Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(10–11) 3042–3056
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218794722
journals.sagepub.com/home/smm
SAGE

Robust identification of target genes and outliers in triple-negative breast cancer data

Pieter Segaert,[1],* Marta B Lopes,[2],* Sandra Casimiro,[3] Susana Vinga[2,4] and Peter J Rousseeuw[1]

Abstract

Correct classification of breast cancer subtypes is of high importance as it directly affects the therapeutic options. We focus on triple-negative breast cancer which has the worst prognosis among breast cancer types. Using cutting edge methods from the field of robust statistics, we analyze Breast Invasive Carcinoma transcriptomic data publicly available from The Cancer Genome Atlas data portal. Our analysis identifies statistical outliers that may correspond to misdiagnosed patients. Furthermore, it is illustrated that classical statistical methods may fail to identify outliers due to their heavy influence, prompting the need for robust statistics. Using robust sparse logistic regression we obtain 36 relevant genes, of which ca. 60% have been previously reported as biologically relevant to triple-negative breast cancer, reinforcing the validity of the method. The remaining 14 genes identified are new potential biomarkers for triple-negative breast cancer. Out of these, *JAM3*, *SFT2D2*, and *PAPSS1* were previously associated to breast tumors or other types of cancer. The relevance of these genes is confirmed by the new DetectDeviatingCells outlier detection technique. A comparison of gene networks on the selected genes showed significant differences between triple-negative breast cancer and non-triple-negative breast cancer data. The individual role of *FOXA1* in triple-negative breast cancer and non-triple-negative breast cancer, and the strong *FOXA1-AGR2* connection in triple-negative breast cancer stand out. The goal of our paper is to contribute to the breast cancer/triple-negative breast cancer understanding and management. At the same time it demonstrates that robust regression and outlier detection constitute key strategies to cope with high-dimensional clinical data such as omics data.

Keywords

Logistic regression, sparsity, cellwise outliers, gene networks

1 Introduction

Triple-negative breast cancer (TNBC) represents 10% to 17% of all diagnosed breast tumors,[1] and has the worst prognosis amongst the different subtypes of breast cancer (BC).[2] TNBC is characterized by the lack of expression of targetable proteins like estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2).[3] Based on this, the current standard of care treatment protocols for TNBC are limited to surgery, radiotherapy, and chemotherapy.[4] Since the BC subtype directly influences the therapeutic options, there is a high demand for the development of methods that not only accurately classify BC patients into BC subtypes but also identify relevant (target) genes that discriminate between TNBC patients and patients with other types of BC. The identification of genes that are either down- or up-regulated in TNBC is expected to play an

[1] Department of Mathematics, KU Leuven, Leuven, Belgium
[2] IDMEC, Instituto de Engenharia Mecânica, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[3] Luís Costa Lab, Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal
[4] INESC-ID, Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento, Lisboa, Portugal

*These authors contributed equally to this work.

Corresponding author:
Pieter Segaert, Celestijnenlaan 200B, bus 2400, Leuven B-3001, Belgium.
Email: pieter.segaert@kuleuven.be

important role in precision medicine, by providing a more in-depth knowledge on the cancer biology, but also yielding diagnostic, prognostic, and therapeutic markers that will ultimately improve patient outcomes.[5]

The classification of BC into TNBC and non-TNBC is dependent on the absence/presence of ER and PR expression, and HER2 amplification. The ER, PR, and HER2 *status* is evaluated by immunohistochemistry (IHC), either in core biopsies and/or surgical resection specimens. Tumors are classified as either ''positive'' or ''negative'' based on the results obtained. However, preanalytic variables, thresholds for positivity, and interpretation criteria may generate inaccurate results. For example, it has been reported that up to 20% of ER and PR test results are false negative or false positive.[6] This is probably mostly related with the difficulty to apply the thresholds for positivity. Guidelines from the European Society of Medical Oncology and the American Society of Clinical Oncology/College of American Pathologists recommend the cut-off to define ER or PR-positive cases as $\geq$1% ER or PR-positive tumour cells, respectively.[6,7] Tumours exhibiting <1% of tumour cells positive for ER or PR should be considered negative.[6]

Oncogenic ERBB2 is overexpressed in 15–20% of primary breast cancers.[8] Determination of positivity for HER2 overexpression by IHC is dependent on the assessment of the intensity of the reaction product, the completeness of membrane staining, and the percentage of positive cells.[8,9] An IHC score of 3+ is categorized as HER2 positive and is defined as strong complete reactivity seen in >10% of tumor cells. An IHC score of 2+ is classified as borderline reactivity (equivocal) and defined as weak to moderate complete membranous reactivity in >10% of tumor cells. IHC scores 0 and 1+ are considered HER2 negative. IHC 1+ is defined as faint, barely perceptible membranous reactivity in >10% of tumor cells. IHC 0 is defined as no reactivity or membranous reactivity in <10% of tumor cells. All equivocal cases, IHC 2+, must be evaluated by fluorescence in situ hybridization (FISH) or chromogenic in situ hybridization (CISH), where also a threshold for positivity is implied, based on average copy number/cell.

The clinical consequences of receptor status are extremely important. A patient given a wrong BC subtype classification will undergo inappropriate cancer treatment, either hormonal based or not, with severe consequences for cancer progression and survival. False negatives for ER and PR could benefit from endocrine therapy, and for false positives the hormonal therapy will fail. On the other hand, while a false positive HER2 assessment leads to the administration of potentially toxic, costly, and ineffective HER2-targeted therapy, a false negative HER2 assessment results in denial of anti-HER2 targeted therapy for a patient who could benefit from it.[8]

In this context, statistical analysis of gene expression data for known BC cases may provide valuable insights. However, real data often contain one or more observations deviating from the main pattern of the data.[10,11] For example, when considering gene expressions from TNBC data, inaccuracies may be due to variations in ER, PR, and HER2 testing, as mentioned above. Wrong TNBC class labels may result from inconsistencies between the IHC and FISH testing technologies. Unfortunately, classical results are highly influenced by these suspicious observations. The effect of outliers may be such that classical statistical techniques no longer detect them. This phenomenon is known as masking in statistics literature.[12] Moreover, outlying observations may even influence classical statistics so much that regular observations are flagged as outliers, a phenomenon known as swamping.[13] In regression models, these observations may compromise the predictive performance. Due to the high dimensional nature of the data, typical regression techniques are no longer valid. For example, the classical least squares fit cannot be computed when there are fewer observations than variables. Therefore, one uses sparsity-inducing methods to select relevant subsets of the original variables. However, these sparse methods might also be impacted by outliers, leading to relevant variables being neglected and irrelevant variables being selected.[14] Moreover, detecting interesting anomalous cases (e.g. a normal patient with deviating expressions of specific genes) may be of particular interest.

The importance of detecting outlying patients is therefore twofold. Outliers corresponding to errors in the labeling must be detected and treated accordingly in order to achieve accurate model predictions. Correctly diagnosing patients is of utmost interest as wrongly diagnosed patients may receive ineffective, expensive, and potentially harmful treatment. Secondly, outliers which are not errors reveal hidden information on the covariates that might play a role in the definition of new therapies based on target genes revealed by outlier analysis.

The remainder of the paper is structured as follows. In the next section, we discuss TNBC data construction from RNA-Seq and clinical data from Breast Invasive Carcinoma (BRCA). We then discuss logistic regression as a tool to decide BC class membership. Due to the high dimensionality of the data and the concern for outliers, we then turn to robust sparse logistic regression which selects relevant variables and flags outlying cases. Also the DetectDeviatingCells (DDC) method[15] is applied, and its results are linked to those of the robust logistic regression which brings new insights. The paper concludes with a discussion of results and model diagnostics along with the biological interpretation of the selected gene set.

