Cancer Informatics

ORIGINAL RESEARCH

An Integrative Genomics Approach for Associating GWAS Information with Triple-Negative Breast Cancer

Chindo Hicks[1,2], Ranjit Kumar[1], Antonio Pannuti[1], Kandis Backus[1], Alexandra Brown[3], Jesus Monico[3] and Lucio Miele[1]

[1]Cancer Institute, University of Mississippi Medical Center, Jackson, MS. [2]Department of Medicine, University of Mississippi Medical Center, Jackson, MS. [3]Department of Pathology, University of Mississippi Medical Center, Jackson, MS.
Corresponding author email: chicks2@umc.edu

Abstract: Genome-wide association studies (GWAS) have identified genetic variants associated with an increased risk of developing breast cancer. However, the association of genetic variants and their associated genes with the most aggressive subset of breast cancer, the triple-negative breast cancer (TNBC), remains a central puzzle in molecular epidemiology. The objective of this study was to determine whether genes containing single nucleotide polymorphisms (SNPs) associated with an increased risk of developing breast cancer are connected to and could stratify different subtypes of TNBC. Additionally, we sought to identify molecular pathways and networks involved in TNBC. We performed integrative genomics analysis, combining information from GWAS studies involving over 400,000 cases and over 400,000 controls, with gene expression data derived from 124 breast cancer patients classified as TNBC (at the time of diagnosis) and 142 cancer-free controls. Analysis of GWAS reports produced 500 SNPs mapped to 188 genes. We identified a signature of 159 functionally related SNP-containing genes which were significantly ($P < 10^{-5}$) associated with and stratified TNBC. Additionally, we identified 97 genes which were functionally related to, and had similar patterns of expression profiles, SNP-containing genes. Network modeling and pathway prediction revealed multi-gene pathways including p53, NFkB, BRCA, apoptosis, DNA repair, DNA mismatch, and excision repair pathways enriched for SNPs mapped to genes significantly associated with TNBC. The results provide convincing evidence that integrating GWAS information with gene expression data provides a unified and powerful approach for biomarker discovery in TNBC.

Keywords: triple negative breast cancer GWAS gene expression

Cancer Informatics 2013:12 1–20

doi: 10.4137/CIN.S10413

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

Introduction

Recent advances in high-throughput genotyping and reduction in genotyping costs have made it possible to identify genetic variants associated with an increased risk of breast cancer possible by using genome-wide association studies (GWAS).[1–3] These findings are providing valuable clues about the genetic susceptibility landscape of breast cancer. However, the association of genetic variants and their associated genes with the most aggressive subset of breast cancer, the triple negative breast cancer (TNBC) remains a central puzzle in molecular epidemiology. Currently, there is a substantial gap between single nucleotide polymorphism (SNP) associations from GWAS and understanding how susceptibility loci contribute to the TNBC phenotype.

TNBC—tumors that do not express estrogen receptors, progesterone receptors, or HER2—are typically high-grade duct carcinomas, although low-grade tumors do occur.[4,5] They represent an important clinical challenge because these cancers do not respond to endocrine therapy or other available targeted therapies.[6,7] TNBC is significantly more aggressive than other subtypes of breast cancer and disproportionately affects younger premenopausal women, with a higher mortality rate among African-American women.[8] Patients with TNBC have a significantly increased risk of relapse and shorter survival rate than patients affected with tumors of other molecular subtypes.[5] In fact, although TNBC accounts for a relatively small proportion of breast cancer cases—about 15% to 20% of all breast cancers diagnosed in the general US population and about 30% in the African-American population-it is responsible for a disproportionate number of breast cancer deaths.[9,10] Additionally, a lower proportion of TNBCs are discovered by mammographic screening, possibly due partly to the age distribution of patients afflicted by this disease.[11,12]

To date, there have been fewer advances in the treatment of TNBC compared to other subtypes of cancer. Although these tumors respond to conventional chemotherapy, which is toxic and affects a wide range of dividing cells, the approach has met with mixed success.[5] TNBC is often aggressive and highly resistant to chemotherapy.[5] TNBC relapses more frequently than hormone receptor-positive, luminal subtypes, and have a worse prognosis.[5] The five-year survival rate for TNBC is about 77%, compared to 93% for other types of breast cancer.[5] An important goal is therefore the identification of molecular markers to reliably identify high and low risk subsets of patients with TNBC, both for different treatment approaches and for the development of novel, more effective therapeutic strategies.

Over the last decade, transcription profiling using gene expression microarray technology has made possible the systematic molecular stratification of TNBC.[8,13] Microarrays have also been used for the histopathological characterization of TNBC.[14] More recently, gene expression profiling has been used to identify triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.[15] However, while these primary analyses have made it possible to identify molecular signatures of TNBC, they have been unsuccessful in identifying which genes and pathways have causative roles as opposed to being consequences of the disease states.[2] Recently, several genetic variants associated with TNBC were reported.[16,17] However, the reported genetic variants associated with TNBC—which include rs2046210 (ESR1), rs12662670 (ESR1), rs3803662 (TOX3), rs999737 (RAD51L1), rs8170 (19p13.1), and rs8100241 (19p13.1)—explain only a small fraction of genetic variation and have modest effects.[16,17] In addition, results from these studies provide no information about the functional roles and the broader context in which the identified susceptibility loci operate leading to the TNBC phenotype.

Because genetic variants that are associated with an increased risk for breast cancer do not lead directly to the disease but instead act on intermediate molecular phenotypes (gene expression), a more optimal approach is to combine GWAS information with gene expression data. Integration of GWAS information with gene expression data provides a unified and powerful approach for identifying potential biomarkers and biological pathways dysregulated in TNBC. The objective of this study was to determine whether genes containing single nucleotide polymorphisms (SNPs) associated with an increased risk of developing breast cancer are both associated with TNBC and could stratify TNBC. A second but equally important objective was to identify molecular pathways and networks enriched for SNPs, which are dysregulated in TNBC. We hypothesized that genes containing SNPs (herein called genetic variants) associated with an increased risk of developing breast cancer are associated with TNBC. We further hypothesized that genes containing genetic variants associated with TNBC are functionally related and interact with each other in biological pathways and networks associated with TNBC. We tested these hypotheses using an integrative genomics approach which combines GWAS information with publicly available gene expression data derived from both breast cancer patients and cancer-free controls. The analysis strategy assumes a gene-centric approach in which the genes containing SNPs are treated as the units of association, with gene expression data derived from patients classified as TNBC at the time of diagnosis acting as the intermediate phenotype. Throughout this report we have used the terms genetic variants and SNPs interchangeably.

Material and Methods

Source of SNP data

Our methods for data collection were based on guidelines proposed by the Human Genome Epidemiology network for the systematic review of genetic association studies; our methods follow the PRISMA guidelines.[18–22] We have previously reported sources of SNP data, including references.[2] Here we provide a brief but detailed description of the methods used in data collection. We mined SNP data and gene information from published reports of GWAS on breast cancer. GWAS were eligible to be included if they met 4 criteria. First, the study design had to be a case-control, cohort or a cross-sectional association study conducted on unrelated individuals in human populations. Second, the study examined the association between breast cancer and the polymorphic phenotype, had a sample case size of greater than 500 subjects and greater than 500 controls, and provided sufficient information such that genotype frequencies for both breast cancer cases and controls could be determined without ambiguity. Third, breast cancer must have been diagnosed by pathological or histological examination. Fourth, publications must have been in peer-review journals or online and published in English on or before June 2012. Only studies published as full-length articles or letters in English language peer-reviewed journals were included in the analysis.

To identify all relevant publications, we used two strategies. First, we queried PubMed with terms (breast cancer, GWAS, GWA, WGAS, WGA, genome-wide, genomewide, whole genome, and all terms + association or + scan) in combination with breast cancer to find all the GWAS published before June 2012. This study yielded 100 publications, which were screened by title, abstract, and full text review in order to identify studies that met our eligibility criteria. The data was manually extracted from both the reported GWAS and the supplementary data in websites accompanying those studies. It was then summarized in a consistent manner.[2] When a study included multiple ethnic populations, we picked the results of the model which adjusted for ethnicity. When possible we considered each subpopulation as an independent study. The search yielded 500 SNPs mapped to 188 genes from a population of over 400,000 cases and over 400,000 controls. The 100 genetic variants mapped to intergenic regions were not included in this analysis. Table D provides supplementary data on the 500 SNPs, including SNP ID, the genes they map to, assigned *P*-values indicating effect size, and the reference sources from which SNP information was extracted.

To address publication bias, we catalogued all available SNPs that showed significant ($P < 0.05$) association with an increased risk of developing breast cancer. The rationale for including all available significant SNPs is that relatively few SNPs have *P*-values sufficiently small enough to give conclusive evidence of association. Conversely, there are usually several hundred SNPs with moderately significant *P*-values ($P \sim 10^{-3}$ to 10^{-4}). While these would likely contain several false-positives, they may also contain genuine effects of small magnitude. We reasoned that the presence of a greater than expected number of associated SNPs mapped to genes of similar biological functions and similar patterns of expression profiles gives a degree of confidence that the associations may be genuine even if none of the SNPs individually is highly significant. The premise is that such SNPs could give insights about the biological process and the broader context in which they operate.[2,23] The SNP, IDs (rs-ID), locations, and gene names were verified using the dbSNP database employing chromosome report build 37.7 and the Human Genome Nomenclature (HGNC) database. SNPs were matched with gene names using SNP ID (rs-IDs) information in the database (dbSNP). For SNPs replicated in multiple independent studies, we combined the *P*-values to estimate the overall effect size by using Fisher's methods, as described in our previous study.[2]

Gene expression data

We used publicly available gene expression data in our analysis. Data selection were based on the TNBC classification guidelines reported by Perou, which define TNBC as tumors which lack expression of estrogen receptor (ER), progesterone receptor (PR), and HER2.[8] A significant proportion of TNBC—50% to 75%—matches the basal-like molecular subtype.[8] However, TNBC is a heterogeneous disease entity encompassing other subtypes of cancer. Synonymous terms of this subtype include basal-type, basal-epithelial phenotype, basal breast cancer, and basaloid breast cancer.[24] Although most basal-like breast cancer is TNBC, not all cases are. In fact, even within basal-like/basal, molecular subtyping has revealed two subtypes,[15,25] underscoring the complexity of the TNBC phenotype. Therefore, focusing on the basal-like subtype of TNBC alone might not be specific enough for biomarker discovery and identification of potential therapeutic targets. Because of inherent heterogeneity in the TNBC phenotype, we decided to include the other two subtypes, normal-like and nonluminal basal, both of which are classified as TNBC.[8] The gene expression data set used in this study comprised of 266 subjects, of which 124 were breast cancer patients. Of the 124 breast cancer patients, 29 were classified as normal-like, 20 were classified as basal-like, and 75 were classified as nonluminal basal as documented by data originators.[25] The remaining 142 subjects were cancer-free controls. These sample sizes were large enough to identify genes associated with TNBC with a statistical of power of 99%. We did not include the Claudin-low subtype[8] because we did not find a suitable data set that matched the other data sets used in this study; therefore, we acknowledge this weakness of the investigation.

