Original article

Comprehensive analysis of gene expression profiles to identify differential prognostic factors of primary and metastatic breast cancer

Sarah Albogami
Department of Biotechnology, College of Science, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

Article info

Keywords:
Breast cancer
Metastasis
Prognostic marker
Overall survival
Differentially expressed genes
Protein-protein interaction

Abstract

Breast cancer accounts for nearly half of all cancer-related deaths in women worldwide. However, the molecular mechanisms that lead to tumour development and progression remain poorly understood and there is a need to identify candidate genes associated with primary and metastatic breast cancer progression and prognosis. In this study, candidate genes associated with prognosis of primary and metastatic breast cancer were explored through a novel bioinformatics approach. Primary and metastatic breast cancer tissues and adjacent normal breast tissues were evaluated to identify biomarkers characteristic of primary and metastatic breast cancer. The Cancer Genome Atlas-breast invasive carcinoma (TCGA-BRCA) dataset (ID: HS-01619) was downloaded using the mRNASeq platform. Genevestigator 8.3.2 was used to analyse TCGA-BRCA gene expression profiles between the sample groups and identify the differentially-expressed genes (DEGs). Networks of protein–protein interactions were constructed to identify the top hub genes with the highest degree of interaction. Additionally, the top hub genes were validated based on overall survival and immunohistochemistry using The Human Protein Atlas. Of the top 20 hub genes identified, four (KRT14, KIT, RAD51, and TTK) were considered as prognostic risk factors based on overall survival. KRT14 and KIT expression levels were upregulated while those of RAD51 and TTK were downregulated in patients with breast cancer. The four proposed candidate hub genes might aid in further understanding the molecular changes that distinguish primary breast tumours from metastatic tumours as well as help in developing novel therapeutics. Furthermore, they may serve as effective prognostic risk markers based on the strong correlation between their expression and patient overall survival.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Breast cancer (BC), which is the most frequently diagnosed cancer type in women, could be treatable if detected in its early stages (Andreopoulou and Hortobagyi, 2008, Jemal et al., 2009, Ozkan, 2022, Masood, 2005, Albogami et al., 2021). Based on its stage, the prognosis of BC is usually classed from I to IV, with various sub-stages (Cianfrocca and Goldstein, 2004). BC is considered malignant when it progresses from stage I and II to advanced disease (stage III), and subsequently metastasizes to the lungs, bone, brain, and other organs (stage IV) (Linde et al., 2018). Numerous aspects of BC can aid in predicting the risk of recurrence and the likelihood of survival following surgery (Maitra and Srivastava, 2022). The traditional prognostic factors of BC include positive axillary nodes, tumour size, histology, grade, lymphovascular invasion, the expression of the hormone receptor, and HER2 status (Ly et al., 2012). These factors have been discussed in detail, with an emphasis on classification issues (Eliyatkın et al., 2015, Kim et al., 2012, Sauerbrei et al., 1999). Modern molecular prognostic markers might be able to provide information beyond that by traditional prognostic factors and this was achieved by a better comprehension of BC pathophysiology, which resulted in the...
development of novel molecular markers, enhanced risk evaluation, improved therapies, and personalised treatment (Faramarzi et al., 2021, Ciliberti et al., 2021, Zhou et al., 2022). Genetic profiling can yield predictive and prognostic gene expression signatures, which might aid in the classification of cancers and provide new targeted therapy (Loughman et al., 2022). The focus on new prognostic indicators arises from the belief that a considerable proportion of patients with BC in the early stages are not diagnosed by accurate methods, usually using microscopic metastases (Rogers et al., 2002).

To date, only a few effective prognostic and predictive indicators have been used clinically to manage patients with BC (Kumar et al., 2012). Although numerous techniques have been developed over the years, none have proven to be clinically useful. With the improvement in next-generation sequencing techniques, RNA sequencing (RNA-Seq) technology has emerged as a novel, straightforward, and helpful tool for assessing the content of cDNA transcripts and analysing differential gene expression utilizing high-throughput sequencing (Negi et al., 2022). RNA-Seq technology has demonstrated a high sensitivity for biomarker discovery, while increasing the resolution and decreasing the associated time and cost (Stark et al., 2019, Ergin et al., 2022, Hwang et al., 2018). Thus, cancer-associated genes and the associated signalling pathways at the genome level can be identified using RNA-Seq and other high-throughput sequencing technologies (Ren et al., 2012, Zhang et al., 2019a).

RNA-Seq technology is not frequently used by molecular diagnostic laboratories for clinical testing, monitoring, or management of patients with BC (Roychowdhury and Chinnaiyan, 2016). Owing to this, a sizable fraction of potentially diagnosable cases could be unresolved at the moment (Marco-Puche et al., 2019). Certain authors have illustrated the value of RNA sequencing in diagnosing 10% of patients and identifying potential genes for the other 90% (Kremer et al., 2017). However, the application of RNA-Seq in clinical settings is severely limited owing to the generation of large datasets that require skilled analysis, the inability to maintain independence, poor validation, lack of methodological standardisation, and the complexity of RNA-Seq data outcomes. (Salto-Tellez and Gonzalez de Castro, 2014, Lightbody et al., 2019, Nazarov et al., 2017). Therefore, further studies are required to ensure that RNA-Seq is consistently and reliably performed in clinics. Although RNA-Seq has been used for cancer transcriptome profiling, limited studies have been conducted to identify differences in the transcriptomes of normal breast, primary BC, and metastatic BC tissues using RNA-Seq. Hence, a technique that is applicable to routine clinical practice must be developed based on differential gene profiling. This is especially critical for the identification of high-risk cancer cells that may undergo metastases (Feng et al., 2007).

