American Journal of Bioscience and Bioengineering
2016; 4(2): 26-33
http://www.sciencepublishinggroup.com/j/bio
doi: 10.11648/j.bio.20160402.12
ISSN: 2328-5885 (Print); ISSN: 2328-5893 (Online)

Using Genetic Algorithms and Sparse Logistic Regression to Find Gene Signatures for Chemosensitivity Prediction in Breast Cancer

Wei Hu

Department of Computer Science, Houghton College, Houghton, New York, USA

Email address:
wei.hu@houghton.edu

To cite this article:
Wei Hu. Using Genetic Algorithms and Sparse Logistic Regression to Find Gene Signatures for Chemosensitivity Prediction in Breast Cancer. *American Journal of Bioscience and Bioengineering.* Vol. 4, No. 2, 2015, pp. 26-33. doi: 10.11648/j.bio.20160402.12

Received: December 15, 2015; **Accepted:** January 5, 2016; **Published:** May 4, 2016

Abstract: Various gene signatures of chemosensitivity in breast cancer have been identified. When used to build predictors of have chemosensitivity, many of them have their prediction accuracy around 80%. Identifying gene signatures to build high accuracy such predictors is a prerequisite for their clinical tests and applications. To elucidate the importance of each individual gene in a signature is another pressing need before such signature could be tested in clinical settings. In this study, Genetic Algorithms (GAs) and Sparse Logistic Regression (SLR) were employed to identify two signatures. The first had 28 probe sets selected by GA from the top 65 probe sets that were highly overexpressed between pathologic compete response (pCR) and residual disease (RD) and was used to build a SLR predictor of pCR (SLR-28). The second had 86 probe sets (Notch-86) selected by GA from Notch signaling pathway and was used to develop a SLR predictor of pCR (SLR-Notch-86). These two predictors tested on a training set (n=81) and validation set (n=52) had very precise predictions measured by accuracy, specificity, sensitivity, positive predictive value and negative predictive value with their corresponding P value all zero. Furthermore, these two predictors discovered 12 important genes in the 28 probe set signature and 14 important genes in the Notch-86 signature. Our two signatures produced superior performance over a signature in a previous study, demonstrating the potential of GA and SLR in identifying robust gene signatures in chemo response prediction in breast cancer.

Keywords: Genetic Algorithm, Gene Signature, Breast Cancer, Sparse Logistic Regression, Predictor, Chemosensitivity

1. Introduction

Breast cancer is a complex disease of different molecular subtypes with distinct genetic alterations and clinical outcomes. In current practice, chemotherapy is applied empirically, and not all patients benefit equally, illustrating the imperative needs for a more personalized approach in cancer treatment. The ability to predict whether an individual patient will benefit from a specific therapy is of great clinical significance. The estrogen receptor status can be used to guide the decisions on hormonal therapy. The gene expression data that reflect subtle differences in tumors can be utilized to build a predictor of response to cancer drugs.

Single clinical or molecular parameters, such as tumor size, histology, hormone receptor or human epidermal growth factor receptor 2 (HER2) expression, and tumor grade, does not always give reliable predictions of response. With microarray data, researchers are able to identify gene expression patterns that are predictive of chemotherapy response.

In [1], t-test for unequal-variance was employed to find a signature of 31 probe sets (27 genes) with highest differentially expressed values between pRC and RD. Based on this signature, a 30-probe set Diagnal Linear Discriminant Analysis (DLDA-30) classifier was constructed to predict pathological response to preoperative paclitaxel/FAC chemotherapy. The value of this type of predictor is the ability to identify those patients most likely to benefit from a particular treatment, the neoadjuvant chemotherapy, in this case. This predictor is able to recognize not all responsive patients but exclusively those that will benefit the most, as defined by attaining a pCR. Other clinical studies also identified gene signatures that predict response to

neoadjuvant therapy of breast cancer [2--15].

As a single variable technique, t-test processes one gene at a time and might miss the interactions between genes. We believed the signature identified by t-test could be optimized with the help of a multivariable technique such as GA. We aimed to search for novel signatures and to use them to develop predictors of pCR that can achieve much better predictions than the DLDA-30. Genetic algorithms have the capacity to explore multiple solutions concurrently, which we used to find interacting and informative genes in this study. After identifying a signature, SLR was employed to further explore the importance of each individual gene's contribution to the prediction of pCR.

2. Patients and Methods

2.1. Patient Cohorts and Clinical Information

One breast cancer patient cohort was obtained from a previous publication [1] (n=133). Needle-biopsy samples were collected from 133 patients with stage I, II, or III breast cancer who received preoperative weekly paclitaxel and a combination of fluorouracil, doxorubicin, and cyclophosphamide (T/FAC). These 133 patients were divided into two subsets, one training set of size 81 and one validation set of size 52. These data contain clinical information including patient age, gender, race, histological classification, stage, nuclear grade, ER (estrogen receptor), PR (progesterone receptor), and HER2 (human epidermal growth factor 2) status, pathologic complete response, and residual disease. These data also contain each patient's genome-scale gene expression profiles generated using Affymetrix U133A chip (Santa Clara, CA). pCR was defined as no residual invasive cancer in the breast or lymph nodes. pCR is presently accepted as a reasonable early indicator for long-term survival.

