Chapter 3

Genome-Wide Gene Expression Analysis to Identify Epistatic Gene-Pairs Associated with Prognosis of Breast Cancer

I-Hsuan Lin and Ming-Ta Hsu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/59462

1. Introduction

Breast cancer is the most common cancer in women in the world [1]. According to the most recent estimates from GLOBOCAN published by the International Agency for Research on Cancer (IARC) [2], there were nearly 1.7 million new breast cancer cases diagnosed in 2012 (25.2% of all cancers in women) and 6.3 millions have been diagnosed with breast cancer in 2007-2012. Breast cancer incidence has been increasing by more than 20% and mortality increased by 14% since 2008 and is the most common cause of cancer death in women in less developed regions (324,000 deaths, 14.3% of total). Breast cancer is less favorable in the underdeveloped countries due to less advanced medical diagnosis and treatments. Therefore a good diagnosis/prognosis would help to prevent as well as provide effective clinical treatments.

Biomarker testing is an essential step in the evaluation of breast cancer and help medical doctors and patients in deciding the best treatment strategy. There are several commercial products or services developed towards this purpose. The Oncotype DX (Genomic Health) measures the expression levels of 21 genes and is most helpful for patients of early stage breast cancer with estrogen receptor (ER) positivity and no cancer cells in the lymph node. The HERmark assay (Monogram Biosciences) can quantitatively measure the HER2 total proteins with greater sensitivity than immunohistochemistry (IHC) which is an important indicator of predicting response of HER2-positive breast cancer patients to trastuzumab therapy. There are also tests for *BRCA1* and *BRCA2* mutations for the hereditary breast cancer patients. The targeted sequencing-based breast cancer panels such as BreastNext (Ambry Genetics) and BROCA (University of Washington) can be used to screen for mutations and copy number variants in genes implicated in breast cancer, including *BRCA1* and *BRCA2*.

Despite the relative success of these tests, there is a need for more efficient biomarkers in specific groups of breast cancer, such as lobular carcinoma [3, 4], triple-negative breast cancer [5] and early-onset breast cancers [6, 7] for diagnostic and/or prognostic application. We believe the discovery of more useful markers using the wealth of gene expression data available publicly nowadays would help medical doctors in the decision of the best way to help breast cancer patients, especially if the markers are correlated with specific therapeutic interventions. In the past decade, the high throughput microarray technique has been widely used to identify potential biomarkers for various cancers [8-12]. Recent years, the employment of RNA sequencing (RNA-seq) allows researchers to obtain transcriptome information and differential gene expression profiling at a much higher resolution. With the huge amount of data generated by these technologies, we are able to study the association of genes with cancer survival and identify novel potentially prognostic biomarkers for cancers with improved estimation. Traditionally, genetic search identifies genes that correlate with poor or good prognosis of patients. However, it is important to consider the epistatic gene-gene interactions underlying gene expression in complex diseases such as cancer. The epistatic (second or higher order) information would allow more refined prognostic evaluation that may help clinical treatments. Furthermore, epistatic analysis could be useful for identifying hub genes involved in prognosis and help to identify the major genetic risk factors and pathways in breast cancer.

In this work, we utilized the breast tumor RNA-seq data from The Cancer Genome Atlas (TCGA) as well as microarray-based expression datasets from Gene Expression Omnibus (GEO) to detect differentially expressed genes in various subsets of breast cancer patients, to identify genes whose expression profile is associated with survival of breast cancer patients and to examine the influence of co-expression of a second gene in the survival of patients. This analysis identifies specific gene groups differentially expressed between early-onset vs. late-onset breast cancer, between ductal vs. lobular carcinoma, between early vs. advanced stage breast tumors and tumor of various receptor status. Furthermore, epistatic interactions among these genes demonstrate the gene-gene interactions in patient survival and identify several hub genes that may be important determinants of breast cancer.

2. Statistical analysis of gene expression data

A global change in gene expression is a common theme in many human cancers. High-throughput techniques such as microarrays and next generation sequencing allow investigators to observe and compare the transcriptional landscapes of tumor cells in different biological states [13-18]. In this work, we integrated multiple gene expression data from several large-scale breast cancer studies to improve the assessment of differential gene expression in breast tumor cells and to effectively increase statistical power.

We collected 3,188 breast cancer related Affymetrix expression microarray data from GEO (http://www.ncbi.nlm.nih.gov/geo) from the following 16 series: GSE2603, GSE4922, GSE2990, GSE3494, GSE6532, GSE9195, GSE7390, GSE20194, GSE20271, GSE20685, GSE25066, GSE16391, GSE19615, GSE42568, GSE45255 and GSE50948. We also obtained 1,172 breast invasive carcinoma (BRCA) RNA-seq Level 3 data from TCGA Data Portal (http://tcga-data.nci.nih.gov/). The demographic and clinicopathological characteristics of the breast cancer patients from each study were also retrieved.

2.1. Processing of gene expression data and differential gene expression analysis

The CEL files obtained from microarray experiments were pre-processed by subjecting to quality check using Bioconductor in the R environment to ensure comparability between the different series and microarray platforms. The following quality measurements from the *simpleaffy* and *affy* packages were performed: average background (avbg), scale factor (sfs), percent present (percent.present), and possible RNA Degradation (AffyRNAdeg) of the array. Additionally, the relative log expression (RLE) and normalized unscaled standard error (NUSE) was also estimated using the *affyPLM* package. 466 arrays that did not pass the quality control tests were removed. For the 2,722 arrays that had sufficient quality, the quantile normalization and background correction were performed using the justRMA (robust multi-array average) algorithm of the *affy* package and the gene (probe set)-level log2-transformed expression values were summarized with Custom CDF file annotations (version 18.0.0. ENSG) [19]. Lastly, the COMBAT method available in the *inSilicoMerging* package was used to remove batch effect when combining the final expression data from the HG-U133A and HG-U133 Plus 2.0 arrays [20]. The RMA-normalized expression values from microarrays and the raw count data from RNA-seq datasets were then analyzed using the *edgeR* package [21]. The differentially expressed genes were selected with a threshold of FDR adjusted P-value < 0.05.

2.2. Chinicopathological characteristics of breast cancer patients

We include 2,722 breast cancer patients from various microarray-based studies (referred as GEO cohort) and 1,052 breast cancer patients from the TCGA project (referred as TCGA cohort) following differential gene expression analyses (Table 1). All patients were women in the GEO cohort with a median age of 53 years. The patients from the TCGA cohort were older with a median age of 58 years and approximately 96% of patients were women. There was a significant amount of clinicopathological data not available from the GEO cohort as noted in Table 1. In both cohorts, there were more stage I/II breast cancer cases than advanced stage cases, and invasive ductal carcinoma (IDC) being the major histological subtype diagnosed. The data also contained status of tumor receptors such as the estrogen receptor (ER), progesterone receptor (PR) and HER2 which are frequently used prognostic factors to aid therapeutic decisions. Many patients were positive for the ER and/or PR, and/or negative for the HER2 receptor.

