

COMPARISONS OF GENE EXPRESSION PATTERNS IN PROGRESSIVE BREAST  
CARCINOMA AND THE ADJACENT STROMAL MICROENVIRONMENT

by

John F. King  
A Thesis  
Submitted to the  
Graduate Faculty  
of  
George Mason University  
in Partial Fulfillment of  
The Requirements for the Degree  
of  
Master of Science  
Bioinformatics

Committee:

  
\_\_\_\_\_

Dr. Donald Seto,  
Thesis Co-Director

  
\_\_\_\_\_

Dr. Anna Baranova,  
Thesis Co-Director

  
\_\_\_\_\_

Dr. Iosif Vaisman,  
Committee Member

  
\_\_\_\_\_

Dr. Saleet Jafri,  
Department Chairperson

  
\_\_\_\_\_

Dr. Peter Becker, Associate Dean  
for Graduate Studies, College of  
Science

  
\_\_\_\_\_

Dr. Vikas Chandhoke, Dean,  
College of Science

Date: 12/7/07

Fall Semester 2007  
George Mason University  
Fairfax, VA

Comparisons of Gene Expression Patterns In Progressive Breast Carcinoma And The  
Adjacent Stromal Microenvironment

A thesis submitted in partial fulfillment of the requirements for the degree of Master  
of Science at George Mason University

By

John F. King  
Bachelor of Science  
Georgia Institute of Technology, 1983

Co-Director: Donald Seto, Associate Professor  
Department of Bioinformatics and Computational Biology

Co-Director: Anna Baranova, Associate Professor  
Department of Molecular and Microbiology

Fall Semester 2007  
George Mason University  
Fairfax, VA

## DEDICATION

This work is dedicated to my loving and patient wife, WandaLynne Souder King, who has made innumerable sacrifices for many years as I retooled my career by going to graduate school while working full time. It is also dedicated to the three beautiful children she has given me, Isaiah John, Elijah Lee, and Alanna Lynne, all of whom had to be told too many times that Daddy could not play with them today because he had to study. They are my greatest joy and my greatest accomplishments.

I would also like to dedicate this humble contribution to the unacceptably large number of brave women who have, are, and will battle the cruel and vicious disease of breast cancer.

## IN MEMORIUM

This work is also dedicated to the loving memory of my mother, Effie Frances Bridges King, who was taken from us way too early at the age of forty-eight by cancer. Because of this dread disease, she saw none of her four children marry and never experienced the joy of holding a grandchild. I can only hope she is smiling down upon this effort and that it is pleasing in her sight.

## ACKNOWLEDGMENTS

This work would not have been possible without the funding of the *Susan G. Komen for the Cure* organization, which is dedicated to finding a cure for breast cancer. I would like to thank my advisors, Dr. Donald Seto and Dr. Annna Baranova, for mentoring me and agreeing to a cross-departmental committee arrangement. I would also like to thank Dr. Lance Liotta and Dr. Emanuel Petricoin III, co-directors of the GMU Center for Applied Proteomics and Molecular Medicine, the principal investigators for the larger Komen project of which this work is a part.

The considerable amount of lab work that made this work possible was performed by Dr. Julia Wulfkuhle, Rosa Gallagher, Aybike Biredinc, and Michael Estep. Their long hours in the lab and talented hands are greatly appreciated. I also appreciate the computational support provided by Ganiraju Manyam who performed the p-value validation runs in R (using tools from Bioconductor, [www.bioconductor.org](http://www.bioconductor.org)), and the mentoring, encouragement, and technical review of SinChMAT by Dr. J. Patrick Vandersluis. Also, Dr. Jennifer Weller provided thoughtful review and valuable input into early manuscripts of the work. SinChMAT makes heavy use of an implementation of the Mann-Whitney U test written in C# (.NET) and generously donated to the open source domain by Sergey Bochkhanov via ALGLIB ([www.alglib.net](http://www.alglib.net)).

I am also much obliged to three individuals from the Volgenau School of Information Technology and Engineering who had significant impacts on my early GMU career: Dr. Alex Brodsky, Dr. Jeff Offutt, and Dr. Larry Kerschberg. All three provided invaluable advice and patient counsel as I casted about trying to figure out which direction to go in before I found my Bioinformatics calling. All were instrumental in my successful completion of two IT&E Graduate Certificates and actively helped me navigate several years of turbulent career changes as I realized my dream of becoming a software engineer. I am forever in their debt.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
ABSTRACT .....	ix
Introduction .....	1
Background and Motivation .....	1
Breast Cancer .....	3
Role of Stroma .....	6
Aims .....	9
Materials and Methods .....	14
Overview .....	14
Sample Generation .....	14
RNA Amplification, Labeling, and Hybridization .....	18
First-Round RNA Amplification .....	19
First cRNA Purification .....	21
Second Round RNA Amplification .....	22
Second cRNA Purification .....	23
Microarray Hybridization .....	25
Array Reading .....	29
Human Breast Cancer Biomarker Microarray .....	30
Results .....	34
Raw Data Generation .....	40
Background Corrections .....	44
Normalization .....	49
Statistical Analysis Approach .....	50
Computational Analysis Approach .....	53
Raw Image Results .....	57
Background Correction Results .....	59
Normalization Results .....	61
Statistical Analysis Results .....	62
Discussion .....	67
Analysis of Selected Genes .....	70
ASNS - Asparagine synthetase .....	71
ERBB2 - V-erb-b2 erythroblastic leukemia viral oncogene homolog 2 .....	73
GRB7 - Growth factor receptor-bound protein 7 .....	75
HMGB3 - High-mobility group box 3 .....	77
IGFBP3 and IGFBP5 - Insulin-like growth factor binding protein 3 and 5 .....	78
KRT18 and KRT19 - Keratin 18 and 19 .....	82
MKI67 - Antigen identified by monoclonal antibody Ki-67 .....	85
MYBL2 - V-myb myeloblastosis viral oncogene homolog avian-like 2 .....	86
WISP1 - WNT1 inducible signaling pathway protein 1 .....	88
Transforming Growth Factors .....	89
Validation and Sensitivity Analyses .....	90
Log Normalization .....	90
P-Value Computation .....	91

Group-wise Normalization.....	91
Background Corrections.....	93
Fold Changes with Absent / Present Threshold .....	93
Conclusions .....	96
References .....	99

## LIST OF TABLES

Table	Page
1. Specimens processed showing disease state and corresponding stroma .....	11
2. Specimens processed showing disease state and corresponding stroma .....	35
3. Summary of hypothesis testing results ( $p \leq 0.05$ ) .....	65
4. Summary of hypothesis testing results ( $p \leq 0.05$ ) following the DCIS With and Without IBC split out (both epithelia and stroma groups) .....	66
5. Fold change ( $\geq 1.5$ ) with Absent/Present (AP) threshold ( $\geq 1.2$ ) applied (Genes in <i>italics</i> had fold changes $\geq 2.0$ ) .....	94

## LIST OF FIGURES

Figure	Page
1. The progression of human female ductal breast cancer .....	1
2. Normal tissue architecture in the human breast (upper panel) showing the milk duct lined with epithelial cells (dark purple nuclei) and the surrounding mesenchymal stromal tissue. In invasive ductal breast carcinoma (lower panel), cells arising from the epithelial cells have invaded the stroma.4	4
3. Breast carcinoma cells invading the stroma .....	5
4. Pre-invasive, intraductal breast carcinoma completely filling the duct but not yet penetrating the basement membrane to invade the stroma .....	6
5. Eight condition groups and possible comparisons of interest .....	12
6. High-level view of study approach .....	14
7. Precision dissection of breast duct epithelial cells (left) and LCM transfer of cancer cell clusters (right) .....	16
8. TrueLabeling-PicoAMP™ Kit Overview.....	19
9. Overview of the Oligo GEArray® System process .....	26
10. Sample image of a single Oligo GEArray® captured by a Kodak 4000 MM Imager CCD camera. ....	30
11. Eight condition groups and possible comparisons of interest.....	36
12. Refined condition groupings, splitting out DCIS that occurs with IBC versus DCIS that does not occur with DCIS. (All non-adjacent “leap frog” comparisons omitted for clarity).....	39
13. Screenshot of SuperArray website used to process images.....	41
14. Preliminary image manipulation just before reading .....	42
15. Image ready to be (re)read.....	44
16. Misalignment of spot on conversion grid leads to compound error .....	45
17. Making spot-level adjustments prior to image reading .....	46
18. Bleeding of high-signal spots (e.g., controls, housekeeping genes) .....	47
19. Actual examples of bleeding.....	47
20. Main analysis page following array reading .....	48
21. High-level overview of the overall computational approach .....	53
22. Raw image examples .....	58
23. Single-Channel Microarray Analysis Tool (SinChMAT) heat map viewer .....	60
24. SinChMAT screen shot showing p-value results for DCIS Stroma-to-DCIS .....	63
25. P-value results following Benjamini-Hochberg multiple test correction .....	64
26. Overview of hypothesis testing results ( $p \leq 0.05$ ) with original groupings.....	68
27. Overview of hypothesis testing results ( $p \leq 0.05$ ) with DCIS With IBC and DCIS Without IBC split out. ....	69
28. ASNS normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	72

29. ERBB2 (NEU/HER-2) normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined) .....	74
30. GRB7 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	76
31. HMGB3 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	77
32. IGFBP3 normalized gene expression (line shows statistical significance) .....	79
33. IGFBP5 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	80
34. KRT18 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	83
35. KRT19 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	84
36. MKI67 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	86
37. MYBL2 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	88
38. WISP1 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined).....	89

## ABSTRACT

### COMPARISONS OF GENE EXPRESSION PATTERNS IN PROGRESSIVE BREAST CARCINOMA AND THE ADJACENT STROMAL MICROENVIRONMENT

John F. King, M.S.

George Mason University, 2007

Thesis Directors: Dr. Donald Seto and Dr. Anna Baranova

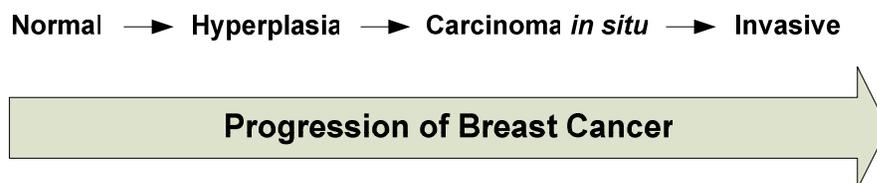
Human breast ductal carcinoma in situ (DCIS) is categorized as Stage 0 because it is noninvasive and limited to the duct lining. However, women diagnosed with DCIS have a 30-40% chance of developing invasive breast cancer (IBC) if it is left untreated. The breast tissue microenvironment of the surrounding stroma plays an important role in the malignant invasion and migration of tumor cells across the basement membrane, which separates the epithelial cells from the stroma in the normal breast. Far from being passive, the stroma plays an active role in invasiveness and perhaps throughout the entire progression of malignancy. Complex signaling networks, both intracellular and extracellular, are activated along with dramatic extracellular matrix (ECM) remodeling and growth factor release, which in turn leads to significant changes in cellular gene expression profiles. This study examines those gene expression profiles across the full range of breast cancer progression from normal to hyperplasia through DCIS and IBC, looking specifically at changes in gene expression between the cancerous epithelial tissue and the

surrounding stroma, using the recent advancement of laser capture microdissection to obtain highly purified, cell type specific samples.

# Introduction

## ***Background and Motivation***

Human breast cancer progresses through stages, forming a spectrum of classifications from normal (non-cancerous) to fully invasive (malignant), as shown in the figure below. These classifications apply equally to ductal and lobular breast cancer. The distinction between ductal versus lobular carcinomas is somewhat controversial because, strictly speaking, there are no anatomical grounds for the distinction. That is, both carcinomas ultimately derive from the terminal duct lobular unit (TDLU). The differences in carcinoma morphology are likely due to different mechanisms of carcinogenesis rather than to anatomical origin [1]. Nevertheless, the conventional terminology will be followed here, since these two types are the most common histological types of breast cancer. This study considers only ductal carcinoma.



**Figure 1.** The progression of human female ductal breast cancer

Hyperplasia is an overactive growth of cells lining the breast ducts (ductal hyperplasia). Hyperplasia is generally benign, unless it is diagnosed as “atypical,” which can progress into carcinoma. For ductal hyperplasia, this is termed Atypical Ductal Hyperplasia (ADH). Ductal carcinoma *in situ* (DCIS) is a proliferation of cells within the human breast ducts that appear malignant but have not breached the ductal basement membrane [2]. Once this barrier has been breached, the well known processes of invasion and metastasis generally follow.

In the United States, between 1975 and 2003, 394,891 cases of invasive breast cancer were diagnosed in women over 40 years old, and 59,837 cases of *in situ* (benign) breast cancer cases were diagnosed [3]. Although recent downward trends are encouraging, due mainly to reduction in hormone replacement therapy [4], breast cancer remains the second leading cause of cancer deaths in women (lung cancer is first), and the sixth leading cause overall. While DCIS is noninvasive and limited to surface cells, women diagnosed with DCIS have a 30-40% chance of developing invasive breast cancer (IBC) if left untreated [5].

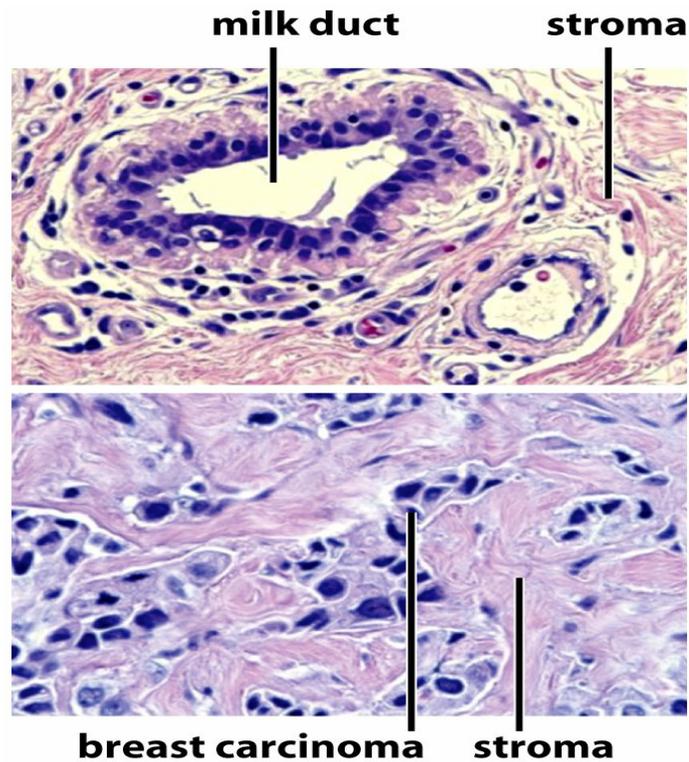
The study of gene expression in primary breast cancer tumors is complicated by two major factors [1]. First, breast cancer tissue consists of many different cell types (tumor, normal epithelial, stromal, adipose, and endothelial cells) and second, tumor cells are morphologically and genetically diverse [6]. For both of these reasons, laser capture microdissection (LCM) is crucial because it permits gene expression analyses from highly homogenous individual cell type populations [7,8]. All samples used in this study were obtained using LCM. The applications of microarray techniques to the study of breast cancer are well established, and have led to gene

expression profiling predicting clinical outcomes [9] and for molecular classifications of various subtypes *in situ* [10, 11]. Most recently, microarray techniques have been combined with laser capture microdissection for “more targeted” gene expression profiling, molecular pathway modeling, and for determining novel biomarkers for differentiating lobular versus ductal invasive carcinomas [1]. Here, an extension of previous work using highly targeted microarrays (breast cancer specific), laser capture microdissection (LCM), and tissue samples from both tumor and normal tissue microenvironments is followed in order to gain further insights into the conditions that might be related to *in situ* carcinomas becoming invasive. This work also broadens the scope of analysis by examining gene expression differences across the full range of mammary cancer progression, including atypical hyperplasia, while simultaneously considering both epithelial and stromal tissue types.

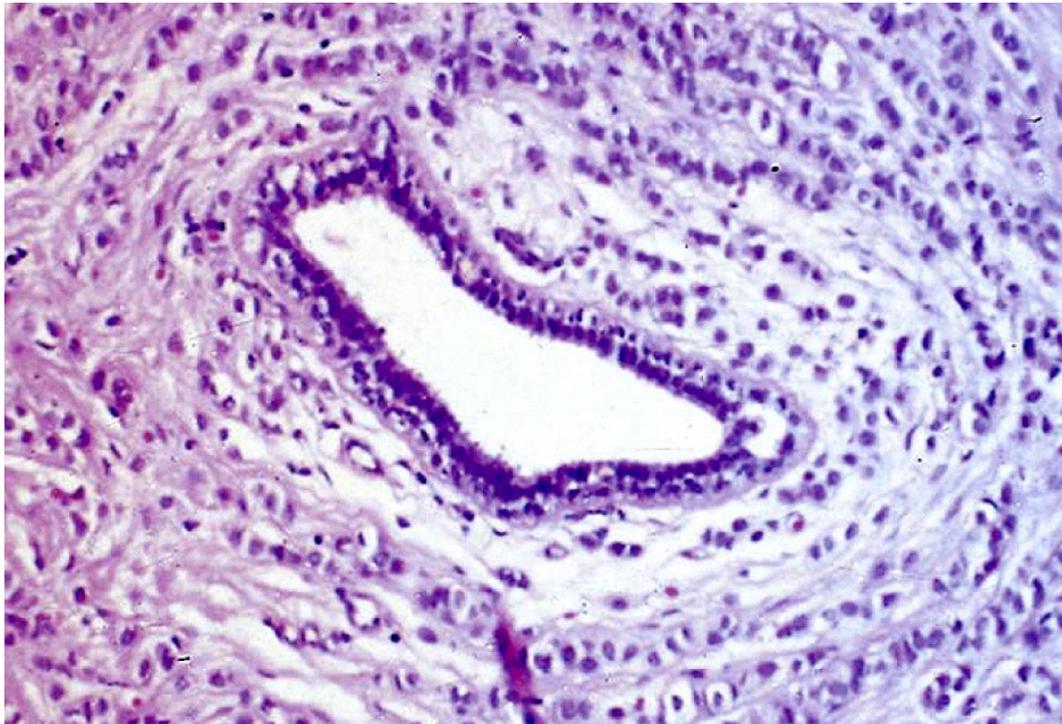
### ***Breast Cancer***

The majority of all human tumors arise from epithelial tissues [12]. In the human breast, the mammary gland contains numerous milk ducts lined with epithelial cells. These ducts are surrounded by mesenchymal tissue called stroma, which consists of fibroblasts, adipocytes, and collagen-based matrix (Figure 2, upper panel). When ductal breast carcinoma invades the stroma (Figure 2, lower panel, and Figure 3), cancer cells arising from the epithelial cells lining the normal ducts display abnormally large nuclei, and no longer form properly structured ducts. That is, they become much less differentiated, a hallmark of cancerous cells. Normally, a basement membrane (or basal lamina) separates the epithelial cells from the underlying, supportive stromal tissue. This basement membrane is a specialized type of extracellular matrix (ECM), formed largely from proteins secreted by the

epithelial cells. For cancer cells to become invasive, this basement membrane barrier must be breached.



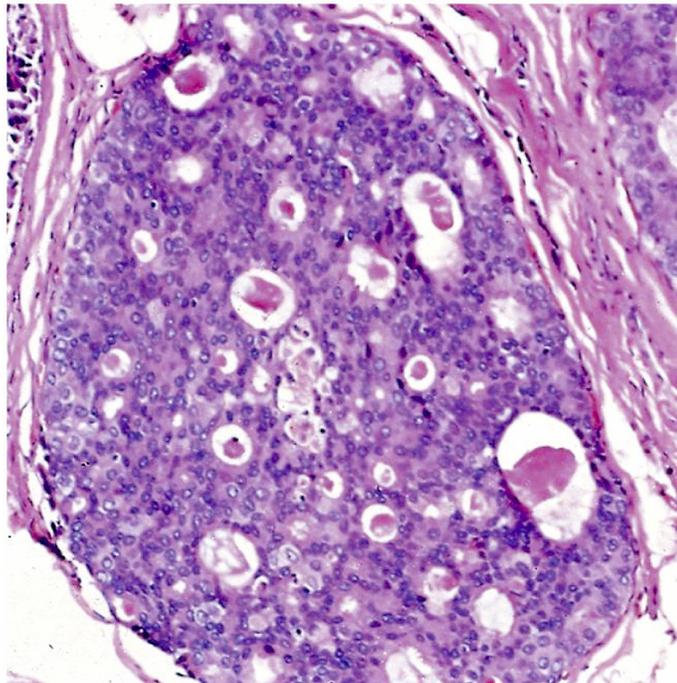
**Figure 2.** Normal tissue architecture in the human breast (upper panel) showing the milk duct lined with epithelial cells (dark purple nuclei) and the surrounding mesenchymal stromal tissue. In invasive ductal breast carcinoma (lower panel), cells arising from the epithelial cells have invaded the stroma. (Source: *The Biology of Cancer*, Robert A. Weinberg, Garland Science, 2007)



**Figure 3.** Breast carcinoma cells invading the stroma (Source: *The Biology of Cancer*, Robert A. Weinberg, Garland Science, 2007)

Understanding this transition from epithelial hyperplasia to dysplasia and to invasiveness is critical for the understanding of the progression to malignant breast carcinoma. Figure 4 shows an advancing intraductal breast carcinoma where dysplastic epithelial cancer cells have almost completely filled a duct and have caused it to swell to an abnormally large size, but have not yet broken through the surrounding basement membrane to invade the stroma. This leads to a question of what specific mechanisms can permit and facilitate the “tipping point” where the clearly abnormal, but not yet invasive, carcinoma crosses the line into invasiveness. For example, at the border of many carcinomas, epithelial cancer cells may change both shapes and gene expression profiles to take on attributes of nearby stromal

cells of mesenchymal origin. This “transdifferentiation” is called the epithelial-mesenchymal transition (EMT), and it enables the invasion by carcinoma cells through the basement membrane. The major focus of this work is to characterize gene expression profiles in the microenvironments before and after this transition, in order to understand better the specific mechanisms involved in breast carcinomas.



**Figure 4.** Pre-invasive, intraductal breast carcinoma completely filling the duct but not yet penetrating the basement membrane to invade the stroma (Source: The Biology of Cancer, Robert A. Weinberg, Garland Science, 2007)

### ***Role of Stroma***

Far from being an innocent bystander, the stroma surrounding breast ducts, and stroma that becomes intermixed with invasive carcinoma, plays an active role in the progression and tumorigenesis of breast cancer [13, 14]. Invasion is facilitated by the exchange of growth factors and cytokines as complex signaling networks develop

between tumor cells and the host stroma that modify the ECM, stimulate motility and migration, and promote proliferation and survival [15, 16].

Epithelial-mesenchymal interactions play a critical role in normal development and differentiation of the germ layers and organogenesis. Specifically, fibroblasts regulate the proliferation and differentiation of epithelial tissues [17], and transformed stroma is known to induce malignancy in both lung and mammary epithelia [18, 19]. Cancer exploits the normal, and reversible, mechanisms of epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET) to aid in its quest to become invasive and spread. EMT is triggered by external signaling and is characterized by cytoskeletal reorganization, loss of contact inhibition, and significant phenotypic changes. The major developmental signaling pathways mediated by, for example, RTK, Notch, Wnt, and TGF- $\beta$  provide primary inputs that drive EMT, resulting in mesenchymal derivatives with enhanced migratory and differentiation capabilities, which is essential for both normal morphogenesis of organs and tissues, as well as wound healing, and for cancer progression [20].

TGF- $\beta$  plays a particularly important and complex role, with carcinoma cells secreting abnormally high doses of bioactive TGF- $\beta$ , which sensitizes both carcinoma and surrounding stroma cells in an autocrine and paracrine fashion, leading to escape from the primary growth suppressive and pro-apoptotic responses to TGF- $\beta$ , but at the same time permitting the establishment of EMT [20]. Thus the current model of the role of TGF- $\beta$  during cancer progression is that it suppresses normal epithelial and benign adenoma cell growth while simultaneously promoting aggressive

carcinoma EMT, invasiveness, and metastasis [21] leading to its characterization as a “double-edged sword.”

Fibroblasts also seem to play a key role in the tumor microenvironment. Once activated, they are referred to as tumor-associated fibroblasts or cancer-associated fibroblasts (CAFs). The exact activation mechanism is not clear, but EMT of the associated cancer cells is likely involved. Particularly in human breast cancer cells that undergo EMT, the cancer appears to form its own nonmalignant stroma that functions reciprocally as a “feeder” of other carcinoma cells, regulating their proliferation [22]. Tumor stroma also contains so-called “activated myofibroblasts,” which are proposed to provide migratory cues for the metastatic carcinoma cells, one of them being TGF- $\beta$  [23] in a manner similar to wound healing.

Perhaps most intriguing are studies that suggest the stroma may actually initiate the invasive process. Invasion of stromal host cells, such as myofibroblasts, into the epithelial cancer compartment may *precede* epithelial cancer invasion into the stroma [23]. In a completely different vein, researchers also have shown that concurrent and independent genetic alterations (e.g., loss of heterozygosity (LOH) and genetic alterations on several chromosomes) in mammary stroma not only occur but that these changes may precede genotypic changes in the epithelial cells [24].

Clearly, the interactions between developing human breast cancer and its host microenvironment are significant and complex. Yet, despite intense recent study, many of the specific mechanisms, both genomic and proteomic, are still poorly characterized and warrant further study using the latest technology and techniques.

## Aims

This study is part of a larger collaborative effort funded under a Grant from the *Susan G. Komen for the Cure* organization. Participants include the George Mason University (GMU) Molecular and Microbiology Department, the GMU Center for Applied Proteomics and Molecular Medicine, and the INOVA Fairfax Hospital (Fairfax, VA). The stated aim of the overall study is to use the latest Laser Capture Microdissection (LCM) technology to obtain highly purified cell type specific samples of both breast epithelial tissue and the surrounding stroma and analyze gene expression and signal pathways for several cellular compartments:

- Ductal carcinoma in situ (DCIS)
- Stroma adjacent to DCIS
- Atypical ductal hyperplasia (ADH)
- Stroma adjacent to ADH
- DCIS accompanying invasive breast carcinoma
- Stroma surrounding DCIS and invasive breast carcinoma

The goal of the overall study is to analyze the disturbances in the local microenvironments surrounding ADH, DCIS, and DCIS accompanying invasive breast cancer using both genomic and proteomic microarray profiling with the specific aim

of distinguishing low grade, possibly non-progressive DCIS from DCIS that will likely progress into invasive cancer. The study presented here is the genomic component of the overall study.