2.2. Sparse Logistic Regression

A standard least squares linear regression solves the following problem:

Given data $\{x_i, y_i\}_{i=1}^{m}$, find $\alpha = (\alpha_0, \alpha_1, \alpha_2, ..., \alpha_n)$ such that

$$\min_{\alpha} \sum_{i} (y_i - f(x_i))^2 \text{ where } f(x_i) = \sum_{j=1}^{n} \alpha_j x_{ij} + \alpha_0 \quad (1)$$

The LASSO regression [16] deals with the following problem:

$$\min_{\alpha} \sum_{i} (y_i - f(x_i))^2 \text{ with } \sum_{j=1}^{n} |\alpha_j| \le t, \quad (2)$$

where t controls the L_1 norm of $(\alpha_1, \alpha_2, ..., \alpha_n)$. This constraint on α produces a sparse model, i.e., many components of α can be zero. Following the idea of LASSO, Shevade *at el*. [17] studied the following problem for sparse logistic regression, when $y_i \in \{-1, +1\}$

$$\min_{\alpha} \sum_{i} g(-y_i f(x_i)) \text{ with } \sum_{j=1}^{n} |\alpha_j| \le t, \quad (3)$$

where g(ξ) = log(1 + e^{ξ}), which is the negative log-likelihood function associated with the probability model

$$\text{Prob}(y \mid x) = \frac{1}{1 + e^{-y \cdot f(x)}} \quad (4)$$

Cawley GC *et al*. [18] utilized a novel technique to solve this sparse logistic regression problem efficiently. In our study, we used +1 to label those cases of RD status, and used -1 to label those cases of pCR status.

2.3. Notch Signaling Pathway

Notch genes encode highly conserved cell surface receptors. The Notch signaling pathway consists of Notch receptors, ligands, negative and positive modifiers, and transcription factors. It plays a key role in the normal development of many tissues and cell types, through diverse effects on cell regulation, proliferation, and differentiation. Aberrant Notch signaling has been observed in several human cancers, including acute T-cell lymphoblastic leukemia, cervical cancer, and breast cancer [19--21]. The Oligo GEArray Human Notch Signaling Pathway Microarray [22] was designed for profiling expression of 113 genes (Table 1) involved in Notch signaling. One of the two signatures identified in this study, Notch-86, was selected from these 113 genes.

Table 1. *Genes Involved in Notch Signaling Pathway as Described in [22].*

Notch Signaling Pathway:
Notch Binding: DLL1 (DELTA1), DTX1, JAG1, JAG2.
Notch Receptor Processing: ADAM10, PSEN1, PSEN2, PSENEN (PEN2).
Notch Signaling Pathway Target Genes:
Apoptosis Genes: CDKN1A, CFLAR (CASH), IL2RA, NFKB1.
Cell Cycle Regulators: CCND1 (Cyclin D1), CDKN1A (P21), IL2RA.
Cell Proliferation: CDKN1A (P21), ERBB2, FOSL1, IL2RA.
Genes Regulating Cell Differentiation: DTX1, PPARG.

Neurogenesis: HES1, HEY1.
Regulation of Transcription: DTX1, FOS, FOSL1, HES1, HEY1, NFKB1, NFKB2, NR4A2, PPARG, STAT6.
Other Target Genes with Unspecified Functions: CD44, CHUK, IFNG, IL17B, KRT1, LOR, MAP2K7, PDPK1, PTCRA.
Other Genes Involved in the Notch Signaling Pathway:
Apoptosis Genes: AXIN1, EP300, HDAC1, NOTCH2, PSEN1, PSEN2.
Cell Cycle Regulators: AXIN1, CCNE1, CDC16, EP300, FIGF, JAG2, NOTCH2, PCAF.
Cell Proliferation: CDC16, FIGF, FZD3, JAG1, JAG2, LRP5, NOTCH2, PCAF, STIL (SIL).
Genes Regulating Cell Differentiation: DLL1, JAG1, JAG2, NOTCH1, NOTCH2, NOTCH3, NOTCH4, PAX5, SHH.
Neurogenesis: DLL1, EP300, HEYL, JAG1, NEURL, NOTCH2, PAX5, RFNG, ZIC2 (HPE5).
Regulation of Transcription: AES, CBL, CTNNB1, EP300, GLI1, HDAC1, HEYL, HOXB4, HR, MYCL1, NCOR2, NOTCH1, NOTCH2, NOTCH3, NOTCH4, PAX5, PCAF, POFUT1, RUNX1, SNW1 (SKIIP), SUFU, TEAD1, TLE1.
Others Genes with Unspecified Functions: ADAM17, GBP2, LFNG, LMO2, MFNG, MMP7, NOTCH2NL, NUMB, SEL1L, SH2D1A.
Other Signaling Pathways that Crosstalk with the Notch Signaling Pathway:
Sonic Hedgehog (Shh) Pathway: GLI1, GSK3B, SHH, SMO, SUFU.
Wnt Receptor Signaling Pathway: AES, AXIN1, CTNNB1, FZD1, FZD2, FZD3, FZD4, FZD6, FZD7, GSK3B, LRP5, TLE1, WISP1, WNT11.
Other Genes Involved in the Immune Response: CXCL9, FAS (TNFRSF6), G1P2, GBP1, IFNG, IL2RA, IL2RG, IL4, IL4R, IL6ST, IRF1, ISGF3G, OAS1, OSM, STAT5A, STUB1.

