ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES IN BREAST CANCER: BRCA1-INDUCED GENE EXPRESSION PROFILES AND META-ANALYSIS GENE SIGNATURE

A THESIS SUBMITTED TO
THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
AND THE INSTITUTE OF ENGINEERING AND SCIENCE OF
BİLKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

By

BALA GÜR DEDEOĞLU

May 2009

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Neşe Atabey

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Işık G. Yuluğ

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Rengül Atalay

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Asst. Prof. Dr. Özlen Konu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Asst. Prof. Dr. Can Akçalı

Approved for the Institute of Engineering and Science

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Science

ABSTRACT

ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES IN BREAST CANCER: BRCA1-INDUCED GENE EXPRESSION PROFILES AND META-ANALYSIS GENE SIGNATURE

BALA GÜR DEDEOĞLU
PhD in Molecular Biology and Genetics
Supervisor: Assoc. Prof. Işık G. Yuluğ
May 2009, 225 Pages

The aim of the first part of this study was to find out the expression profiles of the genes, which were selected from the former BRCA1-induced gene list (OVCA1, OVCA2, ERBIN, RAD21, XRN2, RENT2, SMG1 and MAC30) in normal-matched primary breast tumors and to correlate the gene expression profiles of selected candidate genes with BRCA1 and various pathology parameters. Among the target genes, the expression of ERBIN, SMG1 and RAD21 were found to be highly correlated with that of *BRCA1* both in BRCA1 up- and down-regulated cells and this result was validated with qRT-PCR expression profiling of the eight genes in 32 normal-matched primary breast tumor samples. These genes were found to be discriminative between ER(-) and ER(+) tumors as well as grade 1 and grade 3 tumors. Target genes were also analyzed in independent microarray datasets to assess their predictive power for breast tumor grade, subtype and patient survival. ERBIN, SMG1 and RAD21 were found to have predictive roles in these datasets.

The aim of the second part of the study was to found appropriate reference genes (RGs) for accurate quantification of target gene expressions in breast tumor tissues. The expression patterns of fifteen widely-used endogenous RGs and three candidate genes that were selected through analysis of two independent microarray datasets were determined in 23 primary breast tumors and their matched normal tissues using qRT-PCR. Additionally, 18S rRNA, ACTB, and SDHA were tested using randomly primed cDNAs from 13 breast tumor pairs to assess the rRNA/mRNA ratio. The tumors exhibited significantly lower rRNA/mRNA ratio when compared to their normals. Among the eighteen tested endogenous reference genes, ACTB and SDHA were identified as the most suitable reference genes for the normalization of qRT-PCR data in the analysis of normal-matched tumor breast tissue pairs.

The aim of the third part of this study was to develop a resampling-based meta-analysis strategy. Two independent microarray datasets that contain normal breast, invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC) samples were used for the meta-analysis. The resampling-based meta-analysis has led to the identification of a highly stable set of genes for classification of normal breast samples and breast tumors encompassing both the ILC and IDC subtypes. A subset of this meta-gene list was shown to predict well-established molecular tumor subtypes, e.g., basal vs luminal or ER+/ER-, with high accuracy and sensitivity based on class prediction analysis of existing breast cancer microarray datasets. Expression of selected genes, tested on 10 independent primary IDC samples and matched non-tumor controls by real-time qRT-PCR, supported the meta-analysis results.

ÖZET

MEME KANSERİNDE FARKLILAŞMIŞ İFADE GÖSTEREN GENLERİN ANALİZİ: BRCA1 TARAFINDAN İNDÜKLENEN GEN İFADE PROFİLLERİ VE META-ANALİZ GEN İMZASI

BALA GÜR DEDEOĞLU
Doktora Tezi, Moleküler Biyoloji ve Genetik Bölümü
Supervisor: Doç. Dr. Işık G. Yuluğ
Mayıs 2009, 225 Sayfa

Bu çalışmanın ilk bölümünün amacı normal-eşleştirilmiş primer meme tümörlerinde daha önceki BRCA1 tarafından indüklenen gen listesinden (OVCA1, OVCA2, ERBIN, RAD21, XRN2, RENT2, SMG1 ve MAC30) seçilen genlerin ifade profillerini bulmak ve bu seçilen aday genlerin gen ifade profillerinin BRCA1 ve çeşitli patolojik parametrelerle korelasyonunu araştırmaktır. Hedef genler arasında ERBIN, SMG1 ve RAD21 ifadelerinin BRCA1'in ifadesinin artmış veya azalmış olduğu hücrelerde BRCA1 ile yüksek derecede korelasyon gösterdiği bulundu. Bu korelasyon 32 normal-eşleştirilmiş primer meme tümörü örneğinde sekiz genin qRT-PCR ile ifade profilinin çıkartılmasıyla doğrulanmıştır. Bu genlerin aynı zamanda ER(-) ve ER(+) tümörlerle evre 1 ve evre 3 tümörleri ayırmayı sağladıkları bulunmuştur. Hedef genler ayrıca bağımsız mikrodizin veri setleri kullanılarak meme tümörü evresi, alt tipi ve hasta sağkalımı açısından öngörme güçlerini değerlendirmek üzere analiz edilmişlerdir. Bu veri setleri için ERBIN, SMG1 ve RAD21 öngörme açısından önemli bulunmuştur.

Çalışmanın bu kısmının amacı meme tümörü dokularında hedef gen ifade miktarının hassas bir şekilde belirlenmesi için uygun referans genler (RG'ler) bulmaktı. Sık kullanılan 15 RG'nin ve iki bağımsız mikrodizin veri setinin analizi sonucunda seçilen üç aday genin ifade durumları 23 primer meme tümöründe ve eşleştirilmiş normal dokularında qRT-PCR kullanılarak belirlenmiştir. Ayrıca, 18S rRNA, ACTB, ve SDHA rRNA/mRNA oranını değerlendirmek üzere 13 meme tümörü çiftinden seçkisiz-prime edilmiş cDNA'lar kullanılarak test edildi. Tümörlerde normallerine göre önemli ölçüde düşük rRNA/mRNA oranına sahipti. Test edilen 18 endojen referans geni arasında normal-eşleştirilmiş tümör meme dokusu çiftlerinin analizinde qRT-PCR verilerinin normalizasyonu için en uygun referans genleri olarak *ACTB* ve *SDHA* seçildi.

Çalışmanın üçüncü kısmının amacı örnek sınıfları arasında farklılaşmış ifadenin önemini değerlendirmek için tekrar örnekleme tabanlı bir meta-analiz stratejisi geliştirmekti. Meta-analiz için normal meme, invaziv duktal karsinom (IDC) ve invaziv lobuler karsinom (ILC) örnekleri içeren iki bağımsız mikrodizin veri seti kullanıldı. Tekrar örnekleme tabanlı meta-analiz hem ILC hem IDC alt tiplerini içeren meme tümörleri ve normal meme örneklerinin sınıflandırılması için yüksek düzeyde sabit bir gen seti tanımlanmasını sağlamıştır. Bu meta-gen listesinin bir alt setinin iyi belirlenmiş moleküler tümör alt tiplerini (örn., bazal ve luminal veya ER+/ER-) mevcut meme kanseri mikrodizin veri setlerinin sınıf öngörme analizine dayanılarak yüksek doğruluk derecesiyle ve hassasiyetle öngördüğü gösterilmiştir. Seçilen genlerin 10 bağımsız primer IDC örneği ve eşleştirilmiş tümör olmayan kontrollerinde gerçek zamanlı qRT-PCR ile test edilen gen ifadeleri meta-analiz sonuçlarını desteklemiştir.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Assoc. Prof. Işık G. Yuluğ for her supervision, support and valuable suggestions throughout the course of my studies. It has always been a privilege to work with such a special person and to have her friendship. She was always more than an adviser to me. Her guiding ***light*** always helped me find my way in the science and in my life.

I would like to attend my very special thanks to Assist. Prof. Özlen Konu for sharing her excellent experiences on bioinformatics and for her endless support at every stage of my study. I always felt her friendship throughout the years in Bilkent.

I would like to express my special thanks to Prof. Dr. Mehmet Öztürk for sharing his excellent scientific logic, for his support and instructive comments.

I would like to thank Assoc. Prof. Betül Bozkurt who provided us the surgical materials. In particular, I would like to thank the pathologist, Prof. Dr. Selda Seçkin and Dr. Gülüşan Ergül for their technical support.

I would like to thank to Assoc.Prof. Rengül Atalay and her student Murat İskar for helping me with the promoter analysis.

I am indebted to all my friends for providing a stimulating and fun environment in the lab. I would like to thak them on their friendship and for their endless support at every stage of my thesis work and my life. Many thanks to MBG family, past and present, who were with me during my studies. Thank you for sharing my life.

I would like to thank deeply my husband Rıfat for his endless support in every step of my life and during this thesis study.

Last but not least, I wish to thank my parents and my brother. They gave birth to me, raised me, supported me, taught me, and loved me. To them and to my husband I dedicate this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

APS	Ammonium persulphate
bp	Base Pairs
BCC	Basal Cell Carcinoma
BCC	Breast cancer cells
BRCT	BRCA1 Carboxyl terminus
BSA	Bovine serum albumin
cDNA	Complementary DNA
CHEK	CHK checkpoint homolog (S. pompe)
CP	Crossing Point
Ct	Cycle Threshold
ddH_2O	Double distilled water
DCIS	Ductal Carcinoma *In Situ*
DEPC	Diethylpyrocarbonate
DFA	Discriminant Function Analysis
DMEM	Dulbecco’s Modified Eagle’s Medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
DSB	Double strand break
dNTP	Deoxyribonucleotide triphosphate
ds	Double strand
dsDNA	double-stranded DNA
EGF	Epidermal Growth Factor
ERBB2	v-erb-b2 Erythroblastic Leukemia Viral Oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
ER	Estrogen Receptor
EtBr	Ethidium Bromide
FBS	Fetal Bovine Serum
HER2	ERBB2
IDC	Infiltrating Ductal Carcinoma
ILC	Infiltrating Lobular Carcinoma

LCIS	Lobular Carcinoma *In Situ*
µg	Microgram
mg	Miligram
min	Minute
µl	Microliter
ml	Mililiter
µm	Micrometer
µM	Micromolar
mM	Milimolar
mRNA	Messenger RNA
NHEJ	Nonhomologous end joining
NOS	Not otherwise specified
Oligo(dT)	Oligodeoxythymidylic acid
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
pmol	Picomole
PR	Progesterone Receptor
PTEN	Phosphatase and Tensin homolog
qRT-PCR	Quantitative real time RT-PCR
Rpm	Revolutions Per Minute
RT PCR	Reverse Transcription PCR
Sec	Second
TAE	Tris-Acetate-EDTA buffer
TBP	TATA Box Binding Protein
TDLU	Terminal Duct Lobular Unit
Tm	Melting Temperature
Tris	Tris (Hydroxymethyl)- Methylamine
UV	Ultraviolet
VEGF	Vascular Endothelial Growth Factor
v/v	volume/volume
w/v	weight/volume

CHAPTER 1. INTRODUCTION

1.1 Breast cancer

Breast tumors have been noted since antiquity and were probably first described in the Edwin Smith surgical papyrus originating from Egypt at around 2500 B.C. In this document tumors were described to be "cold and hard to the touch" whereas abscesses were "hot" (cross ref. from Oldenburg *et al*., 2007). Today breast cancer is the most common cancer among women accounting for 22% of all female cancers. A woman's breast is made up of glands that make breast milk (lobules) and ducts (small tubes that connect lobules to the nipple) and breast cancer begins in these breast tissues. The remainder of the breast is made up of fatty and connective tissue, blood vessels, and lymph vessels (Figure1.1).

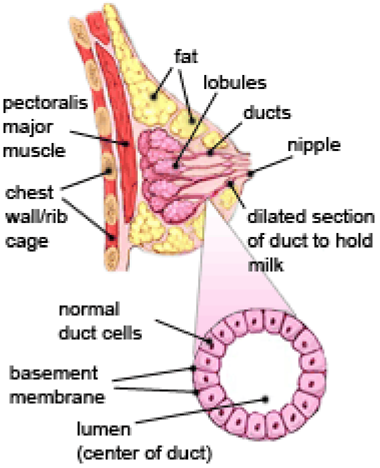


Figure 1.1 Anatomy of the breast. The female breast is formed by ducts, lobules, fatty and connective tissues (http://www.cancer.org).

Breast cancer is one of the major cancer types women suffer from in the United States and Western Europe. After lung cancer, it is the second leading cause of cancer death in women. Nearly 212,920 women in the United States were found to have invasive breast cancer in 2006 and in 2008, nearly 182,460 new cases of invasive breast cancer are expected to be diagnosed among women. The estimated number of death is about 40,460 in 2007 due to breast cancer. The chance of a woman having invasive breast cancer some time during her life is about 1 in 8. The chance of dying from breast cancer is about 1 in 33. Due to earlier diagnosis and improved treatment of breast cancer the death rates because of this disease are going down.

Many risk factors have been identified to date that contributes to the formation of breast cancer. Ethnicity, gender and the age are among the factors that were found to correlate with the incidence rates. Accordingly compared to female breast cancer incidence rate, that of male is much less and the rate of incidence increases up to 10-fold with increasing age (Medina, 2005; http://www.cancer.org). The extent and duration of exposure to sex hormones has also been consistently identified as a risk factor. Early age at menarche, delayed menopause, usage of exogenous hormones and late age of first pregnancy are expected to increase the risk of getting breast cancer (Medina, 2005, Oldenburg *et al*., 2007).

It is thought that breast cancer is a heterogeneous group of diseases with each subtype having its own stable phenotype maintained during tumor progression rather than a single disease with a single tumorigenesis pathway (Mallon *et al.,* 2000; Polyak, 2006). The most important determinants of these subtypes found are estrogen receptor (ER) and progesterone receptor (PR) status of tumor cells and the amplification and overexpression of the HER2 oncogene. Considering these features, breast tumors are divided into pathological and molecular subtypes.

1.1.1 Pathological subtypes of breast cancer

Most breast lumps, areas of thickening, are benign; that is, they are not cancer. Benign breast tumors are abnormal growths, but they do not spread outside of the breast and they are not life-threatening. Some benign breast lumps can increase a woman's risk of getting breast cancer. Lumps are formed by fibrocystic changes in most cases. These changes include stromal fibrosis, cyst formation, and adenosis. Adenomas are also common benign lesions characterized by well-circumscribed benign epithelial elements with a variable amount of stroma.

Epithelial hyperplasia may be one of the initiating steps of breast carcinoma. Atypical hyperplasia is an epithelial proliferation in which some features of ductal carcinoma are seen in epithelial tissues (Beckmann *et al.,* 1997; Mallon *et al.,* 2000).

1.1.1.1 Ductal Carcinoma in Situ (DCIS)

It is the most common type of noninvasive breast cancer. Nearly all women with cancer at this stage can be cured. DCIS is a morphologically identifiable, preinvasive malignant proliferation of the breast epithelial cells (Mallon *et al.,* 2000; http://www.cancer.org). The abnormal cells are contained within the mammary epithelial structures. No invasion of the basement membrane and no infiltration of the breast stroma are apparent. With a true *in situ* carcinoma, malignant epithelial cells do not have access to the lymphatic or vascular channels present within the breast stroma. Classifications are performed according to the degree of nuclear pleomorphism (often graded on a scale of 1–3), the presence or absence of necrosis, and the mitotic activity. The most characteristic feature of DCIS is that the cells composing the intraluminal proliferation are morphologically similar to each other, but have nuclear abnormalities associated with malignancy.

1.1.1.2 Lobular Carcinoma in Situ (LCIS)

LCIS is a neoplastic proliferation of epithelial cells in the terminal duct lobular unit with specific morphological features and therapeutic implications (Beckmann *et al.,* 1997; Mallon *et al.,* 2000). LCIS is a proliferation of neoplastic, epithelial cells which expand the individual acini of the lobular units involving more than 50% of the acini in a lobular unit. Both LCIS and DCIS are observed more in premenopausal women, suggesting that these lesions regress after menopause and that they are hormone dependent. This idea is supported by the ER positivity of these lesions.

1.1.1.3 Infiltrating Ductal Carcinoma (IDC)

IDC is the most common type of breast cancer. It accounts for about 80% of invasive breast cancers (http://www.cancer.org). If a tumor does not show the morphological features of a special type of invasive carcinoma or the characteristics of invasive lobular carcinoma it is called IDC if not otherwise specified (NOS) (Mallon *et al.,* 2000, Weigelt *et al*., 2008). This group of tumors is morphologically heterogeneous. IDC tumors have very variable growth patterns and stromal responses. They are often hard and fibrous.

The stage is determined by spread of the tumor to the body. However, grade is determined by how the tumor cells appear under the microscope, growth rate of the tumor cells, and the tendency of tumor to spread other parts of the body. There are four stages of breast cancer (http://www.cancer.gov). If the tumor size is less than 2 centimeters and there is no metastasis, it is a stage I tumor. As it progresses to stage IV, tumor size and metastasis levels increase. In stage IIIB and IV, the metastasis spreads to other parts of the body rather than lymph nodes. As the stage increases, the severity of the disease increases, as well. It is possible to separate IDC into three grades based on the degree of tubule formation, nuclear pleomorphism, and mitotic activity. Each of the three parameters is given a score of 1–3 and the individual scores are then added together. A score of 3–5 indicates Grade 1, 6–7 indicates Grade 2, and 8–9 indicates Grade 3. The first parameter, tubule formation, is assessed on the basis of percentage of the tumor showing distinct tubules: a score of 1 is assigned if 75% or more, a score of 2 if 10–75%, and a score of 3 if less than 10%. Nuclear pleomorphism is the second component. If the nuclei are small, with regular outlines, uniform chromatin, and little variation in size, they are assigned a score of 1. The cytoplasm of the tumor cells may also show considerable variation, with some cells having little cytoplasm and others having abundant cytoplasm that can be eosinophilic and granular, or foamy and basophilic, or midway between the two. The third parameter is an assessment of the proliferation rate determined by counting the number of mitoses in 10 high-power fields at the periphery of the tumor. The method is standardized for each microscope objective and the tumor scored on a scale of 1–3.

1.1.1.4 Infiltrating Lobular Carcinoma (ILC)

ILC is the second most common type of invasive breast carcinoma and makes about 10-15% of all breast tumors and it is histologically characterized by uniform tumor cells arranged in single-files or concentrically localized around ducts (Yolder *et al*., 2007; Mallon *et al.,* 2000; http://www.cancer.org). The cellular morphology of the tumor and the pattern of infiltration are very important in diagnosis. The tumor cells of lobular carcinoma are found in association with foci of typical LCIS and infiltrate in a very characteristic way with one cell behind the other in a defined pattern called the Indian filing pattern. They often form concentric rings around blood vessels and lobules producing a targetoid pattern. In classical lobular carcinoma, the tumor cells are relatively small. They have regular rounded nuclei with dense, evenly staining chromatin. Nucleoli are not prominent. A high grade aggressive form of ILC is known as pleomorphic lobular carcinoma (PLC) (Simpson *et al*., 2008). Bertucci *et al*. (Bertucci *et al*., 2008) described that IDC and ILC were histologically and genomically distinguishable from each other among the ER(+) grade II invasive breast tumors. Furthermore, ILC molecular subtypes were reported to include the typical and IDC-like ILCs, yet the *CDH1* mutation and/or underexpression was common but not universal to ILCs in general (Yolder *et al*., 2007). ILC tumors mostly metastasize to gastrointestinal, gynecologic and peritoneal tissues and particularly to endocrine related sites (Zhao *et al*, 2004) ILC has a higher incidence of multicentricity and bilaterality than IDC and a slightly better overall survival rate than tumors in the NOS category.

There are also several other less common types of breast cancer including tubular, mucinous, medullary, papillary, invasive cribriform, and secretory carcinoma.

1.1.2 Molecular classification of breast cancer

The classifications of the breast cancer are mostly based on clinical and pathological factors, which unfortunately fail to reflect the heterogeneity of the tumors. There are some histological markers available to decide on the prognosis and treatment of breast cancer. Estrogen receptor (ER) status, as ER-positive or ER-negative, helps to categorize breast cancers into two major classes. ERBB2 (Her-2/Neu) is also routinely used to classify breast cancer into HER-2 amplified or nonamplified categories. There are other single gene markers such as TP53, and cell proliferation markers such as Ki-67, and cyclin D1 that have emerged from detailed molecular analysis (Nielsen *et al*, 2004). While conventional methods were restricted to studying a single locus, current highthroughput techniques have allowed monitoring gene expression or copy number levels of almost all known genes in a single experiment. Molecular profiling has been shown to be well-suited to phenotypic characterization of breast cancer and potentially to discover new molecular classes among cancers with similar histopathological appearance (Sorlie *et al*, 2001; van't Veer *et al*, 2002; van de Vijver *et al*, 2002; Ahr *et al*, 2002; Sotiriou *et al*, 2003; Huang *et al*, 2003)

Several landmark microarray studies have demonstrated that one can build a molecular taxonomy of breast tumors using this technology and can provide a more sophisticated molecular picture together with individualized recurrence risks.

1.1.2.1 Distinguishing tumors on the basis of their gene expression profiles

Gene expression profiling using DNA microarrays has provided an opportunity to perform more detailed and individualized breast tumor characterization leading to classification of breast cancer into distinct new molecular subgroups (Cleator and Ashworth, 2004). The potential advantages of improving tumor classification by expression profiling has been central to several large-scale breast cancer studies over the past few years that have reported identification of signature gene lists with potential for prediction of clinical outcome (Sorlie *et al*, 2001; van't Veer *et al*, 2002; Huang *et al*, 2003; Gruvberger *et al*, 2001; West *et al*, 2001). One of the first comprehensive studies classifying sporadic breast tumors into subtypes distinguished by differences in their expression profiles was performed by Perou *et al.* (Perou *et al*, 2000). Using 40 tumors and 20 matched pairs of samples they identified an "intrinsic geneset' of 476 cDNAs and then used this to cluster and segregate the tumors into four major subgroups: a "luminal-like cells" group expressing estrogen receptor (ER); a "basal-like cells" group; an "ERBB2-positive" group, and a "normal like" epithelial group (Perou *et al.*, 2000). As new samples became available they were re-evaluated the data on 85 new tumors (Sorlie *et al.*, 2001). Not only were the breast cancer subtype definitions modified, clinical outcomes (metastasis, death and survival) were found to be significantly different between the subtypes. Gene sets corresponding to each of these groups were made on one dataset and used to classify the samples of another dataset. All groups were repeatedly found in the second dataset and differences in overall survival between the five groups were significant. Sorlie *et al.* (Sorlie *et al.*, 2003) rexamined 84 of the 85 arrays used in 2001 study and added 38 new breast cancer tumor tissue arrays to this study. Once again they found the same subtypes with significant differences in overall survival between the groups and confirmed their data in independent datasets. Consequently it is now recognized that all breast cancers are not the same in molecular point of view. Subsequent studies confirmed that there are large-scale gene expression differences between ER-positive (mostly luminal-like) and ER-negative (mostly basal-like) cancers and suggested that further molecular subsets also exist (Sorlie *et al*, 2003; Pusztai *et al*, 2003; Sotiriou *et al*, 2003).

The prognosis and chemotherapy sensitivity of the different subgroups are different. The luminal type cancers tend to have the most favorable long-term survival, whereas basal-like and ERBB2-positive tumors are more sensitive to chemotherapy (Sorlie *et al*, 2001; Rouzier *et al*, 2005).

In another molecular classification of breast tumor study, Van't Veer *et al.* have used DNA microarray analysis on the primary breast tumors of 78 lymph node-negative young patients and compared the expression profiles of 34 patients who developed distant metastasis within 5 years and 44 patients who remained diseasefree for at least 5 years (van't Veer *et al*, 2002). Their analysis led to the identification of a 70-gene expression signature that was developed to classify tumors into the good and poor prognosis groups. The results were later confirmed in a larger set of tumors (van de Vijver *et al*, 2002; Buyse *et al*, 2006)

1.1.2.2 Basal-like tumors

Basal-like breast carcinomas are so named because in terms of gene expression, these tumors are generally characterized by high expression of some basal epithelial markers such as KRT5, KRT6, KRT17, KRT23, c-KIT, FOXC1, P-Cadherin and LAMC2 (laminin). In particular, overexpression of LBR, DSC2, MRAS, CDCA7, FABP7, CXCL1, TRIM29, MSN, CCNE1, CCNA2, CCNB1, MYBL2, CDH3, CRYAB, MKI67, MET, AURKB, LYN, FOXM1 have been often observed (Sorlie *et al*., 2001; Nielsen *et al*., 2004).

Basal-like breast carcinomas, as defined by gene expression microarray analysis, are the most undifferentiated breast cancers, frequently lack the expression of hormone receptors (ER) and HER2 (Perou *et al*., 2000; Sorlie *et al*., 2001; Nielsen *et al*., 2004), show p53 immunohistochemical expression and TP53 gene mutations (Sorlie *et al*., 2001). Morphologically, basal-like breast carcinomas are characterized by high histological grade, high mitotic indices and the presence of central necrotic zones (Turner and Reis-Filho, 2006). Finally patients with basal-like tumors experience a much shorter overall-and disease-free survival period. It is the most severe case among the other subtypes of breast cancer.

1.1.2.3 Luminal tumors: Luminal A and Luminal B

Luminal breast tumors were firstly classified according to their hormone receptor status and fell into the group of ER positive tumors. On the other hand none of the tumors in this group found to be expressed Erb-B2 at high levels (Perou *et al*., 2000). They are given the name "luminal tumors" since both luminal A and luminal B subtypes are positive tumors for the expression of luminal cell markers. They show a discriminative expression of certain proteins such as TOPO II, proliferating cell nuclear antigen (PCNA) and cell cycle proteins and they have differential clinical outcomes (Melchor and Benitez, 2008).

1.1.2.4 ErbB2 expressing tumors

ErbB2 tumors show an overexpression of ErbB2 and multiple genes from the 17q11 amplicon. Overexpression of the ErbB2 oncogene was associated with the high expression of a specific subset of genes and these tumors were partially characterized by the high level of expression of this subset of genes. ErbB2 tumors also showed low levels of expression of ER and of almost all of the other genes associated with ER expression, which is a trait they share with the basal-like tumors. (Perou *et al*., 2000)

1.1.2.5 Normal-like tumors

Normal like tumors were the ones clustered with a group of samples that also contained the normal breast specimens (Perou *et al*., 2000). The “normal-like tumor” gene expression pattern was classified by the high expression of genes characteristic of basal epithelial cells and adipose cells, and the low expression of genes characteristic of luminal epithelial cells.

1.1.3 Genes implicated in breast cancer

The existence of a strong predisposition to breast cancer is a well known phenomenon. To date, up to 5-10% of all breast cancers are caused by germline mutations in well-identified breast cancer susceptibility genes. These genes have been divided into high-risk and low-to moderate risk susceptibility genes. The high risk breast cancer susceptibility genes include BRCA1, BRCA2, PTEN, TP53, LKB1/STK11 and CDH1 while CHEK2, TGFβ1, CASP8 and ATM genes belong to the low-to moderate risk susceptibility genes. In one third of the hereditary breast cancers, the germline mutations of BRCA1 and BRCA2 are found to be responsible.

The other genes like p53, PTEN, CHEK2 and ATM account for a small proportion of hereditary breast cancers (Table 1.1) (Palacios *et al.*, 2008).

Table 1.1: List of known high- and moderate to low-risk breast cancer susceptibility genes.

Gene	**Location**	**Frequency**	**Breast cancer risk**
BRCA1	17q21	Rare[a]	46-85% lifetime risk
BRCA2	13q12	Rare[a]	43-84% lifetime risk
TP53	17p13.1	Rare	28-56% by age 45
PTEN	10q23.3	Rare	25-50% lifetime risk
LKB1/STK11	19p13.3	Rare	29-54% lifetime risk
CDH1	16q22.1	Rare	20-40% lifetime risk
ATM	11q22-23	Moderate	RR: 2.2
TGFβ1	19q13.1	Frequent	OR: 1.25 (p=0.009)
CASP8	2q33-34	Frequent	OR: 0.83
CASP10	2q33-34	Frequent	OR: 0.62 (p=0.0076)
CASP8/CASP10		Moderate	OR: 0.37 (p=0.013)
CHEK2	22q12.1	Moderate	RR: 2

Rare: <1% population frequency, moderate 1-5%, frequent >5%, OR: odds ratio and RR: relative risk
[a] In, for example the Ashkenazi Jewish population some mutations have a moderate population frequency. (Palacios *et al.*, 2008)

The TP53 gene encodes a protein involved in many overlapping cellular pathways that control cell proliferation and homeostasis, like cell cycle, apoptosis and DNA repair. The expression of TP53 is activated in response to stress signals including DNA damage (Oldenburg *et al.*, 2007). While the mutations in the p53 gene were found to be responsible for Li-Fraumeni syndrome in nearly 70% of families fulfilling the classical criteria for the disease, they were found to be less common in breast cancer (Frebourg *et al.*, 1995). Somatic mutations are reported in 20-60% of human breast cancers (de Jong *et al.*, 2002) and hypermethylation of the p53 gene seems not to play a major role in breast cancer.

The tumor suppressor gene PTEN is found to be responsible for Cowden syndrome (CS) and found to be mutated in sporadic brain, breast, and prostate cancers (Liaw *et al.*, 1997). Women carrying a PTEN mutation have a 25-50% lifetime risk of developing breast cancer. To date no mutations in the PTEN gene have been detected in breast cancer families without features of CS. Also in sporadic breast cancers germline and somatic mutations are rare. In addition, although LOH at the PTEN locus is found in 11-41% of sporadic breast cancers, no somatic mutations have been observed in the remaining allele (Freihoff *et al.*, 1999; Feilotter *et al.*, 1999).

One of the known low to moderate-risk breast cancer susceptibility gene ATM plays a central role in sensing and signaling the presence of DNA double-strand breaks. The irradiation initiates kinase activity of ATM and phosphorylates the protein products of TP53, CHEK2 and BRCA1 (Bakkenist *et al.*, 2003). Carriers of the ATM gene mutations suffer from a recessive disorder ataxia-telangiectasia (AT). The role for the ATM gene in breast cancer is plausible but the exact association remains unclear, and most probably comprises only a modest role in familial breast cancer susceptibility (Hall, 2005).

Among the well identified breast cancer susceptibility genes BRCA1 is the one to be responsible for 45% of the all hereditary breast cancers.

1.2 BRCA1

Tumor suppressor gene BRCA1 (Breast cancer susceptibility gene 1) plays a central role in the development of breast and ovarian cancers. While inherited mutations of BRCA1 are responsible for 40-45% of the hereditary breast cancers, 71% of the BRCA1 mutation carriers have the risk of developing breast cancer (Rosen *et al,* 2003). Somatic BRCA1 mutations are rare in sporadic breast cancers; however both mRNA and protein expression are downregulated in 30% of sporadic cases (James *et al*, 2007). It is reported that reduced expression or absence of BRCA1 protein in sporadic cases are due to non-mutational mechanisms such as hypermethylation of BRCA1 promoter or incorrect subcellular localization of the BRCA1 protein (Birgisdottir *et al*, 2006, Esteller *et al,* 2000, Rakha *et al*, 2008).

1.2.1 Functions of BRCA1

BRCA1 regulates multiple cellular events including cell cycle regulation and growth control, DNA damage response and repair processes and regulation of transcription (reviewed in Rosen *et al*, 2003, Venkitaraman, 2002). BRCA1 is a component of BASC, a BRCA1-associated genome surveillance complex that includes proteins involved in the recognition and repair of DNA-damage (Wang *et al*, 2000). In addition, the carboxyl terminal of BRCA1 acts as a strong transcriptional activator when fused to a heterologous DNA binding domain (Monterio *et al*., 1996). BRCA1 co-purifies with RNA polymerase II holoenzyme complex, suggesting that it is a component of core transcription machinery (Scully *et al*., 1997a). BRCA1 also interacts with several transcription factors such as p53, CtIP, c-myc, ZBRK1, ATF, E2F, and signal transducer STAT1 (Zheng *et al*., 2000) and modulates their activity. These findings, together with the interaction of BRCA1 with histone deacetylases (HDACs) and the SWI/SNF-related chromatin remodeling complex, imply that transcriptional regulation is one of the main functions of BRCA1 (Bochar *et al*, 2000). In addition, nearly all germ-line BRCA1 mutations involve truncation or loss of the C-terminal BRCT transcriptional activation domain, supporting the transcriptional regulation function of the BRCA1 gene.

1.2.1.1 BRCA1 in damage signaling and DNA repair

Initial evidence suggesting a role of BRCA1 in the repair of damaged DNA was derived from the observation that BRCA1 is hyperphosphorylated in response to DNA damage and relocated to sites of replication forks (Scully *et al*., 1997a; Thomas *et al*., 1997). Additionally BRCA1 has been identified as a target for several upstream nuclear phosphoinositide (PI) like kinases, which are implicated in DNA damage signaling through their protein kinase activities, ataxia-telangiectasia mutated (ATM) and ATM-related (ATR) kinase (Yan *et al*., 2008). The studies show that ATM mediates its phosphorylation role in response to ionizing radiation, while ATR mediates the phosphorylation in response to UV irradiation (Gatei *et al*., 2001). The major target for ATM phosphorylation after ionizing radiation is Ser1387 of

BRCA1. In response to ultraviolet irradiation, Ser1457 residue of BRCA1 is primarily phosphorylated, mainly by ATR (Gatei *et al*., 2001). The G2/M control kinase, CHK2, has also been shown to phosphorylate BRCA1 at Ser988 on exposure to ionizing radiation (Chaturvedi *et al*., 1999) (Figure 1.2).

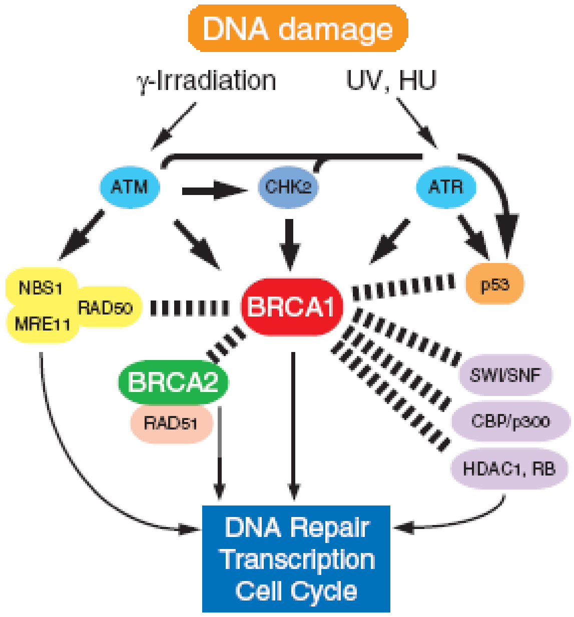


Figure 1.2 Functions of BRCA1 in response to DNA damage. Arrows show the phosphorylations while the dash lines show the interactions through proteins (Chaturvedi *et al*., 1999).

There are two mechanisms for double strand break (DSB) repair; a process known as homologous recombination (HR), which uses homologous sequences on the sister chromatid for repair and non-homologous end joining (NHEJ), which ligates either contiguous or non-contiguous sequences in the genome (Greenberg, 2008). BRCA1 was shown to be involved in the complexes that activate the repair of DSBs and initiate HR. BRCA1 and BRCA2 were found to be localized with Rad51, which is a protein required for recombination during mitosis and meiosis as well as HR repair of DSBs (Scully *et al*., 1997b; Shinohara *et al*., 1992). Co-localization of BRCA proteins with Rad51 shows that BRCA1 and BRCA2 takes role both in the detection and repair of DSBs. However the studies show that BRCA1 may not directly regulate Rad51 since they do not interact with each other.

In mammalian cells NHEJ proceeds in stepwise manner beginning with the end processing by the MRE11/RAD50/NBS1 (MRN) complex and then end joining by Ku70 and Ku80 proteins (Wu *et al*., 2000; Chan *et al*., 1999). It is not clear how BRCA1 promotes NHEJ but studies have shown that BRCA1 co-localizes with MRN complex and negatively regulates end-processing by MRE11 endo- and exonucleases (Zhong *et al*., 1999a). Suppression of MRN-mediated end-processing by BRCA1may enhances NHEJ accuracy (Shrivastav *et al*., 2008).

1.2.1.2 BRCA1 in regulation of transcription

A role for BRCA1 in transcriptional regulation was first suggested by the finding that BRCA1 has a conserved acidic COOH-terminal transcriptional activation domain, which is a globular domain, found in proteins involved in repair and cell-cycle control (Monteiro *et al.,* 1996). Results showing that the BRCA1 C terminus (aa 1560-1863) has the ability to activate transcription when fused to GAL4 DNA-binding domain (DBD) provided the initial experimental evidence of the involvement of BRCT domain of BRCA1 in transcription (Monteiro *et al*., 1996). Series of experiments have demonstrated that the C-terminus of BRCA1 can be used to recruit RNA polymerase II (RNAPII) to synthetic reporters showing that BRCA1 plays some role in transcriptional activation (Monteiro *et al*., 2000). However, direct evidence that BRCA1 binds to promoter regions of genes is lacking. Although BRCA1 is not known to bind to specific DNA sequences, it may regulate transcription through protein:protein interactions. BRCA1 physically associates with many proteins (transcription factors) involved in transcription and is paradoxically involved in both transcriptional activation and repression (Mullan *et al.,* 2006). ER-alpha, p53, STAT1, CtIP, c-Myc and ZBRK1 are the known transcription factors that have interaction with BRCA1 (Mullan *et al*., 2006) (Figure 1.3).

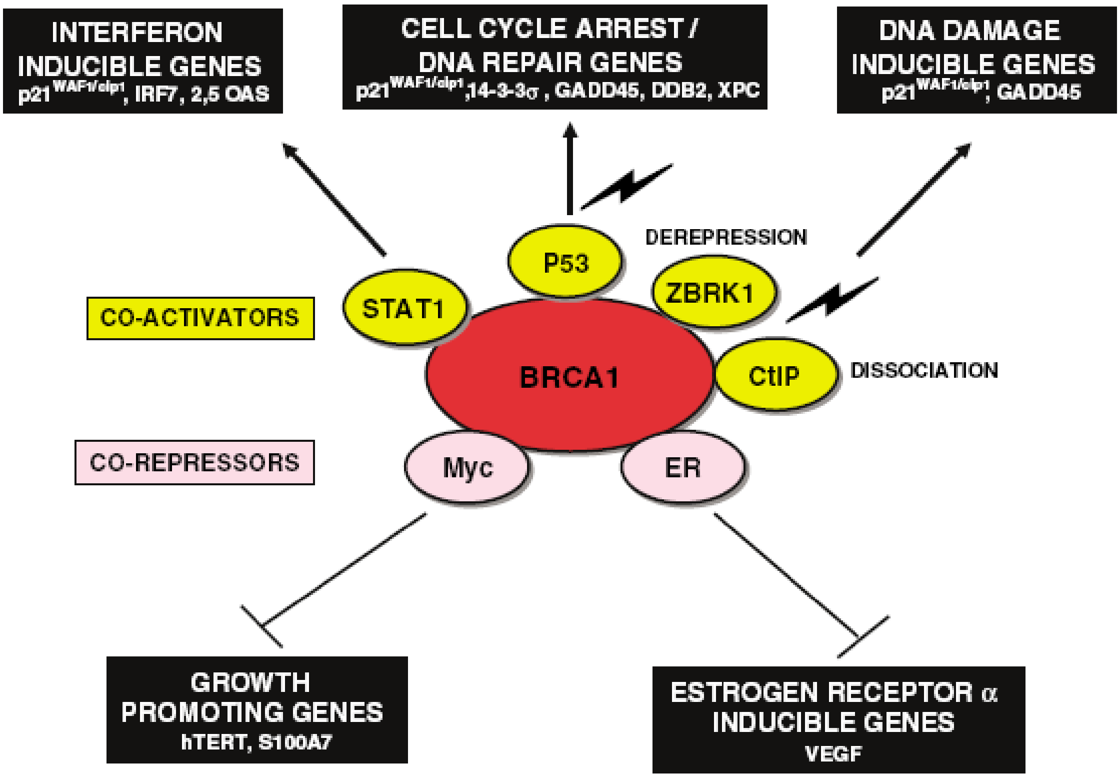

INTERFERON INDUCIBLE GENES
p21^WAF1/cip1, IRF7, 2,5 OAS
CELL CYCLE ARREST / DNA REPAIR GENES
p21^WAF1/cip1,14-3-3σ , GADD45, DDB2, XPC
DNA DAMAGE INDUCIBLE GENES
p21^WAF1/cip1, GADD45
DEREPRESSION
P53
ZBRK1
STAT1
CO-ACTIVATORS
BRCA1
CtIP
DISSOCIATION
CO-REPRESSORS
Myc
ER
GROWTH PROMOTING GENES
hTERT, S100A7
ESTROGEN RECEPTOR α INDUCIBLE GENES
VEGF

Figure 1.3 BRCA1 interacting transcription factors and their roles in the cell. (Mullan *et al.*, 2006)

1.2.1.3 BRCA1 modulation of sequence-specific DNA binding transcription factors

Still no specific DNA binding sequence of BRCA1 was stated but it has been established that it can bind to various sequence specific DNA binding transcription factors to stimulate or inhibit transcription. BRCA1 interacts with tumor suppressor protein p53 and act by both stabilizing and stimulating its transcriptional activity (Chai *et al.*, 1999; Zhang *et al.*, 1998). This stabilization appears to induce a subset of p53-regulated genes involved in DNA repair and cell cycle arrest other than apoptosis which may show that BRCA1:p53 interaction may influence the cell fate decision during the DNA damage (Ongusaha *et al.*, 2003; MacLachlan *et al.*, 2002).

It was found that a physical interaction between BRCA1 and STAT1, which is a transcription factor that transduces the cellular response to interferon-γ (INF- γ), induced a subset of INF- γ responsive genes (Ouchi *et al*., 2000). In addition to its role in inducing gene expression with INF- γ, BRCA1 also potentiated INF- γ mediated apoptosis (Andrews *et al*., 2002).

BRCA1 can also regulate the promoter activity and expression of growth inhibitory genes like p21, Gadd45α and p27. BRCA1 was found to physically interact with sequence specific transcription factors Oct-1 and NF-YA, which directly bind to the OCT-1 and CAAT motifs on GADD45 promoter thus induce the expression of Gadd45α (Fan *et al.,* 2002). On the other hand Zheng *et al.* showed that BRCA1 interacts with a zing finger and KRAB domain protein ZBRK1 to bind a specific DNA sequence on the 3rd intron of GADD45 and in this context ZBRK1 appeared to repress GADD45 transcription in a BRCA1 dependent manner (Zheng *et al,* 2000). Addition to this direct binding of BRCA1 to some transcription factors Cable *et al.* published a specific DNA sequence that can be bound by BRCA1 protein complexes (BRCA1:USF2) to control gene expression (Cable *et al,* 2003). These known sequences have paramount importance to find out the new downstream targets of BRCA1 to explain or to clarify the exact role of BRCA1 in transcriptional regulation.

1.2.2 BRCA1 regulated targets

BRCA1 is known to be a multifunctional protein and it locates at different sites and takes role in many functional events in the cell. Although many of these functions are clear and the mechanism of its action is known some of the functions of BRCA1 and how it controls these functions is still a gap. Hence it is important to find out its exact role in these cellular events to clarify its role in breast cancer development, transcriptional regulation and in other cellular events. Finding out transcriptional targets of BRCA1 is one of the strategies at least to identify the subgroups of genes from different molecular and cellular functions.

In order to identify a series of downstream targets of BRCA1, Harkin *et al.* established an osteosarcoma cell line with tightly regulated BRCA1 expression with the tetracycline inducible system (Harkin *et al*., 1999). High density oligonucleotide arrays were used to analyze the gene expression profiles at various times following BRCA1 induction. They found 20 BRCA1 target genes. Among those genes, they identified GADD45 as one of the major targets of BRCA1 and showed that this activation was p53 independent. Concordant with the previous study Mullan *et al.* used the same tetracycline inducible system with a breast cancer cell line and found GADD45 as the main target of BRCA1 (Mullan *et al*., 2001). Later on MacLachlan *et al*. have overexpressed BRCA1 by using an adenovirus vector and have identified 45 major targets in the SW480 colorectal cancer cell line. Most of the genes they found were DNA damage response genes and the ones involved in cell cycle control (MacLachlan *et al*., 2000). Furthermore Welch *et al*. identified 62 genes that are targets of BRCA1 by using ecdysone inducible expression of BRCA1 in human embryonal kidney epithelial cells (Welch *et al*., 2002). Finally Atalay *et al*. used suppression subtractive hybridization (SSH) technology to generate a library of partial-length cDNAs representing mRNAs in BRCA1-overexpressing MCF7 cells. By this approach 60 genes have been identified that are upregulated as a result of BRCA1 overexpression in breast cancer cells (Atalay *et al*., 2002).

1.3 Genes affected by BRCA1 expression

ERBB2 interacting protein, ERBIN

ERBIN was initially found to interact specifically with ErbB2 by its PDZ domain and acts in the localization of ErbB2 to the basolateral domain in epithelia which is important for its activation and signaling of ERBB2/HER2 in epithelia (Borg *et al*., 2000). It has 16 LLRs and a single PDZ domain in its C-terminus which interacts directly with the C-terminal aminoacids of unphosphorylated ERBB2. It is constitutively associated with ErbB2 in living cells (Borg *et al*., 2000, Huang *et al*., 2001;). They reported that the Erbin PDZ domain binds preferentially to the C terminus of ErbB2, which is non-Tyr1248-phosphorylated (Borg *et al*., 2000). Importantly phosphorylation of this residue following ErbB2 activation is a critical event for the mitogenic signaling and oncogeneity of this receptor (Dittmar *et al*., 2002). Overexpression of ErbB2 correlates with poor prognosis and resistant chemotherapy in breast and ovarian cancers (Klapper *et al*., 2000). Despite the close relation of ERBIN and ErbB2 the functional role of ERBIN has not been studied extensively in breast cancer yet. Recently Liu *et al*. studied the expression and the regulation of ERBIN and its binding partner ErbB2 in the MCF7 breast cancer cell line. One of their finding was that the affinity of Erbin-ErbB2 interaction was reduced by ErbB2 posphorylation (Liu *et al*., 2008).

Furthermore Erbin was found to be a novel suppressor of the Ras signaling (Huang *et al*., 2003). It can inhibit the activation of Ras pathway by disrupting the interaction of Sur-8, which is a positive regulator of the Ras pathway, with Ras and Raf (Dai *et al*., 2006). The requirement of LRRs for this process rather than PDZ domain shows that ERBIN has dual function (functions as a signaling molecule in addition to being a scaffold protein) in cells with its LRR and PDZ domains (Huang *et. al*., 2003, Dai *et al*., 2006). On the other hand a new role of ERBIN was described in inflammatory responses by McDonald *et al*. They found the inhibitory effect of ERBIN by its carboxyl terminus on Nod2-dependent activation of NF-κB and cytokine secretion (McDonald *et al*., 2005). In addition a very recent novel negative regulatory role was added to the functions of ERBIN and expanded the physiological role of it to the regulation of TGFβ signaling through its direct interaction with Smad2/Smad3 (Dai *et al*., 2007).

Two tumor suppressor genes, OVCA1 and OVCA2

OVCA1 and OVCA2 are the two tumor suppressor genes mapped to chromosome 17p13.3, that is most commonly lost in ovarian and breast tumors. The two OVCA genes are expressed from the same genetic locus using two different promoters; however since OVCA2 transcript contains a unique exon and only the 3’ UTR of the OVCA1 transcript, they share no coding sequence (Chen and Behringer, 2004). Northern blot analysis reveled that they are both expressed in normal surface epithelial cells of the ovary but reduced or undetectable in ovarian tumors and tumor cell lines (Schultz *et al*., 1996). In an independent study OVCA1 was shown to be reduced in protein level in breast and ovarian tumors and shown to inhibit growth of ovarian cancer cells (Bruening *et al*., 1999). To address the role of OVCA1 and

OVCA2 in development and proliferation Chen and Behringer (2004) generated OVCA1 disrupted and OVCA1 and OVCA2 disrupted mouse models. The identical results obtained from the two strains showed that OVCA1 was more important than OVCA2 with respect to development and proliferation. They finally concluded that OVCA1 acted as a positive regulator for cell cycle progression and it was a tumor suppressor (Jensen and Helin, 2004; Chen and Behringer, 2004).

Diphthamide is a unique post-translationally modified histidine residue found only on translational elongation factor 2 (EF-2). The biosynthesis of the diphthamide is one of the most complex post-translational modifications. Addition to its tumor suppressor activity Nobokini *et al*. (2005) and Chen *et al*. (2005) showed that OVCA1 was a component of the biosynthetic pathway of diphthamide on EF-2.

Nonsense mediated decay genes, RENT2 and SMG1

Nonsense-mediated mRNA decay (NMD) is a quality control mechanism that selectively degrades mRNAs harboring premature termination (nonsense) codons. The core NMD machinery comprises three trans-acting factors, called up-frameshift (UPF) proteins, which were initially discovered in *S. cerevisiae* and later identified in higher eukaryotes. UPF1, UPF2 also known as RENT2 and UPF3 proteins comprise the core NMD machinery. The SMG-1, SMG-5, SMG-6 and SMG-7 proteins mediate the phosphorylation and dephosphorylation cycle of UPF1 (Chang *et al.*, 2007).

RENT2 (UPF2) is an adapter molecule that brings together UPF1 and UPF3 to elicit NMD. It is a part of large complex of proteins that is deposited on mRNAs at exon-exon junctions during RNA siplicing in the nucleus. With UPF3, RENT2 is a part of the exon-junction complex (EJC), a large dynamic protein complex deposited just upstream of exon-exon junctions during RNA splicing. Many EJC components, including RENT2 and UPF3, remain bound to the mRNA after its export to the cytoplasm, where they function as a second signal to elicit NMD when the mRNA is proofread during translation (Chang *et al.*, 2007).

SMG-1 is a kinase that phosphorylates serine residues. It is a member of the phosphoinositide 3-kinase related kinase (PIKK) family, whose other members function in DNA damage and growth responses. SMG-1 mediated phosphorylation of UPF-1 is likely to be crucial for NMD. SMG-1 forms the SMG-1-Upf1-eRF1-eRF3 complex (SURF) prior to phosphorylation of Upf1. The SURF recognizes downstream Upf2-EJC and they associate to induce Upf-1 phosphorylation. SMG-1 also phosphorylates the p53 protein upon genotoxic stress (Yamashita *et al.*, 2005, Chang *et al.*, 2007). Furthermore, siRNA-mediated depletion of SMG-1 results in accumulation of spontaneous DNA damage and increases cellular sensitivity to ionizing radiation (Brumbaugh *et al.*, 2004). Recently it was found that loss of SMG-1 function dramatically increased the rate and extent of apoptotic cell death induced by tumor necrosis factor-alpha (TNFα). Thus it protects human cells from TNFα induced apoptosis through a mechanism unrelated to its role in NMD (Oliveira *et al.*, 2008).

5' to 3' exonuclease, XRN2

The XRN2 gene is the human homologue of the Saccharomyces cerevisiae RAT1 gene, which encodes a nuclear 5' to 3' exoribonuclease, and is essential for RNA metabolism and cell viability. Xrn2 /Rat1, product of XRN2/RAT1 gene, functions in the mRNA degradation and processing of rRNAs and small nucleolar RNAs (snoRNAs) in the nucleus (Li *et al.*, 2005).

Polymerase II (Pol II) transcriptional termination depends on two independent genetic elements: poly (A) signals and downstream terminator sequences. The latter may either promote cotranscriptional RNA cleavage or pause the elongating of Pol II. It was found that the previously characterized MAZ(4) pause element promotes Pol II termination downstream of a poly(A) signal, dependent on both the proximity of the pause site and poly(A) signal and the strength of the poly(A) signal (Gromak *et al.*, 2006). The 5' to 3' exonuclease Xrn2 facilitates this pause-dependent termination by degrading the 3' product of poly (A) site cleavage. The human beta-actin gene also possesses poly(A) site proximal pause sequences. Xrn2 depletion causes an increase in both steady-state RNA and Pol II levels downstream of the beta-actin poly (A) site. All these data provided new insights into the mechanism of pause site-mediated termination and establish a general role for the 5' to 3' exonuclease Xrn2 in Pol II termination (West *et al.*, 2004; Gromak *et al.*, 2006).

The meningioma associated protein, MAC30

The meningioma associated protein, MAC30 gene is located on 17q11.2, and the protein has a small segment which is similar to an apical gut membrane polyprotein of *Haemonchus contortus*, to olfactory receptor 30 of *Mus musculus*, and to cytochrome b in several organisms. MAC30 mRNA has been found to express as a non-erythropoietic gene in the fatal liver during the early stages of the development but not in the adult liver. In the light of this evidence it was suggested that MAC30 may play a role in growth and differentiation of the liver. MAC30 expression is seen in many types of normal organs including brain, lung, heart, skeletal muscles, testis and ovary but it is changed during tumor development. The decrease in its expression was shown in pancreatic and renal cancers while in meningiomas, ovarian, gastric and colarectal cancers its expression was shown to be increased (Murphy *et al.*, 1993; Moparthi *et al.*, 2007).

In a recent study performed with rectal cancers, MAC30 expression in radiated-primary tumors was related to more aggressive morphological and biological factors that were involved in cell proliferation, invasion and metastasis of the tumors. Stronger MAC30 expression was shown to be an indicator of poor prognosis (Zhang *et al.*, 2007). In another study dealing with the colorectal cancers, MAC30 expression was found to be much stronger in cytoplasm in lymph node metastasis compared to primary tumor and normal mucosa. Addition to overexpression of MAC30 at the invasion margins they suggested that the protein may play an important role in the development and aggressiveness of colorectal cancer (Moparthi *et al.*, 2007).

RAD21

The protein encoded by *RAD21* is highly similar to the gene product of Schizosaccharomyces pombe rad21, a gene involved in the repair of DNA double-strand breaks, as well as in chromatid cohesion during mitosis. This protein is a nuclear phospho-protein, which becomes hyperphosphorylated in cell cycle M phase. The highly regulated association of this protein with mitotic chromatin specifically at the centromere region suggests its role in sister chromatid cohesion in mitotic cells. Sister chromatid cohesion during DNA replication plays a pivotal role in accurate chromosomal segregation in the eukaryotic cell cycle. RAD21 is one of the major cohesin subunits that keeps sister chromatids together until anaphase when proteolytic cleavage by separase allows the chromosomes to separate (Hoque and Isikhawa, 2001). Addition to its roles in chromatid cohesion and repair of DNA double strand breaks, RAD21 was shown to be specifically proteolyzed by caspases into a similarly sized 65-kDa carboxyl-terminal product in cells undergoing apoptosis in response to diverse stimuli. It was also demonstrated that caspase proteolysis of RAD21 precedes apoptotic chromatin condensation and has important functional consequences showing another function of RAD21 in the execution of apoptosis (Chen *et al.*, 2002).

1.4 Combined analysis of microarray data sets: meta-analysis

The extensive use of DNA microarray technology in the characterization of the cell transcriptome is leading to an ever-increasing amount of microarray data from cancer studies.

Different data sets for the same type of cancers are available from different microarray studies and this allows the researchers to carry out a more comprehensive analysis of their existing data set. These studies can be obtained from various public gene expression data repositories including the Stanford Microarray Database (SMD) (Sherlock *et al.*, 2001), the National Cancer Institute's Gene Expression Omnibus (GEO) (Barrett *et al.,* 2005) and Oncomine (Rhodes *et al.*, 2004a). These databases enable researchers to retrieve and perform analyses on various microarray experiments from different laboratories.

Besides individual microarrays, meta-analysis can be used to gather and process the data sets from multiple cancer types to investigate common molecular pathways (Rhodes *et al*., 2004; Choi *et al*, 2007, Xu *et al*. 2007). Meta-analysis of microarray datasets has the potential to lead to more comprehensive measures of the existing differential gene expression data and can therefore provide gene sets with a high diagnostic value.

Several different meta-analysis approaches exist in the literature. In some, each individual study contributes rather independently to the meta-analysis (Moreau *et al*., 2003; Rhodes *et al*., 2002; Choi *et al*., 2003) whereas in others the values are treated as members of a single study thus requiring a generalized normalization step (Toedling and Spang , 2003; Warnat *et al*., 2005). Direct comparison of gene expression values from multiple studies may be relatively more problematic than comparing the effect size obtained from individual studies. Yet, analysis of combined raw data is beneficial when sample sizes of individual studies are small. Another important concern in meta-analysis is determination of the minimum number of samples required to obtain statistically reliable results (Qui *et al*., 2006). One possible solution to this problem is resampling; for example, one can use a *delete-d-jacknife* procedure in which a subset of data is excluded to find out the frequency of selecting a particular gene as differentially expressed (Qui *et al*., 2006). The number of replicates required for producing stable differentially expressed gene lists could also be determined based on a related method known as *leave-one-out* resampling (Pavlidis *et al*., 2005).

Since all cancer cells share some common characteristics such as; loss of growth control, invasion, and metastasis, it is of high importance to identify universal cancer type-independent signatures to better understand cancer pathogenesis and ultimately to improve thereapeutic options. Rhodes *et al*. applied meta-analysis approach to 21 published cancer microarray datasets, spanning 12 distinct cancer types, and identified a set of 67 genes that are universally activated relative to corresponding normal tissues in most cancer types relative to corresponding normal tissues (Rhodes *et al.*, 2004). Meta-analysis of independent microarray datasets generated with the common objective of identifying differentially expressed genes in a certain type of cancer has also been performed for breast cancer. In a very recent meta-analysis study, Smith *et al.* identified differentially expressed genes between ER+ and ER- breast tumors by gathering 9 independent breast cancer microarray studies (Smith *et al*., 2008). Another study used the power of meta-analysis to find out the relation of expression patterns of gene and chromosomal positions. More than 1200 breast tumors were collected from eight independent breast studies and candidate metastasis suppressor and promoting genes were found from a given set of chromosomal regions (Thomassen *et al*., 2008). Similarly, Hu *et al*. were able to identify a new intrinsic gene-set for breast cancer subtype prediction by combining multiple microarray datasets to assess prognosis (Hu *et al*., 2006).

These types of studies have resulted in the identification of gene sets with a high diagnostic value. Schneider and co-workers defined a set of genes that can be used as a diagnostic tool for accurate determination of ER status and to make a decision regarding the therapeutic strategies for breast cancer. (Schneider *et al*., 2006).

The power of meta-analysis is, the approach can provide novel candidates not present in the existing literature allowing reports of multiple genes when neither dataset can report them when analyzed individually (Choi *et al.*, 2004; Grutzmann *et al.*, 2005).

1.5 Measurements of gene expression with quantitative real time RT-PCR

Many applications in medicine or research require detection of the number of specific targets in the specimen (Mocellin *et al.,* 2003). Northern blot analysis, RNase protection assays, and polymerase chain reaction (PCR)-based methods are applied to detect and quantify mRNA amount in the desired specimens. The most sensitive one of these methods is PCR based ones with its combination with reverse transcription (RT) (Fronhoffs *et al.,* 2002). Since a minute amount of mRNA is enough for a reaction to occur it is more advantageous to perform RT-PCR rather than other methods.

PCR is the method to detect as little as a single copy of a particular sequence of DNA and RNA. In theory there is a quantitative relationship between the amount of starting sequence and amount of PCR product at a given cycle. However this is not the case in practice since the PCR product increases exponentially (2^n) in every cycle. Reverse transcription PCR (RT-PCR) is a kind of PCR used to detect expression level of a gene at mRNA level (Pfaffl, 2001). The expressional differences in the genes between tissues, disease states, and treatments can be revealed by using quantitative RT-PCR (Pfaffl, 2001; Bustin *et al.,* 2004a). Quantitative real-time RT-PCR (qRT-PCR) is one of the methods developed for quantitative measurement of the gene expression levels (Pfaffl, 2001; Mocellin *et al.,* 2003). It has many advantages over the traditional RT-PCR (Fronhoffs *et al.,* 2002; Wong *et al.,* 2005). It plays an increasingly important role in high-throughput testing of existing microarray data (Bernard and Wittwer, 2002). qRT-PCR is an accurate and sensitive method quantifying mRNA transcripts and it uses the quantitative relationship between the amount of starting target sample and the amount of PCR product at any given PCR cycle number. It has a wide dynamic range of quantitation and allows high throughput screening at one time. The method allows the detection of amplicon accumulation since it is performed using specific detection chemistries (Livak *et al*, 1995; Heid *et al*, 1996; Tyagi and Kramer, 1996; Whitcombe *et al*, 1999) or more sensitive but less specific intercalating dyes like SYBR Green I. qRT-PCR has the advantages of requiring smaller quantities of sample and producing fast, accurate and easily reproducible quantitative results with little manipulation of the samples (Bustin SA, 2002).

In a PCR reaction there are four major phases: the linear ground phase, early exponential phase, log-linear (exponential) phase, and plateau phase (Wong *et al.,* 2005). In the linear ground phase PCR just begins and fluorescence does not exceed background. Calculation of baseline fluorescence is performed at that phase (Figure 1.4). In early exponential phase fluorescence exceeds the background and threshold cycle (Ct) value begins to be detectable. All the measurement analysis takes place at this stage of the qRT-PCR reaction. The Ct value is used as measurement unit in the real-time RT-PCR. It is the cycle number at which fluorescence reaches a threshold value of ten times the standard deviation of baseline fluorescence emission (Mocellin *et al.,* 2003; Wong *et al.,* 2005). The Ct value is inversely proportional to the amount of starting material. The lower the Ct value the higher the expression of the target gene or the amount of starting material. The threshold value is the point at which a reaction reaches a fluorescent intensity above background. It should be in the linear part of the reaction. The program automatically determines the Ct value of the sample. In the third phase, the PCR reaction reaches its optimal amplification period in which PCR doubling in each cycle is ideal. In the plateau phase reaction components become rate limiting agents. Compared to the traditional PCR reaction in which the amount of PCR product could be detected only after a fixed number of cycles, qRT-PCR assay determines the number of cycles after which amplification of a PCR product is first detected. In a traditional PCR reaction since all the products reach to saturation level it is impossible to quantitate the expression levels at this phase. qRT-PCR overcomes this limitation of traditional PCR by detecting the expression levels at each cycle and monitoring the accumulation of amplicon.

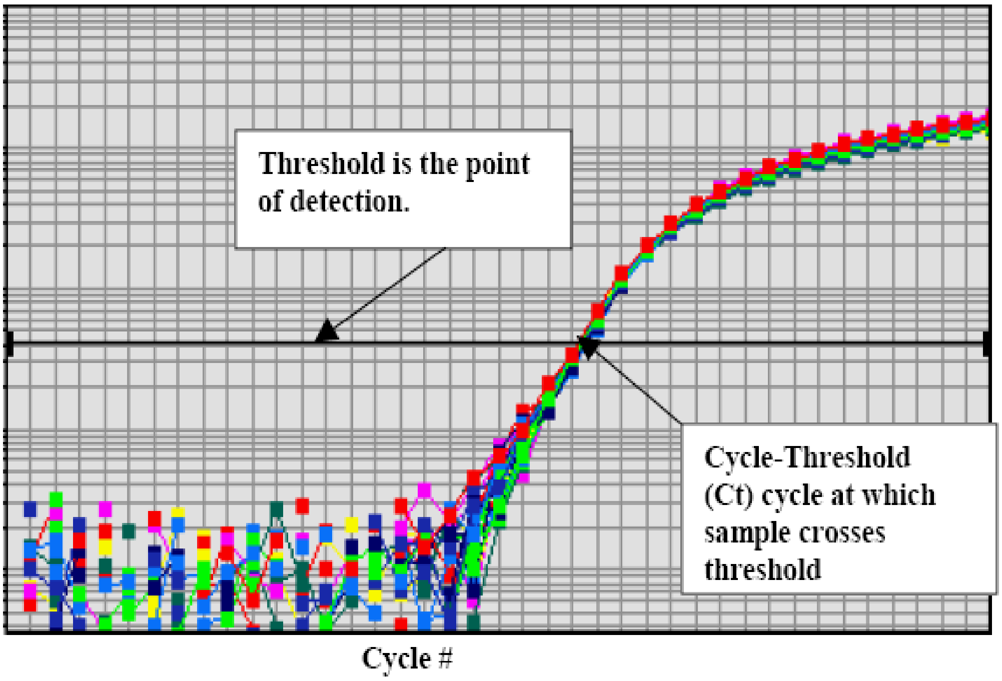


Figure 1.4 Amplification Curve: When fluorescent signal reaches to detectable level it is displayed as an amplification curve. The point at which the amplification curve reaches to the threshold level is called the Ct value (http://www.appliedbiosystems.com)

There are several considerations when performing qRT-PCR. The choice of proper detection chemistry for the design of experiment is the first point to be considered for an accurate performance of the reaction. Inclusion of an endogenous reference gene or genes (RGs) is crucial to standardize initial RNA quantity to overcome bias originating from RNA measurement errors, problems with RNA integrity, and differential cDNA conversion efficiencies (Bustin *et al*., 2005; Stahlberg *et al*., 2004a; Stahlberg *et al*., 2004b). Designing primers for qRT-PCR is another step to be considered to get more accurate results from a qRT-PCR reaction. The optimal PCR product length is approximately 100 to 200 bp for the primers while SYBR Green is used as a detection agent. The forward and the reverse primer should have similar melting temperatures within 0.5^{o}C of each other. The primers should have low or no self complementarities in order to avoid the formation of primer dimers. Primers that span introns or cross intron/exon boundaries are advantageous as they allow the distinction of cDNA from genomic DNA contaminations.

1.5.1 Detection chemistries

The qRT-PCR method allows the detection of amplicon accumulation since it is performed using sensitive fluorogenic Taq-Man Probes (Livak *et al*, 1995, Heid *et al*, 1996), molecular bacons (Tyagi and Kramer, 1996) and scorpions (Whitcombe *et al*, 1999) or more sensitive but less specific intercalating dyes like SYBR Green I which only fluoresce intensely when associated with double stranded DNA). The use of fluorescent dyes allows the amplification and detection steps of the PCR assay to be combined. The fluorescent reagents used with homogeneous fluorescent reporting chemistries can be grouped into two: specific and non-specific detection chemistries. In non-specific method, DNA intercalating dyes are used. They are relatively inexpensive but non-specific and require post-PCR dissociation curve analysis. Probes and molecular beacons are the molecules used in the specific chemistry. They are specific and allow huge choice of chemistries but they are expensive. In the specific chemistry, fluorescent resonance energy transfer (FRET) or similar interactions between the donor and quencher molecules form the basis of detection system (Bustin *et al.,* 2004a).

The most simple detection method in real-time RT-PCR requires a dye that emits fluorescent light when intercalated into double-stranded DNA (dsDNA) but not to single-stranded DNA. The intensity of the fluorescence signal is proportional to the amount of all double-stranded DNA present in the reaction (Wong *et al.,* 2005). In the earlier experiments ethidium bromide and YO-PRO-1 were used as intercalating dyes. However, by ethidium bromide it is not possible to distinguish between the amounts of the ssDNA and the dsDNA. Currently, SYBR Green I is the most frequently used one. SYBR Green I is a binding dye for ds DNA. It binds to the minor groove of the ds DNA and increases the fluorescence over a hundred fold. SYBR Green I is preferred in most of the cancer studies due to the usage of high number of genes and sample input. It is more precise, and produces more linear decay plot than the TaqMan detection system (Bustin *et al.,* 2004b). Although its specificity is less than TaqMan probes it is more sensitive and cheaper because of the lack of probe associated cost (Bustin *et al.,* 2004a).

1.5.2 Quantification strategies

The choice of quantification strategy is the most important part of the real-time RT-PCR to express the qRT-PCR data properly. It depends on the target sequence, expected amount of the mRNA, and degree of accuracy required (Pfaffl, 2001). Two strategies can be performed in real-time RT-PCR: absolute and relative quantification (Pfaffl, 2001). Absolute quantification refers to an analysis where the comparison of unknown samples to an external standard provides an accurate and reliable method for the quantification of mRNA samples. The most accurate way of absolute quantification is to construct standard curves. The reliability of an absolute real-time RT-PCR assay depends on the condition of 'identical' amplification efficiencies for both the native target and the calibration curve in RT reaction and in the following kinetic PCR. Therefore, it should be used in the analysis where the determination of exact copy number is necessary. In most analyses it is enough to determine the relative change in the expression of the gene. In these cases, relative quantification is easier to perform than absolute quantification since there is no need for a calibration curve. It is based on the expression levels of a target gene versus a housekeeping gene (reference gene). The units used to express relative quantities are irrelevant and the relative quantities can be compared across multiple real-time RT-PCR experiments.

Relative quantification determines the changes in steady-state mRNA levels of a gene across multiple samples and expresses it relative to the levels of an internal control RNA (Pfaffl, 2001). This reference gene is often a housekeeping gene and can be co-amplified in the same tube in a multiplex assay or can be amplified in a separate tube (Huggett *et al.,* 2005). Therefore, relative quantification does not require standards with known concentrations and the reference can be any transcript, as long as its sequence is known. To calculate the expression of a target gene in relation to an adequate reference gene various mathematical models are established. Calculations are based on the comparison of the distinct cycle determined by various methods, e.g. crossing points (CP) and threshold values (Ct) at a constant level of fluorescence; or CP acquisition according to established mathematic algorithm. To date, several mathematical models that calculate the relative expression ratio have been developed. Relative quantification model without efficiency correction is given below (equations 1.1).

Equation 1.1: $R= 2^{-\Delta\Delta Ct}$

The equations given are used for the reactions, which have 100% amplification efficiencies. If the efficiency of the reaction is below 100%, the equations have to be corrected. The efficiency of PCR provides information about the amplification rate and varies from 0 to 1. The rate equal 1 (=100%) means that in each cycle the number of copies is doubled. The efficiency of a PCR reaction can be calculated from the slope of a standard curve:

Equation 1.2: Exponential amplification: $10^{(-1/slope)}$
Equation 1.3: Efficiency (E): $(10^{(-1/slope)})-1$

The above equations are valid if the dilution series of the standard curve is prepared by 1/10 dilution series (Rasmussen, 2001).

With kinetic PCR efficiency correction, the relative expression ratio of a target gene is calculated based on its real time PCR efficiency (E) and the difference between Ct (CP) values of the unknown sample versus the control sample. By dividing expression ratio of the target gene to the reference gene relative quantification is performed (Pfaffl, 2001).

1.5.3 Normalization

Housekeeping genes are expressed in all nucleated cells and required for cell survival (Thellin *et al.,* 1999). Thus, accurate normalization of quantitative RT-PCR data requires knowledge of which housekeeping gene or genes (reference genes) are expressed at equal or similar levels within a group of samples (Vandesompele *et al.,* 2002). Inclusion of an endogenous reference gene or genes (RGs) is crucial to standardize initial RNA quantity to overcome bias originating from RNA measurement errors, problems with RNA integrity, and differential cDNA conversion efficiencies (Bustin *et al*., 2005; Stahlberg *et al*., 2004a; Stahlberg *et al*., 2004b). Many techniques developed for mRNA quantification use housekeeping genes as internal standards. Quantification of a target gene requires the use of a proper RG whose expression is relatively stable across samples to estimate the degree of variability within and among experimental groups as well as to standardize the expression to a baseline common to all samples (Vandesompele *et al.,* 2002). Nevertheless, numerous studies show an inherent instability in regard to expression of housekeeping genes, many of which are still commonly used as references (Schmittgen *et al*., 2000; Zhong *et al*., 1999; Barber *et al*., 2005; Selvey *et al*., 2001). Beta actin (ACTB) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) have been historically considered to be adequate housekeepers for normalization of gene expression. However, studies show that their expression may be regulated as well between different tissues (Bereta *et al.,* 1995; Chang *et al.,* 1998). This is partly explained by their participation to other functions in different cell types. GAPDH has been found to be regulated by biphosphonates in breast cancer cell lines (Valenti *et al.,* 2006).

Analysis of gene expression is fundamental for cancer research for the detection of subtle differential expression between tumor and normal tissues or among different tumor types. In particular, recent target validation and disease diagnostic marker selection studies rely primarily on gene expression comparisons between tumor-normal pairs (Chiu *et al*., 2005; Cerutti *et al*., 2007; Jarzabek *et al*., 2005). Moreover, the use of multiple endogenous RGs significantly increases the accuracy of the normalization by reducing the impact of outliers (Vandesompele *et al*., 2002; Bustin

et al., 2005). Accordingly, a plethora of single or combinational usage of two or more RGs has been recommended for relative quantification of expression data for various tumor tissue types (Ohl *et al*., 2005; Ohl *et al*., 2006; Jung *et al*., 2007).
In breast cancer qRT-PCR studies, different single housekeeping genes have been used to quantify the expression level of target genes (Folgueira *et al*., 2006; Morse *et al*., 2005; Kroupis *et al*., 2005; de Cremoux *et al*., 2000; Potemski *et al*., 2006; Iwao *et al*., 2000). Recently, *MRPL19* and *PPIA* were reported as a stable RG combination to analyze benign and malignant breast cancer specimens (McNeill *et al*., 2007). Similarly, Lyng *et al*. (2008) reported an RG panel comprised of *TBP*, *RPLP0* and *PUM1* for normalizing the gene expression levels across the ER+ and ER- breast tumors, and normal breast tissues (Lyng *et al*., 2008).
The researcher should find the most suitable reference gene for the experimental setup. For this purpose, several excel based programs have been developed (Huggett *et al.,* 2005). geNorm and NormFinder are the commonly used programs developed for this purpose. Basically they find the best reference gene or genes by using the geometric mean of the reference gene expression of sample cDNAs.
The geNorm VBA (Visual Basic for Applications) applet for Microsoft Excel determines the most stable reference genes from a set of tested genes in a given cDNA sample panel, and calculates a gene expression normalization factor for each tissue sample based on the geometric mean of a user-defined number of reference genes. geNorm calculates the gene expression stability measure M for a reference gene as the average pairwise variation V for that gene with all other tested reference genes. Stepwise exclusion of the gene with the highest M value allows ranking of the tested genes according to their expression stability (Vandesompele *et al*., 2002).The software is free and can be downloaded from
http://medgen.ugent.be/~jvdesomp/genorm/.
NormFinder is an algorithm for identifying the optimal normalization gene among a set of candidates. It ranks the set of candidate normalization genes according to their expression stability in a given sample set and given experimental design. The algorithm is rooted in a mathematical model of gene expression and uses a solid statistical framework to estimate not only the overall expression variation of the candidate normalization genes but also the variation between sample subgroups of

the sample set e.g. normal and cancer samples. Notably, “NormFinder” provides a stability value for each gene, which is a direct measure for the estimated expression variation enabling the user to evaluate the systematic error introduced when using the gene for normalization (Andersen *et al*., 2004). The software can be downloaded from http://www.mdl.dk.

1.6 Specific aim

The main aim of this study was to analyze differentially expressed genes which may contribute to the breast cancer development.

We used two different approaches to tackle this aim:

1. In a previous study conducted in our research group, the number of candidate target genes that were induced by BRCA1 over expression were identified by using Suppression Subtractive Hybridization (SSH) techniques (Atalay *et al.*, 2002).

We aimed to find out the expression profiles of the selected target genes from a BRCA1-induced gene list (*OVCA1, OVCA2, ERBIN, RAD21, XRN2, RENT2, SMG1* and *MAC30*) in normal-matched primary breast tumors by qRT-PCR and to correlate the gene expression profiles of selected candidate genes with *BRCA1* and various pathology parameters.

The strategies used in this part of the study were as follows:

Ectopic expression of BRCA1 in tissue culture cells

- Transient transfection of the MCF7 cells with full length BRCA1-transcript containing plasmid pCMVmycBRCA1 or empty control plasmid pCMVmyc.
- Induction of BRCA1 expression in U2OS osteosarcoma derived UBR60-bcl2 cells that contain stable BRCA1 cDNA containing construct under the control of tetracycline-regulated promoter.
- Confirmation of BRCA1 over expression in these cell lines by qRT-PCR and western blot analysis.
- Control for up-regulated BRCA1 protein activity by using GADD45 as a positive control.

- Down regulation of BRCA1 gene expression in MCF7 cells by si- and sh-BRCA1 RNA transfection and confirmation of BRCA1 down regulation by qRT-PCR and western blot analysis.
- Designing gene specific primers and calculating amplification efficiencies.
- Analysis of the target genes expression profiles in BRCA1 over expressing cells and in si- and sh BRCA1 down regulated MCF7 cells by qRT-PCR.
- Expression profiles of target genes in eight breast carcinoma cell lines by qRT-PCR.

Expression profiles of target genes in tumor and paired-normal breast tissues determined by qRT-PCR

- Collection and pathological assessment of human primary breast and paired normal tissue samples (n=32).
- RNA preparation and quality assessment.
- Determining the expression profiles in breast samples by qRT-PCR.
- Correlation of target gene expression profiles with each other and to BRCA1.
- Statistical evaluation of target gene expressions with pathological information.
- Evaluation of the target genes in independent microarray datasets for their predictive power to pathological typing and patient survival.
- Promoter analysis of the target genes by bioinformatic analysis.

Identification of suitable endogeneous reference gene(s) for qRT-PCR analysis in breast tumor and normal tissues

Since we were using qRT-PCR technique for relative quantification, we had to identify a suitable RG(s) that can be used as a normalization factor (NF) for more accurate and reliable normalization of paired breast tumor-normal tissues. We search the literature to find out the suitable endogenous reference genes that can be used for the normalization of the target gene expression to reach the relative quantification expression values of the genes in the breast tumor and normal tissues by qRT-PCR and found no consistent information. We tackled this problem by studying 18 candidate reference genes in normal matched breast tumor tissues.

