Investigating Unique Genes of Five Molecular Subtypes of Breast Cancer Using Penalized Logistic Regression

sadegh raoufi (raoufi.sadegh@gmail.com)
Kerman University of Medical Sciences
Saeideh Jafarinejad Farsangi
Kerman University of Medical Sciences
Tania Dehesh
Kerman University of Medical Sciences
Morteza Hadizadeh
Kerman University of Medical Sciences

Research Article

Keywords: Breast cancer, Gene expression, Lasso logistic regression, Adaptive Lasso logistic regression

Posted Date: February 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-249085/v1

Investigating unique genes of five molecular subtypes of breast cancer using penalized logistic regression

Sadegh Raoufi[1], Saeideh Jafarinejad Farsangi[2], Tania Dehesh[3*] & Morteza Hadizadeh[4]

[1]Modeling in Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran.

[2]Physiology Research Center, Institute of Basic and Clinical Physiology Sciences, Kerman University of Medical Sciences, Kerman, Iran.

[3]Department of Epidemiology and Biostatistics, School of Public Health, Kerman University of Medical Sciences, Kerman, Iran.

[4]Cardiovascular Research Centre, Institute of Basic and Clinical Physiology Sciences, Kerman University of Medical Sciences, Kerman, Iran.

Corresponding author: Tania Dehesh, Assistant Professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Kerman University of Medical Sciences, Kerman, Iran. Tel: +983431325069, Email: Tania_dehesh@yahoo.com.

Abstract

Background: Breast cancer is the first cancer and fifth cause of death in women around the world. Exploring unique genes for cancers has become interesting. The aim of this study was to explore unique gens of five molecular subtypes of breast cancer in women using penalized logistic regression models.

Methods: In this study, microarray data of five independent GEO datasets was combined. This combination includes genetic information of 324 women with breast cancer and 12 healthy women. Lasso logistic regression and adaptive lasso logistic regression were used to extract unique genes. Biological process of extracted gens was evaluated in open-source GOnet web-application. R software version 3.6.0 with glmnet package was used for fitting the models.
Results: Totally, 119 genes were extracted among fifteen pairwise comparisons. 17 genes (%14) had overlap between comparative groups. Among 27 genes contributed in positive regulation of cell processes, one gene belonged exclusively to this biological process. Among 46 genes contributed in negative regulation of cell processes, 6 genes belonged exclusively. Among 50 genes that were significant in regulation of metabolism, 4 genes belonged exclusively. Among 32 genes that related to response of stress, 4 genes belonged exclusively.
Conclusions: The most genes selected by lasso logistic regression and adaptive Lasso logistic regression, were diagnosed in negative regulation of cell processes.

Keywords: Breast cancer; Gene expression; Lasso logistic regression; Adaptive Lasso logistic regression

Background

Breast cancer (BC) is the most common cancer worldwide among women [1]. Breast cancer comprises for about 18% of all types of cancers in women and is the fifth reason for death around the world [2]. Several factors are associated with this disease such as genetics, lifestyle, menstrual and reproductive history, long term treatment with estrogens. Genetic abnormalities include germ line mutations, gene amplifications, rearrangements, overexpression, deletions or point mutation [3]. Some changes in the structure and expression of genes can keep cells out of the control of normal growth and turn them into cancer.

During the last two decades, microarray data sets have been helping biological researchers to study the expression of thousands of genes simultaneously [4, 5]. Recently, microarray analysis has revealed new molecular subtypes behind the classical classification of BC. Gene expression profiling in tumor tissues has been shown that breast cancers may be divided into subtypes consisting of two estrogen receptor (ER) positive types (luminal A and luminal B) and three estrogen receptor negative types [human epidermal growth factor receptor 2 (HER2 or ERBB2), triple-negative (TNBC)/basal-like, and unclassified (normal-like)] with distinctive clinical outcomes [6, 7]. These subtypes are clinically meaningful, and can divide patients into groups with distinct tumor morphologies and outcomes [8]. Although the molecular classification of BC has improved treatment outcomes for patients, the heterogeneity among differentially expressed genes requires statistical modeling for selection of unique genes. From the viewpoint of biologists, selection of unique genes can help to find a specific molecular mechanism in each subgroup and improves staging accuracy by elimination irrelevant and noisy genes [9-11]. However, selection of genes with DNA microarray data is a challenge for researchers, because of high number of genes and small number of patients (high dimensional data) [12]. Recently, statistical methods called regularization regression methods have been introduced to solve this problem [13]. The most important regularized models are Least absolute shrinkage and selection operator (LASSO) [14] and Adaptive LASSO [15]. The use of these models in microarray datasets leads to the selection of important genes and the elimination of insignificant genes. Algmal and Lee (2015) proposed regularization regression for gene selection in a microarray analysis that contain three dataset (colon dataset with 2000 genes, prostate dataset with 5966 genes and DLBCL dataset with 7129 genes) [16]. Mostafaei et al. (2018) used Machine-Based Learning Algorithms on 20,097 probes were generated from a small airway epithelium microarray dataset for identification of Novel Genes in Human Airway Epithelial Cells associated with Chronic Obstructive Pulmonary Disease [17].

The present study aimed to employ regularized logistic regression methods called LASSO logistic regression and adaptive LASSO logistic regression in order to introduce unique genes for fifteen comparatives between subtypes of breast cancer in women.