2.4. Top 65 Probe Sets

T-tests for unequal variances for all the probe sets on the Affymetrix U133A chip were carried out to find the genes that were significantly differentially expressed in either the pCR cases or the RD cases. We chose the 60 probe sets with the smallest t-test P values (False Discovery Rate=1%) and 5 probe sets with the most negative t-test statistics in the remaining probe sets to be our first signature (top 65-probe set signature) as presented in Table 1. The top 31 probe set signature in [1] had a FDR=0.5%.

2.5. Genetic Algorithms

Genetic Algorithms (GAs), a particular class of evolutionary algorithms, are search algorithms that adopt some common processes in genetics such as selection, mutation, and inheritance. The GAs outperform other traditional search algorithms in various applications.

The outline of a genetic algorithm is as follows:

Generate an initial population of individuals
Evaluate initial population
Repeat
  Perform selection
  Apply genetic operations such as mutation and crossover to generate a new generation of individuals
  Evaluate individuals in the population
Until some stopping criteria is satisfied

2.6. Prediction Accuracy Evaluation

In order to evaluate the significance of our predictions, we need to compare them with random predictions. For each dataset, a random-label permutation was conducted while keeping the number of instances in each group fixed. The matches between the permuted labels and the original ones were recorded. The standard P value was the percentage of 1000 random predictions with higher accuracy than the calculated predictions.

Table 2. *Top 65 Differentially Expressed Probe Sets by Unequal-Variance t-Test (n=82, probe sets with a * are contained in the top 31 probe sets found in [1]).*

Rank by P value	t-Test	P value	Higher Expression in	Probe Set ID	Gene Symbol	Gene Name
1	6.215265	2.20E-08	RD	203930_s_at*	MAPT	microtubule-associated protein tau
2	6.36741	2.31E-08	RD	203929_s_at*	MAPT	microtubule-associated protein tau
3	6.212778	2.56E-08	RD	212207_at*	THRAP2	thyroid hormone receptor associated protein 2
4	5.804489	1.25E-07	RD	212745_s_at*	BBS4	Bardet-Biedl syndrome 4
5	5.847627	1.42E-07	RD	203928_x_at*	MAPT	microtubule-associated protein tau
6	5.763819	1.67E-07	RD	208945_s_at*	BECN1	beclin 1 (coiled-coil, myosin-like BCL2 interacting protein)
7	5.704523	2.50E-07	RD	206401_s_at*	MAPT	microtubule-associated protein tau
8	5.716982	2.77E-07	RD	205354_at*	GAMT	guanidinoacetate N-methyltransferase
9	5.555817	3.65E-07	RD	219741_x_at*	ZNF552	zinc finger protein 552
10	5.523853	4.08E-07	RD	215304_at*	---	Clone 23948 mRNA sequence
11	5.449088	5.45E-07	RD	209173_at	AGR2	anterior gradient 2 homolog (Xenopus laevis)
12	5.391683	6.89E-07	RD	201508_at*	IGFBP4	insulin-like growth factor binding protein 4
13	5.357545	8.43E-07	RD	217542_at*	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)
14	5.312123	1.30E-06	RD	219044_at*	FLJ10916	hypothetical protein FLJ10916
15	5.26737	1.41E-06	RD	215616_s_at*	JMJD2B	jumonji domain containing 2B
16	5.215414	1.41E-06	RD	204509_at*	CA12	carbonic anhydrase XII
17	5.221534	1.42E-06	RD	202204_s_at*	AMFR	autocrine motility factor receptor
18	5.215809	1.70E-06	RD	214124_x_at*	FGFR1OP	FGFR1 oncogene partner