Therefore, by gene expression profiling and comprehensive bioinformatics analysis of The Cancer Genome Atlas-breast invasive carcinoma (TCGA-BRCA) dataset, this study aimed to compare the primary and metastatic BC tissues to normal breast tissue to screen for the top 20 hub genes, annotate differentially-expressed gene (DEG) profiles, integrate the gene expression profiles with clinical data, and finally identify biomarkers involved in primary and metastatic BC.

2. Materials and methods

2.1. BC transcriptome and clinical profile acquisition and workflow

Human BC transcriptome and clinical profiles [Project title: TCGA-BRCA; Experimenter: NIH national cancer institute (Koboldt et al., 2012, Ciriello et al., 2015); Experiment ID: HS-01619; Experimental technology platform: HS_mRNASeq_HUMAN_GL: mRNA-Seq Gene Level Homo sapiens (ref: Ensembl 97, GRCh38.p12); Study design: DEGs between primary IDC and tumour adjacent tissue, and DEGs between metastatic IDC and tumour adjacent tissue] were analysed and downloaded from Genexpressionist B.1.0 (Zurich, Switzerland) in May 2021. Supplementary Table S1 contains the demographic information for the datasets, and Fig. 1 depicts the study workflow.

2.2. Identification of DEGs between groups

DEGs were analysed in two groups: primary tumour tissue samples obtained from patients with infiltrating duct carcinoma compared with normal breast tissue samples (P vs. N) and metastatic tumour tissue samples obtained from axillary lymph nodes of patients with primary infiltrating duct carcinoma of the breast compared with normal breast tissue samples (M vs. N). DEGs were considered upregulated and downregulated if |Log2 fold change| ≥ 2 and |Log2 fold change| > -2, respectively, and p-value < 0.05 were found between sample groups, in accordance with previously specified criteria (Ward et al., 2013, Luo et al., 2017). Genevestigator 8.3.2 was used to analyse TCGA-BRCA gene expression profiles between sample groups.

2.3. Protein–protein interaction (PPI) network construction, hub gene identification, and module analysis

PPI networks were constructed between DEGs in the P vs. N and M vs. N groups using the search tool for the retrieval of interacting genes (STRING) database (Jensen et al., 2009). The STRING database was used to obtain statistical data for each PPI network. The Network Analyser tool of the Cytoscape software (version: 3.8.2) was employed to visualise as well as analyse the PPI networks to determine the interactional correlations among DEGs (Shannon et al., 2003). The top 20 hub genes from each PPI network group were chosen due to their high degree of connectivity (≥10). Additionally, the Cytoscape plugin molecular complex detection (MCODE) was utilized to identify clusters of genes that were highly connected across the entire PPI network.

2.4. Functional enrichment analysis

The inputs of PPI networks and networks for the top twenty candidate hub genes for both P vs. N and M vs. N groups were applied to find the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and gene ontology (GO) terms. The ClueGo and Cluepedia Cytoscape plugins were used for GO enrichment analyses (Bindea et al., 2013). GO annotations describe the relationship that exists between gene expression and molecular function (MF), cellular component (CC), and biological process (BP).

2.5. Validation of hub genes

Gene expression profiling interactive analysis (GEPIA), a web server, was applied to calculate the Kaplan–Meier curve by correlating gene expression with time and the percent of overall survival (OS) for the top 20 hub gene candidates identified in both P vs. N and M vs. N groups to determine whether they have an effect on survival of patients with BC (Tang et al., 2017). Significant differences were defined as those with a p value of < 0.05, and the median was employed as the group cut-off; high and low cutoffs were 50%. The hazard ratio (HR) with a 95% confidence interval was determined using the Cox PH model, and the BRCA dataset was used as the selection dataset. The Human Protein Atlas (www.proteinatlas.org) database was used to obtain immunohistochemistry
results for the top 20 hub genes in each group; this allowed us to compare the dysregulation of specific hub genes across groups.

2.6. Statistical analysis

GraphPad Prism ver. 9.1.2 (San Diego, CA, USA) was utilized to create a volcano plot to reveal upregulated and downregulated DEGs. The data are expressed as the mean ± SD. The fold-change (FC) was determined using the log2 value of the signal intensity; FC > 1.5 was deemed significant. At a p < 0.05, the differences in gene expression between the datasets were declared significant. Venn Painter 1.2.0 (Toronto, Canada) was used to illustrate the overlap in DEGs between sample groups (Lin et al., 2016). DEGs were mapped to chromosomes using Circa (CA, USA). The RStudio, PBC software (version: 1.3.1056) was used to identify the KEGG pathways in both groups for the complete set of DEGs and the top twenty hub genes. The ggplot2 and tidyr packages in R were used to visualise the KEGG enrichment analysis results. The inputs for KEGG enrichment analysis were generated from the DAVID bioinformatics resource 6.8 online database (Sherman and Lempicki, 2009, Huang et al., 2009).

