FROM DATA TO BENCH TO BEDSIDE – THERAPEUTIC TARGETS IN BREAST CANCER

by

ERHAN BILAL

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Prof. Gyan Bhanot

and approved by

____

____

____

____

New Brunswick, New Jersey

October, 2010

ABSTRACT OF THE DISSERTATION

From Data to Bench to Bedside – Therapeutic Targets in Breast Cancer

By ERHAN BILAL

Dissertation Director:

Prof. Gyan Bhanot

Understanding individualized breast cancer treatment options can help physicians care for their patients by careful selection of personalized therapies. The first steps towards this goal have already been taken by clinicians, with the frequent use of molecular and genetic biomarkers to classify breast cancer into categories which direct treatment. This thesis will propose new therapeutic targets for different breast cancer subtypes, as well as a new set of biomarkers that more efficiently predict hormone resistance in estrogen positive (ER+) breast tumors. A novel methodology for therapeutic target prediction will be proposed, based on a new paradigm called "gene centrality". In addition to being over-expressed, good therapeutic targets should have a high degree of connectivity in the tumor network. Gene centrality encompasses this concept by measuring the connectivity of genes in a network in which each edge is weighted by the level of over-expression of the target gene. Using this method, a series of high centrality SRC proto-oncogenes (LYN, YES1, HCK, FYN, and LCK) were identified in subsets of Basal-like and HER2+ breast cancers. The hypothesis that YES1 is a therapeutic target in breast cancer was experimentally tested. We found that Basal-like breast tumor cell lines showed a significant decrease in fitness upon silencing the expression of YES1. Another validated therapeutic target in breast cancer is the estrogen receptor ESR1, targeted by drugs such as Tamoxifen. However, a significant fraction (~30%) of ER+ cases doesn't respond well to this therapy. A novel outlier analysis method was applied to gene expression data from ER+ breast cancer patients to identify genes highly associated with Tamoxifen resistance. These included cell cycle genes as well as several chromosomal amplification sites. In addition to the well known HER2 amplicon on 17q12, we discovered that amplicons in 8q24.3, 8p11.2 and 17q21.33-q25.1 correlate strongly with early distant metastasis and poor long term survival. As independent biomarkers for Tamoxifen resistance, together these chromosomal regions are predictive for ~75% of patients that suffer early disease relapse.

Acknowledgements and Dedication

My thanks to…

… Gyan Bhanot for his guidance and patience in pushing me to see this process through to its completion.

… Shridar Ganesan for taking the risk in allowing a student with no laboratory experience to work on important problems. It positively changed the course of my career and I learned a lot in the process.

… Andrei Ruckenstein for providing me with the opportunity to study in one of the best computational biology departments in the country.

… Gabriela Alexe a fantastic scientist, a great collaborator and friend, from whom I learned a lot.

… Ming Yao who taught me how to perform experiments without burning down the lab.

… Atul, Vasu, Honeah, Jay and Amal for putting up with me when I needed help with my lab work or just a break from thesis work.

… My mother and sister for their trust and support without which all this would not have been possible.

This thesis is dedicated to the memory of my father, Erol Bilal (1939 - 2003).

Table of Contents

List of Tables

List of Figures

Chapter 1: Breast Cancer – Overview & Facts

"Cancer is a word, not a sentence."
John Diamond

1.1 Molecular origins of breast cancer

Cancer is a generic term for a group of diseases that can affect most tissue types in the human body. It causes cells to lose their normal function and grow out of control. Most types of cancer cells will eventually grow into an abnormal mass of tissue that serves no purpose, and are named after the body part where the tumor originally formed. Eventually cells from the initial tumor site will spread to other organs where they gain the ability to form new tumors, in a process called metastasis. This, in time, leads to multiple organ failure which causes the death of the patient.

Breast cancer is produced by the accumulation of genetic or epigenetic damage in mammary cells. Normal cells go through stages in which they develop, perform their intended function and then eventually die, when they are replaced by other proliferating cells. Cancer occurs when cells escape their normal, regulated program of division and growth and begin to grow in a dysregulated manner. This escape from control can happen through a variety of pathways. One possible mechanism is a sequence of specific mutations that alter the control provided by tumor suppressor genes or proto-oncogenes that normally regulate cellular behavior. As the tumor grows, it eventually accumulates enough genetic damage to allow it to break away from the primary site and establish as a distant tumor (metastasis) in a different part of the body.

Proto-oncogenes are genes that normally function to promote differentiation and proliferation of cells in a regulated manner. A variety of proto-oncogenes exist and play key roles at crucial steps of cell growth, determining when a cell should enter cell cycle, how long it should stay there and when and under what conditions is it appropriate to allow the cell to divide. A mutation in the proto-oncogene's sequence or an increase in the amount of protein it produces (over-expression) can interfere with its normal regulatory role. This can lead to uncontrolled growth, ultimately resulting in a developing tumor. Mutated or over-expressed proto-oncogenes that cause cancer are called oncogenes and are of several types: growth factors, receptor tyrosine kinases, cytoplasmic tyrosine kinases, regulatory GTPases, and transcription factors. An example of an important oncogene in breast cancer is ERBB2 that codes for the human epidermal growth factor receptor 2 (HER2). HER2 is a cell membrane bound receptor tyrosine kinase involved in signaling pathways leading to cell growth and differentiation. Approximately 20-30% of breast tumors over-express ERBB2, leading to a flood of signals to the cell cycle pathway driving it to increase the rate of cell division. This is why HER2-positive breast cancers are aggressive and have poor prognosis if untreated.

Other oncogenes known to be involved in breast cancer include transcriptional regulators MYC, FOS, Cyclin D1 and Cyclin E, involved in cell cycle control, cyclin regulator

CDK-1, G-protein Ras, PI3K and Akt kinases, EIF-4E, an initiator of protein translation (1), and IKBKE, involved in NF-kB activation (2).

The female hormone Estrogen together with the estrogen receptor ESR1, are very important regulators of growth and differentiation in normal mammary glands. Although they are not oncogenes, and are expressed in normal breast tissue, they are important in the development and progression of breast carcinomas because of their involvement in a variety of programs which promote growth of breast tissue. It is also known that specific interactions between Cyclin D1 and estrogen directly stimulate the cell cycle (3; 4).

Tumor suppressor genes are the policemen of the cell, and their role is to prevent cells from becoming tumorigenic. In order for cancer to develop, these genes need to be silenced or their function abrogated in some way (by mutation, methylation, deletion, etc.). One of the main functions of these genes is to establish “check points” during cell cycle. This effectively pauses the cell cycle to allow a variety of cellular programs to perform various checks (such as test for DNA damage, check that the chromosomes have divided correctly in S phase, check that the spindle forms and chromosomes segregate properly in M phase etc). If some of these checks fail, repair programs are initiated and if these fail too, the cell is forced into a program of regulated suicide (apoptosis). The checks controlled by tumor suppressor genes are necessary to avoid damaged chromosomes to be passed to generations of daughter cells, as this may cause them to become cancerous. The key player in the “cell suicide” or apoptotic pathway is the P53 gene, which regulates the delicate balance between survival and death during DNA repair.

Various mechanisms inactivate P53 in approximately 40% of breast cancer cases (5) allowing damaged cells to reproduce.

A small proportion of breast cancer cases (5-10%) are related to inheritance of certain mutated genes that predispose women to cancers of the breast and ovaries. Tumor suppressor genes BRCA1 and BRCA2 are mutated in these cases leading to a disruption of the DNA repair process. Incorrect repair leads to an accumulation of errors that eventually cause cancer. Women with abnormal BRCA1 are estimated to have a 57% risk of developing breast cancer by age 70 while women with abnormal BRCA2 have a corresponding 49% risk (6).

Other tumor suppressor genes known to be involved in breast cancer include the Retino blastoma gene (Rb) which regulates progression from G1 to S phase, the pocket protein p27 which is involved in cell cycle arrest and cyclin dependent protein kinase inhibition, cell cycle checkpoint kinases CHK2 and ATM, and phosphatase PTEN, a negative regulator of AKT kinase (1).

1.2 Breast cancer classification

The breast is made up of glands for milk production, called lobules, and ducts that connect the lobules to the nipple. The remainder of the breast is made up of adipose, connective, and lymphatic tissue. Most breast cancers arise in the epithelial lining of the milk ducts. Some breast cancers are called “in situ” because they are confined, either within the ducts (ductal carcinoma in situ or DCIS) or within lobules (lobular carcinoma in situ or LCIS). Almost all breast cancers which are identified “in situ” can be effectively cured, usually by surgery and radiation alone. However, once the tumor breaks the natural barriers enclosing of the ducts or lobules and invades surrounding tissue, the tumor is called “Infiltrating Ductal Carcinoma or IDC” and is much harder to treat.

The seriousness of breast cancer depends strongly on its stage, a clinical measure assigned by pathologists and used in determining treatment. A commonly used staging system in the US has been defined by the American Joint Committee on Cancer (AJCC) and it classifies tumors based on size, whether the cancer is invasive or non-invasive, whether lymph nodes are involved, and whether the cancer has spread beyond the breast. Based on these features, stages are defined in a manner which is believed to represent disease progression. Stage 0 represents DCIS or LCIS, stage I is assigned if the tumor size is < 2 cm and there is no lymph node involvement, etc. Stage IV is the most advanced stage with lowest survival expectancy (Table 1.1 (7)) and represents breast cancers which have spread to other organs, usually the lungs, liver, bone, or brain.

A number of studies (8; 9) have shown that morphological assessment of the degree of differentiation (histologic grade) also provides useful prognostic information in breast cancer. One of the more commonly used systems for breast tumor grading is the Nottingham Grading System (10), based on a microscopic evaluation of morphologic and cytologic features of tumor cells, including degree of tubule formation, nuclear pleomorphism, and mitotic count. The sum of these scores stratifies breast tumors into grade 1 (well-differentiated, slow-growing), grade 2 (moderately differentiated), and grade 3 (poorly differentiated, highly proliferative) malignancies. There is a relationship between breast cancer stage at diagnosis and tumor grade (8). Stage and grade are somewhat correlated. Tumors assigned a higher stage generally have a larger fraction of high grade tumors than those assigned a lower stage. However, this is not a very strong correlation, and lower stage breast cancers can still be high grade. Thus, both stage and grade are considered to be independent markers of disease progression (Table 1.2 (9)) and both are used by clinicians to determine appropriate treatment.

It is well known that breast cancer is not a single disease, but instead, consists of multiple subtypes with different rates of progression and risk of long term recurrence/survival. The clinical approach to the management of breast cancer depends on their subtype classification. Approximately 60-70% of tumors express the estrogen receptor (ER), and are susceptible to treatments targeting the estrogen signaling pathway (11), such as long term treatment with Tamoxifen or aromatase inhibitors. About 15-30% of breast cancers have amplification of the human epidermal growth factor receptor-2 (HER2) and are treated by Trastuzumab (Herceptin®) and other agents that target the HER2 trans-membrane receptor tyrosine kinase (12). However, there remains significant heterogeneity in both natural history and treatment response in tumors with similar clinical classification (13; 14; 15).

High-throughput gene expression analyses have provided additional insight into this clinical heterogeneity. DNA microarrays are used to measure mRNA expression of thousands of genes simultaneously. In brief, this technology consists of an array of thousands of microscopic spots of DNA oligonucleotides, each containing small amounts of an exact sequence. Each of these probes contains a sequence specific to a single gene that is used to hybridize labeled cDNA (target) from the 3' end of the respective gene. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemi-luminescence-labeling of the target cDNA and used to determine relative abundance of nucleic acid sequences in the target. These intensity values can be obtained in a high throughput format to quantify the mRNA expression of thousands of genes.

Supervised learning methods applied to such gene expression datasets have resulted in several gene panels predictive of risk that are currently being applied to clinical practice (16; 17; 18) (for details see Chapter 3). An alternate approach to analysis of gene expression data is based on unsupervised clustering (for details see Chapter 3) (15; 19) and has successfully identified molecular subtypes of breast cancer with distinct gene expression profiles and risk for disease recurrence and survival. The overall classification that has emerged from the early studies divided breast cancers into Luminal A (ER+ with good prognosis), Luminal B (ER+ with poor prognosis), HER2+ (HER2+, ER-) and Basal-like (ER-, PR-, HER2-). The additional clinical value of this molecular classification is limited by its close correspondence to the status of biomarkers such as ER, PR, HER2 status, and tumor grade (Figure 1.1) and stage, which are routinely measured in the clinic. However, molecular classification has allowed for a deeper understanding of the biology of breast cancer from measurements of over-expressed or under-expressed genes. Gene expression analysis have shown that Luminal tumors express high amounts of luminal cytokeratins and genetic markers of luminal epithelial cells of normal breast tissue (21). In contrast, Basal-like breast cancers express high amounts of basal cytokeratins such as CK5 and a variety of growth factor receptors, including epidermal growth factor receptor (EGFR), c-kit, hepatocyte growth factor (HGF) and insulin growth factors (IGFs) (15; 17). Another feature that differentiates Basal-like tumors from the Luminal type is the dysfunction of the DNA repair mechanism, resulting in tumors with high genomic instability (22). Aberrant genomic patterns in Basal-like tumors are caused in part by the loss or dysregulation of BRCA1 and BRCA2 genes, involved in the repair of double-strand DNA breaks. Additional genomic aberrations associated with Basal-like tumors include the loss of the X chromosome inactivation marker (Xi), loss of heterozygocity and activation of both X chromosomes (23).

1.3 Treatment of breast cancer

There are almost 200,000 cases of invasive breast cancer diagnosed each year in the United States (12). Although advances in diagnosis and treatment have led to improvements in survival, over 40,000 women die each year from this disease (12). Breast cancer prognosis and treatment options are generally based on the stage and biological characteristics of the disease. Lymphovascular spread, histologic grade, ER/PR and HER2 status, as well as patient menopausal status and age are important factors in determining treatment. Table 1.3 outlines typical treatment protocols organized by stage and type. Most women with breast cancer will undergo some type of surgery depending on the size and spread of the tumor. Surgery is often combined with other treatments such as radiation therapy, chemotherapy, endocrine therapy and/or tissue-targeted therapy. Surgery and radiation are considered local therapies, while the rest (treatment with anti-cancer drugs, hormone treatments etc) are classified as systemic therapies, which are based on delivery of the drugs via the blood to all parts of the body. Systemic treatment given to patients before surgery is called neo-adjuvant therapy. It is meant to shrink the tumor enough to make surgery possible or to allow less invasive breast-conserving surgery to be performed. Systemic treatment given to patients after surgery is called adjuvant therapy. After the surgical removal of the tumor, it is important to kill local or circulating tumor cells which may cause recurrence. Systemic therapy often results in substantially decreased cancer recurrence and disease specific death. Lymph node-positive disease benefits most from systemic therapy. Metastatic breast cancers are also treated by systemic therapy, because complete removal of the disseminated tumor foci by surgery is generally not possible.

Estrogen, a hormone produced mainly in the ovaries, promotes the growth of 60-70% of breast cancers. Patients whose tumors test positive for the estrogen receptor ESR1 are administered endocrine therapies such as aromatase inhibitors (AI), selective estrogen receptor modulators (SERMs) such as Tamoxifen, or gonadotropin-releasing hormone agonists. These drugs either block estrogen or prevent estrogen production, thereby preventing stimulation of an estrogen-sensitive tumor. Tamoxifen is a highly popular SERM drug used for both premenopausal and postmenopausal women with ER+ breast cancer. Large clinical trials have shown that Tamoxifen therapy results in a 41%

reduction in annual recurrence and a 33% reduction in cancer related death in the first 5 years (24).

Aromatase inhibitors like Letrozole, Anastrozole and Exemestane are also used to treat ER+ breast cancer cases. However, since AIs work by blocking the conversion of androgens to estrogen, this class of drugs works only for postmenopausal women. Clinical trials have consistently shown that aromatase inhibitors also reduce the risk of relapse both in direct comparison with Tamoxifen or when used after completion of Tamoxifen treatment (25; 26; 27). However, none of these studies have shown an improvement in overall survival compared to Tamoxifen. In spite of this, many doctors prefer AIs over Tamoxifen as the first endocrine treatment for ER+ postmenopausal breast cancer patients, because AIs tend to be better tolerated and have fewer side effects.

Approximately 15-30% of breast cancers over-express ERBB2 (HER2+). Untreated, these cancers are aggressive and have poor prognosis. A humanized anti-ERBB2 monoclonal antibody, Trastuzumab (Herceptin), has been shown to improve recurrence and survival rates when combined with chemotherapy in HER2+ patients. Two large clinical trials showed that the risk of recurrence and death in HER2+ patients treated with Herceptin reduced by 52% and 33%, respectively, compared to chemotherapy alone (28).

1.4 New and upcoming treatment options

With the advent of new high throughput technologies like RNA-seq and next generation single molecule sequencing, targeted therapies and molecular diagnostics, breast cancer treatment has the potential to become personalized to the specifics of each tumor. Estrogen and progesterone receptor expression levels are already used to predict response to hormonal therapy with Tamoxifen or similar drugs, while ERBB2 (HER2) over-expression is used to detect HER2+ tumors that might respond to drugs that target HER2 like Trastuzumab. Since ERBB2 is over-expressed in HER2+ tumors by chromosomal amplification of the 17q12 locus, which is the location of the ERBB2 gene, patients likely to respond to Trastuzumab may be identified by assessing the level of chromosomal amplification of 17q12 by Fluorescence in Situ Hybridization (FISH) with specific probes.

Gene expression profiling has been used to develop genomic tests that may provide better predictions of clinical outcome than the traditional clinical and pathological standards. One of the main purposes is to predict response to Tamoxifen treatment. This is a well tolerated drug with low toxicity and excellent response rates of over 70% in patients with tumors expressing estrogen and progesterone receptors (29). Although most of the cases will eventually develop some form of resistance to anti-estrogen therapy, patients with bad initial response have significantly lower survival expectation. Table 1.4 lists 4 such commercially available tests that are used to predict clinical outcome for these cases. Clinicians use them to decide whether to prescribe chemotherapy in addition to Tamoxifen; i.e. wither it will benefit the patient to undergo more aggressive therapy because of the likelihood of early recurrence.

Besides FDA approved therapies like Tamoxifen and Herceptin, other drugs have shown promising results in treating breast cancer. Zoledronic acid (Zometa$^{®}$) is a bisphosphanate drug used to treat bone metastasis and osteoporosis. It appears to significantly reduce the risk of recurrence in early stage ER+ breast cancer when used in combination with hormonal therapy like Tamoxifen (30). Other drugs that have shown promising results are the so called anti-angiogenic drugs that work by blocking blood supply to the tumor. Preclinical studies showed that when used in combination with a chemotherapeutic agent at lower doses, it slows disease progression in patients with metastatic breast cancer (31; 32).

A promising new class of drugs, called PARP (Poly ADP -ribose polymerase) inhibitors, appears to be effective in treating breast cancers that have inactivating mutations in BRCA1 or BRCA2 genes. BRCA1 and BRCA2 are involved in DNA repair in complementary pathways. When both pathways are compromised, tumor cells become more sensitive to DNA damage induced by chemotherapy and radio therapy (33). A number of PARP inhibitors are currently in clinical testing.

Table 1.1: Breast cancer five-year survival by stage at diagnosis

Five-year survival rates for patients with different diagnosed breast cancer stage. Data for these statistics were collected through 2006 and reported using classifications of situ, localized regional, and distant.

Cancer stage	Classification	Five-year survival rate
0	In situ	100%
I & II	Early invasive	98% (local); 83.6% (regional)
III	Locally advanced	57%
IV	Metastatic	23.4%

Table 1.2: Breast cancer five-year survival by histologic grade at diagnosis

Five-year survival rates for patients with different histologic tumor grades treated with surgery and radiation alone. Data for these statistics were collected from 1977 to 1986.

Tumor grade	Classification	Five-year survival rate
1	Well-differentiated breast cells	93%
2	Moderately-differentiated breast cells	82%
3	Poorly differentiated breast cells	65%

Table 1.3: Typical treatment options for breast cancer by stage

The last three columns list adjuvant therapy options. Table adapted with permission from Maughan et al. (7).

Cancer stage and type	**Primary treatment**	**Hormone receptor negative**	**Hormone receptor positive**	**HER2 over-expression**
Stage 0: in situ Lobular carcinoma	No treatment or consider prophylaxis with Tamoxifen	—	—	—
Ductal Carcinoma in situ	Breast-conserving surgery and radiation therapy	—	—	—
Stages I & II: early stage invasive	Breast-conserving surgery and radiation therapy	Chemotherapy	Chemotherapy and endocrine therapy	Chemotherapy and Trastuzumab (Herceptin)
Stage III: locally advanced Noninflammatory	Induction chemotherapy, followed by breast-conserving surgery and radiation therapy	Induction chemotherapy	Induction chemotherapy and postoperative endocrine therapy	Induction chemotherapy and postoperative Trastuzumab
Inflammatory	Induction chemotherapy, followed by mastectomy and radiation therapy			
Stage IV: metastatic	Radiation therapy or bisphosphonates for bone pain	Chemotherapy	Endocrine therapy with or without chemotherapy	Trastuzumab with or without chemotherapy
Recurrent: Local after breast conserving surgery	Mastectomy	Chemotherapy	Chemotherapy and endocrine therapy	Chemotherapy and Trastuzumab
Local after mastectomy	Wide excision			
Local inoperable	Induction chemotherapy			

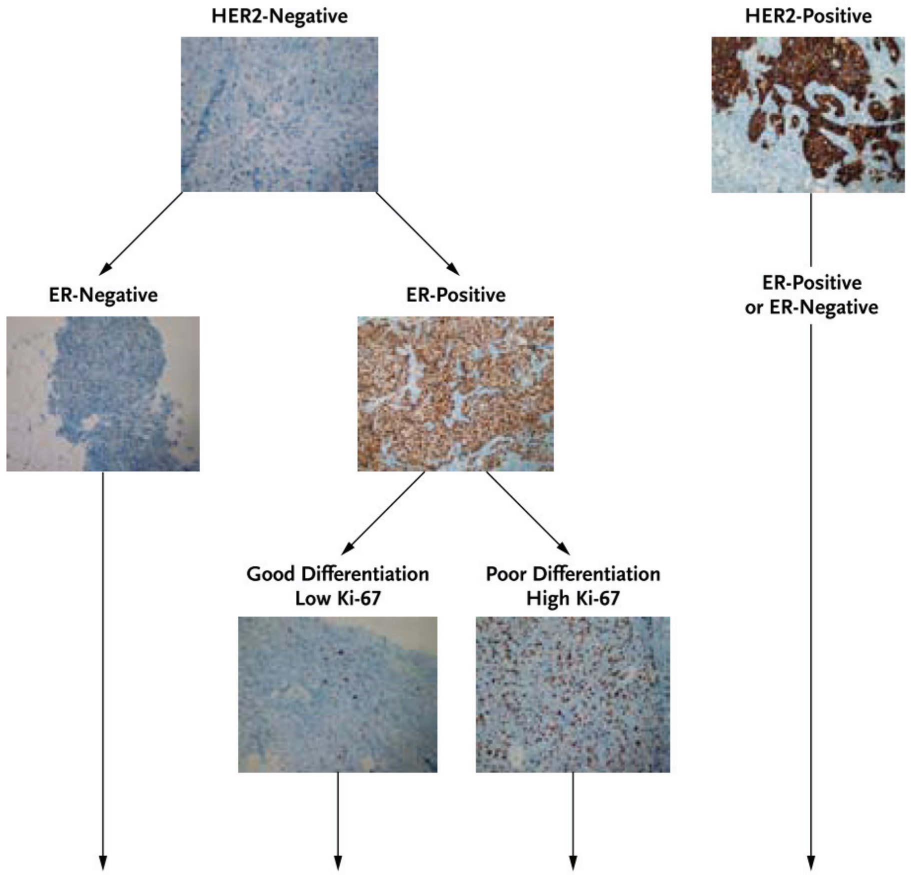
Table 1.4: Commercially available genomic assays for the prediction of clinical outcome in patients with breast cancer

ER denotes estrogen receptor, FDA stands for Food and Drug Administration, and Q-RT-PCR represents the quantitative reverse-transcriptase-polymerase chain reaction. Table adapted with permission from Sotiriou et al. (34).

