

## Research Article

Hicks C, et al. J Oncol Res Ther 6: 10103.

DOI: 10.29011/2574-710X.010103

## Integrating GWAS with Gene Expression Data to Map the Landscape of Luminal Androgen Receptor and Mesenchymal Subtypes of Triple Negative Breast Cancer

Chindo Hicks\*, Jacob Elnaggar, Jiande Wu

Department of Genetics, Louisiana State University Health Sciences Center, School of Medicine, USA

\***Corresponding author:** Chindo Hicks, Department of Genetics, Louisiana State University Health Sciences Center, School of Medicine, 533 Bolivar Street, New Orleans, Louisiana, USA

**Citation:** Hicks C, Elnaggar J, Wu J (2021) Integrating GWAS with Gene Expression Data to Map the Landscape of Luminal Androgen Receptor and Mesenchymal Subtypes of Triple Negative Breast Cancer. J Oncol Res Ther 6: 10103. DOI: 10.29011/2574-710X.010103

**Received Date:** 20 December, 2020; **Accepted Date:** 15 January, 2021; **Published Date:** 18 January, 2021

### Abstract

**Background:** Triple negative breast cancer (TNBC) is the most aggressive and lethal type of breast cancer. It is a heterogeneous disease consisting of many subtypes with distinct molecular and risk profiles. With the exception of cytotoxic chemotherapy, currently there are no effective targeted therapies. There is an urgent need for the discovery of genetic markers that could be used to identify women at high risk of developing subtypes of TNBC at early stages. Here we investigated the potential causal association between genetic susceptibility variants and the two subtypes of TNBC, luminal androgen receptor (LAR) and the mesenchymal (MES) subtypes.

**Methods:** We combined information from genome-wide association studies with gene expression data from the LAR and MES subtypes of TNBC, to identify molecular signatures, gene regulatory networks and signaling pathways enriched for genetic susceptibility variants.

**Results:** The investigation revealed gene signatures, gene regulatory networks and signaling pathways enriched for genetic susceptibility variants associated with the LAR and MES subtypes of TNBC. The networks included genes predicted to be involved in DNA replication, recombination and repair, cell cycle, cell death and cancer. Discovered pathways included the role of BRCA1 in DNA damage response, hereditary breast cancer, aryl hydrocarbon receptor and the molecular mechanisms of cancer signaling pathways.

**Conclusion:** The study revealed that genes containing genetic susceptibility variants are associated with the LAR and MES subtypes of TNBC. Additionally, the study revealed molecular networks and signaling pathways enriched for genetic variants. Further research is recommended to validate the genetic variants in the two subtypes of TNBC.

**Keywords:** Genetic variants; Association; Gene expression; Triple negative breast cancer

### Introduction

Despite remarkable progress in screening and patient management, breast cancer remains the second most diagnosed and the second leading cause of cancer related death in women in the United States [1]. Majority of breast cancers respond to targeted and endocrine therapy. However, a significant proportion (15% - 20%) are triple-negative breast cancers (TNBC), the most aggressive and lethal form of breast cancer [2,3]. TNBC is defined as tumors that lack expression of the oestrogen receptor (ER-), progesterone receptor (PR-) and the human epidermal growth factor (HER-2) [2,3]. It is characterized by poor prognosis, higher

incidences of relapse and poor survival rates [2,3]. TNBC is a heterogeneous disease consisting of many subtypes with distinct risk and molecular profiles [3]. Currently there are no effective targeted therapies, cytotoxic chemotherapy remains the only effective therapeutic modality. Over the last several decades, considerable progress has been made in breast cancer screening using mammography. However, screening for TNBC using this technique has been less effective, in part because TNBC tends to affect younger premenopausal women, primarily African American women [2]. There is an urgent need for the discovery of clinically actionable molecular markers that could be used to identify women at high risk of developing this aggressive and lethal form of breast cancer in early stages to guide therapeutic decision making at the point of care.

Over the last decade, considerable effort has been directed at discovery of genetic variants and genes associated with an increased risk of developing breast cancer using genome-wide association studies (GWAS) [4,5]. Hundreds of genetic variants from GWAS have been reported and are now being incorporated into risk prediction models such as polygenic risk scores to identify individuals at high risk of developing breast cancer [6,7]. However, majority of the genetic susceptibility variants reported thus far, are not breast cancer type or subtype specific. This limited progress must be balanced against the recognition that GWAS were designed as case-controls studies without stratification by breast cancer type and or subtype. Recently, there has been increased uptake in the use of germline genetic testing of TNBC patients using hereditary cancer gene panels [8,9]. This has been necessitated by the high frequency of BRCA1 and BRCA2 mutations in patients diagnosed with TNBC [8,9]. However, there is a paucity of information about the causal association between genetic susceptibility variants and subtypes of TNBC. We recently published a study combining germline, somatic and epigenetic variation information to infer the potential causal association between genetic-epigenetic alterations and TNBC [10,11]. However, those studies did not address associations between genetic variants and individual subtypes of TNBC. Accumulating evidence from published studies suggests that gene variants may confer subtype-specific risks and may affect gene expression [5]. In addition, molecular profiles in TNBC has been shown to be subtype-specific [3]. Therefore, there is need to investigate the association between genetic variants and individual subtypes of TNBC.