2 Data description

The BRCA data set is publicly available from The Cancer Genome Atlas (TCGA) Data Portal[16] and contains genomic and clinical data from BC patients. The RNA-Seq Fragments Per Kilo base per Million (FPKM) data were imported using the "brca.data" R package.[17] The BRCA gene expression data is composed of 57,251 variables for a total of 1222 samples from 1097 individuals. From those samples, 1102 correspond to primary solid tumor, 7 to metastases, and 113 to normal breast tissue. Only samples from primary solid tumor were considered for analysis.

The TNBC gene expression data set was built based on the BRCA RNA-Seq data set available from TCGA. A subset of 19,688 variables, including the three TNBC-associated genes ESR1 (ENSG00000091831), PGR (ENSG00000082175), and ERBB2 (ENSG00000141736), hereafter designated as ER, PR and HER2, was considered, corresponding to the protein coding genes reported from the Ensemble genome browser[18] and the Consensus CDS[19] project. The response variable Y, corresponding to the clinical type, is a binary vector coded with "1" for TNBC individuals and "0" for non-TNBC. This vector was built based on the BRCA clinical data available from TCGA, regarding the individuals' label for ER, PR, and HER2 (either "positive", "negative", or "indeterminate"). When a "negative" label is recorded for all three genes the response is set to "1" (TNBC), whereas in all other cases the *status* is "0" (non-TNBC). However, for assessing the HER2 label, three variables are available from the clinical data: two from the IHC analysis, the HER2 *level* and *status*, and another corresponding to the HER2 *status* obtained by FISH. The IHC *status* was considered, since it was measured for a larger number of individuals. Whenever both IHC *status* and FISH *status* were available for a given patient, the FISH *status* was considered instead, as FISH is a more accurate test for classifying individuals into HER2 "positive" or "negative". Having the final *status* of samples for ER, PR, and HER2 expression, the Y response vector was then built as explained above. Samples showing at least one "negative" label for any of the three genes and the remaining genes' labels set to "indeterminate" were excluded from the dataset, yielding a total of 1019 samples (160 TNBC and 859 non-TNBC) accounted for further data analysis.

A total number of 28 individuals were marked as suspect when no concordance existed between the HER2 IHC *level* and *status*, and between the HER2 IHC *status* and FISH *status*. Special attention will be given to individuals for whom non-concordance between lab testing exists and the choice of one or another determines the final label (TNBC vs. non-TNBC). These suspect individuals are potentially mislabeled and are therefore potentially outliers. We will verify whether they belong to the list of outlying individuals detected in this study. The age and race of the variables were also included as explanatory variables, since they were statistically significant in separate univariate logistic regressions to predict the individuals' *status*. The final dataset consisted of 924 samples (153 TNBC and 771 non-TNBC), after excluding samples with no age and/or race records. This data set will be referred to as the training data with the remaining 95 samples being the test set.

3 Methods

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote the matrix of the predictors and $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$ the response vector. The logistic model assumes that the univariate response Y_i only takes the values 0 or 1, with

$$P[Y_i = 1] = \pi(\boldsymbol{x}_i) \text{ and } P[Y_i = 0] = 1 - \pi(\boldsymbol{x}_i)$$

where $\boldsymbol{x}_i{}^t = (x_{i1}, \ldots, x_{ip})$ is the i-th row of $\boldsymbol{X}$, and $\pi(\boldsymbol{x}_i)$ is given by

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i{}^t\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i{}^t\boldsymbol{\beta})}$$

in which $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ is the column vector containing the p regression coefficients.

Typically, the regression coefficients $\boldsymbol{\beta}$ are estimated using the maximum likelihood estimator which minimizes the negative log-likelihood function

$$\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} d(\boldsymbol{x}_i{}^t, y_i; \boldsymbol{\beta})$$

where the *deviance* is $d(\boldsymbol{x}_i{}^t, y_i; \boldsymbol{\beta}) = \log\left(1 + \exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right)\right) - y_i\boldsymbol{x}_i^t\boldsymbol{\beta}$.

However, when the number of variables p is large, standard maximum likelihood estimators can be very difficult to interpret. Also the predictive power of a model may be impacted when including too many variables.[20,21] Moreover, when $p > n$ the maximum likelihood estimator cannot even be computed because there are more unknowns $\beta_1, \ldots, \beta_p$ than the number of terms n in the objective function $\sum_{i=1}^{n} d(\boldsymbol{x}_i{}^t, y_i; \boldsymbol{\beta})$, which makes the solution underdetermined. A possible solution to this problem is to consider the so-called shrinkage estimators (for a review, see e.g. Tibshirani[22]). For these estimators, a penalty on the regression coefficients is included in the objective function. The penalty adds another p terms to the objective function which then contains more terms than unknowns. In the next section, we will discuss several shrinkage estimators.

3.1 Sparse logistic regression

One of the first shrinkage estimators in the literature is ridge regression.[23,24] The estimator for $\boldsymbol{\beta}$ then becomes the vector $\widehat{\boldsymbol{\beta}}_{ridge}$ minimizing

$$\sum_{i=1}^{n} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + \lambda ||\boldsymbol{v} \cdot \boldsymbol{\beta}||_2^2 = \sum_{i=1}^{n} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + n\lambda \sum_{j=1}^{p} \left[v_j \beta_j\right]^2$$

Here $\cdot$ stands for the elementwise product. The tuning parameter $\lambda > 0$ controls the severity of the penalty and thus the level of shrinkage. A higher value of λ will lead to a higher importance of the penalty and thus a higher percentage of coefficients pulled towards zero. The vector $\boldsymbol{v}$ contains the penalty factors that control how much of the penalty λ affects each coefficient. If the jth component of $\boldsymbol{v}$ is zero, the coefficient β_j is not penalized. On the other hand if $v_j = 1$, the full penalty λ is applied to β_j.

A downside of ridge regression is that it cannot shrink coefficients exactly to zero, thus all variables will be retained in the selected model.[20,21,25] The LASSO estimator,[21] which employs an l_1 penalty instead of the l_2, may be used to solve this problem. It performs variable shrinkage and variable selection at the same time. The LASSO estimate of $\boldsymbol{\beta}$ is the vector $\widehat{\boldsymbol{\beta}}_{LASSO}$ minimizing

$$\sum_{i=1}^{n} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + \lambda ||\boldsymbol{v} \cdot \boldsymbol{\beta}||_1 = \sum_{i=1}^{n} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |v_j \beta_j|$$

Again the tuning parameter $\lambda > 0$ controls the sparsity of the resulting coefficients. A downside of the LASSO estimator is that it tends to randomly select only one variable in a group of highly correlated variables while discarding the other variables.[26]

The elastic net procedure proposed by Zou and Hastie[26] tries to overcome the downsides of both ridge regression and the LASSO. It shrinks the variables and performs variable selection while being able to select groups of correlated variables. The sparse estimate for $\boldsymbol{\beta}$ then becomes the vector $\widehat{\boldsymbol{\beta}}_{enet}$ minimizing

$$\sum_{i=1}^{n} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + n\lambda \left[(1-\alpha) \frac{||\boldsymbol{v} \cdot \boldsymbol{\beta}||_2^2}{2} + \alpha ||\boldsymbol{v} \cdot \boldsymbol{\beta}||_1 \right]$$

The parameter α controls the mixing between the ridge and LASSO penalty and should be chosen in the interval $[0, 1]$. Clearly, when $\alpha = 0$ the ridge estimator is obtained, whereas for $\alpha = 1$ the LASSO estimate is recovered. The optimal values for λ and α may be obtained using k-fold cross-validation techniques. Software implementations of the elastic net method can be found in the free R software[27] package glmnet.[28]

In the next subsection, we discuss how the elastic net procedure may be modified to make it robust to outliers. We will first discuss a robustification of the non-sparse maximum likelihood technique before turning our attention to a robust sparse procedure.