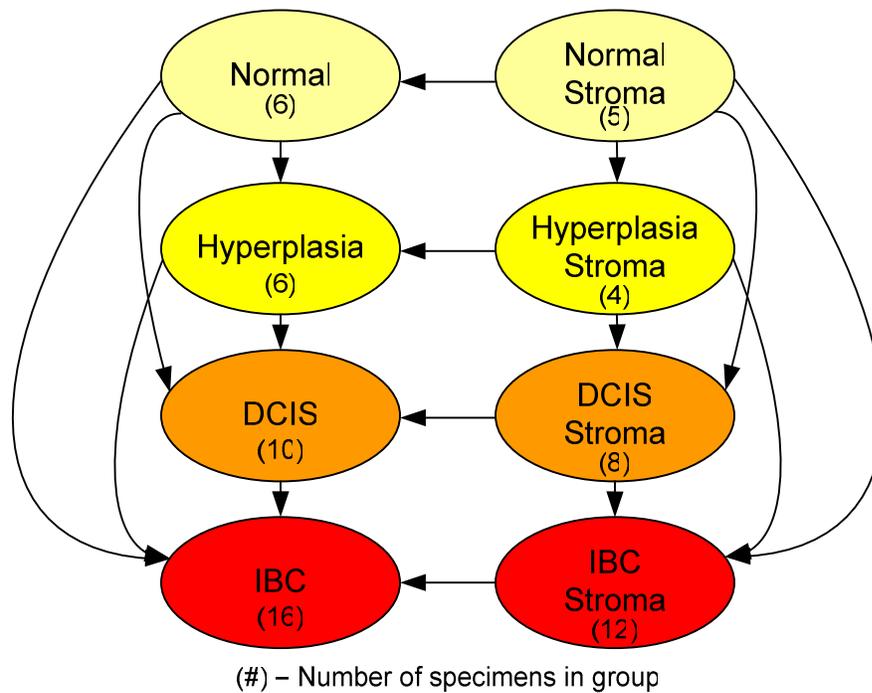
Approximately 65 samples were obtained from INOVA Fairfax Hospital covering the cellular compartments listed above. Table 1 below provides an overview of the specimens. Each gray cell in the table represents a specimen. The specimens are numbered by patient, and pairs of specimens (epithelial and stromal) are correlated by that number. In four cases, two categories of epithelial tissue were obtained, along with stroma, so in these cases there are numbered triplets instead of pairs. In the triplets, the stroma is classified according to the more advanced state of the associated breast tissue samples. Note that some stromal samples do not have associated breast tissue samples; however, the associated disease states for these are known. Also, not all epithelial samples have stromal partners, and not all stromal samples have the associated epithelial.

The aim of the genomics portion is to analyze gene expression between pairs of sample sets representing conditions of interest, for example a disease state and the associated stroma, or progressive disease states, or progressive stroma states. Many group pairings are possible, but not all make meaningful comparisons. The specific pairings analyzed are discussed in the Results section below.

**Table 1.** Specimens processed showing disease state and corresponding stroma

Patient	Normal	Hyperplasia	DCIS	IBC	Stroma	Comments	
1						Stroma => DCIS	1
2							2
6							6
10						No stroma	10
12						No stroma	12
13						No stroma	13
14						No stroma	14
15						No stroma	15
16						No stroma	16
18							18
19							19
22							22
23						Stroma for DCIS	23
24							24
25							25
28						Stroma => IBC	28
30						Stroma for IBC	30
33							33
34						No stroma	34
35							35
36							36
37							37
38							38
39						No stroma	39
41							41
43						Stroma => IBC	43
45							45
48							48
49							49
50						Stroma => IBC	50
52						Stroma for DCIS	52
54						Stroma for DCIS	54
55							55
56						Stroma for DCIS	56
57							57
59							59
62							62
63							63
64							64

Figure 5 shows a general overview of the condition groupings and therefore the sample sets. Arrows indicate comparisons of interest, including both adjacent and non-adjacent group pairings. Further, as the main goal of the overall study is to focus on DCIS with and without associated IBC, this (re)grouping is also of interest (not shown on figure). Refinements such as these are expanded upon and discussed further in the Results section.



**Figure 5.** Eight condition groups and possible comparisons of interest

To achieve the goals of the genomics portion of the overall study presented here, state-of-the-art laser capture microdissection is used to obtain high purity RNA samples, and targeted microarrays are used that are optimized for breast cancer investigations. Note that although gene expression comparisons between disease

states (e.g., DCIS versus IBC) have been investigated in many previous studies, most did not utilize laser capture microdissection (LCM), so these comparisons are worth examining as well, in addition to the main goal of epithelial versus stroma comparisons.

## Materials and Methods

### Overview

Figure 6 below shows an overview of the approach taken in this study. Laser capture microdissections are used to obtain high quality samples for RNA extraction.

TrueLabeling-PicoAMP™ 2.0 kits from SuperArray Bioscience Corporation are used for amplification and biotin labeling is used to prepare biotinylated cRNA target material.

The SuperArray Oligo GEArray® microarray system is used to perform microarray studies. The Oligo GEArray HybPlate Basic protocol is used for hybridization, and a CCD camera is used for chemoluminescent detection. Finally, software tools from the SuperArray GEArray Expression Analysis Suite website are used for preliminary data analysis, followed by additional detailed analysis and study of the results. Each step is described in further detail in this discussion.



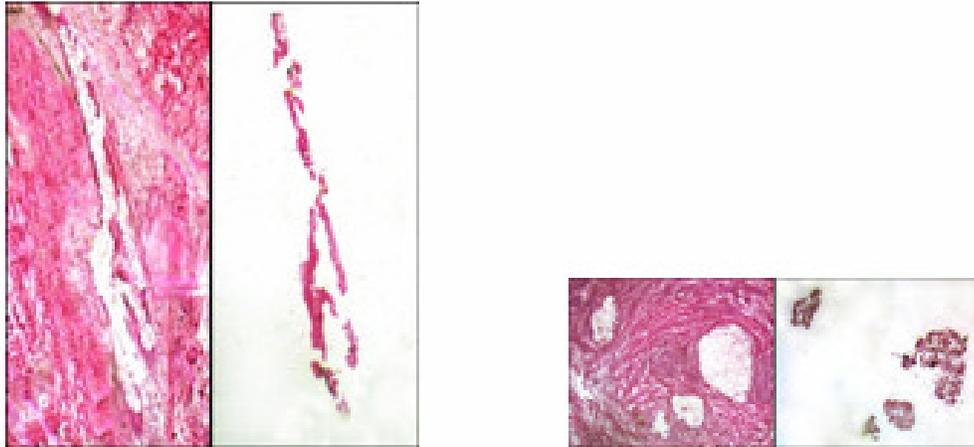
**Figure 6.** High-level view of study approach

### Sample Generation

The study approach begins with the isolation of high quality samples via laser capture microdissections (LCM) on all tissue specimens. LCM is a method for

obtaining cell-level precision in isolating samples from tissue sections [25]. The LCM performed for this study is done according to a recently published protocol [26]. Traditional sectioning approaches for obtaining cells for isolating total RNA can result in obtaining as little as 20% of the desired cell populations, introducing considerable noise into a microarray-based study. A major problem, particularly in studies of disease pathologies, is that the cells of interest (e.g., invading carcinoma cells) may be surrounded by highly heterogeneous tissue elements. In fact, the cells of interest may comprise a very small fraction of the total tissue biopsy sample, requiring careful extraction for testing to obtain low noise to signal results.

In LCM, a transparent transfer film is applied to the surface of the tissue section. Under a microscope, the operator views the thin tissue section through the glass slide on which it is mounted and locates microscopic clusters of the desired cells for the particular study. When the cells of choice are in the center of the field of view, the operator pushes a button which activates a near IR laser diode integrated with the microscope optics. The pulsed laser beam activates a precise spot on the transfer film immediately above the cells of interest. At this location, the film melts and fuses with the underlying cells. When the film is removed, the chosen cells are tightly held by the expanded polymer, while the rest of the tissue is left behind. Figure 7 below shows an example of the precision and resolution of samples achievable with this technology.



**Figure 7.** Precision dissection of breast duct epithelial cells (left) and LCM transfer of cancer cell clusters (right). (Source: NIH NCICGAP website, Introduction to Laser Capture Microdissection, <http://dir.nichd.nih.gov/lcm>)

First, the tissue samples are processed with the HistoGene™ LCM Frozen Section Staining Kit (Arcturus Bioscience, Inc., Mountain View, CA, Catalog #KIT0401) to prepare sections for the LCM. These stained sections are laser microdissected using an Arcturus ProCell IIe LCM device. Typically, 600 – 700 shots are fired on a give sample. The capture efficiency rate is on the order of 85%, so this number of shots conservatively captures approximately 500 cells of the tissue type of that sample. The actual capture is analyzed by eye, following the dissection to ensure this efficiency is achieved.

The resulting LCM caps are processed for RNA extraction using the PicoPure™ RNA Isolation Kit (Arcturus Bioscience, Inc., Mountain View, CA, Catalog #KIT0202 / KIT0204) and the associated protocol [27]. For each cap, dispense Extraction Buffer (XB) and incubate as follows. Pipette 50  $\mu$ L Extraction Buffer (XB) into a 0.5 mL microcentrifuge tube (Applied BioSystems Catalog #N8010611). Insert CapSure

Macro LCM Cap onto the microcentrifuge tube using an LCM Cap Insertion Tool. Invert the CapSure Cap–microcentrifuge tube assembly. Tap the microcentrifuge tube to ensure all Extraction Buffer (XB) is covering the CapSure Macro LCM Cap. Incubate assembly for 30 minutes at 42°C. Centrifuge assembly at 800 x *g* for two minutes to collect cell extract into the microcentrifuge tube. After centrifugation, the microcentrifuge tube contains the cell extract required to complete the protocol. Remove the CapSure Macro LCM Cap and save the microcentrifuge tube with the cell extract in it.

After extraction, the RNA is isolated as follows. First, pre-condition the RNA Purification Column. Pipette 250 µL Conditioning Buffer (CB) onto the purification column filter membrane. Incubate the RNA Purification Column with Conditioning Buffer for 5 minutes at room temperature. Centrifuge the purification column in the provided collection tube at 16,000 x *g* for one minute. Pipette 50 µL of 70% Ethanol (EtOH) into the cell extract from the RNA Extraction. Mix well by pipetting up and down, but do not centrifuge. Pipette the cell extract and EtOH mixture into the preconditioned purification column. The cell extract and EtOH will have a combined volume of approximately 100 µL.

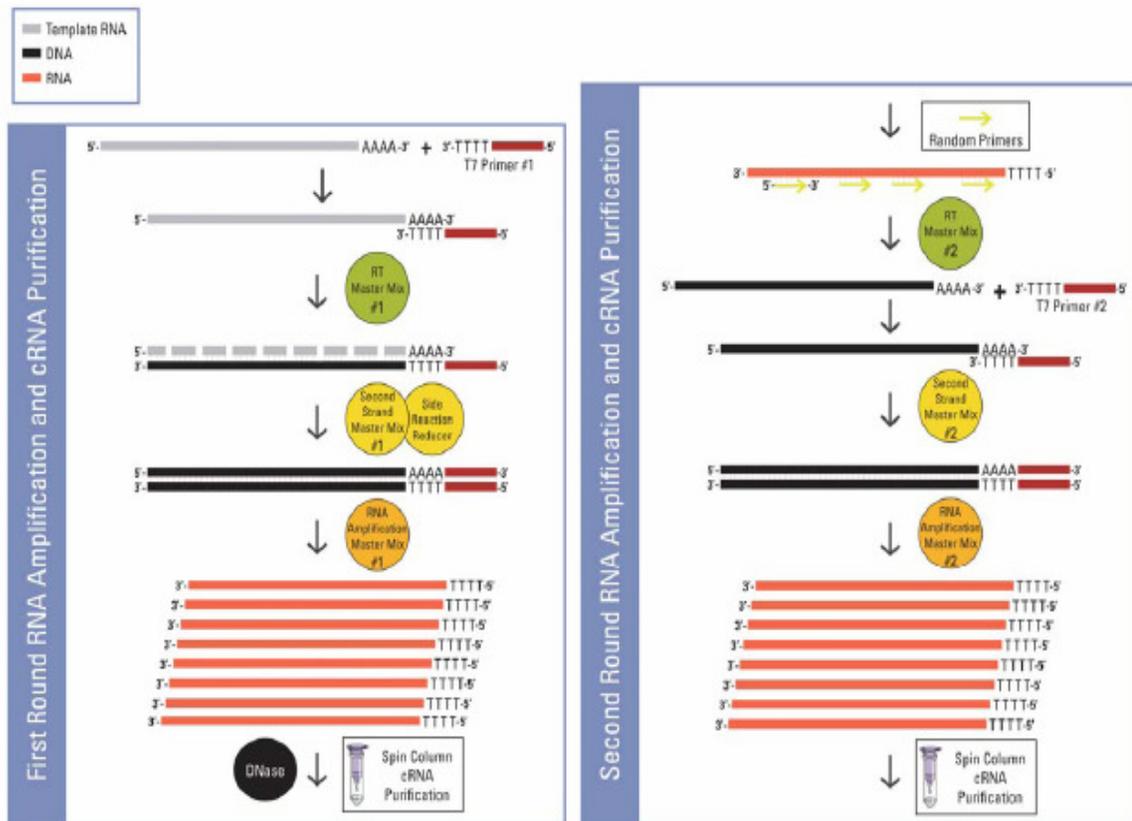
To bind RNA to the column, centrifuge for 2 minutes at 100 x *g*, immediately followed by a centrifugation at 16,000 x *g* for 30 seconds to remove flowthrough. Pipette 100 µL Wash Buffer (W1) into the purification column and centrifuge for one minute at 8,000 x *g*. Pipette 100 µL Wash Buffer 2 (W2) into the purification column and centrifuge for one minute at 8,000 x *g*. Pipette another 100 µL Wash Buffer (W2) into the purification column and centrifuge for two minutes at 16,000 x. Check

the purification column for any residual wash buffer. If wash buffer remains re-centrifuge at 16,000 x *g* for one minute. Transfer the purification column to a new 0.5 mL microcentrifuge tube provided in the kit. Pipette Elution Buffer (EB) directly onto the membrane of the purification column (Gently touch the tip of the pipette to the surface of the membrane while dispensing the elution buffer to ensure maximum absorption of EB into the membrane). Incubate the purification column for one minute at room temperature. Centrifuge the column for one minute at 1,000 x *g* to distribute EB in the column, then for one minute at 16,000 x *g* to elute RNA.

The isolated RNA is now ready for amplification, labeling, and hybridization.

### ***RNA Amplification, Labeling, and Hybridization***

Following RNA preparation, the RNA is amplified from picogram quantities (approximately 500 cells) using the TrueLabeling-PicoAMP™ amplification and labeling kit from SuperArray Bioscience Corporation (Frederick, MD). This kit is designed to amplify and label antisense RNA from picogram quantities of total RNA for hybridization to high-density genome-wide microarrays [28]. Very specific populations of cells are isolated through Laser Capture Microdissection (LCM), however for these samples traditional one-round amplification and labeling methods fail to yield enough target material for microarray applications, so the kit utilizes a two-round RNA amplification procedure (see Figure 8 below) to generate labeled antisense RNA (aRNA), also known as labeled cRNA target. The kit is optimized for use with the Oligo GEArray® arrays (SuperArray Bioscience Corporation, Frederick, MD), which were used in the hybridization step.



**Figure 8.** TrueLabeling-PicoAMP™ Kit Overview (Source: TrueLabeling-PicoAMP™ User Manual, Part#1021A, Version 1.3, July 9, 2007, SuperArray Bioscience Corporation, Frederick, MD.)

The protocol used for amplification, labeling, and hybridization is discussed below [28]. All quantities are per RNA sample.

### First-Round RNA Amplification

First, prepare the annealing mixture by mixing 50 – 500 pg of RNA in up to 2 µl of RNase-free H<sub>2</sub>O with 1.0 µl T7 Primer 1, followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate the mixture at 65°C for 5 min and then chill at 4°C or on ice for at least 1 minute. Centrifuge the mixture briefly (~2

sec) to collect the sample at the bottom of the tube. For the first strand cDNA synthesis, prepare the RT Master Mix using 1.75  $\mu$ l of RT Buffer and 0.25  $\mu$ l cDNA Synthesis Enzyme Mix per sample. Add 2  $\mu$ l of RT Master Mix 1 to each Annealing Mixture. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. The final volume is 5  $\mu$ l which is Incubated at 42 °C for 30 minutes.

To synthesize the second strand of cDNA, add 5  $\mu$ l of Second Strand Master Mix 1 to each completed First Strand cDNA Synthesis reaction (above). Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 65 °C for 10 minutes followed by 80 °C for 3 minutes. Chill at 4 °C or on ice for least 1 minute. For the side reaction reduction, add 1  $\mu$ l of the Side Reaction Reducer to each completed Second Strand cDNA Synthesis reaction from the previous step. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 37 °C for 10 minutes followed by 80 °C for 3 minutes and allow the tubes to cool to room temperature.

The actual RNA amplification is achieved by mixing 25  $\mu$ l RNA Polymerase Buffer 1 and 4  $\mu$ l RNA Polymerase Enzyme to create the RNA Amplification Master Mix 1, and then adding this mix to each completed Side Reaction Reduced sample from the previous step. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 37 °C for 8 hours followed by holding at 4 °C.

To cleanup the DNA, add 1  $\mu$ l of DNase I into each complete RNA Amplification reaction. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 37  $^{\circ}$ C for 15 minutes and immediately continue to the First cRNA Purification.

### **First cRNA Purification**

The first cRNA purification begins with binding the cRNA to the spin column. Set up a spin column in a collection tube for each sample. For each sample, prepare a separate cRNA Binding Mix in individual 1.5-ml RNase-free tubes by mixing 140  $\mu$ l Lysis & Binding Buffer with 140  $\mu$ l ACS-Grade 100% ethanol.

Add each entire reaction mixture volume to its own cRNA Binding Mix. Mix well with a pipettor up and down 5 to 6 times, but do not centrifuge, and immediately proceed to the next step for each sample. Carefully load each sample onto the center of its own Spin Column, and avoid spilling the mixture onto the rim of the spin column. Centrifuge for  $\sim$  30 sec at 10,000  $\times$   $g$ . Remove column from the tube, discard the flow-through, and put the column back into the Collection Tube.

Next, wash the spin column. Apply 200  $\mu$ l of Washing Buffer to each spin column. Centrifuge for  $\sim$  30 sec at 10,000  $\times$   $g$ . Apply 200  $\mu$ l 80% ethanol to each spin column. Prepare 80% ethanol by dilution of molecular-biology grade 100% ethanol with RNase-free H<sub>2</sub>O. Centrifuge for  $\sim$  30 sec at 10,000  $\times$   $g$ . Remove the column from the tube, discard the flow-through and put the column back into the Collection Tube. Centrifuge for  $\sim$  3 min at 16,000  $\times$   $g$ . Rotate the spin column 180 $^{\circ}$  and centrifuge for another 1 min at 16,000  $\times$   $g$ .

Finally, elute the cRNA from the spin column. Transfer each Spin Column to a fresh Elution Tube. To the center of each spin column, carefully add 12  $\mu$ l of room temperature RNase-free H<sub>2</sub>O by gently touching the silica membrane with the pipette tip. Evenly wet the membrane with a briefly vortex of the whole assembly. Incubate at room temperature for 2 min. Centrifuge for  $\sim$  1 min at 16,000 x g. The eluted volume is around 11  $\mu$ l. Use the entire elution volume for Second Round Amplification or store the purified cRNA at -80 °C.

### **Second Round RNA Amplification**

Begin the second round of RNA amplification by preparing the annealing mixture. For each sample, combine the eluted cRNA from the previous step ( $\sim$  11.0  $\mu$ l) with 1.0  $\mu$ l of the Random Primers solution in a sterile PCR tube. Mix the contents well followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 65°C for 5 min and chill at 4°C or on ice for at least 1 minute. Quickly, centrifuge briefly ( $\sim$ 2 sec) to collect the mixture at the bottom of the tube.

To perform the first strand cDNA synthesis, prepare RT Master Mix 2 by adding 7  $\mu$ l of RT Buffer and 1  $\mu$ l of RNase H Minus Reverse Transcriptase to a fresh tube. Add 8  $\mu$ l of RT Master Mix 2 to each Annealing Mixture. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 25 °C for 10 minutes, then at 37 °C for 60 minutes followed by 95 °C for 5 minutes. Chill at 4 °C or on ice for at least 1 minute.

To perform the second strand cDNA synthesis, add 1  $\mu$ l of T7 Primer 2 to each completed First Strand cDNA Synthesis reaction. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at

the bottom of the tube. The final volume is now 21  $\mu$ l. Incubate at 70  $^{\circ}$ C for 5 minutes followed by 42  $^{\circ}$ C for 5 minutes. Prepare Second Strand Master Mix 2 by adding into the tube 3  $\mu$ l of Second Strand Buffer 2 and 1  $\mu$ l DNA Polymerase Mix. Add 4  $\mu$ l of Second Strand Master Mix 2 to each reaction. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 60  $^{\circ}$ C for 10 minutes followed by 80  $^{\circ}$ C for 3 minutes. Cool to room temperature.

The actual RNA amplification is accomplished by first preparing the RNA Amplification Master Mix 2. In a fresh tube, combine 18  $\mu$ l RNA Polymerase Buffer 2, 7  $\mu$ l RNase-free H<sub>2</sub>O, 4  $\mu$ l RNA Polymerase Enzyme, and 6  $\mu$ l 10 mM Biotin-UTP. Add this RNA Amplification Master Mix 2 to each reaction. Mix well but gently with a pipettor up and down 2 to 3 times followed by a brief centrifugation to collect the mixture at the bottom of the tube. Incubate at 37  $^{\circ}$ C for 16 hours followed by holding at 4  $^{\circ}$ C.

### **Second cRNA Purification**

The second cRNA purification starts with binding the cRNA to the spin column. Set up a spin column in a collection tube for each sample. For each sample, prepare a separate cRNA Binding Mix in individual 1.5 ml RNase-free tubes by mixing 210  $\mu$ l Lysis & Binding Buffer and 210  $\mu$ l ACS-Grade 100% ethanol. Add each entire reaction mixture to its own cRNA Binding Mix. Mix well with a pipettor up and down 5 to 6 times and immediately proceed to the next step for each sample. Carefully load each sample onto the center of its own Spin Column. Centrifuge for  $\sim$  30 sec at 10,000  $\times g$ , remove column from the tube, discard the flow-through, and put the column back into the Collection Tube.

Next, wash the spin column. Apply 200  $\mu\text{l}$  Washing Buffer to each spin column. Centrifuge for  $\sim 30$  sec at  $10,000 \times g$ . Apply 200  $\mu\text{l}$  80% ethanol to each spin column. Prepare 80% ethanol by dilution of molecular-biology grade 100% ethanol with RNase-free H<sub>2</sub>O. Centrifuge for  $\sim 30$  sec at  $10,000 \times g$ . Remove column from the tube, discard the flow-through, and put the column back into the Collection Tube. Centrifuge for  $\sim 2$  min at  $16,000 \times g$ . Rotate the spin column  $180^\circ$  and centrifuge for another 30 sec at  $16,000 \times g$ .

To elute the cRNA from the spin column, transfer each Spin Column to a fresh Elution Tube. To the center of each spin column, carefully add 40  $\mu\text{l}$  of room temperature RNase-free H<sub>2</sub>O. Incubate at room temperature for 2 min. Centrifuge for  $\sim 1$  min at  $16,000 \times g$ , and store the eluted cRNA on ice.

Finally, the cRNA quantification and quality assessment is performed using UV spectrophotometry. Prepare a 1:20 dilution (Dilution Factor = 20) of the cRNA target by transferring a small aliquot (2 or 5  $\mu\text{l}$ ) of cRNA into an appropriate volume (38 or 95  $\mu\text{l}$ ) of RNase-free 10 mM Tris buffer pH 8.0. Determine the OD<sub>260</sub>.

Calculate the concentration and yield using the following equations:

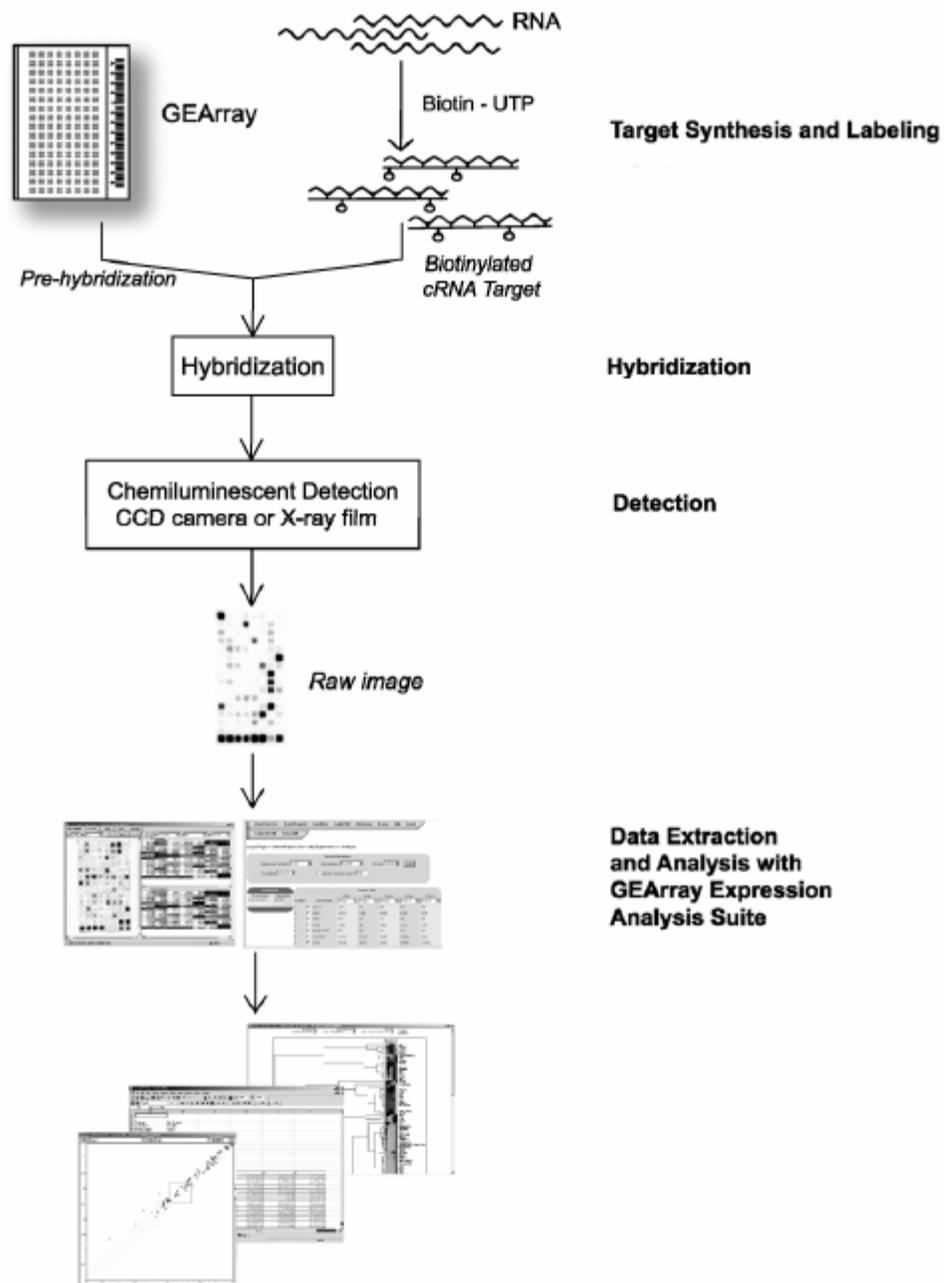
$$\text{Concentration } (\mu\text{g}/\mu\text{l}) = (\text{OD}_{260}) (40 \mu\text{g}/\text{ml}) (\text{Dilution Factor}) (1 \text{ ml} / 1000 \mu\text{l})$$

$$\text{Yield } (\mu\text{g}) = \text{Concentration } (\mu\text{g}/\mu\text{l}) \times 40 \mu\text{l}$$

The quantification measurements are done using the GeneQuant pro spectrophotometer (Biochrom Ltd, Cambridge, UK). This device provides RNA yield and uses  $A_{260}/A_{280}$  and  $A_{260}/A_{230}$  ratios for nucleic acid purity checks

## **Microarray Hybridization**

Approximately 4  $\mu\text{g}$  of cRNA are required for each array to produce a strong enough signal for good chemiluminescent imaging. An overview of the hybridization and array reading is shown in Figure 9 below.



