined immunodeficiency”- strains of mice. Furthermore, xenograft models allow investigation into multicellular interactions that recapitulate more clinically relevant observations. These facets include tumor initiation and growth, microenvironmental effects, and metastatic progression [84]. Albeit, xenograft models have many limitations. First, as mentioned previously, xenografts must be placed into immunocompromised animal models. The immune system has been implicated in tumor progression and development for early stage breast cancer and metastasis [85-87]. Many xenografts are generated by subcutaneous injections, which is not a clinically relevant microenvironment. Conversely, orthotopic transplantation of
cells into mouse mammary glands better approximates the human patient microenvironment. However, mouse and human mammary stroma have differences that need to be considered [88]. Xenografts in mouse models preferentially metastasize to the lungs and rarely grow at other common metastatic sites observed in human breast cancers [89]. Finally, many xenografts originate from a single breast cancer cell line, which fails to mimic the heterogeneity of intratumor. Recently, xenograft studies with entire panels of breast tumor subtype cell lines suggest potential for a suitable translational preclinical model in vivo [68]. Despite these limitations, xenograft models have produced many important insights relevant to breast cancer biology, prognosis, and treatment [90-92].

**Genetically engineered mice.** The spontaneous tumor generation of genetically engineered mice (GEMs) into the correct tissue of interest provides great insight into early tumor initiating events. However, the traditional regulatory elements used to drive transgene expression in GEMs and the mouse microenvironment may not reflect clinical relevance. For instance, GEMs tumor pathology does not perfectly reflect the disease in humans [93, 94]. The attempt to recreate breast cancer subtypes in vivo has led to the generation of lineage specific promoters. To drive basal lineage oncogenic events keratin 14-Cre and keratin 5-Cre have been used. Whey acidic protein promoter (Wap)-Cre and keratin 8-Cre drivers target luminal cells for oncogenic events. To target ER- or ER+ luminal cells beta-lactoglobulin-Cre and Cited1-Cre are used, respectively. Despite these lineage specific promoters, many off-target and unexpected expression has confounded the interpretation of experimental results [95-98]. Regardless of confounding issues, genetic profiling of GEMs aligns to their relevant molecular cancer subtypes [99-101].
Furthermore, studies with GEMs representing subtypes of breast cancer has elucidated tumor cell of origin [102, 103]. For example, the BRCA1 deficiency tumor types was thought to originate from basal stem cells, however GEMs it delineated their cell type of origin to luminal progenitors [104, 105].

Despite the success of targeted therapies in breast cancer, drug resistant relapse remains an issue. Clinically relevant models of genetically engineered mice have been generated to study drug resistance, drug combination therapies, and metastasis. These models have elucidated putative mechanisms of drug resistance and can shutdown critical pathways that underlie drug resistance [106-108]. The use of GEMs in preclinical breast cancer trails for drug combination therapies relies heavily on design and collaboration between clinicians and scientists. When considered thoughtfully such studies have resulted in FDA approved therapeutic options [109, 110]. Although relatively few GEMs recapitulate clinical metastases, a few GEM studies have been instrumental in elucidating lymph and lung sites of breast cancer metastasis [111]. A recent study using CRISPR/Cas9 in combination with GEMs demonstrates their potential as a model of functional relevance to breast cancer tumors highlighting future possibilities with GEMs [112].

1.1.4c Characteristics During Transformation

Despite these models have contributed significantly to understanding breast cancer subtypes in late stages[113], relatively little research investigates the factors underlying early stages of cancer transformation. There are a few hypotheses that describe the transformation of breast cancer subtypes from normal tissue to aggressive metastasis, which are mutation-of-origin, cell-of-origin, and a hybrid of both ideas [114-116].

Briefly, genetic profiles of subsets of normal mammary epithelial cells were derived and compared to gene expression in breast cancer subtypes. These comparisons indicate
there is a high overlap of normal luminal gene expression with luminal breast cancer subtypes [117]. Furthermore, claudin-low expression profiles resembled metaplastic CD10+, epithelial-to-mesenchymal transition, and cancer stem cell profiles [118]. These data indicate a cell-to-lineage tumor origin. Conversely, analysis of basal-like and HER+ subtypes found expression profiles that were similar to luminal progenitor cells alluding to a mutation-of-origin model [119, 120]. Additionally, when luminal and basal mammary epithelial cells were isolated, transformed and implanted into immunocompromised humanized mice, the transformed luminal cells generated both luminal-like and basal-like tumors and the basal transformed cells generated rare metaplastic tumors that resembled the claudin-low subtype [116, 121]. Currently, a hybrid model best explains cellular transformation within the of tumor origin.

In general, metastatic ductal carcinoma begins as epithelial atypical growth inside the breast duct surrounded by myoepithelial cells and a basement membrane. This stage is known as early neoplasia, or more commonly as Atypia (Figure 1.1, second panel from left). As further mutagenic events occur, a heterogeneous population arises filling the

![Figure 1.1 Breast Cancer Progression](image)

Figure 1.1 Breast Cancer Progression Top. Histological representation of ductal breast cancer progression. Bottom. Cartoon representation of ductal breast cancer progression. Left to right. Normal ductal tissue, atypical growth occurs, Ductal Carcinoma *in situ*, and Invasive Ductal Carcinoma. (Unpublished)
duct, which is called as ductal carcinoma in situ (DCIS) (Figure 1.1, third panel from left). As mentioned previously, only 1/3 of these ductal carcinomas escape the basement membrane resulting in invasive ductal carcinoma (IDC) (Figure 1.1).

**1.1.4d MCF10A & Derived Cell Lines**

The numerous models available for breast cancer development and metastasis have provided invaluable information. The MCF10A cell line is the most commonly used model for normal breast tissue *in vitro*. These cells were derived from benign fibrocystic diseased breast tissue that were spontaneous immortalized in culture. They have amplifications of the *Myc* gene and genomic deletions of p16 and p14ARF genes. They are not tumorigenic and are ER\(^{-}\) [122]. Genomic stability studies determined that the MCF10A cell line is fairly stable and diploid with exception to chromosome 17 which had higher percentages of being instable [123]. When cultured on top of Matrigel they form spheroids with a hollow lumen that is covered by a basement membrane [124, 125]. Several cell lines have been derived from MCF10A cells by inducing transformation often with gene expression constructs. As a model of early transformation, MCF10A cells were transformed with H-ras which generated MCF10AT cell line [126, 127]. They grow on 2D environment, similar to the epithelial cells from which they were derived. In 3D culture they form multi-acinar structures [128]. These cells were xenografted into immunocompromised mice and generated nodules that progressed into atypia and then carcinoma *in situ* in a quarter of the mice [126, 129]. The MCF10DCIS.com cell line was generated from tumors that grew from the xenografted MCF10AT cell lines. Injection of the MCF10DCIS.com cells into SCID mice produced tumors that were composed of tightly packed tubular structures with central necrosis and basement membranes surrounding
each node intact. Later passages of the cell line that were in injected into SCID mice generated more aggressive and invasive tumors [130, 131].

1.2 Transposable Elements (TEs)

Transposable elements (TEs) are genetic components capable of transposing or replicating themselves within a host genome. Barbara McClintock discovered transposable elements through her work with maize. Initially distributing them into two categories: TEs that are autonomously transposing, containing DNA to encode the necessary components to reverse transcribe itself and non-autonomous TEs, which lacks the ability to reverse transcribe and relies parasitically upon the reverse transcription machinery from other transposons (Figure 1.2) [132]. A widely adopted proposed classification system for eukaryotic transposable elements can be found in a perspective paper [133]. There are two classes of TEs, class I are described as RNA mediated replication and DNA mediated replication is class II. Both classes have autonomous and non-autonomous elements (Figure 1.2).
1.2.1 Classification of Transposable Elements in Humans

Class I TEs are often referred to as retrotransposons and include subclasses of long terminal repeat elements (LTRs) and non-long terminal repeat elements (non-LTRs) (Figure 1.3). LTR retrotransposons have many sub-classes; however, the most widely known and studied are divided into the following three groups; mammalian apparent LTR retrotransposons (MaLR), endogenous retroviruses (ERVs), and Gypsy-Ty3/Copia-Ty1 elements. The Copia/Ty1 group and the Gypsy/Ty3 group are distinguished by the order of the three protein domains, protease (PR), reverse transcriptase (RT), and integrase (IN), encoded within the polymerase (pol) gene of the elements.

![Figure 1.3 Classification of Human Transposable Elements](Unpublished)

**Figure 1.3 Classification of Human Transposable Elements.** Elements are in hierarchical order starting with generalized transposable elements moving down into classes, types, and subtypes. (Unpublished)
The pol region of Copia/Ty1 elements has the order (PR, IN, RT) whereas the Gypsy/Ty3 group has the more familiar arrangement (PR, RT, IN), which is also the order found in a majority of endogenous and exogenous retroviruses.

The non-LTR retrotransposon class in humans has two sub-classes long interspersed nuclear elements (LINEs or Ls) and short interspersed nuclear elements (SINEs). There only two SINEs (SVA and Alu) and one LINE (L1) thought to be active in primates from these sub-classes. L1 has an autonomous retro-transposition through an RNA intermediate by two open reading frame proteins ORF1p, a nuclear acid chaperon, and ORF2p, an endonuclease and reverse transcriptase [134]. L1s constitute a large portion of the human genome, however only 80 to 100 are thought to be actively replicating and transposing. L1 integration is initiated by ORF2p nicking the target DNA, which serves as a primer for the local reverse transcription of ORF2p. SINEs are short non-autonomous retrotransposons that are amplified in the genome via LINE machinery [135]. However, they do not require the ORF1p for transposition [136]. SINE-VNTR-Alus (SVAs) are on average ~2kbp long with a hexamer repeat region, an Alu-like region consisting of two antisense Alu fragments, a region of variable tandem repeats, a HERV-K10 env and LTR region and a poly-A signal ending with an oligo(dA) – rich tail. There are suggestions of RNA polymerase II being the transcribing factor, however there is no internal promoter and possibly relies on promoter activity in flanking regions [137, 138]. The Alu repeat is > 500 bps long and has a dimeric structure formed by the fusion of two monomers from the 7SL RNA gene [139]. The 5’ region contains an internal RNA polymerase III promoter (A and B boxes) and ends with an oligo(dA) – rich tail of different lengths [140]. Alu’s do not possess RNA poly III termination signals and instead extend until a terminator (usually repeat thymines) is found [141].
Class II transposable elements are DNA transposons (Figure 1.3). These elements are a “cut and paste” mechanism, which means they do not use an RNA intermediate like class I transposons. Most DNA transposons encode a transposase protein flanked by terminal inverted repeats (TIRs). The autonomous group of DNA transposons are further subclassified into several groups, which include Tc1/Mariner [142], piggyBac [143-145], and hATs [146, 147]. The nonautonomous DNA transposons are miniature inverted-repeat transposable elements (MITEs) [147], which do not have transposases and act as microRNAs [148].

1.2.2 Transposition Properties & Consequences of Transposition

It has been reported that about half of the human genome is made up of repeat elements and that the TE's activity in the genome has been in decline [149]. The accumulation and movement of TEs can be attributed to their selfish replicative nature [150, 151]. Many studies have attributed TEs with influencing chromosomal structure and rearrangement. Furthermore, TEs have been implicated in mutagenic activities.

1.2.2a Insertion Mutagenesis and Chromosomal Rearrangement

More than one hundred human inherited diseases have been causally linked to de novo germline TE insertion mutagenesis disrupting a wildtype gene function [152, 153]. For example, a de novo transposition and insertion of Alu Ya5 SINE retrotransposon in the ALMS1 coding sequence was found in a Turkish cohort to drive Alström Syndrome [154]. LINEs make up the majority of these active transposition events [155]. In somatic cells, LINE insertional mutagenesis has been implicated in tumorigenesis and malignancy [156-158]. Additionally, TEs have been linked to several types of cancer by inducing chromosomal rearrangements in somatic cells [159]. There is little evidence for LTR
retrotransposons actively transposing to cause mutations in humans [74, 77, 149], however the potential to cause mutations has not been ruled out. Conversely, in mice LTR retrotranspositional activity inducing mutations is well documented [160, 161].

1.2.2b Insertional Site Preference and Selection

Deep sequencing technologies and classical genetic approaches have provided detailed resolution about the integration site selection of TEs. TEs may integrate in sequence specific locations, chromatin domains, chromosomal regions, or all along chromosomes. Furthermore, TEs may integrate in a given preferential position but can be subsequently selected to integrate in different specific locations.

Integration and selection preferences can vary among of mammals, including primates and humans. In humans, L1 TEs are by far the most active but the apurinic-apyrimidinic endonuclease of L1 does not appear to have a site specific preference in the genome integrating in a dispersed manner with a bias for AT rich regions [162]. The SINE Alu, which utilizes L1 machinery in trans for replication behave in similar manner for AT rich site selection de novo. A majority of L1 transposition sites were rendered inactive due to a 5’ truncation that commonly occurs during its insertion [163]. DNA transposons encode a transposase that bind inverted repeats flanking the coding sequence of the transposon. The transposase is responsible for excising and transposing the DNA TE to double stranded DNA that has an identifiable short nucleotide sequence [164, 165]. LTR retrotransposons also recognize short nucleotide sequences that usually ensure DNA flexibility for the molecular manipulation required for integration [166-169]. Recent studies to elucidate integration patterns of endogenous and exogenous retroviruses found insertions were more likely to occur in gene-rich regions near active histone marks [170, 171].
1.2.2c Recombination at TE loci

Due to the copy and paste mechanisms of many retrotransposons, they create multiple homologous regions in the genome that can induce deletions, duplications, or inversions of chromosomal DNA upon recombination events [172-174]. If the copy of the retrotransposon generated is identical and in the same orientation to the first copy, recombination may delete the sequence in between the two TEs. If the two copies are opposite orientation, then an inversion of the sequence between them may occur due a recombination event.

A. B. C. Key

Figure 1.4 Long Terminal Repeat Retrotransposons Mechanisms of Regulation and Transposition. Top 3A. DNA methylation represses LTR expression. Bottom 3A. Demethylated LTR is actively transcribed. 3B. Reverse transcriptase from LTR turns LTR RNA into a DNA. 3C. Integrase incorporates newly reverse transcribed LTR into novel genomic location. (Unpublished)

1.2.3 Alternative Functions of LTR Retrotransposons

Some of the observable consequences of actively transposing elements are often interpreted by scientists that LTRs act in a selfish manner to propagate its own DNA with potential consequence to be mutagenic. Currently, there is a lack of evidence for retrotransposition activity in humans because generally, LTRs are heavily silenced in the genome by DNA methylation (Figure 1.4A) and older LTRs are found greater than ~5kbps
from native gene promoters, due to the negative impact on expression of proximal genes [175]. Despite their quiescent transposable activity, many LTRs exhibit tissue specific expression patterns [176-178]. For example, in mice dynamic changes in gene expression patterns are associated with the expression of specific TEs [179-181]. Furthermore, Göke et al. observed specific LTR families characterize the different stages of early embryogenesis in humans [182]. Recently, it has been established that many LTRs have been coopted for host use as promoters for coding and noncoding transcripts and ancient viral protein expression. These alternative functions are explored further below.

1.2.3a LTRs as Alternative Promoters for Coding Transcripts

LTRs autonomously recruit cellular transcription factors (TFs) because they contain numerous transcription factor binding sites (TFBSs) within their LTR regions. These LTRs that contain TFBSs contribute to about 20% of functional binding sites in humans [183]. Many LTRs have tissue specific expression and measurable transcription levels. Interestingly, some LTRs act as promoters and drive tissue specific expression of genes, such as (but not limited to), ADH1C in liver, CYP19A1 and ~10% of ENTPD1 in placenta, and MKKs in testis (Figure 1.5A) [184-188]. Many LTRs generate gene variants due to their placement in introns and can act as an alternative promoter. For example, in ~11% of melanoma cases, a LTR insertion between exon 19 and 20 of ALK acts as alternative promoter, which generates an isoform that is constitutively active, promoting oncogenesis [189]. Many other examples of LTRs and other TEs acting as alternative promoters in cancer can be found here (Figure 1.5A) [190].

1.2.3b LTRs Derive Important Noncoding Transcripts

Roughly 70% of lncRNAs are associated with a TE within their transcript and ~30% of lncRNAs are derived (Figure 1.5B) directly from TEs. In a majority of genomes,
the distribution of TEs generating lncRNAs is proportional to the percent of genomic content they represent. Albeit in the human genome, LTRs, account for ~8% of the genome but make up the majority of the lncRNAs derived from TEs. Conversely, LINEs comprise ~40% of the human genome but have disproportionately low contribution in lncRNA expression [191]. Recently, it was found that HERV-H derived lncRNAs are highly and specifically expressed in hESCs [192, 193], define the naïve stem-cell state [194], and are essential for the maintenance of pluripotency [195]. There is evidence that lncRNA are not the only noncoding regions effected by TEs. For example, many microRNAs derived from TEs have also been described as having important regulatory roles for the host cell [196].

![Figure 1.5 Long Terminal Repeat Retrotransposons Alternative Functions](image)

**Figure 1.5 Long Terminal Repeat Retrotransposons Alternative Functions.** A. Alternative promoter for downstream gene with a TFBP that has a higher affinity for the TE promoter than its primary promoter producing high quantities of gene product. B. TE derived long noncoding RNA. TFBP transcription factor binding protein. (Unpublished)

1.2.3c Exaptation of LTR ERV TEs as New Genes

There are a few ERVs from the LTR retrotransposon family that have had their ancient viral genes “domesticated” for host use in placental functions, immune system roles, and neuronal activity (Figure 1.6). An interesting example comes from the
Independently co-opted *syncytin* genes from ERV *env* regions. The products of these genes have roles in cell-cell fusion, formation of the syncytiotrophoblast, and immune suppression of mammal hosts during pregnancy [197-201]. Another example of a viral particle being repurposed for host functions is the neuronal *Arc* gene, which is a remnant of an ERV *gag* gene [202, 203]. The protein product of the *Arc* gene is a multifunctional hub protein responsible for long-term synaptic plasticity which is critical for memory and cognition [204].

**Figure 1.6. Long Terminal Repeat Retrotransposon Viral Protein Exaptation.** Expression of endogenous retroviral genes produce viral protein products, such as, viral envelope proteins (grey circle labelled E) and viral structural proteins (green circle labelled G). (Unpublished)

### 1.2.4 Endogenous Regulation of TEs

To ensure tissue specific expression of specific TEs and minimization of negative impacts on genome fitness TEs require tight regulation. In somatic cells, dense DNA methylation is thought to be the main repressor of TE expression and activity [205-209]. Thus, DNA hypomethylation or demethylation occurs in these cells, TEs are de-repressed [210-213].