Methods

Data collection

The raw data of gene expression profiles was downloaded from the ncbi.nlm.nih.gov website. Among the searches performed, the data considered include enough sample size and the five molecular subtypes in breast cancer and also data contained suitable power. Four independent GEO datasets from independent studies were combined in order to have a comprehensive file. Each of the four studies included microarray datasets of breast cancer in different subtypes. GSE accession number was GSE1456, GSE43358, GSE50428 and GSE57297. The five molecular subtypes in breast cancer in order of disease severity include: Basal, ERBB2, Luminal B, Luminal A and Normal like respectively. The final data file was consisted of 324 patients with breast cancer in five molecular subtypes and 12 normal tissues. The sample size of normal group was small because the microarray data of normal breast tissue was not available compare to cancer tissue in previous studies. Table 1 shows the sample size of five molecular subtypes of breast cancer and control group.

Table 1 The sample size of all four GSE Conducted on five subtypes in breast cancer

subtype	size
Basal	56
ERBB2	35
Luminal B	54
Luminal A	102
Normal like	77
Control	12

Microarray data processing

Samples processes were analyzed with different chip platform. The raw datasets of microarray GSE were handled through the BioConductor affy package in R software [18]. A sensitive step in integration of heterogeneous data is normalization, Therefore, before merging, each dataset were normalized using Limma package in R software [19]. During combining different GSE data, non-biological batch effects were removed. Different studies were done with different procedures and platform. The adjustment was performed to remove non-biological batches and to save meaningful biological effect in combination of datasets. In order to remove non-biological batch effects Surrogate Variable Analysis (SVA) package in R was used [20]. Batch effect removal was checked by PCA and boxplot. The outcome of these data combination was

a unit expression matrix (the combination of four datasets of this study) and then extracted differentially expressed genes (DEGs) from the matrix. Data processing and integrating were performed using R. In order to identify biological process, the DEGs was entered in open-source GOnet web-application (available at http://tools.dice-database.org/GOnet/), with p-value≤0.05 in enrichment analysis options [21]. KEGG pathway analysis was performed by ToppGene Suit (http://toppgene.cchmc.org) web site.

Statistical analysis

Logistic regression is a statistical method to investigate the effect of predictor variables on dichotomous variable (outcome). Let $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ for $i = 1,2, \dots, N$ denote p-predictors for N observation. Assume that responses for the binary logistic regression model can take values $G = 1,2$. Then,

$$Pr = (G = 1|x) = \frac{1}{1+e^{-(\beta_0+\boldsymbol{x_i}^T\boldsymbol{\beta})}}, Pr = (G = 2|x) = \frac{1}{1+e^{(\beta_0+\boldsymbol{x_i}^T\boldsymbol{\beta})}} \quad (1)$$

Where $\beta_0$is the intercept and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is a p-vector of regression parameters. In each pairwise comparison between subtypes, patients in one subtypes coded one and the other coded zero. This dichotomous variable was used as outcome variable in each analysis. Predictor variables are genes. In the presence of several predictors (genes) with a few sample size (a few cases), Logistic regression was not applicable. This situation could be managed with Regularized logistic regression methods, which are able to select effective predictors by shrinking unimportant regression coefficients toward zero. Two famous regularized logistic regressions are: LASSO logistic regression and Adaptive LASSO logistic regression.

LASSO logistic regression

LASSO proposed by Tibshirani [14], is one of the popular regularization method. In LASSO have used $L_1 - norm$ regularization for gene selection and estimation simultaneously by constraining the negative log-likelihood function of gene coefficients. Finally LASSO method then finds parameter values to minimize:

$$-\left[\frac{1}{N}\sum_{i=1}^{N} y_i.(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta}) - \log\left(1 + e^{(\beta_0+\boldsymbol{x_i}^T\boldsymbol{\beta})}\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j| \quad (2)$$

Where λ is a tuning parameter and $\lambda\sum_{j=1}^{p}|\beta_j|$ is the penalty function for the LASSO.

Adaptive LASSO logistic regression

The LASSO method has shown to not always provide consistent variable selection. The LASSO penalizes all coefficients equally, even when the coefficients are large. The Adaptive lasso

regression is the extended of Lasso logistic regression. This method assigns small weight to unimportant coefficients and large weight to important coefficients in order to have more accurate coefficients. Adaptive LASSO method then finds parameter values to minimize:

$$-\left[\frac{1}{N}\sum_{i=1}^{N} y_i.\left(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta}\right) - \log\left(1 + e^{(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta})}\right)\right] + \lambda\sum_{j=1}^{p} w_j\left|\beta_j\right| \qquad (3)$$

Where $w_j = \frac{1}{|\hat{\beta}_j|^v}$, $\left|\hat{\beta}_j\right|$ is the maximum likelihood estimate and $v > 0$. the weighted penalty will allow variables with larger coefficients to receive smaller penalties and thus might provide a more consistent solution.

In the end Gene comparisons in each pair of breast cancer subtype were done by both models and common genes that were extracted with both models simultaneously. Since in current study there were five subtypes of breast cancer and one control group, therefore fifteen pairs comparison were done in the R software using the Glmnet package (https://cran.r-project.org/package=glmnet).