**Figure 9.** Overview of the Oligo GEArray® System process (Source: Oligo GEArray® System User Manual, Part #1018A, Version 3.2, October 20, 2006, SuperArray Bioscience Corporation, Frederick, MD)

The microarray hybridization is done using the Oligo GEArray® System (SuperArray Bioscience Corporation, Frederick, MD, Part #1018A) using the following protocol [29] for HybPlate Basic.

First, prepare the Wash Solution 1 by mixing 10 ml 20X SSC, 5 ml 20 % SDS, and 85 ml ddH<sub>2</sub>O, and Wash Solution 2 by mixing 0.5 ml 20X SSC, 2.5 ml 20% SDS, and 97 ml ddH<sub>2</sub>O. Then perform the Buffer F dilution by diluting 5X Buffer F with ddH<sub>2</sub>O five-fold to prepare enough 1X Buffer F for at least one HybPlate. Eight ml of 1X Buffer F are required for washing each GEArray, so prepare at least 70 ml.

For pre-hybridization, place the GEArrays® into the GEArray® Multi-Chamber HybPlate using a pair of clean, flat forceps. Place only one GEArray® in each hybridization chamber with the bar code facing up along the right side of the array. Wet each GEArray® with 2 ml room temperature, RNase-free H<sub>2</sub>O, cover the HybPlate with the clear plastic cover and incubate for 3 minutes then remove the cover, pour off the water and gently tap the inverted HybPlate on paper towels. Pre-hybridize the array by adding 2 ml pre-warmed 60 °C GEHyb Hybridization Solution (without target) then replace the plastic cover. Gently shake for a few seconds until the membranes are floating free in the buffer and incubate for 1 to 2 hours at 60 °C. Pour off the prehyb buffer, and tap the inverted HybPlate on paper. Carefully transfer the GEArrays to a new HybPlate making sure that no pre-hyb solution touches the tops of the new HybPlate walls.

Now hybridize of the labeled target cRNA to the GEArray®. First, prepare the target hybridization mix by adding 2 µg cRNA target to a 2.0 ml aliquot of warm GEHyb

Hybridization Solution for each sample being analyzed, and keep this mixture at 60 °C. Then carefully add the appropriate Target Hybridization Mix to each hybridization chamber containing a GEMatrix® making sure not to splash any buffer to the wall tops. Gently shake the plate until the GEMatrices are floating free in the buffer. Remove the clear backing film from the GEMatrix® Multi-Chamber Seal. Align the adhesive portion of the seal with the chamber walls and set the seal down onto the top of the HybPlate chambers. Press in place by running a finger over the top of all the chamber walls. Using a tack or fine gauge needle, puncture one small vent hole over the middle of every chamber. Keep the HybPlate level during handling and avoid tipping. To insure a complete seal, carefully and firmly press the seal onto the HybPlate by running the blunt tip of a pen over the top of all the chamber walls and then folding the edges of the seal over the outside walls of the HybPlate. Incubate the GEMatrices overnight at 60 °C. Hybridization time should be limited to 24 hours.

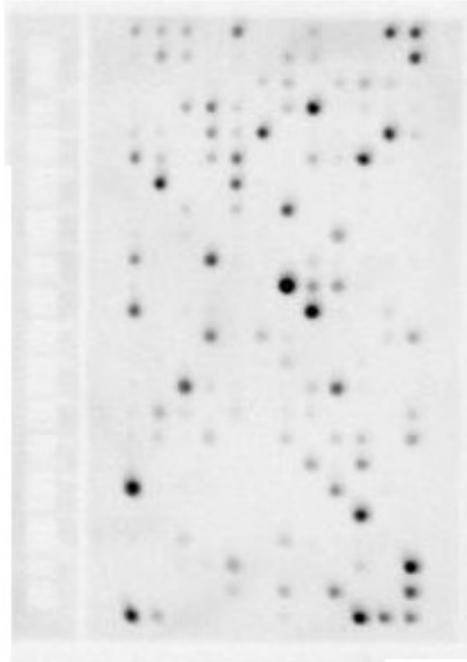
To wash the array, carefully remove the Seal by peeling away from one corner of the HybPlate and pour off the Target Hybridization Mix. Add 4 ml pre-warmed Wash Solution 1 to each chamber and gently swirl the HybPlate by hand until the membrane is floating freely. Incubate the HybPlate at 60 °C for 5 minutes, then pour off the buffer and tap the inverted HybPlate on paper towels. Repeat this wash step two more times. Add 4 ml pre-warmed Wash Solution 2 and gently swirl the HybPlate by hand until the membrane is floating freely. Incubate the HybPlate at 60 °C for 5 min then pour off the buffer and tap the inverted HybPlate on paper towels. Repeat this wash step two more times. Place the HybPlate on the lab bench at room temperature.

For detection, use the Chemiluminescent Detection Kit (D-01). Dilute an aliquot of the AP-Streptavidin stock 1:8000 in GEAblocking Solution Q (2  $\mu$ l into 16 ml) at room temperature. Add 2 ml dilute AP-Streptavidin to each GEMatrix®, gently swirl the HybPlate by hand and incubate for exactly 10 minute on the bench top at room temperature. Pour off the buffer and tap the inverted HybPlate on paper towels. Add 4 ml room temperature 1X Buffer F to each chamber, gently swirl the HybPlate by hand until the membrane is floating freely and incubate for 5 min, then pour off buffer and tap the inverted HybPlate on paper towels. Repeat this wash step 3 more times.

Finally, add 4 ml room temperature Buffer G to each GEMatrix, and incubate for  $\sim$ 1 min. Pour off Buffer G, add 1.0 ml CDP-Star® and incubate for 5 min. It is very important to have the membrane covered evenly with the CDP-Star substrate. Pour off the CDP-Star®, tap the inverted tray on paper towels and immediately acquire the GEMatrix® chemiluminescent image.

### ***Array Reading***

The arrays are read using a Kodak 4000 MM Imager CCD camera. The images are captured using the Kodak Molecular Imaging software (Version 4.04) running on an attached personal computer workstation. A sample image is shown below. A full HybPlate would contain eight of these images, in a grid four across by two down. Note the dark images on two of the corners. These spots represent housekeeping genes for potential use as controls and to provide image orientation and alignment for the image conversion software.



**Figure 10.** Sample image of a single Oligo GEArray® captured by a Kodak 4000 MM Imager CCD camera.

### ***Human Breast Cancer Biomarker Microarray***

The specific microarray used in the procedures described above is the Oligo GEArray® Human Breast Cancer Biomarker Microarray (HybPlate Format, Catalog Number EHS-402). This array has a nylon membrane matrix and is manufactured using non-contact printing with 60-mer oligo probe sets [30]. The array profiles the expression of 264 genes useful as molecular markers in breast cancer diagnosis and prognosis. The genes in the diagnosis markers group are highly associated with breast cancer [31]. The complete set of genes included on the array are summarized and functionally grouped below [32]. Note that for the purposes of this

listing, some genes may appear in more than one category/subcategory, and they are grouped overall by diagnostic versus prognostic potential.

## **Potentially Diagnostic Markers**

### **Cell Cycle**

Cell Cycle Arrest and Checkpoint: MYC, RB1, TP53.

Negative Regulation of the Cell Cycle: ATM, BAX, BRCA1, EGFR, ESR1, NME1, PTEN, RB1, TP53.

Regulation of the Cell Cycle: BCL2, BRCA2, CCND1, CCNE1, CDK4, FGF3, FGF8, IGF2, MAPK3, PCNA, PRKCA, TGFA, TGFB1, TGFB2, TGFB3, VEGF.

DNA Replication: CDK2, EGF, IGF1, PCNA.

### **Cell Growth and Proliferation**

Growth Factors and Cytokines: BMP6, CSF1, CSF3, EGF, FGF18, FGF3, FGF8, IGF1, IGF2, TGFA, TGFB1, TGFB2, TGFB3, TNF, VEGF.

Positive Regulation of Cell Proliferation: CDK2, CSF1, CSF3, EGF, FGF18, FGF3, IGF1, VEGF.

Negative Regulation of Cell Proliferation: BCL2, NME1, ODZ1, PLG.

Regulation of Cell Growth: ESR2, IGFBP3, TP53, TSG101,

Other Genes Involved in Cell Growth and Proliferation: AR, BRCA1, CDK4, EGFR, ERBB2, ERBB4, ESR1, MYC, PCNA, PRKD1, PRL.

### **Cell Differentiation**

CSF1, IGFBP3, TP53.

### **Apoptosis**

Induction of Apoptosis: BAX, MX1, PRKCA, PRKCE, TP53.

Anti-apoptosis: AKT1, BAG3, BCL2, BCL2L1, PRKCZ, TGFB1, TNF.

Other Apoptosis Genes: BRCA1, IGFBP3, VEGF.

### **DNA Repair**

ATM, BRCA1, BRCA2, PCNA, RAD51, TP53, XRCC3.

### **Angiogenesis Factors**

FGF3, VEGF.

### **Cell Adhesion Molecules**

CD34, CDH1, CTNNA1, ITGB3, PECAM1.

### **Extracellular Matrix (ECM) Molecules**

ALB, BRCA1, BRCA2, COL4A2, CSF3, CTSD, EGF, ERBB2, FGF18, FGF3, FGF8, IGF1, IGF2, IGFBP3, INS, KLK13, MMP11, MMP9, ODZ1, PRL, SERPINE1, SHBG, TGFA, VEGF.

### **Protein Kinases**

AKT1, ATM, CDK2, CDK4, EGFR, ERBB2, ERBB3, ERBB4, MAPK3, PDPK1, PRKCA, PRKCB1, PRKCD, PRKCE, PRKCG, PRKCZ, PRKD1, PRKD2, SRC, TYK2.

### **Protein Phosphatases**

IGFBP3, PTEN.

### **Transcription Factors and Regulators**

AR, BRCA1, BRCA2, CTNNB1, EGR3, ESR1, ESR2, FOS, JUN, MYC, NR4A1, PCNA, PGR, RB1, SP1, TNF, TP53, TSG101.

### **Proteases and Protease Inhibitors**

CTSB, CTSC, CTSD, CTSE, CTSL2, KLK13, MMP11, MMP9, PCSK6, PLG, SERPINE1.

### **Other Potential Diagnostic Markers**

ABCB1, ABCG2, AKAP1, CEACAM5, CYB5, CYC1, CYP19A1, GSTM1, GSTM3, KRT18, KRT19, MIB1, MUC1, MUC19, PALM2-AKAP2, VIM.

## **Potentially Prognostic Markers**

### **Cell Cycle**

Cell Cycle Arrest and Checkpoint: CCNE2.

Negative Regulation of the Cell Cycle: ESR1, EXT1.

Regulation of the Cell Cycle: BCL2, CCNB1, CCNB2, CDC25B, CENPF, MKI67, MYBL2, PCTK1, PSMD2, TGFB3, VEGF.

DNA Replication: MCM6, ORC6L, RFC4, RRM2.

Other Cell Cycle Genes: BIRC5, BUB1, CKS2, MAD2L1, SMC4L1, STK6.

### **Cell Growth and Proliferation**

Growth Factors and Cytokines: ESM1, FGF18, TGFB3, VEGF.

Positive Regulation of Cell Proliferation: CDC25B, FGF18, FLT1, VEGF.

Negative Regulation of Cell Proliferation: BCL2, BTG2.

Regulation of Cell Growth: CHPT1, ESM1, IGFBP5, WISP1.

Other Genes Involved in Cell Growth and Proliferation: BUB1, CKS2, ESR1, MAPRE2, MKI67.

### **Cell Differentiation**

NDRG1.

### **Apoptosis**

Anti-Apoptosis: BAG1, BCL2, BIRC5, BNIP3, MYBL2.

Other Apoptosis Genes: RAD21, STK3, VEGF.

### **DNA Repair**

BTG2, RAD21.

### **Angiogenesis Factors**

FLT1, VEGF.

**Cell Adhesion Molecules**

WISP1.

**Extracellular Matrix (ECM) Molecules**

ADM, COL4A2, CP, ESM1, FGF18, FLT1, IGFBP5, MATN3, MMP11, MMP9, RBP3, TFRC, VEGF, WISP1.

**Protein Kinases**

BUB1, CCNE2, CDC42BPA, CKS2, FLT1, MELK, PCTK1, STK3, STK32B, STK6.

**Protein Phosphatases**

CDC25B, MTMR2.

**Transcription Factors and Regulators**

BTG2, ESR1, EZH2, HMGB3, IVNS1ABP, KIAA1442, MCM6, MLLT10, MYBL2, PGR, PIR, SEC14L2, TBX3, TRIP13.

**Proteases and Protease Inhibitors**

BIRC5, CTSL2, GGH, MMP11, MMP9, PCSK6, PITRM1, RBP3, TFRC, UCHL5.

**Other Potential Prognostic Markers**

ACADS, ALDH4A1, ALDH6A1, AP2B1, ASNS, ASPM, BBC3, BM039, C20orf103, C20orf28, C20orf46, CA9, CD68, CENPA, CIRBP, CTPS, DCK, DEGS, DEPDC1, DKFZP434B168, DKFZp762E1312, DLG7, ECT2, EGLN1, EIF2C2, ERP70, EVL, FBP1, FBXO31, FBXO5, FGD6, FLJ10134, FLJ10156, FLJ10511, FLJ10901, FLJ12150, FLJ21924, FLJ22341, FUT8, GBE1, GCN1L1, GMPS, GNAZ, GPR126, GPSM2, GRB7, GSTM1, GSTM3, HRASLS, HRB, IHPK2, ITR, KIAA0882, KIAA1181, KIAA1217, KIAA1324, KIAA1683, KIF14, KIF21A, KIF3B, KNTC2, KRT18, LCHN, LGP2, LOC388134, LOC56901, LYRIC, M160, MCCC1, MGAT4A, MIR, MLF1IP, MRPL13, MS4A7, MYRIP, NMB, NMU, NUSAP1, ODZ3, OXCT, PALM2-AKAP2, PAQR3, PECI, PEX12, PFKP, PGK1, PIB5PA, PLEKHA1, PRAME, PRC1, PRO2000, PSMD7, PTDSS1, PTPLB, QDPR, RAB27B, RAB6B, RAI2, RAMP, RASL11B, RPS4X, RRAGD, SACS, SCUBE2, SERF1A, SLC2A3, SLC7A1, Spc25, ST7, STMN1, STX1A, SYNCRIP, TK1, TMEFF1.

Also on the array are controls, two blanks, and artificial sequences (e.g., four spots with BAS2C - Biotinylated Artificial Sequence 2 Complementary sequence). The controls are (shown with number of spots):

**Controls**

RPS27A (3), GAPDH (3), B2M (3), 18SrRNA (1), HSP90AB1 (1), and ACTB (1)

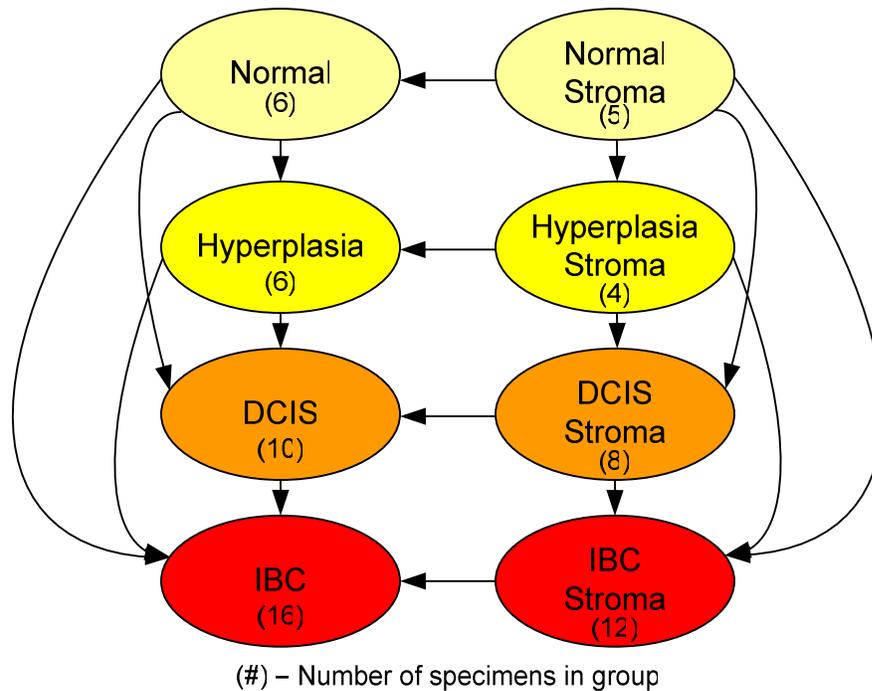
## Results

For convenience, the table of specimens and condition groupings from the Aims section is given again here to set the stage for presenting the results (Table 2 below). The specimens are labeled (not shown in the table) according to the pattern: AA##, where 'AA' is **N** for Normal, **H** for Hyperplasia, **DC** for DCIS, and **I** for IBC, and ## is the patient number. For all stroma samples, 'AA' is **S**, regardless of the associated breast tissue category. Thus, for Patient 1, which happens to be a triplet, there are three specimens: H1, DC1, and S1. Note that the classification of stroma with its associated breast tissue type cannot be inferred simply from the stroma sample label.

The specimens are split into condition groups for the purposes of analysis, resulting in eight distinct condition groups: Normal, Hyperplasia, DCIS, IBC, Normal stroma, Hyperplasia stroma, DCIS stroma, and IBC stroma. These groupings are used extensively throughout the analysis. Consideration of these groups also leads to many possible group-to-group comparisons, as illustrated in the Figure 11 following the table, which is just Figure 5 in the Aims section shown again here for convenience.

**Table 2.** Specimens processed showing disease state and corresponding stroma

Patient	Normal	Hyperplasia	DCIS	IBC	Stroma	Comments	
1						Stroma => DCIS	1
2							2
6							6
10						No stroma	10
12						No stroma	12
13						No stroma	13
14						No stroma	14
15						No stroma	15
16						No stroma	16
18							18
19							19
22							22
23						Stroma for DCIS	23
24							24
25							25
28						Stroma => IBC	28
30						Stroma for IBC	30
33							33
34						No stroma	34
35							35
36							36
37							37
38							38
39						No stroma	39
41							41
43						Stroma => IBC	43
45							45
48							48
49							49
50						Stroma => IBC	50
52						Stroma for DCIS	52
54						Stroma for DCIS	54
55							55
56						Stroma for DCIS	56
57							57
59							59
62							62
63							63
64							64



**Figure 11.** Eight condition groups and possible comparisons of interest

Although more comparisons are possible, the sixteen comparisons indicated in Figure 11 by the arrows are the main ones of interest. Comparisons of adjacent groups are of highest interest, with the other non-adjacent comparisons done for completeness and possible insights. Others, such as Normal breast tissue to DCIS stroma would not appear to add meaningful new information. Of high interest are the pairs composed of a disease condition and the associated stroma (e.g., DCIS-to-DCIS stroma), and pairs of stroma groups, since the role played by the stroma is the central theme here.

The specific arrays used and their groupings are as follows. For Normal (epithelium), six arrays were obtained, denoted N36, N39, N41, N57, N59, and N62. All but one

(N39) have an associated stromal pair, so the Normal Stroma group includes five arrays, denoted S36, S41, S57, S59, S62.

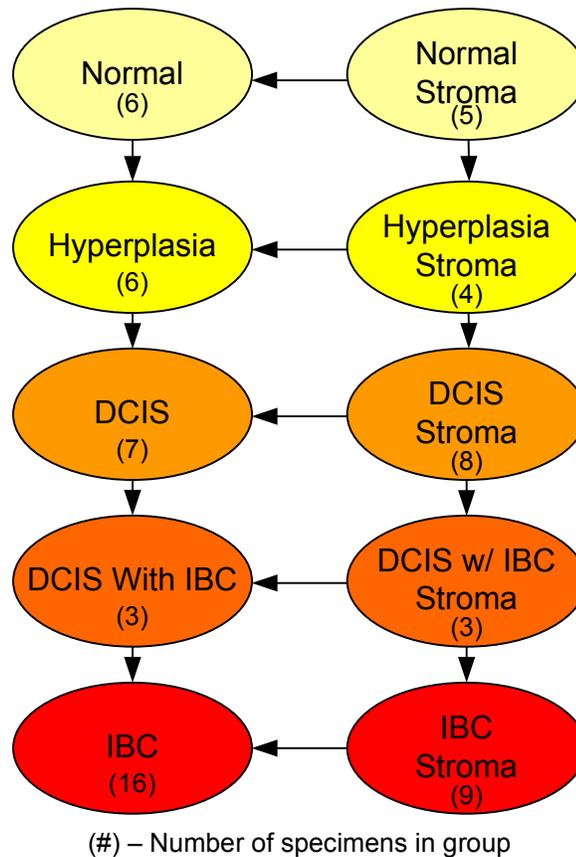
The Hyperplasia group includes six arrays, denoted H1, H2, H6, H12, H22, and H25, and all but one (H12) have a corresponding stroma sample. Also, Patient 1 is a triplet sample, comprised of Hyperplasia, DCIS, and Stroma samples. Thus, the Stroma sample (S1) has to be classified as Hyperplasia Stroma or DCIS Stroma; it cannot be classified as both simultaneously for statistical independence reasons in group-wise comparisons. The rule adopted for triplets is to classify the associated stroma sample as the more advanced or invasive type, so S1 is classified as DCIS Stroma, not Hyperplasia Stroma. Thus, the Hyperplasia Stroma group includes S2, S6, S22, and S25, for a total of four arrays versus the Hyperplasia (epithelium) group, which has six.

Ignoring the split out of DCIS samples in the three remaining triplets (Patients 28, 43, and 50, which have DCIS, IBC, and Stroma samples), the DCIS group includes arrays denoted DC1, DC16, DC24, DC28, DC34, DC43, DC48, DC49, DC50, and DC63. Following the rule of classifying associated stroma samples as the most advanced or invasive type, the stromal samples for triplets 28, 43, and 50 are classified as IBC Stroma, so the corresponding DCIS Stroma group is comprised of S1, S23, S24, S48, S49, S52, S54, and S56, which also reflects that DC16, DC34, and DC63 have no corresponding stromal sample. This group also reflects that, just as there are three DCIS epithelium "solo" samples, some stromal samples have no corresponding epithelium samples, namely S23, S52, and S54. These DCIS Stroma "solo" samples are known to be associated with DCIS from documentation

established during sample gathering. After accounting for all three of these factors, the DCIS Stroma group has a total of eight arrays with a relatively small intersection of patients in both epithelial and stromal groups. However, this small intersection has no adverse effects on group-wise p-value statistical calculations, since all samples within a given group are independent and no sample appears in more than one group.

The IBC epithelial group is the largest group with sixteen members. They are: I10, I13, I14, I15, I18, I19, I28, I33, I35, I37, I38, I43, I45, I50, I55, and I64. Several of these samples have no corresponding stromal samples, namely Patients 10, 13, 14, 15, and 64. The IBC Stroma group is composed of the remaining stromal pairs from the IBC samples, plus one additional solo stroma sample (Patient 30). The resulting IBC Stroma group thus has twelve members: S18, S19, S28, S30, S33, S35, S37, S38, S43, S45, S50, and S55.

As discussed in the Aims section above, a variation of special interest on the groupings shown in Figure 11 is to split out the DCIS samples into two subgroups: those associated with IBC and those that are not associated with IBC. With this refinement, the groupings and comparisons change as shown in Figure 12.



**Figure 12.** Refined condition groupings, splitting out DCIS that occurs with IBC versus DCIS that does not occur with DCIS. (All non-adjacent “leap frog” comparisons omitted for clarity)

Careful inspection of Figure 12 reveals that while the epithelial split out is straightforward (a single DCIS group of 10 becomes two subgroups of 7 and 3), the stromal group split out is more subtle. The DCIS With IBC Stroma group samples come out of the original IBS Stroma group, not the original DCIS Stroma group. To see why, refer back to the specimen table (Table 2) above, and recall there were three “triplet” samples that included both DCIS *and* IBC epithelial samples with a corresponding (single) stromal sample, for a given patient. Before splitting out DCIS With and Without IBC, these stromal samples were classified as IBC Stroma. After the split out, these stromal samples are reclassified as DCIS With IBC Stroma and

must be removed from the IBC Stroma group. The statistical requirement of independence requires that the same sample cannot be included in two groups that are compared against each other, so these stromal samples cannot be in both the DCIS With IBC Stroma and IBC Stroma groups.

Thus, following the DCIS split out, the resulting new groups are:

- DCIS Without IBC (DC1, DC16, DC24, DC34, DC48, DC49, DC63)
- DCIS With IBC (DC28, DC43, DC50, from the DCIS-IBC-Stroma triplet specimens of Patients 28, 43, and 50)
- DCIS Without IBC Stroma (unchanged)
- DCIS With IBC Stroma (S28, S43, and S50, which were formerly in IBC Stroma).

