Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance

Laura Macías-García [b], María Martínez-Ballesteros [a], José María Luna-Romera [a], José M. García-Heredia [c], Jorge García-Gutiérrez [a,*], José C. Riquelme-Santos [a]

[a] Department of Computer Languages and Systems, School of Computer Engineering, University of Seville, Seville, Spain
[b] Department of Citology and Histology, Faculty of Medicine, University of Seville, Seville, Spain
[c] Department of Plant Biochemistry and Molecular Biology, University of Seville, Seville, Spain

Keywords:

Autoencoder
Breast cancer
DNA methylation
Feature generation
Machine learning

A B S T R A C T

Breast cancer is the most frequent cancer in women and the second most frequent overall after lung cancer. Although the 5-year survival rate of breast cancer is relatively high, recurrence is also common which often involves metastasis with its consequent threat for patients. DNA methylation-derived databases have become an interesting primary source for supervised knowledge extraction regarding breast cancer. Unfortunately, the study of DNA methylation involves the processing of hundreds of thousands of features for every patient. DNA methylation is featured by High Dimension Low Sample Size which has shown well-known issues regarding feature selection and generation. Autoencoders (AEs) appear as a specific technique for conducting nonlinear feature fusion. Our main objective in this work is to design a procedure to summarize DNA methylation by taking advantage of AEs. Our proposal is able to generate new features from the values of CpG sites of patients with and without recurrence. Then, a limited set of relevant genes to characterize breast cancer recurrence is proposed by the application of survival analysis and a pondered ranking of genes according to the distribution of their CpG sites. To test our proposal we have selected a dataset from The Cancer Genome Atlas data portal and an AE with a single-hidden layer. The literature and enrichment analysis (based on genomic context and functional annotation) conducted regarding the genes obtained with our experiment confirmed that all of these genes were related to breast cancer recurrence.

1. Introduction

Breast cancer is the most frequent cancer in women [1] and the second most frequent overall after lung cancer (over two million new cases of lung cancer were diagnosed in 2018 [2]). Although the 5-year survival rate of breast cancer is relatively high, recurrence is also common (at approximately 20% to 30%, depending on the initial stage) which often involves metastasis. One of the major challenges in breast cancer management includes the classification of patients into correct risk groups after initial diagnosis for their most appropriate treatment and follow-up (including risk of recurrence). Risk classification is especially important for the improvement of the monitoring of patients, quality care, and use of medical resources.

Furthermore, DNA methylation-derived databases have become an interesting primary source for supervised knowledge extraction regarding breast cancer [3]. DNA methylation is a well-known data source which shows the functioning of the genome. Alterations in DNA methylation have revealed its significant role in tumorigenesis and tumour-suppression [4]. Unfortunately, the study of DNA methylation involves the processing of hundreds of thousands of features for each patient in the study. Thus, feature selection constitutes a key factor not only in the attainment of knowledge from such a vast pod of medical data [5] but also of improvements in techniques to speed up analysis [6].

Machine learning has been profusely applied to tackle the issue of breast cancer prognosis [7–11]. Recently, neural networks have started to play a major role in the extraction of knowledge from genetic databases [12]. Within neural networks, autoencoders (AEs) appear as a specific technique for conducting nonlinear feature fusion [13] with a double-fold strategy: feature selection and noise reduction. AEs have been applied to breast cancer to improve patients' pathological

* Corresponding author.
E-mail addresses: lmacias@us.es (L. Macías-García), mariamartinez@us.es (M. Martínez-Ballesteros), jmluna@us.es (J.M. Luna-Romera), jmgheredia@us.es (J.M. García-Heredia), jorgarcia@us.es (J. García-Gutiérrez), riquelme@us.es (J.C. Riquelme-Santos).

signatures [14], predict their survival [15–17], identify cancer subtypes [18,19], and select precise reference of samples normal tissue for cancer research [20].

Regarding the use of AEs to deal with DNA methylation, Wang et al. [21] recently explored the application of variational AEs on lung cancer DNA methylation data. A later logistic regression classifier, trained with the encoded latent features, was able to accurately classify cancer subtypes. Visakh et al. [22] also proposed an innovative alignment method that made use of AEs to find functionally consistent and topologically sound alignments of epigenetic signatures from pathway networks. Later, those epigenetic signatures were applied to characterise several types and subtypes of breast, lung, colorectal, and prostate cancer. Similarly, Wang et al. [23] developed software named "DeepMethyl" based on stacked denoising AEs to predict the methylation state of DNA using features inferred from three-dimensional genome topology and DNA sequence patterns. They used the experimental data from immortalised myelogenous leukaemia and healthy lymphoblastoid cell lines to train the learning models and assess prediction performance. Moon and Nakai [24] proposed an integrative analysis of gene expression and DNA methylation using normalisation and unsupervised feature extraction by AEs to identify candidate biomarkers of renal-cell carcinoma. Chaudhary et al. [25] used AEs on DNA methylation, RNA sequencing, and microRNAs sequencing to identify survival subgroups of hepatocellular carcinoma (HCC).

On the other hand, the literature related to the processing of DNA methylation by AEs on breast cancer recurrence is, to the best of our knowledge, scarce although results do exist regarding breast cancer survival prediction, such as those reported by Kim et al. [26]. The authors extracted a single pathway profile matrix out of the gene expression and DNA methylation data by following a random path over an integrated graph. They then applied a denoising AE to the pathway profile to further identify significant features and genes for the validation in a survival prediction task for breast cancer patients.

DNA methylation databases are characterised by having a high dimensionality with few cases (patients). Such High Dimension Low Sample Size (HDLSS) databases have very well-known issues regarding feature selection and generation, especially when Principal Components and other similar techniques are applied [27,28]. In this context, AEs could take advantage of their capacity to reduce the dimensionality while maintaining prediction accuracy [29].

With all the previous in mind, the aim of this work is to develop a framework to process DNA methylation to extract meaningful information from relevant genes regarding breast cancer recurrence. Our proposal has been tested on a dataset from The Cancer Genome Atlas (TCGA) data portal. This and the rest of the main scientific contributions of this paper can be summarised as follows:

- An innovative proposal based on AEs to preprocess DNA methylation and generate autoencoded features to characterise breast cancer recurrence;
- a comparative study regarding how the use of autoencoded feature generation could improve recurrence prediction from DNA methylation data;
- an enrichment and literature analysis to provide insights into how the AEs are related to genes regarding breast cancer recurrence and their similarity with their level of importance reported in the breast cancer literature.

The remainder of this paper is organised as follows. Section 2 describes the experimental data used in this work and the methodology applied. Section 3 shows the results achieved. Section 4 discusses the main findings, provides the study of the literature and the enrichment analysis performed regarding the results. Finally, Section 5 is devoted to a summary of the conclusions and the discussion of future lines of research.

2. Material and methods

2.1. Data description

The data used in this study was downloaded from TCGA data portal [30]. We selected the two types of invasive breast cancer provided by TCGA: ductal (the most common type) [31] and lobular carcinoma [32]. Specifically, the profiles of the platform named Illumina Infinium Human DNA Methylation 450 (HumanMethylation450) were selected. HumanMethylation450 provides the methylation status of more than 450,000 CpG sites contained in the human genome [33] for each patient.

In particular, this platform provides the following information on each tissue: methylation value (which is known as beta value (β)), gene symbol, chromosome, and genomic coordinates [6]. The β estimates the methylation level using a ratio of intensities between methylated and unmethylated alleles, which provides values between 0 (unmethylated) and 1 (fully methylated) [4].

In order to obtain clinical information reported by the TCGA data portal, we focused on the follow-up file (version 4.0) that presented information regarding the monitorization of the health of these patients who participated in a clinical study for a period of time. This file provided interesting fields, such as those specified in Table 1. Further information can be found at https://docs.cancergenomicscloud.org/docs/tcga-metadata.

2.2. Data preprocessing

The hazard of recurrence has been shown higher during the first five years after diagnosis [34]. Prediction of patients who would suffer from recurrence in advance could lead to a more effective treatment. With this in mind, we set up our experimental framework which finished with a survival analysis. Unfortunately, an inherent feature that distinguishes survival analysis from other areas in statistics is that survival data (and TCGA data was not an exception) are usually censored. Censoring happens when incomplete information is available about the survival time of some individuals. To overcome such limitation, we planned a study following a type I censoring design. Type I implies a study in which every patient is under observation for a specified period (in our case, five years) or until failure (recurrence). So initially, we focused on the 749 patients included in the HumanMethylation450 platform as stated before, but after applying type I restrictions, data was reduced to 99 cases. The following paragraphs detail the preprocessing pipeline carried out.

The TCGA provides several tissue samples for each case according to sample type (solid tumour, solid tissue normal, etc.), the portion in a

Table 1
Follow-up file description.

Property	Description
Case ID	Patient identifier
New tumuor event	Boolean value which denotes whether a neoplasm developed after the initial treatment had finished (YES or NO)
Days to new tumour event	Number of days to the date of recurrence after initial treatment has finished (if new tumour event is YES)
Vital status	Dead or Alive
Days to last follow-up	Number of days from the date of last follow-up to the date of initial pathologic diagnosis (if vital status is Alive)
Days to death	Number of days from the date of the initial pathological diagnosis to the date of death (if vital status is Dead)

sequence, and analyte codes (DNA, RNA, etc.), among others. Further information can be found at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables. From the 749 patients, we filtered those cases that contained DNA methylation files, their sample type was a solid tumour (01 code) and the analyte code corresponds to DNA (D code). Not all cases provided DNA Methylation files of the same portions, hence, we selected the portion that was presented in the highest number of cases (11 code) to maximise the set of cases under study. Furthermore, we especially focused on 5-year survival for analysis and therefore we were interested in cases treated for a period of time they was either greater than or equal to five years (1825 days) without recurrence, or was less than 5 years with recurrence. After filtering TCGA with HumanMethylation450 in accordance with the previous conditions, the final set of cases was reduced to 99 cases. For every case, two new variables, REC (recurrence) and TREC (time to recurrence) were calculated according to the follow-up information as follows:

- Cases with recurrence (REC as YES):
  –TREC was the number of days to new tumour event if it was less than or equal to 1825;
  –otherwise, TREC was initialised at 1825 and REC changed to NO.
- Cases with no recurrence (REC as NO):
  –TREC was equal to 1825 (since all of them were filtered to be followed up within at least five years).

In order to perform our experiments, we built a large dataset composed of the beta values provided for each CpG site of the methylation files of every case under study. Every case under study had 485,577 CpG sites. We had to process as many files as cases which entailed building a single file and then transposing the data to attain a dataset for a suitable analysis (i.e., cases as rows and CpG sites as columns). The resulting dataset contained a large number of missing CpG sites for many patients. Therefore, the dataset was filtered excluding the features (CpG sites) with null values or incomplete information for at least one case to avoid non-real and doubtful information in the conducted analysis. Additionally, a class (REC variable) and TREC column were added to every case for a later survival analysis. Big Data technologies such as Apache Spark [35] were employed to manage this large amount of data and to prevent efficiency and memory issues. The final dataset resulted in 383,919 CpG sites as features and 99 cases as instances and was structured as shown in Table 2. Furthermore, a mapping file that related each CpG site of the methylation dataset with the corresponding gene symbol was created to help in the analysis of the results.

2.3. Experimental framework

As can be observed in Fig. 1, the experimental framework was divided into two branches. In the upper branch, machine learning classification algorithms were applied to the original data after a feature selection (FS) based on False Positive Rate (this dataset is called Original from now on), Original and autoencoded features together (Original + AE), Original followed by a Best-K FS (Original + FS) where K was the number of patients, and finally, Original + AE followed by Best-K FS (Original + AE + FS). Then, we evaluated the obtained models comparing the accuracy (see Section 2.3.2).

Table 2
Structure of the input dataset containing the CpG sites for each case under study.