Results

In this study after Microarray data processing and integrative meta-analysis, 3948 genes were screened. Lasso and adaptive lasso logistic regression applied on 3948 genes. After fitting two models, some coefficients can become zero and eliminated. Table 2 shows common non-zero gene coefficients that were extracted by two models simultaneously for five molecular subtypes of breast cancer. In total, 102 genes extracted with two models in 10 comparisons between 5 subtypes. 85 genes are unique between comparative groups and they were mentioned by highlights in each comparative columns of table 2. For example, IL17RB gene is unique for comparison between basal and ERBB2. 17 genes were repeated in more than one comparative group. For example, ARSD gene was extracted in comparison groups of ERBB2 vs. basal and normal like vs. Basal.

The comparisons between each subtype and control group were shown in Table 3. Among the final 21 genes, 17 genes were unique between comparative groups and they were highlighted. For example, CXorf36 gene was unique in comparison between basal and control. Four genes (GLTSCR2, RUNX1T1, TLE3 and EGFR) had overlap with table 2. Consequently, 119 candidate genes (102 genes in table 2 and 17 genes in table 3) were associated with either the occurrence or progression of breast cancer. Totally 102 unique genes exist (85 genes in table 2 and 17 genes in table 3). 17 genes (%14) showed overlap between their comparative groups: ARSD, ENTPD5, ERBB2, TLE3, TMEM57, DDX21, EDEM3, EGFR, REL, HSP90AB1, DHTKD1, LHFP, MCC, MYLK, RBMS3, GLTSCR2 and RUNX1T1.

According to GOnet database analysis (Fig. 1), 27 genes contributed in positive regulation and 46 genes contributed in negative regulation of cell processes (showed by green color), 32 genes were associated with response to stress and 50 genes were significant in regulation of metabolism. One gene belonged exclusively to the positive regulation of cell processes (APH1B). 6 genes belonged exclusively to the negative regulation of cell processes (SEMA4F, MCC, NUDT6, SEMA5A, MED28 and TMBIM4). 4 genes belonged exclusively to the response to stress (SORD, MYLK, EDEM3 and IL17RB). 4 genes belonged exclusively to the regulation of metabolism (ENTPD5, CCT5, SMU1 and ZNF652). 13 genes were common in all biological processes (CDC14B, MAP4K4, EGFR, ERBB2, HSP90AB1, PRKAR1A, CR1, TGFBR1, SLC11A1, PRKAA1, SMAD4, TERF2 and RBMS3). KEGG pathway analysis (table 4) showed cancer related pathways such as MAPK, mTOR, FoxO, Wnt signaling. We also observed the adherens junction, cytokine-cytokine receptor interaction and cell adhesion molecules which are related to cell connections that are important in cancer.