3.2 Robust sparse logistic regression

The maximum likelihood estimator is highly susceptible to outliers because both outliers in the predictor space and outliers in the response variable may have an unbounded effect on the log-likelihood. As a possible alternative, more robust procedures have been proposed. For an outlier-contaminated data set they provide a solution that is close to the one that would be obtained on an outlier-free data set using classical methods. One of the ways to achieve robustness is to use a trimmed log-likelihood function. For linear regression the resulting estimator is called the Least Trimmed Squares (LTS) estimator.[29,30]

For logistic regression the LTS estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LTS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta})_{i:n}$$

where the subscript $i:n$ indicates the ith smallest deviance, i.e. the n deviances are first sorted from smallest to largest. The LTS thus minimizes the trimmed deviance, for a subset of h data points out of the full sample of size n. The number h is typically chosen between $\lfloor (n+p+1)/2 \rfloor$ and n. The choice of h determines the robustness of the LTS estimator. In practice one frequently uses a conservative value of h as an initial choice. To improve efficiency one may then increase h to a higher value, while ensuring that h stays below the number of non-outliers found in the data.

The ideas of LTS regression were adapted for sparse robust logistic regression by Kurnaz et al.[31] and were implemented in the R[27] software package enetLTS.[32] They defined the enetLTS logistic estimator which combines the sparsity of the elastic net procedure with the robustness of LTS regression. In that sense their work is an extension of sparse LTS linear regression[14] that combines the LTS estimator for linear regression with the LASSO penalty. The enetLTS estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{enetLTS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{h} d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta})_{i:n} + h\lambda \left[(1-\alpha) \frac{||\boldsymbol{v} \cdot \boldsymbol{\beta}||_2^2}{2} + \alpha ||\boldsymbol{v} \cdot \boldsymbol{\beta}||_1 \right] \right)$$

where $\lambda \in [0, 1]$ as described for the glmnet penalty.

To increase efficiency, LTS regression usually includes a reweighting step.[10] Generally speaking, the reweighting step identifies outliers according to the above fitted robust LTS model. These are then downweighted before fitting a classical model using these weights. Consider the Pearson residuals

$$r_i^s = \frac{y_i - \pi_i(\boldsymbol{x})}{\sqrt{\pi_i(\boldsymbol{x})(1-\pi_i(\boldsymbol{x}))}}$$

Under the logistic model these are known to be approximately normally distributed. Each observation i then receives a weight w_i of 1 when $|r_i^s| < c$ and 0 otherwise. The cutoff value c is determined as the 97.5 quantile of a standard Gaussian distribution, so that 95% of the distribution lies between –1.96 and 1.96. The reweighted enetLTS estimator is then defined as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} w_i d(\boldsymbol{x}_i^t, y_i; \boldsymbol{\beta}) + n_w\lambda \left[(1-\alpha) \frac{||\boldsymbol{v} \cdot \boldsymbol{\beta}||_2^2}{2} + \alpha ||\boldsymbol{v} \cdot \boldsymbol{\beta}||_1 \right] \right)$$

where $n_w = \sum w_i$ is the number of observations receiving weight one.

3.3 Detecting deviating data cells

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ now denote a data matrix of sample size n and dimension p. In statistics, an outlier typically refers to a row (case) that deviates from the bulk of the data. However, it frequently occurs that such a row is only suspicious for q out of the p variables, with $q \ll p$.[33–37] Flagging the entire row as an outlier would thus be too conservative as the remaining $p-q$ measurements of that row still contain valuable information. To work in this paradigm, we no longer see the data as n rows of p variables, but rather as a data matrix of $n \times p$ cells. Cells with possibly deviating behavior are then referred to as cellwise outliers.

Figure 1 illustrates these two different paradigms. The left panel indicates three rowwise outliers in the data. The right panel identifies several contaminated cells in the data matrix. Even though many rows have one or more outlying cells, they still contain valuable information in their remaining cells.

After applying robust sparse logistic regression, we analyze the selected genes (variables) using the DDC method recently proposed by Rousseeuw and Van den Bossche[15] as a second step. This will provide us with additional insight in the role of the selected genes and the nature of flagged outliers.

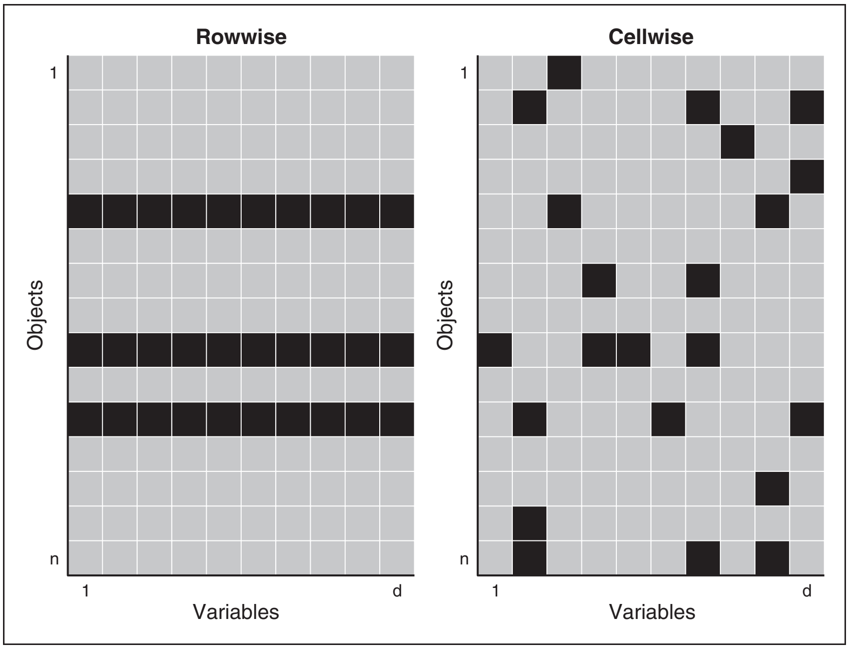

Figure 1. Illustration of the cellwise outlier paradigm versus the typical outlier paradigm.

Table 1. Summary of the fitted models for the robust and non-robust sparse logistic regression methods.

	Sparse logistic regression	Robust sparse logistic regression
α	1.00	0.81
λ	0.005	0.057
# of non-zero coefficients	136	36
Potential outliers	0	43

The DDC procedure uses bivariate correlations between the different variables. It then computes a predicted value for each cell, based on the values of other cells in the same row. Next, it compares the predicted and observed value of each cell. When this robustly standardized difference exceeds a certain cutoff, the cell is flagged. Cells for which the observed value is much lower than the imputed value are colored blue. When the observed value is much higher than the imputed value, the cell is colored red.

4 Results and discussion

We analyzed the TNBC data set using both the sparse logistic and the robust sparse logistic procedures discussed above. In both instances, the parameters α and λ were selected using fivefold cross-validation evaluating the mean of the deviances in the fold. To eliminate randomness in the selection of the folds, the cross-validation was averaged over 10 runs. For the robust sparse logistic regression method, the parameter h was selected as $0.85n$. This parameter was found to be a safe level guarding against outliers after an initial run with $h = 0.5n$. The advantage of choosing a higher $h > 0.5n$ is an increase in efficiency (accuracy), but at the same time one has to make sure that h remains below n minus the number of flagged outliers. The penalty factor $\boldsymbol{v}$ was chosen to be a unity vector penalizing all coefficients equally, except for the coefficients of ER, PR, and HER2 for which $v_j = 0.5$ as these are of special interest in the TNBC context.