Case ID	cg00000029	…	cg00000905	TREC	REC
TCGA-E2-A2P5	0.12020339	…	0.08950129	597	YES
TCGA-LL-A73Z	0.31853254	…	0.88396413	192	YES
TCGA-A2-A0CR	0.441742832	…	0.04866762	1825	NO
…	…	…	…	…	…

In the second branch, an AE from the complete dataset was developed. Then, we applied survival analysis to the nodes (autoencoded features or hidden units in the AEs) to select only the most significant ones. The weights of these nodes were subsequently analysed with the aim of calculating the genes with the highest importance in the AEs (Section 2.3.3). The same methodology was also applied without filtering the nodes by the survival analysis and all nodes were thus considered in the weight analysis stage to study the importance of feature selection after autoencoding. In the following sections, each of the branches is described in detail.

2.3.1. Autoencoders

An autoencoder (AE) is defined as an artificial neural network with a symmetric structure, whose middle layers encode the input data, and aim to build a version of its input onto the output layer. This kind of artificial neural network includes a mechanism which avoids using a direct copy of the data along with the network [13].

Before generating the AE, several preprocessing steps were carried out on the dataset (see feature selection described in Section 2.3.2 for branch 1.A in Fig. 1). In any case, a normalisation step was always done. The data processing in this methodology was carried out with the scikit-learn library [36].

The AEs were developed and executed by using the Keras library [37]. Keras is an open-source neural network library written in Python, which is capable of running on top of Tensorflow [38]. The AEs were established by Keras library and configured with a single hidden layer. Layers in Keras framework for deep neural networks are mainly controlled by three parameters: number of hidden units, batch size, and number of training epochs. Batch size and number of epochs were set up by default with values in the intervals 5–10 and 100–400, respectively, and fixed by trial and error. For the case of the number of hidden units M, an optimisation procedure selected the best value from 5 to 325 (the maximum size our hardware allowed). This procedure calculated the mean square error (loss function) when autoencoders were trained (only on training sets excluding test folds in the case of a validation procedure) and then selected the best value for the experiments according to the "elbow method" [39].

Moreover, Rectified Linear Units (ReLU) were used as a nonlinear activation function because its output values range in the interval $[0, 1]$ which are suitable values for survival analysis, and have also obtained promising results in other studies in the literature [40].

To avoid overfitting issues, an early-stopping condition was introduced in the training phase of the AEs. To this end, 30% of the training set was kept aside to validate learning (validation set). The early-stopping condition stopped the training when it did not produce an improvement in the validation set in three consecutive epochs, acting as a regularisation method to avoid overfitting.

All the experimentation was carried out on an Intel machine, specifically Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz with 12 cores, 64 GB of RAM, and a NVIDIA GeForce Titan Xp Pascal 12 GB GDDR5X GPU and a NVIDIA Geforce RTX 2080 Ti 11 GB GDDR6 GPU. The source code together with the various experiments performed in this study can be found in [41].

2.3.2. Machine learning classification techniques

As mentioned in Section 1, one of the main objectives of this work is to generate autoencoded features to improve the prediction of breast cancer recurrence. To this end, eight classification algorithms from the scikit-learn library (version 0.22.1, https://scikit-learn.org/stable/whats_new/v0.22.html#version-0-22-1) with default parameters (see a deeper description about the classifiers in [42]) were selected to test their performance.

Four types of techniques were used: decision trees (DecisionTreeClassifier, GradientBoostingClassifier, and RandomForestClassifier implementations in scikit-learn), support vector machines (SVC and

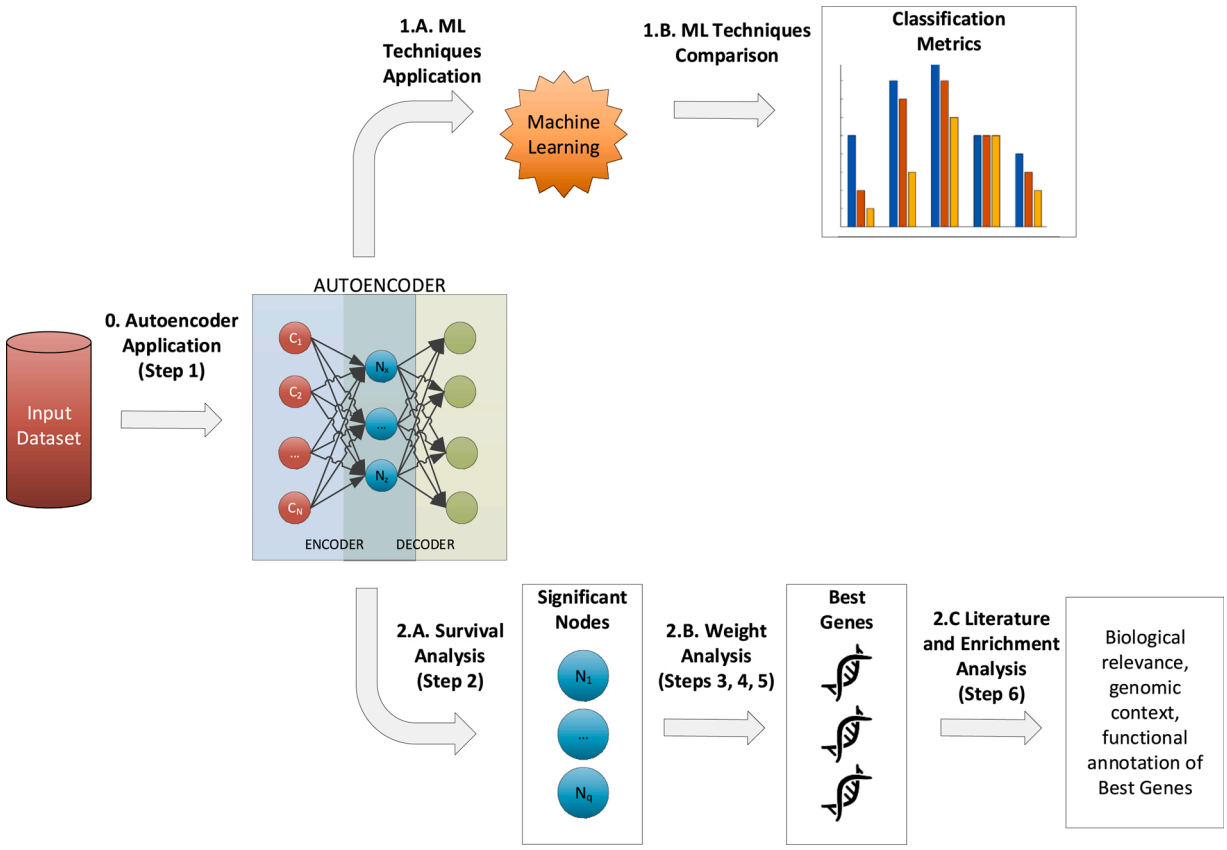

Fig. 1. Experimental framework applied in this work. AEs are applied to the given input dataset (**arrow 0**). Several machine learning algorithms are executed on cross-validated AE outputs (**arrow 1.A**) and their results are compared (**arrow 1.B**). The AE output obtained from the complete dataset is studied through a statistical survival analysis to select the significant hidden nodes of the AE (**arrow 2.A**). The weight analysis uses the weights of CpG sites in significant nodes to calculate the relevance of the genes associated and select the most significant ones (**arrow 2.B**). A literature and enrichment analysis is performed to study the biological relevance, genomic context and functional annotation of the set of genes obtained (**arrow 2.C**). Steps 2-6 correspond to the ones performed in the gene-weight methodology process described in Section 2.3.3 and Process outline outline 1.

NuSVC), meta-learning classifiers (AdaBoostClassifier and GradientBoostingClassifier), Bayesian classifiers (GaussianNB) and the k-nearest neighbours algorithm (KNearestNeighbors) as lazy/instance-based classifier.

A stratified multi-repeated (concretely, ten times to reduce risk of bias by random seed selection) five-fold cross-validation procedure based on accuracy as goodness measure was also applied to evaluate the results of every classifier and to prevent overfitting.

Several algorithms require FS to obtain better results. Such algorithms, according to the Hughes effect, could degrade their performance if the number of features increases in the context of a limited number of instances. Moreover, FS increases parsimony in the models and thus reduces the risk of overfitting or generating artefacts that cannot be applied in real environments [43].

For the comparison of machine learning techniques, a FS was firstly conducted in order to reduce dimensionality to the most-significant CpG sites. For this purpose, we used the widespread False Positive Rate (FPR) test, which selected CpG sites that passed a significance test according to ANOVA F-value between label/feature with a confidence level of 95%. FPR correction applied to each pairwise comparison is a common technique to reduce feature space in medicine [44]. This helped us speed up the development of the compared machine learning models. At the same time, the number of inputs was also reduced, which significantly decreases the computational burden to train the AEs. SelectFpr from scikit-learn with default parameters was responsible for this first level of FS.

The compared classification algorithms were also executed with and without a second level of FS. Namely, after applying the SelectKBest algorithm from the scikit-learn library. This is one of the easiest univariate feature selectors to use in scikit-learn, which provides a fast feature ranking and selection according to the best scores regarding a specific metric (in our case, ANOVA F-value). For the experiments, we fixed the number of features to select as many as the number of training instances. This aims to reduce the risk of overfitting whilst maintaining a sufficient number of representative features from the complete feature space.

With all the previous in mind, cross-validation results from the different algorithms were taken from the original data after a FPR-based FS (Original), Original plus autoencoded features (Original + AE), Original followed by a Best-K FS (Original + FS) and finally, Original + AE followed by Best-K FS (Original + AE + FS). Then, to study the global differences between the use of AEs or not, we aggregated the statistics for every classifier and test in the cross-validation in two categories: the best result for a classifier when AEs (maximum between Original + AE and Original + AE + FS) were applied and when not (maximum between Original and Original + FS). Finally, we studied the differences between both distributions and statistically validated them by the use of a Wilcoxon signed-rank test [45].

2.3.3. *Gene-weight methodology*

This section describes the gene weight analysis performed to discover how AEs consider CpG sites according to whether they belong to specific genes. Thus, it is possible to calculate the weights for a gene in a set of autoencoded features and study the gene relevance in a later biological analysis.

Process outline 1. Steps performed in Gene-weight methodology

```
1:  Input REP: number of AE executions, n: number of CpG sites, p: number of
    patients, dataset: values of n CpG sites for p patients, G: number of genes to
    select, CpGSites: names of CpG sites included in dataset.
2:  Output S: selection of commons genes in REP.
3:  ____________________
4:  genes ⟵ geneSymbols(CpGsites)
5:  S ⟵ genes
6:  for t⟵1 to REP do
7:    //Step 1
8:    hiddenNodes ⟵ AEApplication(dataset)
9:    //Step 2
10:   significantNodes ⟵ survivalAnalysis(hiddenNodes)
11:   q ⟵ |significantNodes|
12:   //Step 3
13:   for l⟵1 to q do
14:     Z_l ⟵ selectWeigthsInAE(CpGsites, significantNodes_l)
15:   end for
16:   for g ∈ genes do
17:     CpGSites_g ⟵ CpGSitesByGene(g)
18:     R_l ⟵ ∅
19:     for l ⟵ 1 to q do
20:       W_g^l ⟵ sumWeights(CpGSites_g,Z_l)
21:       // Step 4
22:       R_g^l ⟵ W_g^l / |CpGSites_g|
23:       R_l ⟵ R_l + {R_g^l}
24:     end for
25:   end for
26:   for l ⟵ 1 to q do
27:     s_l ⟵ bestGenesSelection(R_l, G)
28:   end for
29:   // Step 5
30:   S_t ⟵ s_1
31:   for l ⟵ 2 to q do
32:     S_t ⟵ S_t ∩ s_l
33:   end for
34:   S = S ∩ S_t
35: end for
36: return S
```

The proposed analysis is carried out by applying a methodology composed of six steps depicted as an activity diagram in Fig. 2. As can be observed, the actions of the diagram are processed sequentially and each action corresponds to a step of the proposed methodology.

Table 3 defines the main symbols related with the variables involved in the equations of the different steps of the proposed methodology. Furthermore, the index used for each one is presented. Process outline outline 1 summarises the steps performed in the gene-weight methodology which are described below.

(1) **Step 1–AE application**: As mentioned in Section 2.2, the processed data contained a total of 383,919 features (one for each CpG site). This amount of features constitutes a serious issue since it cannot be processed by using traditional techniques. As can be seen in Fig. 3, in order to reduce the dimensionality of the data, the step 1 of the proposed methodology is the application of the AE (Line 7: Process outline outline 1).