Rank by P value	t-Test	P value	Higher Expression in	Probe Set ID	Gene Symbol	Gene Name
19	5.210207	1.71E-06	RD	219051_x_at*	METRN	meteorin, glial cell differentiation regulator
20	5.194077	1.97E-06	RD	209696_at	FBP1	fructose-1,6-bisphosphatase 1
21	5.052227	2.70E-06	RD	213234_at*	KIAA1467	KIAA1467 protein
22	5.049412	2.74E-06	RD	217838_s_at	EVL	Enah/Vasp-like
23	5.054632	2.95E-06	RD	205074_at	SLC22A5	solute carrier family 22 (organic cation transporter), member 5
24	5.139071	3.06E-06	RD	213623_at*	KIF3A	kinesin family member 3A
25	5.017933	3.38E-06	RD	201413_at	HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4
26	4.908014	5.26E-06	RD	205225_at	ESR1	estrogen receptor 1
27	4.823788	7.04E-06	RD	217016_x_at	FLJ23172	hypothetical LOC389177
28	4.80725	7.18E-06	RD	214053_at*	---	CDNA FLJ44318 fis, clone TRACH3000780
29	4.888899	7.30E-06	RD	213527_s_at	ZNF688	zinc finger protein 688
30	4.819068	7.44E-06	RD	203009_at	LU	Lutheran blood group (Auberger b antigen included)
31	4.865888	9.07E-06	RD	212046_x_at	MAPK3	mitogen-activated protein kinase 3
32	4.854113	9.27E-06	RD	205012_s_at	HAGH	hydroxyacylglutathione hydrolase
33	4.762182	9.56E-06	RD	203675_at	NUCB2	nucleobindin 2
34	4.700102	1.07E-05	RD	203071_at	SEMA3B	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B
35	4.710655	1.07E-05	RD	210129_s_at	TTLL3	tubulin tyrosine ligase-like family, member 3
36	4.671287	1.20E-05	RD	218671_s_at	ATPIF1	ATPase inhibitory factor 1
37	4.689638	1.23E-05	RD	209339_at	SIAH2	seven in absentia homolog 2 (Drosophila)
38	4.629403	1.44E-05	RD	218976_at	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12
39	4.649829	1.44E-05	RD	205734_s_at	AFF3	AF4/FMR2 family, member 3
40	4.634054	1.65E-05	RD	202641_at	ARL3	ADP-ribosylation factor-like 3
41	4.580441	1.68E-05	RD	218259_at	MKL2	MKL/myocardin-like 2
42	4.590716	1.71E-05	RD	220540_at	KCNK15	potassium channel, subfamily K, member 15
43	4.578743	1.71E-05	RD	210831_s_at	PTGER3	prostaglandin E receptor 3 (subtype EP3)
44	4.608731	1.77E-05	RD	218769_s_at	ANKRA2	ankyrin repeat, family A (RFXANK-like), 2
45	4.587999	1.81E-05	RD	218394_at	FLJ22386	leucine zipper domain protein
46	4.568723	1.82E-05	RD	216835_s_at	DOK1	docking protein 1, 62kDa (downstream of tyrosine kinase 1)
47	4.606517	1.98E-05	RD	221728_x_at	XIST	X (inactive)-specific transcript
48	4.582593	2.04E-05	RD	212956_at	KIAA0882	KIAA0882 protein
49	4.531619	2.06E-05	RD	212239_at	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha)
50	4.521411	2.13E-05	RD	212209_at	THRAP2	thyroid hormone receptor associated protein 2
51	4.509765	2.22E-05	RD	204792_s_at	WDTC2	WD and tetratricopeptide repeats 2
52	4.593663	2.45E-05	RD	204862_s_at	NME3	non-metastatic cells 3, protein expressed in
53	4.478137	2.49E-05	RD	206418_at	NOX1	NADPH oxidase 1
54	4.538231	2.74E-05	RD	205059_s_at	IDUA	iduronidase, alpha-L-
55	4.463108	2.74E-05	RD	210958_s_at	MAST4	microtubule associated serine/threonine kinase family member 4
56	4.501318	2.76E-05	RD	202228_s_at	SDFR1	stromal cell derived factor receptor 1
57	4.539226	2.83E-05	RD	212660_at	PHF15	PHD finger protein 15
58	-5.01605	2.96E-05	pCR	213134_x_at*	BTG3	BTG family, member 3
59	4.427751	2.98E-05	RD	203789_s_at	SEMA3C	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C
60	4.484732	3.00E-05	RD	216109_at	THRAP2	Thyroid hormone receptor associated protein 2
61	-5.01538	3.31E-05	pCR	205548_s_at*	BTG3	BTG family, member 3
62	-4.53199	0.000122	PCR	204825_at*	MELK	maternal embryonic leucine zipper kinase
63	-4.00315	0.000496	pCR	205339_at	SIL	TAL1 (SCL) interrupting locus
64	-3.97777	0.000442	PCR	203693_s_at*	E2F3	E2F transcription factor 3
65	-3.94634	0.000361	PCR	216237_s_at	MCM5	MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae)

3. Results

3.1. Two Signatures: The 28 Probe Sets and the Notch-86

To search and select a subset of the 65 probe sets, we represented our solution, referred to as an individual in GA terms, as a binary vector of size 65 to indicate the presence (1) or absence (0) of each probe set in the 65 probe sets. We ran the GA algorithm with population size 200, individual size 65, and 100 generations. In each generation, the top 50% of the individuals with highest fitness values were selected as parents to produce the next generation with crossover and a point mutation was applied to each individual randomly at six genes. Our fitness value was the prediction accuracy of SLR based on the training set. In each generation of GA, we divided the training set (n=82) into five equal subsets and used four subsets as a training set for SLR and one subset as a test set to get the accuracy of SLR on this test set. At the same time, in each generation, we calculated the prediction accuracy of SLR on the validation set (n=51). Our goal was to choose an individual that has similar high accuracy on the

training and validation sets. We found an individual of this quality and its binary representation has 28 ones, which was our first signature (Table 2). A subset of the Notch signature of such quality, Notch-86, was found the same way. Our second signature had 86 probe sets. Table 3 displays the most important probe sets for response prediction in this second signature.

