Identification of Breast Cancer Subtype-Specific Biomarkers by Integrating Copy Number Alterations and Gene Expression Profiles

Claudia Cava [1,*], Mirko Pisati [1,2], Marco Frasca [2] and Isabella Castiglioni [3]

[1] Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Via F. Cervi 93, Segrate-Milan, 20090 Milan, Italy; mirko.pisati@studenti.unimi.it

[2] Department of Computer Science, Università degli Studi di Milano, Via Celoria 18, 20133 Milano, Italy; marco.frasca@unimi.it

[3] Department of Physics "Giuseppe Occhialini", University of Milan-Bicocca Piazza dell'Ateneo Nuovo, 20126 Milan, Italy; isabella.castiglioni@unimib.it

[*] Correspondence: claudia.cava@ibfm.cnr.it

Abstract: *Background and Objectives*: Breast cancer is a heterogeneous disease categorized into four subtypes. Previous studies have shown that copy number alterations of several genes are implicated with the development and progression of many cancers. This study evaluates the effects of DNA copy number alterations on gene expression levels in different breast cancer subtypes. *Materials and Methods*: We performed a computational analysis integrating copy number alterations and gene expression profiles in 1024 breast cancer samples grouped into four molecular subtypes: luminal A, luminal B, HER2, and basal. *Results*: Our analyses identified several genes correlated in all subtypes such as *KIAA1967* and *MCPH1*. In addition, several subtype-specific genes that showed a significant correlation between copy number and gene expression profiles were detected: *SMARCB1*, *AZIN1*, *MTDH* in luminal A, *PPP2R5E*, *APEX1*, *GCN5* in luminal B, *TNFAIP1*, *PCYT2*, *DIABLO* in HER2, and *FAM175B*, *SENP5*, *SCAF1* in basal subtype. *Conclusions*: This study showed that computational analyses integrating copy number and gene expression can contribute to unveil the molecular mechanisms of cancer and identify new subtype-specific biomarkers.

Keywords: copy number alteration; gene expression; breast cancer subtypes

1. Introduction

Breast cancer (BC) is a heterogeneous disease categorized into four subtypes: luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), and basal-like (basal). The different BC subtypes can be distinguished based on the expression level of four significant biomarkers by immunohistochemistry: estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki-67. Indeed, luminal A is characterized by ER positive and/or PR positive and Ki-67 < 14%, lumB by ER positive and/or PR positive and Ki-67 $\geq$ 14%, Her2-enriched by ER negative and PR negative and Her2 positive, and triple negative by ER negative and PR negative and Her2 negative [1]. The molecular classification, such as PAM-50, which considers the expression levels of mRNAs, defines the triple negative subtype with the term basal [2]. However, previous studies reported that there is an overlap of 80% between basal BC defined by molecular classification and triple negative BC defined by immunohistochemistry [3]. Different BC subtypes have different clinical outcomes, such as patient survival, prognosis, and relapse [1].

A correct and early diagnosis of BC subtypes is essential to give effective treatments for patients. Huge volumes of biological data derived by high-throughput sequencing technologies are publicly available in databases, such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). With the availability of these data, many methods

that combined multi-omics data were developed to identify biological patterns and reveal new biological processes [4,5].

Genomic instability, including copy number alteration (CNA), characterizes many cancers, such as BC. Copy number alterations are changes in the DNA where the number of copies of a segment DNA can result in an amplification or deletion of a specific gene.

Recently, more and more studies showed that genomic aberrations are key events in the progression from normal to tumoral tissue [6,7]. In addition, to investigate the crucial role of genomic alterations and gene expression profiles in disease progression, some studies reported novel candidate genes by integrating gene expression and CNA profiles [8,9].

To date, different computational methods based on such integration detected driver genes [10–12]. One example is HIT'Ndrive: its aim was to identify patient-specific altered genes that when combined can change the expression levels of transcripts [13]. HIT'Ndrive was applied to 2200 tumors in a pan-cancer study and detected some important driver genes [13]. A probabilistic model is used by another approach to investigate how mutated genes can influence the expression of other genes [14]. Similarly, Suo et al. defined driver genes as those genes that showed a mutation and interacted with a high number of differentially expressed genes in a gene network [15]. However, only some of the current methods can discover driver genes, as many of them are optimized to identify de-regulated biological modules. In addition, none of them to our knowledge is applied to the BC molecular subtypes. Indeed, the novelty of our study is the integrative analysis of copy number alteration and gene expression profiles to identify genes commonly modulated in all BC subtypes and subtype-specific genes.

CNAs can promote tumor development via the changes of gene expression levels. However, the effects of CNAs on gene expression levels are difficult to study, as gene expression profiling is often performed on biopsies containing tumor and normal cells. Therefore, gene expression levels, which are the average expression of all cell types, are often overshadowed by the influences of non-tumor cells [16,17]. A solution is the use of genomic and gene expression profiles derived from the same patient. Previous studies reported that the integration of mutation and gene expression profiles from the same patient could explain the genetic heterogeneity of cancer [18,19]. As gene expression profiles can detect crucial genes whose expression dynamically changes, for example during the cell cycle, the integration of genomic data from the same patient can give more solid results and reduce the variability of data.

In this study, we analyzed a large dataset of copy number and gene expression profiles to identify BC subtype-specific driver genes. We performed a computational analysis correlating DNA copy number alterations and gene expression profiles in BC samples from the TCGA database. We identified the genes that showed a correlation in all BC subtypes, and BC subtypes-specific genes. Furthermore, we used two other GEO datasets to identify subtype-specific genes that were altered in at least 50% of the samples in TCGA and GEO datasets. Finally, we investigated the biological role of these subtype-specific genes through pathway enrichment analysis.

2. Materials and Methods

The present work introduces an analysis of copy number and gene expression profiles of paired BC samples derived by the TCGA database. We performed a correlation analysis between copy number alterations and gene expression levels for each gene, and we obtained a list of correlated genes for each BC subtype. Furthermore, we selected those genes from the list that are specific for each subtype in TCGA, namely that show a correlation in only one subtype. To validate the results and to obtain a robust signature we included in the study two independent datasets from GEO. Indeed, the final list of correlated genes contained those genes that showed an alteration in at least 50% of the samples in the two GEO datasets and TCGA.