To begin to address this knowledge gap, we recently published a manuscript associating genes containing genetic susceptibility variants with the Basal-like immune activated (BLIA) and the basal-like immune suppressed (BLIS) subtypes of TNBC and identified gene regulatory networks and signaling pathways enriched for genetic variants [12]. However, the potential causal association between genetic susceptibility variants and the other two subtypes, luminal androgen receptor (LAR) and the mesenchymal (MES) subtypes of TNBC has not been reported. The objective of this exploratory study was to determine whether genes containing genetic susceptibility variants are associated with the LAR and MES subtypes of TNBC and to identify gene regulatory networks and signaling pathways driving these associations. Our working hypothesis was that genes containing genetic variants are associated with the LAR and MES subtypes of TNBC. We further hypothesized these genes are functionally related and interact in gene regulatory networks and signaling pathways enriched for genetic variants. We addressed this hypothesis using publicly available information from GWAS [4,5,13] and gene expression data on the two types of TNBC. For the purposes of clarity, throughout this study we defined and considered single nucleotide polymorphisms (SNPs) associated with an increased risk of

developing breast cancer as “genetic variants”, the genes they map to as the “units of association” and gene expression data as the intermediate phenotype. Thus, our analysis approach focuses on genes, molecular networks and signaling pathways rather than individual genetic variants. This comprehensive approach is designed to gain insights about the broader biological context in which genetic variants operate and to establish putative functional bridges between genetic variants and the signaling pathways they regulate in each subtype of TNBC under study.

## **Materials and Methods**

### **Source of genetic susceptibility variants and genes**

Advances in high-throughput genotyping have enabled discovery of genetic variants and genes associated with an increased risk of developing breast cancer using GWAS [4,5,13]. To date, hundreds of genetic susceptibility variants with large, moderate and small effects have been reported [10-13]. However, the variants reported thus far have not been TNBC subtype-specific, and the causal association between them and the subtypes of TNBC remains poorly understood. This investigation was designed with a dedicated focus to determine whether genes containing genetic variants associated with an increased risk of developing breast cancer are associated with the LAR and MES subtypes of TNBC and to identify gene regulatory networks and signalling pathways enriched for genetic variants. We used a comprehensive catalogue of genetic variants and genes associated with an increased risk of developing breast cancer we have developed and published [5,12]. Briefly, the catalogue was developed by manually curating and annotating single nucleotide polymorphisms (SNPs, herein referred to as genetic susceptibility variants) and the genes they map to, from GWAS [5,12]. The catalogue was supplemented with information from the international GWAS catalogue [13]. This curation generated a total of 230 genes containing over 600 genetic variants used in this investigation. Methods of GWAS data collection, curation and annotation have been published elsewhere [5] and followed the international protocol for GWAS [14-18]. Because primary GWAS information was not breast cancer type-specific, we considered all the genetic variants and genes reported to be associated with an increased risk of developing breast cancer.

### **Source of gene expression data**

We used publicly available gene expression data on Caucasian women from the Gene Expression Omnibus (GEO) accession number GSE76124 consisting of 84 Caucasian women diagnosed with LAR (N=37) and MES (N=47) subtypes of TNBC generated at Baylor University [19]. The experimental procedures have been fully described by the data originators [19]. The two subtypes of TNBC represented and met the criteria of the current consensus on TNBC subtype classification [8,9]. As noted earlier in this report, we have previously reported the association of GWAS information

with the other two subtypes of TNBC, BLIA and BLIS [12]. For controls, we used publicly available gene expression data on 100 cancer free breast tissue from Caucasian women generated at Moffitt Comprehensive Cancer Center [20]. The control data set was downloaded from the Gene Expression Omnibus (GEO) database accession number GSE10780 [20]. The experimental procedures and methods of sample processing have been fully described by the data originators [20]. Clinical-pathological data from the patients used in the study included the tumour ER-, PR- and HER-2- status and tumour grade. Both data sets were generated using the Affymetrix platform using the Human GeneChip U133 Plus 2.0, which contains (54,675 probe sets). Gene expression values were calculated using the robust multi-array average (RMA) algorithm as implemented in the Affymetrix platform. All the expression values were on a log scale (log2).

### Data analysis

We performed whole transcriptome analysis comparing gene expression levels between tumour and control samples for each subtype using the Limma package implemented in R [21]. This unbiased approach to analysis was designed to identify genes containing genetic variants as well as other genes associated with each subtype of TNBC under study. Due to the relatively small sample sizes in each subtype of TNBC, we did not partition the data into test and validation sets. Instead, we used the leave-one-out cross-validation procedure as our prediction and validation model to identify genes with predictive power [22]. We used the false discovery rate (FDR) procedure to correct for multiple hypothesis testing [23]. Genes were ranked based on p-values and the FDR, and highly significantly differentially expressed genes were selected for each comparison. Genes containing genetic variants associated with an increased risk of developing breast cancer were identified using gene names and corresponding gene symbols. From these analysis we created two gene lists for each subtype of TNBC, genes containing genetic variants (GWAS genes) and genes without genetic variants (non-GWAS genes). Additional analysis was performed comparing expression levels between the two subtypes of TNBC to identify a signature of genes distinguishing the two types of TNBC.

To determine whether the genes containing genetic susceptibility variants are functionally related and have similar patterns of expression profiles with one another and with non-GWAS genes, we performed two-stage hierarchical clustering separately for LAR and MES. First, we performed analysis for each subtype of TNBC using GWAS derived genes only. In the second step, we performed analysis combining GWAS derived and nonGWAS genes for each subtype. In both analysis strategies, we used the Pearson correlation coefficient as the measure of distance between pairs of genes and complete linkage as the clustering method. Prior to clustering, gene expression data was normalized

using the median normalization, standardized and centered [24]. Hierarchical clustering was performed using GenePattern [25].

To identify the molecular networks and signalling pathways enriched for genetic variants, we performed network and pathway analysis for each subtype of TNBC using the Ingenuity Pathway Analysis (IPA) software (<http://www.ingenuity.com>) [26]. For each subtype of TNBC, a set of GWAS and non-GWAS genes were combined and mapped onto networks and canonical pathways using the network and pathway prediction, build and design modules as implemented in IPA [26]. We computed the probability scores and the log P-values to assess the likelihood and reliability of correctly assigning the genes to the correct networks, functional category and signalling pathway. The molecular networks and biological pathways were ranked based on z-scores and log p-values; respectively. Gene ontology (GO) [27] analysis as implemented in IPA was performed to characterize putative functional relationships between genes and to identify the molecular functions, biological processes and cellular components in which the discovered genes are involved.