Characteristic	No. of Patients			
	Microarray (n = 2722), %		RNA-seq (n = 1052), %	
Sex				
Male	0	0.0%	11	1.0%

Characteristic	No. of Patients			
	Microarray (n = 2722), %		RNA-seq (n = 1052), %	
Female	2722	85.4%	1005	95.5%
Missing data	0	0.0%	36	3.4%
Median Age (range)	53 (24-93)		58 (26-90)	
Younger than 40	300	11.0%	71	6.7%
40 to 55	1184	43.5%	365	34.7%
Older than 55	1224	45.0%	580	55.1%
Missing data	14	0.5%	36	3.4%
Stage				
Early (Stage I and II)	1236	45.4%	751	71.4%
Late (Stage III and IV)	370	13.6%	246	23.4%
Missing data	1116	41.0%	55	5.2%
Histologic Subtype				
IDC	500	18.4%	754	71.7%
ILC	32	1.2%	168	16.0%
Mixed	47	1.7%	29	2.8%
Others	6	0.2%	64	6.1%
Missing data	2137	78.5%	37	3.5%
ER Status				
ER positive	1710	62.8%	749	71.2%
ER negative	647	23.8%	222	21.1%
Missing data	365	13.4%	81	7.7%
PR Status				
PR positive	1017	37.4%	650	61.8%
PR negative	684	25.1%	318	30.2%
Missing data	1021	37.5%	84	8.0%
HER2 Status				
HER2 positive	202	7.4%	150	14.3%
HER2 negative	946	34.8%	524	49.8%
Missing data	1574	57.8%	378	35.9%
Female patients with a least one type of survival data	2294	84.3%	999	36.7%

Table 1. Patient characteristics of the GEO and TCGA cohorts.

2.3. Differentially expressed genes in patients from different age groups

Differential gene expression analysis was performed to identify over- and under-expressed genes specific to tumors derived from young, middle-aged and elderly breast cancer patients. As presented in Figure 1, there were very few middle-aged-specific expression signatures, indicating that the gene expression pattern of middle-aged patients was not significantly different from the young adults and/or elderly patients. In contrast, the elderly breast cancer patients possessed a high number of differentially expressed genes specific to this age group. IPA analysis of the differentially expressed genes from tumor cells obtained from older patients have decreased cell proliferation, movement, migration and cell cycle progression (activation z-score between -2.677 and -1.611) and increased cell death (activation z-score = 1.321). On the contrary, tumor cells from young patients were predicted to have increased proliferation of cells and DNA synthesis (activation z-score between 2.000 and 2.117) and decreased cell death and apoptosis (activation z-score between -0.586 and -0.299).

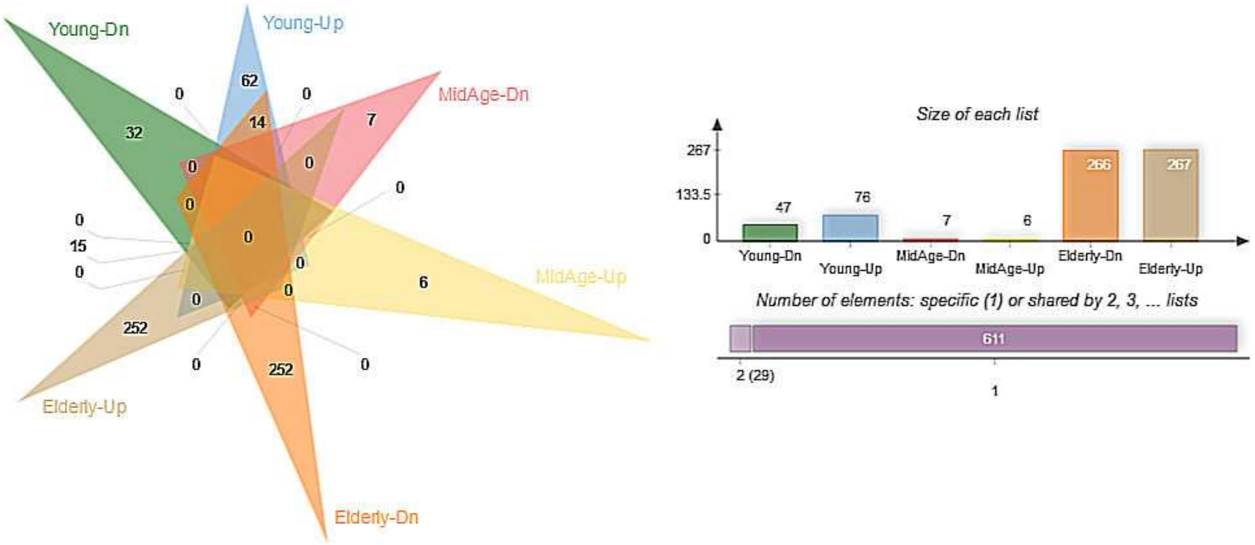

Figure 1. Number of significantly over- and under-expressed genes in the three age groups presented with the jvenn Venn diagram viewer [22].

It is interesting to note that 14 genes that were over-expressed in young patients were under-expressed in elderly patients, and conversely, 15 genes under-expressed in young patients were over-expressed in elderly patients (Table 2). Several of these genes such as *BIRC5* (survivin), *KPNA2*, *PLAC8* (onzin), *TFPI2*, *CITED2*, *NKX3-1*, *PIP* and *ZNF395* have been found to play a role in cancer cell proliferation and cancer progression [23-28].

Type	Symbol	Entrez Gene Name	Location	Type(s)
Young-Up Elderly-Dn	*BIRC5*	baculoviral IAP repeat containing 5	Cytoplasm	Other
	DCX	doublecortin	Cytoplasm	Other
	GAL	galanin/GMAP prepropeptide	Extracellular Space	Other

Type	Symbol	Entrez Gene Name	Location	Type(s)
	HN1	hematological and neurological expressed 1	Nucleus	Other
	KPNA2	karyopherin alpha 2	Nucleus	Transporter
	NOL11	nucleolar protein 11	Nucleus	Other
	NUP85	nucleoporin 85kDa	Cytoplasm	Other
	PLAC8	placenta-specific 8	Nucleus	Other
	POLR3G	polymerase (RNA) III (DNA directed) polypeptide G	Nucleus	Enzyme
	PSMA4	proteasome alpha 4 subunit isoform 1	Cytoplasm	Peptidase
	RAPGEFL1	Rap guanine nucleotide exchange factor	Other	Other
	TFPI2	tissue factor pathway inhibitor 2	Extracellular Space	Other
	UCHL1	ubiquitin carboxyl-terminal esterase L1	Cytoplasm	Peptidase
	XDH	xanthine dehydrogenase	Cytoplasm	Enzyme
Young-Dn Elderly-Up	*ABCC6*	ATP-binding cassette, sub-family C, member 6	Plasma Membrane	Transporter
	ACAA1	acetyl-CoA acyltransferase 1	Cytoplasm	Enzyme
	CCDC28A	coiled-coil domain containing 28A	Other	Other
	CITED2	Cbp/p300-interacting transactivator	Nucleus	Transcription regulator
	CLMN	calmin (calponin-like, transmembrane)	Cytoplasm	Other
	CTDSPL	small CTD phosphatase 3 isoform 1	Nucleus	Other
	CTSF	cathepsin F	Cytoplasm	Peptidase
	FMO5	flavin containing monooxygenase 5	Cytoplasm	Enzyme
	GPC4	glypican 4	Plasma Membrane	Transmembrane receptor
	KIF13B	kinesin family member 13B	Cytoplasm	Other
	NDST1	N-deacetylase/N-sulfotransferase (heparan)	Cytoplasm	Enzyme
	NKX3-1	NK3 homeobox 1	Nucleus	Transcription regulator
	PIP	prolactin-induced protein	Extracellular Space	Peptidase
	ZNF385D	zinc finger protein 385D	Nucleus	Other
	ZNF395	zinc finger protein 395	Cytoplasm	Other

Table 2. Concordant differentially expressed genes identified in the young and elderly breast cancer patients.