Figure 1 shows the analysis performed.
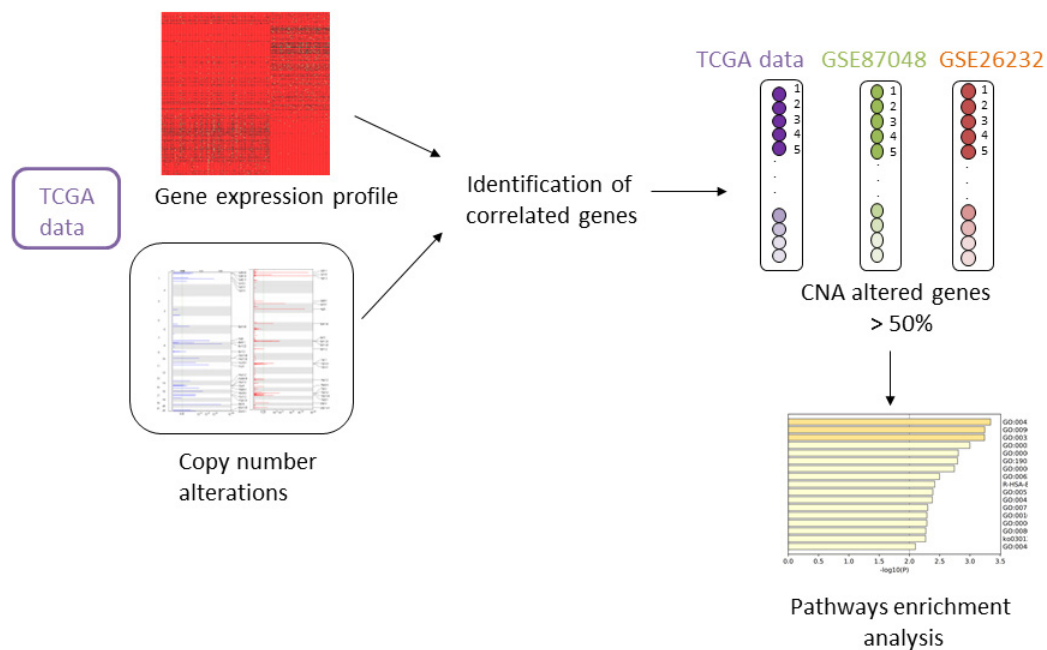
Figure 1. Workflow of the proposed analysis. TCGA: The Cancer Genome Atlas; CNA: Copy number alteration.

2.1. Data

We downloaded and pre-processed, using TCGABiolinks package [20], paired BC samples of copy number and gene expression profiles from TCGA. Copy number and gene expression data contained 548 lumA, 206 lumB, 82 HER2-positive, and 188 basal samples.

Copy number profiles of the BC samples from TCGA were downloaded using the getGistic function of the TCGABiolinks R package. The copy number matrix contained altered genes for each patient indicated with positive and negative numbers that showed amplification and deletion, respectively.

In the validation step, copy number profiles for two independent datasets from the GEO dataset (GSE87048 and GSE26232) were calculated using the package aroma.affymetrix with CRMAv2, the Circular Binary Segmentation model, and the GISTIC tool. Copy number profiles from TCGA and two GEO datasets were estimated for the whole genome. The data derived by bulk sequencing and CNAs were an average of copy numbers from thousands of cells for sample. The number of samples for each subtype is presented in Table 1. Molecular subtypes of BC samples from TCGA and GEO were established by molecular classification PAM50.

Table 1. Number of samples for each breast cancer subtype from The Cancer Genome Atlas (TCGA) and from two Gene Expression Omnibus (GEO) datasets.

Molecular Subtype	TCGA	GSE87048	GSE26232
LumA	548	37	12
LumB	206	20	6
Her2	82	12	2
Basal	188	8	17
Total	1024	77	37

2.2. Integrating Copy Number and Gene Expression

We performed a Pearson correlation test between copy number alterations and gene expression levels for each gene in the TCGA data. Pearson and Spearman correlations are widely used methods to assess the influence of copy number alterations on gene expression profiles in cancer [21,22]. However, we used Pearson's correlation because it assumes that the variables have a linear relationship and derive from a normal distribution [23]. *p*-values

were adjusted with the False Discovery Rate (FDR) method. We considered two genes correlated if the correlation coefficient was >0.6 or <−0.6 and *p*-values adjusted <0.05. The correlation analysis was performed separately for each subtype.

2.3. Identification of Subtype-Specific CNAs

We analyzed the genes that showed a correlation between copy number alterations and gene expression levels in TCGA data, and we selected the subtype-specific genes and those common among different subtypes. For the visualization of the data, we used a Venn diagram with the R package, pheatmap.

Furthermore, we evaluated the subtype-specific genes in TCGA, and we selected those genes that were altered through copy number analysis in at least 50% of the samples in TCGA data, GSE87048, and GSE26232. We used 50% as a cut-off because it is a commonly used threshold in previous studies [24,25]. Additionally, in this step we considered the samples grouped by subtype.

2.4. Survival Analysis

A Kaplan–Meier plotter was used to perform a survival analysis [26]. The tool uses gene expression data and relapse free survival information of datasets downloaded from public datasets, such as GEO and TCGA.

To examine the prognostic value of a gene, the samples were divided into two groups according to median expressions of the proposed biomarkers. The differences between survival curves were estimated using the log-rank test.

We performed the analysis considering all BC samples for the common genes among BC subtypes and BC samples separately for each subtype for subtype-specific genes.

2.5. Pathway Analysis

We performed a pathway enrichment analysis to obtain pathways enriched with subtype-specific genes. Specifically, we used Reactome pathway enrichment analysis to obtain the enriched pathways (*p*-value < 0.01). The Reactome tool is based on a statistical test (hypergeometric distribution) that revealed Reactome pathways that are significantly enriched with subtype-specific genes [27].

3. Results

3.1. Correlation Analysis in TCGA Data

Pearson correlation tests between copy number alterations and gene expression levels for each gene were performed. We found that 439 genes showed statistically significant correlations between DNA copy number and gene expression for the lumA, 970 for the lumB, 1189 for HER2, and 1029 for the basal subtype (correlation coefficient ≥0.6 or ≤−0.6, *p*-values adjusted <0.05). Figure 2 shows the heatmaps of the results of correlations.

In the next steps, we focused on how the genes that showed a correlation between copy number alterations and gene expression levels in TCGA data were distributed among the various BC subtypes. Figure 3 shows with a Venn diagram the common genes among subtypes and the subtype-specific genes.

### 3.1.1. Correlation Analysis: Common Genes among Breast Cancer Subtypes

In this step we presented the genes commonly correlated between copy number and gene expression profiles in all BC molecular subtypes.

We focused on two genes, *KIAA1967* and *MCPH1*, because, as reported in Figure 2, they obtained a high correlation coefficient in all BC subtypes. *KIAA1967*, a tumor suppressor gene, has a high correlation coefficient in all BC subtypes (0.8647 in LumA, 0.849 in LumB, 0.839 in Her2, and 0.778 in basal). Figure 3 shows the distribution of the copy number alterations of *KIAA1967* in the 4 BC subtypes. We noticed that the gene was mostly deleted. Indeed, it was deleted in 50% of LumA samples, and in 35% of the samples, there were no alterations. Only a small fraction of samples (15%) showed a gene amplification.