### Results

#### Associating genetic variants with the LAR subtype

To test the hypotheses that genes containing genetic variants associated with an increased risk of developing breast cancer are associated with the LAR subtype of TNBC, we compared expression levels of the 230 genes containing germline mutations between tumor and control samples. The investigation revealed a signature of 198 significantly differentially expressed genes containing genetic susceptibility variants associated with the LAR, confirming our hypothesis. The signature included genes containing genetic variants reported to be directly associated TNBC [4,12,28,29]. A list of genes containing genetic variants associated with an increased risk of developing TNBC significantly associated with the LAR subtype are presented in Table 1. Among the genes containing genetic variants directly associated with TNBC transcriptionally associated with LAR included the genes BRCA1, BRCA2, TP53, PTEN, STK11, CDH1 with high-penetrance mutations, the genes ATM, BRIP1, CHEK2, PLB2, BARD1, NBN, RD50 with moderate penetrance mutations and the genes MAP3K1, FGFR2, LSP1, TNRC19, H19 with low penetrance mutations [30]. Interestingly, genes containing genetic variants not reported to be directly associated with TNBC were associated with the LAR subtype. A complete list of all the 198 genes containing genetic risk variants (GWAS genes) that were transcriptionally associated with the LAR subtype in this investigation along with their estimates of p-values and False Discovery Rate (FDR) are presented in genes Supplementary Table S1.

One of the challenges in clinical implementation of GWAS information is that genetic variants and genes identified thus far

explain only a small proportion of the phenotypic variation, many of the variants identified to date, may not be causal. To address, this knowledge gap, we investigated the association of non-GWAS genes with the LAR subtype of TNBC. The investigation revealed a signature of 118 highly significantly ( $P < 10^{-6}$ ; FDR  $< 0.01\%$ ) differentially expressed non-GWAS genes associated with the LAR subtype, confirming our hypothesis. A complete list of all the 118 non-GWAS genes highly significantly associated with the LAR subtype is presented in Supplementary Table S1 along with their estimates of P-values and FDR showing the strength and reliability of association.

**Table 1:** List of genes containing genetic variants associated with an increased risk of developing TNBC significantly associated with the LAR subtype in this study.

Gene symbol	Chr. position	GWAS		Gene expression (LAR)
		SNP (rs-ID)	P-value	
BRCA2	13q13.1	rs11531833	$5.00 \times 10^{-8}$	$5.00 \times 10^{-6}$
ESR1	6q25.1	rs11155804	$8.18 \times 10^{-18}$	$0.04 \times 10^{-3}$
LGR6	1q32.1	rs17489300	$9.37 \times 10^{-10}$	$9.0 \times 10^{-1}$
ZNF365	10q21.2	rs10995190	$3.75 \times 10^{-8}$	$5.00 \times 10^{-6}$
TERT	5p15.3	rs2242652	$2.73 \times 10^{-14}$	$5.00 \times 10^{-6}$
TOX3	16q12.1	rs4784227	$1.11 \times 10^{-14}$	$5.00 \times 10^{-6}$
PEX14	1p36.2	rs10864459	$2.13 \times 10^{-9}$	$5.00 \times 10^{-6}$
ADAM29	4q34.1	rs6828523	$1.33 \times 10^{-3}$	$2.50 \times 10^{-5}$
EBF1	19q13.1	rs1432679	$8.62 \times 10^{-3}$	$0.58 \times 10^{-4}$
TCF7L2	10q25.2	rs6585202	$3.35 \times 10^{-5}$	$0.70 \times 10^{-2}$
NTN4	12q22	rs17356907	$7.55 \times 10^{-3}$	$1.60 \times 10^{-3}$
MLK1	14q24.2	rs6001930	$2.5 \times 10^{-3}$	$5.00 \times 10^{-6}$
ANKLE1	19p13.1	rs8170	$2.25 \times 10^{-8}$	$5.00 \times 10^{-6}$
TGFB1	19q13.1	rs1982073	$3.8 \times 10^{-3}$	$5.00 \times 10^{-6}$
MAP3K1	5q11.2	rs889312	$4.60 \times 10^{-20}$	$1.57 \times 10^{-2}$
LSP1	11q15.5	rs4980383	$3.02 \times 10^{-6}$	$5.00 \times 10^{-6}$
TGFBR2	3p24.1	rs1367610	$3.08 \times 10^{-4}$	$5.00 \times 10^{-6}$
FTO	16q12.2	rs11075995	$3.30 \times 10^{-8}$	$6.40 \times 10^{-3}$
ADH1B	19p13.1	rs56069439	$8.25 \times 10^{-19}$	$5.00 \times 10^{-6}$
BABAM1	19p13.11	rs8170	$1.7 \times 10^{-10}$	$5.00 \times 10^{-6}$
CHEK2	22q12.1	rs17879961	$1.00 \times 10^{-8}$	$5.00 \times 10^{-6}$
MYC	8q24.21	rs11780156	$3.00 \times 10^{-11}$	$4.91 \times 10^{-6}$

#### Associating genetic variants with the MES subtype

To test the hypothesis that genes containing genetic variants associated with an increased risk of developing breast cancer are associated with the MES subtype of TNBC, we compared gene expression levels of the same 230 genes between tumor and control samples. The analysis produced a signature of 204 genes associated with the MES subtype. Likewise, among the identified genes included genes containing genetic variants reported to be directly with associated TNBC [28,29]. Table 2 shows a list of genes containing genetic variants associated directly associated with an increased risk of developing TNBC, that were significantly associated with

the MES subtype. The list included genes BRCA1, BRCA2, TP53, PTEN, STK11 and CDH1 with high-penetrance mutations, the genes ATM, BRIP1, CHEK2, PLB2, BARD1, NBN and RD50 with moderate penetrance mutations and the MAP3K1, FGFR2, LSP1, TNRC19 and H19 with low penetrance mutations [30]. The investigation also revealed genes containing genetic variants not reported to be directly associated with TNBC, which were transcriptionally associated with the MES subtype. A complete list of all the 204 genes containing genetic risk variants (GWAS genes) that were transcriptionally associated with the MES subtype in this investigation along with their estimates of p-values and false discovery rate (FDR) are presented in genes Supplementary Table S2.

