Identification of Candidate Genes in Early-Stage Invasive Ductal Carcinoma Patients with High-Risk Mortality Using Genes Commonly Involved in Breast Cancer: A Retrospective Study

Chih-Chiang Hung\textsuperscript{a, b} Hsin-I. Huang\textsuperscript{c} Chao-Ming Hung\textsuperscript{d,e} Sin-Hua Moi\textsuperscript{c}

\textsuperscript{a}Division of Breast Surgery, Department of Surgery, Taichung Veterans General Hospital, Taichung, Taiwan; \textsuperscript{b}Department of Applied Cosmetology, College of Human Science and Social Innovation, Hungkuang University, Taichung, Taiwan; \textsuperscript{c}Center of Cancer Program Development, E-Da Cancer Hospital, I-Shou University, Kaohsiung, Taiwan; \textsuperscript{d}Department of General Surgery, E-Da Cancer Hospital, Kaohsiung, Taiwan; \textsuperscript{e}School of Medicine, College of Medicine, I-Shou University, Kaohsiung, Taiwan

Keywords
Early breast cancer · Invasive ductal carcinoma · Survival analysis · Gene mutation

Abstract

Introduction: Invasive ductal carcinoma (IDC) of the breast is a heterogeneous disease characterized by multiple subtypes. IDC survival is highly impacted by tumor burden, molecular subtypes, and gene profiles. Gene mutation is a type of genomic instability regarded as having a considerable effect on IDC prognosis. Using integrated survival analysis, this study identified candidate genes and a high-risk group of patients with early-stage IDC to provide further understanding of the genetic characteristics associated with poor survival.

Methods: The gene mutation profiles, baseline demographics, clinicopathologic variables, and treatment characteristics of the early-stage IDC subpopulation were downloaded from an open access data platform. These data were analyzed for a total of 444 patients. In total, 40 genes commonly involved in IDC were listed, and the genes exhibiting significant differences (as estimated using the log-rank test) were selected as the candidate genes.

Results: The patients were divided into control, low-risk, and high-risk groups according to their gene mutation profiles. The 5-year overall survival rates of low-risk, control, and high-risk patients were 97.4%, 96.1%, and 73.0%, respectively. The high-risk group had a significantly higher risk of poor overall survival (adjusted hazard ratio = 6.57, 95% confidence interval = 1.51–28.7, \( p = 0.012 \)) than that of the control group, and the low-risk group did not have a significant survival difference compared with control group.

Conclusions: This study proposed an integrative approach for the identification of candidate genes for risk assessment of overall survival in these patients through typical survival analysis methods. The 14 candidate genes selected are particularly involved in cell-cycle processes, deoxyribonucleic acid repair, and drug resistance; their mutations were found to be generally associated with disease progression or therapeutic resistance, which is commonly associated with poor overall survival outcomes in IDC.

© 2021 The Author(s)
Published by S. Karger AG, Basel

Introduction

Invasive ductal carcinoma (IDC), which accounts for approximately 65%–85% of breast cancer cases, is a heterogeneous disease characterized by multiple subtypes [1]. Studies have demonstrated that the treatment op-
Gene mutations are a type of genomic instability considered to have a considerable effect on breast cancer prognosis [8]. Because gene mutations in breast cancer are very broad in scope, identifying all of the related genes is difficult [9]. Therefore, the genes previously reported to be involved in breast cancer were considered appropriate target genes in evaluating the effects of gene mutation on overall survival. TP53, HER2/ERBB2, BRCA1, BRCA2, and PTEN are all well-known genes in breast cancer development and progression [10, 11]. Somatic mutation of TP53 is the most frequent gene alteration associated with human cancer, while BRCA1, PTEN, EGFR, baculoviral IAP repeat containing 5 (BIRC5), PIK3CA, and CHEK2 are frequent co-alterations associated with TP53 and gathered in the p53 signaling pathway [12, 13]. BRCA1/2/3 mutations often participate in the hereditary signaling pathway, and BRCA1/2 is also recently reported playing roles in the biologic response to DNA damage [14, 15]. Moreover, overexpression of EMT was previously reported to be associated with the impairment of BRCA2 [16]. The DNA damage response, mismatch repair, and homologous recombination were highly associated with breast cancer progression, especially in cell-cycle procedure [17]. Besides, TP53, MLH1, PMS1, MSH2/3, BARD1, BRIP1, CHEK2, NBN, NF1, PALB2, RAD51C, and RAD51D also participated in the process of homologous recombination and damage response of DNA [18–26]. The estrogen signaling pathway could be regulated by the genetic markers including EGFR, ESRI, CCND1, IGF1R, MYC, PIK3CA, and SRC, which is also associated with the DNA damage response [27, 28]. The activity of SRC could be regulated by the tumor suppressor gene, RASSF1, which is often found in the tumorogenesis signaling pathway [29]. In addition, the regulation of G2/M cell cycle of the DNA damage checkpoint in breast cancer could be regulated by TOP2A, which is a frequent co-alteration associated with ERBB2/HER2 [30]. PIK3CA is also involved in the HER2 signaling pathway which primarily depends on the expression of ERBB2/HER2 [31]. The roles of PIK3CA also extended to the mTOR signaling pathway and associated with Harvey rat sarcoma viral oncogene homolog (HRAS), Kirsten rat sarcoma 2 viral oncogene homolog (KRAS), HIF1A, and STK11 [32, 33]. PI3K/mTOR pathways are frequently related to cancer progression and drug resistance via cellular transformation [34]. Furthermore, the epithelial-mesenchymal transition (EMT) state is well recognized and associated with a poor survival outcome in breast cancer; the related markers including PTEN, FGFR1, ATM, EGFR, CDH1, HRAS, NRAS, KRAS, Notch homolog 1 (NOTCH1), HIF1A, TWIST1, and RARB were involved in the controlling process of EMT [35, 36]. Collectively, the aforementioned breast cancer progression-associated genes could be considered as survival outcome estimation genetic markers for IDC.