As in lumA, also in other BC subtypes, *KIAA1967* was mostly deleted: 59% of the lumB samples, 66% of the HER2 samples, and 60% of the basal samples.
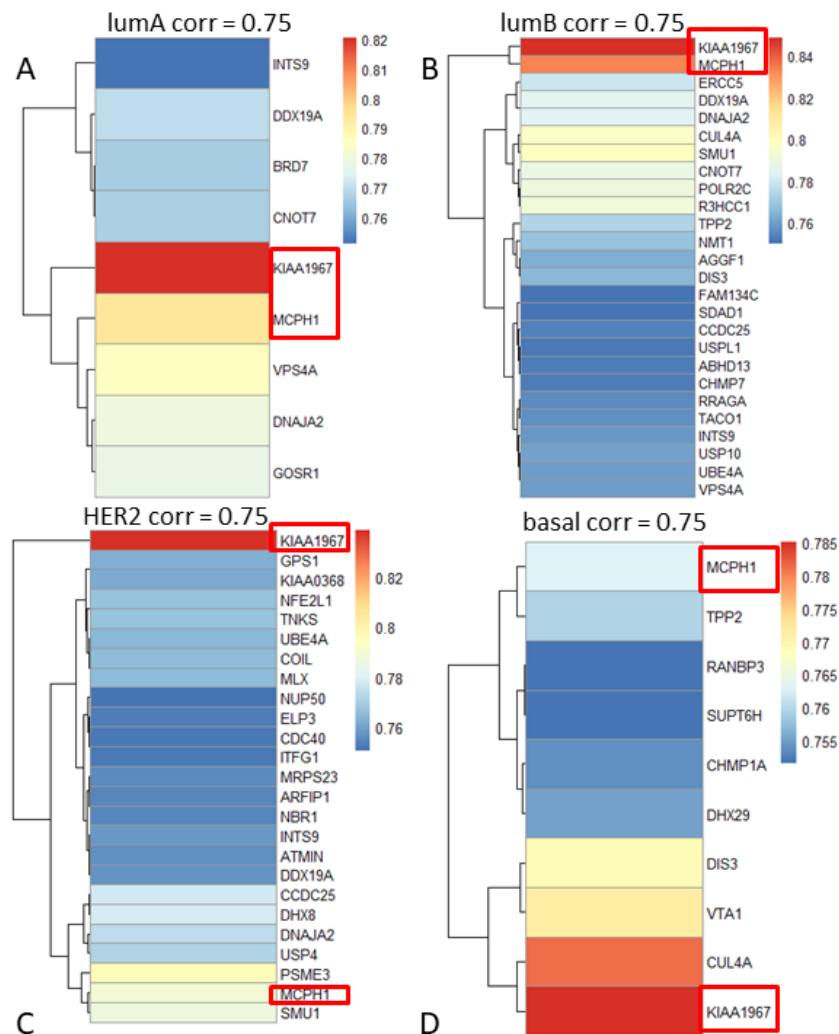
Figure 2. Heatmaps of correlation coefficients for each subtype: (**A**) lumA, (**B**) lumB, (**C**) HER2, and (**D**) basal (correlation coefficient ≥0.75). KIAA1967 and MCPH1 are highlighted as they showed a high correlation in all breast cancer subtypes.
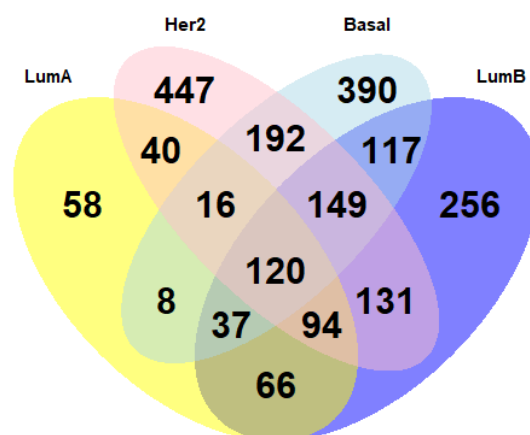
Figure 3. Venn diagram shows how genes whose expression is significantly correlated with the presence of copy number alterations are shared among breast cancer subtypes.

Another gene that, like *KIAA1967*, showed a high correlation value (0.795 in LumA, 0.834 in LumB, 0.789 in Her2, and 0.763 in basal) was the *MCPH1* gene. Like KIAA1967, MCPH1 was mostly deleted: 47% of the lumA samples, 56% of the lumB samples, 66% of the HER2 samples, and 59% of the basal samples.

Figure S1 shows the distribution of the copy number alterations of *KIAA1967* and *MCPH1* in the various BC subtypes.

3.1.2. Correlation Analysis: Subtype-Specific Genes

We obtained 58 subtype-specific genes exclusive for the lumA. Among these, we focused on those genes that had high correlation values in the Pearson test, such as *SMARCB1* (corr = 0.67, *p*-value < 0.001), *AZIN1* (corr = 0.66, *p*-value < 0.001), and *MTDH* (corr = 0.63, *p*-value < 0.001). We performed the same procedure for lumB, HER2, and basal.

For the lumB, we obtained 256 genes including the genes with a high correlation *PPP2R5E* (corr = 0.76, *p*-value < 0.001), *APEX1* (corr = 0.7, *p*-value < 0.001), and *GCN5* (corr = 0.7, *p*-value < 0.001); for the HER2 we identified 447 genes including *TNFAIP1* (corr = 0.80, *p*-value < 0.001), *PCYT2* (corr = 0.77, *p*-value < 0.001), and *DIABLO* (corr = 0.71, *p*-value < 0.001); and for basal we obtained 390 genes such as *FAM175B* (corr = 0.76, *p*-value < 0.001), *SENP5* (corr = 0.72, *p*-value < 0.001), and *SCAF1* (corr = 0.71, *p*-value < 0.001).

3.2. *Survival Analysis*

Survival analysis was applied to genes that showed a significant correlation between copy number and gene expression profiles (Figure 4). *KIAA1967* and *MCPH1*, mostly deleted in all BC subtypes, also demonstrated a prognostic role. Indeed, the low expression of *KIAA1967* and *MCPH1* was associated with a poor prognosis.

Survival analysis was performed on *SMARCB1*, *AZIN1*, and *MTDH*, which are lumA-specific genes. Low expression of *SMARCB1* and high expression of *AZIN1*, and *MTDH* were correlated with a recurrence free survival in patients with lumA. *PPP2R5E*, *APEX1*, and *GCN5*, lumB-specific genes, were associated with a poor prognosis; high expression of *PPP2R5E* and *APEX1*, and low expression of *GCN5* showed poor recurrence free survival in patients with lumB. *FAM175B*, one of the basal-specific genes, was correlated with recurrence free survival in basal BC patients.

