Breast Cancer: Basic and Clinical Research

ORIGINAL RESEARCH

Integrative Analysis of Response to Tamoxifen Treatment in ER-Positive Breast Cancer Using GWAS Information and Transcription Profiling

Chindo Hicks, Ranjit Kumar, Antonio Pannuti and Lucio Miele

Cancer Institute, University of Mississippi Medical Center, 2500 N. State Street, Jackson, MS 39216.
Corresponding author email: chicks2@umc.edu

Abstract: Variable response and resistance to tamoxifen treatment in breast cancer patients remains a major clinical problem. To determine whether genes and biological pathways containing SNPs associated with risk for breast cancer are dysregulated in response to tamoxifen treatment, we performed analysis combining information from 43 genome-wide association studies with gene expression data from 298 ER$^+$ breast cancer patients treated with tamoxifen and 125 ER$^+$ controls. We identified 95 genes which distinguished tamoxifen treated patients from controls. Additionally, we identified 54 genes which stratified tamoxifen treated patients into two distinct groups. We identified biological pathways containing SNPs associated with risk for breast cancer, which were dysregulated in response to tamoxifen treatment. Key pathways identified included the apoptosis, P53, NFkB, DNA repair and cell cycle pathways. Combining GWAS with transcription profiling provides a unified approach for associating GWAS findings with response to drug treatment and identification of potential drug targets.

Keywords: tamoxifen genome-wide association studies gene expression

Breast Cancer: Basic and Clinical Research 2012:6 47–66

doi: 10.4137/BCBCR.S8652

This article is available from http://www.la-press.com.

Introduction

Despite advances in diagnosis and treatment, breast cancer remains a major public health problem and a major cause of death for women worldwide.[1] Excluding cancers of the skin, breast cancer is the most common cancer among women, accounting for nearly 1 in 4 cancers diagnosed in US women.[1] In 2009, an estimated 192,370 new cases of invasive breast cancer were diagnosed among women, as well as an estimated 62,280 additional cases of in situ breast cancer.[1] Out of all breast cancer patients diagnosed in 2009, an estimated 40,170 died from the disease, making it the second most deadly cancer among women after lung cancer.[1] While treatment and control remain the top priorities, identification of molecular markers which are dysregulated in response to treatment is an important long-term goal for the development of more effective therapeutic strategies.

Approximately 70% of the known breast tumors express estrogen receptor α (ER).[2] For women whose tumors are endocrine sensitive, as indicated by the expression of the ER, tamoxifen represents the most important therapeutic modality.[3] Large clinical trials, such as the National Surgical Adjuvant Breast and Bowel Project (NSABP) trials B-14 and B-20 have demonstrated the benefit of tamoxifen treatment in women who have node-negative, estrogen-receptor-positive breast cancer.[4,5] However, response to tamoxifen treatment varies considerably among breast cancer patients. In the advanced setting, approximately 50% of the patients with ER positive breast tumors will not respond to endocrine treatment, and among those who respond some tend to relapse.[6]

Evidence from the published literature using transcription profiling has shown that response and resistance to tamoxifen treatment are both under polygenic control.[7–9] Nevertheless, despite this recognition, the molecular mechanisms underlying variable response and resistance to tamoxifen treatment are poorly understood. Recent advances in high-throughput genotyping and a reduction in genotyping costs have made possible identification of genetic variants (SNPs) associated with risk for breast cancer using genome-wide association studies (GWAS).[10,11] However, although these genetic variants are providing variable clues about the genetic susceptibility landscape of breast cancer, there remain many challenges to overcome in order to understand fully the contribution of genetic polymorphisms to response and resistance to drug treatment and to translate this new knowledge into clinical practice.

Here we use a gene-centric approach to demonstrate the power of combining GWAS information with gene expression data to identify potential candidate genes and biological pathways containing SNPs associated with risk for breast cancer, which are dysregulated in response to tamoxifen treatment. The objectives of this study were three-fold: (1) to investigate whether genes containing SNPs associated with risk for breast cancer are dysregulated in response to Tamoxifen treatment, (2) Within patients treated with tamoxifen identify a molecular signature of genes which stratify patients into distinct groups on the basis of response to tamoxifen treatment, and (3) to identify biological pathways containing SNPs associated with risk for breast cancer, that are dysregulated in response to tamoxifen treatment. Our method focuses on the genes and biological pathways rather than individual SNPs. This holistic approach is aimed at providing insights about the broader biological context in which the SNPs associated with risk for breast cancer operate.

Material and Methods

Source of SNP data

Short of being able to get raw GWAS data to perform traditional single-SNP GWAS analysis to identify SNPs associated with risk for breast cancer, we obtained SNP data by mining data from 43 published GWAS reports and supporting websites containing supplementary data reported through June 2011 using PubMed searches. The studies evaluated involved over 250,000 cases and over 250,000 controls. As an inclusion criteria, only studies with samples of >500 samples were considered. The methods of SNP data collection including sources, along with SNP annotation and gene name validation have been fully described in detail and reported in our previous studies.[10,11]

To address publication bias, we catalogued all the SNPs that showed significant ($P < 0.05$) association with risk for breast cancer. This low threshold is based on the rationale that breast cancer is a polygenic disease involving many genes interacting with each other, with each gene having only a small effect on the observed phenotype. Therefore, restricting the analysis to only those genes containing SNPs with the smallest *P*-values ($P < 10^{-5}$) could potentially miss important biological pathways modulating response to drug treatment. Conversely, use of functionally related genes containing SNPs with the smallest *P*-values and those containing SNPs with moderate *P*-values ($P = 10^{-5}$~$P < 0.05$) together could provide new insights about how the two sets of genes work in concert in response to drug treatment or to produce a particular breast cancer phenotype.

The SNP, IDs (rs-ID), locations and gene names were verified using the dbSNP database using chromosome report build 37.1 and the Human Genome Nomenclature (HGNC) database. SNPs were matched with gene names using SNP IDs (rs-IDs) information contained in the dbSNP database. The analysis yielded a total 500 genetic variants (SNPs), of which 113 mapped to intergenic regions and were not included in further analysis. The remainder 387 SNPs mapped to 110 genes. These genetic variants and genes formed the basis of our analysis to identify the molecular signatures and biological pathways dysregulated in response to tamoxifen treatment, and to identify novel genes which act in concert with genes containing SNPs associated with risk for breast cancer. As noted, a complete catalogue of genetic variants and genes along with references has been reported elsewhere in our previous studies.[10,11]

Source of gene expression data

We used publicly available gene expression data. The data consisted of 298 ER^+ breast cancer patients uniformly treated with Tamoxifen and 125 ER^+ breast cancer patients (controls) not treated with Tamoxifen. All the data was derived from the Caucasian population. The 298 samples were derived from fresh frozen tissue obtained from ER^+ invasive breast cancer patients that were profiled at the Institut Jules Bordet in France.[8] The 125 ER^+ were breast cancer patients with primary operable invasive breast cancer, whose frozen tumor specimens were archived at the John Radcliffe Hospital (Oxford, UK) and the Uppsala University Hospital (Uppsala, Sweden).[12] No patient from the 125 ER^+ had received any adjuvant systematic therapy. The methods of sample preparation and data collection have been fully described by the data originators.[8,12] All samples were assessed for global gene expression profiles using the Affymetrix platform on U133A Human Chips. The data from these samples consisted of the raw probe-level hybridization intensities, which were downloaded from the NCBI's Gene Expression Omnibus (GEO) database http://www.ncbi.nlm.nih.gov/geo/ under accession numbers: GSE17705 and GSE2990, respectively.

In each of the data sets described above, entries in the data matrix were expression values generated by Affymetrix's Microarray Analysis Suite 5.0 (MAS5) statistical algorithm.[13] Following normalization and scaling, MAS5 signal values were summarized by Turkey's biweight estimation of the probe level intensities within each probe set. This was followed by a global normalization (linear scaling) to give all chips the same average intensity. These procedures yield robust weighted means called average-scaled differences that are proportional to the amount of a particular RNA transcript present in the sample after background correction, which we used as the input in this analysis. All the data was already log transformed (log2). Spiked control genes were removed during pre-processing of the data.

Data analysis

Our analysis strategy follows a gene-centric approach. Under this approach we assumed that the genes and pathways containing SNPs are the units of association. This holistic approach has several attractive features: (i) by focusing on the genes and biological pathways instead of individual SNPs, it allows us to make inference about the broader biological context in which the genetic variants operate. (ii) It allowed us to consider the joint effects of all the SNPs including those with small effects and potential rare variants as well as cis regulatory elements which may be impacted by SNPs mapped to the genes under study. (iii) Through co-expression analysis and pathway prediction, this approach allows identification of other genes which are correlated or functionally related with those containing SNPs associated with risk for breast cancer. Such genes could not be identified using traditional single-SNP GWAS analysis. Therefore, it is an optimal analysis strategy. For genes containing multiple SNPs and those containing SNPs replicated in multiple independent studies, we computed their overall effect size (*P*-value) using the procedures described in our previous studies.[10,11]

As a first step, we randomized and partitioned gene expression data into two independent data sets of almost equal sizes, the test set and the validation set. The rationale for partitioning the data set was to determine the reliability and reproducibility of the results by repeating the analysis using the validation set. The test set contained 149 ER^+ breast cancer patients uniformly treated with Tamoxifen and 62 ER^+ controls. The validation set included 149 ER^+ breast cancer patients treated with Tamoxifen and 63 ER^+ controls. The data was normalized using a lowess normalization method, a widely used nonlinear correction technique which allowed us to account for any potential extreme outliers.[14]

We performed a combination of analysis strategies on gene expression data. In the first strategy, supervised analysis, we compared gene expression data between tamoxifen treated patients and controls, using a *t*-test, performed on the test and validation data sets. The goal was to identify significantly differentially expressed genes containing SNPs associated with risk for breast cancer, that are dysregulated in response to drug treatment. We used a false discovery rate procedure[15] to correct for multiple testing. Genes were ranked based on estimates of *P*-values and SNP-containing genes which were significantly ($P < 0.05$) differentially expressed between tamoxifen treated and controls were selected.