3. Results

3.1. Analysis revealed 528 and 955 DEGs in P vs. N and M vs. N groups, respectively

In total, 528 DEGs (134 upregulated and 394 downregulated) were identified between primary tumour tissues and normal breast tissues (P vs. N group) (Fig. 2a, Supplementary Table S2), and 955 DEGs (422 upregulated and 533 downregulated) were identified between metastatic tumour tissue and normal breast tis-

---

Fig. 1. Flowchart depicting the study design and steps involved in the processing, analysis, and validation of data.
sue from patients with BC (M vs. N group) (Fig. 2b, Supplementary Table S2). DEGs in both groups were localised on chromosomes (Fig. 2c, d), and a Venn diagram revealed 99 overlapping upregulated and 201 overlapping downregulated DEGs between the P vs. N and M vs. N groups (Fig. 3a, b).

Among the genes that were not common between the groups, 35 genes were upregulated and 193 genes were downregulated only in the P vs. N group (Fig. 3a, b), while 323 genes were upregulated and 332 were downregulated only in the M vs. N group (Fig. 3a, b). The top 10 unique DEGs (upregulated and downregulated) that did not overlap between the two groups are summarised in Supplementary Tables S2 (P vs. N) and S3 (M vs. N). Supplementary Tables S4 and S5 present the results of the Venn diagram for all DEGs.

3.2. PPI network construction and analysis revealed genes potentially involved in early metastasis

The STRING database was employed to construct the PPI network and analyse its features. The P vs. N PPI network consisted of 219 nodes connecting 679 edges, with an average neighbour count of 7.564, average clustering coefficient of 0.387, network heterogeneity of 1.122, 39 connected components, and a network centralisation of 0.221 (Fig. 4a). In contrast, the M vs. N PPI network consisted of 623 nodes connecting 2120 edges, with an average neighbour count of 7.895, average clustering coefficient of 0.248, network heterogeneity of 1.304, 81 connected components, and network centralisation of 0.096 (Fig. 4b).

---

**Fig. 2.** Differentially expressed genes (DEGs) between P vs. N and M vs. N groups. (a, b) Volcano plots of DEGs, between a) P vs. N, and b) M vs. N groups. Volcano plots present upregulated expression with purple colour, and downregulated expression with blue colour. (c, d) CIRCOS circular plot of DEGs in the genome. (c) Upregulated and downregulated DEGs in P vs. N group. (b, d) Upregulated and downregulated DEGs in M vs. N group. CIRCOS circular plot presents the chromosome number with track 1, DEGs with upregulated expression is shown in purple colour with track 2, and DEGs with downregulated expression is shown with track 3. DEGs were considered upregulated and downregulated if a $|\text{Log}_2 \text{fold change}| \geq 2$ and $|\text{Log}_2 \text{fold change}| \geq -2$, respectively, and p-value < 0.05. P vs. N, primary tumour tissue samples obtained from patients with infiltrating duct carcinoma compared with normal breast tissue samples; M vs. N, metastatic tumour tissue samples obtained from axillary lymph nodes of patients with primary infiltrating duct carcinoma of the breast compared with normal breast tissue samples.
With degree ≥ 10 as the cut-off criterion, 47 and 112 PPI network hub genes were identified among the P vs. N (Supplementary Table S6) and M vs. N (Supplementary Table S7) DEGs, respectively. The top 20 genes in P vs. N and M vs. N groups are listed in Tables 1 and 2, respectively. These hub genes may be involved in early metastasis.

When the cut-off criterion of MCODE was defined as ≥ 5, only one cluster was identified from the P vs. N DEG-PPI network (Fig. 5a), comprising 21 nodes and 95 edges, and 12 seed genes that were common with the top 20 hub genes (DGAT2, FABP4, PCK1, ADIPOQ, COL1A1, SPP1, LPL, LIPE, LEP, FOS, FN1, and PLIN1). Furthermore, five clusters were identified in the M vs. N DEG-PPI network (Fig. 5b–f), and the most significant cluster (score 29.429) consisted of 36 nodes and 515 edges, with 15 seed genes shared with the top 20 hub genes (TTK, TYMS, OIP5, MAD2L1, RACGAP1, NDC80, RAD51, MCM10, KIF15, KIF11, CDC25C, ESPL1, CCNA2, PLK4, and RAD51AP1; Fig. 5b). The second cluster with a score of 8.000 consisted of 12 nodes and 44 edges and contained three seed genes shared with the top 20 hub genes (CCND1, MYC, and BMP4; Fig. 5c). The third (11 nodes and 25 edges), fourth (15 nodes and 35 edges), and fifth (5 nodes and 10 edges) clusters with a score of 5.000 each did not share any seed genes with the top 20 hub genes.
Fig. 4. A diagrammatic representation of the protein–protein interaction (PPI) network. (a) P vs. N PPI network. (b) M vs. N PPI network. The Cytoscape software was used to create the figures. The STRING database was used to obtain statistical data for each PPI network. The Network Analyser tool of the Cytoscape software with indirect parameters was employed to visualise as well as analyse the PPI networks to determine the interactional correlations of DEGs. P vs. N, primary tumour tissue samples obtained from patients with infiltrating duct carcinoma compared with normal breast tissue samples; M vs. N, metastatic tumour tissue samples obtained from axillary lymph nodes of patients with primary infiltrating duct carcinoma of the breast compared with normal breast tissue samples.
functional annotation and KEGG pathways of the top twenty hub genes in both P vs. N and M vs. N analyses. The top 10 GO terms and KEGG pathways for the PPI networks of P vs. N and M vs. N groups are shown in Supplementary Tables S8–S11 and Figures S2–S6.