All gene expression data was generated using the Affymetrix platform and the U133PLUS 2.0 Human Chip. The microarray data from these samples, including the raw probe-level hybridization intensities, were downloaded from the NCBI's Gene Expression Omnibus (GEO) database[26] under accession numbers, GSE21653,[25] and GSE7904,[27] respectively. Methods of sample collection, preparation, and processing have been fully described by the data originators.[25,27] Data on the cancer-free controls GEO accession number GSE10780 has been fully described by the originators.[28] For each data set, the entries in the data matrix were average scaled difference expression values normalized using the RMA suite on a log scale (log2). Spiked control genes were removed from the data during preprocessing.

Data Analysis

We performed both unsupervised and supervised analyses followed by network modeling, visualization, and pathway prediction. The goal of this study was to determine whether genes containing SNPs associated with an increased risk of developing breast cancer are both associated with TNBC and could stratify TNBC. Additional, we aimed to identify gene regulatory networks and biological pathways enriched for SNPs which are dysregulated in TNBC. Therefore, as a first step in the analysis, we partitioned gene expression data into two subsets, a prioritized subset (ie, a data set of 188 genes containing SNPs associated with an increased risk of developing breast cancer) and a nonprioritized set (ie, a data set containing the remainder of the genes not identified by GWAS). Prioritization of SNP-containing genes was aimed at identifying the genes providing good evidence of association with TNBC, amongst a large pool of 188 SNP-containing genes identified by GWAS. The overarching goal was to maximize the yield and the biological relevance of further downstream screens, analysis, refinements, validation, functional analysis, pathway prediction, and network modeling focusing on the most promising candidates associated with TNBC.

We performed unsupervised analysis using hierarchical clustering on the prioritized data set to discern the patterns of gene expression profiles in TNBC for all the 188 genes containing SNPs associated with an increased risk of developing breast cancer. This unbiased class discovery approach included gene expression data on all 188 genes. Prior to clustering, the data was normalized using median normalization, standardized and centered.[29] Pairwise similarity of all the 188 genes was calculated as the Pearson correlation coefficient of the expression levels. The genes were then grouped by hierarchical clustering using the complete linkage method as implemented in GenePattern.[30]

To obtain a more robust analysis on gene expression data and to identify significantly differentially-expressed SNP-containing genes that are both associated with TNBC and could stratify TNBC, we performed supervised analysis. The significant differences in gene expression profiles between cases and controls were tested using a *t*-test. The sample sizes were sufficiently large to identify significantly differentially-expressed genes with a statistical power of 99% at $P < 0.05$. This approach eliminated SNP-containing genes which were not associated with TNBC and narrowed the focus, highlighting the set of genes which were highly significantly associated with TNBC. To investigate gene expression variation among the subtypes of TNBC under study, we performed analysis of variance (ANOVA). We conducted additional analysis comparing gene expression profiles between the three subtypes of TNBC under study. We used a permutation test to calculate empirical *P*-values. The empirical *P*-values and those from the *t*-test did not differ appreciably. We used a false discovery rate (FDR) to correct for multiple hypothesis testing.[31] Because of the small sample sizes for each type of data, we did not divide the data into test and validation sets; instead we used an out of sample validation approach[32] to identify genes with predictive power. Genes were ranked based on estimated *P*-values. Genes that were highly significantly ($P < 10^{-5}$) associated with TNBC were selected. All supervised analysis was performed using Pomello II software package.[33]

We then performed unsupervised analysis based on hierarchical clustering, using gene expression data on genes highly significantly ($P < 10^{-5}$) associated with TNBC. Pairwise similarity of all genes significantly associated with TNBC was calculated as the Pearson correlation coefficient. The data was median normalized, standardized, and centered.[29] The genes and individuals were then grouped by hierarchical clustering using the complete linkage method, as implemented in GenePattern.[30] The goal was to identify functionally related genes consistently showing similar patterns of expression profiles, within and across the TNBC subtypes under study.

To assess biological functional relationships, we performed additional analysis using the gene ontology (GO) information.[34] The GO Consortium has developed three separate categories (molecular 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 event actually occurs or its broader context. Biological process describes the contribution of the gene product to the biological objective. Cellular component refers to where in the cell a gene product functions. Because our goal in this study was to gain biological insights about the broader context in which genetic variants associated with an increased risk of developing TNBC operate, we considered all three GO categories.

One of the limitations of GWAS as noted in this study is that the results of single-SNP GWAS analysis explain only a small fraction of variation. For example, in TNBC only a very small number of risk loci have been reported.[16,17] This begs the question of where the missing variation is located. Realizing that there may be other key driver genes that act in concert with SNP-containing genes to produce the TNBC phenotypes, we performed additional analysis on the remainder of the data set (unprioritized data set) using supervised analysis. We then proceeded with an unsupervised analysis, as described earlier in this report. The data containing genes identified from this analysis was merged with the data set of SNP-containing genes significantly associated with TNBC. The combined data set was then subjected to unsupervised analysis using GenePattern[30] to identify co-expressed genes with similar expression profiles.

Finally, we performed pathway prediction, network modeling, and visualization using the Ingenuity System (IPA) program (http://www.ingenuity.com).[35] The goal was to identify biological pathways that are enriched by the genetic variants associated with breast cancer and which are also involved in TNBC. We hypothesized that genes containing SNPs associated with an increased risk for TNBC interact with each other and other genes within biological pathways. HUGO gene identifiers were mapped to networks available in the Ingenuity database and ranked by score. The score indicates the likelihood of the genes in a network being found together by random chance. Using a 99% confidence interval, scores of ≥ 3 are considered significant. Additional information, validation of predicted pathways, and identification of other downstream target genes was achieved through the literature and database mining module built in the

Ingenuity System. This allowed identification of other functionally related genes not identified by GWAS. The distribution of the overall effect of SNPs in the pathway and replicated SNPs were calculated using the procedures developed by developed by Hicks et al.[2]

Results

Patterns of gene expression profiles for all genes containing SNPs

As a first step, we performed exploratory analysis using unsupervised analysis to assess the patterns of expression profiles for all 188 genes containing 500 SNPs associated with an increased risk of developing breast cancer. Figure 1 shows the patterns of gene expression profiles for all 188 genes containing SNPs associated with an increased risk of developing breast cancer. SNP-containing genes exhibited patterns of expression profiles that were different from the controls (Fig. 1). Overall, patterns of gene expression profiles varied markedly, with basal-like subtypes showing more consistent patterns while basal and normal-like subtypes exhibited more variability and spurious patterns (Fig. 1). Patterns of expression profiles in the normal-like and non-luminal basal subtypes tended to be similar (Fig. 1). Patterns of expression profiles for some genes in the normal-like and basal subtypes were similar to cancer-free controls. The variability in patterns of gene expression was expected given the genetic and phenotypic heterogeneity inherent in TNBC.[5,8,15,36] These results suggest that genetic susceptibility to TNBC could vary, posing challenges in identifying and stratifying breast cancer patients at risk. The spurious patterns in gene expression profiles suggest that some of the genes may not be associated with the TNBC subtypes under study (Fig. 1). Therefore, after this exploratory analysis, we performed more rigorous analyses and refinements as described in the methods section to identify genes that are both significantly associated with TNBC and could stratify TNBC.

Association between GWAS information and TNBC

The primary objective of this investigation was to determine whether genes containing SNPs associated with an increased risk of developing breast cancer are associated with the most aggressive subset of breast cancer, TNBC. We hypothesized that genes containing SNPs associated with an increased risk of developing breast cancer significantly differ in their expression profiles between patients classified as TNBC and cancer-free controls. To test this hypothesis, we performed supervised analysis as described in the methods section of this report. The results

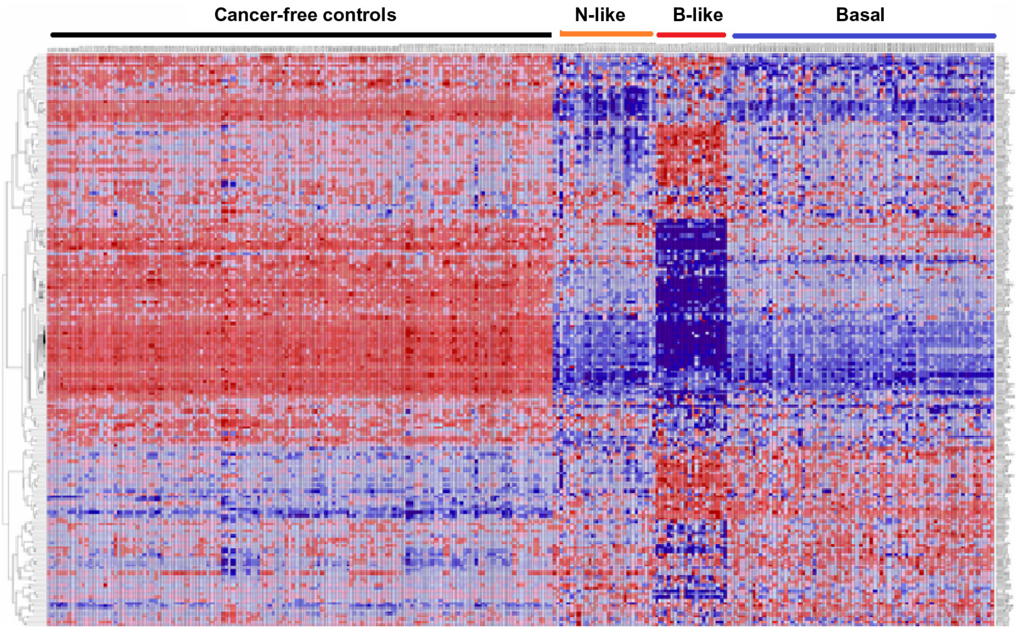

Figure 1. Patterns of gene expression profiles for all the 188 genes containing SNPs associated with an increased risk of developing breast cancer, assessed in 142 controls and subtypes of TNBC (normal-like, N = 29; basal-like, N = 20; and non-luminal basal, N = 75) Tumors.
Notes: Genes are represented in the roles and samples in the columns. Red and blue colors indicate up and down regulation, respectively.

showing estimates of *P*-values, along with the FDR for all 188 genes containing SNPs associated with an increased risk of developing breast cancer, are presented in Table A (provided as supplementary data). Here we present the results of genes which were found to be highly significantly ($P < 10^{-5}$) associated with TNBC.