2.4. Differentially expressed genes in patients with early stage versus advanced stage breast cancer

We compared the gene expression profile of patients diagnosed with early stage (stage I and II) breast cancer with those with advanced stage (stage III and IV) breast cancer. We identified 79 over-expressed and 140 under-expressed genes in early stage breast cancer. IPA analysis showed 121 of the total 219 differentially expressed genes were associated with cancer (*P*-value = 4.81E-02), and 24 were specifically associated with breast cancer (*P*-value = 3.25E-03, Table 3). Also, there were 17 under-expressed genes in early stage breast cancer (i.e. over-expressed in advanced stage tumors) that were found to be cancer recurrence-associated (*ADORA3*, *FLT4*, *GSR*, *HSP90AA1*, *TEK* and *TXNRD1*) and metastasis-associated (*ACP5*, *FLT4*, *FTL*, *GSR*, *HSP90AA1*, *MAPK11*, *MMP9*, *NRAS*, *PGF*, *SCD* and *TEK*). Interestingly, we detected over-expression of the DNA methyltransferase *DNMT1* in early stage tumors. In cancer cells, the over-expression of this gene can lead to hypermethylation of CpG islands and epigenetically silences multiple tumor suppressor genes and hence promotes tumorigenesis in early stage cancers [29-31].

Symbol	Entrez Gene Name	Location	Type(s)	DE Status
ACP5	acid phosphatase 5, tartrate resistant	Cytoplasm	phosphatase	Down
APOE	apolipoprotein E	Extracellular Space	transporter	Down
ARRB1	arrestin, beta 1	Cytoplasm	other	Down
CDKN1A	cyclin-dependent kinase inhibitor 1A	Nucleus	other	Down
ETV1	ets variant 1	Nucleus	transcription regulator	Down
FLT4	fms-related tyrosine kinase 4	Plasma Membrane	transmembrane receptor	Down
GPC3	glypican 3	Plasma Membrane	other	Up
GPR126	G protein-coupled receptor 126	Plasma Membrane	G-protein coupled receptor	Down
HBB	hemoglobin, beta	Cytoplasm	transporter	Down
HIC1	hypermethylated in cancer 1	Nucleus	transcription regulator	Down
HSP90AA1	heat shock protein 90kDa alpha (cytosolic)	Cytoplasm	enzyme	Down
HSPB7	cardiovascular heat shock protein	Cytoplasm	other	Down
MMP15	matrix metalloproteinase 15 preproprotein	Extracellular Space	peptidase	Down
MMP28	matrix metalloproteinase 28 isoform 1	Extracellular Space	peptidase	Down
MMP9	matrix metalloproteinase 9 preproprotein	Extracellular Space	peptidase	Down
NOS3	nitric oxide synthase 3 (endothelial cell)	Cytoplasm	enzyme	Down
PPM1D	protein phosphatase 1D	Cytoplasm	phosphatase	Down
PSIP1	PC4 and SFRS1 interacting protein 1	Nucleus	other	Up
PXN	paxillin	Cytoplasm	other	Down
S100A2	S100 calcium binding protein A2	Nucleus	other	Down
SELL	selectin L	Plasma Membrane	transmembrane receptor	Down
TAL1	T-cell acute lymphocytic leukemia 1	Nucleus	transcription regulator	Down
TEK	TEK tyrosine kinase, endothelial	Plasma Membrane	kinase	Down
TNC	tenascin C	Extracellular Space	other	Up

Table 3. The 25 differentially expressed genes associated with breast cancer in the early versus advanced stage analysis.

2.5. Differentially expressed genes in patients with Invasive Ductal Carcinoma (IDC) versus Invasive Lobular Carcinoma (ILC)

The invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) are the two major histological subtypes of breast cancer. They also represented the main types of breast cancer cases gathered in this study. We compared the gene expression profiles of patients with IDC and ILC and identified 216 over-expressed and 126 under-expressed genes in IDC as compared to patients with ILC. IPA analysis showed 66 genes were related to breast cancer (Table 4), including 12 transcription regulators (*ATF3*, *BTG2*, *EGR1*, *EZH2*, *FOS*, *FOSB*, *JUN*, *JUNB*, *MTDH*, *STAT1*, *ZFP36* and *ZNF423*) and a translation regulator (*EIF4EBP1*). There were also 12 genes annotated as tumor suppressor genes in the TSGene database [32], where *CDH1*, *DKK1* and *S100A2* were over-expressed in IDC and *BTG2*, *CAV1*, *EGR1*, *GPC3*, *MUC1*, *NR4A1*, *SLIT2*, *TGFBR2* and *ZFP36* were under-expressed in IDC. IPA predicted common upstream regulators *KDM5B*, *STUB1*, *CDKN1A*, *HIF1A* and *TGFB1* to be inhibited whereas *FOXM1*, *IFNB1*, *IFNG* and *PPARG* were in activated states. Additionally, the activities of several disease functions such as cell proliferation, invasion and DNA replication were predicted to be increased in IDC (activation z-score between 1.342 and 3.092).

Symbol	Entrez Gene Name	Location	Type(s)	DE Status
ACACB	acetyl-CoA carboxylase beta	Cytoplasm	enzyme	Down
ALDH1A1	aldehyde dehydrogenase 1 family, member A1	Cytoplasm	enzyme	Down
APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic	Cytoplasm	enzyme	Up
ATF3	activating transcription factor 3	Nucleus	transcription regulator	Down
BIRC5	baculoviral IAP repeat containing 5	Cytoplasm	other	Up
BTG2	BTG family, member 2	Nucleus	transcription regulator	Down
CAV1	caveolin 1, caveolae protein, 22kDa	Plasma Membrane	transmembrane receptor	Down
CCL21	chemokine (C-C motif) ligand 21	Extracellular Space	cytokine	Down
CD34	CD34 molecule	Plasma Membrane	other	Down
CD69	CD69 molecule	Plasma Membrane	transmembrane receptor	Down
CDC20	cell division cycle 20	Nucleus	other	Up
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	Plasma Membrane	other	Up
CDH3	cadherin 3, type 1, P-cadherin (placental)	Plasma Membrane	other	Up
CDH5	cadherin 5, type 2 (vascular endothelium)	Plasma Membrane	other	Down
CDK1	cyclin-dependent kinase 1	Nucleus	kinase	Up

Symbol	Entrez Gene Name	Location	Type(s)	DE Status
CXCL14	chemokine (C-X-C motif) ligand 14	Extracellular Space	cytokine	Down
CXCL2	chemokine (C-X-C motif) ligand 2	Extracellular Space	cytokine	Down
CYR61	cysteine-rich, angiogenic inducer, 61	Extracellular Space	other	Down
DKK1	dickkopf WNT signaling pathway inhibitor 1	Extracellular Space	growth factor	Up
DSCC1	DNA replication and sister chromatid cohesion 1	Nucleus	other	Up
DUSP1	dual specificity phosphatase 1	Nucleus	phosphatase	Down
EGR1	early growth response 1	Nucleus	transcription regulator	Down
EIF4EBP1	eukaryotic translation initiation factor 4E	Cytoplasm	translation regulator	Up
EZH2	enhancer of zeste homolog 2 (Drosophila)	Nucleus	transcription regulator	Up
FABP7	fatty acid binding protein 7, brain	Cytoplasm	transporter	Up
FOS	FBJ murine osteosarcoma viral oncogene homolog	Nucleus	transcription regulator	Down
FOSB	FBJ murine osteosarcoma viral oncogene homolog B	Nucleus	transcription regulator	Down
GPC3	glypican 3	Plasma Membrane	other	Down
GRB7	growth factor receptor-bound protein 7	Plasma Membrane	other	Up
HSPB8	heat shock 22kDa protein 8	Cytoplasm	kinase	Up
IER2	immediate early response 2	Cytoplasm	other	Down
IGF1	insulin-like growth factor 1 (somatomedin C)	Extracellular Space	growth factor	Down
IGFBP6	insulin-like growth factor binding protein 6	Extracellular Space	other	Down
ITIH5	inter-alpha trypsin inhibitor heavy chain	Other	other	Down
JUN	jun proto-oncogene	Nucleus	transcription regulator	Down
JUNB	jun B proto-oncogene	Nucleus	transcription regulator	Down
KPNA2	karyopherin alpha 2	Nucleus	transporter	Up
KRT6B	keratin 6B	Cytoplasm	other	Up
MMP1	matrix metalloproteinase 1 preproprotein	Extracellular Space	peptidase	Up
MMP9	matrix metalloproteinase 9 preproprotein	Extracellular Space	peptidase	Up
MRPL13	mitochondrial ribosomal protein L13	Cytoplasm	other	Up
MRPL15	mitochondrial ribosomal protein L15	Cytoplasm	other	Up
MTDH	metadherin	Cytoplasm	transcription regulator	Up