	MammaPrint	**Oncotype DX**	**Theros**	**MapQuant DX**
Provider	Agendia	Genomic Health	Biotheranostics	Ipsogen
Type of assay	70-gene assay	21-gene recurrence score	2-gene ratio of HOXB13 to IL17R and molecular grade index	Genomic grade
Type of tissue sample	Fresh or frozen	Formalin-fixed, paraffin-embedded	Formalin-fixed, paraffin-embedded	Fresh or frozen
Technique	DNA microarrays	Q-RT-PCR	Q-RT-PCR	DNA microarrays
Indication	To aid in prognostic prediction in patients <61 yr of age with stage I or II, node-negative disease with a tumor size < 5 cm	To predict the risk of recurrence in patients with ER+, node-negative disease treated with Tamoxifen; to identify patients with low risk of recurrence who may not need adjuvant chemotherapy	To stratify ER+ patients into groups with a predicted low risk of recurrence and a predicted good or poor response to endocrine therapy	To stratify grade 2 tumors into low-risk grade 1 or high-risk grade 3 tumors, specifically for invasive, primary, ER+ grade 2 tumors
FDA clearance	Yes	No	No	No
Availability	Europe and US	Europe and US	United States	Europe

Figure 1.1: Correspondence between molecular class and clinical features of breast cancer

ER denotes estrogen receptor and HER2 the human epidermal receptor 2. Ki-67 is a nuclear antigen and a marker for cell cycle senescence. Figure reproduced with permission from Sotiriou et al. (34).



Pathological Variables	Basal-like (%)	Luminal A (%)	Luminal B (%)	HER2-like (%)
HER2-positive (IHC)	10	12	20	100
ER-positive (IHC)	12	96	97	46
Grade III	84	19	53	74
Tumor size >2 cm	75	53	69	74
Node-positive	40	52	65	66

Chapter 2: Therapeutic Targets in Breast Cancer

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'."
Isaac Asimov (1920-1992)

2.1 A novel paradigm for therapeutic target identification

The success of Trastuzumab (Herceptin) in treating certain types of breast cancer has heralded a rush for the next targeted therapy in breast cancer. This chapter will present a novel method for estimating therapeutic targets in different classes of breast tumors. Since good therapies such as Tamoxifen and Herceptin, already exist for ER+/PR+ and HER2+ breast cancers, we will first focus on ER-/PR-/HER2- cases and propose new targeted therapies for this type of cancer.

Basal-like breast cancers (BLC) are high grade, invasive tumors characterized by the "triple negative" phenotype, lacking expression of the estrogen receptor (ER), progesterone receptor (PR), and HER2 and hence has no obvious target pathway for adjuvant therapy. BLC account for approximately 15% of breast cancers, tend to occur in younger women and account for a disproportionate amount of breast cancer deaths (35). BLC do not respond to Tamoxifen or Herceptin and there are at present no targeted therapies for their treatment.

Our goal is to identify specific therapeutic targets for BLC. Towards this end, we hypothesize that good targets should be highly expressed proteins that are “important” for the survival of the tumor cell, meaning that down-regulation of the associated gene would lead to a significant impact on the fitness of the cancer cells. Intuitively, important genes are correlated with a large number of other genes in their expression values across multiple samples, acting like hubs (high degree nodes) in the associated gene network. Our assumption is that identifying these “hubs” may lead to appropriate targets for therapy.

Normal cellular behavior is a complex, regulated network of interaction between genes, proteins, transcription factors, microRNA etc. Tumor cells modify this network to allow them to proliferate and avoid detection and apoptosis. This is achieved by altering specific genes to enable them to avoid/ignore apoptotic pathways, proliferate, elicit blood supply, migrate to other tissue and reestablish there as a metastatic tumor. Targeted cancer therapies aim to neutralize specific proteins necessary for the tumor cell to remain viable in-vivo. Ideally, the proteins targeted should be such that their down-regulation has a major impact on the survival/fitness of the tumor cells and, at the same time, has a smaller effect on normal cells. Gene or protein expression levels are not sufficient to identify these targets because the level thresholds for tumorigenic behavior may be different for different genes/proteins and different for each individual. We suggest a novel algorithm and methodology to identify therapeutic targets by using a technique based on a new paradigm which we call *gene centrality*.

The identification of a target gene as one with high centrality is based on our expectation that in addition to being over-expressed, good therapeutic targets must have a high degree of connectivity in the tumor gene network. We identify such genes by computing the eigenvector associated with the largest eigenvalue of a modified gene connectivity network matrix in which each edge is weighted by the over-expression level of the target gene, as shown in Figure 2.1. Genes with high centrality are those with high coefficients in the first eigenvector. We expect such genes to be better therapeutic targets, because their modification would affect a relatively large number of other "important" genes. Simulations on synthetic gene networks (36) show that knocking out such highly connected genes (i.e. genes linked to many other genes) yields a lower fitness compared to knocking out a gene with fewer connections, and this effect is even stronger for genes with high expression values.

We applied this method to two published breast cancer gene expression datasets. Tumors were classified as Luminal, HER2+ and Basal-like based on clinical information on ER, PR and HER2 biomarkers. Molecular subtypes within these classes were identified using consensus clustering and centroid based classification (13). Potential therapeutic targets within each subtype were identified using network analysis as described above. This analysis identified a number of SRC tyrosine kinases LYN, YES1, HCK, FYN, LCK with high centrality scores in subsets of Basal-like breast cancers and HER2+ tumors. Their importance was verified with a growth/survival assay by stably suppressing the expression of YES1 in several breast cancer cell lines. This analysis showed that down-regulation of YES1 has a significant effect on the fitness of cancer cells. It also suggests that several existing drugs, such as SRC inhibitors, might be successfully used in treating an identifiable subset of BLCs.

2.2 Identification of candidate therapeutic targets in breast cancer subtypes

We analyzed gene expression data from Wang et al. (37), consisting of 286 early stage, lymph node negative breast tumor samples from patients treated with surgery and radiation but no adjuvant or neo-adjuvant therapy. Long term recurrence/survival data was available for all patients. Robust unsupervised consensus clustering had previously split this dataset into six core breast cancer subtypes (13; 38; 39), two within each clinical class. The Luminal (ER+) cases split into 28 Luminal A (LA) and 104 Luminal B (LB) samples, HER2+ (HER2+,ER-) cases split into 14 HER2I and 17 HER2NI, while Basal-like (HER2-,ER-,PR-) cases split into 15 BA1 and 22 BA2 samples. LA and LB tumors were both positive for estrogen and progesterone, the main difference between them being that LB cases had a significantly higher recurrence rate. HER2I and HER2NI breast cancers both had amplification of the ERBB2 (HER2) chromosomal region 17q12, but HER2I tumors had a significantly lower recurrence rate correlated with high expression of many lymphocyte associated genes (13). Compared to the BA2 subtype, the BA1 cases were characterized by over-expression of genes associated with the innate immune/defense response pathway. Chapter 3.1 provides a more in depth description of the clustering method used and these breast cancer subtypes.

We also analyzed a second gene expression dataset from Ivshina et al. (40), consisting of 249 samples from primary invasive breast tumors. Samples were classified into subtypes using the core clusters already identified in the Wang et al. (37) dataset by comparing gene expression values for each sample to mean expression values calculated for each of the original core clusters. Centroids for each subtype were identified using normalized gene expression values as described in (13) and distances from the centroids to samples from the new dataset were calculated using several distance metrics (such as Pearson correlation and Euclidean distance). For each distance metric used, the new samples were assigned to the subtype whose centroid they were closest to. Samples that did not consistently classify with the same subtype for all distance measures were discarded. We thus identified 78 LA, 96 LB, 12 HER2I, 24 HER2NI, 11 BA1 and 13 BA2 tumors.

Both gene expression datasets were obtained from the Gene Expression Omnibus (GEO:www.ncbi.nlm.nih.gov/geo) database with accession identifiers GSE2034 and GSE4992 for the first (Wang et al. (37)) and the second (Ivshina et al. (40)) dataset respectively. Table 2.1 summarizes the clinical and pathological characteristics of all patients used in the study. The main difference between the two datasets is in the distribution of lymph node (LN) status and histologic grade, but this does not adversely affect our analysis, because it depends mostly on the subtype assignments, which are also given in Table 2.1.

Outlier scores (θ) and Pearson correlation values (r) were calculated for each gene across all samples. The outlier score is a measure of the relative over-expression of a gene in one subtype compared to all others. It is defined as the percentage of tumor samples that over-express a particular gene. To make the score robust, the outlier score for each subtype was defined as the mean over the distribution of outlier scores across sample bootstrap datasets. To reduce sample size bias, each bootstrap dataset was chosen by random sampling of an equal number of samples from each subtype. Outlier score values were determined separately for the two datasets (GSE2034 and GSE4992) and then merged into one meta-outlier score ($\hat{\theta}$) by taking a weighted mean of the individual outlier scores over bootstrap datasets. Similarly, Pearson correlation values between gene pairs were calculated for each of the six tumor classes for both datasets and then merged into meta-correlation values ($\hat{r}$). Correlations that were significantly different between the two datasets were discarded. Next, centrality scores were calculated for a gene network in which an edge from gene g_i to gene g_j is equal to $\hat{\theta}_j \hat{r}_{ij}^2$. A more detailed analysis in Appendix A shows that the coefficients of the first eigenvector of the corresponding adjacency matrix represent the desired measure of gene centrality.

A more intuitive explanation of the centrality measure would be an analogy to the US highway system. Large cities are connected by a big number of highways as opposed to small towns. Highways also tend to become wider as they approach a large city, to accommodate more traffic. If someone starts driving randomly across the country, they will inevitably end up in a major city, hence the saying *“All roads lead to Rome!”*. One way to identify these hub cities is to calculate the first eigenvector of the highway network, where larger coefficients will correspond to hubs. Similarly, in our gene network, edges connecting over-expressed genes have a higher weight, and if enough connections are present, they will also act like hubs, with high centrality scores.

High correlation scores are transitive (linked across several genes) and could identify cliques of over-expressed genes with similar centrality scores. To find the genes most likely to cause a phenotypic change upon knock-down, we pruned the genes with high centrality scores in each subtype to known oncogenes. These are presented in Table 2.2 along with the associated centrality and outlier scores calculated by meta-analysis over GSE2034 and GSE4992 gene expression tables.

2.3 YES1 is a therapeutic target in basal-like breast cancers

As seen in Table 2.2, our method successfully identified epidermal growth factor receptor ERBB2 as a central gene for HER2+ (HER2I, HER2NI) subtypes and estrogen receptor ESR1 for Luminal subtypes (LA, LB). In addition, most strikingly, we identified a number of SRC protein kinases LYN, YES1, HCK, FYN, and LCK with high centrality scores in either the BA1 subset of basal-like breast cancer and/or the HER2I subset of HER2+ breast cancers, suggesting that they may be potential therapeutic targets for patients in these subtypes. Here we focus on the YES1 (Yamaguchi sarcoma viral oncogene homolog 1) gene, which is also known to be over-expressed in colorectal, head and neck, renal, lung and stomach cancers (41). Figure 2.2A and B show the normalized expression values of YES1 across all subtypes for the two data sets. To avoid sampling bias (due to unequal number of samples in the subtypes), we used the following procedure: Ten samples were chosen from each subtype, the expression value of YES1 was standard normalized across these sixty samples and the normalized value of each sample in this bootstrap dataset was noted. This procedure was repeated 1000 times. The average expression value of YES1 for each sample across these bootstrapped datasets was calculated, keeping track of how often the sample appeared in the bootstrap samplings. Figure 2.2A and B shows the sorted values of YES1 for all samples in each subtype thus obtained for the two datasets. The relative over-expression of YES1 uniquely in the basal-like subtypes is obvious from Figure 2.2A and B.

To further validate YES1 over-expression in a subset of basal-like breast cancers, we analyzed FFPE slides from 13 ER-/PR-/HER2- breast cancer patients. These slides were obtained under an IRB approved protocol from the Tumor Bank at the Cancer Institute of New Jersey. Immunohistochemical analysis of the slides was performed using a YES1-specific antibody and scored as described in Appendix B.1. Figure 2.2C, D and E show staining of the samples identified as having high, medium or low/no expression of YES1 respectively. Of the 13 samples, 2 showed high levels of YES1, 6 had medium expression and 5 had low/no expression.

The analysis described above identified YES1 as a potential therapeutic target in the BA1 subtype of breast cancer. We tested this possibility in-vitro by studying whether suppressing YES1 expression in subsets of breast cancer cell lines has a significant effect on their fitness, as measured by a survival/growth assay (Appendix B.2-5). Appropriate shRNA were purchased and lentiviral vectors constructed to stably suppress the expression of YES1 in breast cancer cell lines: MDA468, MDA231, BT549, MCF10A, SKBR3 and MCF7. Of these, MDA468, MDA231, BT549 and MCF10A are all Basal-like, that is, they are negative for expression of estrogen, progesterone and HER2 proteins (ER-/PR-/HER2-). SKBR3 is ER-/PR- but weakly HER2+ while MCF7 is ER+/PR+, and consistent with the Luminal breast cancer type. Three different shRNA were chosen and their ability to suppress the expression of YES1 was tested on MDA468. Only the most efficient one (shYES1#2) was selected for subsequent experiments. Equal numbers of cells from each cell line infected with either a lentivirus encoding shYES1 or a control scrambled shRNA. After 6 days all cells were counted and the results compared to the controls to assess whether the growth rate of cancer cells was affected by silencing of YES1. We found (see Figure 2.3A, B and D), that all cell lines except the Luminal cell line MCF7 had a significant reduction of cell counts when treated with shYES1#2 compared to the controls. Two additional shRNAs that were less efficient in knocking down YES1 protein levels (shYES1#1 and shYES1#3), could also decrease cell growth in his assay, although not as efficiently as shYES1#2, demonstrating the effect on growth is not likely an off target effect of shYES1#2 (Figure 2.3C).

Our method also successfully identified previously known therapeutic targets like ERBB2 (HER2/neu) and ESR1 that have already led to the development of drugs such as Herceptin® for HER2+ or Tamoxifen® for Luminal (ER+) breast cancers. Interestingly, Tamoxifen treatment is less successful in the case of Luminal B (LB) subtype (15) comparing to Luminal A (LA). The centrality scores of ESR1 are 6.94 in LA and only 3.44 in LB, even though ESR1 is equally over-expressed in both cases. This suggests that perhaps in Luminal B patients, the tumor is not as “addicted” to ESR1 as in Luminal A, and hence does not respond as well to therapy which blocks estrogen.

2.4 Treatment implications for breast cancer

In this chapter we have presented a novel method for analysis of gene expression data that takes into account not only the levels of expression for genes but also a measure of co-relatedness between pairs of genes across multiple samples. The algorithm implemented here, based on outlier scores and correlations, is general enough that can be modified to use different measures of expression, as long as they are positive valued. One such method is the soft-max normalization procedure described in (42). Other estimations of correlation can also be used in place of the Pearson correlation (for example: Spearman rank correlation (43), Kendall tau rank correlation (44) or mutual information (45)).

The rest of the potential therapeutic targets associated with breast cancer subtypes in Table 2.2, are either new or are currently being tested in clinical trials (Clinical Trials: www.clinicaltrials.gov). These targets are: the epidermal growth factor receptor EGFR for high risk ER+ tumors (Luminal B), FOS, TGF beta receptor 2, ETS-related genes ERG, ELK3 and ETS2 for Luminal A tumors, PIM2 and a number of SRC tyrosine kinases predicted to be good therapeutic targets in subsets of Basal-like and HER2+ breast tumors. Among drugs being tested in clinical trials on breast cancer patients are Gefitinib, Cetuximab that target EGRF and Dasatnib that targets SRC kinases.

Our in-vitro confirmation of YES1 as a therapeutic target in a set of Basal-like breast cancers opens the way to new targeted treatments involving SRC kinase inhibitors like Dasatnib® (Bristol-Myers Squibb), AP 23846 (Ariad), TG 100598 (TargeGen), AZD 0539 (AstraZeneca) or SKI-606 (Wyeth). Dasatnib is a drug that inhibits the BCR/ABL pathway in addition to SCR kinases, and has been shown to slow the growth of triple negative (ER-/PR-/HER2-) breast cancer cell lines in vitro (46; 47). It is unclear whether this result is due to the inhibition of BCR/ABL or any of the SRC kinases, but based on centrality scores and subsequent experiments, YES1 is at least partially involved in the phenotypic changes observed upon Dasatnib treatment.

Table 2.1: Microarray datasets used in this study

Clinical and pathological characteristics of all patients, as well as clustering and classification results. Unknown values are not counted.

GEO acc.	No. of samples	Grade ratio (1/2/3)	LN status ratio (+/-)	ER status ratio (+/-)	Luminal class ratio (LA/LB)	HER2+ class ratio (HER2I/HER2NI)	Basal-like class ratio (BA1/BA2)
GSE2034	286	7/42/148	0/286	209/77	28/104	14/17	15/22
GSE4922	249	68/126/55	81/159	211/34	77/96	12/24	11/13

Table 2.2: Top centrality results for cancer genes

Top gene centralities and meta-outlier scores are listed for oncogenes for each breast cancer subtype. High centrality scores are highlighted in red.

The genes corresponding to these high centrality scores are potential drug targets because they are both over-expressed and highly connected, suggesting that the tumor is addicted to them (needs them in an essential way for growth and proliferation).

	BA1		BA2		HER2I		HER2NI		LA		LB	
Gene	Centrality	Outlier score	Centrality	Outlier score	Centrality	Outlier score	Centrality	Outlier score	Centrality	Outlier score	Centrality	Outlier score
LYN	4.35	80%	1.89	38%	3.32	29%	0.21	5%	0.00	0%	0.00	0%
YES1	3.66	70%	1.84	52%	0.00	0%	1.22	24%	0.00	0%	0.23	6%
HCK	3.85	63%	0.38	10%	4.38	47%	0.21	6%	0.57	6%	0.33	8%
FYN	2.42	41%	0.94	32%	7.60	55%	0.37	7%	1.65	13%	0.44	8%
LCK	3.08	52%	0.50	15%	12.01	88%	0.00	0%	0.93	10%	0.41	8%
PIM2	4.12	65%	0.29	10%	5.88	79%	0.00	0%	0.61	9%	0.43	13%
ERBB2	0.00	0%	0.00	0%	6.51	100%	4.51	100%	0.01	0%	0.05	2%
TGFBR2	0.04	1%	0.71	9%	3.32	41%	0.76	12%	13.61	66%	0.46	9%
ERG	0.00	0%	0.71	11%	1.72	21%	2.04	31%	10.57	65%	1.21	26%
ELK3	0.71	13%	1.14	18%	1.16	15%	1.36	23%	6.50	50%	0.72	16%
FOS	0.00	0%	0.10	2%	1.50	28%	0.94	20%	5.76	76%	0.77	34%
ETS2	0.47	11%	1.60	33%	2.39	27%	0.70	19%	5.92	34%	0.50	11%
ESR1	0.00	0%	0.00	0%	0.78	13%	1.54	26%	6.94	69%	3.44	82%
EGFR	0.77	11%	2.36	38%	1.24	18%	1.38	25%	1.57	19%	4.99	40%

Figure 2.1: Example of a gene network with a high centrality gene.

The figure shows an example of a section of the cellular network. The size of the circle representing the node for gene g_i is proportional to the relative expression of the gene. Links between genes represent associations – and transform the network into an adjacency matrix which can be made primitive by eliminating unconnected genes and adding self-loops to all nodes. Undirected edges may be changed into directed edges using relative associated weights which equal the expression level of the target gene. In the configuration shown, the center node (gene) coloured in red would have the highest centrality score because of its high expression and connectivity.

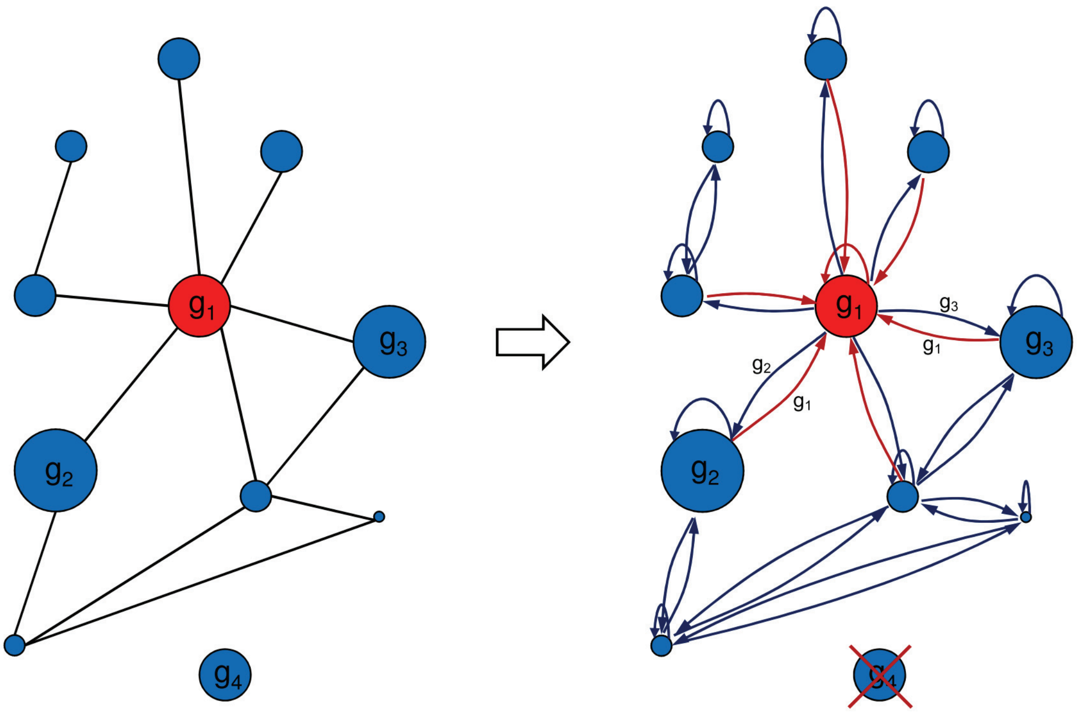


Figure 2.2: YES1 is expressed in a subset of basal-like breast tumors

Bar plots showing relative over-expression of YES1 in a subset of basal-like breast tumors in the GSE2034 (A) and GSE4922 (B) gene expression datasets. To confirm this, 13 ER-/PR-/HER2- paraffin embedded breast cancer tissue slides were probed for expression of YES1 by immunohistochemistry with an appropriate YES1-specific antibody. Of the 13 samples, 2 had high expression levels of YES1, 6 had medium expression and 5 low or no expression. Shown are examples of the staining protocol on slides showing high (C), medium (D) and low-zero (E) expression of YES1 in cancer cells on the slides.