A comparison between normal-like, basal-like and basal subtypes with cancer-free controls produced 98 genes ($P < 5.00 \times 10^{-5}$), 101 genes ($P < 2.50 \times 10^{-5}$), and 142 genes ($P < 7.50 \times 10^{-5}$); respectively (Table A). Among the identified genes that were highly significantly associated with TNBC included 34 genes containing SNPs with large effects ($P < 10^{-5}$) (Table 1). Thirty-one of the identified genes, including *ADH1B, CASP8, CDKN1B, CDKN2A, COMT, EHMT1, ICAM5, IGBP3, LSP1, MAP3 K1, PGR, RB1, RELN, SOD2, SORBS1, TGFB1, ESR1, SLC4A7, TOX3, FGFR2, STXBP4, CDKN1A, LOC643714, TOX3, CCND1, HCN1, TP53, PTEN, CCNE1, RAD51 L1,* and *CHEK1*, contain SNPs replicated in multiple independent GWAS.[2] A subset of these genes (*TOX3, rs3803662; RAD51L1, rs999737; ESR1, rs2046210, rs12662670; CASP8, rs17468277; ANKLE1, rs8170, rs8100241*) contain SNPs which have been recently associated with TNBC.[16,17] Another

Table 1. Estimates of *P*-values for SNP-containing genes highly significantly associated with TNBC subtypes.

Gene symbol	GWAS		Significant expression in TNBC subtypes		
	SNP(rs_ID)	*P*-value	Normal-like *P*-value	Basal-like *P*-value	Basal *P*-value
FGF4	rs1924587	3×10^{-15}	5.00E-06	5.00E-06	5.00E-06
STXBP4	rs6504950	1.4×10^{-8}	0.08	0.3	4.50E-05
RAD51L1	rs999737	1.74×10^{-7}	5.00E-06	5.00E-06	5.00E-06
ABCC4	rs1926657	1.9×10^{-6}	5.00E-06	5.00E-06	5.00E-06
HCN1	rs981782	1.0×10^{-6}	5.00E-06	5.00E-06	5.00E-06
ZNF365	rs10995190	5×10^{-15}	5.00E-06	5.00E-06	5.00E-06
NOTCH2	rs1124933	3.40×10^{-5}	5.00E-06	0.004	5.00E-06
LSP1	rs3817198	3.0×10^{-9}	5.00E-06	0.003	5.00E-06
FGFR2	rs2981582	2×10^{-76}	0.8	0.2	0.001
PPP2R2B	rs9325024	1.7×10^{-5}	1.00E-05	5.00E-06	5.00E-06
GRIK1	rs458685	6.0×10^{-6}	5.00E-06	5.00E-06	5.00E-06
FGF3	rs614367	3×10^{-15}	0.09	5.00E-06	0.1
COL1A1	rs2075555	8.3×10^{-8}	5.00E-06	5.00E-06	5.00E-06
MRPS30	rs7716600	7×10^{-7}	0.2	5.00E-06	5.00E-06
FHOD3	rs9956546	2.9×10^{-6}	0.8	5.00E-06	5.00E-06
BTNL8	rs7711990	8.4×10^{-5}	5.00E-06	5.00E-06	5.00E-06
RNF146	rs2180341	2.9×10^{-8}	5.00E-06	5.00E-06	5.00E-06
FGF19	rs614367	3×10^{-15}	5.00E-06	5.00E-06	5.00E-06
MYEOV	rs614367	3×10^{-15}	0.04	0.01	0.1
ORAOV1	rs614367	3×10^{-15}	5.00E-06	5.00E-06	5.00E-06
K1AA1804	rs1294255	1.9×10^{-5}	0.2	0.0005	5.00E-06
ANKRD16	rs2380205	5×10^{-7}	3.00E-05	0.0006	5.00E-06
ZMIZ1	rs704010	4×10^{-9}	5.00E-06	5.00E-06	5.00E-06
ECHDC1	rs6569480	6.1×10^{-8}	5.00E-06	0.1	0.1
NEK10	rs4973768	4×10^{-23}	5.00E-06	5.00E-06	5.00E-06
LOC643714	rs3803662	1×10^{-36}	5.00E-06	5.00E-06	5.00E-06
CASP8	rs1045485	1.1×10^{-7}	5.00E-06	0.6	5.00E-06
ESR1	rs3020314	8×10^{-5}	0.2	5.00E-06	6.00E-05
FBN1	rs1876206	6.0×10^{-6}	0.005	0.1	0.1
SLC4A7	rs4973768	4.10^{-23}	5.00E-06	5.00E-06	5.00E-06
TGFB1	rs1800470	2.8×10^{-5}	0.72	0.008	0.4
TOX3	rs8051542	1.0×10^{-36}	5.00E-06	5.00E-06	5.00E-06
H19	rs2107425	2.0×10^{-5}	1.00E-05	0.0001	5.00E-06
MAP3K1	rs889312	4.6×10^{-20}	0.003	2.50E-05	5.00E-06

subset of genes found to be significantly associated with TNBC in this study included the genes *P53, PTEN, RB1, BRCA1, BRCA2, ATR, ATM, MAP3K1; CDKN2A, ATR, CHEK1, CCND1, NOTCH2*. These genes are highly mutated and are known to influence gene expression.[37] The genes *ERBB, CCNE1, PTEN, CCND1, CDKN2A, CDKN2B, CHEK1* found in this study were recently associated with different tumor groups.[38] These analyses confirm our hypothesis that genes containing SNPs associated with an increased risk of developing breast cancer are both associated with TNBC and could stratify TNBC.

Interestingly, genes containing SNPs with small to moderate effects were found to be highly significantly associated with TNBC. This is a significant finding given that only a small number of statistically unimpeachable, common low-penetrance breast cancer susceptibility loci have been reported and confirmed in different populations and in TNBC.[16,17] Of particular interest was the weak link or lack of association of both the FGFR2 gene with TNBC and the association of the *ESR1* gene with TNBC (Table 1). The *FGFR2* and *ESR1* genes are the most replicated genes in GWAS and contain SNPs with the largest effect sizes (Table 1). In the published literature, FGFR2 has been associated with the risk of developing ER-positive tumors,[39] while SNPs in ESR1 have been associated with both ER-positive and ER-negative tumors.[40,41]

Overall, there was incomplete overlap in association between TNBC and cancer-free controls. This is attributable to heterogeneity in patterns of gene expression among the three types of TNBC studied. To assess overlap, we used a Venn diagram to delineate the overlapping genes. We sought to group the genes into those exhibiting significant association with all the three subtypes of TNBC, those associated with two subtypes, and those exhibiting subtype-specific association. Figure 2 shows the intersections of normal-like, basal-like, and basal subtypes. Overall, 119 genes were significantly associated with all three TNBC subtypes. Out of the remainder, 22 genes were significantly associated with the basal-like and basal subtypes only, 18 genes with basal and normal-like only, and 12 genes were significantly associated with basal-like and normal-like only (Fig. 2). Very few genes exhibited subtype-specific association (Fig. 2). Only three genes (not included in the Venn diagram) were not associated with any of the TNBC subtypes under study. These results are consistent with previous findings indicating that TNBC are heterogeneous and overlap is incomplete.[8] The heterogeneity and variability in the results suggests that genetic susceptibility in TNBC may not reflect a single disease, but rather a heterogeneous entity with some loci conferring risks to all subtypes, while others confer subtype-specific risks.

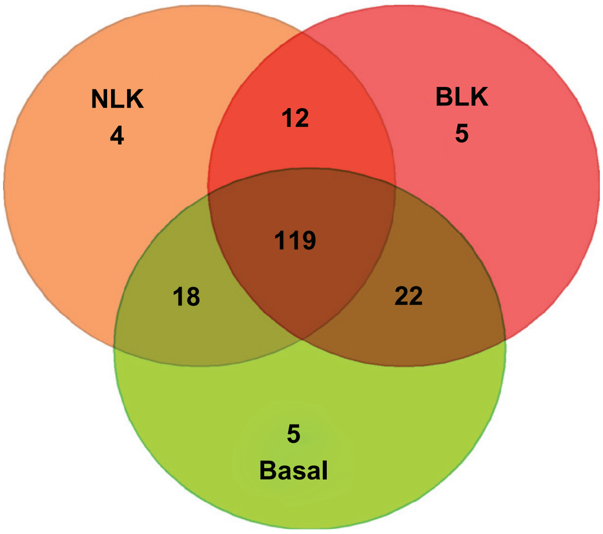

Figure 2. Venn diagram showing the numbers and distribution of genes significantly associated with TNBC subtypes under study.
Notes: The number of overlapping genes are shown in the intersections. NLK indicates normal-like and BLK indicates basal-like.

Having observed heterogeneity and incomplete overlap in association between TNBC and cancer-free controls, we performed ANOVA among the three TNBC subtypes under study in order to quantify the extent of variation and to identify a set of genes which show significant group differences in expression profiles. We hypothesized that gene expression profiles significantly vary within and across the three TNBC subtypes under study, and that a subset of genes contribute to this significant variation. Estimates of *P*-values derived from ANOVA are presented in Table B (provided as supplementary data). ANOVA produced 125 genes which exhibited highly significant ($P < 10^{-5}$) variation among the three subtypes of TNBC studied (Table B). The remaining 63 genes exhibited moderate, little, or no variations among the three subtypes of TNBC studied (Table B).