3.2. Two Predictors: SLR-28 and SLR-Notch-86

In [1] the DLDA-30 was selected as the best predictor after a thorough search of different predictors based on Support Vector Machine (SVM), Diagonal Linear Discrimmant Analysis (DLDA), and K-nearest neighbor (KNN). We developed two SLR based predictors of high precision. One used the 28 probe sets (SLR-28) and one used the Notch-86 (SLR-Notch-86). The evaluation of prediction performance was conducted on both the training set (n=82) and the validation set (n=51). Since the DLDA-30 was evaluated on the training set with five-fold cross validation, we performed five-fold evaluation too for our two predictors. Further, we repeated the five-fold cross validation 10 times and the averaged results were reported in Table 4.

On the training set, the two predictors, SLR-28 and SLR-Notch-86, produced much better predictions across all five measurements than DLDA-30 (Table 5). In [1], authors calculated the P values of their DLDA-30 predictions on the training set in five-fold cross validation and they were all zero. Since our two predictors had higher values in all the five measurements, we concluded that they also have P value zero in all these five measurements.

On the validation set, the two predictors trained on the training set had a much higher accuracy, a much higher specificity, and a much higher PPV than DLDA-30 (Table 6). Our two predictors had P values zero in all five measurements, whereas DLDA-30 had three P values larger than 0.05, especially those for accuracy and specificity. SLR-28 correctly identified all but two who achieved pCR and all but two who achieved RD, and SLR-Notch-86 correctly identified all but three who achieved pCR and all but two who achieved RD (Table 6). Tables 5 and 6 together show that our two predictors can predicts pCR with great precision on the training set and the validation set.

These two predictors also identified the important genes in each signature that had nonzero SLR weights (Figure 1). The genes with zero weight did not contribute to the prediction. The genes with positive weight contribute positively to the RD prediction and those with negative weight contribute positively to pCR prediction. In most cases, genes with positive t-test statistic had positive weight like the three genes in Figure 1, BGT3, MELK, and MCM5. However, there were some exceptions. Two genes, STUB1 and PDPK1, had positive t-test statistic, but negative weight in Table 3, demonstrating that SLR as a multivariable technique can capture some interactions between genes whereas t-test may not. Figure 1 showed that the most discriminative genes measured by SLR as a group were not necessarily those with the smallest P values by t-test as individual genes.

Table 3. *Important Probe Sets in Notch-86 Signature.*

SLR Weight	t-Test	P value	Higher Expression in	Probe set ID	Gene symbol	Gene Name
0.29955	2.585026	0.011947	RD	202221_s_at	EP300	E1A binding protein p300
-0.4945	-0.57563	0.56903	pCR	203393_at	HES1	hairy and enhancer of split 1
-0.02104	-2.73059	0.012681	pCR	203915_at	CXCL9	chemokine (C-X-C motif) ligand 9
-0.27652	-1.92262	0.067834	pCR	204152_s_at	MFNG	Manic fringe homolog
0.50511	2.208649	0.031467	RD	205552_s_at	OAS1	oligoadenylate synthetase 1
-0.71933	-4.00315	0.000496	pCR	205746_s_at	ADAM17	ADAM metallopeptidase domain 17
0.60822	3.119622	0.002969	RD	207760_s_at	NCOR2	nuclear receptor co-repressor 2
-0.49017	-1.3125	0.199077	pCR	211179_at	RUNX1	runt-related transcription factor 1
0.6048	1.873074	0.066922	RD	211209_x_at	SH2D1A	SH2 domain protein 1A
0.5284	0.809656	0.423383	RD	212014_x_at	CD44	CD44 antigen
-0.27566	-1.44514	0.156111	pCR	213523_at	CCNE1	cyclin E1
-0.11683	0.382658	0.704546	RD	217934_x_at	STUB1	STIP1 homology and U-box containing protein 1
0.41896	3.252073	0.001808	RD	218665_at	FZD4	frizzled homolog 4
-0.07142	-0.66774	0.50776	pCR	219683_at	FZD3	frizzled homolog 3
-0.2956	0.212779	0.832697	RD	32029_at	PDPK1	phosphoinositide dependent protein kinase-1

One of the important genes in Figure 1 is CA12, which is a membrane zinc metalloenzyme that is present in different normal tissues but is overexpressed in some cancers such as renal cell and breast cancers. Two studies found that increased CA IX expression is associated with poor relapse free and overall survival in invasive breast cancer [23, 24]. Another study found that *CA12* is regulated by estrogen receptor α (ERα) in breast cancer, and that this regulation involves a distal estrogen-responsive enhancer region [25].