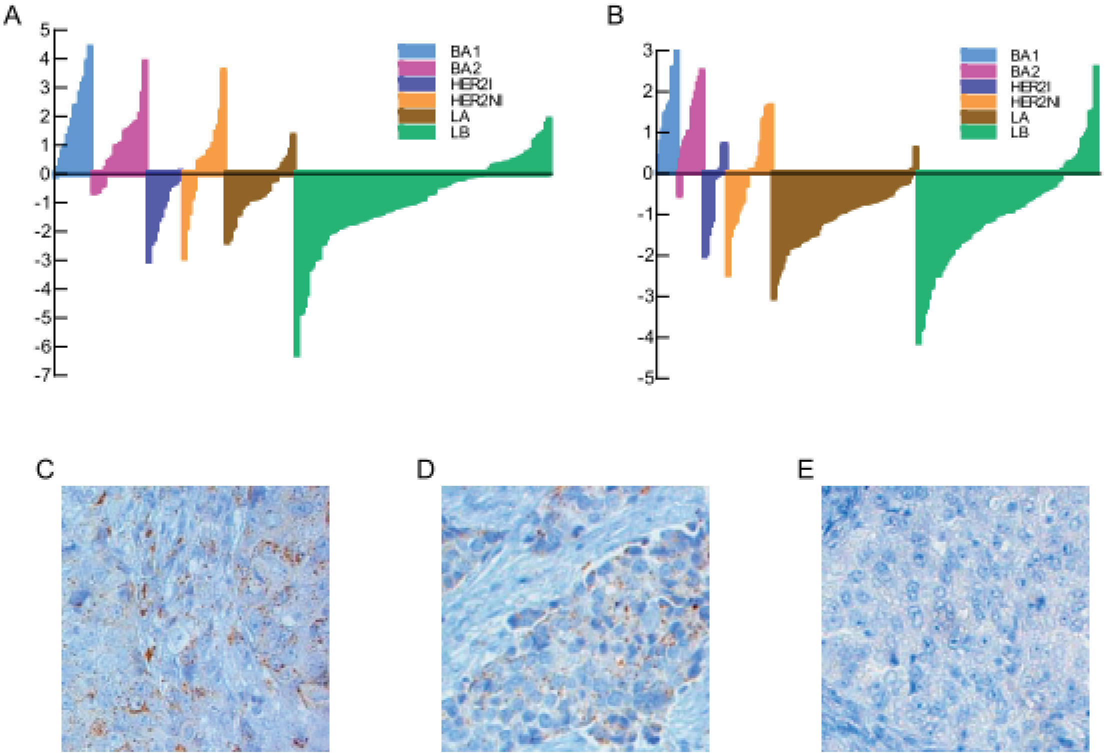


Figure 2.3: YES1 knock-down impairs the growth of breast cancer cell lines

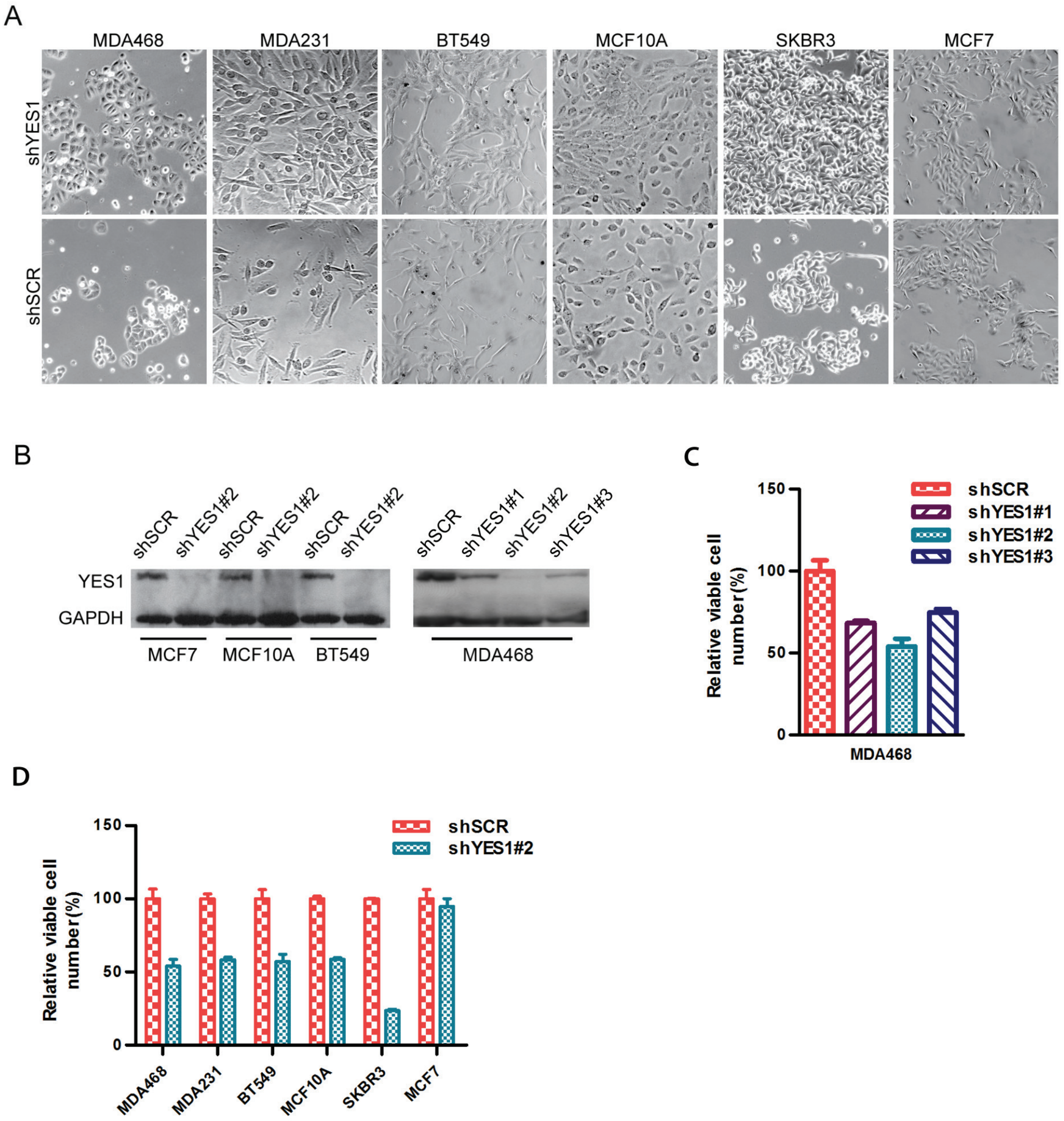
6 breast cancer cell lines were infected with lentiviral constructs designed to suppress expression of YES1 with hairpin shRNA. Equal numbers of these cells were plated in triplicates alongside controls and then counted after 6 days.

(A) Pictures of the cells after 6 days growing in 12 well plates with and without YES1 knockdown.

(B) Western blot of 3 of the cell lines showing the efficacy of YES1 expression knock-down. On MDA468 three different shRNA were used with different efficiencies. The best one was shYES1#2 which was used in further experiments on cell lines.

(C) Average cell counts are normalized to the respective control and shows that compromise of YES1 impairs the growth and survival of MDA468 breast cancer cells. To control for off-target effects three different shRNA constructs were used.

(D) The most efficient lentiviral construct was used on the rest of breast cancer cell lines. Knock-down of YES1 showed a significant effect on basal-like cell lines and no effect on the luminal-like cell line MCF7. All experiments were performed more than once and showed similar results.


A
MDA468
MDA231
BT549
MCF10A
SKBR3
MCF7
shYES1
shSCR
B
shSCR
shYES1#2
shSCR
shYES1#2
shSCR
shYES1#2
shSCR
shYES1#1
shYES1#2
shYES1#3
YES1
GAPDH
MCF7
MCF10A
BT549
MDA468
C
150
100
50
0
Relative viable cell number(%)
shSCR
shYES1#1
shYES1#2
shYES1#3
MDA468
D
150
100
50
0
Relative viable cell number(%)
shSCR
shYES1#2
MDA468
MDA231
BT549
MCF10A
SKBR3
MCF7

Chapter 3:Towards Personalized Therapies

"It is more important to know what sort of person has a disease than to know what sort of disease a person has."
Hippocrates of Kos (ca. 460 BC – ca. 370 BC)

As described in previous chapters, there are numerous drugs for treating breast cancer and even more drugs being tested in clinical trials, waiting for government approval. Considering the molecular diversity of breast tumors, prescribing the right therapy for the patient can be challenging. Diagnosis of cancer and decision about treatment still rely largely on classical histopathological and immunohistochemical techniques. More accurate, quantitative methods are needed that can lead to individualization of treatments. This chapter will present three validated molecular diagnostic tools for predicting therapeutic response in breast cancer, using analysis of gene expression data from both a supervised and an unsupervised perspective.

3.1 Identifying robust subtypes of breast cancer from gene expression data.

The first steps towards understanding the molecular diversity of breast cancer, and how it correlates with response to therapy and other clinical factors, came as a result of the analysis of high throughput breast tumor datasets. Gene expression clustering led to the identification of different subtypes of breast cancer that have distinct biological features, clinical outcomes, and response to treatment (13; 19; 48; 49). Hu el al. (49) showed that hierarchical clustering of a combined set of gene expression datasets (Figure 3.1) first splits the samples into ER+ and ER-, largely on the basis of the difference of expression between genes in the estrogen signaling pathway. ER+ samples further split into two subtypes, Luminal A (ER+ with good prognosis) and Luminal B (ER- with bad prognosis) while ER- samples fell into one of three categories: Basal-like (ER-, PR-, HER2-), HER2+ and IFN (samples that show an enrichment of over-expressed genes from the interferon pathway (IFN)). These authors also found that Basal-like, HER2+, IFN and Luminal B types are in majority high grade, have significantly higher relapse rates and lower overall survival.

The fact that Luminal type breast tumors split into two clinically relevant subtypes with significantly different survival prognosis raised the question whether the same is true for Basal-like or HER2+ cases. Alexe et al. (13) showed that, using the methods and genes originally proposed by Perou et al. (19), HER2+ samples cluster into two groups based on the expression of estrogen pathway genes, which assorts the samples into HER2+/ER- and HER2+/ER+ cases. They also showed that HER2+/ER+ samples cluster with Luminal B samples, and a further analysis of the survival characteristics of these groups shows that this classification does not reflect a clinically useful split. The reason this happens is that there is a disproportionate number of ER+ cases and a large number of genes are co-regulated by ER. Because of these biases in sampling, and in the domination the ER pathway in breast tissue (which in turn biases the choice of genes which represent the dominant variation in the data), all analysis methods will always split samples into ER+ and ER- subsets. However, the fact that this happens does not guarantee that this is a clinically useful classification that adds value beyond the measurement of ER by Immunohistochemistry, which is already a routine part of clinical evaluation of breast cancer. To circumvent this problem, and to take account of the fact that the HER2 pathway is known to cause a more aggressive form of the disease, Alexe et al. (13) removed the HER2+ samples based on IHC measurement of and the over-expression of genes in the HER2 amplicon and clustered HER2+ and HER2- samples separately.

The clustering method used by Alexe et al. (13) was principal component analysis (PCA) for data filtering (identifying significantly variable genes) followed by consensus ensemble clustering of the filtered gene set (51; 52). Figure 3.2 shows a flowchart of the general procedure. After normalization, mRNA levels of four genes in the HER2 amplicon (17q12), ERBB2, GRB7, STARD and PPARBP were used to isolate HER2+ samples. The samples identified as HER2+ were those for at least three of these genes (including ERBB2 or HER2/neu) had high expression. Next, PCA was used to filter out uninformative genes, retaining only those that occur with high coefficients (top 25% in absolute value) in the eigenvectors corresponding to the highest eigenvalues (those representing 85% of the variation in the data).

Consensus ensemble clustering is a procedure which combines results from a number of different clustering algorithms (partitioning, agglomerative, and probabilistic) in a way that improves the quality of the identified clusters. This is achieved by averaging over bootstraps of the data for a given clustering method to find clusters that are stable and robust, i.e. insensitive to perturbations of the choice of samples or genes) and then to combine the results across clustering methods to make them insensitive to the technique used. Samples were assigned to clusters using an agreement matrix constructed whose entries were the fraction of times two samples were in the same cluster across bootstrapped datasets. Core clusters were identified as sets of samples that consistently clustered together across bootstrapped datasets and clustering techniques. Their molecular signatures were then used to classify ambiguous samples (whose class membership fluctuated across bootstrapped datasets or methods). The optimum number of clusters was determined a priori using the gap statistic (53), the Gini index (52) and the silhouette score (54).

The method was demonstrated on the dataset from Wang et al. (37) which consisted of gene expression data from 286 lymph node negative breast cancer samples from patients who were then treated with surgery and radiation alone with clinical median follow up of 86 months. The method described above successfully identified the two ER+ subtypes found previously (Luminal A and Luminal B) which had been confirmed in several previous publications (15; 19; 48; 49). As shown in Figure 3.3A, Luminal B (LB) patients had a significantly worse outcome compared to Luminal A (LA) patients. The analysis also showed that LB samples further clustered into 3 stable subtypes, labeled LB1, LB2 and LB3, with distinct survival expectancies (Figure 3.3B). Within the HER2+ samples, the method identified two core subtypes, with significantly distinct (P = 0.01) long term, distant metastasis free survival rates of 89% for HER2I and 42% for HER2NI (Figure 3.3C). These subtypes were also found by to be distinct by gene set enrichment analysis which showed that the HER2I subtype showed activation of a number of immunity related pathways ($P < 0.01$): T-cell activation, inflammation-mediated chemokine and cytokine signaling, and B-cell activation (55). This correlated well with the analysis of immunohistochemically stained HER2+ samples which found that the HER2I samples had a strong lymphocytic infiltrate compared to HER2NI (13). The clinical value of the classification was prospectively validated by data and slides from small HER2+ neo-adjuvant Herceptin trial which showed that the HER2I samples had a visible immune infiltrate visible in FFPE sections by staining and had better short term response to Herceptin.

Within Basal-like samples, the method also identified two core clusters, labeled BA1 and BA2. Although these subtypes had similar survival curves (Figure 3.3D), pathway enrichment analysis found significant differences between them, with BA1 samples exhibiting up-regulation of genes in the Wnt signaling pathway, immunity and defense and BA2 showing up-regulation of genes in the integrin signaling pathway, cell adhesion, cell structure and motility (55). The conclusion was that Basal-like breast cancers exhibit two molecularly distinct subtypes which are likely to be biological disease entities, but that these subtypes display no significant differences in their risk for progression upon local treatment alone.

The overall result of the study of Alexe et al. (13) was the identification of eight subtypes of breast cancer with distinct molecular and clinical characteristics. In our analysis we used the clustering results by Alexe et al. (13) to classify new breast cancer datasets into BA1, BA2, HER2I, HER2NI, LA and LB subtypes. Sub-classification of LB in LB1, LB2 and LB3 was not used, since this would have introduced a disproportionate number of ER+ subtypes, which would have biased the normalization procedure described in Appendix A.2.

3.2 The Genomic Grade Index: a measure of progression risk

MapQuant DX™ is a molecular diagnostic test provided by Ipsogen Inc. in the European Union. It uses fresh or frozen breast tumor samples to measure the expression of specific genes with the aid of DNA microarray technology. The MapQuant DX genomic grade test is based on the Genomic Grade index (GGi (17)) which measures the expression of 97 genes that best characterize high-grade vs. low-grade tumors. The manufacturer claims that this test can resolve grade 2 breast tumors, which represent 30-60% of all cases, into either grade 1 or grade 3 tumors 80% of the time. Histologic grade, a consensus indicator of tumor proliferation and risk of metastasis, is an important diagnostic factor and aids in deciding treatment course. High grade tumors (grade 3) have bad prognosis and are treated more aggressively as opposed to low grade tumors (grade 1) that have a better prognosis. Grade 3 tumors are also generally chemo-sensitive and are treated by chemotherapy while grade 1 tumors are often chemo-insensitive. This is why choosing the right treatment for intermediate grade 2 tumors is a critical issue. By resolving grade 2 tumors into either low-risk grade 1 or high-risk grade 3, numerous patients can be spared potentially useless and painful chemotherapy.

The current assay was developed from supervised analysis of a gene expression dataset comprising 64 estrogen receptor positive (ER+) breast cancer samples. In addition, a cohort of 597 independent tumors was used to evaluate the association between the Genomic Grade index (GGi) and relapse free survival rates.

Genes that were differentially expressed between histologic grades 1 and 3 in the training set were ranked according to the standardized mean difference (56) between expression levels in the two groups. This statistic is similar to a t-test but better suited when the training dataset comes from different laboratories, such as in this case. A step-down procedure called maxT (57; 58) was used to correct for multiple hypotheses.

The expression pattern of 97 genes was found to be significantly different between 33 grade 1 tumors versus 31 grade 3 tumors, with a majority of them being over-expressed in the high histologic grade group. Amongst them were mostly genes associated with cell cycle progression and proliferation like UBE2C, KPNA2, TPX2, FOXM1, STK6, CCNA2, BIRC5, and MYBL2.

To summarize the similarity between expression profiles of these genes and histologic grade the authors introduced a scored named Genomic Grade index (GGI):

$$\mathrm{GGi} = \mathit{scale}\left(\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j - \mathit{offset} \right),$$

where *scale* and *offset* are transformation parameters to standardize the gene expression grade index values to mean -1 for low grade and mean +1 for high grade. Variables x_j represent logarithmic gene expression measures while G_1 and G_3 represent the set of genes with increased expression in histologic grade 1 and grade 3 tumors, respectively.

Validation of the GGi score was conducted on a combined set of 597 breast tumor samples of various histologic grades, estrogen receptor status and lymph node status. Kaplan-Meier survival curves and hazard ratio (HR) estimates of different risk categories are shown in Figure 3.4. Pathologist scored histologic grade were used to separate the dataset into 3 categories, from grade 1 to 3 with decreasing relapse free survival rates (HR = 3.18, 95% CI = 2.1 – 4.8; P < 0.001) as shown in Figure 3.4A. Patients with histologic grade 2 were then assigned gene expression grade 1 if GGi < 0 and gene expression grade 3 if GGi > 0. Among these intermediate histologic grade patients, the ones assigned gene expression grade 3 had statistically significant difference in survival compared to samples assigned gene expression grade 1 (HR = 3.61, 95% CI = 2.25 – 5.78; P < 0.001). Figure 3.4B shows similar survival curves to those with histologic grades 1 and 3 from Figure 3.4A. When GGi scores were computed for all samples and split into gene expression grades 1 and 3 (Figure 3.4C), a similar survival difference was noticed (HR = 2.83, 95% CI = 2.13 – 3.77; P < 0.001). Furthermore, the differences in survival in the combined 597 samples were conserved among individual datasets as shown by the forest plots in Figure 3.4D-F.

The prognostic power of gene expression grade was assessed in combination with other variables like ER status, lymph node status, histologic grade, tumor size and patient age at diagnosis with a multivariate Cox regression model. Only gene expression grade, lymph node status, and tumor size were found to be statistically significant, with gene expression grade having the strongest association (HR = 1.99, 95% CI = 1.43 – 2.78; P < 0.001). As expected, histologic grade provides little additional information when compared with the gene expression grade index. GGi seems to perform better than traditional pathologist grading at predicting breast cancer prognosis and hence provides a more efficient and consistent way of classifying tumors.

3.3 The Oncotype DX® recurrence score

Oncotype DX® is a molecular diagnostic test provided by Genomic Health Inc. widely in use in Europe and in the United States. It uses paraffin embedded breast tumor samples to measure the expression of specific genes using RT-PCR (59). This assay is used to determine a recurrence score that uses the expression level of 21 genes associated with recurrence in patients with estrogen positive (ER+), node negative breast tumors treated with Tamoxifen (16). The test has been shown to predict the magnitude of chemotherapy benefit for breast cancer patients treated with a variety of different chemotherapy regimens. It is currently used by physicians to decide treatment regiment for early stage ER+/HER2- breast cancer patients. Patients with high Oncotype DX scores will receive hormonal treatment as well as chemotherapy while patients with low scores will be treated only with Tamoxifen or another drug that targets the estrogen pathway.

This method was developed from the meta-analysis of three separate studies (60; 61; 62) that looked for genes that correlated with disease recurrence in breast cancer patients with various disease types and different treatment regiments. 250 cancer related genes were selected from published literature, pathway analysis, genomic databases, and microarray gene expression profiling experiments on breast tumors; and their expression level determined by reverse transcriptase polymerase chain reaction (RT-PCR). Next, the relationship between candidate genes and recurrence over a total of 447 samples was assessed. Imposing consistency in the results between the three separate datasets removed the majority of the candidate genes leaving 16 cancer genes strongly associated with breast cancer recurrence. These were: HER2, GRB7 (from 17q12 amplicon), MMP11,CTSL2 (invasion markers), Ki-67, STK15, Survivin, CCNB1, MYBL2 (proliferation markers), ER, PR, BCL2, SCUBE2 (estrogen pathway), GSTM1, CD68, BAG1; together with reference genes: ACTB, GAPDH, RPLPO, GUS and TFRC. These reference genes are necessary to normalize the quantified RT-PCR expression levels to compensate for sample variation in extracted RNA due to variations resulting from the tissue-fixation processes, the age of the specimen (which affects quality of RNA), and other variables unrelated to gene expression.

Analyses were performed to determine the functional form of the variables to be included in the model. Correlation analysis, dimension reduction, Martingale residual analysis, concordance measure of accuracy, and bootstrap resampling were used for this purpose (63). Intermediate scores were calculated separately for the HER2 group, the ER group, the invasion markers group, and the proliferation markers group. The final recurrence score RS based on these genes was defined as a linear combination of the scores of each group of markers in addition to the normalized expressions of CD68, GSTM1, and BAG1:

$$\begin{aligned}\mathrm{RS} = {} & 0.47 \times \textit{HER2 group score} \\ & - 0.37 \times \textit{ER group score} \\ & + 1.04 \times \textit{Proliferation group score} \\ & + 0.10 \times \textit{Invasion group score} \\ & + 0.05 \times \textit{CD68} \\ & - 0.08 \times \textit{GSTM1} \\ & - 0.07 \times \textit{BAG1}\end{aligned}$$

Details on how the recurrence score (RS) is calculated is given in Paik et al. (16) and Cronin et al. (64).