Symbol	Entrez Gene Name	Location	Type(s)	DE Status
MUC1	mucin 1, cell surface associated	Plasma Membrane	other	Down
NR4A1	nuclear receptor subfamily 4, group A, member 1	Nucleus	ligand-dependent nuclear receptor	Down
ORM1	orosomucoid 1	Extracellular Space	other	Up
PCNA	proliferating cell nuclear antigen	Nucleus	enzyme	Up
PDK4	pyruvate dehydrogenase kinase, isozyme 4	Cytoplasm	kinase	Down
PGK1	phosphoglycerate kinase 1	Cytoplasm	kinase	Up
RFC4	replication factor C (activator 1) 4, 37kDa	Nucleus	other	Up
RRM2	ribonucleotide reductase M2	Nucleus	enzyme	Up
S100A2	S100 calcium binding protein A2	Nucleus	other	Up
SLIT2	slit homolog 2 (Drosophila)	Extracellular Space	other	Down
SPP1	secreted phosphoprotein 1	Extracellular Space	cytokine	Up
SQLE	squalene epoxidase	Cytoplasm	enzyme	Up
STAT1	signal transducer and activator of transcription	Nucleus	transcription regulator	Up
TCP1	t-complex 1	Cytoplasm	other	Up
TGFBR2	transforming growth factor, beta receptor II	Plasma Membrane	kinase	Down
TIMP4	TIMP metallopeptidase inhibitor 4	Extracellular Space	other	Down
TOP2A	topoisomerase (DNA) II alpha 170kDa	Nucleus	enzyme	Up
TPD52	tumor protein D52	Cytoplasm	other	Up
TYMS	thymidylate synthetase	Nucleus	enzyme	Up
UBE2C	ubiquitin-conjugating enzyme E2C	Cytoplasm	enzyme	Up
VEGFA	vascular endothelial growth factor A	Extracellular Space	growth factor	Up
ZFP36	ZFP36 ring finger protein	Nucleus	transcription regulator	Down
ZNF423	zinc finger protein 423	Nucleus	transcription regulator	Down

Table 4. The 66 differentially expressed genes associated with breast cancer in the IDC versus ILC analysis.

2.6. Differentially expressed genes in patients with different receptor status

In the last part of the differential gene expression analysis, we sought to examine the differentially expressed genes of breast cancer patients of different receptor status: (1) estrogen receptor positive (ER+) versus ER negative (ER–), (2) progesterone receptor positive (PR+) versus PR negative (PR–), (3) HER2 receptor positive (HER2+) versus HER2 negative (HER2–), and (4) triple-negative breast cancer (TNBC, also known as basal-like breast cancer) versus non-TNBC. The Venn diagram in Figure 2 summarized the intersections between the differentially expressed genes identified in the four assays. There were 57% and 65% of breast cancer patients that were both ER+ and PR+ in the GEO and TCGA cohorts respectively; hence the patient pools divided by the ER positivity for gene expression assays are similar to that divided by the PR positivity. Because of this fact, it is not surprising to observe genes that were found over- or under-expressed in the ER assay were also differentially expressed in the same direction in the PR assay. Likewise, genes that were over-expressed in TNBC were under-expressed in the ER and PR assays (n = 74) and ER, PR and HER2 assays (n = 35), and vice versa for the under-expressed genes in TNBC (n = 87 and 2 respectively).

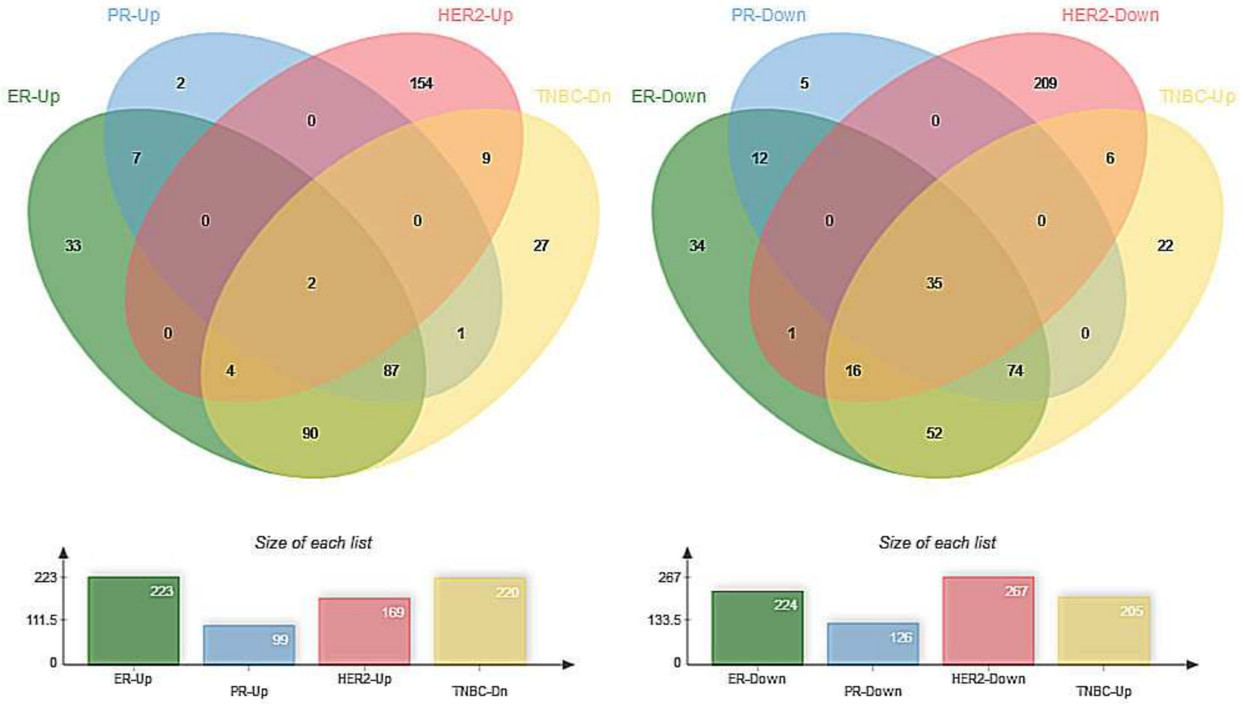

Figure 2. Number of differentially up- and down-regulated genes in the ER, PR, HER2 or TNBC receptor status assays.