Let $x \in R^n$ be the input composed of the CpG sites values for p patients. This forms a matrix with size $p \times n$ of values, each of them hereinafter referred to as x_i^k, where $k = 1...p$ and $i = 1...n$. The AE is applied with a single hidden layer to the data of the aforementioned matrix. Given d as a distance function in R^n, an AE built for m nodes is given by functions f and g with the following properties:

$$\begin{cases} f : R^n \to R^m, & \text{where } z_j = f_j(x) = \text{RELU}\left(\sum_{i=1}^{n} w_{ij} x_i\right) \\ g : R^m \to R^n, & \text{where } y = g(z) \end{cases} \tag{1}$$

f and g will be obtained through an optimisation process that minimises $d(x, y)$. Thus, each node j of the intermediate layer is characterised by a value z_j.

In this case, for each patient $k = 1...p$, the following output is formed:

$$z_j^k = f_j(x^k) = \text{RELU}\left(\sum_{i=1}^{n} w_{ij} x_i^k\right) \tag{2}$$

It should be taken into account that AEs randomly initialise the weights w_{ij} of the hidden units, which are updated iteratively during the training process through the backpropagation mechanism [13]. Therefore, the AE needs to be executed REP times in order to prevent the results from becoming biased (Line 6: Process outline outline 1). The number of executions of the AE is fixed (REP = 10) as a trade-off between a sufficiently large number of AEs to reduce bias and a

Table 3
Definitions of the main symbols related with variables and their indexes.

Symbol	Definition	Index
n	Number of CpG sites as input data	i
p	Number of patients	k
M	Number of hidden nodes or units	j
q	Number of significant hidden output nodes in survival analysis	l
x_i^k	Input value of CpG site i for patient k	
x_i	Vector composed of the x_i^k values, where $k = 1...p$	
w_{ij}	Weight associated to the CpG site i for the hidden node j	
z_j^k	Hidden AE output value of node j for the values of patient k	
W_l^g	Sum of the weights of the CpG sites associated with a gene g for the hidden node l	
C_g	Number of CpG sites associated to the gene g	
R_g^l	Importance of a gene g in the node l (ratio between W_g^l and C_g)	
G	Number of genes with the highest R_g^l	
q_t	Number of significant nodes in the tth AE execution	
s_l	Set of the G genes selected with greater score R_g^l	
S	Intersection between the sets of genes associated to the significant nodes in all the AEs executions	
AE_t	AE execution number t	
REP	Number of total executions of AEs	

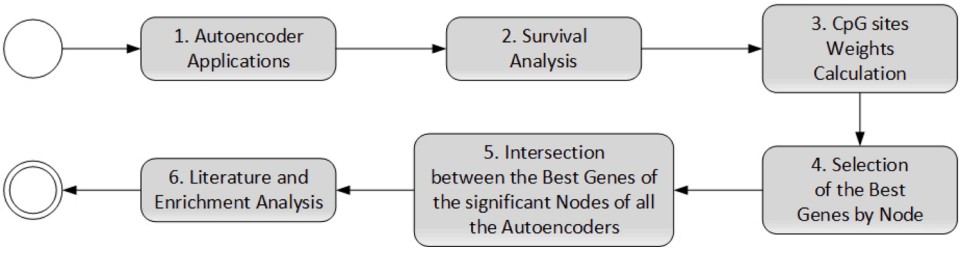

Fig. 2. Activity diagram that shows the complete process of the proposed methodology to analyse gene-weight significance in AEs.

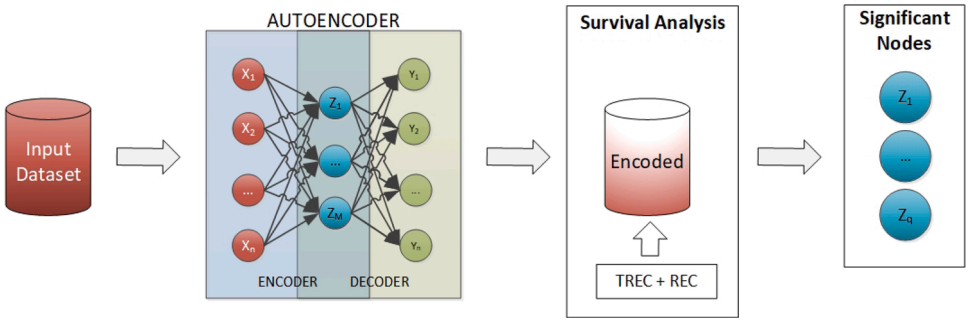

Fig. 3. Steps 1 and 2 – Application of AE to the input dataset followed by survival analysis to discover the significant nodes.

affordable training time (which is of several minutes per AE). We develop REP AEs to later analyse their nodes (hidden units), as can be seen in the next steps, reducing bias due to random selection of initial weights in the AEs. Thus, the nodes can be combined to analyse the most influencing genes. To this end, we just add the weights associated to every CpG site involved in an AE feature, extract the gene each CpG site belongs to and finally obtain the mean weight of each gene in the features of the REP AEs.

(2) **Step 2–Survival analysis**: This step is devoted to supplying survival analysis to the autoencoded dataset obtained in the AE executions from the step 1. The survival analysis discovers which nodes of the AE as independent variables are significant (Line 8: Process outline outline 1) by taking into account the REC variable (recurrence class) and the TREC variable (time from diagnosis to a new tumour appearance) as dependent variables. Specifically, for each AE, we calculate every hidden node output variable. Thus, the following vector $Z_j = [z_j^1, \ldots, z_j^p]$ is obtained applying Eq. (2) where $j = 1 \ldots M$ and p is the number of patients.

We apply a 5-year survival analysis to each Z_j and assume that q variables were significant (Line 9: Process outline outline 1) in the survival analysis of the M variables. To simplify the notation in the next steps and taking into account that the nodes (and the corresponding weights) of the AE can be reordered without change in the transformation, we suppose that the significant variables are the first q in Z_l where $l = 1 \ldots q$.

Survival analysis was carried out by a Cox Regression [46] implemented in JavaStat software. In order to translate continuous biomarkers into clinical decisions, it is often necessary to binarise parameters [14] and so we did. There is no standard method to find an optimal binarization, therefore, we take advantage of the ReLU output ($>= 0$) and the autoencoded datasets are binarised as follows: a parameter value is 0 if the node/autoencoded feature provides the value 0 (non-activated according to the ReLU activation function); otherwise, it was 1. Finally, a node was considered significant if the Cox Regression reports a p-value less than or equal to an $\alpha = 0.05$.

Fig. 3 also summarises the step 2. The methodology starts by applying an AE to the input dataset. Once the encoded data is generated, survival analysis is applied, and the significant nodes ($Z_1, \ldots, Z_q$) are obtained by considering the aforementioned analysis. Note that this step is repeated REP times, once for each AE execution, and hence REP sets of significant nodes are obtained at the end of this step.

(3) **Step 3–CpG site weight calculation**: The weights of the CpG sites of every significant node of the AEs are selected in this step. That is, only the weights of those CpG sites which are included in the significant nodes by the survival analysis were taken into account (Line 11: Process outline outline 1). Fig. 4 shows how the weights of the significant nodes are calculated from the AE solution. As can be seen, each significant node has a list of weights associated where a weight corresponds to a CpG site of the input dataset.

For each of the q nodes and n CpG sites, we obtain the weights w_{il}, where $i = 1 \ldots n$ and $l = 1 \ldots q$, which measure the relationship of the CpG sites with each significant node in the survival analysis. Next, we sum the weights in absolute value of the CpG sites that correspond to the same gene, and obtain a value for each gene (Lines 13–17: Process outline outline 1). It should be noted that each value of the CpG sites is represented by the variable x_i.

In this step, the gene symbols associated with each CpG site involved in the significant nodes are extracted by taking into account the mapping file that relates them (described in Section 2.2). For each node $l = 1 \ldots q$ such that x_i is in gene g, we calculate the following:

$$W_g^l = \sum_{\{i | x_i \in g\}} |w_{il}| \tag{3}$$

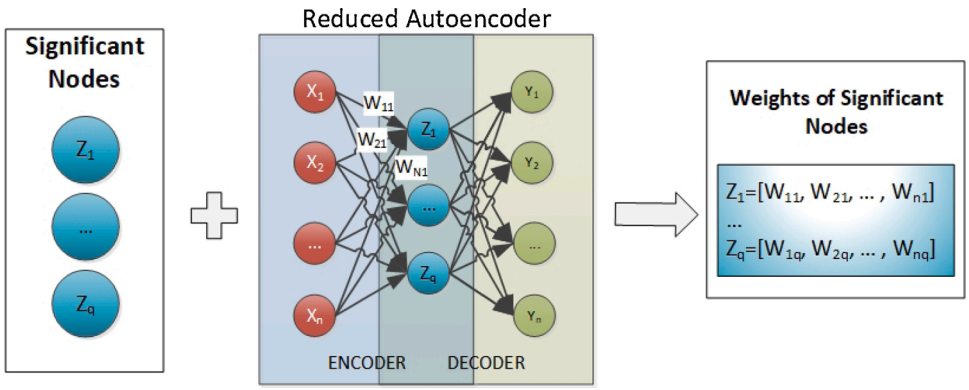

Fig. 4. Step 3 – Calculation of the weights of the significant nodes (hidden units) from an AE.

(4) **Step 4–Selection of the best genes by node**: In this step, the ratio of weights for every gene, defined in Eq. (4), is calculated for every significant node (Lines 18–19: Process outline outline 1). Each significant node has a list of weights associated to every gene obtained in the step 3. From this point onwards, only the best G genes (with the highest ratio of weights) of each significant node are taken into account.

For each gene g the number of CpG sites x_i belonging to g is calculated as $C_g = \#\{x_i | x_i \in g\}$ and the ratio of each gene is defined by Eq. (4).

$$R_g^l = \frac{W_g^l}{C_g} \tag{4}$$

R_g^l gives a measure of the importance of a gene g in the node l of the AE. R_g^j is only influenced by the weight of the gene and not by the total number of apparitions of CpG sites belonging to g.

For each significant node l in an AE we define the set s_l as the G genes with the highest R_g^l (Lines 22–24: Process outline outline 1). Note that G is a parameter of the experimental framework which could vary by increasing or decreasing the number of genes considered relevant.

(5) **Step 5–Intersection between the best genes of all the significant nodes**: The subsequent step is devoted to calculating the intersection S between the sets of genes associated to the significant nodes in all the AEs (Lines 25–29: Process outline outline 1). Thus, let t=1…REPbe thedifferent executionsof anAE andj=1…q_t be the significant nodes of the tth AE, we calculate S as follows:

$$S = \bigcap_{t=1}^{\mathrm{REP}} \bigcap_{l=1}^{q_t} s_l \tag{5}$$

(6) **Step 6–Literature and Enrichment analysis**: Finally, the last step of the proposed methodology involves performing a systematic review of the literature using the well-known PubMed [47] system included within the resources provided by The National Center for Biotechnology Information (NCBI). This step analyses the similarity between the significant genes in the autoencoded features and the relevant genes in the literature regarding breast cancer recurrence. Furthermore, this step also aims at reporting an enrichment analysis of the significant genes found according to genomic context based on information retrieved from NCBI, functional annotation in the context of the Gene Ontology (GO), and pairwise connections linking genes, among others. To this end, several web-interfaces, such as String database [48], GeneMANIA [49] or Database for Annotation, Visualization, and Integrated Discovery (DAVID) tools [50] were queried.

As stated above, when increasing the value of G, the number of selected genes is also increased. In our experimentation, we repeat the steps 4–6 with the following values $G = 1000, 2000, 3000, 4000, 5000$. Each different value give rise to a set of different genes.

3. Results

3.1. Comparison of machine learning classification techniques and study of the importance of feature selection

In this section, the results of the classification algorithms are shown as described in Section 2.3.

Regarding the autoencoded data, loss function to calculate M in our experimental comparison is represented in Fig. 5. In general, there was a noteworthy change of trend when the number of hidden units was 100 and therefore M was fixed.

