5-2015

GENE SET ANALYSIS OF POST-LACTATIONAL MAMMARY GLAND INVOLUTION GENE SIGNATURES IN INFLAMMATORY AND TRIPLE NEGATIVE BREAST CANCER

Arvind Bambhroliya

GENE SET ANALYSIS OF POST-LACTATIONAL MAMMARY GLAND INVOLUTION GENE SIGNATURES IN INFLAMMATORY AND TRIPLE NEGATIVE BREAST CANCER

by

Arvind Bambhroliya, M.B.B.S., M.P.H.

APPROVED:

_____
Wendy Woodward, M.D. Ph.D., Supervisory Professor

_____
Abenaa Brewster, M.D

_____
Arvind Rao, Ph.D.

_____
Erik Sulman, M.D., Ph.D.

_____
Naoto T. Ueno, M.D., Ph.D.

_____
Ralf Krahe, Ph.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

GENE SET ANALYSIS OF POST-LACTATIONAL MAMMARY GLAND INVOLUTION GENE SIGNATURES IN INFLAMMATORY AND TRIPLE NEGATIVE BREAST CANCER

A

THESIS

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M. D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

by

Arvind Bambhroliya, M.B.B.S., M.P.H.

May, 2015

Acknowledgements

I am indebted to many people for helping me attain my M.S. First, I would like to thank my advisor Dr. Wendy Woodward and my mentor Dr. Abenaa Brewster for their invaluable guidance and support during my thesis work. Without their generous help, I believe, I could not able to accomplish all the milestones required to complete the Master of Science degree. I am also greatly indebted to Drs. Arvind Rao, Erik Sulman, Naoto Ueno, and Ralf Krahe for their participation in my committee meetings and their suggestions to improve my thesis work. I am also very thankful to Drs. Steven Van Laere and François Bertucci who provided the data and for their suggestions to improve my thesis work. Additionally, I would like to thank GSBS members, especially Dr. Bill Mattox, Dr. Vicky Knutson, Dr. Thomas Goka, Brenda Gaughan, Carol Helton, Joy Lademora, Bunny Perez, and Lily D'Agostino for their help and guidance in completing my degree. I also want to thank current and former members of Dr. Wendy Woodward's lab, especially Dr. Rachel Atkinson and Dr. Lara Carolina Alvarez de Lacerda Landry for their support in preparation of my thesis project. Lastly, I would like to express my gratitude towards the UT Graduate School of Biomedical Sciences for giving me this unique opportunity to pursue the master program in biomedical sciences.

GENE SET ANALYSIS OF POST-LACTATIONAL MAMMARY GLAND INVOLUTION GENE SIGNATURES IN INFLAMMATORY AND TRIPLE NEGATIVE BREAST CANCER

Arvind Bambhroliya, M.B.B.S., M.P.H.

Supervisory Professor: Wendy Woodward, M.D. Ph.D.

With each cycle of pregnancy and lactation, the mammary gland undergoes dramatic changes. Clarkson et al. (2004) and Stein et al. (2004) conducted a detailed analysis of these changes in the mouse mammary gland. Their results showed that tissue remodeling during post-lactational involution period mimics wound healing and tumorigenesis like pathological conditions. This indicates that post-lactational involution may create the microenvironment that initiates the development of precancerous mammary cells and promotes the progression of precancerous cells into cancer cells. Moreover, epidemiological studies have found that triple-negative (TN) breast cancer (BC) is associated with lower frequency and duration of breastfeeding compared to non-TN BC. Inflammatory breast cancer (IBC) accounts for approximately 1 - 5 % of all breast cancers and has significantly lower 5-year survival rates than those for non-IBC. Approximately75% of the IBC samples belongs to the more aggressive subtypes like TN, HER2-enriched, and luminal B breast cancer. Similar to TN BC, TN IBC had been found to have significantly lower frequency and duration of breastfeeding compared to non-TN IBC. The higher frequency of aggressive forms of breast cancer in IBC and the association of TN BC and TN IBC with lower duration of breastfeeding indicate that molecular changes in the mammary gland during post-lactational involution might play roles in development of IBC and TN BC. We *hypothesized* that gene expression signatures of abrupt post-lactational mammary gland involution correlate with IBC, TN IBC and TN non-IBC. We utilized gene expression data on multiple time-points of the mouse mammary gland development after abrupt weaning to create post-lactational mammary gland involution gene signatures using time-series cluster analysis. Using these gene expression signatures we performed gene set enrichment analysis

(GSEA) on human breast cancer gene expression data to identify specific gene expression signatures that are enriched in IBC compared to non-IBC and in TN compared to non-TN in IBC and non-IBC groups. We divided samples in training and validation sets to validate the results of GSEA. We identified 10 statistically significant time-varied gene expression patterns of post-lactational mammary gland involution. We found significant enrichment of one post-lactational involution gene signature in IBC compared to non-IBC. This enriched signature represents genes showing initial up-regulation and later down-regulation during the involution process and significant overlap with genes up-regulated in vascular smooth muscle cells (VSMC) by c-Jun N-terminal protein kinase (*JNK1*). We identified 3 genes – Involucrin (*IVL*), Cluster of Differentiation 79B (*CD79B*), and leptin (*LEP*) – that were significantly enriched in IBC compared to non-IBC in both training and validation data sets and that are up-regulated in VSMC by *JNK1*. This data indicate that these genes might be playing roles in IBC development.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Background on Post-Natal Mammary Gland Development

The mammary gland undergoes various stages of development during embryonic, pubertal, reproductive, and post reproductive periods of life. Involution is a term that has been described as the reverse of development [1] and, in the context of mammary gland development, describes the different forms of regression of mammary gland. Regression of the gland as a result of senility is described as a senile involution while post-lactational involution is the process by which the mammary gland returns from its lactating state to its non-lactating state [2]. It can occur in a gradual form (after gradual decrease in suckling) or in an abrupt form (after sudden cessation of milking or after sudden weaning of the young) [2]. The abrupt weaning of the young may mimic the shorter duration of breast-feeding in humans. Post-lactational involution of the mammary gland is a complex multistage process which is characterized by regression of mammary gland epithelium through apoptosis and tissue remodeling [3]. Our knowledge of these stages is derived primarily from studies performed in a mouse model system, providing insight into the biology of post-lactational involution in the human mammary gland. Research on involution suggests that there are two phases of involution in mice [3]. The first phase lasts for about 48 hours and is reversible and characterized by apoptosis and alveolar cells detachment but no major morphological changes [3]. The second phase starts at 48 hours and is irreversible and characterized by collapse of alveoli, breakdown of extracellular matrix and activation of proteases [3]. By six days, most of secretary epithelium is removed and replaced by adipocytes [3]. Morphologically, mammary gland at the end of involution process looks similar to that of a virgin state mammary gland. Clarkson et al. [4] and Stein et al. [5] conducted gene expression profiling studies of these changes in the mouse mammary gland with the induction of forced weaning at the peak of lactation. Results of these two studies highlight distinct molecular characteristics between the virgin, pregnant, lactating and involuting states of the mammary gland. Various signaling pathways have been identified that regulate the transition from lactation to involution and the STAT family of proteins has been found to play a major role in this transition. These

results also show that tissue remodeling during post-lactational involution mimics wound healing and tumorigenesis like pathological conditions [6]. Inflammation and would healing responses have been found to be associated with tumor growth and progression [7]. An in-vitro study by McDaniel et al [8] found that mammary gland matrix from rats during the post-lactational involution stage promotes tumor cell invasion and metastasis. A study by Blanchard et al [9] identified genes that showed differential expression during the first day of involution in the mouse using microarray analysis and deregulation (down regulation) in breast cancer cell lines and breast tumor tissues. Stein et al [10] identified that genes showing differential expression at day 3 of mammary gland involution distinguished high and low metastatic breast cancer and these genes were associated with copper ion metabolism and with HIF-1 promoter binding sites. Lyons et al showed that the involuting mammary microenvironment promotes the invasiveness of ductal carcinoma in-situ (DCIS) cells and invasive tumors formed from DCIS cells were dominantly triple negative using a mouse model [11]. Also it has been studied that both luminal and myoepithelial lineages in mammary gland contain long-lived stem cells and pregnancy leads to a transient 11-fold increase in these mammary stem cells [12, 13]. These results indicate that the deregulated involution may create the microenvironment that promotes tumor growth and progression.

1.2 Background on Inflammatory Breast Cancer and Triple Negative Breast Cancer

Breast cancer (BC) is the most common cancer and second leading cause of cancer mortality among women in the United States [14]. Breast cancer has been classified into five subtypes through gene expression profiling [15] with significant differences in epidemiological risk factors, therapeutic approaches and prognosis among subtypes [16]. These subtypes include luminal A, luminal B, human epidermal growth factor receptor 2 (HER2) overexpressing, basal-like, and claudin-low tumor subtypes [15, 17]. Clinically, a panel of immunohistochemical biomarkers is used as a surrogate for gene expression profiling to classify breast cancer subtypes. These biomarkers are estrogen receptor (ER), progesterone receptor (PR) and HER2. Breast cancer with absence of ER expression, PR expression and HER2 overexpression is termed as triple-negative breast cancer. Triple negative (TN) breast cancer is

enriched in the basal-like breast cancer subtype which was identified through gene expression profiling [18]. It accounts for approximate 15% of all breast cancer incidents and has the lowest survival rates among all subtypes of breast cancer [19]. Inflammatory breast cancer (IBC) is a distinct type of breast cancer which is characterized clinically by its rapid onset, erythematous, and edematous presentation of the breast and pathologically by tumor emboli in the dermal lymphatics [20]. It accounts for approximately 1 - 5 % of all breast cancers but at all stages has significantly lower 5-year survival rates than that for non-IBC [20-22]. Gene expression profiling has also identified similar molecular subtypes in IBC as those in non-IBC [23-25] and all molecular subtypes of IBC have been found to have a poor prognosis [26]. One of reproductive risk factors that have been found to be associated with increased risk of triple negative breast cancer is little-to-no breast feeding while having at least one pregnancy have been found to be protective for luminal A breast cancer [27]. Similar to TN BC, TN IBC has been found to have significantly lower frequency and duration of breastfeeding compared to non-TN IBC (ER or PR positive or Her2neu positive) [28].