The strategies used in this part of the study were as follows:

- Matched pairs of normal and tumor breast samples were used for minimization of inter-individual variation and to increase the power of data analysis.
- Total RNA was assessed stringently and only the high quality samples were included in the study.
- 18 candidate RGs were simultaneously analyzed with optimized conditions to determine their expression patterns in samples.
- The tumor and normal matched samples were included in the same run in duplicates for a studied gene in qRT-PCR.
- Established softwares combined with statistical analysis were used to rank the candidate RGs for their expression stability and their suitability as normalization factors (NF).
- To assess the significance of the selected RGs for normalization, the expression level of *GSN* mRNA was measured by qRT-PCR and statistically evaluated in the same set of tumor and matched normal breast tissue samples.
- The suitability of 18s rRNA as a RG gene was also assessed in breast tumor and matched normal breast tissue samples.
- The low level expression of 18S rRNA in the breast tumors led us to perform the bisulfite sequencing analysis to determine the CpG island methylation patterns in the regulatory region of this gene.

2. New high-throughput technologies have opened the possibility to study the gene expression profiles of the breast tumors, find new susceptibility genes and they added valuable information for the molecular sub-typing of breast cancer. The identification of an intrinsic gene-set exhibiting high variability among different tumor clusters has been informative in describing different subtypes of breast cancer samples. Meta-analysis of microarray datasets has the potential to lead to more comprehensive measures of the existing differential gene expression data and can therefore provide gene sets with a high diagnostic value.

In the second part of the study, we primarily aimed to provide gene lists that (a) are discriminative of breast cancer types (IDC, ILC) and normal breast cell populations, (b) may yield breast tumor markers that are invariably expressed across independent experiments, and (c) provide a set of consistently differentially expressed gene candidates with potential discriminative ability for tumor subtypes.

The strategies used in this study were as follows:

- Two comparable independent microarray datasets that contain normal breast, invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC) samples data were downloaded from the Stanford Microarray Database (SMD).
- Expression values that were missing in more than 20% of the data were excluded from the analysis.
- Datasets were combined with respect to probe IDs using a set of customized perl routines.
- Data were filtered separately for ductal and lobular samples to exclude the missing data.
- A resampling based strategy was developed to test the significance of the difference between group medians (e.g. ductal vs lobular) upon a series of resampling schemes from the original and multiple randomly shuffled datasets.
- The genes were listed for classification of normal breast samples and breast tumors encompassing both the ILC and IDC subtypes.

- The gene lists were validated by analysis of existing breast cancer microarray datasets.
- The expression status of the selected genes was tested on 10 independent primary IDC breast samples and matched non-tumor controls by real-time qRT-PCR.

CHAPTER 2. MATERIALS & METHODS

2.1 Materials

2.1.1 General Reagents

The general laboratory chemicals were supplied from Sigma Chemical Co. (St. Louis, USA), Merck (Darmstadt, Germany), Stratagene (Heidelberg, Germany) and AppliChem (Darmstadt, Germany).

2.1.2 Nucleic acids and Proteins

1 kb DNA ladder and pUC Mix Marker 8 were used as DNA molecular weight size markers in this study and supplied from New England Biolabs (U.K.) and MBI Fermentas (Germany). Protein size markers were from Bio-Rad (Broad Range 161-0318), (USA) and New England Biolabs (Broad Range P7708S) (U.K.).

2.1.3 Oligonucleotides

The oligonucleotides used in polymerase chain reactions were synthesized and supplied from Iontek Inc. (Istanbul, Turkey).

2.1.4 Enzymes

Restriction endonucleases were supplied from New England Biolabs. Taq DNA polymerases and SYBR Green Supermix were supplied from MBI Fermentas, Fnzymes (Finland) and (Bio-Rad, California, USA). Reverse Transcriptase was supplied from MBI Fermentas.

2.1.5 Bacterial strains

Bacterial strains were stored at -70°C in LB medium containing 50% (v/v) glycerol for long term storage. Recombinant clones were stored under the same conditions in media supplemented with appropriate antibiotics. Strains were maintained as isolated colonies on LB agar plates at 4 °C for short term storage.

2.1.6 Plasmids

pCMV.myc and pCMV.myc.BRCA1 vectors were gifts from Dr. Tim Crook (Imperial College, London) and pSUPER.retro.pro vector containing the inserts of shRNA sequence specific for BRCA1 and scrambled sequence were gifts from Dr. Luc Gaudreau (Université de Sherbrooke, Quebec, Canada). The BAC Human CTD 2371A15 clone was supplied from Invitrogen (USA).

2.1.7 Protein transfer materials

Immobilen P transfer (PVDF) membrane was from Roche (Germany) and 3MM filter paper was from Whatman International Ltd. (Madison, USA).

2.1.8 Photography and autoradiography

The films used for autoradiography were Kodak and the development of the films were performed with Hyperprocessor (Amersham, UK)

2.1.9 Tissue culture reagents and cell lines

Dulbecco's Modified Eagle's Medium (DMEM), fetal calf serum, L-glutamine, penicillin/streptomycin and trypsin were obtained from Biochrom (UK). Antibiotics used for selection in medium: Hygromycin was from Boehringer Mannheim, and puromycin and tetracycline were from Sigma.

MCF-7 is a human breast carcinoma cell line, UBR60-bcl2 cell line was a gift from Dr. Paul Harkin (Queen's University, Belfast, UK), which expresses BRCA1 under the control of tetracycline-regulated promoter, has been previously described (*Harkin et al.*, 1999).

2.1.10 Transfection reagents

FuGene 6 and Oligofectamine transfection reagents were obtained from Roche and Invitrogen. The medium used for Oligofectamine transfection was OptiMEM I and obtained from Invitrogen.

2.1.11 Kits

Trireagent was from AppliChem (Darmstadt, Germany) and the RNA isolation kit was from Macharel Nagel, (Duren, Germany). MessageClean kit was from GenHunter Co. (MA, USA). RevertAid first strand cDNA synthesis kit was from MBI Fermentas. Miniprep, Midiprep kits and PCR purification and Gel purification kits were from Qiagen (Germany). ECL western blotting detection reagent was from Amersham (UK).

2.1.12 Antibodies

BRCA1 (MS110) antibody was obtained from Oncogene Research products (Darmstadt, Germany). ERBIN antibody was a gift from Dr. Jean-Paul Borg (INSERM, Marseille, France). Calnexin antibody was obtained from Santa Cruz (USA). Anti-mouse and anti-rabbit HRP conjugated secondary antibodies were obtained from Sigma (USA).

2.2 Solutions and media

2.2.1 General solutions

TAE: Stock solution (50XTAE) was prepared by addition of 121g Tris-base, 18,6g EDTA, and 28.55ml glacial acetic acid to 500ml ddH_2O. pH of the stock solution was adjusted to 8.5. Working solution (1XTAE) was prepared by dilution of 50XTAE to 1X with ddH_2O.

Ethidium bromide: 10 mg/ml in water (stock solution), 30 ng/ml (working solution).

6X Agarose Gel Loading Dye: A mixture of 0.009g bromophenol blue (BFB), 0.009g xylene cyanol (XC), 2.8ml ddH_2O, 1.2 ml 0.5M EDTA was prepared. The total volume was brought to 15ml by addition of glycerol.

2.2.2 RNA solutions

DEPC-Treated Water: 1ml DEPC was added to 1lt ddH_2O and stirred under hood overnight. DEPC was inactivated by autoclaving.

FA Gel Buffer: Stock solution (10XFA Gel Buffer) was prepared by dissolving 20.927g MOPS, 3.40g NaAc and 1.86g EDTA in 500ml ddH_2O. pH of the stock solution was adjusted to 7.0. Working solution was prepared by diluting the stock solution to 1X with ddH_2O.

5x RNA Loading Buffer: Bromophenol blue solution 16 µl, 500 mM EDTA, pH 8.0 80 µl, 37% formaldehyde 720 µl, 100% glcerol 2 ml, Formamide 3084 µl, 10x FA gel buffer 4 ml,
RNase free (DEP-C treated) water to 10 ml

2.2.3 Tissue culture solutions

Growth medium: DMEM was supplemented with 10% fetal calf serum, 1mM glutamine and 50 mg/ml penicillin/streptomycin

Freezing solution: 10% DMSO and 90% FCS were mixed freshly.

PBS: Stock solution (10XPBS) was prepared by dissolving 80g NaCl, 2g KCl, 11.5g $Na_2HPO_4.7H_2O$, and 2g KH_2PO_4 in 1lt ddH_2O. Working solution (1XPBS) was prepared by dilution of 10XPBS to 1X with ddH_2O. pH of the working solution was adjusted to 7.4.

Tetracyclin: 0.01 gr tetracycline was dissolved in 10 ml 0f 70% ethanol and stored at -20^{o}C.

Puromycin: 1 mg puromycin was dissolved in 1 ml of DMEM, sterilized by filtrartion and stored at -20^{o}C.

2.2.4 Protein extraction and western blotting solutions

RIPA Buffer: 10mM Tris.Cl pH: 8.0, 1mM EDTA, 150mM NaCl, 1% NP-40 and 1X protease inhibitor mix were mixed in ddH_2O.

Bradford Stock Solution: 17.5 mg Coomassie brilliant blue was dissolved in 4.75 ml ethanol and 10 ml phosphoric acid and completed to 25 ml final volume with ddH_2O.

Bradford Working Solution: 1.5 ml Bradford stock solution was mixed with 0.75 ml 95% Ethanol and 1.5 ml phosphoric acid and completed to final volume up to 25 ml with ddH_2O.

Acrylamide-Bisacrilamide solution: 29 gr acrylamide and 1 gr bisacrylamide were dissolved in 100 ml ddH_2O and stored in the dark.

10% APS: 0.1 gr APS was dissolved in 1 ml of ddH_2O.

5X Loading Buffer: 62.5 mM Tris-HCL, pH:6.8, 5% β-mercaptoethanol, 2% SDS, 15% glycerol and 0.001% bromophenol blue.

5X Running Buffer: 45 g tris, 216 g glycine and 15 g SDS were dissolved in 3 liters of ddH_2O.

Wet Transfer Buffer: 6 g tris and 28.8g glycine was mixed with 1ml 10% SDS and 20% methanol and completed to final volume of 1 liter.

10XTBS: 12.19 g Tris-base and 87.76 g NaCl were dissolved in 1 liter of ddH2O and the pH was adjusted to 8 to prepare 10X TBS stock solution.

TBS-T: 0.3% Tween 20 was added into 1X TBS solution.

Blocking Solution: 3% milk powder in 0.3% TBS-Tween 20 solution

2.3 General methods

2.3.1 Transformation of *E. coli*

E. coli DH5α strain was used for transformation. 200μl of culture was thawed on ice and immediately mixed with 2 μl of plasmid DNA in Greiner transformation tubes. Plasmid-bacteria mixture was incubated on ice for 30 min and mixed gently without vortex. Vortex was not used in any step of this procedure. The mixture was transferred to 42^{o}C water bath for 30 sec and immediately transferred to ice and incubated on ice for 2 min. After 2 min the mixture was added onto 800 μl of LB medium without any antibiotics and incubated in the shaker at 37oC for 1 hour at 220 rpm. 50µl from the 1 hr incubated culture was cultured to LB agar plate containing the desired antibiotics. The remaining LB-plasmid mixture was centrifuged and the pellet was dissolved in 200 µl of LB medium and cultured in LB agar plate. All the plates were incubated at 37^{o}C for 16 hours.

2.3.2 Plasmid DNA preparation

Small scale isolation of plasmid DNA (mini-prep) was performed with Qiagen plasmid isolation kit according to the manufacturer's instructions. This procedure yields approximately 200 ng/µl of plasmid DNA for 1ml of LB culture.
For large-scale preparation of pure plasmid DNA, the Qiagen 100 plasmid isolation kit was used by following the manufacturer's instructions. This procedure yields approximately 1 µg/µl of plasmid DNA for 100 ml of LB culture.

2.3.3 DNA Extraction from tumor and normal tissues

The frozen tumor and normal tissue samples were cut into 5-µm-thick sections and used for DNA isolation (4-5 slices for each sample). The genomic DNA isolation from tumor and normal tissues was performed directly by use of NucleoSpin Tissue kit (Macharel Nagel, Duren, Germany) according to the manufacturer's instructions. The tissue sections were incubated with proteinaseK at 56^{o}C for 24 hours. After the incubation colon purification was performed according to manufacturer's instructions. The DNAs were eluted in a total volume of 100 µl. The concentration of the isolated DNA and the ratio of absorbance at 260 nm to 280 nm were measured with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Montchanin, DE, USA) in triplicate.

2.3.4 RNA Extraction

2.3.4.1 Extraction of total RNA from tissue samples.

The isolation of RNA requires pure reagents and care in preparation due to the sensitivity of RNA to chemical breakdown and cleavage by nucleases. Therefore all the solutions and materials were treated with DEPC (AppliChem, Darmstadt , Germany) in order to avoid RNase contamination and hence degradation of RNA. Total RNA of tumor tissues was isolated with TRI reagent (AppliChem, Darmstadt, Germany). The frozen tumor and normal tissue samples were cut into 5-µm-thick sections and used for RNA isolation (4-5 slices for each sample). Tissue samples were lysed in 1ml TRI reagent with a homogenizer and passed through a 21-gauge needle several times. After 5 min incubation at room temperature, 0.2ml chloroform was added per ml of TRI reagent. Tubes were shaken vigorously by hand for 15 seconds and incubated at room temperature for 2-3 min. After incubation the mixture was centrifuged at 12000xg for 15 min at 4^{o}C and then aqueous phase was collected into a new tube. 0.5ml isopropanol was added onto aqueous phase per 1ml of TRI reagent used. The mixture was incubated at room temperature for 10 min and then centrifuged at 12000xg for 15 min at 4^{o}C to recover RNA. The supernatant was removed and the pellet was washed with 75% ethanol twice, centrifuged at 7500xg for 5 min at 4^{o}C. The pellet was air-dried and dissolved in ddH_2O. The isolated RNA solution was subjected to a second round of isolation by using NucleoSpin RNA II kit (Macharel Nagel, Duren, Germany) to remove any remaining contaminants of DNA.

2.3.4.2 Extraction of total RNA from tissue culture cells

Exponentially growing monolayer cultures were washed twice with ice-cold PBS, scraped witha scraper, pelleted and snap frozen in liquid nitrogen and stored at -70^{o}C until needed for RNA preparation. The total RNA isolation from cell line pellets was performed directly by use of NucleoSpin RNA II kit according to the manufacturer's instructions. The RNAs were eluted in a total volume of 30 µl. The concentration of the isolated RNA and the ratio of absorbance at 260 nm to 280 nm were measured with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Montchanin, DE, USA) in triplicate. The integrity of the isolated RNA samples was measured with the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Isolated RNAs were stored at -80^{o}C.

2.3.5 Quantification of nucleic acids

The concentration of the isolated RNA and DNA samples and plasmid concentrations were measured with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Montchanin, DE, USA). The ratio of absorbance at 260 nm to 280 nm was also measured with the NanoDrop ND-1000 spectrophotometer. The integrity of the isolated RNA samples was measured with the Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

2.3.6 Agarose gel electrophoresis

2.3.6.1 Agarose gel electrophoresis of DNA

DNA fragments were fractionated by horizontal gel electrophoresis in 2% (w/v) agarose gel by using 1xTAE buffer. Agarose was completely dissolved in 1xTAE electrophoresis buffer in the desired percentages and ethidium bromide solution was added to final concentration of 30ng/ml. 2 µl 6X DNA loading dye was added to 10 µl of quantitative real time RT-PCR (q-rt-RT-PCR) products and 12 µl of normal PCR products and total volume was loaded to each well and run at 100 V for 30 minutes. pUC Mix Marker 8 (MBI Fermentas, Ontario, Canada) was used as DNA size marker for the products up to 1000 bps (Figure 2.2) and 1 kb DNA Ladder (New England Biolabs, USA) was used for larger product sizes (Figure 2.2). Nucleic acids were visualized under ultraviolet light (long wave, 340 nm) (Transilluminator, Bio-Rad, California, USA) and MultiAnalyst (Bio-Rad, California, USA) software was used to take photographs of the gels.

2.3.6.2 Agarose gel electrophoresis of RNA

Total RNA samples were fractionated by horizontal gel electrophoresis in 1.2% (w/v) agarose gel by using 1X FA gel buffer. After 1.2% agarose gel solution was prepared the mixture was cooled and 1.8 ml of 37% formaldehyde and 1 µl of ethidium bromide from a stock of 10mg/ml solution were added in 100 ml of gel. The gel was poured in a laminar flow hood. 2µg of total RNA from each sample was mixed with 1 volume of 5x loading buffer per 4 volumes of RNA sample and incubated for 5 minutes at 65 ^{o}C, and chilled on ice then loaded to the gel. The gel was run at 5-7 V/cm in 1xFA gel running buffer. Transilluminator (Bio-Rad, California, USA) was used to visualize the DNA bands under ultraviolet light (long wave, 340 nm). MultiAnalyst (Bio-Rad, California, USA) software was used to take photographs of the gels.

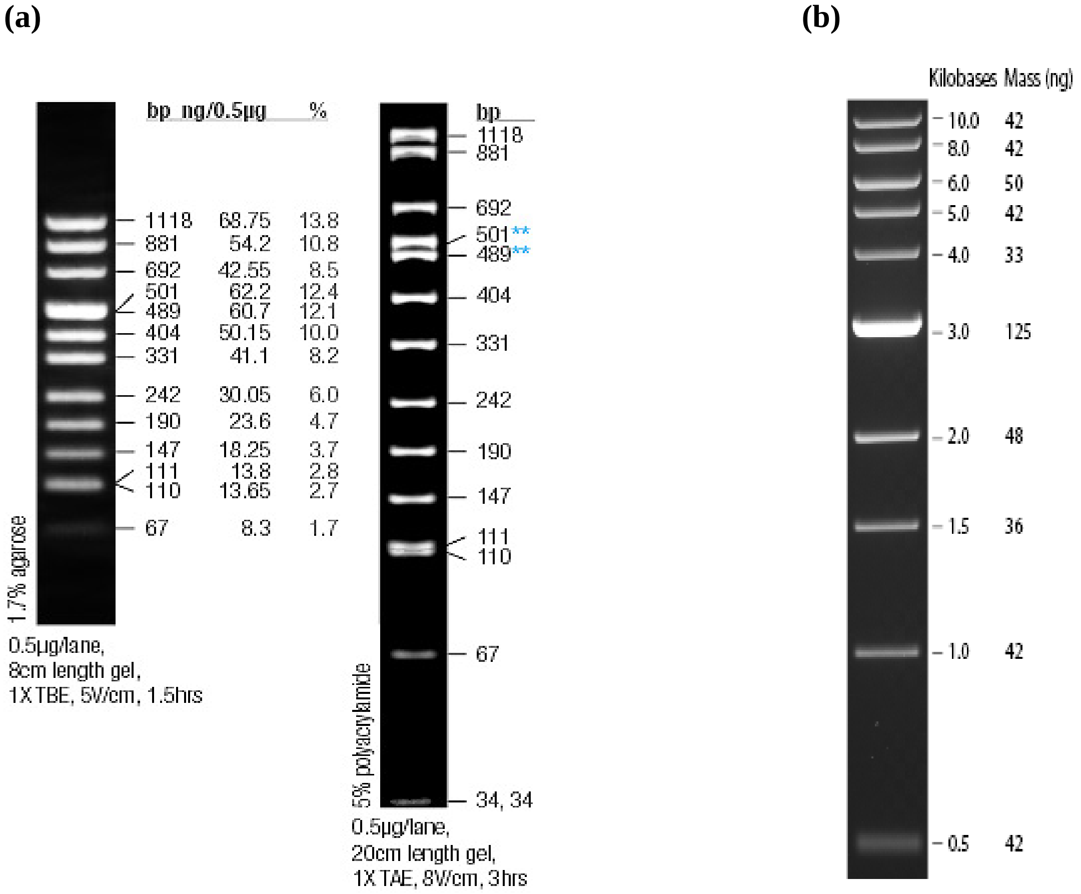


Figure 2.1 pUC Mix Marker 8 and 1kb DNA ladder. (a) Shows the image of pUC Mix Marker 8 on a 1.7% agarose gel and **(b)** shows the 1 kb DNA Ladder visualized by ethidium bromide staining on a 0.8% TAE agarose gel.

2.4 Tissue culture techniques

2.4.1 Growth conditions of cell lines

All cell lines used in this study were maintained in Dulbecco’s Modified Eagle’s Medium (DMEM, Biochrom, UK) supplemented with 10% fetal calf serum, 1 mM glutamine and 50 mg/ml penicillin/streptomycin and appropriate selective antibiotic. The names, types and ER status of the cell lines used in this study were given in Table 2.1.

UBR60-bcl2 cell line which expresses BRCA1 under the control of tetracycline-regulated promoter has been previously described (Harkin *et al,* 1999) and was a gift by Dr. Harkin. UBR60-bcl2 cells were maintained in growth medium containing 500 μg/ml geneticine (G418), 1 μg/ml puromycin, 200 μg/ml hygromycin and 1 μg/ml tetracycline. Upon tetracycline withdrawal the cells express the BRCA1 gene (Tet-off system, Clontech Laboratories Inc.).

The cells were incubated at a 37^{o}C incubator with an atmosphere of 5% CO_2 in air. The cells were passaged before reaching confluence. The growth medium was aspirated and the cells were washed with 1X PBS. Trypsin solution (Biochrom, UK) was added to the flasks to detach the monolayer cells from the surface. Cells were dispersed by pipetting the cells with fresh medium. The cells were transferred to new flasks using different dilutions depending on requirements.

DMEM and PBS were kept at 4^{o}C, trypsin was kept at -20^{o}C. All the solutions were warmed to 37^{o}C before use.

Table 2.1: Breast carcinoma cell line information

Breast Cancer Cell Lines	ATCC Number	Cancer Type	ER Status[a]
MDA MB 231	HTB 26	Adenocarcinoma	N
MDA MB 453	HTB 131	Metastatic carcinoma	N
MDA MB 468	HTB 132	Adenocarcinoma	N
BT 474	HTB 20	Ductal carcinoma	P
BT 20	HTB 19	Carcinoma	P
MCF 7	HTB 22	Adenocarcinoma	P
HCC 1937	CRL 2336	Primary ductal carcinoma	N
T47D	HTB 133	Ductal carcinoma	P
Htert- HME 1	CRL 4010	Epithelial; immortalized with hTERT	N

[a]ER: Estrogen receptor; N: negative; P: positive.

2.4.2 Cryopreservation of cell lines

Exponentially growing cells were harvested by trypsinization and neutralized by adding fresh growth medium. The cells were counted and precipitated at 1500 rpm for 5 minutes. The cells were resuspended with freezing medium at a concentration of 4x10^6 cells/ml/vial. Freezing medium was made up 90% FBS and 10% DMSO and stored in the cryotubes. The cryotubes were stored at -80^{o}C overnight and transferred to the liquid nitrogen tank for long term storage.
To reculture the frozen cells, they were thawed rapidly at 37^{o}C and mixed with 5ml growth medium and centrifuged at 1500 rpm for 5 minutes at room temperature. Supernatant was removed and the cells were resuspended with fresh complete growth medium. Cells were grown in the 10 cm cell culture plates in the incubator.

2.4.3 Transfection of cell lines

2.4.3.1 Transfection of tissue culture cells with BRCA1 containing vector.

Exponentially growing MCF7 cells were plated in 6 well-plates at a concentration of 150.000 cells/well a day before the transfection. The cells were incubated overnight and reached the 70-80% confluency. Next day, the medium was replaced with fresh medium. MCF7 cells were transfected with pCMVmycBRCA1 and pCMVmyc vectors. In separate tubes, 250ng pCMVmycBRCA1 and pCMVmyc vectors were mixed with FuGENE 6 transfection reagent (Roche Applied Science) and incubated for 30 min at room temperature. The mixture was added onto the cells drop by drop. The optimum amount of FuGENE reagent needed for an efficient transfection was specified to be 3µl for 1µg of plasmid DNA. Thus the plasmid-FuGENE 6 mixture was prepared accordingly. The cells were harvested after 24 hours of incubation and used for further experiments.

2.4.3.2 shRNA and siRNA mediated knockdown of BRCA1 expression

shRNA mediated knockdown of BRCA1 expression

The pSUPER.retro.pro vector (Figure 2.1) containing the inserts of shRNA sequence specific for BRCA1 and scrambled sequence were gifts from Dr. Luc Gaudreau (Université de Sherbrooke, Quebec, Canada) (Gaudreau *et al*, 2006). MCF7 cells were grown in 6 well-plates to 60-70% confluency and transfected with 1µg of shRNA expressing vector and with 1 µg of control vector using the FuGENE 6 transfection reagent (Roche Applied Science) as described previously. The selection medium containing 2 µg/ml puromycin was applied 24 hours post-transfection to produce stable clones. Puromycin selection was applied for one week and the surviving cells were diluted and plated to a 150mm plate to obtain stable clones. The independent colonies were picked, transferred to a 6 well-plates, and expanded in tissue culture selective medium.

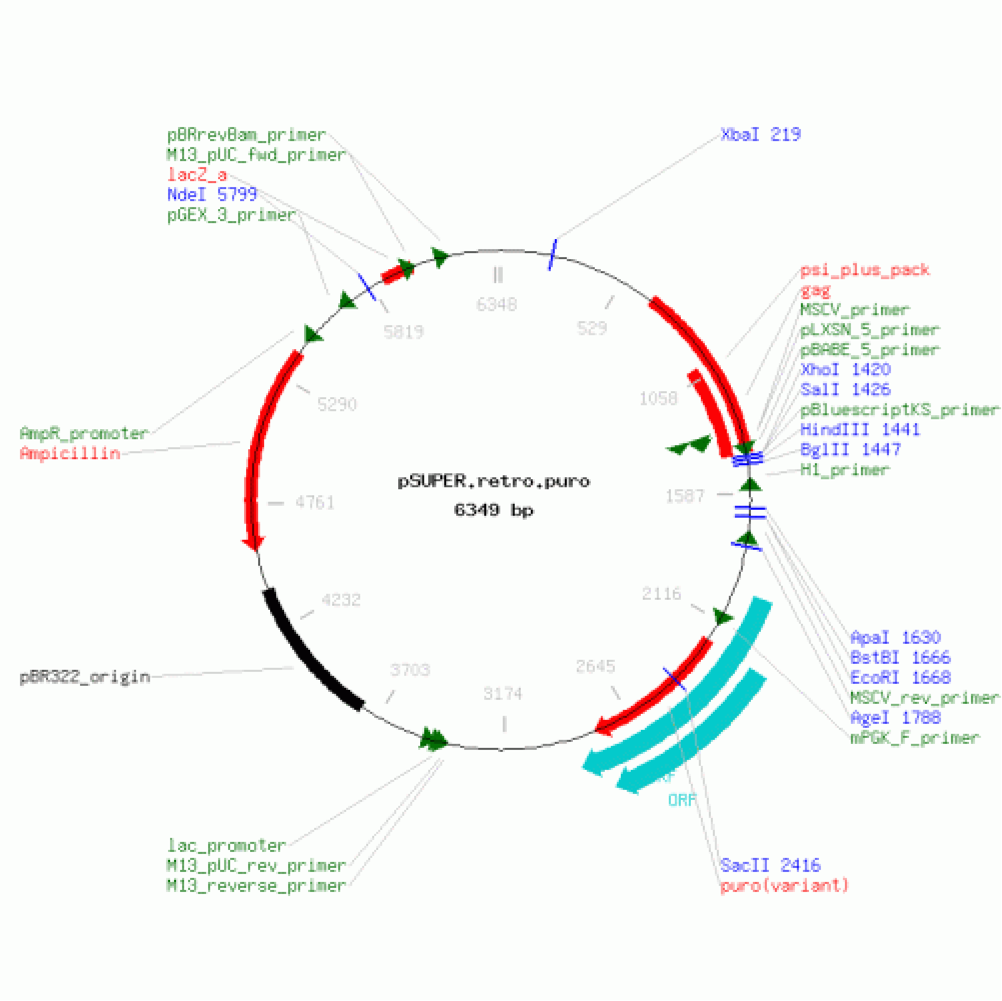


Figure 2.2 Map of pSUPER.retro.puro vector (OligoEngine).

siRNA mediated knockdown of BRCA1 expression

Exponentially growing MCF7 cells were plated in 6 well-plates at a concentration of 150.000 cells/well a day before the transfection. The cells were incubated overnight in growth medium. Next day, the medium was changed with fresh, serum and antibiotic-free medium. Small inhibitory RNA (siRNA) sequence pools directed against BRCA1 and pooled scrambled control siRNA sequences were obtained from Dharmacon Research, Inc. (Lafayette, Colorado). Transfection of siRNAs into MCF7 cells was carried out with Oligofectamine reagent (Invitrogen). 250 nM of siRNA was mixed with 4 µl of Oligofectamine and incubated for 30 min at room temperature. The mixture was added onto the cells drop by drop. 4 hours after transfection 500µl containing 30% FBS was added onto the cells and the cells were placed into incubator. The cells were harvested after 24 hours of transfection for further experiments.

2.5 Patient samples

2.5.1 Collection of tissue samples

Primary tumor samples and matched normal breast tissues samples were obtained from Ankara Numune Research and Teaching Hospital, Ankara, Turkey. These tissues were collected at the time of biopsy or surgery and immediately snap-frozen in liquid nitrogen and stored at -80^{o}C until RNA extraction. The frozen tissue samples were sectioned and mounted on glass slides. The slides were stained with hematoxylin and eosin for histopathological examinations. All of the slides were reviewed by a pathologist to determine the integrity of the specimens. All the tumor samples had been classified as infiltrating ductal carcinoma (IDC). The tumor samples containing more than 90% tumor cells and patient-matched tissue pairs with normal histological examination were included in this study. Tumor grades were determined according to the Bloom-Richardson score. The pathological information of the tumor samples is listed in Table 2.2 and Table 2.3. The use of the tissue material in this project was approved by the Research Ethics Committee of Ankara Numune Research and Teaching Hospital.

2.5.2 Tissue sectioning and staining for pathological examination

Frozen tissue samples were cut in a series of 5 µm sections by using Shandon cryotome cryostat (Thermo Scientific, USA) and transferred on to the adhesive slides (Histobond, Marienfeld, Germany) to prevent detachment of the sample sections from the slides. The slides were stained with hematoxylin for 2 minutes washed under the tap water and rinsed with 1% acidic alcohol. The rinsed slides were gently washed with 1% ammonia water (v/v) and rinsed under the tap water and stained with Eosin for one minute then rinsed with tap water. The slides were immersed in

70%, 90% and 100% ethanol respectively and air-dried. Then the slides were rinsed in xylene and covered with coverslips and mounted with mounting medium. The slides were then sent for pathological examinations.

Table 2.2: Pathological features of the primary breast tumor samples.

Name	Cancer Type	LN	Grade	Stage	Age	ER status	PR status	ErbB2 status	Tumor %
MFT 001	ILC	N		2A	78	P	P	N	70
MFT 007	IDC	N	3	2A	68	N	N	N	90
MFT 011	IDC	P	1	3B	59	P	P	N	90
MFT 014	IC	P		4	76	P	P		90
MFT 016	IDC	N	2	2A	54	P	P	N	100
MFT 021	IDC	P	3	4	71	N	N	N	90
MFT 025	IDC	P	3	3A	51	P	P	N	100
MFT 029	IDC	P	1	2	58	P	P	P	70
MFT 040	IDC	P	3	4	48				100
MFT 041	IDC	P	2	2B	28	N	P	P	100
MFT 049	IDC	P	2	2A	43	P	P	P	100
MFT 059	IDC	N	1	1	34	N	N	P	90
MFT 079	IDC	P	3	2A	24	N	N	P	90
MFT 083	IDC	P	2	2B	47	P	N	P	100
MFT 085	IDC	N	2	2A	45	N	P	P	100
MFT 088	IDC	P	2		60	P	P	N	90
MFT 090	IDC	P	2	2B	32	P	P	P	100
MFT 093	IDC	P	3	2A	42	P	P	P	100
MFT 094	IDC	N	3	2A	37	N	N	N	100
MFT 096	IDC	N	1	2A	39	N	P	N	100
MFT 097	IDC	P	2	3B	30				80
MFT 113	IDC	P	1	3B	43	P	N	P	100
MFT 115	DCIS	N	3	2A					80
MFT 116	IDC	P	1	1	74	N	N	P	90
MFT 117	ILC	P	2	4	30	N	N	P	90
MFT 120	IDC	N	2	2B	50				90
MFT 124	IDC	P	1	2A	57	P	N	P	100
MFT 127	IDC	P	2	2A	30	N	N		100
MFT149	IDC	P	3	2B	59				95
MFT154	IDC	P	3	3B	54	N	N	P	95
MFT155	IDC	N	3	1	57				95
MFT173	IDC	P	3	3B	44	P	P	N	90
MFT174	IDC	P	1	2B	44	N	N	P	90

IDC: Infiltrating ductal carcinoma; ILC: Infiltrating lobular carcinoma; DCIS: Ductal carcinoma in situ; LN: lymph node; P: positive; N: negative; ER: estrogen receptor; PR: progesterone receptor; Tumor %: tumor percentage in pathological sections

Table 2.3: The information of the primary breast tumor samples used in meta-analysis study

Name	Cancer type	LN	Grade	Stage	ER status	PR status	ErbB2 status	Tumor %
MFT 1	IDC	N		2A	P	P	N	80
MFT 14	IDC	P		4	P	P		90
MFT16	IDC	N	2	2A	P	P	N	100
MFT21	IDC	P	3	4	N	N	N	90
MFT41	IDC	P	2	2B	N	P	P	100
MFT49	IDC	P	2	2A	P	P	P	100
MFT93	IDC	P	3	2A	P	P	P	100
MFT94	IDC	N	3	2A	N	N	N	100
MFT97	IDC	P	2	3B				80
MFT113	IDC	P	1	3B	P	N	P	100
MFT116	IDC	P	1	1	N	N	P	90
MFT117	IDC	P	2	4	N	N	P	90
MFT120	IDC	N	2	2B				90
MFT124	IDC	P	1	2A	P	N	P	100
MFT127	IDC	P	2	2A	N	N		100
MFT149	IDC	P	3	2B				95
MFT154	IDC	P	3	3B	N	N	P	95
MFT155	IDC	N	3	1				95

IDC: Infiltrating ductal carcinoma; LN: lymph node; P: positive; N: negative; ER: estrogen receptor; PR: progesterone receptor; Tumor %: tumor percentage in pathological sections

2.6 cDNA SYNTHESIS

First-strand cDNA synthesis was carried out using the Revert Aid First strand cDNA synthesis kit (MBI Fermentas, Ontario, Canada). 1 μg of total RNA and 1 μl oligo(dT) primers or 1 μl random hexamer primers were incubated at 70°C for 5 min in a total volume of 12 μl and chilled on ice. Then, 4 μl of 5X First Strand Buffer, 1 μl of RNase inhibitor and 2 μl of deoxynucleotide triphosphate mix (10 mM) were added and the reaction was incubated at 37°C for 5 min. Finally, the mixture was incubated at 42^{o}C for 1 hr with 1 μl of reverse transciptase enzyme. The reaction was stopped by heating the mixture for 10 min at 70^{o}C. Each cDNA sample was diluted at a ratio of 1:5 with ddH$_2$O and stored at -20^{o}C to be used as a PCR template for further experiments.

The oligo(dT) primed cDNA samples were used for the analysis of all the target and reference genes included in this study. The random hexamer primed cDNA samples were used only for the gene expression analysis of 18S rRNA, ACTH, SDHA, and TBP.

2.7 Amplification of DNA by polymerase chain reaction (PCR)

Each primer pair was first optimized for the amplification conditions and validated for its specificity with a regular thermocycler then the optimal primers, which were primer dimmer and nonspecific binding free were subjected to real-time RT-PCR.

2.7.1 Primers

The gene specific primers used in RT-PCR and real-time qRT-PCR experiments were designed by Primer 3 [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi] and primer.exe programs and purchased from Iontek Inc. (Bursa, Turkey). The primers were designed to include large intronic sequences between the forward and reverse pair or designed from exon-exon boundaries to avoid DNA contamination if any remained in the RNA samples. The sequences of the gene-specific primers were put into the blast search to determine their specificities. None of the primer pairs showed significant homology to other sequences in the genome but their own. The primer sequences and accession numbers of BRCA1 and BRCA1 induced genes are listed in Table 2.4, the primer sequences and accession numbers of genes used for reference gene analysis are listed in Table 2.5 and the primer sequences and accession numbers of genes used for meta-analysis are listed in Table 2.6.

Table 2.4: List of gene-specific primers used for expression analysis, accession numbers and amplicon sizes.

Gene Symbol	**Accession No/** **Primer sequence [5'-3']**	**Amplicon Size (bp)**
BRCA1 Forward Reverse	U14680 acagctgtgtggtgcttctgtg cattgtcctctgtccaggcatc	107
ERBIN Forward Reverse	NM_018695 ctaatcagattgaagagcttcc aactcctgtattccattcttgc	122
OVCA1 Forward Reverse	U34880 gaggccgttgtgtatcttgg catgcgctggtggtcatagt	135
OVCA2 Forward Reverse	NM_080822 atcagacttcgggtcctgcc ccagcctgttcagtgcctgt	154
SMG1 Forward Reverse	AY014957.1 taactcagtggctcaacaggct ctggagatgcagcaatcaacac	101
RENT2 Forward Reverse	NM_080599 cgaagaaaaggtgaaggcag aaatgatgtcgttcccaagc	192
XRN2 Forward Reverse	NM_012255 catcatagtcaactgcgtgg gtcttcaggatgagtacagg	153
RAD21 Forward Reverse	NM_006265 accaatgccaaccatgactg cttcctcttcctcttggctt	114
MAC30 Forward Reverse	BC091504 tcctcaaaggaagctgcaag gggggcatagacagacacaa	184

[a] PCR efficiencies were calculated according to Rasmussen R, (2001).

Table 2.5: List of gene-specific primers used for reference gene analysis, accession numbers and amplicon sizes.

Gene symbol	Accession No/	Amplicon
ACTB	NM_001101	124
Forward	ccaaccgcgagaagatgacc	
Reverse	ggagtccatcacgatgccag	
GAPD	NM_002046	143
Forward	ggctgagaacgggaagcttgtcat	
Reverse	cagccttctccatggtggtgaaga	
TBP	NM_003194	132
Forward	tgcacaggagccaagagtgaa	
Reverse	cacatcacagctccccacca	
SDHA	NM_004168	86
Forward	tgggaacaagagggcatctg	
Reverse	ccaccactgcatcaaattcatg	
HPRT	NM_000194	112
Forward	gctgacctgctggattacat	
Reverse	tcccctgttgactggtcatt	
HMBS	NM_000190	64
Forward	ggcaatgcggctgcaa	
Reverse	gggtacccacgcgaatcac	
B2M	NM_004048	132
Forward	atgagtatgcctgccgtgtga	
Reverse	ggcatcttcaaacctccatg	
PPIA	NM_021130	229
Forward	cgtgtgctattagccatggt	
Reverse	ccattatggcgtgtgaagtc	
GUSB	BC014142	157
Forward	caccagcgtggagcaagaca	
Reverse	ggctgacacctggcacctta	
YWHAZ	NM_003406	193
Forward	aagacggaaggtgctgagaa	
Reverse	acctcagccaagtaacggta	

PGK1 Forward Reverse	NM_000291 aaccagaggattaaggctgc gcctacacagtccttcaaga	195
GSN Forward Reverse	NM_198252 ttcgagtcggccaccttcct tctgcaccaccacctcgttg	108
RPL41 Forward Reverse	NM_001035267 aagatgaggcagaggtccaa tccagaatgtcacaggtcca	248
PUM1 Forward Reverse	NM_001020658 ttcacagacaccacctcctt ctggagcagcagagatgtat	104
RPLP0 Forward Reverse	NM_053275 tcatccagcaggtgttcgac agacaaggccaggactcgtt	194
MRPL19 Forward Reverse	NM_014763 tcgtgttactacagctgacc atctcgacaccttgtccttc	135
TTC22 Forward Reverse	NM_017904 agtgctgaagtccgaggacc ttgccgaagcagtctagagg	150
IL22RA1 Forward Reverse	NM_021258 ccacttagagctccaggtca tctggcagtgtcttcactcg	177
ZNF224 Forward Reverse	NM_013398 agaacttcaggaacctgctc ggaaggaccactcttgatgt	186
18S rRNA Forward Reverse	NR_003286 aaacggctaccacatccaag cctccaatggatcctcgtta	154

Table 2.6: List of gene-specific primers used for resampling based meta-analysis, accession numbers and amplicon sizes.

Gene symbol	**Accession No** **primer sequence [5'-3']**	**Amplicon Size (bp)**
RAD21	NM_006265	114
Forward	accaatgccaaccatgactg	
Reverse	cttcctcttcctcttggctt	
GSN	NM_198252	108
Forward	ttcgagtcggccaccttcct	
Reverse	tctgcaccaccacctcgttg	
COX6C	NM_004374	138
Forward	tcaggaaggacgttggtgtt	
Reverse	gcacgaatgctacagccata	
MAF	NM_005360	228
Forward	tggagtcggagaagaaccag	
Reverse	gcttccaaaatgtggcgtatt	
SFRP1	NM_003012	163
Forward	ccgagatgcttaagtgtgac	
Reverse	ctcgctggcacagagatgtt	
SPTBN1	NM_003128	173
Forward	ggatcacagacctgtacact	
Reverse	tctcaagatggactctctgc	
GSPT1	NM_002094	157
Forward	cacctgtggaatcctctcaa	
Reverse	cctggctctgcttcacttat	
NME1	NM_198175	173
Forward	tgtgagcgtaccttcattgc	
Reverse	aagaatggacggtccttcag	
PTTG1	NM_004219	127
Forward	cctcagatgatgcctatcca	
Reverse	atcatgagaggcactccact	

FN1 Forward Reverse	NM_212476.1 gcaagaggcaggctcagcaa gcggacctacctaggcaatg	112
ID4 Forward Reverse	NM_001546.2 tcctgcagcacgttatcgac tctctagtgctcctggctc	263
EGFR Forward Reverse	NM_005228.3 gcaagaggcaggctcagcaa gcggacctacctaggcaatg	227
ADAMTS1 Forward Reverse	NM_006988.3 ggctgatgttggaactgtgt acacgtggcctaattcatgg	104
ATF3 Forward Reverse	NM_001040619.1 gcactccgtcttctccttct agaacaagcacctctgccac	191
IGFBP6 Forward Reverse	NM_002178.2 attgtgaccatcgaggcttc aggagcttccattgccatct	132
PRNP Forward Reverse	NM_000311.3 gcgagcttctcctctcctca gtgttccatcctccaggctt	140

2.7.2 Reverse transcription polymerase chain reaction (RT-PCR)

Polymerase chain reaction was performed to amplify the desired DNA fragments from cDNAs using the thermal cycler TechGene (Techne Inc., New Jersey, USA). Gene specific primers were first controlled with Polymerase Chain Reaction (PCR) to determine their optimal working conditions and then used for the real-time qRT-PCR experiments. A reaction mixture of 2.5µl 10X reaction buffer, 1.5µl $MgCl_2$ (25mM), 0.5 µl dNTP (10µM), 1µl of each primer (10 pmol), and 0.2 µl Taq DNA polymerase (5u/µL) was prepared per 1µl cDNA and total volume was adjusted to 25µl with ddH$_2$O. The RNA samples used for cDNA synthesis were also used for (-) RT control (no reverse transcriptase enzyme) reactions. These negative RT-PCR controls were also included in the PCR reactions for each set of primers. No genomic DNA contamination was detected.

The optimized PCR condition for all primer pairs used in this study was as follows:

Initial denaturation	95.0^{o}C	5 min	
Denaturation	95.0^{o}C	30 sec	
Annealing	60.0^{o}C	30 sec	30 cycles
Extension	72.0^{o}C	30 sec	
Final extension	72.0^{o}C	5 min	

2.7.3 Quantitative real time RT-PCR

Real-time qRT-PCR was performed on BioRad iCycler (Bio-Rad, California, USA) using the BioRad iQ^{TM} SYBR Green Supermix. The amplification mixtures contained 1.0 µl of 1:5-diluted cDNA template, 6.25 µl SYBR Green PCR Master Mix Buffer (2X), and 10 pmol of forward and reverse primers in a total volume of 12.5 µl. The cycling conditions were as follows: an initial incubation of 95°C for 5 min and then 45 cycles of 95°C for 30 s and 60°C for 30 s during which the fluorescence data were collected. To verify that the used primer pair produced only a single product, a dissociation protocol was added after thermocycling, determining dissociation of the PCR products from 55°C to 95°C in 80 cycles.

12.5µl mineral oil was added to cover top of the mixture to prevent evaporation.

The amplification reactions were performed in 96 well-PCR plates and the plates were sealed with optical sealing tapes (Bio-Rad, California, USA). All PCR reactions were studied in duplicate. Tumor and matched normal samples were always analyzed in the same run to exclude between-run variations and each sample was studied in duplicate. A no-template control of nuclease-free water was included in each run.

Following amplification, a reaction product melt curve was obtained to provide evidence for a single reaction product. The iCycler iQ Optical System Software (version 3, BioRad Laboratories) was used to determine the melting temperatures of the products. The threshold cycle (Ct) value was calculated as the cycle where the fluorescence of the sample exceeded a threshold level.

2.7.3.1 Amplification efficiency calculations

The PCR amplification efficiencies (E) were evaluated by 10-fold dilution series of cDNAs (1-1:100 000 dilution) for each pair of primers used in this study by using a breast carcinoma cell line cDNA pool (MCF7, MDA-MB-231, T47D, HMEC, MCF12A). The primer amplification efficiencies were also tested with reference genes *ACTB*, *GADPH*, and *SDHA* in breast tumor tissue cDNA pools (n=3) to ensure no inhibitory component was present in the tissue samples. No inhibitory effect was observed in amplification efficiencies (E=2.0). A graph of threshold cycle (Ct) versus relative $\log_{10}$ copy number of the calibration sample from the dilution series was produced and the reaction efficiency was determined for each primer set by using the slope of this graph ($E=10^{(-1/\text{slope})}$) and presented at Table 2.2 and Table 2.3 (Rasmussen, 2001).

For the evaluation of the real-time RT-PCR results $2^{-\Delta\Delta Ct}$ method was corrected according to efficiency method (Pfaffl, 2001) as: $[(E_{\text{target}})^{\Delta\text{CtTarget}\,(control\text{-}sample)} / (E_{\text{ref}})^{\Delta\text{CtReference}\,(control\text{-}sample)}]$ using normal pair samples as control. E value was calculated according to the formula; $E=10^{(-1/\text{slope})}$. In this formula, in the place of “reference” the geometric mean of the Ct values of three reference genes, ACTB, SDHA and TBP, was used. “Control” represented the normal samples while “sample” represented the tumor samples. The formula below shows how to adapt the formula to BRCA1 expression evaluation in tumor and matched normal sample.

$$(E_{\text{BRCA1}})^{\Delta\text{CtBRCA1}\,(Normal1\text{-}Tumor1)} / (E_{(\text{ref})})^{\Delta\text{CtGMref}\,(Normal1\text{-}Tumor1)}$$

GMref: geometric mean of ACTB, TBP and SDHA.

2.8 Bisulfite sequencing

Bisulfite conversion was performed with EpiTect Bisulfite kit (Qiagen, Germany). 1 μg genomic DNA was used for each conversion and the reaction was done according to manufacturer's instructions. The cleaning step of the bisulfite converted DNA was performed on the columns supplied with the kit. The methylation specific primers used for PCR are given in Table 2.7.

Table 2.7: Primers specific to 45S rRNA promoter (methylation specific primers)

Gene symbol	Accession No primer sequence [5'-3']	Amplicon Size (bp)
45SrDNA	gi 337380	395
Forward Reverse	gagtcggagagcgtttttgag catccgaaaaccctctccaa	

2.9 Selection of reference genes for normalization of real time RT-PCR data

Reference gene selection experiments were performed with 23 tumor and matched-normal tissues that was a subset of previously described 32 tumor and matched-normal breast samples (Table 2.1). Fifteen commonly used reference genes (*ACTB*, *GAPD*, *TBP*, *SDHA*, *HPRT*, *HMBS*, *B2M*, *PPIA*, *GUSB*, *YWHAZ*, *PGK1*, *RPL41*, *PUM1*, *RPLP0*, *MRPL19)* and three newly selected genes (*TTC22*, *IL22RA1*, *ZNF22,* see in section 2.7.1) which belong to different functional classes were chosen for stability analysis. The primer sequences of these candidate reference genes are listed in Table 2.5. The software geNorm^{TM}, version 3.4 (Vondesompele *et al*., 2002; see in 2.13.2) and NormFinder (Andersen *et al*., 2004; see in 2.13.3), both Visual Basic Applications (VBA) for Microsoft Excel, were used to calculate the stability of candidate reference genes. Ct values were converted to linear expression quantities by $E^{-\Delta Ct}$ to investigate the genes in geNorm and NormFinder.

Gelsolin (*GSN*) gene, whose expression has been reported to be low in breast tumors, was used to validate the results in the same set of 23 pairs of breast tumor samples. When the *GSN* normalization was based on the multiple reference genes, the geometric mean of reference gene Ct values was applied as a normalization factor (NF).

2.9.1 Data retrieval and selection of candidate reference genes from microarray studies.

Two publicly available independent microarray gene expression data sets GDS2635 (Turashvili *et. al.*) and GDS2250 (Richardson *et. al.*, 2006) were downloaded from the Gene Expression Omnibus [GEO, http://www.ncbi.nlm.nih.gov/geo/] and processed by the BRB-ARRAYTOOLS [Biometric Research Branch [http://linus.nci.nih.gov/BRB-ArrayTools.html]. Both of the datasets were generated by using the Affymetrix HGU133 Plus 2.0 platform; thus they were highly comparable. These two independent microarray datasets (GDS2635 and GDS2250) were combined with respect to gene names using a set of customized Perl routines and the genes that were stably expressed between tumor and normal samples were selected by using Student's t-test (p>0.99). A total number of 12 normal and 45 tumor samples and 54674 gene probes were used in this analysis. *TTC22* was one of the top ranked non-differentially expressed gene between tumor and normal samples (p>0.99) and was selected as a candidate RG.

The GDS2635 dataset is the only available dataset that was generated by using matched-normal breast tumor samples. Therefore this set was used independently and the genes that showed no expression differences between tumors and matched-normal samples were determined by using paired Student t-test (p>0.99).

2.10 Protein extraction

2.10.1 Protein extraction from tissue culture cells

Exponentially growing tissue culture cells were washed with ice-cold 1XPBS three times, scrapped, and collected into Falcon tubes. The cell pellet was collected by centrifugation at 1500 rpm for 5 min at 4°C and lysed in the RIPA buffer (see in General Solutions section) by vortexing every 5 min for half an hour. After centrifugation of the samples at 13000 rpm for 30 min at 4°C, the supernatant was collected and stored at -80°C for future use or immediately loaded onto SDS-PAGE gel.

2.10.2 Protein extraction from tissue samples

The 5 µm thick tissue sections were cut with cryostat and collected in an eppendorf tubes prior to protein extraction. 500 µl of RIPA buffer (see in General Solutions section) was added onto the tissue sections and homogenized with a homogenizer. The homogenized mixture was kept on ice and mixed by vortexing every 5 min for half an hour. After centrifugation of the samples at 13000 rpm for 30 min at 4°C, the supernatant was collected and stored at -80°C for future use or immediately loaded onto SDS-PAGE gel.

2.10.3 Quantification of proteins

After the cell lysates were prepared, their concentrations were detected by Bradford assay. As described in Table 2.7, 2 µl of the samples were diluted with 98 µl deionised water and then 900 µl of Bradford working solution was added to the samples and mixed well. After 5 minutes of incubation, the protein amounts of the samples were measured at OD_{595} nm versus blank reagent. Known concentrations of BSA were prepared according to Table 2.9 as a standard. After reading at OD_{595}, samples and standard values were plotted; unknown concentrations were calculated from the standard curve.

Table 2.8: Protein sample preparation for Bradford assay

Tube no	1	2	3	4	5	6
Sample (µl)	0	2	2	2	2	2
ddH2O (µl)	98	98	98	98	98	98
Bradford (µl)	900	900	900	900	900	900
Lysis buffer	2	-	-	-	-	-

Table 2.9: A standard curve preparation with BSA dilution.

Tube no	1	2	3	4	5	6	7	8
BSA (µl)	0	2.5	5	7.5	10	12.5	15	20
ddH2O (µl)	100	97.5	95	92.5	90	87.5	85	80
Bradford (µl)	900	900	900	900	900	900	900	900

OD_{595} were read after samples were incubated at room temperature.

2.11 Western blotting

2.11.1 SDS polyacrylamide gel electrophoresis

5% polyacrylamide gel (40ml) was prepared as below:

H_2O	23.4 ml
30% Acrylamide mix	6.8 ml
1.5 M Tris, pH: 8.8	10 ml
10% SDS	0.4 ml
0.1% APS	0.4 ml
TEMED	32 µl

5% stacking gel (10ml) was prepared as below:

H_2O	6.8 ml
30% Acrylamide mix	1.7 ml
1 M Tris, pH: 6.8	1.25 ml
10% SDS	0.1 ml
0.1% APS	0.1 ml
TEMED	10 µl

Equal amounts of proteins were denatured in 5X loading buffer at 95^{o}C for five minutes prior to loading on gel. Gels were run at 80 V during stacking gel and 140 V during resolving gel in running buffer. The run was stopped as the unnecessary proteins left the gel, following the protein marker, and the gel was prepared for western blotting.

2.11.2 Protein Transfer to Solid Support

BRCA1 and ERBIN immunoblotting was performed with wet transfer. Polyvinylidene difluoride (PVDF) membrane was washed with methanol for 1 min, then rinsed with water until it sinks in the water and kept in wet transfer buffer. Gel and 3M Whatman papers were also kept in transfer buffer for equilibration for 5-10 min. Two whatman papers (prewet by transfer buffer), membrane, gel and two wet Whatman papers were placed on transblotter. Transfer was performed overnight at 14 V at 4^{o}C. After transfer, the membrane was immediately put into blocking solution and immunological detection method was followed as described in section 2.10.3. Alternatively, the membrane was put into a nylon bag and stored at 4^{o}C and kept dry for future use.

Two different pre-stained protein markers were used to determine the molecular weight of the proteins. The sizes of the protein markers were as follows (kiloDaltons, kDa):

Bio-Rad (Broad Range 161-0318): 203, 120, 90, 51.7, 34.1, 28, 20 and 6.4

BioLabs (Broad Range P7708S): 175, 83, 62, 47.5, 32.5, 25, 16.5 and 6.5

2.11.3 Immunological detection of immobilized proteins

After the transfer, the membrane was immersed in the blocking solution for an hour on a slowly rotating platform to inhibit non-specific binding sites. The time of blocking extended to over night if the primary antibody has more non-specific binding capacity. Each primary antibody was diluted in blocking solution in different concentrations. The anti-BRCA1 antibody MS110 was used in 1:100 while anti-ERBIN was used in 1:400 dilutions. Anti-Calnexin antibody, which was used for equal loading control was diluted by 1:2500. The diluted primary antibody was added to the blocking solution and the membrane was left in primary antibody for one hour at room temperature or for overnight at 4^{o}C on a slowly rotating platform. The membrane was washed with TBS-T three times, 10 min for each wash. The appropriate secondary antibody (HRP-conjugated) was diluted in blocking solution as recommended by the supplier (1:5000). The membrane was left in secondary antibody for 1 hour at room temperature and then washed three times with TBS-T for half an hour. The membrane was treated with ECL kit reagents according to the manufacturer's instructions and then wrapped with stretch film. The autoradiography was carried out for various exposure times.

2.12 Bioinformatic analysis

2.12.1 Data retrieval for Meta-anlaysis

The extensive use of DNA microarray technology in the characterization of the cell transcriptome amounts to an ever-increasing amount of microarray data from cancer studies.

Different data sets for the same type of cancers are available from different microarray studies and this allows the researchers to carry out a more comprehensive analysis of their existing data set. These studies can be obtained from various public gene expression data repositories including the Stanford Microarray Database (SMD) (Sherlock *et al.*, 2001), the National Cancer Institute’s Gene Expression Omnibus (GEO) (Barrett *et al.* 2005) and Oncomine (Rhodes *et al*, 2004b). These databases enable researchers to retrieve and perform analyses on various microarray experiments from different laboratories.

2.12.1.1 Data retrieval from Stanford Microarray Database (SMD)

Two independent microarray gene expression data sets, Sorlie *et al*. (Sorlie *et al,* 2003) and Zhao *et al*. (Zhao *et al*., 2004), were downloaded from the Stanford Microarray Database (SMD); http://genome-www5.stanford.edu/. Gene filtering options of SMD were used for log transformation and median centering the data arraywise. Expression values that were missing in more than 20% of the data were excluded from the analysis. Details of tumor specimen histology, available on SMD, were used to restructure the experiments according to breast tumor subtypes as invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC) and normal samples. Since the genes on the arrays were represented with more than one probe, probe IDs were used instead of gene names and the two datasets were combined with respect to probe IDs using a set of customized perl routines (source codes are given in section 2.10). These two data sets combined resulted in an initial list of 4769 IMAGE clones (3465 unique genes) common in both datasets. A total of 139 IDC (38 samples Zhao, 101 samples Sorlie datasets), 29 (21 samples Zhao, 8 samples Sorlie datasets) ILC and 7 (3 samples Zhao, 4 samples Sorlie datasets) normal samples were available for further analysis. The pathological data of the two datasets were given in Table 2.10.

Table 2.10: Pathology information of tumor samples used in Sorlie *et al.*, 2003 and Zhao *et al.*, 2004 studies.

	ER Status		Grade Status			Pathological Subtype		
	ER(+)	ER(-)	1	2	3	IDC	ILC	N
Sorlie	81	32	11	48	54	101	8	4
Zhao	40	14	7	40	12	38	21	3

ER: Estrogen receptor; IDC: Infiltrating ductal carcinoma; ILC: Infiltrating lobular carcinoma; N: normal tissue

2.12.2 Data sets used in resampling-based meta-analysis study

Two independent microarray gene expression data sets, Sorlie *et al*. (Sorlie *et al,* 2003) and Zhao *et al*. (Zhao *et al*., 2004), were used in the study.

2.12.2.1 Sorlie *et al.* dataset

Sorlie *et al*. dataset contains a total of 109 breast carcinomas (101 ductal and 8 lobular) from different individuals and four normal breast tissue samples, three of which were pooled normal breast samples from multiple individuals (CLONTECH) were included. The authors used freshly frozen breast tissue samples and the tumor specimens, which contained more than 50% tumor cells were included in the study. In this study, the 8102 human cDNA genes/clones were used. They were obtained from Research Genetics (Huntsville AB, USA) and chosen from a set of 15,000 cDNA clones that corresponded to the Research Genetics Human Gene Filters sets GF200-202 (http://www.resgen.com/). This set of genes contained some redundancy (approximately 300 genes were printed more than once on each array) and contained approximately 4000 named genes, 2000 genes with homology to named genes in other species, and approximately 2000 ESTs of unknown function.

Each of the experimental samples tested was analyzed by a comparative hybridization, using a common "reference" mRNA pool as a standard composed of equal mixtures of mRNA isolated from 11 established human cell lines (MCF7, Hs578T, OVCAR3, HepG2, NTERA2, MOLT4, RPMI-8226, NB4+ATRA, UACC-62, SW872, and Colo205) and labeled with Cy3. The raw data of this study is publicly available on Stanford Microarray Database (SMD); http://genome-www5.stanford.edu/

2.12.2.2 Zhao *et al.* dataset

The authors used 59 primary breast cancer cases in the Zhao dataset. Of these 38 were IDC and 21 were ILC. They also used 3 normal samples in their study and these

samples were from 3 different IDC patients. The primary breast samples used in the study were frozen in either liquid nitrogen or on dry ice just after devascularization and stored at -80^{o}C; hence the RNA samples were obtained from freshly frozen breast samples. The pathological details of the tumors are given in Table 2.9.
The data were generated by amplification of the RNA samples and hybridization of the samples to cDNA microarray containing 42,000 clones. The amplified total RNA was labelled by Cy5 and the amplified RNA from Universal Human Reference total RNA was labelled by Cy3. The raw data of this study is publicly available on Stanford Microarray Database (SMD); http://genome-www5.stanford.edu/

2.12.3 Data Filtering

Data were filtered separately for ductal and lobular samples. IMAGE clones with more than 50% missing data in either of the Sorlie or Zhao datasets were excluded from the common clone set. Data filtering was further improved by performing two-tailed Student's t-tests with equal variance (Matlab©) between the Sorlie and Zhao datasets for the IDC and ILC samples separately. Those clones with probability values less than 0.05 (after Bonferroni correction), which show differential expression among the IDC samples or ILC samples, were excluded from further analysis. This two-step data filtering has resulted in a common set of 1726 IMAGE clones for the analysis of ductal and normal samples, and 2029 IMAGE clones for the analysis of lobular and normal samples. Upon taking the intersection of the ductal-normal and lobular-normal clone sets, 1522 IMAGE clones were available for the ductal-lobular analysis. The resulting clone subsets were further filtered by removing IMAGE clones with more than 40% missing data for the two groups in comparison (e.g., ductal and normal) in the combined data before application of the resampling steps. In addition, if an IMAGE clone had a sample size (of normal samples) less than the resampling sample size, data on this IMAGE clone was also removed.

2.12.4 Resampling technique

Many studies of gene expression seek genes that are “differentially expressed”, showing changes in expression levels associated with an experimental variable or biological condition, i.e, cancer vs. normal. The variance of the number of differentially expressed genes depends on the chosen statistical tests, the method of multiple testing adjustments and the structure of the data. Stability is an important issue in the selection of differentially expressed genes. Resampling (drawing repeated samples from the given data, or population suggested by the data) is a proven cure for this issue and is now the method of choice for confidence limits, hypothesis tests, and other everyday inferential problems. In statistics, one can estimate, via resampling, the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknife) or drawing randomly with replacement from a set of data points (bootstrapping).

Jackknifing, which is similar to bootstrapping, is used in statistical inferencing to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it. The basic idea behind the jackknife estimator lies in systematically recomputing the statistic estimate leaving out one observation at a time from the sample set. From this new set of "observations" for the statistic an estimate for the bias and an estimate for the variance of the statistic can be calculated.

In this study a resampling technique that resembles delete-d-jackknife method, where d refers to the sample size used for resampling, was used to find out the most stable differentially expressed genes between IDC and ILC and normal samples. In the resampling method used in our study, the chosen samples were never leaved out but put back to the pool for further samplings (like bootstrapping).

2.12.4.1 Resampling and statistical analysis

We have used a resampling method for meta-analysis of microarray data in which the significance of the difference between group medians (e.g. ductal vs lobular) could be tested upon a series of resampling schemes from the original and multiple randomly shuffled datasets (Figure 2.3; code written in Matlab© using Statistics Toolbox is available in Appendix A). The original data referred to the combined data of Sorlie and Zhao while the shuffled data referred to a generated data which was randomly shuffled for multiple times to generate a random data set to test whether the differentially expressed gene sets were obtained by a coincidence. Accordingly, a preset number of samples were selected from each group (i.e., IDC, ILC, normal) of the original dataset, referred herein as the *test*. The p-value was calculated indicating the significance of the difference between the group medians based on the Wilcoxon Rank Sum Test (see in Section 2.13.9). This test was repeated for a series of *i* number of iterations; at the end of each iteration scheme, a set of *p-values (pt)* per IMAGE clone was obtained. The above procedure was also applied to each of the shuffled datasets yielding *pr1* and *pr2*. P-value distributions were then tested in a pair-wise fashion (i.e., *pt* vs. *pr1*; and *pr1* vs. *pr2*) using the two-sample Kolmogorov-Smirnov test (see in Section 2.13.8) for each clone in the dataset (Figure 2.1). The resulting p-values were named as *kst* and *ksr*, respectively. To obtain an estimate of the false discovery rate (FDR), *ksr* values were sorted in the ascending order and the k^{th} value from the top (lowest p-value) was determined as $\mathrm{FDR}_{\mathrm{observed}}$, where *k* equals the expected value of FDR (e.g., 0.01) multiplied by the number of IMAGE clones tested. $\mathrm{FDR}_{\mathrm{observed}}$ was set as the threshold according to which IMAGE clones were assigned as significant or not. If *kst* of a particular gene had a value that was smaller than the $\mathrm{FDR}_{\mathrm{observed}}$, the gene was accepted to be significant.

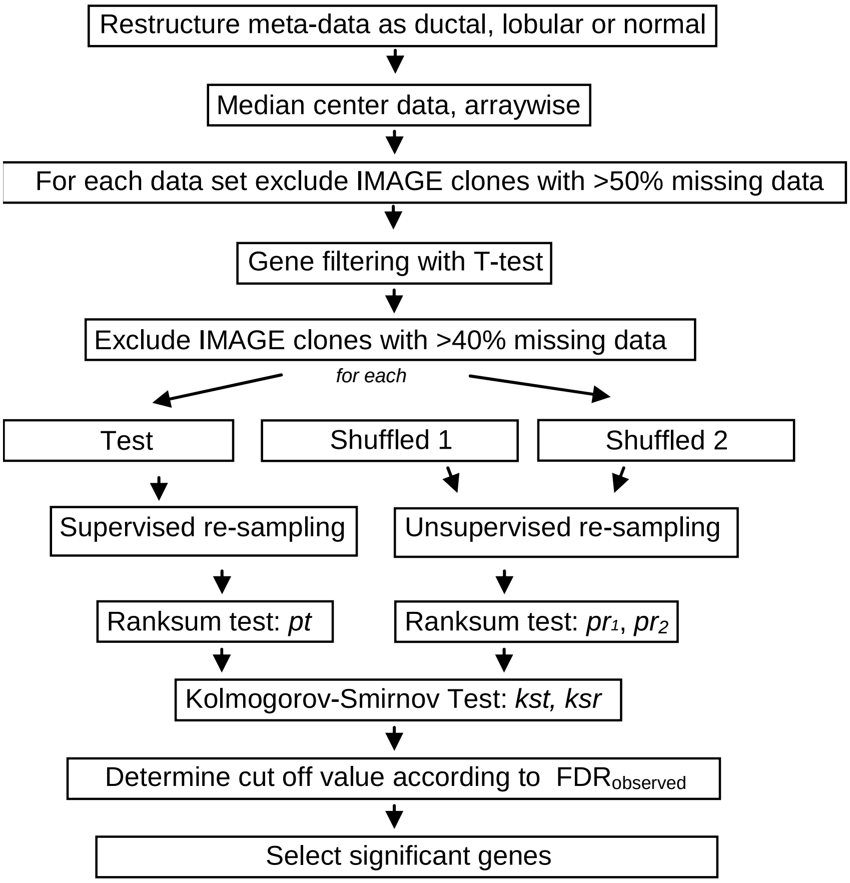

Restructure meta-data as ductal, lobular or normal
Median center data, arraywise
For each data set exclude IMAGE clones with >50% missing data
Gene filtering with T-test
Exclude IMAGE clones with >40% missing data
for each
Test
Shuffled 1
Shuffled 2
Supervised re-sampling
Unsupervised re-sampling
Ranksum test: pt
Ranksum test: pr1, pr2
Kolmogorov-Smirnov Test: kst, ksr
Determine cut off value according to FDRobserved
Select significant genes

Figure 2.3 General meta-analysis flowchart. Workflow is represented by boxes and arrows.

2.12.4.2 Application of resampling to the breast cancer datasets

The above tasks were performed for a particular sample size *n* (e.g., 3), repetitively for *i* number of times, where *i* = 10, 20, 30, …, 100 and 150. For each particular *i*, three parameters were recorded, namely, *kst* values, the mean expression value of each of the two groups compared, and the significance of the differential expression based on *kst* and *ksr*. These above steps were then repeated with different sample sizes: For ductal vs. lobular comparison, *n* was set to be 3, 4, 5, 6, 10, 15 and 20. On the other hand, since the total number of normal samples was 7, the highest sampling value could be set to 6 for ductal vs. normal and lobular vs. normal comparisons, and *n* equaled 3, 4, 5 and 6. These sample size-iteration combinations led to 77 runs for ductal vs. lobular analysis, and 44 runs for ductal vs. normal and lobular vs. normal analyses. At the end, a final differentially expressed gene set was determined for each of the three comparisons (i.e., ductal vs. lobular, DL; ductal vs. normal, DN; lobular vs. normal, LN) by gathering the IMAGE clones that were assigned as significant in 90% or more of these 44 or 77 runs. The mean values of each of the two groups in comparison obtained at n = 20 (or 6, in the case of normal vs. tumor comparisons) and *i* = 150 were used as an estimate of the measure of expression.

2.12.5 Data retrieval and prediction analyses via BRB-Arraytools

In order to assess what proportion of the tumor/normal gene set (ductal/normal and lobular/normal gene sets, separately) was informative in comparison to molecular subtypes of breast cancer such as luminal and basal as well as ER+ and ER- samples, the predictive ability of the meta-gene lists was tested in independent microarray datasets by using the BRB-ARRAYTOOLS (http://linus.nci.nih.gov/BRB-ArrayTools.html). Additionally to investigate if the BRCA1 target genes can predict the ER or grade status of the tumors and if they were informative enough to group the basal tumors, the predictive ability of the BRCA1-target candidate genes were tested by using the BRB-ARRAYTOOLS.

The ".cel files" of the publicly available independent microarray gene expression data sets, GDS2635 (Turashvili *et al.*, 2007), GDS2250 (Richardson *et al.*, 2006) GDS1329 (Farmer *et al.*, 2005), GSE8977 (Karnoub *et al.*, 2007), GSE2990 (Sotiriou *et al.*, 2006) and GSE2034 (Wang *et al.*, 2005) were downloaded from Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and processed by the BRB-ARRAYTOOLS [(http://linus.nci.nih.gov/BRB-ArrayTools.html). All the datasets were obtained using Affymetrix HGU133A or HGU133 Plus 2.0 platform; thus they were highly comparable. The GDS2635 was used in order to identify gene expression profiles of microdissected ductal and lobular carcinomas in relation to their normal ductal and lobular cells (n= 10). The authors identified multiple genes differentially expressed in comparisons between ductal and lobular tumor and their matched-normal cells (Turashvili *et al.*, 2007). In the GDS2250

study, Richardson *et al.* performed gene expression array-based analysis of three breast tumor subtypes, i.e., sporadic basal-like cancer (BLC), BRCA-associated breast cancer, and non-BLC. They used 47 human breast tumor cases to provide insight into the molecular pathogenesis of BLC and BRCA1-associated breast cancer and the contribution of X chromosome abnormalities to the pathogenesis of BLC (Richardson *et al.*, 2006). In GDS1329, Farmer *et al*. performed the analysis of tumors from 49 breast cancer patients that were successfully classified into luminal and basal classes, and a novel molecular apocrine class. Apocrine tumors were estrogen receptor negative ER(-) and androgen receptor positive AR(+), while luminal tumors were ER(+) and AR(+), and basal tumors were ER(-) and AR(-). Sotiriou *et al.* (GSE2990) analyzed microarray data from 189 invasive breast carcinomas and identified 97 differentially expressed genes in a training set of 64 estrogen receptor (ER)-positive tumor samples by comparing expression profiles between histologic grade 3 tumors and histologic grade 1 tumors and used the expression of these genes to define the gene expression grade index. Data from 597 independent tumors were used to evaluate the association between relapse-free survival and the gene expression grade index in a Kaplan-Meier analysis.
Details of the breast specimens, (normal-tumor, non-basal like- basal like, basal-luminal, ER (+)/ER (-) and grade 1, 2 and 3) available from GEO database were used in the supervised class prediction with binary tree algorithm.
Among the independent microarray datsets, GDS2635, GDS2250, GSE8977 and GDS1329 were used for the analysis of meta-analysis genes while GSE2990, GSE2034 and GDS2250 were used to test the BRCA1-target genes.

2.12.6 Promoter analysis

In order to investigate if any common regulatory sequences were present at the promoter regions of the BRCA1 target genes MEME Suite Motif-based sequence analysis tools was used (http://meme.sdsc.edu/meme4_1/intro.html). For ach gene, 1000 bp upstream of the transcription start site (TSS) was downloaded from UCSC genome browser, human march 2006 assembly in FASTA format. These sequences were used as *input* to be analyzed by MEME Suite motif-based sequence analysis tool.

MEME Suite motif-based sequence analysis tool was used to search for any common motifs among the regulatory regions of the selected genes and MAST was used as a tool of MEME for the motif alignment. MAST takes as *input* a file containing the descriptions of one or more motifs and searches a sequence database that you select for sequences that match the motifs. The motif file can be the output of the MEME motif discovery tool or any file in the appropriate format.

Some parameters were considered for MEME and they were as fallows; the motif width was from 6 to 12 bases, maximum number of motifs to search was 100 and reverse complementary strand was considered. Parameters used for MAST were all significant motifs (with motif E value less than 1.0) was returned by MEME, motif was reported if its sequence p value was less than 0.005, correlated motifs were filtered out, both strands were searched and individual sequence compositions was used to calculate p and E values.

2.13 Statistical analysis

2.13.1 Cluster Analysis

Cluster Analysis makes grouping of the objects according to degree of association between the objects. If the degree of association is high between the two objects, they are put in the same group, and if the association is low, they do not end up the same group. We used a hierarchical clustering algorithm in which each case has a separate cluster at the beginning, and then clusters are combined according to their degree of association. This hierarchical clustering process can be represented as a tree, or dendrogram. Average linkage method of hierarchical cluster was chosen for clustering; this algorithm takes into account the average distance between the objects. The clustering method represents the relationships of genes as a tree structure by connecting genes using their similarity scores based on the Pearson correlation coefficient. Both gene and sample based cluster analyses were performed.

Cluster analysis and imaging were carried out by Eisen Lab Cluster Version 2.11 and Treeview Version 1.60 softwares (http://rana.lbl.gov/EisenSoftware.htm).

2.13.2 GeNorm

The geNorm program determines the most stable reference genes from a set of investigated genes in a given set of samples. It calculates the gene expression stability measure (*M*) for a reference gene considering the average pair-wise variation of all other tested reference genes (Vandesopele *et al.*, 2002). The program requires the Ct values to be converted to linear expression quantities by the $E^{-\Delta Ct}$ method using the highest expression level as calibrator. The lowest M value marks the gene(s) with the most stable expression.

In addition to the stability value M, pair-wise variations (V*n/n*+1) are calculated to determine the effect of adding a gene (*n*+1) in normalization. This allows for determination of the optimal number of reference genes required for reliable normalization. A large pair-wise variation value, which is bigger than the cut-off value 0.15 means that the added gene has a significant effect on normalization hence should be included for calculation of reliable normalization.

2.13.3 Normfinder

NormFinder is an add-in for Microsoft Excel and is used for calculating a stability value from a set of candidate reference genes. In this program, the stability value is based on the combined estimate of inter- and intra-group expression variations of the studied gene. The candidate gene with the smallest variability value has higher stability as it shows the lowest variability of inter- and intra-group expression (Anderson *et al.*, 2004). NormFinder also ranks the set of candidate reference genes according to their expression stability from a panel of candidate genes that could be organized in different subgroups (i.e. tumor and matched non-tumor tissues).

2.13.4 Pearson Correlation

Pearson correlation reflects the degree of linear relationship between the two variables (Rodgers and Nicewander, 1988). It ranges from +1 to -1. +1 reflects the perfect positive linear relationship and -1 represents the negative linear relationship between the two variables. Pearson correlation was carried out in Minitab™. The cut off p value used in the Pearson correlation test was 0.05. Bonferroni correction was performed when multiple tests were applied.

2.13.5 Discriminant Function Analysis

Discriminant function analysis shows how independent variables contribute to the categorical dependent variable (Anderson, 1984). All variables were assumed to have the same covariance matrix, thus, linear discriminant analysis was performed. Discriminant analysis was performed in Minitab™ for several clinical properties of the samples used in the analysis.

2.13.6 Mann Whitney Test

The Mann Whitney test is used to identify if the two populations are different from each other (Mann and Whitney, 1947) In the two-tailed test, null hypothesis is that median of the two populations are equal. If the p value is less than the chosen α level there is sufficient evidence to reject null hypothesis. The Mann Whitney test was performed in Minitab™ for ER status of primary breast tumor samples used in this study with a 95% confidence interval for each gene.

2.13.7 Student's t-test

A t-test compares the difference between two means of different groups to determine whether that difference is statistically significant. There are three types of t-tests: one-sample, independent-samples, and paired-samples.

The t-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups.

2.13.8 Kolmogorov-Smirnov test

In statistics, the Kolmogorov-Smirnov test (often called K-S test or the vodka test) is used to determine whether two underlying probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples.

2.13.9 Wilcoxon Rank Sum Test

Wilcoxon rank-sum test is a non-parametric test for assessing whether two samples of observations come from the same distribution (Wilcoxon, 1945). The null hypothesis is that the two samples are drawn from a single population, and therefore that their probability distributions are equal. It requires the two samples to be independent, and the observations to be ordinal or continuous measurements, i.e. one can at least say, of any two observations, which is the greater. In a less general formulation, the Wilcoxon rank-sum test may be thought of as testing the null hypothesis that the probability of an observation from one population exceeding an observation from the second population is 0.5.

CHAPTER 3. RESULTS

3.1 Ectopic expression of BRCA1 in tissue culture cells

The MCF7 breast carcinoma cells contain one copy of the BRCA1 gene and display low expression of this gene. In a previous study, overexpression of BRCA1 was achieved by using two different approaches to study the target genes regulated by BRCA1 (Atalay *et al*., 2002). The MCF7 cells were transfected with a vector containing the full length BRCA1 cDNA and also with the cells that stably express BRCA1 gene at a modest level in an inducible manner.

In this study, our aim was to study whether the overexpression of BRCA1 has any effect on the expression levels of selected target genes. Therefore, both BRCA1 overexpression approaches were applied to achieve this aim.