### 3.3. Functional enrichment analysis identified pathways associated with BC metastasis

Figs. 6–8 illustrate the top significant terms from the GO functional annotation and KEGG pathways of the top twenty hub genes in both P vs. N and M vs. N analyses. The top 10 GO terms and KEGG pathways for the PPI networks of P vs. N and M vs. N groups are shown in Supplementary Tables S8–S11 and Figures S2–S6.

### 3.3.1. GO functional annotation of the top twenty hub genes in the P vs. N PPI group

MF ontology analysis (Fig. 6a–c, Tables S9) revealed that 71.43% of the terms had identical protein-binding groups. Furthermore,
14.29% of the identified terms were in the protein-containing complex binding group and 14.29% of the terms in the protein-binding group. In CC ontology analysis (Fig. 6d–f, Tables S9), 66.67% of the identified terms were in the extracellular space group. In addition, 22.22% of the identified terms were in the endoplasmic reticulum group and 11.11% of the terms in the cytoplasm group. In BP ontology analysis (Fig. 6g–I, Tables S9), 54.55% of the terms were found in the triglyceride metabolic process group and 45.45% of the terms in the positive regulation of chemokine production group.

3.3.2. GO functional annotation of the top 20 hub genes in the M vs. N PPI group

In the MF ontology analysis (Fig. 7a–c, Tables S10), 28.57% of the terms were found in the anion binding group. In total, 28.57% of the terms were associated with the activity of a transferase responsible for the transfer of phosphorus-containing compounds. In the binding group, 14.29% of the terms were identical. Furthermore, 14.29% of the terms were found in the enzyme-binding group and 14.29% of the terms were in the protein-binding group. In the CC ontology analysis (Fig. 7d–f, Tables S10), 28.57% of the terms were determined to be enriched in the intracellular organelle lumen group. Moreover, 28.57% of the terms were identified in the chromosome group, 14.29% of the terms were identified in the intracellular organelle group, and 14.29% were identified in the microtubule cytoskeleton group. An additional 14.29% of the terms were identified in the cytosol group. BP ontology analysis (Fig. 7g–i, Tables S10) identified 53.12% of the terms enriched in mitotic cell cycle phase transition. In addition, 28.12% of the genes were associated with negative regulation of the cell cycle. Moreover, 15.62% of the terms were enriched in regulation of cell cycle. The remaining 3.12% of the terms were associated with cell cycle.

3.3.3. Pathway enrichment of the top twenty hub genes in the P vs. N PPI group

The top 10 KEGG pathways for the top twenty hub genes in the P vs. N PPI group (Fig. 8a) were the PPAR signalling pathway (five genes), AMPK signalling pathway (five genes), adipocytokine signalling pathway (four genes), PI3K-AKT signalling pathway (six genes), regulation of lipolysis in adipocytes (three genes), pathways in cancer (five genes), CM-receptor interaction (three genes), amoebiasis (three genes), Toll-like receptor signalling pathway (three genes), and TNF signalling pathway (three genes).
3.3.4. Pathway enrichment of the top 20 hub genes in the P vs. N PPI group

The top KEGG pathways for the top 20 hub genes in the M vs. N PPI group (Fig. 8b) included cell cycle (seven genes), thyroid hormone signalling pathway (four genes), progesterone-mediated oocyte maturation (three genes), microRNAs in cancer (three genes), oocyte meiosis (three genes), hepatitis B (three genes), Hippo signalling pathway (three genes), pathways in cancer (four genes), proteoglycans in cancer (three genes), thyroid cancer (two genes), bladder cancer (two genes), HTLV-I infection (three genes).

Fig. 6. Gene ontology (GO) term analysis of the top 20 hub genes in P vs. N group. (a–c) Molecular function category (MF). (d–f) Cellular component category (CC). (g–i) Biological processes category (BP). (a, d, g) Networks for GO terms that were unique to the top 20 hub genes in P vs. N group for each GO category. (b, e, h) Pie chart with % terms per group, including particular terms for the top 20 hub genes in each GO category from P vs. N group. (c, f, i) Bars indicate the % of genes per term associated with each GO category from P vs. N group. Figures were generated using the ClueGO and CluePedia plugins in Cytoscape. P vs. N, primary tumour tissue samples obtained from patients with infiltrating duct carcinoma compared with normal breast tissue samples; M vs. N, metastatic tumour tissue samples obtained from axillary lymph nodes of patients with primary infiltrating duct carcinoma of the breast compared with normal breast tissue samples.
Fig. 7. Gene ontology (GO) term analysis of the top 20 hub genes in M vs. N group. (a–c) Molecular function category (MF). (d–f) Cellular component category (CC). (g–i) Biological processes category (BP). (a, d, g) Networks for GO terms that were unique to the top 20 hub genes in M vs. N group for each GO category. (b, e, h) Pie chart with % terms per group, including particular terms for the top 20 hub genes in each GO category from M vs. N group. (c, f, i) Bars indicate the % of genes per term associated with each GO category from M vs. N group. Figures were generated using the ClueGO and CluePedia plugins in Cytoscape. P vs. N, primary tumour tissue samples obtained from patients with infiltrating duct carcinoma compared with normal breast tissue samples; M vs. N, metastatic tumour tissue samples obtained from axillary lymph nodes of patients with primary infiltrating duct carcinoma of the breast compared with normal breast tissue samples.
3. Correlation analysis revealed candidate hub genes associated with overall survival in BC

Next, the correlations among the expression of the top twenty candidate hub genes and the OS of patients with BC in both P vs. M and M vs. N groups were evaluated using Kaplan–Meier curves to ascertain the hub genes’ prognostic significance (Figs. 9 and 10).