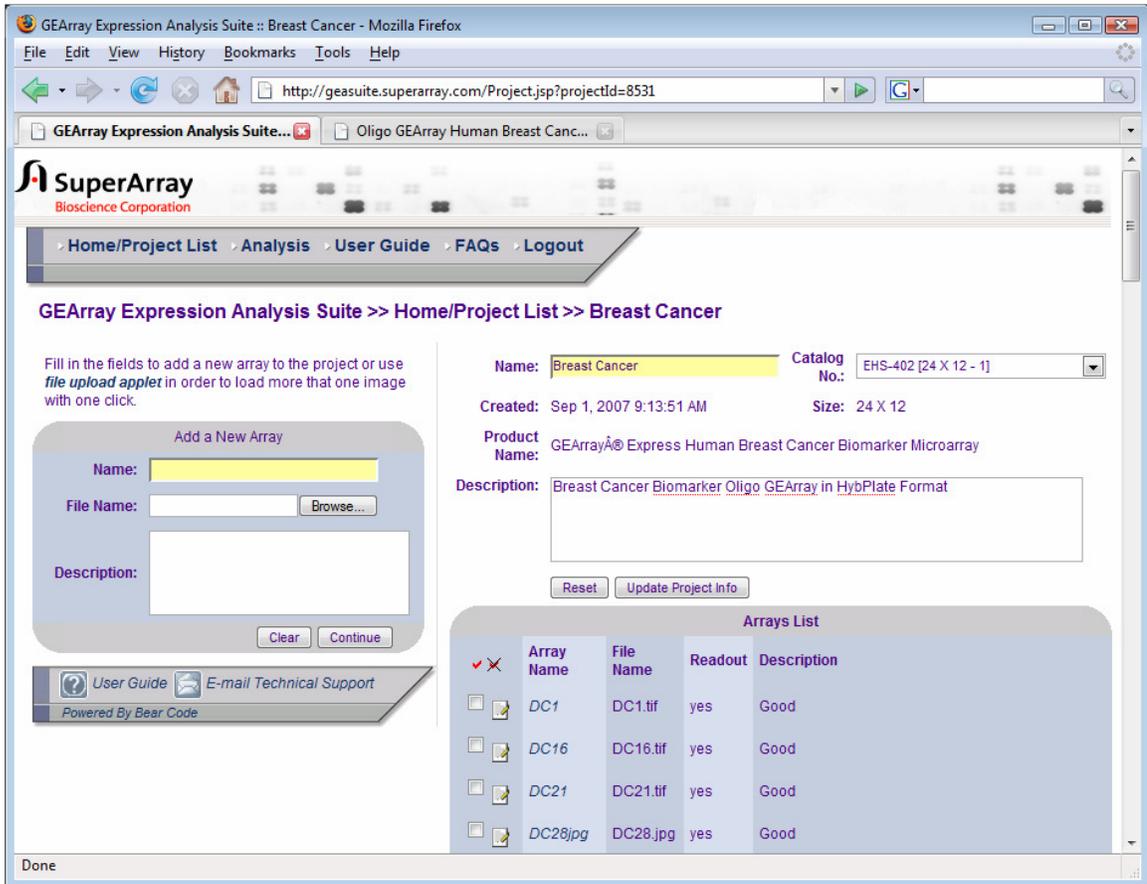
The net reduction in IBC Stroma samples, following reclassification, does not present a problem since this group was the second largest group.

### ***Raw Data Generation***

The analysis approach begins with converting the SuperArray hybridization array images to raw data. The SuperArray Bioscience kits used are designed to be “read” by software specifically designed for this purpose and provided by the vendor (via its website). The vendor tool is called the GEArray Expression Analysis Suite 2.0, and is available to SuperArray customers on a subscription basis at:

<http://geasuite.superarray.com/index.jsp>

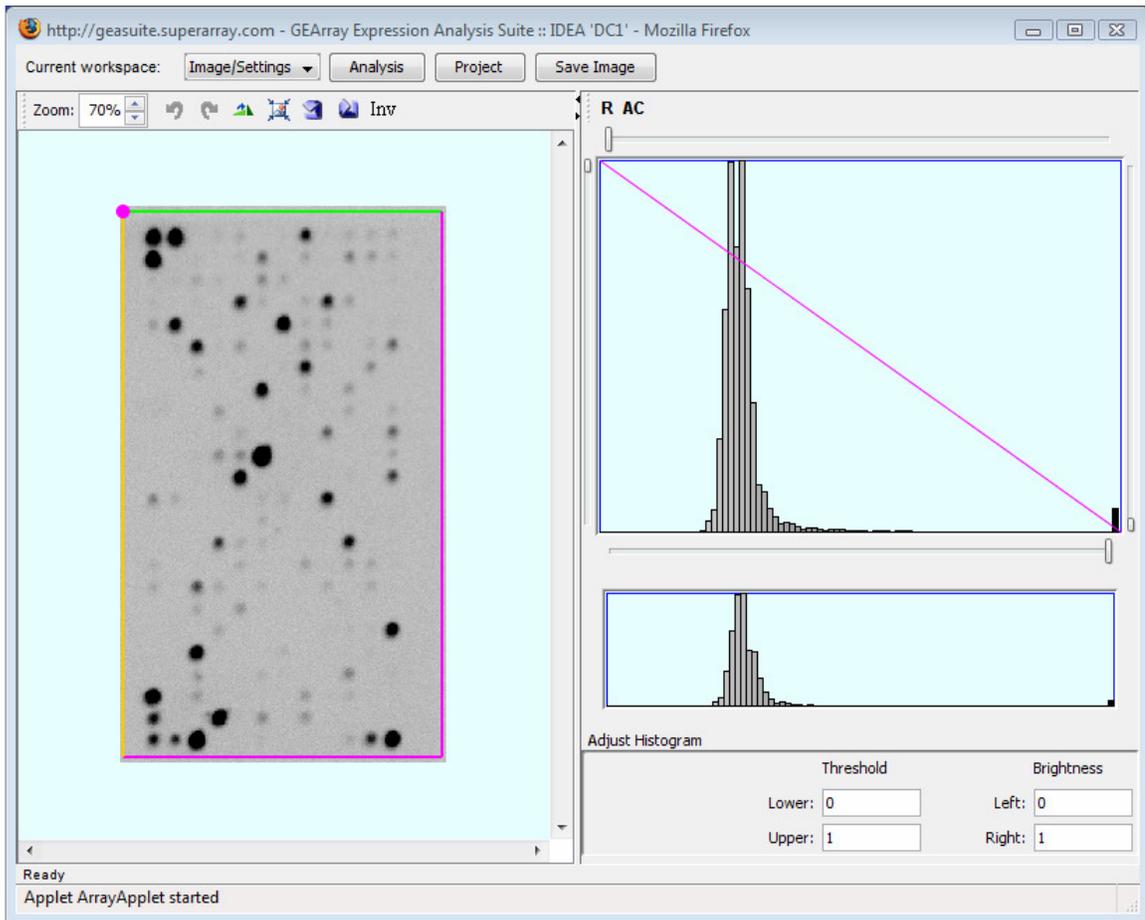
Each array image obtained from the CCD camera at the conclusion of the lab processes described above is uploaded to this website, and this step initiates the analysis phase of the study. Once the project is established on the website, a screen such as the one below is used to upload images to the site.



**Figure 13.** Screenshot of SuperArray website used to process images

The first step in converting images to data, following image upload, is to “read” the array using the automated vendor software to convert the images to raw data. This step is accomplished using a screen such as the one below. This screen can be used to adjust contrast and other image manipulations prior to the actual reading. This screen is also used for the important step of cropping the image tightly around the

corner spots, shown by high signaling controls spots. The image below shows the crop box before it has been used to crop down the image.

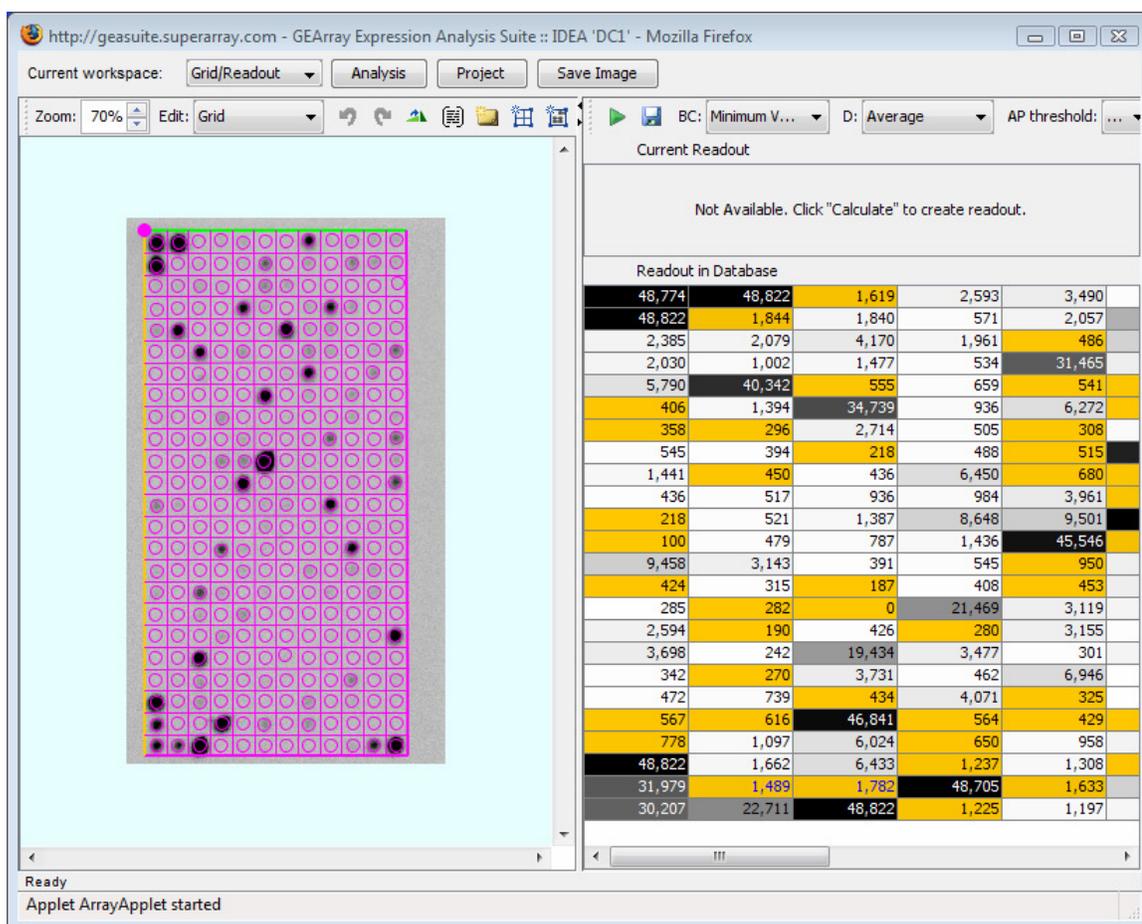


**Figure 14.** Preliminary image manipulation just before reading

At this step, care must be taken to ensure the array is aligned correctly (top-bottom, left-right), because the software uses the corner spots to orient the reading and assign intensities to known array locations (genes). Thus, if the array is reversed, either horizontally or vertically, the spot readings will be incorrectly assigned to the wrong genes. The small pink circle on the upper left of the crop box shows the origin, and spot reading beginning at that corner will be assigned to Position 1, 2, 3,

and so forth reading left to right for 12 spots before wrapping around to the next row, and so on vertically down the array.

Once the array image has been tightly cropped, the drop down control at the top of the screen is used to superimpose the grid, make adjustments (if needed), and generate the raw data as shown in the screenshot below. This step is the actual reading of the array. The screen below shows an array that has already been read (note "Readout in Database" to the right) but could be re-read if desired. Clicking on a spot on the grid image selects the corresponding spot on the tabular layout on the right, and vice versa.



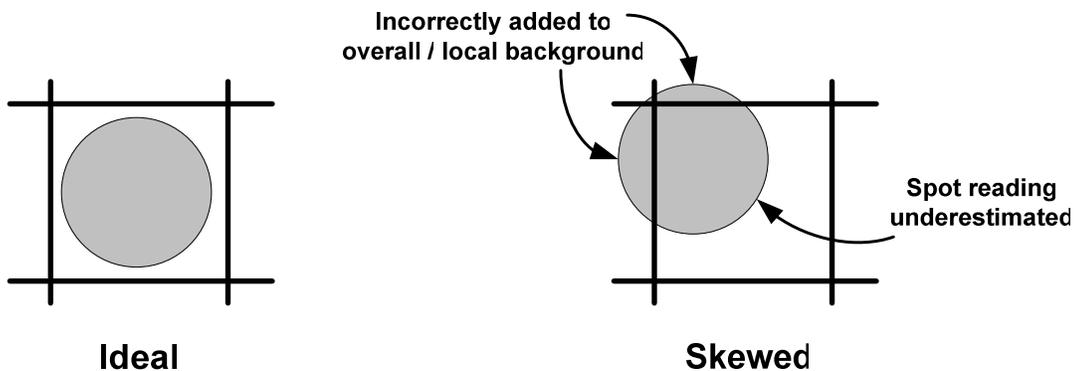
**Figure 15.** Image ready to be (re)read

## ***Background Corrections***

Subtracting out background exposure (noise) correctly, consistently, and accurately is crucial to correct microarray analysis. This step in the analysis can range from trivial to complex, depending on the quality and uniformity of the images obtained. The simplest case would be when the image has crisp, well-defined spots on a pure white background. The spots would vary from pure white (no signal, not even background) to black at the various grid locations across the array. In reality, most readings are a shade of gray with an intensity spanning the continuum from white to black. The vendor's software tools provide various means of performing the

background subtraction. The results of working with these options will be discussed in the Results section below.

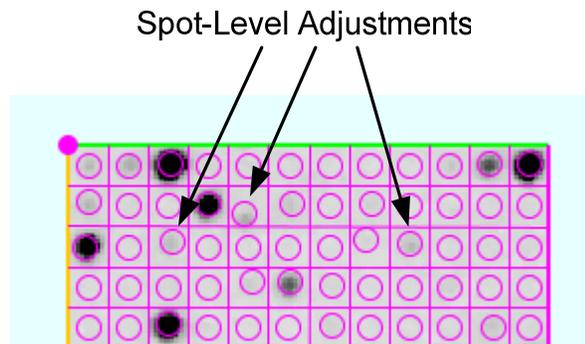
One of the most difficult problems to deal with is misalignment of the probe itself on the array with respect to the rest of the array grid and how this impacts background corrections. The figure below illustrates the misalignment background subtraction problem. When a spot is skewed relative to the grid, not only is the overall (or local) background correction artificially inflated, but also the individual spot reading is under counted. The misalignment has the effect of compounding the pixel-to-numerical conversion error by subtracting too much background from too little true reading. The vendor's tools allow for manual touch up and realignment to correct for these problems, but the analyst is always bound to the confines of the (fixed) grid. That is, the spot reading is required to be done within the grid cell, but the circular reading target can be moved around within the cell.



**Figure 16.** Misalignment of spot on conversion grid leads to compound error

When the spot is simply off center but still confined within the grid square, the vendor software permits spot-level adjustments to the reading by clicking and

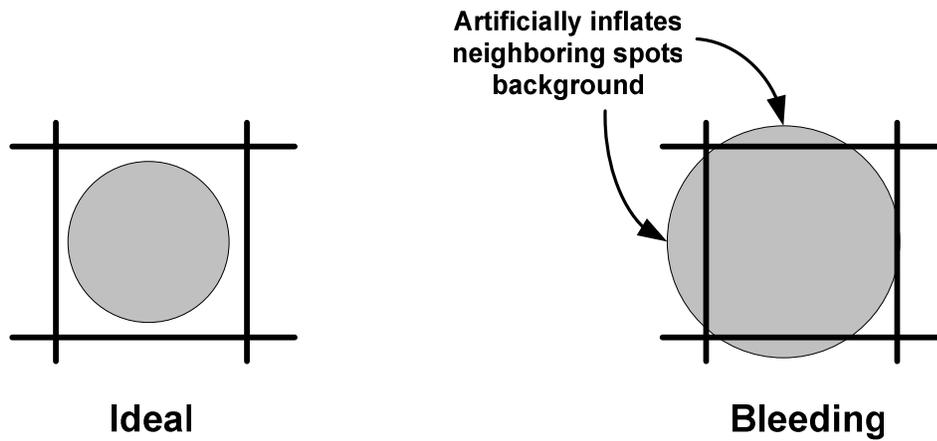
dragging individual target reading circles within the grid square. Adjustments of this type are shown in the figure below.



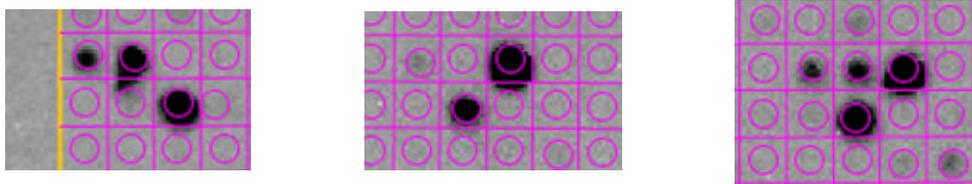
**Figure 17.** Making spot-level adjustments prior to image reading

Another issue of concern in image-to-data conversion and the associated background correction is the issue of “bleeding.” Bleeding refers to the situation where one spot location, typically a highly signaling control probe, has heavy hybridization and therefore a heavy signal. Depending on exposure time, the signal may “bleed” out of the confines of its designated grid square. The vendor software attempts to detect and flag these situations when they occur, but the background correction actions are still the responsibility of the user. Figures 18 and 19 below show a notional sketch and actual bleeding examples, respectively.

Time constraints prevented making spot-level adjustments to the data for bleeding situations in the results presented here. These adjustments would have to be made in future work.

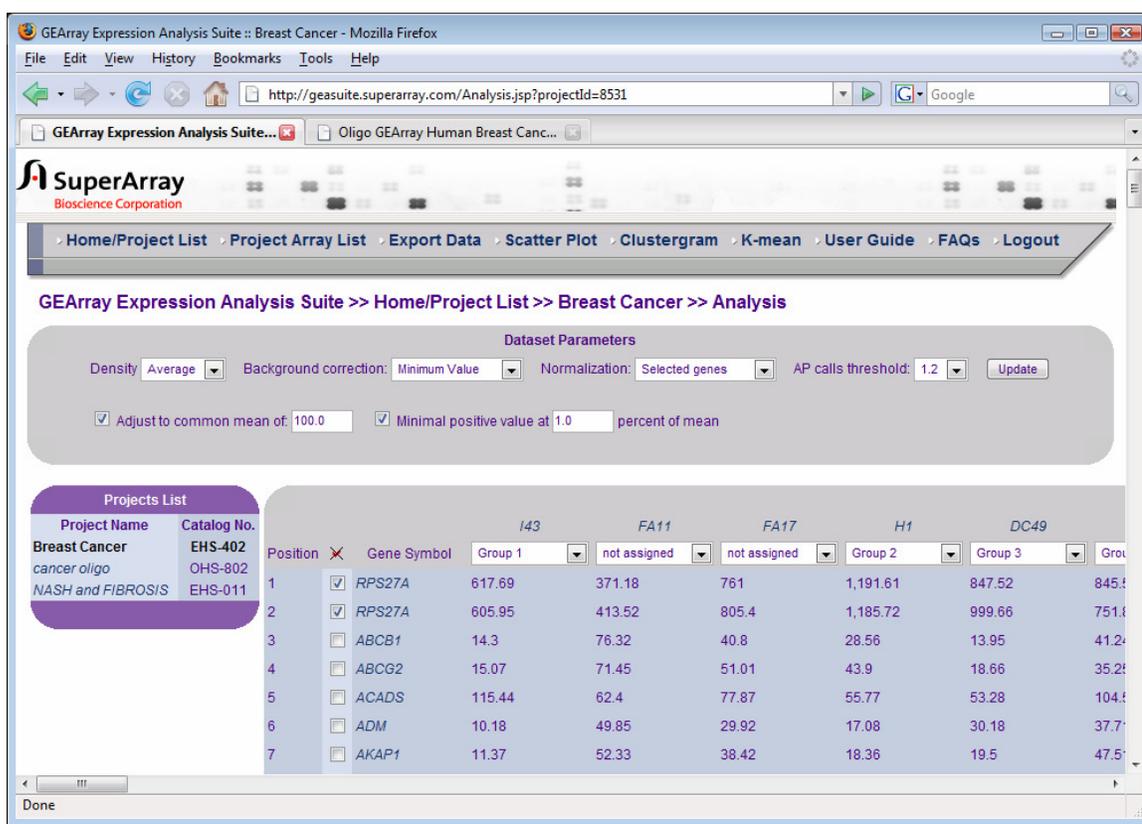


**Figure 18.** Bleeding of high-signal spots (e.g., controls, housekeeping genes)



**Figure 19.** Actual examples of bleeding

Once all the array images have been read, the raw data is stored in the website's database and available for manipulation, online analysis, or download at any time. From the Home/Project List page (the homepage for the project), clicking on "Analysis" navigates to the main analysis page, as shown in Figure 20 below. This page provides options for background correction, normalization, and (online) analysis. Note the Dataset Parameters panel where various parameters related to background corrections and normalizations are specified.



**Figure 20.** Main analysis page following array reading

The vendor website provides four background corrections: Local, Global, Empty Spots, and Minimum Value. The user may also forego any background correction and just work with the raw data. Local background correction subtracts the intensity of the region outside the spot area within each grid cell from the spot intensity. Average density is used in all calculations. Global background correction subtracts the average of all the local backgrounds from each spot. Empty Spots background correction takes the average intensity of the two empty spots on the array, and subtracts this amount from each spot. Minimum Value correction takes the minimum average intensity of all the spots and subtracts this amount from all the spots. In all cases, with the checkboxes for Adjust to common mean and Minimum positive value selected, at least one spot (Minimum) and usually more (Local, Global, and Empty

Spots) will be reset to the minimum positive value following the background correction. All four methods are examined across all arrays to determine the best array-level setting for each individual array. The results are discussed in the Background Corrections Results section below.

### ***Normalization***

In addition to background corrections, normalization is also used to attempt to correct for systemic variation that naturally arises in microarray experiments from variations in sample preparation, hybridization, and array scanning. While background corrections can be thought of as normalization *within* the arrays, normalization attempts to remove as much systemic variation *across* the arrays as possible to allow more accurate measurement of the “true” biological variations.

Normalization is crucial to correct microarray analysis. Unfortunately, the SuperArray website normalization features have some limitations. In particular, the ability it provides to normalize a group of arrays with one subset of controls and another group with another subset of controls is a powerful way to account for control skew between different condition groups. While this is possible on the vendor’s website, it is not practical for two reasons.

First, the controls tend to be the highest signaling spots on the array, so normalizing against them has the effect of compressing the range of signals, which can easily be many orders of magnitude, into the range from zero to around one. This compression results from the normalization process which divides all the readings on an array by what typically are the largest readings, namely the controls spots.

Secondly, when downloading the data from the website following such a group-wise normalization, the vendor's tools exacerbate the compression problem by truncating all values at two decimal places. Since all the data typically has been compressed into the range zero to one, this truncation results in significant information loss. No setting could be found to increase the number of decimal places on downloading.

For these reasons, a classical approach to normalizing across arrays was used whereby each array is scaled such that they all have a common mean. This approach amounts to simple scaling to a common basis across all arrays. The vendor website provides good tools for this approach, allowing the user to choose the common mean as well as (optionally) specifying an arbitrary lower bound to set all values to which would otherwise fall below this bound following the normalization. The minimum value setting was used (typically either one or ten), and the common mean was set to 100. The minimum threshold setting helps prevent negative numbers, zeros, or arbitrarily small values that will only introduce noise into the statistics.

### ***Statistical Analysis Approach***

Before discussing statistical tests, it is important to distinguish between biological replicates and technical replicates. For this study, biological replicates are RNA samples obtained from independent biological sources, for example DCIS samples from different patients. Technical replicates represent repeated sampling of the same biological material and are useful for assessing random errors introduced by laboratory and image processing. For statistical tests, biological replicates are ideal because they represent independent biological samples of a given condition (e.g.,

stroma surrounding cancerous tissue). Biological replicates are used for this study, but technical replicates are not, due mainly to the prohibitive cost of amplification kits. However, good RNA amplification yields generally allow re-running arrays as needed.

The statistical analysis approach used is the usual hypothesis testing and significance testing approach used in microarray analysis [33]:

1. Generate a null and alternate hypothesis
2. Choose a significance level
3. Calculate an appropriate statistic based on the data, and calculate a p-value based on it
4. Apply a Multiple Test Correction to obtain the final (adjusted) p-values
5. Compare these p-values with the significance level and either reject or not reject the null hypothesis

Here, p-value has the usual meaning of the probability of drawing the wrong conclusion by rejecting a true null hypothesis, and choosing a significance level means choosing the maximum acceptable level for this probability. The typical significance level of  $\leq 5\%$  (0.05) is used, which means that a gene expression measurement in a condition of interest has a 5% probability or less of having randomly occurred by chance from the normally distributed control condition's gene expression distribution (the null hypothesis).