The value of this assay to separate ER+, node-negative breast cancer patients, into meaningful clinical classes, which represent the risk of disease recurrence, was tested on 668 patients enrolled in the National Surgical Adjuvant Breast and Bowel Project (NSABP) clinical trial B-14. The patients were selected if they had been randomly assigned to receive Tamoxifen or had received Tamoxifen as members of the registration group of NSABP trial B-14. Each patient was assigned a single risk class based on their recurrence score: low risk (RS < 18), intermediate risk (18 ≤ RS < 31), or high risk (RS ≥ 31). These thresholds were determined based on the recurrence rates of Tamoxifen-only treated patients in the NSABP B-20 clinical trial, one of the three initial datasets on which the recurrence score was trained.

Kaplan-Meier estimates for the proportion of patients who had suffered distant recurrence (Figure 3.5) at 10 years was 6.8% (95% CI = 4.0% – 9.6%) for the low risk group, 14.3% (95% CI = 8.3% – 20.3%) for the intermediate group, and 30.5% (95% CI = 23.6% – 37.4%) for the high risk group. The observed differences in survival were statistically significant at a log-rank P value < 0.001.

A multivariate Cox regression model was used to explore the relation between the recurrence score, age at surgery, clinical tumor size, histologic grade, HER2 amplification, and ER protein levels. Only high tumor grade (HR = 3.34, 95% CI = 1.79 – 6.29; P < 0.001) and the recurrence score (HR = 2.81, 95% CI = 1.70 – 4.64; P < 0.001) were significant predictors of distant metastasis recurrence. The assessment of tumor grade was made by three pathologists separately, with similar results for high grade tumor but substantial differences in assigning consistent labels to low and intermediate grade tumors. However, the recurrence score provided significant (P < 0.001) discriminatory power beyond tumor grade for each of the three pathologists.

3.4 Limitations and challenges in predicting recurrence risk

Molecular diagnostic tests have proven their advantage over traditional laboratory techniques and usefulness in clinical practice. However, questions remain over their accuracy, and whether the high price paid for them is justified. All studies previously described showed that histologic tumor grade is a good marker for disease progression and treatment response. Although numerous claims have been made about the inconsistencies in histological grading between different pathologists (65; 66; 67), unified methods such as Elston and Ellis modification (8) of the Bloom and Richardson method have greatly improved the reproducibility of histologic grading. Furthermore, the problem of undecidable intermediate grade tumors is not completely solved by any of the presented molecular methods. Clustering based methods discovered classes with intermediate recurrence rates such as LB1 and LB3, which interpolate between the good prognosis LA and truly poor prognosis subtype LB2. The Genomic Grade index from

Ipsogen Inc. suffers from “the intermediate value” problem when GGi score is close to zero, when it becomes impossible to determine whether the patient should be assigned Grade 1 or Grade 3. Similarly, for the Recurrence Score of Genomic Health Inc., 22% of patients were declared to have intermediate risk if RS score was between 18 and 31. Patients assigned an intermediate RS have an ambiguous risk assignment of limited clinical value. Since the Oncotype DX assay is expensive (~$3000 per test at this time), its clinical value for this intermediate class is unclear.

Another caveat of these methods is that they are uninformative beyond providing a recurrence score. Compared to them, the use of a well knows breast cancer marker HER2, not only indicates a particular risk group, but also provides a clear molecular pathway of disease progression that can be therapeutically targeted. Moreover, HER2 amplification can be determined by a simple test that can be done cheaply and quickly by any hospital, compared to the expensive diagnostic assays that require fresh, snap-frozen or paraffin embedded tumor samples to be sent to a central laboratory.

Common technologies used to measure gene expression levels, like DNA microarrays or Q-RT-PCR, actually measure RNA levels in a mix of tumor, fat and connective tissue. The specifics of tumor cellularity and within patient variability will bias the result of these measurements, and significantly affect the reliability of the final risk score. Using laser micro-dissection to harvest only tumor tissue, as well as replacing current gene expression measurement techniques with more accurate methods based on sequencing (RNA-seq), has the potential to greatly improve the accuracy and reproducibility of molecular diagnostics.

Figure 3.1: Molecular breast cancer subtypes derived from hierarchical clustering

Hierarchical cluster analysis of the 315-sample combined test set using the Intrinsic/UNC gene set reduced to 306 genes. (A) Overview of complete cluster diagram. (B) Experimental sample-associated dendrogram. Figure reproduced with permission from Hu et al. (49).

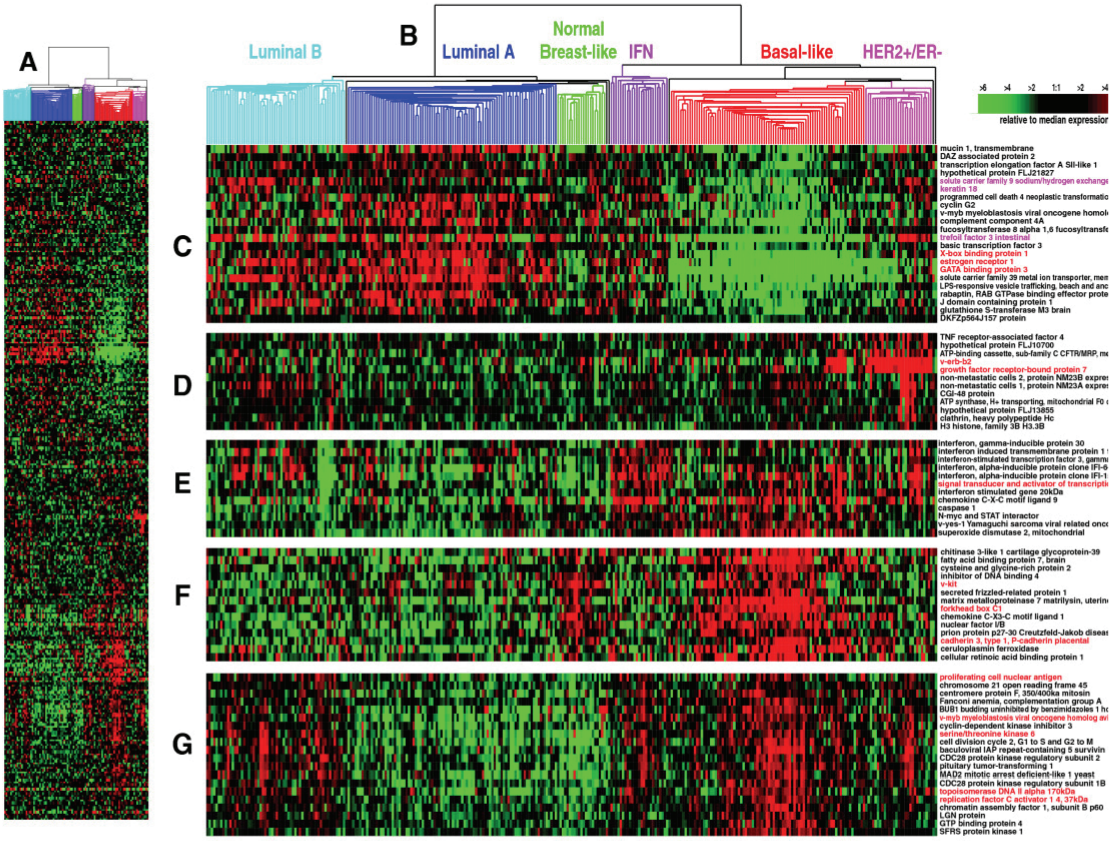


Figure 3.2: Flowchart of the clustering method

After identification of HER2+ samples, PCA and consensus ensemble clustering find 2 HER2+ subtypes (HER2I & HER2NI), 2 Basal-like subtypes (BA1 & BA2) and 4 Luminal subtypes (LA, LB1, LB2, LB3). Figure reproduced with permission from Alexe et al. (13).

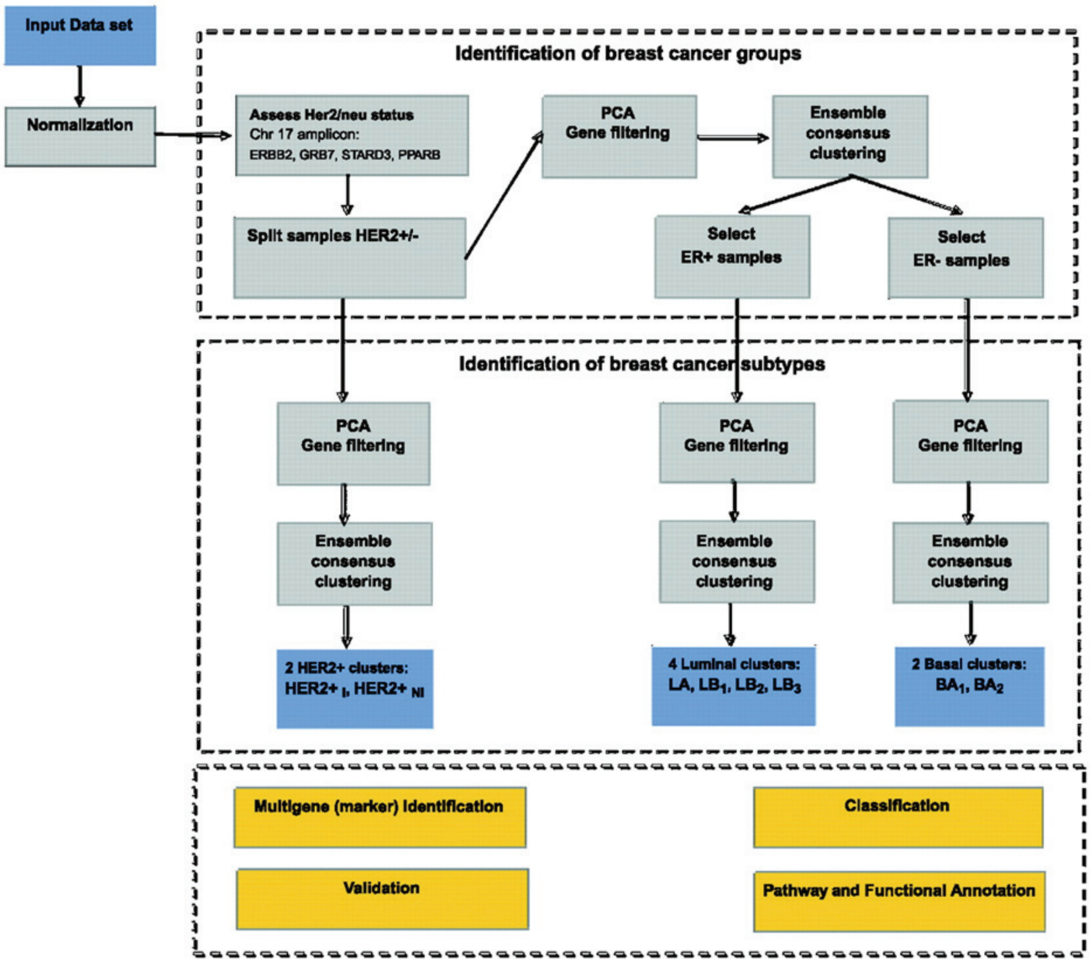


Figure 3.3: Kaplan-Meier curves comparing distant metastasis rates for Luminal, HER2+ and Basal-like breast cancer subtypes.

(A) Luminal B (LB) has a slightly significant (log-rank P value = 0.14) poorer prognosis than Luminal A (LA). (B) LB splits into LB1, LB2 and LB3 with LB2 having the worst prognosis. (C) HER2I has 89% long term distant metastasis free survival rate vs 42% for HER2NI (log-rank P value = 0.01). (D) The log-rank P value for difference in survival is 0.6 so this difference is not significant. However, this does not preclude a biological basis for the two subtypes. Figure reproduced with permission from Alexe et al. (13).

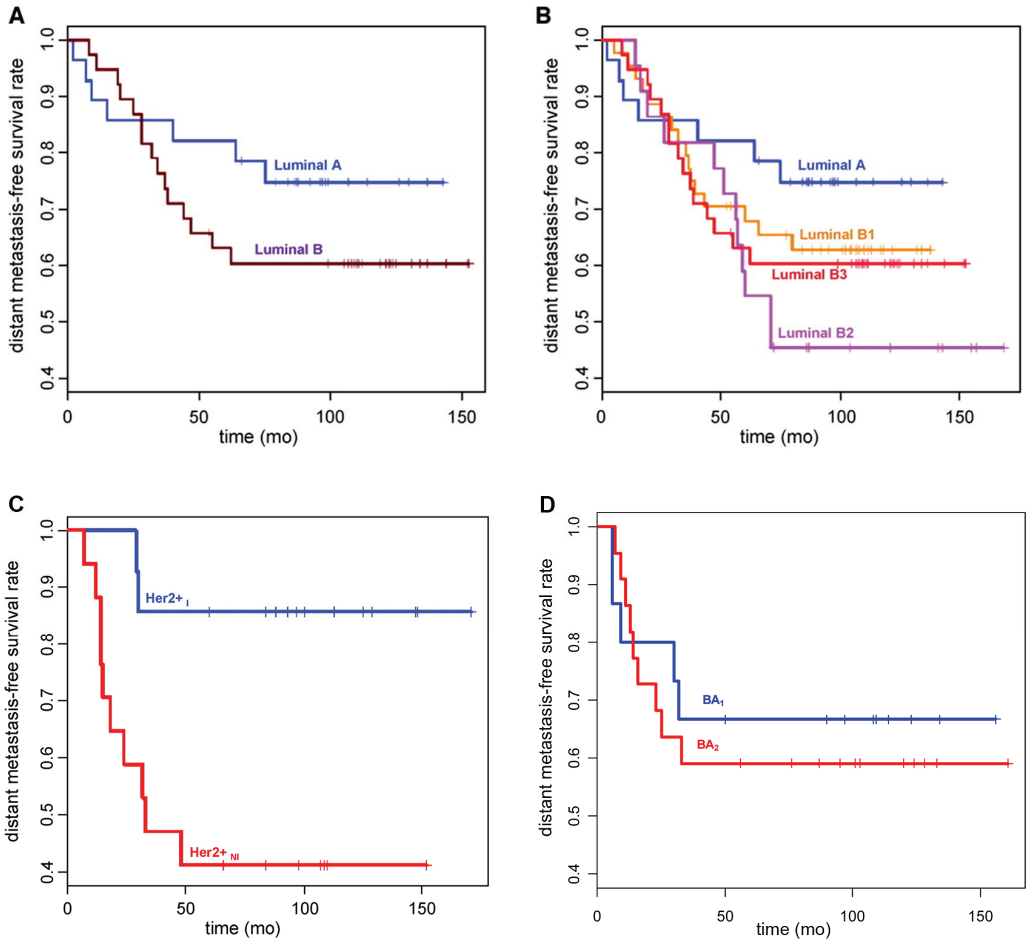
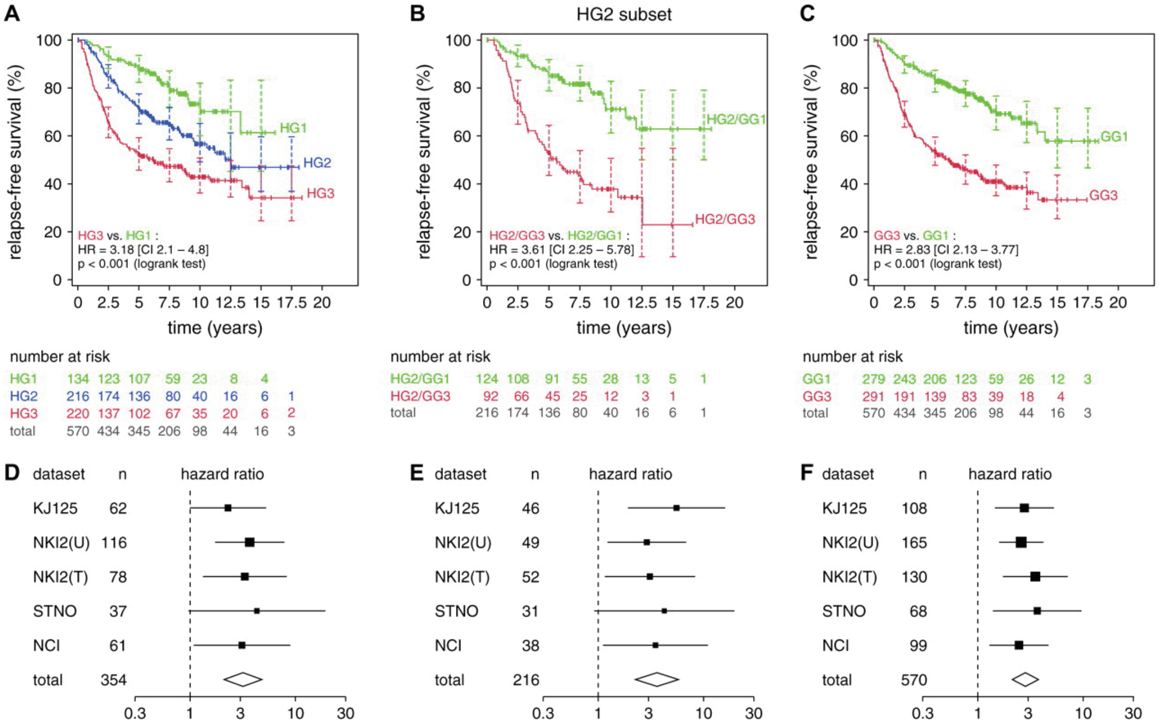


Figure 3.4: Relapse-free survival according to Genomic Grade Index categories

Relapse-free survival analysis for all validation datasets. Only 570 patients with complete histologic grade (HG) and relapse-free survival information were included. Kaplan–Meier analyses were conducted with pooled data. Number of patients at risk and 95% confidence intervals (CIs) for the relapse-free survival estimates (shown as error bars) are indicated at 2.5-year intervals. Difference in relapse-free survival between two groups is summarized by the hazard ratio (HR) for recurrence with its 95% CI. NKI2(U) = untreated subset of dataset NKI2; NKI2(T) = treated subset of dataset NKI2. (A) Analysis of the whole dataset by HG1 (green), HG2 (blue), or HG3 (red). (B) Analysis of patients with HG2 tumors by gene expression grade (GG). The 217 patients with HG2 tumors were separated into low- and high-risk subsets by GG as GG1 (green) and GG3 (red), respectively. (C) Analysis of the whole dataset of 572 patients by GG. GG1 = green; GG3 = red. All statistical tests were two-sided. To show consistency among different datasets, forest plots of the hazard ratios and confidence intervals for individual datasets are shown below the corresponding Kaplan–Meier plots (panels D, E, and F, corresponding to panels A, B, and C, respectively). The difference among the groups is significant (log-rank P value < 0.001). Figure reproduced with permission from Sotiriou et al. (17).



number at risk

HG1	134	123	107	59	23	8	4	
HG2	216	174	136	80	40	16	6	1
HG3	220	137	102	67	35	20	6	2
total	570	434	345	206	98	44	16	3

number at risk

HG2/GG1	124	108	91	55	28	13	5	1
HG2/GG3	92	66	45	25	12	3	1	
total	216	174	136	80	40	16	6	1

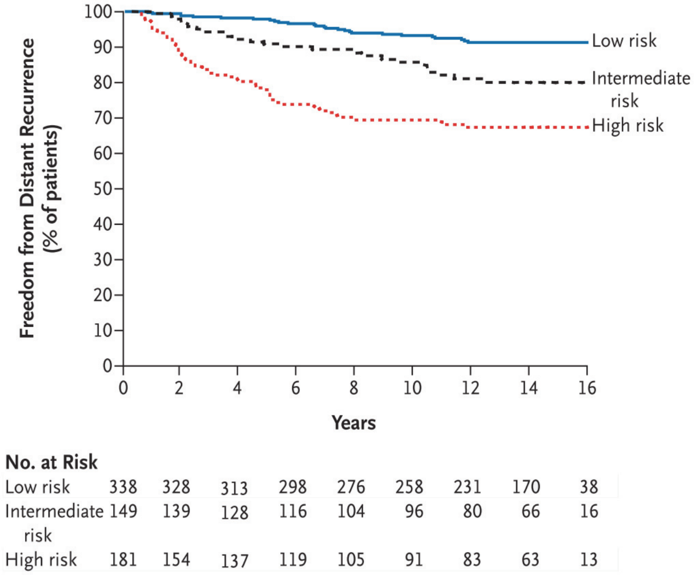
number at risk

GG1	279	243	206	123	59	26	12	3
GG3	291	191	139	83	39	18	4	
total	570	434	345	206	98	44	16	3

dataset	n (D)	n (E)	n (F)
KJ125	62	46	108
NKI2(U)	116	49	165
NKI2(T)	78	52	130
STNO	37	31	68
NCI	61	38	99
total	354	216	570

Figure 3.5: Distance recurrence free survival according to Oncotype DX score categories

A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher. There were 28 recurrences in the low-risk group, 25 in the intermediate-risk group, and 56 in the high-risk group. The difference among the groups is significant (log-rank P value < 0.001).



No. at Risk

Low risk	338	328	313	298	276	258	231	170	38
Intermediate risk	149	139	128	116	104	96	80	66	16
High risk	181	154	137	119	105	91	83	63	13

Chapter 4:Predicting Resistance to Endocrine Therapy

"Prediction is very difficult, especially about the future."
Niels Bohr (1885-1962)

4.1 Motivation and overview

Hormonal therapy is widely prescribed for the treatment of estrogen receptor positive (ER+) breast cancer and has had a great impact on survival in this disease (29). In spite of this, there remains a significant subset of ER+ breast cancer patients have early recurrence despite endocrine therapy. This suggests that a subset of ER+ breast cancers have intrinsic resistance to hormone therapy. A better understanding of the biological mechanisms underlying resistance to hormonal therapy is of considerable clinical significance and may suggest new strategies in the treatment of breast cancer patients.

The best validated assay to identify patients likely to have early recurrence on hormone therapy is the Oncotype DX assay (16) from Genomic Health, Inc., based on RT-PCR measurement of the mRNA of 21 genes. Other assays, such as the Genomic Grade Index (17), and clinical markers such as histological grade, are also used to identify good prognosis and poor prognosis ER+ breast cancer patients. Analysis of gene-expression data can also separate ER+ breast cancers into good prognosis Luminal A cancers, and poor prognosis Luminal B cancers (13; 38; 39).

Several studies have shown that prognostic assays such as Oncotype DX are essentially identifying Luminal A tumors (low grade, ER+ breast cancers) as being good prognosis, and non-Luminal A, ER+ breast cancers (Luminal B which are ER+, non-low grade, some with HER2 amplification) as poor prognosis (15; 74; 75). Moreover, these assays, although they have prognostic and predictive utility, do not identify the biologic pathways driving resistance in the poor prognosis tumors. For example ER+ tumors that have HER2 amplification will have a high Oncotype DX recurrence score (RS), high histological grade, and a high genomic grade and be identified as poor prognosis. But conversely, not all high RS ER+ breast cancers have HER2 amplification; indeed the majority of them do not. If an ER+ tumor identified as being poor prognosis by genomic assays is found to have HER2 amplification, this finding gives insight into the biological pathways mediating ER independence and identifies a therapeutic target that can be successfully exploited. However the majority of poor prognosis ER+ cancers (Luminal B cancers) do not have HER2 amplification (15). As identification of HER2 amplification by FISH is now routinely done for all breast cancers, clinicians do not use tools such as Oncotype DX for HER2+ tumors; such assays are mostly performed on ER+/HER2- tumors. For ER+/HER2- tumors identified by Oncotype DX and similar assays as being poor prognosis (HER2-, Luminal B breast cancers), there is neither great insight into their underlying biology nor is there available targeted therapy.