3.3. *Analysis of GEO Datasets*

As we obtained copy number alterations of TCGA, GSE87048, and GSE26232 samples, we focused on the subtype-specific genes that have an alteration in at least 50% of the samples for each of the 3 datasets. Although RNA-Seq used in TCGA data and microarray used in GEO datasets are different technologies, they showed a high degree of concordance. This suggests that the identification of driver genes that show a consistency of results from different technologies is an important aspect of our study.

We found that 29 out of 58 genes in lumA, 90/256 in lumB, 40/447 in HER2, and 23/390 in basal were altered in at least 50% of the samples in all three datasets. Table S1 shows the list of these genes. We obtained a good percentage of consistency in the three datasets (29/58, 50%) in lumA; 35% (90/256) of genes were reproducible in the three datasets for lumB.

3.4. *Pathway Enrichment Analysis*

We examined the biological pathways associated with subtype-specific genes. Table 2 shows the top 3 pathways enriched with the genes.
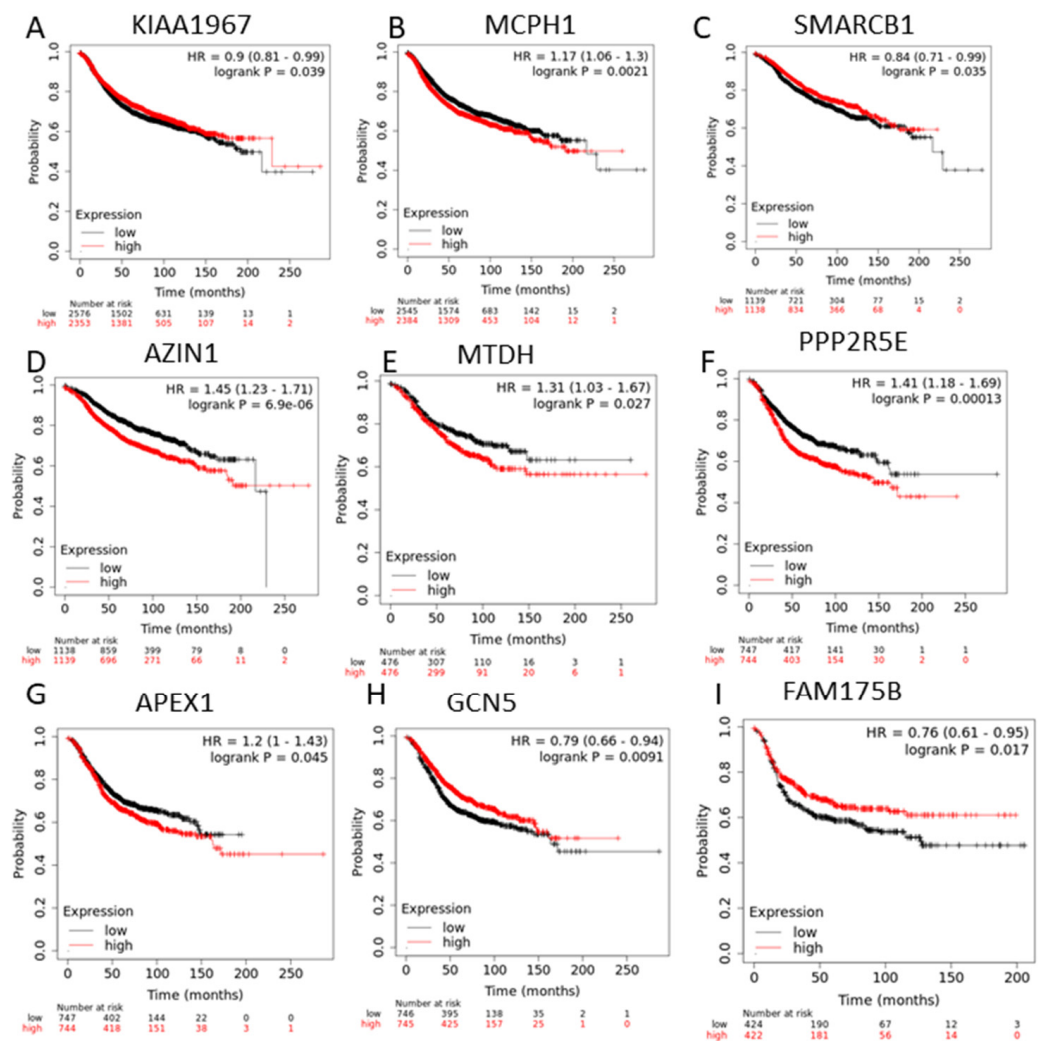
Figure 4. Survival analysis applied to genes whose expression is significantly correlated with the presence of copy number alterations in all breast cancer subtypes (**A**,**B**) *KIAA1967* and *MCPH1*; and subtype-specific genes (**C–I**) *SMARCB1* (**C**), *AZIN1* (**D**), *MTDH* (**E**) in lumA breast cancer patients, *PPP2R5E* (**F**), *APEX1* (**G**), *GCN5* (**H**) in lumBC samples, and *FAM175B* (**I**) in basal breast cancer patients.

We obtained for lumA the following pathways: SUMOylation of SUMOylation proteins, SUMOylation of ubiquitinylation proteins, and defective intrinsic pathway for apoptosis due to p14ARF loss of function.

The pathways cellular senescence, oxidative stress induced senescence, and disassembly of the destruction complex and recruitment of AXIN to the membrane were identified by the analysis for the lumB.

The pathways enriched with 40 HER2-specific genes were aryl hydrocarbon receptor signaling, VxPx cargo-targeting to cilium, and amplification of signal from the kinetochores.

The pathways enriched with 23 basal-specific genes were regulation of cholesterol biosynthesis by SREBP (SREBF), activation of gene expression by SREBF (SREBP), and metabolism of steroids.

Table 2. Pathways enriched by subtype-specific genes using Reactome.

	Pathway	*p*-Value
29 Luminal A-Specific Genes		
	SUMOylation of SUMOylation proteins	0.003
	SUMOylation of ubiquitinylation proteins	0.004
	Defective intrinsic pathway for apoptosis due to p14ARF loss of function	0.004
90 Luminal B-Specific Genes		
	Cellular senescence	0.005
	Oxidative stress induced senescence	0.001
	Disassembly of the destruction complex and recruitment of AXIN to the membrane	0.001
40 HER2-Specific Genes		
	Aryl hydrocarbon receptor signaling	0.0003
	VxPx cargo-targeting to cilium	0.003
	Amplification of signal from the kinetochores	0.004
23 Basal-Specific Genes		
	Regulation of cholesterol biosynthesis by SREBP (SREBF)	0.0002
	Activation of gene expression by SREBF (SREBP)	0.004
	Metabolism of steroids	0.004

4. Discussion

In this study we performed an analysis pipeline integrating copy number, gene expression, and clinical data. This analysis allowed for the identification of common genes among BC subtypes and subtype-specific genes, which may be potential candidates for personalized treatment. Unlike previous approaches, our analysis does not simply produce gene drivers, but we studied the effect of DNA copy number alteration on gene expression levels in the different stages of breast cancer progression defined by molecular subtypes.