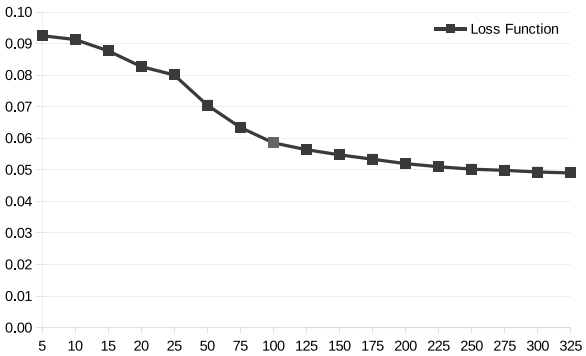

Fig. 5. Mean loss function on the training sets (in the procedure of multi-repeated five-fold cross-validation) according to the number of hidden units in the hidden layer of the AE. In red, the selected value in the experimental comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4 presents the results obtained in the eight classification algorithms described in Section 2.3.2 by using the default parameters provided by the scikit-learn library. This table shows the accuracy of each classifier. The results are in four columns (Original, Original + FS, Original + AE and Original + AE + FS). The results also appear aggregated in two calculated categories: the best result for the classifier when an autoencoder was not applied (Non-AE column) and the best result for the classifier when autoencoded features were included in the classification (AE column).

Fig. 6 provides the distribution of results for the categories Non-AE and AE throughout the 400 training-test experiments (8 classifiers × 5 folds × 10 repetitions). We used those results to statistically validate the reported differences. As we said, we applied a Wilcoxon signed-rank test on the 400 results. According to the software StatService [51], the statistic for Wilconxon was 56605.5 with 1 and 398 degrees of freedom, and with a p-value less than 0.0001. With this p-value, the null hypothesis (i.e., no significant differences between both distributions exist) can be rejected. AE distribution reported a mean accuracy of 0.68 with a standard deviation of 0.09, whilst Non-AE only reached a mean accuracy of 0.63 with a higher standard deviation of 0.11. With these results, it could be possible to conclude that AEs statistically improved the results of the classifiers under study.

3.2. Gene-weight analysis

This section provides the outcomes of the gene-weight analysis after the methodology proposed in Section 2.3.3 has been applied to the results obtained in the REP executions of the AE performed.

Table 5 shows the number of significant nodes obtained by each AE execution which is identified in row ID by AE_t, $\forall\, t \in [1, \mathrm{REP}]$. It should be noted that only six AEs obtained significant nodes when survival analysis was applied with $\alpha = 0.05$.

As mentioned in Section 2.3.3, a set of sets of the best genes (genes with the highest ratio of weights) that are common among all the significant nodes of all the AE executions is obtained after applying the gene-weight methodology to the input dataset.

Table 6 shows the set of genes selected according to the different values of G. It should be taken into account that the genes selected for $G = 1000$ are a subset of those obtained for $G = 2000$, and so on. That is, the most restrictive set of genes is given for $G = 1000$ and is present in the sets of the rest of the G values. The set of genes for $G = 5000$ is the largest since it includes all those obtained for the rest of the G values. Thus, ranking 1 denotes the genes that are in the final set for all the G values tested, and are therefore the most relevant among the significant genes. Ranking 2 includes the genes present for $G >= 2000$, and so on until ranking 5 in which there are only the genes selected for $G = 5000$. Additionally, several references from the literature of the last ten years

Table 4
Mean results of the accuracy obtained by the classification algorithms from the original dataset, original and feature selection, original plus autoencoded features and feature-selected original plus autoencoded features in ten repetitions of five-fold cross-validation results with different random seeds next to the best result when autoencoding was and was not used. In bold, the best values.

Classifier	Original	Original + FS	Original + AE	Original + AE + FS	Non-AE	AE
SVC	0.646	0.651	**0.675**	0.653	0.651	**0.675**
KNearestNeighbors	**0.671**	0.660	0.657	0.659	**0.671**	0.659
NuSVC	0.643	0.650	**0.669**	0.647	0.650	**0.669**
DecisionTreeClassifier	0.621	0.627	**0.633**	0.620	0.627	**0.633**
RandomForestClassifier	0.668	0.669	0.666	**0.687**	0.669	**0.687**
GradientBoosting	0.639	0.647	0.638	**0.651**	0.647	**0.651**
AdaBoost	**0.645**	0.639	0.641	0.632	**0.645**	0.641
GaussianNB	0.511	0.613	0.511	**0.614**	0.613	**0.614**

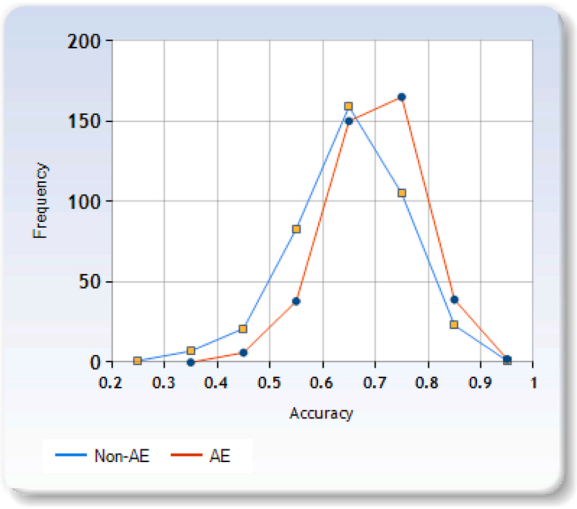

Fig. 6. Distribution of results obtained for Non-AE and AE results in the experimentation.

that relate each gene with breast cancer are also included.

Fig. 7 shows the distribution of the genes related to the CpG sites with the highest number of apparitions in the input dataset. This figure shows that most of the genes associated to CpG sites in the input dataset remain unknown. As can be observed, PTPRN2 is the most repeated gene, followed by PRDM16, MAD1L1, TNXB, and RPTOR. The other genes appear fewer times and the number of repetitions decreases gradually, however they still represent around the 40% of the total number of CpG sites.

Fig. 8 shows similar results to Table 6 via a heatmap. This figure shows the AE executions ID with significant nodes along the *x*-axis (AE_1, AE_2, AE_3, AE_8, AE_9, AE_{10}), and the gene symbols along the *y*-axis. The heatmap values are given by the mean of the weights of the best genes normalised by the AEs, and in this manner, the results of the proposed methodology are shown in a plot. Moreover, the heatmap includes the results of applying the average linkage clustering both to the AEs and to the genes, by using the Euclidean distance. The cluster which contains the subfamilies of the Protocadherin gamma and beta is highlighted for later discussion.

We also provide the results given by the gene-weight analysis methodology but without filtering any nodes of the AEs and thus assuming that all nodes are to be taken into account.

The genes with the highest ratio of weights of the AEs were PTPRN2 (the only gene that appeared in the set of the best 2000 genes), and MAD1L1, PRDM16, RPTOR, and TNXB (only in the set of 5000). Table 6 highlights (in bold) those genes that are in common to all the nodes of all AE executions.

Fig. 9 shows the normalised average ratio of weights of the genes (only PTPRN2, MAD1L1, PRDM16, RPTOR, and TNXB) and selects the significant nodes (the five upper nodes) denoted as Autoencoded-Filtered Selection (AFS), and the same genes while taking into account all the nodes from the AEs (the five bottom nodes) represented as Non-Autoencoded-Filtered Selection (Non-AFS).

3.3. Enrichment analysis

This section presents the results obtained through the enrichment analysis conducted to the set of genes found by the proposed methodology (Table 6) based on genomic context and functional annotation.

First, we collected data from NCBI regarding the chromosome, chromosomal region and position of each identified gene in our study as can be observed in Table 7. Then, we queried several web-interfaces related to gene lists analysis using available genomics and proteomics data as stated in the step 6 of Section 2.3.3. In particular, we applied the String database [48], that provides a method to compute the enrichment analysis for a variety of classification systems such as the Gene Ontology (GO).

The enrichment itself is computed using a Fisher's exact test followed by a correction for multiple testing. Table 8 presents the functional enrichment analysis of the GO Biological Process, Molecular Function and Cellular Component, respectively. Table 8 includes the GO terms found, their descriptions and the false discovery rate (FDR) obtained. We report the results that reached an FDR < 0.05 significance threshold when conducting GO term analysis.

Our set of genes was also queried into DAVID tool [50] in order to detect more functional annotations and we applied the GeneMANIA tool [49] to discover pairwise connections linking the genes. Although the latter was able to find other genes related to a set of input genes, we focused our analysis only in the gene-set obtained by our methodology. Fig. 10 displays the gene interaction network reported according to protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity. It can be noted that most of the discovered interactions are based on co-localisation similarities.

4. Discussion

As can be observed in Table 4, AEs could generally improve the results of the set of classifiers under study. Only two classifiers obtained better results for Non-AE data. This could indicate that they took

Table 5
Number of significant nodes obtained by each AE execution.

ID	AE_1	AE_2	AE_3	AE_4	AE_5	AE_6	AE_7	AE_8	AE_9	AE_{10}
# Nodes	2	3	1	0	0	0	0	3	1	2

Table 6
The best genes obtained in the gene-weight methodology proposed. The ranking for the genes in the significant nodes in the AEs is shown. In addition, references in the literature in the last ten years which directly relates them to breast cancer are presented. Those common genes to all the nodes (i.e., without filtering non-significant nodes) are highlighted in bold.

Rank	Gene Symbol	Ref.
1	**PTPRN2**	[52]
2	**PRDM16**	[54]
3	ATP11A	–
	MAD1L1	[57]
	TNXB	[58]
	TSNARE1	–
4	EIF2B5	[61]
	GABBR1	[63]
	HDAC4	[65]
	MCF2L	[66]
	PCDHGA[1-8]	[68]
	PCDHGB[1-4]	[68]
	PSMB9	[71]
	SHANK2	–
	SNHG14	[72]
	TBCD	[73]
Rank	**Gene Symbol**	**Ref.**
5	AGO2	[53]
	C7orf50	[55]
	CAMTA1	[56]
	CASZ1	–
	CBFA2T3	[59]
	FIP1L1	[60]
	PARD3	[62]
	PRKCZ	[64]
	RASA3	–
	RPTOR	[67]
	SDK1	[69]
	SMOC2	[70]
	SORCS2	–
	TBC1D16	–

advantage of the original CpG sites (firstly selected based on FPR tests) and the later inclusion of more features (autoencoded ones) degraded their performance. Since those classifiers were kNN (which is sensible to a higher number of features due to the curse of dimensionality) and AdaBoost (which could have more difficulties to find better subsets of features with a higher number of features) the former conclusion could be reasonable. On the other hand, the global distribution in our results seen in Fig. 6 proved an improvement when AE features were included in the modelling. This statement was confirmed by a Wilcoxon signed-rank test, suggesting that AEs should be taken into account in future studies for DNA Methylation analysis with classification purposes (specially regarding breast cancer recurrence) by machine learning modelling.

AEs feature generation also seems to be relatively independent on random initialisation as can be seen in Fig. 8. Note that regardless of the AE analysed, there exists a pattern (highlighted in yellow) regarding all the genes from the subfamilies of the Protocadherin gamma (PCDHGA [1-8]) and Protocadherin beta (PCDHGB[1-4]). This pattern could demonstrate that AEs do not provide artifact behaviour but do keep relationships among genes relatively stable. However, the initial random selection of weights still exerts an impact on the results as can be seen in the remaining genes, and should be taken into account in future research.

Differences in the results were detected when feature selection was carried out in certain classification methods whilst most methods presented no significant change. This finding may be due to the fact that the latter classification algorithms do include an internal feature selection/weighting process. In our experimentation, default parameters were used, and thus the results should be taken with caution since most of the classifiers could reach higher levels of accuracy with a proper parameterisation.

Fig. 9 outlines feature selection based on survival analysis could provide an important step to take full advantage of AEs since feature-selected genes provide a higher entropy which could potentially lead to a better performance of the classifiers (since information in the features is higher). On the other hand, non-selected features provide genes with homogeneous weights which could lead to a worse separability and therefore errors in classification. Again, more research is needed to confirm this point.