In this chapter we describe our own efforts to gain insight into the biology of poor prognosis ER+/HER2- breast cancers. Using a novel outlier analysis, we found that other amplicons besides HER2 may be driving the growth of these tumors. To identify potential amplicons, we applied an outlier analysis to published clinically annotated gene expression dataset of ER+ breast cancers treated with Tamoxifen. Instead of identifying individual genes associated with poor outcome, which simply generates meaningless lists of genes, we took a different approach. We focused instead on identifying sets of genes in the same set of patients with similar outlier profiles whose expression correlated with outcome. The idea is to identify clusters of samples with a specific phenotype (poor prognosis patients) who exhibit a set of associated outlier genes (with unusually high or low expression) compared to controls (good prognosis patients). The set of outlier genes was mapped to chromosomal regions and then analyzed these genes to identify enrichment of chromosomal regions whose amplification was correlated with outcome. This analysis had the dual goals of identifying potential amplicons whose presence in ER+ breast cancer directly correlates with poor prognosis as well as identification of "driver" oncogenes that can be therapeutically targeted. Much like the identification of the HER2 gene, one expects "driver" oncogenes in these amplicons who are responsible for the poor prognosis phenotype. The identification of such genes would lead to improved therapies, as was the case in the development of Herceptin which targets HER2 amplification.

Our analysis found that high expression of sets of genes in four distinct regions of the genome is highly predictive of poor prognosis in ER+ breast cancers treated with Tamoxifen. One of the identified chromosomal regions was in 17q12, which is the HER2 amplicon (77; 78). This identification validates our method. The other three regions we found were in 17q21.33-q25.1, 8p11.2 and 8q24.3. Although these regions have been previously identified in subsets of breast cancers (79) and known to contain potential oncogenes, their association with poor prognosis in patients undergoing hormone therapy is novel.

We have validated the presence of these amplicons in ER+ breast cancer and to some extent, their association with poor prognosis in an independent set of clinical samples, using a high Oncotype DX score as a surrogate for poor prognosis.

4.2 Identifying amplicons associated with Tamoxifen resistance

Although most patients with ER+ breast tumors have good outcome with Tamoxifen treatment, a subset of ~30% has disease recurrence despite Tamoxifen treatment (29). In this scenario, resistant ER+ breast tumors act as outliers because they behave markedly different from Tamoxifen responsive patients. Hence patients with the resistant phenotype may contain genes whose expression are "outliers" and whose identification may suggest the biology of resistance. We define an "outlier" as a measurement that deviates significantly from the distribution of the rest of the data. One expects that genes whose over/under-expression is responsible for the more aggressive phenotype should behave as outliers, because the proliferating tumor is either addicted to them or else needs to down-regulate them to survive and grow. Each gene defines its own outlier profile, on sets of samples where its expression levels are unusually high or unusually low compared to the rest.

Three gene expression datasets collected from breast cancer patients published by Loi et al. (74; 80) were obtained from the Gene Expression Omnibus website (GEO:www.ncbi.nlm.nih.gov/geo) accession number GSE6532 (74; 80). The sets are abbreviated with KIT, OXFT and GUYT representing the institutions where they were processed: Uppsala University Hospital, Uppsala, Sweden, John Radcliffe Hospital, Oxford, United Kingdom and Guys Hospital, London, United Kingdom. They comprise of 81, 109 and 87 ER+ breast cancer samples from patients treated with Tamoxifen that included 9 years of median follow-up for relapse free survival and distant metastasis information (Table 4.1). Distant metastatic events were recorded for 26% of the sample population, 92% of patients were over 50 years old, 19% low grade, 51% medium grade, 16% high grade, 47% lymph node negative and 53% lymph node positive (74) (Additional Table 1). The expression data were obtained on Affymetrix (Affymetrix Inc., Santa Clara, CA) microarray platforms U133A/B (KIT & OXFT) and U133Plus2 (GUYT), then MAS5 normalized. In order to combine the three sets into one analysis, probes corresponding to genes that were not present across all platforms were discarded. After taking log2 of each intensity values, multiple probes corresponding to the same gene were compressed to the one with the biggest median expression over all samples.

For each gene, the expression values were median centered and then divided by the median absolute deviation (MAD) as described in Tomlins et al. (81). Median and MAD were used here instead of the usual mean and standard deviation because they are less influenced by the presence of outliers. This step was performed separately for KIT, OXFT and GUYT datasets in order to avoid distribution biases that arise from the merger of separate expression array tables.

We define the outlier cut-off value for a gene as the expression value in the normalized data which is outside of the 90% quantile across genes for each sample for "high" outliers and 10% quantile for "low" outliers. The results presented are not too sensitive to these thresholds and similar results are obtained when the outlier quantile cut-off is varied by +/-5%. Outlier expression limits for high and low expression were identified for each gene and the high/low outlier genes in every sample array were identified using these values. In this way, the dataset, which is a matrix of *genes x samples*, splits into a sum of three matrices: one matrix defining non-outlier samples and genes, and two matrices, B_1 and B_2 defining high and low outlier genes and samples respectively, with rows corresponding to genes and columns corresponding to samples. Each row represents the distribution of outliers for the corresponding gene across samples with $B_{1,2} = 1$ if *gene i* is over-expressed (high outlier) in *sample j* and $B_{1,2} = 0$ if *gene i* is under-expressed (low outlier) in *sample j*. The rest of the elements (for non-outlier samples) were set to zero.

This process was repeated for all three gene expression datasets: KIT, OXFT, GUYT and the results merged by concatenation, resulting in matrices with the same number of rows but with number of columns equal to the total number of samples for the three datasets combined. Next, genes with less than 10 outliers across all samples were discarded because they lack statistical power. In a sample size of ~100 (the approximate number of poor prognosis cases), one expects a standard deviation of ~ 10 samples. Hence, discarding genes which have less than 10 samples in their outlier set controls for statistical fluctuations at the one-sigma level. For each of the remaining genes, the distribution of outliers across samples defines two classes: the set of samples with aberrant (outlier) expression of the corresponding gene and the set of samples where the gene expression is “normal”. For each gene, we generated Kaplan-Meier survival curves for these two classes and compared them for differential survival using a log-rank test. A gene was retained as a true outlier (relevant to prognosis) if its outlier sample class was statistically distinguishable from its complement at a log-rank P = 0.05 or better.

Usual methods based on classification or clustering fail to identify features (biomarkers), associated with prognosis, that are not consistently spread in the dataset. Each case might exhibit different combinations of features, which leads to inconsistent results that are highly dependent on the frequency and distribution of the biomarkers. To overcome this, we ask the question: Do the outlier lists define gene and sample sets which are “collectively” associated with the phenotype or poor prognosis? For this to happen sets of genes must exist with similar outlier sets of samples - i.e. the genes must be over/under-expressed in roughly the same set of samples. This corresponds to the presence of tightly correlated clusters in binary matrices B_1 and B_2. One suitable correlation measure to identify such clusters is the Phi coefficient which is equivalent to a Pearson correlation between pairs of rows of the matrices B_1 and B_2. Let C_1 and C_2 be the covariance matrices between the rows of B_1 and B_2 respectively, then $R_{1,2}(i,j) = C_{1,2}(i,j) \Big/ \sqrt{C_{1,2}(i,i)C_{1,2}(j,j)}$ is the matrix of correlation coefficients between the outlier profiles of the genes in $B_{1,2}$.

Clusters of tightly correlated genes were identified by iteratively removing *row i* and *column i* with $\sum_j \Delta(i,j) \leq 1$ where $\Delta(i,j) = 1$ if $R_{1,2}(i,j) > 0.5$ and $\Delta(i,j) = 0$ otherwise; until a stable set was obtained, where stability means that the size of the reduced matrix *R'* stops changing. PCA plots of the resulting reduced matrices B_1 and B_2 identify distinct groups of highly correlated genes that are now suitable for pathway enrichment analysis. Gene clusters in Figures 4.1A and B are associated with bad/good prognosis based on the survival profiles defined by the genes within each cluster. Further, each gene is labeled with the appropriate pathway information taken from the Gene Ontology (82) database together with chromosomal location information obtained from Affymetrix annotation files. We used a Fisher Exact test to assess the significance of pathways and chromosomal location enrichment for each group of genes (Table 4.2).

Figure 4.1 shows the projection of the outlier gene profiles on the first two principal components of the filtered matrix for high outlier values (A) and low outlier values (B). Each point in the graph corresponds to a gene. By careful examination of the survival curves associated with their outlier profiles we observed that genes associated with good prognosis naturally separate from outliers associated with poor prognosis. The clusters circled in red are correlated with poor prognosis and the ones in blue with good prognosis.

Pathway enrichment analysis using Gene Ontology (82) (GO) revealed that the outlier gene clusters in Figure 4.1 were enriched in specific biological pathways. Chromosomal location information for each gene was collected from the Affymetrix annotation file of the 3' Expression Array HG-133 Plus2. Mapping these to chromosomes, we defined amplicons as continuous regions on the chromosome which were enriched in these outlier genes. Amplicon and pathway enrichment was assessed using Fisher's Exact Test (83). These results are summarized in Table 4.2.

Over-expressed outlier genes associated with good prognosis were enriched in two pathways - the development and cell adhesion pathway and the immune response pathway. In the poor prognosis samples, the outlier genes over-expressed genes in cell cycle pathways and in four chromosomal regions: 17q12, 17q21.33-q25.1, 8p11.2 and 8q24.3. The set of cell cycle pathway genes we identified contained genes associated with proliferation and were almost all were part of the genes used in the Genomic Grade Index (17). Our observation thus confirms that proliferation-associated genes are strong markers of poor prognosis in ER+ breast cancer. One of our identified amplicons was the 17q12 amplicon (84), which contains the gene ERBB2 (HER2), is known to be associated with relative resistance to hormonal therapy and poor prognosis. The three other amplicons:17q21.33-q25.1 (85; 86; 87), 8p11.2 (88; 89) and 8q24.3 (79) have been previously reported as amplified in a subset of breast cancers and probably contain one or more driver oncogenes responsible for the poor prognosis phenotype. The full list of outlier genes identified in the amplified chromosomal regions is given in Table 4.3. Highlighted in red are oncogenes previously identified in the literature: WHSC1L1 (90; 91), CLTC (92; 93; 94), HSF1 (95), and LSM1 (96).

For under-expressed outliers with good prognosis we find enrichment of the cell cycle pathway, while the immune response and cell adhesion are associated with poor prognosis. This mirrors the results from the over-expressed outlier analysis confirming the strong association of the cell cycle, immune response and cell adhesion pathways with prognosis in ER+ breast cancers.

4.3 Gene patterns that predict Tamoxifen resistance

To examine the relationship between the cell cycle pathway and the four potential amplicons identified by our analysis, a correlation matrix of all the genes identified to be associated with poor outcome was computed. This is displayed as a heatmap in Figure 4.2. Correlation between the presence of any one amplicon and the presence of the other amplicons or the cell cycle pathway are shown in Table 4.4. The cell cycle pathway correlates partly with all the amplicons (Figure 4.2, Table 4.4), suggesting that increased expression of the cell cycle pathway is always associated with the presence of the amplicons to some degree. To study this association further, samples with enrichment of any of the four amplicons or with cell cycle pathway enrichment were identified by requiring at least 50% of gene markers in each group to be over-expressed, i.e. is marked as a high outlier in the respective sample. It was found that most samples (90.5%) that over-expressed cell cycle genes displayed at least one of the four chromosomal amplifications, suggesting a direct (causal) relationship between tumor proliferation and the presence of these amplicons. The most likely relationship is that the amplicons are upstream of the cell cycle genes, i.e., driver genes on the amplicons up-regulate pathways which result in amplification of the cell cycle genes.

Our analysis also showed that the amplicons are poorly correlated with each other (see Table 4.4) suggesting that the presence of each amplicon is most likely to be functionally independent of the others. The conclusion is that each of these amplicons *is a separate marker for poor prognosis* and that *each amplicon may be driving the disease process using distinct mechanisms and pathways.*

We next analyzed the effect of the presence of cell cycle pathway amplification and the presence of the four amplicons on survival. These results are shown in Figure 4.3 and Figure 4.4. The up-regulation of cell cycle pathway genes (Figure 4.3) was found to be associated with significantly reduced time to distant metastasis (Hazard Ratio (HR) = 9.71, 95% CI = 3.3 – 28.6; P < 0.0001). The presence of any one of the four amplicons was also associated with significantly increased risk of distant recurrence (Figure 4.4). Hazard ratios for samples with amplicons on 17q12, 17q21.33-q25, 8p11.2 or 8q24.3 vs. no amplicons were (4.09, 3.14, 3.75, 4.29) respectively, while log-rank p-values for the survival differences were (6.3e-07, 3.0e-04, 5.7e-06, 2.2e-06) (Table 4.5). This shows that the three novel amplicons each confer additional risk of disease progression that is similar to that of HER2 amplification.

4.4 Validation of the association of amplicons with poor outcome

To validate the association of amplicons with poor prognosis we analyzed a CGH dataset on a separate set of breast cancer samples published by Jonsson et al. (97) (GEO accession number GSE22133 in GEO: www.ncbi.nlm.nih.gov/geo). This comprised of

359 breast tumor tissues that included 8.1 years of median follow-up survival information, of which 222 were ER+ samples which we used in our analysis. Unfortunately, patients were not uniformly treated and the exact specifics of the treatments, as well as clinical information other than grade, ER and PR status, were unavailable. Copy number estimates were obtained and segmented using circular binary segmentation (CBS) (98) followed by identification of significant amplification peaks with the GISTIC (99) algorithm as described by Jonsson et al. (97). Amplification peaks were detected in 17q12, 17q23.2, 8p11.2 and 8q24.3 that strongly overlapped the regions we had discovered previously in gene expression data. A correlation analysis between samples with these amplicons showed little to medium associations (Table 4.6), similar to the previous obtained values in Table 4.4.

Survival curves were plotted (Figure 4.5) for samples with and without amplicons. As shown in Figure 4.5 and Table 4.7, the presence of an amplicon in any of these four regions was associated with significantly higher death rates. Of note, 17q23.2 as identified by GISTIC is a peak region included in the previously defined amplicon 17q21.33-q25 that contains a number of outlier genes from 17q23.2 from Table 4.3. This suggests that the driver gene for this amplicon may be in this region.

To eliminate the possibility that the amplicons are just a surrogate for high histologic grade, we analyzed the ability of the presence of any amplicon in intermediate grade tumors to distinguish low and high risk breast cancers. Two risk categories were defined: *any amplicon:* a high risk set of samples with chromosomal amplifications at any of the four sites; and no *amplicon:* a low risk set of samples with no amplicons. Kaplan-Meier curves comparing recurrence in intermediate grade tumors for these two classes are shown in Figure 4.6A. It is clear from this figure that the amplicons are able to separate intermediate grade samples into significantly different risk classes. The Hazard ratio for intermediate grade tumors with amplicons versus intermediate grade tumors without amplicons was: (HR = 3.22, 95% CI = 1.6 – 6.5; P = 0.0012). Similar results were found for overall survival for intermediate grade tumors with any of the 4 amplicons versus cases without amplicons (Figure 4.6B: HR = 3.01, 95% CI = 1.2 – 7.6; P = 0.0200). These results show that the defined risk categories have a discriminatory power beyond that of classic histologic grade.

Average rates of distant metastasis at 10 years were calculated for the two risk classes in the gene expression data set of Loi et al. (80) (GSE6532). Similarly, average death rates at 10 years were available for the CGH array data set of Jonsson et al. (97) (GSE2133). Kaplan–Meier estimates (Table 4.8) for the proportion of patients in the high risk group who experienced an event (distant recurrence: 49.3% or death: 43.9%) were much higher than those in the low risk category (distant recurrence: 18.7% or death: 13.0%).

A validated marker of poor outcome in ER+ breast cancers with hormonal treatment is the Oncotype DX assay (16) described previously. This assay uses a linear combination of the expression of 21 genes to generate a single recurrence score, consisting of HER2, GRB7, GSTM1, CD68, BAG1, invasion markers (MMP11,CTSL2), proliferation markers (Ki67,STK15,Survivin,CCNB1,MYBL2) as well as estrogen and some reference markers. When the same gene panel is used to generate a relative Oncotype DX score (Figure 4.8) using normalized expression levels and published weights (16), we found that the presence of any of these amplicons was associated with higher recurrence scores, while tumors lacking the amplicons had lower recurrence scores. The relative Oncotype DX score computed using published weights on normalized gene expression values on the Loi et al. (80) dataset was also able to separate poor prognosis samples from good prognosis samples (Figure 4.7).

To test the hypothesis that regions 17q21.33-q25.1, 8p11.2 and 8q24.3 are likely to be amplified in ER+/HER2- breast tumor samples having high Oncotype DX recurrence scores, a set of 14 ER+/HER2- breast cancer samples with known Oncotype DX scores was evaluated for the presence of 17q21.33-q25.1, 8p11.2 or 8q24.3 amplifications using FISH with validated probes (for details of the procedure see Appendix B.6). Of these, 8 had high recurrence scores (RS) (>30) and 6 had low scores (<18). As shown in Table 4.9 and Figure 4.9, tumors with high RS had amplification of at least one of these regions, while almost all tumors with low RS did not exhibit any amplification at the mentioned chromosomal locations.

4.5 Potential for clinical use

Currently Oncotype DX or other quantitative grading methods are used to predict outcome and guide treatment for early stage ER+/HER2- breast cancer patients. A high recurrence score can identify patients likely to have poor outcome with hormonal therapy alone, who may benefit most from additional chemotherapy. However, Oncotype DX is an expensive assay (~$3000 per test), requires sending RNA to a central lab and about 30% of samples tested are assigned an “intermediate” risk class of dubious prognostic value. Moreover, it does not give any insight into new biological mechanisms driving poor prognosis, nor does it identify potential therapeutic targets.

On the other hand, the results presented here show that the presence of amplification in chromosomal regions 17q21.33-q25.1, 8p11.2 and 8q24.3 are strong markers of poor prognosis in ER+/HER2- breast cancers. Our analysis suggests that each amplicons has an associated risk equal to but independent of the risk of amplification of HER2. In our dataset, out of 44 patients who suffered distant metastasis within the first 4 years after diagnosis, 72.5% were predicted to have at least one of the four amplicons on chromosomes 8 and 17 while only 30% were predicted to have only 17q12 (HER2+) amplification.

The presence of chromosomal amplification on 17q21.33-q25.1, 8p11.2 or 8q24.3 in early ER+/HER2- tumors may be highly predictive of poor outcome in the setting of hormonal treatment. These amplicons are associated with higher expression of genes that drive a high Oncotype DX (ODx) recurrence score. Direct analysis of clinical specimens for amplification of these regions using FISH also demonstrated that the presence of amplification at least one of these loci is associated with high recurrence ODx scores, while tumors that lack any of these amplicons have low recurrence ODx scores. The amplicons may be valuable as strong biomarkers in predicting poor outcome under hormonal therapy of early stage ER+/HER2- breast cancers. They can be identified using a cost effective FISH assay on routine FFPE specimens. Since FFPE samples are routinely collected and archived in all hospitals, the association between the amplicons and poor prognosis can be easily validated in large retrospective and prospective studies and they can be easily identified and used in clinical practice.

In addition to their value as biomarkers, these chromosomal regions may contain driver oncogenes that could be specific therapeutic targets for patients harboring these amplicons. Such targets can be identified using routine knock-out and knock-in experiments on breast cancer cell lines. We discuss below some possible oncogenes which may be responsible for the observed phenotype.

Recent work by Turner et al. (100) identified that amplification of Fibroblast growth factor receptor 1 (FGFR1) is a driver for endocrine therapy resistance and a therapeutic target in breast cancer. FGFR1 is located in chromosomal region 8p11.2 and is part of one of the amplicons that we found associated with Tamoxifen resistance. In the present dataset its outlier profile is associated with poor survival with 1.8 hazard ratio and 0.046 log-rank p-value. This gene may well contribute to the observed effect of the 8p11.2 amplicon on cancer recurrence in ER+/HER2- breast cancer. Other genes in these amplicon regions that have been identified as putative oncogenes and therapeutic targets include U6 snRNA-associated Sm-like protein (LSM1), Wolf-Hirschhorn syndrome candidate 1-like 1(WHSC1L1) in region 8p11.2 and Heat shock transcription factor 1 (HSF1) in 8q24.3 (90; 95; 96).

As seen in Figure 4.2, the majority of outlier genes associated with poor prognosis on the *q* arm of chromosome 8, are clustered in the region 8q24.3, with the rest of them scattered all the way to 8q11.2. This suggests that in some cases, the whole *q* arm is amplified or else there are a number of different amplicons on 8q. Slightly more upstream of 8q24.3 there is a well known oncogene MYC, a key estrogen effector, that has been reported to induce Tamoxifen resistance when over-expressed (101). Although MYC could also contribute to the effect of this amplicon on resistance, it is not as strongly associated with differential survival (log-rank P = 0.042) as more distal genes, suggesting it may contribute to only a minority of cases containing this amplicon.

Another estrogen effector associated with Tamoxifen resistance is Cyclin D1 (CCND1, log-rank P value = 5.7e-6) (99) located on chromosomal band 11q13 another well known amplification site (102). However CCND1 is also a cell cycle marker and its expression is associated with proliferation. Thus its association with poor outcome may in part reflect its role in proliferation and not just as a driver oncogene. Intriguingly there are reports of an association between 11q13 amplification and amplification of 8p12 (97; 103; 104) in breast cancers, with some reports demonstrating a physical association between these domains (103). Thus 8p12 amplification may be functionally linked to 11q13 amplification in a subset of breast cancers.

Of the chromosomal regions identified in this study, 17q21.33-q25.1 is the least understood. Situated downstream of a much better known amplicon 17q12 (HER2+), it is a huge region known to be amplified and correlated with high grade tumors and poor prognosis (86). However, there is still no definite answer on indentifying the driver oncogenes in this region. Possible candidates are CLTC, involved in gene fusions in B-cell lymphomas and non-small cell lung carcinomas, RAD51C involved in DNA repair and homologous recombination, and PPM1D, a protein phosphatase, with only the first two significantly associated with Tamoxifen resistance in this dataset ($P < 0.05$).

In summary, the data presented here suggest that amplification of chromosomal regions 17q21.33-q25.1, 8p11.2 and 8q24.3 is strongly associated with intrinsic hormone resistance in early stage ER+/HER2- breast cancers, and correlates with high Oncotype DX recurrence scores. Similar to the HER2 amplicon, the presence of these amplicons may serve as a biomarker of poor prognosis in Luminal breast cancers. Moreover these chromosomal regions may contain genes whose over-expression may drive hormone independence in ER+ breast cancers.

Table 4.1: Microarray datasets used in this study

Clinical and pathological characteristics of all ER+, Tamoxifen treated patients in the dataset analyzed in our study.