1.3 Central Hypothesis and Specific Aims

Studies in mice indicate that post-lactational involution after abrupt weaning mimics wound healing and tumorigenesis like pathological conditions [6] and involuting mammary microenvironment promotes the invasiveness of ductal carcinoma in-situ (DCIS) cells [11]. Epidemiological studies indicate that lower duration of breast feeding is associated with increased risk of developing TN BC [27] and TN IBC [28]. Findings from these animal and epidemiological studies suggest that changes in the mammary gland during abrupt post-lactational involution might specifically play roles in the development of TN BC and TN IBC. Also, recent studies in mice suggest that mammary gland contains long-lived stem cells residing at the top of the hierarchy which undergoes extensive expansion during several cycles of pregnancy. Abrupt involution could leave persistent likely receptor negative stem cells increasing the chance of an initiating TN BC and TN IBC event. We *hypothesize* that gene expression signatures of the abrupt post-lactational involution stage of mammary gland development are enriched in IBC and TN breast cancer. To test our underlying hypothesis, we proposed three specific aims.

1. To identify gene expression signatures for the abrupt post-lactational involution stage of mammary gland development using gene expression data from post-natal mouse mammary gland development
2. To evaluate whether gene expression signatures for the abrupt post-lactational involution stage of mammary gland development are enriched in IBC compared with non-IBC
3. To evaluate whether gene expression signatures for the abrupt post-lactational involution stage of mammary gland development are enriched in TN breast cancer compared with non-TN breast cancer

2 DEVELOPMENT OF INVOLUTION SPECIFIC GENE SIGNATURES

2.1 Methods

Gene Expression Dataset on Mouse Mammary Gland Involution

Two major studies by Clarkson et al. [4] and Stein et al. [5] conducted time-series microarray experiments on mammary gland at various stages of pregnancy, lactation and involution and identified gene expression changes in post-natal mammary gland development. In the study by Clarkson et al. [4] , genome wide expression profiles were measured with Affymetrix GeneChip MGU74ver2a arrays at the 12 stages of adult mouse mammary gland development (virgin, 8 week; pregnancy days 5, 10 and 15; lactation days 0, 5 and 10; and involution hours 12, 24, 48, 72 and 96 after forced weaning). In the study by Stein et al. [5], gene expression profiles were measured at the 17 stages of adult mouse mammary gland development (virgin, 10 and 12 weeks; pregnancy days 1, 2, 3, 8.5, 12.5, 14.5 and 17.5; lactation days 1, 3 and 7; and involution days 1, 2, 3, 4 and 20 after forced weaning). The gene expression data for both studies can be downloaded from the webpage of the Mammary Apoptosis and Development Group at the University of Cambridge (http://www.path.cam.ac.uk/~madgroup/) and from the NCBI Gene Expression Omnibus (GSE12247) (http://www.ncbi.nlm.nih.gov/geo/).

Preprocessing of Gene Expression Data

Raw gene expression profiles were preprocessed using GCRMA analysis [29] with quantile normalization and probeset-level signals were summarized in log base 2 scale. We selected a custom Chip Definition File (CDF) MGU74Av2_Mm_ENTREZG version 18 for more accurate probe mapping to genome [30]. There are 7952 probe sets with a CDF MGU74Av2_Mm_ENTREZG version 18 representing 7882 genes as per the annotation database available for a CDF MGU74Av2_Mm_ENTREZG version 18 at the BrainArray. After preprocessing gene expression data, further analysis were conducted using information available for 7882 probe sets representing 7882 genes with one probe set-one gene relationship.

Identification of Differentially Expressed Genes across Time Points

Tests of differences in expression were performed with the limma package (version 3.22.1) [31] from the Bioconductor project. The limma package uses the moderated *t*-statistic. A total of 1055 genes were identified as the most significantly differentially expression genes across time points (q-value < 0.05) with greater than two-fold changes in at least 1 pair comparing time points.

Clustering Analysis of Time-Series Expression Data

Ernst et al presented an algorithm specifically designed for clustering time-series expression data [32] and developed the Short Time-series Expression Miner (STEM) program for analysis of time-series gene expression data [33]. The STEM program was obtained from the website of the Systems Biology Group of the School of Computer Science of Carnegie Mellon University (http://gene.ml.cmu.edu/stem/). The STEM program works by first defining a set of representative model profiles corresponding to possible patterns of gene expression across the conditions that are examined in the experiment. Each gene is, then, assigned to the closest profile on the basis of correlation coefficients. The expected number of genes for each profile is also computed using random permutation, renormalization and assignment of original values for each gene to profiles with over 500 repeated permutations. This serves as a basis for the calculation of the statistical significance of each profile. Statistically significant profiles represent the dominant expression profiles in the data set. The parameters used for STEM clustering were set at 50 for a maximum number of model profiles, 3 for a maximum unit change between time points and 0.7 for a minimum correlation for clustering similar profiles. Significant expression profiles were identified with a false discovery rate (FDR) of 0.05.

Ontology Analysis of Significant Clusters

An ontology-based analysis was performed on genes of significant clusters identified through the STEM program. We used gene ontology (GO) annotations for mus musculus gene products (filename:

gene_association.mgi.gz) available from Mouse Genome Informatics (www.informatics.jax.org). Enrichment analysis for GO annotations was performed using a hypergeometric distribution in the STEM program and multiple hypothesis correction was done using a randomization test. *P* values were corrected with 500 randomizations and were considered significant with an FDR of <0.05 for gene-ontology enrichment analysis with the STEM program.

2.2 Results

Differentially Expressed Genes across Time Points

We used the *limma* package [31] from the Bioconductor project to identify differentially expressed genes across time points from lactation day 10 to involution day 4. We identified 1,055 differentially expressed genes across time points from lactation day 10 to involution day 4 in the data from Clarkson et al. (2004). To verify our results, we conducted an analysis for differential expressed genes using data from Stein et al. (2004) and identified 2,567 genes differentially expressed across time points from lactation day 7 to involution day 20. 79% of genes identified as differentially expressed in the data from Clarkson et al. were also again identified as differentially expressed in the data from Stein et al. (2004). Figure 1 shows the Venn diagram of the overlap and discrepancies between genes differentially expressed in Clarkson et al. (2004) and Stein et al. (2004).

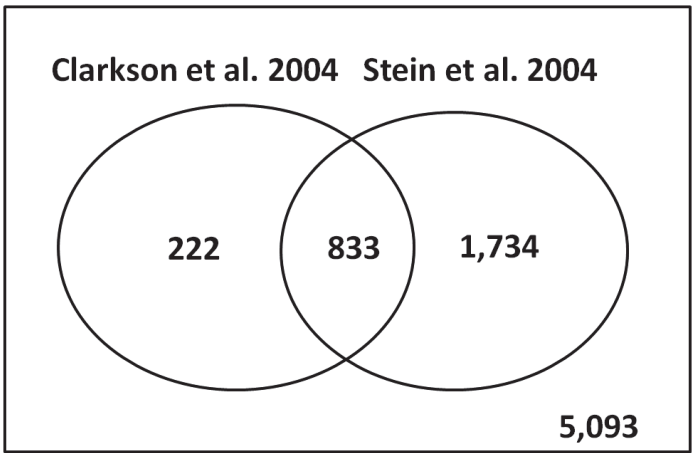


Figure 1: Venn diagram showing the overlap and discrepancies between genes differentially expressed in Clarkson et al. (2004) and Stein et al. (2004).

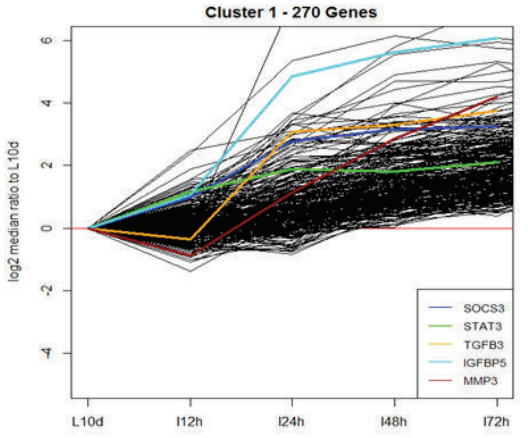
Clusters of Differentially Expressed Genes

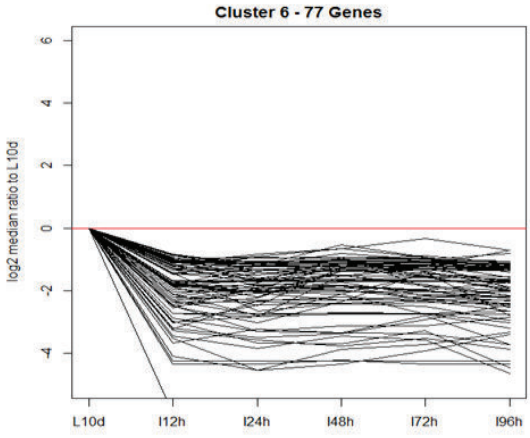
We used STEM algorithm developed by Ernst et al (28,29) to cluster genes identified as differentially expressed between last day lactation and last time point of involution. We used c=3 and m=50 for input parameters where c indicates units of change and m, the number of candidate profiles. This run significantly clustered 774 genes out of 1,055 genes (73.4 % of 1055 genes) differentially expressed genes. Table 1 lists patterns, size and p-value of significant clusters out of 50 possible cluster profiles. Patterns indicate the relative expression of genes in clusters compared to the lactation day 10 levels. Figure 2 shows gene expression profiles for the ten significant clusters.