While ANOVA identified genes which show significant differences in expression profiles among the three TNBC subtypes, it could not identify differentially expressed sets of genes distinguishing the subtypes of TNBC. Therefore, we performed additional supervised analysis comparing gene expression profiles between individual subtypes of TNBC. We hypothesized that gene expression profiles significantly differ between individual subtypes of TNBC under study. Estimates of *P*-values along with FDR are presented in Table B (provided as supplementary data). A comparison between normal-like and basal-like subtypes produced 103 highly significantly ($P < 10^{-5}$) differentially expressed genes distinguishing the subtypes. When we compared normal-like to basal, we identified 37 highly significantly ($P < 10^{-5}$) differentially expressed genes distinguishing the two subtypes. In contrast, a comparison between basal-like and basal produced 95 highly significantly ($P < 10^{-5}$) differentially expressed genes distinguishing the two subtypes (Table B, supplementary data). The results confirmed our hypothesis that genes containing SNPs associated with an increased risk of developing TNBC significantly differ in their expression between the subtypes of TNBC and that genetic susceptibility may be subtype-specific.

Functional relationship

To further refine the genetic susceptibility landscape and identify functionally related genes with similar patterns of expression profiles, we performed unsupervised analysis using hierarchical clustering on the set of genes which were significantly ($P < 10^{-5}$) associated with TNBC. We hypothesized that genes containing SNPs associated with an increased risk of developing breast cancer have similar patterns of expression profiles and are functionally related. The rationale is that genes with similar patterns of expression profiles and similar functions are likely to be regulated via the same regulatory mechanisms.[29] To investigate the biological functions, biological process, and cellular process in which the genes are involved, we performed GO analysis as described in the methods section.

Figure 3 presents patterns of expression profiles of SNP-containing genes for each subtype and the controls. Figure 3A represents patterns of gene expression profiles for the 98 genes in normal-like subtype and controls. Figure 3B depicts patterns of

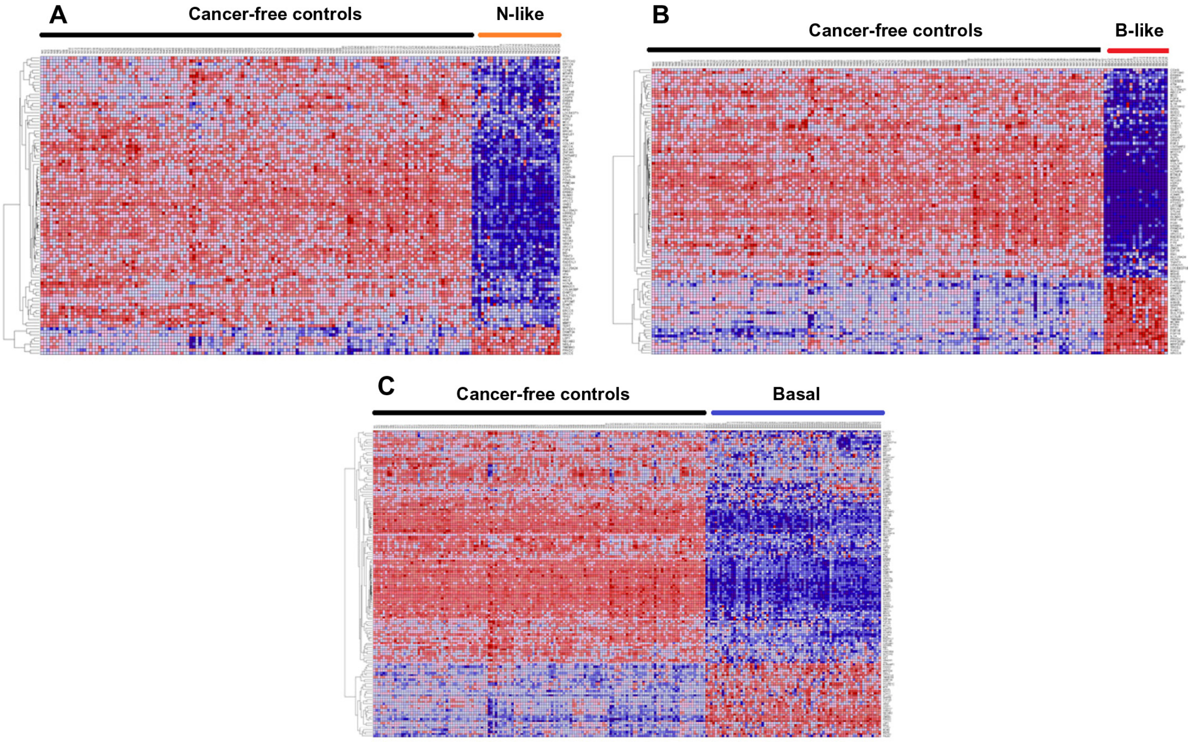

Figure 3. Patterns of gene expression profiles for genes containing SNPs associated with an increased risk associated with an increased risk of developing breast cancer, which were highly significantly ($P < 10^{-6}$) associated with TNBC subtypes. (**A**) shows patterns of expression profiles for the 98 SNP-containing genes obtained by comparing expression profiles between normal-like breast cancer patients (N = 29) and 142 cancer-free controls. (**B**) Patterns of gene expression profiles for the 101 SNP-containing genes in basal-like patients (N = 20) and 142 cancer-free controls. (**C**) Patterns of gene expression profiles for the 142 SNP-containing genes in non-luminal basal patients (N = 75) and 142 cancer-free controls.
Notes: Genes are represented in the roles and samples in the columns. Red and blues colors indicate up and down regulation, respectively.

gene expression profiles for the 101 genes in basal-like subtype and controls. Patterns of gene expression profiles for the 142 genes in basal subtype and controls are presented in Figure 3C. Patterns of gene expression profiles for the 159 genes showing consistent patterns across all the three subtypes are presented in Figure 4. In all four cases examined, genes were co-expressed and exhibited similar patterns of expression profiles. Interestingly, genes containing genetic variants with large effects and genetic variants replicated in multiple independent GWAS studies were co-expressed with genes containing genetic variants with small to moderate effects. Additional analysis using GO information revealed that genes containing SNPs associated with an increased risk of developing breast cancer are functionally related and are involved in the same biological processes and cellular components. A full catalogue of the physiological functions, biological processes, and cellular components in which the genes containing SNPs associated with risk for breast cancer are involved is presented in Table C (provided as supplementary data). These analyses confirmed our hypothesis that genes containing SNPs associated with an increased risk of developing breast cancer are functionally related. This is a significant finding given that traditional single-SNP GWAS analysis does not provide information about the functional relationship of genes containing SNPs associated with risk for developing breast cancer. Importantly, these data indicate that it is reasonable to use gene expression data as an intermediate phenotype to assess the association of GWAS information with TNBC and to gain biological insights about the broader context in which SNPs operate.

To assess the significance of SNP-containing genes as potential biomarkers, we examined their functional relationship with high-penetrance genes (*BRCA1, BRCA2, TP53, PTEN, RB1, CDKN2A, ATR*), moderate-penetrance genes (*ATM, BRIP1, CHEK2*, and *PALB2*), and low-penetrance genes (*FGFR2, TOX3, MAP3K1, LSP1, CASP8*) genes.[1] These genes have been associated with an increased high risk of developing breast cancer.[1] Importantly, these genes contain genetic variants that have been replicated in multiple independent GWAS studies[2] and contain mutations involved in TNBC.[37] The results revealed functional relationships and similarity in patterns of gene expression profiles between these sets of genes and other SNP-containing genes (Fig. 4). This indicates that the mechanisms of action in some common low-risk genetic susceptibility loci are likely to be

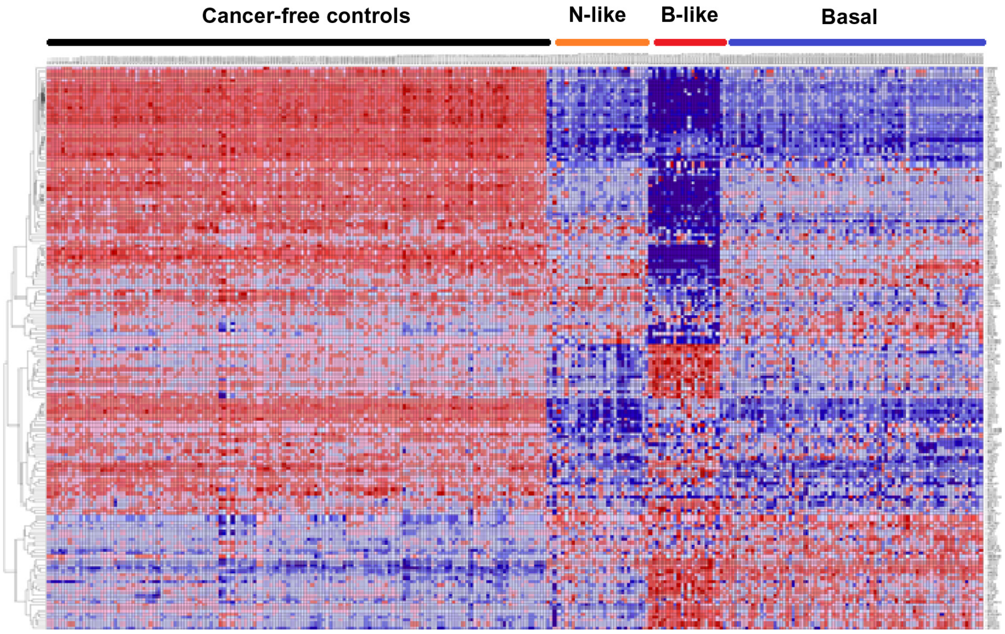

Figure 4. Patterns of gene expression profiles for the 159 genes containing SNPs associated with an increased risk of developing breast cancer, which were highly significantly ($P < 10^{-6}$) associated with the TNBC subtypes under study.
Notes: Genes are represented in the roles and samples in the columns. Red and blues colors indicate up and down regulation, respectively. Note that this is a combination of Figure 3A–C.

activated through activation of these genes. It is also conceivable that these genes could regulate downstream SNP-containing target genes. This is a significant finding given that, for example, *BRCA1* and *BRCA2* confer approximately a 10 to 20 fold relative risk and mutations in these genes account for 16% of the familial risk of breast cancer.[1,42] Notably, most of the *BRCA1* and *BRCA2* cancers are TNBC. For example, in a recent study involving 1, 469 patient aged 20 to 49 from Los Angeles County in California, 48% of *BRCA1* carriers had TNBC compared to 12% of non-carriers.[42] Additionally, recent studies have shown that the *P53* gene could be used to stratify TNBC in subgroups with distinct predictive and prognostic value.[43,44]