Most research studies so far have focused on a single gene or the effect of a specific gene function on breast cancer survival outcomes. An integration method for identification of candidate genes associated with poor survival is required to better understand the effects of gene mutation on overall survival. Typically, clinicians are more familiar with survival analysis methods than complex machine learning methods. Although multiple machine learning algorithms have provided precise and efficient analyses, using these algorithms is not necessarily worthwhile due to the long-running time and the computational complexity – particularly for cases with only slight improvement. Therefore, in this study, we used an integrative approach involving typical survival analysis methods to identify candidate genes in patients with early-stage IDC breast cancer. This study aimed to achieve further understanding of the genetic characteristics associated with poor survival outcomes in breast cancer.

**Materials and Methods**

**Dataset**

The somatic mutation (MuTect2 pipeline) in the DNA-seq level and clinical information of breast invasive carcinoma were obtained from the Pan-Cancer Atlas of The Cancer Genome Atlas (TCGA) research network. All data can be downloaded from the cBioPortal (https://www.cbioportal.org/datasets). The data use certification agreement of TCGA allows for exemption from further approval by Ethics Committee (https://www.cancer.gov/about-nci/organization/ccg/research/strucutral-genomics/tcga/history/policies). The patient inclusion criteria were as follows: (1) female; (2) diagnosis of pathologic stage I, II, or III breast cancer; (3) specifically IDC; and (4) estimable survival status and follow-up interval. A total of 444 patients were included for analysis. Age at diagnosis, race, and ethnicity was included as baseline demographics. Age at diagnosis was categorized as <35, 35–50, and >50 years. Others prognostic variables, including breast cancer subtypes, pathologic stage, and treatment characteristics, were also included in later analysis.
The roles of each involved genes breast cancer and the supportive references were summarized in online suppl. Table 1. The gene mutation profile of each patient was matched with the genes on the involved gene list. The gene mutation status of each patient was determined based on matching results between the gene list and the patient’s gene mutation profile. The proportion of somatic mutation detected at the gene-level in the study population was defined as the mutation rate. For instance, a total of 187 patients obtained TP53 mutants among 444 patients; thence, the mutation rate of TP53 was 0.421 (187/444) in overall patients.

Gene Mutation Risk Groups

Afterward, survival analysis was used to evaluate the effect of the gene mutation statuses on overall survival. The involved genes that displayed significant survival differences between mutated and nonmutated patients were selected as the candidate genes. The patients who had no mutations in any of the 40 involved genes on the list were defined as controls. According to gene mutation status, the remaining patients were divided into low- and high-risk groups. Patients with mutations in the candidate genes were placed into the high-risk group, and the others were assigned to the low-risk group.

Outcome Measurements

Overall survival is considered a suitable endpoint for the evaluation of risk characteristics in patients with cancer [37]. The Kaplan-Meier estimator with the log-rank test and the Cox proportional hazard regression model testing are methods typically used for analyzing overall survival. They are based on a time-to-event model and have been widely used in cancer research [38, 39]. The Cox proportional hazard regression model can singly or partially assess the effect of both single and multiple variables on overall survival [38].

Statistical Analysis

The baseline demographics, clinicopathologic variables, treatment characteristics, and overall survival outcomes of the study population are presented as frequencies and percentages, and the distribution differences in between the control, low-risk, and high-risk groups were estimated using the χ² test or Fisher’s exact test, distribution differences in between the control, low-risk, and high-risk population are presented as frequencies and percentages, and the patient’s gene mutation profile. The proportion of somatic mutation detected at the gene-level in the study population was defined as the mutation rate. For instance, a total of 187 patients obtained TP53 mutants among 444 patients; thence, the mutation rate of TP53 was 0.421 (187/444) in overall patients.

Gene Mutation Risk Groups

Afterward, survival analysis was used to evaluate the effect of the gene mutation statuses on overall survival. The involved genes that displayed significant survival differences between mutated and nonmutated patients were selected as the candidate genes. The patients who had no mutations in any of the 40 involved genes on the list were defined as controls. According to gene mutation status, the remaining patients were divided into low- and high-risk groups. Patients with mutations in the candidate genes were placed into the high-risk group, and the others were assigned to the low-risk group.

Outcome Measurements

Overall survival is considered a suitable endpoint for the evaluation of risk characteristics in patients with cancer [37]. The Kaplan-Meier estimator with the log-rank test and the Cox proportional hazard regression model testing are methods typically used for analyzing overall survival. They are based on a time-to-event model and have been widely used in cancer research [38, 39]. The Cox proportional hazard regression model can singly or partially assess the effect of both single and multiple variables on overall survival [38].