4.1. *Literature analysis*

We carried out an analysis of the most heavily weighted genes (according to the weights associated to their CpG sites) formed part of the autoencoded features generated by the AEs used in this work (see Table 6). A literature search provided previous knowledge regarding its correlation with prognosis in breast cancer. The genes analysed seem to fall into five categories:

- **Confirmed recurrence biomarker**. PTPRN2 is proposed as an upregulator in highly metastatic cancer cells [52] and their increased expression is associated with human metastatic relapse. TNXB is involved in signalling pathways relating to TP53 (a well-known biomarker in breast and other cancers) which leads to preventing apoptosis and metastasis [58]. AGO2 mRNA expression is correlated

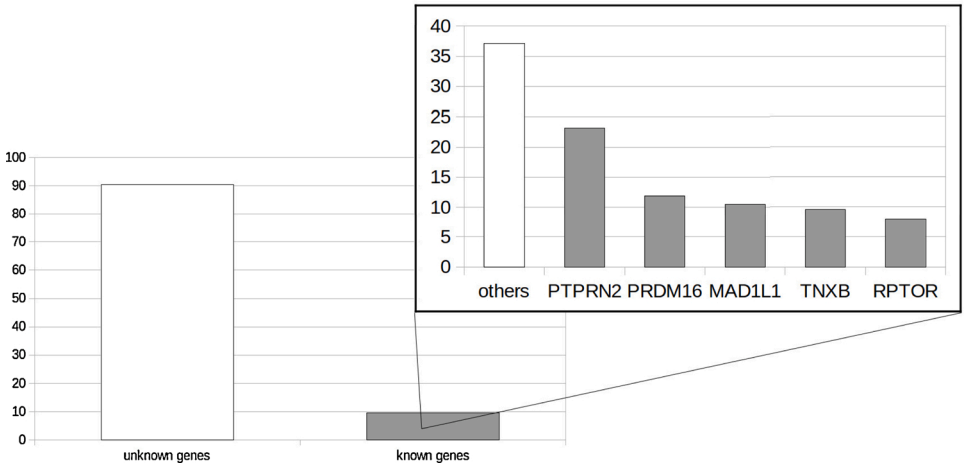

Fig. 7. Percentage of known vs. unknown genes (only the most important genes according to Table 6) related to CpG sites selected by the significant autoencoded features.

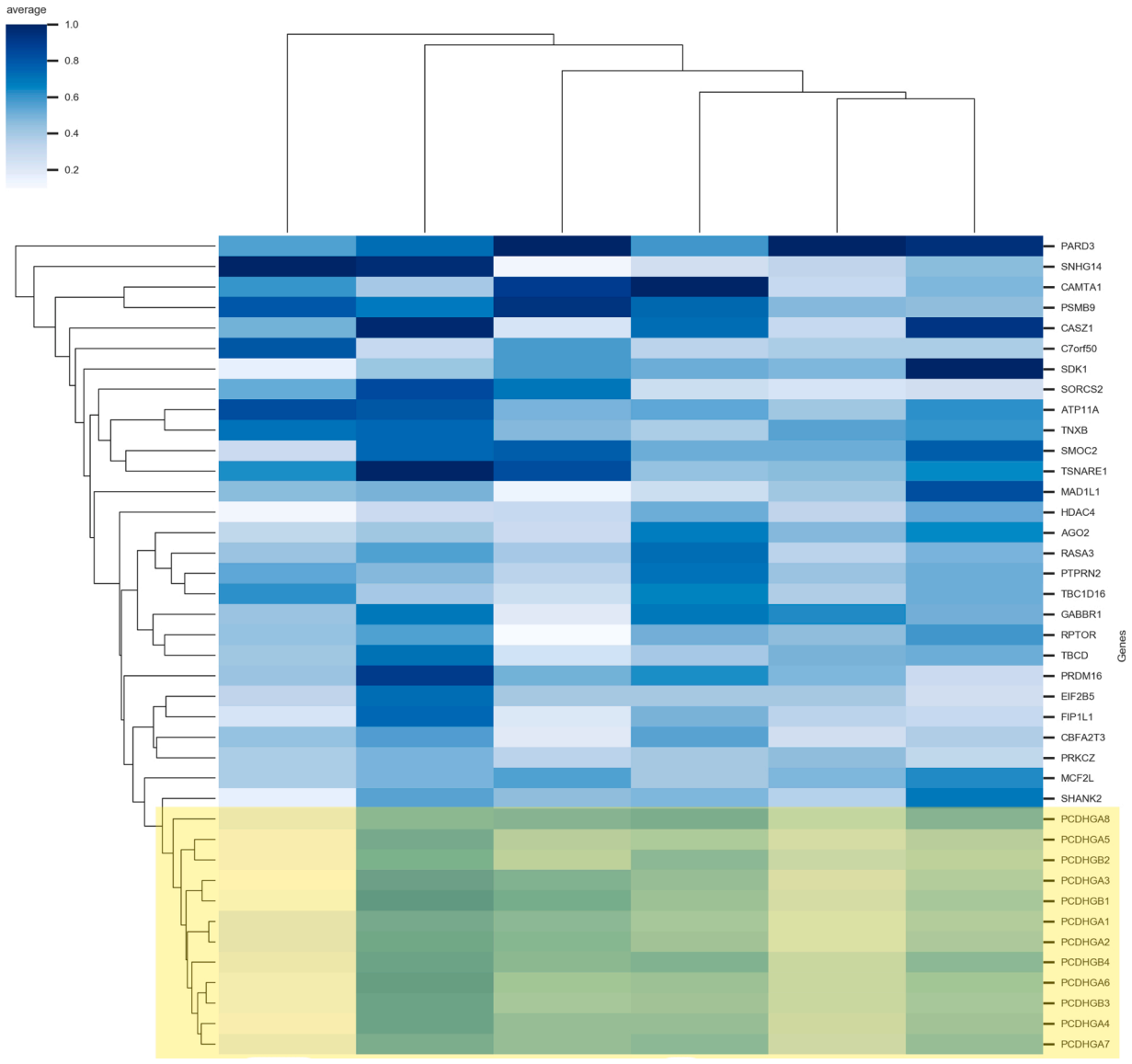

Fig. 8. Heatmap and hierarchical clustering applied both to the best genes normalised by AE (represented in the *y*-axis) and to the AEs (*x*-axis). The cluster which contains the subfamilies of the Protocadherin gamma (PCDHGA[1-8]) and Protocadherin beta (PCDHGB[1-4]) is highlighted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

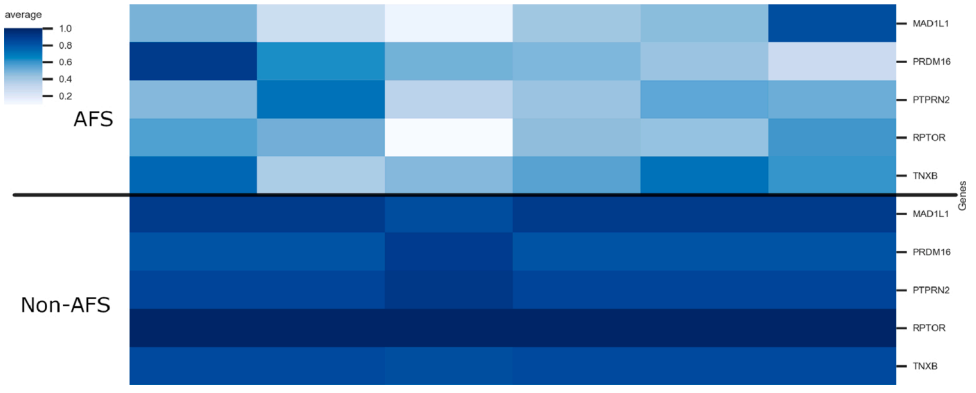

Fig. 9. Heatmap of the best genes of the AEs without filtering the significant nodes, and the same genes in the results of the significant nodes. The AEs are represented along the *x*-axis, and the genes are shown along the *y*-axis.

Table 7
Genomic positions of the identified genes in our study.

Gene	Chromosome	Chromosomal Region	Position
PRDM16	1	1p36.32	NC_000001.11 (3069203..3438621)
CAMTA1	1	1p36.31-p36.23	NC_000001.11 (6785324..7769706)
CASZ1	1	1p36.22	NC_000001.11 (10636604..10796650, complement)
PRKCZ	1	1p36.33	NC_000001.11 (2050411..2185399)
HDAC4	2	2q37.3	NC_000002.12 (239048168..239401649, complement)
EIF2B5	3	3q27.1	NC_000003.12 (184135023..184145311)
FIP1L1	4	4q12	NC_000004.12 (53377641..53462583)
SORCS2	4	4q12	NC_000004.12 (7192538..7742828)
PCDHGA1	5	5q31.3	NC_000005.10 (141330685..141512979)
PCDHGA2	5	5q31.3	NC_000005.10 (141338760..141512975)
PCDHGA3	5	5q31.3	NC_000005.10 (141343829..141512975)
PCDHGA4	5	5q31.3	NC_000005.10 (141355021..141512975)
PCDHGA5	5	5q31.3	NC_000005.10 (141364331..141512979)
PCDHGA6	5	5q31.3	NC_000005.10 (141373891..141512975)
PCDHGA7	5	5q31.3	NC_000005.10 (141382742..141512975)
PCDHGA8	5	5q31.3	NC_000005.10 (141391916..141512979)
PCDHGB1	5	5q31.3	NC_000005.10 (141350261..141512979)
PCDHGB2	5	5q31.3	NC_000005.10 (141360136..141512979)
PCDHGB3	5	5q31.3	NC_000005.10 (141370242..141512975)
PCDHGB4	5	5q31.3	NC_000005.10 (141387698..141512975)
TNXB	6	6p21.33-p21.32	NC_000006.12 (32041153..32109338, complement
GABBR1	6	6p22.1	NC_000006.12 (29602228..29633183, complement)
PSMB9	6	6p21.32	NC_000006.12 (32854192..32859851)
SMOC2	6	6q27	NC_000006.12 (168441153..168667992)
PTPRN2	7	7q36.3	NC_000007.14 (157539052..158587823)
MAD1L1	7	7p22.3	NC_000007.14 (1815795..2232945, complement)
C7ORF50	7	7p22.3	NC_000007.14 (977964..1138325, complement)
SDK1	7	7p22.2	NC_000007.14 (3301252..4269000)
TSNARE1	8	8q24.3	NC_000008.11 (142212080..142403291, complement)
AGO2	8	8q24.3	NC_000008.11 (140520156..140642406, complement)
PARD3	10	10p11.22-p11.21	NC_000010.11 (34109560..34815325, complement)
SHANK2	11	11q13.3-q13.4	NC_000011.10 (70467854..71252724, complement)
ATP11A	13	13q34	NC_000013.11 (112690034..112887168)
MCF2L	13	13q34	NC_000013.11 (112894378..113099742)
RASA3	13	13q34	NC_000013.11 (113977783..114132623, complement)
SNHG14	15	15q11.2	NC_000015.10 (24823608..25419462)
CBFA2T3	16	16q24.3	NC_000016.10 (88874858..88977207, complement)
RPTOR	17	17q25.3	NC_000017.11 (80544838..80966368)
TBCD	17	17q25.3	NC_000017.11 (82752048..82945914)
TBC1D16	17	17q25.3	NC_000017.11 (79932343..80035875, complement)

Table 8
Biological process, molecular function and cellular component go-terms significantly enriched.