3.4. Candidate hub genes associated with OS in P vs. N group

The GEPIA results (Fig. 9) indicated that the upregulated expression of FN1, SPP1, COL1A1, and KRT19 and the downregulated expression of ALB, IL6, FOS, EGR1, PCK1, KRT14, KIT, and DGAT2 markers were correlated with poor OS, which is in accordance with
the log-fold change results generated using the mRNASeq platform in this study. Unexpectedly, the GEPIA results revealed higher expression of LEP, ADIPOQ, LPL, LIPE, FABP4, PLIN1, and SLC2A4 and lower expression of MMP9 than the mRNASeq results, which might correlate with poor OS. The criteria for identifying prognostic risk factors included: log-rank p < 0.05 and an HR score > 1 in upregulated genes and < 1 in downregulated genes (Xu et al., 2020). Based on log-rank p and HR scores, only two genes from the candidate hub genes in the P vs. N group were considered prognostic risk factors; these were KRT14 (HR = 0.63, log-rank p = 0.0049) and KIT (HR = 0.67, log-rank p = 0.016), and the expression of both was downregulated.

3.4.2. No significant candidate hub genes were associated with OS in M vs. N group

The GEPIA findings (Fig. 10) indicated that the upregulated expression of CCND1, CNA2, MAD2L1, ESR1, RAD51, TTK, ESP11, RAD51AP1, CDC25C, MCM10, and TMY5 expression and the downregulated expression of MYC might be correlated with poor OS; these results were supported by mRNA-Seq-generated log-
transformed results. Unexpectedly, GEPIA results showed higher BMP4 expression and lower KIF11, EZH2, RACGAP1, KIF15, NDC80, PLK4, BMP4, and OIP5 expression than the mRNA-Seq results, which correlated with poor OS. Based on the log-rank p and HR scores, these nine candidate hub genes in the M vs. N group could be considered as prognostic risk factors. However, none of these genes were statistically significant, with only two of them approaching significance: RAD51 (HR = 1.3, log-rank p = 0.075) and TTK (HR = 1.3, log-rank p = 0.087); the expression of both was upregulated. Thus, no prognostic risk factors were clearly identified in this study.

3.5. Validation of mRNA and protein expression of the identified potential prognostic risk factors

The Human Protein Atlas database was utilized to examine the protein expression levels of the four hub genes that were considered as potentially prognostic risk factors based on OS. Notably, KRT14 (Fig. 11a) and KIT (Fig. 11b) were not expressed in BC tissues, whereas moderate KRT14 levels and high KIT levels were observed in normal breast tissues, implying that hub gene expression was downregulated at the transcriptional and translational levels in patients with BC. Additionally, RAD51 (Fig. 11c) and TTK
[38x270]were not detected in normal breast tissues but were detected at moderate levels in BC tissues, indicating that the transcript and protein levels of both hub genes were increased in patients with BC.

4. Discussion

Patients with BC who have axillary lymph node metastases have a greater chance of early relapse and a shorter OS compared to those with BC with primary non-metastatic tumours (Panel, 2001). The mechanisms underlying distant metastases are complex and involve a diverse array of genes and signalling pathways (Pachmayr et al., 2017, Nathanson, 2003). Hence, the identification of predictive biomarkers that determine the extent of cancer spread and subsequent development of novel therapeutics could be highly effective in eradicating metastatic cancer. Determination of such biomarkers that can identify the transition point from the initial stage to metastasis would require high-throughput genome-wide association analysis. Numerous research have attempted to explain the molecular pathways behind the transformation from primary cancer to metastasis by comparing gene expression profiles of primary and healthy tissues as well as of primary and metastatic tissues; however, these studies focused exclusively on the changes in the tumour microenvironment and differences in gene expression and proteins between these groups (Ikemura et al., 2017, Hao et al., 2004, Mimori et al., 2005, Feng et al., 2007, Vecchi et al., 2008, Ellsworth et al., 2009).

This study revealed many DEGs common to both primary and metastatic tumours as well as several DEGs that were characteristic of each type. Additionally, the top 20 candidate hub genes uniquely expressed in tissues of primary and lymph node metastatic BC were identified to determine the molecular changes between them. Among these top 20 candidate hub genes, KRT14 and KIT were identified in primary BC tissues in the P vs. N group, whereas RAD51 and TTK were detected in lymph node metastasis BC tissues in the M vs. N group. These four genes were considered to be prognostic risk factors based on the significant association between their expression and OS. The transcriptional and translational levels of KRT14 and KIT were higher, whereas those of RAD51 and TTK were lower, in patients with BC.