Statistical Analysis

The baseline demographics, clinicopathologic variables, treatment characteristics, and overall survival outcomes of the study population are presented as frequencies and percentages, and the distribution differences in between the control, low-risk, and high-risk groups were estimated using the χ² test or Fisher’s exact test, as appropriate. The mutation rates of the 40 involved genes in the patients were determined, and the differences in overall survival by mutation status of each gene were estimated using the log-rank test. The genes that displayed significant differences in survival between the patients with and without mutations in the candidate genes were included in the high-risk gene set. The Kaplan-Meier curve was used to illustrate the differences in overall survival by risk category and single gene mutation status. Cox proportional hazard regression analysis was performed to estimate the effect of gene mutation risk score and covariates on overall survival. All factors were included in the multivariate model. All p values were 2-sided, and the statistical significance was set at 0.05. All analyses were conducted using version 4.0.1 of the computing environment R (R Core Team, 2020).

### Table 1. Baseline demographics and clinicopathologic variables of patients with early-stage IDC breast cancer

| Variables                        | Overall, n = 444 |
|----------------------------------|------------------|
| **Age, years, n (%)**            |                  |
| <35                              | 10 (2.3)         |
| 35–50                            | 141 (31.8)       |
| >50                              | 293 (66.0)       |
| **Race, n (%)**                  |                  |
| American Indian or Alaska native | 1 (0.2)          |
| Asian                            | 36 (8.1)         |
| Black or African American        | 56 (12.6)        |
| White                            | 307 (69.1)       |
| Not reported                     | 44 (9.9)         |
| **Ethnicity, n (%)**             |                  |
| Hispanic or Latino               | 13 (2.9)         |
| Not Hispanic or Latino           | 345 (77.7)       |
| Not reported                     | 86 (19.4)        |
| **Subtypes, n (%)**              |                  |
| Basal-like                       | 101 (22.7)       |
| HER2-enriched                    | 42 (9.5)         |
| Luminal A or normal-like         | 191 (43.0)       |
| Luminal B                        | 110 (24.8)       |
| **Pathologic stage, n (%)**      |                  |
| Stage I                          | 89 (20.0)        |
| Stage II                         | 273 (61.5)       |
| Stage III                        | 82 (18.5)        |
| **Treatment characteristics, n (%)** |            |
| Radiation                        | 243 (54.7)       |
| Pharmaceutical                   | 363 (81.8)       |
| All-cause mortality              | 16 (3.6)         |

HER2, human epidermal growth factor receptor 2; IDC, invasive ductal carcinoma.

### Results

**Patient Characteristics**

Baseline demographics, clinicopathologic variables, and treatment characteristics of the study population are summarized in Table 1. At a median follow-up of 2.2 years (range 1.2–23.4), 16 all-cause deaths (3.6%) had occurred. The study population had a higher proportion of patients who were aged >50 years (293; 66.0%), were Caucasian (69.1%), and not Hispanic or Latino (77.7%). The study population was more likely to have been diagnosed as having the luminal A or normal-like (191; 43.0%) and luminal B (110; 24.8%) molecular subtypes of breast cancer. The basal-like subtype was observed in 101 (22.7%) patients and HER2 overexpression was observed in 42 (9.5%) of patients. In treatment characteristics, 243 (54.7%) patients had received radiation treatment, and 363 (81.8%) patients had received pharmaceutical treatment.
Gene Mutation Profiles

In Table 2, the mutation rate of each gene among overall population, alive, and dead groups was presented. Among the 444 patients, the mutations in 37 of the 40 involved genes were observed, and the highest mutation rate was noted for TP53 (0.421) and PIK3CA (0.372), followed by PTEN (0.043), ATM (0.043), and NF1 (0.043).

No mutations were found for BRCA3, EMSY, and RASSF1. The survival different of mutated and nonmutated subgroup in each gene were compared and tested, and the results showed 14 candidate genes obtained a higher mutation rate and significant poor survival outcome in the mutated subgroup compared to the nonmutated subgroup, including BARD1 (mutation rate in dead vs. alive: 0.005).
0.063 vs. 0, \( p < 0.001 \), BIRC5 (0.063 vs. 0.002, \( p < 0.001 \)), CCND1 (0.063 vs. 0, \( p < 0.001 \), CHEK2 (0.063 vs. 0.007, \( p = 0.005 \)), ESR1 (0.063 vs. 0.014, \( p = 0.026 \), HRAS (0.063 vs. 0, \( p = 0.001 \)), IGFR1 (0.125 vs. 0.014, \( p = 0.002 \), KRAS (0.125 vs. 0.009, \( p < 0.001 \), MLH1 (0.063 vs. 0.014, \( p = 0.007 \), MSH2 (0.125 vs. 0.005, \( p < 0.001 \), NOTCH1 (0.125 vs. 0.009, \( p < 0.001 \), SRC (0.063 vs. 0.007, \( p = 0.002 \), RAD51C (0.063 vs. 0.014, \( p = 0.011 \), and RAD51D (0.125 vs. 0, \( p < 0.001 \). The survival curve comparison in 14 candidate genes is summarized in online suppl. Fig. 1–14, the solid line represents the nonmutated group, and the dashed line represents the specific gene mutated group.