3.1.1 Ectopic expression of BRCA1 in MCF7 cells

In order to achieve the high level expression of BRCA1, the MCF7 cells were transiently transfected with full length BRCA1-transcript containing plasmid pCMVmycBRCA1 or empty control plasmid pCMVmyc. Transfection efficiency of the MCF7 cells was determined by transfecting the MCF7 cells with pEGFP-N2 reporter plasmid under the same experimental conditions. All the cells were collected 24 h after transfection. The transfection efficiency was calculated by trypsinizing and counting the green fluorescent cells with a hematocytometer under 525 nm wavelength fluorescence light and found to be approximately 70%. BRCA1 and control plasmid-transfected cells were used to determine the expression level of BRCA1 at protein and mRNA levels by Western blot and real-time quantitative RT-PCR (qRT-PCR) analysis respectively.

The Western blot result of transfection experiments are shown in Figure 3.1. Transient transfection resulted in a nearly 2.5 folds of increase in BRCA1 protein level in pCMVmyc-BRCA1 transfected cells (Figure 3.1, lane 1) compared to MCF7 cells transfected with control plasmid pCMVmyc. (Figure 3.1, lane 2). The blot was also hybridized with calnexin antibody as a control for determining the equal loading of the proteins.

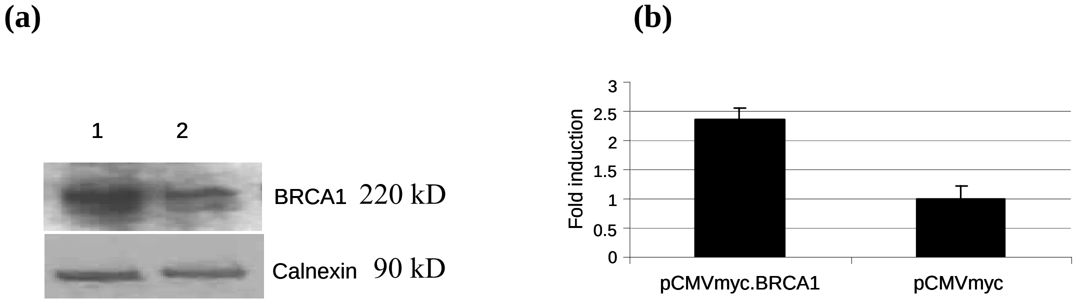


Figure 3.1: Western blot analysis of ectopic expression of BRCA1 in MCF7 breast carcinoma cells. (a) pCMVmycBRCA1 plasmid (lane 1) and pCMVmyc vector transfected cell lysates (lane 2). Transfection of the cells was performed by FuGENE kit. Total protein extracts were analyzed for BRCA1 protein level by using BRCA1 specific monoclonal AB-1 antibody. Anti-calnexin antibody (Santa Cruz, USA) was used for equal loading control of the proteins in each well (bottom panel). **(b)** Transient transfection resulted in a nearly 2.5 fold of increase in BRCA1 protein level in pCMVmyc-BRCA1 transfected cells compared to MCF7 cells transfected with control plasmid pCMVmyc.

3.1.2 Induction of BRCA1 Expression in UBR60-bcl2 cells

UBR60-bcl2 cell line was derived from the U2OS osteosarcoma cell line which expresses BRCA1 under the control of tetracycline-regulated promoter (Harkin *et al*, 1999). This cell line was also used to study the expression levels of selected target genes since it stably expresses BRCA1 gene at modest level in an inducible manner (Figure 3.2). Lane 2 shows that low levels of endogenous BRCA1 protein was detectable in the presence of tetracycline and drug withdrawal led to the 2.5 -fold induction of BRCA1 protein (Figure 3.2, lane 1).

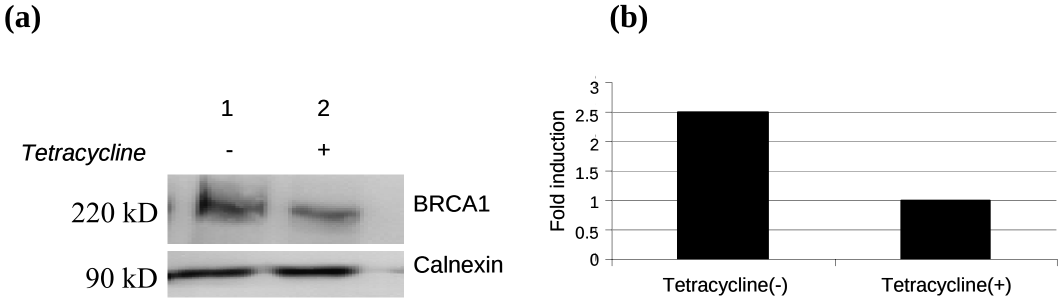


Figure 3.2: Western blot analysis of BRCA1 expression in tetracycline inducible UBR60-bcl2 cells. (a) Lane 1 shows the BRCA1 level in UBR60-bcl2 cells cultured without tetracycline and lane2 shows the cells grown in tetracycline. **(b)** Tetracycline induction resulted in a 2.5 fold of increase in BRCA1 protein level.

BRCA1 protein level was analyzed by using total protein extracts from UBR60 cells grown in the absence (lane 1) or presence (lane 2) of tetracyline. The blot was labeled with BRCA1 specific monoclonal antibody AB-1 (top panel) and with Calnexin specific monoclonal antibody for equal loading control (bottom panel).

3.1.3 Control for upregulated BRCA1 activity

It was previously shown that overexpression of BRCA1 induced the DNA damage-responsive gene GADD45 (Harkin *et al.*, 1999). The BRCA1 activation of the GADD45 promoter is mediated through the OCT-1 and CAAT motifs located at the GADD45 promoter region. This activation requires normal-transcription activating function of the BRCA1 since the BRCA1 mutants lacking the transcription activity were unable to activate the GADD45 promoter (Fan *et al.*, 2002). To control the activity of BRCA1 in BRCA1 induced systems, both in pCMVmyc-BRCA1 transfected and in Tet (-) UBR60 cells, GADD45 gene was used as a positive control. GADD45 expression was shown to be induced 3.5 fold in Tet- cells compared to Tet+ cells while the pCMVmyc-BRCA1 transfected cells had 2.5 fold more *GADD45* expression compared to empty vector transfected MCF7 cells (Figure 3.3).

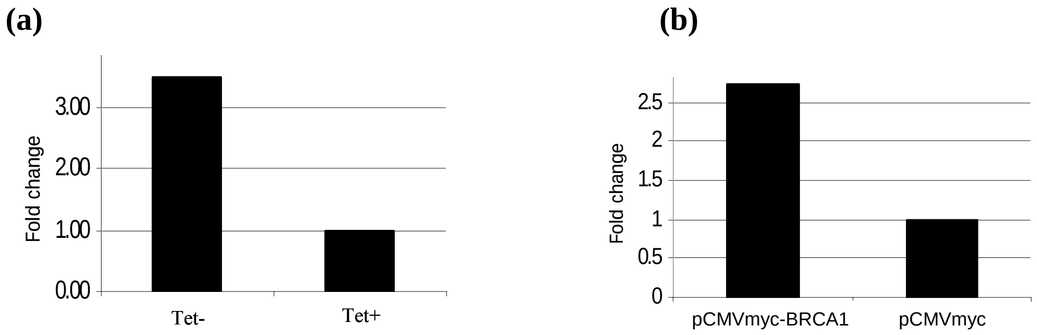


Figure 3.3: Expression profile of GADD45 in BRCA1 induced cells. (a) The expression level of GADD45 in Tetracycline negative (Tet-) and Tetracycline positive (Tet+) UBR60-bcl2 cells. **(b)**The expression level of *GADD45* in pCMVmyc-BRCA1 and pCMVmyc transfected MCF7 cells.

3.2 Quantitative real-time RT-PCR

Quantitative real-time RT-PCR (qRT-PCR) is one of the most sensitive and specific methods for quantification of expression at the mRNA level. To have the efficient amplification reactions and to get reliable results there are some considerations that have to be evaluated before, during and after the amplification step. Before starting the experiments the most important criteria is the selection of the primer pairs. The primer pairs yielding product sizes smaller than 150 bps as well as free of dimers and nonspecific bindings are always recommended to obtain higher efficiencies in amplification reactions. Each primer pair is tested for the amplification efficiencies by generating standard curves. For the analysis of the specificity of the reaction, dissociation curve analysis is run just after each amplification reaction.

3.2.1 Melt curve analysis

The real-time RT-PCR reaction uses SYBR Green I as a detection system. SYBR Green I is a DNA intercalating agent which binds to double stranded DNA during the product amplification at qRT-PCR.
The melt curve analysis is an important step to show the specificity of the qRT-PCR reaction. In a typical melt curve result, we expect to see a single peak for each sample at the same dissociation temperature. Any extra peak means that nonspecific amplicons or primer dimers occurred in the reaction. Therefore, in each qRT-PCR reaction, melt curve analysis was carried out to ensure the specificity of the reaction. Figure 3.4a shows a typical melt curve result, and Figure 3.4b shows a melt curve graph that contains non-specific amplicon and primer dimer in the reaction.

In this study, the PCR conditions for each gene specific primer pairs were optimized to give the desired melt curve with qRT-PCR. If the desired melt curve was not achieved, that primer pair was discarded and new primer pairs were designed for that gene.

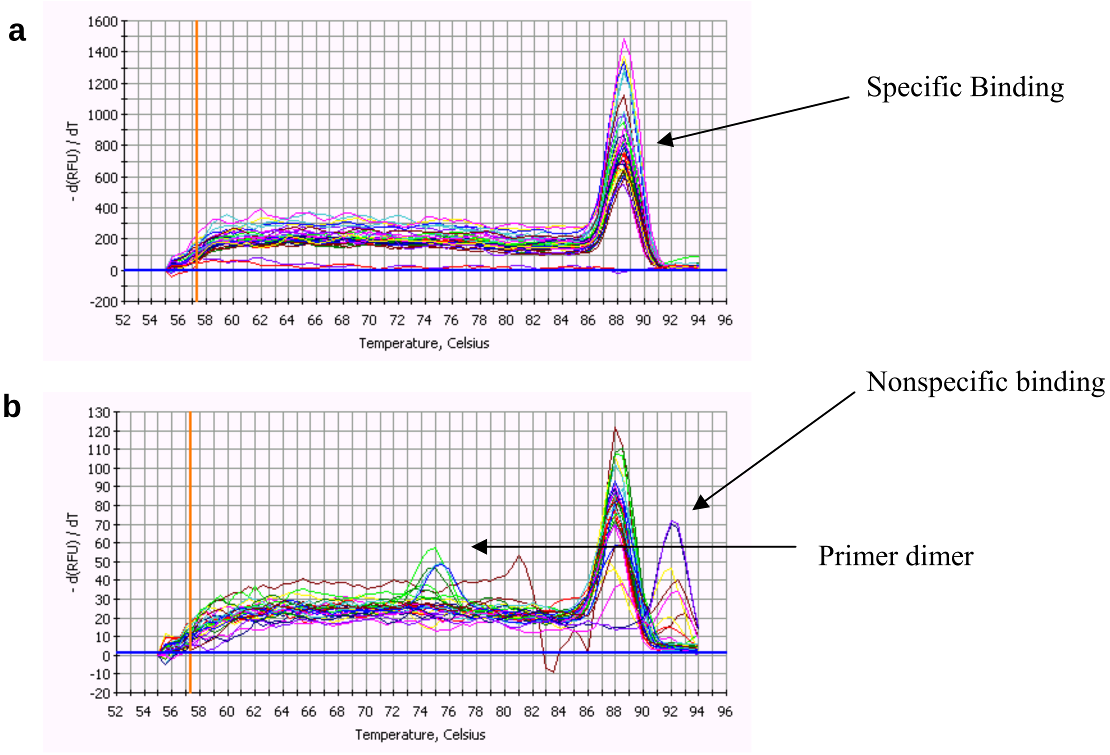


Figure 3.4: Melt curve graph: (a) Melt curve for qRT-PCR reaction of *BRCA1* gene for primary breast tumor samples gave single peak for each sample at the same temperature. **(b)** An example of a melt curve for the primer pair that gave nonspecific PCR product in qRT-PCR reaction for primary breast tumor samples.

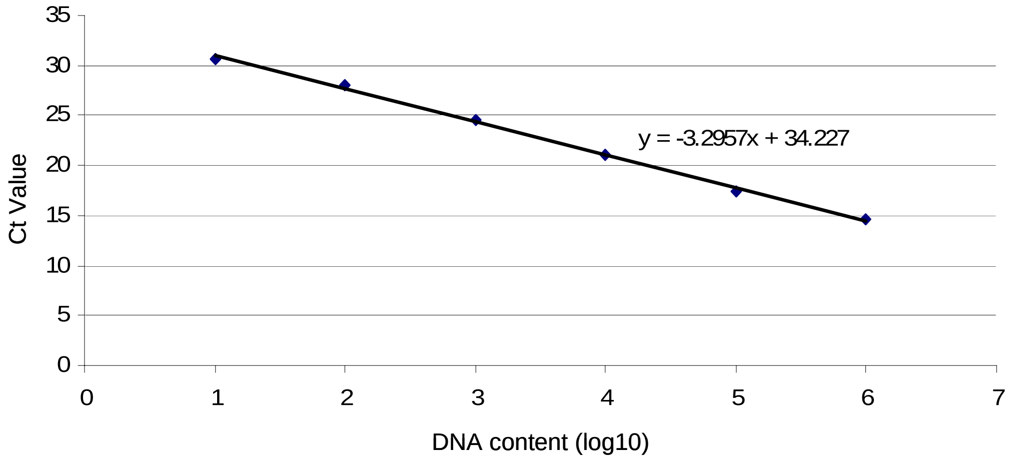
3.2.2 Standard curves and amplification efficiencies

Amplification efficiency calculation is an important step in the qRT-PCR analysis. In relative quantification measurements, it is used in the calculation of the comparative expressional level of the samples. If the amplification efficiency for both reference gene and the target gene was 100%, the $2^{-\Delta\Delta Ct}$ was used for the calculations. In other cases the following equation was used: (Pfaffl, 2001).

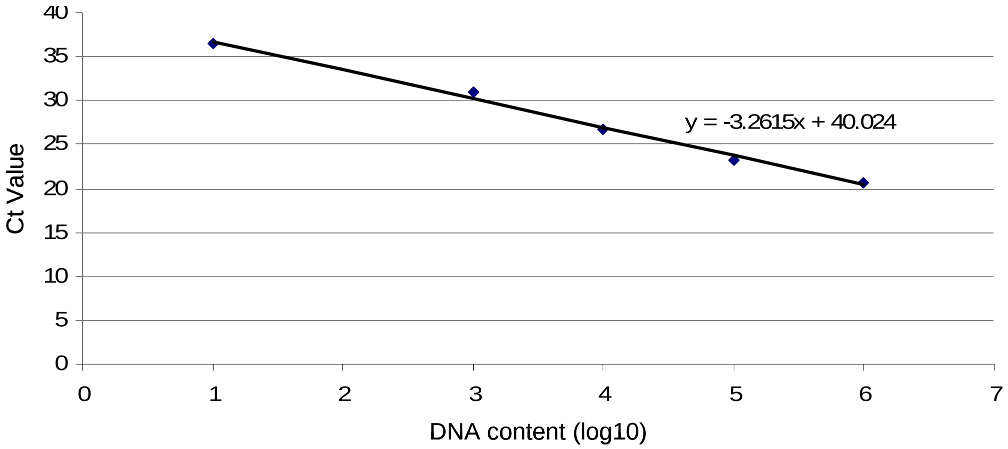
$$\Delta \mathrm{Ct} = [(E_{\mathrm{target}})^{\Delta \mathrm{CtTarget}\ (\mathit{control\text{-}sample})} / (E_{\mathrm{ref}})^{\Delta \mathrm{CtReference}\ (\mathit{control\text{-}sample})}]$$

The PCR efficiencies (E) were evaluated by 10-fold dilution series of cDNAs (1-1:100 000 dilution) for each pair of primers by using a breast carcinoma cell line cDNA pool (MCF7, MDA-MB-231, T47D, HMEC, MCF12A). The primer amplification efficiencies were also tested with reference genes *ACTB*, *GADPH*, and *SDHA* in breast tumor tissue cDNA pools (n=3) to ensure no inhibitory component was present in the tissue samples. No inhibitory effect was observed in amplification efficiencies (E=2.0). The primer amplification efficiencies of all the genes used in this study were tested and the results were given in the Table 3.1. A graph of threshold cycle (Ct) versus relative $\log_{10}$ copy number of the calibration sample from the dilution series was produced and the reaction efficiency was determined for each primer set by using the slope of this graph ($E=10^{(-1/slope)}$). Theoretically, the slope has to be 3.3 for the amplification reactions which have 100% efficiency. Figure 3.5 shows the results of the amplification efficiency curves of some of the genes used in the study.

(a)



(b)



(c)

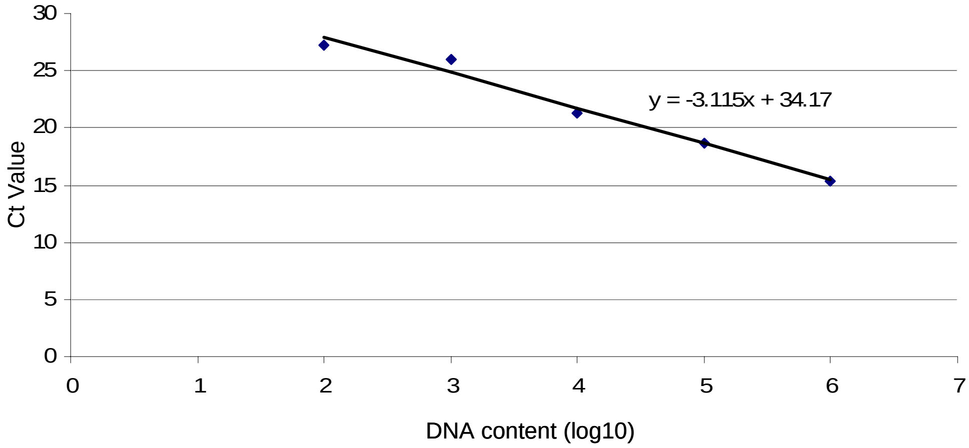


Figure 3.5: Standard curves of ACTB, SDHA and TBP for efficiency calculations. The efficiencies were calculated according to the formula: $E=10^{-(1/slope)}$. Slope is the number that is infront of x. If the slope is 3.3, than the efficiency of the reaction is calculated as 100%. The amplification efficiencies of the former primers were found to be 100%. The efficiencies for primer pairs of the genes (a) ACTB, (b) SDHA, (c) TBP.

Table 3.1: Amplification efficiencies of the gene-specific primer pairs.

Gene	Efficiency	% Efficiency	Gene	Efficiency	% Efficiency
BRCA1	2.0	100	*RPLP0*	1.9	90
ERBIN	1.9	90	*MRPL19*	2.0	100
OVCA1	1.9	90	*TTC22*	1.9	90
OVCA2	1.9	90	*IL22RA1*	1.9	90
SMG1	2.0	100	*ZNF224*	1.9	90
RENT2	1.8	80	*18SrRNA*	2.0	100
XRN2	1.9	90	*GSN*	2.0	100
RAD21	2.0	100	*COX6C*	2.0	100
MAC30	1.8	80	*MAF*	1.97	97
ACTB	2.0	100	*SFRP1*	1.97	97
GAPDH	1.97	97	*SPTBN1*	1.9	90
TBP	1.97	97	*GSPT1*	1.9	90
SDHA	2.0	100	*NME1*	1.9	90
HPRT	2.0	100	*PTTG1*	2.0	100
HMBS	2.3	130	*FN1*	1.9	90
B2M	1.9	90	*ID4*	2.0	100
PPIA	1.9	90	*EGFR*	2.0	100
GUSB	1.9	90	*ADAMTS1*	1.9	90
YWHAZ	2.0	100	*ATF3*	1.9	90
PGK1	1.9	90	*IGFBP6*	1.9	90
RPL41	2.0	100	*PRNP*	2.0	100
PUM1	2.0	100			

3.3 Expression profiles of target genes in BRCA1 over-expressed cells

Eight candidate genes that were found to be upregulated with the overexpression of BRCA1 protein were selected from the list of genes obtained from the previous study (Atalay *et al*, 2002). Real-time quantitative RT-PCR was performed with both of the BRCA1 overexpressing cell lines, MCF7 and UBR60-bcl2 and with their controls to determine the effect of BRCA1 induction on the expression of selected target genes.

The increase in BRCA1 mRNA level was found to be $2^{9.4}$ fold in pCMVmyc-BRCA1 transfected MCF7 cells compared to the ones transfected with the control plasmid pCMVmyc. Figure 3.6 shows the expression patterns of target genes in these BRCA1 upregulated cells. Real-time quantitative RT-PCR demonstrated that BRCA1 induced the expression of two tumor suppressor genes OVCA1 and OVCA2 1.46 and 1.6 folds and RNA surveillance genes RENT2, XRN2 and SMG1 1.86, 1.78, 1.78 fold respectively. Among the 8 selected target genes, RAD21 which functions in sister chromatid alignment as a part of the cohesion complex and also in double strand break repair was induced 1.94 fold with BRCA1 induction. MAC30 and receptor-mediated signaling gene ERBIN were also found to be induced by 1.76 and 1.74 fold respectively with the induction of BRCA1 (Figure 3.6a).

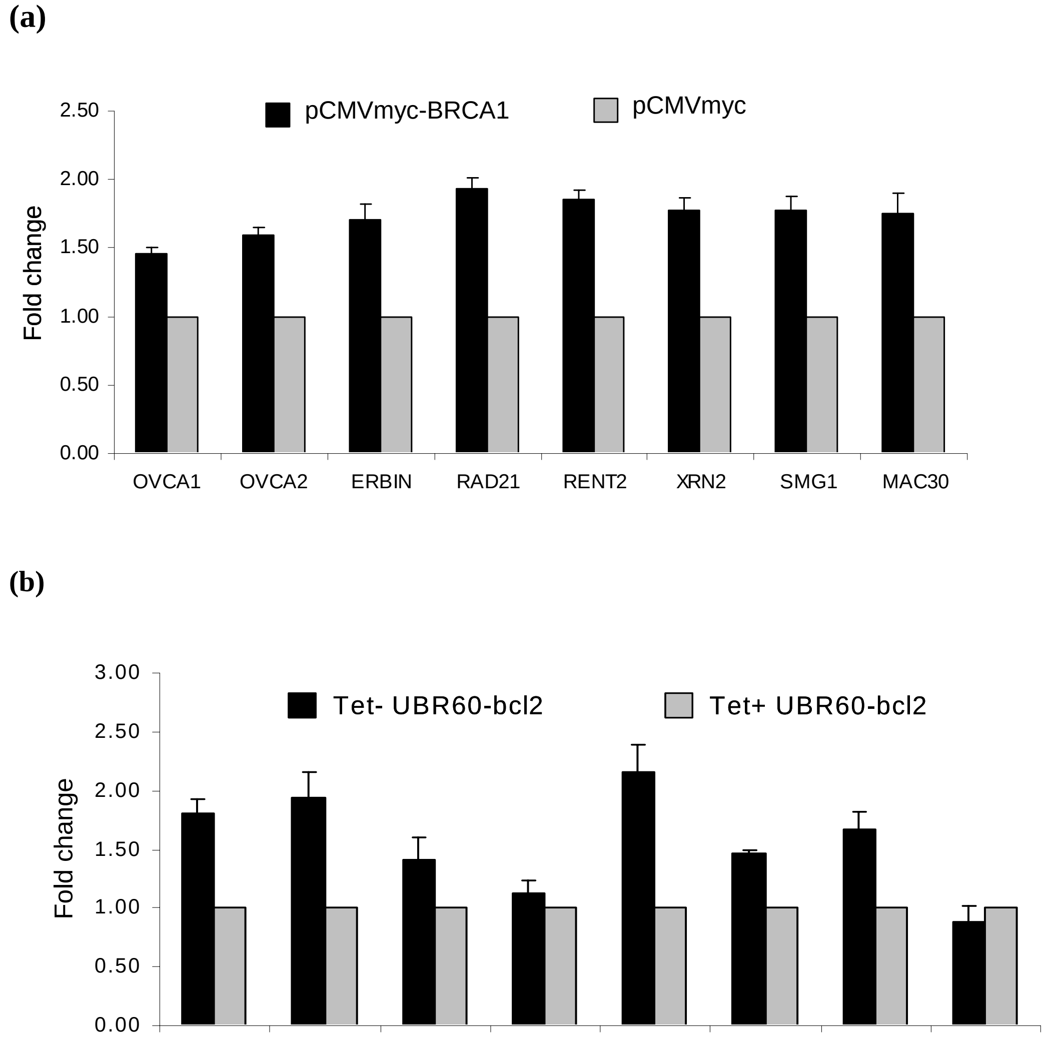


Figure 3.6: Expression profiles of target genes in BRCA1 induced cells. (a) The black bars shows the fold change in the expressions of BRCA1 target genes in pCMVmyc-BRCA1 transfected MCF7 cells compared to control cells while the gray ones, which were set to 1 shows control plasmid pCMVmyc transfected cells. **(b)** The fold change in the expressions levels of 8 target genes in UBR60-bcl2

osteosarcoma cells. BRCA1 expression was upregulated upon withdrawal of the tetracycline. Black bars show the expression profiles of the target genes compared to Tet (-) cells while the gray ones show Tet (+) cells.

The tetracycline withdrawal (Tet-) resulted in a $2^{8.7}$ fold induction of BRCA1 expression in UBR60-bcl2 cells. In this inducible system expression of six genes (OVCA1, OVCA2, ERBIN, RENT2, XRN2, SMG1) was increased while MAC30 and RAD21 genes were not affected (0.8 and 1.3 fold) compared to the UBR60-bcl2 cells grown in the presence of tetracycline (Tet+) (Figure 3.6b). When the Tet(+) cells were used a as a calibration control the fold induction of the six genes in Tet(-) cells were as follows; 1.81, 1.95, and 1.41 for OVCA1, OVCA2 and ERBIN and, 2.16, 1.46 and 1.67 for RENT2, XRN2 and SMG1 respectively.

All the transfection and tetracycline withdrawal experiments were performed three times to demonstrate the reproducibility of the results.

3.4 Expression profiles of target genes in breast cancer cell lines

The expression patterns of eight candidate BRCA1 target genes and BRCA1 were analyzed in 8 breast cancer cell lines (BCC) (T47D, MCF7, MDA-MB-231, MDA-MB-453, MDA-MB-468, BT20, BT474 and HCC1937) and in the hTERT immortalized mammary epithelial cell line HME1. qRT-PCR was used for the determination of expressions and the level of expression was calculated by using two different approaches. All the expression values of the target genes were first normalized with GAPDH, which was used as a reference gene in cell line studies ($2^{\Delta Ct}$). The normalized values then either normalized to the hTERT immortalized HME1 to see the expression profile of the genes in cancer cells compared to normal cells or to the mean expression values in all cell lines to obtain a more global picture ($2^{-(\Delta\Delta Ct)}$).

The results obtained through the former analyses were used for the clustering of BCC and genes. The BCC showed different patterns of clustering with respect to normalization strategy used (http://rana.lbl.gov/EisenSoftware.htm; Figure 3.7).

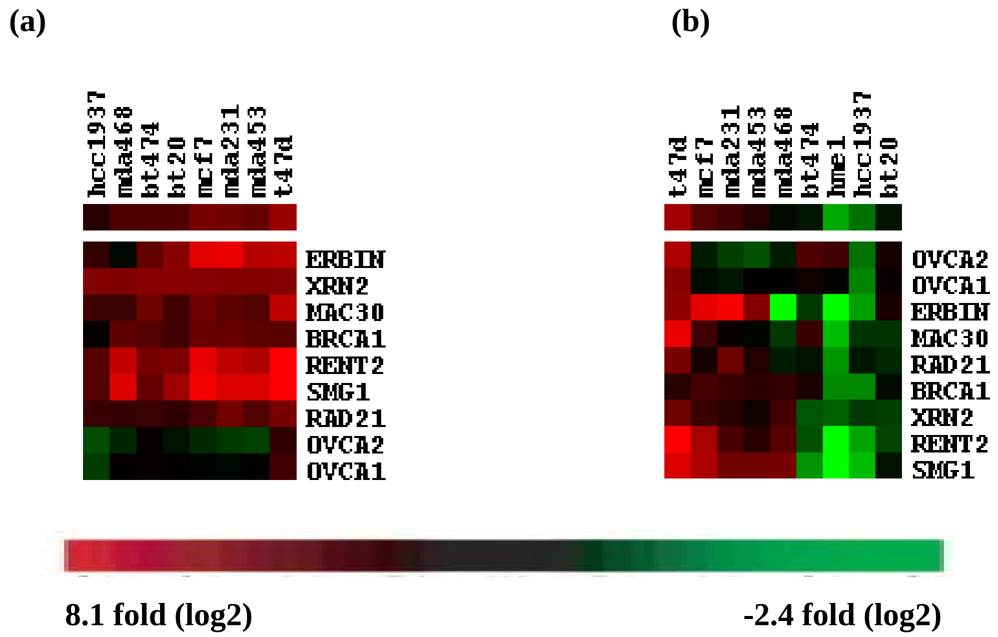
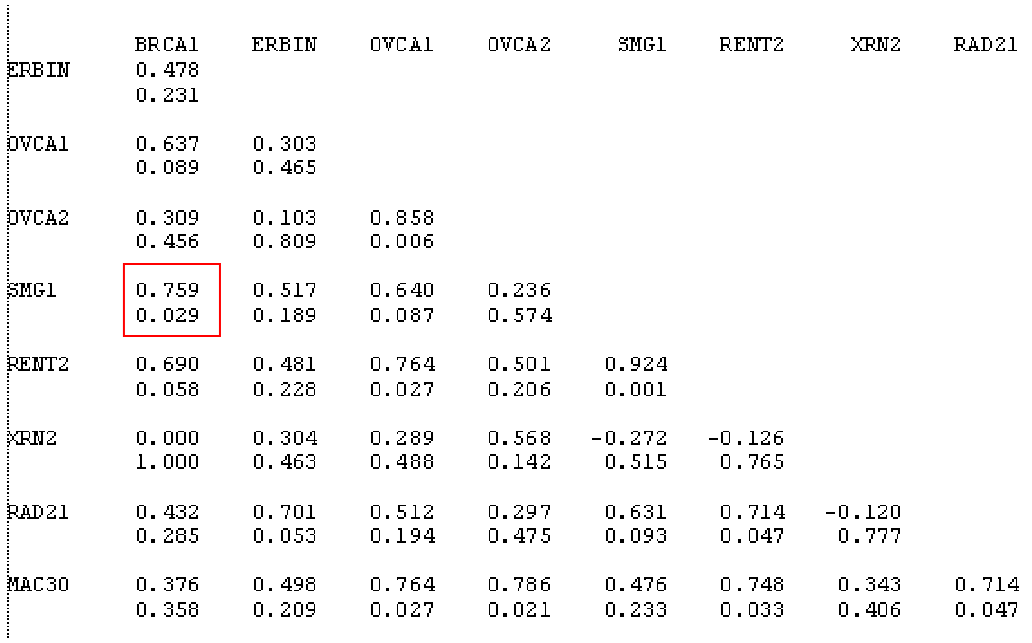


Figure 3.7: Hierarchical clustering of breast cancer cells (BCC). (a) The GAPDH normalized gene expression values of the target genes in the cell lines were normalized with respect to the expression values of the target genes in HME1 for generating the clustergram. **(b)** The GAPDH normalized expression values of the target genes were normalized to the mean expression values of all breast cancer cell lines.

To see the expression correlation between the target genes and BRCA1 gene in BCC, the correlation coefficients between the target gene expression profiles with that of BRCA1 expression profiles were calculated by Pearson correlation analysis. The values obtained through HME1 normalization and global normalization was analyzed independently. When the expression values were normalized to that of HME1, only SMG1, one of the RNA surveillance genes, was found to be highly correlated with BRCA1 (Table 3.2. Pearson correlation, Minitab; N=8; r = 0.759; p = 0.029).

Table 3.2: Correlation coefficients and corresponding p values of eight target genes and BRCA1 expression values. The expression values of the target genes in breast cancer cells were normalized to that of HME1. The normalized values were used for correlation analysis.



```
          BRCA1    ERBIN    OVCA1    OVCA2     SMG1    RENT2     XRN2    RAD21
ERBIN     0.478
          0.231

OVCA1     0.637    0.303
          0.089    0.465

OVCA2     0.309    0.103    0.858
          0.456    0.809    0.006

SMG1      0.759    0.517    0.640    0.236
          0.029    0.189    0.087    0.574

RENT2     0.690    0.481    0.764    0.501    0.924
          0.058    0.228    0.027    0.206    0.001

XRN2      0.000    0.304    0.289    0.568   -0.272   -0.126
          1.000    0.463    0.488    0.142    0.515    0.765

RAD21     0.432    0.701    0.512    0.297    0.631    0.714   -0.120
          0.285    0.053    0.194    0.475    0.093    0.047    0.777

MAC30     0.376    0.498    0.764    0.786    0.476    0.748    0.343    0.714
          0.358    0.209    0.027    0.021    0.233    0.033    0.406    0.047
```

When the expression values were obtained through global normalization, three of the RNA surveillance genes, SMG1, RENT2 and XRN2 and RAD21 were found to be highly correlated with BRCA1 (Pearson correlation, Minitab; N=9; r = 0.864, p = 0.003; r = 0.830, p = 0.006; r = 0.698, p = 0.036; and r = 0.682, p = 0.043) respectively) (Table 3.3).

Table 3.3: Correlation coefficients and corresponding p values of eight target genes and BRCA1 expression values. The expression values of the target genes in breast cancer cells were normalized to the mean expression values of all the genes in all the cell lines. The normalized values were used for correlation analysis.

	BRCA1	ERBIN	OVCA1	OVCA2	SMG1	RENT2	XRN2	RAD21
ERBIN	0.627 0.071							
OVCA1	0.521 0.150	0.288 0.453						
OVCA2	0.089 0.819	-0.017 0.965	0.830 0.006					
SMG1	0.864 0.003	0.669 0.049	0.470 0.202	-0.011 0.978				
RENT2	0.830 0.006	0.648 0.059	0.572 0.107	0.181 0.641	0.962 0.000			
XRN2	0.698 0.036	0.478 0.193	0.478 0.193	0.060 0.877	0.911 0.001	0.919 0.000		
RAD21	0.682 0.043	0.790 0.011	0.378 0.315	0.035 0.928	0.819 0.007	0.855 0.003	0.753 0.019	
MAC30	0.617 0.077	0.640 0.063	0.620 0.075	0.446 0.229	0.699 0.036	0.848 0.004	0.647 0.060	0.831 0.006

3.5 Identification of endogenous reference genes for real time qRT-PCR analysis in normal matched breast tumor tissues.

Inclusion of an endogenous reference gene or genes (RGs) is crucial to standardize initial RNA quantity to overcome bias originating from RNA measurement errors, problems with RNA integrity, and differential cDNA conversion efficiencies. Quantification of a target gene requires the use of a proper reference gene whose expression is relatively stable across samples to estimate the degree of variability within and among experimental groups as well as to standardize the expression to a baseline common to all samples. Therefore we selected fifteen commonly used reference genes (ACTB, GAPD, TBP, SDHA, HPRT, HMBS, B2M, PPIA, GUSB, YWHAZ, PGK1, RPL41, PUM1, RPLP0, MRPL19) and three newly selected genes (TTC22, IL22RA1, ZNF22) which belong to different functional classes to assess their suitability as RGs in breast tumor and normal matched tissues.

3.5.1 Expression patterns of candidate RGs

Expression levels of 18 candidate RGs were determined in 23 breast tumor tissues and their matched normal samples by qRT-PCR using the SYBR Green I dye detection system. Amplification efficiencies calculated based on standard curves from the serial dilutions of breast cancer cell lines indicated that all primer pairs were over 90% efficient (values ranged between 1.97 and 2.3; Table 3.1). Each RG had a different expression range between the tumors and matched normal samples. The RG expression levels displayed a wide range of Ct values between 13 and 33, grouped into three ranges for their mean Ct values. Highly expressed genes were B2M, ACTB, PPIA, RPL41, RPLP0 and GAPDH (mean Ct values below 20 cycles). Genes with moderate expression were YWHAZ, PGK1, SDHA, PUM1, MRPL19 and GUSB (mean Ct values between 20 and 25 cycles). Genes with low expression were TBP, HPRT, IL22RA1, TTC22, ZNF224 and HMBS (mean Ct values over 25 cycles).

The reference genes used in our panel exhibited relatively higher expression in tumor samples than in their normal counterparts (Paired t-test; $p<0.05$). 17 out of 18 reference genes displayed a consistent 1.86±0.7 (log2, mean±std) fold expression difference between breast tumor and normal pairs. The expression range of candidate genes was shown in terms of difference between the Ct values of tumor and normal samples as box-whisker-plots (Figure 3.8a and 3.8b). The box-whisker-plots representing the same data were shown in two different forms.

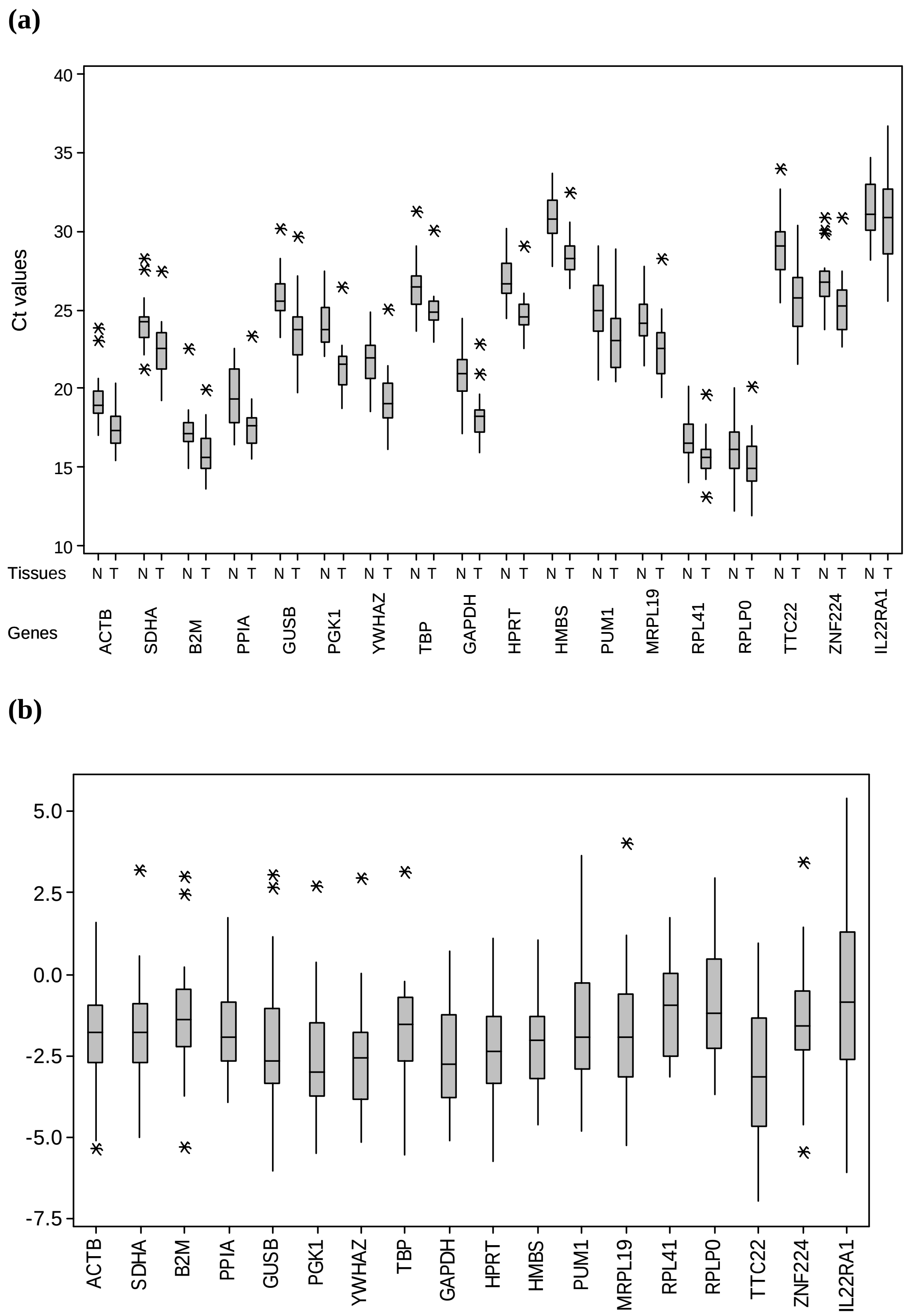


Figure 3.8: Expression range of differences between the Ct values of breast tumor and normal samples for each candidate reference genes. a and b show the

different representation of the same data. Threshold cycle values (Ct_{Tumor}-Ct_{normal}) for each reference gene are shown as medians (lines), 25th to 75th percentile (boxes) and range (whiskers). Whiskers illustrate the data points in Q3+1.5(IQR) and Q1-1.5(IQR)[a].

P values were calculated using the paired Student's t-test ($p<0.05$, significant).

* shows the Ct values that fall beyond the whiskers.

[a] Interquartile range, IQR=Q3-Q1

3.5.2 Expression stability of candidate RGs

The expression stability of each gene was validated using two different software programs, geNorm and NormFinder, to identify the most suitable genes for normalization.

The geNorm program determines the most stable RGs from a set of investigated genes in a given set of samples. It calculates the gene expression stability measure (*M*) for an RG, considering the average pair-wise variation of all other tested RGs (Vandesompele *et al.*, 2002). The lowest M value marks the gene(s) with the most stable expression. The average M value of the 18 candidate RGs are plotted in Figure 3.9a. The curve represents the stepwise exclusion of the least stable genes with higher M values. This result led to the identification of the two most stable genes, ACTB and SDHA, in the tested samples (M=0.7).

In addition to the stability value M, pair-wise variations (V*n/n*+1) were calculated to determine the effect of adding a gene (*n*+1) in normalization (Figure 3.9b). This allowed for determination of a normalization factor (NF) needed to define the optimal number of RGs required for reliable normalization. A large pair-wise variation means that the added gene has a significant effect on normalization and should therefore be included for calculation of reliable normalization (Vandesompele *et al.*, 2002). The most stable six genes, ACTB, SDHA, TBP, PGK1, GUSB, and MRPL19 yielded a V value of 0.147, giving the cut-off value 0.15.

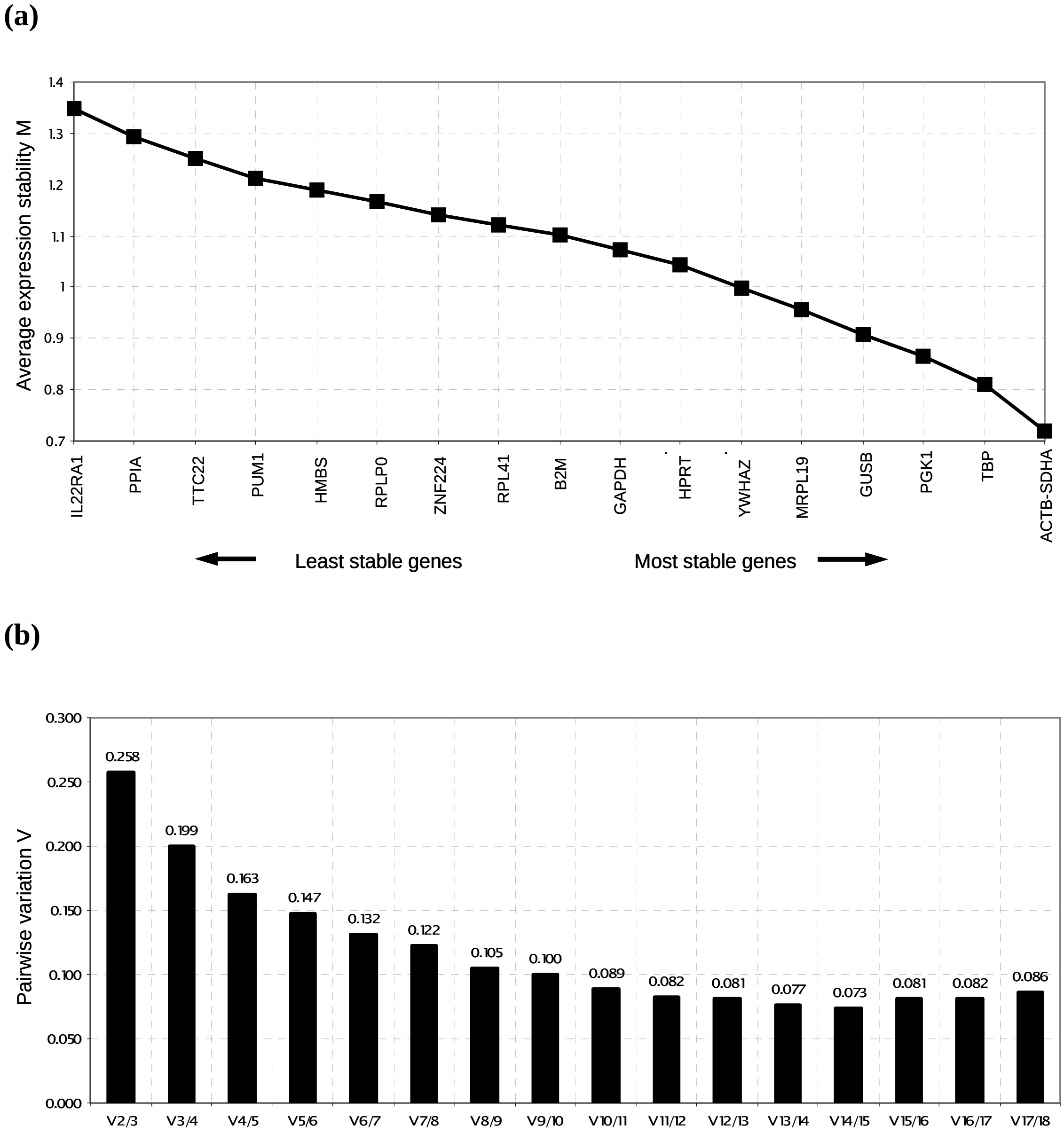


Figure 3.9: Selection of reference genes for normalization in breast tumor samples using geNorm analysis.

(a) The curve represents the stepwise exclusion of the least stable genes according to the M values calculated by geNorm. The genes with the higher M values are eliminated and the remainders represent the two most stable genes, SDHA and ACTB. The genes are ranked on the *x*-axis from left to right according to their expression stability. **(b)** Determination of the optimal number of reference genes for normalization by calculation of the pair-wise variation (V) of normalization factor ratios for different numbers of control genes. Each number on the bars shows the pair-wise variation between two sequential normalization factors. On the left-most side is the pair-wise variation when the number of genes is enlarged from 2 to 3 (V2/3). Stepwise inclusion of less stable genes generates the next data points. Inclusion of the third and the fourth genes (V4/5) nears the V value to the cut-off value of 0.15.

We also used the NormFinder software program for stability evaluation among the candidate RGs. NormFinder is an add-in for Microsoft Excel and is used for calculating a stability value from a set of candidate RGs. In this program, the stability value is based on the combined estimate of inter- and intra-group expression variations of the studied gene. The candidate gene with the smallest variability value has higher stability as it shows the lowest variability of inter- and intra-group expression (Andersen *et al.*, 2004). NormFinder also ranks the set of candidate RGs according to their expression stability from a panel of candidate genes that could be organized in different subgroups (tumor and matched normal tissues). Our findings indicated that the genes occupying the top five ranks, SDHA, ACTB, MRPL19, TBP, and GUSB appeared to be the most stable genes, while IL22RA1 was defined as the least stable gene (Table 3.4). Although NormFinder selected SDHA as the most stable gene with a stability value of 0.135, the best combination of the two genes selected by the program, ACTB and SDHA, improved the stability value to 0.089, indicating a more reliable normalization.

Table 3.4: Rank of candidate reference genes according to the expression stability calculated by Normfinder. The candidate reference genes are listed with decreasing expression stability from 1 to 18. The best combination of the two genes and the stability value were calculated by NormFinder.

Ranking order	**Gene name**	**Stability value**
1	SDHA	0.135
2	ACTB	0.155
3	MRPL19	0.186
4	GUSB	0.196
5	TBP	0.215
6	PUM1	0.271
7	ZNF224	0.289
8	PPIA	0.315
9	HPRT	0.330
10	B2M	0.340
11	PGK1	0.345
12	YWHAZ	0.391
13	GAPDH	0.404
14	RPL41	0.406
15	HMBS	0.456
16	RPLP0	0.478
17	TTC22	0.520
18	IL22RA1	0.574
Best two genes	**ACTB and SDHA**	**0.089**

3.5.3 Assessment of suitable RGs for normalization

GSN is an actin depolymerizing factor acting as the principal intracellular and extracellular actin-severing protein. Expression of GSN was shown to be undetectable or greatly reduced in invasive human breast carcinomas both at the protein and RNA level (Asch *et al.*, 1996). The progressive loss of GSN from benign mammary tissue through different stages of mammary tumorigenesis has also been demonstrated (Dong *et al.*, 2002; Winston *et al.*, 2001). To assess the significance of the selected RGs for normalization, the expression level of GSN mRNA was measured by qRT-PCR and statistically evaluated in the same set of tumor and matched normal breast tissue samples. Since a gene expression NF could either be based on a single gene or a combination of gene expression values (Vandesompele *et al.*, 2002), GSN gene expression levels were normalized using the RGs proposed by the geNorm or NormFinder calculations, i.e., ACTB, SDHA, GUSB, MRPL19, TBP, and PGK1 in combinations (Figure 3.10). We also tested the performance of IL22RA1, the lowest ranked gene both in the geNorm and NormFinder analyses for GSN normalization (Figure 3.10). The median GSN expression values were below zero, which indicated down regulation with respect to matched normal GSN expression, independent of the NF used.

Moreover, statistical analyses indicated that the GSN expression was significantly down regulated in tumor samples when compared with that from normal samples with combinational use of the best RGs (*ACTB* and *SDHA*) proposed both by the geNorm or NormFinder programs ($p<0.05$). In contrast, down regulation of the *GSN* expression was not significant when the least stable gene, *IL22RA1,* was used as NF ($p>0.05$) with on average 39% of the tumor samples being upregulated with respect to their normal counterparts (Figure 3.10). In addition, when GSN expression in tumors was not normalized with RGs but normalized only with the corresponding normal GSN expression (ΔCt; $\mathrm{Ct}_{(\mathrm{GSN\ tumor})}$- $\mathrm{Ct}_{(\mathrm{GSN\ normal})}$), the expression difference was not significant between tumor and normal pairs (0.18 ± 2.2, mean ± SD; $p=0.7$, one sample t-test).

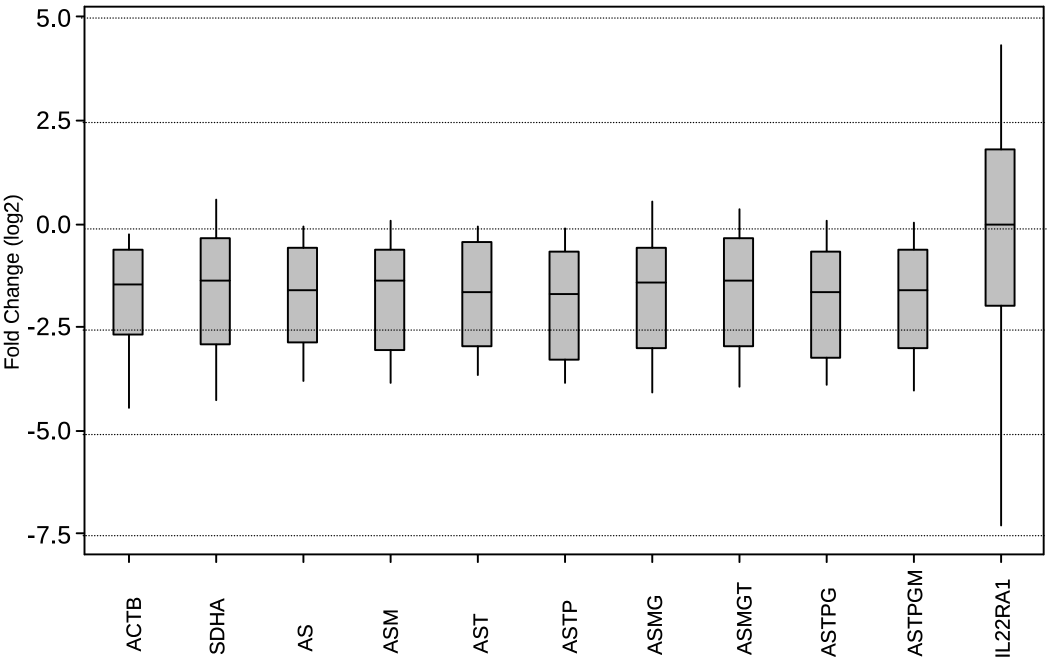


Figure 3.10: The normalization of GSN gene expression with combinations of candidate reference genes in tumor and matched normal breast samples. The gene expression level of GSN in 23 tumor and normal samples was normalized with respect to an individual RG or combinations of RGs and displayed as a box plot of $[(E_{\text{target}})^{\Delta\text{CtTarget}\,(\textit{control-sample})} / (E_{\text{ref}})^{\Delta\text{CtReference}\,(\textit{control-sample})}]$ using matched normal samples as controls. ACTB (A), SDHA (S), GUSB (G), MRPL19 (M), TBP (T), and PGK1 (P) individually or in combinations of two or more gene combinations of the above RGs are used as NFs. GSN normalization by the lowest ranking RG, IL22RA1 was performed. P values were calculated using the paired Student’s t-test (p<0.05, significant). ACTB, p= 0.003; SDHA, p= 0.009; AS, p= 0.005; ASM, p= 0.008; AST, p= 0.008; ASTP, p= 0.007; ASMG, p= 0.014; ASMGT, p= 0.014; ASTPG, p= 0.010; ASTPGM, p= 0.012; IL22RA1, p= 0.236.

Fold change values in GSN expression obtained by using different NFs were significantly correlated with each other, yet the degree of correlation increased when two genes (in combination ACTB and SDHA) were used as NF. For example, the correlation coefficient between tumor samples' GSN expression values normalized with ACTB and those with SDHA ($r_{\text{A vs S}}$) was 0.80 whereas the degree of correlation increased when a combination of the best two RGs was used ($r_{\text{AS vs A}}$ = 0.95 and $r_{\text{AS vs S}}$ = 0.96, where A and S refer to ACTB and SDHA respectively). The addition of the third or the fourth gene to the best two genes did not change the correlation results more than 1% ($r_{\text{AS vs ASM}}$ = 0.96 and $r_{\text{AS vs AST}}$ = 0.97, $r_{\text{AST vs ASTP}}$ = 0.97).

Table 3.5: Correlation coefficients and corresponding p values of GSN expression values normalized with one or combinations of reference genes.

	ACTB	SDHA	AS	ASM	AST	ASTP	ASMG	ASMGT	ASTPG
SDHA	0.807 0.000								
AS	0.955 0.000	0.946 0.000							
ASM	0.872 0.000	0.949 0.000	0.955 0.000						
AST	0.907 0.000	0.957 0.000	0.978 0.000	0.958 0.000					
ASTP	0.902 0.000	0.951 0.000	0.973 0.000	0.958 0.000	0.977 0.000				
ASMG	0.841 0.000	0.944 0.000	0.936 0.000	0.985 0.000	0.936 0.000	0.937 0.000			
ASMGT	0.836 0.000	0.953 0.000	0.938 0.000	0.986 0.000	0.958 0.000	0.949 0.000	0.994 0.000		
ASTPG	0.881 0.000	0.959 0.000	0.966 0.000	0.970 0.000	0.967 0.000	0.985 0.000	0.976 0.000	0.979 0.000	
ASTPGM	0.848 0.000	0.954 0.000	0.945 0.000	0.985 0.000	0.955 0.000	0.972 0.000	0.989 0.000	0.991 0.000	0.992 0.000

3.5.4 Evaluation of 18S rRNA to mRNA ratio

As shown in Figure 3.10 the reference genes used in our panel exhibited relatively higher expression in tumor samples than in their normal counterparts (Paired t-test; $p<0.05$). We concentrated on mRNA/rRNA ratio to explain this expression difference. Since total RNA is represented mostly by rRNA (>90%), even a small decrease in rRNA expression may lead to a disproportional increase in the mRNA pool estimation (Elberg *et al.*, 2006; Spanakis, 1993). In light of the studies that show the methylation of rDNA genes in breast and ovarian cancers when compared with those of normal controls (Chan *et al.*, 2005, Yan *et al.*, 2000), we chose 18SrRNA as a representative of rRNA and ACTB and SDHA for mRNA to investigate their expression levels in tumor and normal samples (n=13).

We quantified 18S rRNA, ACTB, and SDHA mRNA levels in a group of 13 tumor and normal pairs. The mean expression of 18S rRNA was found to be downregulated in tumor samples (9/13) compared to their normal counterparts (log2 difference, 1.16 ± 1.06; mean ± STD) while the expression of ACTB and SDHA genes were consistently high in tumor samples compared to their normal pairs (log2 difference, 1.9±1.4 and 1.8±1.5 respectively; mean ± STD). Our results showed that the 18S rRNA to ACTB or SDHA mRNA ratio was approximately 8-fold lower in tumors than that of normal pairs on average (paired t-test p= 4.2×10^{-5} and p=2.2×10^{-4} respectively) (Figure 3.11).

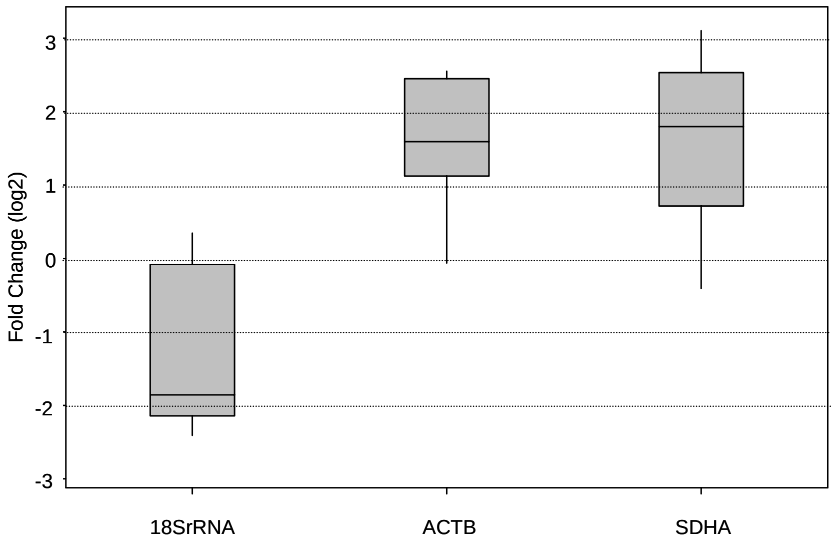
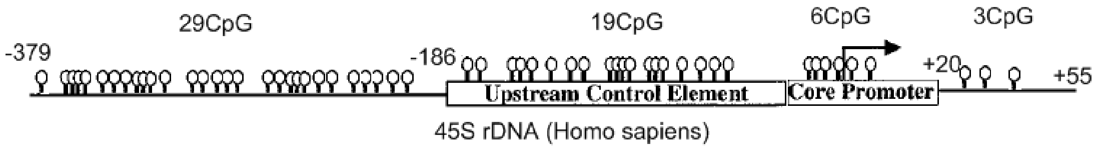


Figure 3.11: The expression levels of 18S rRNA, ACTB and SDHA genes in tumor samples compared to their normal pairs. The gene expression levels of 18S rRNA, ACTB and SDHA in 13 tumor samples were normalized with respect to that of their normal pairs (-ΔCt : -($Ct_{(Tumor)}$-$Ct_{(Normal)}$)) and displayed as a box plot. The 18S rRNA to ACTB or SDHA mRNA ratio was close to 8-fold lower in tumors than that of normals. The significance of this difference was calculated.

3.5.5 Bisulfite sequencing of DNA samples for methylation analysis.

The studies of Chan *et al.* and Yan *et al.* have shown rDNA genes in breast and ovarian cancers to be methylated when compared with those of normal controls (Chan *et al.*, 2005, Yan *et al.*, 2000). Considering these studies and the results shown in Figure 3.13, the methylation status of the 45S rRNA promoter was investigated in two breast tumor samples. The DNA from the tumor samples was isolated, subjected to the bisulfite treatment and used for PCR amplification. The PCR amplified product was then sequenced (bisulfite sequencing) to determine its pattern of methylation. Treatment of DNA with bisulfite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. Eventually the sequence does not change if the cytosine residues are methylated. The promoter region and the CpG status of this region are shown in Figure 3.12. A 395 bp product was amplified spanning the promoter region between -340 and +55 at which 54 CpG islands are located (Figure 3.12a). Figure 3.12b shows the sequencing and alignment of the original 45S rDNA promoter sequence and bisulfite treated and amplified DNA sequence from the tumor sample MFT 148. This tumor sample was found to have 5 fold lower expression of 18SrRNA compared to ACTB and SDHA expression with qRT-PCR. The sequencing results showed that the CpG islands were not converted after bisulfite treatment and displayed full alignment with the original sequence of the 45S rRNA promoter. Among the 54 CpG islands 50 were found to be methylated (Figure 3.12b). Another tumor sample, MFT116, which was found to have 4.5 fold lower expression of 18S rRNA compared to ACTB and SDHA expression was also sequenced and the sequencing results showed that among the 54 CpG islands 46 were found to be methylated.

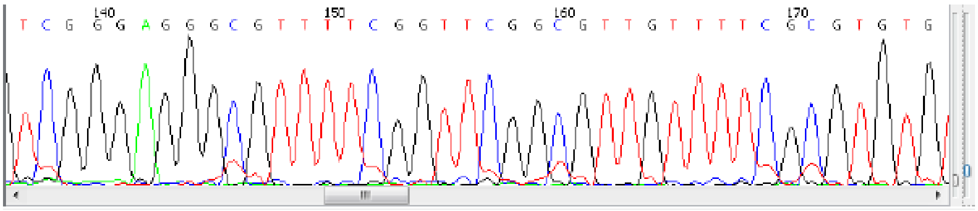
(a)



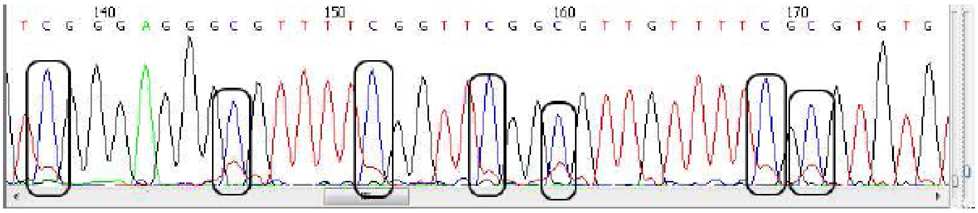
```
GAGTCGGAGAGCGTTTTTTGAGCGCGCGTGCGGTTCGAGA      -340

GGTCGCGTTTGGTCGGTTTTCGGTTTTTCGTGTGTTTCGG      -300

TCGTAGGAGGGGTCGGTCGAAAATGTTTTCGGTTTTCGTT      -260

TTGGAGATACGGGTCGGTTTTTTGCGTGTGGTACGGGCGG      -220

TCGGGAGGGCGTTTTCGGTTCGGCGTTGTTTTCGCGTGTG      -180

TTTTGGGGTTGATTAGAGGGTTTCGGGCGTTTCGTGTGTG      -140
          upstream control element
GTTGCGATGGTGGCGTTTTTGGGGATAGGTGTTCGTGTCG      -100

CGCGTCGTTTGGGTCGGCGGCGTGGTCGGTGACGCGATTT      -60
                     core promoter element
TTCGGTTTCGGGGAGGTATATTTTTCGTTTCGAGTCGGTA      -20
                      +1
TTTTGGGTCGTCGGGTTATTGTTGATACGTTGTTTTTTGG      +21

CGATTTGTCGTTGGAGAGGTTGGGTTTTCGGATG            +55
```

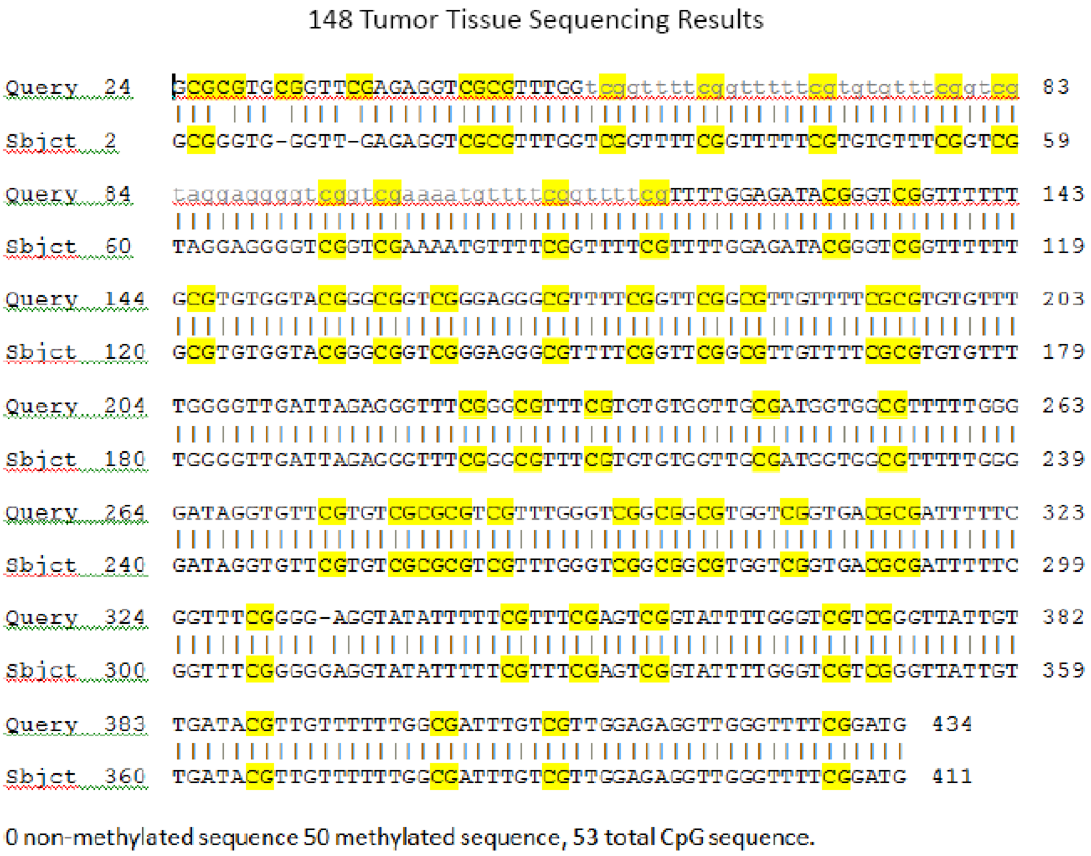
b

148 Tumor Tissue Sequencing Results



C-T overlapping reads in 148 Tumor Tissue Sequencing Results





148 Tumor Tissue Sequencing Results

```
Query  24   GCGCGTGCGGTTCGAGAGGTCGCGTTTGGtcggttttcggtttttcgtgtgtttcggtcg  83
            ||| ||| |||| |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2    GCGGGTG-GGTT-GAGAGGTCGCGTTTGGTCGGTTTTCGGTTTTTCGTGTGTTTCGGTCG  59

Query  84   taggaggggtcggtcgaaaatgttttcggttttcgTTTTGGAGATACGGGTCGGTTTTTT  143
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  60   TAGGAGGGGTCGGTCGAAAATGTTTTCGGTTTTCGTTTTGGAGATACGGGTCGGTTTTTT  119

Query  144  GCGTGTGGTACGGGCGGTCGGAGGGCGTTTTCGGTTCGGCGTTGTTTTCGCGTGTGTTT  203
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120  GCGTGTGGTACGGGCGGTCGGAGGGCGTTTTCGGTTCGGCGTTGTTTTCGCGTGTGTTT  179

Query  204  TGGGGTTGATTAGAGGGTTTCGGCGTTTCGTGTGTGGTTGCGATGGTGGCGTTTTTGGG  263
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  180  TGGGGTTGATTAGAGGGTTTCGGCGTTTCGTGTGTGGTTGCGATGGTGGCGTTTTTGGG  239

Query  264  GATAGGTGTTCGTGTCGCGCGTCGTTTGGGTCGGCGGCGTGGTCGGTGACGCGATTTTTC  323
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  240  GATAGGTGTTCGTGTCGCGCGTCGTTTGGGTCGGCGGCGTGGTCGGTGACGCGATTTTTC  299

Query  324  GGTTTCGGGG-AGGTATATTTTTCGTTTCGAGTCGGTATTTTGGGTCGTCGGTTATTGT  382
            |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  300  GGTTTCGGGGGAGGTATATTTTTCGTTTCGAGTCGGTATTTTGGGTCGTCGGTTATTGT  359

Query  383  TGATACGTTGTTTTTTGGCGATTTGTCGTTGGAGAGGTTGGGTTTTCGGATG  434
            ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  360  TGATACGTTGTTTTTTGGCGATTTGTCGTTGGAGAGGTTGGGTTTTCGGATG  411
```

0 non-methylated sequence 50 methylated sequence, 53 total CpG sequence.

Figure 3.12: The promoter region and the bisulfite sequencing result of the 45S rDNA. (a) The locations of the CpG islands in the 45S rRNA promoter are shown in the figure. The 395 bp product was amplified by using the bisulfite-treated DNA. The bold sequences show the locations of the primers used for both amplification and sequencing reactions. **(b)** The illustration of the bisulfite sequencing result of MFT148 breast tumor tissue is shown in the upper panel. The bottom panel shows the cytosine residues in the CpG islands that were not changed after the bisulfite treatment. The yellow labeled bars show the methylated CpG islands (50 of the 53 CpG islands) in the breast tumor sample.

3.6 Human breast and paired normal tissue samples

Breast tumor samples and their matched-normal samples (n=32) were used to investigate the expression profiles of the selected target genes and also their correlation with BRCA1 expression *in vivo*.

Primary tumor samples and matched-normal breast tissues were obtained from patients during surgery and immediately snap-frozen in liquid nitrogen and stored at -80^{o}C until RNA extraction. The frozen tissue samples were sectioned and mounted on glass slides. The slides were stained with hematoxylin and eosin and sent for histopathological examinations (Figure 3.13). Only those tumor samples with more than 90% of tumor cells and matched tissue pairs with normal histology were included in the study. Table 3.3 shows the percentage of the tumor cells in each tumor sample used in this study.

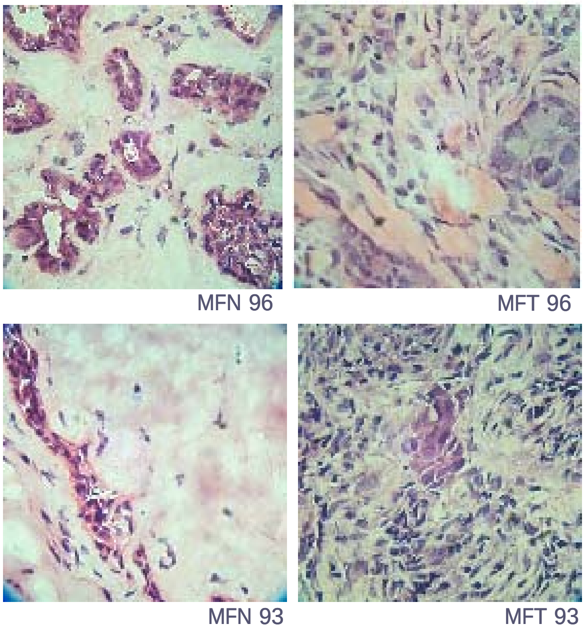


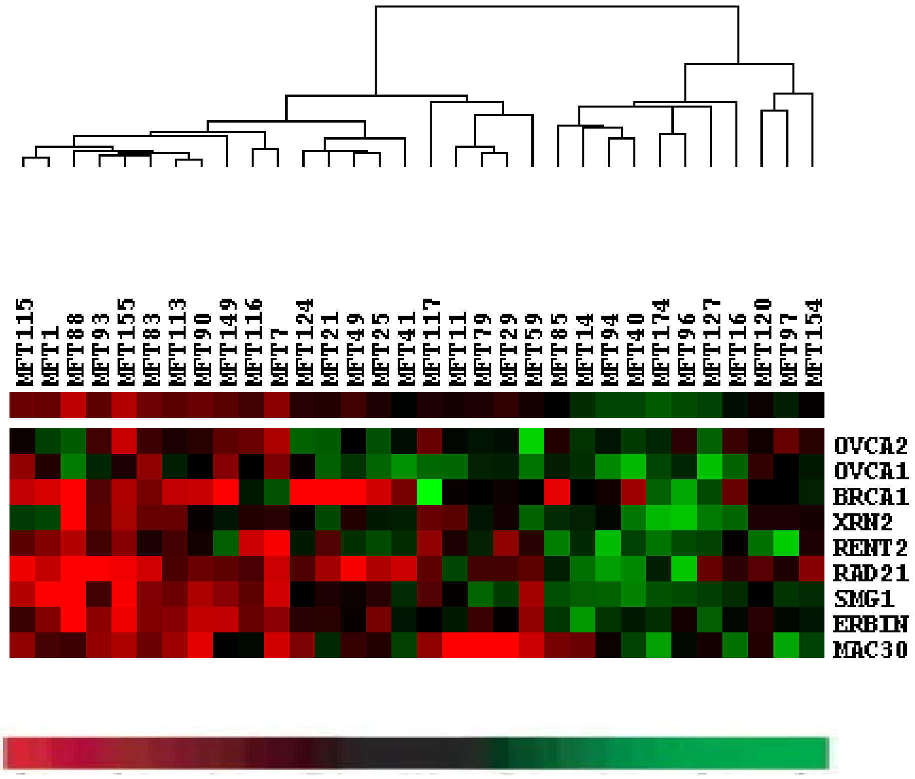
Figure 3.13: Hematoxylin staining of the breast tissue sections. The tissue sections were prepared from the normal and tumor breast tissue biopsies. The tumor cell percentage in breast tumor samples MFT96 and MFT93 were 100%. The breast tissues MFN96 and MFN93 are the matched-normal tissues of these two tumor samples taken from the same patients. The non-tumor tissues displayed normal breast tissue staining.

Table 3.6: Tumor cell percentages of the breast tumor samples used in this study.

Sample ID	**Tumor %**	**Sample ID**	**Tumor %**
MFT001	70	MFT 093	100
MFT007	90	MFT 094	100
MFT011	90	MFT 096	100
MFT014	90	MFT 097	80
MFT016	100	MFT 113	100
MFT021	90	MFT 115	80
MFT025	100	MFT 116	90
MFT029	70	MFT 117	90
MFT040	100	MFT 120	90
MFT041	100	MFT 124	100
MFT049	100	MFT 127	100
MFT059	90	MFT 149	95
MFT079	90	MFT 154	95
MFT083	100	MFT 155	95
MFT085	100	MFT 173	90
MFT088	90	MFT 174	90
MFT 090	100		

3.6.1 Expression profiles of target genes in tumor and paired-normal breast tissues determined by qRT-PCR.

The pathologically evaluated 32 tumor and matched-normal samples were used in this study to find out the expression profiles of eight target genes and BRCA1. The expression levels of each target gene were determined by qRT-PCR with gene specific primers in each tumor and normal matched breast sample. Their expression values were first normalized with respect to the reference gene (normalization factor, NF) expression levels in each sample. The NF normalized tumor sample expression values were then normalized to the NF normalized pair-normal tissue expression values (Pfaffl, 2001). This normalization was an approach to eliminate (subtract) the effect of normal cells from the tumor cells. The values obtained through this analysis were used for the clustering of tumor samples by using an integrated pair of programs, Cluster and TreeView (http://rana.lbl.gov/EisenSoftware.htm). The result of the cluster analysis was shown in Figure 3.14.



3.56 fold (log2) **-2.93 fold**

(log2)

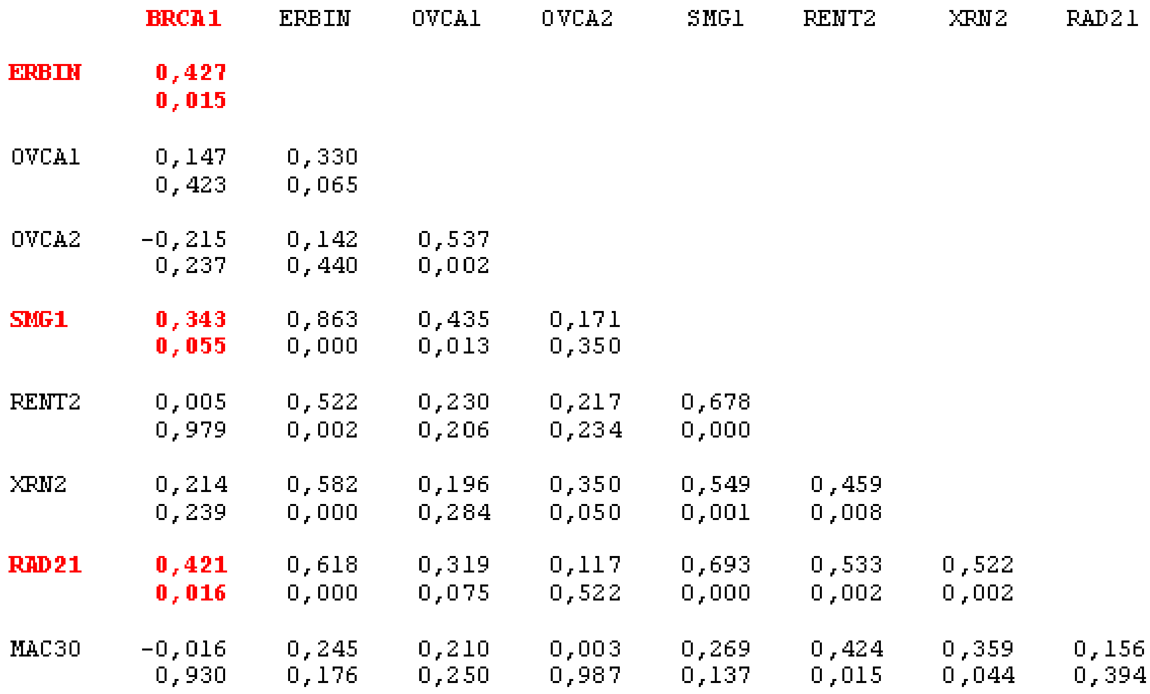
Figure 3.14: Hierarchical clustering of tumor samples. The expression value of each tumor sample was normalized with respect to its normal counterpart and used as an expression value for generating the clustergram heat map.