Association of SNP-containing genes with novel genes

One of the limitations of GWAS is that the susceptibility loci identified thus far are few and explain only a small fraction of the variation. This begs the question of where the missing variation is located. It is plausible that there are potentially many yet-to-be discovered common susceptibility alleles with smaller effects missed by GWAS. Even if the genes containing genetic variants associated with an increased risk of developing breast cancer are identified, it may not be obvious which genes mediate their biological effects. Therefore, key driver genes may be overlooked and important pathways may be missed by focusing solely on genes containing genetic variants associated with an increased risk of developing breast cancer. To address this question, we performed supervised analysis on the non-prioritized data set comparing gene expression profiles between each TNBC subtype and controls in order to identify novel genes which are significantly differentially expressed and co-expressed with SNP-containing genes. This analysis produced a set of 97 significantly ($P < 10^{-6}$) differentially expressed novel genes. We then combined data on this set of genes with data on the 159 SNP-containing genes associated with TNBC and performed hierarchical clustering to determine whether the two sets of genes are functionally related and have similar patterns of expression profiles. The results showing patterns of expression profiles for the combined set of 256 genes are presented in Figure 5. SNP-containing genes were found to have similar patterns of expression profiles and were also found to be functionally related with genes not identified through GWAS analysis. Interestingly, genes containing SNPs with low to moderate effect size were

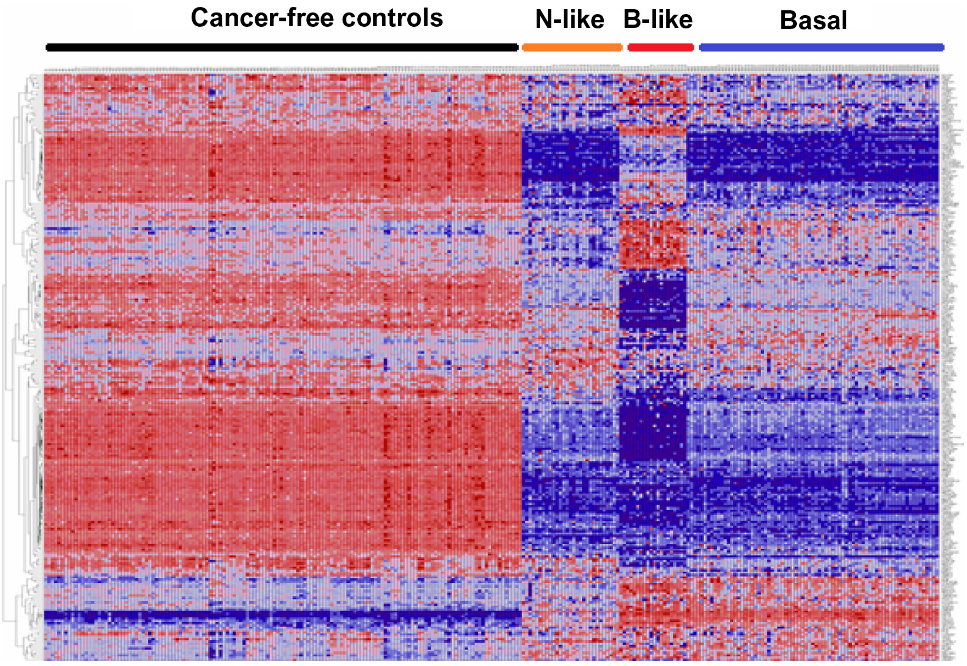

Figure 5. Patterns and clusters of gene expression profiles for the 256 genes (159 SNP-containing genes and 97 novel genes not identified by GWAS, highly significantly associated with TNBC) in different subtypes of TNBC and controls.
Notes: Genes are represented in the roles and samples in the columns. Red and blues colors indicate up and down regulation, respectively.

found to be functionally related with genes not identified through GWAS analysis. This is an interesting finding given that relatively few SNPs have *P*-values sufficiently small enough to give conclusive evidence of association and the fact SNPs identified thus far explain only a small fraction of the variation. The co-expression of genes containing SNPs with genes not identified through GWAS analysis indicates that integrating GWAS information with gene expression data could provide insights about the missing variation in GWAS studies. Overall, the results demonstrate that molecular perturbation leading to the TNBC phenotype involves other genes in the genome acting in concert with SNP-containing genes and that use of GWAS information alone may miss potential biomarkers.

Biological pathways and gene networks

The second objective of this study was to investigate the broader context in which genetic variants and genes associated with TNBC operate and to identify gene regulatory networks and biological pathways enriched for SNPs, which are involved in TNBC. We hypothesized that genes containing SNPs associated with an increased risk of developing breast cancer, which are significantly associated with TNBC, interact with each other and with other genes not identified through GWAS. To test this hypothesis, we performed network analysis and pathway prediction using all 259 genes (SNP-containing and novel genes) as described in the methods section. We identified five top networks with the highest scores (predicted scores ranging from 28 to 51). The five top regulatory networks identified contained genes with multiple overlapping functions and involved multi-gene pathways. The first (network 1, score 51) contained genes involved in DNA repair, DNA mismatch repair, DNA replication and recombination, cell cycle, and nucleic acid metabolism. The second (network 2, score 41) produced genes involved in the DNA replication, recombination and repair, cellular compromise, and cell cycle. The third (network 3, score 33), contained genes involved in cell cycle, cellular development, cellular growth, and proliferation. The fourth (network 4, score 32) contained genes involved in cancer and organismal development. The fifth (network 5, score 28) contained genes involved in apoptosis. In all the networks identified, SNP-containing genes significantly associated with TNBC and other genes were functionally related.

To further refine the genetic susceptibility landscape, we mapped the genes onto the networks focusing on the most significant network. Figures 6–8 show the gene regulatory networks of SNP-containing genes and genes not identified through GWAS analysis. Network modeling and visualization revealed that SNP-containing genes significantly associated with TNBC interact with each other and with other genes not identified by GWAS confirming our hypothesis. Network analysis further revealed that genes containing SNPs with large effects and SNPs replicated in multiple independent studies interact with genes containing SNPs with small effects and not replicated. Among the identified SNP-containing genes were *ATR, ATM, NBN, BLM, WRN, XRCC1, XRCC2, XRCC3, RPA1, BRCA1, BRCA2, RAD51, TP53, H19, RPA2, XPA, ERCC5, FRMD4A, COMT, TMEM43, ERCC6, ERCC2* and *PALB2*. (Fig. 6) These genes were predicted to be involved in DNA repair, DNA mismatch repair, base-excision repair, cell cycle, and nuclei acid metabolism. Additional network analysis revealed the genes *PMS1, ATM, ATR, DCLRE1C, PRKDC, XRCC5, KSR2, CHEK2, ERBB2, ERBB4, ESR1, CHEK1, BRCA1, CYP1B1, PGR, CYP19A1, MSH2, MSH3* and *FANCA* (Fig. 7). The genes were predicted to be involved in DNA replication, recombination, repair, cellular compromise, and cell cycle. Further delineation of the network produced the genes *ALPL, CDK6, NUMA1, CCND1, FOXM1, CDKN2B, SUV39H2, CCNE1, EHMT1, EHMT2, TERT, MYCL1, TYMS, RB1, CDKN1A, MCMB* and *DNMT3A* which are involved in cell cycle, cellular development, cellular growth, and proliferation (Fig. 8).

A close examination of the genes in the networks on the basis of molecular and cellular function revealed 66 genes which are involved in DNA repair, DNA mismatch repair, and base-excision repair to be the most significant ($P = 7.50 \times 10^{-29}$ to 5.41×10^{-7}), followed by 103 genes involved in cell death (1.66×10^{-21} to 6.92×10^{-7}), 75 genes involved in cell cycle (1.07×10^{-19} to 6.92×10^{-7}), 23 genes involved in cellular response to therapeutics (5.99×10^{-15} to 1.19×10^{-9}), and 31 genes involved in cellular compromise (2.37×10^{-14} to 5.41×10^{-7}). Overall, most of the genes identified have multiple

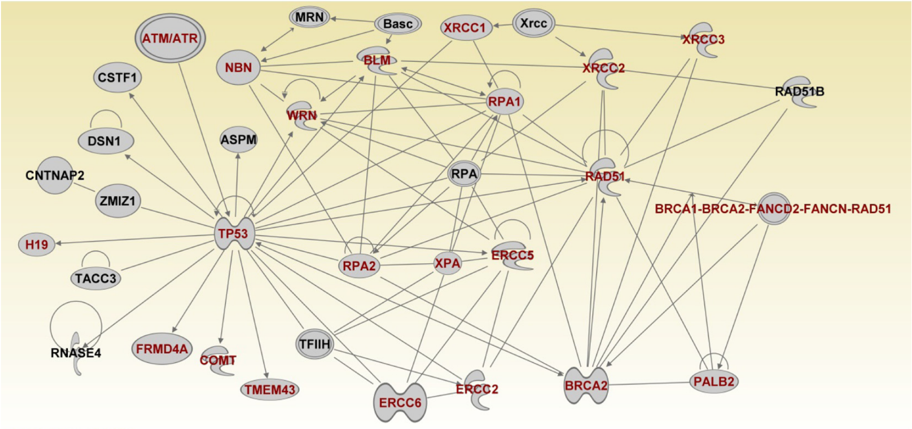

Figure 6. (Network 1). Gene regulatory networks containing multigene pathways involved in DNA replication, recombination, and repair, cell cycle, nucleic acid metabolism.
Note: Red fonts indicate genes containing SNPs.