Evaluation of non-GWAS genes, revealed a signature of 119 highly significantly ( $P < 10^{-6}$ ; FDR  $< 0.01\%$ ) non-GWAS genes associated with the MES subtype, confirming our hypothesis. A complete list of all the 119 non-GWAS genes that were highly significantly associated with the MES subtype is presented in Supplementary Table S2 along with their estimates of P-values and FDR showing the strength and reliability of association. Comparison of genes expression levels between the LAR and MES revealed significant overlap. That is majority of GWAS and non-GWAS genes found to be significantly associated with LAR were also significantly associated with MES subtype.

**Table 2:** List of genes containing genetic variants associated with an increased risk of developing TNBC significantly associated with the MES subtype in this study.

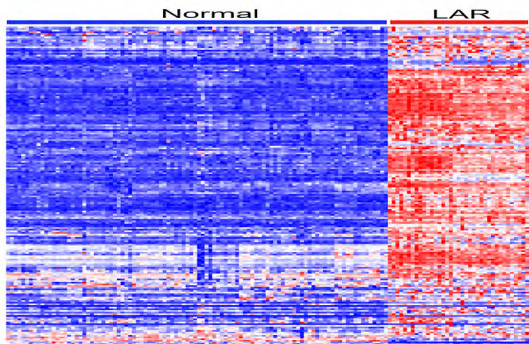
Gene Symbol	Chr. position	GWAS		GeneExpression (LAR)
		SNP (rs-ID)	P-value	
BRCA2	13q13.1	rs11531833	$5.00 \times 10^{-8}$	$5.00 \times 10^{-6}$
ESR1	6q25.1	rs11155804	$8.18 \times 10^{-18}$	$1.80 \times 10^{-3}$
LGR6	1q32.1	rs17489300	$9.37 \times 10^{-10}$	$0.28 \times 10^{-2}$
ZNF365	10q21.2	rs10995190	$3.75 \times 10^{-8}$	$5.00 \times 10^{-6}$
TERT	5p15.3	rs2242652	$2.73 \times 10^{-14}$	$5.00 \times 10^{-6}$
TOX3	16q12.1	rs4784227	$1.11 \times 10^{-14}$	$1.46 \times 10^{-4}$
PEX14	1p36.2	rs10864459	$2.13 \times 10^{-9}$	$5.00 \times 10^{-6}$
ADAM29	4q34.1	rs6828523	$1.33 \times 10^{-3}$	$5.00 \times 10^{-6}$
EBF1	19q13.1	rs1432679	$8.62 \times 10^{-3}$	$5.00 \times 10^{-6}$
TCF7L2	10q25.2	rs6585202	$3.35 \times 10^{-5}$	$5.00 \times 10^{-6}$
NTN4	12q22	rs17356907	$7.55 \times 10^{-3}$	$0.26 \times 10^{-3}$
MLK1	14q24.2	rs6001930	$2.5 \times 10^{-3}$	$5.00 \times 10^{-6}$
ANKLE1	19p13.1	rs8170	$2.25 \times 10^{-8}$	$5.00 \times 10^{-6}$
TGFB1	19q13.1	rs1982073	$3.8 \times 10^{-3}$	$5.00 \times 10^{-6}$
MAP3K1	5q11.2	rs889312	$4.60 \times 10^{-20}$	$1.61 \times 10^{-3}$
LSP1	11q15.5	rs4980383	$3.02 \times 10^{-6}$	$5.00 \times 10^{-6}$
TGFBR2	3p24.1	rs1367610	$3.08 \times 10^{-4}$	$5.00 \times 10^{-6}$
FTO	16q12.2	rs11075995	$3.30 \times 10^{-8}$	$0.23 \times 10^{-3}$
ADH1B	19p13.1	rs56069439	$8.25 \times 10^{-19}$	$4.16 \times 10^{-3}$
BABAM1	19p13.11	rs8170	$1.7 \times 10^{-10}$	$5.00 \times 10^{-6}$



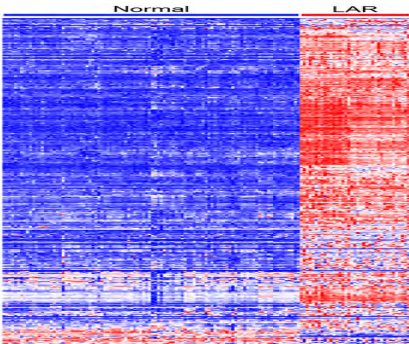
CHEK2	22q12.1	rs17879961	1.00x10 <sup>-8</sup>		5.00x10 <sup>-6</sup>
MYC	8q24.21	rs11780156	3.00x10 <sup>-11</sup>		5.00x10 <sup>-6</sup>

Patterns of gene expression profiles for the LAR and MES

To address the hypothesis that genes containing genetic variants are co-regulated and have similar patterns of gene expression profiles with one another and with non-GWAS genes, we performed two-step hierarchical clustering for LAR and MES separately, as explained in the data analysis section. Figure 1 shows patterns of gene expression profiles for the 198 genes containing genetic variants that were significantly associated with LAR subtype of TNBC. Results showing patterns of expression profiles for the combined set of 198 GWAS and 118 non-GWAS genes are presented in Figure 2. Hierarchical clustering revealed that genes containing genetic variants are co-regulated and have similar patterns of gene expression profiles (Figure 1), confirming our hypothesis. In addition, hierarchical clustering combining GWAS and non-GWAS revealed that genes containing genetic variants are co-regulated and have similar patterns of gene expression profiles with non-GWAS genes (Figure 2), confirming our hypothesis. Importantly, genes containing genetic variants with strong GWAS associations were co-regulated and had similar patterns of expression profiles with genes containing genetic susceptibility variants with weak to moderate associations (Figure 1). In addition, genes containing genetic variants with weak to strong GWAS associations were co-regulated and had similar patterns of expression with non-GWAS genes. Moreover, the analysis revealed that genes containing genetic variants reported to be directly associated with TNBC, were functionally related with other GWAS genes and non-GWAS genes.