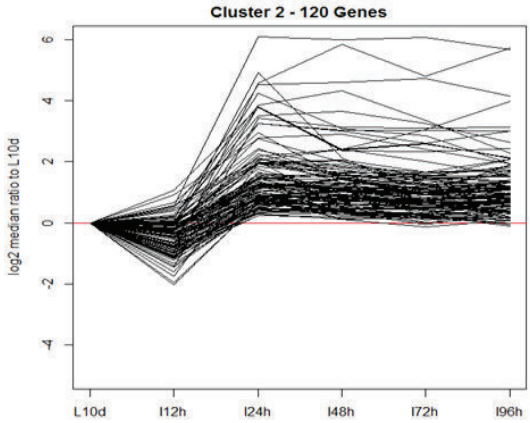
Table 1: Patterns, size and p-values of significant clusters identified through the STEM algorithm using data from Clarkson et al. 2004

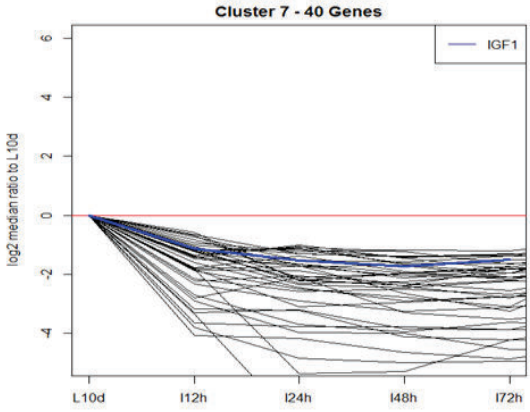
Cluster	Pattern*	Size	P-Value	Number of human orthologus genes identified
#1	0,1,2,3,4,5	270	1.30E-234	256
#2	0,-1,2,1,1,1	120	1.10E-54	118
#3	0,-1,-2.-3.-4.-5	74	3.00E-26	69
#4	0,-3,-1,-1,0,1	50	5.10E-11	47
#5	0,3,0,-3,-3,-2	46	7.30E-08	43
#6	0,-3,-4,-2,-3,-3	77	1.20E-07	76
#7	0,-3,-4,-5,-6,-3	40	2.30E-05	39
#8	0,-2,1,2,4,1	32	1.10E-04	31
#9	0,2,1,0,2,5	39	5.70E-04	37
#10	0,2,5,3,6,3	26	7.30E-04	26

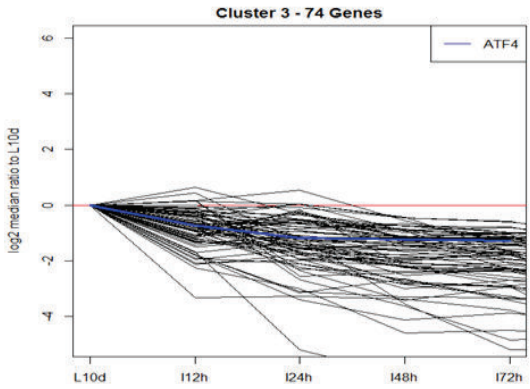
*Pattern indicates relative gene expression levels of lactation day 10 and involution days 0.5, 1, 2, 3 and 4 days compared lactation day 10.

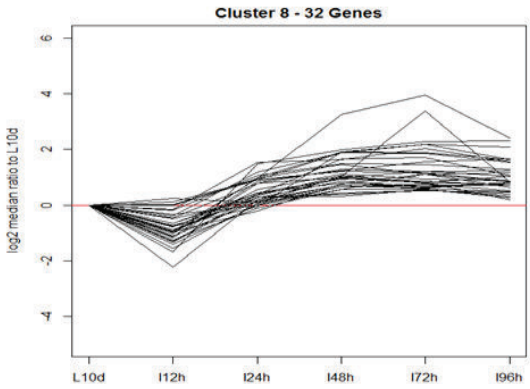

Cluster 1 - 270 Genes
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h
SOCS3
STAT3
TGFB3
IGFBP5
MMP3


Cluster 6 - 77 Genes
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h
I96h


Cluster 2 - 120 Genes
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h
I96h


Cluster 7 - 40 Genes
IGF1
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h


Cluster 3 - 74 Genes
ATF4
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h


Cluster 8 - 32 Genes
log2 median ratio to L10d
6
4
2
0
-2
-4
L10d
I12h
I24h
I48h
I72h
I96h

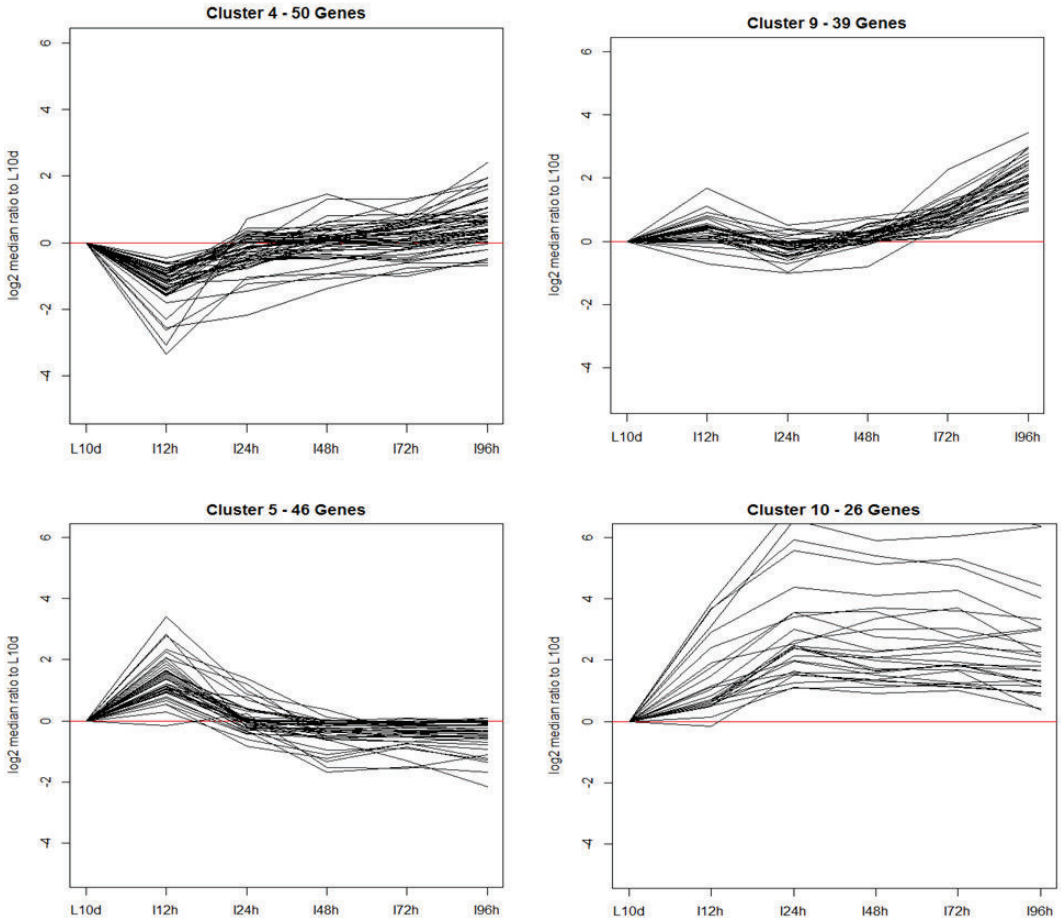


Figure 2: **Relative gene expression profiles for the ten significant clusters identified through the STEM Clustering.**

Y-axis represents the relative gene expression levels of involution days 0.5, 1, 2, 3 and 4 days compared lactation day 10 in log2 scale. X-axis represents the time points (L10, lactation day 10; I12h, involution day 0.5; I24h, involution day 1; I48h, involution day 2; I72h, involution day 3; I96h, involution day 4). SOCS3, suppressor of cytokine signaling 3; IGF1, insulin-like growth factor 1 (somatomedin C); STAT3, signal transducer and activator of transcription 3 (acute-phase response factor); TGFB3, transforming growth factor, beta 3; ATF4, activating transcription factor 4; IGFBP5, insulin-like growth factor binding protein 5; MMP3, matrix metallopeptidase 3

Ontology Analysis for Significant Clusters

We next conducted ontology analysis on genes of significant clusters identified through the STEM using GO annotations for mus musculus gene products available from the Mouse Genome Informatics to understand biologically relevant processes. Gene ontology enrichment showed that only Cluster 1 was statistically significantly enriched for the biological processes while few biological processes were enriched in the other clusters (Table 2).