The hierarchical clustering gave two major groups with high and low expression profiles of the target genes normalized to their matched normal samples in the primary breast tumor samples. Then, we wanted to analyze if these two groups showed any statistically meaningful correlation in respect to gene expression profiles of these target genes with each other and with BRCA1 expression.

3.6.1.1 Correlation of target gene expression profiles

The expression correlation of the target genes and BRCA1 gene in the breast tumor samples were determined with Pearson correlation coefficient analysis by using the normalized expression data. Three of the eigth target genes, ERBIN, RAD21 and SMG1were found to be highly correlated with BRCA1 expression (Pearson correlation, Minitab; N=32; r = 0.427, r = 0.421 and r = 0.343, respectively; p<0,5) (Table 3.3).

Table 3.7: Correlation coefficients and corresponding p values of eight target genes and BRCA1 expression values.



	BRCA1	ERBIN	OVCA1	OVCA2	SMG1	RENT2	XRN2	RAD21
ERBIN	0,427 0,015							
OVCA1	0,147 0,423	0,330 0,065						
OVCA2	-0,215 0,237	0,142 0,440	0,537 0,002					
SMG1	0,343 0,055	0,863 0,000	0,435 0,013	0,171 0,350				
RENT2	0,005 0,979	0,522 0,002	0,230 0,206	0,217 0,234	0,678 0,000			
XRN2	0,214 0,239	0,582 0,000	0,196 0,284	0,350 0,050	0,549 0,001	0,459 0,008		
RAD21	0,421 0,016	0,618 0,000	0,319 0,075	0,117 0,522	0,693 0,000	0,533 0,002	0,522 0,002	
MAC30	-0,016 0,930	0,245 0,176	0,210 0,250	0,003 0,987	0,269 0,137	0,424 0,015	0,359 0,044	0,156 0,394

3.6.1.2 Evaluation of target gene expressions with clinical information

The clinical information of the patients were available (Table 2.2). An analysis was performed to investigate if the expression profiles of the target genes in breast tumors may contribute to the differentiation of any pathological group and whether this was statistically meaningful. Linear Discriminant Function Analysis (DFA) was performed (explained in 2.11.5) for 32 primary breast tumor samples to demonstrate if the target gene expression profiles in these tumors can discriminate and group the tumors for estogen receptor (ER) and grade status.

There were 13 ER(-) and 13 ER(+) tumor samples and 8 grade 1, 11 grade 2 and 11 grade 3. When all the target gene expression values were taken into consideration for ER status, DFA predicted the ER status correctly in 88% of the samples (Figure 3.15). ER(+) and ER(-) samples were correctly predicted in 85% and 92% respectively. The false prediction rate was 15% for ER (+) and 8% for the ER(-) tumor samples.

	True Group	
Group	**ER-**	**ER+**
0	11	1
1	2	12
Total N	13	13
N Correct	11	12
Proportion	**0,846**	**0,923**

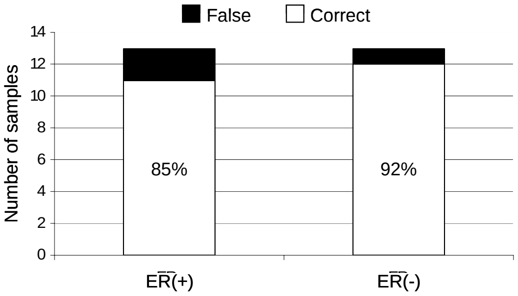


Figure 3.15: Discriminant Function Analysis of primary breast tumor tissues based on their ER status.

When all the target gene expression values were taken into consideration for grade status, grade 1 samples were correctly predicted in 88%, grade 2 samples in 64%, and grade 3 samples in 55%.

	True Group		
Group	**Grade 1**	**Grade 2**	**Grade 3**
1	7	0	1
2	0	7	4
3	1	4	6
Total N	8	11	11
N Correct	7	7	6
Proportion	**0,875**	**0,636**	**0,545**

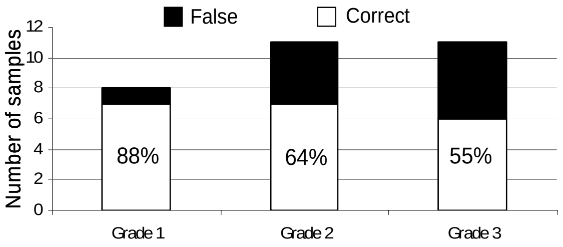


Figure 3.16: Discriminant Function Analysis of primary breast tumor tissues based on their grade status.

The data in the literature supports that the grade 2 breast tumors can be between grade 1 and grade 3 at the molecular level (Sotiriou *et al*. 2006). Linear discriminant function analysis (DFA) was performed only for the grade 1 and 3 tumors to determine if the expression of the target genes in these tumors can discriminate between these two groups. This time the prediction rate was increased and grade 1 samples were correctly predicted in 88% while grade 3 samples were correctly predicted in 100%.

	True Group	
Group	**Grade 1**	**Grade 3**
1	7	0
3	1	11
Total N	8	11
N Correct	7	11
Proportion	**0,875**	**1,000**

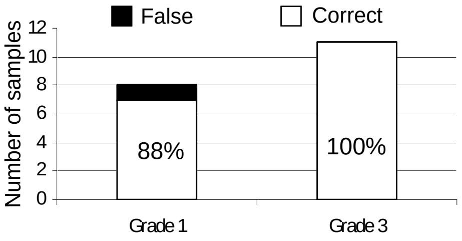


Figure 3.17: Discriminant Function Analysis of primary breast tumor tissues based on their grade 1 and 3 status.

3.7 Expression profiles and pathological contributions of BRCA1 and candidate BRCA1-target genes in independent microarray datasets

3.7.1 Tumor ER Status prediction in independent microarray datasets

The expression profiles of the target genes were able to discriminate the tumor samples from our cohort into ER(+) and ER(-) with high accuracy. To validate this result, two additional independent microarray datasets, which contain microarray data on ER(+) and ER(-) samples were analyzed (Wang *et al.,*2006 and Sitoriou *et al.*, 2006). Accordingly, the BRCA1-target genes were able to predict the ER(+) vs. ER(-) with high accuracies, ranging from 60 to 70%.

Among the genes tested ERBIN, RENT2 and BRCA1 were the ones significantly predicted the ER status (Table 3.8; p<0.05).

Table 3.8: Validation of the prediction power of BRCA1 and BRCA1 target genes on ER status of the tumors by two independent microarray datasets.

	ER(-) vs ER(+)			
	Wang *et al.*		Sitoriou *et al.*	
	p value	fold change (log2)	p value	fold change (log2)
ERBIN	1.00E-07	0.76	4.81E-02	0.87
RENT2	4.11E-04	1.19	1.64E-03	1.18
BRCA1	9.00E-02	0.86	2.55E-03	1.13

3.7.2 Tumor grade status prediction in independent microarray datasets

To determine whether the gene expression pattern of the 8 genes and BRCA1 could consistently predict the histological grade in an independent group of tumors, an independent microarray dataset (Sitoriou *et al.*, 2006; GSE2990), containing grade 1, 2 and 3 tumor samples, was used. Concordant with the literature, the power was not very high (55%) when the prediction was done between 3 classes of histological grade (1, 2, and 3). However, the gene list was able to predict the grades with a higher accuracy of 70% if the discriminant analysis was performed with 2 classes, grade1 and grade 3, (Table 3.9). When stepwise regression analysis was performed with our expression data and the data from Sitoriou *et al*., the RAD21 gene stood out as a predictor of grade status. RAD21 was consistently found to be the leading predictor of grade status as a result of binary tree prediction performed by the BRB array tools.

Table 3.9: Validation of the prediction power of BRCA1 and BRCA1 target genes on grade status of the tumors by two independent microarray datasets.

	Grade 1 vs Grade 3 Sitoriou *et al.*	
	p value	**fold change (log2)**
RAD21	7.80E-06	0.72
MAC30	1.29E-05	0.67
MAC30	2.30E-05	0.59
MAC30	7.79E-05	0.66
BRCA1	1.16E-03	0.88
RAD21	1.65E-03	0.65
RENT2	9.25E-03	0.87
BRCA1	1.31E-02	0.88
OVCA2	3.04E-02	1.07

3.7.3 Basal and non-basal breast tumor subtype prediction in an independent microarray dataset.

The new molecular markers were found in addition to the histopathologic classification of breast cancer and these markers were able to classify breast tumors into 5 molecular subtypes. Basal-like breast carcinomas, as defined by gene expression microarray analysis, are the most undifferentiated breast cancers, frequently lack the expression of hormone receptors (ER) and HER2 and patients with basal-like tumors experience a much shorter overall-and disease-free survival period. To assess if the expression pattern of the 8 genes and BRCA1 could differentiate the basal-like subtype from the non-basal like a microarray dataset (Richardson *et al.*, 2006; GDS2250), containing basal-like and non-basal like tumor samples, was used. As shown in Table 3.10, among the 9 genes analyzed 4 genes, ERBIN, SMG1, RAD21 and RENT2 were able to predict the tumor subtypes significantly with 78% accuracy (p<0.05)

Table 3.10: Prediction power of BRCA1 and BRCA1 target genes on basal and non-basal like breast cancer subtypes.

	Basal vs non-basal Richardson *et al.*	
	P value	**Fold change (log2)**
ERBIN	3.5E-04	0.6
SMG1	0.02	0.7
ERBIN	0.02	0.7
RAD21	0.03	1.4
RAD21	0.04	1.6
RENT2	0.06	1.4

3.7.4 Relapse-free survival analysis for BRCA1 and BRCA1-target genes.

Two independent microarray datasets Sitoriou *et al.*, 2006 (GSE2990) and Wang *et al.*, 2006 (GSE2034) were used to predict if the target genes used in this study has any prediction power in the breast cancer patient survival (Sitoriou *et al.*, 2006; Wang *et al.*, 2006). RAD21 was found to be significantly determinant of relapse free survival in two of the studies analyzed (Table 3.11) and additionally BRCA1 was included to the predictive list with high significance (p=0.009) in the study of Wang *et al.*

Table 3.11: The survival prediction analysis of BRCA1 and BRCA1-target genes

	Survival analysis	
	Wang *et al.*	**Sitoriou *et al.***
	p value	**P value**
RAD21	0.01	0.004
BRCA1	0.009	0.080

The Kaplan-Meier analysis performed with the target genes and BRCA1 showed that RAD21 and BRCA1 were together discriminative between high risk and low risk groups in the case of survival when the analysis was performed with the data of the GSE2034 dataset (Figure 3.18a). On the other hand, RAD21 was found to be the only predictor for the discrimination of the high risk group from the low risk group when the data from GSE2990 was analyzed (Figure 3.18b).

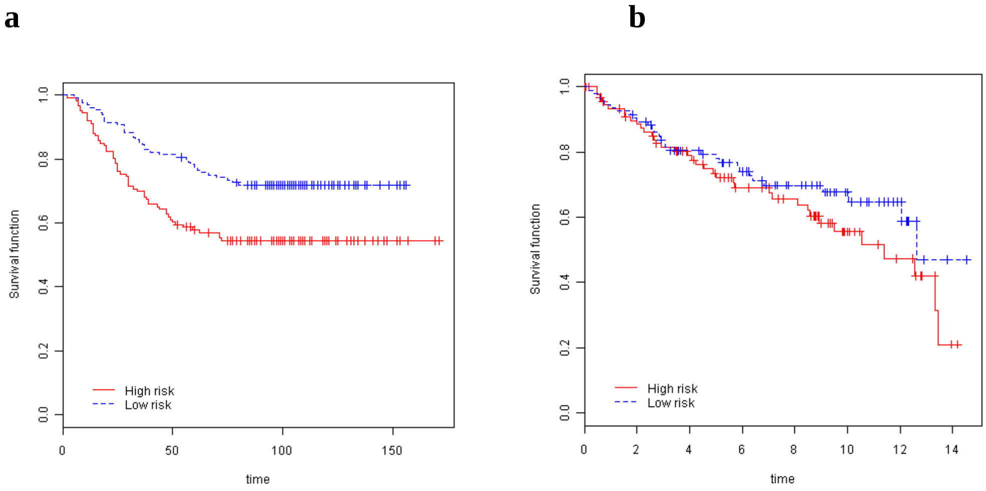


Figure 3.18: Relapse free survival analysis for GSE2034 and GSE2990 datasets. (a) Analysis of the GSE2034 dataset by the target genes. The time line was shown with the months. **(b)** Analysis of the GSE2990 dataset by the target genes. The time line was shown with the years.

3.7.5 Expression profiles of BRCA1 and candidate BRCA1-target genes in tumor and normal samples in independent microarray datasets

The expression profiles of BRCA1 and its eight potential targets were evaluated in tumor and matched normal samples in this study. Three independent microarray datasets, containing tumor and normal samples, were analyzed and the expression levels of these genes were extracted to see the expression patterns of these genes in other tumor and normal samples. Table 3.12 shows the expression levels of the genes and their ratios in logarithmic scale. The results from independent studies seem to be consistent for most of the genes. BRCA1, ERBIN, RAD21 and RENT2 were consistently overexpressed or unchanged while OVCA2 was downregulated in tumors compared to normal samples.

Table 3.12: The results of the expression profiles of 8 genes in GEO breast cancer microarray datasets. The mean expression levels of the genes in normal (N) and tumor (T) samples and their ratio were given on the log 2 scale.

Gene name	GSE8977 (Karnaub *et al.*)			GDS22509 (Richardson *et al.*)			GDS2635 (Turashvili *et al.*)		
	T	N	T/N	T	N	T/N	T	N	N/T
BRCA1	4.77	4.73	**0.03**	4.98	4.22	**0.76**	5.14	4.07	**1.07**
ERBIN	7.02	6.90	**0.12**	7.13	7.02	**0.11**	7.48	7.17	**0.30**
RAD21	11.26	9.63	**1.63**	11.26	9.63	**1.63**	9.58	8.82	**0.76**
XRN2	7.79	8.51	**-0.72**	8.02	8.08	**-0.06**	6.38	6.13	**0.26**
SMG1	7.53	7.53	**0.01**	7.67	7.10	**0.57**	5.70	6.20	**-0.50**
RENT2	6.84	6.75	**0.09**	8.45	8.27	**0.17**	8.29	7.78	**0.51**
OVCA2	6.40	6.50	**-0.09**	4.79	5.68	**-0.89**	6.22	6.87	**-0.65**
MAC30	6.32	7.49	**-1.18**	8.50	7.44	**1.06**	6.52	5.53	**0.98**

3.8 Depletion of BRCA1 gene expression with sh- and si-RNA knockdown

The expressions of eight selected target genes were previously shown to be increased depending on BRCA1 upregulation by SSH previously (Atalay *et al.*, 2002). These previous results were confirmed in two different cell line systems where BRCA1 upregulation resulted in increase in the target gene expressions with qRT-PCR experiments (Section 3.1.3). To have a better understanding of the BRCA1 effect on the target gene expressions, the BRCA1 gene expression was knocked down in MCF7 cells by using both siRNA which was commercially obtained from Dharmacon and shRNA which was kindly provided by Dr. Luc Gaudreau (Moisan *et al*, 2006). The expression of the target genes was then analyzed with qRT-PCR in these two systems.

Gene silencing and knockdown using RNA interference is becoming routine. The introduction of small interfering RNAs (siRNAs) into cultured cells provides a fast and efficient means of knocking down gene expression and has allowed siRNAs to quickly become a ubiquitous tool in molecular biology. siRNA has been shown to be effective for short-term gene inhibition in certain transformed mammalian cell lines, while shRNA offers an opportunity to potently and stably silence gene expression.

3.8.1 shRNA mediated knock down of BRCA1 gene expression

The MCF7 cells were transfected separately with pSUPER.retro.pro vector containing the inserts of shRNA sequence specific against BRCA1 and pSUPER.retro.pro vector containing scrambled sequence control plasmids. The MCF7 clones transfected with shRNA-BRCA1 were named as shB, while the clones transfected with shRNA-scrambled control were named as ctB. Eight cell clones from both transfection experiments were collected after one week of puromycin selection. The clones were expanded in the culture and RNA was prepared from each clone to check BRCA1 expression levels. Figure 3.19 shows the BRCA1 transcript levels in 8 scrambled shRNA control and 8 BRCA1 shRNA transfected MCF7 clones. MCF7 cells without any plasmid transfection were used as a mock control (mock1). Mock 1 was used as a calibrator to calculate the transcript levels of BRCA1 both in shB and ctB clones. Accordingly among the shB clones selected shB7 and shB15 were the ones having the least BRCA1 transcript levels. On the other hand ctB14 and ctB1 were the control clones found to have the highest BRCA1 expressions closed to the mock 1 level (Figure 3.19).

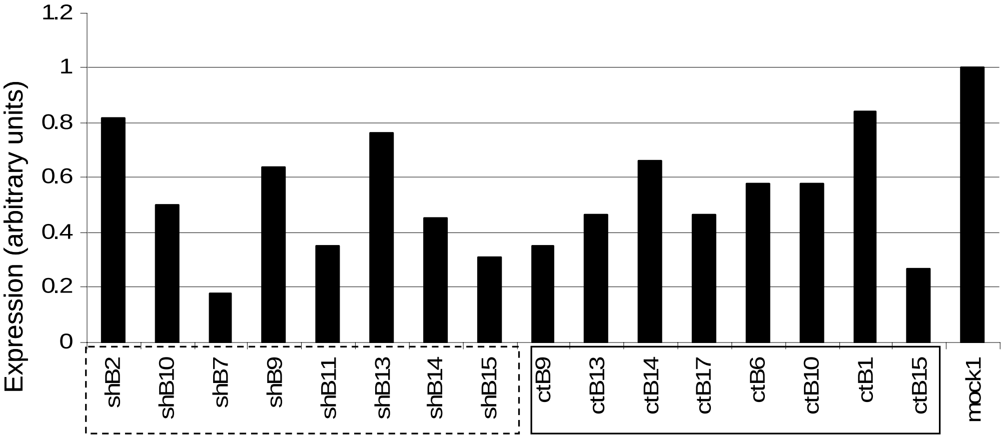


Figure 3.19: Expression level of BRCA1 in shRNA and scrambled control RNA transfected MCF7 cells. shB represents the clones stably transfected with shRNA against BRCA1 and ctB represents the cells transfected with scrambled control RNA. Expression values were calculated by comparing the expression level of BRCA1 in the transfected cells with that in mock cells (mock 1).

It was essential to analyze clones to observe the BRCA1 expression was also down regulated at the protein level. Three clones were selected according to their transcript levels obtained from qRT-PCR experiments and cell lysate was prepared from these clones to analyze the BRCA1 expression at the protein level with western blotting. The specific bands in the blot were quantified with the ImageJ programme. The results were then normalized first by normalizing the protein level of the clones to corresponding vinculin level and then shB7 BRCA1 level was normalized to ctB protein levels. It was found that BRCA1 protein was down regulated nearly 50% in the MCF7-shB7 compared to MCF7-ctB1 and MCF7-ctB14 (Figure 3.20a). Vinculin was used as an equal protein loading control (Figure 3.20b).

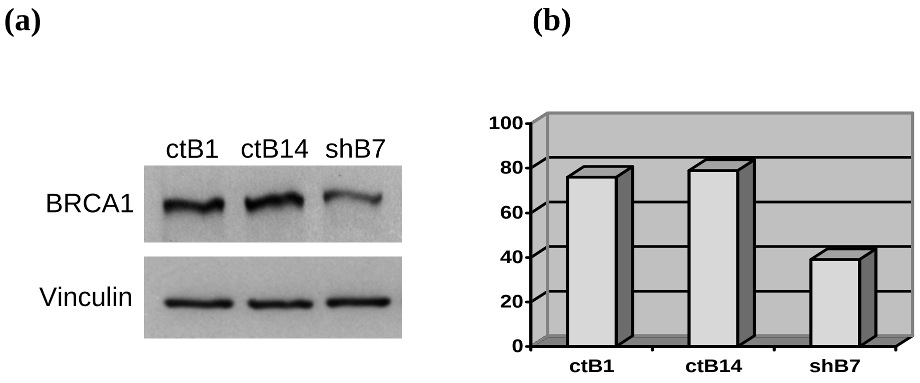


Figure 3.20: Immunoblot analysis of shRNA-BRCA1 stable MCF7 clones. The MCF7 cell extracts were prepared from stably tranfected clones. The ctB1 and ctB14 clones are from the MCF7 cells transfected with scrambled control shRNAs. The shB7 clone is from MCF7 cells transfected with shRNA against BRCA1. (a) BRCA1 protein level was detected with AB1 monoclonal antibody against BRCA1. Vinculin antibody was used for equal loading control. (b) The graphical representation of the quantified results of the immunoblot.

3.8.1.2 Analysis of the target gene expression in shRNA-BRCA1 depleted cells

In order to observe if the BRCA1 down regulation has any effect in the target genes, the shB7, ctB1 and 14 clones were used to analyze the expression levels of the target genes with qRT-PCR experiments.

The expression of ERBIN and SMG1 genes decreased nearly 40% in MCF7-shB7 clones compared to the control clones MCF7-ctB1 and MCF7-ctB14 (Figure 3.21). The expression levels of the other six target genes did not change in shB7 clone with respect to any of the ctB clones.

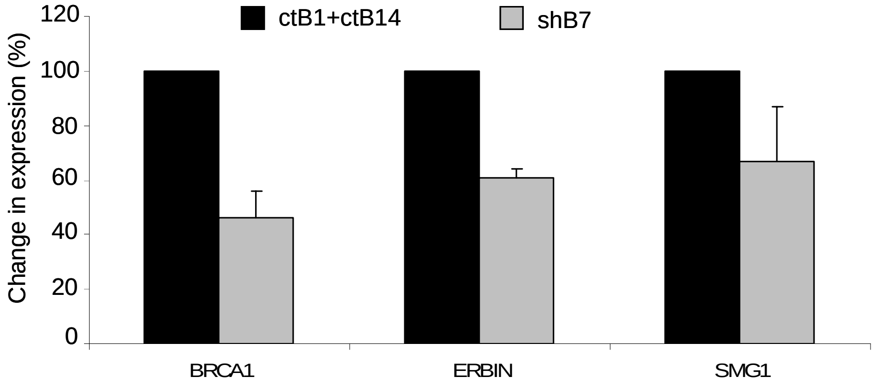


Figure 3.21: Expression levels of BRCA1, ERBIN and SMG1 genes in shRNA-BRCA1 MCF7 stable clones. The grey bars show the results from shB7 clone transfected with shRNA against BRCA1. The black bars show the results from ctB1 and ctB14 clones transfected with scrambled controls. The change in the mean expression levels of the genes in shB7 were calculated by using the expression levels of the genes in ctB1 and ctB14 as normalizers and the expression levels of the controls were set to 100%.

3.8.2 siRNA-mediated knock down of BRCA1 gene expression

The siRNA mediated knock down was used as a second approach to down regulate the BRCA1 expression level. The MCF7 cells were transiently transfected with siRNAs that targets the BRCA1 mRNA in three different regions and with scrambled control siRNA. According to the qRT-PCR results, nearly 80% decrease was observed in BRCA1 mRNA levels in transiently transfected MCF7 cells with siRNA directed against BRCA1 compared to scrambled siRNA transfected cells (Figure 3.22).

3.8.2.1 Analysis of the target gene expression in siRNA-BRCA1 depleted cells

The BRCA1 down regulation has any effect in the target genes, the BRCA1-siRNA transfected MCF7 cells were used to analyze the expression levels of the target genes with qRT-PCR experiments in order to observe if the BRCA1 down regulation has any effect on the target genes.

The expression levels of the genes in siBRCA1 transfected cells were compared to scrambled-siRNA and the quantification was done according to this comparison. This experiment was performed four times independently and the mean values, shown in Figure 3.22 were calculated. The results showed that BRCA1 down regulation mostly affected the expression of ERBIN and SMG1 by decreasing their expression levels nearly 60%. The expression levels of MAC30, RAD21, RENT2 and XRN2 genes were also decreased nearly by 40%. The effect of BRCA1 down regulation was moderate on tumor suppressor gene OVCA1 expression (10%) while no change was observed in OVCA2 expression (Figure 3.22).

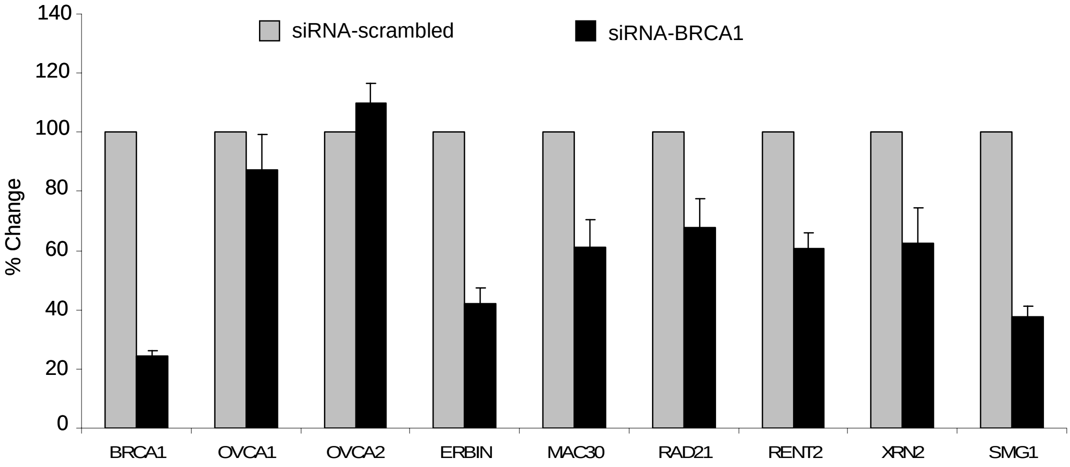


Figure 3.22: Expression profiles of target genes in siRNA-BRCA1 depleted MCF7 cells. The grey bars show the results from scrambled control siRNA transfected cells and their expression levels were set to 100%. The black bars show the results from siRNA-BRCA1 transfected cells. The mean values were calculated from four separate experiments.

Validation of siRNA-BRCA1 mediated knock down activity

The *GADD45* gene, which was found to be regulated by BRCA1, was used as a positive control to validate the functional activity of BRCA1 in BRCA1 down regulating cells (Harkin *et al.*, 1999). BRCA1 has previously been shown to be necessary for the activation of the *GADD45* promoter transcription-activating function. We showed in the BRCA1 induction experiments that the *GADD45* expression level was increased with the increased expression of BRCA1, which confirmed the functional activity of BRCA1.

The expression level of GADD45 gene in mock, siRNA-scrambled transfected and siRNA-BRCA1 transfected MCF7 cells was analyzed with qRT-PCR. The mean values were calculated from four separate experiments.

The down regulation of BRCA1 level was 80% compared to both mock and siRNA-scrambled cells while the GADD45 gene expression was 40% reduced compared to the control cells (Figure 3.23).

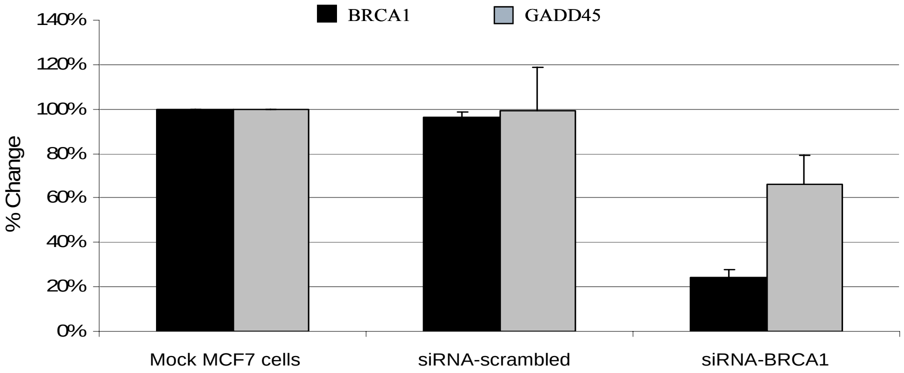


Figure 3.23: GADD45 gene expression was affected by BRCA1 depletion. The expression levels of both BRCA1 and GADD45 were the same in siRNA-scrambled compared to mock cells. BRCA1 level was decreased 80% in siRNA-BRCA1 transfected cells while the down regulation was 40% in GADD45 expression in the same cells. The mean values were calculated from four separate experiments.

The effect of siRNA-BRCA1 mediated knock down activity on an independent gene set

The expression levels of four reference genes and a random gene *CST6* was analyzed in siRNA transfected MCF7 cells in order to investigate whether the effect of silencing of the BRCA1 in MCF-7 cells was specific to BRCA1 targets or systemic (Figure 3.24). It was observed that none of the genes expressed BRCA1 lower than mock MCF7 cells. Compared to control cells *CST6* and *PGK1,* expression levels seemed to be lower in siRNA-BRCA1 transfected ones, but the levels of the genes in control cells were observed to be higher than mock controls.

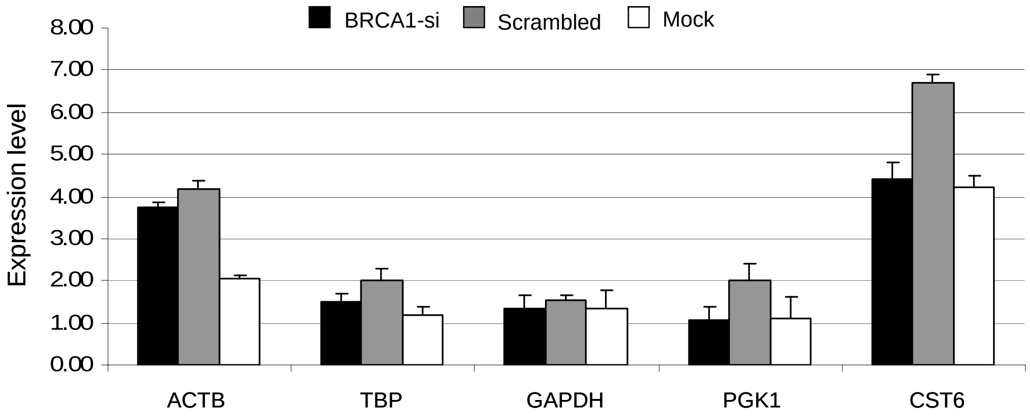


Figure 3.24: The expression patterns of ACTB, TBP, GAPDH, PGK1 and CST6 genes in BRCA1 downregulated MCF7 cells.

The induction and depletion experiments showed that the expression patterns of BRCA1 and the target genes were directly proportional to each other. When there was an increase in the expression of BRCA1 gene, the expression of the target genes was also increased. When the expression of BRCA1 was decreased, the expression of at least seven target genes was decreased (with the exception of OVCA2). This similar behavior in the expression patterns of BRCA1 and the target genes, which was obtained from *in vitro* experiments, directed the study to another level at which the expression patterns of these genes were investigated *in vivo* and their correlation status of them were questioned.

3.9 The effect of BRCA1 on ERBIN expression

We demonstrated that the ERBIN gene expression was up- and down-regulated in the pCMV-mycBRCA1 transfected and sh- or siRNA downregulated MCF7 cells respectively with qRT-PCR and the ERBIN gene expression was consistently correlated with BRCA1 expression independent of the normalization method in breast tumor and normal samples.

An expression analysis was also performed in breast cancer cell lines which have functional BRCA1 (BT20, MDA 453, MDA 231 and T47D) and in HCC1937 that has non-functional BRCA1. A normal mammary epithelial cell line HME1 was also included in the qRT-PCRexperiments. The results showed that there was a significant correlation between the expression of ERBIN and BRCA1 in the breast cancer cell lines (r=0.8; p=0.02) (Figure 3.25).

a

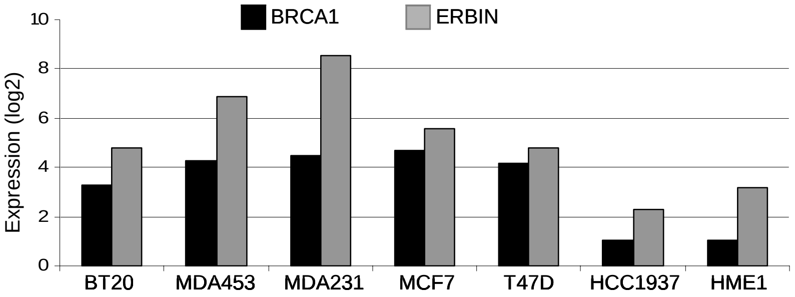


Figure 3.25: Expression pattern of *BRCA1* and *ERBIN* in breast cancer cell line panel. Expression profiles of the ERBIN and BRCA1 in the breast carcinoma cell lines were performed with qRT-PCR. The GADPH expression value was used for normalization of the target gene expression in each cell line.

We were able to obtain a small amount of ERBIN-specific antibody from Dr. Jean-Paul Borg (INSERM, Marseille, France) and used it for the analysis of ERBIN at the protein level with Western blot analysis. The ERBIN expression was shown to be upregulated 2 fold at the protein level in the pCMV-mycBRCA1 transfected MCF7 cells compared to the cells transfected with an empty pCMVmyc vector (Figure 3.26).

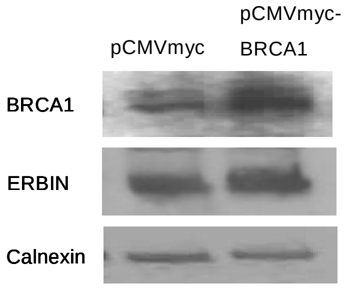


Figure 3.26: BRCA1 expression changes ERBIN expression. Western blot analysis of the BRCA1 and ERBIN in pCMVmyc control and pCMVmycBRCA1 transfected MCF7 cell lysates. AB1 antibody is BRCA1 specific antibody. Calnexin antibody was used for equal loading control of the cell lysates.

3.10 Promoter region analysis of BRCA1-target genes

BRCA1 is known to be a transcription factor. Although a direct binding sequence of BRCA1 on DNA was still not identified it is known to regulate some promoters through other transcription factors. ZBRK1, OCT-1 and NF-YA were the known transcription factors physically interact with BRCA1 (Fan *et al.*, 2002).

In order to find out whether the BRCA1 target genes have any common sequence on their promoter regions, which could be a possible candidate sequence for BRCA1 binding, we analyzed the regulatory regions of selected BRCA1 target genes. The MEME Suite Motif-based sequence analysis tool was used for this purpose [http://meme.sdsc.edu/meme4_1/intro.html]. For each gene, the region spanning -1000 to -1 was downloaded from UCSC genome browser, human March 2006 assembly in FASTA format.

The MEME Suite Motif-based sequence analysis tool was used to search for any common motifs among the regulatory regions (-1000, -1) of the selected genes and

MAST was used as a MEME tool for the motif alignment. MAST takes a file as an input containing the descriptions of one or more motifs and searches a sequence database that you select for sequences that match the motifs. The motif file can be the output of the MEME motif discovery tool or any file in the appropriate format. The parameters used for MEME was as fallows:

- Motif width from 6 to 12
- Maximum number of motifs to search 100 with reverse complementary strands considered
- MAST parameters were considered ($E<1.0$ and $p<0.005$).
- If the p value has $p<0.005$ the sequence motif returned by MEME as significant.

Since we could not come up with any common sequences or motifs among those genes we searched for the regulatory regions of the genes to see if there was any known transcription factor binding motifs or known BRCA1 binding sequences like Oct-1 and CAAT on their promoter regions. Since the proximity (closeness) of the binding sites is important for the transcription factors to be located on the DNA sequences, the proximity of these two motifs was also searched in the regulatory regions. Figure 3.27 shows the locations where Oct-1 and CAAT locate with maximum 50 bp proximity on the regulatory regions of XRN2, SMG1 and ERBIN. The sequence search was performed in the transcriptional regulation region of the genes spanning the -1000 +500 nucleotides. A significant hit was obtained from the XRN2 promoter region and there was 12 bps between the two motifs. ERBIN and SMG1 were also containing these two motifs on their regulatory regions which made them possible targets for further evaluations.

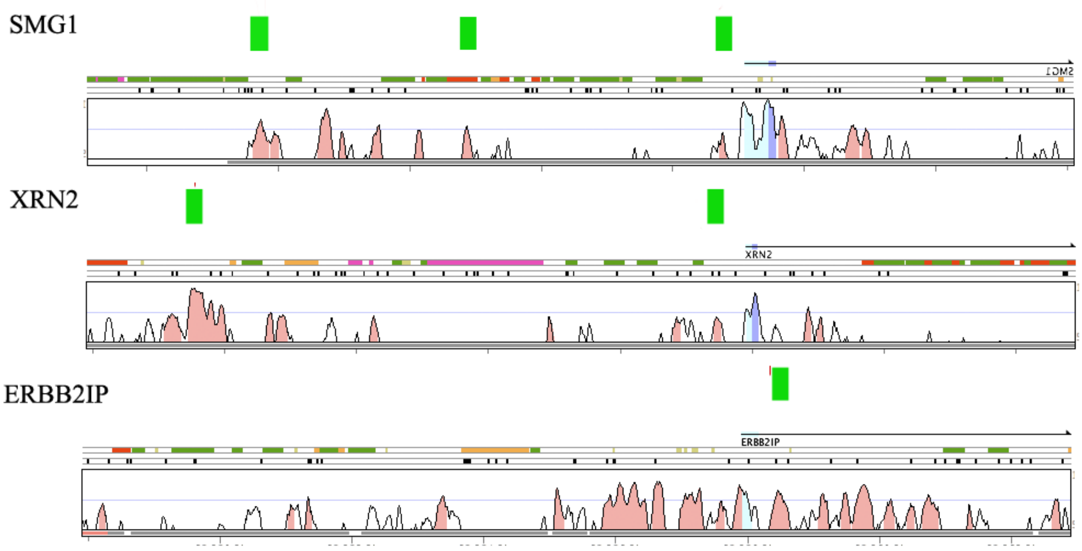


Figure 3.27: Promoter region analysis of the genes. Green bars show the locations where Oct-1 and CAAT are located a maximum of 50 bp apart from each other. The region under survey spans 1500 bp, which is1000bp upstream of the transcription start site (TSS) and 500 bp downstream of the TSS.

Since *ERBIN* was found to be the most significantly correlated gene with *BRCA1*, an additional promoter region analysis was performed for the regulatory region of *ERBIN* and all the known target sequences known to be bound by *BRCA1* targets were searched with FOOTER, a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting (Corcoran *et al.*, 2005). Figure 3.28 shows the locations of the known sequences present on the promoter region of *ERBIN.*

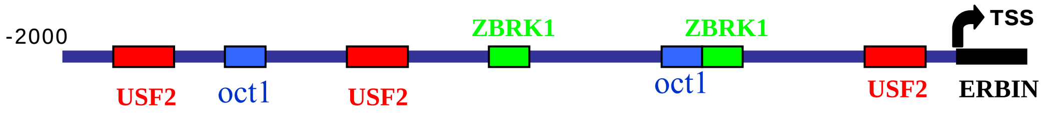


Figure 3.28: Promoter region analysis of ERBIN. The region searched was spanning the 2000 bp upstream of TSS. The boxes show the regions of DNA sequences specific to transcription factors Oct-1, ZBRK1 and USF2.

3.11 The data retrieved from the Stanford Microarray Database (SMD)

Two independent microarray gene expression data sets, Sorlie *et al*. (Sorlie *et al,* 2003) and Zhao *et al*. (Zhao *et al*., 2004), were downloaded from the Stanford Microarray Database (SMD). Both of the data were log transformed and then median centered (median normalized) arraywise by using the gene filtering options of SMD. Genes missing more than 20% of the expression data were excluded from the analysis. Since the genes on the arrays were represented with more than one probe probe IDs were used instead of gene names and the two datasets were combined with respect to probe IDs using a set of customized perl routines (Appendix A). These two data sets combined resulted in an initial list of 4769 IMAGE clones (3465 unique genes) common in both datasets. The Sorlie dataset used for further analysis was contained the expression values of 3465 genes in 101 IDC, 8 ILC and 4 normal samples. This data is available at
http://www.biomedcentral.com/imedia/1073571582428031/supp1.xls
On the other hand the Zhao dataset contained the expression values of 3465 genes in 38 IDC, 21 ILC and 3 normal samples. This data is now available at
http://www.biomedcentral.com/imedia/7518530352428028/supp2.xls.

3.12 Correlation of Sorlie and Zhao Datasets

Combining the datasets in meta-analysis requires that they have similar expressions, both in magnitude and individual variability. To assess whether the Sorlie and Zhao datasets were correlated, the Pearson's correlation coefficient was calculated between the mean expression values of the ductal or lobular samples from each dataset, respectively before and after performing t-tests (Figure 3.29). Even before the removal of IMAGE clones showing significant differences between the studies, the mean expression values of ductal samples from Sorlie were highly correlated with those from Zhao; and a similar result was observed for the lobular samples (r = 0.8329 and 0.8233, respectively). After filtering out the differentially expressed IMAGE clones, the correlations between the aforementioned datasets increased to 0.9389 and 0.8465 for the ductal and lobular samples, respectively. These results ensured that there was significant correlation between the Sorlie and Zhao datasets although they were based on independent tumor and normal samples and different experimental procedures.

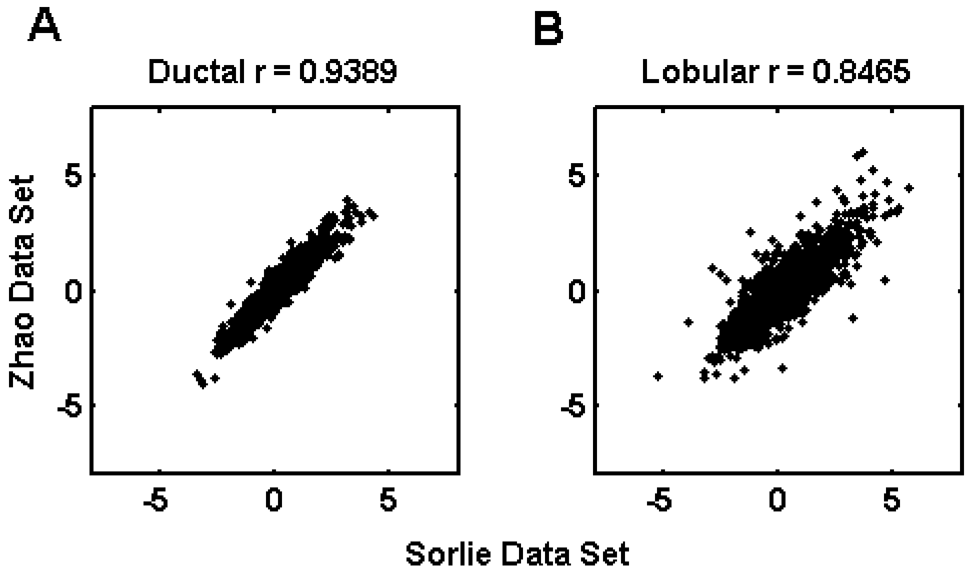


Figure 3.29: Pearson correlation coefficients (r) between Sorlie and Zhao datasets. Correlation plots between datasets after differentially expressed IMAGE clones were filtered out based on t-tests. (A) Correlation between mean expression values of ductal samples ($p<0.05$). (B) Correlation between mean expression values of lobular samples ($p<0.05$).

3.13 The effect of different normalization strategies on the data distribution

Different studies can be normalized and directly compared to each other in meta-analysis. Our comparisons based on Pearson's correlation before and after t-test eliminations ensured that there was a significant correlation between the Sorlie and Zhao datasets although these studies were based on independent tumor and normal samples; and the experimental procedures (e.g., amplification of RNA) also varied considerably between the two studies (Figure 3.29). The choice of normalization type is an important step to overcome the bias and systematic differences between samples; the same slides or between slides which do not represent the true biological variation between samples. Global normalization enforces the chips to have equal median intensity while quantile normalization enforces the chips to have identical intensity distribution. The data were normalized using both global and quantile normalizations in this study. Since the global median-normalized and quantile-normalized data correlated well (Figure 3.30), the former normalization method, with the least number of data manipulation steps was used before combining these two datasets.

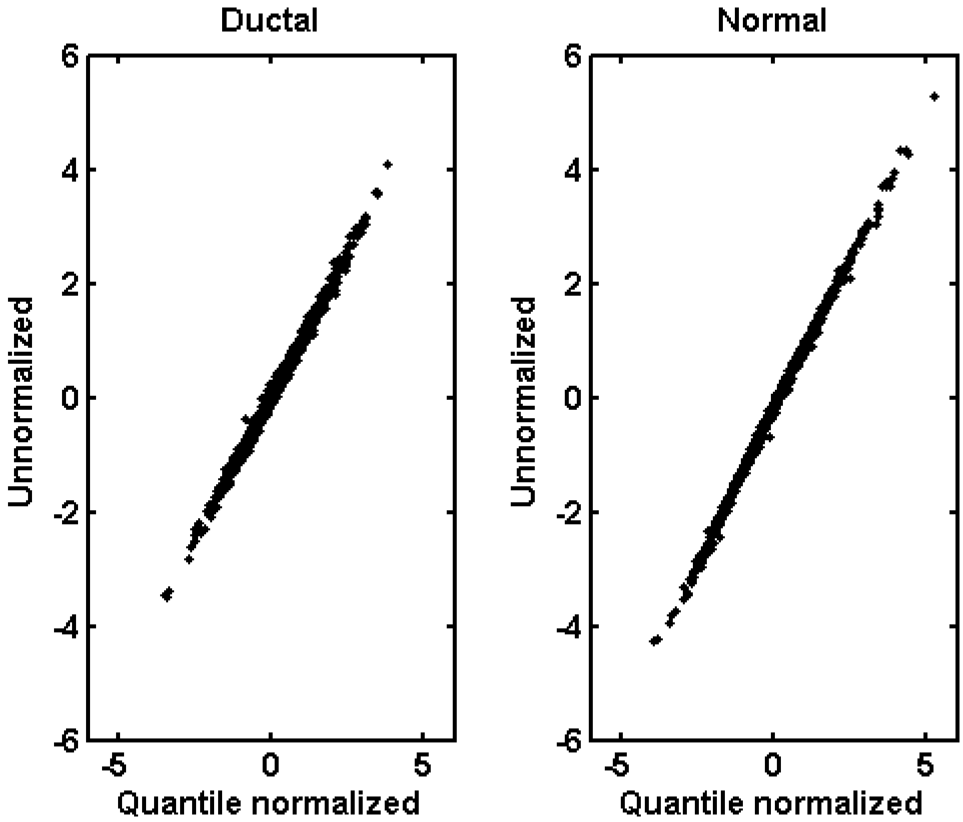


Figure 3.30: Correlation of global median and quantile normalized data. The figure shows the correlation of global median and quantile normalized data of ductal (D) and normal (N) tissue samples.

3.14 Distribution statistics for generation of meta-lists

Global-median normalized and filtered datasets were used in this study since they minimized the number of manipulations performed during gathering of the meta-data (Figure 3.30). Accordingly, assessment of significance was based on p-values obtained from the Kolmogorov-Smirnov analysis between test and random distributions (*pt* and *pr1*, respectively) of a gene in the meta-data. For example, the GSN gene had a highly significant differential expression between ductal and normal samples as evidenced by the highly skewed distribution towards lower p-values whereas the RAP2A gene exhibited a uniform distribution of p-values (Figures 3.31 A, B and 3.31 C, D, respectively).

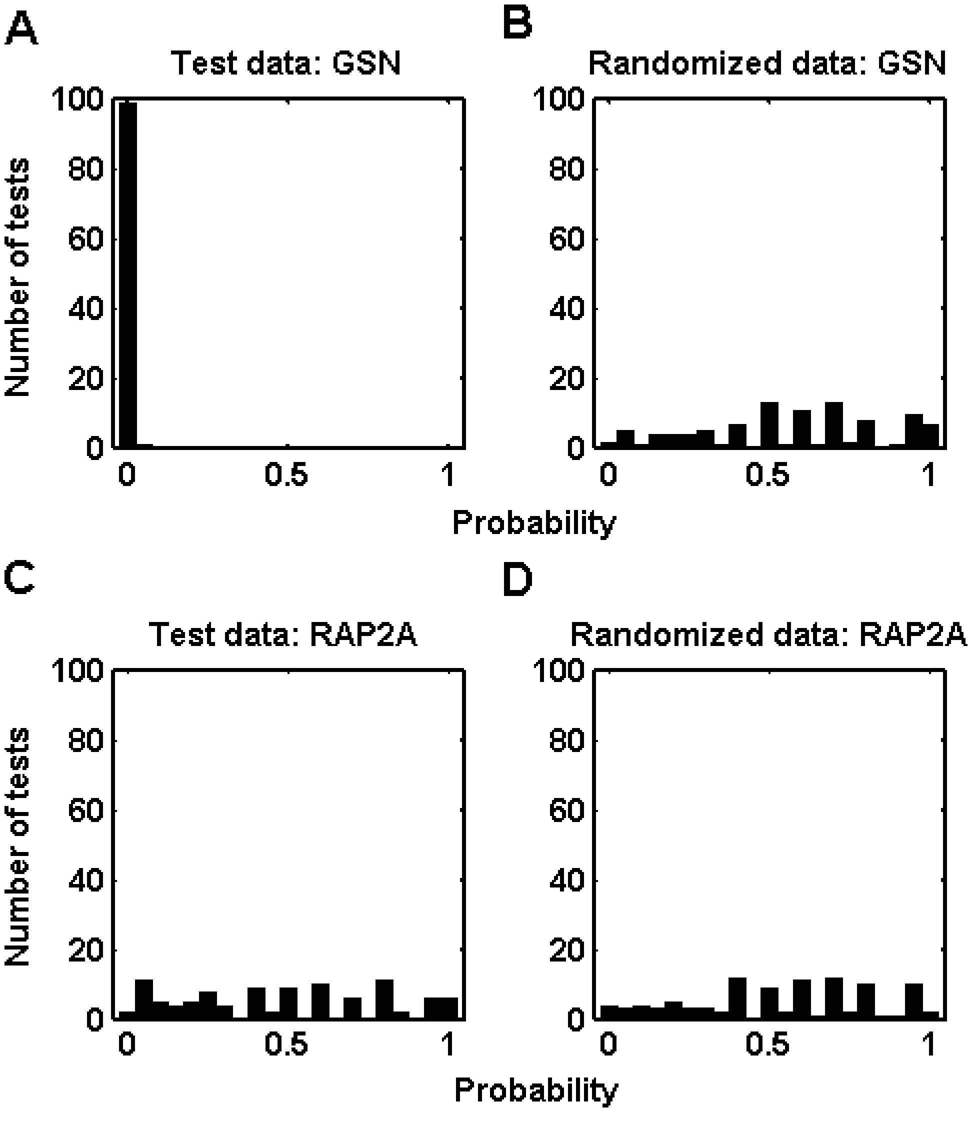


Figure 3.31: Examples for probability distributions of Wilcoxon rank sum tests. Data were obtained where resampling size, *n*, equaled 6 (100 iterations). Assessment of significance was based on p-values obtained from the Kolmogorov-Smirnov test between test and random distributions (*pt* and *pr1*, respectively). (A, C) For test data,

GSN gene had a highly significant differential expression (significant at 100% of iterations, p = 0.00) between ductal and normal samples whereas RAP2A gene did not (significant at 5% of iterations, p = 0.98). (B, D) Probability values of both GSN and RAP2A, obtained from randomized data, were uniformly distributed. GSN; IMAGE: 214990 and RAP2A; IMAGE: 36684.

3.15 Effects of resampling on estimates of expression and differentially expressed gene number

Resampling was performed for a particular sample size *n* (e.g., 3), repetitively for *i* number of times, iteratively, where *i* = 10, 20, 30, …, 100 and 150. The *n* was set to be 3, 4, 5, 6, 10, 15 and 20 for ductal vs. lobular comparison. On the other hand, since the total number of normal samples was 7, the highest sampling value could be set to 6 for ductal vs. normal and lobular vs. normal comparisons, and *n* equaled 3, 4, 5 and 6. These sample size-iteration combinations led to 77 runs for ductal vs. lobular analysis, and 44 runs for ductal vs. normal and lobular vs. normal analyses. The effect of sample size and number of iterations on the estimation of mean expression level and the number of differentially expressed genes were tested. For each run performed with a different sample size, the change in grand mean of expression (i.e., mean expression of all IMAGE clones) as well as the number of differentially expressed IMAGE clones were plotted with respect to the increasing number of iterations (Figure 3.32). As the number of iterations increased, the grand mean became more stabilized. As expected, the magnitude of change in mean values asymptotically decreased as the number of iterations and sampling size increased (Figure 3.32 A and 3.32 C). On the other hand, the number of genes stated as significant increased as a function of the number of iterations and sampling size (Figure 3.32 B, 3.32 D). Significant IMAGE clones made up more than 70% of all analyzed genes at sampling size 6 with the highest iteration in ductal vs. normal analysis whereas the same set-up resulted in only 20% significant IMAGE clones in ductal vs. lobular analysis.

It is reasonable to assume that use of a single sample size and iteration number may not be adequate to understand the variability among the tumor samples (Figure 3.4). It might instead be beneficial to consider all of the information gathered from the individual runs. Accordingly, the significant gene lists reported in this study were obtained by taking only those IMAGE clones that were assigned as significant in a given set of all resampling analyses performed (90% or more for ductal-normal, DN; and lobular-normal, LN; and 80% or more for ductal-lobular, DL comparisons) in an effort to minimize the effects of sampling size and iteration number on p-values.

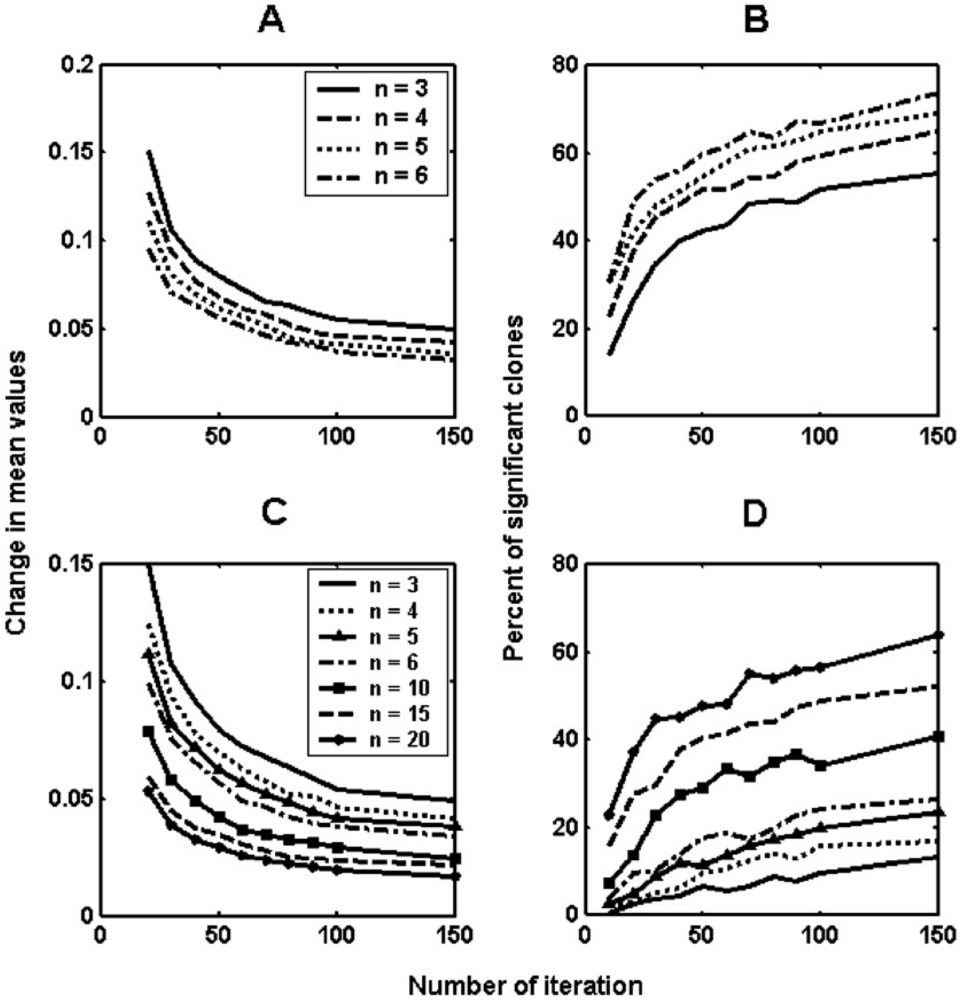


Figure 3.32: Effect of change in sample size and number of iterations on mean expression values and number of significant IMAGE clones. For each of the runs performed with different sample sizes (n), the change in the mean expression value (A, C) and the number of IMAGE clones that were stated as differentially expressed (B, D) were plotted with respect to the increasing number of iterations. A and B refer to the results of ductal vs. normal analysis whereas C and D show the results of ductal vs. lobular analysis.

3.16 Characteristics of differentially expressed meta-gene lists

Differentially-expressed gene lists for DN and LN contained 298 (282 genes) and 216 (202 genes) IMAGE clones, respectively (Appendix B1 and B2). On the other hand, there were only 66 (65 genes) differentially expressed IMAGE clones between the ductal and lobular (DL) datasets for 80% criteria (Appendix B3). The size of these lists was dependent on the False Discovery Rate (FDR) input value (herein set to 0.01) or the percentage of resampling runs considered for significance (i.e., 90% or 80%). In order to obtain a larger number of genes for DL analysis, the significance percentage value was set to 80.

The same resampling procedures were also performed on the individual datasets, Sorlie and Zhao, separately. Compared to our meta-analysis these separate analyses together could provide 91% of IMAGE clones that were present in the significant DN list and LN list and 68% of the IMAGE clones of the DL list. However neither of the studies could supply 9% of the IMAGE clones of the DN and LN list and 32% of the DL list (90% cut-off), each of which corresponds to a novel contribution by our meta-analysis (see Appendix B4 for meta-analysis specific gene lists)**.**

The final DL significant gene list was also compared with the list of 52 genes reported by Zhao *et al*. (Zhao *et al*., 2004). The DL list shared CDH1, AOC3, FADS2, SORBS1, ALDH1A1, LPL, ANXA1 and AKR1C1 with that of Zhao *et al*. (Zhao *et al*., 2004). However, our analysis did not assign reasonable significance to F11 and VWF genes according to the set cut-off criteria (80%). The remaining genes in the Zhao gene list were not encountered since they were not included in the combined dataset used in the present meta-analysis. Meta-analysis of these two datasets provided a total of 36 significant genes not previously reported by Zhao *et al* and when either dataset is analyzed individually (Appendix B3).

3.17 Validation of tumor vs normal meta-gene lists by independent microarray datasets

Recent meta-analysis studies identified common cancer signatures by combining microarray datasets from different tissues for increasing accuracy of tumor vs.

normal class prediction (Rhodes *et al.*, 2004; Xu *et al.*, 2007). In this study, we focused on extracting a stable tumor molecular signature based on two of the existing breast cancer studies that contain microarray data on normal, IDC, and ILC tissue samples. We have also validated the predictive power of the meta-gene lists obtained through the resampling-based meta-analysis using three additional breast cancer datasets, which contain microarray data on 3 or more samples of normal and tumor breast tissues (Table 3.13) (Turashvili *et al.*, 2007; Richardson *et al.*,2006; Karnoub *et al.*, 2007). Subsets of genes from DN and LN meta-gene lists were accordingly able to predict the tumor vs. normal classes with high accuracies, ranging from 80 to 100% (Table 3.13). Strikingly, correlation between expression values obtained from significant discriminators from each of the three normal/tumor datasets and those from the meta-analysis was high (Table 3.13). This indicated that the DN and LN lists harbored a robust expression profile for the breast tumors when compared with normal breast tissue.

3.18 Prediction of tumor-subtypes

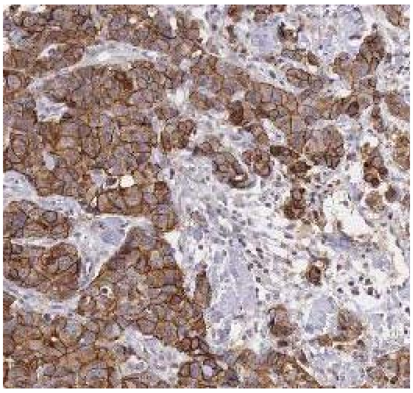
We extracted a small, highly correlated classifier gene subset which was commonly detected among the three microarray studies and the meta-analysis to identify a more conservative gene set differentially expressed between tumor and normal cells (Appendix C). Twenty-eight genes from the DN or LN meta-gene lists intersected with the three other microarray datasets (GDS2635, GDS2250, and GDS1329); 17 of which discriminated between basal vs. non-basal and/or ER status (Appendix C). For example, *ADAMTS1*, *ATF3*, *IGFBP6*, *PRNP*, *EGFR*, *FN1*, *ID4*, *SPTBN1*, and *SFRP1* genes from the DN list were found to significantly discriminate between nonbasal-like vs. basal-like tumors as well as basal and luminal subtypes of the breast tumors ($p<0.05$). Besides subtype prediction, all of the above genes except FN1 were found to be significantly predictive of the tumor ER status ($p<0.05$; Appendix C1).

Table 3.13: Summary of GEO breast cancer microarray datasets and results of class prediction analysis for the meta-gene lists, DN (Ductal/Normal) and LN (Lobular/Normal). Normal (N) and tumor (T) sample sizes, accuracy of prediction from binary tree algorithm (% accuracy), and the number of genes in classifier (number of genes) were shown for each study, separately. Correlation (r_{DN}, r_{LN}) of the classifier expression from each study with the DN and LN meta-gene expressions also were indicated (Pearson correlation, Minitab®; p<0.001).

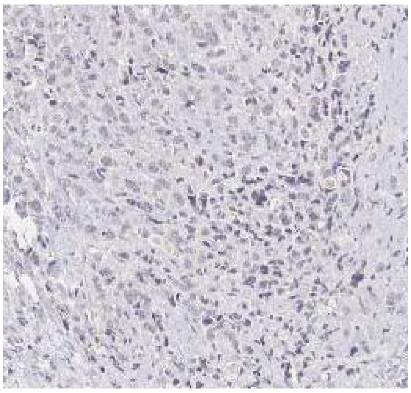
Study GEO ID	Class		Meta gene-list					
			DN			LN		
	N	T	Accuracy (%)	Number of genes	r_{DN}	Accuracy (%)	Number of genes	r_{LN}
Turashvili [5] GDS2635	10	10	93	57	0.85	80	49	0.87
Richardson [7] GDS2250	7	40	100	145	0.86	100	96	0.78
Karnoub [8] GSE8977	15	7	95.5	109	0.72	95.5	89	0.81

3.19 Validation of ductal vs lobular meta-gene list

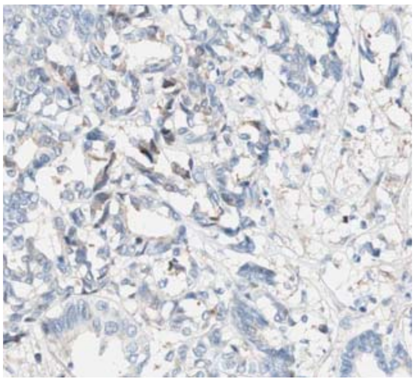
Comparison of fold-change values of the DL meta-gene list consisting of 65 genes with that of the Turashvili's DL list (GDS2635) resulted in a high degree of correlation (r=0.53; p<0.001), suggesting that the direction and magnitude of expression change between the IDC and ILC samples were largely consistent between data from different microarray experiments. Furthermore, we combined published expression data from IDC and ILC samples from experiments performed by Bertucci *et al* (Bertucci *et al.*, 2008) with the meta-analysis results (Appendix D). Some of the members of the 65 meta-gene list were consistently down- or up-regulated also in the Turashvili and Bertucci datasets (i.e., down-regulated ALDH1A1 and RBP4 in IDC; and up-regulated CDH1 and TFAP2A in IDC). Protein expression levels of these four genes were investigated using the Human Protein Atlas, a public resource for immunohistochemistry (IH) of normal and pathological human tissues (http://www.proteinatlas.org/). IH data were available for CDH1, TFAP2A, and RBP4 proteins; and only data from antibodies exhibiting differential expression among breast tumors were reported herein. The representative IHC staining slides taken from Human Protein Atlas were given in Figure 3.33. Accordingly, 2 out of 3 ILC samples exhibited moderate to strong signals for RBP4 (Antibody CAB00455) whereas 7 out of 9 IDC samples were either negative or had weak staining. CDH1 data in Protein Atlas database was not very informative since number of ILC samples were limited, however, a moderate signal was detected for the ILC sample whereas 5 out of 6 IDC samples expressed CDH1 strongly (Antibody CAB000087) Similarly, TFAP2A was weakly or moderately expressed in the two ILC samples examined whereas a moderate to strong staining was observed in 5 of the 9 IDC samples. Although sample size was limited for the ILC samples in the Human Protein Atlas database there was a corresponding trend between the mRNA levels reported by the present study and the protein level assessment obtained from the Human Protein Atlas.



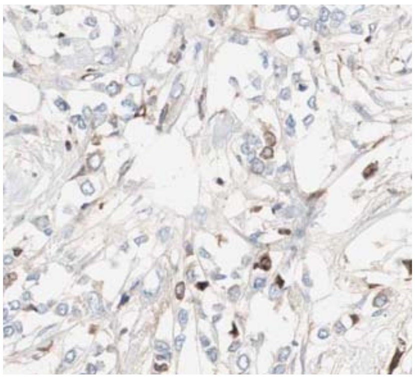
CDH1 staining of IDC sample



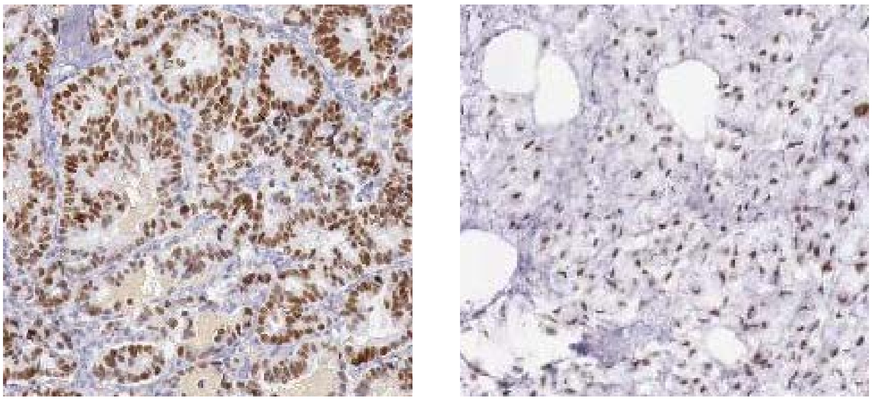
CDH1 staining of ILC sample



RBP4 in IDC



RBP4 in ILC



TFAP2A in IDC TFAP2A in ILC

Figure 3.33: Immunohistochemistry results of CDH1, RBP4 and TFAP2A in breast cancer tissues. The results were obtained from the Human Anatomy Atlas. **A,B:** shows the expression pattern of CDH1 in IDC and ILC samples. Strong reactivity is detected in IDC but no positivity is observed in ILC. **C, D:** Expression pattern of RBP4. ILC sample showed a strong RBP4 staining and no positivity observed in IDC sample. **E.F:** TFAP2A is strongly positive in IDC sections but weakly stained in ILC cells.

3.20 Validation of meta-analysis by real time qRT-PCR

Nine genes from the meta-gene list that were found to be differentially expressed in both the DN and LN lists (except MAF) were selected for validation of the meta-analysis. Expression profiles of these genes were tested in independent paired IDC breast tumor and non-tumor tissue samples through real time qRT-PCR. The results obtained through qRT-PCR were consistent with those of the meta-analysis such that GSN, SPTBN1, SFRP1 and MAF were down-regulated in most tumor samples with respect to their matched non-tumor samples whereas COX6C, RAD21, GSPT1, NME1 and PTTG1 were up-regulated (Figure 3.34). Additionally seven other genes ATF3, ADAMTS1, EGFR, PRNP, IGFBP6, ID4 and FN1, with predictive potential for tumor subtype and ER+/ER- classification were selected from the tumor-specific differentially expressed gene-set. All except *FN1* were found to be down-regulated in tumor samples with respect to their normal counterparts. The meta-analysis results were supported by the real-time qRT-PCR experiments since all tested genes exhibited differences between matched normal and tumor samples in the same direction as expected by the meta-analysis (Pearson correlation coefficient, r = 0.78, p = 0.001).

Among the genes we used for validation through real time qRT-PCR, *ID4* was the gene found to be differentially expressed between DN only by meta-analysis rather than each study alone.

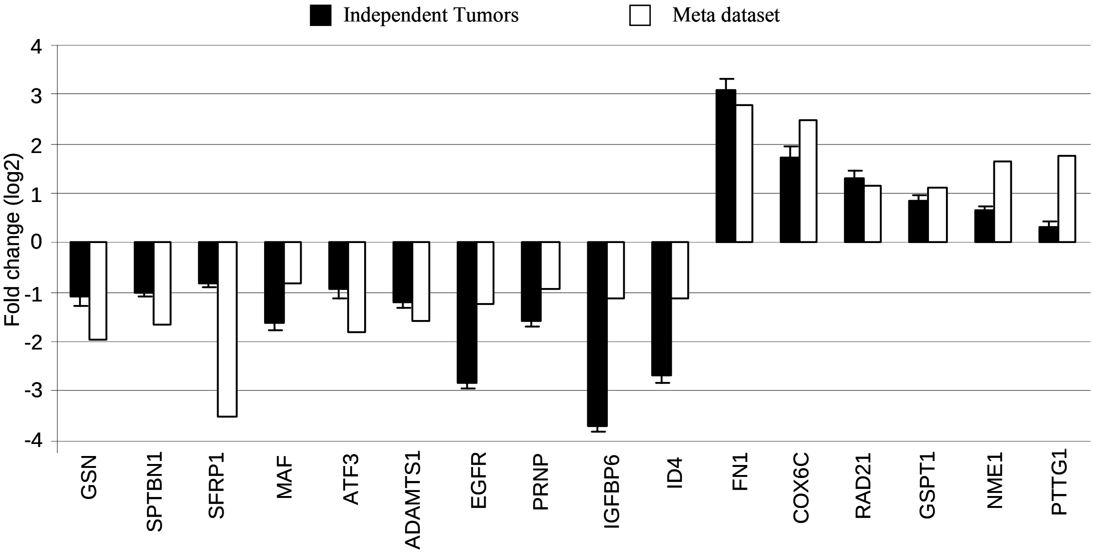


Figure 3.34: Validation of meta-analysis results by real-time qRT-PCR. Sixteen genes were selected from the ductal-normal (DN) significant meta-gene list for real-time qRT-PCR. Solid black bars refer to mean expression values (± SEM) of 10 independent IDC breast tumors normalized to their non-tumor pairs. White bars refer to the mean expression values from the combined meta-gene list.

CHAPTER 4. DISCUSSION

PART 1: BRCA1-INDUCED GENE EXPRESSION PROFILES

BRCA1 possesses a number of features common to transcriptional regulatory proteins, suggesting that it may regulate the expression of one or more downstream genes (Venkitaraman, 2002; Rosen *et al*., 2003). It is important to determine which genes are trancriptionally influenced by BRCA1 *in vivo* to explain its role in tumor suppression and in cancer development. In previous studies BRCA1 over-expression systems enabled the researchers to define the genes whose expression levels were upregulated with the overexpression of BRCA1 (Harkin *et al*., 1999; MacLachlan *et al*., 2000; Aprelikova *et al*., 2001; Atalay *et al*., 2002). Atalay *et al.* generated BRCA1 over-expression systems in MCF-7 breast cancer cells by using the PCR-dependent Suppression Subtractive Hybridization (SSH) technique (Atalay *et al*, 2002) and found 60 genes, which were likely to be regulated by BRCA1.
Herein we investigated the expression profiles of some of the genes, which were selected from the former BRCA1-induced gene list (Atalay *et al*, 2002), (OVCA1, OVCA2, ERBIN, RAD21, XRN2, RENT2, SMG1 and MAC30) in normal-matched primary breast tumors. Furthermore, correlations with the gene expression profiles of selected target genes with *BRCA1* expression and with various pathology parameters, namely, tumor grade, stage, ER, PR and ErbB2 status were analyzed. The same set of genes was also tested in breast carcinoma MCF7 cells in which BRCA1 was stably or transiently down regulated.

The expression profiling of eight genes and BRCA1 was identified in 32 normal-matched tumor samples by real-time qRT-PCR, which is a valuable and powerful technique to get accurate results from expression studies. Getting all the technical steps right is an obvious precondition to achieve a meaningful result in qRT-PCR and all the requirements and necessities needed for a sensitive and specific qRT-PCR reaction were taken into consideration and each critical step was carefully evaluated. The interpretation of the data was done meticulously to get accurate results.

One of the important requirements in accurate interpretation of the results using qRT-PCR is finding the endogenous reference genes that do not change their expression across the tissue type used when target gene expression is evaluated. Therefore we tried to find out the most stable reference genes across tumor and matched-normal breast tissues.

Identification of endogenous reference genes for qRT-PCR analysis in normal-matched breast tumor tissues

Real-time RT-PCR is attractive for clinical use since it can be automated and performed on a variety of tissues, fresh or archived, paired or unpaired. However, accurate quantitative analysis of gene expression levels with qRT-PCR can only be obtained by using appropriate reference genes (RGs) for normalization procedures. As no universal RG exists, it is inevitable to search for stably expressed genes for normalization purposes in each experimental condition, such as tumor versus normal breast specimens, to get reliable results from relative expression experiments (Jung *et al*., 2007; Ohl *et al*., 2006; Saviozzi *et al*., 2006).

In this expression analysis study, matched tumor and normal beast tissues were used. Since there was no study in the literature questioning the best reference genes for matched breast tissues, a systematic comparison of frequently used reference genes (RGs) was performed and their utility as internal controls for accurate relative gene quantification in tumor and matched normal breast tissue samples for qRT-PCR studies assessed.

The following measures were evaluated to increase the accuracy and reliability of the data in this study: (1) matched pairs of normal and tumor breast samples were used for minimization of inter-individual variation and to increase the power of data analysis; (2) total RNA was assessed stringently and only the high quality samples were included in the study; (3) the 18 candidate RGs were simultaneously analyzed with optimized conditions; (4) the tumor and normal-matched samples were included in the same run in duplicates for a studied gene; and (5) established software combined with statistical analysis was used to rank the candidate RGs for their suitability as normalization factors (NFs). Additionally, it was shown that the expression of the RG set in breast tumors did not exhibit differences in terms of grade, ER, or PR status and age of the individuals when normalized to their matched controls. This is important in clinical use since the selected RGs can be used in all malignant samples independent of the tested clinical parameters.

In the present work, 15 of the commonly used RGs and 3 newly selected candidates were analyzed to find out the most suitable ones as NF for relative gene quantification in paired breast tumor/normal gene expression profiling. The candidate reference genes used in this study have independent functions in cellular maintenance. This is important since the selection of genes that share identical biochemical pathways could bias analysis. To constitute the candidate reference gene panel in this study we first searched for the frequently used genes as references for qRT-PCR studies in breast cancer. While ACTB, TBP, and GAPDH were commonly used as normalization factor GUSB, B2M and PPIA have also been used in breast cancer studies (Folgueira *et al*., 2006; Parr *et al*., 2006; Wu *et al*., 2005; Shim *et al*., 2006; Morse *et al*., 2005; Kroupis *et al*., 2005; de Cremoux *et al*., 2000; Potemski *et al*., 2006; Iwao *et al*., 2000; Oshiro *et al*., 2005; Zhang *et al*., 2005). As a second approach we identified candidate genes, SDHA, PGK1, HMBS, HPRT, RPL41, and YWHAZ, as being used in different studies dealing with the identification of suitable reference genes for any human tissues in addition to being also recommended by geNorm. We included 3 more genes, RPLP0, MRPL19 and PUM1 in our study as they were reported to be the stable genes in breast cancers by two other studies that were investigating the endogenous control reference genes for gene expression normalization in breast cancer (McNeill *et al*., 2007; Lyng *et al*., 2008). The genes, TTC22, ZNF224, and IL22RA1 that were selected by analyzing the publicly available breast cancer microarray data-sets were also included in the panel as new candidate reference genes.

Our findings indicated that raw Ct values obtained from this RG set were highly correlated with each other although they were not necessarily functionally related. On the other hand, the raw Ct values obtained by using a set of randomly primed cDNA samples showed that although the correlation between two RNA polymerase

II transcribed genes, ACTB and SDHA, was still reserved (r=0.8, p=0.001), the correlation of expression from either of these two genes with the RNA polymerase I transcribed 18S rRNA gene expression was not significant (r=0.034, p=0.912; r=0.206, p=0.499). Concordant with these results, the previous studies indicated that a large number of housekeeping genes transcribed by RNA polymerase II behaved similarly among themselves (de Kok *et al.*, 2005; Lyng *et al.*, 2008), which may explain the possible reason for this correlation.