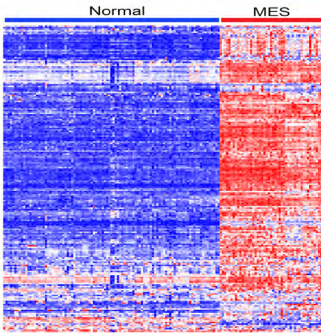


**Figure 1:** Patterns of expression profiles for the 198 GWAS genes containing genetic susceptibility variants significantly associated with the LAR subtype of TNBC. The rows represent genes and columns represent samples. The red and blue colors indicate upregulation and down regulation; respectively.

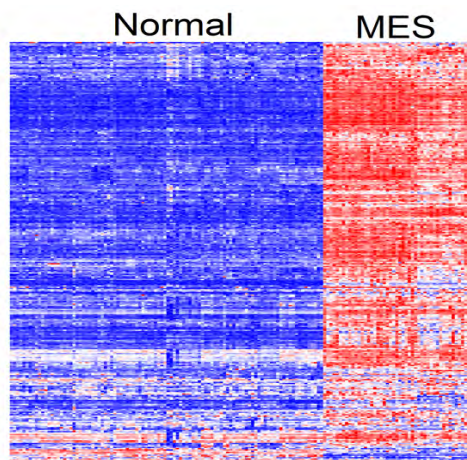


**Figure 2:** Patterns of expression profiles for the 198 GWAS genes containing genetic susceptibility variants and the 118 non-GWAS genes associated with the LAR subtype of TNBC. Columns represents samples and rows represent genes. The red and blue colors indicate upregulation and down regulation respectively.

The results showing patterns of gene expression profiles for the 204 GWAS genes only in tumor and control samples for MES are presented in Figure 3. Figure 4 shows the patterns of gene expression profiles for combined 204 GWAS and 119 non-GWAS genes. Genes containing genetic variants were co-regulated and had similar patterns of expression profiles (Figure 3) regardless of the level of association. Likewise, GWAS and non-GWAS genes were co-regulated and had similar patterns of gene expression confirming our hypothesis [Figure 4]. Interestingly, genes containing genetic variants reported to be directly associated with TNBC were functionally related and had similar patterns of expression profiles with genes containing genetic variants not reported to be directly associated with TNBC (Figure 3). In addition, genes containing genetic variants directly associated with TNBC were co-regulated with non-GWAS (Figure 4).



**Figure 3:** Patterns of gene expression profiles for the 204 GWAS genes transcriptionally associated with the MES subtype of TNBC. Columns represents patients and rows represent genes. The red and blue colors indicate upregulation and down regulation; respectively.



**Figure 4:** Patterns of gene expression profiles for the 204 GWAS genes 119 non-GWAS genes associated with the MES subtype of TNBC. Columns represents patients and rows represent genes. The red and blue colors indicate upregulation and down regulation respectively.

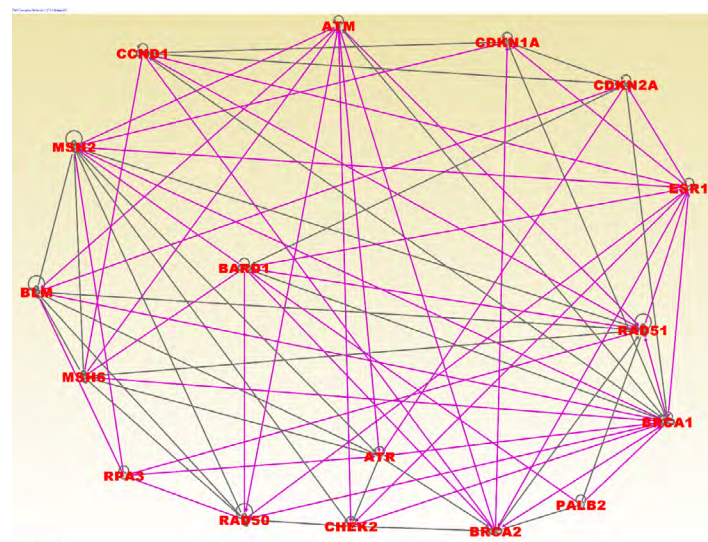
Overall, the investigations in both LAR and MES revealed that genes containing genetic variants are co-regulated and have similar patterns of expression among themselves and with non-GWAS genes. Taken together, the presence of genetic variants in co-regulated genes with similar biological functions could give a degree of confidence that the associations are potentially genuine, even if none of the genetic individually are highly significant. Co-expression analysis provides a framework for the discovery of co-regulated genes.

#### Molecular networks and signaling pathways enriched for genetic variants

To gain insights about the broader biological context in which genetic variants operate and to establish putative functional bridges between genetic variants and the pathways they control in LAR and MES, we performed network and pathways analysis. For these analysis, we combined GWAS and non-GWAS genes and were performed separately for each subtype of TNBC. Our working hypothesis was that GWAS and non-GWAS genes are functionally related and interact in gene regulatory networks and signaling pathways enriched for genetic variants.