Table 2: Results of ontology analysis for the STEM significant clusters

Type	Category ID	Category Name	# Genes Category	STEM Clusters									
				#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
BP	GO:0001568	Blood vessel development	394	**0.002**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0001944	Vasculature development	415	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0002455	Humoral immune response mediated by circulating immunoglobulin	37	**0.038**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006082	Organic acid metabolic process	517	NS	NS	NS	NS	NS	NS	**0.054**	NS	NS	NS
BP	GO:0006631	Fatty acid metabolic process	190	**0.038**	NS	NS	NS	NS	**0.048**	NS	NS	NS	NS
BP	GO:0006635	Fatty acid beta-oxidation	37	**0.038**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006909	Phagocytosis	84	**0.008**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006950	Response to stress	1539	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006952	Defense response	575	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006954	Inflammatory response	322	**0.002**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0006955	Immune response	529	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0009062	Fatty acid catabolic process	47	**0.036**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0009605	Response to external stimulus	932	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0016042	Lipid catabolic process	130	**0.018**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0019395	Fatty acid oxidation	54	**0.026**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0019752	Carboxylic acid metabolic process	490	NS	NS	NS	NS	NS	NS	**0.036**	NS	NS	NS
BP	GO:0019882	Antigen processing and presentation	60	**0.008**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0030036	Actin cytoskeleton organization	279	**0.038**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0030198	Extracellular matrix organization	129	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0032101	Regulation of response to external stimulus	344	**0.014**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0032103	Positive regulation of response to external stimulus	144	**0.046**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0032787	Monocarboxylic acid metabolic process	286	NS	NS	NS	NS	NS	**0.05**	NS	NS	NS	NS
BP	GO:0034097	Response to cytokine	298	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0034341	Response to interferon-gamma	39	**0.008**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0034440	Lipid oxidation	56	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0040012	Regulation of locomotion	435	**0.048**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0042127	Regulation of cell proliferation	869	**0.052**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0042221	Response to chemical	1573	NS	NS	NS	NS	NS	NS	NS	NS	**0.004**	NS
BP	GO:0043062	Extracellular structure organization	129	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0043436	Oxoacid metabolic process	511	NS	NS	NS	NS	NS	NS	**0.05**	NS	NS	NS
BP	GO:0044281	Small molecule metabolic process	1299	NS	NS	NS	NS	NS	**0.006**	**0.012**	NS	NS	NS
BP	GO:0044711	Single-organism biosynthetic process	700	NS	NS	NS	NS	NS	**0.02**	NS	NS	NS	NS
BP	GO:0045087	Innate immune response	233	**0.002**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0048514	Blood vessel morphogenesis	344	**0.032**	NS	NS	NS	NS	NS	NS	NS	NS	NS
BP	GO:0070887	Cellular response to chemical stimulus	1033	NS	NS	NS	NS	NS	NS	NS	NS	**0.024**	NS

BP	GO:1901605	Alpha-amino acid metabolic process	117	NS	NS	**0.052**	NS	NS	NS	NS	NS	NS	NS
BP	GO:2000145	Regulation of cell motility	396	**0.048**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0000502	Proteasome complex	53	NS	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0000932	Cytoplasmic mrna processing body	26	NS	**0.02**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0005578	Proteinaceous extracellular matrix	183	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0005581	Collagen trimer	51	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0005615	Extracellular space	694	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0005783	Endoplasmic reticulum	747	NS	NS	NS	NS	NS	**0.002**	NS	NS	NS	NS
CC	GO:0005838	Proteasome regulatory particle	11	NS	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0009986	Cell surface	473	**0.026**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0012505	Endomembrane system	1633	NS	NS	NS	NS	NS	**0.006**	NS	NS	NS	NS
CC	GO:0016021	Integral component of membrane	1967	NS	NS	NS	NS	NS	**0.008**	NS	NS	NS	NS
CC	GO:0022624	Proteasome accessory complex	18	NS	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0031224	Intrinsic component of membrane	2045	NS	NS	NS	NS	NS	**0.006**	NS	NS	NS	NS
CC	GO:0031982	Vesicle	1944	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0031988	Membrane-bounded vesicle	1828	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	**0.052**
CC	GO:0043227	Membrane-bounded organelle	5489	NS	NS	NS	NS	NS	NS	NS	**0.038**	NS	NS
CC	GO:0043230	Extracellular organelle	1534	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	**0.04**
CC	GO:0044420	Extracellular matrix part	89	**0.004**	NS	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0044421	Extracellular region part	1992	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	NS
CC	GO:0044444	Cytoplasmic part	3297	NS	NS	NS	NS	NS	**0.048**	NS	NS	NS	NS
CC	GO:0065010	Extracellular membrane-bounded organelle	1534	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	**0.04**
CC	GO:0070062	Extracellular vesicular exosome	1528	**<0.001**	**<0.001**	NS	NS	NS	NS	NS	NS	NS	**0.038**

BP, biological process; CC; cellular components

Involution Specific Gene Signatures

To develop the involution specific gene signatures for the conduct of gene set analysis on human IBC and non-IBC gene expression profiles, we identified human orthologus genes for genes of significant clusters identified through STEM. Orthologs data was downloaded from the ENSEMBL website (http://www.ensembl.org/biomart/martview) and orthologous genes were identified using ENSEMBL gene id. Table 3 lists the number of human orthologous genes identified for each of 10 significant clusters.

Table 3: Involution Specific Gene Signatures through the STEM clustering using data from Clarkson et al. 2004

STEM Cluster	Pattern*	# Genes	Number of human orthologous genes identified	Signature name
#1	0,1,2,3,4,5	270	256	Inv1
#2	0,-1,2,1,1,1	120	118	Inv2
#3	0,-1,-2.-3.-4.-5	74	69	Inv3
#4	0,-3,-1,-1,0,1	50	47	Inv4
#5	0,3,0,-3,-3,-2	46	43	Inv5
#6	0,-3,-4,-2,-3,-3	77	76	Inv6
#7	0,-3,-4,-5,-6,-3	40	39	Inv7
#8	0,-2,1,2,4,1	32	31	Inv8
#9	0,2,1,0,2,5	39	37	Inv9
#10	0,2,5,3,6,3	26	26	Inv10

*Pattern indicates relative gene expression fold changes of lactation day 10 and involution days 0.5, 1, 2, 3 and 4 days profiles compared lactation day 10.

3 GENE SET ANALYSIS OF INVOLUTION SPECIFIC GENE SIGNATURES

3.1 Methods

Gene Expression Dataset on IBC and non-IBC cases

Gene expression data for IBC and non-IBC cases were obtained from the NCBI Gene Expression Omnibus (GSE22597) and the EBI ArrayExpress (E-MTAB-1006 and E-MTAB-1547) and collected through the World IBC Consortium [25]. These databases include the largest series of IBC samples ever reported and tumor samples were obtained from patients treated in three institutions: the Institut Paoli-Calmettes (IPC, Marseille, France: 71 IBC and 139 non-IBC cases), the MD Anderson Cancer Center (MDA, Houston, TX, USA: 25 IBC and 58 non-IBC cases), and the General Hospital Sint-Augustinus (TCRU, Antwerp, Belgium: 41 IBC and 55 non-IBC cases) [25].

Preprocessing of Gene Expression Data

Raw gene expression profiles were preprocessed using GCRMA analysis [29] with quantile normalization and probeset-level signals were summarized in log base 2 scale. We selected a custom Chip Definition Files (CDFs) HGU133A_Hs_ENTREZG version 18 for preprocessing GSE22597 data and HGU133Plus2_Hs_ENTREZG version 18 for preprocessing E-MTAB-1006 and E-MTAB-1547 data [30]. There are 12,135 probe sets with a CDF HGU133A_Hs_ENTREZG version 18 with 12,064 probe sets representing 12,064 genes as per the annotation database available at the BrainArray. There are 19674 probe sets with a CDF HGU133Plus2_Hs_ENTREZG version 18 with 19,544 probe sets representing 19544 genes as per the annotation database available at the BrainArray. After preprocessing gene expression data, all 3 data sets were merged using common informative probe sets (N= 12,129). To remove the batch effect, we used the removeBatchEffect function from the *limma* package from the Bioconductor [31]. This function fits a linear model to the data and removes the components due to the batch effects. The principal component analysis plots were generated prior and after removing the batch effect to verify the accuracy of the removeBatchEffect function (Figure 3). The final merged dataset

consisted of 388 samples (137 IBC cases and 251 non-IBC cases) with 12,129 probe sets with 12,063 probe sets representing 12,063 genes with one probe set – one gene relationship.

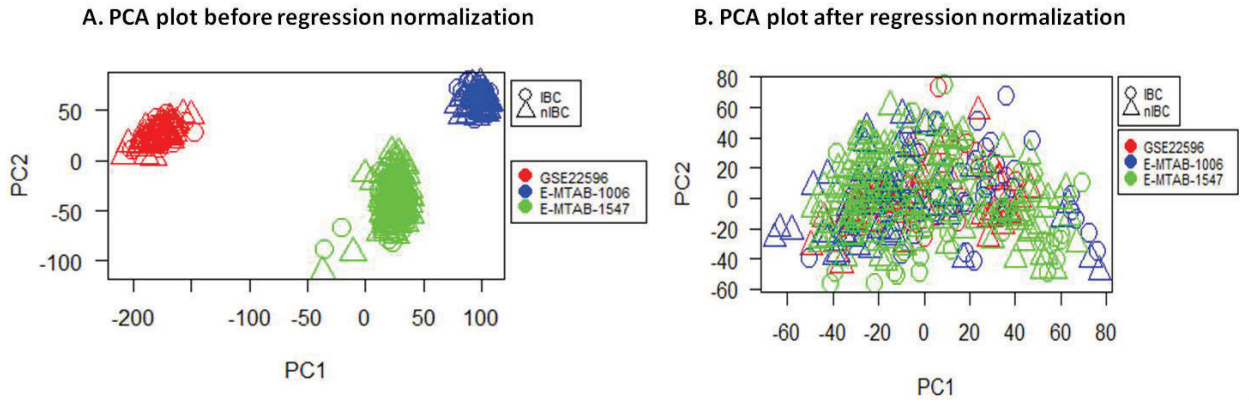


Figure 3: **Principal Component Analysis plots prior and after regression normalization to verify the removal of batch effect.**

Involution Specific Gene Signatures

Involution specific gene signatures were identified from the results of STEM cluster analysis conducted on post-natal mouse mammary gland development as described in Chapter 2. We identified orthologous genes for genes that were found to form significant clusters in STEM cluster analysis by using ENSEMBL gene id on the orthologous data downloaded from the ENSEMBL website. Involution specific gene signatures have also been reported in the study by Stein et al. 2009 [10]. We downloaded those signatures and identified orthologous genes for each signature and used them to evaluate their enrichment in IBC versus non-IBC and TN versus non-TN subtypes.