In the second step, unsupervised analysis, we performed pattern recognition analysis using hierarchical clustering based on complete linkage method and correlation distance. The goal was to identify SNP-containing genes with similar patterns of expression profiles among genes exhibiting significant differences in expression between cases and controls. In addition, this analysis was carried out to determine whether genes containing SNPs with small *P*-values and SNPs replicated in multiple independent studies interact with genes containing SNPs with moderate *P*-values. Generally, SNPs with moderate *P*-values are "often considered not genome-wide significant in traditional GWAS analysis". Prior to performing hierarchical clustering, the data was normalized, standardized and centered using the methods developed by Eisen et al.[16] This analysis was performed on a set of genes which provided good evidence of distinguishing breast cancer patients treated with tamoxifen from controls.

In the third step, we performed unsupervised followed by supervised analysis within the 298 breast cancer patients treated with tamoxifen. The objective was to identify a signature of genes which exhibited significant differences in expression profiles in response to tamoxifen treatment within treated patients. To achieve this objective, we used gene expression data on the set of significantly differentially expressed genes identified using the test and validation sets. Supervised analysis was performed using Pomello II.[17] Unsupervised and correlation analysis were performed using GenePattern software.[18] We used the Pearson correlation coefficient to assess whether genes containing SNPs with small *P*-values and genes containing SNPs replicated in multiple independent studies are co-expressed with genes containing SNPs with moderate *P*-values. The correlation coefficient between pairs of genes was estimated using the SAS System.[19]

In the fourth step, we combined data on genes containing SNPs with the most significant *P*-values and those replicated in multiple independent studies (see Table 1 and Table 2 in this report) with data on 19 genes experimentally confirmed to be predictive of resistance to tamoxifen treatment. The 19 genes experimentally confirmed to be tamoxifen resistant included the genes, *RAD21, BAP1, NAE1, MYC, TNFAIP3, CLLP, CEACAM6, PTEN, RARG, NF1, PAX2, NIPBL, CCND1, UAB3, SMC3, PAK1, ERBB2, NSD1, GPRC5D*. The genes were identified by mining the literature on tamoxifen resistance focusing only on reports which have experimental conformation.[9,20–25] We performed supervised analysis, correlation analysis and pattern recognition analysis on the combined data set. The objectives of these analyses were two-fold: (1) to determine whether genes containing SNPs and genes resistant to tamoxifen are predictive of response to tamoxifen and to identify genes that are resistant to tamoxifen treatment, (2) To determine whether genes containing SNPs are co-regulated and have similar patterns of expression profiles with experimentally confirmed genes predictive of resistance to tamoxifen treatment.

To determine how genes containing SNPs and genes resistant to tamoxifen treatment affect the survival outcome, we compared the expression values of the combined set of genes between 71 breast cancer patients who relapsed and 227 breast cancer patients who did not relapse. Significantly differentially expressed genes were then correlated with the 19 genes that are resistant to tamoxifen treatment. The goal was to determine which of the genes containing SNPs could be potential predictors of outcome, and whether they correlate or are co-expressed with known predictors of tamoxin resistance. It is worth noting that this analysis was carried out in the absence of other variables such as age, tumor grade, as such data was not available. Data analysis in the fourth step was performed using the GenePattern Software Package[18] and the SAS Software Package.[19]

Finally, we performed pathway prediction and network visualization using the Ingenuity System[26] to determine whether genes containing SNPs associated with risk for breast cancer, interact with each other in biological pathways. The goal was to identify biological pathways and gene regulatory networks containing SNPs, which are dysregulated in response to tamoxifen treatment. Therefore, for pathway prediction and network modeling, the inputs were the genes containing SNPs, which were dysregulated in response to tamoxifen treatment. Gene expression values were included in the input to identify up and down regulated genes in the pathways. The same analysis was applied to a set of genes which stratified tamoxifen treated patients. Validation of predicted pathways and identification of other downstream target genes was achieved using the global pathway prediction and network modeling module built in the Ingenuity System. This global approach allowed identification of other functionally related genes and biological pathways, which could not be identified using traditional single-SNP GWAS analysis. Functional relationship among SNP-containing genes and with other genes was assessed using gene ontology (GO) information incorporated in the ingenuity system and the GO database.[27] The overall effect size for the SNPs in the pathways (defined as the average *P*-value of SNPs within a pathway) was computed using the procedure reported in our earlier study.[11]

Results

Identifying genetic variants that increase susceptibility to breast cancer has been the primary aim of genome-wide association studies with application to breast cancer and other common human diseases. However, identification of genetic variants by means of such studies provides limited insights about the biological context in which genetic variants operate. This knowledge gap is hampering translation of genomic discoveries into clinical practice to guide patient treatment. Although in some cases this knowledge immediately illuminates a path towards development of therapeutic strategies,[28,29] to date, there is no information regarding the use of GWAS information to guide drug treatment in breast cancer patients. In this study, we have used the power of integrative genomic and bioinformatics analysis combing GWAS information from 43 genome-wide association studies and gene expression data on 298 ER^+ breast cancer patients uniformly treated with tamoxifen and 125 ER^+ untreated controls to determine whether genes and biological pathways containing SNPs associated with risk for breast cancer are dysregulated in response to tamoxifen treatment. The results of this analysis are presented below.

Response to TAM treatment

One of the primary objectives of this study was to investigate whether genes containing SNPs associated with risk for breast cancer are dysregulated in response to Tamoxifen treatment. We hypothesized that genes containing SNPs associated with risk for breast cancer significantly differ in their expression profiles between ER^+ breast cancer patients treated with tamoxifen and ER^+ control patients not treated with tamoxifen. We tested this hypothesis by comparing the expression of genes containing SNPs associated with risk for breast cancer between tamoxifen treated and untreated patients using the test and validation data sets. The results showing significantly ($P < 0.05$) differentially expressed genes distinguishing tamoxifen treated from untreated ER^+ control breast cancer patients are presented in Figures 1 and 2 for the test and validation sets, respectively. Out of the 110 genes containing SNPs associated with risk for breast cancer tested, we identified 101 significantly ($P < 0.05$) differentially expressed genes, which clearly distinguished tamoxifen treated patients from controls, in the test set (Fig. 1). These genes were also identified and validated in the validation set which produced 97 significantly ($P < 0.05$) differentially expressed genes distinguishing the two groups (Fig. 2). After ranking the genes in the test and validation sets, we identified a signature of 95 highly

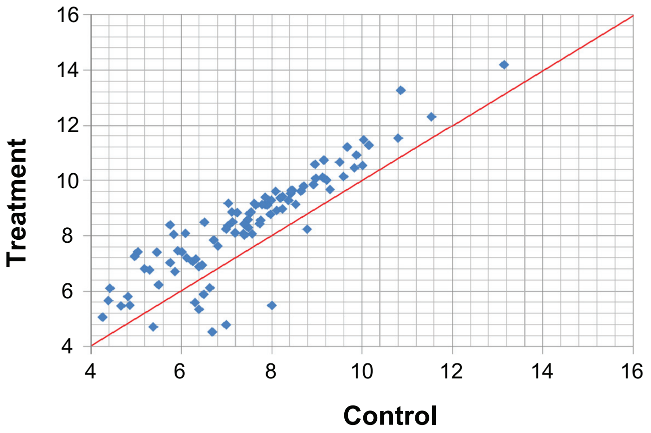

Figure 1. Predictive power of the 92 significantly differentially genes identified by comparing mean gene expression values in the 149 ER+ breast cancer patients uniformly treated with tamoxifen and 62 ER+ control breast cancer patients.
Note: Expression values are based on a log scale (log2).

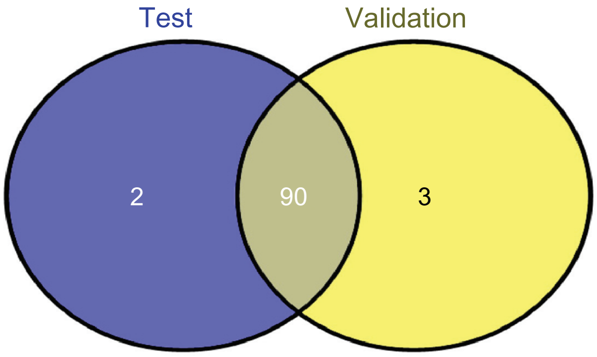

Figure 3. Distribution of genes for the 95 gene signatures identified using the test and validation data sets.
Notes: The number 90 genes in the intersection set indicate the most highly significant genes which overlapped between the test and validation data sets. All genes were selected using the same threshold of $P < 0.000001$ in the test and validation set. The numbers 2 and 3 represent the significantly differentially expressed genes at $P < 0.000001$ only in the test set and only validation set, respectively.

significant ($P < 0.000001$) differentially expressed genes (FDR = 0), which distinguished tamoxifen treated from untreated breast cancer patients (Fig. 3). The results confirmed our hypothesis that genes containing SNPs associated with risk for breast cancer are dyregulated in response to tamoxifen treatment.

As expected from the 95 gene signature identified, 90 genes overlapped between the test and validation sets (Fig. 3) confirming the reproducibility of the results. Out of the 95 gene signature identified, 13 genes contained SNPs with small *P*-values ($P < 10^{-5}$) (Table 1), whereas 19 genes contained SNPs replicated in multiple independent studies (Table 2). The rest of significantly differentially expressed genes contained SNPs with moderate *P*-values ($P = 10^{-4}$~$P < 10^{-2}$). This finding was of particular interest, given that SNPs with moderate *P*-values are often considered to be noise and thus of less value in traditional single-SNP GWAS analysis. Estimates of *P*-values and FDR in the test and validation data sets for all the 110 genes derived from gene expression data comparing breast cancer patients treated with tamoxifen to controls are presented in Table A, in the appendix, provided as supplementary data. The 95 gene signature identified in this study indicates that genes containing SNPs associated with risk for breast cancer could be potential predictors of response to tamoxifen treatment.