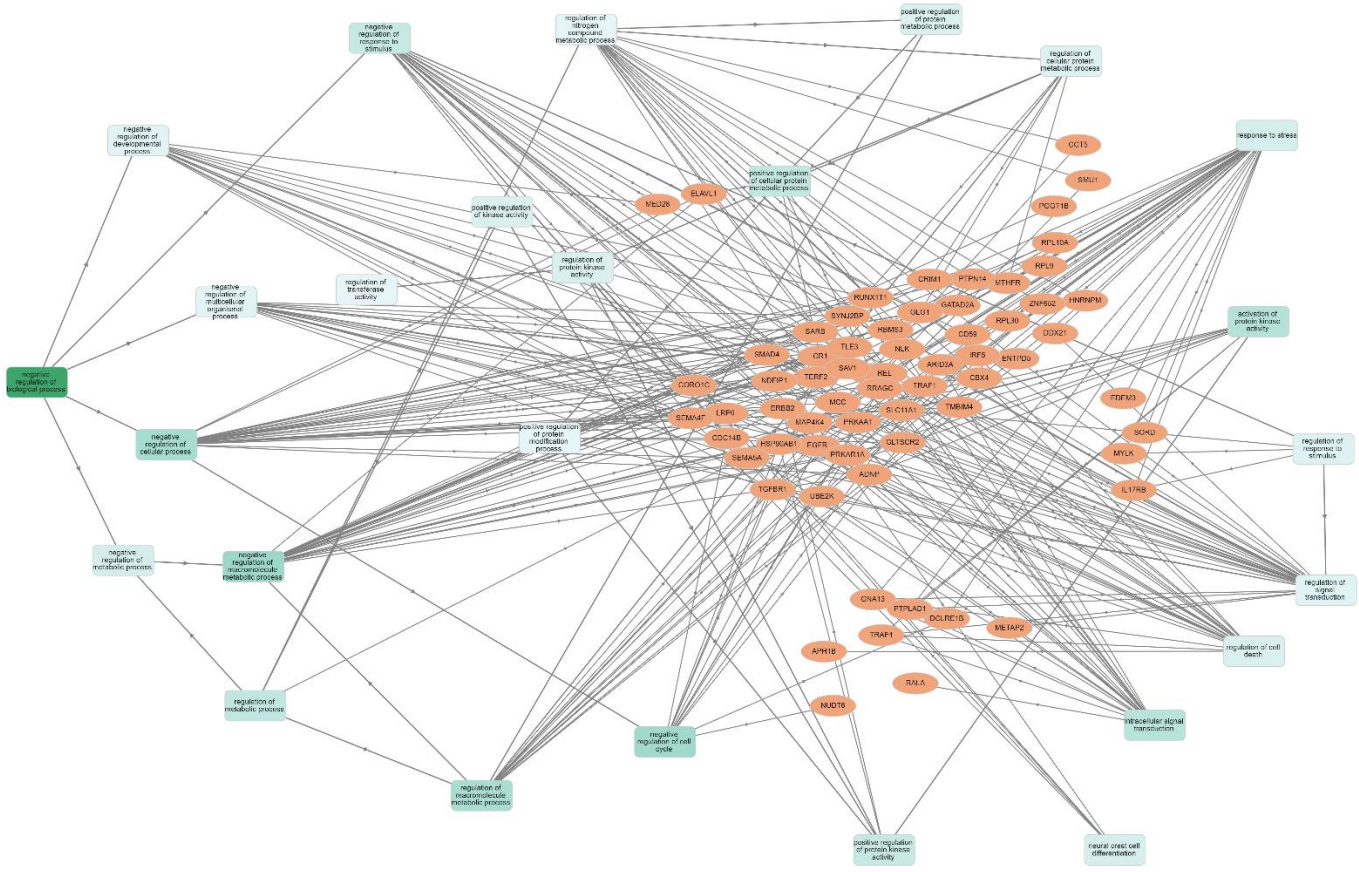

Fig. 1 Biological process of extracted genes in open-source GOnet web-application with q-value ≤ 0.05 in enrichment analysis options.

Table 2 Genes coefficients across ten comparisons of breast cancer subtypes (Unique genes in each two subtypes comparison are shown by highlights)

	ERBB2 vs. Basal		Luminal A vs. Basal		Luminal B vs. Basal		Normal like vs. Basal		Luminal A vs. ERBB2		Luminal B vs. ERBB2		Normal like vs. ERBB2		Luminal B vs. Luminal A		Normal like vs. Luminal A		Normal like vs. Luminal B	
Genes	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso
ARSD	-0.9	-1.6					-0.2	-0.1												
ENTPD5	-0.5	-0.8					-0.2	-0.1												
ERBB2	-0.6	-0.3									0.4	0.1								
IL17RB	0.1	0.1																		
NLK	-0.5	-1.1																		
SMU1	0.0	0.4																		
SORD	-0.2	-0.1																		
TGFBR1	-0.0	-0.6																		
TLE3	-0.0	-0.9	-0.8	-2.5	-0.1	-1.6														
AMACR	-0.1	-0.1																		
MTHFR	-0.6	-0.8																		
SEMA4F	0.7	0.6																		
TMEM57	-0.8	-0.1																		
DDX21			0.3	0.2			0.1	0.8												
RBMS1			0.3	0.4																
CBX4			-0.3	-0.5																
EDEM3			-0.2	-1.0													0.4	0.2		
RPL9			-0.4	-0.3																
UBE2K			-0.7	-1.3																
ARF3					-0.4	-1.5														
CD59					0.5	0.6														
MPZL2					0.3	0.4														
PDXK					0.3	1.6														
PGGT1B					-0.2	-0.1														
PRKAR1A					-0.3	-0.1														
PTPLAD1					-0.7	-1.3														
PTPN14					0.5	0.5														
RPS6KB1					-0.6	-0.5														
RRAGC					0.2	1.0														
CORO1C							0.6	0.3												
CR1							0.6	0.0												
DCLRE1B							0.5	0.5												
KLHL28							-0.4	-0.3												
SYNJ2BP							-0.5	-2.1												
TERF2							1.3	1.6												
TMBIM4							-0.7	-2.7												

ZNF652							-0.5	-0.7												
GLTSCR2							-0.0	-0.9												
APH1B									-0.5	-0.8										
EGFR									0.2	0.7	0.3	1.9					-0.8	-2.0		
METAP2									0.9	1.5										
NUDT6									-0.6	-0.8										
PCCA									0.1	0.5										
RALA									0.7	1.8										
REL									-0.3	-0.1							0.6	1.3		
RUNX1T1									0.4	0.6										
TRAF4									0.1	0.7										
CDC14B									0.2	0.3										
ADNP											-0.5	-0.2								
HNRNPM											-0.8	-2.7								
HSP90AB1											-0.4	-0.4							0.1	0.8
MAP4K4											0.6	1.4								
MED28											-0.6	-3.3								
PRKAA1											0.1	0.7								
RPL30											-0.5	1.8								
VPS8											-0.3	-0.5								
ELAVL1											-0.2	-0.3								
NDFIP1											-0.5	-0.7								
DHTKD1													0.5	1.0			0.7	0.6		
GATAD2A													0.9	0.5						
GUK1													0.3	2.4						
LHFP													-1.0	-1.2					-1.1	-0.3
MCC													-0.3	-0.2					-0.1	-0.2
MYLK													-0.5	-0.7			-1.3	-0.9	-1.2	-1.3
POMP													1.4	2.3						
RBMS3													-1.4	-0.5					-0.7	-0.2
RPL10A													-0.3	-1.1						
SARS													-0.7	-3.0						
SMAD4													-1.1	-1.5						
SPAG7													-0.1	-2.0						
CCT5															0.4	1.6				
CRIM1															-0.7	-1.2				
FAM114A1															-0.6	-1.0				
GNA13															0.2	0.3				
OSMR															-0.1	-0.6				
RPE															0.5	1.0				
SERBP1															0.1	0.6				