Table 1 summarizes the sparse fits. For both procedures, we list the selected α and λ parameters in the glmnet procedure and the resulting number of non-zero coefficients. We also provide the number of observations that are flagged as outliers with respect to the fitted model. The criterion used to flag outliers in the classical sparse logistic model corresponds to the procedure described in the reweighting step of the robust sparse logistic regression method.

The results in Table 1 show that the cross-validation leads to different parameter choices for α and λ between the robust and non-robust sparse logistic regression method. The classical method selects roughly four times as many coefficients (genes) as the robust method. Moreover, only three of the 136 genes selected by the classical method are also selected by the robust method, namely *ER*, *HER2*, and *PPP1R14C*. While the non-robust method fails to identify outlying observations, due to the masking effect described in the Introduction, the robust procedure indicates 43 observations as potential outliers, 12 in common with the set of 24 outliers identified by ensemble outlier detection technique in the study of Lopes et al.[38] It is important to note that all 43 outliers flagged by the robust method are marked as non-TNBC patients in the data, but are predicted to be TNBC according to the fitted model. If indeed true, these patients would receive ineffective, costly, and potentially toxic therapies. It is therefore important to detect all such cases.

The observed differences between the classical sparse logistic regression and its robust counterpart are in concordance with the simulation study performed by Kurnaz et al.[31] They compared the performance of both estimators using clean and contaminated simulated data by several performance measures including the false positive rate (or the proportion of non-informative variables that are incorrectly included in the model) and the false negative rate (or the proportion of informative variables that are incorrectly excluded from the model). Their results indicate that for contaminated data classical sparse logistic regression suffers from a high false negative rate up to almost a 90%. This means that most of the informative variables are excluded from the selection. The number of potential outliers found by the robust procedure provides a strong indication that we are in such a contaminated data case. This explains why there is only a small overlap between the selected genes by the two methods. To compensate for leaving out these informative genes, a larger set of less informative genes is picked up by the classical method. For a more detailed discussion, see Kurnaz et al.[31]

As neither model selects the variables age and race, the 95 observations that were initially left out of the analysis due to missing values for either of these variables may be used for out-of-sample testing. The prediction of TNBC occurrence does not match with the data in 4 out of 95 cases for the classical method and in 6 of the 95 cases for the robust method. These observations may be considered to be discordant cases for which two are predicted differently from the observed response by both the classical and robust procedure. Correspondingly, from the four discordant cases by the classical method, two are only discordant for the classical method, whilst for the robust procedure four out of six cases are only discordant for the robust procedure. Both methods find two additional outliers in the test set corresponding to patients who were attributed a non-TNBC status in the data. From the six cases who did not match according to the robust method, two were borderline cases with a predicted TNBC probability of 0.53 and 0.57, respectively. Even though the robust model selects only a handful of genes compared to the classical method, its out of sample performance is comparable to traditional methods.

4.1 Detailed discussion of identified outliers

We now turn our attention to the observations flagged as potential outliers by the robust method and investigate how these observations differ from the others. The expression of ER, PR, and HER2 for the flagged outliers, along with their clinical label (TNBC vs. non-TNBC), can be found in Table 2.

All observations flagged as outliers are originally labeled as non-TNBC individuals, based on ER and/or PR expression, and/or HER2 amplification. From the list of 43 outliers detected, 7 belonged to the group of 28 observations that were previously identified as suspect, i.e. individuals for whom there is no concordance between lab methods for the HER2 determination (see section 2). This is particularly critical for individuals ER and PR "negative" and for which the HER2 label determines the final TNBC vs. non-TNBC label, like for example individuals 12, 34, and 40.

Several inconsistencies between clinical labeling and gene expression can be also observed for ER and PR. An ER expression value of 0.03, for example, has corresponding positive label for individual 8, while an ER value of 0.63 is translated into a negative label for individual 11. Regarding PR labeling, the same PR expression value of 0.03 corresponds to a negative label for individual 8 and to a positive label for individual 10. Similarly, individual 2 with a PR expression value of 0.25 has a positive PR label, while individual 36 with a PR expression value of 0.26 has a negative PR label. This can be justified by the fact that the clinical threshold for ER and PR positivity is >1% of positive cells within the tumor. Not only IHC scoring has inherent subjectivity, but also receptors positivity can be achieved at very different levels of gene expression. The same happens for HER2. Individual 5 was labeled negative with an HER2 expression value of 15.1, and for the same expression value individual 12 was

Table 2. Summary of the 43 individuals identified as outliers by robust sparse logistic regression regarding ER, PR, and HER2 gene expression and corresponding clinical label (within parentheses).

		Genes						
		ER	PR	HER2				
					(clinical)			
	Individual	FPKM (clinical)	FPKM (clinical)	FPKM	*level* IHC	*status* IHC	*status* FISH	Clinical type
1	TCGA-AR-A1AO	1.47(+)	1.13(−)	14.89	(−)	(−)		non-TNBC
2	TCGA-BH-A6R9	0.59(−)	0.25(+)	8.18		(−)		non-TNBC
3	TCGA-AC-A62X	0.19(+)	0.02(−)	28.53				non-TNBC
4	TCGA-A2-A0YJ	0.09(+)	0.03(−)	240.24	(−)	(−)		non-TNBC
5	**TCGA-LL-A5YP**	**0.16(+)**	**0.05**(−)	**15.10**	(−)	(−)	(+)	**non-TNBC**
6	TCGA-A7-A13D	0.52(−)	0.81(+)	42.28	(Ind)	(Equiv)	(−)	non-TNBC
7	TCGA-E2-A1II	0.14(−)	0.19(+)	10.73	(−)	(−)		non-TNBC
8	TCGA-AR-A1AH	0.03(+)	0.03(−)	34.12		(−)		non-TNBC
9	TCGA-BH-A0DL	6.99(+)	0.04(−)	9.92		(−)		non-TNBC
10	TCGA-E2-A14Y	0.67(+)	0.03(+)	487.90	(Ind)	(Equiv)	(+)	non-TNBC
11	**TCGA-AO-A0JL**	**0.63**(−)	**0.08**(−)	**63.60**	(−)	(−)	(+)	**non-TNBC**
12	**TCGA-AN-A0FL**	**0.09**(−)	**1.07**(−)	**15.07**	(−)	(+)		**non-TNBC**
13	TCGA-AO-A1KO	10.78(+)	9.12(+)	14.91	(−)	(−)		non-TNBC
14	**TCGA-AN-A0FX**	**1.13**(−)	**0.64**(−)	**24.02**	(−)	(+)		**non-TNBC**
15	TCGA-A1-A0SB	3.16(+)	0.03(−)	32.35		(−)		non-TNBC
16	TCGA-D8-A1JM	5.01(+)	0.01(−)	21.85	(−)	(−)		non-TNBC
17	TCGA-E9-A1NC	0.11(−)	0.08(+)	15.91		(+)		non-TNBC
18	TCGA-A2-A25F	0.62(−)	0.23(+)	5.19		(−)		non-TNBC
19	TCGA-A2-A1G1	0.53(−)	0.17(−)	819.76	(Ind)	(Equiv)	(+)	non-TNBC
20	TCGA-LL-A6FR	0.33(−)	0.04(+)	32.13	(Ind)	(Equiv)	(+)	non-TNBC
21	TCGA-A2-A3Y0	2.18(+)	0.03(−)	11.34	(−)	(−)		non-TNBC
22	TCGA-B6-A0IJ	1.18(+)	0.46(+)	11.12				non-TNBC
23	TCGA-AR-A0TP	0.04(+)	0.03(−)	13.39		(−)		non-TNBC
24	TCGA-S3-AA0Z	16.67(+)	0.07(+)	33.07	(−)	(Equiv)	(−)	non-TNBC
25	TCGA-A2-A4S1	0.29(+)	0.01(−)	0.61		(−)		non-TNBC
26	TCGA-A7-A13E	0.82(+)	0.06(−)	46.08	(Ind)	(Equiv)	(−)	non-TNBC
27	TCGA-D8-A1JK	0.40(−)	0.72(+)	22.19	(−)	(−)		non-TNBC
28	TCGA-E9-A1ND	1.44(−)	0.05(−)	13.05		(+)		non-TNBC
29	**TCGA-JL-A3YW**	**0.35(+)**	**0.09(+)**	**31.47**	(−)	**(+)**		**non-TNBC**
30	**TCGA-AN-A0FJ**	**0.08(+)**	**0.04**(−)	**14.28**	(−)	(+)		**non-TNBC**
31	TCGA-D8-A1XW	0.32(−)	0.11(+)	21.03	(−)	(−)		non-TNBC
32	TCGA-UU-A93S	0.30(−)	0.12(−)	1668.35	(+)	(+)		non-TNBC
33	TCGA-OL-A5S0	0.09(+)	0.06(−)	31.92			(+)	non-TNBC
34	TCGA-E9-A22G	0.44(−)	0.02(−)	15.32		(+)		non-TNBC
35	TCGA-AR-A24Q	1.00(+)	0.36(−)	20.67		(−)		non-TNBC
36	TCGA-E2-A1B0	0.14(−)	0.26(−)	563.81	(+)	(+)		non-TNBC
37	TCGA-AR-A251	1.57(+)	0.10(−)	14.02	(Ind)	(Equiv)	(−)	non-TNBC
38	TCGA-A2-A4RX	0.68(+)	0.93(+)	26.64	(−)	(−)		non-TNBC
39	TCGA-AR-A1AJ	1.47(+)	0.07(−)	9.74		(−)		non-TNBC
40	**TCGA-A2-A04U**	**0.02**(−)	**0.02**(−)	**9.64**	(−)	(−)	(+)	**non-TNBC**
41	TCGA-BH-A5IZ	5.12(+)	0.03(−)	28.08		(−)	(−)	non-TNBC
42	TCGA-D8-A13Y	15.48(+)	4.17(+)	4.83	(−)	(−)		non-TNBC
43	TCGA-LL-A8F5	1.08(+)	0.04(−)	11.86	(−)	(−)		non-TNBC