Gene Set Enrichment Analysis (GSEA) of Involution Specific Signatures

We used GSEA algorithm as mentioned in [34] to evaluate enrichment of involution specific gene signatures in IBC cases compared to non-IBC cases and TN BC cases compared to non-TN BC cases. We ranked genes in the GSEA using t-test and all other options in the GSEA kept as default.

Training and Validation Data

We divided the merged dataset into the training set to run the GSEA and into the validation set to validate the GSEA results for reproducibility. We used the stratified random sampling method with inclusion of information on IBC status, TN status and age at diagnosis (<50 years or >=50 years) to divide the merged dataset into the training and validation sets (Table 4).

Table 4: Final merged dataset and training and validation sets

	Total cases	IBC*			Non-IBC*		
		Total	TN	Non-TN	Total	TN	Non-TN
Merged Dataset	388	137	20	101	251	34	197
Training Set	195	68	10	50	127	18	99
Validation Set	193	69	10	51	124	16	98

IBC, inflammatory breast cancer; TN, triple-negative

*15 cases in IBC and 20 cases in non-IBC groups did not have information available on TN status.

3.2 Results

Results of GSEA of Involution Specific Signatures in IBC versus non-IBC

Out of 10 gene signatures developed through the STEM clustering using data from Clarkson et al. (2004), 7 signatures were upregulated in the training set of IBC phenotype with 2 signatures significant at FDR < 25%. In the validation set, we found that 9 gene signatures were upregulated in IBC phenotype with no signature significant at FDR < 25%. Three signatures were upregulated in non-IBC phenotype with no gene signatures significant at FDR <25% in the training set (Table 5). In the validation set, 1 gene signature was upregulated in non-IBC phenotype with no gene signature significant at FDR <25% (Table 5). When comparing the results in the training and validation sets, we found that 6 out of 10 gene signatures were upregulated in IBC in both training and validation sets with only 1 gene signature (Inv 5) significant at nominal p-value of 0.05 in both the training and validation sets. In the merged 3 breast cancer data sets, we found that 2 out of 10 gene signatures (Inv5 and Inv6) were significantly upregulated in IBC versus non-IBC phenotype at nominal p-value of 0.05 (Table 6). Figure 4 represents the enrichment plot from the GSEA for Inv5 signature in IBC versus non-IBC and Table 7 shows the list of genes in Inv5 signature and genes enriched in IBC. For the involution specific signatures reported by Stein et al. (2009), we found that no gene signature was significantly enriched in IBC or non-IBC at FDR <25% or nominal p-value of 0.05 in both the training and validation sets (Table 5) or in the merged 3 breast cancer data sets (Table 6).

Table 5: GSEA results of involution-specific signatures in IBC versus non-IBC in the training and validation sets

Gene Signature	Results on Training Set						Results on Validation Set					
	Enriched in IBC vs non-IBC	Size	ES	NES	Nominal p-value	FDR q-value	Enriched in IBC vs non-IBC	Size	ES	NES	Nominal p-value	FDR q-value
Involution specific signatures developed through STEM clustering using data from Clarkson et al. 2004												
Inv1	IBC	205	0.360	1.064	0.415	0.443	IBC	205	0.212	0.643	0.912	0.927
Inv2	Non-IBC	100	-0.139	-0.552	0.933	0.976	IBC	100	0.169	0.674	0.849	1.000
Inv3	Non-IBC	57	-0.299	-0.917	0.625	1.000	IBC	57	0.322	1.030	0.403	0.836
Inv4	IBC	33	0.392	1.108	0.360	0.469	IBC	33	0.307	0.907	0.581	0.835
Inv5	**IBC**	**30**	**0.514**	**1.492**	**0.014**	**0.263**	**IBC**	**30**	**0.492**	**1.436**	**0.028**	**0.433**
Inv6	IBC	62	0.402	1.399	0.043	0.222	IBC	62	0.326	1.122	0.258	1.000
Inv7	IBC	36	0.295	0.843	0.719	0.650	Non-IBC	36	-0.284	-0.810	0.790	0.709
Inv8	Non-IBC	24	-0.294	-0.883	0.647	0.897	IBC	24	0.294	0.878	0.621	0.774
Inv9	IBC	33	0.611	1.363	0.131	0.184	IBC	33	0.409	0.915	0.592	0.975
Inv10	IBC	21	0.296	0.897	0.583	0.654	IBC	21	0.339	1.030	0.414	1.000
Involution specific signatures reported in Stein et al. 2009												
S.C1	IBC	182	0.42	1.16	0.274	0.818	IBC	182	0.35	0.94	0.522	1
S.C2	Non-IBC	205	-0.27	-0.78	0.737	0.746	Non-IBC	205	-0.2	-0.63	0.945	0.962
S.C3	IBC	252	0.2	0.74	0.766	0.922	Non-IBC	252	-0.22	-0.81	0.695	1
S.C4	IBC	258	0.18	0.74	0.753	0.832	Non-IBC	258	-0.29	-1.25	0.215	0.718
S.C5.I3VL7	IBC	117	0.21	0.8	0.743	0.906	IBC	117	0.23	0.83	0.672	1
S.C6	Non-IBC	100	-0.25	-0.94	0.534	0.905	Non-IBC	100	-0.26	-1	0.438	0.715
S.C7	IBC	225	0.29	1.17	0.19	1	IBC	225	0.24	0.97	0.504	1
S.C8	Non-IBC	153	-0.25	-1.01	0.413	1	IBC	153	0.29	1.17	0.171	1
S.C9	Non-IBC	66	-0.25	-0.8	0.818	0.952	Non-IBC	66	-0.39	-1.24	0.149	0.365
S.I1VL7	IBC	495	0.19	0.85	0.693	1	Non-IBC	495	-0.16	-0.7	0.962	1
S.I2VL7	IBC	612	0.21	0.87	0.647	1	IBC	612	0.2	0.79	0.784	0.977
S.I3VL7	IBC	648	0.2	0.83	0.708	0.968	IBC	648	0.19	0.75	0.836	0.779
S.I4VL7	IBC	894	0.23	0.91	0.579	1	IBC	894	0.2	0.78	0.777	0.835

ES, enrichment score; NES, normalized enrichment score; FDR, false discovery rate; IBC, inflammatory breast cancer

Table 6: GSEA results of involution-specific signatures in IBC versus non-IBC in the merged 3 breast cancer data sets

Gene Signature	Merged 3 breast cancer data sets IBC = 137 and non-IBC = 251 # Genes = 12064					
	Enriched in IBC versus non-IBC	Size	ES	NES	Nominal p-value	FDR q-value
Involution specific signatures developed through STEM clustering using data from Clarkson et al. 2004						
Inv1	IBC	205	0.29	0.88	0.584	0.869
Inv2	IBC	100	0.13	0.52	0.97	0.988
Inv3	IBC	57	0.25	0.81	0.807	0.902
Inv4	IBC	33	0.32	0.94	0.541	0.873
Inv5	**IBC**	**30**	**0.51**	**1.47**	**0.03**	**0.392**
Inv6	**IBC**	**62**	**0.42**	**1.47**	**0.021**	**0.204**
Inv7	IBC	36	0.26	0.75	0.867	0.896
Inv8	Non-IBC	24	-0.3	-0.92	0.573	0.541
Inv9	IBC	33	0.54	1.22	0.279	0.407
Inv10	IBC	21	0.43	1.34	0.122	0.298
Involution specific signatures reported in Stein et al. 2009						
S.C1	IBC	182	0.4	1.08	0.379	0.614
S.C2	Non-IBC	205	-0.25	-0.77	0.735	0.902
S.C3	Non-IBC	252	-0.19	-0.72	0.783	0.849
S.C4	Non-IBC	258	-0.21	-0.89	0.568	1
S.C5.I3VL7	IBC	117	0.29	1.1	0.315	0.818
S.C6	Non-IBC	100	-0.26	-0.99	0.48	1
S.C7	IBC	225	0.28	1.13	0.26	1
S.C8	Non-IBC	153	-0.22	-0.88	0.717	0.835
S.C9	Non-IBC	66	-0.37	-1.21	0.189	0.847
S.I1VL7	IBC	495	0.19	0.83	0.717	0.639
S.I2VL7	IBC	612	0.21	0.87	0.671	0.798
S.I3VL7	IBC	648	0.21	0.84	0.694	0.729
S.I4VL7	IBC	894	0.23	0.9	0.591	0.882