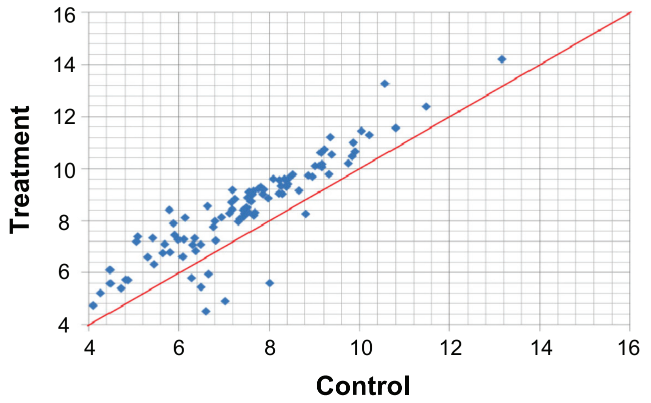

Figure 2. Predictive power of the 93 significantly differentially genes identified by comparing mean gene expression values in the 149 ER+ breast cancer patients uniformly treated with tamoxifen and 63 ER+ control breast cancer patients.
Note: Expression values are based on a log scale (log2).

To determine whether genes containing SNPs associated with risk for breast, which distinguished tamoxifen treated from untreated ER+ breast cancer patients exhibit similar patterns of expression profiles, we performed unsupervised analysis using hierarchical clustering. We hypothesized that genes containing SNPs associated with risk for breast cancer are functionally related and are likely to exhibit similar patterns of expression profiles in response to tamoxifen treatment. Figure 4 shows patterns of gene expression profiles for the top 92 significantly ($P < 0.00001$) differentially expressed up and down regulated genes (FDR = 0), which distinguished tamoxifen treated breast cancer patients from controls for the test set. Patterns of gene expression profiles for the top 93 significantly ($P < 0.00001$) differentially expressed genes (FDR = 0) identified in the validation set are presented in Figure 5. We identified

Table 1. List of genes and SNPs with the smallest *P*-values (large effect size) associated with risk for breast cancer, and estimates of *P*-values for the SNP-containing genes which responded to tamoxifen treatment, in the test and validation data sets.

Gene name	SNP_ID	Estimated GWAS-*P*-value	Estimated *P*-value in test set	Estimated *P*-value in validation set
SLC4A7	rs4973768	4×10^{-23}	0.04	1.00E-05
CASP8	rs1045485	1.1×10^{-5}	5.00E-06	5.00E-06
TGFB1	rs1800470	2.8×10^{-5}	5.00E-06	5.00E-06
ESR1	rs3020314	8.4×10^{-5}	5.00E-06	5.00E-06
LSP1	rs3817198	3×10^{-9}	1.50E-05	5.00E-06
TOX3	rs8051542	1.0×10^{-36}	5.00E-06	5.00E-06
TOX3	rs12443621	2×10^{-19}	5.00E-06	5.00E-06
RNF146	rs2180341	2.9×10^{-8}	5.00E-06	5.00E-06
ECHDC1	rs6569480	6.1×10^{-8}	5.00E-06	5.00E-06
ECHDC1	rs7776136	6.6×10^{-8}	5.00E-06	5.00E-06
ABCC4	rs1926657	1.9×10^{-6}	5.00E-06	5.00E-06
BTNL8	rs7711970	8.4×10^{-5}	5.00E-06	5.00E-06
COLIA1	rs2075555	8.3×10^{-8}	5.00E-06	5.00E-06
GRIK1	rs458685	6.0×10^{-5}	1.00E-05	5.00E-06
RAD51L1	rs999737	1.74×10^{-7}	5.00E-06	5.00E-06

clusters of co-expressed up and down regulated genes with similar patterns of expression profiles (Figs. 4 and 5). This confirms our hypothesis that genes containing SNPs associated with risk for breast cancer are co-regulated and functionally related. As expected, patterns of expression profiles in the test set (Fig. 4) and validation set (Fig. 5) were identical. Interestingly, genes containing SNPs with the smallest *P*-values, and genes containing SNPs replicated in multiple independent studies were co-expressed and exhibited similar patterns of expression with genes containing SNPs with moderate *P*-values. This result

Table 2. List of genes and SNPs associated with risk for breast cancer, replicated in multiple independent studies, and estimates of *P*-values for the SNP-containing genes which responded to tamoxifen treatment, in the test and validation data sets.

Gene name	SNP_ID	Number of replication	Range of estimated GWAS-*P*-values	Estimated *P*-value in test set	Estimated *P*-value in validation set
ADH1B	rs1042026	3	0.02–0.03	5.00E-06	5.00E-06
CASP8	rs1045485	2	0.02–1.1×10^{-7}	5.00E-06	5.00E-06
CDKN1B	rs34330	2	0.01–0.01	5.00E-06	5.00E-06
CDKN2A	rs3731239	2	0.01–0.001	5.00E-06	5.00E-06
COMT	rs4818	2	0.05–0.02	5.00E-06	5.00E-06
EHMT1	rs4634736	2	0.02–0.02	5.00E-06	5.00E-06
ICAM5	rs1056538	2	0.05–0.001	5.00E-06	5.00E-06
IGBP3	rs2854744	5	0.03–0.05	5.00E-06	5.00E-06
LSP1	rs3817198	6	0.05–3×10^{-9}	1.50E-05	5.00E-06
MAP3K1	rs889312	3	0.05–0.05	5.00E-06	5.00E-06
PGR	rs1042838	2	0.02–0.05	0.007	0.0006
RB1	rs198580	2	0.02–0.02	5.00E-06	5.00E-06
RELN	rs17157903	2	0.05–0.0006	0.0006	1.00E-05
SOD2	rs4880	2	0.01–0.05	5.00E-06	5.00E-06
SORBS1	rs10450393	2	0.01–0.03	4.50E-05	1.50E-05
TGFB1	rs1800470	4	0.05–2.8×10^{-5}	5.00E-06	5.00E-06
ESR1	rs3020314	2	8.4×10^{-5}–8×10^{-5}	5.00E-06	5.00E-06
SLC4A7	rs4973768	3	0.05–4×10^{-23}	0.04	1.00E-05
TOX3	rs8051542	2	0.05–1.0×10^{-36}	5.00E-06	5.00E-06

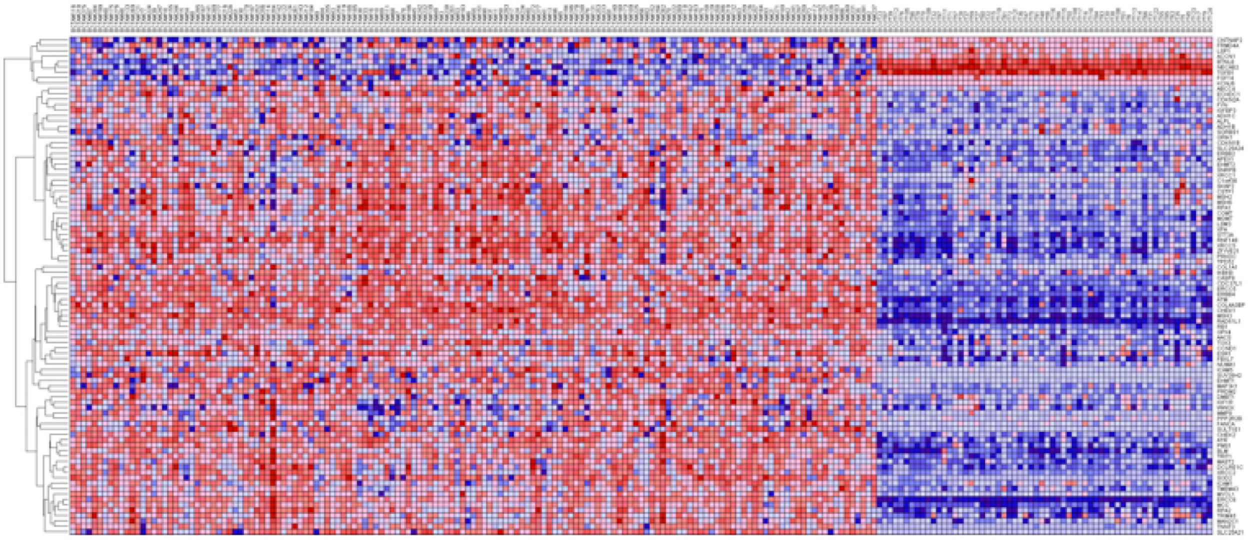

Figure 4. Patterns of gene expression for the 92 significantly ($P \leq 10^{-6}$) differentially expressed genes containing SNPs associated with risk for breasted cancer, distinguishing the 149 tamoxifen treated (right) patients from 62 cancer-free controls (left) in the test set.
Note: Genes are shown in rows and breast cancer patients in columns. Red color indicates upregulation and blue color indicated down regulation.

confirms our hypothesis that genes containing SNPs with small *P*-values and genes containing SNPs replicated in multiple independent studies are likely to act in concert with genes containing SNPs with moderate *P*-values in response to drug treatment.