In the first step of our study, we analyzed the correlation between copy number alterations and gene expression levels for each gene and for each subtype in 1024 BC samples of TCGA data. We obtained 439 genes correlated in lumA, 970 in lumB, 1189 in HER2, and 1029 in basal subtype.

We focused on the genes with a high correlation obtained in all subtypes. Specifically, a high correlation coefficient in all subtypes was reported for two genes, *KIAA1967* and *MCPH1*.

DBC1/KIAA1967 is a tumor suppressor gene that regulates p53-signaling through the inhibition of SIRT1 deacetylase. However, although *SIRT1* plays a vital role in carcinogenesis by regulating cell proliferation, survival, and death, its role in BC remains controversial [28]. *DBC1/KIAA1967* encodes a CCAR2 protein whose role is debated. It was reported that in squamous cell cancer, the loss of CCAR2 in mice results in cell cycle progression, suggesting that CCAR2 may function as a tumor suppressor. In addition, it was hypothesized that CCAR2 plays a role in promoting the stability of the transcription factors RFX1 and CREB1 required for proliferation [29].

However, the role of *DBC1/KIAA1967* is controversial, as in some studies it has been reported to be up-regulated and in others down-regulated within the same tumor [30]. A recent study investigated the clinical value of *DBC1/KIAA1967* in hepatocellular carcinoma, analyzing its prognostic ability with survival analysis; a higher expression of *DBC1/KIAA1967* reduced overall survival and disease free survival [30].

Survival analyses performed in our study demonstrated the prognostic role of *DBC1/KIAA1967* in BC; its low expression is associated with a poor prognosis.

The second gene with a high correlation in all subtype is *MCPH1*. Previous studies demonstrated that the tumor suppressor gene *MCPH1* is significantly associated with BC susceptibility in hereditary and non-hereditary BC [31]. These data therefore suggested that this gene is involved in the development of BC, despite its role needing to be further analyzed. In addition, mutated *MCPH1* down-regulates histone genes and leads to migration and invasion of the cells [32]. Frequent deletions and methylation of *MCPH1* were reported in many cancers, including BC, and were also associated with tumor stages in BC [33]. An important clinical role of *MCPH1* was demonstrated in our study; low expression of *MCPH1* is associated with a poor relapse free-survival.

Furthermore, in our study we focused on the subtype-specific genes that showed a correlation between copy number alterations and gene expression levels in TCGA data. Among 439 genes correlated in lumA, we found 58 lumA specific genes. The top three genes with the highest correlation were *SMARCB1* (corr = 0.67), *AZIN1* (corr = 0.66), and *MTDH* (corr = 0.63).

SMARCB1 was the lumA-specific gene with the highest correlation coefficient. It encodes a member of the ATP-dependent family and plays an important role in chromatin modelling, enabling the entry of transcription factors to DNA. A recent study demonstrated that the most frequent alterations of *SMARCB1* in human cancer are the deletions [34]. However, it is not completely clear what is the role of *SMARCB1* in BC [34,35].

The second gene with a high correlation coefficient is *AZIN1*. A previous study analyzed the role of *AZIN1* in polyamine homeostasis and cell proliferation in BC cells. Polyamines are an important pathway for different cellular functions, including cell growth [36]. Their intracellular concentrations are controlled by a complex network of regulatory mechanisms, in which *AZIN1* plays a key role. *AZIN1* decreases cellular polyamine by downregulating the enzyme that catalyzes the biosynthesis of polyamine, ornithine decarboxylase (ODC), and the absorption of polyamines [36]. The activity of *AZIN1* is repressed by the binding of a protein, called Az inhibitor (AzI), which is an enzymatically inactive homolog of ODC. Two forms of AzI have been described, namely AzI1, which is omnipresent, and AzI2, which is expressed in the brain and testes. The overexpression of *AZIN1* increases cell proliferation with a simultaneous increase in ODC activity and putrescine content [36].

MTDH, an oncogene, has been associated with an aggressive phenotype, poor prognosis, and chemo-resistance in BC [37]. A previous study described its possible mechanism of action in cancer; HIF-1 can bind to the *MTDH* promoter and regulate *MTDH* expression [37]. Other studies reported that *MTDH* can regulate two biological pathways involved in tumorigenesis and metastasis, the NF-κB and MAPK pathways [38].

For lumB we obtained 256 out of 970 specific genes. Additionally, in this case we focused on the genes with the highest correlation, namely *PPP2R5E*, *APEX1*, and *GCN5*.

The *PPP2R5E* gene, encoding protein phosphatase 2A (PP2A), is an important cellular phosphatase and plays key regulatory roles in growth, differentiation, and apoptosis [39]. Given the wide range of cellular functions of PP2A, its activity is tightly regulated to maintain cellular homeostasis [39].

The second lumB-specific gene is *APEX1*. Genetic alterations in genes that code for proteins that play a role in DNA repair pathways and in homologous recombination of DNA such as *APEX1*, *BRCA1*, *BRCA2*, *XRCC2*, *XRCC3*, *ATM*, *CHEK2*, *PALB2*, *RAD51*, and *XPD* have been implicated in BC [40]. APEX1 is a multifunctional protein that plays a central role in the base excision repair pathway. The APEX1 gene is highly polymorphic in cancer patients and has a role in the accumulation of the apurine/apyrimidine site in DNA and consequently may lead to an increased risk of cancer development [41].

GCN5, also known as *KAT2A*, a prototype of histone acetyltransferase (HAT), is involved in a variety of cellular processes and regulates downstream target genes by acetylation of different lysine residues (H3K9, H3K14, H3K18, and H3K23) in the core of histone H3 [42]. Furthermore, *GCN5* modulates the function of non-histone proteins, such

as transcription factors, by transferring an acetyl group to lysine residues. *GCN5* is also involved in cell cycle progression [42].

For HER2 we identified 447 genes, including *TNFAIP1*, *PCYT2*, and *DIABLO*.

A gene highly correlated and specific for the HER-2 tumor subtype is the *TNFAIP1* gene (corr = 0.8). Previous studies suggested that the *CSAGA*, *TNFAIP1*, and *POLDIP2* complex represents a gene module significantly associated with the amplification of the genomic region on 17q11.2 and correlated with the expression of ERBB2 in BC [43]. The co-expression pattern of this complex correlates with histological grades and a poor prognosis in BC when overexpressed [43].

A second gene analyzed, for the same reasons as the previous one, was the *PCYT2* gene (corr = 0.77). A previous study, based on metabolome analyses, demonstrated that glutamine deprivation leads to the accumulation of phosphoethanolamine (PEtn) in cancer cells through downregulation of cytidyltransferase PEtn (PCYT2) [44]. Accumulation of PEtn was correlated with tumor growth in nutrient-deficient conditions. *PCYT2* suppression was partially mediated by downregulation of the transcription factor ELF3. Furthermore, PCYT2 overexpression reduced PEtn levels and tumor growth [44]. *PCYT2* could represent a target in novel metabolic strategies for cancer [45].