Figure 1 shows the gene mutation profiles and clinical features for the patients who have at least 1 mutation of involved genes, thus there was only 317 patients were showed. The waterfall plot summarized the percentage of mutant, mutations per megabase, mutation type, and clinical features of 317 patients. The percentage of mutants was shown in the left panel according to the percentage of single-gene mutation based on 317 patients. The mutations per megabase in the upper panel summarized the tumor mutational burden (TMB) of somatic mutation in 317 patients. In the medium panel, the gene mutation profiles of 317 patients were illustrated using the waterfall plot according to the mutation type detected in each gene. The clinical features including age-group, pathologic stage, breast cancer subtypes, and survival status of 317 patients were summarized in the bottom panel, and the waterfall plot was sorted from left to right according to the survival status of the study population. The results demonstrated that pathogenic mutations of the involved gene were imploled in TP53 (83/317, 26.2% pathogenic mutants) and PIK3CA (57/317, 18.0% pathogenic mutants), other pathogenic mutations were found in KRAS (4/317, 1.2% pathogenic mutants), PTEN (2/317, 0.6% pathogenic mutants), ATM (1/317, 0.3% pathogenic mutants), MLH1 (1/317, 0.3% pathogenic mutants), and MLH2 (1/317, 0.3% pathogenic mutants). Other mu-

---

**Fig. 1.** Gene mutation profiles and clinical features of early-stage IDC breast cancer. IDC, invasive ductal carcinoma.
Identification of High-Risk Mortality Genes in IDC

Fig. 2. Early-stage IDC according to gene mutation risk groups. The solid line represents controls, dotted line represents low-risk group, and dashed line represents high-risk group. IDC, invasive ductal carcinoma.

tation types including deletion, insertion, SNV, and other mutants were also illustrated using different colors as shown in Figure 2. HRAS (1/317, 0.3% pathogenic mutants), CCND1 (1/317, 0.3% SNV mutants), and BIRC5 (1/317, 0.3% SNV mutants) showed an SNV (1/317, 0.3% SNV mutants) gene mutation only in the dead group, but others showed no particular tendency to implode according to survival status. Moreover, the TMB results showed that dead patients have higher TMB than alive patients.

Gene Mutation Risk Groups

Hundred and twenty-seven patients who had no mutations in any of the 40 involved genes were enrolled as controls. The low-risk group consisted of 279 patients with mutations in 23 of 40 involved genes (not including 14 candidate genes and 3 nonmutated genes). Thirty-eight patients with mutations in 14 candidate genes including BARD1, BIRC5, CCND1, CHEK2, ESR1, HRAS, IGF1R, KRAS, MLH1, MSH2, NOTCH1, SRC, RAD51C, and RAD51D constituted the high-risk group. Baseline demographics, clinicopathologic variables, and treatment characteristics of the 127 controls, 279 low-risk patients, and 38 high-risk patients are summarized in Table 3. The 3 groups had similar distributions of age-group, race, ethnicity, subtypes, pathologic stage, and treatment characteristics. With regard to the molecular subtype distribution, the high-risk subgroup had a slightly higher proportion of patients with the basal-like (29.0% in high-risk, 10.2% in low-risk, and 9.5% in control). The high-risk group had a significantly higher mortality rate (5 deaths; 13.2%) than that of the control (4 deaths; 3.1%) or low-risk (7 deaths; 2.5%) group. The between-group survival differences are shown in Figure 2. Five-year overall survival rates of the high-risk, low-risk, and control groups were 73.0%, 96.1%, and 97.4%, respectively. Control and low-risk groups had a similar overall survival rate, and the high-risk group had significantly worse survival than that of the 2 other groups (p < 0.001).

Table 4 presents the results of the Cox proportional hazard regression analysis for overall survival. The effect of the group category (control, low risk, or high risk) and known prognostic factors (i.e., age-group, molecular subtypes, pathologic stages, and treatment characteristics) on overall survival were analyzed. Compared with controls, the high-risk group had a significantly higher risk of poor overall survival. The estimated hazard ratios were 7.22 (95% CI = 1.82–28.70, p = 0.005) and 6.57 (95% CI = 1.51–28.7, p = 0.012) in the univariate and multivariate models, respectively. The low-risk group did not have a significantly higher risk of poor overall survival than did controls.

Discussion/Conclusion

This study adopted an integrative approach in the identification of candidate genes for risk assessment of overall survival in patients with early-stage breast cancer. The low-risk group with mutation in any genes of 23 of 40 involved genes (not including 14 candidate genes and 3 nonmutated genes) showed no survival difference compared to the control group. Mutations in the 14 candidate genes were used to identify the high-risk group and achieved a higher risk of poor overall survival than that of control and low-risk groups. Thence, compared with the 23 mutated involved genes, the gene functions of 14 candidate genes might be more related to cancer activity associated with poor overall survival outcomes. These 14 candidate genes mostly participate in cell-cycle processes (estrogen signaling, p53 signaling, and mTOR signaling pathway), DNA damage response, and drug resistance.