To determine the functional relationship of identified genes we used the gene ontology (GO) information.[21] GO analysis allows characterization of genes according to the GO nomenclature. The GO consortium has developed three separate ontologies-molecular or physiological function, biological process and cellular component to describe the attributes of gene products. Molecular function defines what a gene product does at the biochemical level without specifying where or when the activity occurs; biological process describes the contribution of a gene product to a biological objective; while cellular component refers to where in the cell a gene product functions. Here we characterized the genes according to all three GO categories in which the genes containing SNPs associated

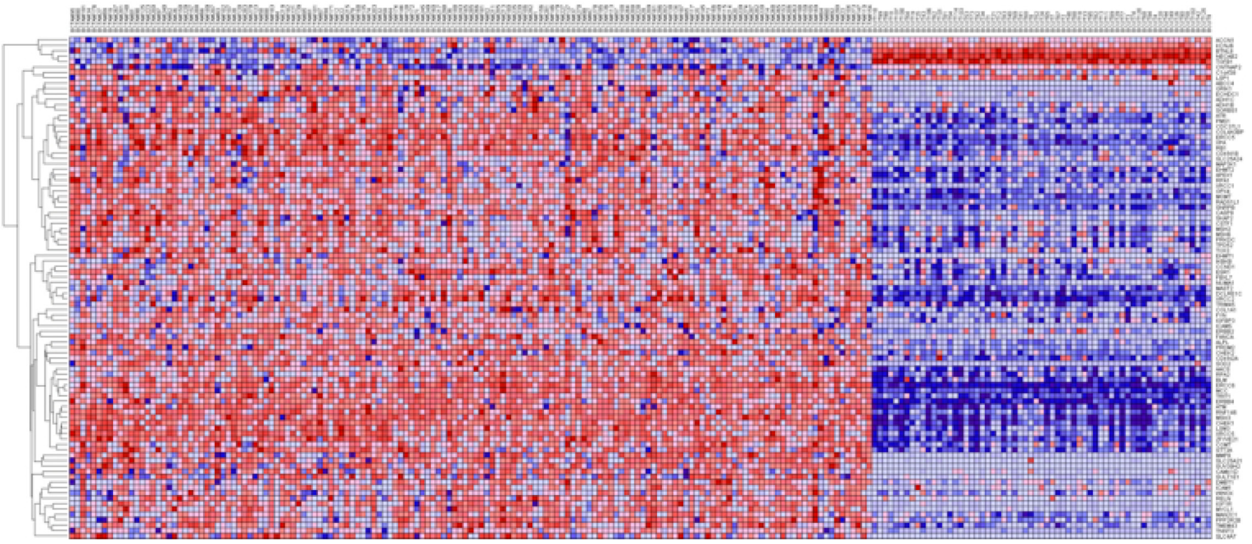

Figure 5. Patterns of gene expression for the 93 significantly ($P \leq 10^{-6}$) differentially expressed genes containing SNPs associated with risk for breasted cancer, distinguishing the 149 tamoxifen treated patients (right) from 63 cancer-free controls (left) in the validation set.
Note: Genes are shown in rows and breast cancer patients in columns. Red color indicates upregulation and blue color indicate down regulation.

with risk for breast cancer are involved. The results of GO classification are presented in Table A for all the 110 genes examined, presented as supplementary data. GO information revealed that genes containing SNPs associated with risk for breast cancer are functionally related, and are involved in the same biological processes. Interestingly, genes containing SNPs with small *P*-values were found to be functionally related with those containing SNPs with moderate *P*-values, suggesting that the two sets of genes could be acting in concert in response to drug treatment. Importantly, genes containing SNPs with large and small *P*-values were found to be interacting with genes containing SNPs replicated in multiple independent studies. This is a significant finding given that replication is often difficult to achieve in traditional single-SNP GWAS analysis. These results show that high-throughput SNP mapping combined with transcription profiling data could lead to identification of potential drug targets.

In both the test and validation sets, we observed variable response to tamoxifen treatment among patients, confirming our original hypothesis that individual patients respond differently to tamoxifen treatment. Part of the observed differences in individual patient's response to tamoxifen treatment can be explained by genetic and phenotypic heterogeneity, although other factors such as age, tumor grade, could not be ruled out. This clinical information was not available in the data sets used and therefore we did not consider them in our analysis. However, the observed variation in response to Tamoxifen treatment is consistent with literature reports.[3] Importantly, the results of this study show that response to tamoxifen treatment is under polygenic control and that genes containing SNPs associated with risk for breast cancer are likely to play an important role in endocrine therapy. This finding provides insights about the functional bridges between GWAS findings and response to drug treatment, and suggests that genes containing SNPs associated with risk for breast cancer could be potential drug targets.

Stratifying patients on the basis of response to tamoxifen treatment

One of the major challenges in endocrine therapy and in particular tamoxifen treatment is variability and resistance in breast cancer patients' response to treatment. Therefore, understanding the molecular mechanisms underlying variable response and resistance to drug treatment is critical. Therefore, our second objective in this study was to investigate whether within the 298 breast cancer patients treated with tamoxifen we could identify a molecular signature of SNP-containing genes, which stratified patients on the basis of variability and differences in response to tamoxifen treatment. The underlying hypothesis supported by evidence from the literature is that in the advanced setting, approximately 50% of the patients with ER^+ breast tumors will not respond to endocrine treatment,[6] and many who do initially respond, subsequently relapse due to the acquisition of endocrine resistance.[30] We reasoned that identifying a signature of genes discriminating patients within tamoxifen treated patients could provide insights about the potential molecular mechanisms underlying variable response and resistance to tamoxifen treatment. Such outcome could have a significant clinical impact by making it possible to stratify breast cancer patients according to various tailored treatments by identifying good responders from poor responders.

To formally test this hypothesis, we performed unsupervised followed by supervised analysis of gene expression data within the 298 breast cancer patients uniformly treated with tamoxifen. This analysis was performed on the 95 genes, which were dysregulated in response to tamoxifen treatment. From the 95 genes tested, we identified 76 genes which exhibited significant ($P < 0.05$) differences in expression profiles. After ranking the genes on the basis of estimated *P*-values, we identified a signature of 54 highly significantly ($P < 0.000001$) differentially expressed genes (FDR = 0), which clearly distinguished patients treated with tamoxifen into two distinct groups (Fig. 6). The 54 gene signature included genes involved in estrogen action, apoptosis, extracellular matrix and immune response. Within the 54 gene signature, the genes *DCLREIC, WWOX, RPA2, BLM, DMBT1, PPP2R2B, TRIM45, IGF1R, MMP8, ICAM1, RELN* were up regulated, whereas the genes *SKAP2, APEX1, CDC37 L1, C1ORF38, MGMT, CASP8, RB1, TGFB1, XRCC5, MSH2, CSFT1, RPA, MSH6, EHMT2, XPA, PRKDC* and *ERCC5* were down regulated (Fig. 6). The 54 gene signature included 10 genes, *CASP8, TGFB1, LSP1, TOX3, ECHDC1, ABCC4, BTNL8, COLIA1, GRIK1* and *RAD51L1* containing SNPs with small *P*-values; and 15 genes, *ADH1B, CASP8,*

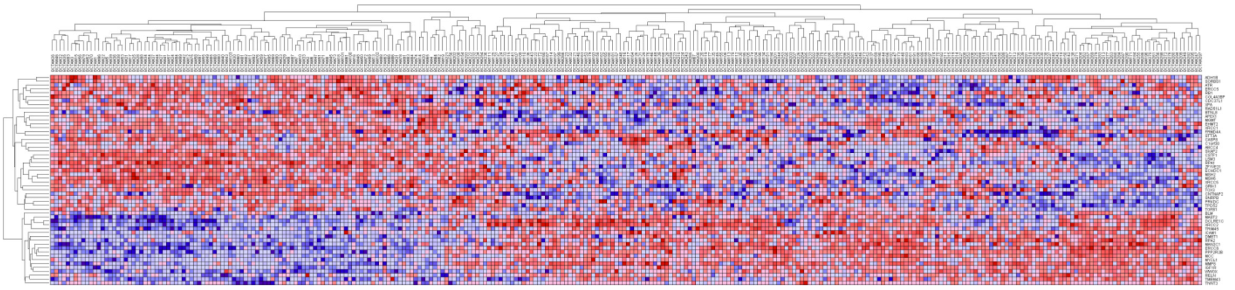

Figure 6. Patterns of gene expression for the top 54 significantly ($P \leq 10^{-6}$) differentially expressed genes containing SNPs associated with risk for breasted cancer, distinguishing poor responders from good reponders within the 298 tamoxifen treated patients.
Note: Genes are shown in rows and breast cancer patients in columns. Red color indicates upregulation and blue color indicate down regulation.

CDKN1B, CDKN2A, LSP1, MAP3K1, PGR, RB1, RELN, SOD2, SORBS1, TGFB1, ESR1, SLC4A7 and *TOX3* containing SNPs replicated in multiple independent studies. The remainder of the genes in the 54 gene signature contained SNPs with moderate *P*-values. A complete list of the 95 genes (including the 54 genes), and their estimates of *P*-values and false discovery rates based on the 298 patients treated with tamoxifen are presented in Table B, provided as supplementary material. Also presented in Table B is information on biological process, molecular function and cellular process in which the 95 genes are involved.

Overall, the results of the 54 gene signature confirm our hypothesis that genes containing SNPs associated with risk for breast cancer could discriminate patients treated with tamoxifen. This is an important finding in that these genes could be used to stratify patients into good and poor responders. Functional information using GO nomenclature revealed that the genes are functionally related (see Table B, supplementary data). Interestingly, co-expression analysis within the 54 gene signature revealed that, genes containing SNPs with small *P*-values and SNPs replicated in multiple independent studies were co-expressed with genes containing SNPs with moderate *P*-values (Fig. 6).

Prediction of clinical outcome and resistance to tamoxifen treatment

One of the major challenges in endocrine therapy is resistance to tamoxifen treatment and how this affects patient outcome. Genetic mechanisms underlying resistance to tamoxifen treatment remain unknown. Therefore, our third objective in this study was to identify genes containing SNPs associated with risk for breast cancer that predictive resistance to tamoxifen treatment and clinical outcome. We reasoned that genes containing SNPs associated with risk for breast cancer either by themselves or acting in concert with other resistant genes confer resistance to tamoxifen treatment and affect clinical outcome. To address this question, we performed supervised and unsupervised analysis combining data on 26 genes containing SNPs with the most significant *P*-values and those replicated in multiple independent studies with data on 19 genes experimentally confirmed to be predictive of resistance to tamoxifen treatment. The genes containing the most highly significant and replicated SNPs included: *SLC4A7, CASP8, TGFB1, ESR1, LSP1, TOX3, RNF146, ECHDC1, ABCC4, BTNL8, COLIA1, GRIK1, RAD51L1, ADH1B, CDKN1B, CDKN2A, COMT, EHMT1, ICAM5, IGBP3, MAP3K1, PGR, RB1, RELN, SOD2,* and *SORBS1*. (see Table 1 and Table 2 in this report). The genes resistant to tamoxifen treatment included (*RAD21, BAP1, NAE1, MYC, TNFAIP3, CLLP, CEACAM6, PTEN, RARG, NF1, PAX2, NIPBL, CCND1, UAB3, SMC3, PAK1, ERBB2, NSD1, GPRC5D*). This approach allowed in silico analysis and validation.