Another gene with a high correlation was the *DIABLO* gene (corr = 0.71).

Since the Smac/DIABLO protein is involved in the mechanisms of apoptosis, it might be expected that the expression of this protein decreases with tumor development. A previous study confirmed this hypothesis, as Smac/DIABLO protein expression was significantly lower in stage 2 and stage 3 of BC than in stage 1 [46]. Furthermore, there was a weak correlation between low Smac/DIABLO protein expression and cancer embolism in minor blood and lymphatic vessels. In conclusion, the study indicated that Smac/DIABLO expression is inversely related to the tumor stage, which may suggest that this protein may play an important role in BC development [46].

For basal subtypes, we obtained 390 specific genes such as *FAM175B*, *SENP5*, and *SCAF1*.

The function of *FAM175B*, also known as *ABRO1* and *KIAA0157*, is largely unknown. Recent studies have revealed that *ABRO1* is a novel tumor suppressor by regulating the stability and functionality of p53 signaling. It plays an important role in tumor suppression and DNA damage response [47]. A recent study demonstrated that *ABRO1* overexpression stabilizes p53 and inhibits the growth of p53-expressing wild-type tumor cells, suggesting that the inhibition of cell growth by *ABRO1* upregulation is dependent on p53 status [47]. Furthermore, it was shown that *ABRO1* overexpression causes cell cycle arrest in the G1 phase, which is p53-dependent. These results suggest that *ABRO1* can suppress tumorigenesis. The role of *ABRO1* may be clinically relevant, because *ABRO1* protein levels are reduced in several cancerous tissues, including liver, kidney, breast, and thyroid cancers, and a higher *ABRO1* expression level correlates with better survival in patients [47].

A second important gene for the basal subtype is the *SENP5* gene. The downregulation of *SENP5* expression is associated with a good prognosis among BC patients. A previous study reported that the silencing of *SENP5* leads to the inhibition of growth, proliferation, and invasion in BC cell lines [48]. These changes are driven by the regulation of TGFβRI levels. One of TGFβRI target genes, *MMP9*, which plays a key role in degrading the extracellular matrix and contributes to invasion, is dramatically under-regulated by the silencing of *SENP5*. These data suggested the involvement of the SENP5-TGFβ-*MMP9* cascade in BC [48].

SCAF1 is one of the basal-specific genes with a higher correlation coefficient between copy number alteration and gene expression levels in TCGA data. Elevated expression of *SCAF1* was observed in a previous study in 31/81 (38.3%) BC tissues and was found to be more frequent in patients with tumors of large size, as well as in patients with lymph node invasion [49].

From the validation analysis using two GEO independent datasets, we found that 29 out of 58 genes in lumA, 90/256 in lumB, 40/447 in HER2, and 23/390 genes in basal were altered in at least 50% of the samples in two GEO dataset and TCGA data. The lowest

percentage of consistency for HER2 and basal can be explained by the low number of samples for the two GEO datasets. Indeed, GSE26232 contains only two HER2 samples, and genes found altered in 50% of the two samples may have affected the results. The same applied for the GSE87048 dataset, although it contained a few more samples (eight basal samples); this aspect could have influenced the obtained results.

We examined the biological pathways associated with 29 lumA-specific genes, and we obtained the following pathways: SUMOylation of SUMOylation proteins, SUMOylation of ubiquitinylation proteins, and defective intrinsic pathway for apoptosis due to p14ARF loss of function.

SUMO proteins are involved in different biological processes such as protein stability, cell growth, and apoptosis. It was previously reported that SUMO proteins are associated with advanced BC [50].

p14 plays a role in the cell cycle and apoptosis. A recent study demonstrated an association between the deletion of p14 and p53 signaling pathway disruption [51].

The pathways cellular senescence, oxidative stress induced senescence, and disassembly of the destruction complex and recruitment of AXIN to the membrane were identified by the enrichment analysis of 40 specific genes for the lumB.

40 HER2-specific genes were involved in 3 main pathways: Aryl hydrocarbon receptor signaling, VxPx cargo-targeting to cilium, and amplification of signal from the kinetochores.

The pathways enriched with 23 basal-specific genes were regulation of cholesterol biosynthesis by SREBP (SREBF), activation of gene expression by SREBF (SREBP), and metabolism of steroids. Previous studies reported that the under-expression of SREBP leads to a downregulation of several enzymes of fatty acids signaling. However, SREBP is upregulated in BC and associated with poor prognosis [52].

5. Conclusions

In conclusion, this study demonstrated the strong correlation between copy number alterations and gene expression levels of several known tumor suppressors and oncogenes. Thus, we revealed that integration analysis is crucial in discovery of therapeutic target genes in BC cancer subtypes.