WHSC1L1															0.3	0.5				
ATF6															-0.2	-1.0				
BCAN															-0.2	-0.2				
CNOT6															0.2	0.1				
COPZ1															-1.0	-1.1				
JTB															-0.2	-0.1				
KPNA4															0.2	0.1				
NUCKS1															-0.2	-0.2				
SEC63															0.3	0.5				
STK4															0.1	0.2				
BMPR1B																	0.1	-1.4		
CLDN11																	-0.4	-0.6		
LTBP2																	-0.9	-1.8		
RPL4																	-0.5	-0.7		
SLC38A2																	-0.6	-2.5		
TCF7L2																	-0.4	-0.3		
TRIOBP																	-0.6	-1.7		
IFNGR1																	-0.1	-0.9		
NDE1																	0.3	0.4		
C20orf24																			0.3	1.3
EEF2																			-0.8	-1.8
GLI2																			-0.1	-0.6
METTL4																			0.4	0.4
PGK1																			0.5	0.8
SGCD																			-1.2	-1.0

Table 3 Genes coefficients for comparison between control group and breast cancer subtypes (Unique genes in each two subtypes comparison are shown by highlights)

	Basal vs. control		ERBB2 A vs. control		Luminal A vs. control		Luminal B vs. control		Normal like vs. control	
Genes	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso	Lasso	A.lasso
CXorf36	-0.6	-0.5								
ECHDC2	0.6	-1.4								
GLTSCR2	-2.5	-4.0								
IRF5	0.3	0.3								
NPL	0.5	0.8								
SLC11A1	1.0	3.0								
TMEM206	0.1	1.9								
TRAF1	0.3	0.5								
LRP6			-0.8	-0.5						
SAV1			-0.8	-2.4						
RUNX1T1					-0.4	-0.7				
SEMA5A					-1.0	-2.2				
SOBP					-0.1	-0.3				
TLE3					0.9	2.1				
ARID3A							0.9	1.2		
EGFR							-0.7	-3.6		
GLG1							-1.1	-2.7		
GNAS									-1.1	-1.6
MDN1									-1.0	-2.2
SLC44A1									0.9	0.5
SRP14									2.3	8.6

Table 4 KEGG pathway analysis for unique genes of fifteen comparative group

Comparative group	Pathway	Source	q-value FDR B&H	Hit Count in Query List	Hit Count in Genome	Hit in Query List
ERBB2 vs. Basal	Adherens junction	KEGG	0.021	2	72	NLK,TGFBR1
	FoxO signaling pathway	KEGG	0.035	2	132	NLK,TGFBR1
	Bile acid biosynthesis, cholesterol => cholate/chenodeoxycholate	KEGG	0.049	1	12	AMACR
	MAPK signaling pathway	KEGG	0.049	2	255	NLK,TGFBR1
	Cytokine-cytokine receptor interaction	KEGG	0.049	2	270	IL17RB,TGFBR1
	Primary bile acid biosynthesis	KEGG	0.049	1	17	AMACR
	One carbon pool by folate	KEGG	0.050	1	20	MTHFR
Luminal A vs. Basal	Ubiquitin mediated proteolysis	KEGG	0.037	1	137	UBE2K
	Ribosome	KEGG	0.037	1	154	RPL9
Luminal B vs. Basal	Autophagy - animal	KEGG	0.020	2	128	RPS6KB1,RRAGC
	Insulin signaling pathway	KEGG	0.020	2	138	PRKAR1A,RPS6KB1
	mTOR signaling pathway	KEGG	0.020	2	151	RPS6KB1,RRAGC
	Vitamin B6 metabolism	KEGG	0.020	1	6	PDXK
Normal like vs. Basal	Malaria	KEGG	0.028	1	49	CR1
	Legionellosis	KEGG	0.028	1	55	CR1
	Leishmaniasis	KEGG	0.028	1	73	CR1
	Complement and coagulation cascades	KEGG	0.028	1	79	CR1
	Hematopoietic cell lineage	KEGG	0.028	1	97	CR1
	Tuberculosis	KEGG	0.043	1	179	CR1
Luminal A vs. ERBB2	Pathways in cancer	KEGG	0.018	3	395	TRAF4,RALA,RUNX1T1
	Propanoyl-CoA metabolism, propanoyl-CoA => succinyl-CoA	KEGG	0.020	1	4	PCCA
Luminal B vs. ERBB2	AMPK signaling pathway	KEGG	0.030	2	121	PRKAA1,ELAVL1
Normal like vs. ERBB2	-	-	-	-	-	-