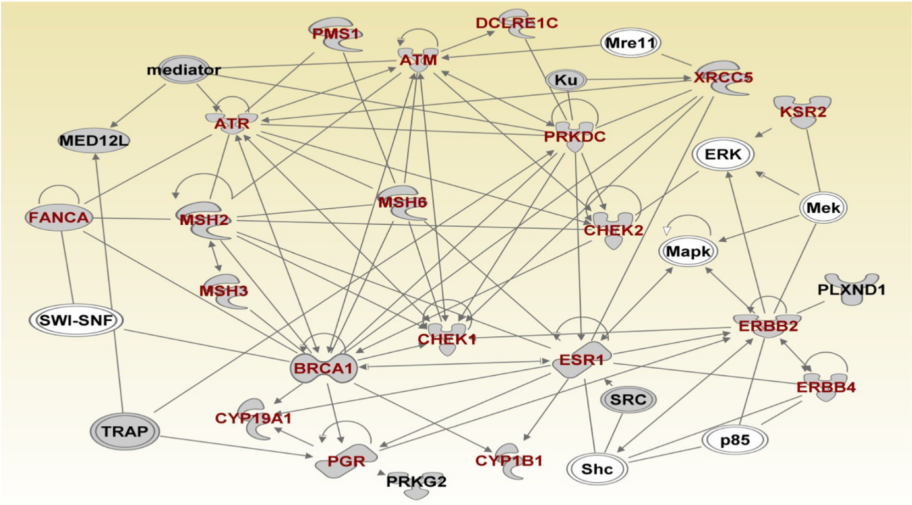

Figure 7. (Network 2). Gene regulatory network containing multigene pathways involved in DNA replication, recombination, and repair, cellular compromise, and cell cycle.
Note: Red fonts indicate genes containing SNPs.

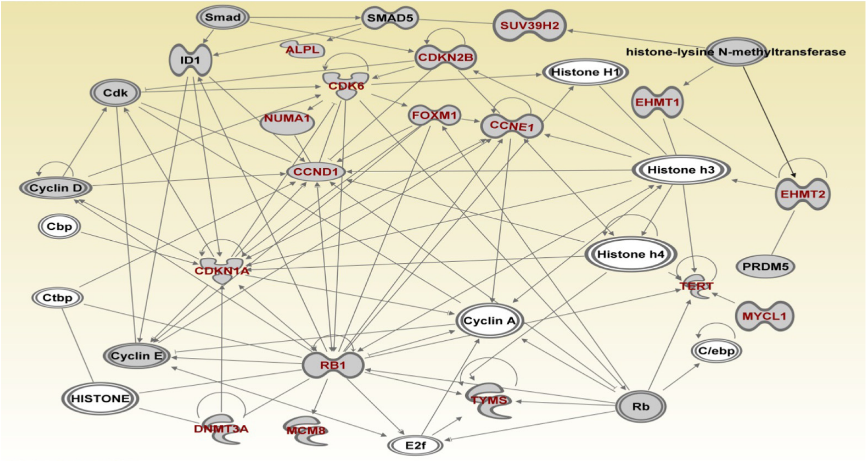

Figure 8. (Network 3). Gene regulatory network containing multigene pathways involved in cell cycle, cellular development, cellular growth and proliferation.
Note: Red fonts indicate genes containing SNPs.

overlapping molecular functions and are involved in similar biological and cellular processes. Of the 258 genes investigated, 130 were directly classified as cancer genes by the IPA and were the most significant ($P = 2.24 \times 10^{-23}$ 6.75×10^{-7}). The information on molecular functions and biological and cellular processes for all 256 genes are presented in Table C (provided as supplementary data to this report).

To further determine the broader context in which genes containing genetic variants operate and to establish the functional bridges between GWAS findings and biological pathways relevant to TNBC, we performed pathway prediction. Specifically, we mapped the genes significantly associated with TNBC onto the pathways in the IPA database. The predicted pathways were ranked on predicted *P*-values, after correcting for multiple testing. Pathway prediction revealed that SNP-containing genes significantly associated with TNBC interact with each and with other genes not identified by GWAS, in cascades of complex multi-gene pathways. We identified many multi-gene pathways. The most highly significant pathways included the hereditary breast cancer signaling pathway ($P = 3.86 \times 10^{-17}$), the role of BRCA1 in DNA repair response ($P = 4.17 \times 10^{-17}$), the Aryl hydrocarbon receptor signaling pathway ($P = 2.44 \times 10^{-13}$), the DNA double-strand break repair by non-homologous end joining ($P = 1.99 \times 10^{-10}$), and the role of CHK proteins in cell cycle checkpoint control ($P = 2.98 \times 10^{-10}$).

Figure 9 shows the SNP-containing genes mapped to the *BRCA1* in the DNA damage response pathway. SNP-containing genes significantly associated with TNBC and which mapped to this pathway included *FANCA, ATM, ATR, CHECK1, CHEK2, P53, BRCA1, BARD1, BLM, MSH2, MSH6* and *RAD50* (Fig. 9). These genes are involved in DNA repair, mismatch repair, and base-excision repair. This is a significant finding given that the majority of BRCA1 tumors are TNBC.[12] Both *BRCA1* and *BRCA2* have been implicated in double-strand DNA repair.[45] The gene *ATM* has been shown to encode a checkpoint kinase that has key functions in DNA repair and which also phosphorylates *P53* and *BRCA1*.[46] Given that impaired base-excision repair functions can give rise to the accumulation of DNA damage and initiation of cancer,[47] and the fact that chemotherapy is the only effective therapeutic modality for TNBC that can cause DNA damage,[45] targeting the DNA repair defects in this pathway could potentially serve as an effective therapeutic strategy.[48]

To further explore the role of SNP-containing genes in biological pathways relevant to TNBC, we explored the role of CHK proteins in the cell cycle

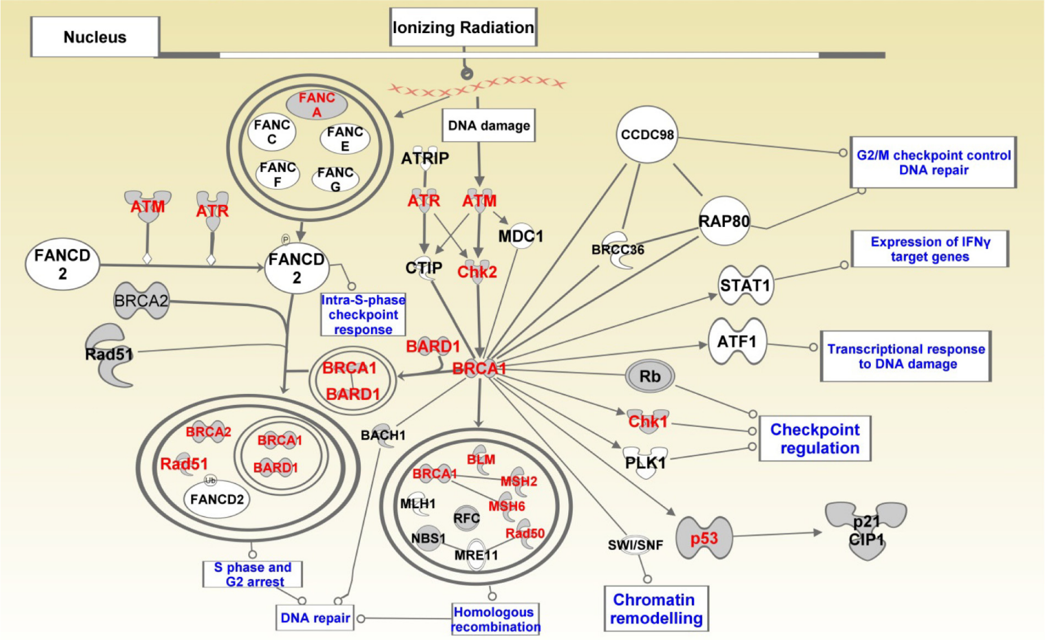

Figure 9. BRCA pathway showing the role of BRCA1 in DNA damage response and crosstalk with other biological pathways, notably the P53 pathway.
Notes: Double circles indicate complex interactions involving multiple genes. Genes containing SNPs are shown in red fonts.

checkpoint control pathway (Fig. 10). This pathway contained 6 SNP-containing genes *RAD50, ATM, CHEK1, CHEK2, BRCA1* and *P53* (Fig. 10). Identification of this pathway was of particular interest because the most recent study has shown that targeting *CHEK1* in P53-deffiecient triple-negative breast cancer is therapeutically beneficial in human-in-mouse tumor models.[49] *CHEK1* and *CHEK2* are checkpoint kinases involved in DNA repair that directly modulate the activities of *TP53* and *BRCA1* through phosphorylation. Thus, targeting these pathways could be therapeutically beneficial. In particular, the *P53* gene has been shown to be a specific prognostic factor in TNBC.[50] In addition, the DNA damage signaling kinase *ATM* has been shown to be aberrantly reduced or lost in BRCA1/BRCA2-deficient breast cancer and TNBC.[46] Overall, the pathway prediction and network modeling confirmed our hypothesis that genes containing SNPs interact with each other and other genes not identified by GWAS in biological pathways. In the predicted pathways and networks, genes containing SNPs were found to interact with genes not identified by GWAS Figures 9 and 10. This is a significant finding regarding the missing variation as it suggests that using GWAS alone may miss potential therapeutic targets.

In all predicted pathways enriched for SNPs, genes containing SNPs with large effects and SNPs replicated in multiple studies were found to be interacting with genes containing SNPs with small to moderate effects. This is a significant finding given that overall, only a small number of statistically unimpeachable, common low-penetrance breast cancer susceptibility alleles have thus far been reported in TNBC in different populations.[16,17] The identification of many multigene pathways enriched for SNPs indicates that many loci and pathway crosstalk may be involved in the pathogenesis of TNBC. The association of the DNA mismatch repair genes and pathways with TNBC is of particular interest both because ensuring fidelity of DNA replication is central to preserving genomic integrity and because DNA mismatch repair is critical for maintaining the fidelity of replication.[51] The functional relationship, similarity in patterns of gene