Interestingly, early germline cells and the early embryo are able to control TE expression activity without DNA methylation. These cells undergo global DNA demethylation and yet still rely on tightly regulated repression and transcription of TEs,
In fact the timing of the expression for specific TEs dictates the pluripotency of the cell [179, 214]. One mechanism of specific TE repression is with Kruppel-associated box zinc finger (KRAB-ZF) proteins whose numbers strongly correlate with the number of LTR elements in mammals [215]. The KRAB domain is responsible for recruiting TRIM28 (also known as KAP1, Tif1β, and KRIP-1), which acts as a scaffold for many proteins, such as SETDB1 a histone methyltransferase responsible for H3K9me3 marks [216, 217]. The major function of ~two-thirds of KRAB-ZFPs, inferred from genome-wide binding profiles or direct loss-of-function studies, is the binding, epigenetic marking, and repression TEs [218-221]. This suggests that KRAB-ZFPs play a major role in initiating stable epigenetic marks via de novo methylation of ERVs throughout embryogenesis [222]. Furthermore, ablation of the KRAB-interacting protein TRIM28 or SETDB1 activates several TEs [223, 224]. Interestingly, there is experimental evidence that some KRAB-ZFPs function beyond early development and even control ERV expression in adult tissues [225]. These findings suggest a tight rope regulation of sequence specific TEs which may be necessary for tissue specific expression and activity.

Further regulation in male gametes is obtained through piwi proteins and their associated piRNAs. Briefly, abundantly expressed TEs are spliced into pools of small RNAs that bind to PIWI proteins and guide them to target DNA sequences. The targeted sequences are silenced and maintained quiescent until de novo DNA methylation occurs [226-228].

Finally, in differentiated cells, evidence of chromatin modifications associated with condensed repressed states at TE insertions provides further regulation beyond DNA methylation. Chromatin modifications that prevent transcription initiation at these TE insertions are methyl marks at histone 3 lysine 9 and 27 (H3K9me and H3K27me).
Additionally, there is a lack of transcription initiation chromatin marks at TE locations, such as histone acetylation [229-231].

1.3 Introduction to Transcriptomics

Transcriptome analysis allows for the detection differentially expressed genes in breast cancer vs. non-diseased tissue that may drive its pathogenesis. The variety of technologies available to researchers makes choosing the most appropriate platform to address and resolve specific scientific problems (or hypotheses) using transcriptome analysis a daunting task. While some researchers believe microarrays are the most reliable due to their maturity, others embrace next-generation sequencing (NGS) as the superior method because it is the current vanguard of molecular technology. Assumptions in data analysis can skew and obscure data interpretation of gene expression if the hypothesis and, most importantly, the experimental design don’t mitigate the shortcomings of each platform. After gene expression has been measured, the researcher must also choose from numerous software programs and analyze expression data. Therefore, the goal of this section is to provide brief explanations of the history, strengths, and limitations of transcriptomics technologies as it pertains to gene expression analysis as a resource guide to interpreting transcriptome experiments within the context of clinical data.

1.3.1 Materials and Methods Used for Transcriptomic Studies

Transcriptome analysis creates a detailed molecular synopsis of cellular physiology by elucidating the mRNA available for translation and/or the abundance of the various
transcripts, such as noncoding RNAs or microRNAs. Techniques used in transcriptome analysis belong to two broad classes; hybridization-based or sequencing-based. These two types of transcriptome analyses are versatile to examine differential gene expression in many cellular physiological contexts, comparing developmental stages, cell cycle, or in disease states. The time and cost of transcriptome analysis has been greatly reduced by the development of microarrays and, more recently, NGS, particularly when compared to older large-scale gene expression analysis technologies such as serial analysis of gene expression (SAGE), expressed sequence tag (EST) libraries, differential display, etc. Given the variety of factors affecting breast cancer and the multiple pathways involved, transcriptomics is a useful tool for diagnostics, discovery science, and pinpointing molecular mechanisms in both clinical and translational models. Transcriptomics also provides a way to identify and test novel treatments and therapeutics that directly correct the underlying breast cancer, in addition to novel therapeutic approaches to alleviate symptomology.

1.3.2 Gene Expression Analysis History of RNA Identification

1.3.2a Sanger sequencing

Sanger sequencing is the invention fundamental for developing modern methods, such as NGS, to sequence for expressed genes in transcriptomic studies. Sanger sequencing is the first-generation method of determining DNA nucleotide sequence based on chain-termination idea developed in 1975. Modern modification of this classic method is based on in vitro DNA elongation of query template, which is interrupted by occasional incorporation into nascent DNA strands of differently labeled di-deoxynucleotide
triphosphates (ddNTPs) present in the reaction mixture along with dNTPs. Unlike dNTP, incorporation of a ddNTP halts extension of synthesized strand; because ddNTP is labeled, the synthesized fragment is detected as a single band of certain color after size separation, and collection of these bands represents DNA sequence [232]. Expressed genes are identified using various methods to harvest RNA to make and use cDNA for expression studies, including subtractive cloning, EST, SAGE, differential display analysis, and microarray analysis. Once identified, gene interactions with other genes can be pursued experimentally [232].

**1.3.2b Subtractive Cloning**

Subtractive cloning is an inexpensive and readily available technique in individual biomedical and clinical laboratories to analyze gene expression using basic molecular biology resources and bioinformatics tools. A sequence, the “tracer” is hybridized with a complementary sequence, the “driver”, which is missing a sequence of interest. The two sequences are combined, and reannealing is dictated by the driver, which is present at least 10 times higher than the tracer. Once annealed, driver-tracer complexes are removed in the subtraction step which is repeated until all common sequences are annealed and removed. The remaining tracer sequences create a tracer specific library that represents differentially expressed genes. Hybridization of cDNA may cause some bias for small fragments of cDNA that hybridize faster than long sequences, but is resolved by PCR amplification [233].

**1.3.2c cDNA Libraries and Expressed Sequence Tags (EST)**

In the late 1970s, cDNA libraries [234] became popular for gene discovery and expression analysis, as the library clones were stable, reproducible, and recoverable
representations of mRNAs isolated from distinct organs and species [235, 236]. The idea to execute Sanger sequencing of many (i.e., >1,000) library clones to understand gene expression in a given tissue spurred the EST project [237]. ESTs are derived from cDNA libraries by random sampling, followed by arraying the sampled clones for future use, and then executing a single sequencing reaction for each clone. Most expressed genes can be unambiguously identified through these relatively short (~ 300-500 nt) EST sequences. Data are collected using automated DNA sequencers and analyzed using bioinformatics tools. ESTs are clustered and assembled to construct consensus sequences using, e.g., CAP3 program [238]. Meaningful data are generated with high throughput preparation of either normalized or non-normalized cDNA libraries [239]. ESTs allow for de novo gene discovery [235, 236], and large-scale prediction of gene products and function [240, 241]. An EST warehouse with high-quality data, including breast epithelium and breast cancer, can be found in Unigene and ENSEMBL. Unigene uses reference sequences along with Genbank mRNAs for cluster generation, and contaminants, low-complexity and repeat sequences are identified with RepeatMasker and DUST [239].

1.3.2d Serial Analysis of Gene Expression (SAGE)

The next development in transcriptomics was Serial Analysis of Gene Expression (SAGE) in 1995. SAGE constructs cDNA libraries in a similar fashion to ESTs but using shorter tags to identify genes. Libraries are constructed by using restriction enzymes and primer “linkers” ligated into the fragments encoding a recognition site for type II restriction enzyme [242]. The cDNA fragments are then amplified and digested resulting in the production of a 13-20 base pair cDNA fragment or “SAGE tag” and are associated with the initial four base pair restriction enzymes. One advantage of the SAGE method is
high-throughput sequencing resulting from its pairing with an automated sequencer although the bioinformatics tools required to analyze the libraries are highly specialized. SAGE analysis can be successfully used for de novo expression profiling, but the short length of the SAGE tag can impair differentiating between highly homologous genes.

**1.3.2e Differential Display Analysis**

Differential Display is a PCR-based method to measure differential gene expression. Differential Display reverse amplifies mRNA transcripts using primers from random sequences and visualized using gel electrophoresis. The presence of a band confirms gene expression and its intensity represents relative amount of transcript [243]. Differential display detects all expressed genes without using specific primers, making it a robust, inexpensive discovery tool. Current advances in Differential Display incorporate the use of fluorescent labels with automation to yield high throughput analyses [244, 245].

**1.3.2f Microarray Analysis**

In 1995, cDNA microarrays superseded the method of Differential Hybridization, introducing the use of miniature spotted DNA probes and fluorescent labeling of sample, reducing the redundancies after hybridization. Pools of known cDNAs (spot) in indexed locations on glass slides represent known genes. Total sample mRNA is reverse transcribed, cRNA amplified by in vitro transcription, and then hybridized to microarray slide. The intensities of the spots produced are then recorded and analyzed by computer software to determine the expression level of a gene [232]. One advantage of cDNA microarrays over EST or SAGE is the ability to analyze gene expression differences under various experimental conditions concurrently by using different fluorophores during the
cRNA transcription. Examples of microarray technology use in breast cancer research are X, Y, and Z. Microarray analysis requires substantially less poly(A) RNA (0.5-2.0 µg) compared to cDNA libraries or Differential Display methods, albeit typical limitations are the quality, specificity and signal discrepancy of the probes on the array. After introduction, microarray analysis became commonplace and made labor-intensive cDNA and SAGE libraries essentially obsolete, despite this method’s restriction to previously discovered transcripts only (e.g. identified via EST libraries) and its inherent inability to discover novel genes, alleles, or splice variants [246]. With its high throughput method requiring low manual labor, low amount of starting RNA, and streamlined bioinformatics processing, microarrays provide an attractive alternative to sequencing for large-scale transcriptome analysis.

1.3.3 Next-Generation Sequencing and Deep Transcriptome Analysis

Second generation sequencing techniques emerged in 2005 and equipment fundamentally differs from first generation sequencers because multiple different DNA molecules are sequenced concurrently. As a result, tens of thousands to hundreds of millions of individual sequencing reads are produced with each run. Different principles underlying sequencing and detection and different chemistries behind various platforms lead to large differences in read length, base call accuracy, and total number of output reads. The largest obstacle for second generation sequencers is obtaining read length to read quality ratios comparable to Sanger sequencing, with most platforms producing average reads with less than 300 bases. In addition, the samples are sequenced in a stop-read-start manner that leads to lengthy processing times, with some platforms requiring
over a week for a single run to complete. To make these platforms economical, the number of reads per run has been increased through the introduction of larger machines, such as the Illumina HiSeq series, or denser chips, in the case of Ion Torrent. However, the larger sequencers have a substantially higher price and require processing at full capacity to benefit from the increased throughput and, consequently, are not typically found in individual laboratories or small research consortia. There are smaller platforms available from Illumina, 454 Roche, and Ion Torrent that produce longer length reads than the larger sequencers, thereby suit the needs of small laboratories [247].

1.3.3a Basic principles of NGS sequencing

All second-generation sequencing platforms require modification and amplification of sample DNA. Samples are fragmented, and adapters are annealed to the ends. For platforms that use emulsion PCR (emPCR) to amplify the samples, the adapters allow the fragments to bind to complementary bases on the emulsion beads. SOLiD sequencing further modifies the fragments after amplification by adding regions that allow the fragments to covalently bond with the sequencer slide. The Illumina platform uses a bridge PCR to amplify the samples, which have been modified with adapters to the base pair with oligonucleotides embedded on the sequencer slide.

Each platform also employs a different method for generating the base calls for each sample, but only Ion Torrent does not use a light-based recording method. The base calls are reported by pyrosequencing in 454 Roche platforms, and by fluorescent tag cleavage in Illumina and SOLiD platforms. The Illumina platform produces forward and reverse reads from each DNA fragment and SOLiD identifies each fragment’s bases twice,
thereby increasing accuracy. Ion Torrent uses a microchip with pH meters incorporated into each well to detect the release of an H\(^+\) ion with each base incorporated.

Extension of fragments occurs during sequential “flooding” of the sequencing reaction chamber with solutions containing specific nucleotides. Illumina differs from other platforms by using a reaction mixture containing all 4 nucleotides. The Illumina nucleotides are modified with a fluorescent group plus a terminator to prevent introduction of additional bases in the cycle. The fluorescence is recorded, and its tag cleaved before flooding the sequencer with the nucleotide-containing reaction mixture again. In pyrosequencing, the nucleotides have a modified pyrophosphate group that is cleaved after addition. SOLiD sequencing uses di-base oligonucleotides with a 3-base extended region and a fluorescent tag. An (n+1)-long primer is added after each round of synthesis which, after 5 repetitions, emits two base signals for each incorporated nucleotide. Nucleotides in Ion Torrent sequencers are added in alternating floods of A, T, C, and G. As each base is paired to the fragment, an H\(^+\) ion is released and detected by the sequencer microchip.

1.3.3b Development of single-cell RNA sequencing strategies

The recent ability to interrogate the transcriptome of individual cells using second generation sequencers has revealed heterogeneity in gene expression of individual cells within a population. As the name implies, single-cell RNA sequencing (scRNA-seq) relies on the isolation and amplification of transcriptomes from individual cells, and many different isolation and amplification strategies have been developed, such as Cel-seq2 [248], Smart-seq2 [249] and Drop-seq [250]. Isolation of individual cells is accomplished by using microfluidic capture chips (Cel-seq2), fluorescence activated cell sorting (Smart-
seq2), or droplet emulsion (Drop-seq). Most scRNA-seq protocols, excluding Smart-seq, incorporate cell-specific barcodes during the reverse transcription reaction that allows for a large amount of multiplexing. Smart-seq, in contrast to other scRNA-seq methods, generates full length cDNA and can more accurately differentiate between splice variants. A side-by-side comparison of these scRNA-seq strategies found that Drop-seq was the most cost-effective method, whereas Smart-seq was the most accurate [251]. Analyzed cells may be clustered based on expression levels of selected genes either to detect changes in cell populations or within a population induced by a disease.

**1.3.3c Strengths & Caveats for Transcriptome Analysis**

Next generation sequencers are powerful tools, but they are not without flaws and errors that can arise at any step of the sequencing process. Firstly, errors may be introduced by polymerase during the amplification of sample cDNA, and research indicates this may be the primary source of errors in second generation sequencing data [252]. Secondly, errors originate from the chemistry used by the various platforms, and often manifest in nucleotide substitutions, insertions, or deletions [247]. The error rates of second-generation sequencers are principally increased in homopolymeric regions caused by incorporation of multiple bases in a single cycle. AT-enriched regions and genomes cause increased error rates in next generation sequencers possibly from PCR artifacts and nonrandom fragmentation of sample DNA [253]. Errors due to AT-richness are most pronounced in the Ion Torrent platforms [254]. Furthermore, when utilizing single-cell sequencing strategies, comparison between samples can be greatly impaired by poor matching of samples, the stages of disease progression, and the variability between individuals can compound the inherent heterogeneity that is present when
comparing individual cells. While the ability to determine the response and contribution of individual cell types to disease progression is important, more samples are necessary to identify and distinguish between inter-individual and intra-individual variation.

For next-generation RNAseq analysis, the most important parameters to consider in experimental design to substantially increase the quality of downstream analysis are: the number of biological replicates, the depth of sequencing (i.e., number of reads produced for each sample), read length, single-end vs. pair-end sequencing (i.e., each sequenced DNA molecule is represented by a single strand read vs. two reads from each strand), and RNA extraction. Under budgetary constraints, tradeoffs between sequencing depth and the number of biological replicates are often made. As consistently reported, the requisite number of biological replicates (n=3-4) is more critical for robust, reliable, and replicable analysis than sequencing depth [255-258]. As technologies improve, sequence lengths increase. For differential expression, little difference is seen if the length is >25 bps, in either single-end or pair-end sequencing. However, for greater accuracy in transcript identification and splice junction detection, reads should be pair-end and ≥100 bp [259]. The RNA extraction method impacts the ratio of RNAs present during sequencing, and a specific strategy should be chosen with the biological or biomedical question of interest in mind. For example, total RNA extraction is useful in capturing unique transcriptome features, such as noncoding RNA. However, ribosomal RNA (rRNA) comprises >90% of total RNA and should be depleted if noncoding, non-ribosomal RNA is to be assessed. Current techniques cannot completely remove rRNA, and ~2%-35% residual remains in the sample. Therefore, greater sequencing depth should be considered when using ribosomal depletion methods to counter the abundance
of rRNA and improve detection of other transcripts. In eukaryotic organisms, if only protein coding genes are of interest, poly(A) selection yields greater accuracy of transcript quantification [260]. These issues are particularly critical for clinical samples from patients, which are routinely processed as formalin-fixed, paraffin-embedded (FFPE) samples, which adversely impact the quality of RNA. Fortuitously, side-by-side comparison of FFPE and flash-frozen samples shows a great degree of concordance (e.g. $r^2$ in the range of 0.90-0.97 in recent studies [261, 262]), proving RNAseq is a viable tool for gene quantification in clinical settings. Controls, depending upon availability, need to be non-diseased tissue, either of the same patient origin or from another individual without the disease [263]. In addition, given breast cancer is a common disease, patients are from genetically diverse, heterogeneous populations with variable symptomology, which requires more samples to detect meaningful changes in the transcriptome truly reflecting disease process. However, in other studies, such as breast cancer, as few as n=9-10 patient samples (plus samples of healthy controls), have been ample to detect specific alleles and molecular pathways [263].

Despite the errors that may occur when using second generation sequencers, several advantages over previous transcriptome technologies warrant their use experimentally and clinically. First of all, second generation sequencers offer orders of magnitude deeper coverage of sample RNA than achieved by Sanger sequencing, via EST libraries, yielding overall faster discovery and more accurate analysis of an entire transcriptome. Also, the length and quality of sequence produced by second generation sequencers are much better than the fragments produced in SAGE, which improves transcriptome accuracy. While EST typically produces fragments of at least 500bp, most
second-generation sequencing produce shorter read lengths, albeit, read length from second generation sequencers can be increased at the expense of read depth. Next generation sequencers have advantages over microarrays because essentially all expressed transcripts and their variants, can be detected, without restriction to the probes present on the microarray chip or beads [264], plus the ability to barcode different samples, or conditions, within a single sequencing procedure permits multiplexing of samples.

### 1.3.4 Third Generation Sequencing

The latest generation of sequencers is distinguished from first and second generations by eliminating sample amplification. Bypassing sample amplification reduces sample preparation time and eliminates signal mismatch and distortion errors introduced during amplification. In addition, these single-molecule sequencers produce extremely long reads, surpassing the lengths achieved by Sanger sequencing. The Pacific Biosciences Single Molecule Real Time (SMRT) sequencer utilizes pyrosequencing (Fig 5) in polymerase-embedded plates, which lower the signal-to-noise ratio to detect real-time signal processing of fluorophore cleavage. The use of pyrophosphate-labeled nucleotides in polymerase-containing plates to extend DNA at near its natural speed facilitates processivity and sequencing length. Another third-generation sequencing platform available now is nanopore sequencing (MinION, Oxford Nanopore Technologies). This technology utilizes electrophoresis of DNA molecules via nanopores (5-8 nm diameter); as the DNA molecules squeeze through the pore, each nucleotide (A, T, G and C) produces
a unique electromagnetic signature detected. Similar to SMRT, nanopore sequencing can produce very long reads, up to 880 kb in a recent report [265].