Identifier	No. of samples	Grade ratio (1/2/3)	LN status ratio +/-()	Treatment
GUYT	87	17/37/16	58/29	Tamoxifen
OXFT	109	21/51/17	37/66	Tamoxifen
KIT	72	12/43/14	48/21	Tamoxifen

Table 4.2: Pathways and amplicons associated with Tamoxifen response and resistance.

Gene Ontology pathways/chromosomal location enrichment analysis results. P values were computed using Fisher's Exact Test.

	Over-expression	P values	Under-expression	P values
Good Outcome with Tamoxifen treatment	Immune response	1.61E-05	Cell cycle	1.10E-03
	Development	7.56E-08		
	Cell adhesion	1.68E-04		
Poor outcome with Tamoxifen treatment	Cell cycle	9.12E-07	Immune response	1.36E-05
	17q21.33-q25.1	3.87E-05	Cell adhesion	2.01E-08
	17q12	1.39E-08		
	8p11.2	1.11E-16		
	8q24.3	2.22E-16		

Table 4.3: Over-expressed genes in chromosomal regions 17q12, 17q21.33-q25.1, 8p11.2 and 8q24.3 associated with Tamoxifen resistance

Genes on chromosomes 8 and 17 associated with Tamoxifen resistance. Genes highlighted in red are known cancer related genes CLTC, WHSC1L1 and oncogenes ERBB2, LSM1 and HSF1.

Gene	Name	Cytoband	Start	End
STARD3	StAR-related lipid transfer (START) domain containing 3	chr17q12	35,046,940	35,073,248
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	chr17q12	35,110,005	35,122,109
GRB7	growth factor receptor-bound protein 7	chr17q12	35,152,029	35,156,782
GSDML	gasdermin B	chr17q12	35,326,079	35,328,194
PSMD3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3	chr17q12	35,390,607	35,407,732
PHB	prohibitin	chr17q21.33	44,836,413	44,847,246
SLC35B1	solute carrier family 35, member B1	chr17q21.33	45,133,688	45,140,281
SUPT4H1	suppressor of Ty 4 homolog 1 (S. cerevisiae)	chr17q22	53,778,283	53,784,556
RAD51C	RAD51 homolog C (S. cerevisiae)	chr17q22	54,124,987	54,127,694
CLTC	clathrin, heavy chain (Hc)	chr17q23.1	55,052,102	55,126,906
PTRH2	peptidyl-tRNA hydrolase 2	chr17q23.1	55,129,449	55,139,638
ABC1	ATP-binding cassette, sub-family A (ABC1), member 1	chr17q23.1	55,475,337	55,499,876
APPBP2	amyloid beta precursor protein (cytoplasmic tail) binding protein 2	chr17q23.2	55,875,300	55,958,365
TRIM37	tripartite motif-containing 37	chr17q23.2	57,059,999	57,184,266
USP32	ubiquitin specific peptidase 32	chr17q23.2	58,254,691	58,469,586
CYB561	cytochrome b-561	chr17q23.3	58,864,245	58,869,052
CCDC44	coiled-coil domain containing 44	chr17q23.3	59,038,377	59,039,456
PSMC5	proteasome (prosome, macropain) 26S subunit, ATPase, 5	chr17q23.3	59,258,832	59,263,111
PSMD12	proteasome (prosome, macropain) 26S subunit, non-ATPase, 12	chr17q24.2	62,764,494	62,793,171
KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	chr17q24.2	66,031,848	66,042,970
ICT1	immature colon carcinoma transcript 1	chr17q25.1	70,520,374	70,528,950
ATP5H	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit d	chr17q25.1	70,546,552	70,548,888
MRPS7	mitochondrial ribosomal protein S7	chr17q25.1	70,769,394	70,773,734
SAP30BP	SAP30 binding protein	chr17q25.1	71,175,038	71,214,431
SPFH2	ER lipid raft associated 2	chr8p11.2	37,713,267	37,734,476
PROSC	proline synthetase co-transcribed homolog (bacterial)	chr8p11.2	37,739,282	37,756,441
ASH2L	ash2 (absent, small, or homeotic)-like (Drosophila)	chr8p11.2	38,082,214	38,116,216
LSM1	LSM1 homolog, U6 small nuclear RNA associated (S. cerevisiae)	chr8p11.2	38,140,017	38,153,183
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1	chr8p11.2	38,293,091	38,358,947
BRF2	BRF2, subunit of RNA polymerase III transcription initiation factor, BRF1-like	chr8p12	37,821,053	37,826,512
DDHD2	DDHD domain containing 2	chr8p12	38,208,356	38,239,442

UBE2V2	ubiquitin-conjugating enzyme E2 variant 2	chr8q11.21	49,083,545	49,136,681
ATP6V1H	ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H	chr8q11.23	54,828,192	54,832,484
MRPL15	mitochondrial ribosomal protein L15	chr8q11.23	55,210,341	55,223,011
COPS5	COP9 constitutive photomorphogenic homolog subunit 5 (Arabidopsis)	chr8q13.2	68,117,869	68,136,905
TCEB1	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)	chr8q21.11	75,020,403	75,047,049
FAM82B	family with sequence similarity 82, member B	chr8q21.3	87,555,453	87,590,037
UQCRB	ubiquinol-cytochrome c reductase binding protein	chr8q22	97,312,308	97,316,963
POLR2K	polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa	chr8q22.2	101,232,001	101,235,407
ATP6V1C1	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	chr8q22.3	104,102,463	104,152,473
EBAG9	estrogen receptor binding site associated, antigen, 9	chr8q23	110,621,485	110,646,565
YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	chr8q23.1	102,001,097	102,033,426
ENY2	enhancer of yellow 2 homolog (Drosophila)	chr8q23.1	110,415,745	110,425,074
RAD21	RAD21 homolog (S. pombe)	chr8q24	117,927,353	117,956,221
SQLE	squalene epoxidase	chr8q24.1	126,100,439	126,102,952
MRPL13	mitochondrial ribosomal protein L13	chr8q24.12	121,477,267	121,526,557
SCRIB	scribbled homolog (Drosophila)	chr8q24.3	144,945,082	144,968,239
SIAHBP1	poly-U binding splicing factor 60KDa	chr8q24.3	144,970,536	144,983,471
GRINA	glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1 (glutamate binding)	chr8q24.3	145,136,247	145,139,570
EXOSC4	exosome component 4	chr8q24.3	145,205,516	145,207,538
CYC1	cytochrome c-1	chr8q24.3	145,221,982	145,224,415
SHARPIN	SHANK-associated RH domain interactor	chr8q24.3	145,225,527	145,230,852
C8orf30A	chromosome 8 open reading frame 30A	chr8q24.3	145,264,659	145,267,608
BOP1	block of proliferation 1	chr8q24.3	145,456,867	145,485,928
HSF1	heat shock transcription factor 1	chr8q24.3	145,497,218	145,498,193
FBXL6	F-box and leucine-rich repeat protein 6	chr8q24.3	145,549,899	145,552,940
GPR172A	G protein-coupled receptor 172A	chr8q24.3	145,553,131	145,555,738
VPS28	vacuolar protein sorting 28 homolog (S. cerevisiae)	chr8q24.3	145,619,807	145,623,174
RPL8	ribosomal protein L8	chr8q24.3	145,985,957	145,988,332
ZNF7	zinc finger protein 7	chr8q24.3	146,023,747	146,043,697
ZNF250	In multiple Geneids	chr8q24.3	146,076,967	146,079,026
C8orf33	chromosome 8 open reading frame 33	chr8q24.3	146,248,629	146,251,814

Table 4.4: Sample correlations between gene patterns associated with Tamoxifen resistance

Correlations amongst sample with up-regulation of the cell cycle pathway and each of the four amplicons associated with Tamoxifen resistance in the gene expression dataset (GSE6532) from Loi et al. (80). Values represent Phi coefficients measuring the strength of association between the group of samples that over-express cell cycle genes and amplicons 17q12, 17q21.33-q25.1, 8p11.2 and 8q24.3. The last column lists the percentage counts of ER+ samples with the associated pathway/amplicons. Marked in red are correlation values significant at $p < 0.01$ except for self correlations.

Note that the cell cycle pathway is correlated with each amplicon. However, the amplicons themselves are not correlated with each other. This suggests that they are independent markers of poor progression.

	cell cycle	**17q12**	**17q21.33-q25.1**	**8p11.2**	**8q24.3**	**Percent samples**
cell cycle	1.00	0.26	0.30	0.17	0.20	7.8%
17q12	0.26	1.00	0.18	0.01	0.00	12.7%
17q21.33-q25.1	0.30	0.18	1.00	0.07	0.23	13.1%
8p11.2	0.17	0.01	0.07	1.00	0.26	13.4%
8q24.3	0.20	0.00	0.23	0.26	1.00	9.0%

Table 4.5: Amplicon survival properties

Metastasis free survival in samples with amplicons 17q12, 17q21.33-q25.1, 8p11.2, 8q24.3 in the gene expression dataset (GSE22133) of Loi et al. (80). Hazard ratio and log-rank P values were computed with reference to samples without any amplicons.

Amplicon	Median time to recurrence (days)	Hazard ratio	95% CI	Log-rank P value
17q12	3355	4.09	3.84 – 21.99	6.3e-07
17q21.33-q25.1	—	3.14	2.17 – 13.62	3.0e-04
8p11.2	3795	3.75	3.18 – 18.31	5.7e-06
8q24.3	3468	4.29	4.32 – 34.08	2.2e-06

Table 4.6: Correlations between samples different amplicons in an independent CGH array dataset

Phi coefficients measuring the association between amplicons 17q12, 17q23.2, 8p11.2 and 8q24.3 in the test CGH dataset (GSE22133) from Jonsson et al. (97). The last column lists the percentage counts of ER+ samples with the associated amplicons. Marked in red are correlation values significant at P < 0.01 not including self correlations.

	17q12	17q23.2	8p11.2	8q24.3	Percent samples
17q12	1.00	0.32	0.12	0.27	51.8%
17q23.2	0.32	1.00	0.15	0.20	41.9%
8p11.2	0.01	0.15	1.00	0.22	45.9%
8q24.3	0.27	0.20	0.22	1.00	68.5%

Table 4.7: Amplicon survival properties in an independent CGH array dataset

Overall survival properties of samples with amplicons 17q12, 17q23.2, 8p11.2, 8q24.3 derived from the CGH array dataset (GSE22133) of Jonsson et al. (97). Hazard ratio and log-rank P values are relative to samples without amplicons.

Amplicon	Median survival (days)	Hazard ratio	95% CI	Log-rank P value
17q12	4356	2.61	1.51 – 5.51	6.8e-04
17q23.2	2879	3.02	1.76 – 5.18	7.3e-05
8p11.2	3813	2.65	1.48 – 4.74	1.3e-03
8q24.3	5800	2.12	1.24 – 3.65	6.7e-03

Table 4.8: Kaplan-Meier estimates of the rate of distant metastasis/death events at 10 years

Average rates of distant metastasis/death are compared between two risk categories: any amplicon vs. no amplicon present. Distant metastasis times are obtained for the gene expression dataset (GSE6532) of Loi et al. (80) while overall survival times are collected from the CGH array dataset (GSE22133) of Jonsson et al. (97).

Risk category	Rate of distant recurrence at 10 years	95% CI	Rate of death at 10 years	95% CI
Any amplicon	49.3%	36.9 - 61.7%	43.9%	36.1 - 51.7%
No amplicon	18.7%	11.7 - 25.7%	13.0%	2.30 - 23.7%

Table 4.9: FISH staining scores for ER+/HER2- breast cancer tissue samples

Fluorescence in situ hybridization (FISH) results for 14 paraffin embedded ER+/HER2- breast cancer samples. Scores are calculated as average number of spots over 20 cancer cells for each chromosomal location and the separated into amplified, not amplified and borderline classes as follows: (>4 amplified; 2-4 borderline; <2 not amplified). The last column lists the associated Oncotype DX score for each sample, 8 have high scores while 6 have low scores. Note that all samples with high Oncotype DX score (in red) have at least one associated amplicon, while samples with low Oncotype DX score have none.

17q23.1	8q24.3	8p11.2	Oncotype DX
amplified	amplified	amplified	46
not amplified	not amplified	amplified	42
borderline	not amplified	amplified	38
amplified	borderline	borderline	36
borderline	amplified	borderline	33
amplified	amplified	amplified	44
amplified	amplified	borderline	42
borderline	borderline	borderline	34
no signal	not amplified	not amplified	13
no signal	not amplified	not amplified	8
not amplified	not amplified	not amplified	5
borderline	not amplified	not amplified	12
not amplified	no signal	no signal	11
not amplified	not amplified	not amplified	11

Figure 4.1: PCA plots of high (A) and low (B) outliers

Outlier profiles of genes associated with differential survival were organized into two binary matrices B_1 and B_2 corresponding to high and low outlier values respectively. For both matrices, $B(i,j) = 1$ if *gene i* was an outlier in *sample j* and $B(i,j) = 0$ otherwise. These matrices were further pruned by iteratively eliminating *row i* and *column i* if *gene i* was not positively correlated with at least one other gene from the remaining set. The figure represents the projection of each gene's outlier profile on the first two principal components of the corresponding matrix. Clusters associated with good prognosis are circled blue while clusters associated with bad prognosis are circled red.

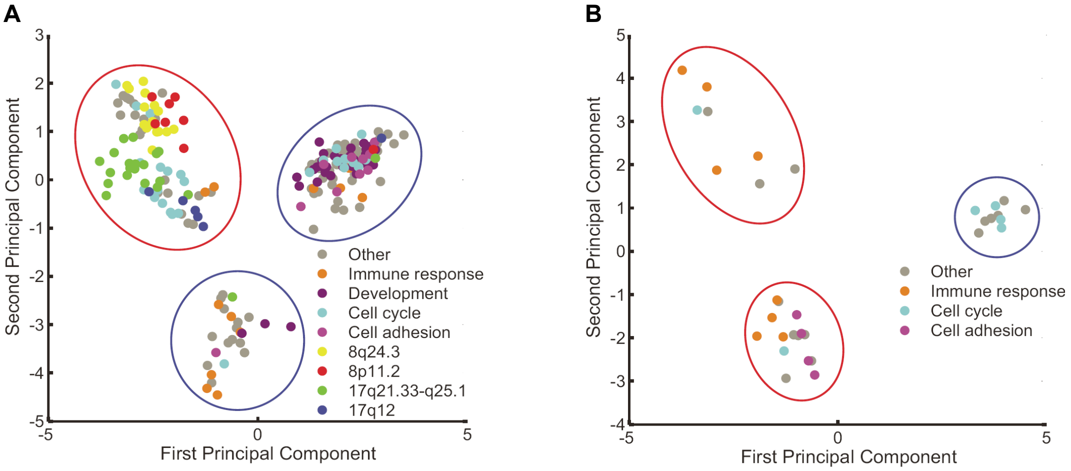


Figure 4.2: Clustergram of the correlation matrix between selected over-expressed genes associated with poor survival under Tamoxifen treatment

Calculating Phi coefficients for the distribution of high outliers between every two genes found to be associated with Tamoxifen resistance in Figure 1A produces a correlation matrix. This figure shows the resulting heatmap of the hierarchical clustering (Pearson correlation distance, complete linkage) of this correlation matrix. Genes in the same pathway or chromosomal region are clustered together as marked.

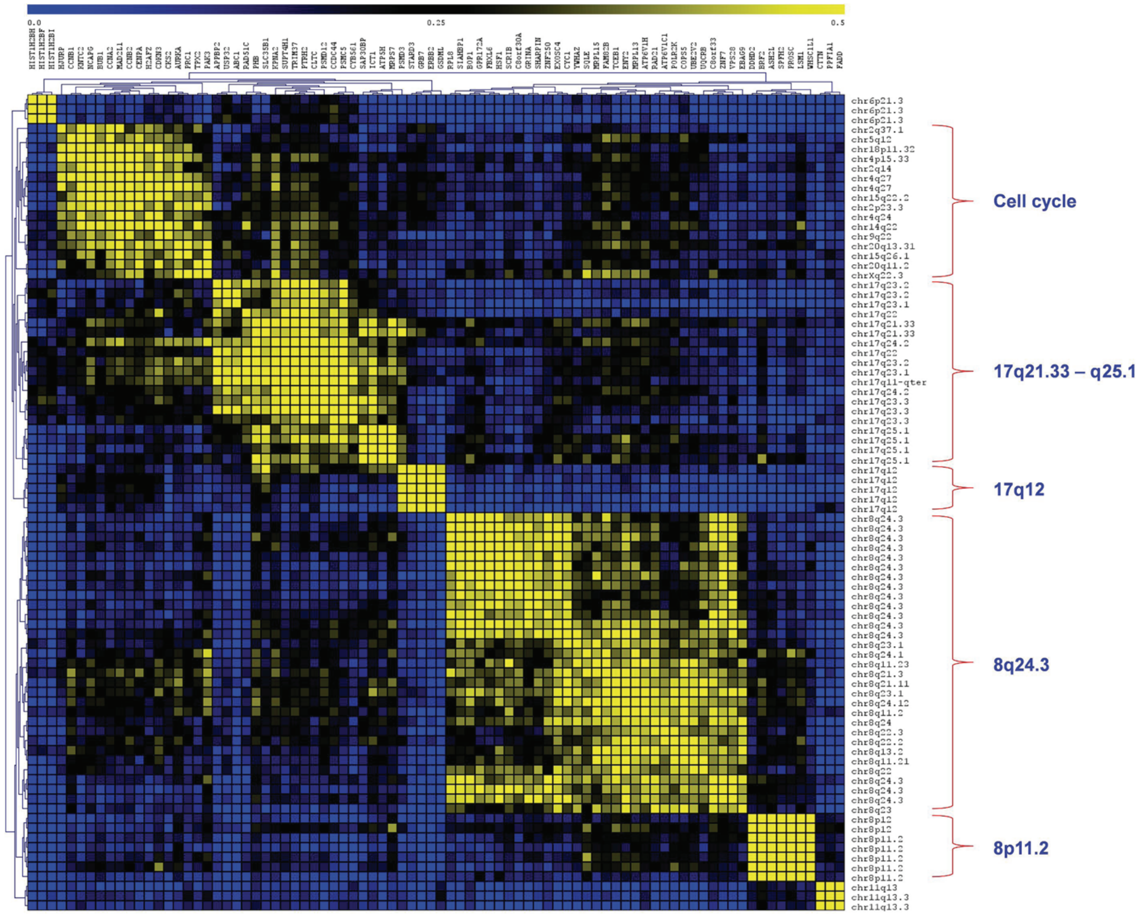


Figure 4.3: Patients with cell cycle pathway activation show poor survival outcome under Tamoxifen treatment

Kaplan-Meier curves of the samples enriched for over-expressed cell cycle genes versus the rest of samples that don't show this feature. Patients with cell cycle activated genes show a significant decrease in relapse free survival rate (HR = 9.71, 95% CI = 3.3 – 28.6; P < 0.0001).

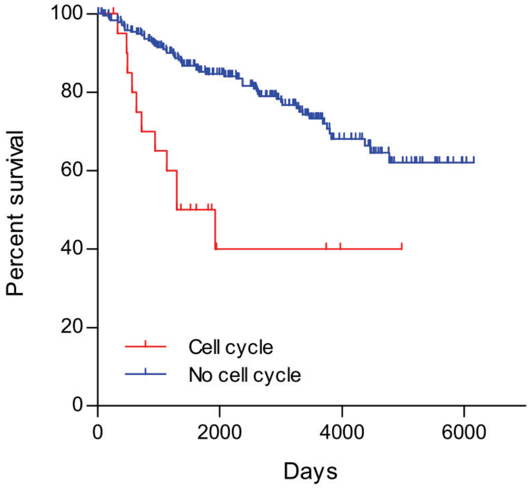


Figure 4.4: Patients with 17q12, 17q21.33-q25.1, 8p11.2 and 8q24.3 amplifications show poor survival outcome under Tamoxifen treatment

Kaplan-Meier curves of the samples with the 4 amplicons versus samples that don't have any of the chromosomal amplifications. Samples with enrichment of any of the four amplicons were identified by requiring at least 50% of gene markers in each group to be over-expressed, i.e. is marked as a high outlier in the respective sample. Patients that show any one of the chromosomal amplifications have significantly higher relapse rates at an overall log-rank P value < 0.0001. See Table 4.4 for additional detail.

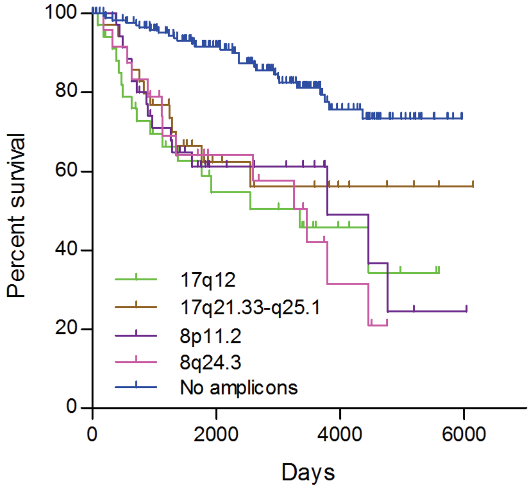


Figure 4.5: Patients with 17q12, 17q23.2, 8p11.2 and 8q24.3 amplifications also show poor survival outcome in an independent CGH array data set.

Kaplan-Meier curves of the samples with the 4 amplicons versus samples without these amplifications. Analysis of the CGH data identified amplification peaks at each of the four regions that overlap with the previously identified loci. Once again, we see that patients with any of the amplifications have significantly higher relapse rates at an overall log-rank P =0.0015. Additional details are in Table 4.6.

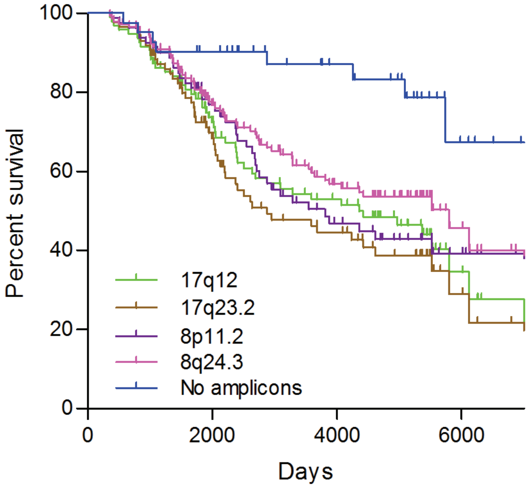


Figure 4.6: Analysis of intermediate grade tumors by presence of amplicons

Kaplan-Meier curves comparing rates of distant metastasis for patients with intermediate grade tumors who harbor one of the 4 amplicons versus patients with intermediate grade tumors without amplicons (A) in the training set GSE6532 (HR = 3.22, 95% CI = 1.6 – 6.5; P = 0.0012) and (B) in the test set GSE22133 (HR = 3.01, 95% CI = 1.2 – 7.6; P = 0.0200).