CCND1 (B-cell leukemia/lymphoma 1) is strongly associated with the expression of estrogen receptors, particularly in node-positive breast cancer [40]. In addition, the co-alterations of CCND1 and EMSY (BRCA2-interacting transcriptional repressor) are associated with poor treatment outcome in ER-positive breast cancer [41]. ESR1 (estrogen receptor 1) encodes ER-alpha, is an estro-
gen receptor, and the overexpression of ER alpha protein could be used as a prognostic factor for ER-positive breast cancer and is predictive of endocrine therapy responsiveness [42]. The V-Src avian sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (SRC) encodes the nonreceptor protein tyrosine kinase, which plays a role in the regulation of cell proliferation, migration, and survival [43]. The SRC mutation is known to be involved in malignant progression [44, 45]. Increased SRC activation is correlated with trastuzumab resistance and interacts with estrogen receptors. Therefore, the combined use of SRC inhibitors has been suggested to overcome drug resistance in specific breast cancers [46]. Insulin-like growth factor 1 receptor (IGF1R) is a receptor with tyrosine kinase activity, and its expression is associated with tumorigenesis [47]. Moreover, IGF1R overexpression is more common in ER-positive breast cancer. IGF1R expression is generally low in the HER2-enriched subtype and is heterogeneous in the basal-like subtype [47]. In addition, IGF1R may be used as an alternative therapeutic target for HER2-enriched, trastuzumab-resistant, or basal-like breast cancer [48, 49].

The intracellular actions mediated by IGF1R are also involved in RAS signaling pathways [47], which are associated with the oncogenes HRAS and KRAS in the RAS family. RAS mutations also participate in mTOR signaling and EMT pathways, which are associated with cancer progression and drug resistance via cellular transformation [34, 36]. Moreover, ESR1 and NOTCH1 are also involved in the EMT pathway. NOTCH1 is a protein-coding gene contributes to multiple cellular processes, including cell differentiation, proliferation, and survival [50]. High expression of NOTCH1 may induce breast carcinoma and is strongly associated with poor prognosis and survival outcomes in breast cancer [51, 52].

**Table 3.** Baseline demographics, clinicopathologic variables, and treatment characteristics of gene mutation risk groups in patients with early-stage IDC breast cancer

| Variables | Controls (n = 127) | Low-risk (n = 279) | High-risk (n = 38) | p value |
|-----------|-------------------|-------------------|-------------------|---------|
| Age, years, n (%) | | | | |
| <35 | 4 (3.1) | 4 (1.4) | 2 (5.3) | |
| 35–50 | 47 (37.0) | 87 (31.2) | 7 (18.4) | 0.074 |
| >50 | 76 (59.8) | 188 (67.4) | 29 (76.3) | |
| Race, n (%) | | | | |
| American Indian or Alaska native | 1 (0.2) | 0 (0.0) | 1 (0.4) | |
| Asian | 36 (8.1) | 5 (3.9) | 26 (9.3) | |
| Black or African American | 56 (12.6) | 22 (17.3) | 30 (10.8) | 0.576 |
| White | 307 (69.1) | 87 (68.5) | 195 (69.9) | |
| Not reported | 44 (9.9) | 13 (10.2) | 27 (9.7) | |
| Ethnicity | | | | |
| Hispanic or Latino | 5 (3.9) | 7 (2.5) | 1 (2.6) | 0.680 |
| Not Hispanic or Latino | 98 (77.2) | 220 (78.9) | 27 (71.1) | |
| Not reported | 24 (18.9) | 52 (18.6) | 10 (26.3) | |
| Subtypes, n (%) | | | | |
| Basal-like | 101 (22.7) | 13 (10.2) | 81 (29.0) | 0.430 |
| HER2-enriched | 42 (9.5) | 3 (2.4) | 37 (13.3) | |
| Luminal A or normal like | 191 (43.0) | 74 (58.3) | 102 (36.6) | |
| Luminal B | 110 (24.8) | 37 (29.1) | 59 (21.1) | |
| Pathologic stage | | | | |
| Stage I | 23 (18.1) | 56 (20.1) | 10 (26.3) | |
| Stage II | 79 (62.2) | 174 (62.4) | 20 (52.6) | 0.759 |
| Stage III | 25 (19.7) | 49 (17.6) | 8 (21.1) | |
| Treatment | | | | |
| Radiation | 70 (55.1) | 152 (54.5) | 21 (55.3) | 0.990 |
| Pharmaceutical | 110 (86.6) | 226 (81.0) | 27 (71.1) | 0.081 |
| All-cause mortality | 4 (3.1) | 7 (2.5) | 5 (13.2) | 0.014 |

*p values were estimated using Fisher’s exact or the χ² test. HER2, human epidermal growth factor receptor 2; IDC, invasive ductal carcinoma.*
lating inhibitor of apoptosis proteins and has been indicated as an independent prognostic factor in breast cancer [53, 54]. *CHEK2* is reported to be associated with inherited *TP53* mutation, and also response in DNA damage which interacted with BRCA1 phosphorylation [55, 56]. Besides, *BARD1*, named *BRCA1*-associated RING domain protein 1, was indicated to be involved in the regulation of the recombinase RAD51 family [21]. The overexpression of RAD51 members, including *RAD51C* and *RAD51D* are involved in DNA homologous recombination and DNA damage response, which are also suggested to be associated with tumor progression [25]. *MLH1* and *MSH2* are mismatch repair genes involved in repairing DNA replication errors. Defects in *MLH1* can contribute to breast cancer progression [57].