1.3.4a Strengths & Caveats for Transcriptome Analysis

The Nanopore and SMRT sequencer both have ~10-15% error rate, distributed evenly over the length of the read [265]. Fortunately, the lack of location bias in SMRT and Nanopore reads provide sufficient coverage to extrapolate highly accurate consensus sequences. Third generation sequencers are not yet ubiquitous, but they promise several advantages over previous generation sequencers. The lack of sample amplification allows for quicker, cheaper analysis and avoids the polymerase errors caused by amplification. The long reads generated by third generation sequencers allow for more accurate assembly of large contiguous sequences, such as whole chromosomes, complete sequencing of whole genes in a single read [266], and identification of novel transcript isoforms. These platforms are excellent for whole-genome and whole-transcriptome assemblies [267, 268], including complex genomes such as gorilla [269] and human [265]. However, at this time, third generation sequencers are at a disadvantage for use in transcriptome analysis for quantification of expression due to the relatively low number (e.g. ~50,000 for RSII sequencer) of output reads generated with each run comparing to, e.g., Illumina sequencers (current typical low-end is 20,000,000+ reads per sample). The long reads greatly improve de novo assembly and transcriptome analysis for gene isoform identification, and the emerging technology in the field of metagenomics, which may be important for investigating the role of microbiome imbalance in breast cancer patients. Longer reads are also useful when assembling genomes that include large stretches of repetitive regions. These technologies are recommended for whole genome assembly and
splice variant detection albeit given the error rate currently not recommended for transcript quantification.

### 1.3.5 Results of Transcriptome Analysis: Unbiased Data Mining

#### 1.3.5a Differential expression analysis

In most cases, comparison of one or more conditions will result in a ranked list of transcripts with either relative or absolute levels of expression. The typical approaches include (1) raw data collection (processing of image files to collect intensities for individual probes on microarrays, counts of number of reads per transcript for RNAseq data, etc.); (2) data normalization, often followed by transformation [270]; (3) statistical analysis to identify transcripts whose expression differences between conditions are significant, and most importantly, (4) downstream analysis.

Microarrays of any platform are substantially more rapid to process using the manufacturers’ software suites, such as Affymetrix’s Expression Console and Transcriptome Analysis Console, or Illumina’s GenomeStudio. Alternative open-source, peer-reviewed, and publicly available software for microarray analyses using the R programming language, such as affy [271], lumi [272] and limma [273], are available as installation packages from the Bioconductor portal [233].

For next-generation RNAseq analysis, the most important parameters to consider in experimental design that substantially increases the quality of downstream analysis are depth of sequencing (i.e., number of reads produced for each sample, also referred to as “coverage”), read length, and single-end vs. paired-end sequencing. These parameters
vary based on the goal of the biological or clinical experiment. For example, comparison of expression between samples requires far less read depth than identification of novel transcripts or splice variants. Journals that publish RNAseq studies sometimes also have their own requirements for read depth. Furthermore, the length of sequencing reads varies depending on experimental design, with longer reads typically being used in novel transcript identification or *de novo* assembly generation [268]. Sequence read lengths as low as 75 bases are sufficient for differential expression analysis [256]. Finally, paired-end sequencing from both ends of a single mRNA fragment facilitates identifying splice variants and alignment [259].

Once the sequence is obtained from the raw signals, the quality of the output must be assessed, based on sequence read lengths and processing direction (single-end vs. paired-end sequencing) with either FastQC [274] or NGSQC [275]. These tools will provide GC content, overrepresented reads, PCR artifacts, and sequence quality to detect potential PCR bias or DNA contamination. It is usual for sequence quality to weaken at 3' end and software programs, such as Trimmomatic [276] or FastQ trimmer [277], can remove these low-quality 3' ends. Alignment is a critical step in RNA sequencing analysis because raw sequence reads must be mapped precisely to an annotated reference genome or transcriptome for the species. While it is possible to analyze RNAseq data without a reference, e.g. by using Trinity software [278], most clinical and translational models of breast cancer have assembled genomes available. The most common software platforms to align RNA sequence to a reference genome are TopHat [279], HiSAT [280], and STAR [281]. These platforms differ with respect to speed, memory usage, and their algorithms for handling base and splice junction alignment precision, with HiSAT and STAR
optimized to process large datasets (>10^8 reads), whereas TopHat is designed for smaller datasets (<2×10^7 reads).

Measurement of transcript expression in RNAseq data is based on quantifying raw counts at each genetic locus along the chromosomes using an assembled genome with programs such as HTSeq-count [282] or featureCounts [283]. This approach uses a GFF (Generic Feature Format) or GTF (General Transfer Format) file that contains gene coordinates, identifiers, and descriptions in a strict predefined format [284]. All the reads that map within the genomic coordinates of a given feature (e.g., gene, exon) contribute to the count number of this feature. The counts from the RNAseq data are corrected for sequencing depth and often for length of gene transcripts because smaller datasets will have fewer count numbers, with the consequence that longer transcripts will have higher representation among raw RNAseq reads. The majority of normalization methods report the amount of transcript expression as reads per kilobase of exon per million reads (RPKM), fragments per kilobase of exon per million of reads (FPKM), transcripts per million (TPM), or counts per million (CPM) [285-287].

1.3.5b Enrichment Analysis: Overview of Biological Ontologies

Description of gene functions in scientific literature can vary significantly between authors, even if both are describing the same phenomenon. Consequently, unbiased grouping of genes by functional similarities may become a daunting endeavor. To facilitate the task of describing the universe of genes, the methods of formal ontology were applied to create the first controlled vocabulary to standardize gene descriptions across species and disciplines. The resulting Gene Ontology (GO), and GO Consortium were formed in 1998 to create a framework for standardizing gene products description [288].
Since inception, GO was used to annotate millions of genes, with over 1,350,000 annotations for *H. sapiens*, *R. norvegicus*, and *M. musculus* genes alone. The highest-level annotations for genes in GO are a trinity of molecular function, cellular component, and biological process. Currently, GO uses 29,623 “Biological Process”, 11,139 “Molecular Function”, and 4,189 “Cellular Component” terms, and strict rules to describe evidence linking a gene to a term (from relatively vague “Inferred from Sequence or structural Similarity” to strong “Inferred from Experiment”), to annotate genes across the tree of life; taking into account the total number of annotated genes in species, currently average number of annotations ranges from 5 for *E. coli* to 21 for *R. norvegicus*. GO is organized as a graph, with individual terms being nodes, and relationships between terms being edges. Currently, there are 8 types of relationships between terms, and the “is_a” relationship gives this ontology a loose hierarchy, with more general terms being “parent” to more specific “child” terms [288]. Curation remains an ongoing process, including the field of breast cancer, and new annotations, and new GO terms are added frequently as scientific and specific knowledge expands. The dynamic nature of GO catalyzes new discoveries to be readily integrated into the existing ontology while older annotations are updated with new information as it becomes available. Following the success of GO, other ontologies began to emerge to formalize biological and biomedical knowledge to assist in large-scale data analysis and discovery of new treatment avenues. Relevant examples include Mammalian Phenotype Ontology [289] and Human Phenotype Ontology [290], both used to formalize descriptions of normal and breast cancer phenotypes. Another example, Protein Ontology, describes evolutionary relation, isoforms, and complexes of proteins [291, 292]. These and other ontologies collectively form an Open Biological and
Biomedical Ontology (OBO) Foundry and share common goals to facilitate curation, management, distribution, and analysis of data [293].

Analyzing the data produced from transcriptome analysis facilitates researchers to explore gene functions, expression levels, differential gene expression, organismal responses to environmental and developmental changes, and more. Understanding these characteristics can allow for the identification of highly specific drugs or disease biomarkers. When analyzing transcriptomes of samples, the key focus is the difference of expression levels of various groups of genes.

1.3.5c Using Ontologies & Pathway Analysis for Precision Medicine in Breast Cancer

Precision medicine classifies individuals according to their underlying susceptibility, prognosis, or targeting potential treatment response. Unlike DNA sequencing focusing on genome, RNA sequencing produces the snapshot of the full transcriptome, and has the capability to fulfill precision medicine to classify patients at both molecular and cellular levels when used in conjunction with programs for ontologies and pathway analysis. Development of RNA sequencing pipelines is important for implementation of transcriptomics as precision medicine [294], which can be used successfully to classify patient or model attributes and predict therapeutic response and ultimate outcomes. Classifying patients based on symptoms is limited because symptoms often arise from numerous origins or multimodal pathways, as the case with breast cancer. Biomedical researchers in both clinical and basic research settings need to choose transcriptome analysis to the specific characteristics of disease, and its pathology, to detect changes in the target molecular, cellular and physiological pathways under
scientific scrutiny. Transcriptomics is a robust method to measure both common and unique pathways simultaneously. For unbiased detection in molecular and cellular pathways, researchers need to use a variety of tools, from read alignment to ontological analysis.

The first step of ontological analysis of genes is its annotation for its biological process, cellular component, and molecular function. Once all the gene annotations have been collected, they are grouped by category, and these categories are analyzed for enrichment or depletion against a “universe set” of all the genes of an organism. The number of annotations to a distinct ontological term in a list of genes, for example, a list of downregulated genes in invasive ductal breast cancer (IDC) vs. normal breast epithelium is compared to the number of annotations to this term among genes in the universe set (i.e., all genes in the genome) to identify if the occurrence of this term in the experimental results is higher or lower than expected from a random sampling of the universe set. This analysis facilitates discovery of common biological themes, based on ontologies, within the lists of genes. Multiple tools exist for determining pathway enrichment; among preferred tools in our laboratory is the VisuaL Annotation Display (VLAD [295]), which allows to define the “universe set”, simultaneously use more than one query set, as well as operate with any ontology within OBO Foundry, rather than only GO.

In particular, the ability to upload own “universe set” of genes allows for more precise identification of over- and underrepresented ontologies, while the ability to upload any ontology from OBO Foundry allows exploration of additional ontologies such as Mammalian Phenotype [289]. Importantly, in the online version of VLAD, GO
annotations, as well as nomenclature of mouse and human genes are updated weekly automatically [295] although local installment of VLAD requires the individual laboratory to manually update gene annotations from GO. Similar tools, such as AmiGO [296], BiNGO [297], DAVID [298], GOrilla [299], are also very popular free public resources to identify GO term overrepresentations in the lists of genes, however, many of these excellent tools lag behind in updating their gene annotations by as much as 3-4 years. Similar idea of measuring and testing overrepresentation within a group of genes of interest is implemented in Gene Set Enrichment Analysis (GSEA) [300], and commercial platforms such as Ingenuity Pathway Analysis (IPA) [301] and Pathway studio [302].

Another useful tool to identify specific pathways in the large-scale gene expression data is MetaCyc [303], which contains a collection of curated biochemical pathways, annotated with organism-specific data on genes, pathways, proteins and compounds. MetaCyc tool, Cellular Overview, allows the user to upload gene expression data and visualize the expression upon the entire metabolic map while simultaneously retaining the ability to focus on individual pathways affected by disease or condition, such as atypia, DCIS, or IDC samples from breast biopsies [85]. For mammals, curated databases include human [304], mouse [305] and cattle [306]. Differentially expressed gene lists can also be overlaid onto existing cellular pathways using portals such as Reactome [307] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) [308] to explore potential secondary pathways, and dysregulated pathways specific to breast cancer pathology or healthy samples. Importantly, research community involvement in the process of gene annotation and curation, including creation of disease-specific ontology terms, improves the precision and quality of these resources to breast tissue and breast cancer research [309].
1.3.6 Summary of Transcriptomics Approaches

Transcriptome analysis is a dynamic tool, whose efficacy and efficiency are continually improving. The variety of platforms available to perform such analyses is a great advantage to laboratories both large and small, and the high-throughputs for some of these technologies provide rapid results with great accuracy. Identification of affected pathways using transcriptomics bioinformatics tools allows researchers and clinicians to make a focused and informed decision on the genes to concentrate on as potential therapeutic targets in precision medicine. Application of transcriptomics can facilitate the exploration of underlying pathogenic mechanisms, identification of genetic variants, determination of treatment effects, including screening for molecular biomarkers. Importantly, expression signatures in diseased phenotypes may pinpoint precise interventions required to alleviate the disease state, a goal of precision medicine, without a need for the cost prohibitive personalized assembly and deep analysis of patient’s genome. Thus, transcriptomics can classify individuals while simultaneously facilitate discovery, testing, and validation of new therapeutics for patients with breast cancer, defined at the cellular and molecular levels.

1.4 Objectives, Hypotheses and Aims of Studies

Great efforts have been made to determine the etiology of human invasive breast cancer. Currently, many studies have identified several risk factors and mutated genes associated with breast cancer. However, the components contributing to malignancy progression from DCIS to IDC are unknown and there is unmet need for prognostic markers to identify patients that will develop malignancy. Interestingly, a recent study, using spatially resolved single-cell genome sequencing, posits that in DCIS to IDC breast
cancer, genome evolution occurs early in tumor progression and prior to invasion \[310\]. This study is counter to the bottleneck theory that suggests a single clone is selected for basal membrane escape or invasion into surrounding tissue \[311\]. This early heterogeneity suggests there could be additional factors involved in basement membrane escape. It is well known that cancer genomes display a decrease in epigenetic regulatory marks, except for in CpG rich promoter regions of tumor suppressor genes where hypermethylation occurs. Recently, such dysregulation was described in breast cancer progression, as well \[312\]. It is well known that LTR TEs, when deregulated through demethylation, drive cell fate and potency as each pluripotent state is linked with the activation of distinct classes of TEs \[313\]. We hypothesize that a change in the LTR expression signature plays a role in the cell fate shift of breast tissue transformation to metastasis. To test this hypothesis, we aimed to develop a bioinformatics pipeline to accurately analyze gene and TE expression from RNA-Seq data derived from formalin-fixed, paraffin-embedded (FFPE) clinical samples. Next aim is to, identify the genome-wide retrotransposon expression signature during breast cancer progression. Lastly, we aimed to demonstrate that TE expression contributes to the transition from pre-malignant to malignant breast cancer by altering their repression and mobile activity through direct and indirect inhibition of DNA methylation and reverse transcriptase and integrase inhibition.

1.5 Significance and Impact on Health

Current clinical imaging and detection technologies have increased the number of diagnoses of very early breast cancer, including atypia (early neoplasia) and ductal carcinoma in situ (DCIS). As previously mentioned only one-third of these early
diagnosed breast cancer cases will progress into an invasive form. Our current understanding lacks the capability to identify early breast cancer diagnoses that will most likely transform into invasive carcinoma. Current treatments for individuals diagnosed at pre-malignant stages of breast cancer involve breast conserving surgery with radiation therapy or a mastectomy. While there are clear benefits to receiving radiation therapy with breast-conserving surgery, there are also drawbacks [314]. For example, radiation therapy is administered to the whole breast, requires daily treatments, and is associated with an increased risk of secondary cancers [315, 316]. Mastectomy is the second most likely treatment for DCIS, especially with patients who have a mass of 4-5cm or more [317]. Patients who undergo mastectomy have a very low probability of a recurrence. However, patients are still at an increased risk of an untreated breast developing DCIS or IDC [318]. Some patients who elect for mastectomy have lifelong side effects that adversely affects their quality of life, such as phantom pain, fatigue, lymphedema and hormonal dysregulation issues. Understanding the mechanisms of breast cancer transformation is imperative to treatment and care of the ~80,000 patients diagnosed at early stages of breast cancer before IDC each year [319].

1.6 References


104. Molyneux, G., et al., BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. 2010. 7(3): p. 403-417.


Chapter Two:
Bioinformatics optimization of clinical samples

2.1 Introduction

After next-generation sequencing was introduced in 2005, many high-throughput “next-generation” sequencing (NGS) approaches followed [1]. One of the approaches introduced in 2008, RNA sequencing (RNA-seq), captures transcriptomes from collections of cells or tissue samples. Reads generated by RNA-seq can then be used to assess single nucleotide polymorphisms (SNPs), detect splice variants, detect fusion genes, and measure individual transcript abundance in the samples for differential expression analysis. Thusly, this exciting technology has since been used for diagnoses, prognoses, and therapeutic selections [2-4]. In comparison to previous popular technologies, microarrays, RNA-seq provides better transcriptome coverage and is more suitable for discovery science, as the identification of expressed genes is not limited to the probes present on the array [5, 6]; on the flip side, gene expression analysis becomes computationally more challenging due to the requirement to identify each and every read output in RNA-seq dataset. However, as NGS technologies and computers able to run the analyses have become cheaper, RNA-seq has expanded rapidly producing massive amounts of data to be analyzed.
2.1.1 Research Purpose and Approach

Many everyday experimentalists and clinicians, who utilize RNA-seq in their experiments, rely on outsourcing to cores or companies which generate and analyze their RNA-seq data. This practice is common with many niche experiments, particularly omics based. The purpose of the research is to bridge the gap between those in the RNA-seq niche and the everyday experimentalist or clinician by showcasing the importance of understanding biological conditions and the bioinformatics analysis applied.

For an ideal RNA-seq experiment, researchers and clinicians would require freshly frozen tissue samples with minimal contamination from other tissues (e.g. blood, fat). However, most clinical research relies on archived tissue samples (formalin fixed paraffin embedded, FFPE), or are hindered by limited sample material from biopsy collection. It is worth noting that FFPE samples have increased nucleotide degradation and decreased poly(A) binding affinity [7-9]. We analyze RNA-seq data from formalin fixed paraffin embedded tissue, which is highly variable for quality and depth of reads. We believe the quality and depth of these data to represent the quality and depth of data the everyday researcher or clinician will likely work with. One of the most common biological question asked when analyzing RNA-seq data is, “What genes have different levels of transcription in one experimental condition when compared to another experimental condition?” To determine differential expression, RNA-seq reads need to be assessed for quality and then aligned to a reference genome.

After quality control checks, alignment is most often the first step in RNA-seq analysis and any analysis thereafter relies heavily upon this initial step. Generally, when available and well annotated, reads obtained from sequencing will be mapped to either a reference genome or a transcriptome of a species. For simplicity, we will be describing
tools and settings for mapping to a reference genome. For in-depth discussion, see a recent review by Conesa et al [10]. The biggest challenge for aligning RNA-seq reads to a reference genome is that most reads have exon-exon splice junctions. The most common software platforms available for mapping to a reference genome, TopHat [11], HISAT2 [12], and STAR [13], identify splice junctions. Where these platforms differ is computationally, with respect to speed and memory usage, and their algorithms for handling base and splice junction alignment precision. TopHat is currently not maintained and has been superseded by HISAT2 due to computational inefficiencies, both of which are built on the short read mapping program Bowtie2 [14]. While all three aligners are considered fast, HISAT2 and STAR consistently outperform TopHat with respect to computational speed [13, 15, 16]. Baruzzo et al [16] and Engstrom et al [17], provide insight into the major differences in aligner performance. All three aligners performed well in placing a read onto the respective genomic locus. However, significant discrepancies and deficiencies were found for TopHat in getting a read to align or total percent mapped. Here we evaluate the alignment performance of STAR and HISAT2 due to the discrepancies in TopHat performance in previous comparison papers.