References

1. Timms, K.M.; Abkevich, V.; Hughes, E.; Neff, C.; Reid, J.; Morris, B.; Kalva, S.; Potter, J.; Tran, T.V.; Chen, J.; et al. Association of BRCA1/2defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **2014**, *16*, 1–9. [CrossRef] [PubMed]
2. Sorlie, T.; Perou, C.M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10869–10874. [CrossRef] [PubMed]
3. Pusztai, L.; Mazouni, C.; Anderson, K.; Wu, Y.; Symmans, W.F. Molecular Classification of Breast Cancer: Limitations and Potential. *Oncology* **2006**, *11*, 868–877. [CrossRef]
4. Cava, C.; Zoppis, I.F.; Mauri, G.; Ripamonti, M.; Gallivanone, F.; Salvatore, C.; Gilardi, M.C.; Castiglioni, I. Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 608–611. [CrossRef]
5. Cava, C.; Bertoli, G.; Castiglioni, I. In silico identification of drug target pathways in breast cancer subtypes using pathway cross-talk inhibition. *J. Transl. Med.* **2018**, *16*, 154. [CrossRef]
6. Pinkel, D.; Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* **2005**, *37*, S11–S17. [CrossRef]
7. Li, D.; Xia, H.; Li, Z.-Y.; Hua, L.; Li, L. Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling. *BioMed Res. Int.* **2015**, *2015*, 1–17. [CrossRef] [PubMed]
8. Cava, C.; Bertoli, G.; Colaprico, A.; Bontempi, G.; Mauri, G.; Castiglioni, I. In-Silico Integration Approach to Identify a Key miRNA Regulating a Gene Network in Aggressive Prostate Cancer. *Int. J. Mol. Sci.* **2018**, *19*, 910. [CrossRef]
9. Cava, C.; Sabetian, S.; Castiglioni, I. Patient-Specific Network for Personalized Breast Cancer Therapy with Multi-Omics Data. *Entropy* **2021**, *23*, 225. [CrossRef] [PubMed]
10. Zhao, J.; Cheng, F.; Zhao, Z. Tissue-Specific Signaling Networks Rewired by Major Somatic Mutations in Human Cancer Revealed by Proteome-Wide Discovery. *Cancer Res.* **2017**, *77*, 2810–2821. [CrossRef] [PubMed]
11. Porta-Pardo, E.; Garcia-Alonso, L.; Hrabe, T.; Dopazo, J.; Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* **2015**, *11*, e1004518. [CrossRef] [PubMed]
12. Korthauer, K.D.; Kendziorski, C. MADGiC: A model-based approach for identifying driver genes in cancer. *Bioinformatics* **2015**, *31*, 1526–1535. [CrossRef]
13. Shrestha, R.; Hodzic, E.; Sauerwald, T.; Dao, P.; Wang, K.; Yeung, J.; Anderson, S.; Vandin, F.; Haffari, G.; Collins, C.C.; et al. HIT'nDRIVE: Patient-specific multidriver gene prioritization for precision oncology. *Genome Res.* **2017**, *27*, 1573–1588. [CrossRef]
14. Ding, J.; McConechy, M.K.; Horlings, H.M.; Ha, G.; Chan, F.C.; Funnell, T.; Mullaly, S.C.; Reimand, J.; Bashashati, A.; Bader, G.D.; et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.* **2015**, *6*, 8554. [CrossRef]
15. Suo, C.; Hrydziuszko, O.; Lee, D.; Pramana, S.; Saputra, D.; Joshi, H.; Calza, S.; Pawitan, Y. Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics* **2015**, *31*, 2607–2613. [CrossRef]
16. Bhattacharya, A.; Bense, R.D.; Urzúa-Traslaviña, C.G.; De Vries, E.G.E.; Van Vugt, M.A.T.M.; Fehrmann, R.S.N. Transcriptional effects of copy number alterations in a large set of human cancers. *Nat. Commun.* **2020**, *11*, 1–12. [CrossRef]
17. Fehrmann, R.S.N.; Karjalainen, J.M.; Krajewska, M.; Westra, H.-J.; Maloney, D.J.; Simeonov, A.; Pers, T.H.; Hirschhorn, J.N.; Jansen, R.C.; Schultes, E.A.; et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **2015**, *47*, 115–125. [CrossRef] [PubMed]
18. Fleck, J.L.; Pavel, A.B.; Cassandras, C.G. Integrating mutation and gene expression cross-sectional data to infer cancer progression. *BMC Syst. Biol.* **2016**, *10*, 1–12. [CrossRef] [PubMed]
19. Cava, C.; Zoppis, I.; Gariboldi, M.; Castiglioni, I.; Mauri, G.; Antoniotti, M. Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. *J Clin Bioinform.* **2014**, *4*, 2. [CrossRef] [PubMed]
20. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef]
21. Ng, S.; Collisson, E.A.; Sokolov, A.; Goldstein, T.; Gonzalez-Perez, A.; Lopez-Bigas, N.; Benz, C.; Haussler, D.; Stuart, J.M. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **2012**, *28*, i640–i646. [CrossRef] [PubMed]
22. Bussey, K.J.; Chin, K.; Lababidi, S.; Reimers, M.; Reinhold, W.C.; Kuo, W.-L.; Gwadry, F.; Ajay; Kouros-Mehr, H.; Fridlyand, J.; et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Ther.* **2006**, *5*, 853–867. [CrossRef]
23. Sales, G.; Coppe, A.; Bisognin, A.; Biasiolo, M.; Bortoluzzi, S.; Romualdi, C. MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res.* **2010**, *38*, W352–W359. [CrossRef]
24. Lygirou, V.; Latosinska, A.; Makridakis, M.; Mullen, W.; Delles, C.; Schanstra, J.P.; Zoidakis, J.; Pieske, B.; Mischak, H.; Vlahou, A. Plasma proteomic analysis reveals altered protein abundances in cardiovascular disease. *J. Transl. Med.* **2018**, *16*, 1–12. [CrossRef]

25. Zhao, Z.; Jinde, S.; Koike, S.; Tada, M.; Satomura, Y.; Yoshikawa, A.; Nishimura, Y.; Takizawa, R.; Kinoshita, A.; Sakakibara, E.; et al. Altered expression of microRNA-223 in the plasma of patients with first-episode schizophrenia and its possible relation to neuronal migration-related genes. *Transl. Psychiatry* **2019**, *9*, 1–11. [CrossRef]

26. Nagy, Á.; Lánczky, A.; Menyhárt, O.; Győrffy, B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci. Rep.* **2018**, *8*, 1–9. [CrossRef]

27. Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, D649–D655. [CrossRef] [PubMed]

28. Zhao, W.; Kruse, J.-P.; Tang, Y.; Jung, S.Y.; Qin, J.; Gu, W. Negative regulation of the deacetylase SIRT1 by DBC1. *Nat. Cell Biol.* **2008**, *451*, 587–590. [CrossRef]

29. Best, S.A.; Nwaobasi, A.N.; Schmults, C.D.; Ramsey, M.R. CCAR2 Is Required for Proliferation and Tumor Maintenance in Human Squamous Cell Carcinoma. *J. Investig. Dermatol.* **2017**, *137*, 506–512. [CrossRef] [PubMed]

30. Li, C.; Liao, J.; Wu, S.; Fan, J.; Peng, Z.; Wang, Z. Overexpression of DBC1, correlated with poor prognosis, is a potential therapeutic target for hepatocellular carcinoma. *Biochem. Biophys. Res. Commun.* **2017**, *494*, 511–517. [CrossRef]

31. Richardson, J.; Shaaban, A.M.; Kamal, M.; Alisary, R.; Walker, C.; Ellis, I.O.; Speirs, V.; Green, A.R.; Bell, S.M. Microcephalin is a new novel prognostic indicator in breast cancer associated with BRCA1 inactivation. *Breast Cancer Res. Treat.* **2010**, *127*, 639–648. [CrossRef]

32. Tervasmäki, A.; Mantere, T.; Eshraghi, L.; Laurila, N.; Tuppurainen, H.; Ronkainen, V.; Koivuluoma, S.; Devarajan, R.; Peltoketo, H.; Pylkäs, K. Tumor suppressor MCPH1 regulates gene expression profiles related to malignant conversion and chromosomal assembly. *Int. J. Cancer* **2019**, *145*, 2070–2081. [CrossRef] [PubMed]

33. Bhattacharya, N.; Mukherjee, N.; Singh, R.K.; Sinha, S.; Alam, N.; Roy, A.; Roychoudhury, S.; Panda, C.K. Frequent Alterations of MCPH1 and ATM are Associated with Primary Breast Carcinoma: Clinical and Prognostic Implications. *Ann. Surg. Oncol.* **2012**, *20*, 424–432. [CrossRef] [PubMed]