The mitochondrial ATPase inhibitory factor 1(ATPIF1) is another important gene in Figure 1. Several studies discovered a close link between metabolic and genetic changes observed during malignant growth [26, 27]. The large positive weight of this gene in Figure 1 is also in agreement with this observation.

One study in [28] revealed that MAPT is the best single gene discriminator of pCR to preoperative chemotherapy with paclitaxel, 5-fluorouracil, doxorubicin, and

cyclophosphamide. There are four probe sets of MAPT selected in Table 2 and MAPT is one of the 16 informative genes in the top 65 probe sets in Figure 1.

SDFR1 encodes a cell surface protein of the immunglobulin superfamily that regulates cell adhesion and process outgrowth. BTG3 is tumor suppressor [29-31] and its large negative weight implies that its presence enhances chemo sensitivity. The functions of the important genes in Notch-86 signature in Figure 1 can be found in Table 3.

Table 4. *Prediction Measures (Five-fold cross validation) of DLDA-30, SLR-28 and SLR-Notch-86 on the Training Set with all P Values Zero in the Five Measurements.*

Measures	DLDA-30	SLR-28	SLR-Notch-86
Accuracy	83	0.90	0.96
Sensitivity	75	0.90	0.96
Specificity	73	0.89	0.96
PPV	50	0.75	0.91
NPV	90	0.96	0.99

Table 5. *Prediction Measures of DLDA-30, SLR-28, and SLR-Notch-86 on the Validation Set along with Their P values.*

Measure	DLDA-30	P value	SLR-28	P value	SLR-Notch-86	P value
Accuracy	0.76	0.1900	0.92	0	0.90	0
Sensitivity	0.92	0	0.85	0	0.77	0
Specificity	0.71	0.96	0.95	0	0.95	0
PPV	0.52	0.0920	0.85	0	0.83	0
NPV	0.96	0	0.95	0	0.92	0

Table 6. *Confusion Matrices for DLDA-30, SLR-28, and SLR-Notch-86 on the Validation Set.*

DLDA-30	Predicted as pCR	Predicted as RD	SLR-28	Predicted as pCR	Predicted as RD	SLR-Notch-86	Predicted as pCR	Predicted as RD
Observed pCR	n=12	n=1	Observed pCR	n=11	n=2	Observed pCR	n=10	n=3
Observed RD	n=11	n=27	Observed RD	n=2	n=36	Observed RD	n=2	n=36

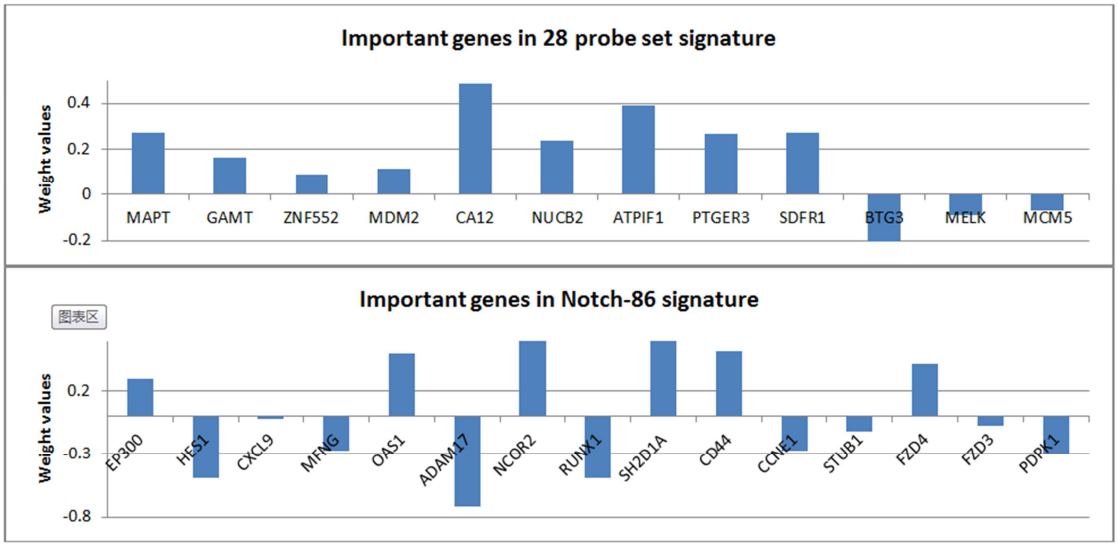

Figure 1. *Two plots to show the important genes in the 28 probe set signature and the Notch-86 probe set signature.*

4. Conclusion

In this study, we intended to uncover gene signatures for developing predictors that have a much higher accuracy than DLDA-30. With the ability to account for multiple gene interactions, the multivariable techniques, such as genetic algorithms and SLR, have demonstrated their potential utility in identifying robust gene signatures of clinical relevance.