The relative expression level of genes is estimated based on the number of mapped reads. These counts are subjected to statistical tools to assess significant differences between groups. Here we explain and assess the strengths and weaknesses of the two most widely used differential expression analysis tools, DESeq2 [18] and edgeR [19]. A simplified bioinformatics pipeline for raw data to differential expression analysis can be found in Figure 2.1. We performed differential gene expression analysis using the series of breast cancer progression RNA-seq data from FFPE samples. In our comparison we assessed similarities and differences in the results, to test if (and how) different aligners
affect the gene expression level counts or location of mapping, and if different algorithms used for assessing differential gene expression affect the “final list” of statistically significant transcripts, as well as downstream analysis for Gene Ontology [20] enrichment. We focused on two of the most popular aligners, HISAT2 and STAR, and two differential gene expression testing tools, DESeq2 and edgeR, all available for users via the Galaxy platform, a portal designed to fulfill the bioinformatics needs of researchers with very modest bioinformatics and programming experience.

### 2.2 Methods

#### 2.2.1 Breast cancer samples

A dataset of 72 RNA sequencing experiments is deposited in NCBI SRA database (project number PRJNA205694 [21]). Datasets represent transcriptomes of biopsies from different stages of breast cancer: 24 normal tissues, 25 early neoplasia (EN), 9 ductal carcinoma in situ (DCIS) and 14 infiltrating ductal carcinoma (IDC), from 25 patients. Briefly, RNA was extracted from core punches of FFPE specimens after histological confirmation of the cancer stage; only samples that possessed >90% of luminal cells with

![Figure 2.1. Bioinformatics pipeline for RNA-Seq tool analysis](Unpublished)
the appropriate diagnosis were used. Directional cDNA libraries were constructed and sequenced using Illumina GAIIx to obtain 36-base single-end reads. We used these data to compare two common aligners, STAR and HISAT2, and differential expression tools, DESeq2 and edgeR (Figure 2.1).

2.2.2 RNAseq reads alignment

We used two different programs, STAR and HISAT2, to align the RNAseq reads to the reference human genome assembly (hg19). For more precise alignment, both programs use a dataset of known splice sites for correct identification of potentially spliced reads in RNAseq data; this dataset, in “gene transfer format” (gtf), was obtained from ENSEMBL (release 87, 12/8/2016).

2.2.2a STAR

STAR’s algorithm [13] uses a two-step approach. STAR aligns the first portion, or seed, of a read to a reference genome up to the maximum mappable length (MML) of the read then aligns the left-over portion, second seed, of the read, up to its MML. After the read is completely aligned, STAR stitches the two, or more, seeds together and scores them based off a user-defined penalty for mismatches, insertions, and deletions. The “stitched” seeds with the highest score are chosen as the correct alignment of a read. This approach allows for quick and easy annotation of multi-mapping reads with their own alignment scores. --seedSearchStartLmax 50 -seedSearchStartLmaxOverLread 1.0 -seedSearchLmax 0 -seedMultimapNmax 10000 -seedPerReadMax 1000 -seedPerWindowMax 50 -seedNonoLociPerWindow 10 -alignIntronMin 21 -alignIntronMax 0 -alignMatesGapMax 0 -alignSJoverhangMin 5 -
alignSJDBoverhangmin 3 –alignSpliceMateMapLmin 0 –
alignSplicedMateMapLminOverLmate 0.66 –alignWindowsPerReadNmax 10000 –
alignTranscriptsPerWindowNmax 100 –alignTranscriptsPerReadNmax 10000 –
alignEndsType Local

2.2.2b HiSAT2

HiSAT2 uses the Bowtie2 [14] algorithm to construct and search a Ferragina-Manzini (FM) index [22]. However, HISAT2 employs two types of indexes for aligning: a whole-genome FM index to anchor alignments, and numerous overlapping local FM indexes for extension of the alignment. --mp MX=6, MN=2 –sp MX=2, MN=1 –np 1 –rdg 5.3 –rgf 5.3 –score-min L,0,-0.2 –pen-cansplice 0 –pen-noncansplice 12 –pen-
canintronlen G,-8,1 –pen-noncanintronlen G,-8,1 –min-intronlen 20 –max-intronlen 500000

2.2.3 Gene expression counts

The simplest method for estimating transcript expression is to count the raw reads for each annotated genomic locus in the genome assembly. This approach uses a gene transfer format (GTF) file that contains coordinates (i.e., positions in the genome for each exon, transcription start site, transcription termination site, etc.) and nomenclatures of genes. We used FeatureCounts [23], a program that extracts information from bam files for reads overlapping with features in an input gtf file containing exon coordinates for all transcripts in the genome assembly, with the following parameters: --t ‘exon’ –g ‘gene_id’ –M –fraction –Q 12 –minOverlap 30.
2.2.4 Data normalization and quality control

2.2.4a Data normalization

There is much confusion in the literature when reporting units of expression for RNA-seq data. The confusion stems from the different forms of normalization required for within sample comparison vs between samples comparison. Many methods for within sample comparison attempt to correct for sequencing depth and gene length. These methods produced the most frequently reported unit of expressions for RNA-seq data, which are read per kilobase of exon per million reads (RPKM), fragments per kilobase of exon per million of reads (FPKM), and transcripts per million (TPM)[24]. The order in which RPKM and FPKM normalize the read counts causes differences within samples that should not be ignored. Instead, when comparing within samples one should use TPM values which eliminates the invariance within samples[24]. A relationship between RPKM/FPKM and TPM is derived here[25]. Alternative normalization methods, which allow for comparisons between samples or differential expression analysis, are available and explained in the next section. We used CPM normalization of gene count data for quality control step using the following formula:

\[
CPM_i = \frac{R_i}{T_{Ra}} \times 1,000,000
\]

where CPM\(_i\) is a CPM value of a gene in a biological replicate; \(R_i\) is the number of reads mapping to all exons of this gene in this biological replicate; \(T_{Ra}\) is the total number of reads aligned (anywhere in the genome) from this biological replicate (i.e., the number
of aligned reads in either STAR or HISAT2 output “binary alignment map” bam files). This procedure also transforms data from counts to a continuous scale.

2.2.4b Quality control

For quality control, we used ClustVis [26], a statistical tool for clustering of complex data such as RNAseq, based on principal component analysis and visualization of results. Any samples that fell outside of the initial 95% confidence interval on the two-dimensional PCA plot were flagged as outliers and removed before further analysis. ClustVis is an intuitive user interface built on several R packages that provides Principal Component Analysis (PCA) and heatmap plots of high-dimensional data from a data matrix. Data may be uploaded by a delimited file or copy and pasted into the ClustVis text box. Dimensions (e.g. genes) and observations (e.g. samples) may have multiple annotations that can be detected automatically or input manually to provide additional depth to the PCA and heatmap plots [26].

2.2.5 Differential gene expression analysis

According to citation reports, edgeR and DESeq2 are the leading tools for differential gene expression analysis of RNAseq data (9,411 and 7,318 citations in Google Scholar, respectively; retrieved 12/2018). Both tools are R packages and require raw read counts in a data matrix which they normalize to account for differences in sequencing depth and low count variability. Both tools assume the data will have variance beyond what is expected in random sampling, what’s known as overdispersion. Therefore, they assume the data will fit a negative binomial distribution and further attempt to shrink the
raw counts into this distribution through unique Empirical Bayes methods. To display
differential expression outputs uniformly we used the R package ViDGER [27].

2.2.5a DESeq2

DESeq2 normalizes each gene using a generalized linear model [28]. Then,
DESeq2 uses an Empirical Bayes shrinkage to detect and correct for dispersion. Unlike
early versions of edgeR, DESeq2 also uses an Empirical Bayes shrinkage on log₂-fold
change estimates. Within R, the “raw count data” was imported as a matrix along with a
“DataFrame” that has a column which annotates the columns of the imported matrix. A
DESeqDataSet was generated with these imported files using the DESeq2 function
DESeqDataSetFromMatrix(countData="raw count data", colData="DataFrame",
design=~0 + Stage). The names in the quotation marks are arbitrary and assigned by the
user when importing the corresponding files all quotation marks should be removed when
using the functions in R. The design argument within the function is a formula which
expresses how the counts for each gene depend on the variables in colData. Here, the
design formula of ~0 + Stage will build a results table and plot all experimental group
comparisons. The data was filtered to keep only genes that had at least 4 counts between
at least 10 samples generating the function keep <-
rowSums(counts("name_DESEQDataSetFromMatrix")>4) >=10 and applying it to
“name DESEQDataSetFromMatrix” with name DESEQDataSetFromMatrix <-
name DESEQDataSetFromMatrix[keep,]. The DE analysis was called using
DESeq("name DESEQDataSetFromMatrix").
2.2.5b EdgeR

EdgeR’s default method of normalization is called trimmed mean of M-values, or TMM, obtained with the function calcNormFactors. This method of normalization estimates the ratio of RNA production through a weighted trimmed mean of the log expression ratios. There are alternative normalization methods available in edgeR to account for data that does not fit the negative binomial distribution assumed with TMM. To control for false discovery rate (FDR) we applied the estimateDisp function.

2.2.6 Gene enrichment analysis using Visual Annotation Display (VLAD)

VLAD [29], accessible via MGI web portal, is a powerful tool to find common functional themes in the lists of genes by analyzing statistical over- or underrepresentation of ontological annotations. Currently, users can choose among Gene Ontology (GO) [20] annotations for human genes, Gene Ontology and Mammalian Phenotype Ontology (MP) [30] annotations for mouse genes, or upload a file of own annotations (in open biomedical ontology [31] ‘obo’ format). Unlike other packages for ontological enrichment, VLAD allows analysis of more than one query (i.e., several lists of genes may be analyzed and visualized simultaneously), as well as permits user to provide own “universe set”, i.e. gene list to test queries. For GO analysis, we searched for overrepresentation among terms with experimental evidence (i.e., codes EXP, “Inferred from experiment”; IDA, “Inferred from direct assay”; IMP, “Inferred from mutant phenotype”; TAS, “Traceable author statement”).
2.3 Results

2.3.1 Output of aligners

All reads for all samples were aligned to the human genome assembly (hg19). Overall, STAR significantly outperformed HISAT2 in aligning the FASTQ reads to the genome (Figure 2.2). The generally low proportion of aligned reads to all input reads for both programs is likely due to the quality of the libraries, as a significant number of input reads were poly(A) sequences, Illumina adapter sequences, and reads corresponding to the very 3’-ends of mRNAs, which are too uninformative for correct mapping.

2.3.2 Quality control

To eliminate sample outlier biases, we performed Principal Component Analysis of gene expression counts for each sample by each stage for both aligners. Gene expression counts were collected using featureCounts and normalized to the total number of aligned reads for each sample, and PCA completed on these data using ClustVis (large edition). For all subsequent analysis,

Figure 2.2. Performance of HISAT2 and STAR aligners on the breast cancer series data. (Unpublished)

Figure 2.3. PCA visualization of gene expression data from HISAT2 and STAR alignments. A, B: Clustering of HISAT2 samples on the first two principal components before (A) and after (B) outlier removal. C, D: Clustering of STAR data before (C) and after (D) outlier removal. (Unpublished)
any samples that fell outside the 95% confidence ellipse in their respective stages (Normal, Atypia, DCIS and IDC) were removed. For both HISAT2 and STAR, the same samples fell outside of the 95% confidence ellipse in each stage (Figure 3). In total, we identified six outlier samples in the RNAseq dataset, which were: SRX286949 (normal tissue), SRX286945 and SRX286964 (atypia), SRX286961 (DCIS) and SRX286951 (IDC). Interestingly, the PCA plots for all stages in HISAT2 data had the Atypia stage cluster well removed from other stages (Figure 2.3A & B). Overall, Atypia stage presented more heterogeneity than any other stage, irrespective of the aligner.

2.3.3 Gene expression profiling

2.3.3a Highly expressed genes

To determine how concordant the alignment tools were in mapping the reads to the genome, we compared the highest expressed genes that correspond to 50% of all reads mapped to exons. In the normal samples, 50% of the mapped reads came from 330 and 305 genes for STAR and HISAT2 respectively and they shared 263 of those genes (Figure 2.4). In atypia samples, 50% of the mapped reads came from 417 and 406 genes for STAR and HISAT2 respectively and they shared 40 of those genes. In DCIS samples, 50% of the exon-mapped reads came from 469 and 416 genes for STAR and HISAT2, respectively; of

![Figure 2.4. Overlap between the highest expressed genes in the RNAseq breast cancer datasets aligned by HISAT2 or STAR. HISAT2-identified genes are in red; STAR genes are in green; overlapping genes are in yellow. (Unpublished)](image-url)
those, 383 genes were shared. In IDC samples, 50% of the exon-mapped reads came from 384 and 366 genes for STAR and HISAT2, respectively, and the lists shared 319 of those genes. The high amount of discrepancies in atypia convinced us to look further into what were the major differences in alignment.

2.3.3b Alignment to pseudogenes.

Retrogenes are intronless gene copies produced by reverse transcription of a “parent” gene mRNA and insertion of the cDNA copy elsewhere in the genome. Retrogenes are often non-functional and are generally assigned to the category of “pseudogenes”, i.e. genomic loci harboring similarity to a protein-coding gene but not having any recognized function. The sequence similarity among retrogenes and their parent genes poses a problem for aligners, whose algorithms must decide when assigning a read to a specific locus in the genome. To determine what the differences in alignments were, we analyzed the numbers of reads mapped to pseudogenes by HISAT2, and STAR. Between two aligners tested, HISAT2 consistently had significantly higher amounts of reads aligned to pseudogenes when compared to STAR (Figure 2.5A). Furthermore, for Atypia stage, HISAT2 had drastically higher amounts of reads aligned to pseudogenes than the other stages. To
determine what portion of the top 50% of mapped reads were pseudogenes, we obtained a list of pseudogenes from the hg19 gtf annotation file we used and compared this list with the top 50% of mapped genes for each stage and each aligner. A single pseudogene was in the gene list for each stage which represented the top 50% of mapped reads for STAR. Conversely, HISAT2 consistently had higher amounts of pseudogenes represented in the top 50% of mapped reads (Figure 2.5B).

**2.3.4 Differential gene expression analysis**

The differential expression comparison on data from different alignment tools was done to further explore the consequences of previously described alignment tool biases, as well as to compare the two popular tools used for the purpose of identification and quantification of gene expression differences between different conditions, edgeR and DESeq2.
Figure 2.6. MA plots of pairwise comparisons of all stages. A. edgeR. HISAT2 (top) and STAR (bottom). B. DESeq2. HISAT2 (top) and STAR (bottom) gene counts for all samples were analyzed to identify differentially expressed genes. Each gene is represented by a single dot. Blue dots represent genes whose expression difference between conditions is both significant and at least two-fold. Green dots represent genes whose expression difference between conditions is significant, but less than two-fold; grey dots represent genes whose expression differences between conditions is not statistically significant. Y-axis (all plots): log2 of expression fold change; X-axis (all plots): log of gene expression mean value. (Unpublished)
2.3.4a edgeR of HISAT2 and STAR bams

Overall patterns of differential gene expression performed by edgeR on HISAT2 and STAR data were similar for all pairwise stage comparisons, except atypia vs any other stage (Figure 2.6). HISAT2 consistently had the atypia stage comparisons produce > 15,000 statistically significant differentially expressed transcripts with a log$_2$ fold change (LFC) $\geq$ 1 (i.e., difference in expression between condition is at least two-fold) (Figure 2.7). Conversely, differential expression pairwise comparisons with STAR atypia stage vs each other stage identified 350 to 2,496 differentially expressed genes (Figure 2.7).

2.3.4b DESeq2 of HISAT2 and STAR bams

Differential expression analysis using DESeq2 on pairwise comparisons of STAR alignments Normal vs Atypia revealed 255 transcripts having LFC > 1 and 177 transcripts with LFC < 1. Normal vs DCIS and Normal vs IDC analysis revealed 1,677 LFC > 1, 482 LFC < 1, and 2,304 LFC > 1, 1,417 LFC <1 DE transcripts, respectively (Figure 2.6B, bottom panels). Similarly, to edgeR differential expression analysis of HISAT2 with DESeq2 consistently produced high numbers of DE transcripts in atypia pairwise comparisons, > 19,000 transcripts with LFC > 1 (Figure 2.6B, top panels). Normal vs Atypia, Normal vs DCIS, and Normal vs IDC analysis revealed 19,419 LFC > 1, 1,212 LFC
< 1, 1,585 LFC > 1, 190 LFC < 1, and 2,196 LFC > 1, 732 LFC < 1 DE transcripts, respectively.

**2.3.4c Comparison of DESeq2 and edgeR results**

To compare how similar the most common differential tools, edgeR and DESeq2, are on the same aligners we produced Venn diagrams of all DE genes each produced, for each pairwise cancer stage comparison to normal samples. DESeq2 and edgeR shared 341, 1,678, and 2,809 of the differentially expressed transcripts on the STAR alignment pairwise comparisons for Normal vs Atypia, Normal vs DCIS, and Normal IDC, respectively (Figure 2.8). DESeq2 and edgeR shared 14,220, 1,433, and 2,137 of the differentially expressed genes on the HISAT2 alignment pairwise comparisons for Normal vs Atypia, Normal vs DCIS, and Normal vs IDC, respectively.

![Venn diagrams showing overlap of genes identified as differentially expressed](image)

**Figure 2.8. Overlap of genes identified as differentially expressed.** Top row is HISAT2-aligned data. Bottom row is STAR-aligned data. (Unpublished)
2.4 Discussion

STAR had the highest average rates of mapped reads for each stage. Conversely, HISAT2 not only had lower rates of alignment but also had increased rates of pseudogene alignment which may have compromised alignment fidelity. We recommend STAR alignment over this version of HISAT2. It’s important to note that no alignment tool is right for every job and that alignment parameters play a large role in alignment outcomes. For in depth explanations and comparisons of aligners we refer the reader to [16, 17].

Next-generation technology has improved greatly in a short amount of time. Initially, read lengths being generated were between 25 – 36 nucleotides. Currently, nucleotide reads can be generated at a length of >300 nucleotides. Chhanawala et al [32] reported reads with a length of >25 nucleotides had negligible differences with differential expression detection. Conversely to what was previously reported, we believe the short nucleotide sequences (36nts) may have been an issue for the FM index generation utilized by this version of HISAT2 with the annotated settings, thus propagating misalignments to pseudogenes. However, further testing needs to be done to verify. To improve concordance between aligners we recommend ensuring greater read lengths during experimental design to increase alignment accuracy.