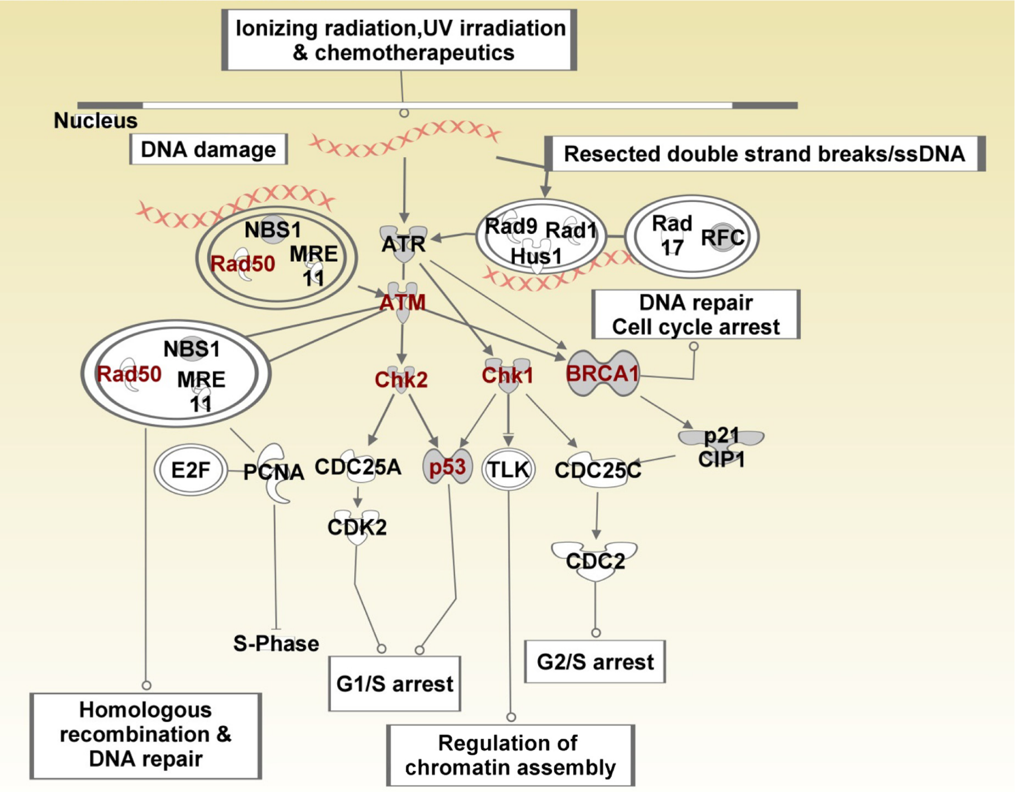
Ionizing radiation,UV irradiation & chemotherapeutics
Nucleus
DNA damage
Resected double strand breaks/ssDNA
NBS1
Rad50
MRE 11
ATR
Rad9
Rad1
Hus1
Rad 17
RFC
ATM
DNA repair Cell cycle arrest
NBS1
Rad50
MRE 11
Chk2
Chk1
BRCA1
p21 CIP1
E2F
PCNA
CDC25A
p53
TLK
CDC25C
CDK2
CDC2
S-Phase
G1/S arrest
G2/S arrest
Homologous recombination & DNA repair
Regulation of chromatin assembly

Figure 10. CHEK kinase pathway showing the role of CHEK proteins in cell cycle checkpoint control, and crosstalk with other pathways notably the P53 and BRCA pathways.
Notes: Double circles indicate complex interactions involving multiple genes. Genes containing SNPs are shown in red fonts.

expression profiles, and interactions in gene regulatory networks and pathways between different sets of genes, highlights the complexity of the molecular mechanisms involved in TNBC. An exciting result with potential therapeutic and clinical significance in this study is that genes with high penetrance (*BRCA1, BRCA2, TP53, PTEN, STK11* and *CDH1*), moderate penetrance (*PALB2, BRIP1, ATM,* and *CHEK1*) and genes with lower penetrance (*CASP8, FGFR2, TOX3, MAP3K1*, and *LSP1*), all of which contained genetic variants, were significantly associated with TNBC. Importantly, these genes were found to be functionally related and interacted with other SNP-containing genes and novel genes not identified by GWAS. This is a significant finding in that although we do not fully comprehend the genetic basis of TNBC, and perhaps rare variants are yet-to-be found, the mechanism of action for at least some SNP-containing genes associated with TNBC may be through activation or regulation of breast cancer-predisposing loci. Although we did not integrate findings with clinical outcomes, evidence from the literature suggests that some of the genes identified in this study—notably *MAP3K1, DAPK1, LSP1, MMP7, TOX3*, and *ESR1*—are associated with survival.[52]

Discussion

The primary goal of this study was to determine whether genes containing SNPs associated with an increased risk of developing breast cancer are associated with and could stratify different subtypes of TNBC. In addition, we sought to identify molecular pathways and networks relevant to TNBC. Our analysis establishes the association between genes containing SNPs associated with an increased risk of developing breast cancer and TNBC. It further identifies gene regulatory networks and biological pathways enriched for SNPs, which are involved in TNBC. This is a significant finding because to date, very few genetic variants and genes associated with an increased risk of developing TNBC have been reported.[16,17] Many examples highlight the power of transcription profiling (signatures) in informing an understanding of the molecular basis and stratify subtypes of TNBC,[8,13,14] as well as in the prediction of clinical outcomes.[8,14,15,52,53] However, although these studies have made great strides in deciphering the molecular basis of TNBC, they have been unsuccessful in determining which genes have causative roles as opposed to being consequences of breast cancer state.[2] Recently our group reported integration of GWAS with gene expression data.[2,54,55] However, this is the first study to link genes containing SNPs associated with an increased risk of developing breast cancer both with the TNBC intermediate phenotype and with the identification of biological pathways enriched for SNPs, which are involved in TNBC.

Importantly, this study indicates that combining GWAS information with gene expression data provides a powerful approach to identification of potential predictive biomarkers involved in TNBC. Predictive markers in TNBC will be particularly important because in the absence of effective therapy, these tumor subtypes tend to have poor prognosis.[8] Although, the ability to interpret the direct effects of the genetic variants on genes and pathways remains a challenge, this does not diminish the power of integrative analysis presented here to provide insights about the broader context in which genetic variants operate. It also establishes functional bridges between GWAS findings and the TNBC intermediate phenotype. Another novel feature of this integrative genomics approach is that it allows identification of additional genes which could not be identified using GWAS alone.

The practical significance of our approach is that it can be used to identify candidate genes to prioritize for sequencing. By prioritizing and evaluating SNP-containing genes using gene expression profiles, we were able to identify not only the most promising genes but also candidate pathways. This study therefore has the important goal of maximizing the yield and biological relevance of further downstream screens, experimental validation, functional studies, and targeted sequencing (by focusing on the most promising genes and biological pathways). Although we did not perform experimental validation, we used functional and co-expression analysis along with pathways prediction and network modeling in order to prioritize the genes on the basis of putative links to other genes—notably high, moderate, and low penetrance genes that have more established roles as key drivers of breast cancer.[1] Future research directions for prioritization will focus primarily on broadening the scope of this study beyond transcription profiling of SNP-containing genes to include sequence information.

A key opportunity and component will be the prioritization of genetic variants and associated genes for the purposes of next generation sequencing and elucidating the impact of genetic variants on gene and pathway function. Such work was beyond the scope of this report, but it is ongoing and will be reported elsewhere. Although we did not perform sequencing, some of the genes identified in this study (notably *TP53, RB1, PTEN, ATM, ATR, MAP3K1, BRCA1, BRCA2*, and *ERBB2*) were reported in the most recent and first study to perform deep-sequencing on 104 primary TNBC tumors (with the goal of identifying mutations).[37]

One caveat is important in this study. We observed significant heterogeneity in patterns of expression profiles and overlap was incomplete. This suggests to the research community that to make headway against TNBC, we researchers must first come to grips with the burgeoning data from this and other studies,[8] showing that this subgroup of breast cancer is highly heterogeneous. The heterogeneity observed in this study indicates that TNBC is not a single disease entity and that genetic susceptibility could potentially be TNBC subtype-specific. Thus, identification of predictive risk markers must be conducted with that in mind.

It is worth noting that until recently, the only candidates for defining TNBC were mutations in *BRCA1, TP53*, and *RB1* genes. However, in this study SNP-containing genes were functionally related, co-expressed, and interacted with these genes in biological pathways. This indicates that the molecular basis of TNC is far more complex and involves many genes and multi-gene pathways, with each gene likely contributing a small effect. These findings coupled with the observed heterogeneity in patterns of expression profiles indicate that there are many other genes acting in concert with these genes in order to produce the TNBC phenotype. We argue that associating GWAS information with the TNBC phenotypes is the first step to understanding the broader context in which genetic variants operate.

The results found in this study provide convincing evidence that genes containing SNPs associated with an increased risk of developing breast cancer are associated with TNBC and the identification of biological pathways enriched for SNPs, which are involved in TNBC. However, several limitations of this study must be acknowledged. First, we used publicly available GWAS information and gene expression data. The results could potentially be influenced by factors contained in use of such information and are beyond the scope of this report. Second, we did not investigate allele-specific gene expression. Current knowledge about how SNPs identified by GWAS—particularly those mapped to noncoding and inter-genic regions—regulate gene expression and pathways remains sketchy at best. However, we can now at least begin to understand the broader context in which they operate. Moreover, because the disease-causing alleles are likely uncommon, it is unlikely that they will be identified by association studies.[1] Integrative genomics provides a powerful approach for identifying candidate genes for further downstream screening. Although we did not investigate allele-specific expression, an earlier report on breast cancer confirmed that alleles in genes *CASP8, TOX3* (previously known as *TNRC9*) and *ESR1* affect gene expression.[52] Additionally, there is anecdotal evidence from literature that allele-specific gene expression is widespread across the genome.[56,57] In fact, allele-specific variation and gene expression differences in humans have been reported.[58,59]

Importantly, the majority of the GWAS information, as well as gene expression data used in this study, were derived from Caucasian populations. TNBC preferentially affects young African-American women.[60] While TNBC represents about 15% to 20% of all diagnosed breast cancers in the general US population, it constitutes about 30% in the African-Americans.[61] In fact, a recent study revealed populations differences and over-representation of TNBC in indigenous African women.[62] It is conceivable that genetic variants may confer population-specific risks depending on exposure. In the absence of gene expression data on the African-American population, we were not able to address that question, though it warrants investigation in future. Accordingly, we view this study as exploratory and the results found here cannot be generalized to different ethnic populations.

In conclusion, this study provides convincing evidence that genes containing SNPs associated with an increased risk of developing breast cancer are significantly associated with and could stratify the TNBC intermediate phenotypes. The study further reveals molecular pathways and networks enriched for SNPs, which are involved in TNBC. Based on these results we recommend that an integrative genomics approach combining GWAS information and gene expression data provides a unified and powerful approach to identification of potential biomarkers and molecular pathways in TNBC. More studies directed at understanding how the genetic variants regulate gene expression in target populations, notably the African-American population, are needed.

Acknowledgements

This study was supported by funding from the University of Mississippi Cancer Institute, to whom the authors are very thankful.