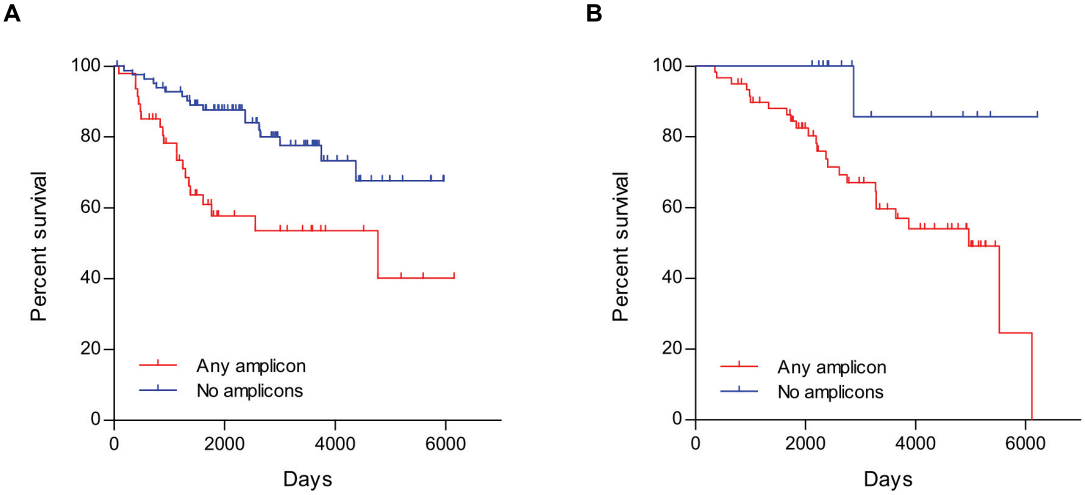


Figure 4.7: Relative Oncotype DX scores separate low risk from high risk breast cancers

Kaplan-Meier curves showing significantly lower survival (HR = 2.81, 95% CI = 1.7 – 4.5; P < 0.0001) for tumor samples with high Oncotype DX scores (ODx score > 0) versus low Oncotype DX scores (ODx score < 0). The scores were computed from normalized gene expression data using published weights (16).

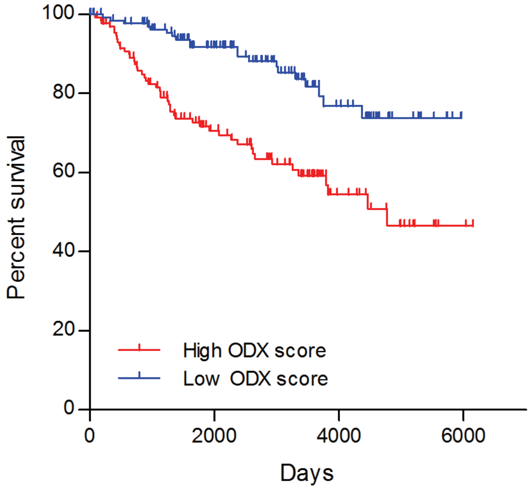


Figure 4.8: Relative Oncotype DX recurrence scores vs. presence/absence of amplicons.

Oncotype DX scores were inferred from the gene expression data using published weights and normalized expression values for the 21 genes in the Oncotype DX panel. The figure shows the range of these scores in sets of patients with each amplicon and in patients without any amplicons. Note that the Oncotype DX scores are highest in patients with amplicons in 17q12 or with amplification of cell cycle genes, because HER2 and proliferation genes are included in the Oncotype DX panel. However, patients with amplifications at 17q23.2, 8p11.2 and 8q24.3 have intermediate recurrence scores. This suggests that they are enriched in samples in the "intermediate" risk class of Oncotype DX. This shows that identification of these amplicons gives useful clinical information beyond what is the risk score of Oncotype DX.

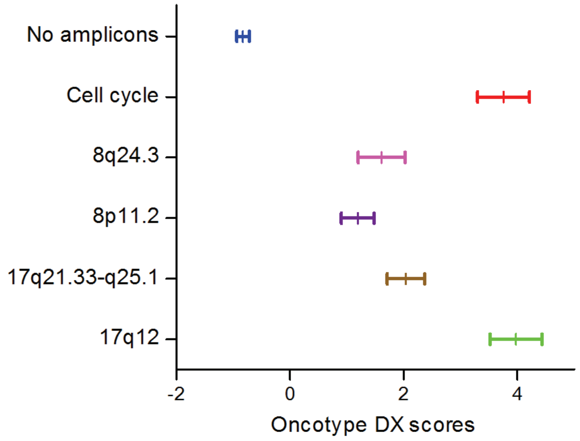
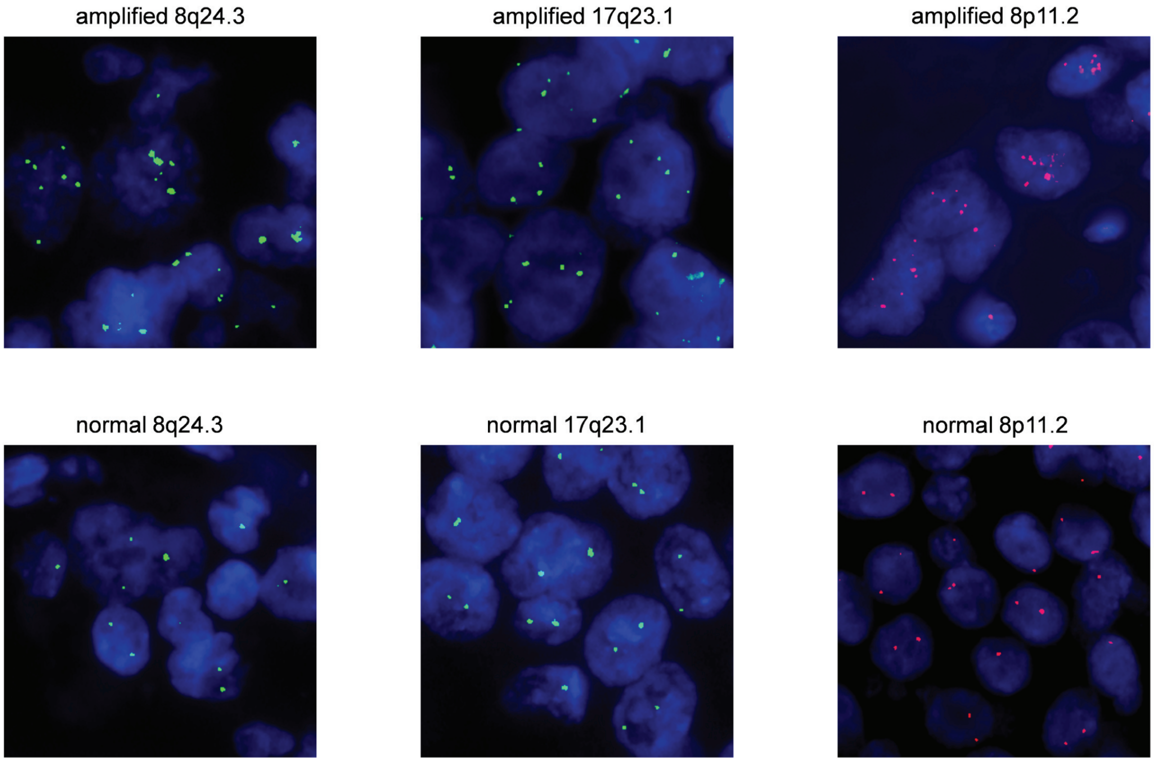


Figure 4.9: Fluorescent in situ hybridizations (FISH)

FISH images from experiments performed on breast cancer tissue slides with and without each amplicon. The FISH assay was done using probes specific to the amplicon chromosomal regions 17q23.2, 8p11.2 and 8q24.3.



Chapter 5:Conclusions and Outlook

"There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."
Charles Darwin (1809-1882)

5.1 Conclusions

Unraveling the intricate mechanisms that drive the growth of breast tumor cells has proven to be a formidable task. The current classification of breast cancer based on clinical markers correlates well with molecular classes based on gene expression analyses, and this in turns drives treatment. However, in clinical practice, the distinctions between different types of breast cancers are often not clear-cut, and cases often exhibit molecular patterns associated with a combination of the known subtypes. For example, Luminal (ER+) breast tumors can also be HER2+, showing chromosomal amplification at the 17q12 loci. Other known genomic aberrations like 8p12, 8q24.21, 8q24.3 and 11q13.3, identified in this thesis as relevant in clinical management of disease, are present in patients across all subtypes (97). This indicates that there is considerable more heterogeneity between breast cancers than previously thought, and treatment should be directed accordingly, by careful molecular profiling of each case.

ERBB2 (HER2) is one of a handful of molecular targets for which FDA approved drugs exist. Monoclonal antibodies like Trastuzumab, Cetuximab and tyrosine kinase inhibitor Lapatinib are common targeted therapies for HER2+ breast cancers. The success of these treatments, verified in large scale clinical trials, led to an avalanche of new candidate molecular targets in different subsets of breast tumors. Some of them have been predicted by the work presented here in Chapter 2, like epidermal growth factor receptor EGFR for high risk ER+ tumors (Luminal B); FOS, TGF beta receptor 2, ETS-related genes ERG, ELK3 and ETS2 for Luminal A tumors; PIM2 and a number of SRC tyrosine kinases predicted to be good therapeutic targets in subsets of Basal-like and HER2+ breast tumors.

EGRF is already being used as a partial target together with ERBB2 (HER2) for the treatment of HER2+ breast cancers. Lapatinib is a drug that targets both proteins expressed by these genes and has shown its efficacy in HER2+ cases. However, other drugs like Cetuximab or Gefitinib are in different phases of clinical trials for the treatment of other classes of breast tumors. Another class of drugs that are in clinical trials for breast cancer treatment is the one that targets SRC kinases. Dasatnib is one such cancer drug that has shown promising early results on triple negative (ER-/PR-/HER2-) breast cancer cell lines (46; 47).

An equally important issue in managing breast cancer cases, besides the availability of adjuvant, neoadjuvant or systemic therapies, is the ability to predict disease drug response. Classic methods based on IHC or FISH staining of tumors tissue slides have been very useful in predicting response to HER2 targeted therapies as well as estrogen pathway inhibitors. In addition, these methods are cheap and the assay can be done quickly, in the same hospital where the patient is being treated. A pathologist usually scores the amount of HER2, ER and PR protein expressed in the cancer cells. An equally efficient alternate method to identify ERBB2 over-expression is to use a FISH assay to assess chromosomal amplification of the 17q12 locus.

Until a few years ago, the standard of care in the evaluation of breast cancers to guide the clinician in determining appropriate treatment were standard techniques such as histo-pathological examination of the tumor and immunohistochemical measurements of ER, PR, and HER2. The advent of high throughput gene expression analysis and more recently, sequencing techniques, has given the clinician additional information about the underlying molecular features of the tumor and the possible risk of progression and possible failure of hormone therapy. Some of these new methods, such as the Oncotype DX® assay by Genomic Health Inc. and MapQuant DX™ by Ipsogen Inc. are already in use in the clinic. However, as we have shown in Chapter 4, these assays do not adequately identify all the risk associated molecular events in breast cancer. We have identified three additional regions of chromosomal amplifications in 8p11.2, 8q24.3 and 17q23.2 that confer additional risk of progression, similar to the HER2 amplicon on 17q12, which are not currently assessed in the clinic, which can be easily identified by relatively inexpensive methods from FFPE specimens. Although the biology of these amplicons remains to be understood, they can be used, with potentially tremendous benefit to the patients, as markers of risk. Their identification would help identify patients currently unidentifiable, who may benefit from additional chemotherapy. An additional benefit may result from an understanding of their biology, which may reveal gene and protein targets for the development of novel therapeutics.

5.2 Outlook

There is a long road winding from research laboratories to clinical practice. In this thesis, we developed techniques for the analysis of large high throughput gene expression datasets and used them to identify therapeutic targets in Basal-like breast cancers, for which no systemic therapy currently exists. We also discovered markers associated with differential survival which identifies patients likely to have early recurrence under standard therapy, who may benefit from additional chemotherapy. Our discoveries can easily be validated in larger retrospective and prospective datasets and easily and cost effectively implemented in the clinic, to the benefit of the patients.

The discoveries in this thesis need to be further verified in large independent datasets and clinical trials. The gene targets listed in Table 2.2 could be tested in breast cancer cell lines, using techniques similar to the methods used in Chapter 2 and Appendix B to validate YES1 as a target in Basal-like breast cancers. This could be followed by validation/testing in a mouse models, followed by drug development and a clinical trial before use in the clinic.

We are hoping to be able to obtain experimental validation of the chromosomal markers (amplicons) discovered in Chapter 4. To this end, FISH probes corresponding to 8p11.2, 8q24.3 and 17q23.2 loci will be hybridized to a large number of tumor tissue samples from ER+, Tamoxifen only treated patients with long term follow up information. Relapse risk will be assessed as a function of the presence/absence of each of the amplicons, and the value of this assessment will be compared to standard clinical markers such as stage, grade and assays like Oncotype DX.

We hope that some of the discoveries in this thesis will eventually be incorporated into clinical practice because they have the potential to assist the clinician in the management of breast cancers and to markedly improve the quality of life of patients. As more data becomes available for other types of cancers, our hope is that collaborations between researchers from diverse backgrounds and clinicians, that led to this thesis, will become routine and will reveal many new and useful tools for diagnosis, prognosis and the development of effective therapies.

Appendix A: Derivation of the Gene Centrality Score

A.1 Datasets and pre-processing

Previously published breast cancer microarray datasets (accession numbers GSE2034 (37) and GSE4922 (40)) were downloaded from the Gene Expression Omnibus website (GEO:www.ncbi.nlm.nih.gov/geo). The first dataset (GSE2034) comes from a study of Wang et al. (37) and consists of gene expression data from 286 lymph node negative patients treated with surgery and radiation alone and followed for up to 150 months after treatment, with recorded events for distant metastasis. ER status was available; HER2 status and histologic grade were known but not provided. The second dataset (GSE4922) from (40) consisted of 249 primary invasive breast tumors. In this cohort, 64% of patients were lymph node negative and were treated with surgery and radiation alone. The remaining were lymph node positive and received systemic adjuvant polychemotherapy consisting of intravenous cyclophosphamide, methotrexate and 5-fluorouracil (105). Histological grade, tumor size, ER and P53 biomarker information were available for each sample together with up to 153 months of follow up information for distant metastasis.

The arrays were MAS 5.0 normalized and only probes present in both datasets were retained. Multiple probes corresponding to the same gene were compressed to the one with the biggest median over all arrays after taking log2 of each intensity value. In addition, every array was scaled to median zero by subtracting the median of each array from every expression value.

Robust unsupervised consensus ensemble clustering methods previously applied to the data of Wang et al. (37) identified six core breast cancer subtypes(13; 38; 39): Two ER+,HER2- subtypes labeled Luminal A (LA) and Luminal B (LB), two Basal subtypes BA1 and BA2, both ER-,HER2- and two HER2+ subtypes labeled HER2I and HER2NI. The samples in the Ivshina et al. (40) dataset were assigned subtype as follows: HER2+ samples were identified based on Chr-17q12 amplification using expression levels of ERBB2, GRB7, STARD3 and PPARBP. Gene expression values in both datasets were normalized by subtracting the median and dividing by the median absolute deviation. HER2+ samples were identified as those over-expressing ERBB2 and at least two others from the set GERB7, STARD3 and PPARBP. After HER2 samples were identified, the two datasets were merged using a method called Distance Weighted Discrimination (DWD) (106) which corrects for biases arising from different experimental conditions. The assignment of samples in (15) to subtypes was done by comparison to mean expression profiles (centroids) across all genes for each subtype, using the classification of the Wang dataset as the standard. This method, called Single Sample Predictor (49), calculates a "distance" from each sample to mean expression values of samples in labeled sets using Euclidean distance or Pearson correlation and assigns them to the set for which this distance the smallest. Samples with inconsistent class labels for different distance metrics were discarded.

A.2 Meta-analysis of outliers

To minimize sample size bias, 10 arrays were randomly picked from each breast cancer subtype and combined into a reduced gene expression table $\mathbf{G} = [g_{ij}]_{nx60}$ where n is the total number of genes in each array. For each gene, the expression values were median centered and then divided by the median absolute deviation (MAD) as described in Tomlins et al. (81): $g'_{ij} = \frac{g_{ij} - median(g_i)}{MAD(g_i)}$. Median and MAD were used here instead of the usual mean and standard deviation because they are less influenced by the presence of outliers. Outlier scores (θ) were defined for each gene and class as the percentage of high outlier values across each breast cancer subtype: $\theta = \frac{1}{N}\sum_j^N \Delta_j$ where $N = 10$, $\Delta_j = 1$ if $g'_j > 1$ and $\Delta_j = 0$ otherwise.

The sampling procedure was repeated 1000 times, separately for the two datasets (GSE2034 and GSE4922), and in each sampling, outlier scores was generated for each gene in each subtype. At the end of this analysis, every gene had two associated distributions of outlier scores for each subtype that could now be combined into a single consensus score. This meta-outlier score was calculated, using the method of Cochran (109), as a weighted mean of the average outlier scores from the two distributions ($\overline{\theta}_1$ and $\overline{\theta}_2$), where the weights are the inverse of the corresponding variances σ_k^2:

$$\hat{\theta} = \sum_k^2 w_k \overline{\theta}_k \Big/ \sum_k^2 w_k \ , \ w_k = 1/\sigma_k^2 \qquad [2.1]$$

Each gene was now assigned a meta-outlier score for each of the 6 breast cancer classes (BA1, BA2, HER2I, HER2NI, LA and LB) which assesses whether it is over or under expressed in that subtype.

A.3 Meta-analysis of correlations

For each dataset, Pearson correlations were computed between all pairs of genes within each subtype. Assuming a common underlying population correlation between every two genes in each class, we calculated meta-correlation values by first transforming each Pearson correlation r with a Fisher z-transform $z = \frac{1}{2}\ln\frac{1+r}{1-r}$. The method usually used to estimate a common correlation value across multiple datasets (110) is to calculate the weighted average $\hat{z} = \sum_k^2 w_k z_k \Big/ \sum_k^2 w_k$ [2.2] where z_1 and z_2 are z-transformed Pearson correlations between any two genes from datasets GSE2034 and respectively GSE4922. The weights are $w_k = n_k - 3$ where n_1 and n_2 are number of samples used to calculate the correlations in the two datasets. Since correlation values calculated from gene expression arrays are often noisy (111), a homogeneity chi-squared statistic $Q = \sum_k^2 (n_k - 3)(z_k - \hat{z})^2$ was used to reject inconsistent correlation values. This statistic is chi-squared distributed (110) with $K-1$ degrees of freedom, where $K = 2$ is the total number of studies. Based on this statistic, the degree of inconsistencies can be measured as $I^2 = 100\% \times (Q - df)/Q$ where $df = K - 1$is the number of degrees of freedom. The measure I^2 describes the percentage of total variation across studies that is due to actual heterogeneity (signal) rather than chance (112).

Meta-correlation values were calculated using the inverse Fisher z-transform: $\hat{r} = \frac{\exp(2\hat{z}) - 1}{\exp(2\hat{z}) + 1}$ and the ones for which $I^2 > 50\%$ were discarded. This ensures that more than 50% of the observed variations were due to true heterogeneity.

A.4 Gene centrality

Eigenvector centrality (113; 114) is a measure of the importance of a node in a network. Relative scores are assigned to each node based on the idea that connections to nodes with high scores should contribute more to the score of the node in question than equal connections to low scoring nodes. Similarly, gene centrality is a measure of the importance of a gene in a modified gene network, where directed edges between nodes (genes) are weighted by a positive measure of the over-expression of the target gene as shown in the toy gene network from Figure 2.1. More generally, connections between nodes can be real positive numbers representing connection strengths.

Let $\mathbf{A} = [a_{ij}]_{nxn}$ be an adjacency matrix where every element $a_{ij} = \hat{r}_{ij}^2$ is the square of the meta-correlations between all genes within a subtype. (For more detailed explanation of the material here, refer to (115) and (110)). This is the inverse of $\hat{z}$ from equation [2.2] and measures how much of the variance in the expression of gene g_i can be explained by gene g_j. It provides an intuitive measure of the "connection" strength between the two genes. Let s_i be the centrality of gene g_i with associated meta-outlier score $\hat{\theta}_i$ as described in equation [2.1]. Then the centrality of gene g_i is proportional to the sum of scores of all genes modulated by the "connection" strength with each one of them; and also proportional to its own measure of over-expression:

$$s_i = \frac{1}{\lambda}\hat{\theta}_i \sum\nolimits_j^n a_{ij} s_j \qquad [2.3]$$

where λ is the constant of proportionality to be determined. Let $\mathbf{\Theta} = diag(\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta}_n)$ be the diagonal matrix with meta-outlier scores of all genes on the main diagonal and

$\mathbf{s} = [s_i]_{nx1}$ be a column vector of all gene centrality scores, then the previous equation [2.3] can be rewritten as an eigenvector problem:

$$\mathbf{\Theta A s} = \lambda \mathbf{s} \qquad [2.4]$$

Equation [2.4] identifies λ as an eigenvalue of the product of matrices $\mathbf{\Theta}$ and $\mathbf{A}$. In general, there will be many different eigenvalues λ for which an eigenvector solution $\mathbf{s}$ exists, and they describe the behavior of the discrete linear dynamical system:

$$\mathbf{x}_{m+1} = \mathbf{\Theta A x}_m \qquad [2.5]$$

Where,

$$\mathbf{x}_m = c_1 \lambda_1^m \mathbf{s}_1 + c_2 \lambda_2^m \mathbf{s}_2 + \ldots + c_n \lambda_n^m \mathbf{s}_n \qquad [2.6]$$

The linear system defined in [2.5] is completely characterized by the matrix $\mathbf{\Theta A}$ which can be viewed as an adjacency matrix of a directed graph whose nodes represent genes and an edge from gene g_i to gene g_j is equal to $\hat{\theta}_j \hat{r}_{ij}^2$.

If $\mathbf{\Theta A}$ is a primitive matrix (see below for a definition), Perron-Frobenius Theorem (115) states that it has a unique positive largest eigenvalue whose eigenvector has only positive entries. This guarantees that the maximal eigenvalue in equation [2.6] will dominate the long term behavior of the system defined by equation [2.5]. This property justifies choosing the corresponding eigenvector as a measure of gene centrality. Each element in this vector is a centrality score and is proportional to the long term “state” of the associated node in the gene network.

A primitive matrix $\mathbf{M}$ is a non-negative square matrix such that there is a number k for which all elements of $\mathbf{M}^k$ are strictly positive. Since $\mathbf{\Theta A}$ is not always a primitive matrix, minor modification in its structure need to be made for the analysis above to apply. A sufficient condition for a non-negative matrix to be primitive is that the matrix must be irreducible and have strictly positive elements along the main diagonal. An irreducible matrix is equivalent, in graph theoretic terms, to a fully connected network. In the case of a graph it is thus sufficient to eliminate unconnected nodes until the remaining ones are fully connected and add self loops to one or all nodes as shown in Figure 2.1. Similarly, to transform $\mathbf{\Theta A}$ to a primitive matrix, it is sufficient to make all elements on the principal diagonal positive, in this case equal to $\hat{\theta}_i > 0$, and discard unconnected nodes.