KRT14 is a filamentous intermediate protein that is predominantly produced by epithelial progenitor cells (Chu et al., 2001). This cell population exhibits migratory behaviour, replication, and differentiation, and controls branching morphogenesis throughout organ development (Papaioyiou et al., 2016, Abashev et al., 2017, Rock et al., 2009). The increased expression of KRT14 in primary BC tissues (P vs. N) is comparable to that observed in

![Image]
a previous study, wherein a > 100-fold higher expression of KRT14 was found in RNA isolated from 20 patients with primary BC (Ellsworth et al., 2009). Similarly, Hanley et al. found a significant correlation between an extracellular matrix organisation module mediated by stromal cells and KRT14 expression. KRT14 expression and the basal transcription factor TP63 are upregulated in cancer cells with genetic autophagy inhibition. KRT14–TP63 expression has been found to be highly pro-metastatic in human and mouse mammary tumours (Marsh and Debnath, 2020). Similarly, an in vitro model of epithelial ovarian cancer metastasis and imaging mass spectrometry revealed KRT14 expression to be key for identifying the invasive potential of ovarian cancer cells (Bilandzic et al., 2019). Furthermore, although KRT14 is required for invasion, it has no effect on the viability or proliferation of cells (Bilandzic et al., 2019). KRT14 is likely to perform a distinctive role during invasion, additionally to maintaining cytoskeletal stability and integrity (Montor et al., 2018, Karimnia et al., 2021, Pearson, 2019, Werner et al., 2020). KRT14 is expressed in numerous types of tumours, particularly in invasive tumour cells as opposed to non-invasive tumour cells (Chu et al., 2001, Papafotiou et al., 2016, Cheah et al., 2015, Volkm et al., 2012, Cheung et al., 2016, Cheung et al., 2013, Lichtner et al., 1991, Gordon et al., 2003, Petrocca et al., 2013). In fact, in 2020, a study demonstrated that the expression of a 34-gene profile, including KRT14, represents a favourable prognostic factor for squamous non-small cell lung carcinoma (Theelen et al., 2020). In another study, a comprehensive bioinformatics analysis was conducted to identify potential genes associated with prognosis and treatment in BC in four GEO datasets. They identified seven genes, including KRT14, as critical prognostic candidate genes that were strongly related with OS in BC (Xu et al., 2020).

KIT, a tyrosine kinase of transmembrane receptors, plays a vital function in cell proliferation, survival, and migration (Liu et al., 2017, Hernández-Rojas et al., 2022, Wintheiser and Silberstein, 2021, Montor et al., 2018). Notably, tumour progression is linked to the downregulation of c-KIT expression, which is thought to be involved in human BC development during the initial stages (Janostiak et al., 2018). Talaezadeh et al. evaluated KIT expression in 60 BC and normal breast specimens via immunohistochemistry and concluded that there is an association between downregulation of KIT and the transformation of breast epithelial cells into cancer cells. KIT has also been assessed for its prognostic and predictive utility in patients with high-risk BC. Specifically, its expression was evaluated in 236 patients using tissue microarrays and its downregulation was detected in 12% of patients with BC. Furthermore, its expression was determined to be strongly related with poorer OS (Diallo et al., 2006). Thus, KIT expression may be a separate negative prognostic factor in patients with high-risk recurrence of BC. Comparable to the findings of the current study, a previous study of approximately 667 patients with BC showed that individuals with basal-like BC (28%) and those with nodal metastases (218%) had higher KIT positive tumours than healthy controls. They concluded that KIT may be a prognostic marker for individuals with basal-like BC and a potential biological target for treatment (Kashiwagi et al., 2013).

RAD51 is necessary for the homologous recombination mechanism to repair damaged DNA (Grundy et al., 2020, Inano et al., 2017, Gachechiladze et al., 2017). Thus, RAD51 mutations are directly correlated with genome instability and a predisposition for cancer development (Prakash et al., 2015, Sullivan and Bernstein, 2018). Moreover, RAD51 has been found to have novel functions in pancreatic cancer that, mechanistically, may involve the regulation of aerobic glycolysis, which is associated with OS outcome (Zhang et al., 2019b). In fact, patients with pancreatic cancer and elevated RAD51 levels show poorer prognosis, suggesting that RAD51 serves as a prognostic biomarker for pancreatic cancer and regulates cell proliferation (Zhang et al., 2019b). Nonetheless, RAD51 expression in cancer continues to be investigated both clinically and biologically. Specifically, RAD51 has come under intense scrutiny owing to its involvement in tumour progression and resistance to chemotherapy (Raderschall et al., 2002). Moreover, RAD51 is overexpressed in invasive ductal BC, and the degree of overexpression is connected to the histological categorisation of BC (Maacke et al., 2000). Additionally, another study indicated that RAD51 expression may be prognostic for individuals with surgically treated non-small cell lung cancer, as the lower survival of patients with non-small cell lung cancer with high levels of RAD51 expression was associated with an increased proclivity of tumour cells for survival, therapeutic resistance, and resistance to apoptosis. These findings suggest that RAD51 expression at a high level could be used as a prognostic marker of survival in patients with BC.