Recently, Schurch et al [33] compared the performance and accuracy of the top 11 differential tools available for RNA-Seq analysis. DESeq2 and edgeR out performed other tools with the lowest FDR and highest true discovery rate (TDR). Further, they report DESeq2 slightly outperformed edgeR with respect to FDR when there were more than 12 biological replicates. Here, we report DESeq2 having a higher number of significantly differentially expressed genes. This may be counter to earlier reports of edgeR’s propensity to a higher FDR at higher number of biological replicates. The addition of the
estimateDisp function may have applied heavier weighted likelihood empirical Bayes methods to obtain the posterior dispersion estimates [34]. Therefore, when dealing with higher number of replicates we recommend using DESeq2 or edgeR with the estimateDisp function to control FDR. When dealing with lower than 12 replicates we believe edgeR and DESeq2 are both viable options however we recommend edgeR as it may have an edge in false negative rates (FNR), or TDR.

These data clearly demonstrate the need for careful considerations and purposeful intent when generating and utilizing a bioinformatics pipeline to assess differential expression of clinically applicable RNA seq runs. Due to the increase use of RNA seq to diagnose, prognose, and generate therapeutic options we feel that clinicians and everyday experimentalists should have a strong foundation and understanding of the bioinformatics tools being utilized to generate and analyze their data. This study highlights possible limitations of this version of HISAT2 for older RNA seq read generation technologies, poor quality sample reads, and short RNA seq reads. Thus, providing clinicians with insights into the “right” tool for the job.

2.5 References

3.1 Introduction

Etiology of breast cancer includes two major, independent, cellular decisions: establishment of a small population of atypical cells in healthy tissue and malignant transformation of ductal carcinoma in situ (DCIS) stage cells into invasive breast carcinoma [1]. The pathways and the regulators underlying these cellular decisions are currently poorly understood but clearly have genetic, environmental, and age-dependent components. Importantly, neither of these cellular decisions is definitive because not all healthy breast tissues will develop atypia, nor all DCIS will become deadly breast cancer [2-4]. Detection of breast cancer at early, pre-invasive stages greatly increases the odds for long-term survival. Early diagnosis of breast cancer often results in unnecessary treatment for the majority of women. In fact, a recent prospective study revealed only 5-30% of cases of atypia will develop aggressive, invasive forms of breast cancer, whereas the majority, 75%-95% will not [5]. The absence of diagnostic biomarkers predicting invasive ductal carcinoma (IDC) leads to costly, unnecessary treatment [6].

There is little understanding why some patients diagnosed with atypia and DCIS remain cancer-free for years, while in others the disease progresses rapidly to malignant invasive ductal carcinoma. A ‘one gene at a time’ approach does not sufficiently predict whether an early lesion will become invasive because it is based on genetic markers
associated with developed tumors. For instance, poor prognosis for human breast cancer strongly associates with expression and/or mutations in \textit{ER\alpha}, \textit{ER\beta}, \textit{HER2}, \textit{PR}, \textit{EGFR}, \textit{BRCA1}, \textit{BRCA2} and several other genes in \textit{malignant} IDC. In contrast, the genetic players of atypia and DCIS, and their prognostic value, remain unknown. Therefore, considerable interest lies among clinical and biomedical research community to identify the gene expression changes driving cell fate decisions at the atypia and DCIS stages, which ultimately skew the balance to malignant transformation.

Changes in cell phenotype from normal to atypical to malignant are reflected in changes of gene expression, also known as “gene expression signatures”. Large-scale studies using microarray and more recently, next-generation sequencing technologies are proving useful in defining gene expression signatures of breast cancer progression from early atypia to invasive ductal carcinoma. Over the last decade, a large number of microarray gene expression studies were conducted to uncover the gene expression signature of atypia and DCIS in human patients, mouse models of breast cancer, and mouse-human comparisons [7-12]. Cross-comparison of microarray studies performed by different research teams often yield little to no overlap [13], leaving the gene expression signature of atypia and DCIS elusive. Thus, a fundamental goal of translational bioinformatics is identification of early stage expression changes for impending invasive breast cancer from biopsies of premalignant atypia and DCIS compared to control (non-cancerous).
### 3.1.1 Determining Transposable Element Expression During Breast Cancer Progression

The role of dysregulated LTR retrotransposons as mutagenic agents is widely known, as accumulation of retrotransposition-induced mutations in critical genes contributes to oncogenesis [14]. Retrotransposition is a hallmark of advanced cancer driving genome instability [15, 16]. Of note, HERV-K elements have been detected in the plasma of people diagnosed with invasive breast cancer [17]. However, it is unknown if specific LTR retrotransposons are expressed in the early stages of atypia and DCIS. TEs, notably LTRs, expression at early stages indicate the mechanisms driving genomic instability in cancer become active quite early in mammary tumor progression. Moreover, TE expression is linked to the undifferentiated state of the chromatin from stem cell studies, which has been implicated as a driver in cellular transformation and oncogenesis.

### 3.1.2 Research Purpose and Approach

The purpose of this research is to investigate, using *in silico* methods, the expression signature of transposable elements during breast cancer progression. We hypothesize that transcriptional re-activation of endogenous retrotransposons occurs during transformation of atypia and DCIS. To test this hypothesis, we designed a Transposon Enrichment Set Analyses (TESA) based on open-source, published tools freely available to the scientific and research community. Furthermore, we also measured transcript expression from transcriptome data to detect and corroborate known oncogenic pathways.
3.2 Methods

3.2.1 Transcript & TESA Pipeline

Classic bioinformatic pipelines generated to analyze RNA expression rely on a reference genome file (.gtf) during the alignment process and a gene annotation file (.gtf or .fasta), which specifies the name of the gene for the genomic location from read alignment. Here, we generated a unique transposable element annotation file from the RepeatMasker program [18]. The RepeatMasker screens DNA sequences for repeats and low complexity DNA sequences then outputs a detailed annotation from the query. The annotation file generated from RepeatMasker was then used in a RNA-Seq bioinformatics pipeline further explained below (figure 3.1).

Briefly, RNA-Seq data of biopsies from formalin-fixed, paraffin-embedded breast tissue during oncogenic progression were obtained from NCBI’s national database GEO DataSets [19]. These datasets were aligned using STAR aligner to human reference file hg19. Read summarization of transposable element counts were generated using featureCounts [20] with the alignment file generated with RepeatMasker, as mentioned.

![Figure 3.1 Transposon Enrichment Set Analyses (TESA) pipeline/workflow. Orange boxes are bioinformatic tools. Light blue boxes are output files. Dark blue boxes are input files. (Unpublished)
previously, and transcript counts were generated using an annotation file from ENSEMBL. To assess intrasample and intergroup quality raw counts were normalized using counts per million, as previously described (Chapter 2). PCA plots were generated with the normalized samples and group outliers were identified and removed. The normalized count files that fell within the 95% of the confidence interval were further processed through DESeq2 to detect differential expression (DE) of TEs and transcripts among all stages. Gene ontology terms were then assigned to DE transcripts using VisuAL Annotation Display.

3.2.2 Alignment of RNA-Seq Data Obtained from Clinical and Cell Line Models Using STAR

We aligned the raw breast cancer progression data to the human genome reference build HG19 using STAR aligner. STAR’s algorithm [21] uses a two-step approach. STAR aligns the first portion, or seed, of a read to a reference genome up to the maximum mappable length (MML) of the read then aligns the left-over portion, second seed, of the read, up to its MML. After the read is completely aligned, STAR joins the two, or more, seeds together and scores them based off a user-defined penalty for mismatches, insertions, and deletions. The “merged” seeds with the highest score are chosen as the correct alignment of a read. This approach allows for quick and easy annotation of multi-mapping reads with their own alignment scores. --seedSearchStartLmax 50 – seedSearchStartLmaxOverLread 1.0 –seedSearchLmax 0 –seedMultimapNmax 10000 –seedPerReadMax 1000 –seedPerWindowMax 50 –seedNonoLociPerWindow 10 –alignIntronMin 21 –alignIntronMax 0 –alignMatesGapMax 0 –alignSJoverhangMin 5 –alignSJDBoverhangmin 3 –alignSpliceMateMapLmin 0 –
alignSplicedMateMapeLminOverLmate 0.66 –alignWindowsPerReadNmax 10000 –
alignTranscriptsPerWindowNmax 100 –alignTranscriptsPerReadNmax 10000 –
alignEndsType Local

3.2.3 Quality Control Procedure

To detect and ultimately eliminate outliers of the normalized count files produced from the featureCounts program we utilized ClustVis. This tool is a web graphical interface built on many R packages that produces Principal Component Analysis plots from a data matrix. The tool can handle high dimensional data with many ways to input, such as copy and paste or upload [22].

3.2.5 Differential Expression Analysis – DESeq2

DESeq2 is an R package that requires raw counts in a data matrix. The program normalizes the raw counts through a general linear model to account for differences in sequencing depth [23]. DESeq2 assumes the normalize counts fit a negative binomial distribution, then adjusts the distribution through unique Empirical Bayes methods to account for low count variability. DESeq2 also uses Empirical Bayes on log-fold change estimates to reduce false positives [24]. Within R, the “raw count data” was imported as a matrix along with a “DataFrame” that had a column which annotates the columns of the imported matrix. A DESeqDataSet was generated with these imported files using the DESeq2 function DESeqDataSetFromMatrix(countData=”raw count data”,
colData=”DataFrame”, design=~0 + Stage). The names in the quotation marks are arbitrary and assigned by the user when importing the corresponding files all quotation marks should be removed when using the functions in R. The design argument within the
function is a formula which expresses how the counts for each gene depend on the variables in colData. Here, the design formula of ~0 + Stage will build a results table and plot all experimental group comparisons. The data was filtered to keep only genes that had at least 4 counts between at least 10 samples generating the function keep <- rowSums(counts(“name_DESEQDataSetFromMatrix”)>4) >=10 and applying it to “name_DESEQDataSetFromMatrix” with name_DESEQDataSetFromMatrix <- name_DESEQDataSetFromMatrix[keep,]. The DE analysis was called using DESeq(“name_DESEQDataSetFromMatrix”).

3.2.6 VLAD Gene Ontology Analysis of DE Output

VLAD [25], accessible via MGI web portal, is a powerful tool to find common functional themes in the lists of genes by analyzing statistical over- or underrepresentation of ontological annotations. Currently, users can choose among Gene Ontology (GO) [26] annotations for human genes, Gene Ontology and Mammalian Phenotype Ontology (MP) [27] annotations for mouse genes, or upload a file of custom annotations (in open biomedical ontology [28] known as ‘obo’ format). Unlike other packages for ontological enrichment, VLAD allows analysis of more than one query (i.e., several lists of genes may be analyzed and visualized simultaneously), as well as permits user to provide own “universe set”, i.e. gene list to test queries. For GO analysis, we searched for overrepresentation among terms with experimental evidence (i.e., codes EXP, “Inferred from experiment”; IDA, “Inferred from direct assay”; IMP, “Inferred from mutant phenotype”; TAS, “Traceable author statement”). Here we analyzed the genes that were differentially expressed in the disease state when compared to normal. The output settings for VLAD was set to show the top 10 GO term nodes. Additional GO analysis was
done on only the shared genes that had significantly differentially expression between all three disease stages.

### 3.3 Results

#### 3.3.1 Expression of Transcripts Across Cancer Progression

The number of downregulated, when compared to normal, differentially expressed transcripts increased at each stage starting with atypia (224) to DCIS (1026) to IDC (1797). The number of upregulated when compared to normal increased at each stage starting with atypia (208) to DCIS (1133) to IDC (1923). The number of upregulated genes were greater than the number of downregulated genes at each stage except for atypia. All three disease states had 71 genes in common that were upregulated and 107 downregulated. Atypia shared 41 and 23 upregulated genes and 28 and 29 downregulated genes.

![Figure 3.2 Venn Diagram of Differentially Expressed Transcripts](image)

*Figure 3.2 Venn Diagram of Differentially Expressed Transcripts.* The number in each circle represents the amount of significant differentially expressed genes between comparisons to normal. The overlapping number is the mutual differentially expressed genes between the comparisons. The nonoverlapping numbers are the unique differentially expressed genes to the stage **A.** Increased expression. **B.** Decreased expression. (Unpublished)
genes with DCIS and IDC, respectively. DCIS and IDC shared 480 upregulated and 652 downregulated genes (Figure 3.2).

### 3.3.2 Expression of All TEs Across Cancer Progression

The percent of transcriptome made up by TEs varied from each stage. Normal samples had ~14% of their transcriptome derived from TEs. Atypia, DCIS and IDC stages had 16%, 31%, and 24% of their transcriptome derived from TE sequences, respectively (Figure 3.3). The most abundant TE in the transcriptome was by far SINE Alus, followed by LINE L1 elements (Figure 3.3). Differential expression analysis with DESeq2 revealed a total of 91 TEs were differentially expressed in atypia when compared to normal, 50 upregulated and 41 downregulated (Table 3.1). DCIS had 93 DE TEs with 52 downregulated and 41 upregulated. IDC had 90 DE TEs with 38 downregulated and 52 upregulated (Table 3.1).

![Figure 3.3 Gene transcriptome and TE transcriptome.](image)

Total transcriptome represents of all reads mapped for each stage. TE transcriptome applies to only the portion of TEs mapped. Labels inside the donut charts N, A, D, and I are normal, atypia, DCIS, IDC, respectively. (Unpublished)
All three stages had 14 TEs downregulated and 8 TEs upregulated in common. Atypia...
shared 6 upregulated TEs and 6 downregulated TEs with DCIS and only 3 upregulated TEs with IDC (Figure 3.4). The most differentially expressed TEs are derived from LTR retrotransposons (Figure 3.5).

Table 3.1 Differentially Expressed TEs for Each Stage

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Decreased Expression</th>
<th>Increased Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean counts</td>
<td>log2fc</td>
</tr>
<tr>
<td>Arthur1A</td>
<td>88.31144</td>
<td>-1.19761</td>
</tr>
<tr>
<td>LTR26</td>
<td>42.94001</td>
<td>-0.84414</td>
</tr>
<tr>
<td>MER9B</td>
<td>3.946216</td>
<td>-0.71463</td>
</tr>
<tr>
<td>LTR47B</td>
<td>38.77274</td>
<td>-0.70031</td>
</tr>
<tr>
<td>MER75A</td>
<td>5.174508</td>
<td>-0.67639</td>
</tr>
<tr>
<td>LTR45B</td>
<td>20.86345</td>
<td>-0.61464</td>
</tr>
<tr>
<td>LTR31</td>
<td>33.38377</td>
<td>-0.64127</td>
</tr>
<tr>
<td>MER92A</td>
<td>26.71537</td>
<td>-0.63421</td>
</tr>
<tr>
<td>MER9a1</td>
<td>5.039478</td>
<td>-0.63007</td>
</tr>
<tr>
<td>UCON8</td>
<td>1.339151</td>
<td>-0.61849</td>
</tr>
<tr>
<td>LTR33C</td>
<td>30.45387</td>
<td>-0.60741</td>
</tr>
<tr>
<td>MER30</td>
<td>3304.641</td>
<td>-0.58869</td>
</tr>
<tr>
<td>Kanga1</td>
<td>31.79583</td>
<td>-0.55766</td>
</tr>
<tr>
<td>LTR5</td>
<td>11.39286</td>
<td>-0.50638</td>
</tr>
<tr>
<td>MER4B</td>
<td>126.7963</td>
<td>-0.50568</td>
</tr>
<tr>
<td>Charlie26a</td>
<td>48.28848</td>
<td>-0.48566</td>
</tr>
<tr>
<td>AluYc3</td>
<td>422.5673</td>
<td>-0.48151</td>
</tr>
<tr>
<td>PABL_B-int</td>
<td>26.668</td>
<td>-0.47848</td>
</tr>
<tr>
<td>MER30B</td>
<td>93.93234</td>
<td>-0.4441</td>
</tr>
<tr>
<td>LTR22A</td>
<td>32.48954</td>
<td>-0.43086</td>
</tr>
<tr>
<td>LTR57</td>
<td>149.023</td>
<td>-0.42003</td>
</tr>
<tr>
<td>MLT1G1</td>
<td>317.0708</td>
<td>-0.41623</td>
</tr>
<tr>
<td>LTR2C</td>
<td>70.10864</td>
<td>-0.40906</td>
</tr>
<tr>
<td>LTR7Y</td>
<td>13.69006</td>
<td>-0.40684</td>
</tr>
<tr>
<td>LTR71A</td>
<td>15.19133</td>
<td>-0.39458</td>
</tr>
<tr>
<td>MamSINE1</td>
<td>30.62804</td>
<td>-0.37619</td>
</tr>
<tr>
<td>LTR71B</td>
<td>36.38883</td>
<td>-0.37396</td>
</tr>
<tr>
<td>MER67B</td>
<td>22.97493</td>
<td>-0.36667</td>
</tr>
<tr>
<td>LFSINE_Vert</td>
<td>34.7053</td>
<td>-0.35311</td>
</tr>
</tbody>
</table>
Table 3.1 Differentially Expressed TEs for Each Stage continued

<table>
<thead>
<tr>
<th>Normal vs Atypia TE Differential Expression</th>
<th>Normal vs DCIS TE Differential Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decreased Expression</strong></td>
<td><strong>Increased Expression</strong></td>
</tr>
<tr>
<td>Gene ID</td>
<td>mean counts</td>
</tr>
<tr>
<td>LTR19C</td>
<td>23.57059</td>
</tr>
<tr>
<td>MER34A1</td>
<td>36.36152</td>
</tr>
<tr>
<td>MER57A-int</td>
<td>117.5749</td>
</tr>
<tr>
<td>MLT2B3</td>
<td>137.7282</td>
</tr>
<tr>
<td>MER4B-int</td>
<td>142.0717</td>
</tr>
<tr>
<td>HERVE-int</td>
<td>64.99913</td>
</tr>
<tr>
<td>MER4D1</td>
<td>86.04822</td>
</tr>
<tr>
<td>LTR33</td>
<td>335.7643</td>
</tr>
<tr>
<td>LRT2B3</td>
<td>174.62155</td>
</tr>
<tr>
<td>MER115</td>
<td>104.8611</td>
</tr>
<tr>
<td>MER34A</td>
<td>62.62155</td>
</tr>
<tr>
<td>MER57A-int</td>
<td>58.07753</td>
</tr>
<tr>
<td>L1MEd</td>
<td>775.946</td>
</tr>
<tr>
<td>L1MC1</td>
<td>1230.854</td>
</tr>
</tbody>
</table>