Author Contributions

Conceived and designed the experiments: CH. Analysed the data: CH, RK ,KB. Wrote the first draft of the manuscript: CH. Contributed to the writing of the manuscript: CH, AP, AB, JM, KB, LM. Agree with manuscript results and conclusions: CH, RK, KB, AP, AB, JM, LM . Jointly developed the structure and arguments for the paper: CH, RK, LM. Made critical revisions and approved final version: CH, AB, LM. All authors reviewed and approved of the final manuscript.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Stratton MS, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet*. 2008;40(1):17–22.
2. Hicks C, Asfour R, Pannuti A, Miele L. An integrative genomics approach to biomarker discovery in breast cancer. *Cancer Inform*. 2011;10:185–204.
3. Zhang B, Beeghly-Fadiel A, Long J, Zheng W. Genetic variants associated with breast-cancer risk: Comprehensive research synopsis, meta-analysis, and epidemiology evidence. *Lancet Oncol*. 2011;12:477–88.
4. Dent R, Trudeau M, Pritchard KI, et al. Triple-negative breast cancer: Clinical features and patterns of recurrence. *Clin Cancer Res*. 2007;13:4429–34.
5. Carey L, Winer E, Viale G, Cameron D, Gianni L. Triple-negative breast cancer: disease entity or title of convenience. *Nat Rev Clin Oncol*. Dec 2010; 7(12):683–92.
6. Bosch A, Eroles P, Zaragoza R, Vina JR, Lluch A. Triple-negative breast cancer: Molecular features, pathogenesis, treatment and current lines of research. *Cancer Treat Rev*. 2010;36:206–15.
7. Hudis CA, Gianni L. Triple-negative breast cancer: An unmet medical need. *Onclogist*. 2011;16(Suppl 1):1–11.
8. Perou CM. Molecular stratification of triple-negative breast cancers. *Oncologist*. 2010;15(Suppl 5):39–48.
9. Schneider BP, Winer EP, Foulkes WD, et al. Triple negative breast cancer: risk factors to potential targets. *Clin Cancer Res*. 2008;14:8010–8.
10. Curigliano G, Goldhirsch A. The triple-negative subtype: New ideas for the poorest prognosis breast cancer. *J Natl Cancer Inst Monogr*. 2011;43: 108–10.
11. Yang W-T, Dryden M, Broglio K, et al. Mammographic features of triple-negative primary breast cancers in young premenopausal women. *Breast Cancer Res Treat*. Oct 2008;111(3):405–10. Epub Nov 17, 2007.
12. Dawood S. Triple-negative breast cancer. Epidemiology and management options. *Drugs*. 2010;70(17):2247–58.
13. Turner N, Lambros MB, Horlings HM, et al. Integrative molecular profiling of triple-negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene*. 2010;29(14):2013–23.
14. Kreike B, van Kouwenhove M, Horlings H, et al. Gene expression profiling and histopathology characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Res*. 2007;9:R65.
15. Lehmann BD, Bauer JA, Chen X, et al. Identification of triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*. 2012. doi:10.1172/JCI45014.
16. Stevens KN, Vachon CM, Lee AM, et al. Common breast cancer susceptibility loci are associated with triple-negative breast cancer. *Cancer Res*. 2011;71(19):6240–9.
17. Stevens KN, Fredericksen Z, Vachon CM, et al. 19p13.1 Is a triple-negative-specific breast cancer susceptibility locus. *Cancer Res*. 2012;72(7): 1975–803.
18. Ioannidis JP, Boffetta P, Little J, et al. Assessment of cumulative evidence on genetic associations: Interim guidelines. *Intl J Epidemiol*. 2008;37: 120–32.
19. Khoury MJ, Bertram I, Boffetta P, et al. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation in human diseases. *Am J Epidemiol*. 2009;170:269–79.
20. Sagoo GS, Little J, Higgins JP. Systematic reviews of genetic association studies. *PLoS Med*. 2009;6:e28.
21. Moher D, Liberati A, Tetzlaff J, Altman DG; for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151:264–9.
22. Liberati A, Altman DG, Tetzlaff J, Mulrow C, et al. The PRISMA statement for reporting systematic reviews and meta-analyses studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med*. 2009; 6(7):e1000100.
23. Holmans P, Green EK, Pahwa JS, et al. Gene ontology analysis of GWA study data sets provides insights into biology of bipolar disorder. *American J Hum Genet*. 2009;85:13–24.
24. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. *N Eng J Med*. 2010;363:1938–48.

25. Sabatier R, Finetti P, Cervera N, et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res Treat*. 2011;126:407–20.
26. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl Acids Res*. 2002;30(1):207–10.
27. Richardson AL, Wang ZC, De Nicolo A, et al. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*. 2006;9:121–32.
28. Chen D, Nasir A, Culhane A, et al. Proliferative genes dominate malignancy-risk gene signature in historically-normal breast cancer: *Breast Cancer Res Treat*. 2010;119(2):335–46.
29. 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.
30. Reich M, Liefield T, Gould J, Lerner J, Tamayo P. GenePattern 2.0. *Nat Genet*. 2006;38(5):500–1.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc*. *Series B Methodology*. 1995;57(1):289–300.
32. RadMacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol*. 2002;9(3):505–11.
33. Morrissey ER, Diaz-Uriarte R. Pomello II: Finding differentially expressed genes. *Nucl Acids Res*. 2009;37:W581–6.
34. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontolgy Consortium. *Nat Genet*. 2000;25(1):25–9.
35. Ingenuity Pathways Analysis (IPA) system. Ingenuity Incorporated, California.
36. Rakha EA, Elsheikh SE, Aleskandarany MA, et al. Triple-negative breast cancer: Distinguishing between basal and nonbasal subtypes. *Clin Cancer Res*. 2009;15(7):2302–10.
37. Shah SP, Roth A, Goya R, Oloumi A, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*. Apr 4, 2012; 486(7403):395–9.
38. Curtis C, Shah SP, Chin SF, et al. The genomic transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature*. Apr 18, 2012;486(7403):346–52.
39. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet*. 2008;40:703–6.
40. Figueroa JD, Garcia-Closas M, Humphries M, et al. Association of common variants at 1p11.2 and 14q24.1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: Findings from the Breast Cancer Association Consortium. *Hum Mol Genet*. 2011;20:4693–706.
41. Broekes A, Schmidt MK, Sherman ME, et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: Findings from Breast Cancer Association Consortium. *Hum Mol Genet*. 2011;20:3289–303.
42. Lee E, McKean-Cowdin R, Ma H, et al. Characteristics of triple-negative breast cancer in patients with a BRCA1 mutation: Results from a population-based study of young women. *J Clin Oncol*. 2011;29:4373–80.
43. Alberti S, Ambrogi F, Pedriali M, et al. p53 status splits triple-negative breast cancer in subgroups with distinct predictive and prognostic potential value. *Cancer Res*. 2009;69(24 Suppl 3).
44. Biganzoli E, Coradini D, Ambrogi F, et al. p53 status identifies two subgroups of triple-negative breast cancers with distinct biological features. *Jpn J Clin Oncol*. 2011;41(2):172–9.
45. Silver DP, Richardson AL, Eklund AC. Efficacy of Neoadjuvant cisplatin in triple-negative breast cancer. *J Clin Oncol*. 2010;28(7):1145–53.
46. Tommiska J, Bartkova J, Heinonen M, et al. The DNA damage signaling kinase ATM is aberrantly reduced or lost in BRCA1/BRAC2-diffient and ER/PR/ERBB2-triple-negative breast cancer. *Oncogene*. 2008;27: 2501–6.
47. Zhang Y, Newcomb PA, Egan KM, et al. Genetic polymorphisms in base-excision repair pathway genes and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2006;15(2):353–8.
48. Farmer H, McCabe N, Lord CJ, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*. 2005;434:917–20.
49. Ma CX, Cai S, Li S, et al. Targeting Chk1 in p53-diffient triple-negative breast cancer is the therapeutically beneficial in human-in-mouse tumor models. *J Clin Invest*. 2012;122(4):1541–52.
50. Chae BJ, Bae JS, Lee A, et al. p53 as a specific prognostic factor in triple-negative breast cancer. *Jpn J Clin Oncol*. 2009;39(4)217–24.
51. Hsieh P, Yamane K. DNA mismatch repair: Molecular mechanisms, cancer, and ageing. *Mech Age Develop*. 2008;129:391–407.
52. Tapper W, Hammond V, Gerty S, et al. The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Res*. 2008;10(6):R108.
53. Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*. 2010;12:R68.
54. Hicks C, Pannuti A, Miele L. Associating GWAS information with the Notch signaling pathway using transcription profiling. *Cancer Inform*. 2011;10:93–108.
55. Hicks C, Kumar R, Pannuti A, Miele L. Integrative analysis of response to tamoxifen treatment in ER-positive breast cancer using GWAS information and transcription profiling. Breast Cancer: *Basic Clin Res*. 2012:47–66.
56. Paracios R, Gazave E, Goni J, et al. Allele-specific gene expression is widespread across the genome and biological processes. *PLoS One*. 2009;4(1):e4150.
57. Cheung VG, Conclin LK, Weber TM, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*. 2003;33:422–3.
58. Buckland PR. Allele-specific gene expression differences in humans. *Hum Mol Genet*. 2004;13(2):R255–60.
59. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. *Science*. 2002;297:1143.
60. Reis-Filho JS, Tutt ANJ. Triple negative tumors: a critical review. *Histopathology*. 2008;52:108–18.
61. Stead LA, Lash TL, Sobieraj JE, et al. Triple-negative cancers are increased in black women regardless of age or body mass index. *Breast Cancer Res*. 2009;11:R18.
62. Huo D, Ikpatt F, Khramtsov A, et al. Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple negative breast cancer. *J Clin Oncol*. 2009;27(27):4515–21.

Supplementary data

Table A. Estimates of *P*-values and false discovery rate (FDR) for SNP-containing genes for individual subtypes of breast cancer compared to the cancer-free controls (Supplementary data).

Abbreviations: N, cancer-free controls; NLK, normal-like; BLK, basal-like; basal is non-luminal basal.

Table B. Estimates of *P*-values and FDR for all the three subtypes and between subtypes (Supplementary data).

Table C. GO functional relationships of the 256 SNP-containing and novel genes found to be highly significantly ($P < 10$-5) associated with TNBC subtypes.

Table D. List of genes, SNPs, *P*-values and references of genome-wide association studies for GWAS data used in this study.