Note: Individuals highlighted in bold correspond to individuals previously identified as suspicious as described in the Data description section. FPKM: fragments per kilobase million; Ind: indeterminate; Equiv: equivocal; ER: estrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2.

labeled as HER2 positive. Wrong labeling in one or more variables clearly impacts final labeling of individuals into TNBC and non-TNBC, with serious consequences in clinical decision and prognosis.

For other individuals, however, the outlyingness cannot be explained by mislabeling of the three TNBC-associated gene expressions, as they seem to have concordant gene expression and label (see e.g. individuals 32, 36, and 42). This suggests that these individuals are correctly classified but might exhibit abnormal expression values for some of the measured genes.

4.2 Discussion of selected genes with relation to outlier identification

To gain more insight, we ran the DDC algorithm on the 36 selected genes only, without telling DDC anything about the clinical response variable or which rows were flagged as outliers. The result is a cell map with over 1000 rows, which is hard to visualize. Therefore, Figure 2 instead shows (from top to bottom) the first 30 non-TNBC patients, 30 TNBC patients, and the 43 outliers found by the robust logistic fit. The indices of the outliers correspond to the row numbers in Table 2. A blue cell in Figure 2 indicates an unexpectedly low gene expression value whereas a red cell indicates an unexpectedly high value, relative to the other cells in its row and using the correlations between the columns. We see that the overall pattern detected by the DDC algorithm for the patients flagged as potential outliers (all originally labeled as non-TNBC) matches the pattern observed for the TNBC patients. This is a very strong indication that indeed these patients have an erroneous label in the data.

Genes for which most of the cells of the TNBC patients are colored are of particular interest. Additional evidence for their role in classifying TNBC patients is provided by the sign and size of their coefficients in the robust sparse logistic model. Figure 3 depicts the coefficient in the robust sparse logistic model for each of the genes, using the same color coding as in the DDC map. The coefficients of the ER, PR, and HER2 genes have been colored black. We indeed see that the genes standing out in the DDC map are mostly those genes with higher coefficients, in absolute value, in the model. The red-colored genes turn out to get positive coefficients, the blue genes get negative coefficients, and the yellow ones get coefficients closer to zero. The selected genes may thus be of particular biological and medical interest for the understanding and diagnosis of TNBC.

Table 3 lists the genes selected by the robust sparse logistic method. It also includes the corresponding coefficients estimated by the robust sparse logistic method, rounded to three digits. The color coding as determined by the DDC map is also noted and corresponds to the color coding in Figure 3.

4.3 Biological interpretation of selected genes and correlation structures

Among the 36 genes listed in Figure 3, 13 (36.1%) were down-regulated in TNBC and 23 (63.9%) were up-regulated. The majority of genes found to be down-regulated in TNBC (11/13) were previously reported to be down-regulated in this particular subtype of BC or overexpressed in ER+/HER2+ breast tumors. These include *ESR1*, *PGR*, and *HER2*, but also *AGR2*,[39] *CA12*,[40] *FOXA1*,[41] *GATA3*,[42,43] *MLPH*,[44,45] *SPDEF*,[46] *SPARCL1*,[47] and *TGFB3*.[48] Also 11 of the 23 genes up-regulated in TNBC were previously described as such, namely *ART3*,[49] *B3GNT5*,[50] *EN1*,[51,52] *FOXC1*,[53] *FZD9*,[54] *HORMAD1*,[55] *POU5F1*,[56] *ROPN1*,[57] *TMSB15A*,[58] *UGT8*,[59] and *VGLL1*[60].

Our analysis has led to the identification of 14 genes that were not previously reported as specifically involved in TNBC or (breast) cancer overall, therefore contributing to the search for new interest biomarkers to further validate and functionally study. These include *JAM3* and *PODN*, down-regulated in TNBC; and *SFT2D2*, *CDCA2*, *CHODL*, *CT83*, *FANCE*, *NKX1-2*, *PPP1R14C*, *SRSF12*, *TBC1D22B*, *TMCC2*, *TTLL4*, and *PAPSS1*, up-regulated in TNBC. *JAM3* has been previously identified as a biomarker for cervical cancer[61] and was found to be up-regulated and associated with higher cancer risk in the offspring from mice with high-fat diet intake during pregnancy.[62] Amongst the up-regulated genes, *SFT2D2* was previously described as down-regulated in a bone (specific) metastasis-related gene set.[63] *PAPSS1*, involved in estrogen metabolism, was not directly implicated in TNBC before but found to be overexpressed in breast tumor tissues in comparison to adjacent normal tissue, independently of ER status.[64]

Finally, we consider a graphical representation of the correlation-based network structure between the selected genes depicted in Figure 4. For clarity, only correlations above 0.6 in absolute value are shown. The thickness of the connecting lines represents the strength of the correlation, whereas green represents a positive correlation and red signals a negative correlation. Figure 4(a) shows the correlation plot for the non-TNBC individuals, whereas Figure 4(b) is the plot for TNBC individuals. The patterns in the left and right panels are strikingly different.