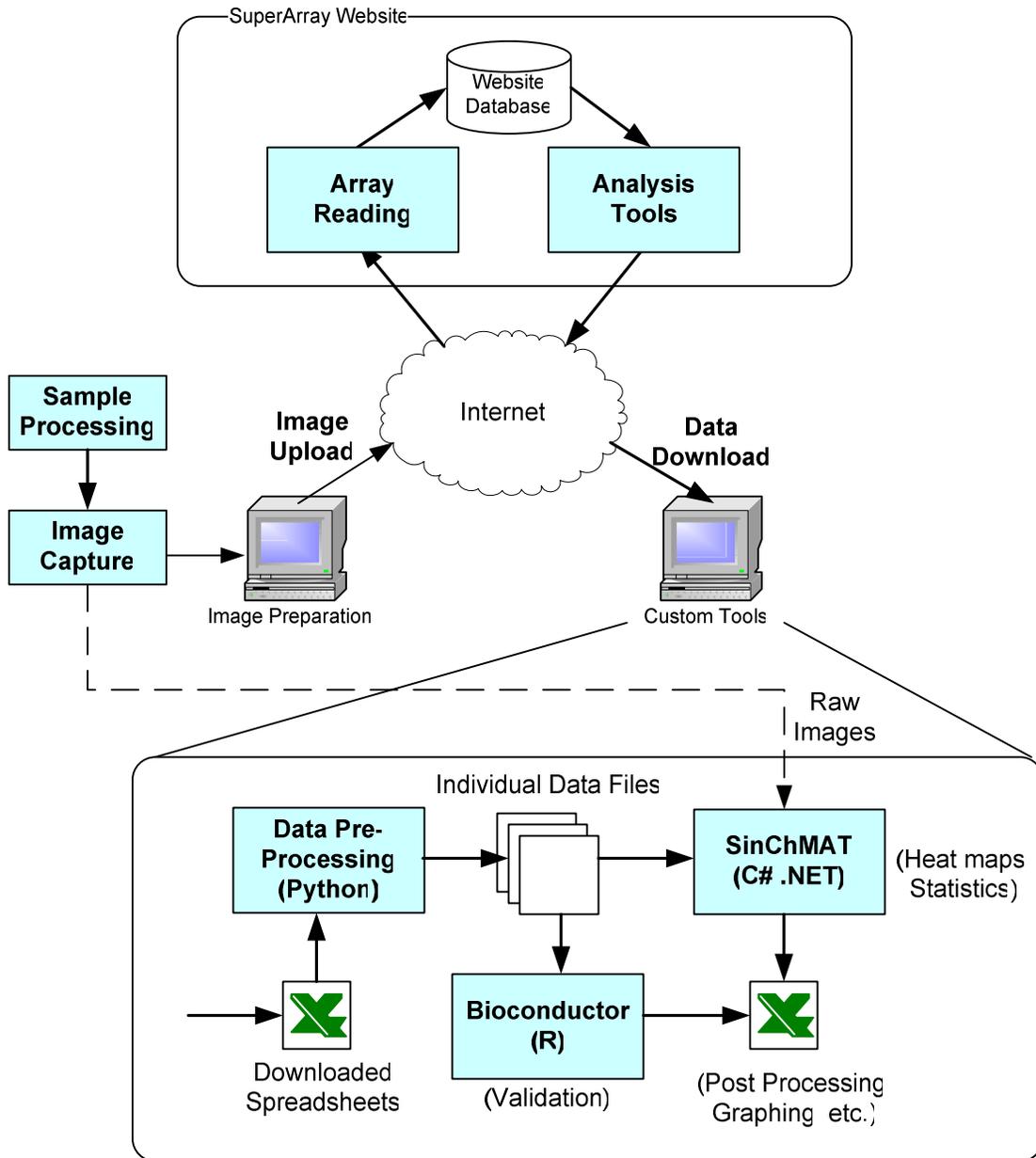
Choosing the appropriate statistic based on the data must be done with care. Statistical tests fall into two broad categories: parametric and non-parametric, depending on the number of samples. Parametric tests such as the Student's t-test for example, require about thirty samples minimum as a rule of thumb and assume a normal distribution. As discussed in the Results below, most condition groups of interest here had far fewer than this minimum value. Thus, non-parametric tests are more appropriate. For non-parametric testing, the statistic used is the Mann-Whitney U test [34], which is the Wilcoxon [35] rank-sum test extended to non-equal samples sizes. This statistic not only is appropriate for smaller sample sizes, but it also does not rely on the variable having a normal distribution.

All genes on all arrays are compared, computing p-values for all. Additional quality control checks are done to look for functional anomalies (biologically based trends going in the wrong direction, e.g., tumor-suppressor genes), typical fractions of genes up- or down-regulated out of the total, and symmetry of statistically significant gene differences (roughly the same up and down).

Finally, to correct for type I errors, Multiple Test Correction techniques are applied. The multiple comparisons problem occurs when one subjects a number of independent observations to the same acceptance criterion that would be used when considering a single event [36]. Since the Bonferroni method is known to be overly strict, the Benjamini – Hochberg False Discovery Rate [37] correction is used instead.

## Computational Analysis Approach

The overall approach involves several tools. Figure 21 below shows a high-level view of the specific computational analysis approach used.



**Figure 21.** High-level overview of the overall computational approach

The procedures performed on the SuperArray website were discussed in the previous section. Picking up where the data are downloaded from the website, all data files come down as Microsoft® Excel® spreadsheets. These spreadsheets contain several sheets of data for all arrays selected on the website (nominally, all of them). The main sheet of interest is the listing of the spot readings, background corrected and normalized as specified on the website analysis page, and indexed by array position, which is in turn correlated to a gene list. From the spreadsheet, the array readings are dumped to tab-delimited text files which are processed by custom Python scripts to extract and split up the array readings into separate files (to allow, for instance, mixing and matching the best individual background corrections at the array level). The individual groups of array-level data files are then read in user-determined combinations into a custom tool developed specifically for this analysis called Single Channel Microarray Analysis Tool (SinChMAT). SinChMAT is a Microsoft Windows® application developed in C# (.NET).

Although many microarray analysis tools exist in the public domain, most are geared towards dual-channel microarray analysis. For maximum flexibility and for study-specific features, SinChMAT was developed. It performs two main functions. First, it is used to generate multiple, side-by-side heat maps for a given array's data files resulting from the various background corrections, for viewing along side the raw image. This feature is useful for assessing which background correction is best for a given array, particularly since the raw image can be viewed concurrently to visually assess any smudging or background variations. Secondly, SinChMAT allows rapid computation of all p-values, with and without multiple test corrections, for any

combination of the condition groups desired. Many different pairs of condition groups are of interest as will be discussed in the Results section.

SinChMAT uses an open source library from ALGLIB [38] to compute Mann-Whitney based p-values. The Bonferroni and Benjamini – Hochberg False Discovery Rate implementations are taken from an Agilent Technologies White Paper [39]. Both are simple, conservative implementations as given below.

Bonferroni Correction:

$$\text{Corrected P-value} = p\text{-value} * n \text{ (number of genes in test)} < 0.05$$

Benjamini – Hochberg Correction:

- 1) The p-values of each gene are ranked from the smallest to the largest.
- 2) The largest p-value remains as it is.
- 3) The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant.

$$\text{Corrected p-value} = p\text{-value} * (n/n-1) < 0.05, \text{ if so, gene is significant.}$$

- 4) The third p-value is multiplied as in step 3:

$$\text{Corrected p-value} = p\text{-value} * (n/n-2) < 0.05, \text{ if so, gene is significant.}$$

And so on.

The statistical calculations from ALGLIB are validated by computing p-values independently in R using Bioconductor libraries ([www.bioconductor.org](http://www.bioconductor.org)).

Specifically, the `stats` library is used.

Finally, fold change analysis is used as a “sanity check.” Examining fold changes is simple and intuitive, and thus is frequently used in microarray gene expression analyses. However, the fold change method has important disadvantages [33]. The most important drawback is that the fold change threshold is chosen arbitrarily, without the more rigorous significance level basis used in hypothesis testing with multiple test corrections. Thresholds such as 1.2, 1.5, or 2.0 are frequently used, but these are merely rules of thumb. If the arbitrarily chosen threshold is too high, there is poor sensitivity, and if the threshold is chosen too low, false positives creep in. Another important disadvantage is that microarray technology tends to have poor signal to noise ratios for genes with low expression levels. This fact results in genes with higher expression levels potentially being more reliable even though they have relatively lower fold change differences than genes with lower expression levels but higher fold changes that are actually less reliable. Since the fold change technique uses a constant threshold, this distinction is lost. Worse, this effect tends to introduce false positives at the low end, thus reducing specificity, while simultaneously missing true positives at the high end, thus reducing sensitivity.

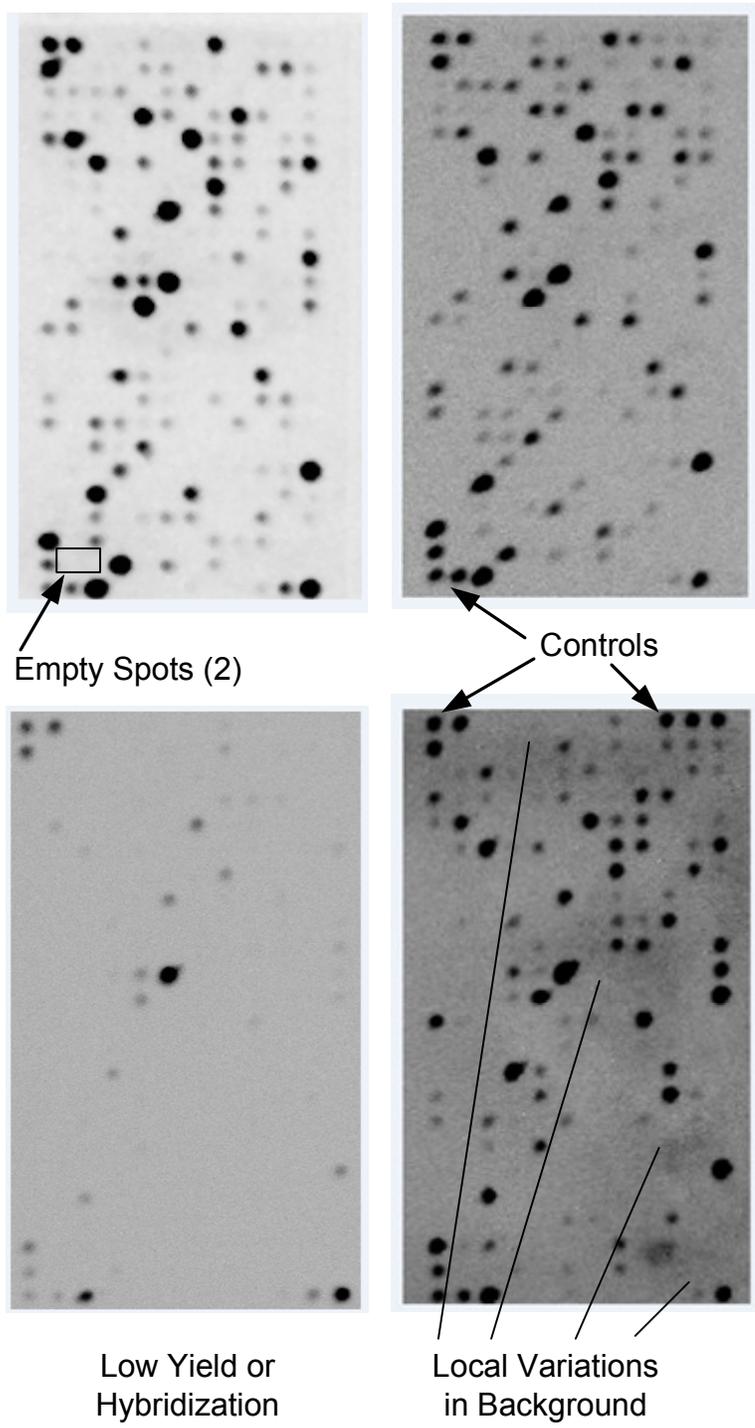
Nonetheless, the SuperArray website automatically computes fold changes and produces scatter plots with many user-controllable parameters, so these tools are used as yet another validation technique, albeit informal, on the results obtained from the more rigorous statistical techniques.

## ***Raw Image Results***

Figure 22 below shows a sampling of raw images obtained from the Kodak 4000 MM Imager CCD camera (see Materials and Methods section). Due to the normal variability in laboratory processing, hybridization, and exposure and image capture, the raw images show a range of overall lightness or darkness, contrast, and local background variation. A delicate balance must be struck, particularly in the face of low yielding or hybridizing samples, between increasing exposure times to detect low signals versus decreasing exposure times to reduce background. Higher exposures reveal more (faint) signals, helping to reduce false negatives at the expense of possibly increasing background to the point of introducing false positives.

Figure 22 also shows the two Empty Spot locations and the controls (corners). Unfortunately, as the figure indicates, the empty spots are directly adjacent to control locations, which tend to be the highest intensity locations and thus are the most likely to suffer bleeding.

Arrays from two groups (Normal Stroma and Hyperplasia epithelium) had to be read using a ChemiDoc system by Bio-Rad Laboratories ([www.bio-rad.com](http://www.bio-rad.com)) instead of the Kodak CCD, due to the latter's unavailability. These samples produce lower quality images, due either to the different imager or perhaps low hybridization (or both), and thus these two groups had to be dropped out of the analysis.



**Figure 22.** Raw image examples

## ***Background Correction Results***

The SuperArray website can compute four possible background corrections for each individual array: Local (L), Global (G), Empty Spots (E), and Minimum Value (M). Each was chosen successively, and the results were downloaded and preprocessed into individual array level files in separate file directories on a Microsoft Windows® workstation for use by the SinChMAT tool. This tool permits viewing the raw array image side-by-side with up to three different background correction heat maps simultaneously, as shown in the screen shot below.



**Figure 23.** Single-Channel Microarray Analysis Tool (SinChMAT) heat map viewer

The specimen tree on the left is used to navigate and select the specimen of interest. This tree is built dynamically from an input file and is therefore configurable for other experiments. Once selected, the specimen's raw image appears in the upper left pane. The dropdown controls in the remaining three panes are used to open and display heat maps for the various background corrections. Each array in the experiment was analyzed using this tool in order to select the most appropriate

background correction for it. The results of this analysis are captured in a SinChMAT input file that is used in the statistics calculations. This input file acts as a switch which directs SinChMAT to choose the specific array level data file for all downstream computations. The tool uses several internal data structures to keep track of group membership and the chosen array data vectors as well as intermediate and final results.

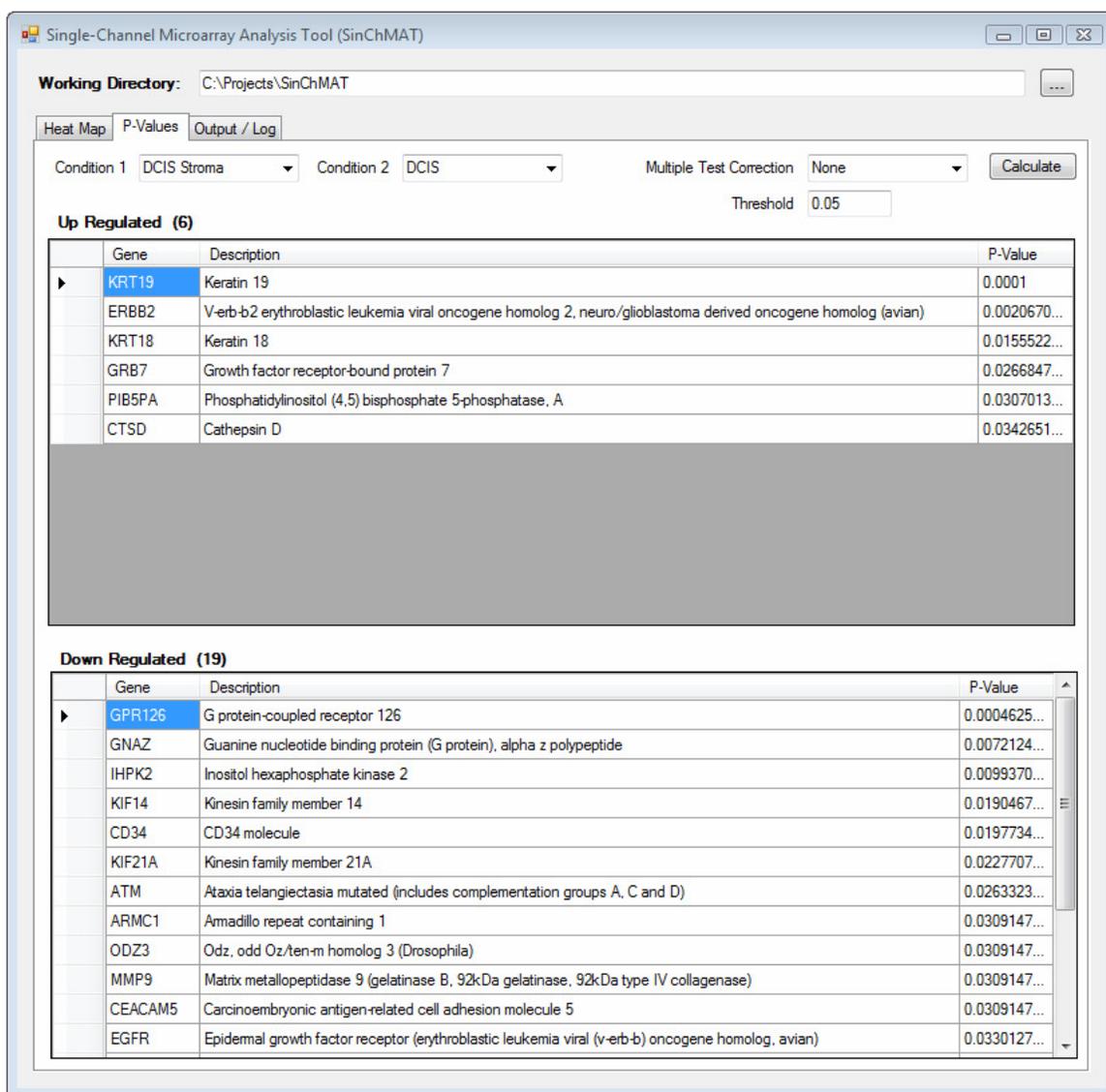
The background correction results were consistent with common sense. Empty Spot or Global background correction worked best for low background images with good signaling and contrast. (Recall that Global correction subtracts the average of all Local (cell level) corrections equally from all spots.) Empty correction did worse when there was obvious bleeding from the neighboring control locations, as would be expected. Global correction worked best with arrays with heavy background, and Local correction worked best when noticeable local variations were apparent within the array, both consistent with what would be expected. Minimum Value usually was not the best choice; however, this correction worked better for some low intensity or poorer contrasting images.

### ***Normalization Results***

As discussed in the previous section, limitations on the vendor's website made group-wise or control based normalization across arrays difficult. Thus, the classical approach of simply scaling all arrays to a common mean was used. The common mean chosen was 100 with a minimum value cutoff of one. This approach resulted in ranges typically on the order of one to several thousands for an individual array, allowing for reasonable statistics and many potential fold changes.

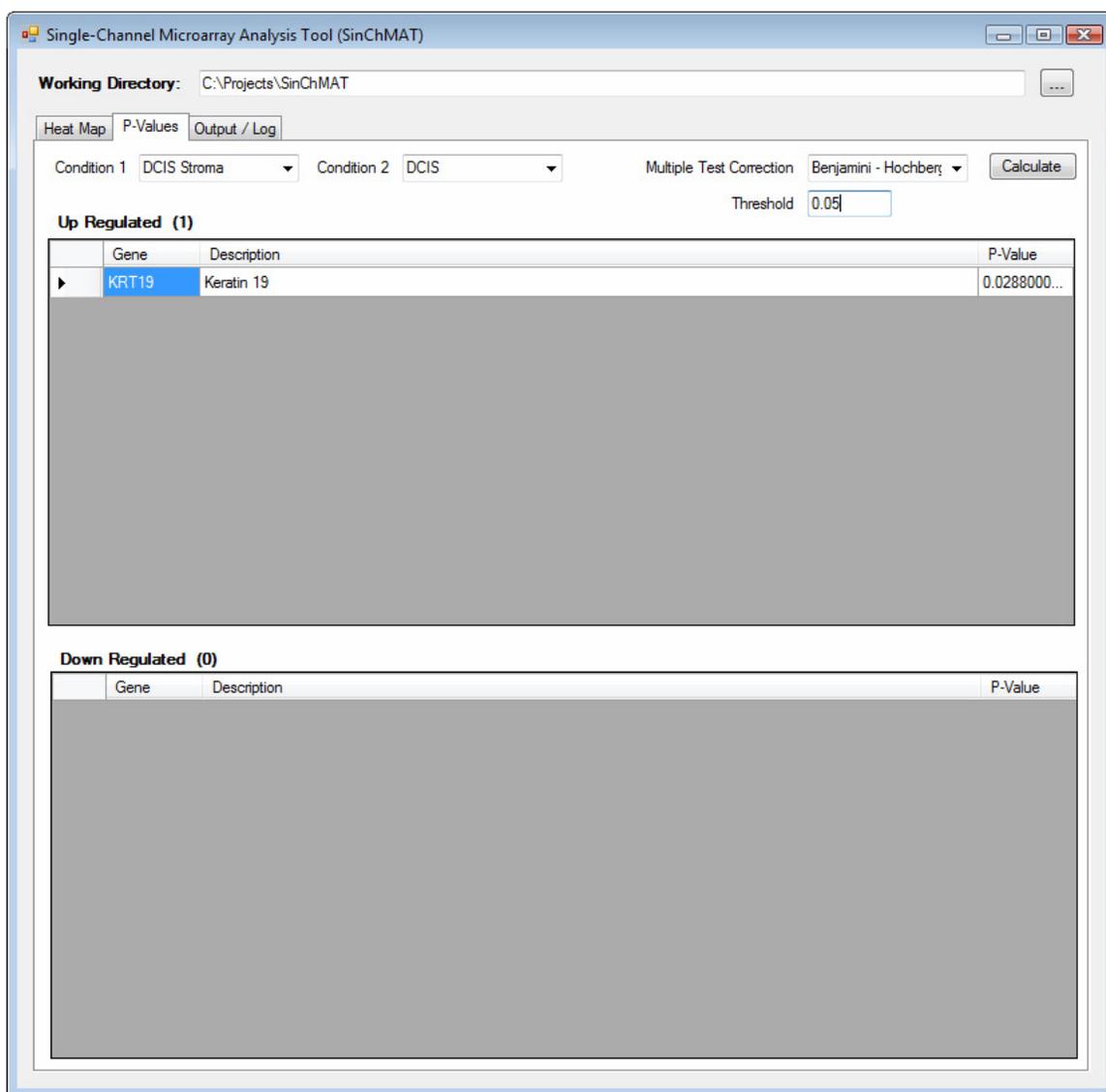
## ***Statistical Analysis Results***

Following array-level background correction choices, and using the globally applied common mean normalization, the statistics are computed. The SinChMAT tool was developed to facilitate these calculations, in particular to allow rapid calculations and recalculations of any arbitrary group-wise pair comparison. SinChMAT uses an open source implementation of Mann-Whitney based p-values with both Bonferroni and Benjamini-Hochberg (False Discovery Rate) multiple test corrections. The typical 5% ( $p \leq 0.05$ ) significance level is applied as a filter on the displayed results. All 288 array locations are considered, although blanks, controls, and artificial sequences are filtered from the results display.



**Figure 24.** SinChMAT screen shot showing p-value results for DCIS Stroma-to-DCIS

The screen shot in Figure 24 shows the results of a DCIS Stroma versus DCIS comparison, before application of the multiple test correction. Figure 25 below shows the results after applying the Benjamini-Hochberg multiple test correction.



**Figure 25.** P-value results following Benjamini-Hochberg multiple test correction

Another useful feature of SinChMAT (not shown) is that the results of most statistics calculations are dumped in tab delimited form to a window on the "Output / Log" tab for easy copy-and-paste operations into spreadsheets, which facilitates post-processing, further analyses, and plotting.

Tables 3 and 4 below summarize the hypothesis testing results for the original groupings and for the groupings following the DCIS With and Without IBC break out.

**Table 3.** Summary of hypothesis testing results ( $p \leq 0.05$ )

Condition 1	Condition 2	Before Multi-test Correction		After B-H (FDR) Corr.	
		Up	Down	Up	Down
Stroma	Epithelium				
DCIS Stroma (N = 8)	DCIS (N = 10)	KRT19, ERBB2, KRT18, GRB7, PIB5PA	GPR126, ARMC1, ALDH4A1, CD34, KIF21A, EGFR, GNAZ, IHPK2, CEACAM5, INS, ECT2, MIB1, PCSK6, PLG	KRT19	-
IBC Stroma (N = 12)	IBC (N = 15)	KRT18, CYC1, ERBB2, KRT19, GRB7, TK1, BAX, ST7, BCL2L1, BAG3, BBC3	PRAME, WISP1, ... (32 total, see Note 1)	KRT18, CYC1, ERBB2, KRT19, GRB7, TK1	PRAME, WISP1, IGFBP3
Stroma	Stroma				
Hyperplasia Stroma (N = 4)	DCIS Stroma (N = 8)	-	ASNS	-	-
Hyperplasia Stroma (N = 4)	IBC Stroma (N = 12)	CENPN, TNF, ERGIC1, TP53	RPS4X	-	-
DCIS Stroma (N = 8)	IBC Stroma (N = 12)	DCK, MYBL2, MKI67, CENPN, PLEKHA1, TK1, PCNA, CYC1, ERGIC1, PGK1, STMN1	IGRBP5	-	-
Epithelium	Epithelium				
Normal (N = 5)	DCIS (N = 15)	CD68, DEGS, MX1, TMEM45A, NME1, CTSD, IGF2, AKT1	TBX3, MAD2L1, ... (14 total, see Note 1)		-
Normal (N = 5)	IBC (N = 15)	CD68, BRCA1, BAG3, CCNE1, CTSD, MX1, ... (25 total, see Note 1)	RPS4X, MAD2L1, IGFBP5, KIF21A, MARCH8	CD68	-
DCIS (N = 10)	IBC (N = 15)	C20orf28, TBX3, CDC25B, BBC3, , MUC1, CSF1, ASNS, BTG2, CYC1	RPS4X	-	-

Note 1 – Several genes significant at  $p \leq 0.05$ . Only those with  $\leq 0.01$  are shown.

**Table 4.** Summary of hypothesis testing results ( $p \leq 0.05$ ) following the DCIS With and Without IBC split out (both epithelia and stroma groups)

Condition 1	Condition 2	Before Multi-test Correction		After B-H (FDR) Corr.	
		Up	Down	Up	Down
Stroma	Epithelium				
DCIS w/o IBC (N = 8)	DCIS w/o IBC (N = 7)	KRT19, ERBB2, PIB5PA, GRB7, KRT18, BAG3	GPR126, CD34	-	-
DCIS with IBC (N = 3)	DCIS with IBC (N = 3)	-	-	-	-
IBC (N = 9)	IBC (N = 15)	KRT18, KRT19, ERBB2, CYC1, GRB7, BAX, PRKCG, TK1	MKI67, IGFBP3, PCNA, ESR2, ... (~80 total, see Note 1)	-	-
Stroma	Stroma				
Hyperplasia (N = 4)	DCIS w/o IBC (N = 8)	-	ASNS	-	-
Hyperplasia (N = 4)	DCIS with IBC (N = 3)	ERGIC1, DEGS1 (Note 2)	FBXO31, ASNS	ERGIC1, DEGS1 (Note 2)	FBXO31, ASNS
Hyperplasia (N = 4)	IBC (N = 9)	TNF, TK1, BNIP3, CENPN, BIRC5, SP1, CTSD	RPS4X	-	-
DCIS w/o IBC (N = 8)	DCIS with IBC (N = 3)	ERGIC1	GRB7	-	-
DCIS w/o IBC (N = 8)	IBC (N = 9)	MKI67, K1, STMN1, PCNA, MYBL2, ... (15 total, see Note 1)	IGFBP5, RP5-860F19.3, EGLN1	-	IGFBP5
DCIS with IBC (N = 3)	IBC (N = 9)	ASNS, GRB7, MKI67, ERBB2, BIRC5, CDC25B, MYBL2, FBXO31	RB1, HMGB3, EGLN1, RP5-860F19.3, IGFBP5	-	-
Epithelium	Epithelium				
Normal (N = 5)	DCIS w/o IBC (N = 7)	CD68, NME1, DEGS1, BAG3, MX1, CTSD	IGFBP5, TBX3	-	-
Normal (N = 5)	DCIS w/ IBC (N = 3)	CD68, BRCA1, DEGS1	-	-	-
DCIS w/o IBC (N = 7)	DCIS w/ IBC (N = 3)	HMGB3	FBXO31	-	-
DCIS w/o IBC (N = 7)	IBC (N = 15)	C20orf28	PRAME	-	-
DCIS w/ IBC (N = 3)	IBC (N = 15)	CYC1, ASNS, MTDH, BBC3, KRT18, BTG2, CDC25B, PDPK1	-	-	-

Note 1 – Several (> 10) genes significant at  $p \leq 0.05$ . Only those with  $\leq 0.01$  are shown.