Term ID	Term description	FDR
Biological process GO-terms		
GO:0007156	Homophilic cell adhesion via plasma membrane adhesion molecules	8.81e−15
GO:0007399	Nervous system development	3.35e−08
GO:0007155	Cell adhesion	1.13e−07
GO:0007267	Cell–cell signaling	1.73e−06
GO:0048731	System development	5.11e−05
GO:0007275	Multicellular organism development	0.00011
GO:0023052	Signaling	0.0016
GO:0007154	Cell communication	0.0021
GO:0032501	Multicellular organismal process	0.0071
GO:0030812	Negative regulation of nucleotide catabolic process	0.0286
GO:0045820	Negative regulation of glycolytic process	0.0286
GO:0051198	Negative regulation of coenzyme metabolic process	0.0286
GO:2001170	Negative regulation of ATP biosynthetic process	0.0355
Molecular function GO-terms		
GO:0005509	Calcium ion binding	1.44e−07
GO:0046872	Metal ion binding	0.00076
GO:0043167	Ion binding	0.0050
Cellular component GO-terms		
GO:0031226	Intrinsic component of plasma membrane	3.54e−05
GO:0044459	Plasma membrane part	3.54e−05
GO:0005887	Integral component of plasma membrane	7.91e−05
GO:0071944	Cell periphery	0.0015
GO:0005886	Plasma membrane	0.0034
GO:0033267	Axon part	0.0232
GO:0030054	Cell junction	0.0249
GO:0044425	Membrane part	0.0267
GO:0005923	Bicellular tight junction	0.0419
GO:0043005	Neuron projection	0.0419
GO:0150034	Distal axon	0.0419

with reduced relapse-free survival in human breast cancer [53], and C7orf50 as a heritable DNA methylation mark (associated with breast cancer in multiple-case families) was proved to be associated with a higher risk in the general population (Melbourne Collaborative Cohort Study) [55]. Recent work [56] also demonstrated that CAMTA1 expression promoted cell viability and migration/invasion and was therefore a potential innovative therapeutic target for treatment. Rossett et al. [59] provided the first demonstration that loss of CBFA2T3 function in the nucleolus of breast epithelial cells could induce morphological and molecular changes typical of cancer initiation. Upregulation of EIF2B5 was also associated with breast

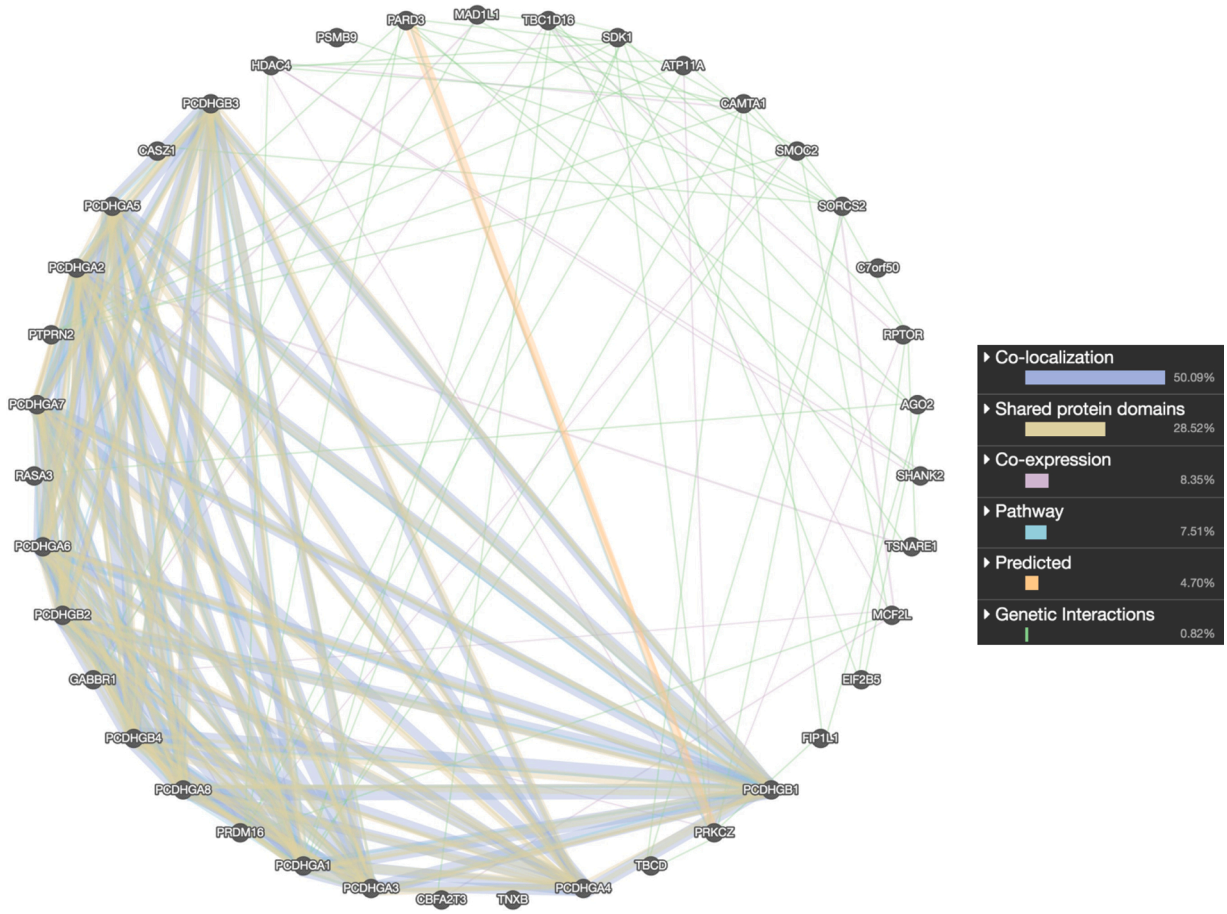

Fig. 10. Genes interaction network (GeneMANIA).

cancer [61] and PARD3 was reported as a regulator of signalling pathways relevant to invasive breast cancer [62]. FIP1L1 has been reported as a candidate synthetic lethal interaction associated with RB1 defects which lead to the origin of triple-negative breast cancer [60]. PRKCZ mediates epidermal growth factor (EGF)-stimulated chemotactic signalling pathways which is a key factor in the process of tumour development and metastasis [64]. MCF2L (also known as DBS) regulates cell motility in tumour-derived, human breast epithelial cells [66]. The protocadherin family (PCDHGA and PCDHGB) frequently acts as tumour suppressor genes [74,75] and their inactivation through promoter methylation is closely correlated with tumour development. Downregulation of SMOC2 was also associated with advanced tumour stage in breast cancer [70], and the expression of immunoproteasome genes such as PSMB9 was associated with longer survival [71]. Finally, TBCD has recently been reported as an SLC27A4-correlated gene associated with poor prognosis in breast cancer [73].

- **Possible recurrence factor**. Although no breast cancer reference has been found regarding CASZ1, it proved to be bound to the nucleosome remodelling and histone deacetylase complex [76] which was linked directly to breast cancer oncogenesis and specifically with TWIST [77] (a very well-known breast cancer biomarker [14]). Along the same lines, RASA3 has recently been reported as a critical regulator of Rap1 in endothelial cells [78]. Active Rap1 (as well as the endothelium in general) plays an important role in breast cancer tumour invasion and metastasis [79]. The AEs in this work could have used RASA3 due to its relation with Rap1. Additionally, SDK1 has been identified as a breast cancer heritability gene [69]. Although most cancer heritability genes are mostly tumour suppressors and control cell proliferation, invasion, and metastasis, the SDK1 function remains unknown regarding breast cancer. An association of serum IGF-I and IGFBP-3 concentrations with breast cancer risk, particularly for women with a later diagnosis of cancer, was established in the 1980s [80]. Since SORCS2 is highly associated to both molecules [81], its AEs selection appears fully justified. A hypomethylation event that reactivates TBC1D16 was reported to be a characteristic feature of the metastatic cascade [82]. Finally, SHANK2 interacts with the actin-binding protein cortactin which, when it is overexpressed, has been implicated in the progression of tumours [83].
- **Obesity-related biomarkers**. Obesity has been established as a risk factor for cancer incidence and cancer-related mortality [84]. However, the absence of brown adipose tissue has recently been found to be a possible risk factor for breast cancer recurrence [85]. Moreover, a recent therapy proposal suggests the use of a combination of drugs to force post-mitotic adipogenesis [86]. PRDM16 is a well-known master regulator of brown adipocyte differentiation [54], and therefore its presence could provide an important tool for prognosis. Moreover, GABBR1 is present in the top 32 most highly significantly differentially expressed genes that associate obesity with triple-negative breast cancer in premenopausal women [63]. Furthermore, the fact that higher levels of PTPRN2 have been found in children with obesity [87] could lead to a future line of research into the relationship between early-age obesity and cancer recurrence in future patients.
- **Chemotherapeutic drug inhibitors**. Over the last several decades, farnesyltransferase inhibitors (FTIs) have been used as anti-cancer agents [88]. Unfortunately, an increased expression of ATP11a [89] indicates the resistance to FTIs in Bcr/Abl-positive lymphoblastic leukemia. Our results suggest similar results regarding breast

cancer patients although no reference in the literature has been found to directly confirm this finding. On the other hand, MAD1L1 is associated with poor prognosis and insensitivity to taxol treatment in breast cancer [57]. Another important gene that seems to fit in this category is HDAC4. Downregulation of HDAC4 has proved to lead to the acquisition of tamoxifen resistance (for patients with estrogen-receptor-positive tumours, treatment with tamoxifen is the gold standard) [65]. Sic You et al. [67] suggest that RPTOR mediates, at least partially, the resistance to epidermal growth factor receptor inhibition (a common target for chemotherapeutic treatments) in triple-negative breast cancer cells. Along the same lines, SNHG14 was found to be upregulated in resistant cells against trastuzumab (an inhibitor used in the treatment of advanced HER2-positive breast cancer) when compared with breast cancer cells before treatment [72].

- **Probable pesticide exposure indicator**. Diazinon remains one of the most widely used insecticides in the U.S. for household as well as for agricultural pest control. It is also suspected to be a carcinogen [90] and has been related to breast and ovarian cancer [91,92]. Taking into account that TCGA data was obtained in the U.S., it is no wonder that TSNARE1 was selected since it was highlighted as a gene with unknown functions related with Diazinon exposure [93].

4.2. Enrichment analysis

This section provides the main findings discovered after performing the enrichment analysis based on genomic context and functional annotation previously described.

From the list of identified genes regulated by CpG methylation, whose genomic positions and chromosomal features were shown in Table 7, we found four genes located in chromosomes 1, 6 and 7, and three in chromosomes 13 and 17 (Fig. 11). In chromosomes 1, 7, 13 and 17, the genes are located in distal or telomeric regions, whilst in chromosome 6, three of the genes are located at the central region of p-arm and the other at the distal region of q-arm. This distribution is interesting because it has been described that deletions in chromosomal 1 p-arm are usual in cancer [94–96]. Additionally, other genes located on the chromosomal region 1p36 have been previously found to be methylated in breast cancer, suggesting the bona fide of our found genes [97]. Additionally, the 5q31.3 region, that includes the protocadherin gene cluster PCDHA, B and G, the first two also found in our study, is usually silenced in cancer [98]. Regarding the chromosomal region 6p21, its hypomethylation has been described as better prognostic in epithelial ovarian cancer [99], suggesting that methylation of genes in this region could favour the appearance of metastases. Also, methylation of genes from chromosomal region 17q25 has been connected with a higher risk of cancer development [100,101]. In chromosomal region 13q34, other genes have been also found methylated, like SOX1 [102].

Additionally, as can be observed in Table 8, the found GO-terms significantly enriched were: 13 for biological process, 3 for molecular function, 11 for cellular component. Due to the regulation established by CpG methylation, it is probable that all genes obtained in Table 6 would be downregulated. This result suggests that, from the list of significant GO terms obtained, cell adhesion is modified in tumour cells, a typical event in metastasis. Thus, methylation of genes connected to cell adhesion, like cadherins (PCDHGA[1-4] and PCDHGA[5-8]), can diminish cell adhesion properties.