Currently there is an urgent need to develop knowledge to identify groups of patients who will derive benefit from the different chemotherapy agents. Molecular profiling of individual tumors will help to predict the most appropriate therapy for each cancer patient. One study already suggests that responses to neoadjuvant chemotherapy correlate with gene expression profile, with tumors displaying the ER-positive gene signatures being less likely to respond than other types of breast cancer [32].

References

[1] Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N. and Pusztai, L. (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer, J. ClinOncol, Vol. 24, pp. 4236–4244.

[2] Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, Mac Grogan G, Mauriac L, Katsaros D, Molina F, Theillet C, Darbon JM. (2008) A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. Clin Cancer Res. 14(6): 1744-52.

[3] Linke SP, Bremer TM, Herold CD, Sauter G, Diamond C. (2006) A multimarker model to predict outcome in tamoxifen-treated breast cancer patients. Clin Cancer Res. 12(4): 1175-83.

[4] P. E. Lønning, S. Knappskog, V. Staalesen, R. Chrisanthar & J. R. Lillehaug (2007) Breast cancer prognostication and prediction in the postgenomic era, Annals of Oncology 18: 1293–1306.

[5] Folgueira MA, Carraro DM, Brentani H, Patrão DF, Barbosa EM, Netto MM, Caldeira JR, Katayama ML, Soares FA, Oliveira CT, Reis LF, Kaiano JH, Camargo LP, Vêncio RZ, Snitcovsky IM, Makdissi FB, e Silva PJ, Góes JC, Brentani MM. (2005) Gene expression profile associated with response to doxorubicin-based therapy in breast cancer. Clin Cancer Res. 11(20): 7434-43.

[6] Holly K. Dressman, Christopher Hans6, Andrea Bild, John A. Olson, Eric Rosen, P. Kelly Marcom, Vlayka B. Liotcheva, Ellen L. Jones, Zeljko Vujaskovic, Jeffrey Marks, Mark W. Dewhirst, Mike West, Joseph R. Nevins and Kimberly Blackwell (2006) Gene Expression Profiles of Multiple Breast Cancer Phenotypes and Response to Neoadjuvant Chemotherapy, Clinical Cancer Research Vol. 12, 819-826.

[7] Olaf Thuerigen, Andreas Schneeweiss, Grischa Toedt, Patrick Warnat, Meinhard Hahn, Heidi Kramer, Benedikt Brors, Christian Rudlowski, Axel Benner, Florian Schuetz, Bjoern Tews, Roland Eils, Hans-Peter Sinn, Christof Sohn, Peter Lichter (2006) Gene Expression Signature Predicting Pathologic Complete Response With Gemcitabine, Epirubicin, and Docetaxel in Primary Breast Cancer, Journal of Clinical Oncology, Vol 24, No 12 pp. 1839-1845.

[8] Goldstein NS, Decker D, Severson D et al. (2007) Molecular classification system identifies invasive breast carcinoma patients who are most likely and those who are least likely to achieve a complete pathologic response after neoadjuvant chemotherapy, Cancer 110: 1687–1696.

[9] Chang, J. C. et al. (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 362, 362–369.

[10] Gianni, L. et al. (2005) Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer, J. Clin. Oncol. 23, 7265–7277.

[11] Hannemann, J. et al. (2005) Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. J. Clin. Oncol. 23, 3331–3342.

[12] Singer CF, Klinglmüller F, Stratmann R, Staudigl C, Fink-Retter A, Gschwantler D, et al. (2013) Response Prediction to Neoadjuvant Chemotherapy: Comparison between Pre-Therapeutic Gene Expression Profiles and In Vitro Chemosensitivity Assay. PLoS ONE 8(6): e66573.

[13] Torsten NielsenEmail author, Brett Wallden, Carl Schaper, Sean Ferree, Shuzhen Liu, Dongxia Gao, Garrett Barry, Naeem Dowidar, Malini Maysuria and James Storhoff (2014) Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens, BMC Cancer 201414: 177.

[14] Brian David Lehmann, Yan Ding, Daniel Joseph Viox, Ming Jiang, Yi Zheng, Wang Liao, Xi Chen, Wei and Yajun Yi (2015) Evaluation of public cancer datasets and signatures identifies TP53 mutant signatures with robust prognostic and predictive value, BMC Cancer 201515: 179.

[15] Prat A, Lluch A, Albanell J, Barry WT, Fan C, Chacón JI, Parker JS, Calvo L, Plazaola A, Arcusa A, Seguí-Palmer MA, Burgues O, Ribelles N, Rodriguez-Lescure A, Guerrero A, Ruiz-Borrego M, Munarriz B, López JA, Adamo B, Cheang MC, Li Y, Hu Z, Gulley ML, Vidal MJ, Pitcher BN, Liu MC, Citron ML, Ellis MJ, Mardis E, Vickery T, Hudis CA, Winer EP, Carey LA, Caballero R, Carrasco E, Martín M, Perou CM, Alba E (2014) Predicting response and survival in chemotherapy-treated triple-negative breast cancer, Br J Cancer. 2014 Oct 14; 111(8): 1532-41.