95
Table 3.1 Differentially Expressed TEs for Each Stage continued

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Decreased Expression</th>
<th>Increased Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR14C</td>
<td>17.30651</td>
<td>-0.70115</td>
</tr>
<tr>
<td>LTR14B</td>
<td>32.79084</td>
<td>-0.70005</td>
</tr>
<tr>
<td>MER11A</td>
<td>82.26047</td>
<td>-0.6841</td>
</tr>
<tr>
<td>LTR31</td>
<td>33.38377</td>
<td>-0.66541</td>
</tr>
<tr>
<td>LTR77</td>
<td>11.17067</td>
<td>-0.63033</td>
</tr>
<tr>
<td>HERVK3-int</td>
<td>42.23356</td>
<td>-0.61771</td>
</tr>
<tr>
<td>LTR45B</td>
<td>20.86345</td>
<td>-0.61327</td>
</tr>
<tr>
<td>HERVK14C-int</td>
<td>29.48786</td>
<td>-0.60756</td>
</tr>
<tr>
<td>PABL_B-int</td>
<td>26.668</td>
<td>-0.60702</td>
</tr>
<tr>
<td>LTR2B</td>
<td>64.26932</td>
<td>-0.57812</td>
</tr>
<tr>
<td>Ricksha_b</td>
<td>6.586193</td>
<td>-0.56754</td>
</tr>
<tr>
<td>Charlie5</td>
<td>140.7848</td>
<td>-0.56399</td>
</tr>
<tr>
<td>LTR47A</td>
<td>122.1144</td>
<td>-0.56255</td>
</tr>
<tr>
<td>LTR18B</td>
<td>554.2769</td>
<td>-0.56142</td>
</tr>
<tr>
<td>Tigger16a</td>
<td>27.83332</td>
<td>-0.52133</td>
</tr>
<tr>
<td>LTR7</td>
<td>269.9551</td>
<td>-0.52074</td>
</tr>
<tr>
<td>MLT1J</td>
<td>1437.407</td>
<td>-0.50306</td>
</tr>
<tr>
<td>LTR52</td>
<td>74.46954</td>
<td>-0.50299</td>
</tr>
<tr>
<td>LTR71B</td>
<td>36.38883</td>
<td>-0.48756</td>
</tr>
<tr>
<td>LTR5A</td>
<td>276.3234</td>
<td>-0.48653</td>
</tr>
<tr>
<td>MamSINE1</td>
<td>30.62804</td>
<td>-0.48147</td>
</tr>
<tr>
<td>MLT2B3</td>
<td>137.7282</td>
<td>-0.47711</td>
</tr>
<tr>
<td>HERVE-int</td>
<td>64.99913</td>
<td>-0.46512</td>
</tr>
<tr>
<td>LTR19C</td>
<td>23.57059</td>
<td>-0.45116</td>
</tr>
<tr>
<td>MER113B</td>
<td>15.63274</td>
<td>-0.44184</td>
</tr>
<tr>
<td>LTR16B</td>
<td>29.38154</td>
<td>-0.4357</td>
</tr>
<tr>
<td>MER74C</td>
<td>15.92237</td>
<td>-0.43204</td>
</tr>
<tr>
<td>MER34A1</td>
<td>36.36152</td>
<td>-0.42523</td>
</tr>
<tr>
<td>Tigger2a</td>
<td>120.1334</td>
<td>-0.4221</td>
</tr>
<tr>
<td>MIRb</td>
<td>9179.493</td>
<td>-0.39734</td>
</tr>
<tr>
<td>MER4D1</td>
<td>86.04822</td>
<td>-0.38151</td>
</tr>
<tr>
<td>MER63C</td>
<td>92.03893</td>
<td>-0.37102</td>
</tr>
<tr>
<td>L1MEg2</td>
<td>52.67794</td>
<td>-0.35629</td>
</tr>
<tr>
<td>THE1D</td>
<td>290.7215</td>
<td>-0.33133</td>
</tr>
<tr>
<td>MER5A1</td>
<td>216.3702</td>
<td>-0.32085</td>
</tr>
<tr>
<td>MER68</td>
<td>171.6801</td>
<td>-0.31175</td>
</tr>
<tr>
<td>Charlie19a</td>
<td>41.869</td>
<td>-0.31096</td>
</tr>
</tbody>
</table>
Table 3.1 Differentially Expressed TEs for Each Stage continued

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>mean counts</th>
<th>log2fc</th>
<th>Gene ID</th>
<th>mean counts</th>
<th>log2fc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthur1A</td>
<td>88.31144</td>
<td>-2.22411</td>
<td>MSR1</td>
<td>83.43339</td>
<td>1.459245</td>
</tr>
<tr>
<td>LTR33C</td>
<td>30.45387</td>
<td>-1.43662</td>
<td>LTR10B1</td>
<td>29.21936</td>
<td>1.313723</td>
</tr>
<tr>
<td>LTR47B</td>
<td>38.77274</td>
<td>-1.38331</td>
<td>HERVI</td>
<td>-1.459245</td>
<td>1.18361</td>
</tr>
<tr>
<td>LTR26</td>
<td>42.94001</td>
<td>-1.19</td>
<td>HERV3-int</td>
<td>59.02318</td>
<td>1.05755</td>
</tr>
<tr>
<td>LTR11</td>
<td>4.07613</td>
<td>-1.09968</td>
<td>LTR12B</td>
<td>59.98547</td>
<td>1.01489</td>
</tr>
<tr>
<td>UCON2</td>
<td>6.113835</td>
<td>-1.02234</td>
<td>LTR35B</td>
<td>2.305686</td>
<td>0.923133</td>
</tr>
<tr>
<td>LTR22B</td>
<td>9.025771</td>
<td>-0.94438</td>
<td>UCON28b</td>
<td>5.92568</td>
<td>0.923133</td>
</tr>
<tr>
<td>LTR14C</td>
<td>17.30651</td>
<td>-0.93774</td>
<td>REP522</td>
<td>85.42611</td>
<td>0.904658</td>
</tr>
<tr>
<td>Kanga1</td>
<td>31.79583</td>
<td>-0.85578</td>
<td>MER57D</td>
<td>3.76455</td>
<td>0.857979</td>
</tr>
<tr>
<td>UCON8</td>
<td>1.339151</td>
<td>-0.76292</td>
<td>MER126</td>
<td>119.5002</td>
<td>0.72054</td>
</tr>
<tr>
<td>LTR31</td>
<td>33.38377</td>
<td>-0.7259</td>
<td>MLT2B4</td>
<td>175.1764</td>
<td>0.704449</td>
</tr>
<tr>
<td>MER136</td>
<td>7.400865</td>
<td>-0.71672</td>
<td>MLT1E1</td>
<td>43.42278</td>
<td>0.731794</td>
</tr>
<tr>
<td>HERVK14C-int</td>
<td>29.48786</td>
<td>-0.71089</td>
<td>L1P3b</td>
<td>1.846388</td>
<td>0.731274</td>
</tr>
<tr>
<td>LTR43B</td>
<td>5.763048</td>
<td>-0.63699</td>
<td>ACRO1</td>
<td>2.914907</td>
<td>0.723884</td>
</tr>
<tr>
<td>LTR18B</td>
<td>554.2769</td>
<td>-0.62973</td>
<td>MamRep4096</td>
<td>119.5002</td>
<td>0.72054</td>
</tr>
<tr>
<td>LTR5A</td>
<td>276.3234</td>
<td>-0.60201</td>
<td>MER117</td>
<td>474.3747</td>
<td>0.704449</td>
</tr>
<tr>
<td>LTR81AB</td>
<td>20.83903</td>
<td>-0.59248</td>
<td>MER72B</td>
<td>17.07958</td>
<td>0.691256</td>
</tr>
<tr>
<td>ORSL-2a</td>
<td>4.846807</td>
<td>-0.55552</td>
<td>LSAU</td>
<td>11.58124</td>
<td>0.665175</td>
</tr>
<tr>
<td>LTR77</td>
<td>11.17067</td>
<td>-0.5368</td>
<td>LTR3B</td>
<td>83.12914</td>
<td>0.633166</td>
</tr>
<tr>
<td>LTR45B</td>
<td>20.86345</td>
<td>-0.53585</td>
<td>LTR48B</td>
<td>50.26954</td>
<td>0.612501</td>
</tr>
<tr>
<td>HAL1-3A_ME</td>
<td>133.1612</td>
<td>-0.51514</td>
<td>MLT-int</td>
<td>19.4105</td>
<td>0.608322</td>
</tr>
<tr>
<td>LTR71B</td>
<td>36.38883</td>
<td>-0.51507</td>
<td>L1P</td>
<td>0.98082</td>
<td>0.590872</td>
</tr>
<tr>
<td>MER57E1</td>
<td>124.6372</td>
<td>-0.50986</td>
<td>LTR89</td>
<td>91.29133</td>
<td>0.57424</td>
</tr>
<tr>
<td>MER11A</td>
<td>82.26047</td>
<td>-0.50613</td>
<td>MER91B</td>
<td>56.85263</td>
<td>0.564587</td>
</tr>
<tr>
<td>MamSINE1</td>
<td>30.62804</td>
<td>-0.48073</td>
<td>MER34B-int</td>
<td>58.07753</td>
<td>0.568957</td>
</tr>
<tr>
<td>HERVE-int</td>
<td>64.99913</td>
<td>-0.44387</td>
<td>Charlie7a</td>
<td>491.628</td>
<td>0.559075</td>
</tr>
<tr>
<td>LTR71A</td>
<td>15.19133</td>
<td>-0.43595</td>
<td>LTR45C</td>
<td>95.3888</td>
<td>0.554385</td>
</tr>
<tr>
<td>MER74C</td>
<td>15.92237</td>
<td>-0.43385</td>
<td>MER4A1</td>
<td>209.7975</td>
<td>0.54857</td>
</tr>
<tr>
<td>LTR24B</td>
<td>16.62667</td>
<td>-0.4284</td>
<td>UCON4</td>
<td>13.45966</td>
<td>0.505048</td>
</tr>
<tr>
<td>MER106B</td>
<td>34.97379</td>
<td>-0.42594</td>
<td>SVA_F</td>
<td>177.8978</td>
<td>0.48302</td>
</tr>
<tr>
<td>L1M3f</td>
<td>84.27394</td>
<td>-0.41599</td>
<td>MamGypLTR2c</td>
<td>47.37201</td>
<td>0.474557</td>
</tr>
<tr>
<td>Tigger10</td>
<td>114.5806</td>
<td>-0.40669</td>
<td>MER54A</td>
<td>106.9699</td>
<td>0.467797</td>
</tr>
<tr>
<td>Charlie5</td>
<td>140.7848</td>
<td>-0.39116</td>
<td>LTR66</td>
<td>58.15395</td>
<td>0.460548</td>
</tr>
<tr>
<td>MER34A1</td>
<td>36.36152</td>
<td>-0.3804</td>
<td>LTR10E</td>
<td>26.02065</td>
<td>0.449156</td>
</tr>
<tr>
<td>MER51-int</td>
<td>56.46602</td>
<td>-0.37199</td>
<td>MER63B</td>
<td>273.1816</td>
<td>0.447306</td>
</tr>
</tbody>
</table>
Table 3.1 Differentially Expressed TEs for Each Stage continued

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>mean counts</th>
<th>log2fc</th>
<th>Gene ID</th>
<th>mean counts</th>
<th>log2fc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOR1b</td>
<td>55.24626</td>
<td>-0.33469</td>
<td>MER113A</td>
<td>135.9744</td>
<td>0.445428</td>
</tr>
<tr>
<td>MSTB</td>
<td>583.0843</td>
<td>-0.25025</td>
<td>Arthur1C</td>
<td>12.71432</td>
<td>0.437378</td>
</tr>
<tr>
<td>SVA_D</td>
<td>170.6297</td>
<td>-0.22132</td>
<td>MER1B</td>
<td>437.3959</td>
<td>0.414795</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER31B</td>
<td>73.26301</td>
<td>0.412004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MLT2C1</td>
<td>91.98793</td>
<td>0.401928</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER61C</td>
<td>110.6387</td>
<td>0.401604</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tigger11a</td>
<td>440.1543</td>
<td>0.399371</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER68-int</td>
<td>14.18787</td>
<td>0.38088</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LTR79</td>
<td>87.83257</td>
<td>0.380088</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Charlie16a</td>
<td>109.2027</td>
<td>0.374402</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER102b</td>
<td>527.5156</td>
<td>0.354318</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MLT1C</td>
<td>1671.165</td>
<td>0.343778</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MLT1G</td>
<td>106.6917</td>
<td>0.304594</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER6A</td>
<td>97.45217</td>
<td>0.303702</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MER115</td>
<td>104.8611</td>
<td>0.290959</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L2c</td>
<td>6367.266</td>
<td>0.233836</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L1MB5</td>
<td>967.9758</td>
<td>0.163403</td>
</tr>
</tbody>
</table>
3.3.3 VLAD Gene Ontology Output

3.3.3a Examples of Validation of Cancer Pathways

The top 10 GO molecular functions terms for transcripts that were increased in atypia compared to normal centered around transmembrane transported activity (Table 3.2 and Figure 3.6). In DCIS the GO analysis of transcripts that were increased compared to normal revealed terms for not only transmembrane transporter activity, but also GTPase binding to include ras GTPase binding and intracellular trafficking (Table 3.2 and Figure 3.7). The increased expressed transcripts from IDC had an overwhelming number of cellular components GO terms for extracellular pathways. Furthermore, the molecular function GO terms for IDC involved cell adhesion, extra cellular matrix constituents, and collagen binding, as well as, growth factor binding and protein kinase binding (Table 3.2 and Figure 3.8).

Figure 3.6 Top Ten Overrepresented Gene Ontology Terms of Upregulated Differentially Expressed Genes in Atypia. A. Biological Processes. B. Cellular Components. C. Molecular Functions (Unpublished)
Figure 3.6 (continued) Top Ten Overrepresented Gene Ontology Terms of Upregulated Differentially Expressed Genes in Atypia. A. Biological Processes. B. Cellular Components. C. Molecular Functions (Unpublished)
Figure 3.7 Top Ten Overrepresented Gene Ontology Terms of Upregulated Differentially Expressed Genes in DCIS. A. Biological Processes. B. Cellular Components. C. Molecular Functions (Unpublished)
Figure 3.8 Top Ten Overrepresented Gene Ontology Terms of Upregulated Differentially Expressed Genes in IDC. A. Biological Processes. B. Cellular Components. C. Molecular Functions (Unpublished)
**3.3.3b Examples of Novel Pathways**

For example, GO analysis of DE gene in DCIS and IDC when compared to normal produced outputs that reveal genes in the category to “Response to Virus” under the biological process hierarchy and under the molecular function hierarchy, genes belonging to Interferon Activation Response terms are significantly overrepresented (Table 3.2 and Figures 3.7 & 3.8). Additionally, the 2’-5’-oligoadenylate synthetase activity term for molecular function is present for both DCIS and IDC, which indicates this pathway is upregulated at a precancerous stage before malignancy occurs (Table 3.2 and Figures 3.7 & 3.8). This is a novel target pathway to inhibit and mitigate cellular transformation to malignancy.

**3.3.3c Conserved Transcripts**

A GO analysis of the 71 conserved transcripts that had an increase in expression across all stages of cancer compared to normal, revealed many terms associated with regulating the differentiation processes, transport activity, and hormone binding.
<table>
<thead>
<tr>
<th>Biological Processes</th>
<th>Atypia TermID</th>
<th>Term</th>
<th>DCIS TermID</th>
<th>Term</th>
<th>IDC TermID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0060971</td>
<td>embryonic heart tube left/right pattern formation</td>
<td>GO:0071357 cellular response to type I interferon</td>
<td>GO:0030198 extracellular matrix organization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048468</td>
<td>cell development</td>
<td>GO:0060337 type I interferon signaling pathway</td>
<td>GO:0043062 extracellular structure organization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0060562</td>
<td>epithelial tube morphogenesis</td>
<td>GO:0034340 response to type I interferon</td>
<td>GO:0002376 immune system process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0060972</td>
<td>left/right pattern formation</td>
<td>GO:0046903 secretion</td>
<td>GO:0019221 cytokine-mediated signaling pathway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0030856</td>
<td>regulation of epithelial cell differentiation</td>
<td>GO:0032879 regulation of localization</td>
<td>GO:0046903 secretion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008016</td>
<td>regulation of heart contraction</td>
<td>GO:0006811 ion transport</td>
<td>GO:0032879 regulation of localization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006811</td>
<td>ion transport</td>
<td>GO:0051179 localization</td>
<td>GO:0045055 regulated exocytosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:1990778</td>
<td>protein localization to cell periphery</td>
<td>GO:0006629 lipid metabolic process</td>
<td>GO:0032940 secretion by cell</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0051239</td>
<td>regulation of multicellular organismal process</td>
<td>GO:0050790 regulation of catalytic activity</td>
<td>GO:0051239 regulation of multicellular organismal process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0099637</td>
<td>neurotransmitter receptor transport</td>
<td>GO:0009615 response to virus</td>
<td>GO:0034097 response to cytokine</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cellular Components</th>
<th>Atypia TermID</th>
<th>Term</th>
<th>DCIS TermID</th>
<th>Term</th>
<th>IDC TermID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005891</td>
<td>voltage-gated calcium channel complex</td>
<td>GO:0005737 cytoplasm</td>
<td>GO:0031982 vesicle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0097458</td>
<td>neuron part</td>
<td>GO:0044444 cytoplasmic part</td>
<td>GO:0044421 extracellular region part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:1902495</td>
<td>transmembrane transporter complex</td>
<td>GO:0031982 vesicle</td>
<td>GO:0005615 extracellular space</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 Top 10 GO Terms of DE Genes for Each Stage Compared to Control and Each Category continued

<table>
<thead>
<tr>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:1990351</td>
<td>transporter complex</td>
<td>GO:0044425</td>
<td>membrane part</td>
<td>GO:0031012</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td>GO:0034704</td>
<td>calcium channel complex</td>
<td>GO:0044459</td>
<td>plasma membrane part</td>
<td>GO:0005576</td>
<td>extracellular region</td>
</tr>
<tr>
<td>GO:0034702</td>
<td>ion channel complex</td>
<td>GO:0016020</td>
<td>membrane</td>
<td>GO:0005737</td>
<td>cytoplasm</td>
</tr>
<tr>
<td>GO:0008282</td>
<td>inward rectifying potassium</td>
<td>GO:0031410</td>
<td>cytoplasmic vesicle</td>
<td>GO:0070062</td>
<td>extracellular exosome</td>
</tr>
<tr>
<td>GO:0044459</td>
<td>plasma membrane part</td>
<td>GO:0097708</td>
<td>intracellular vesicle</td>
<td>GO:1903561</td>
<td>extracellular vesicle</td>
</tr>
<tr>
<td>GO:0098590</td>
<td>plasma membrane region</td>
<td>GO:0070062</td>
<td>extracellular exosome</td>
<td>GO:0043230</td>
<td>extracellular organelle</td>
</tr>
<tr>
<td>GO:0043005</td>
<td>neuron projection</td>
<td>GO:0012505</td>
<td>endomembrane system</td>
<td>GO:0044444</td>
<td>cytoplasmic part</td>
</tr>
<tr>
<td>GO:0034703</td>
<td>cation channel complex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0022857</td>
<td>transmembrane transporter activity</td>
<td>GO:0001730</td>
<td>2'-5'-oligoadenylate synthetase activity</td>
<td>GO:0005515</td>
<td>protein binding</td>
</tr>
<tr>
<td>GO:0005245</td>
<td>voltage-gated calcium channel activity</td>
<td>GO:0022857</td>
<td>transmembrane transporter activity</td>
<td>GO:0005201</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>structural constituent</td>
</tr>
<tr>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>GO:0016822</td>
<td>hydrolase activity, acting on acid carbon-carbon bonds</td>
<td>GO:0008092</td>
<td>cytoskeletal protein binding</td>
</tr>
<tr>
<td>GO:0022843</td>
<td>voltage-gated cation channel activity</td>
<td>GO:0016823</td>
<td>hydrolase activity, acting on acid carbon-carbon bonds, in ketonic substances</td>
<td>GO:0050839</td>
<td>cell adhesion molecule binding</td>
</tr>
</tbody>
</table>
Table 3.2 Top 10 GO Terms of DE Genes for Each Stage Compared to Control and Each Category continued