TTK is an important member of a number of mitotic stages in a wide range of eukaryotic cells (Wang et al., 2018, Pangou and Sumara, 2021). It is a major kinase implicated in the localisation of kinetochores and the spindle assembly checkpoint (Liu et al., 2020, Pachis and Kops, 2018). TTK is overexpressed in several different kinds of cancer and protects against chromosomal anomalies (Xie et al., 2017, Maire et al., 2013, Liu et al., 2015, Suyal et al., 2022, Yao et al., 2022, Jiao et al., 2022). According to earlier studies, TTK is upregulated in BC tissue and cells, notably in HER2-positive and triple-negative BC subtypes (Daniel et al., 2011, Salvatore et al., 2007, Tannous et al., 2013, Cui and Guadagno, 2008). Previously published research addressed the role of TTK in cancer utilizing in vitro pharmacological inhibitors. For instance, in the colorectal and glioblastoma in vitro models, the inhibition of TTK expression pharmacologically decreases cell viability, resulting in abnormal cell cycle progression, elevated aneuploidy, and induced apoptosis (Yao et al., 2021). TTK activity is crucial for pancreatic cancer cell proliferation, making TTK a potential therapeutic target for new treatments (Liu et al., 2021). Additionally, TTK was evaluated as a potential prognostic biomarker for triple-negative BC and it was demonstrated that increased TTK expression is associated with improved disease-free survival and OS, with a better prognosis (Xu et al., 2016). Similarly, the decreased TTK mRNA expression was positively correlated with worse OS, elevated incidence of metastasis, and shorter disease-free survival, all of which are indicators of poor prognosis (Maire et al., 2013). Another study using RNA microarray showed that more aggressive tumours have higher TTK expression, resulting in extremely low survival of less than two years (Al-Ejhe et al., 2014). TTK is variably expressed in normal and cancerous tissues, highlighting its viability as a genetic marker for diagnosis. Nevertheless, the fact that it correlates strongly with relapse and overall survival implies that it also might function as an independently prognostic marker (Xie et al., 2017). Based on the findings of this study, it is recommended to use TTK expression as a predictive marker in patients with IDC to aid in diagnosis as well as for the creation of individualised treatment plans for patients.

5. Conclusion

The current study used a novel approach to identify DEGs associated with the prognosis of primary and metastatic BC. Of the top 20 hub genes identified, four genes [KRT14, KIT (upregulated in BC) and RAD51, TTK (downregulated in BC)] could be considered for risk prognosis based on OS. However, the clinical importance of KRT14, KIT, RAD51, and TTK as prognostic factors should be further investigated prospectively. Future analysis of these four proposed prognostic genes may also provide insights into the molecular differences between primary breast tumours and metastatic lymphomas.
node metastases. In addition, the genes whose expression is altered in BC metastases may be used to design innovative therapies that precisely target metastatic colonisation.

**Funding**

This work was supported by the Taif University Researchers Supporting Project, Taif University, Taif, Saudi Arabia (Grant number TURSP-2020/202).

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Appendix A. Supplementary material**

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sjbs.2022.103318.

**References**

Abashev, T.M., Metzler, M.A., Wright, D.M., Sandell, L.L., 2017. Retinoic acid signaling regulates Krt5 and Krt14 independently of stem cell markers in submandibular salivary gland epithelium. Dev. Dyn. 246, 135–147.

Aftane, S.,牛. Pipson, P., Sanus, J., Klein, K., Kalimuthu, M., Shi, W., Miranda, M., Kutasovic, J., Raghavendra, A. & Madore, J. 2014. Meta-analysis of the global gene expression profile of triple-negative breast cancer identifies genes for the prognostication and treatment of aggressive breast cancer. Oncogenesis, 3, e100-e100.

Albogami, S.M., Asiri, Y., Asiri, A., Alnejadi, A.A., Alnejadi, S., 2021. Effects of neoadjuvant therapies on genetic regulation of targeted pathways in ER primary ductal breast carcinoma: a meta-analysis of microarray datasets. Saudi J. Pharm. J. 29, 656-669.

Andropoulou, E., Hortobagyi, G.N., 2008. Prognostic factors in metastatic breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 26, 3660–3662.

Bilanzidze, M., Rainczuk, A., Green, E., Fairweather, N., Jobling, T.W., Plebsianski, M., Stephens, A., 2019. Keratin-14 (KRT14) positive leader cells mediate mesothelial clearance and invasion by ovarian cancer cells. Cancers 11, 1228.

Bindea, G., Calon, J., Mlecnik, B., 2013. CrieVeda Cytosis plugin: pathway insights using integrated experimental and in silico data. Bioinformatics (Oxford, England) 29, 661–663.

Cheah, M.T., Chen, Y.J., Sainto, D., Contreras-Trujillo, H., Volkner, A.K., Scheeren, F.A., Volkner, J.P., Weissman, L.L. 2015. CD44-expressing cancer cells establish the inflammatory and proliferative tumor microenvironment in bladder cancer. Proc. Natl. Acad. Sci. 112, E4154–E4163.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Chen, S., Sato, K., Katsuki, Y., Kobayashi, W., Tanaka, H., Nakajima, K., Nakada, S., Miyoshi, H., Kines, K., Katoaki-Kondo, A. 2017. RFWD3-mediated ubiquitination promotes timely removal of both RPA and RAD51 from DNA damage sites to facilitate homologous recombination. Mol. Cell 62, 622–634.e8.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Chen, S., Sato, K., Katsuki, Y., Kobayashi, W., Tanaka, H., Nakajima, K., Nakada, S., Miyoshi, H., Kines, K., Katoaki-Kondo, A., 2017. RFWD3-mediated ubiquitination promotes timely removal of both RPA and RAD51 from DNA damage sites to facilitate homologous recombination. Mol. Cell 62, 622–634.e8.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Circular, G., Ianni, S., and M. Merlo, 2021. Targeting Leader Cells in Ovarian Cancer as an Effective Therapeutic Option. Ovarian Cancer-Updates in Management. Current Issues in Clinical Medicine, 4, 59-71.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Chen, S., Sato, K., Katsuki, Y., Kobayashi, W., Tanaka, H., Nakajima, K., Nakada, S., Miyoshi, H., Kines, K., Katoaki-Kondo, A., 2017. RFWD3-mediated ubiquitination promotes timely removal of both RPA and RAD51 from DNA damage sites to facilitate homologous recombination. Mol. Cell 62, 622–634.e8.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Circular, G., Ianni, S., and M. Merlo, 2021. Targeting Leader Cells in Ovarian Cancer as an Effective Therapeutic Option. Ovarian Cancer-Updates in Management. Current Issues in Clinical Medicine, 4, 59-71.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Chen, S., Sato, K., Katsuki, Y., Kobayashi, W., Tanaka, H., Nakajima, K., Nakada, S., Miyoshi, H., Kines, K., Katoaki-Kondo, A., 2017. RFWD3-mediated ubiquitination promotes timely removal of both RPA and RAD51 from DNA damage sites to facilitate homologous recombination. Mol. Cell 62, 622–634.e8.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Circular, G., Ianni, S., and M. Merlo, 2021. Targeting Leader Cells in Ovarian Cancer as an Effective Therapeutic Option. Ovarian Cancer-Updates in Management. Current Issues in Clinical Medicine, 4, 59-71.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Circular, G., Ianni, S., and M. Merlo, 2021. Targeting Leader Cells in Ovarian Cancer as an Effective Therapeutic Option. Ovarian Cancer-Updates in Management. Current Issues in Clinical Medicine, 4, 59-71.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.