Note 2 – Only  $\leq 0.05$  and surviving the B-H correction in one of the two independent computational methods, and that method experienced difficulty on some sets with  $N < 4$ . However, the second method found the same four genes as the top four with  $p$  just over 0.05, so they are reported with qualification.

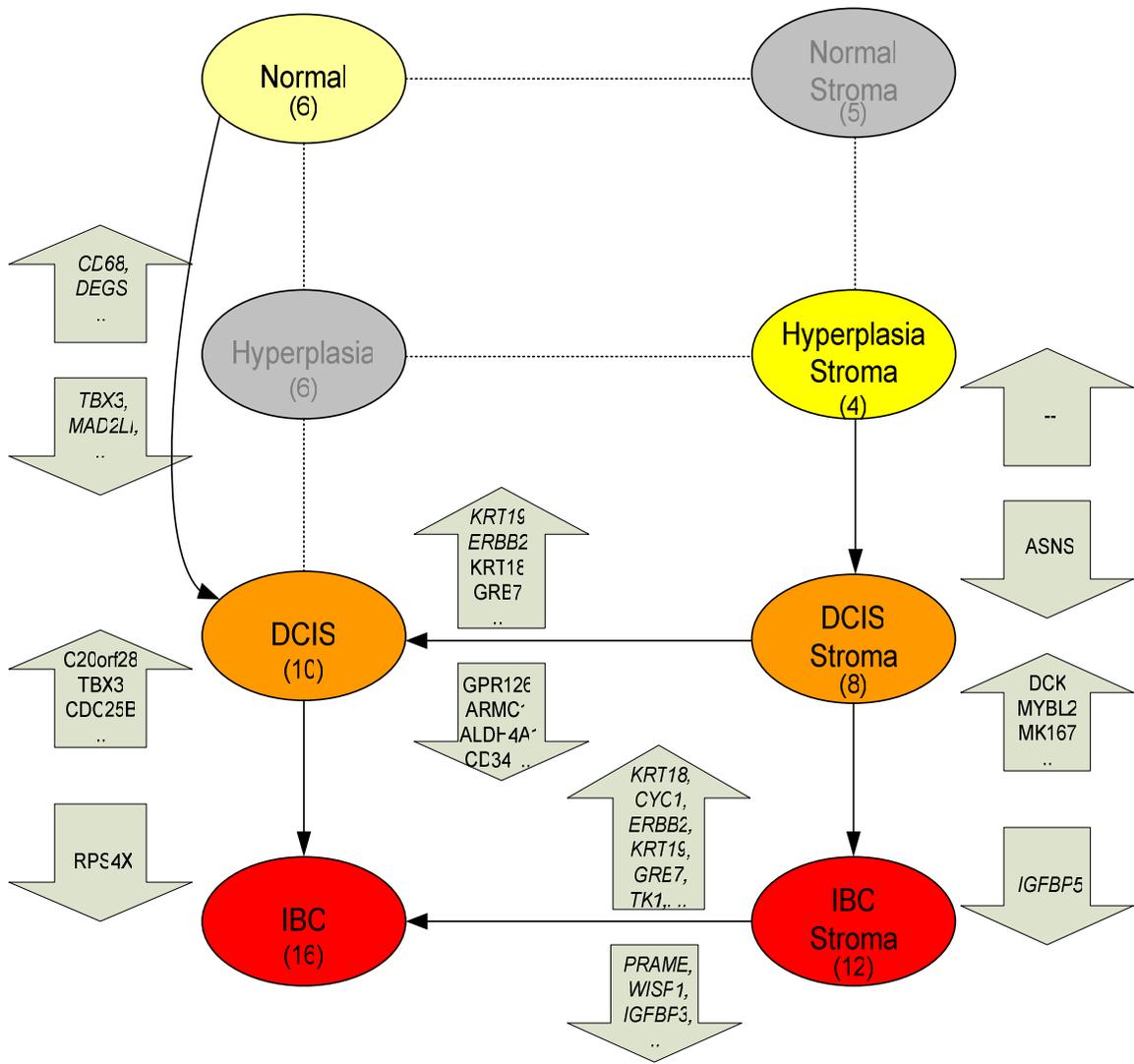
## Discussion

The data and results of the hypothesis testing presented are discussed and rationalized in a biological context. Due to the large number of group-wise comparisons and the resulting large numbers of genes of interest, the discussion will focus on a subset of genes that proved particularly significant, that appeared in multiple comparison results, or both. Recall that the normalization method used for all arrays was basic scaling to a common mean. Though this technique is simple and has known limitations, it does allow the analysis of a gene's expression across all groups, representing both disease progression and epithelium versus stroma, simultaneously and on a common scale. For these results, this common scaling technique proves quite valuable for gaining insights at a "systems" level.

Figure 26 below is a graphical summary of the raw results in Table 3 above, before splitting out DCIS with and without IBC, and Figure 27 below provides the same "bird's eye" view of Table 4 after the DCIS split out.

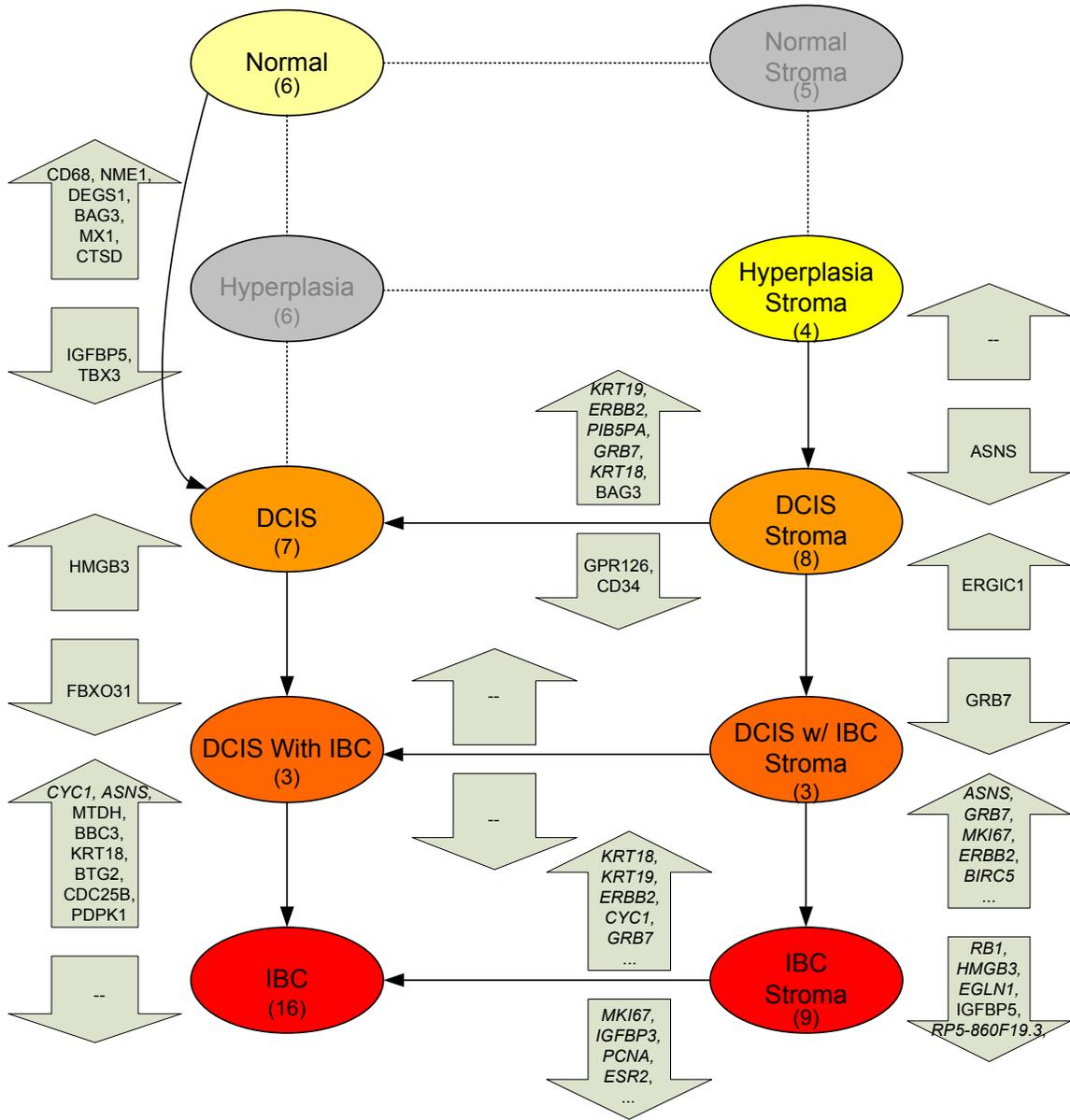
Most samples in the Normal Stroma and Hyperplasia groups did not yield usable images, probably due to equipment unavailability issues that required reading these arrays on a ChemiDoc system instead of the Kodak CCD used for all other samples. These two groups are shown grayed out on the figures below, and subsequent plots reflect that these groups are missing. Thus, for the stromal groups, the comparisons

begin with Hyperplasia Stroma instead of Normal Stroma. Likewise, the missing Hyperplasia epithelial group is simply skipped over by making comparisons directly between the Normal and DCIS epithelial groups.



*Italics indicates either survival of Benjamin-Hochberg FDR or  $p \leq 0.01$*

**Figure 26.** Overview of hypothesis testing results ( $p \leq 0.05$ ) with original groupings.



*Italics indicates either survival of Benjamini-Hochberg FDR or  $p \leq 0.01$*

**Figure 27.** Overview of hypothesis testing results ( $p \leq 0.05$ ) with DCIS With IBC and DCIS Without IBC split out.

## ***Analysis of Selected Genes***

Due to the large number of group-wise comparisons, and the resulting significant gene lists, the following discussion will focus on individual gene expression across all groups simultaneously rather than group pair comparisons. This is done not only for brevity and clarity but also to give a more comprehensive, high-level snapshot of gene expression in both tissue types and across all disease conditions. Since many genes discussed appear across several groupings, the genes are presented simply in alphabetical order. They are as follows:

- ASNS - Asparagine synthetase
- ERBB2 (NEU/HER-2) - V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog
- GRB7 - Growth factor receptor-bound protein 7
- HMGB3 - High-mobility group box 3
- IGFBP3 and IGFBP5 – Insulin-like growth factor binding proteins 3 and 5
- KRT18 and KRT19 – Keratin 18 and Keratin 19
- MKI67 – Antigen identified by monoclonal antibody Ki-67
- MYBL2 - V-myb myeloblastosis viral oncogene homolog avian-like 2
- WISP1 - WNT1 inducible signaling pathway protein 1

Note that the “direction” of the group-wise comparisons follows the general notion and convention of a “control” condition versus a “treated” condition. In the analysis space of interest here, the progression of disease from Normal to IBC (invasive, fully malignant) provides an obvious step-wise ordering of comparisons. This ordering is parallel on both the epithelial and stromal sides (Figures 26 and 27 above) with each

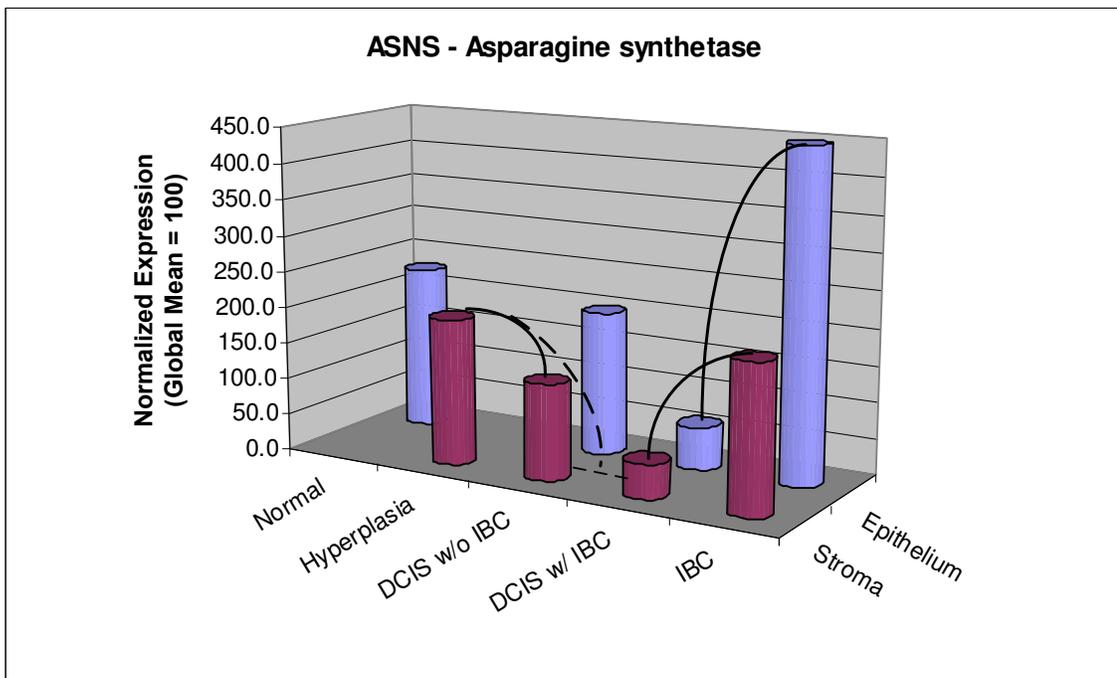
step acting as a treated condition to the previous step's control. In addition, comparisons are made between epithelia and stroma at each disease progression step. Again striving to follow convention, the stroma is treated as the "from" or control condition and the epithelial tissue is the "to" or treated condition. In this case however, it may be more illustrative to view the stroma as the condition in some instances, since the stroma is of particular interest here. Of course, in either case the direction is irrelevant, since reversing the direction of comparison simply reverses up regulation and down regulation.

It is important to note that the plots in all of the following discussion are group means on a data set globally scaled to a common mean. While convenient and illustrative, these plots must be used with care and cannot be used in isolation to draw conclusions about gene regulation. Only when combined with the rigorous statistical hypothesis testing results in Tables 3 and 4 (summarized in Figures 26 and 27), can conclusions be drawn about statistically significant up and down regulation of genes. All other insights gained from these plots are necessarily speculative. The following discussion will take care to clearly distinguish between the statistically significant findings, which can be found in Tables 3 and 4, versus general trends of the (universally scaled) group means. For clarity, the plots have curves connecting the group mean bars for those pairs found to have changes in gene expression that are statistically significant.

### **ASNS - Asparagine synthetase**

ASNS is asparagine synthetase and is listed under Other Prognostics Markers on the breast cancer biomarker array used for all specimens. Its appearance in stromal

samples is not surprising as its use as a biomarker derives from its protective role in the microenvironments of certain cancers, particularly ovarian [40] and acute lymphoblastic leukemia (ALL) [41]. In the latter case, bone marrow-derived mesenchymal cells with high ASNS expression acted to form a protective microenvironment where leukemia cells could grow. However, no such study was found for breast cancer, and ASNS's utility seems to be mainly as a biomarker for pharmaceutical efficacy.



**Figure 28.** ASNS normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

In ALL, the leukemia cells are extremely sensitive to asparagine depletion because asparagine synthetase expression, and therefore asparagine biosynthesis, is low. Thus, asparaginase is a major component of ALL therapy. For the breast cancer samples examined here, the results show a significant drop in ASNS expression in

breast epithelium right at the point of transformation, suggesting possible therapeutic application for asparaginase to breast cancer similar to ALL. No reference to this could be found in the literature.

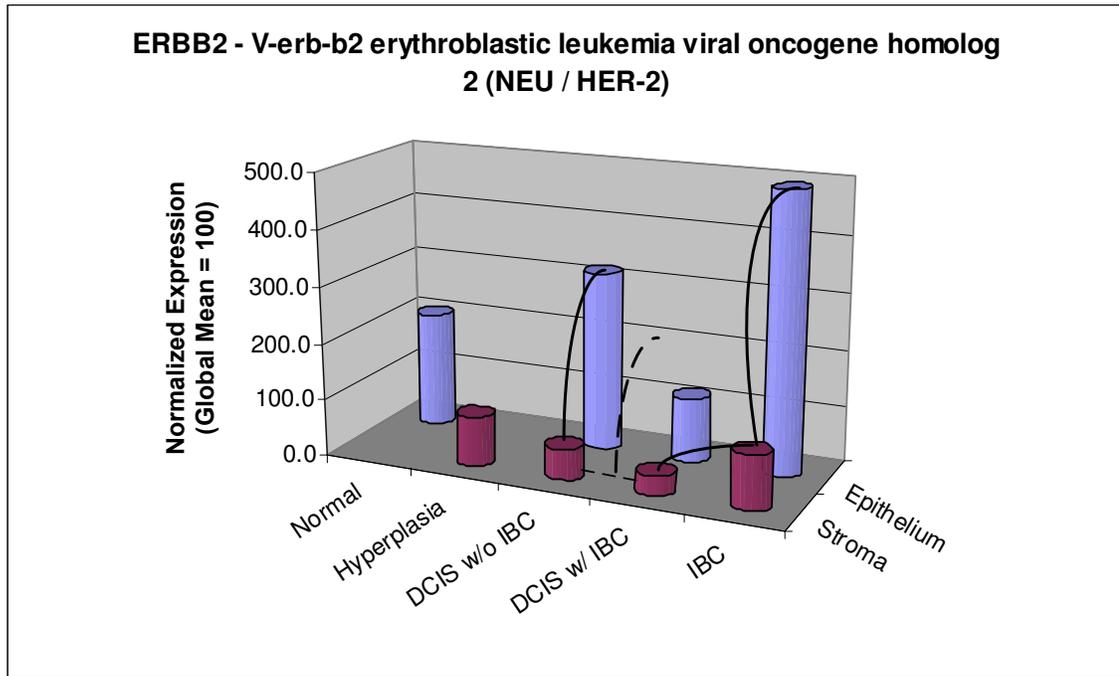
Interestingly, as the disease progresses, ASNS expression is likewise suppressed in the corresponding stromal groups, falling steadily until the invasive point is reached and then rebounding to the same level as Hyperplasia. As shown in Figure 27 and Table 4, ASNS was found significantly down regulated between the Hyperplasia Stroma versus DCIS (Without IBC) Stroma group and significantly up regulated moving from DCIS With IBC Stroma to IBC Stroma. Also from Figure 27, ASNS is significantly up regulated moving from DCIS With IBC (epithelium) to IBC.

It should be noted that the DCIS With IBC groups (both epithelium and stroma) are the smallest groups (N = 3), so the apparent drop at that disease progression step could be spurious despite the fact that the image quality for those six samples is quite good. It turns out this possibility is irrelevant because even with the DCIS groups combined, the difference in ASNS expression between DCIS and IBC epithelial groups is still statistically significant, supporting the notion that ASNS expression is much lower in DCIS and may provide a therapeutic opportunity similar to ALL prior to the cancer becoming invasive.

## **ERBB2 - V-erb-b2 erythroblastic leukemia viral oncogene homolog 2**

ERBB2, better known as NEU or HER-2, has been widely reported as over expressed or amplified in numerous cancers, certainly breast carcinomas included [42], so some up regulation (or amplification) of this gene in breast epithelium as disease

progresses is not surprising. This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases.



**Figure 29.** ERBB2 (NEU/HER-2) normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

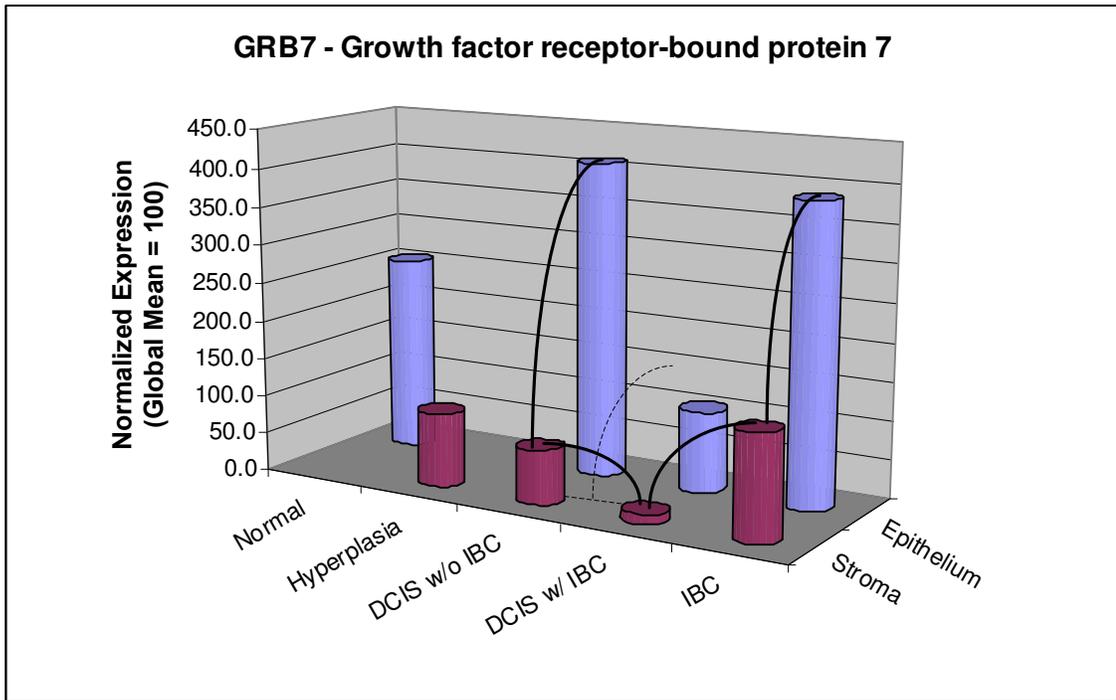
The apparent drop in ERBB2 at the DCIS split out may be due to the small sample size of DCIS With IBC (N = 3) mentioned above, although again image quality of these samples is actually quite good. Indeed, if the DCIS groups are collapsed, the significance on the DCIS With IBC versus IBC stroma-to-stroma comparison (Figure 27) disappears. Also, the appearance of significant up regulation between multiple stroma-to-epithelia group pairs is consistent with this gene's specific up regulation (or amplification) in the breast epithelium tissue proper, not the stroma. Although ERBB2 up regulation is well established in certain subtypes of breast cancers, it is

not a general biomarker for all breast cancers. In fact, ERBB2 is amplified in only about 30% of all human breast cancers (and indicates a poor prognosis). Thus the results presented here must be considered in a context in which the exact fractions of samples in the epithelial groups that are ERBB2 (NEU/HER-2) positive have not been established.

### **GRB7 - Growth factor receptor-bound protein 7**

GRB7 encodes a gene whose product belongs to a small family of adapter proteins that interact with receptor tyrosine kinases and signaling molecules. GRB7's product specifically interacts with epidermal growth factor receptor (EGFR) receptors and contains a Src homology 2 (SH2) domain. It is involved in integrin signaling and cell migration via binding to focal adhesion kinase (FAK), and its up regulation in breast cancer is well established [43].

As with ERBB2 (NEU/HER-2), we may discount the apparent drop at DCIS With IBC as possibly spurious due to low sample number ( $N = 3$ ); however the significant differences between epithelial and stromal groups are as expected and lend additional confidence to successful laser capture microdissection of the distinct cell populations. Of possible interest is the trend across epithelial groups during disease progression. Unlike ERBB2, which monotonically ramps up with disease progression, GRB7 appears to rise sharply at DCIS Without IBC (epithelium) and remain high.



**Figure 30.** GRB7 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

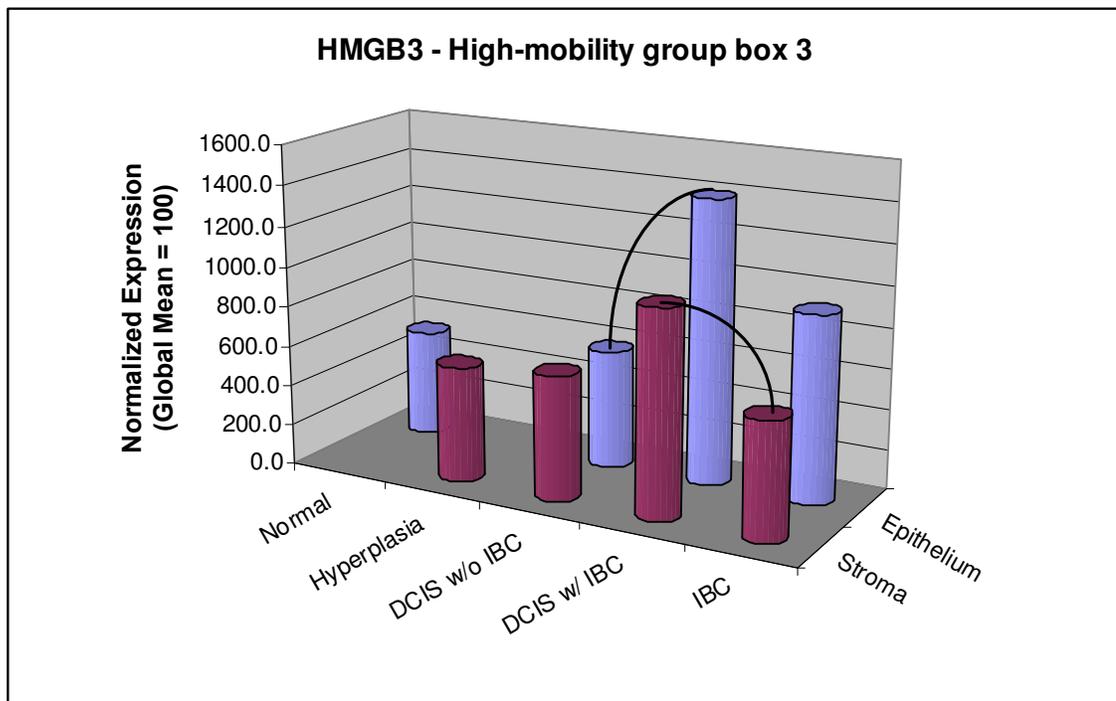
This finding is of potential interest for two reasons. First, GRB7 is generally associated with invasiveness (motility, cell migration) and metastasis, yet its expression in epithelia seems to be quite high well before IBC is reached. Second, GRB7 is a therapeutic target of significant interest, so again its elevation apparently earlier than expected is of interest. No reference was found in the literature for elevated GRB7 in DCIS.