Figure 7 shows patterns of expression for the 45 gene signature in treated and controls. Estimates of *P*-values derived from supervised analysis comparing tamoxifen treated with controls for the combined set of genes are presented in Table C provided as additional supplementary data. All the 45 genes were significantly differentially expressed between tamoxifen treated and controls, suggesting that these genes are predictive of response to tamoxifen. Pattern recognition analysis revealed that genes containing SNPs and genes resistant to tamoxifen were co-expressed and had similar patterns of expression profiles (Fig. 7), confirming our hypothesis that genes containing

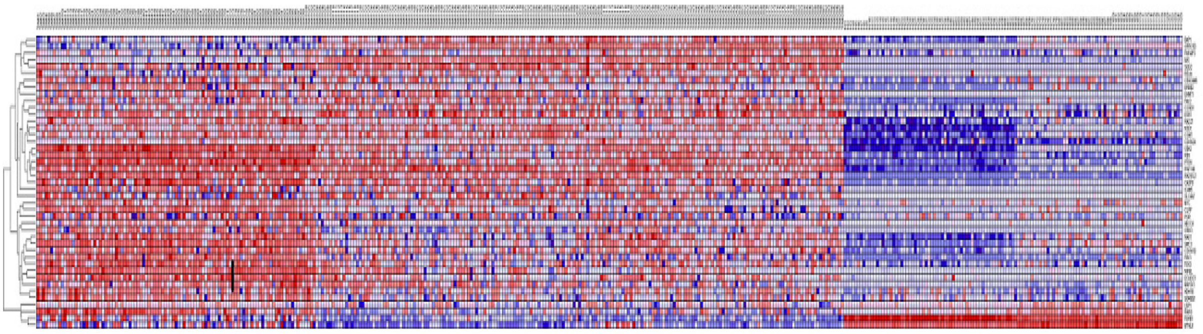

Figure 7. Patterns of gene expression for the genes containing SNPs with smallest *P*-values and replicated in multiple independent studies with 19 genes resistant to tamoxifen treatment.
Note: The left side indicates the 298 patients (in columns) treated with tamoxifen while the right side indicates the 125 patients not treated with tamoxifen herein referred to as controls. Mixed blue and red color in the same gene on the left indicates resistance in some patients. The red color indicates up regulation while the blue indicates down regulation. Analysis based on original raw data.

SNPs either alone or in concert with genes resistant to tamoxifen treatment confer resistance to tamoxifen. This indicates that genes containing SNPs could be potential therapeutic targets, although the specific role of SNPs warrants further investigation. In general, both tamoxifen resistant genes and genes containing SNPs associated with risk for breast cancer exhibited significant variation in expression in both tamoxifen treated patients and controls. The variability in patterns of expression can be explained in part by the introduction of the 19 tamoxifen resistance genes in the analysis. This analysis also demonstrates that use of GWAS information alone may miss other important genes that are predictive of clinical outcomes, notably, genes conferring resistance to tamoxifen treatment.

To further evaluate the association between tamoxifen resistant genes and genes containing SNPs, we performed correlation analysis between the two sets of genes within the 298 patients treated with tamoxifen. This analysis yielded significant correlations between tamoxifen resistant genes and genes containing SNPs. Among the most significantly correlated or co-expressed genes included: *BAP1** vs. *SORBS1*, $r = -0.30$; *UBA3** vs. *SORBS1*, $r = 0.4$; *PTEN** vs. *SORBS1*, $r = 41$; *UBA3** vs. *CASP8*, $r = 0.25$; *PTEN** vs. *ADH1B*, $r = 0.26$; *CASP8* vs. *NF1**, $r = -0.23$; (* *indicates resistant gene*). The correlations between genes containing SNPs and genes resistant to tamoxifen suggest that the two sets of genes likely act in concert to confer resistance or dysregulation to tamoxifen treatment.

To determine whether genes containing SNPs could be predictive of clinical outcome we compared gene expression values between the 71 breast cancer patients who relapsed and 227 breast cancer patients who did not relapse. We identified genes containing SNPs associated with risk for breast cancer which exhibited significant differences in expression between the breast cancer patients who relapsed and breast cancer patients who did not relapse. Among these genes included *CASP8* ($P < 0.0009$), *ADH1B* ($P < 0.0006$), *SORBS1* ($P < 0.0008$), *PGR* ($P < 0.005$), *RNF146* ($P < 0.04$); and genes resistant to tamoxifen *RAD21* ($P < 0.04$) and *MYC* ($P < 0.05$). These results tend to suggest that genes containing SNPs could be potential predictors of clinical outcome. Co-expression analysis of genes containing SNPs and tamoxifen resistant genes revealed significant correlations suggesting that the two sets of genes likely act in concert to affect clinical outcome.

Pathway analysis and network modeling

In a clinical setting, drug treatment may be aimed at targeting specific key biological pathways instead of individual genes, in order to be effective. Identification of candidate biological pathways containing SNPs, which are dysregulated in response to drug treatment is critical. Therefore, our third objective in this study was to identify biological pathways containing SNPs mapped to genes dysregulated in response to tamoxifen treatment. We hypothesized that response to tamoxifen treatment is regulated by many genes interacting within biological pathways and gene regulatory networks containing SNPs associated with risk for breast cancer. The results showing pathways and gene regulatory networks for the SNP-containing genes dysregulated in response to Tamoxifen treatment are presented in Figures 8 and 9 for the

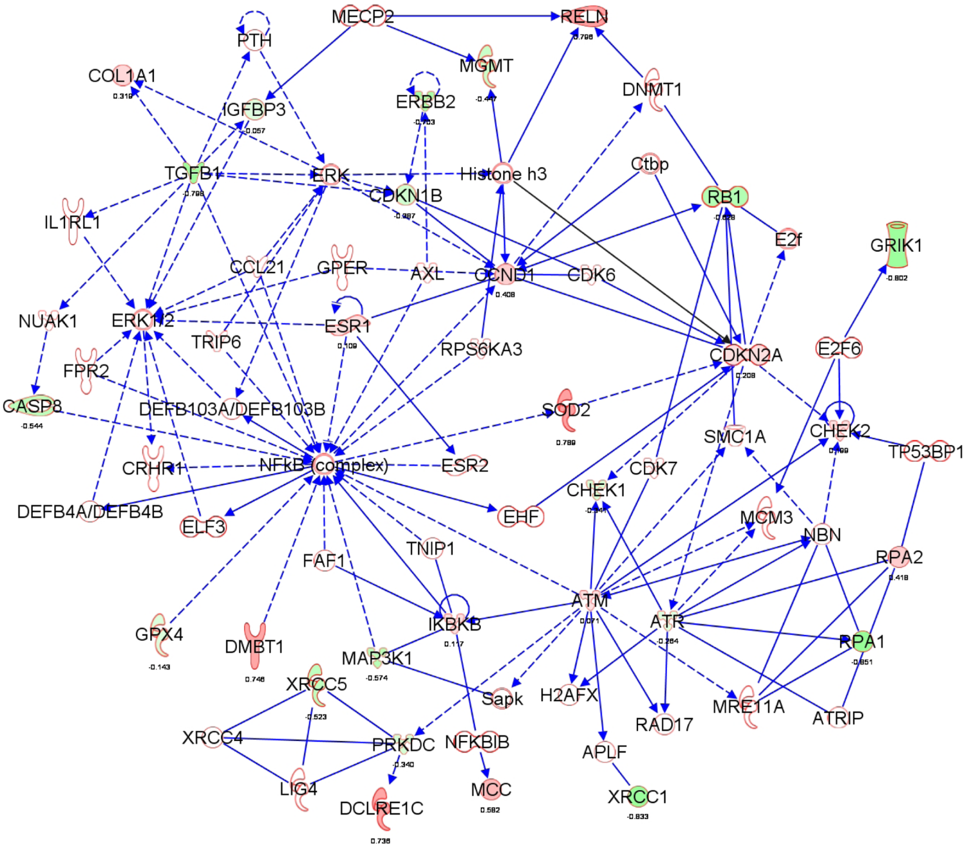

Figure 8. Network1 of the 95 gene signature. Pathways and gene regulatory networks for the 95 genes that exhibited significant differential expression.
Note: Genes mapped to legend symbols in red contain SNPs and are upregulated. Genes mapped to legend symbols in green contain SNPs and are down regulated. All other genes are predicted genes no reported in GWAS studies. Prediction based on genes associated with response to tamoxifen treatment.