Finally, and according to the analysis conducted through DAVID tool, we identified an enrichment term with p-value <0.05 named R-HSA-2173791 that belongs to the Reactome Pathway category. This term is a TGF-beta receptor signalling in EMT (epithelial to mesenchymal transition). In normal cells and in the early stages of cancer development, signalling by TGF-beta plays a tumour-suppressive role. However, in advanced cancers, TGF-beta signaling promotes metastasis

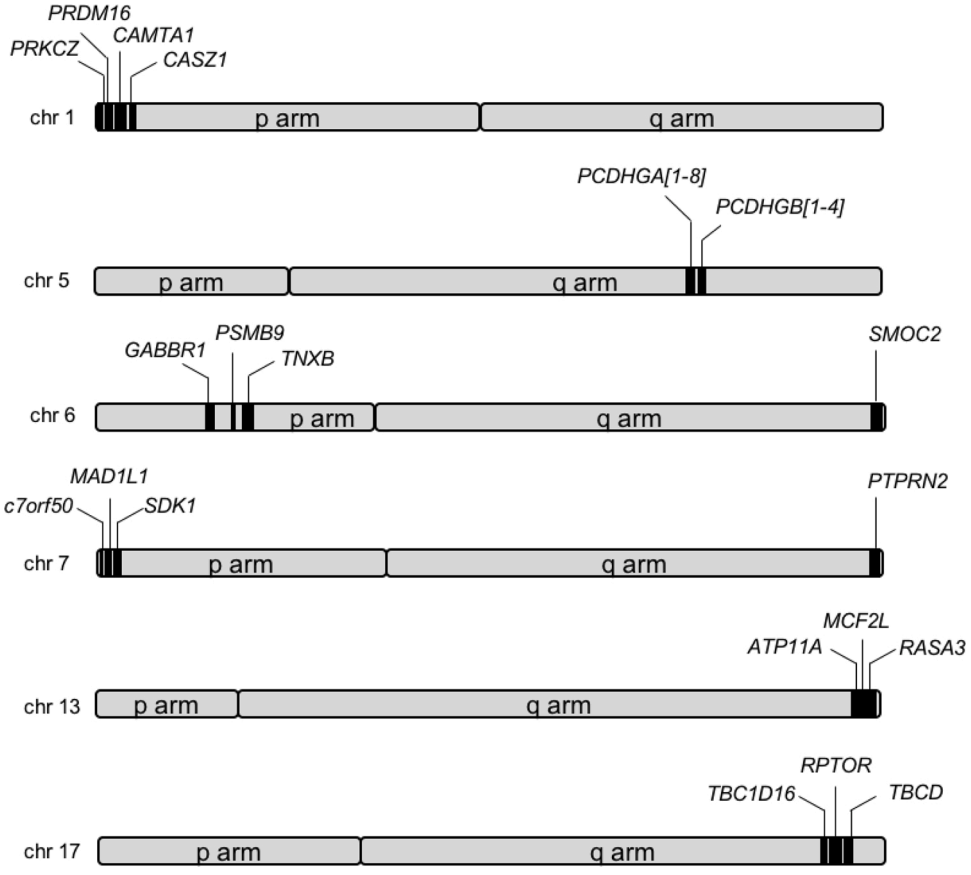

Fig. 11. Subset of genes grouped by chromosomal locations (chromosomes are not at the same scale for easy viewing).

by stimulating EMT. The DAVID tool also integrates several tissue expression data from different sources to identify enriched gene expression patterns cross hundreds of normal/disease tissues for any given gene lists. In particular, we identified several cancer-associated gene expression signatures significantly enriched. With regards to the problem under study, it highlighted terms related to tumour grade II, third mammary gland with status ER positive, PR positive and HER2 negative, third mammary gland breast carcinoma myoepithelium and third mammary gland neoplasia, among others.

4.3. Threats to validity

Our experimentation thus seems to indicate that feature generation by AEs may improve the prediction of breast cancer recurrence from DNA methylation data. Improvement seems to be based on the correct selection of new autoencoded features based on genetically significant CpG site weight. Although our conclusions are based on the findings in the experimentation, certain limitations may affect their validity:

- **Limited data.** The results presented in this paper were obtained from TCGA after a rigorous filtering to study similar cases (5-year follow-up, primary tumour, ductal and lobular cancer, etc.). Although this number of instances is comparable to the sizes used in other studies, it still constitutes a limitation since we cannot state that this set of patients is sufficiently large to conclude that autoencoding could reach similar results in any other dataset.
- **Lack of parameterisation and dependence of initial random selection of weights.** AEs are neural networks with a minimal set of parameters (such as activation units) or other regularisation parameters (such as weight decay terms on hidden unit weights). None of these parameters have been taken into account in this study, and default parameterisation provided by the selected implementation has been used. Furthermore, bias due to random initial selection of weights was partially controlled but could still have influenced the results.

  Default parameterisation has also been applied to the classifiers tested, which could have hampered their performance. In the future, a proper optimisation procedure should be established for the classifiers and the autoencoders also.
- **A limited set of competitors.** In this study, a limited set of machine learning techniques have been taken into account. There exist many other related options that have not been covered in our experimentation. In the future, we should provide insights into other machine learning techniques. Moreover, results of techniques that could deal with DNA methylation directly should especially be taken into account and their result should be duly reported.
- **Limited validation of results.** Fig. 7 provides a global view of the most important genes in the AEs. As can be seen, there exists a large set of weights related to CpG sites with no associated gene symbols. Those hitherto "unknown" CpG sites could provide new hints about breast cancer recurrence.
- **Lack of data fusion.** Related to the limitation outlined above, the data in this study is limited to DNA methylation data. However, other sources of data exist which could complete such types of data (e.g., RNA sequences from TCGA) to enrich our conclusions with a multiomic approach.

5. Conclusions

In this paper, we proposed a methodology to handle HDLSS datasets of DNA methylation based on the use of AEs and survival analysis. In particular, this process was applied to extract meaningful information based on relevant genes regarding breast cancer recurrence. A study was also provided on the relationship between autoencoded features learnt by AEs and the genetic knowledge regarding breast cancer in the literature.

The most heavily weighted genes in the autoencoded features developed by the AEs were all related to the literature on breast cancer and could be classified into five categories (confirmed-recurrence biomarkers, probable-recurrence biomarkers, obesity-related biomarkers, chemotherapeutic inhibitors, and probable pesticide exposure indicators). The enrichment analysis conducted based on genomic context and functional annotation revealed that the proposed methodology characterised the underlying information, discovering relevant genes for the problem under study and agreeing with prior biological knowledge. Functional annotation enrichment analysis found several enriched terms and linked connections between the genes, being an additional result that validates our methodology. Altogether, AEs could be used to find significant genes not only in breast cancer recurrence, but in other diseases, using a similar approach.

Although the results confirmed that autoencoding could select meaningful information from DNA methylation, further research is needed to verify this finding. The limited set of patients under study, the influence of a default parameterisation and initial selection of weights, and its effects on a larger set of machine learning techniques should be taken into account in future work. Furthermore, additional and preferably multiomic data should be used in future work in order to confirm the results in this study. Finally, CpG sites with unknown gene symbols outlined by the AEs should be studied in order to explore AE skills for the discovery of a new epigenetic signature related to recurrence in breast cancer.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The results shown here are based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga). This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects TIN2014-55894-C2-R and TIN2017-88209-C2-2-R. J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness. We also gratefully acknowledge the support of the NVIDIA Corporation with the donation of the Titan Xp and RTX 2080 Ti GPUs used in this research.

References

[1] International Agency for Research on Cancer – World Health Organization. Cancer fact sheets. 2019. https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics.

[2] World Cancer Research Fund. Breast cancer statistics. 2019. http://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf.

[3] Wu J, Zhang Y, Li M. Identification of methylation markers and differentially expressed genes with prognostic value in breast cancer. J Comput Biol 2019;26:1394–408.

[4] Celli F, Cumbo F, Weitschek E. Classification of large DNA methylation datasets for identifying cancer drivers. Big Data Res 2018;13:21–8.

[5] Yamada M, Tang J, Lugo-Martinez J, Hodzic E, Shrestha R, Saha A, et al. Ultra high-dimensional nonlinear feature selection for big biological data. IEEE Trans Knowl Data Eng 2018;30:1352–65.

[6] Cappelli E, Felici G, Weitschek E. Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. BioData Min 2018;11:22.

[7] Abreu PH, Santos MS, Abreu MH, Andrade B, Silva DC. Predicting breast cancer recurrence using machine learning techniques: a systematic review. ACM Comput Surv 2016;49:52:1–52:40.

[8] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.

[9] Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. Cancers 2019;11.

[10] Mihaylov I, Nisheva M, Vassilev D. Application of machine learning models for survival prognosis in breast cancer studies. Information (Switzerland) 2019;10.

[11] Fu B, Liu P, Lin J, Deng L, Hu K, Zheng H. Predicting invasive disease-free survival for early stage breast cancer patients using follow-up clinical data. IEEE Trans Biomed Eng 2019;66:2053–64.