Note also that GRB7 happens to lie very close to ERBB2 on chromosome 17q12, so the fact that Figures 29 and 30 have very similar overall trends is confirmatory and validating since the main mechanism of ERBB2 up regulation is thought to be gene

amplification. It would appear that GRB7, lying in close proximity to ERBB2 chromosomally, is likewise amplified.

### HMGB3 - High-mobility group box 3

HMGB3 regulates the balance between hematopoietic stem cell self-renewal and differentiation [44]. It is an X-linked member of a family of chromatin-binding proteins, and is listed under Prognostic – Transcription Factors and Regulators on the breast cancer biomarker array used for all samples. The plot of HMGB3 expression across all groups is shown in Figure 31 below.



**Figure 31.** HMGB3 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

HMGB3 shows an intriguing pattern of up regulation specifically in the DCIS With IBC groups, both epithelial and stromal. With the DCIS samples separated according to

With and Without IBC, HMGB3 is the lone gene with statistically significant up regulation between DCIS With IBC and DCIS Without IBC (epithelial) groups. The same pattern is evident in the corresponding stromal groups and indicates that the up regulation occurs in the stroma as well, with the statistical significance showing up when the gene's expression falls sharply back down between DCIS With IBC Stroma and IBC Stroma. Of the genes examined here, HMGB3 exhibits the most similar trends between epithelia and stroma groups across all disease conditions. The striking similarity of trends across the entire analysis space hints at, but certainly does not demonstrate, the possibility of a common mechanism up regulating this gene in lock step in both epithelia and stroma.

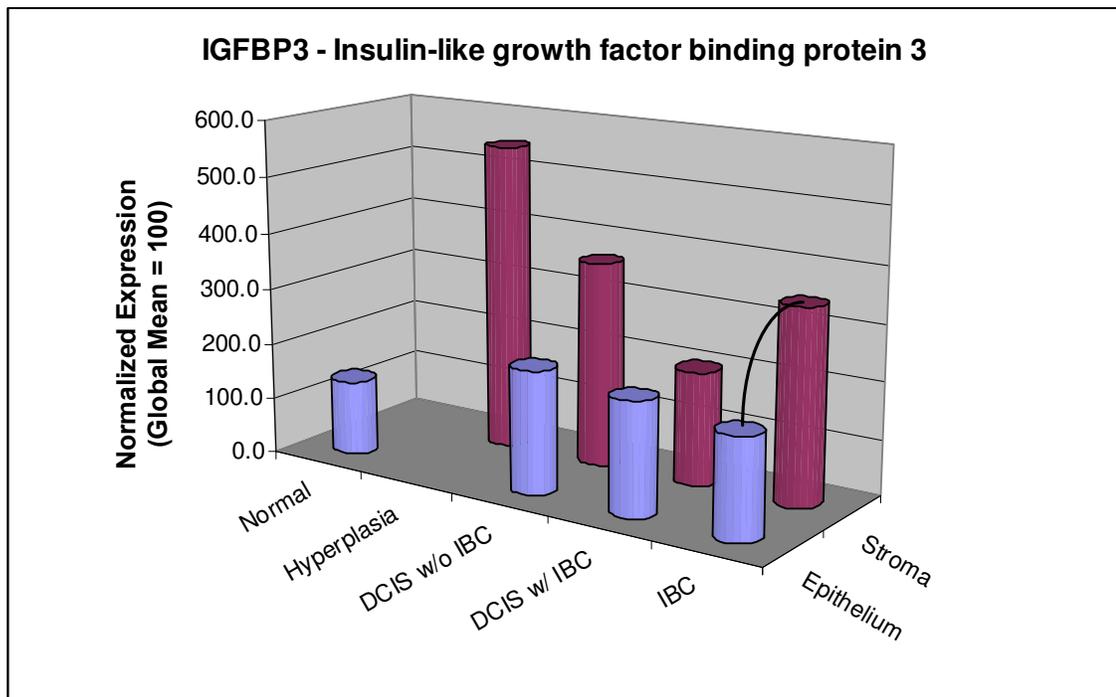
Finally, note that in marked contrast to the previous three genes discussed, the expression level is highest in the DCIS With IBC groups (both epithelial and stromal), providing confidence that these groups are not simply biased towards low signals as the previous three plots may lead one to believe. Inspection of the raw images also support these samples have strong, clean signals.

### **IGFBP3 and IGFBP5 – Insulin-like growth factor binding protein 3 and 5**

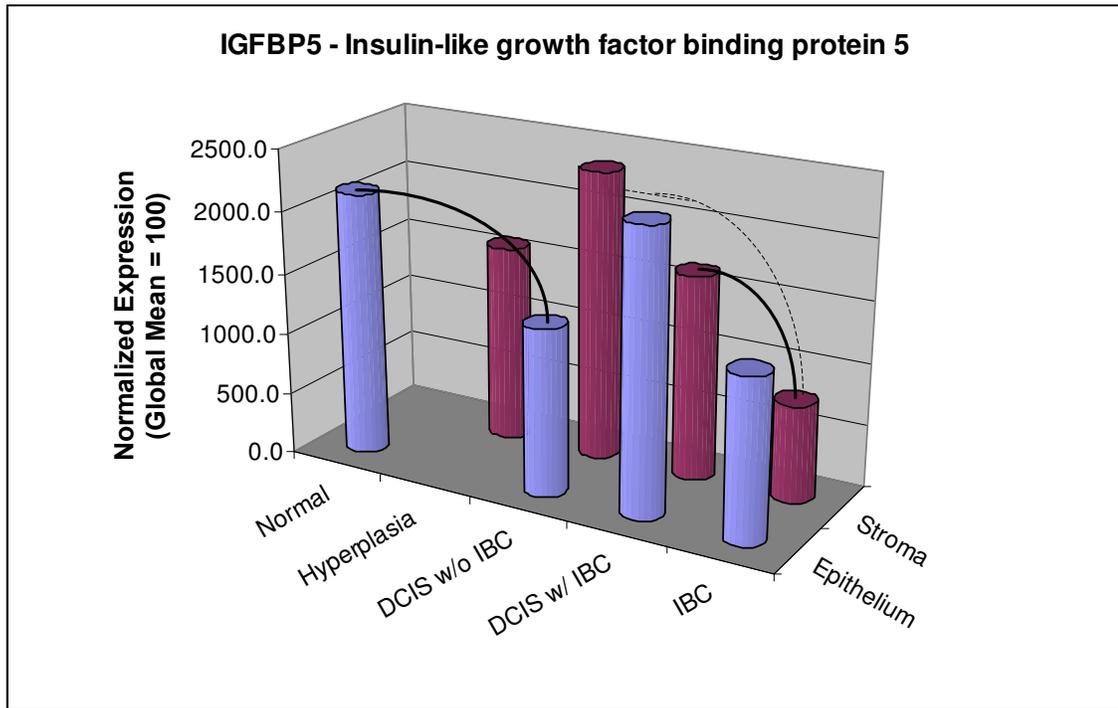
Cancer cells use a variety of extracellular signaling strategies to protect against their apoptotic programs getting triggered [45]. Cancer cells may secrete abnormally high levels of insulin-like growth factors-1 and -2 (IGF-1 and IGF-2), which are trophic (survival) signals in the extrinsic apoptotic pathway, in an autocrine fashion to protect themselves from externally triggered apoptosis. Another common strategy is to reduce the level of IGF-binding proteins (IGFBPs) in the extracellular

space, which bind and sequester IGFs, in order to reduce the threat to the cancer cell of losing trophic signaling. Indeed, suppression of IGFBP serves multiple purposes for the cancer cell since expression of IGF-binding protein-3 (IGFBP-3) and IGFBP-5 in human breast cancer cells may also induce apoptosis directly via modulations in Bcl-2 proteins, suggesting that these IGFBPs induce an intrinsic apoptotic pathway as well [46]. Also, a very recent proteomics study [47] has further established that IGFBP3 is expressed in the normal breast epithelial cells (not stroma) where it plays a paracrine inhibitory role in breast tumor development.

From the current literature, the expected result of IGFBP3 and IGFBP5 gene expression would be down regulation in the epithelial groups. Figures 32 and 33 show the results obtained in this study.



**Figure 32.** IGFBP3 normalized gene expression (line shows statistical significance)



**Figure 33.** IGFBP5 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

In the case of IGFBP3 (Figure 32), gene expression across the spectrum of epithelial disease states appears flat, and certainly not down regulated relative to normal tissue. Indeed, IGFBP3 expression is somewhat higher in the disease states versus Normal epithelium. However, IGFBP3 down regulation in IBC relative to IBC Stroma is statistically significant, which is consistent with the carcinoma taking measure to protect itself from extrinsic apoptosis triggered by loss of IGF topographic factors. Note that the group means in the IBC group pair and the DCIS Without IBC group pair shown in Figure look very similar, but only the IBC pair is statistically significant (Table 4), reinforcing the notion that these plots must be used with care and cannot be used in isolation to draw conclusions about gene regulation.

For IGFBP5 (Figure 33), the downward trend in group means is striking and clear in the stromal groups but less so in the epithelial groups. Statistically, the down regulation of IGFBP5 in the stromal groups is only significant between the DCIS and IBC groups. This statistical significance holds regardless of whether the DCIS With and Without IBC are split out or combined. This result would be expected in the epithelial groups, but its finding in the stromal groups is surprising.

In contrast to its irrelevance for stromal groups, the DCIS With and Without IBC split does make a difference in the results for the epithelial groups. Down regulation in the epithelial groups would be expected and indeed is found, but only when DCIS is split into the two subgroups of DCIS With and Without IBC. In this case, the IGFBP5 down regulation is statistically significant between the Normal and DCIS Without IBC. However, combining the DCIS subgroups into one causes IGFBP5 to disappear from the significance list. This result make sense intuitively from the plot of means (Figure 33) which would indicate combining these subgroups would blunt the sharp drop that otherwise appears when they are split out.

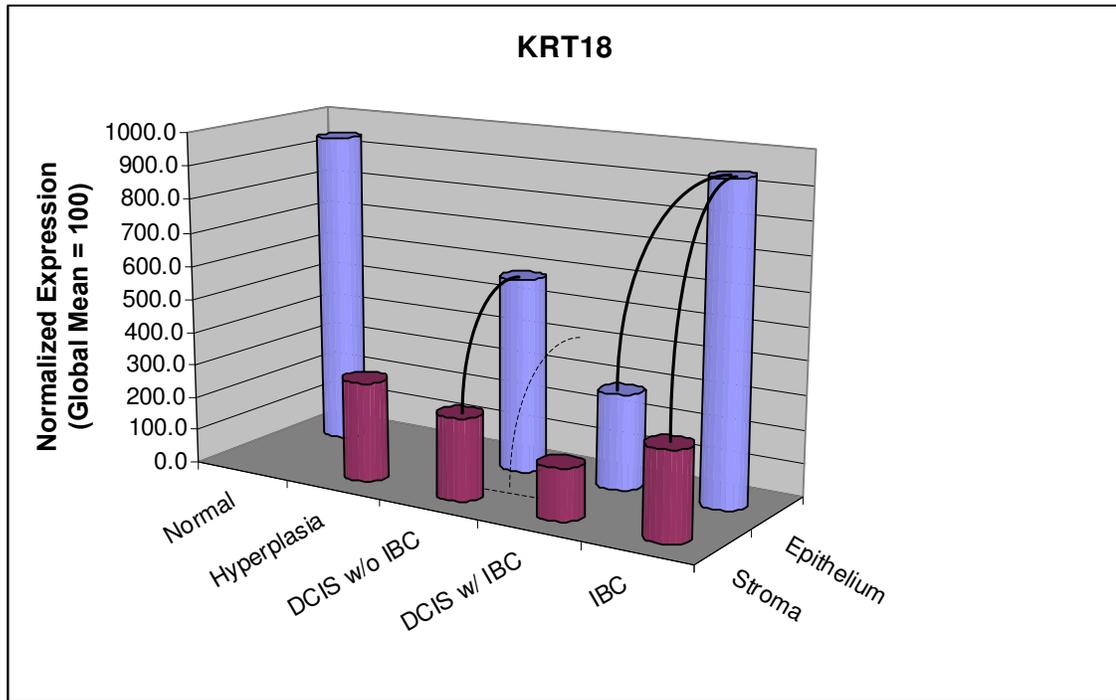
With the DCIS With and Without IBC groups split out, an intriguing scenario becomes possible. Recall that IGFBPs are released to the extracellular space where they bind to trophic IGFs and sequester them. Thus, the cancer benefits no matter where the down regulation of IGFBPs occurs, that is, regardless of whether it occurs in the epithelial tissues or the nearby stroma. With this and Figure 8 in mind, a possible scenario becomes clear where the statistically significant down regulation of IGFBP5 occurs initially in the epithelial tissue (significant between Normal and DCIS Without IBC). Then, through an undetermined mechanism, the stroma is induced to down

regulate IGFBP5 as the disease progresses from DCIS to IBC, which is significant with either grouping. The findings here would support (but certainly not prove) just such a “hand off” scenario between the DCIS epithelium just before IBC and the stroma just after invasion, and it would be consistent with the well established intensifying of signaling activity between epithelial tissue and stroma as the cancer becomes invasive.

### **KRT18 and KRT19 – Keratin 18 and 19**

The results for these two genes are presented here because they appear so frequently as significantly up regulated between the stromal and epithelial groups. They appear as consistently “up regulated,” statistically speaking, in epithelial groups mainly because these proteins are simply much more common in epithelial tissue than in stroma. Recall that the same breast cancer biomarker array is used throughout on both breast cancer and stromal samples.

KRT18 and KRT19 are cytokeratins (CKs), which have long been recognized as structural markers specific to epithelial cells. Thus, when comparing epithelial groups to stromal groups, epithelial-specific genes should have noticeably higher expression. The results bear this out with nearly order-of-magnitude differences in relative expression levels for some parallel group comparisons, when normalized across all groups, as shown in Figures 10 and 11 below.

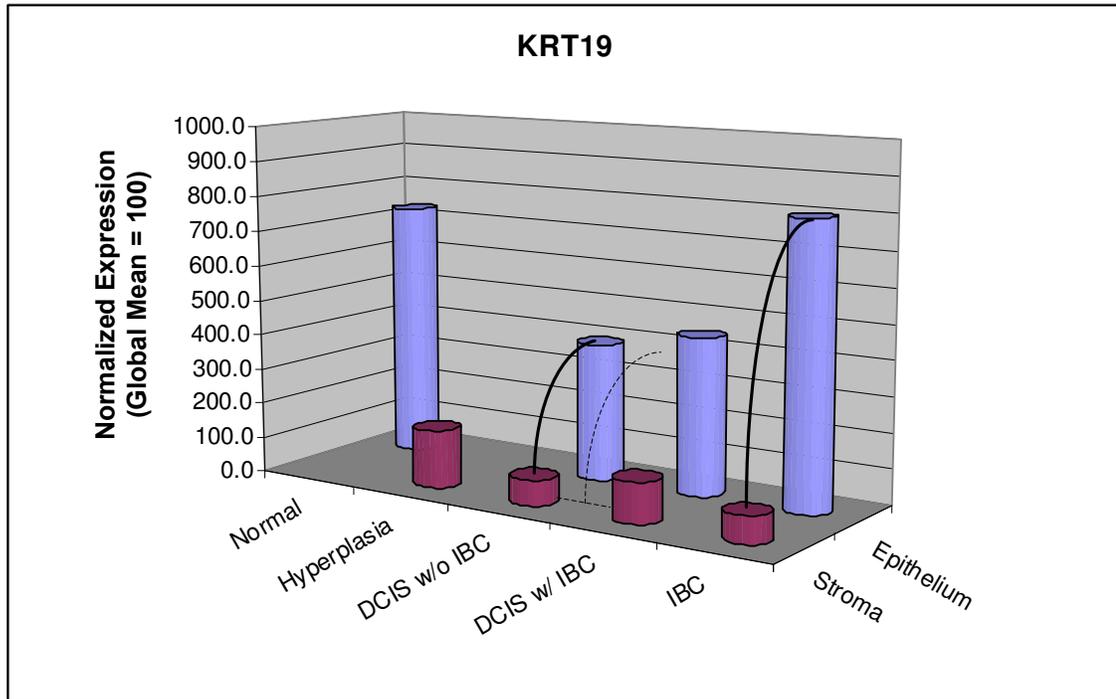


**Figure 34.** KRT18 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

Despite the apparently wide swings in gene expression in the epithelium-to-epithelium comparisons, only one yielded statistically significant differences for KRT18 and KRT19, namely up regulation between DCIS With IBC and IBC for KRT18. All other significance was between epithelial and stromal group pairs, which would be expected due to the CKs being epithelial markers.

KRT19 (also known CK19) has been identified as being highly expressed in HER-2/neu-positive breast tumors [48]. The sample sets represented in Figures 34 and 35 have not been classified with respect to HER-2/neu (ERBB2) status, although Figure 4 would suggest some degree of HER-2/neu positive status. Note in Figure 35 that mean KRT19 expression in the IBC epithelium group is essentially equal to the

Normal epithelium group. Perhaps more interesting is the apparent dip in KRT19 epithelial expression during the progression from Normal to IBC, although no epithelial group pair differences proved statistically significant.



**Figure 35.** KRT19 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

Likewise, KRT18 shows a noticeable dip and a trend similar to KRT19. Even with the DCIS groups combined, there would be a noticeable (roughly a full fold down) dip in the progression from Normal to IBC epithelia. This trend may explain some apparent discrepancies in the literature. Some recent investigators have suggested KRT18 gene expression is down regulated in breast cancer [49, 50] while previous proteomic studies [51, 52] indicated that the reduction of KRT18 in cancer cells is a proteomic process involving cancer-associated cleavage and accelerated protein

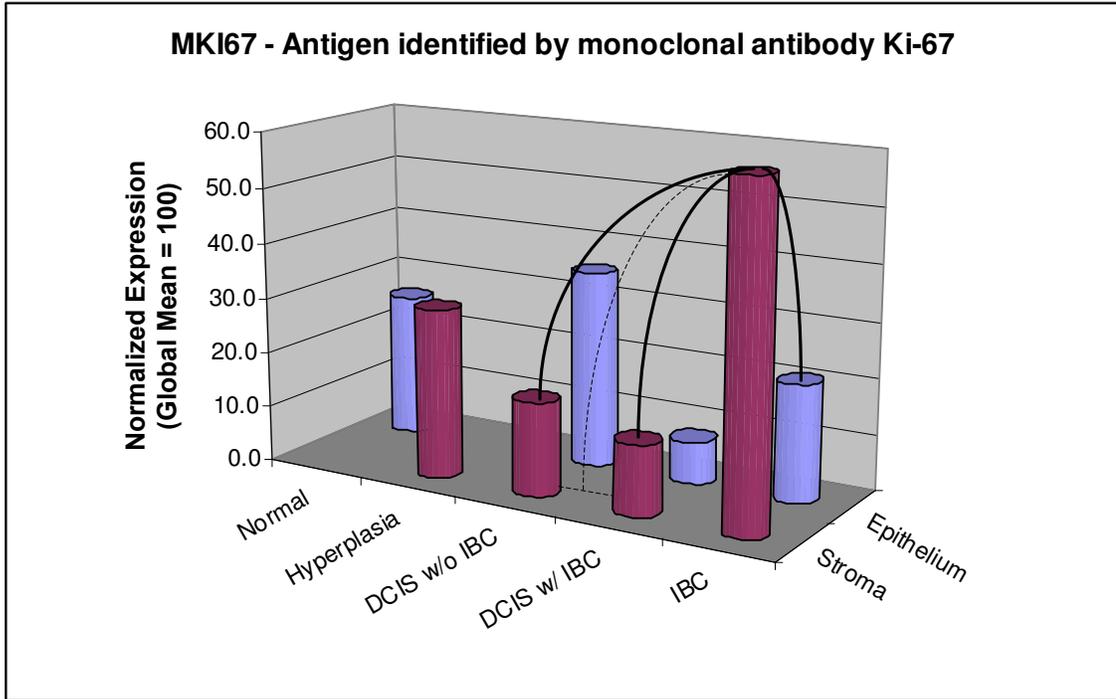
degradation via a ubiquitin-dependent proteasome pathway in breast cancer, a reminder of the danger of ascribing too much causality to gene expression results.

The results presented here certainly do not settle the issue, but they do show an interesting, and consistent, pattern for both genes across the epithelial groups' disease spectrum. The genes' mean expression drops noticeably from Normal to DCIS, consistent with the down regulation of KRT18 reported in the literature, and then rises sharply to return very close to the Normal level in the IBC state. This rebound between DCIS With IBC and IBC is statistically significant.

### **MKI67 – Antigen identified by monoclonal antibody Ki-67**

MKI67, better known as Ki-67, is a well know proliferation marker in human breast and other cancers. However, even very recent studies examining the role of breast tumor stroma have only reported correlations between the presence of stromal myofibroblasts in tumor stroma and expression of Ki-67 (and HER-2) “in breast cancer cells [53].” The results presented in the figure below clarify that the strong correlation of up regulated (or amplified) MKI67 expression that has been in use for some time as a prognostic marker for advanced breast carcinoma is actually occurring in the stroma, not the carcinoma. The up regulation is statistically significant between DCIS and IBC Stroma groups, regardless of the DCIS With and Without IBC split out. Again, the precision of laser capture microdissection (LCM) reveals the specific tissue cell type responsible for the gene expression observed historically in homogenized sections. As LCM comes into more routine use, this finding may be valuable to future investigators as it suggests MKI67 will not be as useful a prognostic indicator as it has been when dealing with highly purified

epithelial carcinoma cell samples, unless corresponding stromal samples are also examined.



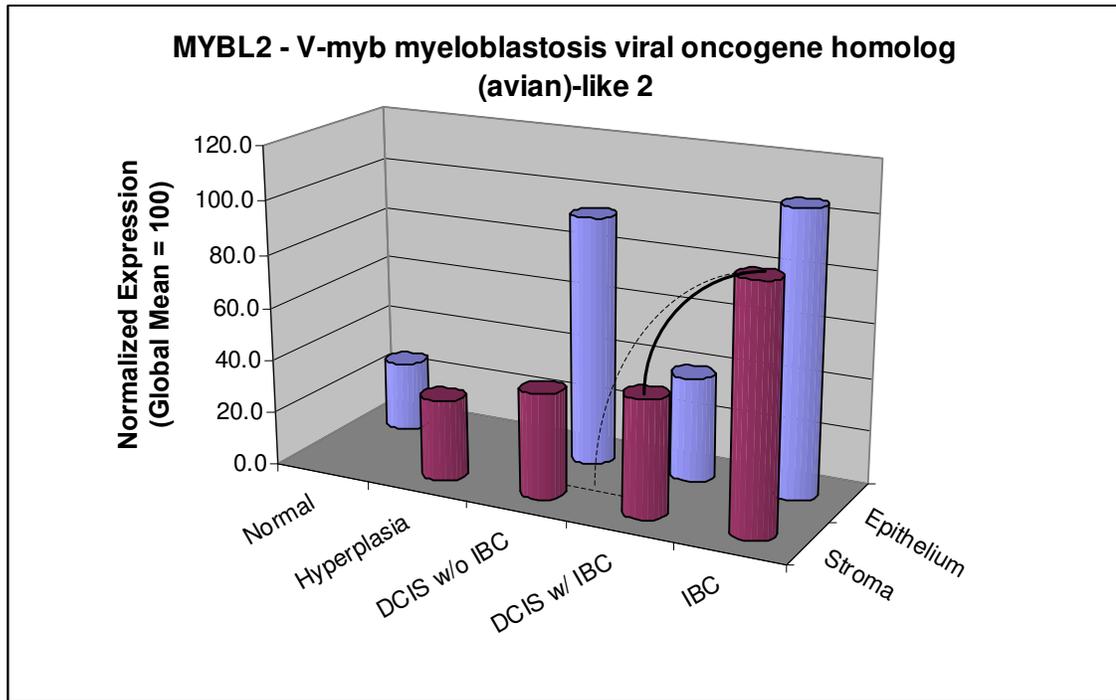
**Figure 36.** MKI67 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

### **MYBL2 - V-myb myeloblastosis viral oncogene homolog avian-like 2**

MYBL2 is also a well-established prognostic marker for breast and other types of carcinoma. High levels of MYBL2 expression are generally associated with amplified 20q12-q13 regions, which is common in breast cancer [54]. As Figure 37 below shows, the findings here agree with high levels of MYBL2 expression from DCIS through IBC epithelial groups as expected. The drop at DCIS With IBC epithelium could be spurious, due to small sample size, but the same elevated levels are seen at DCIS generally, whether split out or not. However, these results were not

statistically significant. In contrast, the up regulation of MYBL2 in the stroma of IBC versus DCIS stroma was statistically significant, regardless of the DCIS With and Without IBC split out. After staying relatively flat from Hyperplasia Stroma through DCIS With IBC Stroma, MYBL2 expression in IBC stroma rises sharply.

Since high MYBL2 expression is generally associated with amplified 20q12-q13 regions common in breast cancer, the findings here raise the question of what causes the sharply higher expression in IBC Stroma. Further, since the measurements presented here are all RNA based, high transcript levels must result from either increased transcription of normal chromosomal DNA or amplified DNA (or both). At least one study has shown that concurrent and independent genetic alterations (e.g., loss of heterozygosity (LOH) and genetic alterations on several chromosomes) in mammary stroma not only occur but that these changes may precede genotypic changes in the epithelial cells [55]. However, that study did not specifically examine 20q12-q13 amplification. Instead, it focused on LOH at other chromosomal locations, and the findings here certainly do not suggest the stroma is leading the carcinoma with respect to MYBL2 expression levels. Nonetheless, the findings do raise the interesting question of what is causing statistically significant higher levels of MYBL2 expression in the stroma (of IBC) when this gene's over expression is generally associated with gene amplification, but gene amplification in turn is normally associated with the carcinoma, not the stroma.