95 gene signature. Also presented in the pathways are the color codes red and green indicating the direction of change up and down regulation, respectively, as assessed by gene expression. We identified key biological pathways dysregulated in response to tamoxifen treatment, including the Estrogen receptor, apoptosis, DNA replication, DNA missmatch repair, DNA repair, cell cycle, NFKB, Mapkinase, and P53 pathways. Within the identified pathways, the genes *COL1A1, RELN, SOD2, RPA2, MCC, DCLRE1C, DMBT1, ESR1, WWOX, PPP2R2B, IGF1R, FANCA, MMP8,* and *ICAM1* were upregulated; whereas the genes *ERBB2, RB1, GRIK1, CDKN1B, TGFB1, CASP8, MAP3K1, XRCC5, XRCC1, RPA1, GPX4, CHEK1, XPA, ERCC5, CSTF1, EHMT2, MSH2, MSH3, C4ORF38, APEX1, SKAP2, CDC37L1, MSH6,* and *ALP1* were down regulated (Figs. 8 and 9). The pathways included genes involved in DNA replication, recombination, DNA repair, apoptosis, cell morphology, cell growth and proliferation. The identified pathways and gene regulatory networks included genes that are upregulated *ESR1, COL1A1, IGFBP3, PPP2R2B*; and down regulated *CASP8, GRIK1, MAP3K1, TGFB1* genes containing SNPs with small *P*-values ($P < 10^{-5}$) (Figs. 8 and 9). Also identified in the pathways and gene regulatory networks were the genes *ESR1, TGFB1, CASP8, IGFBP3, CDKN1B, RB1* and *MAP3K1* containing SNPs replicated in multiple independent studies. With the exception of *ESR1* which was up regulated, the rest of the genes containing SNPs associated with risk for breast cancer were down regulated (Fig. 8). Interestingly, genes containing SNPs with the smallest *P*-values and genes containing SNPs replicated in multiple independent studies were found to interact with each other and with genes containing SNPs with moderate *P*-values. This result demonstrates that genes containing SNPs with small *P*-values and SNPs replicated in multiple independent studies are coordinately regulated with those containing SNPs with moderate *P*-values, in response to drug treatment. Importantly, pathway prediction and network modeling also identified

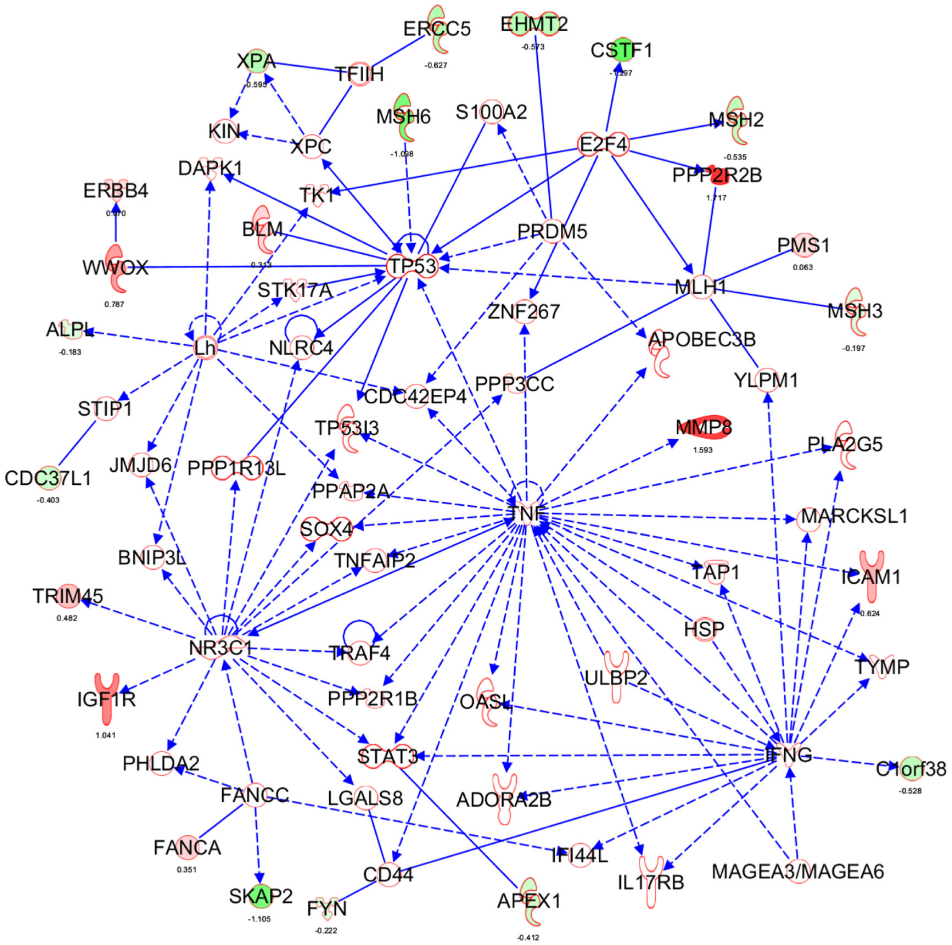

Figure 9. Network 2 of the 95 gene signature.
Note: Pathways and gene regulatory networks for the 95 genes that exhibited significant differential expression. Genes mapped to legend symbols in red contain SNPs and are upregulated. Genes mapped to legend symbols in green contain SNPs and are down regulated. All other genes are predicted genes no reported in GWAS studies. Prediction based on genes associated with response to tamoxifen treatment.

other genes that are functionally related and interact with genes containing SNPs associated with risk for breast cancer. Notably among these genes included genes involved in the NFKB complex, TNF complex, FNG complex, NR3C1 complex, the immune system, and P53 pathway (Figs. 8 and 9). The results confirm our hypothesis that integrative analysis combining GWAS information with gene expression data provides a unified approach to identifying other genes which could not be identified using traditional single-SNP GWAS alone. The identification of multiple biological pathways dysregulated in response to tamoxifen treatment suggests that global pathway crosstalk may be involved in regulating response to tamoxifen treatment.

To determine whether genes in the 54 gene signature interact with each other in pathways and gene regulatory networks, and to identify other genes which may interact with these genes, we performed pathway prediction and network modeling. Figure 10 shows the pathways and gene interaction networks obtained from global pathway prediction and network modeling using the 54 gene signature. We identified multiple biological pathways including the apoptosis, P53, DNA repair, cell cycle and the NFkB pathways. Within the identified pathways and gene regulated networks were seventeen down regulated genes *RPA, CSTF1, MSH2, RB1, XRCC5, CASP8, SKAP2, APEX1, CDC37L1, C1ORF38, TGFB1, MGMT, EHMT2, ERCC5, XPA, MSH6*; and ten up regulated genes *TRIM45, IGF1R, MMP8, ICAM1, RELN, DMBT1, PPP2R2B, DCLRE1C, WWOX, BLM* and *RPA2*. Among the identified genes were genes involved in DNA replication, recombination, DNA repair, cell death, cell morphology, cellular development, and cellular function and maintenance. In addition, we identified other genes

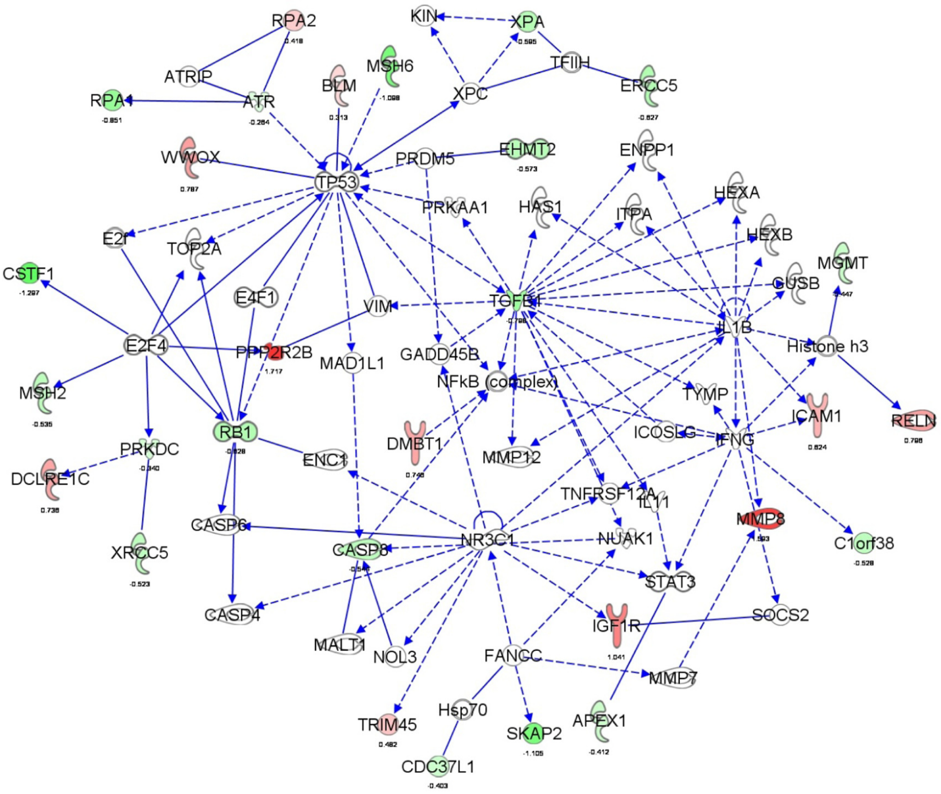

Figure 10. Pathways and gene regulatory networks for the signature of 54 genes, which exhibited significant differential expression.
Note: Genes mapped to legend symbols in red contain SNPs and are upregulated. Genes mapped to legend symbols in green contain SNPs and are down regulated. All other genes are predicted genes not reported in GWAS studies. Prediction based on genes associated with good and poor response to tamoxifen treatment.

which interact with genes containing SNPs associated risk for breast cancer, which have not been reported in GWAS. Considerable variation in response to tamoxifen treatment was observed, which is consistent with previous reports involving transcription profiling.[8]

Consistent with our analysis, a wealth of information about the potential clinical applications of the genes and biological pathway identified in this study exist. Estrogen is a pivotal regulator of cell proliferation in breast cancer. Therefore, endocrine therapies targeting the Estrogen receptor pathways such as tamoxifen could be effective molecular markers. Although the primary action of tamoxifen is believed to be through inhibition of estrogen receptor, our research shows that additional non ER mediated molecular mechanisms exist, including the modulation of the signaling proteins *PKC, TGFB1, CASP8, NFkB* and *Map3K1*, which play a critical role in TAM-induced apoptosis. For example, there are at least eleven isoforms of PKC that either cooperate or exert opposite effects on the process of apoptosis.[31] MAPK are activated by upstream MAP2K, which are, in turn, activated by MAP3K.[31] Caspases are believed to be terminal executors of apoptosis, their activation mediated through cell death receptors. The NFkB is a potential prognostic marker capable of identifying a high-risk subset of ER^+, primary breast cancer destined for early relapse despite adjuvant endocrine therapy with tamoxifen.[32] Additionally, initial studies have suggested that treatment strategies designed to prevent or interrupt activation of NFkB in cell line models of more aggressive ER^+ breast cancers can restore their sensitivity to tamoxifen treatment.[32] The P53 gene is a trascription factor that normally inhibits cell growth and stimulates cell death when induced by cellular stress.[33,34] The most common way to disrupt the P53 pathway is through a point mutation that inactivates its capacity to bind specifically to its cognate recognition sequence.