Luminal B vs. Luminal A	-	-	-	-	-	-
Normal like vs. Luminal A	-	-	-	-	-	-
Normal like vs. Luminal B	Glycolysis, core module involving three-carbon compounds	KEGG	0.047	1	12	PGK1
	Gluconeogenesis, oxaloacetate => fructose-6P	KEGG	0.047	1	17	PGK1
	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	KEGG	0.047	1	25	PGK1
	Hedgehog signaling pathway	KEGG	0.047	1	47	GLI2
	Basal cell carcinoma	KEGG	0.047	1	55	GLI2
	Viral myocarditis	KEGG	0.047	1	59	SGCD
	Glycolysis / Gluconeogenesis	KEGG	0.047	1	67	PGK1
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	KEGG	0.047	1	72	SGCD
	Biosynthesis of amino acids	KEGG	0.047	1	75	PGK1
	Hypertrophic cardiomyopathy (HCM)	KEGG	0.047	1	83	SGCD
	Dilated cardiomyopathy	KEGG	0.047	1	90	SGCD
	HIF-1 signaling pathway	KEGG	0.048	1	101	PGK1
	Carbon metabolism	KEGG	0.049	1	114	PGK1
	AMPK signaling pathway	KEGG	0.049	1	121	EEF2
Basal vs. Control	-	-	-	-	-	-
ERBB2 A vs. control	Hippo signaling pathway -multiple species	KEGG	0.023	1	29	SAV1
	Wnt signaling pathway	KEGG	0.025	1	143	LRP6
	Breast cancer	KEGG	0.025	1	144	LRP6
	mTOR signaling pathway	KEGG	0.025	1	151	LRP6
	Hippo signaling pathway	KEGG	0.025	1	154	SAV1
Luminal A vs. control	Axon guidance	KEGG	0.014	1	175	SEMA5A
Luminal B vs. control	Cell adhesion molecules (CAMs)	KEGG	0.023	1	145	GLG1
	Protein export	KEGG	0.048	1	23	SRP14
	Vasopressin-regulated water reabsorption	KEGG	0.048	1	44	GNAS
	Endocrine and other factor-regulated calcium reabsorption	KEGG	0.048	1	47	GNAS
	Cocaine addiction	KEGG	0.048	1	49	GNAS
	Ovarian steroidogenesis	KEGG	0.048	1	50	GNAS
	Vibrio cholerae infection	KEGG	0.048	1	51	GNAS

Normal like vs. control	Regulation of lipolysis in adipocytes	KEGG	0.048	1	54	GNAS
	Long-term depression	KEGG	0.048	1	60	GNAS
	Renin secretion	KEGG	0.048	1	65	GNAS
	Amphetamine addiction	KEGG	0.048	1	68	GNAS
	Bile secretion	KEGG	0.048	1	71	GNAS
	Thyroid hormone synthesis	KEGG	0.048	1	74	GNAS
	Gastric acid secretion	KEGG	0.048	1	75	GNAS
	Aldosterone synthesis and secretion	KEGG	0.048	1	82	GNAS
	Insulin secretion	KEGG	0.048	1	85	GNAS
	Gap junction	KEGG	0.048	1	88	GNAS
	Salivary secretion	KEGG	0.048	1	90	GNAS
	Dilated cardiomyopathy	KEGG	0.048	1	90	GNAS
	Morphine addiction	KEGG	0.048	1	91	GNAS
	GnRH signaling pathway	KEGG	0.048	1	92	GNAS
	Pancreatic secretion	KEGG	0.048	1	96	GNAS
	Circadian entrainment	KEGG	0.048	1	96	GNAS
	Amoebiasis	KEGG	0.048	1	96	GNAS
	Endocrine resistance	KEGG	0.048	1	96	GNAS
	Inflammatory mediator regulation of TRP channels	KEGG	0.048	1	97	GNAS
	Estrogen signaling pathway	KEGG	0.048	1	98	GNAS
	Choline metabolism in cancer	KEGG	0.048	1	99	SLC44A1
	Melanogenesis	KEGG	0.048	1	101	GNAS
	Chagas disease (American trypanosomiasis)	KEGG	0.048	1	102	GNAS
	Glucagon signaling pathway	KEGG	0.048	1	103	GNAS
	Ribosome biogenesis in eukaryotes	KEGG	0.048	1	106	MDN1
	Serotonergic synapse	KEGG	0.048	1	113	GNAS
	Glutamatergic synapse	KEGG	0.048	1	114	GNAS
	Vascular smooth muscle contraction	KEGG	0.049	1	121	GNAS
	Platelet activation	KEGG	0.049	1	123	GNAS

Discussion

The present study was performed in order to combine differentially expressed genes (DEGs) from different microarray studies and identify important and unique genes for five molecular subtype of breast cancer using penalized logistic regression models. Unique genes have been selected with minimal overlapping between comparative groups. The extracted genes belonged to different biological processes including negative and positive regulation of cell processes, regulation of metabolism and response to stress. According to our results, no significant hub genes were identified, but most genes selected by two models related to the negative regulation of cell processes.

Among the selected genes, MYLK [22], TLE3 [23] and LHFP [24] have been previously reported to be expressed in all breast cancer subtypes. We observed that Epidermal growth factor receptor (EGFR) [25] was highlighted in luminal A group while its role as a biomarker was investigated in triple negative/basal subtype. Erb-B2 receptor tyrosine kinase (ERBB2) is a strong prognostic biomarker and overexpresses in about 15-30% of breast tumors [26]. Inhibition of this receptor is the main therapeutic strategy in ERBB2-positive subtypes and we observed it in ERBB2 group compared to Basal and Luminal B groups. Furthermore, our model, classified most DEG to specific comparative groups without overlapping. ER degradation enhancing alpha-mannosidase like protein 3 (EDEM3) is a protein involved in glycan degradation which is a sign of tumor malignancy [27]. Seifi-Alan et al. has shown differential expression of EDEM3 in Luminal A subtype [28]. Our model highlighted EDEM3 in Luminal A-Basal comparative group.

myelin protein zero like-2 (MPZL2) is one of the PACE4 targets which is a proprotein convertase and play a key role in tumor cell migration and malignancy especially in basal breast tumors [29]. We observed MPZL2 in luminal B-basal comparative group. Coronin-1C another gene with differential expression in normal-like versus basal comparative group, have been shown to be the direct target of Y-box binding protein-1 (YB-1) protein. YB-1 down regulates the expression of CORO1C and promotes invasion, migration and metastasis of triple negative/basal breast tumor cells [30].