Briefly, the candidate genes included in mortality risk estimation model are associated with estrogen signaling pathway (*CCND1*, SRC, *IGF1R*, and *ESR1*), mTOR signaling pathway (*HRAS* and *KRAS*), p53 signaling pathway (*BIRC5*, *CCND1*, and *CHEK2*), EMT pathway (*HRAS*, *KRAS*, *ESR1*, and *NOTCH1*), and DNA damage response (*MLH1*, *MSH2*, *BARD1*, *CHEK2*, *RAD51C*, and *RAD51D*). Some candidate genes were overlapped in multiple pathways and interact with others genes. According to the past studies, the candidate genes are generally involved in breast cancer progression. Some are related to drug resistance in specific subtypes. Disease progression and drug resistance are commonly associated with poor overall survival due to the uncontrollable nature of cancer cells [58–60]. Previously, several studies have demonstrated the gene expression in the RNA-seq level of IDC using machine learning approaches [61, 62] and indicated that IDC has differed gene mutation profiles compared to invasive lobular carcinoma. The current study used somatic mutation data in the DNA-seq level to identify mortality risk-associated gene mutation profiles. Although the identified candidate genes were not completely suitable with previous findings, a similar pathway associated with cancer progression for IDC was found, including DNA damage response and p53 signaling [61].

This study had several limitations. The retrospective nature of this study prevented the inclusion of certain covariates that might be associated with overall survival. Despite this, we have included information on the baseline demographic, tumor characteristics, molecular subtypes, and treatment characteristics used in most breast cancer studies. The Cox proportional hazard regression analysis for overall survival in patients with early-stage IDC breast cancer is shown in Table 4.

### Table 4. Cox proportional hazard regression analysis for overall survival in patients with early-stage IDC breast cancer

| Variables                           | Crude-HR (95% CI)         | p value | Adjusted-HR (95% CI) | p value |
|-------------------------------------|--------------------------|---------|----------------------|---------|
| Gene mutation risk subgroup         |                          |         |                      |         |
| Controls                            | Ref                      |         | Ref                  |         |
| Low-risk                            | 1.08 (0.31–3.77)         | 0.900   | 0.85 (0.21–3.40)     | 0.823   |
| High-risk                           | 7.22 (1.82–28.70)        | **0.005** | 6.57 (1.51–28.70)    | **0.012** |
| Age-group, years                    |                          |         |                      |         |
| <35                                 | Ref                      |         | Ref                  |         |
| 35–50                               | 0.20 (0.02–1.97)         | 0.166   | 0.29 (0.03–3.31)     | 0.321   |
| >50                                 | 0.54 (0.07–4.24)         | 0.560   | 0.57 (0.06–5.07)     | 0.612   |
| Subtypes                            |                          |         |                      |         |
| Luminal A or normal like            | Ref                      |         | Ref                  |         |
| Luminal B                           | 2.82 (0.77–10.34)        | 0.118   | 1.62 (0.40–6.57)     | 0.498   |
| HER2-enriched                       | 3.59 (0.83–15.60)        | 0.088   | 3.7 (0.70–19.50)     | 0.123   |
| Basal-like                          | 1.30 (0.31–5.50)         | 0.722   | 1.24 (0.26–5.91)     | 0.785   |
| Pathologic stage                    |                          |         |                      |         |
| Stage I                             | Ref                      |         | Ref                  |         |
| Stage II                            | 4.06 (0.52–31.82)        | 0.183   | 5.96 (0.65–55.00)    | 0.115   |
| Stage III                           | 5.56 (0.62–50.28)        | 0.127   | 9.14 (0.82–101.00)   | 0.072   |
| Treatment                           |                          |         |                      |         |
| Radiation versus without radiation  | 0.75 (0.28–2.03)         | 0.577   | 0.66 (0.22–1.97)     | 0.460   |
| Pharmaceutical versus without Pharma| 0.33 (0.10–1.07)         | 0.065   | 0.33 (0.10–1.14)     | 0.081   |

HER2, human epidermal growth factor receptor 2; HR, hazard ratio; CI, confidence interval; Ref, reference; IDC, invasive ductal carcinoma.
cancer studies. In addition, we included only patients with early-stage IDC breast cancer; therefore, the generalizability of our findings is limited to this subpopulation.

This study used integrated survival analysis to identify candidate genes in a group of patients with early-stage IDC breast cancer at high risk for poor survival outcomes compared to control and low-risk group and achieved further understanding of the effects of their gene mutation profiles on overall survival. The candidate genes are particularly involved in cell-cycle processes, DNA repair, and drug resistance, and their mutations are generally associated with disease progression or therapeutic resistance, which are commonly associated with poor survival outcomes. Current findings conducted by a simple integrated survival analysis enable a quick way to find higher mortality risk-associated candidate genes for IDC of the breast. Future studies should investigate the integrative functionality of candidate genes in a different genomic level. Moreover, the combined use of machine learning algorithms should be the focus of further research on the nonlinear or high-order interaction effects between clinicopathologic variables and the involved genes.

Acknowledgments

All data were obtained from the Pan-Cancer Atlas of TCGA research network.