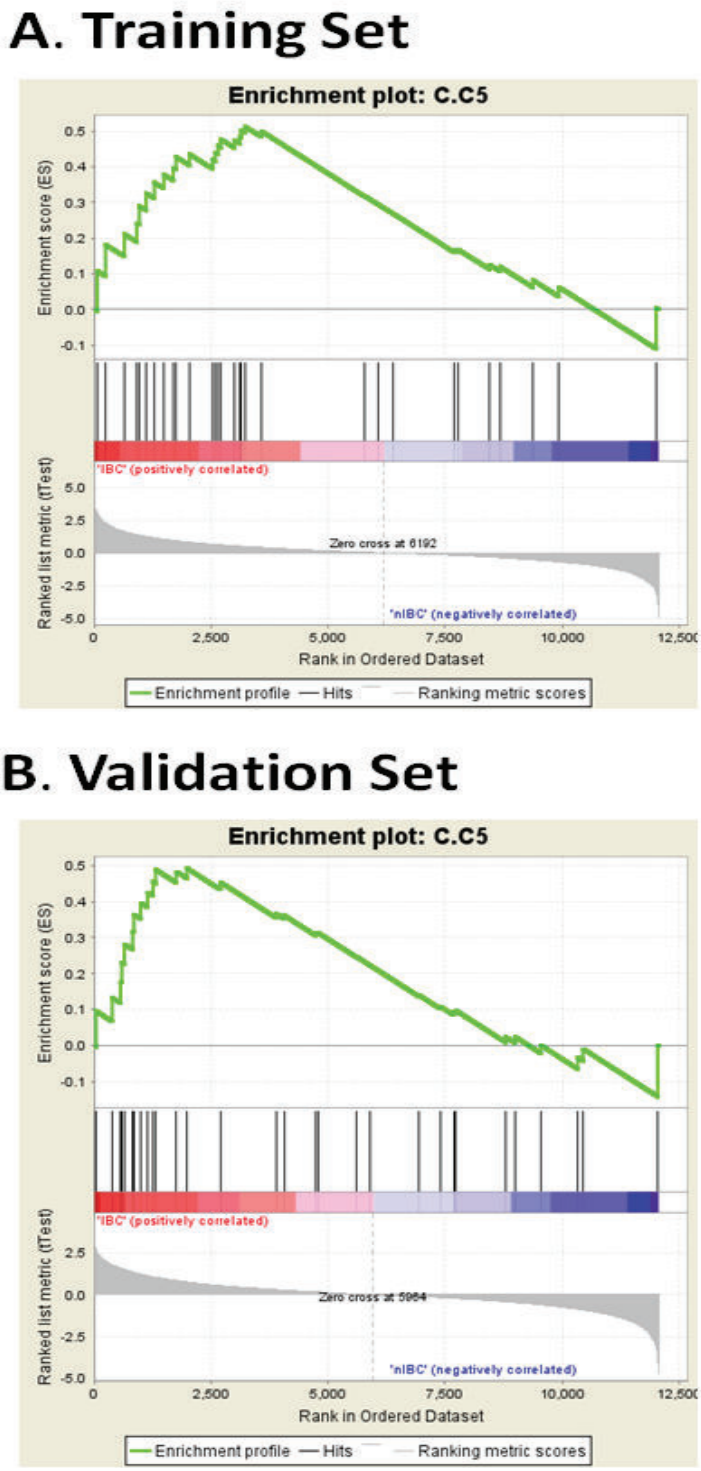


Figure 4: Enrichment plots form GSEA for Inv5 signature for IBC versus non-IBC in the training and validation sets.

IBC, inflammatory breast cancer; nIBC, non-Inflammatory breast cancer.

Table 7: List of genes in Inv5 signature and genes enriched in IBC versus non-IBC

Gene Symbol	Gene Name	Enrichment in IBC (Training Set)	Enrichment in IBC (Validation Set)
CD79B	CD79b molecule, immunoglobulin-associated beta	Yes	Yes
IVL	involucrin	Yes	Yes
KIF2C	kinesin family member 2C	Yes	Yes
NOP2		Yes	Yes
LDHB	lactate dehydrogenase B	Yes	Yes
LEP	leptin (obesity homolog, mouse)	Yes	Yes
AVIL	advillin	Yes	Yes
DKK2	dickkopf homolog 2 (Xenopus laevis)	Yes	Yes
ARAP3		Yes	Yes
YBX2	Y box binding protein 2	Yes	No
CEP250	centrosomal protein 250kDa	Yes	No
RPS6KB2	ribosomal protein S6 kinase, 70kDa, polypeptide 2	Yes	No
STX3	syntaxin 3	Yes	No
NFKBIE	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon	Yes	No
GSTO1	glutathione S-transferase omega 1	Yes	No
WT1	Wilms tumor 1	Yes	No
DYNC1I1	dynein, cytoplasmic 1, intermediate chain 1	Yes	No
RET	ret proto-oncogene (multiple endocrine neoplasia and medullary thyroid carcinoma 1, Hirschsprung disease)	Yes	No
TIPIN	TIMELESS interacting protein	Yes	No
LLGL2	lethal giant larvae homolog 2 (Drosophila)	No	Yes
TG	thyroglobulin	No	Yes
DDIT4	DNA-damage-inducible transcript 4	No	Yes
HPCA	hippocalcin	No	Yes
GRAMD3	GRAM domain containing 3	No	No
PAX4	paired box gene 4	No	No
KCNJ4	potassium inwardly-rectifying channel, subfamily J, member 4	No	No
NPY	neuropeptide Y	No	No
CHRNA6	cholinergic receptor, nicotinic, alpha 6	No	No
FRAT2	frequently rearranged in advanced T-cell lymphomas 2	No	No
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	No	No

Results of GSEA of Involution Specific Signatures in TN subtype versus non-TN subtype in IBC

Out of 10 gene signatures developed through STEM clustering using data from Clarkson et al. (2004), we found that no gene signature was significantly enriched in TN subtype versus non-TN subtype in IBC cases at FDR <25% or nominal p-value of 0.05 in both the training and validation sets. For involution specific signatures reported by Stein et al. (2009), we also did not find any gene signature that was significantly enriched in TN subtype versus non-TN subtype in IBC cases at FDR <25% or nominal p-value of 0.05 in both the training and validation sets (Table 8).

Table 8: GSEA results of involution-specific signatures in TN subtype versus non-TN subtype in IBC

Gene Signature	Results on Training Set						Results on Validation Set					
	Enriched in TN vs non-TN	Size	ES	NES	Nominal p-value	FDR q-value	Enriched in TN vs non-TN	Size	ES	NES	Nominal p-value	FDR q-value
Involution specific signatures developed through STEM clustering using data from Clarkson et al. 2004												
Inv1	Non-TN	205	-0.215	-0.620	0.906	1.000	TN	205	0.389	1.192	0.270	1.000
Inv2	Non-TN	100	-0.256	-1.009	0.434	0.977	TN	100	0.183	0.718	0.818	0.863
Inv3	Non-TN	57	-0.476	-1.362	0.039	0.271	Not-TN	57	-0.401	-1.162	0.217	0.856
Inv4	TN	33	0.211	0.598	0.938	0.950	Not-TN	33	-0.295	-0.897	0.602	1.000
Inv5	TN	30	0.336	1.029	0.391	1.000	Not-TN	30	-0.315	-0.837	0.785	0.692
Inv6	TN	62	0.255	0.872	0.717	0.906	TN	62	0.269	0.915	0.632	0.652
Inv7	Non-TN	36	-0.517	-1.371	0.077	0.515	Not-TN	36	-0.313	-0.871	0.717	0.849
Inv8	Non-TN	24	-0.309	-0.962	0.485	0.847	TN	24	0.311	0.979	0.475	0.878
Inv9	Non-TN	33	-0.281	-0.592	0.915	0.962	TN	33	0.431	0.959	0.528	0.707
Inv10	Non-TN	21	-0.242	-0.719	0.827	1.000	TN	21	0.333	0.988	0.488	1.000
Involution specific signatures reported in Stein et al. 2009												
S.C1	Non-TN	182	-0.23	-0.59	0.965	1	TN	182	0.43	1.13	0.31	1
S.C2	Non-TN	205	-0.2	-0.56	0.969	0.989	Non-TN	205	-0.31	-0.9	0.605	1
S.C3	Non-TN	252	-0.18	-0.59	0.937	1	TN	252	0.2	0.79	0.762	0.93
S.C4	Non-TN	258	-0.29	-1.13	0.327	0.908	TN	258	0.17	0.76	0.787	0.832
S.C5.I3VL7	Non-TN	117	-0.35	-1.28	0.116	1	Non-TN	117	-0.31	-1.08	0.349	0.933
S.C6	Non-TN	100	-0.32	-1.12	0.31	0.696	TN	100	0.2	0.83	0.836	1
S.C7	TN	225	0.23	0.93	0.572	0.472	Non-TN	225	-0.22	-0.8	0.822	0.729
S.C8	Non-TN	153	-0.32	-1.19	0.149	1	Non-TN	153	-0.3	-1.12	0.243	1
S.C9	Non-TN	66	-0.38	-1.11	0.301	0.497	TN	66	0.27	0.93	0.615	1
S.I1VL7	Non-TN	495	-0.27	-1.12	0.275	0.558	Non-TN	495	-0.2	-0.88	0.676	1
S.I2VL7	Non-TN	612	-0.26	-0.97	0.491	0.645	Non-TN	612	-0.21	-0.81	0.805	0.817
S.I3VL7	Non-TN	648	-0.27	-1.01	0.446	0.64	Non-TN	648	-0.22	-0.84	0.745	0.911
S.I4VL7	Non-TN	894	-0.25	-0.9	0.604	0.724	TN	894	0.22	0.88	0.653	1

ES, enrichment score; NES, normalized enrichment score; FDR, false discovery rate; IBC, inflammatory breast cancer; TN, triple negative

Results of GSEA of Involution Specific Signatures in TN subtype versus non-TN subtype in non-IBC

Out of 10 gene signatures developed through STEM clustering using data from Clarkson et al. (2004), we found that no gene signature was significantly enriched in TN subtype versus non-TN subtype in non-IBC cases at FDR <25% or nominal p-value of 0.05 in both the training and validation sets. For involution specific signatures reported by Stein et al. (2009), we also did not find any gene signature that was significantly enriched in TN subtype versus non-TN subtype in non-IBC cases at FDR <25% or nominal p-value of 0.05 in both the training and validation sets (Table 9).