The results of network analysis for LAR are presented in Figure 5. The analysis produced networks containing genes with overlapping functions. The networks contained genes predicted to be involved in DNA replication, recombination and repair, cell cycle and cancer (z-score = 49); cell cycle and cell death and

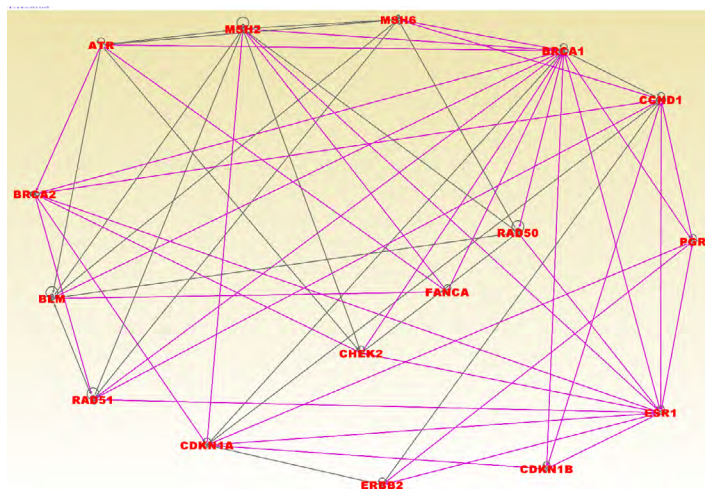
survival (z-score =28); DNA replication, recombination and repair, gene expression and cellular development (z-score = 26); and cell morphology and inflammatory response (z-score = 19) (Figure 5). The investigations revealed signaling pathways enriched for genetic variants including, the role of BRCA1 in DNA damage response ( $P=1.10 \times 10^{-25}$ ), Hereditary breast cancer ( $P=1.26 \times 10^{-22}$ ), Aryl hydrocarbon receptor ( $P=1.48 \times 10^{-14}$ ) and the molecular mechanisms of cancer ( $P=2.44 \times 10^{-10}$ ) signaling pathways, all of which have been implicated in TNBC [5,12].



**Figure 5:** Molecular networks for genes containing genetic variants that were associated with LAR subtype. The gene symbols in red font represent genes containing SNPs associated with an increased risk of developing cancer. The solid lines indicate functional relationships.

The results of network analysis for the MES subtype are presented in Figure 6. The analysis produced molecular networks containing genes predicted to be involved in cellular development, cellular growth and proliferation, organ development (z=32), cell cycle and cancer (z=32), DNA replication, recombination and repair, cell cycle survival, cellular compromise and cellular assembly and morphology (z=30). Majority of the genes were predicted to be significantly involved in DNA replication and repair, cell cycle, cell death and survival, cellular compromise and cellular assembly and organization for the MES subtype. Pathways enrichment analysis reveals role of BRCA1 in DNA damage response ( $P=2.05 \times 10^{-25}$ ), Hereditary breast cancer signaling ( $P=2.52 \times 10^{-22}$ ), aryl hydrocarbon receptor signaling ( $P=1.49 \times 10^{-16}$ ) and the molecular mechanisms of cancer ( $P=2.43 \times 10^{-12}$ ) signaling pathways enriched for genetic variants.





**Figure 6:** Molecular networks for genes containing genetic variants that were associated with mesenchymal (MES) subtype. The gene symbols in red font represent genes containing SNPs associated with an increased risk of developing cancer. The solid lines indicate functional relationships.

In both the LAR and MES subtypes, network and pathway analysis revealed that genes with low to moderate GWAS associations are functionally related and interact with genes containing genetic variants with strong GWAS associations. There was considerable overlap in the functions of genes, molecular networks and pathways associated with either type of TNBC (that is molecular networks and pathways discovered in LAR were also discovered in MES), although their ranking differed between the two subtypes of TNBC. Overall, the investigation revealed that in the context of TNBC, the LAR and MES subtypes can be considered as emergent properties of gene regulatory networks and signaling pathways controlled by many genetic variants and genes, rather than individual genetic variants or a small number of genes. Thus, integrating GWAS information using gene expression data from LAR and MES as the intermediate phenotype holds promise for establishing the causal association between genetic susceptibility and the two subtypes of TNBC.

## Discussion

GWAS have revealed genetic variants associated with an increased risk of developing breast cancer. However, majority of the genetic variants have not been cancer type and subtype-specific, rendering their clinical implementation in heterogeneous disease entities like TNBC a challenge. Here we integrated GWAS information with gene expression data from the LAR and MES subtypes of TNBC. The goal was to infer the potential causal association between genetic susceptibility and the two subtypes of

TNBC; and to identify molecular networks and signaling pathways to gain insights about the broader biological context in which genetic variants and associated genes operate. The investigation revealed that genes containing genetic variants are associated with the LAR and MES subtypes of TNBC. In addition, the investigation revealed molecular networks and signaling pathways enriched for genetic variants. These findings establish putative functional bridges between GWAS discoveries and the signaling pathways they control. More, importantly, this demonstrates that integration of GWAS information using gene expression data as the intermediate phenotype provides a framework for addressing knowledge gaps not addressed by GWAS. To our knowledge this is the first study to infer the potential causal association between GWAS and the two subtypes of TNBC. The results of this investigation are consistent with our earlier investigation in which we associated GWAS information with the BLIA and BLIS subtypes of TNBC [12]. While we did not investigate individual genetic variants, the aggregation of genetic variants through co-expression, functional, network and pathways analysis provides convincing evidence that some of the genetic variants may play a role in the pathogenesis of LAR and MES subtypes of TNBC. The practical significance of these findings is that understanding the biological context in which genetic variants operate is a necessary step towards clinical implementation and identifying potential drug targets [31,32].

An important limitation in GWAS studies is that SNP–trait associations reported thus far, do not necessarily lead directly to the identification of the causal gene(s), or much less elucidating the context in which the genetic variants operate [33,34]. However, combining GWAS with non-GWAS genes through co-expression, functional, network and pathways analysis provides a framework for uncovering complex oncogenic interactions likely to drive and shape clinical phenotypes. The discovery of gene regulatory networks and signaling pathways enriched for genetic variants captures both *cis* and *trans* regulatory mechanisms in which the genetic variants may be involved. The approach demonstrates that the missing variation in GWAS and potential causal genes may be inferred by layering in gene expression as the intermediate phenotype for LAR and MES subtypes of TNBC [35,36].