The proto-oncogene *AGR2* (anterior gradient homology 2) is among the down-regulated genes in TNBC, and a strong correlation between *AGR2* and *FOXA1* was found in TNBC. Moreover, *AGR2* was correlated with other

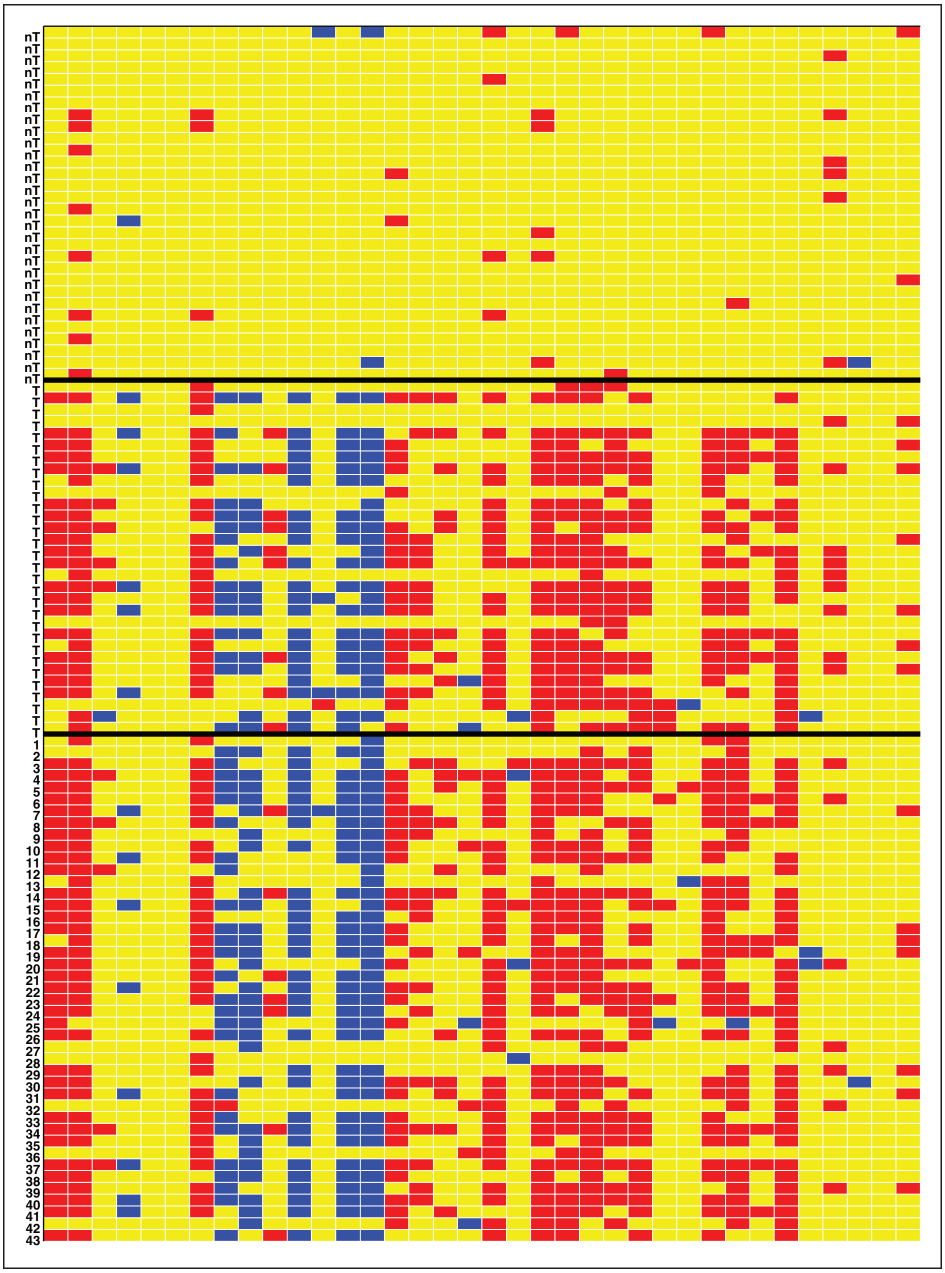

Figure 2. Cellwise outlier map. The columns correspond to the genes selected by the robust sparse logistic model. The rows correspond to 30 non-TNBC patients (label nT), 30 TNBC patients (label T), and the 43 outliers found by the robust fit.

genes, suggesting a particular relevance for this gene. *AGR2* is a known biomarker of poor prognosis in ER+ BC.[65] Accordingly, different studies have reported that expression of the proto-oncogene *AGR2* is induced by estrogen and tamoxifen in BC cells,[66] and that *AGR2* is required for the growth and migration of ER+ cells. The transcription factor *FOXA1* is implicated in the regulation of many ERα-target genes, including *AGR2*. This justifies the multiple correlations we found between *FOXA1* and other genes in non-TNBC. However, in tamoxifen-resistant cells, the expression of *AGR2* occurs in a constitutive manner, requiring *FOXA1*, but loses its dependence on ER, suggesting a mechanism where changes in *FOXA1* activity obviate the need for ER in the regulation of this gene.[67] It is hypothesized that *AGR2* may be involved in the folding of the extracellular domains of proteins that influence cell growth and survival, and that *AGR2* may represent an important biomarker and therapeutic target in BC.[68] It thus appears that the *FOXA1–AGR2* link in TNBC may be of particular relevance and deserves further study.

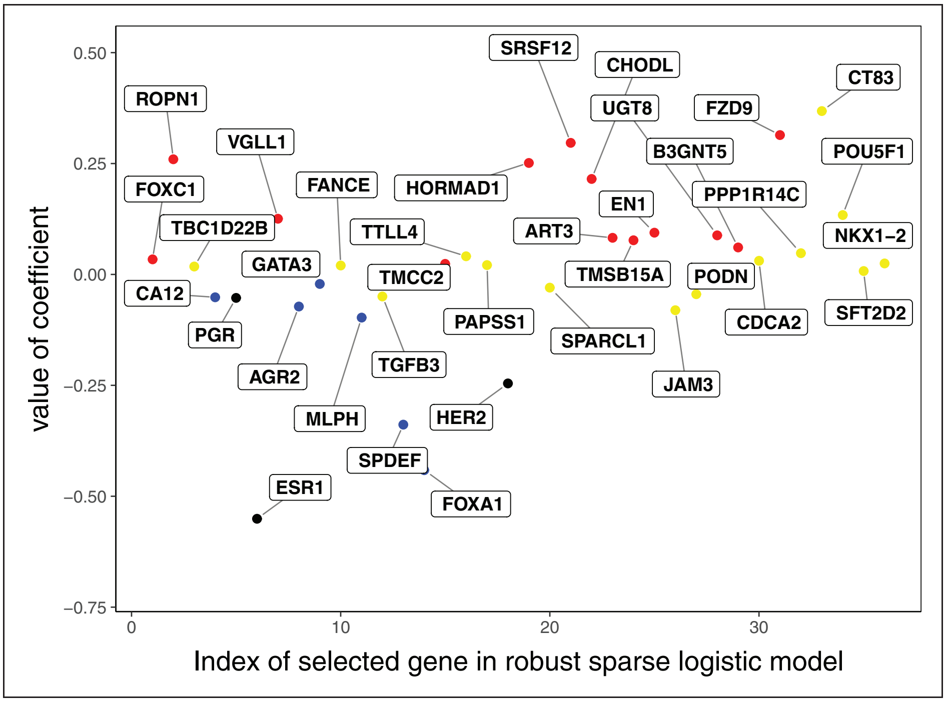

Figure 3. Interpretation of genes selected in the robust sparse logistic model. The color coding corresponds to the color determined by the DDC map.

Table 3. Genes selected by the robust sparse logistic method, corresponding coefficients (rounded to 3 digits) and their color coding.