According to the results, it seems regularized logistic regression methods called LASSO logistic regression and adaptive LASSO logistic regression can select important and unique genes which show a good consistency with experimental studies several limitations of this study should be considered. First, we can refer to the small number of samples in control group due to the nature of the data. Second, our data was extracted from different studies hence precision is not equal in different countries then we tried with normalization to resolve the heterogeneity.

Conclusion

The present study, performed in order to combine differentially expressed genes (DEG) from different microarray studies to improve heterogeneity of DEG list among different studies and identify important and unique genes for five molecular subtypes of breast cancer using penalized logistic regression models. Penalized logistic regression models are efficient for extracting genes in high-dimensional genetic studies. The result of this study did not show specific hub genes but most genes founded by two models related to the negative regulation of cell processes.

Abbreviations

GEO: Gene Expression Omnibus; GSE: Genetics Selection Evolution; DEG: Differentially Expressed Genes; BC: Breast Cancer; LASSO: Least Absolute Shrinkage and Selection Operator.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data used to run the models and the datasets generated during the current study are available in the GEO (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE1456, GSE43358, GSE50428 and GSE57297.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Authors' Contributions

T.D. conceived and designed the study. M.H. collected data, performed Microarray analysis, and prepared the figure. S.R. performed the statistical analysis and wrote the paper. S.J.F. wrote the discussion of the paper. S.J.F. and T.D. reviewed and revised the manuscript. All authors have read and approved the manuscript for publication.

Acknowledgements

Not applicable.

References

1. Azamjah, N., Y. Soltan-Zadeh, and F. Zayeri, *Global Trend of Breast Cancer Mortality Rate: A 25-Year Study.* Asian Pacific journal of cancer prevention: APJCP, 2019. **20**(7): p. 2015.
2. Zhang, Y., B. Zhang, and W. Lu, *Breast cancer histological image classification with multiple features and random subspace classifier ensemble*, in *Knowledge-Based Systems in Biomedicine and Computational Life Science*. 2013, Springer. p. 27-42.
3. Tekmal, R.R. and N. Keshava, *Role of MMTV integration locus cellular genes in breast cancer.* Front Biosci, 1997. **2**: p. d519-526.
4. Kalina, J., *Classification methods for high-dimensional genetic data.* Biocybernetics and Biomedical Engineering, 2014. **34**(1): p. 10-18.
5. Kastrin, A. and B. Peterlin, *Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data.* Expert Systems with Applications, 2010. **37**(7): p. 5178-5185.
6. Perou, C.M., et al., *Molecular portraits of human breast tumours.* nature, 2000. **406**(6797): p. 747.
7. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(14): p. 8418-8423.
8. Nguyen, P.L., et al., *Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy.* Journal of clinical oncology, 2008. **26**(14): p. 2373-2378.
9. Kamkar, I., et al., *Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso.* Journal of biomedical informatics, 2015. **53**: p. 277-290.
10. Yu, L., Y. Han, and M.E. Berens, *Stable gene selection from microarray data via sample weighting.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011. **9**(1): p. 262-272.
11. Peng, H., et al., *Optimal gene subset selection using the modified SFFS algorithm for tumor classification.* Neural Computing and Applications, 2013. **23**(6): p. 1531-1538.
12. Chen, S.X. and Y.-L. Qin, *A two-sample test for high-dimensional data with applications to gene-set testing.* The Annals of Statistics, 2010. **38**(2): p. 808-835.
13. Nan, X., et al., *Biomarker discovery using 1-norm regularization for multiclass earthworm microarray gene expression data.* Neurocomputing, 2012. **92**: p. 36-43.
14. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
15. Zou, H., *The adaptive lasso and its oracle properties.* Journal of the American statistical association, 2006. **101**(476): p. 1418-1429.

16. Algamal, Z.Y. and M.H. Lee, *Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification.* Expert Systems with Applications, 2015. **42**(23): p. 9326-9332.
17. Mostafaei, S., et al., *Identification of Novel Genes in Human Airway Epithelial Cells associated with Chronic Obstructive Pulmonary Disease (COPD) using Machine-Based Learning Algorithms.* Scientific reports, 2018. **8**(1): p. 1-20.
18. Gautier, L., et al., *affy—analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-315.
19. Altman, N.S., *Differential Expression Analysis using LIMMA.* 2013.
20. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-883.
21. Pomaznoy, M., B. Ha, and B. Peters, *GOnet: a tool for interactive Gene Ontology analysis.* BMC bioinformatics, 2018. **19**(1): p. 470.
22. Hu, K., et al., *Small interfering RNA library screen identified polo-like kinase-1 (PLK1) as a potential therapeutic target for breast cancer that uniquely eliminates tumor-initiating cells.* Breast Cancer Research, 2012. **14**(1): p. R22.
23. Kashiwagi, S., et al., *Identification of predictive markers of the therapeutic effect of eribulin chemotherapy for locally advanced or metastatic breast cancer.* BMC cancer, 2017. **17**(1): p. 604.
24. Chen, J., et al., *Competing endogenous RNA network analysis identifies critical genes among the different breast cancer subtypes.* Oncotarget, 2017. **8**(6): p. 10171.
25. Nogi, H., et al., *EGFR as paradoxical predictor of chemosensitivity and outcome among triple-negative breast cancer.* Oncology reports, 2009. **21**(2): p. 413-417.
26. Fragomeni, S.M., A. Sciallis, and J.S. Jeruss, *Molecular subtypes and local-regional control of breast cancer.* Surgical Oncology Clinics, 2018. **27**(1): p. 95-120.
27. Potapenko, I.O., et al., *Glycan-related gene expression signatures in breast cancer subtypes; relation to survival.* Molecular oncology, 2015. **9**(4): p. 861-876.
28. Seifi-Alan, M., et al., *MIR-206 target prediction in breast cancer subtypes by bioinformatics tools.* International Journal of Cancer Management, 2018. **11**(7).
29. Wang, F., L. Wang, and J. Pan, *PACE4 regulates proliferation, migration and invasion in human breast cancer MDA-MB-231 cells.* Molecular medicine reports, 2015. **11**(1): p. 698-704.
30. Lim, J.P., et al., *YBX1 gene silencing inhibits migratory and invasive potential via CORO1C in breast cancer in vitro.* BMC cancer, 2017. **17**(1): p. 201.