[12] Daoud M, Mayo M. A survey of neural network-based cancer prediction models from microarray data. Artif Intell Med 2019;97:204–14.
[13] Charte D, Charte F, García S, del Jesus MJ, Herrera F. A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. Inf Fusion 2018;44:78–96.
[14] Macías-García L, Luna-Romera JM, García-Gutiérrez J, Martínez-Ballesteros M, Riquelme-Santos JC, González-Cámpora R. A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation. J Biomed Inform 2017;72:33–44.
[15] Liu Q, Hu P. Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer. Cancers 2019;11.
[16] Tan J, Ung M, Cheng C, Greene C. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pacific symposium on biocomputing 2015:132–43.
[17] Zhang D, Zou L, Zhou X, He F. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. IEEE Access 2018;6:28936–44.
[18] Guo Y, Qi Y, Li Z, Shang X. Improvement of cancer subtype prediction by incorporating transcriptome expression data and heterogeneous biological networks. BMC Med Genomics 2018;11.
[19] Guo Y, Shang X, Li Z. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing 2019;324:20–30.
[20] Zeng W, Glicksberg B, Li Y, Chen B. Selecting precise reference normal tissue samples for cancer research using a deep learning approach. BMC Med Genomics 2019;12.
[21] Wang Z, Wang Y. Exploring DNA methylation data of lung cancer samples with variational autoencoders. Proceedings – 2018 IEEE international conference on bioinformatics and biomedicine 2018:1286–9.
[22] Visakh R, Abdul Nazeer K. Deepaligner: deep encoding of pathways to align epigenetic signatures. Comput Biol Chem 2018;72:87–95.
[23] Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo Y-Y, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep 2016;6.
[24] Moon M, Nakai K. Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. J Bioinform Comput Biol 2018;16.
[25] Chaudhary K, Poirion O, Lu L, Garmire L. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res 2018;24:1248–59.
[26] Kim S, Kim T, Jeong H-H, Sohn K-A. Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. BMC Med Genomics 2018;11.
[27] Shen D, Shen H, Zhu H, Marron J. The statistics and mathematics of high dimension low sample size asymptotics. Stat Sin 2016;26:1747.
[28] Yata K, Aoshima M. Effective pca for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. J Multivar Anal 2010;101:2060–77.
[29] Samani FS, Zhang H, Stadler R. Efficient learning on high-dimensional operational data. 2019 15th international conference on network and service management (CNSM) 2019:1–9.
[30] TCGA. The cancer genome atlas program (TCGA). 2020. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
[31] The Cancer Genome Atlas Network, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–90.
[32] Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015;163:506–19.
[33] Walker DL, Bhagwate AV, Baheti S, Smalley RL, Hilker CA, Sun Z, et al. DNA methylation profiling: comparison of genome-wide sequencing methods and the infinium human methylation 450 bead chip. Epigenomics 2015;7:1287–302.
[34] Colleoni M, Sun Z, Price KN, Karlsson P, Forbes JF, Thürlimann B, et al. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: results from the international breast cancer study group trials i to v. J Clin Oncol 2016;34:927–35. PMID: 26786933.
[35] Apache Spark. Apache spark: lightning-fast cluster computing. 2019. http://spark.apache.org/.
[36] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[37] Chollet F, et al. Keras. 2015. https://keras.io.
[38] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org; 2015.
[39] Han J, Kamber M, Pei J, Han J, Kamber M, Pei J. Cluster analysis: basic concepts and methods. Data mining. 2012. p. 443–95.
[40] Gulcehre C, Bengio Y. Knowledge matters: importance of prior information for optimization. J Mach Learn Res 2013;17.
[41] Garcia-Gutierrez J. Github source code. 2020. https://github.com/jorgeBIGS/breast-aim.git.
[42] Buitinck L, et al. Scikit api. 2020. https://scikit-learn.org/stable/modules/classes.html.
[43] Burnham K, Anderson D. Model selection and multimodel inference: a practical information-theoretic approach. Springer New York; 2003.
[44] Riffenburgh R. Chapter 10 – risks, odds, and roc curves. In: Riffenburgh R, editor. Statistics in medicine (Third Edition). 3rd ed. San Diego: Academic Press; 2012. p. 203–19.
[45] Rey D, Neuhäuser M. Wilcoxon-signed-rank test. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 1658–9.
[46] Cox D, Oakes D. Analysis of survival data, monographs on statistics and applied probability. Chapman & Hall; 1996.
[47] Pubmed. Pubmed resource. 2020. http://www.ncbi.nlm.nih.gov/pubmed/.
[48] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2018;47:D607–13.
[49] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 2010;38:W214–20.
[50] Sherman BT, Lempicki RA, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc 2009;4:44.
[51] Parejo JA, García J, Ruiz-Cortés A, Riquelme JC. Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas. Actas del VIII Congreso Expa nol sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados. 2012.
[52] Sengelaub CA, Navrazhina K, Ross J, Halberg N, Tavazoie SF. PTPRN2 and PLC 1 promote metastatic breast cancer cell migration through PI(4,5)P2-dependent actin remodeling. EMBO J 2015;35.
[53] Casey M, Prakash A, Holian E, McGuire A, Kalinina O, Shalaby A, et al. Quantifying argonaute 2 (AGO2) expression to stratify breast cancer. BMC Cancer 2019;19:712.
[54] Singh R, Parveen M, Basgen JM, Fazel S, Meshesha MF, Thames EC, et al. Increased expression of beige/brown adipose markers from host and breast cancer cells influence xenograft formation in mice. Mol Cancer Res 2016;14:78–92.
[55] Joo E, Dowty JG, Milne RL, Wong E, Dugué P-A, English D, et al. Heritable DNA methylation marks associated with susceptibility to breast cancer. Nat Commun 2018;9.
[56] Lu P, Gu Y, Li L, Wang F, Yang X, Yang Y. Long noncoding RNA CAMTA1 promotes proliferation and mobility of human breast cancer cell line MDA- MB-231 via targeting miR-20b. Oncol Res 2017;26.
[57] Sun Q, Zhang X, Liu T, Liu X, Geng J, He X, et al. Increased expression of Mitotic Arrest Deficient-Like 1 (MAD1L1) is associated with poor prognosis and insensitive to Taxol treatment in breast cancer. Breast Cancer Res Treat 2013;140:323–30.
[58] Zhang J, Zhang S. Discovery of cancer common and specific driver gene sets. Nucleic Acids Res 2017;45. e86–e86.
[59] Rossetti S, Hoogeveen AT, Esposito J, Sacchi N. Loss of MTG16a (CBFA2T3), a novel rDNA repressor, leads to increased ribogenesis and disruption of breast acinar morphogenesis. J Cell Mol Med 2010;14:1358–70.
[60] Brough R, Gulati A, Haider S, Kumar R, Campbell J, Knudsen E, et al. Identification of highly penetrant Rb-related synthetic lethal interactions in triple negative breast cancer. Oncogene 2018;37:5701–18.
[61] Yang S, Zhang H, Guo L, Zhao Y, Chen F. Reconstructing the coding and non-coding RNA regulatory networks of miRNAs and mRNAs in breast cancer. Gene 2014;548:6–13.
[62] McCaffrey LM, Montalbano J, Mihai C, Macara IG. Loss of the Par3 polarity protein promotes breast tumorigenesis and metastasis. Cancer Cell 2016;30:351–2.
[63] Mamidi TKK, Wu J, Tchounwou PB, Miele L, Hicks C. Whole genome transcriptome analysis of the association between obesity and triple-negative breast cancer in caucasian women. Int J Environ Res Public Health 2018;15:2338.
[64] Wu J, Liu S, Fan Z, Zhang L, Tian Y, Yang R. A novel and selective inhibitor of PKC ζ potently inhibits human breast cancer metastasis in vitro and in mice. Tumor Biol 2016;37:8391–401.
[65] Ahmad A, Ginnebaugh KR, Yin S, Bollig-Fischer A, Reddy KB, Sarkar FH. Functional role of miR-10b in tamoxifen resistance of ER-positive breast cancer cells through down-regulation of HDAC4. BMC Cancer 2015;15. 540–540.
[66] Liu Z, Adams HC, Whitehead IP. The rho-specific guanine nucleotide exchange factor dbs regulates breast cancer cell migration. J Biol Chem 2009;284:15771–80.
[67] Sic You K, Yi YW, Kwak S-J, Seong Y-S. Inhibition of RPTOR overcomes resistance to EGFR inhibition in triple-negative breast cancer cells. Int J Oncol 2018;52.
[68] Shima J, Delaney J, Umesh A, Park J, Wall G, Su Q, et al. Disruption of protocadherin function and correlation with metastasis and cancer progression in TCGA patients. J Clin Oncol 2012;30. 70.
[69] Fanfani V, Citi L, Harris AL, Pezzella F, Stracquadanio G. Gene-level heritability analysis explains the polygenic architecture of cancer. 2019. bioRxiv.
[70] Fidalgo F, Rodrigues TC, Pinilla M, Silva AG, Maciel MdS, Rosenberg C, et al. Lymphovascular invasion and histologic grade are associated with specific genomic profiles in invasive carcinomas of the breast. Tumour Biol: J Int Soc Oncodev Biol Med 2015;36:1835–48.
[71] Rouette A, Trofimov A, Haberl D, Boucher G, Lavallée V-P, D'Angelo G, et al. Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers. Sci Rep 2016;6.
[72] Dong H, Wang W, Chen R, Zhang Y, Zou K, Ye M, et al. Exosome-mediated transfer of lncRNA-SNHG14 promotes trastuzumab chemoresistance in breast cancer. Int J Oncol 2018;53:1013–26.
[73] Yen M-C, Chou S-K, Kan J-Y, Kuo P-L, Hou M-F, Hsu Y-L. Solute carrier family 27 member 4 (SLC27A4) enhances cell growth, migration, and invasion in breast cancer cells. Int J Mol Sci 2018;19:3434.
[74] Banelli B, Romani M. Quantitative methylation analysis of the PCDHB gene cluster. New York, NY: Springer New York; 2015. p. 189–200.

[75] Sui X, Wang D, Geng S, Zhou G, He C, Hu X. Methylated promoters of genes encoding protocadherins as a new cancer biomarker family. Mol Biol Rep 2012; 39:1105–11.
[76] Liu Z, Lam N, Thiele CJ. Zinc finger transcription factor CASZ1 interacts with histones, DNA repair proteins and recruits NuRD complex to regulate gene transcription. Oncotarget 2015;6:27628–40.
[77] Fu J, Qin L, He T, Qin J, Hong J, Wong J, et al. The TWIST/Mi2/NuRD protein complex and its essential role in cancer metastasis. Cell Res 2011;21:275–89.
[78] Molina-Ortiz P, Orban T, Martin M, Habets A, Dequiedt F, Schurmans S. Rasa3 controls turnover of endothelial cell adhesion and vascular lumen integrity by a Rap1-dependent mechanism. PLOS Genetics 2018;14:1–25.
[79] Zhang Y-L, Wang R-C, Cheng K, Ring BZ, Su L. Roles of Rap1 signaling in tumor cell migration and invasion. Cancer Biol Med 2017;14:90–9.
[80] Rinaldi S, et al. IGF-I, IGFBP-3 and breast cancer risk in women: the european prospective investigation into cancer and nutrition (EPIC). Endocr-Relat Cancer 2006;13:593–605.
[81] Kaplan RC, et al. A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3. Hum Mol Genet 2011;20:1241–51.
[82] Vizoso M, et al. Epigenetic activation of a cryptic TBC1D16 transcript enhances melanoma progression by targeting EGFR. Nat Med 2015;21.
[83] Srikant CB. Somatostatin, endocrine updates. Springer US; 2004.
[84] Duong MN, Geneste A, Fallone F, Li X, Dumontet C, Muller C. The fat and the bad: mature adipocytes, key actors in tumor progression and resistance. Oncotarget 2017;8:57622–41.
[85] Fujii T, Yajima R, Tatsuki H, Oosone K, Kuwano H. Implication of atypical supraclavicular F18-fluorodeoxyglucose uptake in patients with breast cancer: association between brown adipose tissue and breast cancer. Oncol Lett 2017;14: 7025–30.
[86] Ishay-Ronen D, Diepenbruck M, Kalathur RKR, Sugiyama N, Tiede S, Ivanek R, et al. Gain fat-lose metastasis: converting invasive breast cancer cells into adipocytes inhibits cancer metastasis. Cancer Cell 2019;35. 17 – 32.e6.
[87] Lee S. The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. Sci Rep 2019;9:4855.
[88] Wang J, Yao X, Huang J. New tricks for human farnesyltransferase inhibitor: cancer and beyond. Med Chem Commun 2017;8:841–54.
[89] Zhang B, Groffen J, Heisterkamp N. Resistance to farnesyltransferase inhibitors in Bcr/Abl-positive lymphoblastic leukemia by increased expression of a novel ABC transporter homolog ATP11a. Blood 2005;106:1355–61.
[90] Guyton KZ, Loomis D, Grosse Y, El Ghissassi F, Benbrahim-Tallaa L, Guha N, et al. Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate. Lancet Oncol 2015;16:490–1.
[91] Lerro CC, Koutros S, Andreotti G, Friesen MC, Alavanja MC, Blair A, et al. Organophosphate insecticide use and cancer incidence among spouses of pesticide applicators in the agricultural health study. Occup Environ Med 2015; 72:736–44.
[92] Engel LS, Werder E, Satagopan J, Blair A, Hoppin JA, Koutros S, et al. Insecticide use and breast cancer risk among farmers' wives in the agricultural health study. Environ Health Perspect 2017;125. 097002-097002.
[93] Zhang X, Wallace AD, Du P, Lin S, Baccarelli AA, Jiang H, et al. Genome-wide study of DNA methylation alterations in response to Diazinon exposure in vitro. Environ Toxicol Pharmacol 2012;34.
[94] Thorstensen L, Qvist H, Heim S, Liefers G-J, Nesland JM, Giercksky K-E, et al. Evaluation of 1p losses in primary carcinomas, local recurrences and peripheral metastases from colorectal cancer patients. Neoplasia (New York, NY) 2000;2: 514.
[95] Arlt M, Herzog T, Mutch D, Gersell D, Liu H, Goodfellow P. Frequent deletion of chromosome 1p sequences in an aggressive histologic subtype of endometrial cancer. Hum Mol Genet 1996;5:1017–21.
[96] Qazilbash MH, Saliba RM, Ahmed B, Parikh G, Mendoza F, Ashraf N, et al. Deletion of the short arm of chromosome 1 (del 1p) is a strong predictor of poor outcome in myeloma patients undergoing an autotransplant. Biol Blood Marrow Transplant 2007;13:1066–72.
[97] Titus AJ, Way GP, Johnson KC, Christensen BC. Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes. Sci Rep 2017;7:1–9.
[98] Novak P, Jensen T, Oshiro MM, Watts GS, Kim CJ, Futscher BW. Agglomerative epigenetic aberrations are a common event in human breast cancer. Cancer Res 2008;68:8616–25.
[99] Wang C, Cicek MS, Charbonneau B, Kalli KR, Armasu SM, Larson MC, et al. Tumor hypomethylation at 6p21. 3 associates with longer time to recurrence of high-grade serous epithelial ovarian cancer. Cancer Res 2014;74:3084–91.
[100] Iwaya T, Sawada G, Amano S, Kume K, Ito C, Endo F, et al. Downregulation of ST6GALNAC1 is associated with esophageal squamous cell carcinoma development. Int J Oncol 2017;50:441–7.
[101] McRonald FE, Liloglou T, Xinarianos G, Hill L, Rowbottom L, Langan JE, et al. Down-regulation of the cytoglobin gene, located on 17q25, in tylosis with oesophageal cancer (TOC): evidence for trans-allele repression. Hum Mol Genet 2006;15:1271–7.
[102] Jerónimo C, Henrique R, Hoque MO, Mambo E, Ribeiro FR, Varzim G, et al. A quantitative promoter methylation profile of prostate cancer. Clin Cancer Res 2004;10:8472–8.