34. Mimori, K.; Inoue, H.; Shiraishi, T.; Ueo, H.; Mafune, K.-I.; Tanaka, Y.; Mori, M. A Single-Nucleotide Polymorphism of SMARCB1 in Human Breast Cancers. *Genomics* **2002**, *80*, 254–258. [CrossRef]

35. Sévenet, N.; Lellouch-Tubiana, A.; Schofield, D.; Hoang-Xuan, K.; Gessler, M.; Birnbaum, D.; Jeanpierre, C.; Jouvet, A.; Delattre, O. Spectrum of hSNF5IINI1 Somatic Mutations in Human Cancer and Genotype-Phenotype Correlations. *Hum. Mol. Genet.* **1999**, *8*, 2359–2368. [CrossRef] [PubMed]

36. Silva, T.M.; Cirenajwis, H.; Wallace, H.M.; Oredsson, S.; Persson, L. A role for antizyme inhibitor in cell proliferation. *Amino Acids* **2015**, *47*, 1341–1352. [CrossRef] [PubMed]

37. Chu, P.-Y.; Wang, S.-M.; Chen, P.-M.; Tang, F.-Y.; Chiang, E.-P.I. Expression of MTDH and IL-10 Is an Independent Predictor of Worse Prognosis in ER-Negative or PR-Negative Breast Cancer Patients. *J. Clin. Med.* **2020**, *9*, 3153. [CrossRef] [PubMed]

38. Nguyen, Y.T.-K.; Moon, J.Y.; Ediriweera, M.K.; Cho, S.K. Phenethyl Isothiocyanate Suppresses Stemness in the Chemo- and Radio-Resistant Triple-Negative Breast Cancer Cell Line MDA-MB-231/IR Via Downregulation of Metadherin. *Cancers* **2020**, *12*, 268. [CrossRef] [PubMed]

39. Dupont, W.D.; Bs, J.P.B.; Bs, K.M.B.; Rn, P.A.S.; Bs, W.D.P.; Sanders, M.E.; Page, D.L.; Smith, J.R. Protein phosphatase 2A subunit gene haplotypes and proliferative breast disease modify breast cancer risk. *Cancer* **2009**, *116*, 8–19. [CrossRef] [PubMed]

40. Ali, K.; Mahjabeen, I.; Sabir, M.; Baig, R.M.; Zafeer, M.; Faheem, M.; Kayani, M.A. Germline variations of apurinic/apyrimidinic endonuclease 1 (APEX1) detected in female breast cancer patients. *Asian Pac. J. Cancer Prev.* **2014**, *15*, 7589–7595. [CrossRef]

41. Coughlin, S.S. Epidemiology of Breast Cancer in Women. *Adv. Exp. Med. Biol.* **2019**, *1152*, 9–29. [CrossRef]

42. Oh, J.H.; Lee, J.-Y.; Kim, K.H.; Kim, C.Y.; Jeong, D.S.; Cho, Y.; Nam, K.T.; Kim, M.H. Elevated GCN5 expression confers tamoxifen resistance by upregulating AIB1 expression in ER-positive breast cancer. *Cancer Lett.* **2020**, *495*, 145–155. [CrossRef]

43. Taghavi, A.; Akbari, M.E.; Hashemi-Bahremani, M.; Nafissi, N.; Khalilnezhad, A.; Poorhosseini, S.M.; Hashemi-Gorji, F.; Yassaee, V.R. Gene expression profiling of the 8q22-24 position in human breast cancer: TSPYL5, MTDH, ATAD2 and CCNE2 genes are implicated in oncogenesis, while WISP1 and EXT1 genes may predict a risk of metastasis. *Oncol. Lett.* **2016**, *12*, 3845–3855. [CrossRef] [PubMed]

44. Osawa, T.; Shimamura, T.; Saito, K.; Hasegawa, Y.; Ishii, N.; Nishida, M.; Ando, R.; Kondo, A.; Anwar, M.; Tsuchida, R.; et al. Phosphoethanolamine Accumulation Protects Cancer Cells under Glutamine Starvation through Downregulation of PCYT2. *Cell Rep.* **2019**, *29*, 89–103.e7. [CrossRef]

45. Zhu, L.; Bakovic, M. Breast cancer cells adapt to metabolic stress by increasing ethanolamine phospholipid synthesis and CTP:ethanolaminephosphate cytidylyltransferase-Pcyt2 activity. *Biochem. Cell Biol.* **2012**, *90*, 188–199. [CrossRef]

46. Pluta, P.; Cebula-Obrzut, B.; Ehemann, V.; Pluta, A.; Wierzbowska, A.; Piekarski, J.; Bilski, A.; Nejc, D.; Kordek, R.; Robak, T.; et al. Correlation of Smac/DIABLO protein expression with the clinico-pathological features of breast cancer patients. *Neoplasma* **2011**, *58*, 430–435. [CrossRef] [PubMed]

47. Zhang, J.; Cao, M.; Dong, J.; Li, C.; Xu, W.; Zhan, Y.; Wang, X.; Yu, M.; Ge, C.; Ge, Z.; et al. ABRO1 suppresses tumourigenesis and regulates the DNA damage response by stabilizing p53. *Nat. Commun.* **2014**, *5*, 5059. [CrossRef] [PubMed]

48. Cashman, R.; Cohen, H.; Ben-Hamo, R.; Zilberberg, A.; Efroni, S. SENP5 mediates breast cancer invasion via a TGFβRI SUMOylation cascade. *Oncotarget* **2014**, *5*, 1071–1082. [CrossRef]

49. Adamopoulos, P.G.; Raptis, G.D.; Kontos, C.K.; Scorilas, A. Discovery and expression analysis of novel transcripts of the human SR-related CTD-associated factor 1 (SCAF1) gene in human cancer cells using Next-Generation Sequencing. *Gene* **2018**, *670*, 155–165. [CrossRef]

50. Alshareeda, A.T.; Negm, O.H.; Green, A.R.; Nolan, C.; Tighe, P.; AlBarakati, N.; Sultana, R.; Madhusudan, S.; Ellis, I.O.; Rakha, E.A. SUMOylation proteins in breast cancer. *Breast Cancer Res. Treat.* **2014**, *144*, 519–530. [CrossRef]

51. Wazir, U.; Jiang, W.G.; Yasaei, H.; Linne, H.; Newbold, R.F.; Mokbel, K. P14ARF is down-regulated during tumour progression and predicts the clinical outcome in human breast cancer. *Anticancer. Res.* **2013**, *33*, 2185–2189.

52. Bao, J.; Zhu, L.; Zhu, Q.; Su, J.; Liu, M.; Huang, W. SREBP-1 is an independent prognostic marker and promotes invasion and migration in breast cancer. *Oncol. Lett.* **2016**, *12*, 2409–2416. [CrossRef] [PubMed]