Circular, G., Ianni, S., and M. Merlo, 2021. Targeting Leader Cells in Ovarian Cancer as an Effective Therapeutic Option. Ovarian Cancer-Updates in Management. Current Issues in Clinical Medicine, 4, 59-71.

Cheung, K.J., Gabrielson, E., Werb, Z., Ewald, A.J., 2013. Collective invasion in breast cancer: successes and challenges toward individualized therapy. J. Clin. Oncol. 31, 11024–11026.
Nazarov, P.V., Muller, A., Kaoma, T., Nicot, N., Maximo, C., Birembaut, P., Tran, N.L., Nathanson, S.D., 2003. Insights into the mechanisms of lymph node metastasis.

Marsh, T., Debnath, J., 2020. Autophagy suppresses breast cancer metastasis by

Maacke, H., Opitz, S., Jost, K., Hamdorf, W., Henning, W., Krüger, S., Feller, A.C., Ly, A., Lester, S.C., Dillon, D., 2012. Prognostic factors for patients with breast cancer: portraits of human breast tumours.

Lichtner, R., Julian, J., North, S., Glasser, S., Nicolson, G., 1991. Coexpression of

Kumar, R., Sharma, A., Tiwari, R.K., 2012. Application of microarray in breast cancer: attractive therapeutic target for triple-negative breast cancer. PLoS ONE 8, e73519.

Lichtner, R., Julian, J., North, S., Glasser, S., Nicolson, G., 1991. Coexpression of

Kumar, R., Sharma, A., Tiwari, R.K., 2012. Application of microarray in breast cancer: attractive therapeutic target for triple-negative breast cancer. PLoS ONE 8, e73519.

Lichtner, R., Julian, J., North, S., Glasser, S., Nicolson, G., 1991. Coexpression of

Kumar, R., Sharma, A., Tiwari, R.K., 2012. Application of microarray in breast cancer: attractive therapeutic target for triple-negative breast cancer. PLoS ONE 8, e73519.
Xie, Y., Wang, A., Lin, J., Wu, L., Zhang, H., Yang, X., Wan, X., Miao, R., Sang, X., Zhao, H., 2017. Mps1/TTK: a novel target and biomarker for cancer. J. Drug Target. 25, 112–118.

Xu, Q., Xu, Y., Pan, B., Wu, L., Ren, X., Zhou, Y., Mao, F., Lin, Y., Cuan, J., Shen, S., Zhang, X., Wang, C., Zhong, Y., Zhou, L., Liang, Z., Zhao, H., Sun, Q., 2016. TTK is a favorable prognostic biomarker for triple-negative breast cancer survival. Oncotarget 7, 81815–81829.

Xu, Y.-H., Deng, J.-L., Wang, L.-P., Zhang, H.-B., Tang, L., Huang, Y., Tang, J., Wang, S.-M., Wang, G., 2020. Identification of candidate genes associated with breast cancer prognosis. DNA Cell Biol. 39, 1205–1227.

Yao, Z.-P., Zhu, H., Shen, F., Gong, D., 2021. Hsp90 regulates the tumorigenic function of tyrosine protein kinase in osteosarcoma. Clin. Exp. Pharmacol. Physiol. 49 (3), 380–390.

Yao, Z.P., Zhu, H., Shen, F., Gong, D., 2022. Hsp90 regulates the tumorigenic function of tyrosine protein kinase in osteosarcoma. Clin. Exp. Pharmacol. Physiol. 49, 380–390.

Zhang, S., Cao, H., Ye, L., Wen, X., Wang, S., Zheng, W., Zhang, Y., Huang, D., Gao, Y., Liu, H., 2019a. Cancer-associated methylated IncRNAs in patients with bladder cancer. Am. J. Transl. Res. 11, 3790.

Zhang, X., Ma, N., Yao, W., Li, S., Ren, Z., 2019b. RAD51 is a potential marker for prognosis and regulates cell proliferation in pancreatic cancer. Cancer Cell Int. 19, 1–11.

Zhou, Z., Zhang, C., Ma, Z., Wang, H., Tuo, B., Cheng, X., Liu, X., Li, T., 2022. Pathophysiologica role of ion channels and transporters in HER2-positive breast cancer. Cancer Gene Ther., 1–8.