All the RGs studied here exhibited relatively higher expression in tumors than their normal counterparts. Similarly, it was reported that breast biopsy samples exhibited great intra- and inter-individual variability and mean expression values of tumors measured in copy numbers were greater than those of their normal counterparts (Tricarico *et al.*, 2002). Because of the extensive variability in RG expression, total RNA-based (or mRNA copy numbers when available) normalization was suggested as an NF for tumor samples (Tricarico *et al.*, 2002). However, since total RNA is represented mostly by rRNA (>90%), even a small decrease in rRNA expression may lead to a disproportional increase in the mRNA pool estimation (Elberg *et al.*, 2006; Spanakis, 1993). Moreover, studies have shown that rDNA genes were methylated in breast and ovarian cancers when compared with those of normal controls (Chan *et al.*, 2005; Yan *et al.*, 2000). In fact our finding of low tumor rRNA to mRNA ratio suggests that normal and tumor samples are heterogeneous in total RNA fractions. We found that 69% of breast tumors (9/13) exhibit dramatically lower expression of 18S rRNA as compared to their non-tumor pairs while mRNA expression of widely used housekeeping genes ACTB and SDHA in the same set of tumors was higher (84%, 11/13).

These recent findings suggest that normalization based on a proper set of endogenous RGs obtained from equal amounts of total RNA/input material might be the optimal approach for comparing tumor specimens. Our findings indicated that estimation of mRNA from total RNA represented an important issue requiring further investigation in qRT-PCR studies. Since rDNA hyper-methylation holds considerable possibility in breast tumors and total RNA is largely made up of rRNA, the use of poly(A)+ RNA as a starting material may be another approach for studying tumor and their matched normal samples.

Moreover, studies have shown that rDNA genes were methylated in breast and ovarian cancers when compared with those of normal controls (Chan *et al*., 2005; Yan *et al*., 2000).

Since it has been reported in the literature that rDNA genes were methylated in breast cancer when compared with those of normal controls, we wanted to investigate our the breast tumor tissue samples for methylation status. Initially, we selected two tumor samples that showed low expression of 18S rRNA expression with qRT-PCR. These tumor DNA samples were used for bisulfite tratment and amplified with the methylation specific primers that were designed for the 45S rRNA gene promoter region. We found that these two samples showed a high level of methylation in the CpG islands of the 45S rRNA promoter. Although these results are very encouraging and indicate that the low expression level of 18S rRNA might indeed be due to the methylation of the 45S rRNA promoter, the number of breast tumor samples should be increased and also the normal breast tissue CpG islands should be evaluated.

In order to increase the reliability of the endogenous RG selection process, we analyzed the expression stability of the 18 selected RGs with two different statistical models: a pair-wise comparison model, geNorm, and an ANOVA-based model, NormFinder. The results obtained from the two programs were consistent for the most and least stable gene selection. ACTB and SDHA were found to be the most stable RGs while IL22RA1 was the least stable among the 18 genes selected for these analyses.

17 out of 18 reference genes in our panel displayed a consistent 1.86±0.7 (log2, mean±std) fold expression difference between breast tumor and normal pairs suggesting that there might be a more generalized mechanism reflected in the breast samples. One possibility is that all these genes although with unrelated functions and chromosomal locations are upregulated in tumors but such global deregulation is unlikely considering that many of these genes have been reported previously as stable housekeeping genes. Alternatively tumor and normal samples might consist of heterogeneous rRNA and mRNA compartments affecting estimation of the amount of mRNA from the total RNA pool. In support of this possibility we found that a significant portion of tumors had lower levels of 18S rRNA than normals.

Furthermore, recent literature has supported our finding as RNA hypermethylation has been shown in breast tumors (Chan *et al*., 2005; Yan *et al*., 2000).

Recent studies suggested that the variation in the average of multiple genes was smaller than the variation in individual genes. Therefore, it is an optimal approach to use multiple RGs rather than a single gene as NF. Normalization to geometric mean of more than one control gene compensates for outlying values of single RGs in individual samples and may therefore more accurately reflect transcript abundances of target genes (Vandesompele *et al*.,2002).

Our results suggested that increasing the number of RGs stabilized the ranks of tumor samples among normalized gene expression values yet adding a third gene was not as critical as adding the second gene. This is in accordance with the findings of Vandesompele *et al.*, who states when NF_n and NF_{n+1}, where *n* represents the number of genes used in normalization, do not significantly differ in their effect, using NF_n might offer a more economical choice (Vandesompele *et al*.,2002). Accordingly, the two best genes, ACTB and SDHA, can be used as NF, and additionally more genes, MRPL19, GUSB, TBP and PGK1, identified by both programs might be combined with the two best genes to be used as NF.

In the present study, we compared the expression values of the gelsolin gene by using single or different combinations of the best-ranked RGs. When the GSN expression was normalized with ACTB and SDHA alone, the fold change values were significantly correlated with each other, yet the degree of correlation increased when two best performing genes ACTB and SDHA were used as NF. Addition of more best-performing RGs (MRPL19, GUSB, TBP, and PGK1) did not improve the degree of correlation results more than 1%.

GSN expression is known to decrease in breast tumors when compared with normal breast tissues. The adverse effect of using the least stable RG (IL22RA1) was highly significant, and there was a substantial error associated with the estimation of the relative GSN gene expression in breast tumors compared to their normal counterparts.

Considering that the housekeeping mRNA expression studied here might not actually be unregulated but over-estimated due to a rRNA bias, exclusion of this bias may actually correct the potential underestimation of mRNA amount estimation between tumors and their matched normals. We calculated this possible error as 1.16 (log2 difference) for tumor-non tumor bias from the expression data obtained by using 18S rRNA from randomly primed subset of tumor-non tumor pairs. 17 out of 18 RGs in our panel displayed on average, a 1.86 fold expression difference between tumor and normal pairs, of which 1.16 fold might be attributable to rRNA/mRNA bias. If RG normalization not performed, then it is likely that GSN expression in tumors would be overestimated at least 1.16 fold.

The present study focused on identification of RGs for paired tumor/normal breast tissue based on the ranking agreement between commonly referred normalization software, geNorm and NormFinder and expression results of GSN, a well-known down regulated target gene in breast tumors. Although this panel is highly comprehensive and consists of frequently used reference genes, they may still not be the best applicable reference genes for breast cancer normalization studies unless there is a bias due to RNA estimation or breast tissue heterogeneity since all the genes in our panel showed higher expression in tumors than in their normal pairs. However, ACTB and SDHA were consistently found to be the least variable genes between tumor and normal pairs with two programs, geNorm and NormFinder, in this panel.

In conclusion, the results indicated that normalization of target gene expression levels to a normalization factor consisting of the geometric mean of two best performing genes, ACTB and SDHA, offers increased accuracy and resolution in the relative quantification of gene expression in breast tumors with respect to their matched normal tissues. Future studies are needed to establish the percentage of tumors with such rRNA/mRNA bias and the underlying causes such as methylation patterns of rDNA.

Expression pattern analysis of BRCA1 induced genes in BRCA1 up- and down-regulated cells

Eight candidate BRCA1 target genes (OVCA1, OVCA2, ERBIN, SMG1, RENT2, XRN2, RAD21, MAC30) were selected from the list of genes obtained from a previous study (Atalay *et al*, 2002), which were found to be upregulated with the overexpression of BRCA1 in breast cancer cells. To validate the effect of BRCA1 on the selected targets in expression level, two different BRCA1-overexpressing systems were established. The transient one was pCMVmyc-BRCA1 transfected MCF7 cells and the stable cells were, UBR60-bcl2, osteosarcoma cells with tightly regulated BRCA1 expression with the tetracycline inducible system. In both of the systems we could induced the expression of BRCA1 and the target genes were observed to be upregulated at least 1.5 fold (except MAC30 in UBR60-bcl2 cells.). However the effect of BRCA1 on its targets seemed to be more stable in MCF7 cells while the response of the targets were fluctuating and even the expression of MAC30 was not altered in the tetracycline inducible BRCA1-osteosarcoma cell system. This could be due to the fact that this effect of BRCA1 could be tissue or breast cancer specific.

In order to evaluate the expression profile of the target genes in the absence of BRCA1, sh-RNA and si-RNA mediated knock down of BRCA1 were performed. The reduction of BRCA1 expression was more than 60% both in the transcript and protein level. Among the candidate BRCA1 targets, ERBIN and SMG1 were the two genes most affected from this depletion. Since SMG1 is one of the members of the nonsense mediated decay (NMD) process (Yamashita *et al.*, 2001; Ivanov *et al.*, 2008), the regulatory effect of BRCA1 on SMG1 may suggest that BRCA1 may alter the NMD process upon DNA damage leading to accelerated degradation of truncated proteins that could be toxic to cells.

On the other hand ERBIN is an ErbB2 binding protein and the PDZ domain of the ERBIN binds preferentially to the C terminus of ErbB2, which is non-Tyr1248-phosphorylated (Borg *et al.*, 2000). It is important that phosphorylation of this residue following ErbB2 activation is a critical event for the mitogenic signaling and oncogenicity of this receptor (Dittmar *et al.*, 2002). Thus the downregulation of ERBIN by BRCA1 silencing may inhibit the binding of ERBIN to ErbB2 and since the Tyr1248 residue of the ErbB2 is free from Erbin it tends to be phosphorylated which may explain the proliferation of the cancer cells in the absence of BRCA1.

BRCA1 depletion may effect different genes' expressions since it may cause a global gene repression in the cells. In order to investigate this, we used five independent genes that take role in different pathways to investigate their expression profile in si-BRCA1 depleted cells. There was no change in their expression compared to the scrambled controls in MCF7 cells. Although we can not rule out its effect in number of genes that play role in different pathways, we may conclude that down regulation of BRCA1 expression does not affect the gene expression globally.

Our results showed that both shRNA and siRNA-mediated reduction of BRCA1 expression down regulated the expression levels of target genes while induction of BRCA1 expression induced them. This two-sided effect of BRCA1 suggests that BRCA1 is a possible regulator of these selected target genes.

qRT-PCR based expression profiling of BRCA1 induced genes in primary breast tumors

Despite being expressed ubiquitously in adult tissues, germline mutations in *BRCA1* predispose individuals only to breast and ovarian tumors with only minor effects on the predisposition to cancer in other sites (Hu Y., 2009). Finding out the reason for this tissue specificity of BRCA1 may be possible by investigating its potential transcriptional targets in breast tissues. To clarify this idea in more detail, the expression profiles of these eight genes and BRCA1 were investigated in breast tumor and matched normal samples. This would be valuable to understand the role of BRCA1 more in breast cancer development or progression.

The expression levels and profiles of the BRCA1 target genes in primary breast tumors and matched-normal samples were detected by qRT-PCR. Since the normalization was an important step for accurate data analysis, reference gene selection was done carefully and ACTB, TBP and SDHA genes were used in combination as a normalization factor (Pfaffl, 2001; Vandesompele *et al*., 2002). Next, the normalized expression values of the genes from the tumor samples were normalized to that of their normal counterparts. The aim was to subtract the noise coming from the normal cell population. The results obtained through this analysis were used for the clustering of tumor samples. Primary breast tumors were hierarchically clustered into two major groups based on their gene expression profiles normalized to their matched normal counterparts (Figure 3.15). The pathological parameters were tested among these two groups but none of the parameters, which were grade, ER, PR, ErbB2, stage, and lymph node status, were able to explain this classification.

On the other hand, the correlation coefficients were calculated between BRCA1 and target genes to see their relation in primary breast tumors. As a result, three of the target genes, ERBIN, RAD21 and SMG1, were found to be highly correlated with BRCA1 (Pearson correlation, Minitab; N=32; r = 0.427, r = 0.421 and r = 0.343, respectively; p<0,5).

Such a strong correlation of BRCA1 with SMG1, one of the nonsense-mediated decay (NMD) genes, may suggest the presence of a role of BRCA1 in the regulation of NMD process in cancer cells. On the other hand RAD21 is a nuclear phosphoprotein that repairs the double strand breaks (DSB) and is a component of the cohesion complex that holds sister chromatids during mitosis (Chen *et al*., 2002). It thus plays a role in DNA damage response and chromosomal structure maintenance. As BRCA1 deficient cells suffer from both DSB repair deficiency and chromosomal instability (Wang *et al*., 2000), the ectopic overexpression of BRCA1 or its downregulation may be responsible for the inefficient expression of RAD21, which is a potential target of BRCA1 in those pathways.

Since ERBIN could be a potential regulator protein in breast cancer because of its role in the localization of ErbB2 to the basolateral domain in epithelia, which is important for its activation and signaling of ERBB2/HER2 in epithelia, its correlation with BRCA1 was discussed more in the following sections.

Expression profile predicts the ER status of the primary breast tumors

It is well known that estrogen signaling plays a significant role in the development and progression of breast cancer. The role of estrogen in sporadic breast cancer development has been widely studied but there are still gaps in the relation of estrogen signaling and BRCA1-related tumors. The striking restriction of BRCA1-

related tumors to hormone-responsive tissues (breast and ovary) is a strong clue for the connection between BRCA1 and hormone signaling (Rosen *et al.*, 2003).

In order to investigate how gene expression profiles of the target genes contribute to the ER status classification, linear discriminant function analysis (DFA) was performed (explained in 2.11.5) for 32 primary breast tumor samples. There were 13 ER (-) and 13 ER (+) tumor samples in the tissue panel used. When all the target gene expression values were taken into consideration, DFA predicted the ER status correctly in 88% of the samples (Figure 3.17). ER(+) and ER(-) samples were correctly predicted with 85% and 92% respectively. The false prediction rate was 15% for ER (+) and 8% for ER(-) tumor samples. Concordant with the literature, the samples expressing high levels of BRCA1 were ER(+) and the tumors which had downregulation in the expression of BRCA1 were ER(-) and this difference between the expression levels of BRCA1 in ER(+) and ER(-) tumor samples were statistically significant (t-test, p=0.005).

Although the crosstalk between the BRCA1 and ER signaling pathway is still not clear, many studies were done to identify the reason of the ER negative nature of BRCA1-related breast cancers. In one model, it was suggested that the ER(-) breast cancer cells may be derived from ER(+) cells (Hu Y., 2009). In support of this idea, Li *et al.* reported that mice carrying conditional Brca1 knockout in their mammary gland tumors that were ER(+) at early stages became ER(-) at later stages (Li *et al.*, 2007). The genomic instability of the breast cancer cells in early stages may lead to estrogen independent proliferation and result in an increase in the ER(-) cell population as they have growth advantage over hormone dependent cells.

Additionally, the reciprocal activation of BRCA1 and ER expression was reported in two different models. In the "ER activates BRCA1" model, it was suggested that ER(+) cells have additional protection in the DNA damage response because of elevated levels of BRCA1. In contrast ER(-) cells have no added genome protection from ER-mediated BRCA1 elevation and have the advantage in cancer initiation (Marquis *et al.*, 1995; Gudas *et al.*, 1995; Spillman *et al.*, 1996). On the other hand, the "BRCA1 activates ER expression" model proposes that loss of BRCA1 in precancerous cells would lead to genome instability as well as ER negativity. Concordant with the result of Li *et al.*, they suggested the possibility that ER negativity is a result of cancer rather being a cause for initiation (Hosey *et al*., 2007). It could also be stated that lacking ER could be a byproduct of the cells lacking functional BRCA1.

All these models proposed to explain the interplay between BRCA1 and ER pathways, may be inspirational to explain the prediction power of ER status of the tumor samples with the expression profiles of BRCA1 and its transcriptional targets.

Expression profile predicts grade 1 and grade 3 in primary breast tumors

Histological grade in breast cancer provides clinically important prognostic information. About half of the breast cancers are assigned grade 1 or 3 and a substantial percentage of tumors, 30-60%, are classified as grade 2 (Sotiriou *et al*. 2006). Grade 2 is found to be associated with an intermediate risk of recurrence and is thus not informative for clinical usage. We examined whether the histological grade was associated with gene expression profiles of eight BRCA1 target genes and with BRCA1. DFA results showed that BRCA1-target genes were able to discriminate grade 1 and grade 3 tumors. Stepwise regression analysis was performed in order to find out which gene was more powerful in that discrimination and the RAD21 gene was found to be the one responsible for grade discrimination ($p<0.05$).

Analysis of target genes in independent microarray datasets for grade prediction

We also used an independent microarray study (Sotiriou *et al*., 2006) to analyze if the target genes and BRCA1 expression values can predict grade 1, 2, and 3 in this study. The ID numbers of the target genes were found and expression values were pulled out for each tumor sample from this study and used for the binary tree prediction analysis. The analysis by using the expression values of these genes from this study also gave the same results and RAD21 was found to be one of the genes significantly predictive for grade 1 and 3 tumor discrimination (7.80E-06). Additionally MAC30, BRCA1, RENT 2 and OVCA2 were also found to be significant ($p<0.03$). Although additional confirmation studies are needed, these concordant results showed that RAD21 could be a powerful candidate for molecular grade prediction in breast cancer.

The significant contribution of BRCA1 and its targets in different molecular events in breast cancer progression such as being an ER status marker or histological grade predictor may elucidate the unreported roles of BRCA1 in breast cancer.

Analysis of target genes in independent microarray datasets for basal and non-basal breast cancer subtypes

We wanted to analyze if the target genes used in this study has any prediction power in basal- and non-basal breast cancer subtypes. The basal type of tumors were characterized with low BRCA1 expression, ER(-), and ERBB2 (-).

The expression microarray study reported by Richardson *et al*. was able to define gene sets that can predict the basal and non-basal breast tumor subtypes (Richardson *et al.*, 2006). We extracted the expression values of the target genes from this study and used them to define tumor subtype information from the same data and applied binary tree algorithm to analyze their prediction power for basal and non-basal subtypes of breast cancer.

The genes which were found to be highly correlated with BRCA1 (ERBIN, SMG1 and RAD21) in our study were also able to classify the tumors into the basal subtype of breast cancer, which has been shown to have the same pathologic characters as BRCA1 related tumors (Turner and Reis-Filho, 2006).

Analysis of target genes in independent microarray datasets for survival prediction

Additionally, we also performed an analysis using the results of the Sitoriou *et al.*, 2006 (GSE2990) and Wang *et al.*, 2006 (GSE2034) microarray studies to predict if the target genes used in this study had any prediction power in breast cancer patient survival (Sitoriou *et al.*, 2006; Wnag *et al.*, 2006). In this study, the authors used 189 number of patients diagnosed with breast cancer and followed these patients between the date of surgery and the date of diagnosis of any type of relapse to predict the genes and their expression level that could contribute to the survival rate of the patients. First, we found the eight target genes and BRCA1 gene ID numbers used in the study and pulled out their expression data for each patient included in this microarray analysis study. Since the survival data for these patients were also available, we used the expression values of these genes to be able to predict their contribution to the survival of the patients by using the survival prediction tool of BRB-ARRAYTOLS. The RAD21 gene that was found to be highly correlated with BRCA1 in our study was also found to predict the patient survival as a result of this analysis (p<0.01).

It is also noteworthy to mention that the survival prediction power of BRCA1 and RAD21, which was confirmed with two independent microarray studies, again emphasizes the significant role of BRCA1 and its targets in breast cancer progression as well as its development.

Correlated expression of BRCA1 and ERBIN

Among the list of selected genes, the expression of ERBIN was found to be highly correlated with that of *BRCA1* both in BRCA1 over-expressing and down-regulating cell lines and this correlation was validated with the experiments performed with a breast cancer cell line (BCC) panel. ERBIN was found to be highly expressed with BRCA1 in the BCC panel. In HCC1937, which has no functional BRCA1 protein, the expression level of ERBIN was also decreased with that of BRCA1 to a level that was under the expression level of both genes in HME1 and BCC panel.

Additionally the correlation was repeatedly observed when their expression profiles were analyzed in normal-matched tumor samples. After normalization, ERBIN expression was found to be highly correlated with that of BRCA1 in breast cancer tissues. Erbin was expressed at a higher level than their matched normal tissues in more than half of the tumors. Concordantly Liu *et al*. reported the same finding for an independent breast tissue panel (Liu *et al*., 2008)

In light of these findings, the target ERBIN gene promoter was analyzed to obtain some clues on the binding sequences of BRCA1 or BRCA1 interacting proteins that may indicate ERBIN is a transcriptional target of BRCA1. It was previously reported that BRCA1 interacts with a zing finger and KRAB domain protein ZBRK1 to bind a specific DNA sequence on the 3rd intron of GADD45 (Zheng *et al*., 2000). In addition to this direct binding of BRCA1 to ZBRK1, Cable *et al.* published a specific DNA sequence that can be bound by BRCA1 protein complexes (BRCA1:USF2) to control gene expression (Cable *et a*l, 2003). These two specific sequences were found with more than 95% similarity at the predicted promoter region of ERBIN. ERBIN was initially found to be interacting specifically with ErbB2 by its PDZ domain and acts in the localization of ErbB2 to the basolateral domain in epithelia which is important for its activation and signaling of ERBB2/HER2 in epithelia (Borg *et al*., 2000). They reported that the Erbin PDZ domain binds preferentially to the C terminus of ErbB2, which is non-Tyr1248-phosphorylated (Borg *et al*., 2000). Importantly phosphorylation of this residue following ErbB2 activation is a critical event for the mitogenic signaling and oncogenicity of this receptor (Dittmar *et al*., 2002). Overexpression of ErbB2 correlates with poor prognosis and resistance to chemotherapy in breast and ovarian cancer cases (Klapper *et al*., 2000). Despite the close relation of ERBIN and ErbB2, the functional role of ERBIN has not been studied extensively in breast cancer yet. Recently Liu *et al*. studied the expression and the regulation of ERBIN and its binding partner ErbB2 in the MCF7 breast cancer cell line. One of their findings was that the affinity of Erbin-ErbB2 interaction was reduced by ErbB2 posphorylation (Liu *et al*, 2008). BRCA1 is known to be down regulated by methylation or mutation or mislocalized and become unfunctional in most the breast cancer cases. We showed that BRCA1 and ERBIN are highly correlated and the downregulation of BRCA1 caused a decrease in the expression of ERBIN in breast cancer cell lines. Considering these findings, it can be suggested the lack of BRCA1 in breast tumor cells may cause the downregulation of ERBIN and Erbin can not bind to ErbB2 anymore; since the Tyr1248 residue of the ErbB2 is free from Erbin, it tends to be phosphorylated which leads to proliferation of the cancer cells.

It was inevitable to analyze the regulatory regions of selected BRCA1 target genes to see if there was any common transcription factor binding motifs or known BRCA1 binding sequences on their promoter regions. Although there were no novel common motifs or sequences on the promoter regions of the eight BRCA1-target genes when the known sequences were searched, OCT-1 and CAAT motifs were found at close

proximity on the regulatory regions of XRN2, SMG1 and ERBIN. It was previously found that BRCA1 physically interacts with sequence specific transcription factors Oct-1 and NF-YA, which directly bind to the OCT-1 and CAAT motifs on GADD45 promoter thus inducing the expression of Gadd45α (Fan *et al*., 2002).
Among the genes we studied, one of the RNA processing genes XRN2 was found to have specific OCT-1 and CAAT motif, which is a direct binding site of specific transcription factors Oct-1 and NF-YA, on its promoter region (Fan *et al*., 2002). This specific motif was also found to be localized on the promoter region of only 55 genes when genome wide analysis was performed. This made the XRN2 a possible transcriptional target of BRCA1 or showed that BRCA1 could have a regulatory role on the expression of this gene. Since XRN2 (Gromak *et al*., 2006; West *et al*., 2006) was one of the proteins playing a role in RNA processing like SMG1, these results supported our previous suggestion that BRCA1 could be one of the mediators of RNA surveillance and RNA processing in the cell. According to these findings, SMG1 and XRN2 and also ERBIN were worth studying further.

PART II: A RESAMPLING BASED META-ANALYSIS FOR DETECTION OF DIFFERENTIAL GENE EXPRESSION IN BREAST CANCER

Microarrays allow high-throughput analysis of expression for thousands of genes and provide valuable information for tumor studies. For example, individual microarray studies have identified differentially expressed gene lists for distinguishing breast cancer subtypes and normal breast tissue (Turashvili *et al*., 2007; Grigoriadis *et al*., 2006, Karnoub *et al*., 2007; Tripathi *et al*., 2008). Meta-analysis, on the other hand, might increase the knowledge by gathering and processing individual microarray datasets. In the present study, we provided highly stable lists of differentially expressed genes based on meta-analysis of two breast cancer datasets (Sorlie *et al*., 2003; Zhao *et al*., 2004). We have used a resampling-based strategy in which the effects of number of iterations and sample size were minimized by using a voting scheme in which each IMAGE clone, at each run, was voted as either significantly- or non-differentially expressed and the significant counts then were added up. A percentage value was obtained by dividing the number of significant votes by the total number of votes and a threshold of 80-90% for each IMAGE clone was chosen as a cut-off value for this meta-analysis. The meta-analysis was able to report multiple genes (i.e., 29, 21, and 6 genes for DN, LN, and DL, respectively) which neither dataset could report when analyzed individually.

Sample size greatly influences the reproducibility of the significant gene lists, such that the lower the sample size the less stable the gene lists become (Pavlidis *et al*., 2004). In addition, Qui *et al*. (Qui *et al*., 2006) have shown that the stability of genes identified as differentially expressed varies: some genes are consistently stable whereas others are not, independent of the statistical methodology used. Along these considerations, our voting scheme provided an advantage for extracting highly stable gene lists.

Different statistical methods are available for assessing differential expression. Among these, non-parametric tests allow for comparison of low sample size and distribution-independent comparisons. Our choice of rank-sum test was based on this idea; similarly, previous studies reported the use of the Kolmogorov-Smirnov test to compare the reference and sample distributions in the context of Gene Ontologies (Ben-Shaul *et al*., 2005). We used the Kolmogorov-Smirnov test for comparison of test and random distributions of p-values obtained from rank sum tests. In generating random datasets, we applied a gene-wise permutation algorithm that preserved the expression level information. Based on gene-wise permutations, a set of probability values that compare the actual and randomized distributions allowed for the assessment of the significance of the difference between groups tested using the Kolmogorov-Smirnov tests.

Different studies can be normalized and directly compared to each other in meta-analysis. Our comparisons ensured that there was a significant correlation between the Sorlie and Zhao datasets although these studies were based on independent tumor and normal samples; and the experimental procedures (e.g., amplification of RNA) also varied considerably between the two studies. Median rank scores (Toedling and Spang, 2003) or quantile discretization algorithms have frequently been used to transform gene expression values from different studies to a common numerical range (Warnat *et al*., 2005). Since the global median-normalized and quantile-normalized data correlated well, we have used the former normalization method, with the least number of data manipulation steps, before combining these two datasets.

Due to the large number of comparisons involved in microarray data analysis, it is important to take into account the false positive error rate and control it for the number of tests performed. FDR is a well-known methodology for multiple-test correction; its estimation relies on calculation of the number of false positives in a randomly permuted set of experiments (Benjamini and Hochberg, 1995). Therefore, we made comparisons between randomly shuffled datasets to obtain an estimate of FDR; and kept the value of FDR low (% 0.01) to reduce the number of false positives.

Invasive breast tumors comprise of 18 different histological types (Weigelt *et al.*, 2008), most of which were classified as invasive ductal carcinoma not otherwise specified (IDC NOS). ILC, on the other hand, makes about 10-15% of all breast tumors and it is histologically characterized by uniform tumor cells arranged in single-files or concentrically localized around ducts (Yoder *et al.*, 2007). As IDC, ILC exhibit heterogeneity; and a high grade aggressive form of ILC known as pleomorphic lobular carcinoma (PLC) exists (Simpson *et al.*, 2008). Bertucci *et al.* (Bertucci *et al.*, 2008) described that IDC and ILC were histologically and genomically distinguishable from each other among the ER(+) grade II invasive breast tumors. Furthermore, ILC molecular subtypes were reported to include the typical and IDC-like ILCs, yet the *CDH1* mutation and/or underexpression was common but not universal to ILCs in general (Yoder *et al.*, 2007). Low-grade breast tumors were generally characterized by ER(+), PR(+) and with limited genomic aberrations whereas high grade tumors were generally ER(-) and PR(-) and had complex karyotypic changes. However, molecular differences among subtypes may not surpass the differences between any tumor cell and the normal since the degree of genomic stability in normal cells would be relatively higher.

The other three studies presenting data on ILC and IDC, Turashvili *et al.* (Turashvili *et al.*, 2007), Sorlie *et al.* (Sorlie *et al.*, 2003) and Zhao *et al.* (Zhao *et al.*, 2004) have used a more diverse selection of tumor samples. Although IDC and ILC have distinctive clinical and pathological characteristics and differ in their ER status and metastatic behaviors (Arpino *et al*., 2004), meta-analysis of Zhao and Sorlie datasets indicated that a small number of genes were distinguished between the expression profiles of IDC and ILC patients. On the hand, the number of genes that was differentially expressed between normal and IDC or normal and ILC samples were much greater. Indeed, Turashvili *et al*. (Turashvili *et al*., 2007) also has reported only 28 genes that were significantly differentially expressed between IDC and ILC samples, which were extracted using laser-dissection, a more recent methodology allowing for precise collection of a given cell population. These findings suggest that the degree of molecular differences between IDC and ILC are indeed smaller than those between the tumor and normal classes.

Comparisons among the meta-analysis, Turashvili and Bertucci studies pointed out to CHD1, TFAP2A, RBP4, and ALDH1A1 genes as commonly modulated. Indeed, CDH1 is one of the best studied discriminators for ductal/lobular breast cancer specimens in the literature by immunohistochemistry and at the genomic level. In breast cancer, reduced CDH1 expression has been found in 50% of invasive ductal carcinomas, whereas CDH1 expression was almost always absent in infiltrating lobular carcinoma (ILC) (Sorlie *et al*., 2003; Zhao *et al*., 2004; Turashvili *et al*., 2007, Bertucci *et al*., 2008; Sarrió *et al*., 2003; Caldeira *et al*., 2006). TFAP2A was shown to be highly expressed in ductal tumor cells while normal cells expressed TFAP2A in the inner glandular cell layer (Friedrichs *et al*., 2005). On the other hand, nuclear TFAP2 expression was shown to be higher in lobular than ductal breast carcinomas (Pellikainen *et al*., 2002). There is no report on RBP4 in the literature in connection with ductal vs. lobular breast cancer distinction while ALDH1A1 protein levels were shown to exhibit differences among the ductal carcinoma patients (Sládek, 1999). The candidates identified in the meta-analysis then are likely to be discriminatory at the mRNA level rather than the protein level since protein localization and variability in intensity might make the ductal vs. lobular tissue discrimination less clear.

Analyses of Sorlie, Zhao, and Turashvili data showed that tumor cells were remarkably distinct from their respective normals in their transcription profiles implicating that whatever the subtype structure underneath, most of the variability among samples was due to changes during tumorigenesis. Accordingly, the idea that genes discriminating tumor from normal in a stable manner may also provide information on the state of the tumorigenesis is a valid one.

Breast tumor subtype classification remains a complicated issue due to the difficulties associated with the presence of multiple interacting factors such as the presence or absence of node-filtration, ER-positivity, metastatic potential, different degrees of genomic instability, and tumor cell origin. For example, basal like cancers have distinct molecular expression profiles and histological differences when compared with the luminal type (Fadare and Tavassoli, 2007). Nielsen *et al.* (Nielsen *et al.*, 2004) have categorized basal like breast cancer tumors as having variable levels of expression of one of the three stem/basal markers, namely CK5/6, EGFR, and c-kit. Luminal cell markers, on the other hand, include CK8, CK18, CK19, mostly characteristic of glandular and/or lobular epithelial cells (Abd El-Rehim *et al.*,2004). However, both basal and luminar histochemical markers may exist simultaneously suggesting that breast cancer is rather a heterogeneous tissue (Moriya *et al.*, 2006). It is also evident that tumors with a triple negative status (ER-, PR-, HER2-) are more likely to belong to the basal type (Nielsen *et al.*, 2004, Liu *et al.*, 2008). In general, gene expression studies have associated the basal-like breast tumors with high proliferative abilities and thus having the worse prognosis when compared with the luminal subtype of breast cancers (Sorlie *et al.*, 2003; Sotiriou *et al.*, 2003). Thus identification of genes best classifying breast cancer into intrinsic molecular subtypes like luminal, HER2+/ER- and basal-like also allow determination of risk-factors and likely prognosis for the patients. The importance of identification of these different subtypes is that they differ in clinical outcome thus molecular subtype signatures help predict clinical outcome and response to therapy.

Genes differentially expressed between tumor and normal states (DN and LN) also keep information about intrinsic subtypes. Accordingly, meta-analysis identified ATF3, ADAMTS1, EGFR, PRNP, IGFBP6, ID4, SFRP1, SPTBN1, and FN1 with ability to classify tumors into basal and luminal subclasses. Additionally, most of them accurately differentiated ER(+) and ER(-) tumors (Additional file 9).

Among the abovementioned genes, ID4 was found to be a novel tumor suppressor gene in normal human breast tissues and epigenetically silenced in breast cancer cell lines and primary breast tumors (Noetzel *et al*., 2008; Umetani *et al*., 2005). As supporting information for our data, de Candia *et al*. suggested that the expression of ID4 in the mammary duct epithelium may be regulated by estrogen depending on the differential expression pattern of ID4 in ER(+) and ER(-) breast tumors (de Candia *et al*., 2003). SFRP1 on the other hand is a frizzled-related protein taking role in a variety of cellular processes, including control of cell polarity, cell fate determination, and malignant transformation. Loss of *SFRP1* was found to be associated with cancer progression and poor prognosis in breast cancer in previous studies (Klopocki *et al*.,2004; Kawano and Kypta, 2003). EGFR is known to be a positive immunohistochemical marker for basal-like breast cancers and it was shown to accurately identify basal-like tumors from microarray data with potential therapeutic implications (Cheang *et al*., 2008; Arnes *et al*., 2008). Activating transcription factor 3 (ATF3) is a member of the ATF/cyclic AMP response element-binding family of transcription factors. It was shown to enhance apoptosis in the untransformed mammary epithelial cells while protecting the aggressive cells and enhancing cell motility. Array analyses indicated that ATF3 upregulated the expression of several genes in the tumor necrosis factor pathway in the untransformed mammary epithelial cells. However, the expression of several genes implicated in tumor metastasis including fibronectin (FN1) was upregulated in aggressive cells. ATF3 was also shown to regulate the transcription of FN1, one of the genes obtained in the present study. *ATF3* gene copy number was at least doubled in 80% of the breast tumors examined; protein levels also were elevated in close to 50% of these tumors (Yin *et al*., 2008).

Since the normal vs. tumor classification was strikingly distinct based on meta-analysis, and a gene-set with the capacity for breast cancer subtype classification, we further analyzed a set of normal-matched tumors for selected genes from the meta-gene list using real-time qRT-PCR. The selected 16 significant genes were shown to have expression profiles similar to those found from the meta-analysis. Our findings also suggested that these genes could be used as predictors of tumor status regardless of the origin of the reference samples, i.e., a matched or pooled reference tissue. Furthermore, there was a high level of correlation between fold changes obtained from the DL meta-genes and those from the Turashvili dataset, regardless of the different sample extraction methods used in each study (i.e., frozen sections and laser-dissection, respectively). These findings indicate that our proposed methodology is robust in predicting the tumor or non-tumor status as well as ductal and lobular cancer expression signatures in breast samples. Future studies might concentrate on whether meta-analysis specific genes can also be helpful in the prediction of the level of prognosis and time to disease-free survival.

CHAPTER 5. FUTURE PERSPECTIVES

Reference gene selection for breast tumor sample studies

For the accurate interpretation of the results of qRT-PCR experiments endogenous gene selction was performed. The expression patterns of fifteen widely-used endogenous RGs (ACTB, TBP, GAPDH, SDHA, HPRT, HMBS, B2M, PPIA, GUSB, YWHAZ2, PGK1, RPLP0, PUM1, MRPL19 and RPL41), and three candidate genes that were selected through analysis of two independent microarray datasets (IL22RA1, TTC22 and ZNF224) were determined in 23 primary breast tumors and their matched normal tissues using qRT-PCR.

All the reference genes studied in this study exhibited relatively higher expression in tumors than their normal counterparts. Similarly, it was reported that breast biopsy samples exhibited great intra- and inter-individual variability and mean expression values of tumors measured in copy numbers were greater than those of their normal counterparts (Tricarico *et al.*, 2002). Because of the extensive variability in RG expression, total RNA-based (or mRNA copy numbers when available) normalization was suggested as an NF for tumor samples (Bustin, 2002, Tricarico *et al.*, 2002). However, since total RNA is represented mostly by rRNA (>90%), even a small decrease in rRNA expression may lead to a disproportional increase in the mRNA pool estimation (Spanakis *et al.*, 1993, Elberg *et al.*, 2006). Moreover, studies have shown that rDNA genes were methylated in breast and ovarian cancers when compared with those of normal controls (Yan *et al.*, 2000, Chan *et al.*, 2005). In fact our finding of low tumor rRNA to mRNA ratio suggests that normal and tumor samples are heterogeneous in total RNA fractions.

Although the panel used in this study is highly comprehensive and consists of frequently used reference genes, they may still not be the best applicable reference genes for breast cancer normalization studies unless there is a bias due to RNA estimation or breast tissue heterogeneity. Our findings indicated that estimation of mRNA from total RNA represented an important issue requiring further investigation in qRT-PCR studies. Since the rDNA hyper-methylation holds considerable possibility in breast tumors and the total RNA is largely made up of rRNA, the use of poly(A)+ RNA as a starting material may be another approach for studying tumor and their matched normal samples. This way we may be able to establish the percentage of tumors with such rRNA/mRNA bias and the underlying causes such as methylation patterns of rDNA.

We found out that the mean expression of 18S rRNA was down regulated in tumor samples (9/13) compared to their normal counterparts. Therefore it is essential to analyze the breast tumor samples for the promoter region of the 45S rRNA gene since this promoter controls both 18S rRNA and 28S rRNA expression. Indeed, we determined high methylation pattern in the region that regulates 18S rRNA in two breast tumor samples. This study needs to be extended in a large panel of tumor samples and also for normal breast tissues. It is also essential to perform clonal selection of the bisulfite DNA PCR amplified products from each tumor and normal breast tissue DNA samples. Then, at least five clones should be selected for each sample and the bisulfite sequencing analysis should be applied to estimate the accurate percentage of methylation in the CpG islands in the promoter region for each sample before reaching the final conclusion.

The expression level of 28S rRNA has not been assessed in the breast samples. It would be wise to perform expression analysis for the 28s rRNA transcript in the same samples.

Target gene expression profiles in normal matched breast tumors

The tumor suppressor gene BRCA1 (Breast cancer susceptibility gene 1) plays a central role in the development of breast and ovarian cancers. The role of gene in the maintenance of chromosomal integrity is linked to a number of biological properties of its protein product including transcriptional regulation. We aimed to find out the expression profiles of the genes, which were selected from the former BRCA1-induced gene list (*OVCA1, OVCA2, ERBIN, RAD21, XRN2, RENT2, SMG1* and *MAC30*) in normal-matched primary breast tumors and to correlate the gene expression profiles of selected candidate genes with *BRCA1* and various pathology parameters.

The target genes regulated by BRCA1 expression analysis showed that ERBIN, SMG1 and RAD21 were highly correlated with BRCA1 expression in breast samples used in this study.

Taking into account all the findings in this study, the target gene regulatory regions should be analyzed more extensively to be able obtain more clues on the binding sequences of BRCA1 or BRCA1 interacting proteins or any common transcription factor binding motifs that may indicate these genes can be transcriptional targets of BRCA1.

Finding out a common potential binding sequence for BRCA1 in its target genes may explain the function of BRCA1 in tumor suppression in breast cancers.

Among the list of selected genes, the expression of *ERBIN* was found to be highly correlated with that of *BRCA1* both in BRCA1 over-expressing and down-regulating cell lines and this correlation was validated with the experiments performed with a breast cancer cell line (BCC) panel. Additionally, the correlation was repeatedly observed when their expression profiles were analyzed in normal-matched tumor samples. We searched all the target gene promoters for the previously reported binding sequences of BRCA1 or BRCA1 interacting proteins and found that the ERBIN promoter contains ZBRK1, USF2, and Oct1 binding sequences that were also reported to be present in the other BRCA1 regulated genes, such as GADD45.

It will be important to clone the regulatory region of the ERBIN into a reporter vector and use it to transfect the cells which have inducible controlled BRCA1 expression. This approach may provide important findings if the target promoter is regulated by the BRCA1 expression. If so, it is possible to identify the BRCA1 responsive sites in the regulatory regions of the target gene by using deletion-mapping strategies of the region with new reporter constructs.

The expression profiles of the target genes used in this study were determined with the qRT-PCR approach. It is important to show their expression at the protein level in breast tumor and normal samples in order to have a better understanding and more comprehensive evaluation of their role for prediction of clinical parameters.

Resampling-based meta-analysis gene signature

Diagnostic gene-sets at the mRNA expression level have recently been reported to be better predictors of disease state and classification of cancer subtypes. However, breast cancer is heterogeneous in nature; thus extraction of differentially expressed gene-sets that stably distinguish normal tissue from various pathologies poses challenges. Meta-analysis of high-throughput expression data using a collection of statistical methodologies leads to the identification of robust tumor gene expression signatures, which can be further explored for their ability to discriminate between cancer subtypes and/or provide valuable prognostic information.

In the meta-analysis part of this study, meta-analysis of two independent comparable microarray data sets allowed us to provide genes that are able to discriminate IDC and ILC and normal mammary cells from the tumors. We also provided highly generalized and stable gene lists that could be used for prediction of tumor or normal status. The resampling approach proposed herein has the ability to detect a set of differentially expressed genes, with the least amount of within-group variability. This meta-analytic approach thus provides a method to combine two or more independent cancer data sets leading to the identification of differentially expressed gene sets for better understanding of cancer development and progression.

Due to the lack of ILC samples, the confirmation of the IDC/ILC (DL) meta-genes could only be done by correlation analysis with independent microarray datasets rather than qRT-PCR experiments. The meta-gene list discriminating between ductal and lobular breast tumor samples at the mRNA level requires further confirmation at the protein level to better assess discriminatory power. Future validation studies might concentrate on whether meta-analysis specific genes also participate in prediction of level of prognosis and/or time to disease-free survival.

Since the normal vs. tumor classification was strikingly distinct based on meta-analysis and a gene-set with the capacity for breast cancer subtype classification, we further analyzed a set of normal matched tumors for selected genes from the meta-gene list using real-time qRTPCR. The selected 16 significant genes were shown to have expression profiles similar to those found from the meta-analysis. Our findings also suggested that these genes could be used as predictors of tumor status regardless of the origin of the reference samples, i.e., matched or pooled reference tissue. The number of samples used in qRT-PCR was relatively small and increasing the sample size may help generalize our results to a wider range of breast tumor samples.

REFERENCES

Abd El-Rehim DM, Pinder SE, Paish CE, Bell J, Blamey RW, Robertson JF, Nicholson RI, Ellis IO: Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol* 2004, 203: 661-671.

Ahr A, Karn T, Solbach C: Identification of high risk breast-cancer patients by gene expression profiling. *Lancet* 2002, 359: 131-132.

Andersen CL, Jensen JL, Orntoft TF: Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004, 64: 5245-5250.

Anderson T.W. (1984). An Introduction to Multivariate Statistical Analysis, Second Edition, John Wiley & Sons.

Andrews HN, Mullan PB, McWilliams S, Sebelova S, Quinn JE, Gilmore PM, McCabe N, Pace A, Koller B, Johnston PG, Haber DA, Harkin DP: BRCA1 regulates the interferon gamma-mediated apoptotic response. *J Biol Chem* 2002, 277: 26225-32.

Aprelikova O, Pace AJ, Fang B, Koller BH, Liu ET: BRCA1 is a selective co-activator of 14-3-3 sigma gene transcription in mouse embryonic stem cells. *J Biol Chem* 2001, 276: 25647-25650.

Arnes JB, Bégin LR, Stefansson IM, Brunet JS, Nielsen TO, Foulkes WD, Akslen LA: Expression of EGFR in relation to BRCA1 status, basal-like markers and prognosis in breast cancer. *J Clin Pathol* 2009, 62: 139-146.

Arpino G, Bardou VJ, Clark GM, Elledge RM: Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. *Breast Cancer Res* 2004, 6:R149-156.

Atalay A, Crook T, Ozturk M, Yulug IG: Identification of genes induced by BRCA1 in breast cancer cells. *Biochem Biophys Res Commun* 2002, 299: 839-846.

Bakkenist CJ, Kastan MB: DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* 2003, 421: 499-506.

Barber, R.D., Harmer, D.W., Coleman, R.A., Clark, B.J: GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genom* 2005, 21: 389-395.

Barrett T, Suzek TO, Troup DB: NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res* 2005, **33**: D562-566.

Beckmann MW, Niederacher D, Schnurch HG, Gusterson BA, Bender HG: Multistep carcinogenesis of breast cancer and tumor heterogeneity. *J Mol Med* 1997, 75: 429-439.

Benjamini, Y. and Hochberg, Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statis. Soc Ser* 1995, 57:289-300.

Ben-Shaul Y, Bergman H, Soreq H: Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics* 2005, 21:1129-1137.

Bereta J, Bereta M: Stimulation of glyceraldehyde-3- phosphate dehydrogenase mRNA levels by endogenous nitric oxide in cytokine-activated endothelium. *BBRC* 1995, 217: 363–369.

Bernard PS, Wittwer CT: Real-time PCR technology for cancer diagnostics. *Clin Chem* 2002, 48: 1178-1185.

Bertucci F, Orsetti B, Nègre V, Finetti P, Rougé C, Ahomadegbe JC, Bibeau F, Mathieu MC, Treilleux I, Jacquemier J, Ursule L, Martinec A, Wang Q, Bénard J, Puisieux A, Birnbaum D, Theillet C: Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles. *Oncogene* 2008, 27: 5359-5372.

Biometric Research Branch [http://linus.nci.nih.gov/BRB-ArrayTools.html]

Birgisdottir V, Stefansson OA, Bodvarsdottir SK, Hilmarsdottir H, Jonasson JG, Eyfjord JE: Epigenetic silencing and deletion of the BRCA1 gene in sporadic breast cancer. *Breast Cancer Res* 2006, 8: R38.

Bochar DA, Wang L, Beniya H, Kinev A, Xue Y, Lane WS, Wang W, Kashanchi F, Shiekhattar R: BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell* 2000,102: 257-65.

Borg JP, Marchetto S, Le Bivic A, Ollendorff V, Jaulin-Bastard F, Saito H, Fournier E, Adélaïde J, Margolis B, Birnbaum D. ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor. *Nat Cell Biol* 2000, 2: 407-14.

Bruening W, Prowse AH, Schultz DC, Holgado-Madruga M, Wong A, Godwin AK: Expression of OVCA1, a candidate tumor suppressor, is reduced in tumors and inhibits growth of ovarian cancer cells. *Cancer Res* 1999, 59: 4973-4983.

Brumbaugh KM, Otterness DM, Geisen C, Oliveira V, Brognard J, Li X, Lejeune F, Tibbetts RS, Maquat LE, Abraham RT: The mRNA surveillance protein hSMG-1 functions in genotoxic stress response pathways in mammalian cells. *Mol Cell* 2004 14:585-598.

Bustin S A: Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol* 2002, 29: 23-39.

Bustin SA, and Nolan T: Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomol Tech* 2004a, 15: 155-166.

Bustin SA: A-Z of Quantitative PCR. La Jolla, CA: International University Line, 2004b.

Bustin SA, Benes V, Nolan T, Pfaffl MW: Quantitative real-time RT-PCR-a perspective. *J Mol Endocrinol* 2005, 34: 597-601.

Buyse M, Loi S, Van't Veer L: Validation and clinical utility of a 70-Gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006, 98: 1183-1192.

Cable PL, Wilson CA, Calzone FJ, Rauscher FJ 3rd, Scully R, Livingston DM, Li L, Blackwell CB, Futreal PA, Afshari CA: Novel consensus DNA-binding sequence for BRCA1 protein complexes. *Mol Carcinog* 2003, 38: 85-96.

Caldeira JR, Prando EC, Quevedo FC, Neto FA, Rainho CA, Rogatto SR: CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. *BMC Cancer* 2006, 6: 48.

Cerutti JM, Oler G, Michaluart P Jr, Delcelo R, Beaty RM, Shoemaker J, Riggins GJ: Molecular profiling of matched samples identifies biomarkers of papillary thyroid carcinoma lymph node metastasis. *Cancer Res* 2007, 67: 7885-7892.

Chai YL, Cui J, Shao N, Shyam E, Reddy P, Rao VN: The second BRCT domain of BRCA1 proteins interacts with p53 and stimulates transcription from the p21WAF1/CIP1 promoter. *Oncogene* 1999, 18: 263-268.

Chan DW, Ye R, Veillette CJ, Lees-Miller SP: DNA-dependent protein kinase phosphorylation sites in Ku 70/80 heterodimer. *Biochemistry* 1999, 38: 1819-1828.

Chan MW, Wei SH, Wen P, Wang Z, Matei DE, Liu JC, Liyanarachchi S, Brown R, Nephew KP, Yan PS, Huang TH: Hypermethylation of 18S and 28S ribosomal DNAs predicts progression-free survival in patients with ovarian cancer. *Clin Cancer Res* 2005, 11: 7376-7383.

Chang YF, Imam JS, Wilkinson MF: The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 2007, 76: 51-74.

Chang TJ, Juan CC, Yin PH, Chi CW, Tsay HJ: Up-regulation of beta-actin, cyclophilin and GAPDH in N1S1 rat hepatoma. *Oncology and Reproduction* 1998, 5: 469–471.

Chaturvedi P, Eng WK, Zhu Y, Mattern MR, Mishra R, Hurle MR, Zhang X, Annan RS, Lu Q, Faucette LF, Scott GF, Li X, Carr SA, Johnson RK, Winkler JD, Zhou BB: Mammalian Chk2 is a downstream effector of the ATM-dependent DNA damage checkpoint pathway. *Oncogene* 1999, 18: 4047-4054.

Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, Perou CM, Nielsen TO: Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res* 2008, 14:1368 1376.

Chen CM, Behringer RR: Ovca1 regulates cell proliferation, embryonic development, and tumorigenesis. *Genes Dev* 2004, 18: 320-332.

Chen CM, Behringer RR: OVCA1: tumor suppressor gene. *Curr Opin Genet Dev* 2005,15: 49-54.

Chen F, Kamradt M, Mulcahy M, Byun Y, Xu H, McKay MJ, Cryns VL: Caspase proteolysis of the cohesin component RAD21 promotes apoptosis. *J Biol Chem* 2002 277: 16775-16781.

Chiu ST, Hsieh FJ, Chen SW, Chen CL, Shu HF, L, H: Clinicopathologic Correlation of Up-regulated Genes Identified Using cDNA Microarray and Real-time Reverse Transcription-PCR in Human Colorectal Cancer. *Cancer Epidemiol Biomarkers & Preven* 2005 14: 437-443.

Choi H, Shen R, Chinnaiyan AM: A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* 2007, 8: 364.

Choi JK, Choi JY, Kim DG: Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett* 2004, 565: 93-100.

Choi JK, Yu U, Kim S, Yoo OJ: Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 2003, 19 Suppl 1:i84-90.

Cleator S and Ashworth A: Molecular profiling of breast cancer: clinical implications. *Br J Cancer* 2004, 90:1120-1124.

Corcoran DL, Feingold E, Benos PV: FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res* 2005, 33 (Web Server issue): W442-6.

Dai F, Chang C, Lin X, Dai P, Mei L, Feng XH: Erbin inhibits transforming growth factor beta signaling through a novel Smad-interacting domain. *Mol Cell Biol* 2007, 27: 6183-6194.

Dai P, Xiong WC, Mei L: Erbin inhibits RAF activation by disrupting the sur-8-Ras-Raf complex. *J Biol Che*. 2006, 281: 927-933.

de Candia P, Akram M, Benezra R, Brogi E: Id4 messenger RNA and estrogen receptor expression: inverse correlation in human normal breast epithelium and carcinoma. *Hum Pathol* 2006, 37: 1032-1041.

de Cremoux, P., Bieche, I., Tran-Perennou, C., Vignaud, S., Boudou, E., Asselain, B., Lidereau, R., Magdelénat, H., Becette, V., Sigal-Zafrani, B., Spyratos, F: Inter-laboratory quality control for hormone-dependent gene expression in human breast tumors using real-time reverse transcription-polymerase chain reaction. *Endocr Relat Cancer* 11:489-495, 2000.

de Jong MM, Nolte IM, te Meerman GJ, van der Graaf WT, Oosterwijk JC, Kleibeuker JH, Schaapveld M, de Vries EG: Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. *J Med Genet* 2002, 39: 225-42.

de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, Span PN: Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab Invest* 2005, 8:154-159.

Dittmar T, Husemann A, Schewe Y, Nofer JR, Niggemann B, Zänker KS, Brandt BH: Induction of cancer cell migration by epidermal growth factor is initiated by specific phosphorylation of tyrosine 1248 of c-erbB-2 receptor via EGFR. *FASEB J* 2002, 16: 1823-1825.

Elberg G, Elberg D, Logan CJ, Chen L, Turman MA: Limitations of commonly used internal controls for real-time RT-PCR analysis of renal epithelial mesenchymal cell transition. *Nephron Exp Nephrol* 2006, 102: 113-122.

Esteller M, Silva JM, Dominguez G, Bonilla F, Matias-Guiu X, Lerma E, Bussaglia E, Prat J, Harkes IC, Repasky EA, Gabrielson E, Schutte M, Baylin SB, Herman JG: Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors.*J Natl Cancer Inst* 2000, 92: 564-569.

Fadare O, Tavassoli FA: The phenotypic spectrum of basal-like breast cancers: a critical appraisal. *Adv Anat Pathol* 2007, 14: 358-373

Fan W, Jin S, Tong T, Zhao H, Fan F, Antinore MJ, Rajasekaran B, Wu M, Zhan Q: BRCA1 regulates GADD45 through its interactions with the OCT-1 and CAAT motifs. *J Biol Chem* 2002, 277: 8061-8067.

Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, Macgrogan G, Bergh J, Cameron D, Goldstein D, Duss S, Nicoulaz AL, Brisken C, Fiche M, Delorenzi M, Iggo R: Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005, 24:4660-4671.

Feilotter HE, Coulon V, McVeigh JL, Boag AH, Dorion-Bonnet F, Duboué B, Latham WC, Eng C, Mulligan LM, Longy M: Analysis of the 10q23 chromosomal region and the PTEN gene in human sporadic breast carcinoma. *Br J Cancer* 1999, 79: 718-723

Folgueira MA, Brentani H, Katayama ML, Patrao DF, Carraro DM, Mourao Netto M, Barbosa EM, Caldeira JR, Abreu AP, Lyra EC, Kaiano JH, Mota LD, Campos AH, Maciel MS, Dellamano M, Caballero OL, Brentani MM: Gene expression profiling of clinical stages II and III breast cancer. *Braz J Med Biol Res* 2006, 39:1101-1113.

Frebourg T, Barbier N, Yan YX, Garber JE, Dreyfus M, Fraumeni J Jr, Li FP, Friend SH. Germ-line p53 mutations in 15 families with Li-Fraumeni syndrome. *Am J Hum Genet* 1995, 56: 608-615.

Freihoff D, Kempe A, Beste B, Wappenschmidt B, Kreyer E, Hayashi Y, Meindl A, Krebs D, Wiestler OD, von Deimling A, Schmutzler RK: Exclusion of a major role for the PTEN tumour-suppressor gene in breast carcinomas. *Br J Cancer* 1999, 79: 754-758.

Friedrichs N, Jäger R, Paggen E, Rudlowski C, Merkelbach-Bruse S, Schorle H, Buettner R: Distinct spatial expression patterns of AP-2alpha and AP-2gamma in non-neoplastic human breast and breast cancer. *Mod Pathol* 2005, 18:431-438.

Fronhoffs S, Totzke G, Stier S, Wernert N, Rothe M, Bruning T, Koch B, Sachinidis A, Vetter H, Ko Y: A method for the rapid contruction of cRNA standard curves in quantitative real-time reverse transcription polymerase chain reaction. *Mol Cell Probes* 2002, 16: 99-110.

Gatei M, Zhou BB, Hobson K, Scott S, Young D, Khanna KK: Ataxia telangiectasia mutated (ATM) kinase and ATM and Rad3 related kinase mediate phosphorylation of Brca1 at distinct and overlapping sites. In vivo assessment using phospho-specific antibodies. *J Biol Chem* 2001, 276: 17276-80.

Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo/]

Greenberg RA: Recognition of DNA double strand breaks by the BRCA1 tumor suppressor network. *Chromosoma* 2008, 117: 305-317.

Grigoriadis A, Mackay A, Reis-Filho JS, Steele D, Iseli C, Stevenson BJ, Jongeneel CV, Valgeirsson H, Fenwick K, Iravani M, Leao M, Simpson AJ, Strausberg RL, Jat PS, Ashworth A, Neville AM, O'Hare MJ: Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res* 2006, 8:R56.

Gromak N, West S, Proudfoot NJ. Pause sites promote transcriptional termination of mammalian RNA polymerase II: *Mol Cell Biol* 2006, 26: 3986-3996.

Grutzmann R, Boriss H, Ammerpohl O: Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 2005, 24: 5079-88.

Gruvberger S, Ringner M, Chen Y: Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001, 61: 5979-5984.

Gudas JM, Nguyen H, Li T, Cowan KH: Hormone-dependent regulation of BRCA1 in human breast cancer cells. *Cancer Res* 1995, 55: 4561-4565.

Hall J: The Ataxia-telangiectasia mutated gene and breast cancer: gene expression profiles and sequence variants. *Cancer Lett* 2005, 227:1 05-114

Harkin DP, Bean JM, Miklos D, Song YH, Truong VB, Englert C, Christians FC, Ellisen LW, Maheswaran S, Oliner JD, Haber DA: Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* 1999, 97: 575-86.

Heid CA, Stevens J, Livak KJ, Williams PM: Real time quantitative PCR. *Genome Res* 1996, 6: 986-994.

Hoque MT, Ishikawa F: Human chromatid cohesin component hRad21 is phosphorylated in M phase and associated with metaphase centromeres. *J Biol Chem* 2001, 276: 5059-5067.

Hosey AM, Gorski JJ, Murray MM, Quinn JE, Chung WY, Stewart GE, James CR, Farragher SM, Mulligan JM, Scott AN, Dervan PA, Johnston PG, Couch FJ, Daly PA, Kay E, McCann A, Mullan PB, Harkin DP. Molecular basis for estrogen receptor alpha deficiency in BRCA1-linked breast cancer. *J Natl Cancer Inst* 2007 Nov 21,99(22):1683-94.

http://www.cancer.org

Hu Y: BRCA1, hormone, and tissue-specific tumor suppression. *Int J Biol Sci* 2009, 5: 20-27.

Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM: The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 2006, 7: 96

Huang E, Cheng SH, Dressman H: Gene expression predictors of breast cancer outcomes. *Lancet* 2003, 361: 1590-1596.

Huang YZ, Wang Q, Xiong WC, Mei L. Erbin is a protein concentrated at postsynaptic membranes that interacts with PSD-95. *J Biol Chem* 2001, 276: 19318-19326.

Huang YZ, Zang M, Xiong WC, Luo Z, Mei L: Erbin suppresses the MAP kinase pathway. *J Biol Chem* 2003, 278: 1108-1114.

Huggett J, Dheda K, Bustin S, and Zumla A: Real-time RT-PCR normalisation, strategies and considerations. *Genes and Immunity* 2005, 6: 279-284.

Ivanov PV, Gehring NH, Kunz JB, Hentze MW, Kulozik AE: Interactions between UPF1, eRFs, PABP and the exon junction complex suggest an integrated model for mammalian NMD pathways. *EMBO J* 2008, 27: 736-747.

Iwao K, Miyoshi Y, Egawa C, Ikeda N, Tsukamoto F, Noguchi S: Quantitative analysis of estrogen receptor-alpha and -beta messenger RNA expression in breast carcinoma by real-time polymerase chain reaction. *Cancer* 2000, 89: 1732-1738.

James CR, Quinn JE, Mullan PB, Johnston PG, Harkin DP: BRCA1, a potential predictive biomarker in the treatment of breast cancer. *Oncologist* 2007, 12: 142-150.

Jarzabek K, Koda M, Kozlowski L, Mittre H, Sulkowski S, Kottler ML, Wolczynski S: Distinct mRNA, protein expression patterns and distribution of oestrogen receptors alpha and beta in human primary breast cancer: correlation with proliferation marker Ki-67 and clinicopathological factors. *Eur J Cancer* 2005, 41: 2924-2934.

Jensen MR, Helin K: OVCA1: emerging as a bona fide tumor suppressor. *Genes Dev* 2004, 18: 245-248.

Jung M, Ramankulov A, Roigas J, Johannsen M, Ringsdorf M, Kristiansen G, Jung K: In search of suitable reference genes for gene expression studies of human renal cell carcinoma by real-time PCR. *BMC Mol Biol* 2007, 8: 47.

Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, Richardson AL, Polyak K, Tubo R, Weinberg RA: Mesenchymal stem cells within tumor stroma promote breast cancer metastasis. *Nature* 2007, 449: 557-563.

Kawano Y, Kypta R: Secreted antagonists of the Wnt signalling pathway. *J Cell Sci* 2003, 116: 2627-2634.

Klapper LN, Kirschbaum MH, Sela M, Yarden Y: Biochemical and clinical implications of the ErbB/HER signaling network of growth factor receptors. *Adv Cancer Res* 2000, 77: 25-79.

Klopocki E, Kristiansen G, Wild PJ, Klaman I, Castanos-Velez E, Singer G, Stohr R, Simon R, Sauter G, Leibiger H, Essers L, Weber B, Hermann K, Rosenthal A, Hartmann A, Dahl E: Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors. *Int J Oncol* 2004, 25: 641-649.

Kroupis C, Stathopoulou A, Zygalaki E, Ferekidou L, Talieri M, Lianidou ES: Development and applications of a real-time quantitative RT-PCR method (QRT-PCR) for BRCA1 mRNA. *Clin Biochem* 2005, 38: 50-57.

Li J, Zheng H, Ji C, Fei X, Zheng M, Gao Y, Ren Y, Gu S, Xie Y, Mao Y: A novel splice variant of human XRN2 gene is mainly expressed in blood leukocyte. *DNA Seq* 2005, 16: 143-146.

Li W, Xiao C, Vonderhaar BK, Deng CX: A role of estrogen/ERalpha signaling in BRCA1-associated tissue-specific tumor formation. *Oncogene* 2007, 26: 7204-7212

Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacocke M, Eng C, Parsons R: Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* 1997, 16: 64-67.

Liu N, Zhang J, Zhang J, Liu S, Liu Y, Zheng D: Erbin-regulated sensitivity of MCF-7 breast cancer cells to TRAIL via ErbB2/AKT/NF-kappaB pathway. *J Biochem* 2008, 143: 793-801.

Liu ZB, Liu GY, Yang WT, Di GH, Lu JS, Shen KW, Shen ZZ, Shao ZM, Wu J: Triple-negative breast cancer types exhibit a distinct poor clinical characteristic in lymph node-negative Chinese patients. *Oncol Rep* 2008, 20: 987-994.

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K: Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 1995, 4: 357-362.

Lyng MB, Laenkholm AV, Pallisgaard N, Ditzel HJ: Identification of genes for normalization of real-time RT-PCR data in breast carcinomas. *BMC Cancer* 2008, 8: 20.

MacLachlan TK, Somasundaram K, Sgagias M, Shifman Y, Muschel RJ, Cowan KH, El-Deiry WS: BRCA1 effects on the cell cycle and the DNA damage response are linked to altered gene expression. *J Biol Chem* 2000, 275: 2777-2785.

MacLachlan TK, Takimoto R, El-Deiry WS: BRCA1 directs a selective p53-dependent transcriptional response towards growth arrest and DNA repair targets. *Mol Cell Biol* 2002, 22: 4280-4292.

Mallon E, Osin P, Nasiri N, Blain I, Howard B, and Gusterson B: The basic pathology of human breast cancer. *J Mammary Gland Biol Neoplasia* 2000, 5: 139-163.

Mann HB, Whitney D R: On a test of whether one of two random variables is stochastically larger than the other. *Annals Math Stat* 1947, 18: 50–60.

Marquis ST, Rajan JV, Wynshaw-Boris A, Xu J, Yin GY, Abel KJ, Weber BL, Chodosh LA: The developmental pattern of Brca1 expression implies a role in differentiation of the breast and other tissues. *Nat Genet* 1995, 11: 17-26

McDonald C, Chen FF, Ollendorff V, Ogura Y, Marchetto S, Lécine P, Borg JP, Nuñez G: A role for Erbin in the regulation of Nod2-dependent NF-kappaB signaling. *J Biol Chem* 2005, 280: 40301-40309.

McNeill RE, Miller N, Kerin MJ: Evaluation and validation of candidate endogenous control genes for real-time quantitative PCR studies of breast cancer. *BMC Mol Biol* 2007, 8: 107.

Medina, D: Mammary developmental fate and breast cancer risk. *Endocr Relat Cancer* 2005, 12: 483-495.

Melchor L, Benítez J: An integrative hypothesis about the origin and development of sporadic and familial breast cancer subtypes. *Carcinogenesis* 2008, 29: 1475-1482.

Mocellin S, Rossi CR, Pilati P, Nitti D, Marincola M: Quantitative real-time PCR: a powerful ally in cancer research. *Trends Mol Med* 2003, 9: 189-195.

Monteiro AN, August A, Hanafusa H: Evidence for a transcriptional activation function of BRCA1 C-terminal region. *Proc Natl Acad Sci U S A.* 1996, 93: 13595-13599.

Monteiro AN. BRCA1: exploring the links to transcription. *Trends Biochem Sci* 2000 Oct,25(10):469-74.

Moparthi SB, Arbman G, Wallin A, Kayed H, Kleeff J, Zentgraf H, Sun XF: Expression of MAC30 protein is related to survival and biological variables in primary and metastatic colorectal cancers. *Int J Oncol* 2007, 30: 91-95.

Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet* 2003, 19: 570-577.

Moriya T, Kasajima A, Ishida K, Kariya Y, Akahira J, Endoh M, Watanabe M, Sasano H: New trends of immunohistochemistry for making differential diagnosis of breast lesions. *Med Mol Morphol* 2006, 39: 8-13.

Morse DL, Carroll D, Weberg L, Borgstrom MC, Ranger-Moore J, Gillies RJ: Determining suitable internal standards for mRNA quantification of increasing cancer progression in human breast cells by real-time reverse transcriptase polymerase chain reaction. *Anal Biochem* 2005, 342: 69-77.

Mullan PB, Quinn JE, Harkin DP: The role of BRCA1 in transcriptional regulation and cell cycle control. *Oncogene* 2006, 25: 5854-5863.

Murphy M, Pykett MJ, Harnish P, Zang KD, George DL: Identification and characterization of genes differentially expressed in meningiomas. *Cell Growth Differ* 1993, 4: 715-722.

Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, van de Rijn M, Perou CM: Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 2004, 10: 5367-5374.

Nobukuni Y, Kohno K, Miyagawa K: Gene trap mutagenesis-based forward genetic approach reveals that the tumor suppressor OVCA1 is a component of the biosynthetic pathway of diphthamide on elongation factor 2. *J Biol Chem* 2005, 280: 10572-10577.

Noetzel E, Veeck J, Niederacher D, Galm O, Horn F, Hartmann A, Knüchel R, Dahl E: Promoter methylation-associated loss of ID4 expression is a marker of tumour recurrence in human breast cancer. *BMC Cancer* 2008, 8: 154

Ohl F, Jung M, Radonic A, Sachs M, Loening SA, Jung K: Identification and validation of suitable endogenous reference genes for gene expression studies of human bladder cancer. *J Urol* 2006, 175: 1915-1920.

Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G, Loening SA, Radonic A, Jung K: Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? *Mol Med* 2005, 83: 1014-1024.

Oldenburg RA, Meijers-Heijboer H, Cornelisse CJ, Devilee P: Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol* 2007, 63: 125-149.

Oliveira V, Romanow WJ, Geisen C, Otterness DM, Mercurio F, Wang HG, Dalton WS, Abraham RT: A protective role for the human SMG-1 kinase against tumor necrosis factor-alpha-induced apoptosis. *J Biol Chem* 2008, 283: 13174-13184.

Ongusaha PP, Ouchi T, Kim KT, Nytko E, Kwak JC, Duda RB, Deng CX, Lee SW. BRCA1 shifts p53-mediated cellular outcomes towards irreversible growth arrest. *Oncogene* 2003, 22: 3749-3758.

Oshiro MM, Kim CJ, Wozniak RJ, Junk DJ, Munoz-Rodriguez JL, Burr JA, Fitzgerald M, Pawar SC, Cress AE, Domann FE, Futscher BW: Epigenetic silencing of DSC3 is a common event in human breast cancer. *Breast Cancer Res* 2005, 7: 669-680.

Ouchi T, Lee SW, Ouchi M, Aaronson SA, Horvath CM: Collaboration of signal transducer and activator of transcription 1 (STAT1) and BRCA1 in differential regulation of IFN-gamma target genes. *Proc Natl Acad Sci U S A* 2000, 97: 5208-5213.

Palacios J, Robles-Frías MJ, Castilla MA, López-García MA, Benítez J: The molecular pathology of hereditary breast cancer. *Pathobiology* 2008,75: 85-94.

Parr C, Gan CH, Watkins G, Jiang WG: Reduced vascular endothelial growth inhibitor (VEGI) expression is associated with poor prognosis in breast cancer patients. *Angiogenesis* 2006, 9:73-81.

Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 2004, 29:1213-1222.

Pellikainen J, Kataja V, Ropponen K, Kellokoski J, Pietiläinen T, Böhm J, Eskelinen M, Kosma VM: Reduced nuclear expression of transcription factor AP-2 associates with aggressive breast cancer. *Clin Cancer Res* 2002, 8: 3487-3495.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: Molecular portraits of human breast tumours. *Nature* 2000, 406: 747-752.

Pfaffl, M. W: A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001, 29: e45.

Polyak K: Pregnancy and breast cancer: the other side of coin. *Cancer Cell* 2006, March: 151-153.

Potemski, P., Pluciennik, E., Bednarek, A.K., Kusinska, R., Kubiak, R., Jesionek-Kupnicka, D., Watala, C., Kordek, R. Ki-67 expression in operable breast cancer: a comparative study of immunostaining and a real-time RT-PCR assay. *Pathol Res Pract* 202:491-495, 2006.

Pusztai L, Ayers M, Stec J: Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal largescale molecular differences between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res* 2003, 9: 2406–2415.

Qiu X, Xiao Y, Gordon A, Yakovlev A: Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics* 2006, 7: 50.

R. Rasmussen. Quantification on the LightCycler. In: Meuer S., Witter C., Nakagawara K, eds. Rapid cycle real-time PCR, methods and applications. Berlin: Springer, 2001: 21-34.

Rakha EA, El-Sheikh SE, Kandil MA, El-Sayed ME, Green AR, Ellis IO: Expression of BRCA1 protein in breast cancer and its prognostic significance. *Hum Pathol* 2008, 39: 857-865.

Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 2002, 62: 4427-4433.

Rhodes DR, Yu J, Shanker K: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* 2004,101: 9309-9314.

Rhodes DR, Yu J, Shanker K: ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 2004a, 6: 1-6.

Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* 2006, 9: 121-132.

Rodgers JL and Nicewander WA: Thirteen ways to look at the correlation coefficient. *The American Statistician* 1988, 42: 59–66.

Rosen EM, Fan S, Pestell RG, Goldberg ID: BRCA1 gene in breast cancer. *J Cell Physiol* 2003, 196:19-41

Rouzier R, Perou CM, Symmans WF: Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* 2005, 11: 5678-5685.

Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, Gamallo C, Esteller M, Palacios J: Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. *Int J Cancer* 2003, 106: 208-215.

Saviozzi S, Cordero F, Loacono M, Novello S, Scagliotti GV, Calogero RA: Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. *BMC Cancer* 6:200-209, 2006.

Schmittgen TD, Zakrajsek BA: Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. *J Biochem Biophys Methods* 2000, 46: 69-81.

Schneider J, Ruschhaupt M, Buness A: Identification and meta-analysis of a small gene expression signature for the diagnosis of estrogen receptor status in invasive ductal breast cancer. *Int J Cancer* 2006,119: 2974-2979.

Schultz DC, Vanderveer L, Berman DB, Hamilton TC, Wong AJ, Godwin AK: Identification of two candidate tumor suppressor genes on chromosome 17p13.3. *Cancer Res* 1996, 56: 1997-2002.

Scully R, Chen J, Ochs RL, Keegan K, Hoekstra M, Feunteun J, Livingston DM: Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell* 1997a, 90: 425-435.

Scully R, Chen J, Plug A, Xiao Y, Weaver D, Feunteun J, Ashley T, Livingston DM: Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* 1997, 88: 265-275.

Selvey S, Thompson EW, Matthaei K, Lea RA, Irving MG, Griffiths LR: Beta-actin-an unsuitable internal control for RT-PCR. *Mol Cell Probes* 2001, 15:307-311.

Sherlock G, Hernandez-Boussard T, Kasarskis A: The Stanford microarray database. *Nucleic Acids Res* 2001,**29**: 152-155.

Shim H, Lau SK, Devi S, Yoon Y, Cho HT, Liang Z: Lower expression of CXCR4 in lymph node metastases than in primary breast cancers: potential regulation by ligand-dependent degradation and HIF-1alpha. *Biochem Biophys Res Commun* 2006, 346: 252-258.

Shinohara A, Ogawa H, Ogawa T. Rad51 protein involved in repair and recombination in S. cerevisiae is a RecA-like protein. *Cell* 1992, 69: 457-470.

Shrivastav M, De Haro LP, Nickoloff JA: Regulation of DNA double-strand break repair pathway choice. *Cell Res* 2008,18:134-147.

Simpson PT, Reis-Filho JS, Lambros MB, Jones C, Steele D, Mackay A, Iravani M, Fenwick K, Dexter T, Jones A, Reid L, Da Silva L, Shin SJ, Hardisson D, Ashworth A, Schmitt FC, Palacios J, Lakhani SR: Molecular profiling pleomorphic lobular carcinomas of the breast: evidence for a common molecular genetic pathway with classic lobular carcinomas. *J Pathol* 2008, 215: 231-244.

Sládek NE: Aldehyde dehydrogenase-mediated cellular relative insensitivity to the oxazaphosphorines. *Curr Pharm Des* 1999, 5: 607-625.

Smith DD, Saetrom P, Snøve O Jr, Lundberg C, Rivas GE, Glackin C, Larson GP: Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics* 2008, 9: 63.

Sorlie T, Perou CM, Tibshirani R: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001,98: 10869-10874.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Bodstein D: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003, 100: 8418-8423.

Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A.* 2003, 100:10393-10398.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006, 98: 262-72

Spanakis, E: Problems related to the interpretation of autoradiographic data on gene expression using common constitutive transcripts as controls. *Nucleic Acids Res* 1993, 21:3809-3819.

Spillman MA, Bowcock AM: BRCA1 and BRCA2 mRNA levels are coordinately elevated in human breast cancer cells in response to estrogen. *Oncogene* 1996, 13:1639-1645.

Stahlberg A, Hakansson J, Xian X, Semb H, Kubista M: Properties of the reverse transcription reaction in mRNA quantification. *Clin Chem* 2004, 50:509-515.

Stahlberg, A., Kubista, M., Pfaffl, M. Comparison of reverse transcriptases in gene expression analysis. *Clin Chem* 2004, 50:1678-1680.

Stanford Microarray Database [http://genome-www5.stanford.edu/]

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, and Heinen, E: Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology* 1999, 75: 291-295.

Thomas JE, Smith M, Tonkinson JL, Rubinfeld B, Polakis P: Induction of phosphorylation on BRCA1 during the cell cycle and after DNA damage. *Cell Growth Differ* 1997, 8: 801-809.

Thomassen M, Tan Q, Kruse TA: Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis. *Breast Cancer Res Treat* 2008, Feb 22, doi: 10.1007/s10549-008-9927-2.

Toedling J, Spang R: Assessment of Five Microarray Experiments on Gene Expression Profiling of Breast Cancer. Poster Presentation RECOMB 2003. [http://citeseer.ist.psu.edu/611350.html].

Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distante V, Pazzagli M, Bustin S.A., Orlando C: Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. *Anal Biochem* 2002, 309: 293-300.

Tripathi A, King C, de la Morenas A, Perry VK, Burke B, Antoine GA, Hirsch EF, Kavanah M, Mendez J, Stone M, Gerry NP, Lenburg ME, Rosenberg CL: Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int J Cancer* 2008, 122:1557-1566.

Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, Klein J, Fridman E, Skarda J, Srovnal J, Hajduch M, Murray P, Kolar Z: Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* 2007, 7:55.

Turner NC, Reis-Filho JS: Basal-like breast cancer and the BRCA1 phenotype. *Oncogene* 2006, 25: 5846-5853

Tyagi S, Kramer FR: Molecular beacons: probes that fluoresce upon hybridization. *Nat Biotechnol* 1996, 14: 303-308.

Umetani N, Mori T, Koyanagi K, Shinozaki M, Kim J, Giuliano AE, Hoon DS: Aberrant hypermethylation of ID4 gene promoter region increases risk of lymph node metastasis in T1 breast cancer. *Oncogene* 2005, 24:4721-4727

Valenti MT, Bertoldo F, Dalle Carbonare L, Azzarello G, Zenari S, Zanatta M, Balducci E, Vinante O, and Lo Cascio V: The effect of biphosphates on gene expression: GAPDH as a housekeeping or a new target gene? *BMC Cancer* 2006, 6: 49.

van de Vijver MJ, He YD, van't Veer LJ: A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, 347: 1999-2009.

van't Veer LJ, Dai H, van de Vijver MJ: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, 415: 530-536.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002,3: research0034.1– research0034.

Venkitaraman AR: Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 2002, 108:171-182.