The *GALNT6* (polypeptide N-acetylgalactosaminyltransferase) and *SCGB2A2* (secretoglobin) are the two genes consistently over-expressed in ER+, PR+, HER2+ breast tumors but under-expressed in TNBC. There were also 87 genes over-expressed in ER+ and PR+ breast tumors and under-expressed in TNBC, including seven transcription regulators (*EGR3*, *FOXA1*, *GATA3*, *INSM1*, *NRIP1*, *TBX3* and *XBP1*) and 11 breast cancer associated genes (*ABAT*, *AGR2*, *CXCL14*, *GSTM3*, *HSPB8*, *MUC1*, *NR4A2*, *PGR*, *PIP*, *PLAT* and *PSD3*). On the other hand, there were more under-expressed genes (n = 35) shared among ER+, PR+ and HER2+ breast tumors that were over-expressed in TNBC. Among these are four transcription regulators (*ELF5*, *EN1*, *FOXC1* and *ZIC1*) and 12 extracellular proteins (*CHI3L1*, *CHI3L2*, *COL2A1*, *COL9A3*, *CRLF1*, *KLK6*, *KLK7*, *MMP12*, *MMP7*, *PTX3*, *SERPINB5* and *SOSTDC1*), and some of these genes are known TNBC-associated markers [33-37].

3. Identifying survival-related genes of patients with breast cancer

In cancer survival analysis, survival time is often defined as the period of time from the beginning of the medical process (treatment, surgery, etc.) until the death (or some other events such as development of a particular symptom or to relapse after the remission of disease) of the observed patient or until the end of observation. The goal of such analysis is to link the time to event (i.e. survival time) to certain explanatory variables. New methodologies were developed for calculating the survival probabilities using gene expression profiles when genome-wide expression data becomes increasingly available in the past two decades [38-42].

In this work, we analyzed associations between breast cancer patient survival and gene expression of breast tumors from published microarray and the RNA-seq datasets, denoted as the GEO and TCGA cohorts respectively. Survival analysis was performed separately for each cohort and the median times from diagnosis to death or last follow-up were 99.5 and 21.4 months in the GEO and TCGA cohorts respectively. We transform the expression values into gene expression status (i.e. 0 for low expression and 1 for high expression) using the modified auto_cutoff function of the R script available from the Kaplan Meier-plotter website (http://kmplot.com/). The survival probability is calculated using the "survival" package and modified *kmplot* function (http://biostat.mc.vanderbilt.edu/wiki/Main/TatsukiRcode#kmplot) is used to plot Kaplan-Meier curves. The hazard ratio with 95% confidence intervals and log-rank *P*-value are estimated using the Cox proportional hazards model. All analyses were conducted within the R statistical environment.

3.1. Univariate gene selection and survival analysis

We extract the gene expression profiles of 1,694 genes that were found differentially expressed (consistently in microarray and RNA-seq datasets) in any one type of the assays discussed in section 2.3 to 2.6. We calculated the overall survival (OS), relapse-free survival (RFS) and the distant metastasis-free survival (DMFS) of breast cancer patients with respect to the expression status. DMFS is not calculated for the TCGA cohort due to unavailability of the time to distant metastases information from patients in this cohort. The log-rank *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) method and were used to select genes expression profiles significantly associated with survival.

We summarized the survival statistics, including the hazard ratios (an estimate of the ratio of the hazard rate in the highly versus the lowly expressed patient group) and the estimated 2- and 10-year survival rates in Table 5. There were about 24% OS-associated, 48% RFS-associated and 51% DMFS-associated genes that have adjusted log-rank *P*-value < 0.01 in the GEO cohort. There were 23% OS-associated genes but only 1.3% RFS-associated genes in the TCGA cohort, due to much fewer relapse/recurrence information in this cohort (adjusted log-rank *P*-value < 0.05). The breast cancer patients in the TCGA cohorts have lower overall survival (10-year) than those from the GEO and both cohorts have similar RFS rates. In the DMFS analysis, 51% of the differentially expressed genes were significant predictor of DMFS.

Statistics	GEO	TCGA
Adjusted log-rank P-value cutoff	0.01	0.05
Overall survival (OS)		
No. of genes associated with OS	414 (24.4%)	386 (22.8%)
Hazard Ratio > 1		
No. of genes	192	134
2-year survival (low / high expression)	0.969, 0.923	0.968, 0.947
10-year survival (low / high expression)	0.798, 0.658	0.562, 0.369
Hazard Ratio < 1		
No. of genes	222	252
2-year survival (low / high expression)	0.918, 0.968	0.945, 0.970
10-year survival (low / high expression)	0.654, 0.787	0.382, 0.558
Relapse-free survival (RFS)		
No. of genes associated with RFS	811 (47.8%)	22 (1.3%)
Hazard Ratio > 1		
No. of genes	344	7
2-year survival (low / high expression)	0.901, 0.840	0.949, 0.829
10-year survival (low / high expression)	0.685, 0.586	0.754, 0.555
Hazard Ratio < 1		
No. of genes	467	15
2-year survival (low / high expression)	0.845, 0.900	0.826, 0.946
10-year survival (low / high expression)	0.588, 0.684	0.538, 0.749
Distant metastasis-free survival (DMFS)		
No. of genes associated with DMFS	856 (50.5%)	NA
Hazard Ratio > 1		
No. of genes	384	NA
2-year survival (low / high expression)	0.923, 0.863	
10-year survival (low / high expression)	0.755, 0.663	
Hazard Ratio < 1		
No. of genes	472	NA
2-year survival (low / high expression)	0.858, 0.926	
10-year survival (low / high expression)	0.667, 0.754	

Table 5. Survival statistics according to gene expression profiles of breast cancer patients.

3.2. Cox regression analysis using the expression profiles of two genes

From the three survival data, i.e. OS, RFS and DMFS, we selected the top 500 most significantly survival associated gene expression profiles consistent in both cohorts to generate 124,750 two-gene combinations and perform Cox regression analysis with two covariates (i.e. using the expression status of each gene as a covariate).

There were 81,902 (65.7%) and 78,136 (62.6%) pairs whose expression signatures of both genes remained predictors of OS in the GEO and TCGA cohorts respectively (P-value of the coefficient estimates < 0.05). Of these, 31,189 pairs were mutual in the two cohorts and 234 gene-pairs (consisting of 131 genes) had survival probability patterns greatly shifted compared with the previous single-gene model. The strongest predictor pairs were *COL16A1-ARHGEF3*, *IGF1R-LTB*, *IGF1R-PTGDS*, *NPY1R-ARHGEF3* and *SERPINA1-ACADSB*, where the lower expression of both genes in each pair was associated with lowest survival probabilities in all five cases. The results were presented as Kaplan Meier plots in Figure 3A to E.