We identified genes MSH2 and MSH6 involved in DNA repair and mismatch repair (MMR). Nuclear mismatch repair has been initiated by the heterodimeric complexes hMSH2-hMSH6 (hMutSα) and hMSH2-hMSH3 (hMutSβ).[35] The MSH2 gene identified in this study is one of the crucial proteins in MMR. The response of the cell to DNA damage and its ability to maintain genomic stability by DNA repair are crucial in preventing cancer initiation and progression. Hence, genetic variants of DNA repair genes may affect the process of carcinogenesis. Although the role of genetic variants mapped to DNA mismatch repairs genes in breast cancer is unknown, the importance of genetic variability of the components of mismatch repair genes is well documented in colorectal cancer.[35] In addition DNA repair genes are known to be sensitive or responsive to changes in environment. Various DNA alterations can be caused by exposure to environmental and endogeneous carcinogenes. Most of these alterations if not repaired, can result in genetic instability, mutagenesis and cell death. Ensuring fidelity of DNA replication is central to preserving genomic integrity, and DNA mismatch repair genes are critical for maintaining the fidelity of replication. The response of the cell to DNA damage and its ability to maintain genomic stability by DNA repair are crucial in preventing cancer initiation and progression. Therefore, genetic variants in DNA repair genes may affect the process of carcinogenesis.

The results in this study provide proof-of-concept that genes and pathways containing SNPs associated with risk for breast cancer are dysregulated in response to tamoxifen treatment. Short of being able to sequence the entire genome of every tamoxifen treated patient and performing allele-specific profiling, how does one rationally go about identifying genetic polymorphisms that influence drug response?. The results in this study demonstrate that integrating GWAS information with gene expression data provides a holistic approach to identifying candidate genes and candidate pathways to make an informed prediction of the genes in which polymorphisms might affect the predisposition or response to tamoxifen treatment. Additionally, the results show that combining GWAS with gene expression data can assist in the identification of as-yet-unrecognized potential drug targets.

Discussion

We describe an integrative genomics approach that combines GWAS information with gene expression data to identify molecular signatures, biological pathways and gene regulatory networks, which are dysregulated in response to tamoxifen treatment in ER$^+$ breast cancer patients. Our work has the clinical goal of better understanding the molecular mechanisms underlying variable response to tamoxifen treatment in ER$^+$ breast cancer patients. Key findings from this study can be summarized as follows: (a) Genes containing SNPs associated with risk for breast cancer are dysregulated in response to tamoxifen treatment and could distinguish treated from untreated breast cancer patients. (b) Within tamoxifen treated patients, genes containing SNPs associated with risk for breast cancer were able to stratify patients into two groups. This is an important finding in that it could guide stratification of patients into poor and good responders a key step in guiding patient treatment at point of care. This finding could also allow identification of genes which contribute to variable response to tamoxifen treatment a major step in identifying molecular mechanisms which may contribute to patients' resistance to treatment and guide personalized treatment. (c) Pathways and gene regulatory networks containing SNPs associated with risk for breast cancer were found to be dysregulated in response to tamoxifen treatment. This is an important finding in that such candidate pathways if confirmed could be targeted for therapy. To the best of our knowledge such findings have not been previously reported in breast cancer.

Clearly, application of GWAS information in clinical practice has been complicated by the fact that studies of risk alleles for breast cancer have shown that although GWAS can identify novel genes that contribute to risk, the odds ratios and effect sizes as determined by *P*-values are relatively small.[10,11] However, the results in this study demonstrate that genes containing SNPs with small *P*-values and SNPs replicated in multiple independent studies interact with genes containing SNPs with moderate *P*-values in biological pathways which influence response to tamoxifen treatment. This finding has two important implications: first, it allows identification of potential candidate genes and candidate pathways that could serve as drug targets if confirmed. Second, it demonstrates that genes containing SNPs with small *P*-values act in concert with those containing SNPs with moderate *P*-values. This is an important aspect of these results given that most of the loci found to date are small and replication in most GWAS studies tends to be elusive.

Currently, few clinically relevant genome-wide association studies of drug response phenotypes on breast cancer have been reported that it is impossible to effectively compare our results. However, the results reported in this study are consistent with the 12 pharmacogenomics-based GWAS on various diseases reported and summarized by Crowley et al.[36] The main difference between our study and these studies is that rather than identifying polymorphisms altering drug response, our focus was on identifying genes and key biological pathways containing SNPs, which are key drivers of response to tamoxifen treatment. The results demonstrate that integration of GWAS information with gene expression data on breast cancer patients uniformly treated with tamoxifen is an essential tool to identifying genes and key biological pathways in which polymorphisms might affect the disposition or response to tamoxifen treatment. Although this study focuses on tamoxifen, this approach could be applied to any given drug where GWAS information and gene expression are available.

The results in this study show that response to tamoxifen involves many genes and multiple pathways. These results are consistent with earlier reports based on expression profiling,[6,8] and those reported recently by Mendes-Pereira et al.[9] However, one caveat is important in this study. The results in this study do not show how individual SNPs or alleles contribute to variable response to tamoxifen treatment. This is a key limitation of this study and is acknowledge here. However, the results of this study provide a proof-of-concept and information about the biological pathways in which SNPs associated with risk for breast cancer operate. The practical clinical utility of that type of information is that it could guide future experimental designs to identify candidate genes and candidate pathways containing SNPs, which are key drivers of drug response and could be potential targets for therapy. In fact, although we did not investigate allele-specific expression or the effect of genetic variants on gene expression, previous studies have demonstrated that individual alleles could affect gene expression in humans.[37,38] For example, a recent study showed that allele specific up-regulation of the FGFR2 (the most replicated gene in GWAS) increased susceptibility to breast cancer.[39]

The potential clinical relevance of the results reported in this study can be summarized as follows: (1). High-throughput SNP-mapping combined with transcription profiling could potentially allow cancer associated drug targets to be identified thereby reducing attrition in early-phase clinical trials. For example, the cost of additional clinical trials might be reduced if the population of responders and non-responders could be segmented on the basis of their genetic profiles in early phases of clinical trials.[40] SNPs identified from candidate genes in early phases of clinical trials could allow non-responders to be excluded from subsequent clinical trial studies, therefore potentially allowing enriched, smaller, faster, less expensive clinical studies on patients with better chance of responding favorably.[40] (2) Although classic response to tamoxifen treatment has been assessed by polymorphisms in the CYP2D6 gene, the results from this study and others studies,[6,8] show that response to tamoxifen treatments is under polygenic control. The identification of multiple pathways that are dysregulated in response to tamoxifen treatment tends to suggest that global pathway crosstalk may be involved. (3) The results show that integrative analysis combining GWAS information with gene expression data could potentially identify candidate genes and potential drug targets that lie outside of the current range of knowledge. This approach could also potentially provide novel biological insights into the mechanisms of drug action and resistance. However, further studies will be required to determine how individual SNPs influence gene expression and response to tamoxifen treatment before firm conclusions of the practical utility of GWAS information in a clinical setting can be drawn. Such investigation though warranted, was beyond the scope of this report, but is the subject of our future studies.

Several studies from our own group[10,11] and others breast cancer[41,42] have reported pathway-based approaches to dissection of the genetic susceptibility architecture of breast cancer. To our knowledge, this is the first study to associate GWAS information with response to tamoxifen treatment and to identify genes and biological pathways dysregulated in response to tamoxifen treatment. Recently, Genomic Health (Redwood City, CA) developed the Oncotype DX diagnostic assay based on candidate gene selection (not genome wide) approach.[7] The multiplex 21-gene test includes genes associated with proliferation, estrogen, and HER2 action, invasion, and five control genes. This 21-gene recurrence score assay provides a recurrence score for node-negative breast cancer patients with ER$^+$ tumors who have received adjuvant tamoxifen.[7] The association between this 21-gene recurrence assay and risk of locoregional recurrence in node-negative estrogen receptor-positive breast cancer has been established using results from NSABP B-14 and NSABP B-20.[43] The utility of the Oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers has also be been established.[44] Although our study was not designed to develop a therapeutic index or to evaluate the Oncotype DX assay, the findings are consistent with those reported in the Oncotype DX assay in that regulation and response to tamoxifen is under polygenic control. In fact, we indentified two genes ESR1 and ERBB2, which are also found in the Oncotype DX assay.

One of the major challenges in endocrine therapy is resistance to tamoxifen treatment, and many patients who respond tend to relapse. In this study, 71 patients relapsed out of the 298 treated with tamoxifen, while the rest exhibited significant variation in response to tamoxifen. The genetic mechanisms underlying resistance to tamoxifen treatment remain poorly understood. Our analysis revealed that genes containing SNPs are co-regulated with genes that are predictive of tamoxifen. This is a significant finding given the urgent need to identify predictive markers and potential targets for developing novel therapeutic strategies. The association of tamoxifen resistant and SNP-containing genes is consistent with recent studies.[20,21] However, more research is needed to ascertain the role of SNPs in tamoxifen treatment. Although such a study would provide more insights about the genetic mechanisms underlying variability and resistance to tamoxifen treatment, it is beyond the scope of this paper, but is the subject of our ongoing investigation and will be reported elsewhere.

