

RESEARCH

Open Access



# Merging microarray studies to identify a common gene expression signature to several structural heart diseases

Olga Fajarda<sup>1</sup> , Sara Duarte-Pereira<sup>1,2</sup>, Raquel M. Silva<sup>1,2,3</sup> and José Luís Oliveira<sup>1</sup>

\*Correspondence:

[olga.oliveira@ua.pt](mailto:olga.oliveira@ua.pt)

<sup>1</sup>IEETA/DETI, University of Aveiro,

3810-193 Aveiro, Portugal

Full list of author information is available at the end of the article

## Abstract

**Background:** Heart disease is the leading cause of death worldwide. Knowing a gene expression signature in heart disease can lead to the development of more efficient diagnosis and treatments that may prevent premature deaths. A large amount of microarray data is available in public repositories and can be used to identify differentially expressed genes. However, most of the microarray datasets are composed of a reduced number of samples and to obtain more reliable results, several datasets have to be merged, which is a challenging task. The identification of differentially expressed genes is commonly done using statistical methods. Nonetheless, these methods are based on the definition of an arbitrary threshold to select the differentially expressed genes and there is no consensus on the values that should be used.

**Results:** Nine publicly available microarray datasets from studies of different heart diseases were merged to form a dataset composed of 689 samples and 8354 features. Subsequently, the adjusted  $p$ -value and fold change were determined and by combining a set of adjusted  $p$ -values cutoffs with a list of different fold change thresholds, 12 sets of differentially expressed genes were obtained. To select the set of differentially expressed genes that has the best accuracy in classifying samples from patients with heart diseases and samples from patients with no heart condition, the random forest algorithm was used. A set of 62 differentially expressed genes having a classification accuracy of approximately 95% was identified.

**Conclusions:** We identified a gene expression signature common to different cardiac diseases and supported our findings by showing their involvement in the pathophysiology of the heart. The approach used in this study is suitable for the identification of gene expression signatures, and can be extended to different diseases.

**Keywords:** Heart disease, Random forest, Gene expression signature, Microarray data



## Background

Heart disease is the leading cause of death worldwide [1] and in particular in the United States [2] and Europe [3]. The 2017 Global Burden of Disease estimated that ischaemic heart disease alone is responsible for approximately 8.93 million deaths globally, which represents an increase of 22.3% compared to 2007. Hypertensive heart disease, in turn, is estimated to be responsible for approximately 0.93 million deaths and has increased by 46.6% compared to 2007 [4]. A thorough understanding of heart disease can lead to the development of more efficient diagnosis and treatments that may prevent premature deaths.

A gene expression signature (GES) is a set of genes whose altered expression can distinguish patients with different conditions, e.g. healthy vs. diseased [5, 6]. GES can be used for diagnosis, prognosis or prediction of therapeutic response [7] and it can also assist drug discovery by helping to identify a new potential target [6]. Several studies identified GESs for specific heart conditions. Barth et al. [8] identified 27 genes that can distinguish patients with dilated cardiomyopathy from patients with nonfailing hearts. Kittleson et al. [9] compared the gene expression of hearts from patients with nonischemic and ischemic cardiomyopathy with those from patients with nonfailing hearts. They identified 257 genes differentially expressed in nonischemic cardiomyopathy and 72 genes in ischemic cardiomyopathy. Tan et al. [10] reported 103 genes that were differentially expressed between failing and nonfailing hearts in patients with end-stage dilated cardiomyopathy.

Microarray technology is widely used to measure, in a single experiment, the expression levels of thousands of genes simultaneously [11]. The development of next-generation sequencing led to the conception of a new technology to measure gene expression, the RNA-Sequencing (RNA-seq) [12]. Despite the advantages of RNA-seq, microarray technology continues to be widely used, due to its lower cost and the existence of mature, reliable and robust processes and analysis tools [13, 14]. Furthermore, as the scientific community recommends that the data generated should be publicly available [13], several repositories were created. The Gene Expression Omnibus (GEO) [15] at the National Center for Biotechnology Information (NCBI) and the ArrayExpress [16] at the European Bioinformatics Institute (EMBL-EBI) nowadays provide a tremendous amount of microarray data available for further analysis.

Most microarray datasets are composed of a limited number of samples, and therefore, have low statistical power to identify a GES [17]. A way to obtain more reliable results is by merging microarray datasets from independent studies, since this leads to an increase of sample size [18]. However, merging microarray datasets is challenging, since most of the datasets were originated using different platforms measuring the expression of diverse sets of genes [19]. Furthermore, combining microarray datasets from different experiments introduces a batch effect to the data. Batch effect is the term used to identify technical, non-biological, variations introduced in the measurements due to the use of different processes, protocols and platforms [20]. This technical variation can obscure and confound true biological variation, leading to erroneous results [21].