Figures

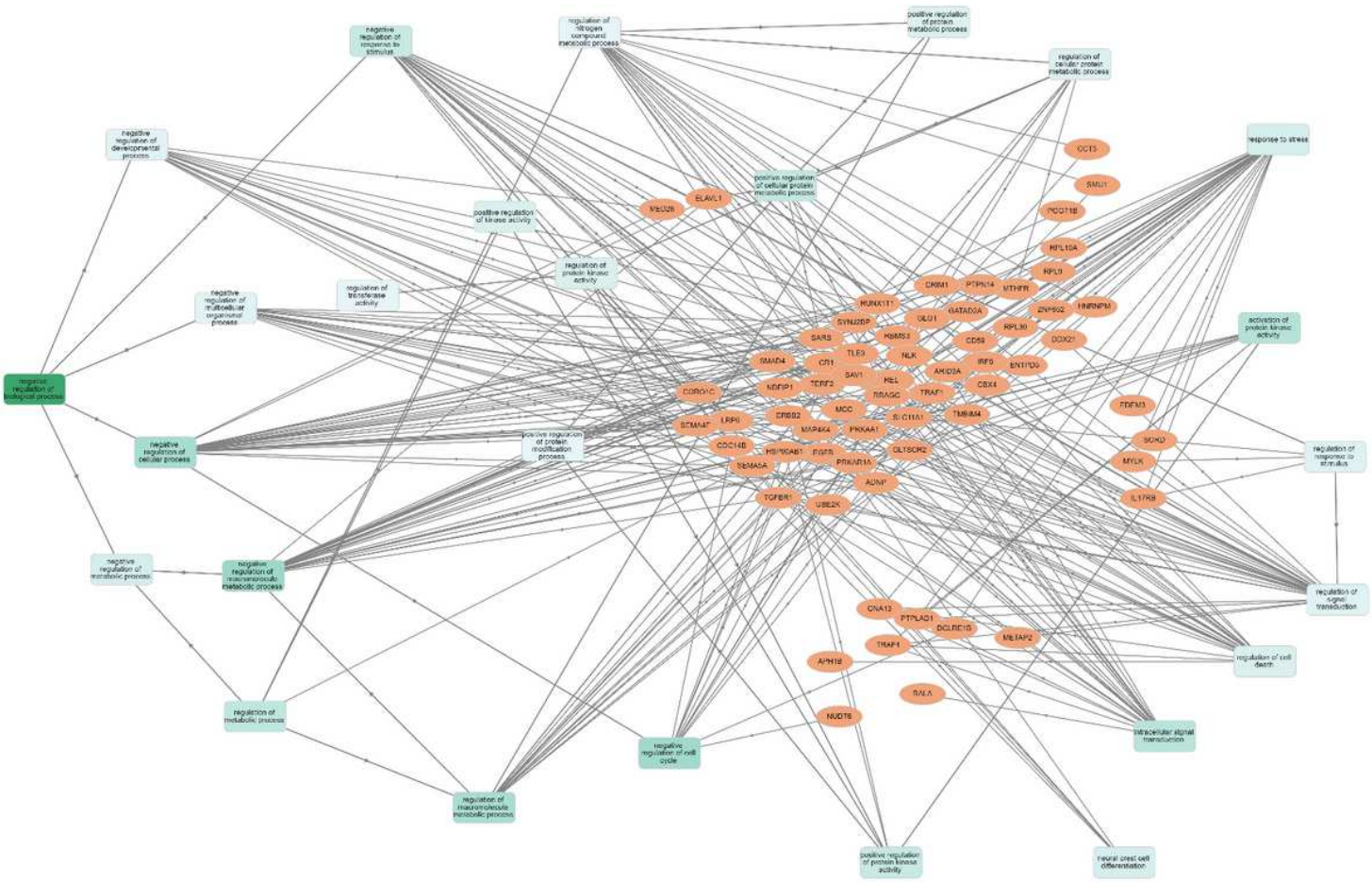

Figure 1

Biological process of extracted genes in open-source GOnet web-application with q-value ≤ 0.05 in enrichment analysis options.