Table 9: GSEA results of involution-specific signatures in TN subtype versus non-TN subtype in non-IBC

Gene Signature	Results on Training Set						Results on Validation Set					
	Enriched in TN or non-TN	Size	ES	NES	Nominal p-value	FDR q-value	Enriched in TN or non-TN	Size	ES	NES	Nominal p-value	FDR q-value
Involution specific signatures developed through STEM clustering using data from Clarkson et al. 2004												
Inv1	Non-TN	205	-0.207	-0.621	0.926	0.945	TN	205	0.443	1.303	0.160	0.431
Inv2	TN	100	0.193	0.740	0.731	1.000	TN	100	0.270	1.016	0.397	0.375
Inv3	Non-TN	57	-0.328	-0.950	0.537	1.000	Non-TN	57	-0.289	-0.842	0.732	0.640
Inv4	TN	33	0.228	0.693	0.864	1.000	TN	33	0.407	1.174	0.264	0.278
Inv5	TN	30	0.385	1.166	0.159	1.000	Non-TN	30	-0.403	-1.059	0.388	0.674
Inv6	TN	62	0.263	0.893	0.667	0.989	TN	62	0.365	1.260	0.108	0.364
Inv7	Non-TN	36	-0.333	-0.923	0.583	0.795	TN	36	0.375	1.063	0.336	0.367
Inv8	TN	24	0.361	1.093	0.356	0.666	TN	24	0.398	1.179	0.250	0.328
Inv9	TN	33	0.308	0.689	0.843	0.868	TN	33	0.685	1.504	0.048	0.291
Inv10	TN	21	0.384	1.138	0.284	0.825	TN	21	0.389	1.181	0.258	0.407
Involution specific signatures reported in Stein et al. 2009												
S.C1	Non-TN	182	-0.24	-0.66	0.912	0.924	TN	182	0.55	1.47	0.047	0.302
S.C2	Non-TN	205	-0.38	-1.16	0.297	1	TN	205	0.28	0.85	0.645	0.594
S.C3	TN	252	0.16	0.62	0.933	0.952	TN	252	0.29	1.07	0.352	0.355
S.C4	TN	258	0.22	0.87	0.597	1	TN	258	0.35	1.46	0.088	0.164
S.C5.I3VL7	Non-TN	117	-0.28	-1.07	0.365	0.978	TN	117	0.26	0.97	0.477	0.456
S.C6	TN	100	0.21	0.8	0.772	0.825	TN	100	0.28	1.12	0.271	0.34
S.C7	TN	225	0.24	0.96	0.523	1	TN	225	0.22	0.9	0.621	0.539
S.C8	Non-TN	153	-0.21	-0.78	0.883	1	Non-TN	153	-0.2	-0.77	0.902	0.774
S.C9	TN	66	0.28	0.94	0.564	1	TN	66	0.3	0.98	0.467	0.485
S.I1VL7	TN	495	0.2	0.86	0.7	0.845	TN	495	0.28	1.22	0.179	0.365
S.I2VL7	Non-TN	612	-0.21	-0.85	0.684	1	TN	612	0.29	1.18	0.208	0.366
S.I3VL7	Non-TN	648	-0.22	-0.87	0.652	1	TN	648	0.28	1.15	0.253	0.35
S.I4VL7	Non-TN	894	-0.2	-0.75	0.845	0.924	TN	894	0.32	1.24	0.148	0.43

ES, enrichment score; NES, normalized enrichment score; FDR, false discovery rate; IBC, inflammatory breast cancer; TN, triple negative

4 DISCUSSION

We reanalyzed the previously published expression profiling data set obtained from mammary glands derived from mice at various stages of post-lactational mammary gland involution (12 gene expression profiles with two hybridizations for each of 10 day lactation time point, and 12, 24, 48, 72 and 96 hour involution time points). We focused on genes that were differentially expressed during time periods spanning from the last day of lactation (day 10) to 4th day of involution by greater than 2 fold ($P < 0.05$) and performed STEM cluster analysis to discern time varied expression patterns. We identified 10 separate and significant time varied gene expression patterns of post-lactational mammary gland involution.

Up to now, the two most comprehensive studies examining the global gene expression on the post-lactational mammary gland involution have been conducted by Clarkson et al [4, 6] and Stein et al [5]. Clarkson et al [4] used the K means clustering method and Stein et al used the self-organizing map in [5] and the hierarchical ordered partitioning and collapsing hybrid (HOPACH) method in [10] to find the patterns among differentially expressed genes in the post-lactational involution period. They observed the induction of apoptotic pathways and immunomodulatory signals in the post-lactational involution. During our reanalysis, we used the STEM clustering method on the dataset by Clarkson et al [4] to find the time-varied patterns among differentially expressed genes in the post-lactational involution period. We identified 10 separate and significant time varied expression patterns from 774 genes out of 1,055 significantly differentially expressed genes. Gene ontology analyses of these clusters showed that cluster 1, which represented genes showing gradual up-regulation during first 4 days of involution, had over-representation of biological processes like fatty acid beta-oxidation, blood vessel morphogenesis, phagocytosis, response to interferon-gamma, humoral immune response mediated by circulating immunoglobulin, and regulation of cell motility. Our findings of over-representation of biological processes are in general agreement with findings by Clarkson et al. [4] who used a different analytical

approach, and with findings by Stein et al. [5, 10] who used a different mouse system and a different analytical approach.

We examined the enrichment of post-lactational mammary gland involution gene expression patterns in TN BC and IBC using the GSEA method. First, we used 10 significant time varied gene expression patterns that we found in the dataset from Clarkson et al [4]. We found that only 1 gene expression pattern was enriched in IBC compared to non-IBC in both training and validation sets at the nominal p-value. None of these gene expression patterns was enriched in TN BC compared to non-TN BC for both IBC and non-IBC groups in both trading and validation sets. Second, we used 13 gene expression patterns on post-lactational involution as reported by Stein et al [10] and found that none of these gene expression patterns was significantly enriched in IBC compared to non-IBC and TN BC compared to non-TN BC in both training and validation sets. To further investigate, we examined the overlap between the involution-specific signatures and the IBC-like signature (79 genes) [25]. We found that there was minimal overlap between the involution-specific signatures and the IBC-like signature and no gene overlapped between Inv5 signature and the IBC-like signature (Table 10).

Major strength of our study is the use of the largest series of IBC samples ever reported by the World IBC Consortium and the novelty of our study is the use of updated Chip Definition Files from the BrainArray [30] during preprocessing of gene expression data for accurate probe mapping to the genome. Also, we used the gene set enrichment analysis (GSEA) [34] to examine enrichment of involution signatures in IBC and TN BC phenotype. The GSEA gives more statistical power to detect smaller changes in genes of a gene set compared to other methods of enrichment analysis. Major limitations of our study are cross-sectional analysis of enrichment of involution specific signatures in breast cancer and array-based measurement of gene expression profiles which limits the detection of differentially expressed genes with lower levels of expression.

One gene signature that showed enrichment in IBC compared to non-IBC contained genes that showed initial up-regulation and later down-regulation during the involution process. This might suggest that genes that up-regulate during an initial phase of involution after abrupt weaning might not be turning off and could be responsible for development of IBC like phenotype later in life. We examined the overlap of this gene expression pattern with the existing gene signatures using the Molecular Signatures Database (MSigDB) v4.0 which is available online at http://www.broadinstitute.org/gsea/msigdb/index.jsp [34]. We found that the genes up-regulated in vascular smooth muscle cells (VSMC) by c-Jun N-terminal protein kinase 1 (JNK1) [35] showed the most significant overlap (FDR q value = 6.07E-5). Among these overlapped genes, 3 genes - Involucrin (IVL), Cluster of Differentiation 79B (CD79B), and leptin (LEP) –were significantly enriched in IBC compared to non-IBC in both training and validation data sets in our analysis. Involucrin is a transglutaminase substrate protein present in keratinocytes of epidermis and other stratified squamous epithelia [36]. Tsuda et al (1997) investigated the expression of Involucrin in breast cancer and found that Involucrin expression was detected in 27% of breast cancer cases and was associated with high-grade atypia, a solid-nest pattern, cancer cell necrosis on histology and negative oestrogen receptor status [37]. Leptin is a product of the obese (*ob*) gene, an important regulator of energy balance and necessary for normal mammary gland development [38]. In ER-positive breast cancer cell lines, leptin has been shown to stimulate cell growth through activation of multiple signaling pathways including the Janus Kinase/Signal Transducer and Activator of Transcription (JAK/STAT) pathway [39].

Table 10: Overlap of genes between involution-specific signatures and IBC-like signature

	Involution signature	# Genes total	# Genes overlapped with IBC-like signature	Symbols of Genes overlapped with IBC-like signature
Clarkson et al. 2004	Inv1	256	1	MARCKS
	Inv2	118	2	ARPC2, HSP90B1
	Inv3	69	2	FOLR1, TMC6
	Inv4	47	0	
	Inv5	43	0	
	Inv6	76	1	DAB2
	Inv7	39	0	
	Inv8	31	0	
	Inv9	37	0	
	Inv10	26	0	
Stein et al. 2009	S.c1	242	1	PRKCB
	S.c2	262	1	HSP90B1
	S.c3	325	4	MARCKS, PNP, SP3, ZFR
	S.c4	343	7	ACTG1, ARPC2, ATF2, IQGAP1, RYK, TBL1XR1, TNPO1
	S.c5.i3vl7	146	4	ACOX1, CTBP2, CTSA, IQGAP1
	S.c6	137	1	PAK2
	S.c7	282	2	DAB2, TMC6
	S.c8	212	3	BCKDK, DAB2, FOLR1
	S.c9	81	1	JMJD6
	S.i1vl7	643	7	ACOX1, ARPC2, CTBP2, CTSA, FOLR1, IQGAP1, RYK
	S.i2vl7	790	10	ACOX1, ARPC2, CTBP2, CTSA, DAB2, FOLR1, IQGAP1, PNP, RYK, TMCO1
	S.i3vl7	834	9	ACOX1, ARPC2, CTBP2, CTSA, DAB2, FOLR1, HSP90B1, IQGAP1, RYK
	S.i4vl7	1160	14	ACOX1, ARPC2, ATF2, BCKDK, CTBP2, CTSA, DAB2, FOLR1, IQGAP1, MARCKS, PNP, RYK, TMCO1, ZFR

5 CONCLUSION

In summary, we have reanalyzed the previously published gene expression data set using the STEM clustering method and identified 10 separate and significant time-varied gene expression patterns for the abrupt post-lactational mammary gland involution. We found significant enrichment of one post-lactational involution gene signature in IBC compared to non-IBC. This enriched signature represented genes showing initial up-regulation and later down-regulation during the involution process and there was significant overlap with genes up-regulated in vascular smooth muscle cells (VSMC) by c-Jun N-terminal protein kinase (*JNK1*). We identified three genes – Involucrin (*IVL*), Cluster of Differentiation 79B (*CD79B*), and leptin (*LEP*) – that were significantly enriched in IBC compared to non-IBC in both training and validation data sets and that are up-regulated in VSMC by *JNK1*. Our results suggest that these genes might be playing a role in IBC development. Whether or not this signature is upregulated in the normal tissues around IBC tumors and if breast feeding or abrupt cessation of breast feeding contributes to the persistence of this signature in the normal breast will be investigated in future studies.