Separate $\mathbf{\Theta A}$ matrices were calculated for each breast cancer subtype (BA1, BA2, HER2I, HER2NI, LA and LB) and the principal eigenvector determined. Genes that were eliminated to make $\mathbf{\Theta A}$ primitive were assigned centrality score zero, while the rest were assigned scores from the dominant eigenvector. To allow the comparison of centrality scores between subtypes, the scores for each subtype were normalized by dividing by the median score across all genes.

Appendix B: Experimental methods and conditions

B.1 Immunohistochemistry

Anti-YES1 antibody (Santa Cruz) was first optimized on human breast tissue microarray slides using Discovery XT (Ventana Medical Systems) automated immunostainer. Before hybridization, breast cancer tissue slides were deparaffinized in a 60°C oven for 1 hour followed by 3x5 minutes in xylene, and hydrated in 100%, 80%, 70% ethanol and dH_2O. Antigen retrieval was performed by using Cell Conditioning Solution (Ventana Medical Systems) for 72 minutes. Anti-c-Yes antibody was applied at a dilution of 1:30 and incubated at 37°C for 1 hour, followed by 12 minutes with a universal secondary antibody (Ventana Medical Systems). DABMap (Ventana Medical Systems) was used for chromogenic detection after which slides were counterstained with Hematoxylin (Richard-Allan Scientific) and dehydrated in 70%, 80%, and 100% ethanol.

B.2 Cell culture conditions

MDA468 and MDA231 cell were maintained in DMEM/F12 (Gibco) supplemented with 5% Fetal Bovine Serum(FBS) (Gibco), 1% amino acids (Cellgro), 1% sodium pyruvate (Sigma); BT549 and SKBR3 cells were maintained in RPMI 1640 (ATCC) with 10% FBS; MCF7 and HEK-293T in DMEM (Gibco) with 10% FBS and MCF10A were grown in DMEM/F12 to which the following were added: 5% horse serum (Invitrogen), 20 ng/ml epidermal growth factor (Invitrogen), 100 ng/ml cholera toxin (Sigma), 0.01 mg/ml insulin (Sigma) and 500 ng/ml hydrocortisone (Sigma). With the exception of

HEK-293T cell culture media, all presented solutions had an addition of 1% penicillin/streptomycin (Gibco).

B.3 Immunoblotting

After incubation, cells were washed in cold (4□°C) PBS solution then kept on ice with NETN buffer (20 mM Tris, 150 mM NaCl, 1mM EDTA, 0.5% NP40, 1x Protease inhibitor cocktail (Sigma)) for 15 minutes. Cells were then scraped and collected in 1.5 ml tubes, incubated on ice for an Supplementary 5 minutes. Whole cell protein was extracted by sonication followed by 14,000 rpm centrifugation for 10 minutes. The supernatant was then collected and quantified by using a Bradford(116) based protein assay (Bio-Rad). After loading 25-50 μg protein onto 10% polyacrylamide gels they were subject to electrophoresis, transferred to PVDF membranes (Bio-Rad) and probed with antibodies against YES1 (1:1000, BD Transduction Laboratories) and GAPDH (1:5000, Abcam).

B.4 Lentivirus production

To suppress YES1 we introduced shRNA specific for the following sequences using pLKO.1 lentiviral vectors (117) acquired from Open Biosystems:

shYES1 #1	CCAGCCTACATTCACTTCTAA
shYES1 #2	ACCACGAAAGTAGCAATCAAA
shYES1 #3	CCTCGAGAATCTTTGCGACTA

A standard 18bp non-hairpin control (CCGCAGGTATGCACGCGT) was also acquired from Addgene together with psPAX2 packaging plasmid and pMD2.G envelope plasmid. Lentiviruses were produced by transiently transfecting individual shRNA constructs together with packaging and envelope plasmids into HEK-293T cells using Fugene 6 (Roche). Viral supernatants were collected and passed through 0.45 µm syringe filters.

B.5 Cell proliferation assays

Cells were plated in 6 cm culture dishes and grown in the incubator until they were 70% confluent. After changing to fresh culture media, 8 µg/ml of polybrane (Millipore) was added together with 0.5 ml of each of the previously prepared lentiviral solutions to separate dishes: one for the lentivirus containing the scrambled shRNA (shSRC) and one corresponding to the lentivirus designed to knock-down the expression of YES1 (shYES1). After 24 hours the media containing viral particles was replaced with fresh media to which 3 µg/ml puromycin (Sigma) was added in order to select for infected cells. The cells were kept on growing for 3-4 days until a stable population was obtained. Cells expressing shYES1 and shSCR were separately plated in triplicates in 12-well plates in the following quantities: 50x10^3 cells for MDA231, BT549, MCF10A; 25x10^3 cells for MDA468, SKBR3; and 10x10^3 cells for MCF7. After 6 days of growing in specific media supplemented by 3 µg/ml puromycin, cells in each well were collected and counted by trypan blue exclusion using a Beckman Coulter counter.

B.6 Fluorescence in situ hybridization (FISH)

Prelabeled FISH probes for BAC clones RP11-1065N2, RP11-90P5 and RP11-1136N16 were purchased from Empire Genomics, Buffalo, NY and tested on metaphase chromosome spreads. They successfully hybridized to corresponding chromosomal locations 17q23.1, 8p11.2 and 8q24.3. FISH experiments were further performed on 14 4µm paraffin embedded breast cancer tissue slides, collected from women diagnose with ER+/HER2- breast cancers between 2007 and 2009 at Robert Wood Johnson University Hospital, New Jersey, USA. Out of all samples 8 had high Oncotype DX scores (>30) and 6 had low scores (<18). Oncotype DX scores for these samples were independently determined by Genomic Health, Redwood City, CA.

Before hybridization, tissue sections were deparaffinized in pepsin solution, fixed with formaldehyde, and dehydrated in 70%, 80%, and 100% ethanol followed by denaturation at 83°C for 3 min on hybrite (Vysis, Downers Grove, IL, USA). Hybridization was performed on hybrite for 16-24 hours at 37□°C, and then slides were washed first with 4x SSC for 3 min at 37°C then with 0.1% NP-40 (Vysis, Downers Grove, IL, USA) for 30 sec at room temperature.

Slides were scored for chromosomal amplification by counting signals in 20 tumor cells and then reporting the average number of spots/cell.

References

1. Osborne C. Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. The Oncologist. 2004;9(4):361-377.

2. Boehm JS, Zhao JJ, Yao J, et al. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. Cell. 2007;129(6):1065-79.

3. Neuman E, Ladha MH, Lin N, et al. Cyclin D1 stimulation of estrogen receptor transcriptional activity independent of cdk4. Molecular and cellular biology. 1997;17(9):5338-47.

4. Zwijsen RM, Wientjens E, Klompmaker R, van Der Sman J, Bernards R, Michalides RJ. CDK-independent activation of estrogen receptor by cyclin D1. Cell. 1997;88(3):405-15.

5. Coles C, Condie A, Chetty U, Michael Steel C, John Evans H, Prosser J. p53 Mutations in Breast Cancer. Cancer Res. 1992;52(19):5291-5298.

6. Chen S, Parmigiani G. Meta-analysis of BRCA1 and BRCA2 penetrance. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25(11):1329-33.

7. Maughan KL, Lutterbie MA, Ham PS. Treatment of breast cancer. American family physician. 2010;81(11):1339-46.

8. Elston C, Ellis I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-410.

9. Henson DE, Ries L, Freedman LS, Carriaga M. Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. The basis for a prognostic index. Cancer. 1991;68(10):2142-9.

10. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. British journal of cancer. 1957;11(3):359-77.

11. Wood AJ, Osborne CK. Tamoxifen in the Treatment of Breast Cancer. New England Journal of Medicine. 2009;339(22):1609-1618.

12. American Cancer Society. Breast Cancer Facts & Figures 2009-2010. 2010;

13. Alexe G, Dalgin GS, Scanfeld D, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer research. 2007;67(22):10669-76.

14. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

15. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(19):10869-74.

16. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine. 2004;351(27):2817-26.

17. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute. 2006;98(4):262-72.

18. Van De Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine. 2002;347(25):1999-2009.

19. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

20. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(19):10869-74.

21. Rakha EA, El-Sayed ME, Green AR, et al. Biologic and clinical characteristics of breast cancer with single hormone receptor positive phenotype. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25(30):4772-8.

22. Melchor L, Honrado E, García MJ, et al. Distinct genomic aberration patterns are found in familial breast cancer associated with different immunohistochemical subtypes. Oncogene. 2008;27(22):3165-75.

23. Richardson AL, Wang ZC, De Nicolo A, et al. X chromosomal abnormalities in basal-like human breast cancer. Cancer cell. 2006;9(2):121-32.

24. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet. 2005;365(9472):1687-1717.

25. The Breast International Group (BIG). A Comparison of Letrozole and Tamoxifen in Postmenopausal Women with Early Breast Cancer. New England Journal of Medicine. 2009;353(26):2747-2757.

26. Howell A, Cuzick J, Baum M, et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. The Lancet. 2005;365(9453):60-62.

27. Goss PE, Ingle JN, Martino S, et al. A Randomized Trial of Letrozole in Postmenopausal Women after Five Years of Tamoxifen Therapy for Early-Stage Breast Cancer. New England Journal of Medicine. 2009;349(19):1793-1802.

28. Romond EH, Perez EA, Bryant J, et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. The New England journal of medicine. 2005;353(16):1673-84.

29. Clark GM, McGuire WL. Steroid receptors and other prognostic factors in primary breast cancer. Seminars in oncology. 1988;15(2 Suppl 1):20-5.

30. Gnant M, Mlineritsch B, Schippinger W, et al. Endocrine therapy plus zoledronic acid in premenopausal breast cancer. The New England journal of medicine. 2009;360(7):679-91.

31. Dellapasqua S, Bertolini F, Bagnardi V, et al. Metronomic cyclophosphamide and capecitabine combined with bevacizumab in advanced breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(30):4899-905.

32. Colleoni M, Orlando L, Sanna G, et al. Metronomic low-dose oral cyclophosphamide and methotrexate plus or minus thalidomide in metastatic breast cancer: antitumor activity and biological effects. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2006;17(2):232-8.

33. Fong PC, Boss DS, Yap TA, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. The New England journal of medicine. 2009;361(2):123-34.

34. Sotiriou C, Pusztai L. Gene-Expression Signatures in Breast Cancer. New England Journal of Medicine. 2009;360(8):790-800.

35. Cheang MC, Voduc D, Bajdik C, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008;14(5):1368-76.

36. Siegal ML, Promislow DE, Bergman A. Functional and evolutionary inference in gene networks: does topology matter? Genetica. 2007;129(1):83-103.

37. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005;365(9460):671-679.

38. Dalgin GS, Alexe G, Scanfeld D, et al. Portraits of breast cancer progression. BMC bioinformatics. 2007;8291.

39. Alexe G, Dalgin GS, Ramaswamy R, Delisi C, Bhanot G. Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns. Cancer informatics. 2007;243-74.

40. Ivshina AV, George J, Senko O, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer research. 2006;66(21):10292-301.

41. Sugawara K, Sugawara I, Sukegawa J, et al. Distribution of c-yes-1 gene product in various cells and tissues. British journal of cancer. 1991;63(4):508-13.

42. Han J, Kamber M. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann; 2001.

43. Spearman C. The proof and measurement of association between two rings. Ameri. J. Psychol. 1904;(15):72-101.

44. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938;30(1):81 - 93.

45. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics. 2006;7 Suppl 1(Suppl 1):S7.

46. Huang F, Reeves K, Han X, et al. Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection. Cancer research. 2007;67(5):2226-38.

47. Finn RS, Dering J, Ginther C, et al. Dasatinib, an orally active small molecule inhibitor of both the src and abl kinases, selectively inhibits growth of basal-type/“triple-negative” breast cancer cell lines growing in vitro. Breast Cancer Research and Treatment. 2007;105(3):319-326.

48. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8418-23.

49. Hu Z, Fan C, Oh D, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics. 2006;7(1):96.

50. Alexe G, Dalgin GS, Scanfeld D, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer research. 2007;67(22):10669-76.

51. Strehl A, Ghosh J. Cluster ensembles --- a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research. 2003;583.

52. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning. 2003;52(1):91.

53. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411-423.

54. Boryczka U. Finding groups in data: Cluster analysis with ants. Applied Soft Computing. 2009;9(1):61-70.

55. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.[Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.Available from: http://www.pnas.org/cgi/content/abstract/102/43/15545

56. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. London: Academic Press; 1985.

57. Korn E, Troendle J, Mcshane L, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference. 2004;124(2):379-398.

58. Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment (Wiley Series in Probability and Statistics). New York: Wiley-Interscience; 1993.

59. Cronin M, Pho M, Dutta D, et al. Measurement of Gene Expression in Archival Paraffin-Embedded Tissues: Development and Performance of a 92-Gene Reverse Transcriptase-Polymerase Chain Reaction Assay. Am. J. Pathol. 2004;164(1):35-42.

60. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner R, Walker M, Watson D, Park T BJ. Multi-gene RT-PCR assay for predicting recurrence in node

negative breast cancer patients-NSABP studies B-20 and B-14. In:Antonio Breast Cancer Symposium.2003. p.A16-A16.

61. M. A. Cobleigh, P. Bitterman, J. Baker, M. Cronin, M.-L. Liu, R. Borchik, B. Tabesh, J.-M. Mosquera, M. G. Walker SS. Tumor gene expression predicts distant disease-free survival (DDFS) in breast cancer patients with 10 or more positive nodes: High throughput RT-PCR assay of paraffin-embedded tumor tissues. In:Proc Am Soc Clin Oncol.2003. p.850-850.

62. Esteban J, Baker J, Cronin M, et al. Tumor gene expression and prognosis in breast cancer: Multi-gene RT-PCR assay of paraffin-embedded tissue. In:Prog Proc Am Soc Clin Oncol.2003. p.850-850.

63. Tableman M, Kim JS. Survival Analysis Using S: Analysis of Time-to-Event Data. Chapman and Hall/CRC; 2005.

64. Cronin M, Sangli C, Liu M, et al. Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clinical chemistry. 2007;53(6):1084-91.

65. Theissig F, Kunze KD, Haroske G, Meyer W. Histological grading of breast cancer. Interobserver, reproducibility and prognostic significance. Pathology, research and practice. 1990;186(6):732-6.

66. Robbins P, Pinder S, de Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. Human pathology. 1995;26(8):873-9.

67. Singletary SE. Revision of the American Joint Committee on Cancer Staging System for Breast Cancer. Journal of Clinical Oncology. 2002;20(17):3628-3636.

68. Clark GM, McGuire WL. Steroid receptors and other prognostic factors in primary breast cancer. Seminars in oncology. 1988;15(2 Suppl 1):20-5.

69. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine. 2004;351(27):2817-26.

70. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute. 2006;98(4):262-72.

71. Alexe G, Dalgin GS, Ramaswamy R, Delisi C, Bhanot G. Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns. Cancer informatics. 2007;243-74.

72. Alexe G, Dalgin GS, Scanfeld D, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer research. 2007;67(22):10669-76.

73. Dalgin GS, Alexe G, Scanfeld D, et al. Portraits of breast cancer progression. BMC bioinformatics. 2007;8291.

74. Loi S, Haibe-Kains B, Desmedt C, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25(10):1239-46.

75. Cheang MC, Chia SK, Voduc D, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. Journal of the National Cancer Institute. 2009;101(10):736-50.

76. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(19):10869-74.

77. Arpino G, Green SJ, Allred DC, et al. HER-2 amplification, HER-1 expression, and tamoxifen response in estrogen receptor-positive metastatic breast cancer: a southwest oncology group study. Clinical cancer research : an official journal of the American Association for Cancer Research. 2004;10(17):5670-6.

78. Arpino G, Wiechmann L, Osborne CK, Schiff R. Crosstalk between the estrogen receptor and the HER tyrosine kinase receptor family: molecular mechanism and clinical implications for endocrine therapy resistance. Endocrine reviews. 2008;29(2):217-33.

79. Naylor T, Greshock J, Wang Y, et al. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. Breast cancer research : BCR. 2005;7(6):R1186-98.

80. Loi S, Haibe-Kains B, Desmedt C, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC genomics. 2008;9239.

81. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, N.Y.). 2005;310(5748):644-8.

82. The Gene Ontology project in 2008. Nucleic acids research. 2008;36(Database issue):D440-4.

83. Fisher RA. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. 1922;85(1):87 - 94.

84. Borg A, Baldetorp B, Fernö M, et al. ERBB2 amplification is associated with tamoxifen resistance in steroid-receptor positive breast cancer. Cancer letters. 1994;81(2):137-44.

85. Kallioniemi A, Kallioniemi O, Piper J, et al. Detection and Mapping of Amplified DNA Sequences in Breast Cancer by Comparative Genomic Hybridization. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(6):2156 - 2160.

86. Orsetti B, Courjal F, Cuny M, Rodriguez C, Theillet C. 17q21-q25 aberrations in breast cancer: combined allelotyping and CGH analysis reveals 5 regions of allelic imbalance among which two correspond to DNA amplification. Oncogene. 1999;18(46):6262-70.

87. Parssinen J, Kuukasjarvi T, Karhu R, Kallioniemi A. High-level amplification at 17q23 leads to coordinated overexpression of multiple adjacent genes in breast cancer. British Journal of Cancer. 2007;96(8):1258-1264.

88. Gelsi-Boyer V, Orsetti B, Cervera N, et al. Comprehensive profiling of 8p11-12 amplification in breast cancer. Molecular cancer research : MCR. 2005;3(12):655-67.

89. Cingoz S, Altungoz O, Canda T, Saydam S, Aksakoglu G, Sakizli M. DNA copy number changes detected by comparative genomic hybridization and their association with clinicopathologic parameters in breast tumors. Cancer genetics and cytogenetics. 2003;145(2):108-14.

90. Bernard-Pierrot I, Gruel N, Stransky N, et al. Characterization of the recurrent 8p11-12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer. Cancer research. 2008;68(17):7165-75.

91. Stec I, van Ommen GJ, den Dunnen JT. WHSC1L1, on human chromosome 8p11.2, closely resembles WHSC1 and maps to a duplicated region shared with 4p16.3. Genomics. 2001;76(1-3):5-8.

92. De Paepe P, Baens M, van Krieken H, et al. ALK activation by the CLTC-ALK fusion is a recurrent event in large B-cell lymphoma. Blood. 2003;102(7):2638-41.

93. Argani P, Lui MY, Couturier J, Bouvier R, Fournet J, Ladanyi M. A novel CLTC-TFE3 gene fusion in pediatric renal adenocarcinoma with t(X;17)(p11.2;q23). Oncogene. 2003;22(34):5374-8.

94. Patel AS, Murphy KM, Hawkins AL, et al. RANBP2 and CLTC are involved in ALK rearrangements in inflammatory myofibroblastic tumors. Cancer genetics and cytogenetics. 2007;176(2):107-14.

95. Dai C, Whitesell L, Rogers AB, Lindquist S. Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis. Cell. 2007;130(6):1005-18.

96. Streicher KL, Yang ZQ, Ethier SP. Transforming function of the LSM1 oncogene in human breast cancers with the 8p11–12 amplicon. Oncogene. 2007;2104-2114.

97. Jonsson G, Staaf J, Vallon-Christersson J, et al. Genomic subtypes of breast cancer identified by array comparative genomic hybridization display distinct molecular and clinical characteristics. Breast cancer research : BCR. 2010;12(3):R42.

98. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics (Oxford, England). 2007;23(6):657-63.

99. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(50):20007-12.

100. Turner N, Pearson A, Sharpe R, et al. FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. Cancer research. 2010;70(5):2085-94.

101. Butt AJ, McNeil CM, Musgrove EA, Sutherland RL. Downstream targets of growth factor and oestrogen signalling and endocrine resistance: the potential roles of c-Myc, cyclin D1 and cyclin E. Endocrine-related cancer. 2005;1247-59.

102. Karlsson E, Ahnstrom Waltersson M, Bostner J, et al. Comprehensive Genomic and Transcriptomic Analysis of the 11q13 Amplicon in Breast Cancer. Cancer Res. 2009;695166.

103. Bautista S, Theillet C. CCND1 and FGFR1 coamplification results in the colocalization of 11q13 and 8p12 sequences in breast tumor nuclei. Genes, chromosomes & cancer. 1998;22(4):268-77.

104. Kwek SS, Roy R, Zhou H, et al. Co-amplified genes at 8p12 and 11q13 in breast tumors cooperate with two major pathways in oncogenesis. Oncogene. 2009;28(17):1892-903.

105. Bergh J, Norberg T, Sjogren S, Lindgren A, Holmberg L. Complete sequencing of the p53 gene provides prognostic information in breast cancer patients,

particularly in relation to adjuvant systemic therapy and radiotherapy. Nature Medicine. 1995;1(10):1029-1034.

106. Benito M, Parker J, Du Q, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004;20(1):105-114.

107. Hu Z, Fan C, Oh D, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics. 2006;7(1):96.

108. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, N.Y.). 2005;310(5748):644-8.

109. Cochran WG. The Combination of Estimates from Different Experiments. Biometrics. 1954;10(1):101 - 129.

110. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Academic Press; 1985.

111. Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics (Oxford, England). 2007;23(13):282-8.

112. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ (Clinical research ed.). 2003;327(7414):557-60.

113. Bonacich P. Simultaneous group and individual centralities. Social Networks. 1991;13(2):155-168.

114. Bonacich P. Power and Centrality: A Family of Measures. The American Journal of Sociology. 1987;92(5):1170 - 1182.

115. Seneta E. Non-negative Matrices and Markov Chains. Springer; 2006.

116. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical biochemistry. 1976;7248-54.

117. Moffat J, Grueneberg DA, Yang X, et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell. 2006;124(6):1283-98.

Curriculum Vitae

ERHAN BILAL

EDUCATION

Politehnica University of Bucharest 1998-2004
B.S., M.Sc. in Control Systems and Industrial Informatics

Rutgers University 2004-2010
Ph.D. in Computational Biology and Molecular Biophysics

PROFESSIONAL EXPERIENCE

Intens IT, Bucharest, Romania 2002-2003
Software Developer

CSC Card Systems, Bucharest, Romania 2003-2004
Software Developer

PUBLICATIONS

Bilal E, Vasallo K, Toppmeyer D, Barnard N, Levine AJ, Bhanot G, Ganesan S. Amplified loci on chromosomes 8 and 17 predict hormone resistance in ER-positive breast tumors. *Submitted*

Bilal E, Alexe G, Yao M, Cong L, Kulkarni A, Ginjala V, Ganesan S, Bhanot G. Identification of YES1 as a therapeutic target in basal-like breast cancers. *Submitted*

Bilal E, Rabadan R, Alexe G, Fuku N, Ueno H, Nishigaki Y, Fujita Y, Ito M, Arai Y, Hirose N, Ruckenstein A, Bhanot G, Tanaka M. Mitochondrial DNA haplogroup D4a is a marker for extreme longevity in Japan**.** PLoS One 2008

Alexe G, Fuku N, Bilal E, Ueno H, Nishigaki Y, Fujita Y, Ito M, Arai Y, Hirose N, Bhanot G, Tanaka M. Enrichment of longevity phenotype in mtDNA haplogroups D4b2b, D4a, and D5 in the Japanese population. Human Genetics 2007