GESs are commonly identified using statistical methods. Fold change and statistical tests like the *t* test are frequently used methods [22, 23]. The R/Bioconductor [24] software package `limma` [25] is also widely used and is considered one of the best methods to identify differentially expressed genes in comparison studies [26, 27]. This package imple-

ments linear models and empirical Bayes methods for microarray data analysis [28]. All these methods are based on the definition of an arbitrary threshold to select the GES and there is no consensus as to the values that should be used.

More recently, supervised machine learning algorithms have been applied to identify differentially expressed genes [29–32]. These algorithms have the leverage to construct a prediction model that can be applied to classify new samples. However, in a microarray experiment the number of features (the genes) is substantially higher than the number of samples and supervised machine learning algorithms can be inefficient when applied to high-dimensional datasets [33]. One way around the high-dimensional problem is reducing the number of features before applying supervised machine learning algorithms. Random forest is a supervised learning algorithm developed by Breiman [34] that constructs various decision trees, using for each split a random subset of features, and makes a prediction by combining the predictions of the different decision trees. It is dependent on only two tuning parameters, provides measures of variable importance and can be used directly for high-dimensional problems without reducing the number of features [35]. An empirical comparison of ten supervised learning algorithms performed by Caruana and Niculescu-Mizil [36] concluded that random forest was one of the algorithms that gave the best average performance.

The objective of this study was to identify a common GES in heart disease. To achieve this goal, we first merged nine publicly available microarray datasets from studies of different heart diseases. Then, we randomly divided the merged dataset into a training set and a test set and repeated this procedure 30 times, obtaining 30 training sets and 30 test sets. Subsequently, we used the R/Bioconductor software package `limma` to determine the adjusted  $p$ -value and the fold change for every training set. A set of adjusted  $p$ -value cutoffs combined with a list of different fold change thresholds were used to obtain several differentially expressed gene sets. To obtain a GES for every combination of adjusted  $p$ -value and fold change cutoff, we intersected the 30 sets of differentially expressed genes obtained using the 30 training sets. Afterwards, we evaluated the performance of every GES, on the 30 test sets, using the random forest algorithm and identified the one which had the best accuracy in classifying samples from patients with heart diseases and samples from patients with no heart condition. We identified a set of 62 differentially expressed genes with a classification accuracy of approximately 95%.

## Methods

The methodology used to obtain a GES for heart disease is described in this section, as well as the functional analysis performed.

### Data selection

All the datasets used are publicly available and were downloaded from GEO. The query: *((heart) OR cardio) AND (((disease) OR pathology) OR failure)* and the following filter criteria were used:

- Species: *Homo sapiens*;
- Sample types: heart tissue;
- Number of samples: more than 23 diseased or control samples (i.e. samples collected from heart donors with no previous history of heart disease);
- Access to unprocessed data (.cel files).

Nine gene expression datasets, with the following accession numbers, were selected: GSE1145 [37], GSE1869 [9], GSE2240 [38], GSE17800 [39], GSE21610 [40], GSE22253 [41], GSE42955 [42], GSE57338 [43] and GSE115574 [44]. A summary of the datasets is presented in Table 1, where, for each dataset, the platform, the number of samples and the heart diseases of the diseased samples can be found.

Some original datasets had more samples than those used in this study and the reasons for excluding some samples are given below.

Data of the original dataset GSE1145 were collected using the Affymetrix Human Genome U133 Plus 2.0 array (GPL96) and the Affymetrix Human Genome U95 Version 2 array (GPL8300). The seventeen samples of platform GPL8300 were not used in this study because the gene list of this platform is substantially different from the gene list of the remaining platforms used. Regarding the dataset GSE2240, we did not use the five samples of patients which had diabetes mellitus, because this may alter the gene expression patterns. We also exclude the 30 samples of the original dataset GSE21610 which were collected after the implementation of a ventricular assist device (VAD), since the use of a VAD can alter the gene expression patterns.

Concerning the dataset GSE22253, we did not use the 21 samples which have rs1333049 genotype CC, because Pilbrow et. al [41] conclude that the risk allele associated with coronary heart disease is C. Finally, the original dataset GSE115574 is composed of 29 samples obtained from the left atrial tissue and 30 obtained from the right atrial tissue. Since the samples obtained from the left and right atrial tissue came from the same patients and most of the samples from the other datasets were obtained from the left ventricular tissue, this study only used the samples obtained from the left atrial tissue.