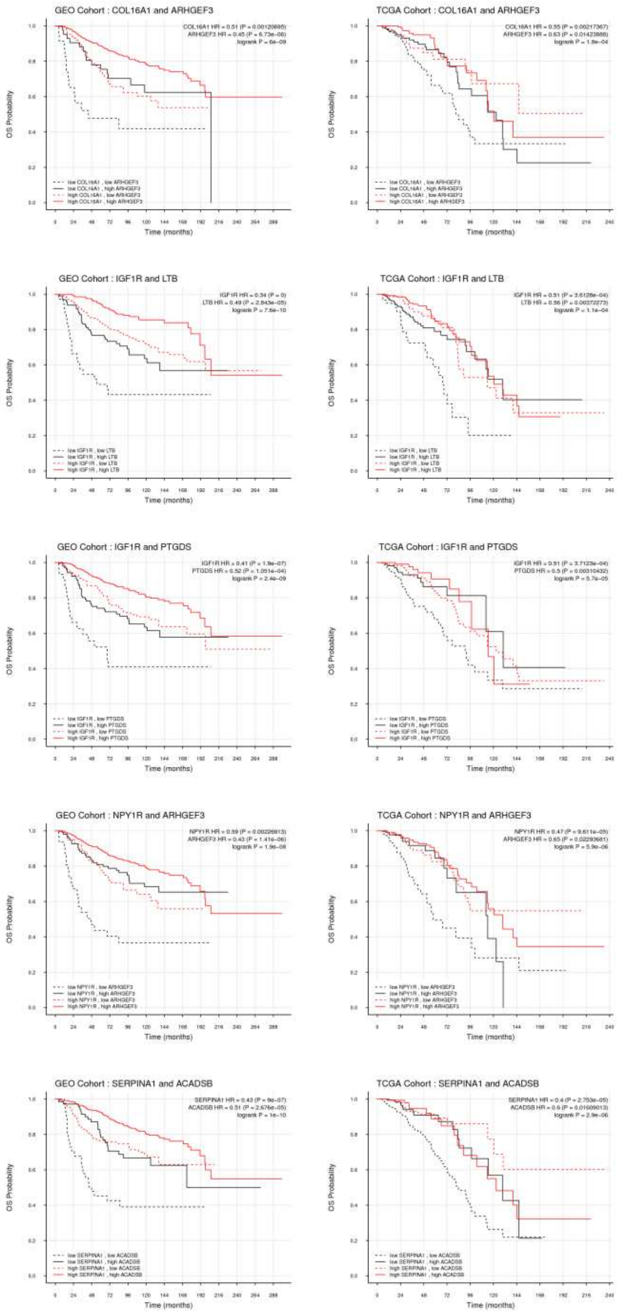

Figure 3. The Kaplan Meier plots of five OS-associated gene-pairs that also gained most changes in survival probabilities compared to the matching univariate approach.

The same analysis was also performed for RFS in the GOE and TCGA cohorts, and 85,244 (68.3%) and 64,049 (51.3%) pairs were significant predictors of RFS respectively (*P*-value of the coefficient estimates < 0.05). We found 22,165 significant pairs common in the two cohorts and 1,130 gene-pairs (consisting of 276 genes) whose survival probability patterns had greatly shifted compared with the single-gene model. The most significant RFS-associated pairs were *ADM-MYBPC1*, *DIRAS3-TANC2*, *KIFC1-ADORA3*, *PDSS1-DIRAS3*, *STMN1-ADORA3* (Figure 4).

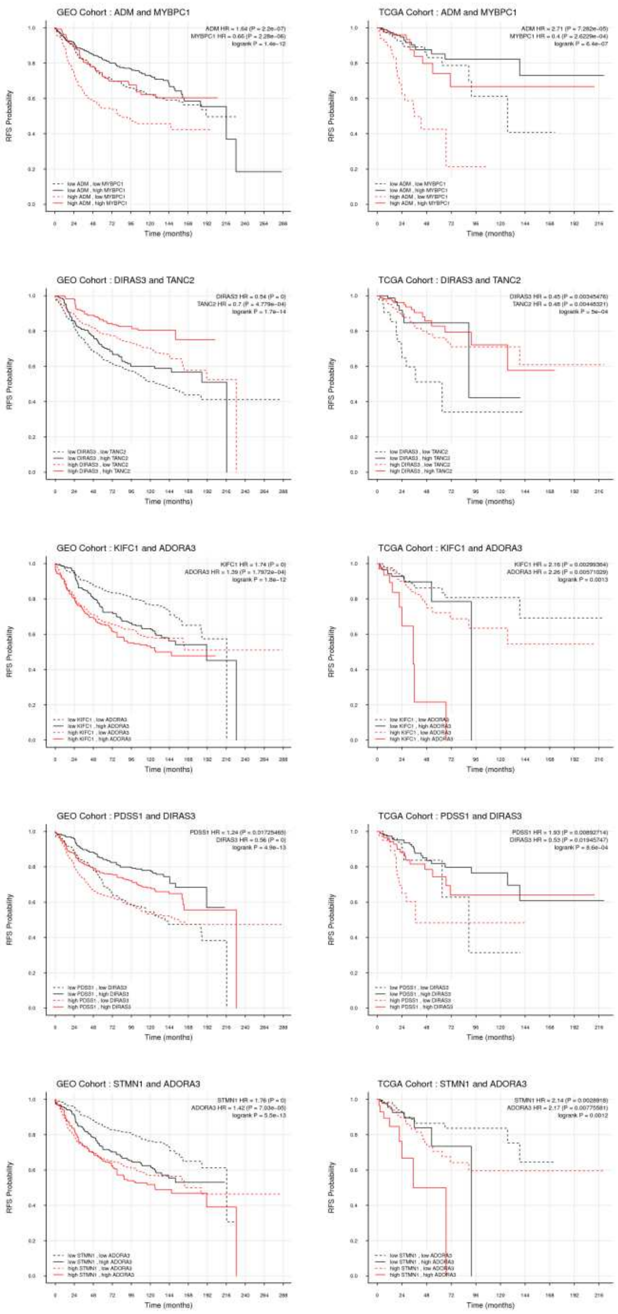

Figure 4. The Kaplan Meier plots of five RFS-associated gene-pairs that also gained most changes in survival probabilities compared to the matching univariate approach.

Lastly, we also make use of the DMFS data available from the GEO cohort to demonstrate the improvement of epistatic gene-pair approach in predicting survival probabilities. Of the 124,750 two-gene combinations, 122,751 (98.4%) were significant predictors of DMFS in breast cancer patients. The high percentage of strong two-gene predictors derived from the DMFS analysis was most likely due to the already high numbers of strong single-gene predictors as shown in Table 5. We further distinguished 228 gene-pairs (consisting of 138 genes) whose survival probability patterns had greatly shifted compared with the single-gene model. Six most significantly improved DMFS-associated pairs were *MMP15-SPDEF*, *TRIB3-ETV1*, *TRIB3-PLD1*, *TRIB3-TRIM2*, *TRIM2-KRT14* and *XBP1-TRIB3* (Figure 5).