Wang Y, Cortez D, Yazdi P, Neff N, Elledge SJ, Qin J: BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev* 2000, 14: 927-939

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005, 365: 671-9.

Warnat P, Eils R, Brors B: Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 2005, 6: 265.

Weigelt B, Horlings H, Kreike B, Hayes M, Hauptmann M, Wessels L, de Jong D, Van de Vijver M, Veer LjV, Peterse J: Refinement of breast cancer classification by molecular characterization of histological special types. *J Pathol* 2008, 216:141-150.

West M, Blanchette C, Dressman H: Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001,98: 11462-67.

West S, Gromak N, Proudfoot NJ: Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* 2004, 432:522-525.

Whitcombe D, Theaker J, Guy SP, Brown T, Little S: Detection of PCR products using self-probing amplicons and fluorescence. *Nat Biotechnol* 1999, 17: 804-807.

Wilcoxon F: Individual comparisons by ranking methods. *Biometrics Bulletin* 1945, 1: 80–83.

Wong ML, and Medrano JF: Real-time PCR for mRNA quantitation. *Biotechniques* 2005, 39:1.

Wu X, Petrini JH, Heine WF, Weaver DT, Livingston DM, Chen J: Independence of R/M/N focus formation and the presence of intact BRCA1. *Science* 2000, 289:11.

Wu G, Xing M, Mambo E, Huang X, Liu J, Guo Z, Chatterjee A, Goldenberg D, Gollin SM, Sukumar S, Trink B, Sidransky D: Somatic mutation and gain of copy number of PIK3CA in human breast cancer. *Breast Cancer Res* 2005, 7:609-616.

Xu L, Geman D, Winslow RL: Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics* 2007,**8**: 275.

Yamashita A, Kashima I, Ohno S: The role of SMG-1 in nonsense-mediated mRNA decay. *Biochim Biophys Acta* 2005, 754: 305-315

Yamashita A, Ohnishi T, Kashima I, Taya Y, Ohno S: Human SMG-1, a novel phosphatidylinositol 3-kinase-related protein kinase, associates with components of the mRNA surveillance complex and is involved in the regulation of nonsense-mediated mRNA decay. *Genes Dev* 2001, 15: 2215-28

Yan Y, Black CP, Cao PT, Haferbier JL, Kolb RH, Spieker RS, Ristow AM, Cowan KH: Gamma-irradiation-induced DNA damage checkpoint activation involves feedback regulation between extracellular signal-regulated kinase 1/2 and BRCA1. *Cancer Res* 2008,68: 5113-5121

Yan PS, Rodriguez FJ, Laux DE, Perry MR, Standiford SB, Huang TH: Hypermethylation of ribosomal DNA in human breast carcinoma. *Br J Cancer* 2000, 82: 514-517.

Yin X, Dewille JW, Hai T: A potential dichotomous role of ATF3, an adaptive response gene, in cancer development. *Oncogene* 2008, 27:2118-2127.

Yoder BJ, Wilkinson EJ, Massoll NA: Molecular and morphologic distinctions between infiltrating ductal and lobular carcinoma of the breast. *Breast J* 2007, 13: 172-179.

Zhang H, Somasundaram K, Peng Y, Tian H, Zhang H, Bi D, Weber BL, El-Deiry WS: BRCA1 physically associates with p53 and stimulates its transcriptional activity. *Oncogene* 1998,16: 1713-1721.

Zhang Z, Yamashita H, Toyama T, Sugiura H, Ando Y, Mita K, Hamaguchi M, Hara Y, Kobayashi S, Iwase H: Quantitation of HDAC1 mRNA expression in invasive carcinoma of the breast. *Breast Cancer Res Treat* 2005*,* 94: 11-16.

Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, Bukholm IK, Karesen R, Botstein D, Borresen-Dale AL, Jeffrey SS: Different Gene Expression Patterns in Invasive Lobular and Ductal Carcinomas of the Breast. *Mol Biol Cell* 2004, 15:2523-2536.

Zhang ZY, Zhao ZR, Adell G, Jarlsfelt I, Cui YX, Kayed H, Kleeff J, Wang MW, Sun XF: Expression of MAC30 in rectal cancers with or without preoperative radiotherapy. *Oncology* 2006, 71: 259-65.

Zheng L, Pan H, Li S, Flesken-Nikitin A, Chen PL, Boyer TG, Lee WH: Sequence-specific transcriptional corepressor function for BRCA1 through a novel zinc finger protein, ZBRK1. *Mol Cell* 2000,6: 757-768.

Zhong Q, Chen CF, Li S, Chen Y, Wang CC, Xiao J, Chen PL, Sharp ZD, Lee WH: Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. *Science* 1999a, 285: 747-750.

Zhong H, Simons JW: Direct comparison of GAPDH, β-Actin, Cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. *Biochem Biophys Res Commun* 1999, 259: 523-526.

Appendix A: Main script for “Resampling” analysis by using Matlab program

Appendix B: Gene lists obtained through meta-analysis of Sorlie and Zhao datasets

1. The list of 298 IMAGE clones differentially expressed between ductal (D) and normal (N) samples with 90% significance.
2. The list of 216 IMAGE clones differentially expressed between lobular (L) and normal (N) samples with 90% significance.
3. List of image clones which are differentially expressed between ductal and lobular samples with 80% significance.
4. List of meta-analysis specific genes differentially expressed in DN, LN and DL.

Appendix C: Validation of meta-analysis gene lists by three independent microarray datasets.

1. Validation of DN meta-gene list by three independent microarray datasets.
2. Validation of LN meta-gene list by three independent microarray datasets.

Appendix D: Comparison of DL list with other published data sets

Appendix A: Main script for “Resampling” analysis by using Matlab program

2.14 Main script for “Resampling” analysis by using Matlab program

```
function f=multirandomsampler(x1,x2,y1,y2,g1,g2,gr1,gr2,n,t,fexp);
% inputs: x1-Sorlie22, x2-Zhao22, y1-SorlieID22, y2-ZhaoID22, g1-starting gene, g2-
finishing gene, gr1-first group(1 2 or 4)
%gr2-second group, n-sampling sizes like [3 4 5], t-array of iteration numbers like
[10 20 30 40 50] fexp=expected FDR value.

%read the xls files
[X1A,X1B]=xlsread(x1);  %read sorlie data
[X2A,X2B]=xlsread(x2);  %read zhao data
[Y1A,Y1B]=xlsread(y1);  %read sorlieID
[Y2A,Y2B]=xlsread(y2);  %read zhaoID

if mean(X1A(1,:))==1    %if the first row of numerical data is all 1 then remove it
    X1A(1,:)=[];
end
if mean(X2A(1,:))==1    %if the first row of numerical data is all 1 then remove it
    X2A(1,:)=[];
end

%set the genes to be analyzed
X=[X1A(g1:g2,:) X2A(g1:g2,:)];  %g1 and g2 determine the gene indexes; set the data
for these set of genes
%set the id information for the genes selected
XG=[X1B(g1:g2+1,:) X2B(g1:g2+1,:)];

%set the first row as group 1=ductal 2=lobular 4=normal
Y=[Y1A(1,:) Y2A(1,:)];
%call multiplerandomrep for each of the entered n values

for i=1:length(n)
    fprintf('n=%d\n',n(i));   %give information about the progress
    evaluator(X,Y,g1,g2,gr1,gr2,n(i),t,fexp);
end

return

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
%commands to be run:,
%multirandomsampler('sorlieD.xls','zhaoD.xls','sorlieID22.xls','zhaoID22.xls',1,1726,
1,4,[3 4 5 6],[10 20 30 40 50 60 70 80 90 100 150],0.01);
```

```
%multirandomsampler('sorlieL.xls','zhaoL.xls','sorlieID22.xls','zhaoID22.xls',1,2029,
2,4,[3 4 5 6],[10 20 30 40 50 60 70 80 90 100 150],0.01);
%multirandomsampler('sorlieDL.xls','zhaoDL.xls','sorlieID22.xls','zhaoID22.xls',1,152
2,1,2,[3 4 5 6],[10 20 30 40 50 60 70 80 90 100 150],0.01);
%multirandomsampler('sorlieDL.xls','zhaoDL.xls','sorlieID22.xls','zhaoID22.xls',1,152
2,1,2,[10 15 20],[10 20 30 40 50 60 70 80 90 100 150],0.01);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
```

Functions that the main script calls

“evaluator” funtion

```
function f = evaluator(X,Y,g1,g2,gr1,gr2,n,t,fexp);
% %X=sorlie+zhao numerical data, Y= sorlie+zhao ID data, g1-starting gene, g2-
finishing gene, gr1-first group(1 2 or 4)
%gr2-second group, n-sampling size, t-array of iteration numbers like [10 20 30 40
50]

%invoke randomrep3m5 for the first iteration value and get the count value
%in order to create allmeana and allmeanb in exact sizes
[ksvalues meana meanb fobs count wokay]=randomsampler(X,Y,g1,g2,gr1,gr2,n,t(1),fexp);
%invoke randomrep31 and get mean values
allmeana=zeros(count,length(t));  %empty matrix to store all mean values of group 1
allmeanb=zeros(count,length(t));  %empty matrix to store all mean values of group 2
validnum=zeros(g2-g1+1,length(t)); %empty matrix to store wokay values
validnum(:,1)=wokay;
counts=zeros(1,length(t));
counts(1)=count;

k=1;
for j=1:length(meana)
        if(isnan(meana(j))==0)&(isnan(meanb(j))==0)  %if a row is not NaN then store
it in allmeana or allmeanb
            allmeana(k,1)=meana(j);
            allmeanb(k,1)=meanb(j);
            k=k+1;   %this is a count for row number of allmeana and allmeanb
        end
end
fprintf('%d/%d is completed...\n',1,length(t));  %information about the progress
       i=2;  %continue with second iteration value and make sure that the same
process is repeated until all iteration calues are used
while(i<=length(t))
    [ksvalues meana meanb fobs count
wokay]=randomsampler(X,Y,g1,g2,gr1,gr2,n,t(i),fexp);  %invoke randomrep3n5 and get
mean values
    validnum(:,i)=wokay;
    counts(i)=count;
    k=1;
    for j=1:length(meana)
        if(isnan(meana(j))==0)&(isnan(meanb(j))==0)  %if a row is not NaN then store
it in allmeana or allmeanb
            allmeana(k,i)=meana(j);
            allmeanb(k,i)=meanb(j);
            k=k+1;     %this is a count for row number of allmeana and allmeanb
        end
    end
    fprintf('%d/%d is completed...\n',i,length(t));   %information about the progress
    i=i+1;
end

valids=zeros(1,length(t)); %empty vector to store sum of each column of validnum
derva=zeros(count,1);   %empty matrix for derivatives of group 1 mean values
dervb=zeros(count,1);   %empty matrix for derivatives of group 2 mean values
tt=[];  %empty array
for i=1:length(t)-1  %subtruct mean values of (n)th iteration from those of (n+1)th
iteration to obtain derivatives
    derva(:,i)=allmeana(:,i+1)-allmeana(:,i);
    dervb(:,i)=allmeanb(:,i+1)-allmeanb(:,i);
    valids(i)=sum(validnum(:,i));  %sum each column of validnum
    if(i==1)
        tt=[(t(i+1)-t(i))];     %create the y axis of the final plot
    else
% for 10 20 30 40 50 100  it should be 10 20 30 40 90
        tt=[tt (tt(i-1)+t(i+1)-t(i))];
    end
end
valids(length(t))=sum(validnum(:,length(t))); %sum the last column of validnum

stda=std(derva);   %standard deviation
stdb=std(dervb);   %standard deviation

meanderva=mean(abs(derva));    %get the absolute mean of derivatives
meandervb=mean(abs(dervb));

if(gr1==1)           %in order to set appropriate title for each graph
    ss1='DUCTAL';
elseif(gr1==2)
    ss1='LOBULAR';
elseif(gr1==4)
    ss1='NORMAL';
end
if(gr2==1)
    ss2='DUCTAL';
elseif(gr2==2)
```

```
    ss2='LOBULAR';
elseif(gr2==4)
    ss2='NORMAL';
end

figure
H1=errorbar(tt,meanderva,stda,stda);  %plot for group 1
H1=title([ss1 ' N=' num2str(n)]);     %set title for the graph
saveas(gcf,[ss1 ' N_' num2str(n)],'jpg');  %save the graph
close all

figure
H2=errorbar(tt,meandervb,stdb,stdb);  %plot for group 2
H2=title([ss2 ' N=' num2str(n)]);     %set title for the graph
saveas(gcf,[ss2 ' N_' num2str(n)],'jpg');  %save the graph
close all

percounts=(100.*valids)./counts;
figure
H3=plot(t,percounts);    %plot valids(number of valid genes) versus iteration number
H3=title([ss1 '/' ss2 ' N=' num2str(n)]);   %set title for the graph
saveas(gcf,[ss1 '_' ss2 '__' 'N_' num2str(n)],'jpg');  %save the graph
close all

A=textread('Masterfile2.txt');   % Read the master file and add into it
[z1,z2]=size(A);
if(z2>1)
    A(:,z2-1)=[];   %remove zero columns which appears for an unknown reason
    A(:,z2-1)=[];
end

id=gr1+10*gr2+100*n;  %determine an id
wmeanderva=[id; meanderva'];
wmeandervb=[id; meandervb'];
wstda=[id; stda'];
wstdb=[id; stdb'];


W=[A t' wmeanderva wstda wmeandervb wstdb valids' counts' percounts'];  %create a
matrix to write into the master file
save Masterfile2.txt W -ascii -TABS;
return
```

“randomsampler” Function

```
function [ksvalues,meana,meanb,fobs,count,wokay] =
randomsampler(X,Y,g1,g2,gr1,gr2,n,t,fexp);
```

```
%X=sorlie+zhao numerical data, Y= sorlie+zhao ID data, g1 gene1 to g2
gene2,gr1=group1 and gr2=group2(1 2 or 4),
%n number of samples,t number of times, fexp=expected FDR value.

%find the first group = gr1
k=find(Y==gr1);
%find the second group = gr2
z=find(Y==gr2);
%recreate data for groups Xa and Xb for depending on their type (ductal, lobular,
normal)
Xa=X(:,k);
Xb=X(:,z);

[c1,c2]=size(X);

%generate zeros matrix for pvalues (to be calculated based on randomly selected
actual data) to be stored for each gene t times calculated
pvalues=zeros(c1,t);
mavalues=zeros(c1,t);
mbvalues=zeros(c1,t);

pvaluesr1=zeros(c1,t); %p values for random data
ksvalues=zeros(c1,1);
pvaluesr2=zeros(c1,t);
ksvaluesrr=zeros(c1,1);

count=0;  %to count how many valid genes with less Nans
%for each gene
for i=1:c1
    %for each time t
    %NEWCODE
    %find gene IDs with too many NaNs
        naa=length(k)-sum(isnan(Xa(i,:)));
        nab=length(z)-sum(isnan(Xb(i,:)));
        if ((naa./length(k))<0.6 | (nab./length(z))<0.6 | naa<n | nab<n)
            pvalues(i,:)=NaN;
            pvaluesr(i,:)=NaN;
            pvaluesr2(i,:)=NaN;
            mavalues(i,:)=NaN;
            mbvalues(i,:)=NaN;
        else
            count=count+1;  %count excepted genes

            tempx=[];
            tempy=[];
        for b=1:c2                          % for each row of X and Y
            if(isnan(X(i,b))==0)
```

```
                tempx=[tempx X(i,b)];       %create tempx which is one row of X and
has no NaN
                tempy=[tempy Y(1,b)];       %create tempy which does not contain the
types of tumors that are missing in X
           end
        end
        tempk=find(tempy==gr1);      %create tempk which contains indices of group 1
        tempz=find(tempy==gr2);      %create tempy which contains indices of group 2
        tempxa=tempx(tempk);         %tempxa equals to one Xa row that has no NaNs
        tempxb=tempx(tempz);         %tempxb equals to one Xb row that has no NaNs
        tempxx=[tempxa tempxb];      %tempxb equals to one XX row that has no NaNs

        for j=1:t

        xtai=randperm(length(tempk));   %generate random numbers of length of the
first group
        Xai=tempxa(xtai(1:n));          %randomly select from the first group, n
number of samples, call it Xai

        xtbi=randperm(length(tempz));   %generate random numbers of length of the
second group
        Xbi=tempxb(xtbi(1:n));          %randomly select from the second group, n
number of samples

        p=ranksum(Xai(1:n),Xbi(1:n));   %perform ranksum test for real data;
        ma=mean(Xai(1:n));              %calculate the mean of the group 1 for gene i
for each time t
        mb=mean(Xbi(1:n));              %calculate the mean of the group 2 for gene i
for each time t
        pvalues(i,j)=p;                 %allocate to pvalues for ij actual data n
samples rank test probabilty
        mavalues(i,j)=ma;               %allocate to matrix for mean of group1--
>mavalues
        mbvalues(i,j)=mb;               %allocate to matrix for mean of group2--
>mbvalues

        %randomized data

        ll=length(tempk)+length(tempz);              %call the length of both groups
ll
        lk=length(tempk);                        %length of first group
        lz=length(tempz);                        %length of second group

        XXti=randperm(ll);                              %generate random numbers of
length of both groups
        XXa=tempxx(XXti(1:ll));                           %shuffle the combined group
data fully
```

```
        XXai=XXa(1:n);                                    %pick the first n as the
first group assign it as XXai
        XXbi=XXa(lk+1:lk+n);

        %significance test for randomized actual data for each gene
        p=ranksum(XXai(1:n),XXbi(1:n)); %do ranksum test for n samples of shuffled
data for each time t
        pvaluesr1(i,j)=p;                %assign to pvalues matrix for fully
randomized data, rank test probability

        %second randomized data
       XXti2=randperm(ll);                              %generate random numbers of
length of both groups
       XXa2=tempxx(XXti2(1:ll));                           %shuffle the combined
group data fully
       XXai2=XXa2(1:n);                                  %pick the first n as the
first group assign it as XXai
       XXbi2=XXa2(lk+1:lk+n);

       %significance test for randomized data for each gene
       p=ranksum(XXai2(1:n),XXbi2(1:n)); %do ranksum test for n samples of shuffled
data for each time t
       pvaluesr2(i,j)=p;                %assign to pvalues matrix for fully
randomized data, rank test probability
   end
   end

   %calculate kolmogorov between random and real data for each gene
   if (isnan(sum(pvalues(i,:))) | isnan((pvaluesr1(i,:))))
       ksvalues(i)=NaN;
   else
       [H,P,ST]=kstest2(pvalues(i,:),pvaluesr1(i,:));   %ks test between actual data
pvalues and randomized data p values
       ksvalues(i)=P;                                   %assign the P, probability,
to ksvalues matrix to report
   end

   %calculate kolmogorov between random and random data for each gene
   if (isnan(sum(pvaluesr1(i,:))) | isnan((pvaluesr2(i,:))))
       ksvaluesrr(i)=NaN;
   else
      [H,P,ST]=kstest2(pvaluesr1(i,:),pvaluesr2(i,:));   %ks test between randomized
data 1 pvalues and randomized data 2 p values
      ksvaluesrr(i)=P;                                   %assign the P, probability,
to ksvalues matrix to report
  end
end
```

```
meana=mean(mavalues')';   %take the mean of mavalues
meanb=mean(mbvalues')';   %take the mean of mbvalues
num=fexp*count;       %expected FDR times counted gene number
sortedksv=sort(ksvaluesrr); %sort ksvaluesrr so that smaller ones are at the top
fobs=sortedksv(round(num)); %determine the observed FDR value


A=textread('Masterfile.txt');   % Read the master file and add into it
[z1,z2]=size(A);
if(z2>1)
    A(:,z2-1)=[];   %remove zero columns which appears for an unknown reason
    A(:,z2-1)=[];
end

id=gr1+10*gr2+100*n+1000*t;  %determine an id

wfobs=zeros(g2-g1+1,1);
wfobs(:,1)=fobs;    %create a column which is g2 long and all the rows contains the
same fobs value
wokay=zeros(g2-g1+1,1);
wwokay=zeros(g2-g1+2,1);

for i=g1:g2;
    if ksvalues(i)<fobs    %if ksvalues of gene is smaller then fobs value then for
this gene ksvalue is significant
        wokay(i)=1;        %so set as 1
    elseif (isnan(ksvalues(i)))  %if ksvalue is NaN
        wokay(i)=0;
    else               %or ksvalue is larger than fobs then this ksvalue is not
acceptable
        wokay(i)=0;    %so set as 0
    end
end

wwokay=[id;wokay];      %add the same id to the first row all the variables that are
going to be written in the master file
wfobs=[id;wfobs];
wmeana=[id;meana];
wmeanb=[id;meanb];
wksvalues=[id;ksvalues];

W=[A wksvalues wmeana wmeanb wfobs wwokay];  %create a matrix to write into the
master file
save Masterfile.txt W -ascii -TABS;

return
```

Appendix B: Gene lists obtained through the meta-analysis of Sorlie and Zhao datasets

1. The list of 298 IMAGE clones differentially expressed between ductal (D) and normal (N) samples with 90% significance.

Clone ID	Gene name	Mean expression of ductal samples	Mean expression of normal samples	Expression difference
IMAGE:377461	CAV1	-2.298	0.235	-2.533
IMAGE:845363	NME1	-0.902	-2.532	1.630
IMAGE:739126	TSTA3	0.937	-0.490	1.427
IMAGE:782635	NDUFAB1	-0.320	-1.306	0.986
IMAGE:509495	PSMA6	-0.544	-1.710	1.166
IMAGE:123614	C20orf55	1.433	0.485	0.948
IMAGE:813698	SPRY2	-1.368	0.125	-1.493
IMAGE:897770	CRABP2	2.042	0.120	1.922
IMAGE:2407433	DPYSL3	0.432	2.012	-1.580
IMAGE:753104	DCT	-0.075	2.597	-2.672
IMAGE:590774	MAPK13	-1.652	-3.040	1.388
IMAGE:362409	GAD1	-0.886	-2.063	1.178
IMAGE:781089	PTTG1	-1.966	-3.727	1.760
IMAGE:785933	SRPX	-1.540	0.425	-1.965
IMAGE:841149	TGFBR2	0.055	1.597	-1.542
IMAGE:727251	CD9	1.197	-0.039	1.236
IMAGE:195129	SFRS2IP	-0.222	-0.813	0.592
IMAGE:815526	MYBL2	-2.061	-4.260	2.199
IMAGE:139009	FN1	-0.029	-2.818	2.788
IMAGE:743804	SEC23B	-0.240	-1.240	1.000
IMAGE:882510	KPNA2	-1.365	-2.748	1.384
IMAGE:511632	POLR3K	-0.322	-2.123	1.801
IMAGE:795198	SLC39A1	0.508	-0.515	1.024
IMAGE:739109	AP2S1	-0.541	-1.607	1.066
IMAGE:884655	GARS	-1.192	-2.291	1.099
IMAGE:484535	AOC3	0.571	2.933	-2.363
IMAGE:472185	ADAMTS1	-0.486	1.108	-1.594
IMAGE:769686	THY1	0.397	-0.732	1.129
IMAGE:840691	STAT1	1.175	-0.277	1.452
IMAGE:362483	SPTBN1	-0.922	0.753	-1.675
IMAGE:755145	VIL2	0.510	-0.853	1.363
IMAGE:838568	COX6C	0.590	-1.886	2.476
IMAGE:196866	NARG1	-0.596	-1.335	0.739
IMAGE:214990	GSN	1.020	2.970	-1.950
IMAGE:825583	RALY	-0.390	-1.119	0.729
IMAGE:144786	BGN	2.001	-0.044	2.045
IMAGE:782718	APOA1BP	0.295	-0.959	1.254
IMAGE:842825	GSPT1	-0.547	-1.651	1.105
IMAGE:855390	MCM6	-1.484	-2.648	1.165

IMAGE:739183		0.715	-0.467	1.181
IMAGE:595213	FAM49B	-0.189	-1.262	1.072
IMAGE:488839	ARID4B	0.514	-0.622	1.136
IMAGE:143966		-0.546	-1.508	0.962
IMAGE:724615	CHC1	-0.885	-1.705	0.820
IMAGE:43229	PCNA	-1.486	-3.052	1.566
IMAGE:252259	COL17A1	0.254	1.992	-1.738
IMAGE:785707	PRC1	0.645	-1.520	2.165
IMAGE:789182	PCNA	-1.358	-2.853	1.495
IMAGE:246869	ZNF207	-0.364	-1.203	0.839
IMAGE:795538	SLC25A22	-0.060	-1.133	1.073
IMAGE:244974	FLJ22875	-0.344	-1.483	1.139
IMAGE:22918	EBNA1BP2	-0.762	-1.907	1.145
IMAGE:739901	CYP51A1	0.847	-0.320	1.167
IMAGE:810156	DTYMK	-0.990	-1.944	0.954
IMAGE:898035	CTSB	0.530	-0.628	1.158
IMAGE:52930	PLEKHC1	-1.023	0.278	-1.301
IMAGE:795936	TSN	-1.549	-2.800	1.251
IMAGE:179276	FASN	0.298	2.159	-1.862
IMAGE:770614	RGS19IP1	0.458	-0.693	1.152
IMAGE:612274	TUBA1	-0.692	-1.708	1.016
IMAGE:856447	IFI30	1.292	-1.007	2.298
IMAGE:80410	FDPS	-0.591	-1.502	0.911
IMAGE:85497	C2	0.620	-0.742	1.361
IMAGE:841340	TAP1	1.393	0.088	1.305
IMAGE:855624	ALDH1A1	-0.850	0.888	-1.738
IMAGE:245277	ITM2A	-1.433	0.752	-2.184
IMAGE:809473	PTRF	-0.126	1.476	-1.602
IMAGE:232612	LMAN1	-0.344	-1.677	1.333
IMAGE:489208	C16orf34	0.165	-1.242	1.407
IMAGE:432564	SF3B4	-0.240	-1.267	1.027
IMAGE:626502	ARPC1B	-0.329	-1.645	1.316
IMAGE:884425	CCT5	-1.185	-2.578	1.393
IMAGE:154185	CD68	0.543	-0.650	1.193
IMAGE:757873	CDK5R1	0.995	3.065	-2.070
IMAGE:713660	GPM6B	-0.811	0.922	-1.732
IMAGE:868368	TMSB4X	0.780	-0.168	0.948
IMAGE:265060	KIT	0.814	2.982	-2.168
IMAGE:884719	HSPA8	-0.421	-1.429	1.009
IMAGE:687875	CTSS	1.592	0.282	1.310
IMAGE:897531	MCAM	-0.588	-2.297	1.709
IMAGE:244147		1.766	-0.172	1.938
IMAGE:300044	SLC35B3	0.818	0.222	0.596
IMAGE:129725	RBPSUH	0.443	-0.523	0.966
IMAGE:767817	POLR2F	-0.695	-1.327	0.631
IMAGE:753313	LAPTM5	0.639	-0.943	1.582
IMAGE:970613	RAD21	-0.317	-1.480	1.163
IMAGE:855521	KRT18	-0.386	-1.483	1.097
IMAGE:207358	SLC2A1	-1.267	-2.680	1.413
IMAGE:811999	ETF1	-0.642	-1.682	1.040
IMAGE:47665	C1orf2	-0.106	-1.139	1.033
IMAGE:431805	SIAT1	0.324	-0.535	0.859
IMAGE:358643	FARSLA	-0.734	-1.821	1.087

IMAGE:753620	IGFBP6	-0.444	0.687	-1.131
IMAGE:135221	S100P	-0.124	-2.915	2.791
IMAGE:810017	PLAUR	0.303	-0.977	1.279
IMAGE:841679	CIB1	0.574	-0.760	1.333
IMAGE:469412	FH	-0.707	-1.761	1.054
IMAGE:39313	MRPS31	0.041	-0.585	0.626
IMAGE:724588	ISGF3G	1.931	0.710	1.221
IMAGE:782811	HMGA1	-1.857	-3.166	1.309
IMAGE:470061	SIAH2	0.367	-0.745	1.112
IMAGE:730638	FNBP3	-0.659	-1.705	1.046
IMAGE:51448	ATF3	0.466	2.273	-1.807
IMAGE:810038	GRINA	0.999	-0.203	1.202
IMAGE:488488	TXNIP	0.540	2.040	-1.500
IMAGE:343167	SARA2	0.066	-0.962	1.028
IMAGE:897636	SEC13L1	-0.214	-1.207	0.993
IMAGE:626206	FLJ10719	-0.998	-1.993	0.995
IMAGE:782513	G1P3	1.405	0.114	1.291
IMAGE:46897	PMVK	0.747	-0.145	0.892
IMAGE:358531	JUN	1.054	2.887	-1.833
IMAGE:346134	CARHSP1	-0.871	-1.690	0.819
IMAGE:291880	MFAP2	0.936	-0.717	1.653
IMAGE:85060	SQRDL	1.354	0.482	0.872
IMAGE:770838	SLC12A7	0.843	0.022	0.821
IMAGE:744917	NINJ1	1.079	0.124	0.955
IMAGE:815501	LMNB2	-1.300	-2.347	1.046
IMAGE:81316	ARG99	-0.104	1.249	-1.353
IMAGE:669485	EGFR	-0.679	0.550	-1.229
IMAGE:753400	ACTL6A	-0.536	-1.287	0.750
IMAGE:340558	ARPC5	-0.116	-1.295	1.179
IMAGE:110467	CAV2	-0.746	1.485	-2.231
IMAGE:853906	HCG4P6	1.290	0.155	1.135
IMAGE:198093	EIF2S2	-0.972	-1.730	0.758
IMAGE:308484	UBE2D3	-0.417	-1.416	1.000
IMAGE:153743	UBE2Q	0.220	-0.592	0.812
IMAGE:810600	ZNF286	-1.384	-2.630	1.246
IMAGE:856135	SRPK1	-0.758	-1.517	0.759
IMAGE:545503	STAT1	1.016	-0.428	1.445
IMAGE:545242	STAT1	0.941	-0.750	1.691
IMAGE:626385	PGM3	-0.300	-1.383	1.084
IMAGE:130884	PCYT2	-0.409	-1.188	0.779
IMAGE:897690	TRA1	-0.495	-1.675	1.180
IMAGE:358936	C9orf75	0.586	-1.355	1.941
IMAGE:809466	TOMM40	-0.657	-1.620	0.963
IMAGE:789376	TXNRD1	-1.386	-2.220	0.834
IMAGE:742082	PCK1	-0.669	1.186	-1.855
IMAGE:525926	SDC1	0.942	-0.594	1.537
IMAGE:49351	PLXNA3	0.808	-0.060	0.868
IMAGE:753215	GNAI1	-0.738	0.359	-1.098
IMAGE:773286	SLC9A3R1	0.900	-0.550	1.449
IMAGE:199663	CCL15	0.501	2.107	-1.606
IMAGE:251685	CDH11	1.628	0.345	1.283
IMAGE:35191	SDF2	0.277	-0.643	0.920
IMAGE:208718	ANXA1	-0.621	0.541	-1.162

IMAGE:40017		-0.990	-2.198	1.207
IMAGE:39884	IMPDH1	-0.296	-1.053	0.756
IMAGE:196992	AKR1C1	-2.088	0.287	-2.375
IMAGE:950482	SNRPB	-1.051	-1.874	0.823
IMAGE:72050	CLNS1A	-0.958	-1.878	0.921
IMAGE:842765	IQWD1	0.857	0.047	0.811
IMAGE:856167	TARS	-1.045	-2.089	1.044
IMAGE:1032796	LSM1	0.134	-0.756	0.891
IMAGE:292212	ABCC5	0.794	-0.358	1.153
IMAGE:49595	THUMPD3	-0.617	-1.165	0.548
IMAGE:626016	NT5C3	-0.557	-1.283	0.727
IMAGE:358456	SEC61G	-0.089	-1.205	1.116
IMAGE:44292	LOC90133	-0.620	-1.674	1.054
IMAGE:323577	SLC4A2	-0.356	-1.180	0.824
IMAGE:72395	MSTP9	1.388	0.566	0.822
IMAGE:361974	PTN	0.096	2.883	-2.787
IMAGE:292731	HBXAP	-0.311	-1.213	0.902
IMAGE:731118	PTPNS1	0.102	1.073	-0.971
IMAGE:399898	CSPG2	1.023	-0.329	1.352
IMAGE:321492	TNPO1	-0.730	-1.390	0.660
IMAGE:786609	COL12A1	2.351	1.000	1.351
IMAGE:772437	PHLDB1	0.798	1.587	-0.789
IMAGE:754046	DXS9879E	0.017	-1.704	1.721
IMAGE:782476	GULP1	-1.024	-0.018	-1.006
IMAGE:341317	ZNF516	0.001	0.812	-0.811
IMAGE:45544	TAGLN2	0.285	-1.015	1.300
IMAGE:786067	CDC25B	-1.159	-2.035	0.876
IMAGE:752631	FGFR3	-0.605	-1.729	1.125
IMAGE:897781	KRT8	-0.487	-1.814	1.327
IMAGE:813410	POLR2K	0.591	-0.438	1.030
IMAGE:203347	CLTC	0.002	-0.806	0.809
IMAGE:39453	LASS6	0.585	-0.588	1.173
IMAGE:789204	TLOC1	-0.222	-1.700	1.478
IMAGE:858153	NFIL3	-0.659	0.387	-1.045
IMAGE:811015	FOS	0.071	2.095	-2.024
IMAGE:345833	HNRPAB	-0.413	-1.403	0.989
IMAGE:321354	MRPL15	-0.525	-1.595	1.070
IMAGE:344589	LCP1	-0.901	-1.864	0.963
IMAGE:429222	C20orf45	-0.455	-1.228	0.774
IMAGE:781017	EGR2	0.780	2.175	-1.395
IMAGE:204214	CDC6	-1.855	-3.505	1.650
IMAGE:712840	STAT5B	-0.612	0.187	-0.799
IMAGE:510845	DTX3L	0.950	0.172	0.778
IMAGE:841195	FRAS1	-1.148	-1.871	0.723
IMAGE:757404	VBP1	-0.725	-1.588	0.863
IMAGE:428100	SFRP1	-0.496	3.006	-3.502
IMAGE:26616	RPA2	-0.780	-1.508	0.728
IMAGE:625234	KDELR3	0.154	-1.097	1.251
IMAGE:321658	PRLR	2.919	1.655	1.264
IMAGE:796323	ADD3	-0.549	0.576	-1.125
IMAGE:260336	ORC3L	-0.453	-1.396	0.943
IMAGE:248256	TTYH3	-0.118	-1.300	1.182
IMAGE:756549	GAA	1.112	0.068	1.044

IMAGE:592540	KRT5	1.772	3.683	-1.912
IMAGE:321510	SRP72	-0.199	-0.797	0.598
IMAGE:811842	SRP72	-0.404	-1.333	0.930
IMAGE:812048	PRNP	-1.137	-0.215	-0.921
IMAGE:856489	RRM1	-1.094	-1.942	0.848
IMAGE:796542	ETV5	-1.560	-0.552	-1.008
IMAGE:366558	CTNNA1	-0.442	-1.357	0.914
IMAGE:487793	MAF	-1.360	-0.529	-0.830
IMAGE:366889	KRT17	0.230	2.238	-2.008
IMAGE:785571	DNAJC1	1.137	-0.142	1.279
IMAGE:769921	UBE2C	-0.919	-2.789	1.871
IMAGE:743114	HSPBP1	-0.336	-1.328	0.992
IMAGE:810734	POLD4	0.532	-0.458	0.990
IMAGE:840821	SSR4	0.499	-0.546	1.045
IMAGE:84786	FOXA1	2.239	0.715	1.524
IMAGE:262231	CALR	-0.833	-1.704	0.871
IMAGE:784593	ARHE	-1.113	-0.295	-0.818
IMAGE:511832	PPARG	-0.415	1.232	-1.646
IMAGE:128159	TPR	-0.270	-1.496	1.225
IMAGE:866694	HSPC121	-0.019	-1.093	1.074
IMAGE:362059	LAMA3	-0.204	1.052	-1.256
IMAGE:154493	SP110	0.922	0.053	0.869
IMAGE:767769	SLC7A6	0.718	-0.115	0.833
IMAGE:85614	LEPROTL1	0.052	1.659	-1.606
IMAGE:43884	PPIF	-0.559	-1.587	1.028
IMAGE:142788	SERPINH1	-0.402	-1.284	0.882
IMAGE:179163	GRIN2C	-0.142	-0.755	0.613
IMAGE:196501	ARMET	-0.059	-0.856	0.796
IMAGE:271952	ARL7	-0.802	-0.077	-0.726
IMAGE:842973	PA2G4	-1.210	-1.821	0.611
IMAGE:85805	RBP4	-2.524	0.501	-3.025
IMAGE:43833	DGKG	-0.609	-1.315	0.706
IMAGE:203275	RFC4	-0.961	-1.626	0.665
IMAGE:377701	CSPG2	1.091	-0.323	1.413
IMAGE:80948	IGJ	-1.220	0.897	-2.117
IMAGE:772304	SLC25A5	-0.717	-1.603	0.887
IMAGE:624360	PSMB8	0.344	-0.533	0.877
IMAGE:774751	NEDD4	-0.116	-0.973	0.858
IMAGE:755578	SLC7A5	-2.226	-3.220	0.994
IMAGE:246079	ALS2CR3	0.146	-0.588	0.734
IMAGE:786504	SMC4L1	-0.656	-1.902	1.246
IMAGE:128329	FADS2	0.028	-1.546	1.574
IMAGE:840944	EGR1	0.164	2.764	-2.600
IMAGE:745138	H2-ALPHA	-0.823	-1.711	0.888
IMAGE:755750	NME2	-0.385	-1.050	0.665
IMAGE:725680	TFAP2C	0.214	1.403	-1.190
IMAGE:827132	RAC2	-0.917	-1.657	0.740
IMAGE:725454	CKS2	-0.930	-2.135	1.205
IMAGE:510466	KRT19	1.206	-0.053	1.259
IMAGE:789369	ID4	1.270	2.393	-1.124
IMAGE:853562	TRIM28	-0.697	-1.463	0.766
IMAGE:299664	C14orf147	0.229	-0.829	1.058
IMAGE:191603	TUBB	-1.164	-2.021	0.857

IMAGE:853066		-1.816	-2.420	0.604
IMAGE:73638		0.156	1.360	-1.204
IMAGE:126406		1.190	0.463	0.727
IMAGE:345525	GTF2H2	-0.560	-1.440	0.880
IMAGE:810131	KRT19	1.558	0.290	1.268
IMAGE:79726	C17orf28	2.076	1.139	0.937
IMAGE:725395	UBE2L6	0.715	-0.233	0.949
IMAGE:590500	PIGQ	0.621	-0.048	0.669
IMAGE:264938	SASH1	0.635	1.795	-1.160
IMAGE:121275	HLA-DQB2	2.575	1.485	1.090
IMAGE:785744	ZDHHC5	0.466	1.817	-1.351
IMAGE:530310	KIAA0143	1.070	0.480	0.589
IMAGE:66694	TTL	-0.782	-1.570	0.788
IMAGE:896962	ACADS	-0.050	0.637	-0.686
IMAGE:825085	ST14	1.010	0.094	0.916
IMAGE:270560	POP1	0.992	0.300	0.692
IMAGE:565235	SMS	-0.591	-1.440	0.849
IMAGE:37449	GAS2L1	-0.119	0.577	-0.696
IMAGE:809526	SEMA3F	1.333	0.265	1.068
IMAGE:376941	MAL2	1.510	0.120	1.390
IMAGE:884301	NPM1	-0.996	-1.748	0.752
IMAGE:510464	PHB	-0.562	-1.323	0.761
IMAGE:284022		-0.600	0.172	-0.771
IMAGE:195525	NAT1	1.367	-0.047	1.414
IMAGE:826350	GPS1	-0.424	-1.046	0.622
IMAGE:486676	LCP1	-0.860	-1.670	0.810
IMAGE:813751	SIAT4C	-0.633	-1.502	0.868
IMAGE:812965	MYC	-1.669	-0.395	-1.274
IMAGE:240766	TIMP1	-1.292	-2.067	0.775
IMAGE:262053	GNL3	-0.661	-1.307	0.645
IMAGE:256664	H2AFX	-0.959	-2.013	1.054
IMAGE:741977	BF	1.993	0.115	1.878
IMAGE:276547	DNMT1	-0.948	-1.615	0.667
IMAGE:230376	BF	2.203	0.063	2.140
IMAGE:213502	CD53	0.848	-0.185	1.033
IMAGE:39808	PIK3R1	0.723	1.832	-1.108
IMAGE:295986	EBP	-1.035	-1.962	0.927
IMAGE:130201	ICAM2	-0.032	0.663	-0.696
IMAGE:179211	GPR160	0.690	-0.443	1.133
IMAGE:125134	CD48	1.051	0.202	0.850
IMAGE:321708	TFDP1	-0.547	-1.140	0.594
IMAGE:127194		-0.483	-0.995	0.512
IMAGE:711857	FGFR1	-0.260	0.800	-1.060
IMAGE:210687	AGTR1	0.122	0.735	-0.613
IMAGE:796284	IRS1	0.157	1.122	-0.965
IMAGE:487777		-0.412	-1.021	0.608
IMAGE:76196	FTSJ3	-0.586	-1.307	0.721

2. The list of 216 IMAGE clones differentially expressed between lobular (L) and normal (N) samples with 90% significance.

Clone ID	Gene name	Mean expression of lobular samples	Mean expression of normal samples	Expression difference
IMAGE:244147		1.9833	-0.1553	2.1386
IMAGE:845363	NME1	-1.0901	-2.5317	1.4416
IMAGE:739126	TSTA3	0.5908	-0.4900	1.0808
IMAGE:213651	ENC1	0.3272	-1.1079	1.4352
IMAGE:248531	GMPS	-1.2613	-2.1968	0.9354
IMAGE:590774	MAPK13	-1.9266	-3.0400	1.1134
IMAGE:362409	GAD1	-1.0718	-2.0633	0.9915
IMAGE:781089	PTTG1	-2.1522	-3.7267	1.5745
IMAGE:431805	SIAT1	0.6492	-0.5350	1.1842
IMAGE:358456	SEC61G	-0.4263	-1.2050	0.7787
IMAGE:815526	MYBL2	-2.8401	-4.2600	1.4199
IMAGE:33478	FPGS	0.8279	-0.7555	1.5835
IMAGE:416386	CDH1	-0.6288	1.3172	-1.9460
IMAGE:511632	POLR3K	-0.6065	-2.1233	1.5168
IMAGE:782730	ALDH1A2	-2.8791	-1.5883	-1.2907
IMAGE:135221	S100P	-0.3663	-2.9150	2.5487
IMAGE:739109	AP2S1	-0.5886	-1.6067	1.0181
IMAGE:884655	GARS	-1.2861	-2.2950	1.0090
IMAGE:754046	DXS9879E	-0.2773	-1.7005	1.4232
IMAGE:362483	SPTBN1	-0.7729	0.7533	-1.5262
IMAGE:755145	VIL2	0.3583	-0.8605	1.2188
IMAGE:470061	SIAH2	0.3232	-0.7450	1.0682
IMAGE:838568	COX6C	0.1591	-1.8992	2.0583
IMAGE:196866	NARG1	-0.7192	-1.3350	0.6158
IMAGE:144786	BGN	2.4191	-0.0660	2.4851
IMAGE:855390	MCM6	-1.5019	-2.6483	1.1464
IMAGE:358531	JUN	1.2322	2.9081	-1.6759
IMAGE:43229	PCNA	-1.6637	-3.0517	1.3879
IMAGE:251019	CDH1	-0.2889	1.4683	-1.7573
IMAGE:770838	SLC12A7	0.8720	0.0217	0.8503
IMAGE:487988	CORO1A	-0.8678	-2.2550	1.3872
IMAGE:789182	PCNA	-1.5346	-2.8533	1.3188
IMAGE:204214	CDC6	-2.2622	-3.5050	1.2428
IMAGE:897952	PSMA5	-0.9123	-1.9298	1.0175
IMAGE:812965	MYC	-2.0913	-0.3947	-1.6966
IMAGE:669485	EGFR	-1.1611	0.5500	-1.7111
IMAGE:510845	DTX3L	0.8716	0.1717	0.6999
IMAGE:260336	ORC3L	-0.5115	-1.4076	0.8960
IMAGE:770192	LGALS9	1.4495	0.0902	1.3592
IMAGE:795936	TSN	-1.9092	-2.8000	0.8908
IMAGE:757265	MGC4399	-1.0145	-0.3433	-0.6712
IMAGE:856447	IFI30	1.1763	-1.0068	2.1831
IMAGE:321488	CDC42EP4	-0.0626	1.2317	-1.2943
IMAGE:295986	EBP	-1.2895	-1.9477	0.6582
IMAGE:245277	ITM2A	-0.9466	0.7474	-1.6940
IMAGE:232612	LMAN1	-0.5467	-1.6767	1.1299
IMAGE:743114	HSPBP1	-0.5310	-1.3283	0.7973

IMAGE:626502	ARPC1B	-0.3653	-1.6450	1.2797
IMAGE:884425	CCT5	-1.4607	-2.5792	1.1186
IMAGE:713660	GPM6B	-0.8611	0.9217	-1.7828
IMAGE:840821	SSR4	0.1230	-0.5582	0.6811
IMAGE:795378	KIAA1238	-0.0976	0.9283	-1.0260
IMAGE:265060	KIT	1.1347	2.9817	-1.8470
IMAGE:35191	SDF2	0.2068	-0.6433	0.8502
IMAGE:300044	SLC35B3	0.7965	0.2217	0.5748
IMAGE:753313	LAPTM5	0.7845	-0.9446	1.7292
IMAGE:509495	PSMA6	-0.7468	-1.7092	0.9624
IMAGE:323371	APP	-0.2992	0.6567	-0.9558
IMAGE:813698	SPRY2	-0.7991	0.1250	-0.9241
IMAGE:753104	DCT	0.7962	2.5967	-1.8005
IMAGE:1032796	LSM1	0.3650	-0.7595	1.1245
IMAGE:154493	SP110	0.8492	0.0533	0.7959
IMAGE:358643	FARSLA	-0.7816	-1.8288	1.0472
IMAGE:882510	KPNA2	-1.9797	-2.7483	0.7686
IMAGE:594382	ESR1	2.8971	1.2650	1.6321
IMAGE:81475	NOTCH3	1.8163	0.7750	1.0413
IMAGE:810017	PLAUR	0.1267	-0.9767	1.1034
IMAGE:841679	CIB1	0.2328	-0.7419	0.9747
IMAGE:840691	STAT1	0.8205	-0.2801	1.1006
IMAGE:203347	CLTC	-0.0655	-0.8009	0.7354
IMAGE:214990	GSN	1.8742	2.9700	-1.0958
IMAGE:595213	FAM49B	-0.3589	-1.2591	0.9001
IMAGE:346134	CARHSP1	-1.1066	-1.6855	0.5789
IMAGE:486676	LCP1	-0.4437	-1.6700	1.2263
IMAGE:85060	SQRDL	1.2857	0.4817	0.8040
IMAGE:252259	COL17A1	0.5423	1.9917	-1.4494
IMAGE:814378	SPINT2	1.0147	-0.2286	1.2433
IMAGE:344589	LCP1	-0.4970	-1.8558	1.3588
IMAGE:744917	NINJ1	0.8043	0.1205	0.6837
IMAGE:22918	EBNA1BP2	-1.0974	-1.9067	0.8093
IMAGE:81316	ARG99	0.1796	1.2368	-1.0573
IMAGE:712840	STAT5B	-0.8098	0.1781	-0.9879
IMAGE:810156	DTYMK	-1.3427	-1.9414	0.5986
IMAGE:725454	CKS2	-1.1547	-2.1350	0.9803
IMAGE:204257	ADAM9	-0.0809	-1.5250	1.4441
IMAGE:593183	EVI2B	0.4117	-0.7283	1.1400
IMAGE:869187	EPAS1	0.9220	2.1700	-1.2480
IMAGE:856135	SRPK1	-0.9583	-1.5167	0.5583
IMAGE:545503	STAT1	0.6314	-0.4283	1.0597
IMAGE:770614	RGS19IP1	0.1599	-0.6933	0.8532
IMAGE:545242	STAT1	0.8763	-0.7593	1.6355
IMAGE:811842	SRP72	-0.6893	-1.3333	0.6440
IMAGE:119384		1.9174	0.3500	1.5674
IMAGE:358936	C9orf75	-0.1200	-1.3550	1.2350
IMAGE:856489	RRM1	-1.2591	-1.9417	0.6825
IMAGE:789376	TXNRD1	-1.5011	-2.2204	0.7193
IMAGE:757873	CDK5R1	1.3467	3.0650	-1.7183
IMAGE:868368	TMSB4X	0.9266	-0.1683	1.0949
IMAGE:687875	CTSS	1.4967	0.2817	1.2150
IMAGE:147744	CDKN1C	-0.2303	1.6483	-1.8787

IMAGE:377461	CAV1	-1.2378	0.2111	-1.4489
IMAGE:72050	CLNS1A	-1.0793	-1.8756	0.7963
IMAGE:970613	RAD21	-0.8175	-1.4800	0.6625
IMAGE:897770	CRABP2	1.2320	0.1038	1.1283
IMAGE:418193	COL1A1	3.1611	0.6367	2.5244
IMAGE:245990	MT1F	-0.3268	0.6133	-0.9401
IMAGE:292212	ABCC5	0.5109	-0.3583	0.8693
IMAGE:121275	HLA-DQB2	2.5634	1.4850	1.0784
IMAGE:132911	PPP1CB	0.0319	0.7513	-0.7194
IMAGE:292731	HBXAP	-0.3422	-1.2210	0.8788
IMAGE:486208	TGFB3	2.1965	1.2317	0.9649
IMAGE:796984	CYBB	0.4490	-0.7388	1.1877
IMAGE:142788	SERPINH1	-0.1663	-1.2851	1.1188
IMAGE:115292		-0.3550	-1.0574	0.7024
IMAGE:795198	SLC39A1	0.2801	-0.5177	0.7977
IMAGE:399898	CSPG2	1.2939	-0.3454	1.6393
IMAGE:769686	THY1	0.6067	-0.7344	1.3410
IMAGE:855547	HLA-DRB1	3.5925	1.7067	1.8859
IMAGE:42118	P2RX4	0.6918	-0.0473	0.7391
IMAGE:39313	MRPS31	-0.0508	-0.5850	0.5342
IMAGE:782811	HMGA1	-2.3253	-3.1798	0.8545
IMAGE:141845	CRIM1	0.6708	1.6683	-0.9975
IMAGE:855395	SCP2	0.8451	0.1905	0.6546
IMAGE:713145	CD44	-0.9895	-1.9767	0.9872
IMAGE:195525	NAT1	1.8556	-0.0498	1.9054
IMAGE:858153	NFIL3	-0.5465	0.3867	-0.9331
IMAGE:79624	SP110	0.4575	-0.1717	0.6292
IMAGE:143966		-0.6091	-1.5083	0.8993
IMAGE:345833	HNRPAB	-0.6117	-1.3940	0.7824
IMAGE:246869	ZNF207	-0.5970	-1.2033	0.6063
IMAGE:199945	TGM2	-0.5156	-1.4667	0.9510
IMAGE:52930	PLEKHC1	-0.9123	0.2721	-1.1844
IMAGE:428100	SFRP1	0.3540	3.0114	-2.6574
IMAGE:110467	CAV2	0.0834	1.4897	-1.4063
IMAGE:625234	KDELR3	0.0754	-1.1099	1.1853
IMAGE:855786	WARS	-0.7080	-1.4153	0.7073
IMAGE:810600	ZNF286	-1.7475	-2.6235	0.8760
IMAGE:66560	IGLC2	-1.3685	0.3736	-1.7421
IMAGE:276547	DNMT1	-1.0675	-1.6150	0.5475
IMAGE:810057	CSDA	-1.4104	-0.3830	-1.0274
IMAGE:380394	EIF1AY	-1.0091	-1.8233	0.8143
IMAGE:789369	ID4	1.1796	2.3933	-1.2138
IMAGE:897690	TRA1	-0.8161	-1.6888	0.8727
IMAGE:85497	C2	0.5114	-0.7417	1.2531
IMAGE:796542	ETV5	-1.4876	-0.5517	-0.9360
IMAGE:841059	CAPG	1.3460	0.1331	1.2129
IMAGE:49351	PLXNA3	0.6537	-0.0600	0.7137
IMAGE:260052	HCLS1	0.2941	-0.7014	0.9956
IMAGE:40017		-1.3616	-2.1977	0.8361
IMAGE:897531	MCAM	-0.6222	-2.3005	1.6782
IMAGE:502151	SLC16A3	-1.7222	-2.7633	1.0411
IMAGE:842765	IQWD1	0.7301	0.0467	0.6835
IMAGE:345525	GTF2H2	-0.3223	-1.4400	1.1177

IMAGE:306841		-2.4817	-3.3486	0.8670
IMAGE:840708		0.5902	1.7000	-1.1098
IMAGE:856167	TARS	-1.3331	-2.1072	0.7742
IMAGE:814701	MAD2L1	-2.1150	-2.6567	0.5417
IMAGE:139009	FN1	-0.5762	-2.8459	2.2697
IMAGE:1680549	MXI1	0.4445	2.0950	-1.6505
IMAGE:743804	SEC23B	-0.5654	-1.2391	0.6737
IMAGE:825085	ST14	1.1019	0.1093	0.9926
IMAGE:323238	CXCL1	-0.5071	1.2104	-1.7176
IMAGE:810603	FLJ14525	0.8053	1.3383	-0.5330
IMAGE:472185	ADAMTS1	-0.4632	1.1181	-1.5812
IMAGE:270560	POP1	0.9552	0.3000	0.6552
IMAGE:341317	ZNF516	-0.0098	0.8117	-0.8215
IMAGE:51448	ATF3	0.5866	2.3026	-1.7160
IMAGE:813410	POLR2K	0.5133	-0.4383	0.9516
IMAGE:825583	RALY	-0.4157	-1.1244	0.7087
IMAGE:782513	G1P3	1.2782	0.1152	1.1629
IMAGE:752625	SLC7A2	-0.8918	-0.2700	-0.6218
IMAGE:291880	MFAP2	0.5390	-0.7167	1.2556
IMAGE:795544	WASPIP	-0.1745	-1.1817	1.0071
IMAGE:739901	CYP51A1	0.6666	-0.3246	0.9912
IMAGE:827132	RAC2	-0.7621	-1.6567	0.8946
IMAGE:811024	BST2	0.6277	-0.2263	0.8541
IMAGE:308484	UBE2D3	-0.7160	-1.4072	0.6912
IMAGE:741977	BF	1.6765	0.1150	1.5615
IMAGE:811582	GOLPH2	1.2748	0.1521	1.1227
IMAGE:809466	TOMM40	-1.0382	-1.6200	0.5818
IMAGE:853066		-1.9981	-2.4200	0.4219
IMAGE:432564	SF3B4	-0.5026	-1.2688	0.7662
IMAGE:549933	IL8	1.2889	0.5275	0.7614
IMAGE:950482	SNRPB	-1.3454	-1.8789	0.5336
IMAGE:767817	POLR2F	-0.7591	-1.3267	0.5675
IMAGE:123614	C20orf55	1.0809	0.4850	0.5959
IMAGE:122091	GALNTL4	-0.9454	-1.7050	0.7596
IMAGE:128159	TPR	-0.4532	-1.4947	1.0414
IMAGE:50503	ITGB2	1.3396	-0.0531	1.3927
IMAGE:727251	CD9	0.8807	-0.0573	0.9380
IMAGE:814119	DHX8	-0.3641	-0.8033	0.4393
IMAGE:377701	CSPG2	1.2379	-0.3251	1.5630
IMAGE:284022		-0.4531	0.1717	-0.6248
IMAGE:786504	SMC4L1	-1.1448	-1.9017	0.7569
IMAGE:842825	GSPT1	-1.2032	-1.6493	0.4461
IMAGE:811015	FOS	0.5127	2.1343	-1.6217
IMAGE:739183		0.8327	-0.4667	1.2993
IMAGE:781017	EGR2	1.1282	2.1750	-1.0468
IMAGE:725680	TFAP2C	-0.2175	1.4450	-1.6625
IMAGE:244974	FLJ22875	-0.7516	-1.4833	0.7318
IMAGE:364448	IKBKE	0.2307	-0.6417	0.8724
IMAGE:70152		0.3424	1.3218	-0.9794
IMAGE:198093	EIF2S2	-1.2186	-1.7300	0.5114
IMAGE:47908	LOC91137	-0.7972	-1.3717	0.5744
IMAGE:154720	ARD1	-1.1836	-2.2762	1.0926
IMAGE:248256	TTYH3	-0.2266	-1.3000	1.0734

IMAGE:179276	FASN	0.6783	2.1319	-1.4535
IMAGE:39808	PIK3R1	0.9588	1.8289	-0.8701
IMAGE:121948	IFRD1	-1.2765	-0.7600	-0.5165
IMAGE:416676	PELI1	0.3401	1.2731	-0.9330
IMAGE:144849	COTL1	-0.7776	-1.7567	0.9790
IMAGE:130100	RRAS2	-1.0992	-0.4583	-0.6408
IMAGE:525926	SDC1	0.4188	-0.6006	1.0194
IMAGE:825478		-0.0123	-0.5917	0.5794
IMAGE:262231	CALR	-0.8009	-1.7146	0.9137
IMAGE:76196	FTSJ3	-0.7938	-1.3124	0.5186

3. List of image clones which are differentially expressed between ductal and lobular samples with 80% significance.

The IMAGE clones differentially expressed between IDC and ILC with 90% significance are marked with (a)

The IMAGE clones marked with (b) are the genes common with the findings of Zhao *et al.* DL list [1].

The IMAGE clones marked with (c) are the novel genes that were obtained through meta-analysis.

	Clone ID	Gene name	Mean expression of ductal samples	Mean expression of lobular samples	Expression difference
c	IMAGE:183476	ACDC	0.7087	2.0751	-1.3663
a	IMAGE:753400	ACTL6A	-0.5325	-1.2567	0.7242
c	IMAGE:307231	ADH1B	0.9144	2.3536	-1.4392
a,b	IMAGE:196992	AKR1C1	-2.1656	-0.7145	-1.4511
b	IMAGE:855624	ALDH1A1	-0.8152	0.1380	-0.9532
b	IMAGE:208718	ANXA1	-0.6364	0.2303	-0.8667
a,b	IMAGE:484535	AOC3	0.5679	1.9453	-1.3774
c	IMAGE:39677	ARL10C	0.4365	-0.4556	0.8921
c	IMAGE:796694	BIRC5	-1.4277	-2.1303	0.7026
c	IMAGE:48167	C14orf130	-0.2495	-0.8204	0.5710
a	IMAGE:489208	C16orf34	0.1243	-0.8280	0.9523
c	IMAGE:811149	C9orf3	0.1320	0.8217	-0.6897
c	IMAGE:358936	C9orf75	0.5628	-0.1194	0.6821
	IMAGE:377461	CAV1	-2.2556	-1.2375	-1.0181
c	IMAGE:321488	CDC42EP4	0.6572	-0.0639	0.7211
a,b	IMAGE:251019	CDH1	1.5701	-0.3345	1.9046
a,b	IMAGE:416386	CDH1	1.5979	-0.5350	2.1329
a	IMAGE:839736	CRYAB	1.1036	2.8977	-1.7942
c	IMAGE:295843	CYP27A1	0.4584	1.1922	-0.7338
c	IMAGE:43833	DGKG	-0.5832	-1.0069	0.4237
a,c	IMAGE:271006	DLAT	-0.8072	-1.5086	0.7014
c	IMAGE:785571	DNAJC1	1.1635	0.1846	0.9790
c	IMAGE:188036	DST	0.9694	2.2770	-1.3076
c	IMAGE:809828	E2F5	0.1876	-0.6673	0.8549
c	IMAGE:840944	EGR1	0.1566	1.4439	-1.2873
b	IMAGE:128329	FADS2	-0.0071	-1.4385	1.4314
	IMAGE:813266	FHL1	-1.3999	-0.2262	-1.1737
a	IMAGE:511428	FXYD3	0.9963	-0.1387	1.1350
	IMAGE:742132	G1P2	1.9374	0.6866	1.2508
c	IMAGE:810326	GRP58	1.6154	0.9103	0.7051
	IMAGE:214990	GSN	1.0504	1.8416	-0.7912
a	IMAGE:842825	GSPT1	-0.5710	-1.2325	0.6615
c	IMAGE:342721	ITGB2	0.3660	1.5011	-1.1351

	IMAGE:783836	JDP2	0.7146	1.3686	-0.6540
c	IMAGE:788256	KIF23	-1.3987	-2.2058	0.8071
	IMAGE:85614	LEPROTL1	0.0677	0.8326	-0.7648
a,b	IMAGE:868169	LPL	0.4099	2.1989	-1.7890
c	IMAGE:126239	MRPL16	-0.9056	-1.5399	0.6343
c	IMAGE:782635	NDUFAB1	-0.3347	-1.0348	0.7001
c	IMAGE:154790	NGFR	0.3096	1.1625	-0.8530
a	IMAGE:179232	PP	-0.3649	-1.2248	0.8599
a,c	IMAGE:361974	PTN	0.1618	1.8125	-1.6507
	IMAGE:731118	PTPNS1	0.1317	0.7561	-0.6243
c	IMAGE:809473	PTRF	-0.1612	0.6250	-0.7862
c	IMAGE:230261	RALA	-0.4301	-0.9523	0.5222
a	IMAGE:85805	RBP4	-2.5315	-1.1455	-1.3860
c	IMAGE:624627	RRM2	-2.2992	-3.0392	0.7400
	IMAGE:35620	SEN54L	-0.1481	-0.8881	0.7400
c	IMAGE:770794	SHB	-0.0657	-0.4778	0.4121
c	IMAGE:115143	SLC4A1	-0.3052	-0.8558	0.5506
c	IMAGE:752625	SLC7A2	0.0065	-0.8888	0.8953
a,b	IMAGE:796406	SORBS1	1.0031	2.4358	-1.4326
c	IMAGE:813698	SPRY2	-1.4394	-0.8128	-0.6266
c	IMAGE:129865	STK6	-2.0430	-2.8399	0.7969
c	IMAGE:141495	SULT2B1	0.5558	-0.3302	0.8860
	IMAGE:346696	TEAD4	-0.7110	-1.3124	0.6014
c	IMAGE:884822	TEBP	-0.5391	-1.0901	0.5510
c	IMAGE:137387	TFAP2A	1.2146	0.3112	0.9034
c	IMAGE:321708	TFDP1	-0.4770	-1.1171	0.6400
c	IMAGE:841149	TGFBR2	0.0774	0.8484	-0.7710
c	IMAGE:814306	TPD52	1.4283	0.2399	1.1883
a	IMAGE:898312	TRAF4	0.1613	-0.6781	0.8394
	IMAGE:731023	WDR5	-0.9780	-1.6413	0.6633
a	IMAGE:66714		0.9464	1.7427	-0.7963
a	IMAGE:308478		-0.4999	-1.3984	0.8985
	IMAGE:430186		-0.8329	-2.0651	1.2322

4. List Of Meta-Analysis Specific Genes Differentially Expressed In DN, LN And DL

DN n=3		LN n=3		DL n=3,4,5	
Clone ID	**Gene name**	**Clone ID**	**Gene name**	**Clone ID**	**Gene name**
IMAGE:159608	APOD	IMAGE:120749	VANGL1	IMAGE:188036	DST
IMAGE:179163	GRIN2C	IMAGE:1476320	NCF2	IMAGE:271006	DLAT
IMAGE:191826	MSCP	IMAGE:195525	NAT1	IMAGE:308478	
IMAGE:200136	TOPBP1	IMAGE:204257	ADAM9	IMAGE:361974	PTN
IMAGE:213502	CD53	IMAGE:209137	GABRE	IMAGE:484535	AOC3
IMAGE:243202	CTSE	IMAGE:244355	IL2RG	IMAGE:66714	
IMAGE:271989	RALGPS2	IMAGE:272018	TIM14		
IMAGE:283398	TM4SF10	IMAGE:358456	SEC61G		
IMAGE:287749	CDC7	IMAGE:365641	PRIM1		
IMAGE:343987	DPP4	IMAGE:429352	DJ971N18.2		
IMAGE:363103	HMGB2	IMAGE:48631	KCNAB2		
IMAGE:365641	PRIM1	IMAGE:503617	CXCL9		
IMAGE:431655	CD37	IMAGE:513077	CD47		
IMAGE:46896	ADAM19	IMAGE:52930	PLEKHC1		
IMAGE:531036		IMAGE:564803	FOXM1		
IMAGE:69935	ADRA2A	IMAGE:626531	SYNCRIP		
IMAGE:724615	CHC1	IMAGE:726035	JUN		
IMAGE:725395	UBE2L6	IMAGE:753104	DCT		
IMAGE:741977	BF	IMAGE:757873	CDK5R1		
IMAGE:755578	SLC7A5	IMAGE:951142	FEN1		
IMAGE:755750	NME2	IMAGE:970613	RAD21		
IMAGE:789369	ID4				
IMAGE:796323	ADD3				
IMAGE:809828	E2F5				
IMAGE:811096	ITGB4				
IMAGE:838716	COX11				
IMAGE:853066					
IMAGE:896962	ACADS				
IMAGE:951142	FEN1				

Appendix C: Validation of meta-analysis gene lists by three independent microarray datasets.

1. Validation of DN meta-gene list by three independent microarray datasets.

DN	Expression differences between tumor and normal samples				Significance of subgroup prediction power of the genes and expression profiles					
					NBL* vs BL*		B* vs L*		ER+* vs ER-*	
Gene Name	meta gene-set	Karnoub	Turashvili	Richardson	p value	expression	p value	expression	p value	expression
ADAMTS1	-1.59	-1.25	-2.15	-1.90	**0.0300**	1.03	**0.0003**	0.86	**0.0020**	0.65
ATF3	-1.81	-1.18	-2.98	-2.06	**0.0212**	0.87	**0.0006**	0.71	**0.0161**	0.47
BGN	2.05	2.00	1.92	1.28	0.1797	-0.62	0.9134	0.02	0.2721	0.05
CAV1	-2.53	-0.64	-2.19	-2.98	0.5470	-0.24	0.1430	-0.27	**0.0237**	-0.39
EGFR	-1.23	-1.75	-2.18	-0.22	**0.0011**	2.01	**4.60E-06**	0.83	**5.00E-07**	0.47
FN1	2.79	1.73	2.98	2.01	**3.28E-05**	-1.58	**0.0219**	-0.31	0.0743	-0.21
GSN	-1.95	-2.15	-1.09	-2.62	0.8275	-0.05	0.0679	-0.32	**0.0148**	-0.37
ID4	-1.12	-0.79	-2.78	-2.49	**7.00E-07**	3.10	**0.0003**	0.47	**0.0018**	1.03
IGFBP6	-1.13	-1.91	-1.24	-4.25	**0.0338**	-0.70	**0.0011**	-0.57	**0.0000**	-0.56
JUN	-1.83	0.30	-1.54	-1.44	0.1732	-0.33	0.5305	0.14	0.5550	0.11
MFAP2	1.65	1.92	2.58	2.16	0.0927	0.71	**0.0120**	0.51	0.1912	0.25
POLR2K	1.03	0.58	1.06	1.09	0.8041	0.06	0.0440	0.33	0.5246	0.09
PRNP	-0.92	-1.15	0.74	-1.41	**0.0005**	1.10	**1.16E-05**	0.78	**7.77E-05**	0.63
SFRP1	-3.50	-2.80	-3.70	-3.68	**1.00E-07**	4.53	**4.10E-06**	1.79	**0.0002**	1.33
SPTBN1	-1.67	-0.71	-1.77	-1.30	**0.0286**	0.57	**0.0016**	0.16	**0.0095**	0.12
STAT1	1.53	1.60	2.38	1.92	0.6182	0.23	0.4435	0.26	0.4930	-0.13
THY1	1.13	1.25	1.50	0.87	0.0725	-0.58	0.1378	-0.30	0.1640	-0.25
TPR	1.23	-1.25	0.98	1.41	0.2837	-0.35	0.0941	0.29	0.5834	0.07

2. Validation of LN meta-gene list by three independent microarray datasets.

LN	Expression differences between tumor and normal samples				Significance of subgroup prediction power of the genes and expression profiles					
Gene Name	meta gene-set	Karnoub	Turashvili	Richardson	NBL* vs BL*		B* vs L*		ER+* vs ER-*	
					p value	expression	p value	expression	p value	expression
BGN	2.49	2.00	2.60	1.28	0.1786	-0.54	0.6999	0.02	0.2721	0.05
CDKN1C	-1.88	-0.83	-1.28	-1.82	0.1369	0.52	0.7346	0.05	0.4072	-0.17
COL1A1	2.52	1.29	2.66	0.64	**5.24E-05**	-1.99	**0.0457**	0.03	**0.0457**	-0.33
EGFR	-1.71	-1.75	-1.98	-0.22	**0.0011**	2.01	**4.60E-06**	0.01	**5.00E-07**	0.06
ESR1	1.63	0.75	-1.37	-3.38	**4.60E-06**	-2.25	**1.00E-07**	0.03	**1.00E-07**	-0.09
ETV5	-0.94	-0.43	-0.81	-1.50	**0.0371**	0.71	**0.0128**	0.04	0.0772	0.23
FAM49B	0.90	1.10	1.71	1.26	0.5968	-0.07	**0.0088**	0.02	**0.0197**	0.27
FN1	2.27	1.73	2.96	2.01	**3.28E-05**	-1.58	**0.0219**	-0.18	0.0961	-0.20
GPM6B	-1.78	-1.55	-2.03	-1.90	**7.00E-07**	2.35	**4.00E-07**	0.63	**4.22E-05**	0.51
ID4	-1.21	-0.79	-1.85	-2.49	**7.00E-07**	3.10	**0.0003**	1.47	**0.0018**	0.36
JUN	-1.68	0.30	-1.52	-1.44	0.6148	-0.17	0.3962	0.15	0.1765	0.21
MT1F	-0.94	-1.11	-1.78	-0.80	0.4972	0.23	**0.0213**	0.37	0.0754	0.28
PPP1CB	-0.72	-0.46	-1.12	-0.04	**0.0001**	0.82	**2.00E-07**	0.58	**9.30E-06**	0.76
SFRP1	-2.66	-2.80	-2.51	-3.68	**1.00E-07**	4.53	**4.10E-06**	1.79	**0.0002**	1.33
SLC7A2	-0.62	-1.37	-2.37	-3.30	**0.0002**	-3.18	**0.0351**	-0.11	**0.0208**	-0.11
STAT1	1.27	1.60	2.19	1.92	0.6182	0.23	0.4435	0.26	0.4930	-0.13
STAT5B	-0.99	-0.27	-1.38	-1.10	**0.0213**	0.53	**0.0332**	-0.16	**0.0013**	-0.20
THY1	1.34	1.25	1.51	0.87	0.1938	-0.33	0.1378	-0.30	0.0665	-0.29

Appendix D: Comparison of DL list with other published data sets

Comparison of the expression profiles of the genes in DL list with other published data sets.

Gene name	meta-gene set (D/L)	Sorlie et al, 2001 (D/L)	Zhao et al, 2004 (D/L)	Turashvili et al, 2007 (D/L)	Bertucci et al, 2008 (D/L)
ACDC	2.08				
ACTL6A	0.72		0.68	0.30	
ADH1B	-1.44			-0.38	
AKR1C1	-1.45		-1.08	-0.58	
ALDH1A1	-0.95		-1.16	-1.24	Down
ANXA1	-0.87		-1.20	-0.65	
AOC3	-1.38			-1.26	
ARL10C	0.89				
BIRC5	0.70			0.52	
C14orf130	0.57			1.19	
C16orf34	0.95		0.61		
C9orf3	-0.69			0.07	
C9orf75	0.68			0.46	
CAV1	-1.02		-1.55	0.45	Down
CDC42EP4	0.72			0.27	
CDH1	1.90	1.77	2.21	1.66	Up
CDH1	2.13				
CRYAB	-1.79	-1.66		-1.75	
CYP27A1	-0.73			0.36	
DGKG	0.42			-0.49	
DLAT	0.70			0.52	
DNAJC1	0.98			0.69	
DST	-1.31			-0.01	
E2F5	0.85			0.40	
EGR1	-1.29			-0.71	
FADS2	1.43		1.58	0.42	
FHL1	-1.17	-1.32		-0.08	
FXYD3	1.13	0.57	1.23	-0.15	
G1P2	1.25				
GRP58	0.71				
GSN	-0.79	-1.13		-0.11	
GSPT1	0.66		0.55	-0.31	
ITGB2	-1.14			0.30	
JDP2	-0.65		-0.73	-2.39	
KIF23	0.81			0.04	
LEPROTL1	-0.76		-0.87	0.72	
LPL	-1.79	-1.55	-2.29	-0.53	
MRPL16	0.63			-0.17	
NDUFAB1	0.70			-0.15	

NGFR	-0.85			0.75	down
PP	0.86		1.05		
PTN	-1.65			0.15	
PTPNS1	-0.62	-0.54			
PTRF	-0.79			-0.12	
RALA	0.52			0.06	
RBP4	-1.39	-1.56	-1.25	-1.91	Down
RRM2	0.74			2.14	
SEN54L	0.74	0.76			
SHB	0.41			0.25	
SLC4A1	0.55			-0.23	
SLC7A2	0.90			1.17	
SORBS1	-1.43		-1.30	-1.13	
SPRY2	-0.63			0.23	
STK6	0.80				
SULT2B1	0.89			2.28	
TEAD4	0.60	0.68		-1.09	
TEBP	0.55				
TFAP2A	0.90			0.41	Up
TFDP1	0.64			0.30	
TGFBR2	-0.77			-0.02	Down
TPD52	1.19			0.17	
TRAF4	0.84		0.87	0.31	
WDR5	0.66		0.88	-0.30	

Pearson correlation of meta-gene set and Turashvilli=0.53, p=0.000

BMC Cancer

BioMed Central

Research article

A resampling-based meta-analysis for detection of differential gene expression in breast cancer

Bala Gur-Dedeoglu†[1], Ozlen Konu†[1], Serkan Kir[1], Ahmet Rasit Ozturk[1], Betul Bozkurt[2], Gulusan Ergul[3] and Isik G Yulug*[1]

Address: [1]Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, TR-06800, Ankara, Turkey, [2]Department of General Surgery, Ankara Numune Research and Teaching Hospital, TR-06100, Ankara, Turkey and [3]Department of Pathology, Ankara Numune Research and Teaching Hospital, TR-06100, Ankara, Turkey

Email: Bala Gur-Dedeoglu - bala@fen.bilkent.edu.tr; Ozlen Konu - konu@fen.bilkent.edu.tr; Serkan Kir - serkankir@gmail.com; Ahmet Rasit Ozturk - ahmetrasit@bilkent.edu.tr; Betul Bozkurt - b2bozkurt@yahoo.com; Gulusan Ergul - gulusanergul@yahoo.com; Isik G Yulug* - yulug@fen.bilkent.edu.tr

* Corresponding author   †Equal contributors

Published: 30 December 2008

BMC Cancer 2008, **8**:396   doi:10.1186/1471-2407-8-396

Received: 21 May 2008
Accepted: 30 December 2008

This article is available from: http://www.biomedcentral.com/1471-2407/8/396

Abstract

Background: Accuracy in the diagnosis of breast cancer and classification of cancer subtypes has improved over the years with the development of well-established immunohistopathological criteria. More recently, diagnostic gene-sets at the mRNA expression level have been tested as better predictors of disease state. However, breast cancer is heterogeneous in nature; thus extraction of differentially expressed gene-sets that stably distinguish normal tissue from various pathologies poses challenges. Meta-analysis of high-throughput expression data using a collection of statistical methodologies leads to the identification of robust tumor gene expression signatures.

Methods: A resampling-based meta-analysis strategy, which involves the use of resampling and application of distribution statistics in combination to assess the degree of significance in differential expression between sample classes, was developed. Two independent microarray datasets that contain normal breast, invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC) samples were used for the meta-analysis. Expression of the genes, selected from the gene list for classification of normal breast samples and breast tumors encompassing both the ILC and IDC subtypes were tested on 10 independent primary IDC samples and matched non-tumor controls by real-time qRT-PCR. Other existing breast cancer microarray datasets were used in support of the resampling-based meta-analysis.

Results: The two independent microarray studies were found to be comparable, although differing in their experimental methodologies (Pearson correlation coefficient, R = 0.9389 and R = 0.8465 for ductal and lobular samples, respectively). The resampling-based meta-analysis has led to the identification of a highly stable set of genes for classification of normal breast samples and breast tumors encompassing both the ILC and IDC subtypes. The expression results of the selected genes obtained through real-time qRT-PCR supported the meta-analysis results.

Conclusion: The proposed meta-analysis approach has the ability to detect a set of differentially expressed genes with the least amount of within-group variability, thus providing highly stable gene lists for class prediction. Increased statistical power and stringent filtering criteria used in the present study also make identification of novel candidate genes possible and may provide further insight to improve our understanding of breast cancer development.

Background

Microarray studies aiming to identify differentially expressed as well as co-regulated gene sets and signaling pathways involved in different cellular states have greatly improved our understanding of breast cancer at the molecular level. The power of expression profiling using cDNA or DNA microarrays for distinguishing subgroups of breast cancers has been demonstrated by several groups [1-4].

The identification of an intrinsic gene-set exhibiting high variability among different tumor clusters has been informative in describing different subtypes of breast cancer samples. However, only a few papers have been published on gene expression profiles of normal cell populations in breast tissue [5-9]. Therefore, it is of paramount importance for the research community in the field of tumor biology to have access to gene lists that exhibit low variability in expression among tumors and yet are distinguishable from a normal tissue profile.