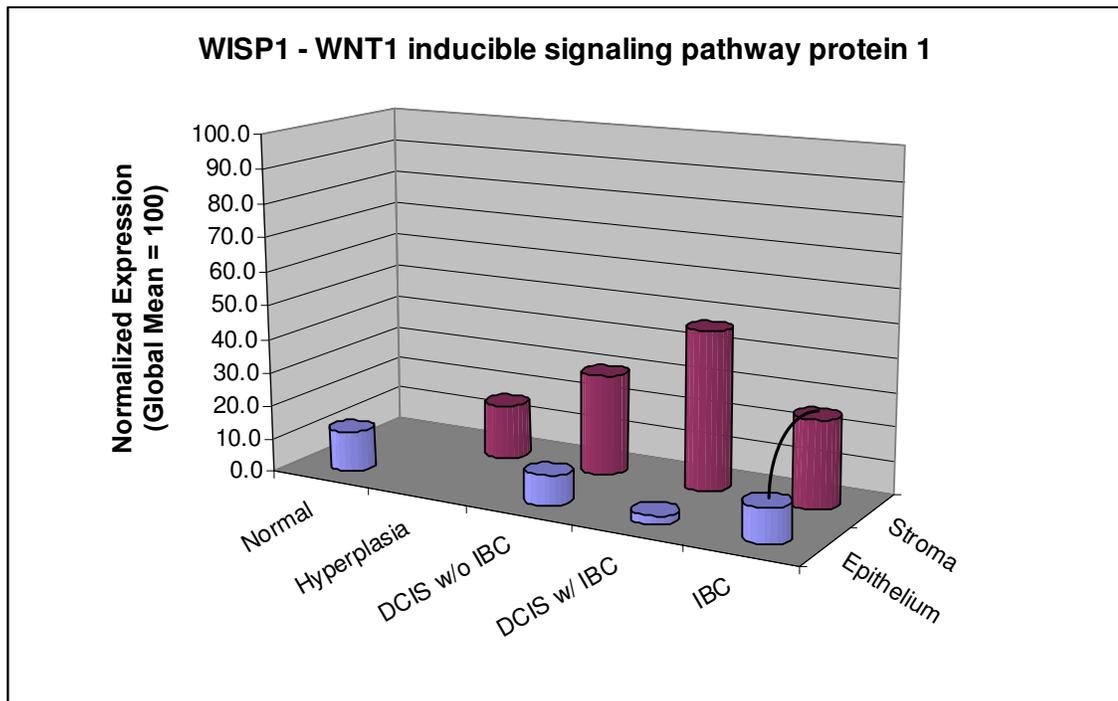


**Figure 37.** MYBL2 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

### **WISP1 - WNT1 inducible signaling pathway protein 1**

WISP1 encodes a protein that belongs to the connective tissue growth factor (CTGF) family and may be downstream in the WNT1 signaling pathway that is relevant to malignant transformation. It is expressed in high levels in fibroblasts, which would be consistent with higher expression in stroma. The encoded protein binds to proteoglycans in the ECM and possibly prevents their inhibitory activity in tumor cell proliferation. It also attenuates p53-mediated apoptosis. Two studies [56, 57] established the association of WISP1 with advanced features of breast carcinoma. However, neither of these studies (circa 2001) used LCM, and both refer to up regulation of WISP1 specifically in breast epithelial cells. The results here show that the elevated WISP1 expression is actually occurring in the stroma, not the epithelium

(Figure 38). Although the difference in the means looks more dramatic at DCIS With IBC, the statistically significant finding is actually at the IBC step. Although not statistically significant, it is worth noting that rising WISP1 expression is not only occurring in the stroma, it appears to start rising noticeably during DCIS, peaking in relative terms just as DCIS transitions into IBC.



**Figure 38.** WISP1 normalized gene expression (lines indicate statistical significance, dashed lines indicate statistical significance with DCIS groups combined)

### Transforming Growth Factors

The absence of some genes from the significance lists may be as notable as those present. Notably absent from the significance lists are the Transforming Growth Factors (TGFs). The breast cancer biomarker array used includes the genes TGF $\alpha$ , TGF $\beta$ 1, TGF $\beta$ 2, and TGF $\beta$ 3. TGF $\beta$  is known to play a key role in breast cancer [58],

but its role is complex, often ambiguous, and even paradoxical [59, 60] Further, its role does not necessarily translate into straightforward up or down regulation in one or the other tissue type examined here. TGF $\beta$  is produced by many cell types in an inactive form that is released and subsequently resides in the ECM until it is liberated by matrix metalloproteinase (MMP) cleavage or other mechanisms.

The measurements of the four TGF genes included on the breast cancer biomarker array did not yield any statistically significant up or down regulation results for any group-wise comparison examined. Further, in absolute terms, the signals for these genes were very low. These results suggest that either the TGF $\beta$  is being produced elsewhere or its main activation mechanism is driven more by proteomic or signaling effects, either intracellular, extracellular, or both, rather than by gene regulation. Further analyses of the samples using alternative techniques, such as Enzyme-Linked ImmunoSorbent Assay (ELISA) testing, would be required to investigate this issue further.

## ***Validation and Sensitivity Analyses***

### **Log Normalization**

Since the main results presented here all resulted from using non-log normalized data (albeit scaled to a common mean), several comparisons are recomputed using  $\log_2$  transformed data. This log normalization had no meaningful effect on the results, so using the non-log normalized data was deemed acceptable.

## **P-Value Computation**

All p-values are computed by two independent means, for validation purposes, as discussed in the Results section (see Figure 21). The first method uses an open source C# (.NET) implementation of Mann-Whitney from ALGLIB, and the second uses the widely used Bioconductor `stats` library in R. The R results are generally greater than (more conservative than) the ALGLIB results, thus genes are not reported as statistically significant ( $\leq 5\%$ ) unless both methods reported a value of  $\leq 0.05$ . Thus, the R results tend to act as a filter on the ALGLIB results, reducing the set of significant genes in either direction (up or down regulation). A simple and conservative implementation of the Benjamini – Hochberg False Discovery Rate is applied separately, following both computational methods. For the ALGLIB method, SinChMAT applies the correction programmatically. For the R results, the correction is applied via a simple spreadsheet manipulation. Since the FDR (and the Bonferroni correction) tend to eliminate most genes from the significance list, if *either* p-value computational method generates values that survive the multi-test correction, it is retained and reported.

## **Group-wise Normalization**

Since simple scaling is used across all arrays for normalization, a group-wise normalization is attempted as a sensitivity analysis, to assess the impact of using simple scaling. As noted, the SuperArray (SA) website tools available to perform group-wise normalization have significant limitations. Nevertheless, group-wise normalizations are attempted by first analyzing the statistical variability of the control spots, which would be used in a group-wise normalization, with respect to background corrections. With simple scaling applied, the controls with the lowest

coefficient of variability (mean divided by standard deviation) across all background corrections are selected for each condition group (before the DCIS With and Without IBC split out, for simplicity).

With the set of selected controls for each group, the SA website tools are used to perform normalization against the specified controls, and the results are downloaded and parsed to isolate the array results for just those members of the given condition group. This process is repeated for each group. Unfortunately, since the SA website tools artificially compress all values down to 2 significant digits, many values cluster around zero and 1.00, losing resolution at the important endpoints, particularly around 1.00. Not surprisingly, this compression results in a much greater number of “significant” genes, many of which are likely false positives. Also, the resulting sets of supposedly significant genes are much larger relative to total number of genes on the array, further calling into question their meaningfulness. Finally, the results are much more asymmetrical (up versus down), which further indicates poor quality of these results. Thus the group-wise normalization was rejected and not reported, and would likely have to be done manually as part of future work to produce meaningful results.

Interestingly, despite these issues, the few genes surviving the FDR correction using the group-wise normalized data still matched very closely to the simple scaling results after FDR correction, with occasionally an addition gene or two, providing further confidence in those most significant findings using simple scaling.

## **Background Corrections**

Careful background correction clearly affects the results of microarray analyses. To assess the sensitivity of the results presented here to background correction, a simple sensitivity analysis is performed where a given background correction technique (e.g., Empty Spots, Minimum, Local, or Global) is applied uniformly to all arrays and the p-values recomputed. The results are indeed sensitive to the background correction. The numbers and identities of the genes reported as significant change noticeably when the same background correction is used for all arrays, confirming the importance of the array-level background corrections that are used in all results presented.

## **Fold Changes with Absent / Present Threshold**

With the limitations of fold change analysis in mind (see Computational Analysis Approach section), the SuperArray website software was used to perform a confirmatory fold change check on the more rigorous statistical significance results. A fold change threshold of  $\geq 1.5$  is used along with an Absent / Present (AP) threshold of 1.2, as determined by the website software's image analysis algorithm. A spot is considered "Present" if its average pixel density is greater than its local background, and the average density of the spot is greater than 1.5x of the mean value of the local backgrounds of the lower 75<sup>th</sup> percentile of all non-bleeding spots. The results are shown in the table below. Note that the fold change analysis is performed using group means.

**Table 5.** Fold change ( $\geq 1.5$ ) with Absent/Present (AP) threshold ( $\geq 1.2$ ) applied (Genes in *italics* had fold changes  $\geq 2.0$ )

Condition 1	Condition 2	Up	Down
Stroma	Epithelium		
DCIS w/o IBC (N = 8)	DCIS w/o IBC (N = 7)	-	EGLN1, IGFBP5, RP5-860F19.3
DCIS with IBC (N = 3)	DCIS with IBC (N = 3)	JUN, KRT18	CD68, CDC25B, PDPK1
IBC (N = 9)	IBC (N = 15)	-	-
Stroma	Stroma		
Hyperplasia (N = 4)	DCIS w/o IBC (N = 8)	EGLN1, IGFBP5, RP5-860F19.3	
Hyperplasia (N = 4)	DCIS with IBC (N = 3)	EGLN1, RB1, RP5-860F19.3	<i>IGFBP3</i>
Hyperplasia (N = 4)	IBC (N = 9)	-	IGFBP3, IGFBP5, RPS4X
DCIS w/o IBC (N = 8)	DCIS with IBC (N = 3)	-	IGFBP5
DCIS w/o IBC (N = 8)	IBC (N = 9)	-	EGLN1, IGFBP5, RP5-860F19.3
DCIS with IBC (N = 3)	IBC (N = 9)	CDK4, EVL, IGFBP3, <i>PFKP</i>	BAX, EGLN1, HMGB3, IGFBP5, RP5-860F19.3
Epithelium	Epithelium		
Normal (N = 5)	DCIS w/o IBC (N = 7)	MAPK3	CIRBP, IGFBP5, KRT18, RPS4X
Normal (N = 5)	DCIS w/ IBC (N = 3)	EGLN1, <i>HMGB3</i> , IGFBP5, MAPK3, RB1, RP5-860F19.3	CIRBP, <i>KRT18</i> , RPS4X
Normal (N = 5)	IBC (N = 15)	HMGB3	IGFBP5, RPS4X
DCIS w/o IBC (N = 7)	DCIS w/ IBC (N = 3)	DEGS1, EGLN1, <i>HMGB3</i> , IGFBP5, RB1, RP5-860F19.3	CDK4, <i>KRT18</i>
DCIS w/o IBC (N = 7)	IBC (N = 15)	-	-
DCIS w/ IBC (N = 3)	IBC (N = 15)	CDK4, EVL, <i>KRT18</i>	EGLN1, HMGB3, IGFBP5, RB1, RP5-860F19.3, RPS4X

Comparison of these results to Table 4 (summary of hypothesis testing results) and the gene-level results figures in the Discussion section shows generally confirmatory overlap. Although fold change analysis lacks the statistical rigor of the formal hypothesis testing, the fold change results are consistent with the figures in the Discussion section, which show the corresponding changes up or down in group means. However, the fact that there is significant disjointedness between the two

sets of results reinforces the danger of relying solely on fold change results without formal statistical analysis and reinforces the warnings in the Discussion section about drawing conclusions based solely on group mean trends. Nonetheless, the fold change results make biological sense to first order and they are generally confirmatory of the p-value results where they overlap. Further, the fold change results incorporate AP calls, which in turn adjust for bleeding. Since these refinements are lacking in the p-value results presented, the fold change results are presented here without further discussion as possible leads for refinements and future work by other investigators.

## Conclusions

It is difficult to draw general conclusions using only gene expression results. To do so would ignore complex and important proteomic and signaling effects.

Nonetheless, it is possible to draw novel conclusions within the realm of gene expression by taking advantage of laser capture microdissection (LCM) to refine gene expression analysis using highly purified, cell-type-specific samples across a range of disease states and tissue types. The results presented here are for single channel microarray experiments where a small, targeted set of genes, namely breast cancer biomarkers, are measured directly in groups of samples sharing a common biological condition and composed of cell type specific, highly homogeneous cell extracts (epithelium or stroma) relevant to breast cancer. Various combinations of these condition-tissue type groups are then compared pair-wise for insights into changes in gene expression patterns in both epithelia and stroma in progressive breast carcinoma.

The specific genes selected by the array manufacturer as breast cancer biomarkers are based on past studies that generally used breast cancer sections rather than LCM-derived samples and thus probably included stromal cells intermixed with cancerous epithelial cells. The results presented here allow further resolution of which specific cell populations were responsible for the changes historically observed in "homogenized" samples. The results also provide a more comprehensive or

“system level” view of both cell type populations of interest over a spectrum of progressive disease conditions.

In summary, the findings are:

1. Some breast cancer biomarker genes previously classified as up regulated in “breast carcinoma” are actually up regulated in the stroma, not the epithelium. They are: MKI67, MYBL2, and WISP1.
2. The “system level” view suggests (but does not prove) some intriguing possibilities in the changes in gene expression patterns in a two dimensional space of disease progression and tissue type (epithelium versus stroma) that may warrant further investigation. They are:
  - a. Although GRB7 is generally associated with invasiveness (motility, cell migration), its expression in epithelia seems to be quite high well before IBC is reached. Also, it appears to be up regulated in both epithelia and stroma.
  - b. With DCIS sub-divided into With and Without IBC groups, HMGB3 displays a pattern of expression where both epithelial and stromal groups move in nearly lock step across the entire disease progression spectrum, with a statistically significant peak at DCIS With IBC for both cell types.
  - c. IGFBP5 trends across the entire space could be (speculatively) interpreted as exhibiting a “hand off” effect where the gene is initially down regulated in the epithelium from Normal through DCIS and then is somehow down regulated in the stroma from DCIS to IBC.

- d. Both KRT18 and KRT19 (epithelial groups) display a noticeable and pronounced trough between Normal and IBC, with these endpoints providing the “walls” of the trough. This pattern may explain why some studies report KRT19 up regulated while others report KRT18 down regulated in breast cancer.
3. More generally, the results presented here suggest that the gene sets used on targeted breast cancer biomarker arrays (and perhaps those of other cancer arrays) should be revised, or possibly split into pairs of arrays, to account for the now well-established role of tumor microenvironments and to facilitate the increasing use of LCM in microarray studies.

## References

1. Gulisa T, Bouchal J, et al.: **Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis**, *BMC Cancer*, 2007. **7**:55.
2. Leonard, G.D., et al., **Ductal carcinoma in situ, complexities and challenges**. *J Natl Cancer Inst*, 2004. **96**:906-920.
3. Jemal A, Ward E, and Thun MJ: **Recent trends in breast cancer incidence rates by age and tumor characteristics among U.S. women**, *Breast Cancer Research*, 2007. **9**-R28.
4. **A sharp decrease in breast cancer incidence in the United States in 2003**, *Proceedings from the 2006 annual San Antonio Breast Cancer Symposium (SABCS)*.
5. Sanders, M.E., et al., **The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up**, *Cancer*, 2005. **103**:2481-2484.
6. Tavssoli FA, Schnitt SJ: *Pathology of the breast*, New York: Elsevier; 1992.
7. Zhu G, et al.: **Combination of microdissection and microarray analysis to identify gene expression changes between differentially located tumor cells in breast cancer**, *Oncogene*, 2003. **22**:3742-3748.
8. Schuetz CS, et al.: **Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis**, *Cancer Res*, 2006. **66**:5278-5286.
9. Vant Veer, LJ, et al.: **Gene expression profiling predicts clinical outcome of breast cancer**, *Nature*, 2002. **415**:530-535.
10. Raju M, et al.: **Molecular classification of breast carcinoma in situ**, *Curr Genomics*, 2006. **7**(8):523-532.
11. ALexe G, et al., **Analysis of breast cancer progression using principal component analysis and clustering**, *J Biosci*. 2007. **32**(5):1027-39.
12. Weinberg, Robert A., *The Biology of Cancer*, Garland Science, Taylor and Francis Group, LLC: 2007
13. Shekhar MP, et al., **Breast stroma plays a dominant regulatory role in breast epithelial growth and differentiation: implications for tumor development and progression**, *Cancer Res*. 2001. **61**(4):1320-6.

14. Bissell MJ and Radisky D, **Putting tumours in context**, *Nat Rev Cancer*. 2001. **1**(1):46-54.
15. Liotta, L.A., et al., **The microenvironment of the tumour-host interface**, *Nature* 2001. **411**:375-379.
16. Littlepage LE, et al., **Coevolution of cancer and stromal cellular responses**, *Cancer Cell*. 2005. **7**:499-500.
17. Ronnov-Jensen, L, et al., **Cellular Changes Involved in Conversion of Normal to Malignant Breast: Importance of the Stromal Reaction**, *Physiol Rev*, 1996. **76**:69-125.
18. Nakamura T, et al., **Induction of hepatocyte growth factor in fibroblasts by tumor-derived factors affects invasive growth of tumor cells: in vitro analysis of tumor-stromal interactions**, *Cancer Res*, 1997. **57**(15):3305-13.
19. Barcellos-Hoff MH and Ravani SA, **Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells**, *Cancer Res*, 2000. **60**(5):1254-60.
20. Moustakas A and Heldin CH, **Signaling networks guiding epithelial-mesenchymal transitions during embryogenesis and cancer progression**, *Cancer Sci*, 2007. **98**(10):1512-20.
21. Akhurst RJ and Derynck R, **TGF-beta signaling in cancer--a double-edged sword**, *Trends Cell Biol*, 2001. **11**(11):S44-51.
22. Petersen OW, et al., **Epithelial to mesenchymal transition in human breast cancer can provide a nonmalignant stroma**, *Am J Pathol*, 2003. **162**(2):391-402.
23. De Wever O, et al., **Critical role of N-cadherin in myofibroblast invasion and migration in vitro stimulated by colon-cancer-cell-derived TGF-beta or wounding**, *J Cell Sci*, 2004. **117**(Pt 20):4691-703.
24. Moinfar F, et al., **Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis**, *Cancer Res*, 2000. **60**(9):2562-6.
25. National Institutes of Health, NCICGAP website, Introduction to Laser Capture Microdissection, <http://dir.nichd.nih.gov/lcm>, June 2007.
26. Espina V, et al., **Laser-capture microdissection**, *Nature Protocols*, 2006. **1**(2):586-602.
27. PicoPure™ RNA Isolation Kit User Guide, Catalog #KIT0202 / KIT0204, Version D (For Research Use Only), Arcturus Bioscience, Inc., Mountain View, CA.

28. TrueLabeling-PicoAMP™ User Manual, Part#1021A, Version 1.3, July 9, 2007, SuperArray Bioscience Corporation, Frederick, MD.
29. Oligo GEArray® System User Manual, Part #1018A, Version 3.2, October 20, 2006, SuperArray Bioscience Corporation, Frederick, MD.
30. Fox-Brashears H, et al., *Oligo GEArrays®: The Pathway-Focused DNA Microarray System for Every Laboratory*, SuperArray Bioscience Corporation, Frederick, MD.
31. Hu Y., et al.: **Analysis of genomic and proteomic data using advanced literature mining**, *J Proteome Res*, 2003. **2**(4):405-12.
32. Oligo GEArray® Human Breast Cancer Biomarker Microarray ([http://www.superarray.com/gene\\_array\\_product/HTML/OHS-402.html](http://www.superarray.com/gene_array_product/HTML/OHS-402.html)), SuperArray Bioscience Corporation, Frederick, MD.
33. Draghici, Sorin, *Data Analysis for DNA Microarrays*, Chapman & Hall / CRC: 2003.
34. Mann HB, Whitney DR: **On a test of whether one of two random variables is stochastically larger than the other**, *Annals of Mathematical Statistics*, 1947. **18**:50-60.
35. Wilcoxon F: **Individual comparisons by ranking methods**, *Biometrics Bulletin*, 1945. **1**:80-83
36. Abdi, H.: **Bonferroni and Sidak corrections for multiple comparisons**. In *Salkind NJ (Ed.): Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA.
37. Benjamini, Y, and Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**, *Journal of the Royal Statistical Society, Series B (Methodological)* 1995, **57**:125-133
38. ALGLIB open source mathematical libraries, C# Mann-Whitney implementation by Sergey Bochkanov, [www.alglib.net/statistics/hypothesistesting/mannwhitneyu.php](http://www.alglib.net/statistics/hypothesistesting/mannwhitneyu.php)
39. Multiple Testing Corrections (White Paper), Agilent Technologies, Inc., 2005, [www.chem.agilent.com/cag/bsp/products/gsgx/Downloads/pdf/mtc.pdf](http://www.chem.agilent.com/cag/bsp/products/gsgx/Downloads/pdf/mtc.pdf)
40. Lorenzi PL, et al., **Asparagine synthetase as a causal, predictive biomarker for L-asparaginase activity in ovarian cancer cells**, *Mol Cancer Ther*, 2006. **5**(11):2613-23.

41. Iwamoto S, et al., **Mesenchymal cells regulate the response of acute lymphoblastic leukemia cells to asparaginase**, *J Clin Invest*, 2007. **117**(4):1049-57.
42. Layfield LJ and Lewis C, **In situ and invasive components of mammary adenocarcinoma: comparison of Her-2/neu status**, *Anal Quant Cytol Histol*, 2007. **29**(4):239-43.
43. Vinatzer U, et al., **Expression of HER2 and the coamplified genes GRB7 and MLN64 in human breast cancer: quantitative real-time reverse transcription-PCR as a diagnostic alternative to immunohistochemistry and fluorescence in situ hybridization**, *Clin Cancer Res*, 2005. **11**(23):8348-57.
44. Nemeth MJ, et al., **Hmgb3 regulates the balance between hematopoietic stem cell self-renewal and differentiation**, *Proc Natl Acad Sci*, 2006. **103**(37):13783-8
45. Weinberg, Robert A., *The Biology of Cancer*, Garland Science, Taylor and Francis Group, LLC: 2007
46. Butt AJ, et al., **Enhancement of tumor necrosis factor-alpha-induced growth inhibition by insulin-like growth factor-binding protein-5 (IGFBP-5), but not IGFBP-3 in human breast cancer cells**, *Endocrinology*, 2005. **146**(7):3113-22.
47. Toillon RA, et al., **Proteomics demonstration that normal breast epithelial cells can induce apoptosis of breast cancer cells through insulin-like growth factor-binding protein-3 and maspin**, *Mol Cell Proteomics*, 2007. **6**(7):1239-47
48. Zhang DH, et al., **Proteomics of breast cancer: enhanced expression of cytokeratin19 in human epidermal growth factor receptor type 2 positive breast tumors**, *Proteomics*, 2005. **5**(7):1797-805.
49. Buhler H and Schaller G, **Transfection of keratin 18 gene in human breast cancer cells causes induction of adhesion proteins and dramatic regression of malignancy in vitro and in vivo**, *Mol Cancer Res*, 2005. **3**(7): 365-71.
50. Woelfle U, et al., **Down-regulated expression of cytokeratin 18 promotes progression of human breast cancer**, *Clin Cancer Res*, 2004. **10**(8):2670-4.
51. Iwaya K, et al., **Ubiquitin-immunoreactive degradation products of cytokeratin 8/18 correlate with aggressive breast cancer**, *Cancer Sci*, 2003. **94**(10):864-70.

52. Ditzel HJ, et al., **Cancer-associated cleavage of cytokeratin 8/18 heterotypic complexes exposes a neoepitope in human adenocarcinomas**, *J Biol Chem*, 2002. **277**(24):21712-22.
53. Surowiak P., et al., **Stromal myofibroblasts in breast cancer: relations between their occurrence, tumor grade and expression of some tumour markers**, *Folia Histochem Cytobiol*, 2006. **44**(2):111-6
54. Ginestier C, et al., **Prognosis and gene expression profiling of 20q13-amplified breast cancers**, *Clin Cancer Res*, 2006. **12**(15):4533-44.
55. Moinfar F, et al., **Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis**, *Cancer Res*, 2000. **60**(9):2562-6.
56. Xie D, et al., **Elevated levels of connective tissue growth factor, WISP-1, and CYR61 in primary breast cancers associated with more advanced features**, *Cancer Res*, 2001. **61**(24):8917-23.
57. Saxena N, et al., **Differential expression of WISP-1 and WISP-2 genes in normal and transformed human breast cell lines**, *Mol Cell Biochem*. 2001. **228**(1-2):99-104.
58. Ehata S, et al., **Transforming growth factor-beta promotes survival of mammary carcinoma cells through induction of antiapoptotic transcription factor DEC1**, *Cancer Res*, 2007. **67**(20):9694-703.
59. Micke P and Ostman A, **Tumour-stroma interaction: cancer-associated fibroblasts as novel targets in anti-cancer therapy?**, *Lung Cancer*, 2004. **45** Suppl 2:S163-75.
60. Pardali K and Moustakas A, **Actions of TGF-beta as tumor suppressor and pro-metastatic factor in human cancer**. *Biochim Biophys Acta*, 2007. **1775**(1):21-62

## CURRICULUM VITAE

John F. King was born August 2, 1961, in Macon, Georgia and is an American citizen. He graduated with honors from Jones County High School, Gray, Georgia in 1979. He attended the Georgia Institute of Technology in Atlanta, Georgia, completing a Bachelor of Science (1983) degree and a Master of Science (1984) degree, both in Nuclear Engineering. He was awarded an Institute of Nuclear Power Operations (INPO) Fellowship to support his graduate work. Mr. King has over twenty years experience as an engineer, analyst, software developer, and manager, working for several different companies in South Carolina, Georgia, and Virginia. He was certified as a Professional Engineer (PE) in Georgia and South Carolina before changing fields to software engineering. Mr. King completed two Graduate Certificates at George Mason University in 2006, one in Information Engineering and the other in Software Engineering. He is currently employed at Future Technologies, Inc., in Fairfax, Virginia, where he is a senior software engineer working on forensic DNA software systems, which are used by criminal justice and missing and unidentified persons organizations. He received his second Master of Science degree in Bioinformatics and Computational Biology from George Mason University in 2007.