Recently, development of risk prediction algorithms such as polygenic risk scores has come into sharper focus in breast cancer research [6,7]. Polygenic risk scores are poised to improve outcomes via precision medicine and potentially precision prevention [6,7]. However, polygenic risk scores available to date are not accurate enough to support patient stratification by subtype. One way to address this knowledge gap and critical unmet need may be by leveraging GWAS information and integrating it with gene expression data to refine current polygenic risk scores.



Knowing that a specific risk profile is associated with a subtype-specific TNBC may lead to subtype-specific tailored genetic screening.

The discovery of important signalling pathways associated with the two subtypes types of TNBC including the role of BRCA1 in DNA damage response, Hereditary breast cancer, Aryl hydrocarbon receptor and the molecular mechanisms of cancer signalling pathways enriched for genetic susceptibility variants was of particular interest. Majority of TNBC tumors harbour BRCA1 and BRCA2 mutations [37]. Germline mutations in BRCA1 and BRCA2 genes predispose individuals to TNBC by impairing homologous recombination (HR) system, thereby causing genomic instability [38]. In clinical practice, homologous recombination deficiency serves as both a predictive and prognostic factor in different settings of TNBC patients treated with DNA damaging drugs and poly ADP ribose polymerase (PARP) inhibitors [38]. This renders the role of BRCA1 in DNA damage response signalling pathway a potential therapeutic target. In addition, the high prevalence of pathogenic mutations in BRCA1 and BRCA2 in sporadic TNBC renders use of these genes prime for genetic testing [39,40]. The discovery of the Aryl hydrocarbon receptor signalling pathway is consistent with literature reports [41]. This signalling pathway mediates DNA damage in breast cancer cells [41], which renders it a potential therapeutic target by itself or through the reactive oxygen species (ROS) whose expression levels correlate with the activity of this pathway [41]. Likewise, the AR signalling pathway could serve as a therapeutic target [42]. The discovery of the hereditary breast cancer signalling pathway has clinical application potential, because inherited germline mutations considered in this study may interact with somatic mutations to drive tumorigenesis in TNBC [43,44]. Although currently there is limited evidence that cancer susceptibility regions are preferential targets for somatic mutations [45], there is compelling evidence that hereditary breast cancer is due to mutations in BRCA1 and BRCA2 genes reported in this study and previous studies [46].

Overall, our investigation revealed that genes containing genetic susceptibility variants are associated with the LAR and MES subtypes of TNBC. The clinical significance of these findings is that genetic testing for TNBC using panels of genes such as BRCA1 and BRCA2 evaluated here are being used routinely in a diagnostic setting [47,48]. In addition, germline mutations considered here are being used in the development of polygenic risk scores to identify individuals at high risk of developing TNBC that could be prioritized for treatment [49,50]. Thus, information on genetic variants and genes when combined with gene expression data has the potential to improve outcomes in TNBC via precision medicine and precision prevention [51]. An limitation of our study and others [52] is the lack of ethnic diversity in genomic studies,

which if not addressed has the potential to exacerbate racial disparities in TNBC.

## Conclusions

The study revealed signatures of genes containing genetic variants are associated with the LAR and MES subtypes of TNBC. In addition, the study revealed molecular networks and signaling pathways enriched for genetic variants. Further research involving multiple racial/ethnic populations to validate genetic variants in LAR and MES subtypes of TNBC is recommended and is the potential subject of our future research.

## Data Availability and sharing

GWAS information is available at the GWAS catalogue managed by the European Bioinformatics Institute: <https://www.ebi.ac.uk/gwas/>. Additional GWAS data is available in our catalogue published earlier in reference # 5.

Original gene expression data is available at the Gene Expression Omnibus (GEO) <https://www.ncbi.nlm.nih.gov/geo/> under accession number GSE76124 for TNBC samples and GSE10780 for control samples.

**Conflicts of Interest:** The author(s) have no conflict of interest to declare. All authors have read and approved this manuscript.

**Acknowledgments:** The authors wish to thank the school of medicine for providing startup funds in support of this research and the patients who volunteered and provide the tumor samples used to generate both GWAS and gene expression data in GEO database used in this study.

## References

1. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. *CA Cancer J Clin* 69: 7-34.
2. Dietze EC, Sistrunk C, Miranda-Carboni G, O'Regan R, Seewaldt V (2015) Triple-negative breast cancer in African American women: disparities versus biology. *Nature Reviews* 15: 240-254.
3. Perou CM (2010) Molecular stratification of triple-negative breast cancers. *Oncologist* 15: 39-48.
4. Stevens KN, Vachon CM, Couch FJ (2013) Genetic susceptibility to triple-negative breast cancer.
5. Hicks C, Kumar R, Pannuti A, Backus K, et al. (2013) An Integrative Genomics Approach for Associating GWAS Information with Triple-Negative Breast Cancer. *Cancer Inform* 12: 1-20.
6. Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, et al. (2019) Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* 104: 21-34.
7. Nasim Mavaddat, Paul D P Pharoah, Kyriaki Michailidou, et al. (2015) Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst.* 107: djv036.