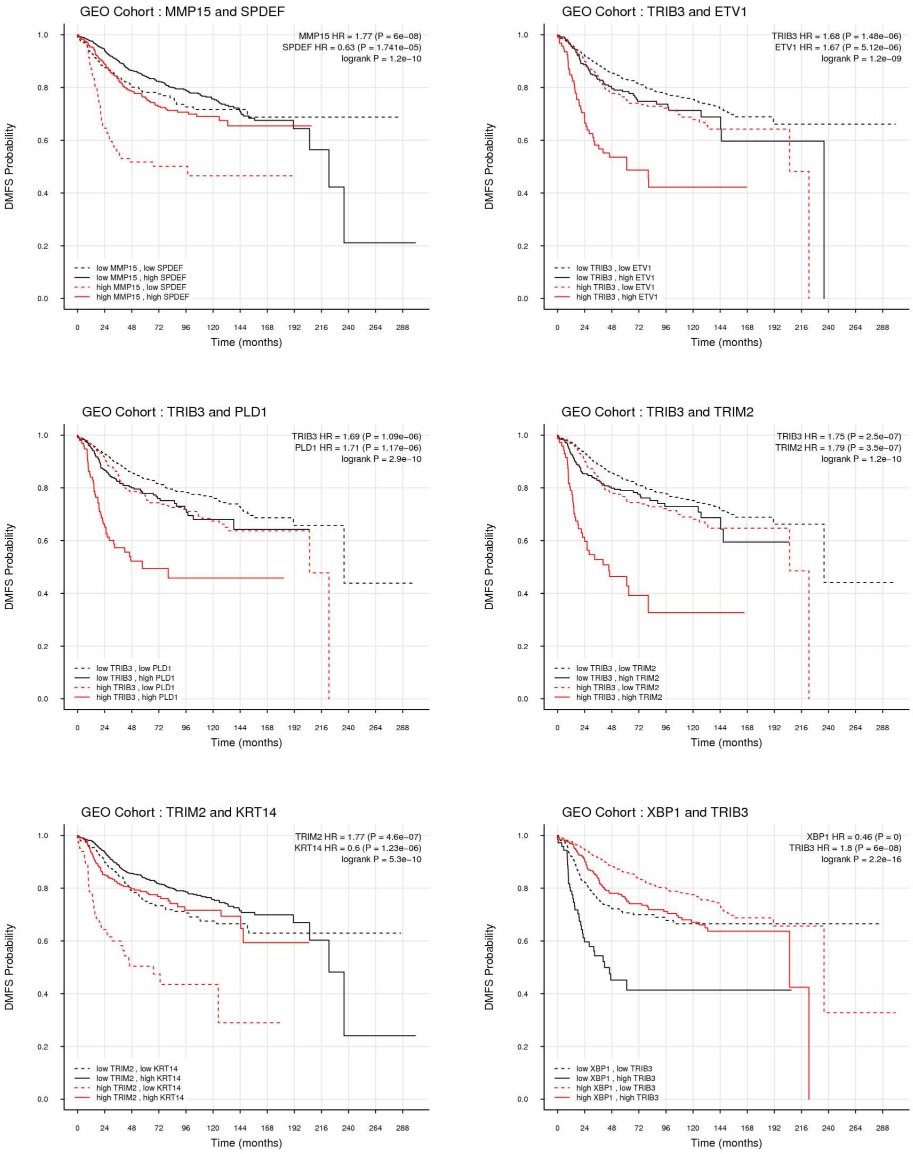

Figure 5. The Kaplan Meier plots of six DMFS-associated gene-pairs that also gained most changes in survival probabilities compared to the matching univariate approach.

3.3. Differentially expressed survival-associated hub genes and gene-pair candidates

As mentioned in the section 3.2, we have identified 234, 1,130 and 228 OS-, RFS- and DMFS-associated gene-pairs (consisting of 131, 276 and 138 genes respectively) that showed improved predictive performance. Some of these genes may be paired with many partners while remaining highly significant. In Table 6, we list five genes that have high number of pairing possibilities and also common in OS, RFS and DMFS analysis.

Genes	No. of Gene-pairs			GEO RFS log-rank *P*	
	OS	RFS	DMFS	single covariate	multiple covariates
MEOX1	6	18	0	4.94E-08	3.43E-11 (C3orf18) ~ 4.02E-08
PPAP2B	37	14	0	1.27E-04	1.07E-10 (ADM) ~ 7.17E-05
PRPF38B	8	49	0	3.41E-02	1.04E-12 (DIRAS3) ~ 1.17E-02
SERPINA1	7	20	22	3.21E-05	2.30E-12 (CDT1) ~ 2.34E-05
XBP1	0	11	5	1.97E-03	2.47E-13 (DIRAS3) ~ 1.11E-03

Table 6. Five hub genes that associated with more than one type of survival data.

Gene-Pairs	OS		RFS		DMFS		DE Status of Gene 1	DE Status of Gene 2
	HR1	HR2	HR1	HR2	HR1	HR2		
C3orf18-PPAP2B	0.51	0.51	0.66	0.64	–	–	Her2+ Down	Elderly Down
IGF1R-KLRB1	0.39	0.60	0.72	0.73	–	–	ER+ Up / TNBC Down	IDC Down
NME5-PPAP2B	0.51	0.56	0.77	0.67	–	–	ER+ Up / PR+ Up / Her2+ Down / TNBC Down	Elderly Down
PRPF38B-RAMP3	0.56	0.58	0.80	0.62	–	–	Early Stage Up (i.e. Advanced Stage Down)	ER+ Up / PR+ Up / Her2+ Down / TNBC Down
GATA3-SERPINA1	0.62	0.41	–	–	0.56	0.56	ER+ Up / PR+ Up / TNBC Down	Young Down / ER+ Up / PR+ Up / Her2+ Down / TNBC Down
PSAT1-SERPINA1	1.61	0.44	–	–	1.53	0.56	Elderly Down / ER+ Down / PR+ Down / TNBC Up	Young Down / ER+ Up / PR+ Up / Her2+ Down / TNBC Down
MMP15-SLC44A4	–	–	1.29	0.75	1.69	0.60	Early Stage Down (i.e. Advanced Stage Up)	ER+ / TNBC Down

Table 7. The gene-pairs that associated with more than one type of survival data.

There were also seven gene-pairs that were significantly associated with more than one type of survival data (Table 7). By incorporating the differential expression information we derived previously, we may observe that the TNBC patients were noticeably having worse survival outcomes than non-TNBC patients as TNBC is known to be an aggressive breast cancer subtype [43, 44]. For example, both *GATA3* [45, 46] and *SERPINA1* were found significantly under-expressed in TNBC cases and the low expressions of both genes were correlated with poor OS and DMFS. Additionally, the over-expression of *PSAT1* and the under-expression of *SERPINA1* in TNBC patients also correlated with poor OS and DMFS. Moreover, the over-expression of *MMP15* relating to advanced stage breast cancer and the under-expression of *SLC44A4* associated with TNBC are predictors of cancer recurrence as well as distant metastases.

4. Conclusion and perspectives

In section 2 of this chapter, we identified 1,694 genes that were differentially expressed in breast cancer patients of three age groups, early versus advanced stage breast cancers, invasive ductal versus invasive lobular breast cancers, and patients of various receptor status. While some of these genes are known to participate in the biological and genetic pathways that lead to breast cancer and many are novel findings. In section 3, we showed that more than 20% the differentially expressed genes were associated with at least one type of survival data. Our data indicated improved predictive performance when using a multivariate approach of combining the expression of two genes in the assessment of survival data. Perceivably, the gene pairs found in the epistatic analysis could provide useful pictures in gene interactions in breast carcinogenesis.

Breast cancer is a heterogeneous and complex disease where researchers and doctors have implemented different classifications (be it molecular, pathological, genetic or prognostic) to aid disease diagnosis and treatment decision. In the future, we hope to use the gene expression profiles of multiple survival-associated biomarkers to sub-classify patients of different types of breast cancer, and ultimately allow medical practitioners to derive better disease assessment and treatment decision.

Acknowledgements

The authors acknowledge financial support from the Ministry of Science and Technology, Taiwan (MOST 103-2811-B-010-020), and from the Ministry of Education, Taiwan (Aim for the Top University Plan; 103AC-T903). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

I-Hsuan Lin and Ming-Ta Hsu*

*Address all correspondence to: mth@ym.edu.tw

Institute of Biochemistry and Molecular Biology, School of Life Science, National Yang-Ming University, Taipei, Taiwan

References

[1] Breast cancer: prevention and control. World Health Organization (WHO) Web site. http://www.who.int/cancer/detection/breastcancer/en/. Accessed June 2014. [http://www.who.int/cancer/detection/breastcancer/en/]

[2] Breast Cancer - Estimated Incidence, Mortality and Prevalence Worldwide in 2012. GLOBOCAN 2012 Fact Sheets. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=breast. Accessed June 2014. [http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=breast]

[3] Cristofanilli M, Gonzalez-Angulo A, Sneige N, Kau SW, Broglio K, Theriault RL, Valero V, Buzdar AU, Kuerer H, Buchholz TA, Hortobagyi GN: Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. *J Clin Oncol* 2005, 23:41-48.

[4] Reis-Filho JS, Simpson PT, Turner NC, Lambros MB, Jones C, Mackay A, Grigoriadis A, Sarrio D, Savage K, Dexter T, et al: FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas. *Clin Cancer Res* 2006, 12:6652-6662.