6 BIBLIOGRAPHY

1. Weber AF: Fundamentals of the Histology of Domestic Animals. Alfred Trautmann and Josef Fiebiger; trans. and rev. from the 8th and 9th German eds. of 1949 by Robert E. Habel and Ernst L. Biberstein. Ithaca, N. Y.: Comstock Pub., Cornell Univ. Press, 1952. 426 pp. Illus. $8.75. *Science* 1952, 116(3024):671-671.
2. Larson BL: Lactation: A Comprehensive Treatise : Vol. 4 : the Mammary Gland : Human Lactation : Milk Synthesis: Academic Press; 1978.
3. Macias H, Hinck L: Mammary gland development. *Wiley interdisciplinary reviews Developmental biology* 2012, 1(4):533-557.
4. Clarkson RW, Wayland MT, Lee J, Freeman T, Watson CJ: Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. *Breast cancer research : BCR* 2004, 6(2):R92-109.
5. Stein T, Morris JS, Davies CR, Weber-Hall SJ, Duffy MA, Heath VJ, Bell AK, Ferrier RK, Sandilands GP, Gusterson BA: Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3. *Breast cancer research : BCR* 2004, 6(2):R75-91.
6. Clarkson RW, Watson CJ: Microarray analysis of the involution switch. *Journal of mammary gland biology and neoplasia* 2003, 8(3):309-319.
7. Balkwill F, Mantovani A: Inflammation and cancer: back to Virchow? *Lancet* 2001, 357(9255):539-545.
8. McDaniel SM, Rumer KK, Biroc SL, Metz RP, Singh M, Porter W, Schedin P: Remodeling of the mammary microenvironment after lactation promotes breast tumor cell metastasis. *Am J Pathol* 2006, 168(2):608-620.

9. Blanchard A, Shiu R, Booth S, Sorensen G, DeCorby N, Nistor A, Wong P, Leygue E, Myal Y: Gene expression profiling of early involuting mammary gland reveals novel genes potentially relevant to human breast cancer. *Frontiers in bioscience : a journal and virtual library* 2007, 12:2221-2232.

10. Stein T, Salomonis N, Nuyten DS, van de Vijver MJ, Gusterson BA: A mouse mammary gland involution mRNA signature identifies biological pathways potentially associated with breast cancer metastasis. *Journal of mammary gland biology and neoplasia* 2009, 14(2):99-116.

11. Lyons TR, O'Brien J, Borges VF, Conklin MW, Keely PJ, Eliceiri KW, Marusyk A, Tan AC, Schedin P: Postpartum mammary gland involution drives progression of ductal carcinoma in situ through collagen and COX-2. *Nature medicine* 2011, 17(9):1109-1115.

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

13. Asselin-Labat ML, Vaillant F, Sheridan JM, Pal B, Wu D, Simpson ER, Yasuda H, Smyth GK, Martin TJ, Lindeman GJ *et al*: Control of mammary stem cell function by steroid hormone signalling. *Nature* 2010, 465(7299):798-802.

14. American Cancer Society: Cancer Facts & Figures 2014. In.; 2014.

15. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: Molecular portraits of human breast tumours. *Nature* 2000, 406(6797):747-752.

16. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S *et al*: Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA : the journal of the American Medical Association* 2006, 295(21):2492-2502.

17. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research : BCR* 2010, 12(5):R68.

18. Foulkes WD, Smith IE, Reis-Filho JS: Triple-negative breast cancer. *The New England journal of medicine* 2010, 363(20):1938-1948.

19. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, Lickley LA, Rawlinson E, Sun P, Narod SA: Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2007, 13(15 Pt 1):4429-4434.

20. Woodward WA, Cristofanilli M: Inflammatory breast cancer. *Seminars in radiation oncology* 2009, 19(4):256-265.

21. Anderson WF, Schairer C, Chen BE, Hance KW, Levine PH: Epidemiology of inflammatory breast cancer (IBC). *Breast disease* 2005, 22:9-23.

22. Hance KW, Anderson WF, Devesa SS, Young HA, Levine PH: Trends in inflammatory breast carcinoma incidence and survival: the surveillance, epidemiology, and end results program at the National Cancer Institute. *Journal of the National Cancer Institute* 2005, 97(13):966-975.

23. Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, Nasser V, Loriod B, Camerlo J, Tagett R, Tarpin C, Houvenaeghel G *et al*: Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer research* 2004, 64(23):8558-8565.

24. Bertucci F, Finetti P, Birnbaum D, Viens P: Gene expression profiling of inflammatory breast cancer. *Cancer* 2010, 116(11 Suppl):2783-2793.

25. Van Laere SJ, Ueno NT, Finetti P, Vermeulen PB, Lucci A, Robertson FM, Marsan M, Iwamoto T, Krishnamurthy S, Masuda H *et al*: Uncovering the molecular secrets of Inflammatory Breast Cancer biology: An integrated analysis of three distinct Affymetrix gene expression data sets. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2013.

26. Masuda H, Brewer TM, Liu DD, Iwamoto T, Shen Y, Hsu L, Willey JS, Gonzalez-Angulo AM, Chavez-MacGregor M, Fouad TM *et al*: Long-term treatment efficacy in primary inflammatory breast cancer by hormonal receptor- and HER2-defined subtypes. *Ann Oncol* 2014, 25(2):384-391.

27. Gaudet MM, Press MF, Haile RW, Lynch CF, Glaser SL, Schildkraut J, Gammon MD, Douglas Thompson W, Bernstein JL: Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. *Breast cancer research and treatment* 2011, 130(2):587-597.

28. RL A, K S, NT U, S K, W W, R E-Z, AM B: Epidemiological Risk Factors Associated with Inflammatory Breast Cancer Triple Negative Subtype. *Cancer Epidemiology Biomarkers & Prevention* 2012, 21(3):564.

29. Wu J: Background adjustment using sequence information. *R package version 20* 2005.

30. Sandberg R, Larsson O: Improved precision and accuracy for microarrays using updated probe set definitions. *BMC bioinformatics* 2007, 8.

31. Smyth GK: Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor.* edn.: Springer; 2005: 397-420.

32. Ernst J, Nau GJ, Bar-Joseph Z: Clustering short time series gene expression data. *Bioinformatics* 2005, 21 Suppl 1:i159-168.

33. Ernst J, Bar-Joseph Z: STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics* 2006, 7:191.

34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005, 102(43):15545-15550.

35. Yoshimura K, Aoki H, Ikeda Y, Fujii K, Akiyama N, Furutani A, Hoshii Y, Tanaka N, Ricci R, Ishihara T *et al*: Regression of abdominal aortic aneurysm by inhibition of c-Jun N-terminal kinase. *Nat Med* 2005, 11(12):1330-1338.

36. Djian P, Phillips M, Easley K, Huang E, Simon M, Rice RH, Green H: The involucrin genes of the mouse and the rat: study of their shared repeats. *Molecular biology and evolution* 1993, 10(6):1136-1149.

37. Tsuda H, Sakamaki C, Fukutomi T, Hirohashi S: Squamoid features and expression of involucrin in primary breast carcinoma associated with high histological grade, tumour cell necrosis and recurrence sites. *British journal of cancer* 1997, 75(10):1519-1524.

38. Garofalo C, Surmacz E: Leptin and cancer. *Journal of cellular physiology* 2006, 207(1):12-22.

39. Yin N, Wang D, Zhang H, Yi X, Sun X, Shi B, Wu H, Wu G, Wang X, Shang Y: Molecular mechanisms involved in the growth stimulation of breast cancer cells by leptin. *Cancer research* 2004, 64(16):5870-5875.

7 VITA

Arvind Bambhroliya was born in Gujarat, India, the son of Bhanubhai Bambhroliya and Savitaben Bambhroliya. After completing his higher secondary school at the Rajratna P. T. Patel Higher Secondary School in 2001, he entered at The Maharaja Sayajirao University of Baroda, India and received the degree of Bachelor of Medicine and Bachelor of Surgery in September, 2007. Then he worked as a medical officer for about a year at the Apara Nursing Home in Anand, India. Then he joined The University of Texas School of Public Health at Houston in 2008 to pursue the Master of Public Health program and completed it in August, 2011. From December 2011 to August 2012, he worked as a Research Technician in the Department of Neurology at The University of Texas Medical School. In August, 2012, he entered The University of Texas Graduate School of Biomedical Sciences to pursue the Master of Science program.

Permanent address:

Raghuvir Tenaments – 22/A

Vallabh Vidyanagar

Anand, Gujarat, India 388120