8. Lehmann BD and Pietenpol JA (2014) Identification and use of biomarkers in treatment strategies for triple-negative breast cancer subtypes. *J Pathol* 232: 142-50.
9. Lehmann BD, Jovanović B, Chen X, et al. (2016) Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* 11: e0157368.
10. Jiande Wu, Tarun Karthik Kumar Mamidi, Lu Zhang, Chindo Hicks (2019) Deconvolution of the Genomic and Epigenomic Interaction Landscape of Triple-Negative Breast Cancer. *Cancers (Basel)* 11: 1692
11. Jiande Wu, Tarun Karthik Kumar Mamidi, Lu Zhang, Chindo Hicks (2019) Integrating Germline and Somatic Mutation Information for the Discovery of Biomarkers in Triple-Negative Breast Cancer. *Int J Environ Res Public Health* 16: 1055.
12. Jacob Elnaggar, Fern Tsien, Clayton Yates, Melisa Davis, Lucio Miele et al. (2019) An Integrative Genomics Approach for Associating Genetic Susceptibility with the Tumor Immune Microenvironment in Triple Negative Breast Cancer.
13. The NHGRI-EBI Catalog of published genome-wide association studies.
14. Ioannidis JP, Boffetta P, Little J, et al. (2008) Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol* 37: 120–132.
15. Khoury MJ, Bertram I, Boffetta P, et al. (2009) Genome-wide association studies, field synopsis, and the development of the knowledge base on genetic variation and human diseases. *Am J Epidemiol* 170: 269–279.
16. Sagoo GS, Little J, Higgins JP (2009) Systematic reviews of genetic association studies. *Human Genome Epidemiology Network. PLoS Med* 6: e28.
17. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 151: 264–269.
18. Liberati A, Altman DG, Tetzlaff J, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 6: e1000100.
19. Burstein MD, Tsimelzon A, Poage GM, Covington KR, et al. (2015) Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 21: 1688-98.
20. Chen DT, Nasir A, Culhane A, Venkataramu C, et al. (2010) Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res Treat* 119: 335-46.
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43: e47.
22. Radmacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. *J. Comput. Biol* 9: 505–511.
23. Benjamini Y, Yosef H (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Society* 57: 289–300.
24. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863–14868.
25. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P (2006) GenePattern 2.0. *Nat Genet* 38: 500–501.
26. Ingenuity Pathways Analysis (IPA) System. Redwood, CA: Ingenuity Systems.
27. Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.
28. Kristen S Purrington, Susan Slager, Diana Eccles et al. (2014) Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* 35: 1012-9.
29. Maria-Ancuta Jurj, Mihail Buse, Alina-Andreea Zimta et al. (2020) Critical Analysis of Genome-Wide Association Studies: Triple Negative Breast Cancer Quae Exempli Causa. *Int J Mol Sci* 21: 5835
30. Darrell L Ellsworth, Clesson E Turner, Rachel E Ellsworth (2019) A Review of the Hereditary Component of Triple Negative Breast Cancer: High- and Moderate-Penetrance Breast Cancer Genes, Low-Penetrance Loci, and the Role of Nontraditional Genetic Elements. *J Oncol*: 4382606.
31. Lum PY, Derry JM, Schadt EE (2009) Integrative genomics and drug development. *Pharmacogenomics* 10: 203-212.
32. Schadt EE, Friend SH, Shaywitz DA (2009) A network view of disease and compound screening. *Nature Rev. Drug Discov* 8: 286–295.
33. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322: 881–888.
34. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. *Nature* 452: 423–428.
35. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet* 37: 710–717.
36. Emmanuelle Génin (2020) Missing heritability of complex diseases: case solved? *Hum Genet* 139: 103-113.
37. Nadine Tung, Judy E Garber, Michele R Hacker, et al. (2016) Prevalence and predictors of androgen receptor and programmed death-ligand 1 in BRCA1-associated and sporadic triple-negative breast cancer. *NPJ Breast Cancer* 2: 16002.
38. Belli C, Duso BA, Ferraro E, Curigliano G (2019) Homologous recombination deficiency in triple negative breast cancer. *Breast* 45: 15-21.
39. De Leeneer K, Coene I, Crombez B, et al. (2012) Prevalence of BRCA1/2 mutations in sporadic breast/ovarian cancer patients and identification of a novel de novo BRCA1 mutation in a patient diagnosed with late onset breast and ovarian cancer: implications for genetic testing. *Breast Cancer Res Treat* 132: 87-95.
40. Engel C, Rhiem K, Hahnen E, Loibl S et al. (2018) Prevalence of pathogenic BRCA1/2 germline mutations among 802 women with unilateral triple-negative breast cancer without family cancer history. *BMC Cancer* 18: 265.
41. Lancelot S McLean, Cheri N Watkins, Petreena Campbell, et al. (2015) Aryl Hydrocarbon Receptor Ligand 5F 203 Induces Oxidative Stress That Triggers DNA Damage in Human Breast Cancer Cells. *Chem Res Toxicol* 28: 855-71.

42. Fangyuan Shao, Heng Sun, Chu-Xia Deng (2017) Potential therapeutic targets of triple-negative breast cancer based on its intrinsic subtype. *Oncotarget*. 8: 73329-73344.
43. Carter H, Marty R, Hofree M, Gross AM, et al. (2017) Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov* 7: 410-423.
44. Zhu B, Mukherjee A, Machiela MJ et al. (2016) An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer. *Br J Cancer* 115: 752-60.
45. Machiela MJ, Ho BM, Fisher VA, Hua X, Chanock SJ (2015) Limited evidence that cancer susceptibility regions are preferential targets for somatic mutation. *Genome Biol* 16: 193.
46. Petrucelli N, Daly MB, and Gerald L. Feldman GL (2010) Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genetics Medicine* 12: 245-259.
47. Hauke J, Horvath J, Groß E, Gehrig A, Honisch E, et al. (2018) Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German Consortium for Hereditary Breast and Ovarian Cancer. *Cancer Med* 7: 1349-1358.
48. Shimelis H, LaDuca H, Hu C, Hart SN, Na J, et al. (2018) Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst* 110: 855-862.
49. Mavaddat N, Michailidou K, Dennis J, et al. (2019) Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* 104: 21-34.
50. Hahnen E, Lederer B, Hauke J, Loibl S, et al. (2017) Germline Mutation Status, Pathological Complete Response, and Disease-Free Survival in Triple-Negative Breast Cancer: Secondary Analysis of the GeparSixto Randomized Clinical Trial. *JAMA Oncol* 3: 1378-1385.
51. Njiaju UO, Olopade O (2012) Genetic determinants of breast cancer risk: a review of current literature and issues pertaining to clinical application. *Breast J* 18: 436-42.
52. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL (2018) Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Aff (Millwood)* 37: 780-785.