Meta-analysis of microarray datasets has the potential to lead to more comprehensive measures of the existing differential gene expression data and can therefore provide gene sets with a high diagnostic value. Meta-analysis of independent microarray datasets generated with the common objective of identifying differentially expressed genes in a certain type of cancer has also been performed for breast cancer. In a very recent meta-analysis study, Smith *et al*. identified differentially expressed genes between ER+ and ER- breast tumors by gathering 9 independent breast cancer microarray studies [10]. Another study used the power of meta-analysis to find out the relation of expression patterns of gene and chromosomal positions. More than 1200 breast tumors were collected from eight independent breast studies and candidate metastasis suppressor and promoting genes were found from a given set of chromosomal regions [11]. Similarly, Hu *et al*. were able to identify a new intrinsic gene-set for breast cancer subtype prediction by combining multiple microarray datasets to assess prognosis [12].

Several different meta-analysis approaches exist in the literature. In some, each individual study contributes rather independently to the meta-analysis [13-15] whereas in others the values are treated as members of a single study thus requiring a generalized normalization step [16,17]. Direct comparison of gene expression values from multiple studies may be relatively more problematic than comparing the effect size obtained from individual studies. Yet, analysis of combined raw data is beneficial when sample sizes of individual studies are small. Another important concern in meta-analysis is the determination of the minimum number of samples required to obtain statistically reliable results [18]. One possible solution to this problem is resampling; for example, one can use a *delete-d-jacknife* procedure in which a subset of data is excluded to find out the frequency of selecting a particular gene as differentially expressed [18]. The number of replicates required for producing stable differentially expressed gene lists could also be determined based on a related method known as *leave-one-out* resampling [19].

Existing meta-analytic approaches applied to different types of cancer show the power of a combined study for identifying novel genes not present in the existing literature (e.g., liver cancers) [20,21]. Invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) make up to 95% of all breast tumors (IDC: 50–80% and ILC: 10–15%) [2,22-24]. Although recent studies suggest differences between the expression profiles of IDC and ILC, the clinical progress, therapeutic response, and molecular signature, there are also many similarities between IDC and ILC tumors distinguishing them from normal breast tissue [2,5,23,25]. However, meta-analysis of gene expression differences between normal breast tissue and such a generalized set of breast tumors has not been reported to date.

In the present study, we primarily aimed to develop a novel methodology for the meta-analysis of independent microarray datasets. Using this methodology, we provide gene lists that (a) are discriminative of breast cancer types (IDC, ILC) and normal breast cell populations, (b) may yield breast tumor markers that are invariably expressed across independent experiments, and (c) provide a set of consistently differentially expressed gene candidates with potential discriminative ability for tumor subtypes. Using a method similar to *delete-d-jacknife*, a series of *d* sample size values have been tested to assess the extent of variability across the tumor samples and the stability of differential expression. Comparison of probability value distributions obtained for the test and randomized samples has led to determination of the degree of differential expression between groups tested. Accordingly, we report that the Sorlie *et al*. [1] and Zhao *et al*. [2] datasets were highly comparable. Our resampling-based meta-analysis led to the identification of genes not differentially expressed when analyzed independently. Predictive ability of the meta-gene set was independently supported in three other breast cancer microarray studies with information on breast normal and tumor tissues [5,7,8] using BRB-TOOLS [26]. A subset of the meta-gene-list was also used as a classifier to accurately predict different molecular subtypes, such as luminal/basal and ER+/ER- based on microarray datasets in which patient subtype classification was available [7,8]. Moreover, selected candidates from stable gene sets obtained from the meta-analysis were validated by real-time qRT PCR. Use of resampling-based meta-analysis combined with class prediction via available microarray datasets pointed to the existence of a tumor-specific differentially expressed gene-set with predictive potential for tumor subtype classification.

Methods

Data retrieval for resampling-based meta-analysis

Two independent microarray gene expression data sets, Sorlie *et al.* [1] and Zhao *et al.* [2], were downloaded from the Stanford Microarray Database (SMD); http://genome-www5.stanford.edu/[27]. Gene filtering options of SMD were used for log transformation and median centering the data arraywise. Expression values that were missing in more than 20% of the data were excluded from the analysis. Details of tumor specimen histology, available on SMD, were used to restructure the experiments according to breast tumor subtypes as invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC) and normal samples. Datasets were combined with respect to probe IDs using a set of customized perl routines (source codes are available upon request). These two data sets combined resulted in an initial list of 4769 IMAGE clones (3465 unique genes) common in both datasets (see Additional file 1; Zhao dataset and Additional file 2; Sorlie dataset). A total of 139 IDC (38 samples Zhao, 101 samples Sorlie datasets), 29 (21 samples Zhao, 8 samples Sorlie datasets) ILC and 7 (3 samples Zhao, 4 samples Sorlie datasets) normal samples were available for further analysis.

Data Filtering

Data were filtered separately for ductal and lobular samples. IMAGE clones with more than 50% missing data in either of the Sorlie or Zhao datasets were excluded from the common clone set. Data filtering was further improved by performing two-tailed Student's t-tests with equal variance (Matlab®) between the Sorlie and Zhao datasets for the IDC and ILC samples separately. Those clones with probability values less than 0.05 (after Bonferroni correction) were excluded from further analysis. This two-step data filtering resulted in a common set of 1726 IMAGE clones for the analysis of ductal and normal samples, and 2029 IMAGE clones for the analysis of lobular and normal samples. Upon taking the intersection of the ductal-normal and lobular-normal clone sets, 1522 IMAGE clones were available for the ductal-lobular analysis. The resulting clone subsets were further filtered by removing IMAGE clones with more than 40% missing data for the two groups in comparison (e.g., ductal and normal) in the combined data before application of the resampling steps. In addition, if an IMAGE clone had a sample size (of normal samples) less than the resampling sample size, data on this IMAGE clone was also removed.

Resampling and statistical analysis

We have used a resampling method for meta-analysis of microarray data in which the significance of the difference between group medians (e.g. ductal vs. lobular) could be tested upon a series of resampling schemes from the original and multiple randomly shuffled datasets (Figure 1; code written in Matlab® using Statistics Toolbox is available upon request). Accordingly, a preset number of samples was selected from each group (i.e., IDC, ILC, normal) of the original dataset, referred herein as the *test*. The p-value was calculated indicating the significance of the difference between the group medians based on the Wilcoxon Rank Sum Test. This test was repeated for a series of *i* number of iterations; at the end of each iteration scheme, a set of *p-values (pt)* per IMAGE clone was obtained. The above procedure was also applied to each of the shuffled datasets yielding *pr1* and *pr2*. P-value distributions were then tested in a pair-wise fashion (i.e., *pt* vs. *pr1*; and *pr1* vs. *pr2*) using the two-sample Kolmogorov-Smirnov test for each clone in the dataset (Figure 1). The resulting p-values were named as *kst* and *ksr*, respectively. To obtain an estimate of the false discovery rate (FDR), *ksr* values were sorted in the ascending order and the k^{th} value from the top (lowest p-value) was determined as $\mathrm{FDR}_{\mathrm{observed}}$, where *k* equals the expected value of FDR (e.g., 0.01) multiplied by the number of IMAGE clones tested. $\mathrm{FDR}_{\mathrm{observed}}$ was set as the threshold according to which IMAGE clones were assigned as significant or not. If *kst* of a particular gene had a value that was smaller than the $\mathrm{FDR}_{\mathrm{observed}}$, the gene was accepted to be significant.

Application to the breast cancer datasets

The above tasks were performed for a particular sample size *n* (e.g., 3), repetitively for *i* number of times where *i* = 10, 20, 30, ..., 100 and 150. For each particular *i*, three parameters were recorded, namely, *kst* values, the mean expression value of each of the two groups compared, and the significance of the differential expression based on *kst* and *ksr*. These above steps were then repeated with different sample sizes: For ductal vs. lobular comparison, *n* was set to be 3, 4, 5, 6, 10, 15 and 20. On the other hand, since the total number of normal samples was 7, the highest sampling value could be set to 6 for ductal vs. normal and lobular vs. normal comparisons, and *n* equaled 3, 4, 5 and 6. These sample size-iteration combinations led to 77 runs for ductal vs. lobular analysis, and 44 runs for ductal vs. normal and lobular vs. normal analyses. At the end, a final differentially expressed gene set was determined for each of the three comparisons (i.e., ductal vs. lobular, DL; ductal vs. normal, DN; lobular vs. normal, LN) by gathering the IMAGE clones that were assigned as significant in 90% or more of these 44 or 77 runs. The mean values of each of the two groups in comparison obtained at n = 20 (or 6, in the case of normal vs. tumor comparisons) and *i* = 150 were used as an estimate of the measure of expression.

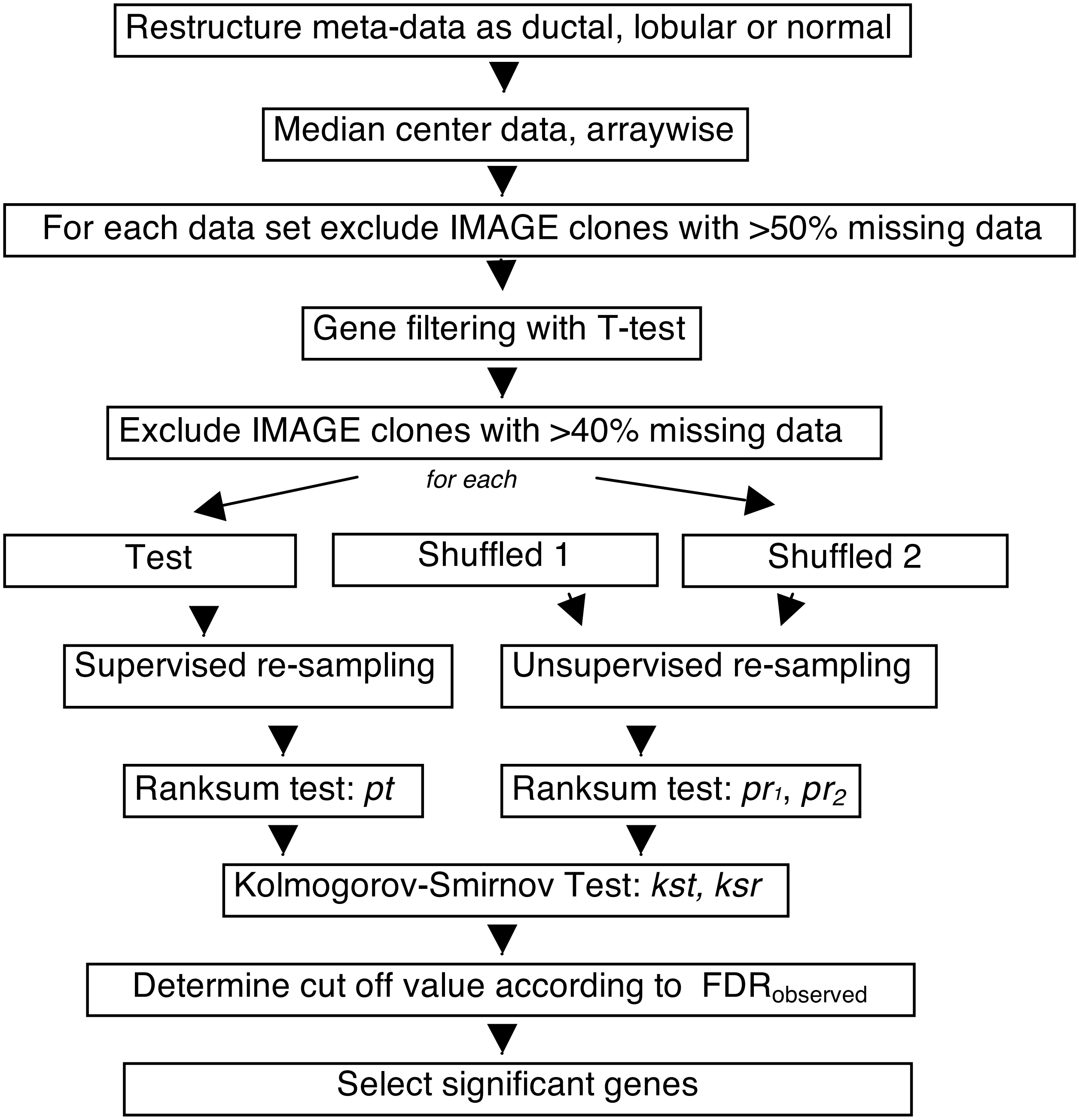

Restructure meta-data as ductal, lobular or normal
Median center data, arraywise
For each data set exclude IMAGE clones with >50% missing data
Gene filtering with T-test
Exclude IMAGE clones with >40% missing data
for each
Test
Shuffled 1
Shuffled 2
Supervised re-sampling
Unsupervised re-sampling
Ranksum test: pt
Ranksum test: pr1, pr2
Kolmogorov-Smirnov Test: kst, ksr
Determine cut off value according to FDRobserved
Select significant genes

Figure 1
General meta-analysis scheme. Workflow is represented by boxes and arrows.

Data retrieval and analysis for validation studies

The ".cel files" of the three publicly available independent microarray gene expression data sets, GDS2635 [5], GDS2250 [7] and GDS1329 [4], were downloaded from GEO [28] and processed by the BRB-ARRAYTOOLS [26]. All three datasets were obtained using the Affymetrix HGU133A or HGU133 Plus 2.0 platform; thus they were highly comparable. In GDS2635 the aim was to identify gene expression profiles of microdissected ductal and lobular carcinomas in relation to their normal ductal and lobular cells (n = 10). The authors identified multiple genes differentially expressed in comparisons between ductal and lobular tumor and normal cells [5]. In the GDS2250 study, a gene expression array-based analysis of three breast tumor subtypes, i.e., sporadic basal-like cancer (BLC), BRCA-associated breast cancer, and non-BLC, was performed. They used 47 human breast tumor cases to provide insight into the molecular pathogenesis of BLC and BRCA1-associated breast cancer and the contribution of X chromosome abnormalities to the pathogenesis of BLC [7]. In GDS1329, Farmer *et al*. performed an analysis of tumors from 49 breast cancer patients that were successfully classified into luminal and basal classes, and a novel molecular apocrine class. Apocrine tumors were estrogen receptor negative ER(-) and androgen receptor positive AR(+), while luminal tumors were ER(+) and AR(+), and basal tumors were ER(-) and AR(-). Details of the breast specimens (normal-tumor, non-basal like-basal like, basal-luminal and ER (+)/ER (-)) available from GEO database were used in the supervised class prediction with a binary tree algorithm [26]. The common genes between the re-analyzed microarray studies and the meta-gene-lists were combined with respect to gene symbols (perl source codes are available upon request).

Clinical Samples

Primary tumor samples and matched non-tumor breast tissues were obtained from patients (n = 10) during surgery and immediately snap-frozen in liquid nitrogen and stored at -80°C until RNA extraction. The frozen tissue samples were sectioned and mounted on glass slides. The slides were stained with hematoxylin and eosin for histopathological examinations. Only those tumor samples with more than 90% of tumor cells and matched tissue pairs with normal histological examination were included in this study. These frozen tissues were cut into 5-μm-thick sections and used for RNA isolation and cDNA synthesis. All the tumor samples had been classified as infiltrating ductal carcinoma. The use of the tissue material in this project was approved by the Research Ethics Committee of Ankara Numune Research and Teaching Hospital and consents were obtained in accordance with the Helsinki Declaration.

RNA extraction and cDNA synthesis

The frozen breast specimens were put into Trizol reagent (AppliChem, Darmstadt, Germany), disrupted with a homogenizer and total RNA was isolated according to the manufacturer's instructions. Genomic DNA contaminations were removed by on-column DNaseI treatment (Macharel Nagel, Duren, Germany). The concentration of the isolated RNA and the ratio of absorbance at 260 nm to 280 nm were measured with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Montchanin, DE, USA) in triplicate.

First-strand cDNA was synthesized from 1 μg total RNA using oligo(dT) primers using Revert Aid First strand cDNA synthesis kit according to the manufacturer's instructions (Fermentas, MD, USA). The cDNA was diluted at a ratio of 1:5 before being used as a PCR template and stored at -20°C until further use.

Real-Time quantitative RT-PCR

Real-time qRT-PCR analysis was performed using gene-specific primer pairs (Additional file 3). Real-time qRT-PCR was performed on the BioRad iCycler Instrument (BioRad Laboratories, Hercules, CA, USA). The amplification mixtures contained 1.0 μl of 1:5-diluted cDNA template, 6.25 μl SYBR Green PCR Master Mix Buffer, and 10 pmol of forward and reverse primers in a total volume of 12.5 μl. Cycling conditions were as follows: an initial incubation of 95°C for 5 min and then 45 cycles of 95°C for 30 s and 60°C for 30 s during which the fluorescence data were collected. To verify that the used primer pair produced only a single product, a dissociation protocol was added after thermocycling, determining dissociation of the PCR products from 55°C to 95°C. Tumor and matched normal samples were always analyzed in the same run to exclude between-run variations and each sample was studied in duplicate. A no-template control of nuclease-free water was included in each run. An initial set of randomly selected genes from the DN list was used for real-time qRT-PCR validation studies. RAD21, GSN, COX6C, MAF, SFRP1, SPTNB1, GSPT1, NME1, PTTG1 but not MAF were also present in the LN list. Furthermore, seven other genes with potential predictive power for tumor subtype classification were studied by real-time qRT-PCR. These genes included *FN1*, *ID4*, *EGFR*, *ADAMTS1*, *ATF3*, *IGFBP6*, and *PRNP*. The geometrical mean of *ACTB*, *TBP* and *SDHA1* gene expression values were used as internal control for relative gene expression quantitation [29]. Primer sequences and accession numbers of these genes were given in Additional file 3. The mean expression values obtained in resampling meta-analysis runs were used as a measure for comparing with the fold-change results obtained from the real-time qRT-PCR validation studies; a Pearson correlation coefficient was also calculated (Matlab®).

Results

Correlation of Sorlie and Zhao Datasets

Combining the datasets in meta-analysis requires that they have similar expressions, both in magnitude and individual variability. To assess whether the Sorlie and Zhao datasets were correlated, a Pearson's correlation coefficient was calculated between the mean expression values of the ductal or lobular samples from each dataset, respectively before and after performing t-tests (Figure 2). Even before the removal of IMAGE clones showing significant differences between the studies, the mean expression values of ductal samples from Sorlie were highly correlated with those from Zhao; and a similar result was observed for the lobular samples (r = 0.8329 and 0.8233,

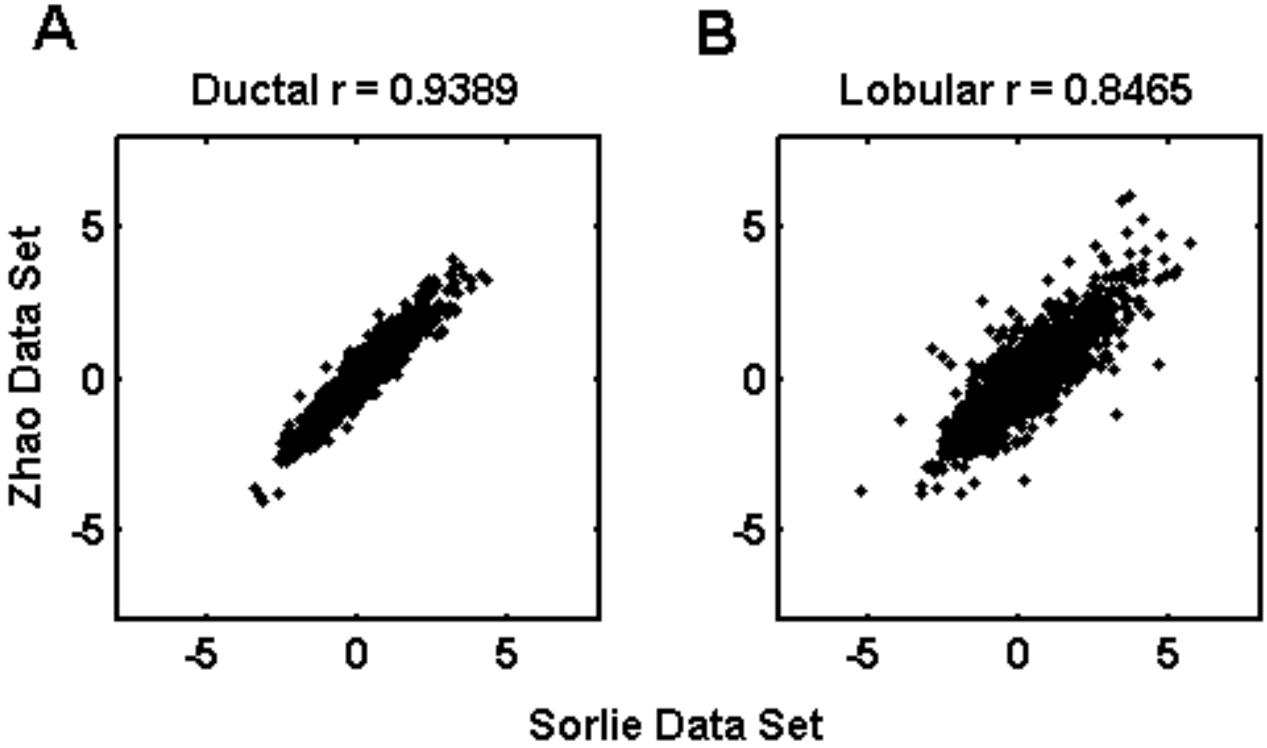


Figure 2
Pearson correlation coefficients (r) between Sorlie and Zhao datasets. Correlation plots between datasets after differentially expressed IMAGE clones were filtered out based on t-tests. (A) Correlation between mean expression values of ductal samples ($p < 0.05$). (B) Correlation between mean expression values of lobular samples ($p < 0.05$).

respectively). After filtering out the differentially expressed IMAGE clones, the correlations between the aforementioned datasets increased to 0.9389 and 0.8465 for the ductal and lobular samples, respectively. These results ensured that there was significant correlation between the Sorlie and Zhao datasets although they were based on independent tumor and normal samples.

Distribution statistics for generation of meta-lists

In this report, we used global-median normalized and filtered datasets since they minimized the number of manipulations performed during gathering of the meta-data (see Additional file 4). Accordingly, assessment of significance was based on p-values obtained from the Kolmogorov-Smirnov analysis between test and random distributions (*pt* and *pr1*, respectively) of a gene in the meta-data. For example, the *GSN* gene had a highly significant differential expression between ductal and normal samples as evidenced by the highly skewed distribution towards lower p-values whereas the *RAP2A* gene exhibited a uniform distribution of p-values (Figures 3A, B and 3C, D, respectively).

Effects of resampling on estimates of expression and differentially expressed gene number

We tested the effect of sample size and number of iterations on the estimation of mean expression level and the number of differentially expressed genes. For each run performed with a different sample size, the change in grand mean of expression (i.e., mean expression of all IMAGE clones) as well as the number of differentially expressed IMAGE clones were plotted with respect to the increasing number of iterations (Figure 4). As the number of iterations increased, the grand mean became more stabilized. Expectedly, the magnitude of change in mean values asymptotically decreased as the number of iteration and sampling size increased (Figure 4A and 4C). On the other hand, the number of genes stated as significant increased as a function of the number of iterations and sampling size (Figure 4B, 4D). Significant IMAGE clones made up more than 70% of all analyzed genes at sampling size 6 with the highest iteration in ductal vs. normal analysis whereas the same set-up resulted in only 20% significant IMAGE clones in ductal vs. lobular analysis.

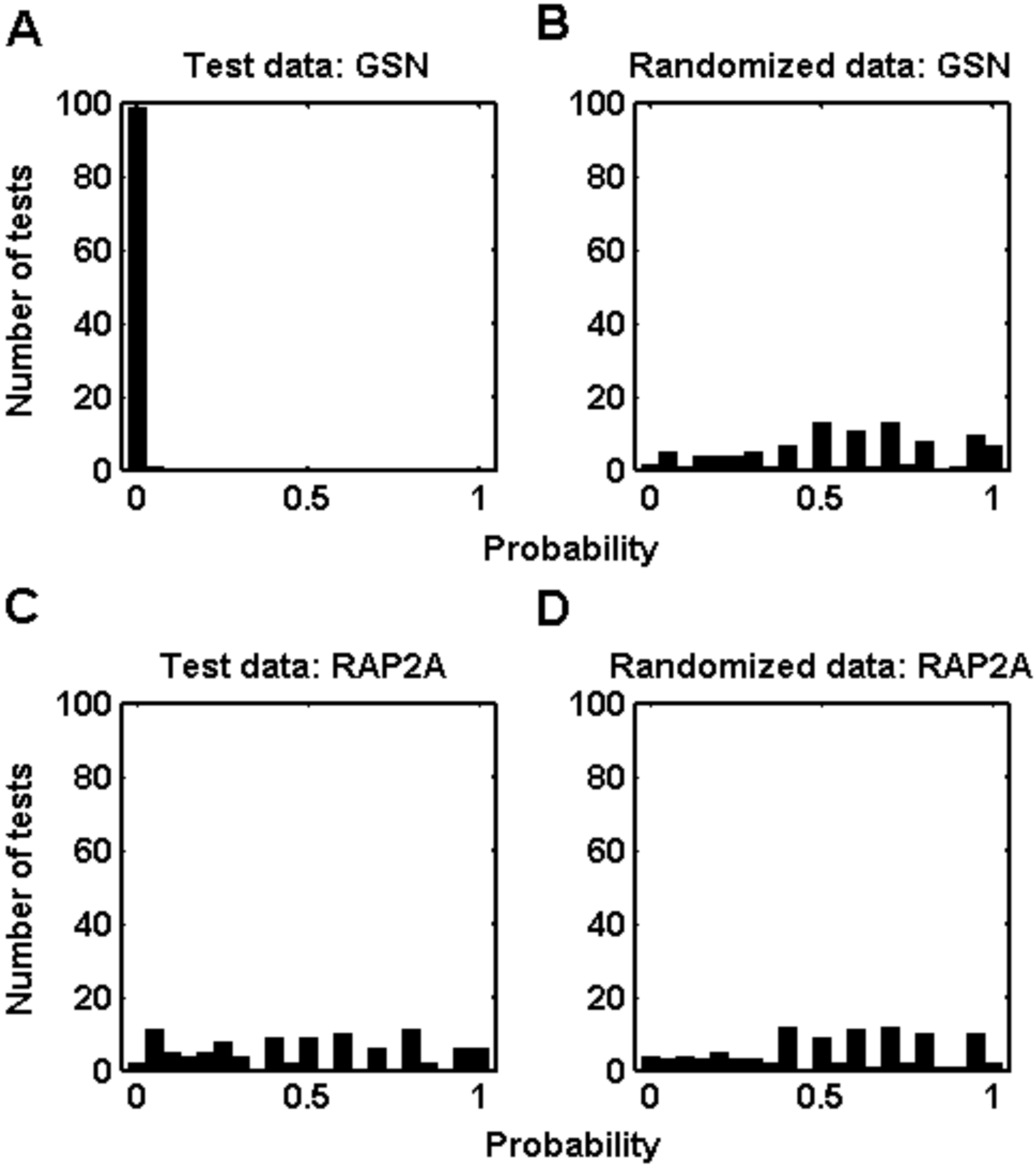


Figure 3

Examples for probability distributions of Wilcoxon rank sum tests. Data were obtained where resampling size, *n*, equaled to 6 (100 iterations). Assessment of significance was based on p-values obtained from the Kolmogorov-Smirnov test between test and random distributions (*pt* and *pr1*, respectively). (A, C) For test data, *GSN* gene had a highly significant differential expression (significant at 100% of iterations, p = 0.00) between ductal and normal samples whereas *RAP2A* gene did not (significant at 5% of iterations, p = 0.98). (B, D) Probability values of both *GSN* and *RAP2A*, obtained from randomized data, were uniformly distributed. *GSN*; IMAGE: 214990 and *RAP2A*; IMAGE:36684.

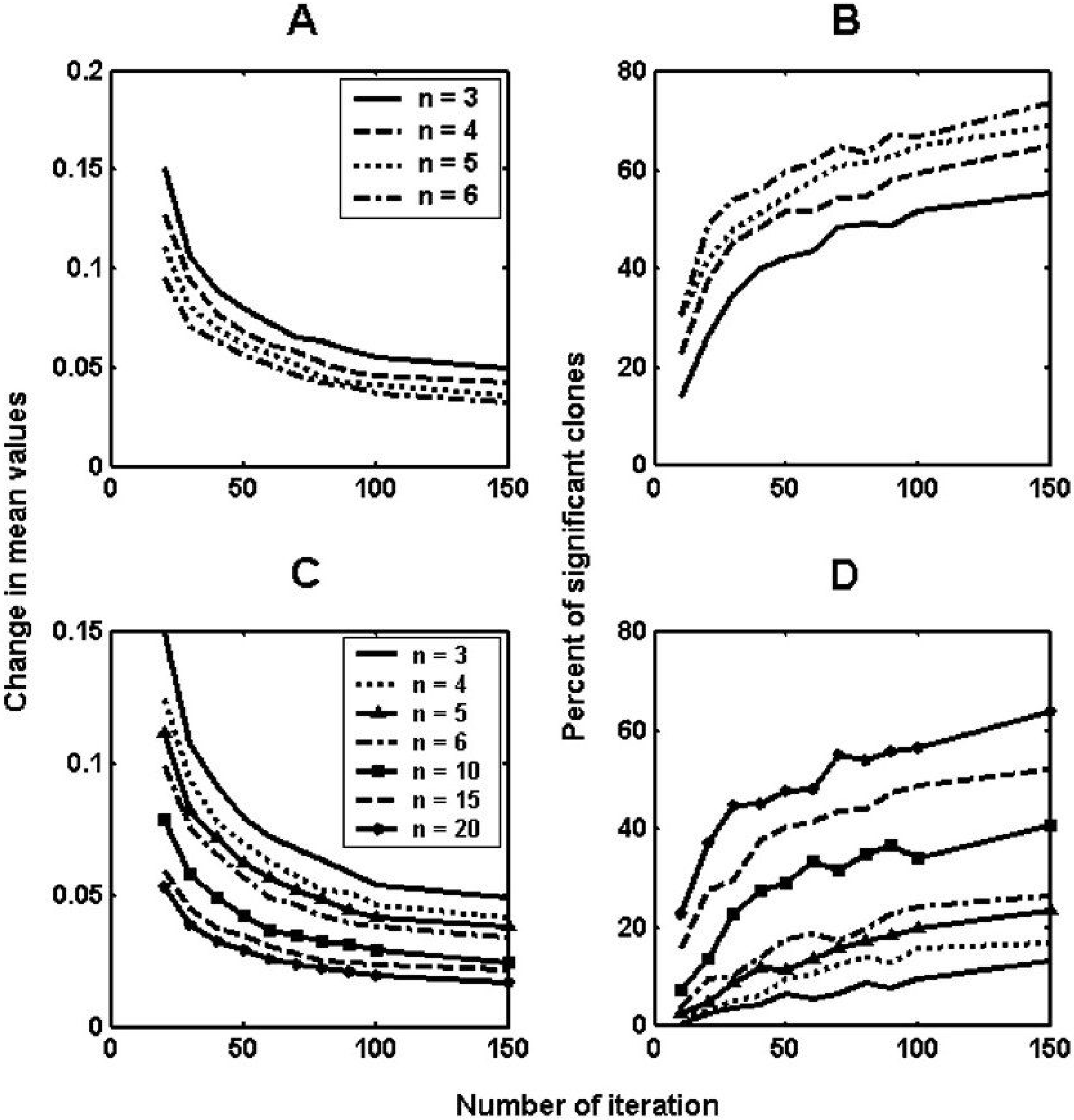


Figure 4

Effect of change in sample size and number of iterations on mean expression values and number of significant IMAGE clones. For each of the runs performed with different sample sizes (n), the change in the mean expression value (A, C) and the number of IMAGE clones that were stated as differentially expressed (B, D) were plotted with respect to the increasing number of iterations. A and B refer to the results of ductal vs. normal analysis whereas C and D show the results of ductal vs. lobular analysis.

It is reasonable to assume that use of a single sample size and iteration number may not be adequate to understand the variability among the tumor samples (Figure 4). It might instead be beneficial to consider all of the information gathered from the individual runs. Accordingly, the significant gene lists reported in this study were obtained by taking only those IMAGE clones that were assigned as significant in a given set of all resampling analyses performed (90% or more for ductal-normal, DN; and lobular-normal, LN; and 80% or more for ductal-lobular, DL comparisons) in an effort to minimize the effects of sampling size and iteration number on p-values.

Characteristics of differentially expressed meta-gene lists

Differentially-expressed gene lists for DN and LN contained 298 (282 genes) and 216 (202 genes) IMAGE clones, respectively (see Additional files 5 and 6). On the other hand, there were only 66 (65 genes) differentially expressed IMAGE clones between the ductal and lobular (DL) datasets for 80% criteria (see Additional file 7). The size of these lists was dependent on the False Discovery Rate (FDR) input value (herein set to 0.01) or the percentage of resampling runs considered for significance (i.e., 90% or 80%). In order to obtain a larger number of genes for DL analysis, the significance percentage value was set to 80.

The same resampling procedures were also performed on the individual datasets, Sorlie and Zhao, separately. Compared to our meta-analysis these separate analyses together could provide 91% of IMAGE clones that were present in the significant DN list and LN list and 68% of the IMAGE clones of the DL list. However neither of the studies could supply 9% of the IMAGE clones of the DN and LN list and 32% of the DL list (90% cut-off), each of which corresponds to a novel contribution by our meta-analysis (see Additional file 8 for meta-analysis specific gene lists).

We also compared the final DL significant gene list with the list of 52 genes reported by Zhao *et al.* [2]. The DL list shared *CDH1*, *AOC3*, *FADS2*, *SORBS1*, *ALDH1A1*, *LPL*, *ANXA1* and *AKR1C1* with that of Zhao *et al.* [2]. However, our analysis did not assign reasonable significance to the F11 and VWF genes according to the set cut-off criteria (80%). The remaining genes in the Zhao gene list were not encountered since they were not included in the combined dataset used in the present meta-analysis. Meta-analysis of these two datasets provided a total of 36 significant genes not previously reported by Zhao *et al* and when either dataset is analyzed individually (see Additional file 7).

Validation of tumor vs. normal meta-gene lists by independent microarray datasets

Recent meta-analysis studies identified common cancer signatures by combining microarray datasets from different tissues for increasing accuracy of tumor vs. normal class prediction [30,31]. In this study, we focused on extracting a stable tumor molecular signature based on two of the existing breast cancer studies that contain microarray data on normal, IDC, and ILC tissue samples. We also have validated the predictive power of the meta-gene lists obtained through the resampling-based meta-analysis using three additional breast cancer datasets, which contain microarray data on 3 or more samples of normal and tumor breast tissues (Table 1) [5,7,8]. Accordingly, subsets of genes from DN and LN meta-gene lists were able to predict the tumor vs. normal classes with high accuracies, ranging from 80 to 100% (Table 1). Strikingly, correlation between expression values obtained from significant discriminators from each of the three

Table 1: Summary of GEO breast cancer microarray datasets and results of class prediction analysis for the meta-gene lists, DN (Ductal/Normal) and LN (Lobular/Normal).

Study GEO ID	Class		Meta gene-list					
			DN			LN		
	N	T	Accuracy (%)	Number of genes	r_{DN}	Accuracy (%)	Number of genes	r_{LN}
Turashvili [5] GDS2635	10	10	93	57	0.85	80	49	0.87
Richardson [7] GDS2250	7	40	100	145	0.86	100	96	0.78
Karnoub [8] GSE8977	15	7	95.5	109	0.72	95.5	89	0.81

Normal (N) and tumor (T) sample sizes, accuracy of prediction from binary tree algorithm (% accuracy), and the number of genes in classifier (number of genes) were shown for each study, separately. Correlation (rDN, rLN) of the classifier expression from each study with the DN and LN meta-gene expressions were also indicated (Pearson correlation, Minitab®; $p < 0.001$).

normal/tumor datasets and those from the meta-analysis was high (Table 1). This indicated that the DN and LN lists harbored a robust expression profile for the breast tumors when compared with normal breast tissue.

Prediction of tumor-subtypes

We extracted a small, highly correlated classifier gene subset, which was commonly detected among the three microarray studies and the meta-analysis, to identify a more conservative gene set differentially expressed between tumor and normal cells (Additional file 9). Twenty-eight genes from the DN or LN meta-gene lists intersected with the three other microarray datasets (GDS2635, GDS2250, and GD1329); 17 of which were differentially expressed between basal vs. non-basal and/or ER status (Additional file 9). For example, *ADAMTS1*, *ATF3*, *IGFBP6*, *PRNP*, *EGFR*, *FN1*, *ID4*, *SPTBN1*, and *SFRP1* genes from the DN list were found to significantly different in expression between nonbasal-like vs. basal-like tumors as well as basal and luminal subtypes of the breast tumors ($p < 0.05$). All of the above genes except FN1 were found to be significantly associated with the tumor ER status ($p < 0.05$; Additional file 9).

Validation of ductal vs. lobular meta-gene list

Comparison of fold-change values of the DL meta-gene list consisting of 65 genes with that of the Turashvili's DL list (GDS2635) resulted in a high degree of correlation ($r = 0.53$; $p < 0.001$), suggesting that the direction and magnitude of expression change between the IDC and ILC samples were largely consistent between data from different microarray experiments. Furthermore, we combined published expression data from IDC and ILC samples from experiments performed by Bertucci *et al* [32] with the meta-analysis results (Additional File 7). Some of the members of the 65 meta-gene list were consistently down- or up-regulated also in the Turashvili and Bertucci datasets (i.e., down-regulated *ALDH1A1* and *RBP4* in IDC; and up-regulated *CDH1* and *TFAP2A* in IDC). Protein expression levels of these four genes were investigated using the Human Protein Atlas, a public resource for immunohistochemistry (IH) of normal and pathological human tissues http://www.proteinatlas.org/. IH data were available for CDH1, TFAP2A, and RBP4 proteins; and only data from antibodies exhibiting differential expression among breast tumors were reported herein. Accordingly, 2 out of 3 ILC samples exhibited moderate to strong signals for RBP4 (Antibody CAB00455) whereas 7 out of 9 IDC samples were either negative or had weak staining. CDH1 data in the Protein Atlas database was not very informative since the number of ILC samples were limited, but a moderate signal was detected for the ILC sample whereas 5 out of 6 IDC samples expressed CDH1 strongly (Antibody CAB000087). Similarly, TFAP2A was weakly or moderately expressed in the two ILC samples examined whereas a moderate to strong staining was observed in 5 of the 9 IDC samples. Although sample size in the ILC samples in the Human Protein Atlas database was limited, there was a corresponding trend between the mRNA levels reported by the present study and the protein level assessment obtained from the Human Protein Atlas. Future studies should include testing of the genes extracted by meta-analysis using protein level studies such as Western blotting or immunohistochemistry on a large set of IDC and ILC samples to confirm their predictive power.

Validation of meta-analysis by real time qRT-PCR

We first selected nine genes that were found to be differentially expressed in both the DN and LN lists (except *MAF*) from the meta-gene list for validation of the meta-analysis. Expression profiles of these genes were tested in independent paired IDC breast tumor and non-tumor tissue samples through real time qRT-PCR. Our results were consistent with those of the meta-analysis such that *GSN*, *SPTBN1*, *SFRP1* and *MAF* were down-regulated in most tumor samples with respect to their matched non-tumor samples whereas *COX6C*, *RAD21*, *GSPT1*, *NME1* and *PTTG1* were up-regulated (Figure 5). Additionally we selected seven other genes, *ATF3*, *ADAMTS1*, *EGFR*, *PRNP*, *IGFBP6*, *ID4* and *FN1*, found to be differentially expressed according to tumor subtype and ER+/ER- classification from the tumor-specific differentially expressed gene-set. All except *FN1* were found to be down-regulated in tumor samples with respect to their normal counterparts. The meta-analysis results were supported by the real-time qRT-PCR experiments since all tested genes exhibited differences between matched normal and tumor samples in the same direction as expected by the meta-analysis (Pearson correlation coefficient, $r = 0.78$, $p = 0.001$).

Among the genes we used for validation through real time qRT-PCR, *ID4* was the gene found to be differentially expressed between DN only by meta-analysis rather than each study alone.

Discussion

Microarrays allow high-throughput analysis of expression for thousands of genes and provide valuable information for tumor studies. For example, individual microarray studies have identified differentially expressed gene lists for distinguishing breast cancer subtypes and normal breast tissue [5,6,8,9]. Meta-analysis, on the other hand, might increase the knowledge by gathering and processing individual microarray datasets.

In the present study, we provided highly stable lists of differentially expressed genes based on meta-analysis of two breast cancer datasets [1,2]. We have used a resampling-based strategy in which the effects of number of iterations

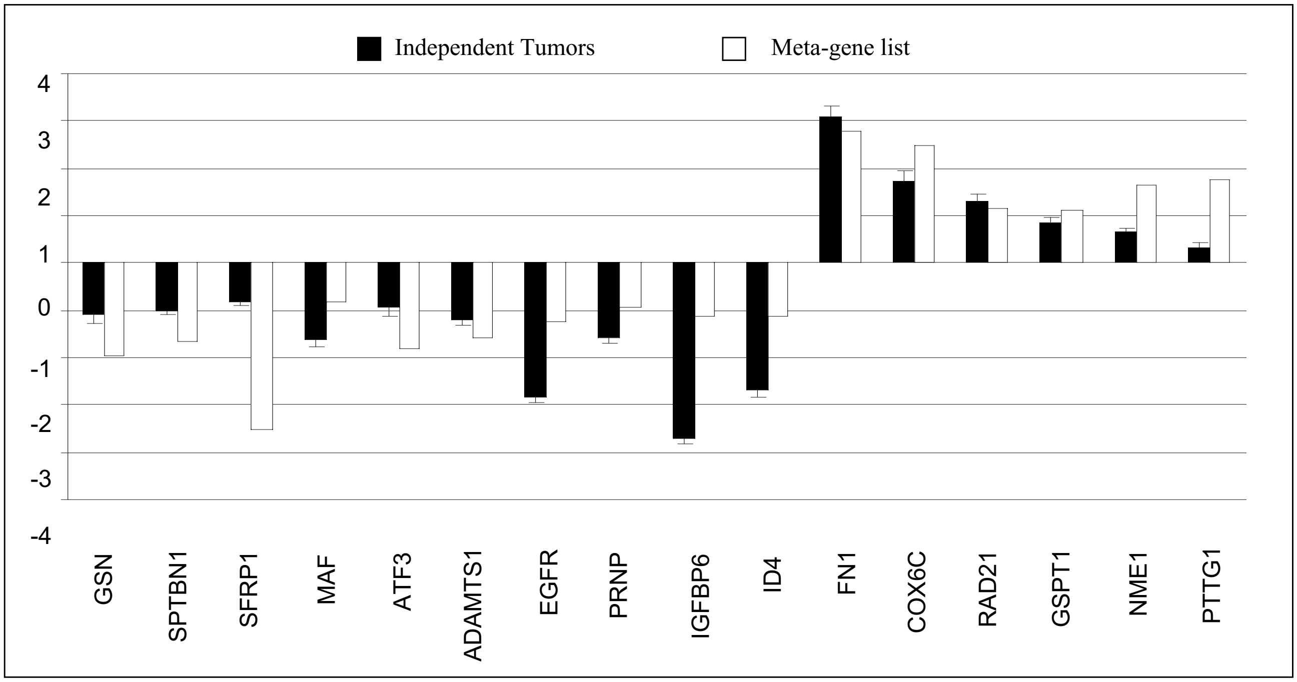


Figure 5

Validation of meta-analysis results by real-time qRT-PCR. Sixteen genes were selected from the ductal-normal (DN) significant meta-gene list for real-time qRT-PCR. Solid black bars refer to mean expression values (± SEM) of 10 independent IDC breast tumors normalized to their non-tumor pairs. White bars refer to the mean expression values from the combined meta-gene list.

and sample size were minimized by using a voting scheme in which each IMAGE clone, at each run, was voted as either significantly- or non-differentially expressed and the significant counts then were added up. A percentage value was obtained by dividing the number of significant votes by the total number of votes and a threshold of 80–90% for each IMAGE clone was chosen as a cut-off value for this meta-analysis. The meta-analysis was able to report multiple genes (i.e., 29, 21, and 6 genes for DN, LN, and DL, respectively) which neither dataset could report when analyzed individually.

Sample size greatly influences the reproducibility of the significant gene lists, such that the lower the sample size the less stable the gene lists become [19]. In addition, Qui *et al.* [18] have shown that the stability of genes identified as differentially expressed varies: some genes are consistently stable whereas others are not, independent of the statistical methodology used. Along these considerations, our voting scheme provided an advantage for extracting highly stable gene lists.

Different statistical methods are available for assessing differential expression. Among these, non-parametric tests allow for comparison of low sample size and distribution-independent comparisons. Our choice of rank-sum test was based on this idea; similarly, previous studies reported the use of Kolmogorov-Smirnov test statistics to compare the reference and sample distributions in the context of Gene Ontologies [33]. We used a Kolmogorov-Smirnov test statistic for comparison of test and random distributions of p-values obtained from rank sum tests. In generating random datasets, we applied a gene-wise permutation algorithm that preserved the expression level information. Based on gene-wise permutations, a set of probability values that compare the actual and randomized distributions allowed for the assessment of the significance of the difference between groups tested using the Kolmogorov-Smirnov tests.

Different studies can be normalized and directly compared to each other in meta-analysis. Our comparisons ensured that there was a significant correlation between the Sorlie and Zhao datasets although these studies were based on independent tumor and normal samples; and the experimental procedures (e.g., amplification of RNA) also varied considerably between the two studies. Median rank scores [16] or quantile discretization algorithms have frequently been used to transform gene expression values from different studies to a common numerical range [17].

Since the global median-normalized and quantile-normalized data correlated well (see Additional file 4), we have used the former normalization method, with the least number of data manipulation steps, before combining these two datasets.

Due to the large number of comparisons involved in microarray data analysis, it is important to take into account the false positive error rate and control it for the number of tests performed. FDR is a well-known methodology for multiple-test correction; its estimation relies on calculation of the number of false positives in a randomly permuted set of experiments [34]. Therefore, we made comparisons between randomly shuffled datasets to obtain an estimate of FDR; and kept the value of FDR low (% 0.01) to reduce the number of false positives.

Invasive breast tumors comprise 18 different histological types [24], most of which are classified as invasive ductal carcinoma not otherwise specified (IDC NOS). ILC, on the other hand, makes about 10–15% of all breast tumors and it is histologically characterized by uniform tumor cells arranged in single-files or concentrically localized around ducts [35]. ILC exhibit heterogeneity just like IDC; and a high-grade aggressive form of ILC known as pleomorphic lobular carcinoma (PLC) exists [36]. Bertucci *et al.* [32] reported that IDC and ILC were histologically and genomically distinguishable from each other among the ER(+) grade II invasive breast tumors. Furthermore, ILC molecular subtypes were reported to include the typical and IDC-like ILCs, yet the *CDH1* mutation and/or under-expression was common but not universal to ILCs in general [35]. Low-grade breast tumors were generally characterized by ER(+), PR(+) and with limited genomic aberrations whereas high grade tumors were generally ER(-) and PR(-) and had complex karyotypic changes. However, molecular differences among subtypes may not surpass the differences between any tumor cell and the normal since the degree of genomic stability in normal cells would be relatively higher.

The other three studies presenting data on ILC and IDC, Turashvili *et al.* [5], Sorlie *et al.* [1] and Zhao *et al.* [2] have used a more diverse selection of tumor samples. Although IDC and ILC have distinctive clinical and pathological characteristics and differ in their ER status and metastatic behaviors [22], meta-analysis of Zhao and Sorlie datasets indicated that a small number of genes distinguished between the expression profiles of IDC and ILC patients. On the other hand, the number of genes that were differentially expressed between normal and IDC or normal and ILC samples was much greater. Indeed, Turashvili *et al.* [5] has also reported only 28 genes that were significantly differentially expressed between IDC and ILC samples, which were extracted using laser-dissection, a more recent methodology allowing for precise collection of a given cell population. These findings suggest that the degree of molecular differences between IDC and ILC are indeed smaller than those between the tumor and normal classes.

Comparisons between the meta-analysis and the Turashvili and Bertucci studies pointed out to *CHD1*, *TFAP2A*, *RBP4*, and *ALDH1A1* genes as commonly modulated. Indeed, *CDH1* is one of the best-studied discriminators for ductal/lobular breast cancer specimens in the literature by immunohistochemistry and at the genomic level. In breast cancer, reduced *CDH1* expression has been found in 50% of invasive ductal carcinomas, whereas *CDH1* expression was almost always absent in infiltrating lobular carcinoma (ILC) [1,2,5,32,37,38]. *TFAP2A* was shown to be highly expressed in ductal tumor cells while normal cells expressed *TFAP2A* in the inner glandular cell layer [39]. On the other hand, nuclear TFAP2 expression was shown to be higher in lobular than ductal breast carcinomas [40]. There is no report on *RBP4* in the literature in connection with ductal vs. lobular breast cancer distinction while ALDH1A1 protein levels were shown to exhibit differences among the ductal carcinoma patients [41]. The candidates identified in the meta-analysis then are likely to be discriminatory at the mRNA level rather than the protein level since protein localization and variability in intensity might make the ductal vs. lobular tissue discrimination less clear. Therefore, it is of paramount importance that future confirmatory studies include use of independent ILC and IDC samples for quantitative expression profiling of the selected candidate genes.

On the other hand, analyses of Sorlie, Zhao, and Turashvili data showed that tumor cells were remarkably distinct from their respective normals in their transcription profiles implicating that whatever the subtype structure underneath, most of the variability among samples was due to changes during tumorigenesis. Accordingly, the idea that genes discriminating tumor from normal in a stable manner also may have information on the state of the tumorigenesis is a valid one.

Breast tumor subtype classification remains a complicated issue due to difficulties associated with the presence of multiple interacting factors such as the presence or absence of node-filtration, ER-positivity, metastatic potential, different degrees of genomic instability, and tumor cell origin. For example, basal like cancers have distinct molecular expression profiles and histological differences when compared with the luminal type [42]. Nielsen *et al.* [43] have categorized basal like breast cancer tumors as having variable levels of expression of one of the three stem/basal markers, namely CK5/6, EGFR, and c-kit. Luminal cell markers, on the other hand, include CK8, CK18, CK19, mostly characteristic of glandular and/or lobular epithelial cells [44]. However, both the basal and luminar histochemical markers may exist simultaneously suggesting that breast cancer is rather a heterogeneous tissue [45]. It is also evident that tumors with a triple negative status (ER-, PR-, HER2-) are more likely to belong to the basal type [43,46]. In general, gene expression studies associated the basal-like breast tumors with high proliferative abilities and thus having a worse prognosis when compared with the luminal subtype of breast cancers [1,47]. Thus identification of genes best classifying breast cancer into intrinsic molecular subtypes like luminal, HER2+/ER- and basal-like also allows determination of risk factors and likely prognosis for the patients. The importance of identification of these different subtypes is that they differ in clinical outcome and molecular subtype signatures thus help predict clinical outcome and response to therapy.

Differentially expressed genes between tumor and normal states (DN and LN) also keep information about intrinsic subtypes. Accordingly, meta-analysis identified *ATF3*, *ADAMTS1*, *EGFR*, *PRNP*, *IGFBP6*, *ID4*, *SFRP1*, *SPTBN1*, and *FN1* with ability to classify tumors into basal and luminal subclasses. Additionally most of them accurately differentiated ER(+) and ER(-) tumors (Additional file 9).

Among those genes, *ID4* was found to be a novel tumor suppressor gene in normal human breast tissues and epigenetically silenced in breast cancer cell lines and primary breast tumors [48,49]. As supporting information for our data, de Candia et al. suggested that the expression of *ID4* in the mammary duct epithelium may be regulated by estrogen depending on the differential expression pattern of ID4 in ER(+) and ER(-) breast tumors [50]. *SFRP1* on the other hand is a frizzled-related protein that plays a role in a variety of cellular processes, including control of cell polarity, cell fate determination, and malignant transformation. In previous studies, loss of *SFRP1* was found to be associated with cancer progression and poor prognosis in breast cancer [51,52]. EGFR is known to be a positive immunohistochemical marker for basal-like breast cancers and it was shown to accurately identify basal-like tumors from microarray data with potential therapeutic implications [53,54]. Activating transcription factor 3 (*ATF3*) is a member of the ATF/cyclic AMP response element-binding family of transcription factors. It was shown to enhance apoptosis in the untransformed mammary epithelial cells while protecting the aggressive cells and enhancing cell motility. Array analyses indicated that ATF3 upregulated the expression of several genes in the tumor necrosis factor pathway in the untransformed mammary epithelial cells. However, the expression of several genes implicated in tumor metastasis including fibronectin (*FN1*) was upregulated in aggressive cells. ATF3 was also shown to regulate the transcription of FN1, one of the genes obtained in the present study. *ATF3* gene copy number was at least doubled in 80% of the breast tumors examined; protein levels also were elevated in close to 50% of these tumors [55].

Since the normal vs. tumor classification was strikingly distinct based on meta-analysis, and a gene-set with the capacity for breast cancer subtype classification, we further analyzed a set of normal matched tumors for selected genes from the meta-gene list using real-time qRT-PCR. The selected 16 significant genes were shown to have expression profiles similar to those found from the meta-analysis. Our findings also suggested that these genes could be used as predictors of tumor status regardless of the origin of the reference samples, i.e., a matched or pooled reference tissue. Since the number of samples used in qRT-PCR was relatively small, increasing the sample size may help generalize our results to a wider range of breast tumor samples.

There was a high level of correlation between fold changes obtained from the DL meta-genes and those from the Turashvili dataset, regardless of the different sample extraction methods used in each study (i.e., frozen sections and laser-dissection, respectively). The meta-gene list discriminating between ductal and lobular breast tumor samples at the mRNA level requires further confirmation at the protein level to better assess discriminatory power. Future validation studies might concentrate on whether meta-analysis specific genes also participate in prediction of level of prognosis and/or time to disease-free survival.

Conclusion

In this study, meta-analysis of two independent comparable microarray data sets allowed us to provide genes that are able to discriminate IDC and ILC and normal mammary cells from the tumors that either study by itself was not able to identify. We also provided highly generalized and stable gene lists that could be used for prediction of tumor or normal status. The meta-gene list for tumor/normal comparison had a striking predictive ability based on comparisons made with three independent microarray datasets. The resampling approach proposed herein has the ability to detect a set of differentially expressed genes, with the least amount of within-group variability. This meta-analytic approach thus provides a method to combine two or more independent cancer data sets leading to the identification of differentially expressed gene sets for better understanding of cancer development and progression.

Abbreviations

DL: Ductal and Lobular; DN: Ductal and Normal; FDR: False Discovery Rate; IDC: Invasive Ductal Carcinoma; ILC: Invasive Lobular Carcinoma; LN: Lobular and Normal; qRT-PCR: Quantitative Reverse Transcriptase Polymerase Chain Reaction; SMD: Stanford Microarray Database.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IGY, BGD and OK originally conceptualized the meta-analysis method and application to breast cancer datasets, and wrote the manuscript. OK and SK developed the meta-analysis algorithm; SK contributed to manuscript writing. BGD and SK applied the algorithm to breast cancer datasets and generated figures. BDG performed all BRB-tool analyses and carried out expression studies. ARO and OK wrote algorithms for data extraction and compilation. BB carried out the surgical removal of the sample tissues and GE, the pathological assessment of the surgical tissue materials. IGY coordinated the study and participated in its design. All authors read and approved the final manuscript.

Additional material

Additional file 1

Complete list of common IMAGE clones. *The data represents 4769 common IMAGE clones from combined Sorlie and Zhao datasets from the study of Zhao* et al. *[2].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S1.xls]

Additional file 2

Complete list of common IMAGE clones. *The data represents 4769 common IMAGE clones from combined Sorlie and Zhao datasets from the study of Sorlie* et al. *[1].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S2.xls]

Additional file 3

Genes used for real-time qRT-PCR analysis. *Gene names, accession numbers and gene specific primer pairs used for real-time qRT-PCR analysis of the selected genes.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S3.pdf]

Additional file 4

Correlation of global median and quantile normalized data. *The figure shows the correlation of global median and quantile normalized data of ductal (D) and normal (N) tissue samples.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S4.pdf]

Additional file 5

Gene set differentially expressed between ductal (D) and normal (N) breast tissue samples. *The data provided represents the list of 298 IMAGE clones differentially expressed between ductal (D) and normal (N) samples with 90% significance.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S5.xls]

Additional file 6

Gene set differentially expressed between lobular (L) and normal (N) breast tissue samples. *The data provided represents the list of 216 IMAGE clones differentially expressed between lobular (L) and normal (N) samples with 90% significance.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S6.xls]

Additional file 7

Gene set differentially expressed between ductal (D) and lobular (L) breast tissue samples. *The data provided represents the list of 66 IMAGE clones differentially expressed between ductal (D) and lobular (L) samples with 80% significance.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S7.xls]

Additional file 8

List of meta-analysis specific genes. *The file includes the list of genes differentially expressed between DN, LN, and DL samples.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S8.xls]

Additional file 9

Validation of meta-analysis gene lists by three independent microarray datasets. *The file represents the gene lists obtained by comparing the DN and LN meta-gene list to the gene expression profiles of normal and breast tumor tissues from three independent microarray datasets.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-8-396-S9.xls]

Acknowledgements

This work was funded by a grant from the Scientific and Technical Research Council of Turkey (107T181) and supported by Bilkent University.

References

1. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Bodstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100:**8418-8423.
2. Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, Bukholm IK, Karesen R, Botstein D, Borresen-Dale AL, Jeffrey SS: **Different Gene Expression Patterns in Invasive Lobular and Ductal Carcinomas of the Breast.** *Mol Biol Cell* 2004, **15:**2523-2536.
3. van 't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene**

expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, **415(6871):**530-536.

4. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, Macgrogan G, Bergh J, Cameron D, Goldstein D, Duss S, Nicoulaz AL, Brisken C, Fiche M, Delorenzi M, Iggo R: **Identification of molecular apocrine breast tumours by microarray analysis.** *Oncogene* 2005, **24:**4660-4671.
5. Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, Klein J, Fridman E, Skarda J, Srovnal J, Hajduch M, Murray P, Kolar Z: **Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis.** *BMC Cancer* 2007, **7:**55.
6. Grigoriadis A, Mackay A, Reis-Filho JS, Steele D, Iseli C, Stevenson BJ, Jongeneel CV, Valgeirsson H, Fenwick K, Iravani M, Leao M, Simpson AJ, Strausberg RL, Jat PS, Ashworth A, Neville AM, O'Hare MJ: **Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data.** *Breast Cancer Res* 2006, **8:**R56.
7. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: **X chromosomal abnormalities in basal-like human breast cancer.** *Cancer Cell* 2006, **9:**121-132.
8. Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, Richardson AL, Polyak K, Tubo R, Weinberg RA: **Mesenchymal stem cells within tumor stroma promote breast cancer metastasis.** *Nature* 2007, **449:**557-563.
9. Tripathi A, King C, de la Morenas A, Perry VK, Burke B, Antoine GA, Hirsch EF, Kavanah M, Mendez J, Stone M, Gerry NP, Lenburg ME, Rosenberg CL: **Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients.** *Int J Cancer* 2008, **122:**1557-1566.
10. Smith DD, Saetrom P, Snøve O Jr, Lundberg C, Rivas GE, Glackin C, Larson GP: **Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation.** *BMC Bioinformatics* 2008, **9:**63.
11. Thomassen M, Tan Q, Kruse TA: **Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis.** *Breast Cancer Res Treat* 2009, **113(2):**239-249.
12. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7:**96.
13. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: **Comparison and meta-analysis of microarray data: from the bench to the computer desk.** *Trends Genet* 2003, **19:**570-577.
14. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res* 2002, **62:**4427-4433.
15. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling interstudy variation.** *Bioinformatics* 2003, **19(Suppl 1):**i84-90.
16. Toedling J, Spang R: **Assessment of Five Microarray Experiments on Gene Expression Profiling of Breast Cancer.** *Poster Presentation RECOMB* 2003 [http://citeseer.ist.psu.edu/611350.html].
17. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.** *BMC Bioinformatics* 2005, **6:**265.
18. Qiu X, Xiao Y, Gordon A, Yakovlev A: **Assessing stability of gene selection in microarray data analysis.** *BMC Bioinformatics* 2006, **7:**50.
19. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: **Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex.** *Neurochem Res* 2004, **29:**1213-1222.
20. Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S: **Integrative analysis of multiple gene expression profiles applied to liver cancer study.** *FEBS Lett* 2004, **565:**93-100.
21. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, Kloppel G, Saeger HD, Pilarsky C: **Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.** *Oncogene* 2005, **24:**5079-5088.
22. Arpino G, Bardou VJ, Clark GM, Elledge RM: **Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome.** *Breast Cancer Res* 2004, **6:**R149-156.
23. Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep AL, Chen YY, Chew KL, Dairkee SH, Jensen RM, Waldman FM: **Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis.** *Cancer Res* 2003, **63:**7167-7175.
24. Weigelt B, Horlings H, Kreike B, Hayes M, Hauptmann M, Wessels L, de Jong D, Vijver M Van de, Veer LjV, Peterse J: **Refinement of breast cancer classification by molecular characterization of histological special types.** *J Pathol* 2008, **216:**141-150.
25. Li CI, Malone KE, Porter PL, Weiss NS, Tang MT, Daling JR: **Reproductive and anthropometric factors in relation to the risk of lobular and ductal breast carcinoma among women 65–79 years of age.** *Int J Cancer* 2003, **107:**647-651.
26. **Biometric Research Branch** [http://linus.nci.nih.gov/BRB-ArrayTools.html]
27. **Stanford Microarray Database** [http://genome-www5.stanford.edu/]
28. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
29. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Res* 2001, **29:**e45.
30. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101:**9309-9314.
31. Xu L, Geman D, Winslow RL: **Large-scale integration of cancer microarray data identifies a robust common cancer signature.** *BMC Bioinformatics* 2007, **8:**275.
32. Bertucci F, Orsetti B, Nègre V, Finetti P, Rougé C, Ahomadegbe JC, Bibeau F, Mathieu MC, Treilleux I, Jacquemier J, Ursule L, Martinec A, Wang Q, Bénard J, Puisieux A, Birnbaum D, Theillet C: **Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles.** *Oncogene* 2008, **27:**5359-5372.
33. Ben-Shaul Y, Bergman H, Soreq H: **Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression.** *Bioinformatics* 2005, **21:**1129-1137.
34. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Roy Statis Soc Ser* 1995, **57:**289-300.
35. Yoder BJ, Wilkinson EJ, Massoll NA: **Molecular and morphologic distinctions between infiltrating ductal and lobular carcinoma of the breast.** *Breast J* 2007, **13:**172-179.
36. Simpson PT, Reis-Filho JS, Lambros MB, Jones C, Steele D, Mackay A, Iravani M, Fenwick K, Dexter T, Jones A, Reid L, Da Silva L, Shin SJ, Hardisson D, Ashworth A, Schmitt FC, Palacios J, Lakhani SR: **Molecular profiling pleomorphic lobular carcinomas of the breast: evidence for a common molecular genetic pathway with classic lobular carcinomas.** *J Pathol* 2008, **215:**231-244.
37. Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, Gamallo C, Esteller M, Palacios J: **Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability.** *Int J Cancer* 2003, **106:**208-215.
38. Caldeira JR, Prando EC, Quevedo FC, Neto FA, Rainho CA, Rogatto SR: **CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer.** *BMC Cancer* 2006, **6:**48.
39. Friedrichs N, Jäger R, Paggen E, Rudlowski C, Merkelbach-Bruse S, Schorle H, Buettner R: **Distinct spatial expression patterns of AP-2alpha and AP-2gamma in non-neoplastic human breast and breast cancer.** *Mod Pathol* 2005, **18:**431-438.
40. Pellikainen J, Kataja V, Ropponen K, Kellokoski J, Pietiläinen T, Böhm J, Eskelinen M, Kosma VM: **Reduced nuclear expression of transcription factor AP-2 associates with aggressive breast cancer.** *Clin Cancer Res* 2002, **8:**3487-3495.
41. Sládek NE: **Aldehyde dehydrogenase-mediated cellular relative insensitivity to the oxazaphosphorines.** *Curr Pharm Des* 1999, **5:**607-625.
42. Fadare O, Tavassoli FA: **The phenotypic spectrum of basal-like breast cancers: a critical appraisal.** *Adv Anat Pathol* 2007, **14:**358-373.

43. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, Rijn M van de, Perou CM: **Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.** *Clin Cancer Res* 2004, **10:**5367-5374.

44. Abd El-Rehim DM, Pinder SE, Paish CE, Bell J, Blamey RW, Robertson JF, Nicholson RI, Ellis IO: **Expression of luminal and basal cytokeratins in human breast carcinoma.** *J Pathol* 2004, **203:**661-671.

45. Moriya T, Kasajima A, Ishida K, Kariya Y, Akahira J, Endoh M, Watanabe M, Sasano H: **New trends of immunohistochemistry for making differential diagnosis of breast lesions.** *Med Mol Morphol* 2006, **39:**8-13.

46. Liu ZB, Liu GY, Yang WT, Di GH, Lu JS, Shen KW, Shen ZZ, Shao ZM, Wu J: **Triple-negative breast cancer types exhibit a distinct poor clinical characteristic in lymph node-negative Chinese patients.** *Oncol Rep* 2008, **20:**987-994.

47. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: **Breast cancer classification and prognosis based on gene expression profiles from a population-based study.** *Proc Natl Acad Sci USA* 2003, **100:**10393-10398.

48. Noetzel E, Veeck J, Niederacher D, Galm O, Horn F, Hartmann A, Knüchel R, Dahl E: **Promoter methylation-associated loss of ID4 expression is a marker of tumour recurrence in human breast cancer.** *BMC Cancer* 2008, **8:**154.

49. Umetani N, Mori T, Koyanagi K, Shinozaki M, Kim J, Giuliano AE, Hoon DS: **Aberrant hypermethylation of ID4 gene promoter region increases risk of lymph node metastasis in T1 breast cancer.** *Oncogene* 2005, **24:**4721-4727.

50. de Candia P, Akram M, Benezra R, Brogi E: **Id4 messenger RNA and estrogen receptor expression: inverse correlation in human normal breast epithelium and carcinoma.** *Hum Pathol* 2006, **37:**1032-1041.

51. Klopocki E, Kristiansen G, Wild PJ, Klaman I, Castanos-Velez E, Singer G, Stohr R, Simon R, Sauter G, Leibiger H, Essers L, Weber B, Hermann K, Rosenthal A, Hartmann A, Dahl E: **Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors.** *Int J Oncol* 2004, **25:**641-649.

52. Kawano Y, Kypta R: **Secreted antagonists of the Wnt signalling pathway.** *J Cell Sci* 2003, **116:**2627-2634.

53. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, Perou CM, Nielsen TO: **Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype.** *Clin Cancer Res* 2008, **14:**1368-1376.

54. Arnes JB, Bégin LR, Stefansson IM, Brunet JS, Nielsen TO, Foulkes WD, Akslen LA: **Expression of EGFR in relation to BRCA1 status, basal-like markers and prognosis in breast cancer.** In *J Clin Pathol* 2008.

55. Yin X, Dewille JW, Hai T: **A potential dichotomous role of ATF3, an adaptive-response gene, in cancer development.** *Oncogene* 2008, **27:**2118-2127.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2407/8/396/prepub

Oncology Research, Vol. 17, pp. 1–100
Printed in the USA. All rights reserved.
Copyright © 2009 Cognizant Comm. Corp.

0965-0407/09 $90.00 + .00
E-ISSN 1555-3906
www.cognizantcommunication.com

Identification of Endogenous Reference Genes for qRT-PCR Analysis in Normal Matched Breast Tumor Tissues

Bala Gur-Dedeoglu,* Ozlen Konu,* Betul Bozkurt,† Gulusan Ergul,‡ Selda Seckin,‡ and Isik G. Yulug*

*Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, Ankara, Turkey
†Department of General Surgery, Ankara Numune Research and Teaching Hospital, Ankara, Turkey
‡Department of Pathology, Ankara Numune Research and Teaching Hospital, Ankara, Turkey

Quantitative gene expression measurements from tumor tissue are frequently compared with matched normal and/or adjacent tumor tissue expression for diagnostic marker gene selection as well as assessment of the degree of transcriptional deregulation in cancer. Selection of an appropriate reference gene (RG) or an RG panel, which varies depending on cancer type, molecular subtypes, and the normal tissues used for interindividual calibration, is crucial for the accurate quantification of gene expression. Several RG panels have been suggested in breast cancer for making comparisons among tumor subtypes, cell lines, and benign/malignant tumors. In this study, expression patterns of 15 widely used endogenous RGs (*ACTB, TBP, GAPDH, SDHA, HPRT, HMBS, B2M, PPIA, GUSB, YWHAZ2, PGK1, RPLP0, PUM1, MRPL19*, and *RPL41*), and three candidate genes that were selected through analysis of two independent microarray datasets (*IL22RA1, TTC22*, *ZNF224*) were determined in 23 primary breast tumors and their matched normal tissues using qRT-PCR. Additionally, 18S rRNA, *ACTB*, and *SDHA* were tested using randomly primed cDNAs from 13 breast tumor pairs to assess the rRNA/mRNA ratio. The tumors exhibited significantly lower rRNA/mRNA ratio when compared to their normals, on average. The expression of the studied RGs in breast tumors did not exhibit differences in terms of grade, ER, or PR status. The stability of RGs was examined based on two different statistical models, namely GeNorm and NormFinder. Among the 18 tested endogenous reference genes, *ACTB* and *SDHA* were identified as the most suitable reference genes for the normalization of qRT-PCR data in the analysis of normal matched tumor breast tissue pairs by both programs. In addition, the expression of the gelsolin (*GSN*) gene, a well-known downregulated target in breast tumors, was analyzed using the two most suitable genes and different RG combinations to validate their effectiveness as a normalization factor (NF). The GSN expression of the tumors used in this study was significantly lower than that of normals showing the effectivity of using *ACTB* and *SDHA* as suitable RGs in this set of tumor–normal tissue panel. The combinational use of the best performing two RGs (*ACTB* and *SDHA*) as a normalization factor can be recommended to minimize sample variability and to increase the accuracy and resolution of gene expression normalization in tumor–normal paired breast cancer qRT-PCR studies.

Key words: Real-time quantitative RT-PCR; Endogenous reference genes; Normalization factor; Breast cancer

INTRODUCTION

Real-time quantitative RT-PCR (reverse-transcription polymerase chain reaction) is one of the most sensitive and specific methods for quantification of expression at the mRNA level (1–4). Inclusion of an endogenous reference gene or genes (RGs) is crucial to standardize initial RNA quantity to overcome bias originating from RNA measurement errors, problems with RNA integrity, and differential cDNA conversion efficiencies (5–7). Different options exist to quantify expression from the results of a qRT-PCR run, such as the relative quantification by the $2^{-\Delta\Delta Ct}$ method or mRNA copy number estimation (8). Quantification of a target gene requires the use of a proper RG whose expression is relatively stable across samples to estimate the degree of variability within and among experimental groups as well as to standardize the expression to a baseline common to all samples (5–7,9). Nevertheless, numerous studies show an inherent instability in regard to expression of housekeeping genes, many of which are still commonly used as references (10–16).

Analysis of gene expression is fundamental for cancer research for the detection of subtle differential expression between tumor and normal tissues or among different tumor types. In particular, recent target valida-

Address correspondence to Isik G. Yulug, Ph.D., Department of Molecular Biology and Genetics, Bilkent University, Faculty of Science, TR-06800 Ankara, Turkey. Tel: +90-312-290-2506; Fax: +90-312-266-5097; E-mail: yulug@fen.bilkent.edu.tr

tion and disease diagnostic marker selection studies rely primarily on gene expression comparisons between tumor–normal pairs (17–21).

Moreover, the use of multiple endogenous RGs significantly increases the accuracy of the normalization by reducing the impact of outliers (5,9). Accordingly, a plethora of single or combinational usage of two or more RGs has been recommended for relative quantification of expression data for various tumor tissue types (22–29).

Breast cancer is the most common cancer affecting women worldwide. New high-throughput technologies have opened the possibility to study the gene expression profile of the tumors. The validation of differentially expressed genes using independent methodology such as qRT-PCR is often desirable. In breast cancer qRT-PCR studies, different single housekeeping genes have been used to quantify the expression level of target genes (30–40). Recently, *MRPL19* and *PPIA* were reported as a stable RG combination to analyze benign and malignant breast cancer specimens (41). Similarly, Lyng et al. reported an RG panel comprised of *TBP*, *RPLP0*, and *PUM1* for normalizing the gene expression levels across the ER+ and ER– breast tumors, and normal breast tissues (42).

However, there are yet no systematic studies reporting on the expression of commonly used RGs in tumor-matched normal breast samples.

The aim of this study was to identify a suitable RG(s) that can be used as a normalization factor (NF) for more accurate and reliable normalization of paired breast tumor–normal tissue gene expression studies with qRT-PCR.

We evaluated 18 potential candidate RGs listed in Table 1 for their expression profile in 23 normal paired breast tumor tissue specimens. The genes *ACTB* and *SDHA* were calculated as the most stable RGs by two dedicated validation programs, geNorm and NormFinder. Furthermore, the suitability of these RGs as NF individually or in combination was validated based on the relative expression quantification of gelsolin (*GSN*). Correlation coefficients between *GSN* expression values that were normalized either to a single RG or combinations of RGs in breast tumors were also assessed. We also determined the expression of 18S rRNA to ACTB or SDHA mRNA from randomly primed cDNA on a subset of tumor–normal samples ($n = 13$ pairs) and found that tumors exhibited significantly lower rRNA/mRNA ratio.

The results of the present study showed that using the geometric mean of the combinations of two of the best performing RGs as NF can be used to reduce the variability between tumor samples and their normal counterparts while studying their expression by qRT-PCR in breast tumor samples.

MATERIALS AND METHODS

Patients and Samples

Primary tumor samples and matched normal breast tissues were obtained from patients ($n = 23$, mean age 48 years, range 24–74 years) during surgery and immediately snap-frozen in liquid nitrogen and stored at –80°C until RNA extraction. The frozen tissue samples were sectioned and mounted on glass slides. The slides were stained with hematoxylin and eosin for histopathological examinations. The tumor samples with more than 90% tumor cells and patient-matched tissue pairs with normal histological examination were included in this study. These frozen tissues were cut into 5-µM-thick sections and used for RNA isolation and cDNA synthesis. All the tumor samples had been classified as infiltrating ductal carcinoma. Tumor grade was determined according to the Bloom-Richardson score. Eight of the 23 tumors studied were grade 1 and the number of grade 2 and grade 3 tumors was 7 and 8, respectively. Eleven of the samples were estrogen receptor positive (ER+) while 10 of them were estrogen receptor negative (ER–). The number of progesterone positive (PR+) tumors was 11 and that of progesterone negative (PR–) tumors was 10.

The use of the tissue material was approved by the Research Ethics Committee of Ankara Numune Research and Teaching Hospital and consents were obtained in accordance with the Helsinki Declaration.

RNA Extraction and cDNA Synthesis

The frozen breast specimens were put into Trizol reagent (AppliChem, Darmstadt, Germany), disrupted with a homogenizer, and total RNA was isolated according to the manufacturer's instructions. Genomic DNA contaminations were removed by on-column DNaseI treatment (Macharel Nagel, Duren, Germany). The concentration of the isolated RNA and the ratio of absorbance at 260 nm to 280 nm were measured with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Montchanin, DE, USA) in triplicate. The mean $OD_{260/280}$ ratio for RNA samples was 2.03 ± 0.12 (range 1.92–2.15; $n = 46$). An aliquot of 1 µg total RNA from each sample was electrophoresed on a 1.2% RNA agarose gel to confirm integrity of the RNA. First-strand cDNA was synthesized from 1 µg total RNA using oligo(dT) or random hexamer primers by using the Revert Aid First strand cDNA synthesis kit (Fermentas, MD, USA). The random hexamer primed cDNA samples ($n = 13$, tumor and normal pair) were used for the analysis of 18S rRNA gene expression together with SDHA and ACTB in the same samples. All cDNAs were diluted 1:5 times before being used as a PCR template and stored at −20°C until further use.

Real-Time Quantitative RT-PCR

Expression levels of 18 RGs [*ACTB, GAPD, TBP, SDHA, HPRT, HMBS, B2M, PPIA, GUSB, YWHAZ, PGK1, RPL41, PUM1, RPLP0, MRPL19, TTC22, IL22RA1, ZNF224*, and the gelsolin (*GSN*) gene] were quantified with qRT-PCR by using the SYBR Green I dye detection system on the BioRad iCycler Instrument (BioRad Laboratories, Hercules, CA, USA). In order to test whether the ratio of the mRNA to rRNA was stable across tumor and matched normal samples, 18S rRNA, *ACTB*, and *SDHA* genes were quantified with qRT-PCR by using randomly primed cDNA samples. The primers were designed to include large intronic sequences between the forward and reverse pair or designed from exon–exon boundaries to avoid DNA contamination if any remained in the RNA samples. The sequences of the gene-specific primers were put into the blast search to determine their specificities. None of the primer pairs showed significant homology to other sequences in the genome but their own. The primer sequences and accession numbers of the RGs are listed in Table 1.

The amplification mixtures contained 1.0 μl of 1:5 diluted cDNA template, 6.25 μl SYBR Green PCR Master Mix Buffer (BioRad, Hercules, CA, USA), and 10 pmol forward and reverse primers in a final volume of 12.5 μl. The cycle conditions were as follows: an initial incubation of 95°C for 5 min and then 45 cycles of 95°C for 30 s and 60°C for 30 s, during which the fluorescence data were collected. Following amplification, a reaction product melt curve was obtained to provide evidence for a single reaction product. The iCycler iQ Optical System Software (version 3, BioRad Laboratories) was used to determine the melting temperatures of the products. The threshold cycle (Ct) value was calculated as the cycle where the fluorescence of the sample exceeded a threshold level. Tumor and matched normal samples were always analyzed in the same run to exclude between-run variations and each sample was studied in duplicate. The stability between duplicates was evaluated by taking the standard deviations of the average differences of all duplicate pairs (95% CI, −0.3 ± 0.8, $n = 984$). A no-template control of nuclease-free water was included in each run. The RNA samples used for cDNA synthesis were also used for (−)RT control (no reverse transcriptase enzyme) reactions. These negative RT-PCR controls were also included in the PCR reactions for each set of primers. No genomic DNA contamination was detected.