<table>
<thead>
<tr>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
<th>TermID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0015075</td>
<td>ion transmembrane transporter activity</td>
<td>GO:0017016</td>
<td>Ras GTPase binding</td>
<td>GO:0019838</td>
<td>growth factor binding</td>
</tr>
<tr>
<td>GO:0005216</td>
<td>ion channel activity</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>GO:0048407</td>
<td>platelet-derived growth factor binding</td>
</tr>
<tr>
<td>GO:0022838</td>
<td>substrate-specific channel activity</td>
<td>GO:0031267</td>
<td>small GTPase binding</td>
<td>GO:0042802</td>
<td>identical protein binding</td>
</tr>
<tr>
<td>GO:0022839</td>
<td>ion gated channel activity</td>
<td>GO:0051020</td>
<td>GTPase binding</td>
<td>GO:0001730</td>
<td>2'-5'-oligoadenylate synthetase activity</td>
</tr>
<tr>
<td>GO:0022836</td>
<td>gated channel activity</td>
<td>GO:0016491</td>
<td>oxidoreductase activity</td>
<td>GO:0005518</td>
<td>collagen binding</td>
</tr>
<tr>
<td>GO:0046873</td>
<td>metal ion transmembrane transporter activity</td>
<td>GO:0019899</td>
<td>enzyme binding</td>
<td>GO:0019901</td>
<td>protein kinase binding</td>
</tr>
<tr>
<td>GO:0015267</td>
<td>channel activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0022803</td>
<td>passive transmembrane transporter activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0015318</td>
<td>inorganic molecular entity transmembrane transporter activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4 Discussion

Through *in silico* methods, we determined that TEs made up a large portion of the transcriptome with the highest proportion in DCIS. The highest number of statistically significant differentially expressed TEs were LTRs. The highest number of LTRs were found to be differentially expressed at the DCIS stage.
Conserved transcripts, such as, HOTAIR [29-31], TMEM8A [32, 33], S100A7 [34, 35], and PAX genes [36-39] had increased expression in disease state compared to normal, are known oncogenes. Furthermore, GO analysis revealed oncogenic processes and pathways activated before and during the disease state, as well as novel pathways that may highlight interesting research to be further explored. Atypia was heavily populated with transporter activity GO terms, a hallmark of cancer stem cells [40]. Many of these terms were also found in DCIS and IDC. However, DCIS also produced terms involving GTPase binding, specifically ras GTPase, a well-known oncogene, and intracellular trafficking. Earlier reports of intracellular trafficking dysregulation in DCIS corroborates these findings [41]. The diversification of IDC GO terms into well-known oncogenic terms, such as, kinase binding, extracellular organization, growth factor binding, cytokine signaling, cell adhesion, and collagen binding indicates the invasive nature of the tissues sampled.

Novel pathways include the activity of 2'-5'-oligoadenylate synthetase which is known to induce an intrinsic antiviral state [42]. Additionally, many interferon response terms are present, which the expression of ERVs are known to induce, followed by an antiviral cell state [43-45]. These findings corroborate ERVs being the most differentially expressed transposable element and may indicate important implications in the role ERVs are playing in breast cancer progression.

Therefore, we have demonstrated a TE signature during breast cancer progression at the early stages, which indicates LTRs as the most abundantly differentially expressed. Furthermore, interferon responses as indicated VLAD GO analysis of abundant differentially expressed coding regions corroborates LTR endogenous retroviral activity, which was further substantiated in previous studies.
3.5 References

Chapter Four:

Regulation of Retrotransposon Activity in an In Vitro Model

4.1 Introduction

Transposable elements make up ~45% of the human genome. Interestingly, retrotransposons, a subgroup of transposable elements, make a larger portion of the human genome than protein coding genes. They are mobile genetic elements which can insert themselves into other genomic locations using a copy and paste mechanism, as reviewed in chapter 1. If left unchecked, the retrotransposition of these elements may become deleterious and hinder host survival. The role of retrotransposons as a mutagenic agent is well known and the accumulation of retrotransposon-induced mutations in important genes contributes to oncogenesis [1, 2]. One of the first lines of defense against retrotransposon activity is DNA methylation[3-7]. Along with other epigenetic repressive and posttranscriptional mechanisms these host defenses are quite effective[8-10]. However, because of the copy and paste mechanisms utilized by retrotransposons, our genome is riddled with multiple repeats of identical or highly homologous inserts. A recent study, using genome-wide microarray approaches that measured DNA methylation changes in cancerous tissue compared to adjacent normal tissue, found that disease tissue had highly variable disruption of epigenetic control. Specifically, certain members of transposon families, such as SVA, HERV, LINE-1-P, ALU, and MaLR, had a loss of methylation appeared in a stochastic fashion [11]. Here we attempt to gain insight
into the role of retrotransposons contributions to cancerous events by examining MCF10A and MCF10DCIS cell lines using pharmacological approaches.

### 4.1.1 Testing the Roles of Retrotransposons in the Transformation of Pre-Malignant to Malignant Breast Cancer using pharmacological approach

#### 4.1.1a Pharmacological approach

Until further advancements are made in gene targeted therapy to differentiate among the numerous copies of transposable elements, alternative approaches must be examined. As previously mentioned, TEs are primarily regulated by epigenetic mechanisms. One approach to induce expression of cryptic transcription start sites encoded by LTRs is to inhibit methyl group availability or DNMT activity [12]. Furthermore, there is experimental evidence that antiretroviral drugs, such as integrase inhibitors and reverse transcriptase inhibitors, influences endogenous retrovirus activity [13-16].

#### 4.1.1b Genetic Engineering Approaches

The application of genetic engineering to knockout, knockdown, or even overexpress genes of interest has revolutionized medical research. Genetic approaches utilize DNA sequence targeting techniques to regulate the loci of interest. Unfortunately, TEs have multiple locations within the host genome and a sequence targeted approach would damage substantial areas of the genome. For example, a single long terminal repeat family may have thousands of genomic inserts with the same or very highly similar sequences. For example, the LTR that acts as an alternative promoter for ALK [17] and is responsible for the transcription initiation of a constitutently active isoform found in many
cancers [18] has > 1,000 non-redundant and 5,000 redundant locations in the human genome (Figure 4.1) [19].

**Figure 4.1 Alternative Promoter and Nonredundant Genomic Locations of LTR16B2.**


### 4.1.2 Research Purpose and Approach

The goal of this study is to demonstrate TEs role in the transition from premalignant to malignant breast cancer. We hypothesize that retrotransposon TEs contribute to the transformation of breast cancer. To test this hypothesis, we utilized novel and previously established pharmacological techniques to induce TE expression and inhibit classical TE retrotransposition activity in an immortalized breast cell line and a breast cancer cell line which has characteristics of an early stage ductal carcinoma. The
expected outcomes of this study are the characterization of TE expression via effects on proliferation, migration, and anchorage independent growth.

4.2 Methods

4.2.1 Cell culture conditions of both cell lines

MCF10a and MCF10DCIS cells were cultured in DMEM (Gibco, 316000-034) with 5% horse serum (ATCC 30-2040), EGF (Sigma, E9644) at 20 ng/ml, hydrocortisone (Sigma, H0888) at 0.5 mg/ml, cholera toxin (Sigma, C8062) at 100 ng/ml, insulin (Sigma, I6634) at 10ug/ml, and 1% Penicillin-Streptomycin (Gibco, 15140). Cells used for experiments were passaged every 3 to 4 days and at a passage number between 5 to 30, the ranges arbitrarily chosen to maintain continuity.

4.2.2 Drugs

To reduce methyl group availability and retrotransposon repression we used SAM cycle inhibitors 3-Deazanplanocin A (Cayman Chemical Company, 11102), 3-Deazaadenosine (Cayman Chemical Company, 9000785), and DNMTi 5-Azacytidine (Sigma, A2385). To decrease retrotransposon activity, we used integrase inhibitors Raltegravir (Sigma, CDS023737) and Elvitegravir (Cayman Chemical Company, 17798), and reverse transcriptase inhibitor Azidothymidine (Sigma, PHR1292). Many of our drugs used dimethyl sulfoxide (DMSO, Sigma, D4540) as a vehicle.

4.2.2a 3-Deazaneplanocin A (DZNEP)

*In vitro* studies revealed DZNEP inhibits S-adenosyl-L-3H-methylmethionine and 3-thymidine incorporation at concentration of 700 nM (200ng/mL). At concentrations
between 500 nM and 1 µM (150 ng/mL and 300 ng/mL) DZNEP also depletes EZH2 and inhibits trimethylation of lysine 27 on histone 3 (H3k27) [20, 21].

### 4.2.2b 3-Deazaadenosine (DZA)

The half maximal inhibitory concentration (IC50) of DZA against S-adenosyl-homocysteine hydrolase is 4 µM. DZA has unspecific inhibitory activity at concentrations near 100 µM [22].

### 4.2.2c 5-Azacytidine (AZA)

At a concentration of 40 µM, AZA replaced ~8% of the cytosines in inhibitory DNA resulting in 0.6 – 1.8 nM ranges for DNMT inhibition [23].

### 4.2.2d Raltegravir (RAL)

Endogenous retroviral DNA remains in a pre-integration complex with integrase after reverse transcription. Integrase snips target DNA at the integration site. Raltegravir effectively inhibits HIV integrase and HTLV-1 cell-cell infection at concentrations of 20nM to 90nM (IC95 at 31±20 nM) [14, 24]. Furthermore, Raltegravir was effective at inhibiting HERV-K infection with an IC90 of 0.075 µM [16].

### 4.2.2e Elvitegravir (ELV)

Second generation integrase inhibitor has a broad antiretroviral activity and inhibits integrase at concentrations of 0.5–5.8 nM [15].

### 4.2.2f Azidothymidine (AZT)

Endogenous retroviruses are reverse transcribed from RNA to DNA through their own viral protein reverse transcriptase. AZT was shown to inhibit recombinant HIV RT at 0.32 ± 0.11 µM concentrations and endogenous retrotransposons at 16.4 ± 4.21 nM [13, 25, 26].
4.2.3 Cell Viability & Density Dose-Response Curves

To determine dosage for future experiments, cell viability and density assays were performed. Briefly, $25 \times 10^3$ of MCF10A or MCF10DCIS cells were seeded with drug into each well of a 24 well plate. All drugs were coded with an alphanumeric nomenclature before experiments to keep experimenter unaware of treatment groups. After 48 hours in treatment cells were trypsinized and pelleted then media was aspirated. The cell pellet was resuspended in fresh media. A 15 µl sample of the resuspension was mixed with tryphan blue at a 1:1 ratio and cell viability and density was analyzed using a hemocytometer. The experimental design is clarified in the table below. Cell viability was determined by dividing the number of live cells by the total number of cells. Cell density is the total number of live cells.

**Table 4.1 Cell viability and density dose-response curve experimental design**

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Drug Concentrations</th>
<th>n of Each Drug Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle H$_2$O</td>
<td>Volume of 10 µl/ml</td>
<td>4</td>
</tr>
<tr>
<td>Vehicle DMSO</td>
<td>10 µM, 100 µM</td>
<td>4, 2</td>
</tr>
<tr>
<td>DZNEP</td>
<td>0.1, 1, 10 µM</td>
<td>2</td>
</tr>
<tr>
<td>DZA</td>
<td>1, 10, 100 µM</td>
<td>2</td>
</tr>
<tr>
<td>5-Azacytidine</td>
<td>1, 10, 100 µM</td>
<td>2</td>
</tr>
<tr>
<td>Elvitegravir</td>
<td>1, 10, 100 µM</td>
<td>2</td>
</tr>
<tr>
<td>Raltegravir</td>
<td>1, 10, 100 µM</td>
<td>2</td>
</tr>
<tr>
<td>Azidothymidine</td>
<td>1, 10, 100 µM</td>
<td>2</td>
</tr>
</tbody>
</table>
4.2.4 Scratch Assay

4.2.4a Pilot Study

To examine growth rates and migration patterns as an indicator of the metastatic phenotype we performed scratch/wound healing assays. Pilot studies of the scratch assay were done in 24 well plates with each cell line, MCF10A and MCF10DCIS, seeded at $5 \times 10^4$ in cell culture media and allowed to reach confluency. When cells reached confluency, a scratch was made down the center of each well with a sterile 200 µl pipette tip. Cells were washed and media with treatment was added to labelled wells. Each treatment had 3 wells in the 24 well plate (Table 4.2). Images were taken at time point 0 and every 6 hours till the 24-hour time point to analyze area of scratch.

Table 4.2 Scratch assay pilot study experimental design

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Drug Concentration</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle H₂O</td>
<td>Volume of 10 µl/ml</td>
<td>3</td>
</tr>
<tr>
<td>Vehicle DMSO</td>
<td>10 µM</td>
<td>3</td>
</tr>
<tr>
<td>DZNEP</td>
<td>1 µM</td>
<td>3</td>
</tr>
<tr>
<td>DZA</td>
<td>10 µM</td>
<td>3</td>
</tr>
<tr>
<td>5-Azacytidine</td>
<td>10 µM</td>
<td>3</td>
</tr>
<tr>
<td>Elvitegravir</td>
<td>10 µM</td>
<td>3</td>
</tr>
<tr>
<td>Azidothymidine</td>
<td>10 µM</td>
<td>3</td>
</tr>
</tbody>
</table>

4.2.4b WoundMaker™

To further examine growth rates and migration or migration (drug + Mitomycin) alone we performed a wound/scratch assay using the Essen WoundMaker™, a 96-pin wound making tool. Briefly, $25 \times 10^3$ cells per well were seeded in cell culture media in all
96 wells. After 18hrs in an incubator at 37°C with 5% CO₂ the WoundMaker™ was utilized to create wounds in all wells. Cells were immediately washed twice with culture media (100 µl per well). Drugs labeled in a way to ensure blind experimental settings was applied to respectively labeled wells. The 96 well plate was then allowed to equilibrate in an IncuCyte® incubator before being imaged for time point 0 and every 6 hours for 24 hours. To block proliferation in the WoundMaker™ scratch assay for migration only, confluent cells were treated with mitomycin C (10µg/mL) for 2 hours (Table 4.3).

**Table 4.3 WoundMaker™ scratch assay experimental design**

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Drug Conc.</th>
<th>n of Drug Only</th>
<th>n of Drug + Mitomycin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle H₂O</td>
<td>Volume of 10 µl/ml</td>
<td>6</td>
<td>11/12</td>
</tr>
<tr>
<td>Vehicle DMSO</td>
<td>10 µM</td>
<td>6</td>
<td>11/12</td>
</tr>
<tr>
<td>DZNEP</td>
<td>1 µM</td>
<td>5/6</td>
<td>11/12</td>
</tr>
<tr>
<td>DZA</td>
<td>10 µM</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>5-Azacytidine</td>
<td>10 µM</td>
<td>5/6</td>
<td>11/12</td>
</tr>
<tr>
<td>Elvitegravir</td>
<td>10 µM</td>
<td>5/6</td>
<td>11/12</td>
</tr>
<tr>
<td>Raltegravir</td>
<td>100 µM</td>
<td>6</td>
<td>11/12</td>
</tr>
<tr>
<td>Azidothymidine</td>
<td>10 µM</td>
<td>6</td>
<td>11/12</td>
</tr>
</tbody>
</table>

**4.2.5 Transformation Assay**

To study the *in vitro* anti-tumorigenesis effect of integrase and reverse transcriptase inhibitors and the tumorigenesis potential of dysregulated epigenome induce TE expression through SAM cycle inhibitors and DNA methyltransferase inhibition we utilized transformation assays. Transformation assays require cell growth
in an anchorage independent way. Non-transformed cells require adequate and appropriate cell-matrix interactions, otherwise the undergo anoikis [27]. Anchorage independent growth is a hallmark of transformation and is a stringent in vitro assay for detecting nonmalignant to malignant transformation. All transformation assays were performed according to a previously established protocol [28]. Briefly, the protocol requires three layers of cell culture media and agarose in a 6 well cell culture plate. The bottom layer is 2ml of 0.6% 2-hydroxyethyl agarose solution (Sigma, A4018) in media. The middle layer is a cell-containing layer with 0.3% 2-hydroxyethyl agarose solution in media. The last layer is a feeder layer of 0.3% 2-hydroxyethyl agarose solution in media to provide extended access to nutrients for cells. In treatment wells, the appropriate drug dose is added to the cell-containing layer and feeder layer and is further explained below. Feeder layers are added once a week. It is recommended that images be taken between one and two weeks of growth.

Here we plated 2 x 10^4 cells per well in four 6-well plates for each cell line, MCF10A and MCF10DCIS, and treated with respective drug in concentrations (Table 4.4).

**Table 4.4 Transformation assay experimental design**

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Drug Concentration</th>
<th>n of MCF10A</th>
<th>n of MCF10DCIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle H₂O</td>
<td>Volume of 10 µl/ml</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>Vehicle DMSO</td>
<td>10 µM</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DZNEP</td>
<td>1 µM</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>DZA</td>
<td>10 µM</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Raltegravir</td>
<td>100 µM</td>
<td>NA</td>
<td>3</td>
</tr>
<tr>
<td>Elvitegravir</td>
<td>10 µM</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
After 10 days a grid template was generated, images were taken with a 10mm lens microscope camera at the west, north, east, south, and middle of each well. Colonies that had a diameter of 50 µm or more were counted using ImageJ software.