	Gene	Coef	Color		Gene	Coef	Color
0	Intercept	0.225	None	19	*TMCC2*	0.024	Red
1	*CT83*	0.368	Yellow	20	*PAPSS1*	0.021	Yellow
2	*FZD9*	0.314	Red	21	*FANCE*	0.020	Yellow
3	*SRSF12*	0.297	Red	22	*TBC1D22B*	0.018	Yellow
4	*ROPN1*	0.260	Red	23	*SFT2D2*	0.008	Yellow
5	*HORMAD1*	0.252	Red	24	*GATA3*	−0.021	Blue
6	*CHODL*	0.215	Red	25	*SPARCL1*	−0.030	yellow
7	*POU5F1*	0.134	Yellow	26	*PODN*	−0.044	Yellow
8	*VGLL1*	0.125	Red	27	*TGFB3*	−0.050	Yellow
9	*EN1*	0.094	Red	28	*CA12*	−0.051	Blue
10	*UGT8*	0.088	Red	29	*PGR*	−0.053	Yellow
11	*ART3*	0.083	Red	30	*AGR2*	−0.072	Blue
12	*TMSB15A*	0.077	Red	31	*JAM3*	−0.080	Yellow
13	*B3GNT5*	0.061	Red	32	*MLPH*	−0.097	Blue
14	*PPP1R14C*	0.048	Yellow	33	*HER2*	−0.246	Yellow
15	*TTLL4*	0.041	Yellow	34	*SPDEF*	−0.338	Blue
16	*FOXC1*	0.034	Red	35	*FOXA1*	−0.441	Blue
17	*CDCA2*	0.031	Yellow	36	*ESR1*	−0.551	Yellow
18	*NKX1-2*	0.025	Yellow				

Note: The genes are sorted by their coefficient.

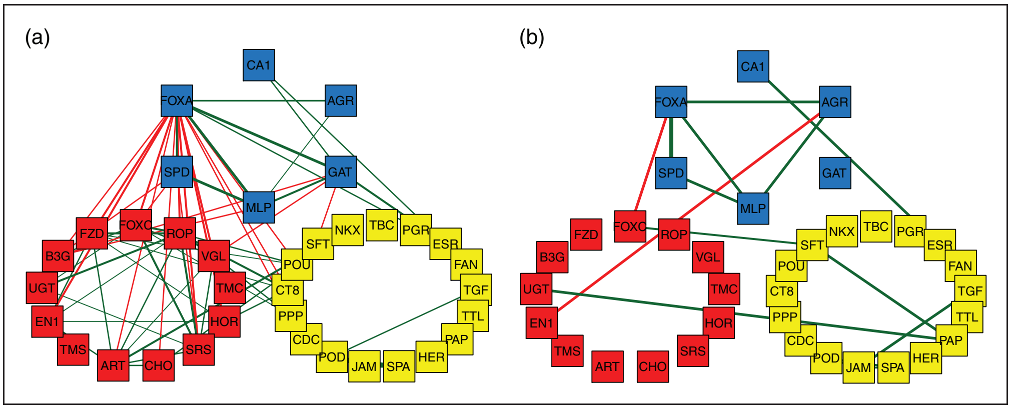

Figure 4. Representation of the correlation between the genes selected by the robust sparse logistic model. The color coding corresponds to the color determined by the DDC map. (a) Correlations for non-TNBC patients. (b) Correlations for TNBC patients.

The fact that the biological role in TNBC of approximately 60% of the selected genes has been previously reported strengthens our analysis and fosters investigation on the potential role of the remaining selected genes in BC and in particular TNBC.

5 Conclusion

This work shows that robust sparse logistic regression can be a powerful tool in precision medicine. It enables accurate prediction of the BC subtype, irrespective of the possible presence of outliers situated in either the clinical label or in the gene expression data. In contrast, classical sparse logistic regression methods are severely affected by the outliers in the data. At the same time, robust methodology allows to inspect the detected outliers which may lead to the correct status of the patient and the prescription of the appropriate treatment. Due to the sparse nature of this robust regression, genes included in the model may be highly relevant to the understanding of TNBC. While 60% of the selected genes were previously reported to be related to TNBC or BC, the remaining identified genes deserve further attention as potential biomarkers for the disease. Among the selected genes, biologically relevant gene networks could be identified both for TNBC and non-TNBC patient data, particularly the strong *FOXA1–AGR2* link in TNBC. These results are intended to contribute to BC/TNBC understanding, the definition of new therapies and ultimately more effective TNBC management.

Acknowledgments

The authors thank André Veríssimo and Eunice Carrasquinha for insightful discussions during problem identification and data processing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the International Funds KU Leuven (grant C16/15/068); the European Union Horizon 2020 research and innovation program (grant no. 633974 (SOUND project)), and the Portuguese Foundation for Science & Technology (FCT), through projects UID/EMS/50022/2013 (IDMEC, LAETA), UID/CEC/50021/2013 (INESC-ID), PTDC/EMS-SIS/0642/2014, and IF/00653/2012. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Flemish Research Foundation (FWO) and the Flemish Government, department EWI.

ORCID iD

Pieter Segaert http://orcid.org/0000-0003-3219-7242
Susana Vinga http://orcid.org/0000-0002-1954-5487

References

1. Reis-Filho JS and Tutt ANJ. Triple negative tumours: a critical review. *Histopathology* 2008; **52**: 108–118.
2. Dent R, Trudeau M, Pritchard KI, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res* 2007; **13**: 4429–4434.
3. Foulkes WD, Smith IE and Reis-Filho JS. Triple-negative breast cancer. *N Engl J Med* 2010; **363**: 1938–1948.
4. Wang Q, Gao S, Li H, et al. Long noncoding rnas (lncrnas) in triple negative breast cancer. *J Cell Physiol* 2017; **232**: 3226–3233.
5. Vucic EA, Thu KL, Robison K, et al. Translating cancer 'omics' to improved outcome. *Genome Res* 2012; **22**: 188–195.
6. Hammond MEH, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 2010; **28**: 2784–2795.
7. Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2013; **24**: vi7–vi23.
8. Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J Clin Oncol* 2013; **31**: 3997–4013.
9. Payne SJ, Bowen RL, Jones JL, et al. Predictive markers in breast cancer – the present. *Histopathology* 2013; **52**: 82–90.
10. Rousseeuw PJ and Leroy A. *Robust regression and outlier detection*. New York: Wiley-Interscience, 1987.
11. Maronna RA, Martin DR and Yohai VJ. *Robust statistics: theory and methods*. New York: Wiley, 2006.
12. Rousseeuw PJ and van Zomeren BC. Unmasking multivariate outliers and leverage points. *J Am Stat Assoc* 1990; **85**: 633–639.
13. Serfling R and Wang S. General foundations for studying masking and swamping robustness of outlier identifiers. *Stat Methodol* 2014; **20**: 79–90.
14. Alfons A, Croux C and Gelper S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat* 2013; **7**: 226–248.
15. Rousseeuw PJ and Van den Bossche W. Detecting deviating data cells. *Technometrics* 2018; **60**: 135–145.
16. TCGA. The cancer genome atlas, https://cancergenome.nih.gov/ (2017, accessed May 2017).
17. Veríssimo A. *TCGA.DATA R package*, https://github.com/averissimo/tcga.data/releases/tag/2016.12.15-brca (2017, accessed 3 August 2017).
18. Ensembl. The ensembl genome browser, www.ensembl.org/index.html (2017, accessed May 2017).
19. CCDS. The consensus cds (ccds) project, www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi (2017, accessed May 2017).
20. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics* 1995; **37**: 373–384.
21. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 1996; **58**: 267–288.
22. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Ser B Stat Methodol* 2011; **73**: 273–282.
23. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
24. Le Cessie S and Van Houwelingen J. Ridge estimators in logistic regression. *J R Stat Soc Ser C Appl Stat* 1992; **41**: 191–201.
25. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
26. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B* 2005; **67**: 301–320.
27. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, www.R-project.org/ (2017, accessed 3 August 2018).
28. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1–22.
29. Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc* 1984; **79**: 971–880.
30. Rousseeuw PJ and Van Driessen K. Computing LTS regression for large data sets. *Data Min Knowl Discov* 2006; **12**: 29–45.
31. Kurnaz FS, Hoffmann I and Filzmoser P. Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometr Intell Lab Syst* 2018; **172**: 211–222.
32. Kurnaz FS, Hoffmann I and Filzmoser P. *enetLTS: robust and sparse methods for high dimensional linear and logistic regression*. R package version 0.1.0, https://CRAN.R-project.org/package=enetLTS (2018, accessed 3 August 2018).
33. Alqallaf F, Van Aelst S, Yohai VJ, et al. Propagation of outliers in multivariate data. *Ann Stat* 2009; **37**: 311–331.