The results reported in this study explain the broader context in which genes containing SNPs associated with risk for breast cancer operate in response to tamoxifen treatment. However, the limitations of the study must be acknowledged. First, our study relies on use of publicly available data, which could have some deficiencies, such as sampling errors, genetic and phenotypic heterogeneity, and environmental factors which were not corrected for. In addition, the data did not include other factors such as age, tumor grade, tumor size. Therefore, these results cannot be generalized and their interpretation should be conservative. Majority of the GWAS studies and gene expression data used in this study are based on Caucasian population. Given the emerging evidence that genetic susceptibility loci may confer population-specific risk,[45] and the plausibility that response to tamoxifen could potentially differ between populations, these results cannot be generalized to all populations. Use of association results diagnostically will require explicit evaluation of how well they can be transferred across different population groups. Additionally, further research is needed to determine the effects of genetic variants on gene expression in different populations. One such approach would involve assessment of allelic variation in gene expression among breast cancer patients.[31,39] Such analysis was beyond the scope of this report.

However, despite these limitations, the results from this study provide insights about the global biological context in which SNPs associated with risk for breast cancer operate in tamoxifen treated patients. This is a major step towards translating GWAS discoveries into clinical practice. The results of this study could guide future experimental designs in breast cancer to identify targets for the development of more effective therapeutic strategies.

In conclusion, our data shows that combining gene expression profiling with GWAS information provides a unified approach to identifying candidate genes and candidate pathways containing SNPs, which are dysregulated in response to tamoxifen treatment in ER$^+$ breast cancer. Furthermore, our analysis demonstrates that genes containing SNPs act in concert with experimentally confirmed tamoxifen resistant genes to confer resistance to tamoxifen. Additional studies are needed to determine how individual or all polymorphisms collectively contribute to variability to endocrine therapy in different ethnic populations, and to determine whether these polymorphisms could serve as potential biomarkers for stratifying breast cancer patients to individualized therapies.

Acknowledgments

This research was supported by funding from UMMC Cancer Institute, to which the authors are gratefully thankful.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. American Cancer Society. Breast cancer facts and figures 2010. *Atlanta American Cancer Society*. 2010.
2. Early breast cancer trials' collaborative group (EBCTCG). Tamoxifen for early breast cancer: An overview of the randomized trials. *Lancet*. 1998;351:1451–67.
3. Ingle JN. Pharmacogenetics and pharmacogenomics of endocrine agents for breast cancer. *Breast Cancer Res*. 2008;10(Suppl 4):S17.
4. Fisher B, Constantino J, Redmond C, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have extrogen-receptor-positive tumors. *N Engl J Med*. 1989;320:479–84.
5. Fisher B, Jeong JH, Bryant J, et al. Treatment of lymph-node-negative, estrogen-receptor-positive breast cancer: long-term findings from the National surgical adjuvant breast and bowel project randomized clinical trials. *Lancet*. 2004;364:858–68.
6. Jansen MPHM, Foekens JA, van Staveren IL, et al. Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. *J Clin Oncol*. 2005;23:732–40.
7. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Eng J Med*. 2004;351:2817–26.
8. Symmans WF, Hatzis C, Sotiriou C, Andre F, et al. Genomic index of sensitivity to endocrine therapy for breast cancer. *J Clin Oncol*. 2010;28:4111–9.
9. Mendes-Pereira AM, Sims D, Dexter T, et al. Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen. *Proc Natl Acad U S A*. 2011.
10. Hicks C, Pannuti A, Miele L. Association of GWAS information with the Notch signaling pathway. *Cancer Informatics*. 2011.
11. Hicks C, Asfour R, Pannuti A, Miele L. An integrative genomics approach to biomarker discovery in breast cancer. *Cancer Informatics*. 2011.
12. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J Nat Cancer Inst*. 2006;98:262–72.
13. Microarray Analysis Suite 5.0, *Affymetrix Inc*. Santa Clara, California.
14. Berger JA, Hautaniemi S, Jarvinen A-K, Edgren H, Mitra SK, Astola J. Optimized lowess normalization parameter selection for DNA microarray data. *BMC Bioinformatics*. 2004;5:194.
15. Benjamini Y, Hochberg Yosef. Controlling the false discovery rat: a practical and powerful approach to multiple testing. *J Royal Stat Society. Series B Methodology*. 1995;57(1):289–300.
16. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Nat Acad Sci U S A*. 1998;95:14863–8.
17. Morrissey ER, Diaz-Uriarte R. Pomello II: Finding differentially expressed genes. *Nucl Acids Res*. 2009;37:W581–6.
18. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P. GenePattern 2.0. *Nature Genetics*. 2006;38(5):500–1.
19. SAS statistical System. *SAS Cary*, NC.
20. Giamas G, Filipova A, Jacob J, et al. Kinome screening for regulators of the estrogen receptor identifies LMTK3 as new therapeutic target in breast cancer. *Nature Med*. 2011;17(6):715–9.
21. Maraqa L, Cummings M, Peter MB, et al. Carcinoembryonic antigen cell adhesion molecule 6 predicts breast cancer recurrence following adjuvant tamoxifen. *Clin Cancer Res*. 2008;14(13):4355–6.
22. Bostner J, Waltersson MA, Fornander T, et al. Amplification of CCND1 and PAK1 as predictors of recurrence and tamoxifen resistance in post menopausal breast cancer. *Oncogene*. 2007;26:6997–7005.
23. Vendrell JA, Ghayad S, Ben-Larbi S, et al. A20/TNFAIP3, a new estrogen-regulated gene that confers tamoxifen resistance in breast cancer cells. *Oncogene*. 2007;26:4656–67.
24. Shoman N, Klassen S, McFadden A, et al. Reduced PTEN expression predicts relapse in patients with breast cancer treated tamoxifen. *Modern Pathology*. 2005;18:250–9.
25. Hurtado A, Holmes KA, Geistlinger TR, et al. Regulation of ERBB2 by oestrogen receptor-PAX2 determine response to tamoxifen. *Nature*. 2008;456:663–67.
26. Ingenuity System. *Ingenuity Inc*. California.
27. Gene Ontology Consortium. Creating the gene ontology resource: Design and implementation. *Genome Res*. 2001;11:1425–33.
28. Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*. 1999;286:487–91.
29. Evans WE, McLeod HL. Pharmacogenomics-drug disposition, drug targets and side effects. *New Eng J Med*. 2003;348:538–49.
30. Musgrove EA, Sergio CM, Loi S, Inman CK, Anderson LR, et al. Identification of functional networks of estrogen-and c-Myc-responsive genes and their relationship to response to tamoxifen therapy in breast cancer. *PLoS one*. 2008;3(8):e2987.
31. Mandlekar S, Kong A-NT. Mechanisms of tamoxifen-induced apoptosis. *Apoptosis*. 2001;6:469–77.
32. Zhou Y, Eppenberger-Castori S, Eppenberger U, Benz CC. The NFkB pathway and endocrine-resistant breast cancer. *Endocrine-Related Cancer*. 2005;12:S37–46.
33. Elledge RM, Green S, Howes L, Clark GM, et al. bcl-2, p53, and response to tamoxifen in estrogen receptor-positive metastatic breast cancer: A southwest oncology group study. *J Clin Oncol*. 1997;15(5):1916–22.
34. Berns EMJJ, Klijn JGM, van Putten WLJ, et al. P53 protein accumulation predicts poor response to tamoxifen therapy of patients with recurrent breast cancer. *J Clin Oncol*. 1998;16(1):121–7.
35. Poplawski T, Zadrozny M, Kolacinska A, et al. Polymorphisms of the DNA mismatch repair gene HMSH2 in breast cancer occurence and progression. *Breast Cancer Res Treatment*. 2005;94:199–204.
36. Crowley JJ, Sullivan PF, McLeod HL. Pharmacogenomic of genome-wide association studies: lessons learned thus far. *Pharmacogenomics*. 2009;10(2):161–3.
37. Yan H, Yuan W, Velculescu VE, et al. Allelic variation in gene expression. *Science*. 2002;297:1143.
38. Buckland PR. Allele-specific gene expression differences in humans. *Human Mol Genet*. 2004;13(2):R255–60.
39. Meyer KB, Maia A, O'Reilly M, et al. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biology*. 2008;6(5):e108.

40. Roses AD. Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nature Reviews Genetics*. 2004;5:645–56.
41. Haiman CA, Hsu C, de Bakker PIW, et al. Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. *Hum Mol Genet*. 2008;17(6):825–34.
42. Menashe I, Maeder D, Garcia-Closas M, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res*. 2010;7(11):4453–9.
43. Manounas EP, Tang G, Fisher B, et al. Association between the 21-gene recurrence score assay and risk of locoregional recurrence in node-negative, estrogen receptor-positive breast cancer: Results from NSABP B-14 and NSABP B-20. *J Clin Oncol*. 2010;28(10):1677–83.
44. Kelly CM, Krishnamurthy S, Bianchini G, et al. Utility of Oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. *Cancer*. 2010; 116(22):5161–7.
45. Helgadottir A, Manolescu A, Helgason A, et al. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet*. 2006;38:68–74.

Supplementary Tables

Table A. List and estimates of *P*-values for the test and validation data sets, and GO information for all the 110 of all the 110 genes containing SNPs associated with risk for breast cancer, provided as supplementary material. Estimates of *P*-values are based on comparing ER$^+$ patients treated with tamoxifen to ER$^+$ patients not treated with tamoxifen.

Table B. List and estimates of *P*-values and GO information for all the 95 genes containing SNPs associated with risk for breast cancer, provided as supplementary material. Estimates of *P*-values are based on comparison within the 298 ER$^+$ patients treated with tamoxifen only. This analysis was carried out on the 95 genes dysregulated in response to tamoxifen to identify a gene signature which stratified tamoxifen treated patients.

Table C. Supplementary data. Estimates of *P*-values and false discovery rate for tamoxifen resistant genes (*) and genes containg highly significant SNPs and replicated in mutiple independent studies. Estimates based on comparing tamoxifen treated to controls.