[5] Lehmann BD, Pietenpol JA: Identification and use of biomarkers in treatment strategies for triple-negative breast cancer subtypes. *J Pathol* 2014, 232:142-150.

[6] Fredholm H, Eaker S, Frisell J, Holmberg L, Fredriksson I, Lindman H: Breast cancer in young women: poor survival despite intensive treatment. *PLoS One* 2009, 4:e7695.

[7] Assi HA, Khoury KE, Dbouk H, Khalil LE, Mouhieddine TH, El Saghir NS: Epidemiology and prognosis of breast cancer in young women. *J Thorac Dis* 2013, 5 Suppl 1:S2-8.

[8] Clarke PA, te Poele R, Workman P: Gene expression microarray technologies in the development of new therapeutic agents. *European journal of cancer* 2004, 40:2560-2591.

[9] Midorikawa Y, Makuuchi M, Tang W, Aburatani H: Microarray-based analysis for hepatocellular carcinoma: from gene expression profiling to new challenges. *World journal of gastroenterology : WJG* 2007, 13:1487-1492.

[10] Lacroix L, Commo F, Soria JC: Gene expression profiling of non-small-cell lung cancer. *Expert review of molecular diagnostics* 2008, 8:167-178.

[11] Nannini M, Pantaleo MA, Maleddu A, Astolfi A, Formica S, Biasco G: Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. *Cancer treatment reviews* 2009, 35:201-209.

[12] Sorensen KD, Orntoft TF: Discovery of prostate cancer biomarkers by microarray gene expression profiling. *Expert review of molecular diagnostics* 2010, 10:49-64.

[13] Oostlander AE, Meijer GA, Ylstra B: Microarray-based comparative genomic hybridization and its applications in human genetics. *Clinical genetics* 2004, 66:488-495.

[14] Shih Ie M, Wang TL: Apply innovative technologies to explore cancer genome. *Current opinion in oncology* 2005, 17:33-38.

[15] Dupuy A, Simon RM: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 2007, 99:147-157.

[16] Morozova O, Hirst M, Marra MA: Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics* 2009, 10:135-151.

[17] Marguerat S, Bahler J: RNA-seq: from technology to biology. *Cellular and molecular life sciences : CMLS* 2010, 67:569-579.

[18] Costa V, Aprile M, Esposito R, Ciccodicola A: RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European journal of human genetics : EJHG* 2013, 21:134-142.

[19] Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005, 33:e175.

[20] Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8:118-127.

[21] Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26:139-140.

[22] Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C: jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 2014, 15:293.

[23] Chou YT, Wang H, Chen Y, Danielpour D, Yang YC: Cited2 modulates TGF-beta-mediated upregulation of MMP9. *Oncogene* 2006, 25:5547-5560.

[24] Sohn DM, Kim SY, Baek MJ, Lim CW, Lee MH, Cho MS, Kim TY: Expression of survivin and clinical correlation in patients with breast cancer. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 2006, 60:289-292.

[25] Naderi A, Meyer M: Prolactin-induced protein mediates cell invasion and regulates integrin signaling in estrogen receptor-negative breast cancer. *Breast cancer research : BCR* 2012, 14:R111.

[26] Noetzel E, Rose M, Bornemann J, Gajewski M, Knuchel R, Dahl E: Nuclear transport receptor karyopherin-alpha2 promotes malignant breast cancer phenotypes in vitro. *Oncogene* 2012, 31:2101-2114.

[27] Xu C, Wang H, He H, Zheng F, Chen Y, Zhang J, Lin X, Ma D, Zhang H: Low expression of TFPI-2 associated with poor survival outcome in patients with breast cancer. *BMC cancer* 2013, 13:118.

[28] Li C, Ma H, Wang Y, Cao Z, Graves-Deal R, Powell AE, Starchenko A, Ayers GD, Washington MK, Kamath V, et al: Excess PLAC8 promotes an unconventional ERK2-dependent EMT in colon cancer. *The Journal of clinical investigation* 2014, 124:2172-2187.

[29] Vertino PM, Yen RW, Gao J, Baylin SB: De novo methylation of CpG island sequences in human fibroblasts overexpressing DNA (cytosine-5-)-methyltransferase. *Molecular and cellular biology* 1996, 16:4555-4565.

[30] Sun L, Hui AM, Kanai Y, Sakamoto M, Hirohashi S: Increased DNA methyltransferase expression is associated with an early stage of human hepatocarcinogenesis. *Japanese journal of cancer research : Gann* 1997, 88:1165-1170.

[31] Damiani LA, Yingling CM, Leng S, Romo PE, Nakamura J, Belinsky SA: Carcinogen-induced gene promoter hypermethylation is mediated by DNMT1 and causal for transformation of immortalized bronchial epithelial cells. *Cancer Res* 2008, 68:9005-9014.

[32] Zhao M, Sun J, Zhao Z: TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res* 2013, 41:D970-976.

[33] Bauer JA, Chakravarthy AB, Rosenbluth JM, Mi D, Seeley EH, De Matos Granja-Ingram N, Olivares MG, Kelley MC, Mayer IA, Meszoely IM, et al: Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2010, 16:681-690.

[34] Ray PS, Wang J, Qu Y, Sim MS, Shamonki J, Bagaria SP, Ye X, Liu B, Elashoff D, Hoon DS, et al: FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res* 2010, 70:3870-3876.

[35] Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, Gaetje R, Solbach C, Ahr A, Metzler D, et al: A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast cancer research : BCR* 2011, 13:R97.

[36] Umekita Y, Ohi Y, Souda M, Rai Y, Sagara Y, Tamada S, Tanimoto A: Maspin expression is frequent and correlates with basal markers in triple-negative breast cancer. *Diagnostic pathology* 2011, 6:36.

[37] Beltran AS, Graves LM, Blancafort P: Novel role of Engrailed 1 as a prosurvival transcription factor in basal-like breast cancer and engineering of interference peptides block its oncogenic function. *Oncogene* 2013.

[38] Jenssen TK, Kuo WP, Stokke T, Hovig E: Associations between gene expressions in breast cancer and patient survival. *Hum Genet* 2002, 111:411-420.

[39] Gordon GJ, Richards WG, Sugarbaker DJ, Jaklitsch MT, Bueno R: A prognostic test for adenocarcinoma of the lung from gene expression profiling data. *Cancer Epidemiol Biomarkers Prev* 2003, 12:905-910.

[40] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, et al: Gene expression predictors of breast cancer outcomes. *Lancet* 2003, 361:1590-1596.

[41] Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003, 100:10393-10398.

[42] Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD: Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med* 2006, 3:e13.

[43] Bernardi R, Gianni L: Hallmarks of triple negative breast cancer emerging at last? *Cell Res* 2014, 24:904-905.

[44] Al-Ejeh F, Simpson PT, Sanus JM, Klein K, Kalimutho M, Shi W, Miranda M, Kutasovic J, Raghavendra A, Madore J, et al: 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* 2014, 3:e100.

[45] Yoon NK, Maresh EL, Shen D, Elshimali Y, Apple S, Horvath S, Mah V, Bose S, Chia D, Chang HR, Goodglick L: Higher levels of GATA3 predict better survival in women with breast cancer. *Human pathology* 2010, 41:1794-1801.

[46] Chu IM, Michalowski AM, Hoenerhoff M, Szauter KM, Luger D, Sato M, Flanders K, Oshima A, Csiszar K, Green JE: GATA3 inhibits lysyl oxidase-mediated metastases of human basal triple-negative breast cancer cells. Oncogene 2012, 31:2017-2027.