4.3 Results

4.3.1 Cell Viability & Density Dose-Response Curves

A dose-response curve on viability and density was conducted for each drug to determine the minimal effective dose in our cell culture assays using viability and density assays. SAM cycle inhibitors DZA and DZNEP doses chosen were 10 µM and 1 µM, respectively. At the 10 µM dose DZA had no change in cell viability compared to control in both MCF10A and MCF10DCIS cell lines (Figure 4.2A & C). However, at 10 µM, DZA in MCF10A cell line had increased cell density when compared to control (Figure 4.2B). At the 1 µM dose DZNEP had a slight decrease in cell viability for both cell lines, when compared to control, and a deficit in both cell lines for cell density (Figure 4.2). The DNMT inhibitor AZA dose chosen was 10 µM. Little to no difference was detected for 1 µM of AZA treatment when compared to control (Figure 4.3). In contrast, 100 µM treatment with AZA resulted in large deficits in cell density and a large decrease in cell viability (Figure 4.3 A & C). 10 µM treatment with AZA resulted in no change of viability with MCF10A and a slight decrease in viability in MCF10DCIS when compared to control (Figure 4.3A & C). However, both cell lines experienced a deficit in density (Figure 4.3B & D). Integrase inhibitors ELV and RAL doses chosen were 10 µM and 100 µM respectively. ELV at 1 µM showed no real difference in cell viability and highly variable results for density for both cell lines. At the 100 µM dose of ELV MCF10DCIS cells had a decrease in cell viability and both MCF10A and MCF10DCIS had large deficits in cell
density when compared to vehicle (Figure 4.4). Overall, RAL had little change in viability and density when compared to control regardless of dose. Reverse transcriptase inhibitor AZT dose chosen was 10 µM. There was little effect on cell viability when compared to control for both MCF10A and MCF10DCIS cells line regardless of the treatment dose (Figure 4.5). Cell density had varying results for treatment doses. At 1 µM MCF10DCIS cells had a much greater density when compared to control (Figure 4.5D). Conversely, at 100 µM MCF10DCIS cells experience slight decrease in cell density (Figure 4.5D). MCF10A cells had a slight increase in density at 1 µM and a slight decrease in cell density at 100 µM (Figure 4.5B).

Figure 4.2 Initial Dose-Response Curve for SAM Cycle Inhibitors. A. MCF10A cell viability for DZA and DZNEP doses. B. MCF10A cell density for DZA and DZNEP doses. C. MCF10DCIS cell viability for DZA and DZNEP doses. D. MCF10DCIS cell density for DZA and DZNEP doses. Red star indicates doses chosen for future studies, not an indicator of significance. (unpublished)
Figure 4.3 Initial Dose-Response Curve for DNA Methyltransferase Inhibitor. A. MCF10A cell viability for AZA doses. B. MCF10A cell density for AZA doses. C. MCF10DCIS cell viability for AZA doses. D. MCF10DCIS cell density for AZA doses. Red star indicates doses chosen for future studies, not an indicator of significance. (unpublished)

Figure 4.4 Initial Dose-Response Curve for Reverse Transcriptase Inhibitor. A. MCF10A cell viability for AZT doses. B. MCF10A cell density for AZT doses. C. MCF10DCIS cell viability for AZT doses. D. MCF10DCIS cell density for AZT doses. Red star indicates doses chosen for future studies, not an indicator of significance. (unpublished)
4.3.2 Scratch Assay

4.3.2a Pilot Study

To measure migration and growth during treatments to induce TE expression or reduce TE activity scratch assays were performed. No significant results were found for any of the treatments in either cell line until the 24-hour time point. At 24 hr post-stratch, AZA and DZNEP had significantly less wound healed when compared to their vehicle controls in MCF10A cell line. Nonsignificant increases of wound healing were seen in MCF10DCIS cell lines for AZT, ELV, DZA, and AZA.
Figure 4.6 Pilot Study of Wound Assay for MCF10A cell line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Reverse transcriptase inhibitor AZT. (unpublished)

Figure 4.7 Pilot Study of Wound Assay for MCF10DCIS cell line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Reverse transcriptase inhibitor AZT. (unpublished)
4.3.2b WoundMaker™

To confirm and extend pilot study results, we performed WoundMaker™ scratch assays with drug treatment only. For MCF10A cells at 6hrs after the wound was generated and treatment was applied only AZA had significant results when compared to control with a decrease in the amount of wound healed (Figure 4.8B). AZA continued to have significantly less amount of wound healed compared to vehicle control at each time point. DZNEP also had significantly less wound healed when compared to vehicle control for all time points except 6hrs (Figure 4.8A & B). ELV and DZA had significantly less wound healed at 12 hours when compared to control (Figure 4.8A & C). For MCF10DCIS cell line no significant difference was determined for any drug treatment at 6-hour timepoint. At all other timepoints only AZA had significantly less wound healed when compared to control (Figure 4.9B).

To determine migration only effects we performed WoundMaker™ scratch assay with drug treatment + mitomycin treatment. In MCF10A cells, AZA had significantly less migration at each timepoint when compared to vehicle control (Figure 4.10B). AZT had nonsignificant increases in migration for the 6-hour and 12-hour timepoint (Figure 4.10E). In MCF10DCIS cells, again only AZA had significantly less migration when compared to control at each timepoint (Figure 4.11B). Starting from 12 hours through 18 hours, RAL had significantly less migration when compared to control (Figure 4.11D).
Figure 4.8 WoundMaker™ Scratch Assay for MCF10A cell line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Integrase inhibitor RAL. E. Reverse transcriptase inhibitor AZT. (unpublished)
Figure 4.9  WoundMaker™ Scratch Assay for MCF10DCIS cell line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Integrase inhibitor RAL. E. Reverse transcriptase inhibitor AZT. (unpublished)
Figure 4.10  WoundMaker™ Scratch Assay Migration Only for MCF10A Cell Line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Integrase inhibitor RAL. E. Reverse transcriptase inhibitor AZT. (unpublished)
Figure 4.11  WoundMaker™ Scratch Assay Migration Only for MCF10DCIS Cell Line. A. SAM cycle inhibitors DZA and DZNEP. B. DNMTi AZA. C. Integrase inhibitor ELV. D. Integrase inhibitor RAL. E. Reverse transcriptase inhibitor AZT. (unpublished)
4.3.5 Transformation Assay

To determine anchorage-independent growth during treatments to induce TE expression or reduce TE activity, transformation assays were performed. No significant differences were found for any of the treatments when compared to control for MCF10A cells (Figure 4.12). In MCF10DCIS cells, DZNEP had significantly fewer colonies with a diameter of 50 µm or more when compared to control (Figure 4.13).

![MCF10A Transformation Assay](image1)

**Figure 4.12 Transformation Assay for MCF10A Cell Line.** A. SAM cycle inhibitors DZA and DZNEP. B. Integrase inhibitor ELV. (unpublished)

![MCF10DCIS Transformation Assay](image2)

**Figure 4.13 Transformation Assay for MCF10DCIS Cell Line.** A. SAM cycle inhibitors DZA and DZNEP. B. Integrase inhibitor ELV. (unpublished)
**Table 4.5 Results Summary for MCF10A.** Outcomes are based on treatment compared to vehicle controls. ND no significant difference, NA not available, ↑ increased percent of wound healed, ↓ decrease percent of wound healed.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Wound Healing (Pilot)</th>
<th>Wound Healing (WoundMaker®)</th>
<th>Wound Healing (WoundMaker® + Mitomycin C)</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DZA</td>
<td>ND</td>
<td>12hr↓</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>DZNEP</td>
<td>24hr↓</td>
<td>≥12hr↓</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>AZA</td>
<td>24hr↓</td>
<td>↓</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>ELV</td>
<td>ND</td>
<td>12hr↓</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>RAL</td>
<td>NA</td>
<td>ND</td>
<td>ND</td>
<td>NA</td>
</tr>
<tr>
<td>AZT</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table 4.6 Results Summary for MCF10DCIS. Outcomes are based on treatment compared to vehicle controls. ND no significant difference, NA not available, ↑ increased percent of wound healed, ↓ decrease percent of wound healed.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Wound Healing (Pilot)</th>
<th>Wound Healing (WoundMaker®)</th>
<th>Wound Healing (WoundMaker® + Mitomycin C)</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DZA</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>DZNEP</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>AZA</td>
<td>ND</td>
<td>≥12hr↓</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>ELV</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>RAL</td>
<td>NA</td>
<td>ND</td>
<td>12hr, 18hr ↓</td>
<td>ND</td>
</tr>
<tr>
<td>AZT</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>NA</td>
</tr>
</tbody>
</table>

4.4 Discussion

These studies tested the role of retrotransposons in breast cancer progression using pharmacological approaches and classic invasive and metastasis analysis assays. Counter to our expected outcomes for DNMTi AZA, the drug consistently caused significantly more cell death and decreased invasiveness when compared to control. Additionally, SAM cycle inhibitor DZNEP consistently had significantly less migration in MCF10A cells and less anchorage independent growth in MCF10DCIS cells. While initially unexpected, these data corroborate recently published research. In 2015, Hurst and Magiorkinis speculated that endogenous retroviruses activate the innate immune response [29]. Later studies confirmed this and further determined the inhibition of DNA methylation induced an interferon response due to endogenous retroviral expression
I believe future studies with smaller doses of our DNA methylation disruptors may further determine roles of retrotransposon expression in breast cancer progression by allowing an increase in retrotransposon expression below the threshold necessary to induce the innate immune response.

The nonsignificant increase in migration for MCF10A cells, when treated with the reverse transcriptase AZT, is worth further exploration as it may suggest alternative activities of RNA transcripts for retrotransposons. Conversely, the integrase inhibitor RAL’s anti-migratory results may suggest a decrease in genomic destabilization prevents metastatic progression. A decrease in genomic instability upon inhibition would support the classic transposon activity in cancer model. Additional soft agar assay experiments should be performed to further address trends not significant.

4.5 References

Chapter Five:
Perspectives

5.1 Pipeline for the Common Clinical RNA-Seq Dataset

With the introduction and advancement of next-generation sequencing, biologists and clinicians have made an enormous amount of discoveries in a short amount of time when compared to classic sequencing approaches. These next-generation sequencing tools produce massive amounts of data and require special attention to their analysis because of the potential effects of biological sample preparation, method of sequencing, and questions of interest. An ideal translational experiment would involve fresh, or freshly frozen, tissue samples micro-dissected from non-target tissues, immediately prepped and sequenced. However, circumstances for clinicians and many biologists working with specimens from human patients must rely on preservation methods, such as, formalin fixed paraffin embedded (FFPE) tissue. These methods of preservation have been known to increase nucleotide degradation [1-3].

Presented in this thesis, the analysis and comparison of common bioinformatics tools study proved successful in the generation of a pipeline that is well-suited for low-quality, short nucleotide sequence reads data. As described in Chapter 2, this study started by examining two common alignment tools, HISAT2 and STAR, to analyze RNA-Seq datasets derived from FFPE tissue samples. We further developed the pipeline by comparing the two most common and robust differential expression analysis tools, edgeR and DESeq2. We were able to discover that the version of HISAT2 used in the comparison
had significantly less reads mapped to the human genome. Furthermore, the reads aligned by HISAT2, were assigned to pseudogene loci instead of genes, which may indicate compromised alignment fidelity in preserved specimens that degrade nucleotides. During the comparison of the differential expression tools, DESeq2 and edgeR programs yielded similar results although it appears that edgeR’s estimateDisp function had increased FDR correction for datasets with 12 or more replicates.

Many clinical and experimental laboratories often have their RNA-Seq samples brought to a core or commercial analytics company for processing and analysis without knowing the pipeline utilized on their data. Equivalently, many bioinformaticians process RNA-seq read output without knowledge or understanding of the experimental preparations executed at the clinician’s or experimentalist’s laboratory that can bias sequencing output. Our study provides evidence that no single bioinformatics tool is appropriate to apply to all experiments. Furthermore, parameters setting contribute to outcomes for both mapping to a reference genome and measuring differential expression. The studies support for STAR aligner to obtain the most robust, accurate alignment for RNA-Seq data generated from FFPE samples. Furthermore, we recommend either DESeq2 or edgeR if estimateDisp function is applied for studies with 12 or more replicates and edgeR without estimateDisp function applied for studies with less than 12 replicates.

5.2 Novel In Silico Method for the Discovery of TE Expression in Breast Cancer Progression

For decades it has been widely accepted that early pre-malignant breast cancer is a precursor to the invasive form. Yet, there is little knowledge as to distinguish the cohort of patients diagnosed with pre-malignant breast cancer that never progress to the invasive
form. Few advances have been made in the evolutionary bottleneck theory of “one gene mutation at a time” until a selected clone escapes the basement membrane. As previously mentioned, a recent study using advanced single cell spatially resolved next-generation sequencing provided evidence that clonal selection occurs before tumor cells escape the basement membrane, subclones are derived from the same or a similar parent population, and multiclonal migration occurs. Early heterogeneity and multiclonal escape suggest there could be additional factors involved in basement membrane escape.

After optimizing the bioinformatics pipeline tools, breast cancer progression RNA-Seq data were analyzed to determine the TE expression signature. This analysis revealed that the LTR retrotransposons family, ERV expression was the most differentially expressed TE at each stage of breast cancer when compared to normal. Additionally, DCIS had the highest percentage of TEs differential expressed at ERV genomic regions compared to the percent of DE ERVs in other stages.

Interestingly, when gene coding regions were analyzed for differential expression and gene enrichment analysis with gene ontology terms the outputs revealed interferon responses and antiviral activity starting at the DCIS stage. These data corroborate the higher percentage of ERVs differentially expressed in DCIS. Furthermore, our transcript analysis revealed progressive diversification of GO terms associated with metastasis from atypia to IDC. This in silico study provides a foundation of evidence for TE, specifically ERV, involvement in breast cancer progression from pre-malignant to malignancy. However, replicate findings in silico analysis of RNA-seq data from other independently collected specimens and experimental evidence are needed to substantiate our discoveries.
5.3 Regulating Retrotransposon Expression and Activity Effects on MCF10A and MCFDCIS Cell Line Models

This study provides convincing introductory data for a role of retrotransposons in breast cancer progression. Retrotransposition induced genetic instability is a hallmark of advanced cancers [4]. HERV-K particles have been discovered in the plasma of patients diagnosed with invasive breast cancer. An important regulator and repressor of retrotransposon expression is DNA methylation and other epigenetic mechanisms. Previous microarray studies postulate epigenetic dysregulation at retrotransposon loci in cancerous tissue when compared to adjacent normal tissue. This study examines the role of retrotransposons in the establishment of metastatic phenotypes of breast cancer using MCF10A and MCF10DCIS cell lines as a model of progression.

Using a pharmacological approach, the contribution of retrotransposon expression to metastatic phenotypes was examined, such as increased migration and anchorage independent growth, by treating with DNA methylation disruptors and retroviral reverse transcriptase and integrase inhibitors. Our results indicate DNA methylation inhibitor 5-Azacytidine (AZA) consistently significantly decreased and retarded cell migration and growth. These results are not caused by toxicity because cell viability assays showed MCF10DCIS cells with decreased viability only at doses higher than used for these experiments. While these findings were unexpected; they corroborate a few recent studies. In 2015, a paper reported the possibility that ERVs contributes to an increase in the innate immune response. That same year it was found that AZA induced an interferon response in cancer by depressing ERV expression. These studies have provided data supporting the possibility of using AZA, and other epigenetic deregulators, as a therapeutic option. Within the last year several studies have started testing known
epigenetic deregulators to induce ERV expression beyond a threshold of tolerance, for a “viral mimicry”, and initiate an immunotherapy response [5-7]. It’s important to note that these studies are providing evidence of an ERV expression threshold to induce immune response. These do not discount or discredit the diverse roles TEs play in oncogenesis as determined from prior studies [8-16]. The significant mitigation of migration on MCF10DCIS cells when treated with an integrase inhibitor RAL corroborates the contribution of insertional mutagenesis to advanced cancers. However, further pharmacokinetic studies must be performed to determine ERV activity and ensure the significant decrease isn’t due to cellular toxicity. Many of the drugs at the doses chosen did not produce significant differences. Additional dose-dependent response curve studies should be done to explore effective concentrations.

A major limitation to this study is in the time of exposure during the wound healing assays. Drug induced expression of transposable element insertional mutagenesis may require weeks of treatment exposure to ensure oncogenic clonal selection occurs. Likewise, pretreatment may also provide insight into how TEs drive transformation, whether through classical TE mutagenesis or alternative TE activities, like alternative promoter exaptation driving oncogenes. To overcome this limitation further experiments with pretreatments for 1 to 2 weeks in drug may be performed. Additional limitations include the current spectrum of doses tested, the number of transformation assays completed, the lack of RNA-seq data for cell line models concordance with clinical RNA-seq data, and the lack of direct experimental evidence to suggest transposable activity is driving transformation.
5.4 Future Directions

5.4.1 Transposable Element Expression in Breast Cancer Cell Lines

The use of MCF10A cell line and its derivations of cancer cell lines as a model for breast cancer progression has been extensively studied [17-20]. However, little is known as to its concordance of transposable element expression during MCF10A transformation and derived cancer lines with clinical samples of breast cancer progression. Therefore, it is imperative to determine the TE expression profile of the specific cell lines we use as models of oncogenic phenotypes during TE dysregulation and TE activity inhibition.

5.4.2 Activity of Transposable Elements

The use of pharmaceuticals within a dose range known to induce TE expression or inhibit TE activity to measure metastatic potential is necessary. However, it is imperative these studies be done in conjunction with studies that can measure activity of transposable elements. One such approach is the reverse transcriptase activity assay. Briefly, purified bacteriophage MS2 RNA are incubated with cell lysate aliquots. Mixtures are exposed to PCR amplification using MS2-specific primer pairs to determine if MS2 cDNA sequences were synthesized. The presence of amplified MS2 cDNA is a measure of reverse transcriptase activity. The assay is roughly based on the protocol established by Voisset et al., 2001 [21]. Versions of the assay have previously described endogenous reverse transcriptase activity in preimplantation embryos and cancer [22, 23].

5.4.3 Patient-Derived Xenograft Models

Patient-derived tumor xenograft (PDX) mouse models offer a unique translational prospect for breast cancer progression research. These models seem to be quite similar to
their donor, retaining important tumor histomorphology, imaging and gene expression characteristics of the donor [24-26]. Furthermore, they have been used to study predictive clinical outcomes with regards to drug efficacy, biomarker analysis, and patient outcomes, reviewed here [27], with high concordance between patients and their PDX models. However, there is a lack of published studies with PDX models recapitulating breast cancer progression from non-obligate in situ carcinomas to metastasis. Not only would these studies provide valuable insight into why only 1/3 of patients progress into invasive breast cancer they may also indicate biomarkers or factors for invasiveness. Furthermore, currently there are no studies revealing the transposable element or non-coding RNA expression profile for breast cancer PDX models.

### 5.5 Final Thoughts

The idea that TEs were fundamental and continue to contribute to the evolution of regulating gene networks started with the pioneering work of Dr. Barbara McClintock. The selfish replicative nature of TEs has predisposed them to the co-option of host gene regulation. However, this co-option of TEs is widely seen as a double-edged sword. As described in Chapter 3, there is a large body of evidence for the diversity of domesticated TE mechanisms employed by whole organism, tissue-specific, and cellular systems. Contrarily, strong evidence linking aberrant TE activity to disease states, such as, cancer, ageing, neurological disorders, and autoimmunity is increasing. The double-edge nature of parasitic elements becoming integral components of many host functions presses for more granular experimental evidence of TE activity at the individual element level. We are in the gene-targeted therapy revolution. Current technologies, such as CRISPR-Cas systems, are becoming increasingly more accurate at genetic and RNA manipulation. As
technologies advance toward functionally testing non-coding and repeat regions of the genome, it will provide greater, much needed, insight into the roles of TEs in disease states.

5.6 References