References

1. Goh CW, Wu J, Ding S, Lin C, Chen X, Huang O, et al. Exome sequencing identifies tumor suppressor genes unique to breast cancer patients. Genome Med. 2020;12(1):81.
2. Kelemen G, Farkas V, Debrah J, Ormandi K, Voros A, Kaizer L, et al. The relationship of Lon gene expression to breast cancer survival: a meta-analysis. Oncotarget. 2017;8:13792.
3. Yang SX, Polley EC. Systemic treatment and radiotherapy, breast cancer subtypes, and survival after long-term clinical follow-up. Breast Cancer Res Treat. 2019;175:287–95.
4. Xiao Q, Zhou Y, Winter S, Büttner F, Schaef feler E, Schwab M, et al. Germline variant burden in multigene resistance transporters is a therapy-specific predictor of survival in breast cancer patients. Int J Cancer. 2020;146:2475–87.
5. Ding S, Wu J, Lin C, Chen W, Li Y, Shen K, et al. Predictors for survival and distribution of 21-gene recurrence score in patients with pure mucinous breast cancer: a SEER population-based retrospective analysis. Clin Breast Cancer. 2019;19:e66–73.
6. Christgen M, Bartels S, Radner M, Raap M, Rieger L, Christgen H, et al. ERBB2 mutation frequency in lobular breast cancer with pleomorphic histology or high-risk characteristics by molecular expression profiling. Genes Chromosomes Cancer. 2019;58:175–85.
7. Figg WD II, Reid J. Monitor tumor burden with circulating tumor DNA. Cancer Biol Ther. 2013;14:697–8.
8. Al-Kuraya K, Schraml P, Torhorst J, Tapia C, Zaharieva B, Novotny H, et al. Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Res. 2004;64:8534–40.
9. Jovanovic J, Ronneberg JA, Tost J, Kristensen V. The epigenetics of breast cancer. Mol Oncol. 2010;4:242–54.
10. Carcangiu ML, Casalini P, Ménard S. Breast tumors: an overview. Atlas Genet Cyto genet Oncol Haematol. 2005.
11. Kirkegaard T, Nielsen KV, Jensen LB, Campbell FM, Mü ller S, Tovey SM, et al. Genetic alterations of CCND1 and EMSY in breast cancers. Histopathology. 2008;52:698–705.
12. Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families with high risk of breast cancer. JAMA. 2006;295:1379–88.
13. Veggan F, Boidot R, Oudin C, Defrain C, Re bucki M, Lizard-Naloc S. Association of p53 gene alterations with the expression of anti-apoptotic survivin splice variants in breast cancer. Oncogene. 2007;26:290–7.
14. Rosen EM, Fan S, Pestell RG, Goldberg ID. BRCA1 gene in breast cancer. J Cell Physiol. 2003;196:19–41.
15. Thompson D, Szabo CI, Mangion J, Oldenburg RA, Odefrey F, Seal S, et al. Evaluation of linkage age of breast cancer to the putative BRCA3 locus on chromosome 13q21 in 128 multiple case families from the breast cancer linkage consortium. Proc Natl Acad Sci USA. 2002;99:827–31.
16. Jelicic P, Eccles LA, Tseng J, Cybulski P, Wielgos M, Powell SN, et al. The EMSY threonine 207 phospho-site is required for EMSY-driven suppression of DNA damage repair. Oncotarget. 2017;8:13792.

Statement of Ethics

The data use certification agreement of TCGA allowing for exemption from further approval by the Ethics Committee (https://www.cancer.gov/about-nci/organization/ccg/research/structuralgenomics/tcga/history/policies).

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This work was supported by grants from the E-Da Cancer Hospital (EDCHT110005).

Author Contributions

C.-C.H. and S.-H.M. drafted the manuscript. C.-C.H., H.-I.H., and C.-M.H. were responsible for data processing and interpretation. S.-H.M. performed statistical analyses. S.-H.M. conceived and designed the study. S.-H.M. and C.-M.H. fully revised the final manuscript. All the authors read and approved the final manuscript.

Data Availability Statement

All data can be downloaded from the cBioPortal (https://www.cbioportal.org/datasets).
Identification of High-Risk Mortality Genes in IDC

Li Z, Pearlman AH, Hsieh P. DNA mismatch repair and the DNA damage response. DNA Repair. 2016;38:2–10.

Lin DP, Wang Y, Scherer SJ, Clark AB, Yang K, Avdievich E, et al. An Muh2 point mutation uncouples DNA mismatch repair and apoptosis. Cancer Res. 2004;64:517–22.

Peltomäki P. DNA mismatch repair and cancer. Mutat Res. 2001;488(1):77–85.

Stojic I, Brun R, Jiricny J. Mismatch repair and DNA damage signalling. DNA Repair. 2004;3:1091–101.

Schulz E, Valentin A, Utkan NZ, et al. Association of the nibrin gene (NBN) variants with breast cancer. Biomed Rep. 2016;4:369–73.

Zhang F, Ma J, Wu J, Ye L, Cai H, Xia B, et al. PALB2 links BRCA1 and BRCA2 in the DNA damage response. Curr Biol. 2009;19:524–9.

Polak P, Kim J, Braunstein LZ, Karlic R, Ha-radhavala NJ, Tiao G, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat Genet. 2017;49:1476.