Data Retrieval and Selection of Candidate Reference Genes From Microarray Studies

Two publicly available independent microarray gene expression data sets GDS2635 (43) and GDS2250 (44) were downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and processed by the BRB-ARRAYTOOLS (Biometric Research Branch [http://linus.nci.nih.gov/BRB-ArrayTools.html]. Both of the datasets were generated by using the Affymetrix HGU133 Plus 2.0 platform; thus, they were highly comparable. These two independent microarray datasets (GDS2635 and GDS2250) were combined with respect to gene names using a set of customized Perl routines and the genes that were stably expressed between tumor and normal samples were selected by using Student's *t*-test ($p > 0.99$). A total number of 12 normal and 45 tumor samples and 54,674 gene probes were used in this analysis. *TTC22* was one of the top ranked nondifferentially expressed genes between tumor and normal samples ($p > 0.99$) and was selected as a candidate RG.

The GDS2635 dataset is the only available dataset that was generated by using matched normal breast tumor samples. Therefore, we used this set independently and determined the genes that showed no expression differences between tumors and matched normal samples by using paired Student's *t*-test ($p > 0.99$). *IL22RA1* and *ZNF224* were selected from the list as top ranked genes and used as candidate RGs ($p > 0.99$).

Data Analysis

The PCR efficiencies (E) were evaluated by 10-fold dilution series of cDNAs (1–1:100 000 dilution) for each pair of primers by using a breast carcinoma cell line cDNA pool (MCF7, MDA-MB-231, T47D, HMEC, MCF12A). The primer amplification efficiencies were also tested with reference genes *ACTB*, *GADPH*, and *SDHA* in breast tumor tissue cDNA pools ($n = 3$) to ensure no inhibitory component was present in the tissue samples. No inhibitory effect was observed in amplification efficiencies ($E = 2.0$). A graph of threshold cycle (Ct) versus relative $\log_{10}$ copy number of the calibration sample from the dilution series was produced and the reaction efficiency was determined for each primer set by using the slope of this graph ($E = 10^{(-1/\text{slope})}$) and presented at Table 1 (45). The amplification efficiency of each primer pair was corrected accordingly (2).

The gene expression level of *GSN* was normalized with respect to RGs and expressed as the ratio of ΔCts $[(E_{\text{target}})^{\Delta\text{CtTarget (control}-\text{sample)}}/(E_{\text{ref}})^{\Delta\text{CtReference (control}-\text{sample)}}]$ using the corresponding normal pair as a control (2). When the *GSN* normalization was based on multiple RGs, the geometric mean of RG Ct values was applied as NF. The

Table 1. Information on the Gene-Specific Primers and Their Real-Time PCR Efficiencies

Gene Symbol	Gene Name	Accession No./ Primer Sequence (5′–3′)	Amplicon Size (bp)	PCR Efficiency*	Exon–ExonCrossing
ACTB	Beta-actin	NM_001101	124	1.97	yes
Forward		ccaaccgcgagaagatgacc			
Reverse		ggagtccatcacgatgccag			
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	NM_002046	143	2	yes
Forward		ggctgagaacgggaagcttgtcat			
Reverse		cagccttctccatggtggtgaaga			
TBP	TATA box binding protein	NM_003194	132	1.97	yes
Forward		tgcacaggagccaagagtgaa			
Reverse		cacatcacagctccccacca			
SDHA	Succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	NM_004168	86	2	yes
Forward		tgggaacaagagggcatctg			
Reverse		ccaccactgcatcaaattcatg			
HPRT	Hypoxanthine phosphoribosyltransferase I	NM_000194	112	2	yes
Forward		gctgacctgctggattacat			
Reverse		tcccctgttgactggtcatt			
HMBS	Hydroxymethylbilane synthase	NM_000190	64	2.3	yes
Forward		ggcaatgcggctgcaa			
Reverse		gggtacccacgcgaatcac			
B2M	Beta-2-microglobin	NM_004048	132	1.9	yes
Forward		atgagtatgcctgccgtgtga			
Reverse		ggcatcttcaaacctccatg			
PPIA	Cyclophilin A	NM_021130	229	1.9	yes
Forward		cgtgtgctattagccatggt			
Reverse		ccattatggcgtgtgaagtc			
GUSB	Glucuronidase, beta	BC014142	157	1.9	yes
Forward		caccagcgtggagcaagaca			
Reverse		ggctgacacctggcacctta			
YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	NM_003406	193	2	yes
Forward		aagacggaaggtgctgagaa			
Reverse		acctcagccaagtaacggta			
PGK1	Phosphoglycerate kinase	NM_000291	195	1.9	yes
Forward		aaccagaggattaaggctgc			
Reverse		gcctacacagtccttcaaga			
GSN	Gelsolin	NM_198252	108	2.0	yes
Forward		ttcgagtcggccaccttcct			
Reverse		tctgcaccaccacctcgttg			
RPL41	Ribosomal protein L41	NM_001035267	248	2.0	yes
Forward		aagatgaggcagaggtccaa			
Reverse		tccagaatgtcacaggtcca			
PUM1	Pumilio homolog 1 (Drosophila)	NM_001020658	104	2.0	yes
Forward		ttcacagacaccacctcctt			
Reverse		ctggagcagcagagatgtat			
RPLP0	Ribosomal protein, large, P0	NM_053275	194	1.9	yes
Forward		tcatccagcaggtgttcgac			
Reverse		agacaaggccaggactcgtt			
MRPL19	Mitochondrial ribosomal protein L19	NM_014763	135	2.0	yes
Forward		tcgtgttactacagctgacc			
Reverse		atctcgacaccttgtccttc			

Table 1. Continued

Gene Symbol	Gene Name	Accession No./ Primer Sequence (5′–3′)	Amplicon Size (bp)	PCR Efficiency*	Exon–ExonCrossing
TTC22	Tetratricopeptide repeat domain 22	NM_017904	150	1.9	yes
Forward		agtgctgaagtccgaggacc			
Reverse		ttgccgaagcagtctagagg			
IL22RA1	Interleukin 22 receptor, alpha 1	NM_021258	177	1.9	yes
Forward		ccacttagagctccaggtca			
Reverse		tctggcagtgtcttcactcg			
ZNF224	Zinc finger protein 224	NM_013398	186	1.9	yes
Forward		agaacttcaggaacctgctc			
Reverse		ggaaggaccactcttgatgt			
18S rRNA	18S ribosomal RNA	NR_003286	154	2.0	no
Forward		aaacggctaccacatccaag			
Reverse		cctccaatggatcctcgtta			

*PCR efficiencies were calculated according to Rasmussen (45).

statistical analyses were performed using Minitab® software. The two-tailed paired Student's *t*-test was used when comparing tumor and matched normal expression values; and values of $p < 0.05$ with Bonferroni correction were considered statistically significant. One-way ANOVA was performed to investigate whether tumor samples, which were normalized to their matched normal counterparts, differed in terms of grade, estrogen (ER), and progesterone receptor (PR) status and the effect of the age at diagnosis was analyzed with regression analysis. Bonferroni correction was performed when multiple tests were applied.

The software geNorm^{TM}, version 3.4 (9) and Norm Finder (25), both Visual Basic Applications (VBA) for Microsoft Excel, were used to calculate the stability of candidate RGs and to find the best normalizer(s) for a given set of reference genes. Ct values were converted to linear expression quantities by $E^{-\Delta Ct}$ to investigate the genes in geNorm and NormFinder. Tissue samples were categorized into normal ($n = 23$) or tumor groups ($n = 23$) according to standard histopathological examinations for the NormFinder analysis.

RESULTS

Expression Patterns of Candidate RGs

Expression levels of 18 candidate RGs were determined in 23 breast tumor tissues and their matched normal samples by qRT-PCR using the SYBR Green I dye detection system. Amplification efficiencies calculated based on standard curves from the serial dilutions of breast cancer cell lines indicated that all primer pairs were over 90% efficient (values ranged between 1.97 and 2.3) (Table 1). Each RG had a different expression range between the tumors and matched normal samples. The RG expression levels displayed a wide range of Ct values between 13 and 33, grouped into three ranges for their mean Ct values. Highly expressed genes were *B2M, ACTB, PPIA, RPL41, RPLP0*, and *GAPDH* (mean Ct values below 20 cycles). Genes with moderate expression were *YWHAZ, PGK1, SDHA, PUM1, MRPL19*, and *GUSB* (mean Ct values between 20 and 25 cycles). Genes with low expression were *TBP, HPRT, IL22RA1, TTC22, ZNF224*, and *HMBS* (mean Ct values over 25 cycles).

The stability between duplicate measurements of each RG used in the study was very high (95% CI, –0.3 ± 0.8, $n = 984$), suggesting high experimental measurement accuracy.

RGs used in this study did not exhibit differences in terms of grade, ER, PR status ($p > 0.05$), or age ($R^2 = 0.001$ to 0.139; $p > 0.05$) in breast cancer. Furthermore, raw Ct values of the 18 RGs were found to be moderately to highly correlated with each other ($p < 0.05$, Pearson's correlation coefficient range 0.516–0.929, $n = 46$).

The reference genes used in our panel exhibited relatively higher expression in tumor samples than in their normal counterparts (paired *t*-test; $p < 0.05$). Seventeen out of 18 reference genes displayed a consistent 1.86 ± 0.7 ($\log_2$, mean ± SD) fold expression difference between breast tumor and normal pairs. The expression range of candidate genes was shown in terms of difference between the Ct values of tumor and normal samples as box-whisker-plots (Fig. 1).

Expression Stability of Candidate RGs

The expression stability of each gene was validated using two different software programs, geNorm and

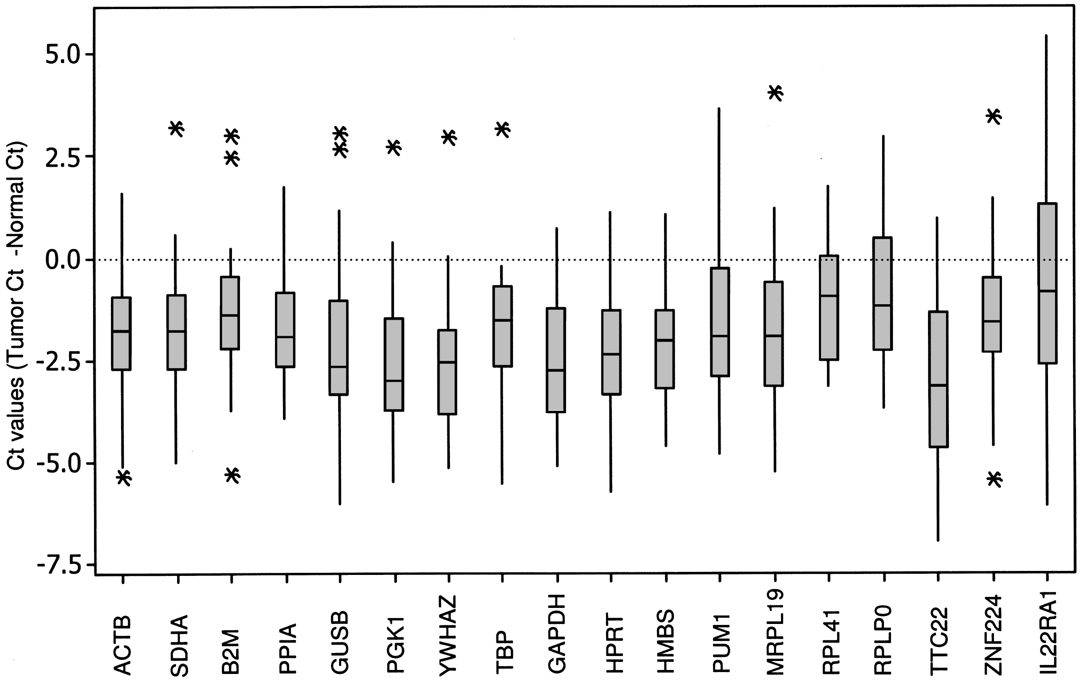


Figure 1. Expression range of differences between the Ct values of breast tumor and normal samples for each candidate reference genes. Threshold cycle values ($Ct_{Tumor}–Ct_{normal}$) for each reference gene are shown as medians (lines), 25th to 75th percentile (boxes), and range (whiskers). Whiskers illustrate the data points in Q3+1.5 (IQR) and Q1−1.5(IQR) [interquartile range (IQR) = Q3–Q1). *p*-Values were calculated using the paired Student's *t*-test ($p < 0.05$, significant). ACTB, $p = 8.7 \times 10^{-6}$; SDHA, $p = 5.7 \times 10^{-5}$; B2M, $p = 0.001$; PPIA, $p = 4.7 \times 10^{-6}$; GUSB, $p = 0.000$; PGK1, $p = 1.0 \times 10^{-6}$; YWHAZ, $p = 1.0 \times 10^{-6}$; TBP, $p = 0.000$; GAPDH, $p = 4.3 \times 10^{-8}$; HPRT, $p = 1.9 \times 10^{-7}$; HMBS, $p = 8.2 \times 10^{-5}$; PUM1, $p = 0.000$; MRPL19, $p = 0.000$; RPL41, $p = 0.005$; RPLP0, $p = 0.036$; TTC22, $p = 3.5 \times 10^{-7}$; ZNF224, $p = 0.001$; IL22RA1, $p = 0.358$. *The Ct values that fall beyond the whiskers.

NormFinder, to identify the most suitable genes for normalization.

The geNorm program determines the most stable RGs from a set of investigated genes in a given set of samples. It calculates the gene expression stability measure (M) for an RG, considering the average pair-wise variation of all other tested RGs (9). The lowest M value marks the gene(s) with the most stable expression. The average M value of the 18 candidate RGs are plotted in Figure 2A. The curve represents the stepwise exclusion of the least stable genes with higher M values. This result led to the identification of the two most stable genes, *ACTB* and *SDHA*, in the tested samples (M = 0.7).

In addition to the stability value M, pair-wise variations (V*n*/*n* + 1) were calculated to determine the effect of adding a gene (*n* + 1) in normalization (Fig. 2b). This allowed for determination of an NF needed to define the optimal number of RGs required for reliable normalization. A large pair-wise variation means that the added gene has a significant effect on normalization and should therefore be included for calculation of reliable normalization (9). The most stable six genes, *ACTB, SDHA, TBP, PGK1, GUSB*, and *MRPL19* yielded a V value of 0.147, giving the cut-off value 0.15.

We also used the NormFinder software program for stability evaluation among the candidate RGs. Norm Finder is an add-in for Microsoft Excel and is used for calculating a stability value from a set of candidate RGs. In this program, the stability value is based on the combined estimate of inter- and intragroup expression variations of the studied gene. The candidate gene with the smallest variability value has higher stability as it shows the lowest variability of inter- and intragroup expression (25). NormFinder also ranks the set of candidate RGs according to their expression stability from a panel of candidate genes that could be organized in different subgroups (tumor and matched normal tissues). Our findings indicated that the genes occupying the top five ranks, *SDHA, ACTB, MRPL19, TBP*, and *GUSB* appeared to be the most stable genes, while *IL22RA1* was defined as the least stable gene (Table 2). Although Norm Finder selected *SDHA* as the most stable gene with a stability value of 0.135, the best combination of the two

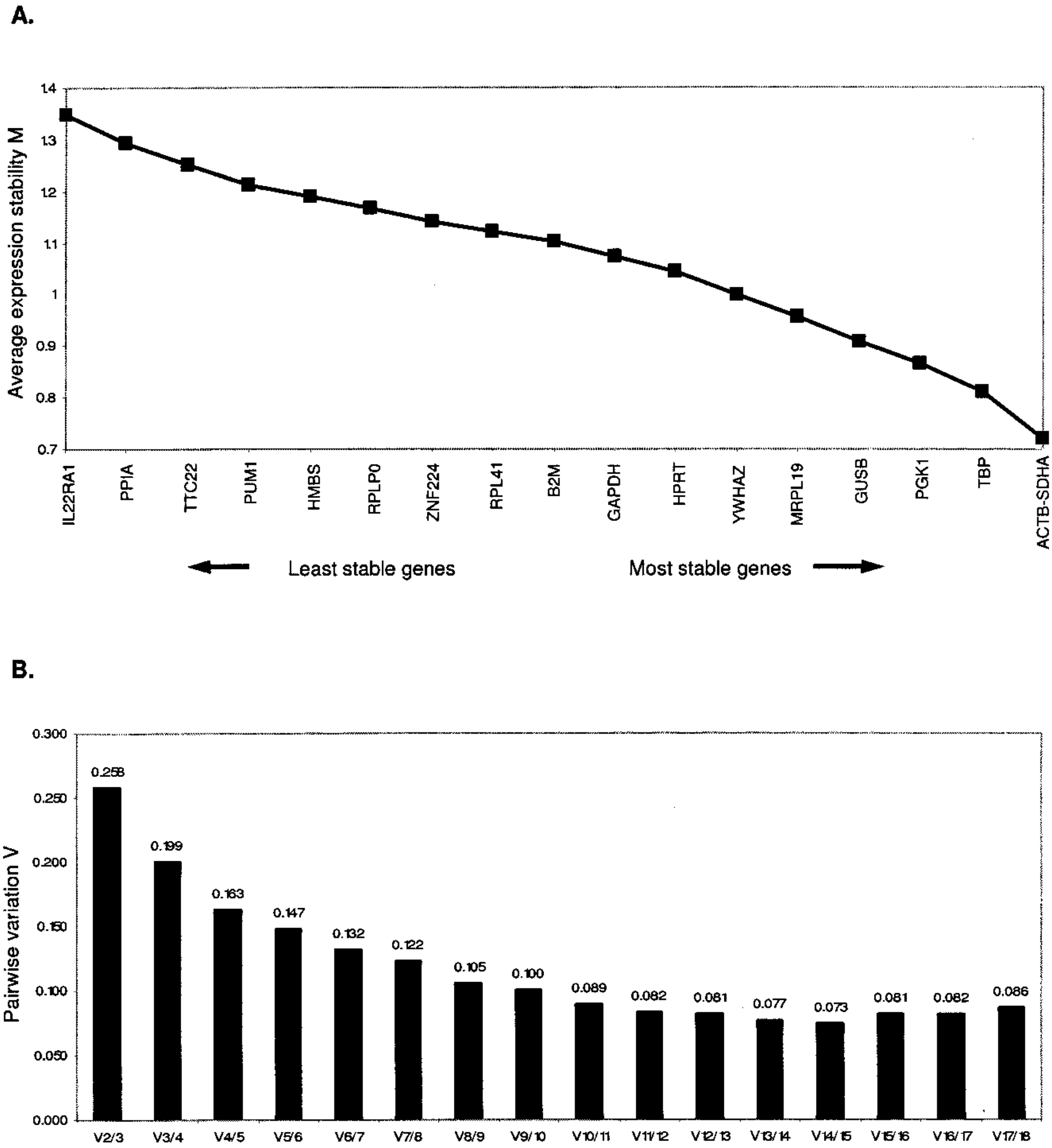


Figure 2. Selection of reference genes for normalization in breast tumor samples using geNorm analysis. (A) The curve represents the stepwise exclusion of the least stable genes according to the M values calculated by geNorm. The genes with the higher M values are eliminated and the remainders represent the two most stable genes, *SDHA* and *ACTB*. The genes are ranked on the *x*-axis from left to right according to their expression stability. (B) Determination of the optimal number of reference genes for normalization by calculation of the pair-wise variation (V) of normalization factor ratios for different numbers of control genes. Each number on the bars shows the pair-wise variation between two sequential normalization factors. On the left-most side is the pair-wise variation when the number of genes is enlarged from 2 to 3 (V2/3). Stepwise inclusion of less stable genes generates the next data points. Inclusion of the third and the fourth genes (V4/5) nears the V value to the cut-off value of 0.15.

Table 2. Rank of Candidate Reference Genes According to the Expression Stability Calculated by Normfinder

Ranking Order	Gene Name	Stability Value
1	*SDHA*	0.135
2	*ACTB*	0.155
3	*MRPL19*	0.186
4	*GUSB*	0.196
5	*TBP*	0.215
6	PUM1	0.271
7	*ZNF224*	0.289
8	*PPIA*	0.315
9	*HPRT*	0.330
10	*B2M*	0.340
11	*PGK1*	0.345
12	*YWHAZ*	0.391
13	*GAPDH*	0.404
14	*RPL41*	0.406
15	*HMBS*	0.456
16	*RPLP0*	0.478
17	*TTC22*	0.520
18	*IL22RA1*	0.574
Best two gene combination	*ACTB* and *SDHA*	**0.089**

The candidate reference genes are listed with decreasing expression stability from 1 to 18. The best combination of the two genes and the stability value were calculated by NormFinder.

genes selected by the program, *ACTB* and *SDHA*, improved the stability value to 0.089, indicating a more reliable normalization.

Assessment of Suitable RGs for Normalization

GSN is an actin depolymerizing factor acting as the principal intracellular and extracellular actin-severing protein. Expression of *GSN* was shown to be undetectable or greatly reduced in invasive human breast carcinomas both at the protein and RNA level (46). The progressive loss of *GSN* from benign mammary tissue through different stages of mammary tumorigenesis has also been demonstrated (47,48). To assess the significance of the selected RGs for normalization, the expression level of *GSN* mRNA was measured by qRT-PCR and statistically evaluated in the same set of tumor and matched normal breast tissue samples. Because a gene expression NF could either be based on a single gene or a combination of gene expression values (9), *GSN* gene expression levels were normalized using the RGs proposed by the geNorm or NormFinder calculations (i.e., *ACTB, SDHA, GUSB, MRPL19, TBP*, and *PGK1* in combinations) (Fig. 3). We also tested the performance of *IL22RA1*, the lowest ranked gene both in the geNorm and NormFinder analyses, for *GSN* normalization (Fig. 3). The median *GSN* expression values were below zero, which indicated downregulation with respect to matched normal GSN expression, independent of the NF used.

Moreover, statistical analyses indicated that the GSN expression was significantly downregulated in tumor samples when compared with that from normal samples with combinational use of the best RGs (*ACTB* and *SDHA*) proposed both by the geNorm or NormFinder programs ($p < 0.05$). In contrast, downregulation of the GSN expression was not significant when the least stable gene, *IL22RA1*, was used as NF ($p > 0.05$) with on average 39% of the tumor samples being upregulated with respect to their normal counterparts (Fig. 3). In addition, when *GSN* expression in tumors was not normalized with RGs but normalized only with the corresponding normal *GSN* expression [ΔCt; $\text{Ct}_{(\text{GSN tumor})} - \text{Ct}_{(\text{GSN normal})}$], the expression difference was not significant between tumor and normal pairs (0.18 ± 2.2, mean ± SD; $p = 0.7$, one-sample *t*-test). Fold change values in *GSN* expression obtained by using different NFs were significantly correlated with each other, yet the degree of correlation increased when two genes (in combination *ACTB* and *SDHA*) were used as NF. For example, the correlation coefficient between tumor samples' *GSN* expression values normalized with *ACTB* and those with *SDHA* ($r_{\text{A vs. S}}$) was 0.80 whereas the degree of correlation increased when a combination of best two RG was used ($r_{\text{AS vs. A}} =$ 0.95 and $r_{\text{AS vs. S}}$0.96, where A and S refer to *ACTB* and *SDHA*, respectively). The addition of the third or the fourth gene to the best two genes did not change the correlation results more than 1% ($r_{\text{AS vs. ASM}} = 0.96$ and $r_{\text{AS vs. AST}} = 0.97$, $r_{\text{AST vs. ASTP}} = 0.97$).

Evaluation of 18S rRNA to mRNA Ratio

In the present study we quantified 18S rRNA, *ACTB*, and *SDHA* mRNA levels in a group of 13 tumor and normal pairs. The mean expression of 18S rRNA was found to be downregulated in tumor samples (9/13) compared to their normal counterparts ($\log_2$ difference, 1.16 ± 1.06; mean ± SD) while the expression of *ACTB* and *SDHA* genes were consistently high in tumor samples compared to their normal pairs ($\log_2$ difference, 1.9 ± 1.4 and 1.8 ± 1.5, respectively; mean ± SD). Our results showed that the 18S rRNA to ACTB or SDHA mRNA ratio was approximately eightfold lower in tumors than that of normal pairs on average (paired *t*-test $p = 4.2 \times 10^{-5}$ and $p = 2.2 \times 10^{-4}$, respectively) (Fig. 4).

DISCUSSION

To our knowledge this is the first systematic comparison of frequently used RGs and their utility as internal controls for accurate relative gene quantification in tumor and matched normal breast tissue samples for qRT-PCR studies.

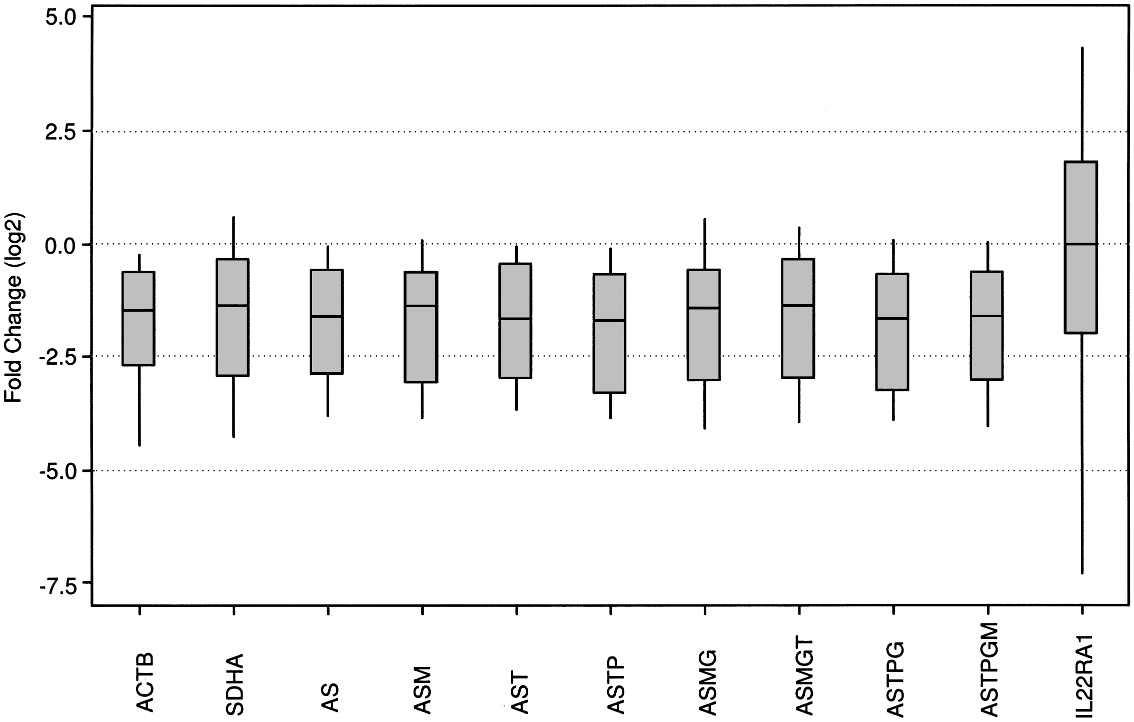


Figure 3. The normalization of *GSN* gene expression with combinations of candidate reference genes in tumor and matched normal breast samples. The gene expression level of *GSN* in 23 tumor and normal samples was normalized with respect to an individual RG or combinations of RGs and displayed as a box plot of $[(E_{\text{target}})^{\Delta\text{CtTarget (control-sample)}}/(E_{\text{ref}})^{\Delta\text{CtReference (control-sample)}}]$ using matched normal samples as controls. *ACTB* (A), *SDHA* (S), *GUSB* (G), *MRPL19* (M), *TBP* (T), and *PGK1* (P) individually or in combinations of two or more gene combinations of the above RGs are used as NFs. *GSN* normalization by the lowest ranking RG, *IL22RA1* was performed. *p*-Values were calculated using the paired Student's *t*-test ($p < 0.05$, significant). ACTB, $p = 0.003$; SDHA, $p = 0.009$; AS, $p = 0.005$; ASM, $p = 0.008$; AST, $p = 0.008$; ASTP, $p = 0.007$; ASMG, $p = 0.014$; ASMGT, $p = 0.014$; ASTPG, $p = 0.010$; ASTPGM, $p = 0.012$; IL22RA1, $p = 0.236$.

We took the following measures to increase the accuracy and reliability of our data in this study: 1) matched pairs of normal and tumor breast samples were used for minimization of inter-individual variation and to increase the power of data analysis; 2) total RNA was assessed stringently and only the high-quality samples were included in the study; 3) the 18 candidate RGs were simultaneously analyzed with optimized conditions; 4) the tumor and normal matched samples were included in the same run in duplicates for a studied gene; and 5) established software combined with statistical analysis was used to rank the candidate RGs for their suitability as NFs. Additionally, we showed that the expression of the RG set in breast tumors did not exhibit differences in terms of grade, ER, or PR status and age of the individuals when normalized to their matched controls. This is important in clinical use because the selected RGs can be used in all malignant samples independent of the tested clinical parameters.

In this study, we analyzed 15 of the commonly used RGs and 3 newly selected candidates to find out the most suitable ones as NF for relative gene quantification in paired breast tumor/normal gene expression profiling. The candidate reference genes used in this study have independent functions in cellular maintenance. This is important because the selection of genes that share identical biochemical pathways could bias analysis. To constitute the candidate reference gene panel in this study we first searched for the frequently used genes as references for qRT-PCR studies in breast cancer. While *ACTB*, *TBP*, and *GAPDH* were commonly used as normalization factor, *GUSB*, *B2M*, and *PPIA* have also been used in breast cancer studies (30–40). As a second approach we identified candidate genes, *SDHA, PGK1, HMBS, HPRT, RPL41*, and *YWHAZ*, as being used in different studies dealing with the identification of suitable reference genes for any human tissues in addition to being also recommended by geNorm. We included three more genes, *RPLP0*, *MRPL19*, and *PUM1*, in our study as they were reported to be the stable genes in breast cancers by two other studies that were investigating the endogenous control reference genes for gene ex-

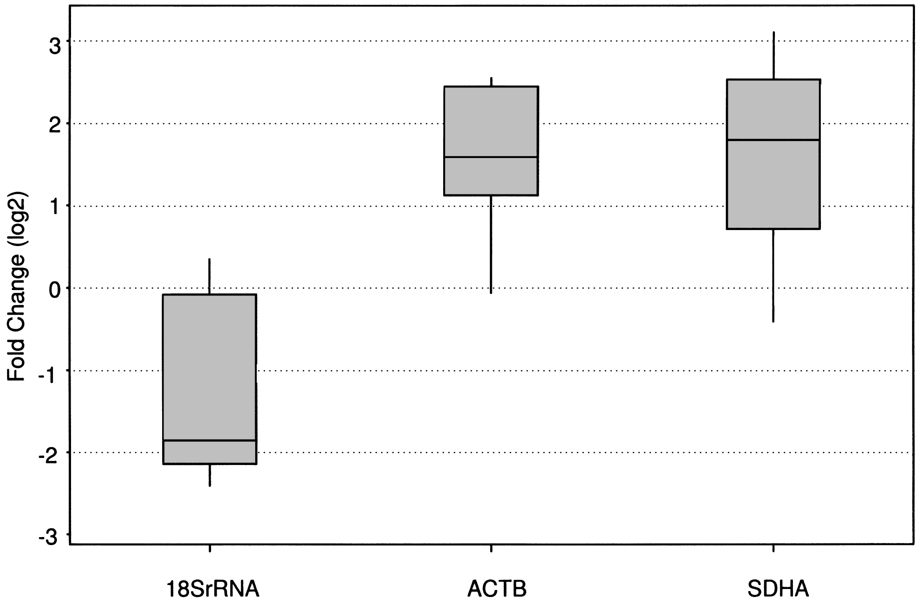


Figure 4. The expression levels of 18S rRNA, *ACTB*, and *SDHA* genes in tumor samples compared to their normal pairs. The gene expression levels of 18S rRNA, *ACTB*, and *SDHA* in 13 tumor samples were normalized with respect to that of their normal pairs [$-\Delta$Ct: $-(\text{Ct}_{(\text{Tumor})} - \text{Ct}_{(\text{Normal})})$] and displayed as box plot. The 18S rRNA to ACTB or SDHA mRNA ratio was close to eightfold lower in tumors than that of normals. The significance in this difference was calculated by paired the Student's *t*-test. *p*-Values were found to be 4.2×10^{-5} for ACTB versus 18S rRNA and 2.2×10^{-4} for SDHA versus 18S rRNA ($p < 0.05$, significant).

pression normalization in breast cancer (41,42). The genes, *TTC22, ZNF224*, and *IL22RA1*, that were selected by analyzing the publicly available breast cancer microarray data sets were also included in the panel as new candidate reference genes.

Our findings indicated that raw Ct values obtained from this RG set were highly correlated with each other, although they were not necessarily functionally related. On the other hand, the raw Ct values obtained by using a set of randomly primed cDNA samples showed that although the correlation between two RNA polymerase II transcribed genes, *ACTB* and *SDHA*, was still reserved ($r = 0.8$, $p = 0.001$), the correlation of expression from either of these two genes with the RNA polymerase I transcribed 18S rRNA gene expression was not significant ($r = 0.034$, $p = 0.912$; $r = 0.206$, $p = 0.499$). Concordant with these results, the previous studies indicated that a large number of housekeeping genes transcribed by RNA polymerase II behaved similarly among themselves (29,42), which may explain the possible reason for this correlation.

All the RGs studied here exhibited relatively higher expression in tumors than their normal counterparts. Similarly, it was reported that breast biopsy samples exhibited great intra- and interindividual variability and mean expression values of tumors measured in copy numbers were greater than those of their normal counterparts (14). Because of the extensive variability in RG expression, total RNA-based (or mRNA copy numbers when available) normalization was suggested as an NF for tumor samples (1,14). However, because total RNA is represented mostly by rRNA (>90%), even a small decrease in rRNA expression may lead to a disproportional increase in the mRNA pool estimation (49,50). Moreover, studies have shown that rDNA genes were methylated in breast and ovarian cancers when compared with those of normal controls (51,52). In fact, our finding of low tumor rRNA to mRNA ratio suggests that normal and tumor samples are heterogeneous in total RNA fractions. We found that 69% of breast tumors (9/13) exhibit dramatically lower expression of 18S rRNA compared to their nontumor pairs, while mRNA expression of widely used housekeeping genes *ACTB* and *SDHA* in the same set of tumors was higher (84%, 11/13).

These recent findings suggest that normalization based on a proper set of endogenous RGs obtained from equal amounts of total RNA/input material might be the optimal approach for comparing tumor specimens. Our findings indicated that estimation of mRNA from total RNA represented an important issue requiring further investigation in qRT-PCR studies. Because rDNA hypermethylation holds considerable possibility in breast tumors and total RNA is largely made up of rRNA, the use of poly(A)+ RNA as a starting material may be another approach for studying tumor and their matched normal samples.

In order to increase the reliability of the endogenous RG selection process, we analyzed the expression stability of the 18 selected RGs with two different statistical models: a pair-wise comparison model, geNorm, and an ANOVA-based model, NormFinder. The results obtained from the two programs were consistent for the most and least stable gene selection. *ACTB* and *SDHA* were found to be the most stable RGs while *IL22RA1* was the least stable among the 18 genes selected for these analyses.

Seventeen out of 18 reference genes in our panel displayed a consistent 1.86 ± 0.7 ($\log_2$, mean ± SD) fold expression difference between breast tumor and normal pairs, suggesting that there might be a more generalized mechanism reflected in the breast samples. One possibility is that all these genes, although with unrelated functions and chromosomal locations, are upregulated in tumors but, considering many of these genes have been reported previously as stable housekeeping genes, such global deregulation is unlikely. Alternatively, tumor and normal samples might consist of heterogeneous rRNA and mRNA compartments affecting estimation of the amount of mRNA from the total RNA pool. In support of this possibility we found that a significant portion of tumors had lower levels of 18S rRNA than normals. Furthermore, recent literature has supported our finding such that RNA hypermethylation has been shown in breast tumors (52).

Recent studies suggested that the variation in the average of multiple genes was smaller than the variation in individual genes. Therefore, it is an optimal approach to use multiple RGs rather than a single gene as NF. Normalization to geometric mean of more than one control gene compensates for outlying values of single RGs in individual samples and may therefore more accurately reflect transcript abundances of target genes (9).

Our results suggested that increasing the number of RGs stabilized the ranks of tumor samples among normalized gene expression values yet adding a third gene was not as critical as adding the second gene. This is in accordance with the findings of Vandesomple et al., who state when NF_n and NF_{n+1}, where n represents the number of genes used in normalization, do not significantly differ in their effect, using NF_n might offer a more economical choice (9). Accordingly, two best genes *ACTB* and *SDHA* can be used as NF, and additionally more genes, *MRPL19, GUSB, TBP*, and *PGK1*, identified by both programs might be combined with the two best genes to be used as NF.

In the present study, we compared the expression values of the gelsolin gene by using single or different combinations of the best ranked RGs. When the *GSN* expression was normalized with *ACTB* and *SDHA* alone, the fold change values were significantly correlated with each other, yet the degree of correlation increased when two best performing genes *ACTB* and *SDHA* were used as NF. Addition of more best performing RGs (*MRPL19, GUSB, TBP*, and *PGK1*) did not improve the degree of correlation results more than 1%.

GSN expression is known to decrease in breast tumors when compared with normal breast tissues. The adverse effect of using the least stable RG (*IL22RA1*) was highly significant, and there was a substantial error associated with the estimation of the relative *GSN* gene expression in breast tumors compared to their normal counterparts.

Considering that the housekeeping mRNA expression studied here might not actually be unregulated but overestimated due to a rRNA bias, exclusion of this bias may actually correct the potential underestimation of mRNA amount estimation between tumors and their matched normals. We calculated this possible error as 1.16 ($\log_2$ difference) for tumor–nontumor bias from the expression data obtained by using 18S rRNA from randomly primed subset of tumor–nontumor pairs. Seventeen out of 18 RGs in our panel displayed on average, a 1.86-fold expression difference between tumor and normal pairs, of which 1.16-fold might be attributable to rRNA/mRNA bias. If RG normalization is not performed, then it is likely that *GSN* expression in tumors would be overestimated at least 1.16-fold.

Real-time RT-PCR is attractive for clinical use because it can be automated and performed on a variety of tissues, fresh or archived, paired or unpaired. However, accurate quantitative analysis of gene expression levels with qRT-PCR can only be obtained by using appropriate RGs for normalization procedures. As no universal RG exists, it is inevitable to search for stably expressed genes for normalization purposes in each experimental condition, such as tumor versus normal breast specimens, to get reliable results from relative expression experiments (22,23,27).

The present study focused on identification of RGs for paired tumor/normal breast tissue based on the ranking agreement between commonly referred normalization software, geNorm and NormFinder, and expression results of *GSN*, a well-known downregulated target gene in breast tumors. Although this panel is highly comprehensive and consists of frequently used reference genes, they may still not be the best applicable reference genes for breast cancer normalization studies unless there is a bias due to RNA estimation or breast tissue heterogeneity because all the genes in our panel showed higher expression in tumors than in their normal pairs. However, *ACTB* and *SDHA* were consistently found to be the least variable genes between tumor and normal pairs with two programs, geNorm and NormFinder, in this panel.

In conclusion, our results indicated that normalization of target gene expression levels to a normalization factor consisting of the geometric mean of two best performing genes, *ACTB* and *SDHA*, offers increased accuracy and resolution in the relative quantification of gene expression in breast tumors with respect to their matched normal tissues. Future studies are needed to establish the percentage of tumors with such rRNA/mRNA bias and the underlying causes such as methylation patterns of rDNA.

ACKNOWLEDGMENTS: This work was funded by a grant from the Scientific and Technical Research Council of Turkey (SBAG-2355) and supported by Bilkent University.

REFERENCES

1. Bustin, S. A. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): Trends and problems. J. Mol. Endocrinol. 29:23–39; 2002.
2. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 29:e45; 2001.
3. Peters, I. R.; Helps, C. R.; Hall, E. J.; Day, M. J. Real-time RT-PCR: Considerations for efficient and sensitive assay design. J. Immunol. Methods 286:203–217; 2004.
4. Bustin, S. A.; Nolan, T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. J. Biomol. Tech. 15:155–166; 2004.
5. Bustin, S. A.; Benes, V.; Nolan, T.; Pfaffl, M. W. Quantitative real-time RT-PCR-a perspective. J. Mol. Endocrinol. 34:597–601; 2005.
6. Stahlberg, A.; Kubista, M.; Pfaffl, M. Comparison of reverse transcriptases in gene expression analysis. Clin. Chem. 50:1678–1680; 2004.
7. Stahlberg, A.; Hakansson, J.; Xian, X.; Semb, H.; Kubista, M. Properties of the reverse transcription reaction in mRNA quantification. Clin. Chem. 50:509–515; 2004.
8. Livak, K. J.; Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 24:402–408; 2001.
9. Vandesompele, J.; De Preter, K.; Pattyn, F.; Poppe, B.; Van Roy, N.; De Paepe, A.; Speleman, F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3:0034.1–0034.11; 2002.
10. Schmittgen, T. D.; Zakrajsek, B. A. Effect of experimental treatment on housekeeping gene expression: Validation by real-time, quantitative RT-PCR. J. Biochem. Biophys. Methods 46:69–81; 2000.
11. Zhong, H.; Simons, J. W. Direct comparison of GAPDH, β-actin, cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. Biochem. Biophys. Res. Commun. 259:523–526; 1999.
12. Barber, R. D.; Harmer, D. W.; Coleman, R. A.; Clark, B. J. GAPDH as a housekeeping gene: Analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol. Genom. 21:389–395; 2005.
13. Selvey, S.; Thompson, E. W.; Matthaei, K.; Lea, R. A.; Irving, M. G.; Griffiths, L. R. Beta-actin—an unsuitable internal control for RT-PCR. Mol. Cell. Probes 15:307–311; 2001.
14. Tricarico, C.; Pinzani, P.; Bianchi, S.; Paglierani, M.; Distante, V.; Pazzagli, M.; Bustin, S. A.; Orlando, C. Quantitative real-time reverse transcription polymerase chain reaction: Normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. Anal. Biochem. 309:293–300; 2002.
15. Ke, L. D.; Chen, Z.; Yung, W. K. A reliability test of standard-based quantitative PCR: exogenous vs endogenous standards. Mol. Cell Probes 14:127–135; 2000.
16. Valenti, M. T.; Bertoldo, F.; Dalle Carbonare, L.; Azzarello, G.; Zenari, S.; Zanatta, M.; Balducci, E.; Vinante, O.; Lo Cascio, V. The effect of bisphosphonates on gene expression: GAPDH as a housekeeping or a new target gene? BMC Cancer 6:49–55; 2006.
17. Chiu, S. T.; Hsieh, F. J.; Chen, S. W.; Chen, C. L.; Shu, H. F.; Li, H. Clinicopathologic correlation of up-regulated genes identified using cDNA microarray and real-time reverse transcription-PCR in human colorectal cancer. Cancer Epidemiol. Biomarkers Prev. 14:437–443; 2005.
18. Cerutti, J. M.; Oler, G.; Michaluart, Jr., P.; Delcelo, R.; Beaty, R. M.; Shoemaker, J.; Riggins, G. J. Molecular profiling of matched samples identifies biomarkers of papillary thyroid carcinoma lymph node metastasis. Cancer Res. 67:7885–7892; 2007.
19. Jarzabek, K.; Koda, M.; Kozlowski, L.; Mittre, H.; Sulkowski, S.; Kottler, M. L.; Wolczynski, S. Distinct mRNA, protein expression patterns and distribution of oestrogen receptors alpha and beta in human primary breast cancer: Correlation with proliferation marker Ki-67 and clinicopathological factors. Eur. J. Cancer 41:2924–2934; 2005.
20. Chao, A.; Wang, T. H.; Lee, Y. S.; Hsueh, S.; Chao, A. S.; Chang, T. C.; Kung, W. H.; Huang, S. L.; Chao, F. Y.; Wei, M. L.; Lai, C. H. Molecular characterization of adenocarcinoma and squamous carcinoma of the uterine cervix using microarray analysis of gene expression. Int. J. Cancer 119:91–98; 2006.
21. Nakamura, Y.; Tanaka, F.; Nagahara, H.; Ieta, K.; Haraguchi, N.; Mimori, K.; Sasaki, A.; Inoue, H.; Yanaga, K.; Mori, M. Opa interacting protein 5 (OIP5) is a novel cancer-testis specific gene in gastric cancer. Ann. Surg. Oncol. 14:885–892; 2007.
22. Jung, M.; Ramankulov, A.; Roigas, J.; Johannsen, M.; Ringsdorf, M.; Kristiansen, G.; Jung, K. In search of suitable reference genes for gene expression studies of human renal cell carcinoma by real-time PCR. BMC Mol. Biol. 8:47; 2007.
23. Ohl, F.; Jung, M.; Radonic, A.; Sachs, M.; Loening, S. A.; Jung, K. Identification and validation of suitable endogenous reference genes for gene expression studies of human bladder cancer. J. Urol. 175:1915–1920; 2006.
24. Ohl, F.; Jung, M.; Xu, C.; Stephan, C.; Rabien, A.; Burkhardt, M.; Nitsche, A.; Kristiansen, G.; Loening, S. A.; Radonic, A.; Jung, K. Gene expression studies in prostate cancer tissue: Which reference gene should be selected for normalization? Mol. Med. 83:1014–1024; 2005.
25. Andersen, C. L.; Jensen, J. L.; Orntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 64:5245–5250; 2004.
26. Fischer, M.; Skowron, M.; Berthold, F. Reliable transcript

quantification by real-time reverse transcriptase-polymerase chain reaction in primary neuroblastoma using normalization to averaged expression levels of the control genes HPRT1 and SDHA. J. Mol. Diagn. 7:89–96; 2005.

27. Saviozzi, S.; Cordero, F.; Lo,acono, M.; Novello, S.; Scagliotti, G. V.; Calogero, R. A. Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. BMC Cancer 6:200–209; 2006.
28. Liu, D. W.; Chen, S. T.; Liu, H. P. Choice of endogenous control for gene expression in nonsmall cell lung cancer. Eur. Respir. J. 26:1002–1008; 2005.
29. de Kok, J. B.; Roelofs, R. W.; Giesendorf, B. A.; Pennings, J. L.; Waas, E. T.; Feuth, T.; Swinkels, D. W.; Span, P. N. Normalization of gene expression measurements in tumor tissues: Comparison of 13 endogenous control genes. Lab. Invest. 8:154–159; 2005.
30. Folgueira, M. A.; Brentani, H.; Katayama, M. L.; Patrao, D. F.; Carraro, D. M.; Mourao Netto, M.; Barbosa, E. M.; Caldeira, J. R.; Abreu, A. P.; Lyra, E. C.; Kaiano, J. H.; Mota, L. D.; Campos, A. H.; Maciel, M. S.; Dellamano, M.; Caballero, O. L.; Brentani, M. M. Gene expression profiling of clinical stages II and III breast cancer. Braz. J. Med. Biol. Res. 39:1101–1113; 2006.
31. Parr, C.; Gan, C. H.; Watkins, G.; Jiang, W. G. Reduced vascular endothelial growth inhibitor (VEGI) expression is associated with poor prognosis in breast cancer patients. Angiogenesis 9:73–81; 2006.
32. Wu, G.; Xing, M.; Mambo, E.; Huang, X.; Liu, J.; Guo, Z.; Chatterjee, A.; Goldenberg, D.; Gollin, S. M.; Sukumar, S.; Trink, B.; Sidransky, D. Somatic mutation and gain of copy number of PIK3CA in human breast cancer. Breast Cancer Res. 7:609–616; 2005.
33. Shim, H.; Lau, S. K.; Devi, S.; Yoon, Y.; Cho, H. T.; Liang, Z. Lower expression of CXCR4 in lymph node metastases than in primary breast cancers: Potential regulation by ligand-dependent degradation and HIF-1alpha. Biochem. Biophys. Res. Commun. 346:252–258; 2006.
34. Morse, D. L.; Carroll, D.; Weberg, L.; Borgstrom, M. C.; Ranger-Moore, J.; Gillies, R. J. Determining suitable internal standards for mRNA quantification of increasing cancer progression in human breast cells by real-time reverse transcriptase polymerase chain reaction. Anal. Biochem. 342:69–77; 2005.
35. Kroupis, C.; Stathopoulou, A.; Zygalaki, E.; Ferekidou, L.; Talieri, M.; Lianidou, E. S. Development and applications of a real-time quantitative RT-PCR method (QRT-PCR) for BRCA1 mRNA. Clin. Biochem. 38:50–57; 2005.
36. de Cremoux, P.; Bieche, I.; Tran-Perennou, C.; Vignaud, S.; Boudou, E.; Asselain, B.; Lidereau, R.; Magdelénat, H.; Becette, V.; Sigal-Zafrani, B.; Spyratos, F. Inter-laboratory quality control for hormone-dependent gene expression in human breast tumors using real-time reverse transcription-polymerase chain reaction. Endocr. Relat. Cancer 11:489–495; 2000.
37. Potemski, P.; Pluciennik, E.; Bednarek, A. K.; Kusinska, R.; Kubiak, R.; Jesionek-Kupnicka, D.; Watala, C.; Kordek, R. Ki-67 expression in operable breast cancer: A comparative study of immunostaining and a real-time RT-PCR assay. Pathol. Res. Pract. 202:491–495; 2006.
38. Iwao, K.; Miyoshi, Y.; Egawa, C.; Ikeda, N.; Tsukamoto, F.; Noguchi, S. Quantitative analysis of estrogen receptor-alpha and -beta messenger RNA expression in breast carcinoma by real-time polymerase chain reaction. Cancer 89:1732–1738; 2000.
39. Oshiro, M. M.; Kim, C. J.; Wozniak, R. J.; Junk, D. J.; Munoz-Rodriguez, J. L.; Burr, J. A.; Fitzgerald, M.; Pawar, S. C.; Cress, A. E.; Domann, F. E.; Futscher, B. W. Epigenetic silencing of DSC3 is a common event in human breast cancer. Breast Cancer Res. 7:669–680; 2005.
40. Zhang, Z.; Yamashita, H.; Toyama, T.; Sugiura, H.; Ando, Y.; Mita, K.; Hamaguchi, M.; Hara, Y.; Kobayashi, S.; Iwase, H. Quantitation of HDAC1 mRNA expression in invasive carcinoma of the breast. Breast Cancer Res. Treat. 94:11–16; 2005.
41. McNeill, R. E.; Miller, N.; Kerin, M. J. Evaluation and validation of candidate endogenous control genes for real-time quantitative PCR studies of breast cancer. BMC Mol. Biol. 8:107; 2007.
42. Lyng, M. B.; Laenkholm, A. V.; Pallisgaard, N.; Ditzel, H. J. Identification of genes for normalization of real-time RT-PCR data in breast carcinomas. BMC Cancer 8:20; 2008.
43. Turashvili, G.; Bouchal, J.; Baumforth, K.; Wei, W.; Dziechciarkova, M.; Ehrmann, J.; Klein, J.; Fridman, E.; Skarda, J.; Srovnal, J.; Hajduch, M.; Murray, P.; Kolar, Z. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer 7:55; 2007.
44. Richardson, A. L.; Wang, Z. C.; De Nicolo, A.; Lu, X.; Brown, M.; Miron, A.; Liao, X.; Iglehart, J. D.; Livingston, D. M.; Ganesan, S. X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9:121–132; 2006.
45. Rasmussen, R. Quantification on the LightCycler. In: Meuer, S.; Witter, C.; Nakagawara, K., eds. Rapid cycle real-time PCR, methods and applications. Berlin: Springer; 2001:21–34.
46. Asch, H. L.; Head, K.; Dong, Y.; Natoli, F.; Winston, J. S.; Connolly, J. L.; Asch, B. B. Widespread loss of gelsolin in breast cancers of humans, mice, and rats. Cancer Res. 56:4841–4845; 1996.
47. Dong, Y.; Asch, H. L.; Ying, A.; Asch, B. B. Molecular mechanism of transcriptional repression of gelsolin in human breast cancer cells. Exp. Cell Res. 276:328–336; 2002.
48. Winston, J. S.; Asch, H. L.; Zhang, P. J.; Edge, S. B.; Hyland, A.; Asch, B. B. Downregulation of gelsolin correlates with the progression to breast carcinoma. Breast Cancer Res. Treat. 65:11–21; 2001.
49. Elberg, G.; Elberg, D.; Logan, C. J.; Chen, L.; Turman, M. A. Limitations of commonly used internal controls for real-time RT-PCR analysis of renal epithelial-mesenchymal cell transition. Nephron Exp. Nephrol. 102:113–122; 2006.
50. Spanakis, E. Problems related to the interpretation of autoradiographic data on gene expression using common constitutive transcripts as controls. Nucleic Acids Res. 21: 3809–3819; 1993.
51. Chan, M. W.; Wei, S. H.; Wen, P.; Wang, Z.; Matei, D. E.; Liu, J. C.; Liyanarachchi, S.; Brown, R.; Nephew, K. P.; Yan, P. S.; Huang, T. H. Hypermethylation of 18S and 28S ribosomal DNAs predicts progression-free survival in patients with ovarian cancer. Clin. Cancer Res. 11:7376–7383; 2005.
52. Yan, P. S.; Rodriguez, F. J.; Laux, D. E.; Perry, M. R.; Standiford, S. B.; Huang, T. H. Hypermethylation of ribosomal DNA in human breast carcinoma. Br. J. Cancer 82: 514–517; 2000.

doi:10.1093/bfgp/eln009

Functional genomics in translational cancer research: focus on breast cancer

Isik G. Yulug and Bala Gur-Dedeoglu

Advance Access publication date 7 March 2008

Abstract

Conventional molecular and genetic methods for studying cancer are limited to the analysis of one locus at a time. A cluster of genes that are regulated together can be identified by DNA microarray, and the functional relationships can uncover new aspects of cancer biology. Breast cancer can be used to provide a model to demonstrate the current approaches to the molecular analysis of cancer. Meta-analysis is an important tool for the identification and validation of differentially expressed genes to increase power in clinical and biological studies across different sets of data. Recently, meta-analysis approaches have been applied to large collections of microarray datasets to investigate molecular commonalities of multiple cancer types not only to find the common molecular pathways in tumour development but also to compare the individual datasets to other cancer datasets to identify new sets of genes. Several investigators agree that microarray results should be validated. One commonly used method is quantitative reverse transcription PCR (qRT-PCR) to validate the expression profiles of the target genes obtained through microarray experiments. qRT-PCR is attractive for clinical use, since it can be automated and performed on fresh or archived formalin-fixed, paraffin-embedded tissue samples. The outcome of these analyses might accelerate the application of basic research findings into daily clinical practice through translational research and may have an impact on foreseeing the clinical outcome, predicting tumour response to specific therapy, identification of new prognostic biomarkers, discovering targets for the development of novel therapies and providing further insights into tumour biology.

Keywords: *breast cancer; gene expression; microarrays; functional genomics; meta-analysis; qRT-PCR*

INTRODUCTION

Components and behaviours of biological systems can be studied using the many tools of genomics, such as SNPs [1, 2], CGH [3], SSH [4], SAGE [5, 6], proteomics [7] and siRNA technology [8]. It is widely believed that functional genomics will transform our understanding of the mechanisms underlying cellular function, and in combination with bioinformatics promises to accelerate the application of basic discoveries into clinical practice despite the natural cautions associated with the implementation of new technologies in the clinical arena.

The human genome sequence is now available and we have started to understand genomic complexity at the DNA and gene expression levels. The development of new technologies for the large-scale analysis of the genome, transcriptome, proteome and metabolome has enabled functional genomics to have a profound impact on clinical medicine [9, 10]. An astounding amount of molecular data resulting from rapid usage of these techniques have accumulated and a multitude of sophisticated methods and algorithms have been developed for comprehensive analysis of these data [11].

The application of genomics technologies to the study of cancer is rapidly shifting toward the analysis of clinically relevant samples derived from patients to discover new biomarkers for early detection of cancers. Since characteristic patterns of gene expression can be measured in parallel by using microarrays, gene expression profiling with DNA microarrays has emerged as a powerful approach to study the transcriptome of individual cancers. Molecular biologists work with clinicians and pathologists to obtain

Corresponding author. Isik G. Yulug, PhD, Faculty of Science, Department of Molecular Biology and Genetics, Bilkent University, Ankara 06800, Turkey. Tel: +90-312-290-2506; Fax: +90-312-266-5097; E-mail: yulug@fen.bilkent.edu.tr

Isik G. Yulug is an Associate Professor at Bilkent University, Department of Molecular Biology and Genetics. Her main area of research is molecular genetics of breast cancer and expression profiling.

Bala Gur-Dedeoglu is a senior PhD student and works with Dr Yulug in breast cancer gene expression profiling.

samples from patients with a known medical history, so that the molecular characteristics of samples can be correlated with the clinical presentation. This approach provides an insight into molecular mechanisms of the different cancer types, and also helps to find novel cancer biomarkers.

There are several published studies that highlight the remarkable impact of DNA microarrays on cancer research [12–18]. For example, gene expression signatures for the major cancer types and the correlation with various tumour characteristics that determine tumour grade or differentiation, metastasis and survival have been identified through these studies [19–22].

MOLECULAR PROFILING OF BREAST CANCER

Breast cancer is a major problem in developed countries and the different classifications of this disease are mostly based on clinical and pathological factors, which unfortunately fail to reflect the heterogeneity of the tumours. There are some histological markers available to decide on the prognosis and treatment of breast cancer. Estrogen receptor (ER) status, as ER-positive or ER-negative, helps to categorize breast cancers into two major classes. ERBB2 (Her-2/Neu) is also routinely used to classify breast cancer into HER-2 amplified or non-amplified categories. There are other single gene markers such as TP53, and cell proliferation markers such as Ki-67, and cyclin D1 that have emerged from detailed molecular analysis [23].

While conventional methods were restricted to studying a single locus, current high-throughput techniques have allowed monitoring gene expression or copy number levels of almost all known genes in a single experiment. Molecular profiling has been shown to be well-suited to phenotypic characterization of breast cancer and potentially to discover new molecular classes among cancers with similar histopathological appearance [24–29]. Several landmark microarray studies have demonstrated that one can build a molecular taxonomy of breast tumours using this technology and can provide a more sophisticated molecular picture together with individualized recurrence risks.

Distinguishing tumours on the basis of their gene expression profiles: impact on the future of breast cancer research

Gene expression profiling using DNA microarrays has provided an opportunity to perform more detailed and individualized breast tumour characterization leading to classification of breast cancer into distinct new molecular subgroups [30]. The potential advantages of improving tumour classification by expression profiling has been central to several large-scale breast cancer studies over the past few years that have reported identification of signature gene lists with potential for prediction of clinical outcome [24, 25, 29, 31, 32].

One of the first comprehensive studies classifying sporadic breast tumours into subtypes distinguished by differences in their expression profiles was performed by Perou *et al.* [33]. Using 40 tumours and 20 matched pairs of samples they identified an 'intrinsic gene set' of 476 cDNAs and then used this to cluster and segregate the tumours into four major subgroups: a 'luminal-like cells' group expressing ER; a 'basal-like cells' group expressing cytokeratins 5 and 17, integrin 4 and laminin, but lacking ER expression; an 'ERBB2-positive' group, and a 'normal like' epithelial group [33]. Subsequent studies confirmed that there are large-scale gene expression differences between ER-positive (mostly luminal-like) and ER-negative (mostly basal-like) cancers and suggested that further molecular subsets also exist [28, 34, 35]. The prognosis and chemotherapy sensitivity of the different subgroups are different. The luminal type cancers tend to have the most favourable long-term survival, whereas basal-like and ERBB2-positive tumours are more sensitive to chemotherapy [24, 36].

van't Veer *et al.* have used DNA microarray analysis on the primary breast tumours of 78 lymph node-negative young patients and compared the expression profiles of 34 patients who developed distant metastasis within 5 years and 44 patients who remained disease-free for at least 5 years [25]. Their analysis led to the identification of a 70-gene expression signature that was developed to classify tumours into the good and poor prognosis groups. The results were later confirmed in a larger set of tumours [26, 37]. The genes significantly up-regulated in the poor prognosis signature included those involved in cell cycle, invasion and metastasis, angiogenesis and signal transduction. This 70-gene marker set is now commercially available on the MammaPrint array (Agendia BV, Amsterdam, The Netherlands) and the prospective MINDACT clinical trial is underway to evaluate whether use of the 70-gene classifier is associated with clinical benefit.

Wang *et al*. reported a promising study showing the use of DNA microarray data for improving risk assessment for patients with lymph node-negative breast cancer. The investigators identified a diagnostic test based on expression values from a set of 76 genes. They specified 76 genes (60 genes for ER-positive, 16 for ER-negative breast tumours) that distinguished lymph node-negative patients who developed distant metastasis within 5 years [38]. The genes included in this prognostic signature belong to many functional classes, including transcriptional regulation, immune response, cell death, cell cycle, growth and proliferation, suggesting that different pathways can influence disease progression [38].

Paik *et al*. [39] used a different approach to show the clinical utility of the OncotypeDx classifier of prognosis for node-negative, ER-positive patients who received tamoxifen following local therapy for primary breast cancer. It analyses the expression of a panel of 21 genes, including ER mRNA, downstream ER-regulated genes, HER2 and proliferation-related gene expression levels, which can help in the diagnosis of ER-positive breast cancer that can be treated with tamoxifen [39].

These studies show that the molecular classification of breast cancer may have an impact on the prognosis and prediction, and provide further insights into tumour biology, providing information to both clinicians and scientists. The molecular signatures that define particular groups may lead to the discovery of new therapeutic targets and treatments that are effective in particular molecular subsets.

The power of joint analysis of microarray datasets: meta-analysis

The extensive use of DNA microarray technology in the characterization of the cell transcriptome is leading to an ever-increasing amount of microarray data from cancer studies. Different datasets for the same type of cancers are available from different microarray studies and this allows the researchers to carry out a more comprehensive analysis of their existing dataset. Besides individual microarrays, meta-analysis can be used to gather and process the datasets from multiple cancer types to investigate common molecular pathways [40–42].

Microarray datasets can be obtained from various public gene expression data repositories including the Stanford Microarray Database (SMD) [43], the National Cancer Institute's Gene Expression Omnibus (GEO) [44] and Oncomine [45]. These databases enable researchers to retrieve and perform analyses on various microarray experiments from different laboratories.

Since all cancer cells share some common characteristics such as loss of growth control, invasion and metastasis, it is very important to identify universal cancer type-independent signatures to better understand cancer pathogenesis and ultimately to improve therapeutic options. Rhodes *et al*. applied the meta-analysis approach to 21 published cancer microarray datasets, spanning 12 distinct cancer types and identified a set of 67 genes that are universally activated relative to corresponding normal tissues in most cancer types [40].

Collection of independent microarray datasets generated with the common objective of identifying differentially expressed genes in a certain type of cancer has also been performed for breast cancer [46–48]. These types of studies have resulted in the identification of gene sets with a high diagnostic value. In a microarray study with invasive ductal carcinoma samples, a reliable set of 10 genes were identified that can be used as a diagnostic tool for accurate determination of ER status and to make a decision regarding the endocrine therapeutic strategies for breast cancer. The robustness and reliability of these classifiers was confirmed after further testing on three independent microarray gene expression datasets [49].

Meta-analysis approach can provide novel candidates not present in the existing literature allowing reports of multiple genes when neither dataset can report them when analysed individually [50, 51].

Large-scale real-time quantitative reverse transcription PCR (qRT-PCR)

Microarray studies allow high-throughput analysis of expression for thousands of genes and add valuable data to tumour studies. However, once cancer target genes have been identified through this technology, validation of existing microarray data becomes inevitable.

qRT-PCR, also known as real-time PCR, plays an increasingly important role in high-throughput testing of existing microarray data [52]. qRT-PCR is an accurate and sensitive method quantifying mRNA transcripts that uses the quantitative relationship between the amount of starting target sample and the amount of PCR product at any given PCR cycle number. The method allows the detection of amplicon accumulation since it is performed using sensitive fluorogenic Taq-Man Probes, molecular beacons, and scorpions or more sensitive but less specific intercalating dyes like SYBR Green I which only fluoresce intensely when associated with double-stranded DNA [53]. The amount of fluorescence produced from the fluorogenic probes is measured at each amplification cycle. qRT-PCR has the advantages of requiring smaller quantities of sample and producing fast, accurate and easily reproducible quantitative results with little manipulation of the samples [54].

qRT-PCR is attractive for clinical use since it can be automated and performed on fresh or archived formalin-fixed, paraffin-embedded (FFPE) tissue samples [55, 56]. The biological classification formed using microarray data has been validated with freshly frozen breast tissues from multiple patient cohorts by qRT-PCR. Sorlie *et al.* [57] validated and characterized two previously defined clinically relevant subtypes of early stage breast carcinomas, luminal A and basal-like, by using three different microarray platforms. The set of 54 predictor genes identified in this study were validated by qRT-PCR using the RNA isolated from the same fresh frozen breast tumour samples that were used in microarray platforms. These genes were defined as potential prognostic molecular markers for these subtypes of breast cancer [57]. Perreard *et al.* [58] used the power of qRT-PCR to make the clinical distinction between ER-positive and ER-negative breast tumours and identified additional subtypes of breast tumours that have prognostic value. In another study, the results obtained with a 70-gene expression profile described previously in breast cancer [25] were reproduced with qRT-PCR by using a different set of frozen breast cancer samples [59]. Urban *et al.* [60] used two different microarray platforms and qRT-PCR in their study and identified the uPA gene associated with distant metastasis-free survival in ErbB2-positive breast tumours that can be used as a powerful prognostic indicator.

Recently, an exhaustive analysis of popular microarray platforms by the multi-centre consortium, the MicroArray Quality Control (MAQC) consortium, delivered reassuringly impressive results for the accuracy and reproducibility of commonly used microarray platforms [61]. The focus on their study has been the identification of common transcripts that are mutually represented among the various microarray platforms included in the analysis. Based on the MAQC dataset, Canales *et al.* [55] used three different RT-PCR methods to profile the same RNA samples to determine the concordance between the microarray-based measurements and RT-PCR results. They found that the correlation coefficients were very high for several hundred genes examined with both methods.

FFPE tissue samples are well-suited for qRT-PCR expression studies [62]. It has been shown that it is feasible to extract and purify RNA from FFPE tissue and to perform gene expression experiments although fragmentation of RNA can occur during the fixation process [56]. Retrospective clinical studies generally use FFPE tissue, as it is the most widely available material. These tissues have been used extensively and provide a major resource for understanding disease mechanisms and using the power of differentially expressed genes to evaluate possible new diagnostic or therapeutic approaches.

A diagnostic assay, OncotypeDX^{TM} (Genomic Health Inc., Redwood City, CA, USA), has been developed as an RT-PCR-based assay performed with FFPE tumour tissue. It analyses the expression of a panel of 21 genes, which can help in the diagnosis of ER-positive breast cancer [39]. The genes identifying molecular subtypes of breast cancer with prognostic significance obtained from microarrays with fresh-frozen tissues were also used to diagnose biological subtypes of breast cancer in FFPE tissues by qRT-PCR. The subtype classifications of the breast tumour with the diagnostic gene set were highly comparable between FF and FFPE tissue samples [63]. Collectively, these studies show the reproducibility of microarray data with the qRT-PCR technique.

Although limited to quantification of mRNA transcripts, the sensitivity, reproducibility, expandability and cost-effectiveness of qRT-PCR make it a benchmark technology for integration with microarray technology.

CONCLUDING REMARKS

Recent advances in genomics and genomic technologies have made it possible to study cancer in many novel ways. Microarray technology has been used to study all aspects of cancer biology that help to uncover the molecular mechanism of cancer development and has an impact on diagnosis, prognosis, drug responses and new therapeutic approaches in cancer. Another important aspect in cancer studies is establishing the epigenetic profile of a cancer type, since modification of proteins associated with chromatin and methylation of CpG sites in the DNA has a profound effect on gene expression. Such studies defining methylation signature could not only help staging of cancer cases but also help to identify the potential molecular markers for early cancer detection, assess cancer risk and improve monitoring of cancer prognosis. New genome-wide, high-throughput tools, such as Chromatin-immunoprecipitation (ChIP)-on-microarray (or ChIP–Chip) are also becoming very useful for studying epigenetic modifications in cells. ChIP combined with high-resolution microarray analysis allows the examination of genome-wide nucleosome occupancy and histone modification status [64]. Genome-wide chromatin status can then be compared with global gene expression patterns to reveal connections between specific patterns of histone modifications and the resulting gene expression in the normal or malignant phenotype of a cell. Genomic studies examining tumour sets with multiple complementary technologies, including comparative genomic hybridization (CGH), single nucleotide polymorphisms (SNPs), serial analysis of gene expression (SAGE), ChIP–chip data, proteomics and gene expression array can provide a multitude of opportunities for cancer research. The large amount of discoveries by these high-throughput techniques could then be integrated with emerging bioinformatics to increase our knowledge in cancer development. Combining the results of these multidisiplinary approaches will contribute to a better biological understanding of, and, therefore, to the improvement of the clinical management of cancer.

Key Points

- Gene expression profiling of breast tumours can distinguish subtypes of breast cancer and has an impact on the diagnosis, prognosis and treatment of the disease when combined with clinical data.
- Meta-analysis uses the power of combining different microarray datasets to identify common molecular signatures among individual or multiple cancer types.
- qRT-PCR can be used to validate gene expression profiles and performed on both fresh or FFPE tissue samples.

Acknowledgements

We would like to thank Bilkent University and the Scientific and Technical Research Council of Turkey for supporting work in our laboratory.

References

1. Zhao X, Li C, Paez JG, *et al*. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* 2004;**64**: 3060–71.
2. Bignell GR, Huang J, Greshock J, *et al*. High-resolution analysis of DNA copy number using oligonucleotidemicroarrays. *Genome Res* 2004;**14**:287–95.
3. Pinkel D, Segraves R, Sudar D, *et al*. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 1998;**20**: 207–11.
4. Atalay A, Crook T, Ozturk M, *et al*. Identification of genes induced by BRCA1 in breast cancer cells. *Biochem Biophys Res Commun* 2002;**299**:839–46.
5. Porter DA, Krop IE, Nasser S, *et al*. A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res* 2001;**61**:5697–702.
6. Allinen M, Beroukhim R, Cai L, *et al*. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 2004;**6**:17–32.
7. Kolch W, Mischak H, Pitt AR. The molecular make-up of a tumour: proteomics in cancer research. *Clinical Sci* 2005; **108**:369–83.
8. Fuchs F, Boutros M. Cellular phenotyping by RNAi. *Brief Funct Genomic Proteomic* 2006;**5**:52–6.
9. Molloy MP, Witzmann AF. Proteomics: technologies and applications. *Brief Funct Genomic Proteomic* 2002;**1**:23–39.
10. Claudino WM, Quattrone A, Biganzoli L, *et al*. Metabolomics: available results, current research projects in breast cancer, and future applications. *J Clin Oncol* 2007; **25**:2840–6.
11. Hanauer DA, Rhodes DR, Sinha-Kumar C, *et al*. Bioinformatics approaches in the study of cancer. *Curr Mol Med* 2007;**7**:133–41.
12. Dhanasekaran SM, Barrette TR, Ghosh D, *et al*. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001; **412**:822–6.
13. Chen X, Cheung ST, So S, *et al*. Gene expression patterns in human liver cancers. *Mol Biol Cell* 2002;**13**:1929–39.
14. Luo J, Duggan DJ, Chen Y, *et al*. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 2001;**61**:4683–8.
15. Iacobuzio-Donahue CA, Maitra A, Olsen M, *et al*. Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am J Pathol* 2003;**162**:1151–62.
16. Higgins JPT, Shinghal R, Gill H, *et al*. Gene expression patterns in renal cell carcinoma assessed by complementary DNA microarray. *Am J Pathol* 2003;**162**:925–32.
17. Bittner M, Meltzer P, Chen Y, *et al*. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;**406**:536–40.

18. Xu L, Tan AC, Naiman DQ, *et al*. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics* 2005;**21**:3905–11.

19. Badve S, Turbin D, Thorat MA, *et al*. FOXA1 expression in breast cancer-correlation with luminal subtype A and survival. *Clin Cancer Res* 2007;**13**:4415–21.

20. Sotiriou C, Wirapati P, Loi S, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;**98**:262–72.

21. Loi S, Haibe-Kains B, Desmedt C, *et al*. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 2007;**25**:1239–46.

22. Weigelt B, Hu Z, He X, *et al*. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* 2005;**65**: 9155–8.

23. Nielsen TO, Hsu FD, Jensen K, *et al*. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clinl Cancer Res* 2004;**10**: 5367–74.

24. Sorlie T, Perou CM, Tibshirani R, *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001;**98**: 10869–74.

25. van't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature (London)* 2002;**415**:530–6.

26. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**:1999–2009.

27. Ahr A, Karn T, Solbach C, *et al*. Identification of high risk breast-cancer patients by gene expression profiling. *Lancet* 2002;**359**:131–2.

28. Sotiriou C, Neo SY, McShane LM, *et al*. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA* 2003; **100**:10393–8.

29. Huang E, Cheng SH, Dressman H, *et al*. Gene expression predictors of breast cancer outcomes. *Lancet* 2003;**361**: 1590–6.

30. Cleator S, Ashworth A. Molecular profiling of breast cancer: clinical implications. *Br J Cancer* 2004;**90**:1120–4.

31. Gruvberger S, Ringner M, Chen Y, *et al*. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001;**61**:5979–84.

32. West M, Blanchette C, Dressman H, *et al*. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001;**98**: 11462–7.

33. Perou CM, Sorlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. *Nature* 2000;**406**:742–52.

34. Sorlie T, Tibshirani R, Parker J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003;**100**:8418–23.

35. Pusztai L, Ayers M, Stec J, *et al*. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res* 2003;**9**: 2406–15.

36. Rouzier R, Perou CM, Symmans WF, *et al*. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* 2005;**11**:5678–85.

37. Buyse M, Loi S, Van't Veer L, *et al*. Validation and clinical utility of a 70-Gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;**98**: 1183–92.

38. Wang Y, Klijn JGM, Zhang Y, *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671.

39. Paik S, Shak S, Tang G, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;**351**:2817–26.

40. Rhodes DR, Yu J, Shanker K, *et al*. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* 2004;**101**:9309–14.

41. Choi H, Shen R, Chinnaiyan AM, *et al*. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* 2007; **8**:364.

42. Xu L, Geman D, Winslow RL. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics* 2007;**8**:275.

43. Sherlock G, Hernandez-Boussard T, Kasarskis A, *et al*. The Stanford microarray database. *Nucleic Acids Res* 2001;**29**: 152–5.

44. Barrett T, Suzek TO, Troup DB, *et al*. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res* 2005;**33**:D562–6.

45. Rhodes DR, Yu J, Shanker K, *et al*. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 2004;**6**:1–6.

46. Rhodes DR, Barrette TR, Rubin MA, *et al*. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 2002;**62**:4427–33.

47. Jiang H, Deng Y, Chen HS, *et al*. Joint analysis of two microarray gene-expression data sets to select lung adeno-carcinoma marker genes. *BMC Bioinformatics* 2004;**5**:81.

48. Shen D, He J, Chang HR. In silico identification of breast cancer genes by combined multiple high throughput analyses. *Int J Mol Med* 2005;**15**:205–12.

49. Schneider J, Ruschhaupt M, Buness A, *et al*. Identification and meta-analysis of a small gene expression signature for the diagnosis of estrogen receptor status in invasive ductal breast cancer. *Int J Cancer* 2006;**119**:2974–9.

50. Choi JK, Choi JY, Kim DG, *et al*. Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett* 2004;**565**:93–100.

51. Grutzmann R, Boriss H, Ammerpohl O, *et al*. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 2005;**24**: 5079–88.

52. Bernard PS, Wittwer CT. Real-time PCR technology for cancer diagnostics. *Clin Chem* 2002;**48**:1178–85.

53. Peters IR, Helps CR, Hall EJ, *et al*. Real-time RT-PCR: considerations for efficient and sensitive assay design. *J Immunol Methods* 2004;**286**:203–17.
54. Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol* 2002;**29**:23–39.
55. Canales RD, Luo Y, Willey JC, *et al*. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* 2006;**24**:1123–31.
56. Specht K, Richter T, Muller U, *et al*. Quantitative gene expression analysis in microdissected archival tissue by real-time RT-PCR. *J Mol Med* 2000;**78**:B27.
57. Sorlie T, Wang Y, Xiao C, *et al*. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics* 2006;**7**:127–41.
58. Perreard L, Fan C, Quackenbush JF, *et al*. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res* 2006;**8**:R23.
59. Espinosa E, Vara JA, Redondo A, *et al*. Breast cancer prognosis determined by gene expression profiling: a quantitative reverse transcriptase polymerase chain reaction study. *J Clin Oncol* 2005;**23**:7278–85.
60. Urban P, Vuaroqueaux V, Labuhn M, *et al*. Increased expression of urokinase-type plasminogen activator mRNA determines adverse prognosis in ErbB2-positive primary breast cancer. *J Clin Oncol* 2006;**24**:4245–53.
61. MAQC Consortium. The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;**24**:1151–61.
62. Ma XJ, Patel R, Wang X, *et al*. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch Pathol Lab Med* 2006; **130**:465–73.
63. Mullins M, Perreard L, Quackenbush JF, *et al*. Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues. *Clin Chem* 2007;**53**:1273–9.
64. Kim TH, Ren B. Genome-wide analysis of protein-DNA interactions. *Annu Rev Genomics Hum Genet* 2006;**7**: 81–102.