34. Van Aelst S, Vandervieren E and Willems G. A Stahel–Donoho estimator based on huberized outlyingness. *Comput Stat Data Anal* 2012; **56**: 531–542.

35. Agostinelli C, Leung A, Yohai VJ, et al. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 2015; **24**: 441–461.

36. Öllerer V, Alfons A and Croux C. The shooting s-estimator for robust regression. *Comput Stat* 2016; **31**: 829–844.

37. Leung A, Zhang H and Zamar R. Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Comput Stat Data Anal* 2016; **99**: 1–11.

38. Lopes MB, Veríssimo A, Carrasquinha E, et al. Ensemble outlier detection and gene selection in triple-negative breast cancer data. *BMC Bioinformatics* 2018; **19**: 168.

39. Tian SB, Tao KX, Hu J, et al. The prognostic value of agr2 expression in solid tumours: a systematic review and meta-analysis. *Sci Rep* 2017; **7**: 1–10.

40. Christgen M, Geffers R, Kreipe H, et al. Iph-926 lobular breast cancer cells are triple-negative but their microarray profile uncovers a luminal subtype. *Cancer Sci* 2013; **104**: 1726–1730.

41. Guiu S, Charon-Barra C, Vernerey D, et al. Coexpression of androgen receptor and foxa1 in nonmetastatic triple-negative breast cancer: ancillary study from pacs08 trial. *Future Oncol* 2015; **11**: 2283–2297.

42. Krings G, Nystrom M, Mehdi I, et al. Diagnostic utility and sensitivities of gata3 antibodies in triple-negative breast cancer. *Hum Pathol* 2014; **45**: 2225–2232.

43. Cimino-Mathews A, Subhawong AP, Illei PB, et al. Gata3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Hum Pathol* 2013; **44**: 1341–1349.

44. Thakkar A, Raj H, Ravishankar, et al. High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. *Biomark Insights* 2015; **10**: 103–112.

45. Thakkar AD, Raj H, Chakrabarti D, et al. Identification of gene expression signature in estrogen receptor positive breast carcinoma. *Biomark Cancer* 2010; **2**: 1–15.

46. Turcotte S, Forget MA, Beauseigle D, et al. Prostate-derived ets transcription factor overexpression is associated with nodal metastasis and hormone receptor positivity in invasive breast cancer. *Neoplasia* 2007; **9**: 788–796.

47. Cao F, Wang K, Zhu R, et al. Clinicopathological significance of reduced sparcl1 expression in human breast cancer. *Asian Pac J Cancer Prev* 2013; **14**: 195–200.

48. Chen C, Zhao KN, Masci PP, et al. Tgfβ isoforms and receptors mrna expression in breast tumours: prognostic value and clinical implications. *BMC Cancer* 2015; **15**: 1010.

49. Tan L, Song X, Sun X, et al. Art3 regulates triple-negative breast cancer cell function via activation of akt and erk pathways. *Oncotarget* 2016; **7**: 46589–46602.

50. Potapenko IO, Lüders T, Russnes HG, et al. Glycan-related gene expression signatures in breast cancer subtypes: relation to survival. *Mol Oncol* 2015; **9**: 861–876.

51. Beltran AS, Graves LM and 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* 2014; **33**: 4767–4777.

52. Kim YJ, Sung M, Oh E, et al. Engrailed 1 overexpression as a potential prognostic marker in quintuple-negative breast cancer. *Cancer Biol Ther* 2018; **15**: 1–11.

53. Jensen TW, Ray T, Wang J, et al. Diagnosis of basal-like breast cancer using a foxc1-based assay. *J Natl Cancer Inst* 2015; 107(8): 148.

54. Conway K, Edmiston SN, May R, et al. DNA methylation profiling in the carolina breast cancer study defines cancer subclasses differing in clinicopathologic characteristics and survival. *Breast Cancer Res* 2014; **16**: 450.

55. Watkins J, Weekes D, Shah V, et al. Genomic complexity profiling reveals that hormad1 overexpression contributes to homologous recombination deficiency in triple-negative breast cancers. *Cancer Discov* 2015; **5**: 488–505.

56. Liu CG, Lu Y, Wang BB, et al. Clinical implications of stem cell gene oct-4 expression in breast cancer. *Ann Surg* 2011; **253**: 1165–1171.

57. Ivanov SV, Panaccione A, Nonaka D, et al. Diagnostic sox10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. *Br J Cancer* 2013; **109**: 441–451.

58. Darb-Esfahani S, Kronenwett R, von Minckwitz G, et al. Thymosin beta 15a (tmsb15a) is a predictor of chemotherapy response in triple-negative breast cancer. *Br J Cancer* 2012; **107**: 1892–1900.

59. Dziegiel P, Owczarek T, Plazùk E, et al. Ceramide galactosyltransferase (ugt8) is a molecular marker of breast cancer malignancy and lung metastases. *Br J Cancer* 2010; **103**: 524–531.

60. Castilla MA, López-García MA, Atienza MR, et al. Vgll1 expression is associated with a triple-negative basal-like phenotype in breast cancer. *Endocr Relat Cancer* 2014; **21**: 587–599.

61. Eijsink JJ, Lendvai A, Deregowski V, et al. A four-gene methylation marker panel as triage test in high-risk human papillomavirus positive patients. *Int J Cancer* 2012; **130**: 1861–1869.

62. Nguyen M, Andrade FO, Jin L, et al. Maternal intake of high n-6 polyunsaturated fatty acid diet during pregnancy causes transgenerational increase in mammary cancer risk in mice. *Breast Cancer Res* 2017; **19**: 77.

63. Savci-Heijink CD, Halfwerk H, Koster J, et al. A novel gene expression signature for bone metastasis in breast carcinomas. *Br Cancer Res Treat* 2016; **156**: 249–259.

64. Xu Y, Liu X, Guo F, et al. Effect of estrogen sulfation by sult1e1 and papss on the development of estrogen-dependent cancers. *Cancer Sci* 2012; **103**: 1000–1009.
65. Barraclough DL, Platt-Higgins A, Rudland SS, et al. The metastasis-associated anterior gradient 2 protein is correlated with poor survival of breast cancer patients. *Am J Pathol* 2009; **175**: 1848–1857.
66. Hrstka R, Nenutil R, Fourtouna A, et al. The pro-metastatic protein anterior gradient-2 predicts poor prognosis in tamoxifen-treated breast cancers. *Oncogene* 2010; **29**: 4838–4847.
67. Wright TM, Wardell SE, Jasper JS, et al. Delineation of a foxa1/erα/agr2 regulatory loop that is dysregulated in endocrine therapy-resistant breast cancer. *Mol Cancer Res* 2014; **12**: 1829–1839.
68. Salmans ML, Zhao F and Andersen B. The estrogen-regulated anterior gradient 2 (agr2) protein in breast cancer: a potential drug target and biomarker. *Breast Cancer Res* 2013; **15**: 204.