Kretova M, Sabova L, Hodny Z, Bartek J, Kolarovic G, Nelson BD, et al. TGF-β/NF1/NF2-Smad-mediated suppression of ATM contributes to oxidative stress in cellular senescence. Cell Signal. 2014;26:2903–11.

Zhuang T, Yu S, Zhang L, Yang H, Li X, Hou Y, et al. Sharpin stabilizes estrogen receptor α and promotes breast cancer cell proliferation. Oncotarget. 2017;8:77137.

Caldon CE. Estrogen signaling and the DNA damage response in hormone dependent breast cancers. Front Oncol. 2014;4:106.

Dammann R, Schagdarsurengin U, Seidel C, Struzik ZV, Rastetter M, Baier K, et al. The tumor suppressor RASSF1A in human carcinogenesis: an update. Histol Histopathol. 2005.

Weigelt B, Kreike B, Reis-Filho JS. Metastatic breast carcinomas are basal-like breast cancers: a genomic profiling analysis. Breast Cancer Res. 2010;12:273–80.

Olayioye MA. Update on HER-2 as a target for cancer therapy: intracellular signaling pathways of ErbB2/HER-2 and family members. Breast Cancer Res. 2001;3:385–9.

Campeau PM, Foulkes WD, Tischkowitz MD. Hereditary breast cancer: new genetic developments, new therapeutic avenues. Hum Genet. 2008;124:31–42.

Danczy J. miTOR signalling and drug development in cancer. Nat Rev Clin Oncol. 2010;7:209.

Guerrero-Zotano A, Mayer IA, Arteaga CL. PI3K/AKT/mTOR: role in breast cancer progression, drug resistance, and treatment. Cancer Metastasis Rev. 2016;35:515–24.

Wang Y, Zhou BP. Epithelial-mesenchymal transition in breast cancer progression and metastasis. Chin J Cancer. 2011;30:603.

Foroni C, Broggiini M, Generali D, Damia G. Epithelial – mesenchymal transition and breast cancer: role, molecular mechanisms and clinical impact. Cancer Treat Rev. 2012;38:689–97.

Driscoll JJ, Rixe O. Overall survival: still the gold standard? It is possible that loss of wild-type p16RASSF1A survival remains the definitive end point in cancer clinical trials. Cancer J. 2009;15:401–5.

Motakis E, Ivshina AV, Kuznetsov VA. Data-driven approach to predict survival of cancer patients: estimation of microarray genes’ prediction significance by Cox proportional hazard regression model. IEEE Eng Med Biol Mag. 2009;28:58–66.

Zaré R, Makaroff C, Mohammad K, Zer-aati H, Hosseinie M, Holakouie Naieni K. A comparison between Kaplan-Meier and weighted Kaplan-Meier methods of five-year survival estimation of patients with gastric cancer. Acta Med Iran. 2014;52:764–7.

Jensen LB, Bartlett JM, Witton CJ, Kirkgaard T, Brown S, Müller S, et al. Frequent amplifications and deletions of G1/S-phase transition genes, CCND1 and MYC in early breast cancers: a potential role in G1/S escape. Cancer Biomark. 2009;5:41–9.

Brown LA, Johnson K, Leung S, Bismar TA, Benitez J, Foulkes WD, et al. Co-amplification of CCND1 and EMSY is associated with an adverse outcome in ER-positive tamoxifen-treated breast cancers. Breast Cancer Res Treat. 2010;121:347–54.

Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE Jr, Davidson NE, et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. N Engl J Med. 2005;353:1673–84.

Zhang S, Huang WC, Li P, Guo H, Poh SB, Zhang S, Huang WC, Li P, Guo H, Poh SB, et al. Notch signaling pathway and cancer metastasis. Adv Exp Med Biol. 2012;727:186–98.

Roeb W, Higgins J, King MC. Response to DNA damage of CHEK2 missense mutations in familial breast cancer. Hum Mol Genet. 2012;21:2738–44.

Nevanlinna H, Bartek J. The CHEK2 gene and inherited breast cancer susceptibility. Oncogene. 2006;25:3912–9.

Malik SS, Zia A, Mukarik S, Massod N, Rashid S, Sherrard A, et al. Correlation of MLH1 polymorphisms, survival statistics, in silico assessment and gene downregulation with clinical outcomes among breast cancer cases. Mol Biol Rep. 2020;47:683–92.

Lim HS, Sun W, Parivar K, Wang D. Predicting overall survival and progression-free survival using tumor dynamics in advanced breast cancer patients. AAPS J. 2019;21:22.

Minchin RF, Butler NJ. Tridimensional doubling of arylnitrone N-acetyltansferase 1 mRNA in breast cancer tumors: association with overall survival and drug resistance. BMC Genomics. 2018;19:513.

Liu Y, Yu C, Wu Y, Sun X, Su Q, You C, et al. CD44+ fibroblasts increases breast cancer cell survival and drug resistance via IGFBP3-CD44-IGF1 receptor signalling. J Cell Mol Med. 2017;21:1979–88.

Roy S, Kumar R, Mittal V, Gupta D. Classification models for invasive ductal carcinoma progression, based on gene expression data-trained supervised machine learning. Sci Rep. 2020;10:4113–5.

Zhao H, Langerød A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, et al. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. Mol Biol Cell. 2004;15:2523–36.