[16] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. B, Vol. 58, pp. 267–288.

[17] Shevade, S. K. and Keerthi, S. S. (2003) A simple and efficient algorithm for gene selection using sparse logistic regression, Bioinformatics, Vol. 19, pp. 2246–2253.

[18] Cawley, G. C. and Talbot, L. C. (2006) Gene selection in cancer classification using sparse logistic regression with bayesian regularization, Bioinformatics, Vol. 22, pp. 2348–2355.

[19] Brennan, K. and Anthony Brown, M. C. (2003) Is there a role for Notch signaling in human breast cancer? Breast Cancer Res, 5(2), 69-75.

[20] Stylianou, S., Clarke, R. B. et al. (2006) Activation of notch signaling in human breast cancer, Cancer Research, 66, 1517-1525.

[21] Xiaolin Huang Li Wang He Zhang Haibo Wang Xiaoping Zhao Guanxiang Qian Jifan Hu Shengfang Ge Xianqun Fan (2012) Therapeutic Efficacy by Targeting Correction of Notch1-Induced Aberrants in Uveal Tumors, PLoS ONE, 7(8): e44301.

[22] GEArray, O. Human notch signaling pathway microarray. http://www.sabiosciences.com/gene_array_product/HTML/OHS-059.html

[23] Chia SK, Wykoff CC, Watson PH, Han C, Leek RD, Pastorek J, Gatter KC, Ratcliffe P, Harris AL. (2001) Prognostic significance of a novel hypoxia-regulated marker, carbonic anhydrase IX, in invasive breast carcinoma. J Clin Oncol. 19(16): 3660-8.

[24] Hussain SA, Ganesan R, Reynolds G, Gross L, Stevens A, Pastorek J, Murray PG, Perunovic B, Anwar MS, Billingham L, James ND, Spooner D, Poole CJ, Rea DW, Palmer DH (2007) Hypoxia-regulated carbonic anhydrase IX expression is associated with poor survival in patients with invasive breast cancer, Br J Cancer. 96(1): 104-9.

[25] Daniel H. Barnett, Shubin Sheng, Tze Howe Charn, Abdul Waheed, William S. Sly, Chin-Yo Lin, Edison T. Liu and Benita S. Katzenellenbogen (2008) Estrogen Receptor Regulation of Carbonic Anhydrase XII through a Distal Enhancer in Breast Cancer, Cancer Research, 68, 3505.

[26] Martin C Abba, Yuhui Hu, Carla C Levy, Sally Gaddis, Frances S Kittrell, Yun Zhang, Jamal Hill, Reid P, Daniel Medina, Powel H Brown and C Marcelo Aldaz (2008) Transcriptomic signature of Bexarotene (Rexinoid LGD1069) on mammary gland from three transgenic mouse mammary cancer models, BMC Medical Genomics, 1: 40.

[27] Antonio Isidoro 1, Enrique Casado 2, Andrés Redondo 2, Paloma Acebo 1, Enrique Espinosa 2, Andrés M. Alonso 4, Paloma Cejas 2, David Hardisson 3, Juan A. Fresno Vara 2, Cristobal Belda-Iniesta 2, Manuel González-Barón 2 and José M. Cuezva (2005) Breast carcinomas fulfill the Warburg hypothesis and provide metabolic markers of cancer prognosis, Carcinogenesis, 26 (12): 2095-2104.

[28] Rouzier R, Rajan R, *et al.* (2005) Microtubule associated protein tau is a predictive marker and modulator of response to paclitaxel-containing preoperative chemotherapy in breast cancer. Proc Natl Acad Sci U S A, 102, 8315-8320.

[29] Ou YH, Pei-Han Chung PH, et al. (2007) The candidate tumor suppressor BTG3 is a transcriptional target of p53 that inhibits E2F1. The EMBO Journal, 26, 3968–3980.

[30] Majid S, Dar AA, Ahmad AE, Hirata H, Kawakami K, Shahryari V, Saini S, Tanaka Y, Dahiya AV, Khatri G, Dahiya R. (2009) BTG3 tumor suppressor gene promoter demethylation, histone modification and cell cycle arrest by genistein in renal cancer, Carcinogenesis, 30 (4): 662-70.

[31] Cheng YC1, Lin TY, Shieh SY. (2013) Candidate tumor suppressor BTG3 maintains genomic stability by promoting Lys63-linked ubiquitination and activation of the checkpoint kinase CHK1, Proc Natl Acad Sci U S A. 110 (15): 5993-8.

[32] Goldstein NS, Decker D, Severson D et al. (2007) Molecular classification system identifies invasive breast carcinoma patients who are most likely and those who are least likely to achieve a complete pathologic response after neoadjuvant, chemotherapy. Cancer, 110: 1687–1696.