### Data pre-processing

Before merging the microarray datasets, the raw data (.cel files) must go through pre-processing. The raw data of the same platform were merged and we used the `oligo` package [45], of the R/Bioconductor software package, which implements the robust mul-

**Table 1** Summary of the nine datasets used in this study

Dataset	Platform	No. of samples (diseased/control)	Diseases
GSE1145	GPL570	<b>90</b> (79/11)	Idiopathic dilated cardiomyopathy; ischemic cardiomyopathy; familial cardiomyopathy; hypertrophic cardiomyopathy; post-partum cardiomyopathy ; viral cardiomyopathy
GSE1869	GPL96	<b>25</b> (25/0)	Ischemic cardiomyopathy; nonischemic cardiomyopathy
GSE2240	GPL96	<b>30</b> (25/5)	Aortic and mitral regurgitation; aortic and mitral stenosis; dilated cardiomyopathy; coronary artery disease
GSE17800	GPL570	<b>48</b> (40/8)	Dilated cardiomyopathy
GSE21610	GPL570	<b>38</b> (30/8)	Dilated cardiomyopathy; ischemic cardiomyopathy
GSE22253	GPL6244	<b>87</b> (0/87)	None
GSE42955	GPL6244	<b>29</b> (24/5)	Dilated cardiomyopathy; ischemic cardiomyopathy
GSE57338	GPL11532	<b>313</b> (177/136)	Idiopathic dilated cardiomyopathy; ischemic cardiomyopathy
GSE115574	GPL570	<b>29</b> (29/0)	Severe mitral regurgitation
Total		<b>689</b> (429/260)	

tichip average (RMA) pre-processing method [46], to perform background correction, normalization and probe summarization.

The microarrays used in this study are oligonucleotide microarrays and each probe corresponds to one or a set of short oligonucleotide sequences. In such arrays a gene can be represented by multiple sequences, i.e. multiple probes, and the expression measurements of these probes which represent the same gene may be very different [13]. We decided to remove these conflicting expression measurements, but in order not to significantly reduce the number of genes used in the study, we used, as probe identifier, the GenBank sequence accession identifier, which uniquely identifies a biological sequence. In Affymetrix Human Genome U133A Array (GPL96) and Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), a probe is associated with a unique GenBank identifier, but in Affymetrix Human Gene 1.0 ST Array (GPL6244) and Affymetrix Human Gene 1.1 ST Array (GPL11532) a probe is associated with a list of GenBank identifiers. So, firstly, for each dataset obtained using platforms GPL96 and GPL570, all the probes corresponding to multiple or no GenBank identifier were removed. Concerning the dataset obtained using platforms GPL6244 and GPL11532, only the lists containing unique GenBank identifiers were maintained, all other lists being removed. Besides, we assigned to each GenBank identifier of a list the corresponding expression measurements of that list.

At this stage, we identified the common GenBank identifier across the different platforms and merged the datasets using only these common GenBank identifiers. The resulting merged dataset has 689 samples and 8354 features (the common GenBank identifiers).

### Feature selection

As observed previously, GESs are commonly identified using statistical methods, but these methods depend on the definition of an arbitrary threshold to select the GES. The cutoffs for the adjusted  $p$ -value commonly used are 0.01 and 0.05 [47]. Concerning the fold change, the cutoffs normally used are 1.5 and 2 [48] and between 2 and 3 [13, 23]. In this study, the merged dataset was randomly divided into a training set (70% of the samples) and a test set (the remaining 30%). This procedure was repeated 30 times and this way 30 different training sets and 30 different test sets were obtained. Next, the R/Bioconductor software package `limma` was used to determine in each training set, for each feature, the adjusted  $p$ -value (adjusted using Benjamini and Hochberg's method to control the false discovery rate [49]) and the fold change. In this study, we used the adjusted  $p$ -value and the fold change combined and instead of a threshold, we used a list of thresholds to identify features which represent differentially expressed genes. Thus for the adjusted  $p$ -value, we used as thresholds the values 0.01 and 0.05 and for fold change, we used the values within the range 1.5–3, which correspond to  $\log_2$  fold changes in the range of 0.585 and 1.585, approximately. In this way we obtained, for each training set, several sets of features that represent differentially expressed genes. For every combination of adjusted  $p$ -value and fold change cutoff, we intersected the 30 sets of features obtained using the 30 training set and get a feature set for every combination of adjusted  $p$ -value and fold change cutoff.

### Batch effect removal

The gene expression measurements may vary according to biological factors as well as

non-biological ones, i.e. technical sources of variation, such as the use of different platforms or different processing times [50]. These non-biological variations are also called the batch effect. Several approaches exist to deal with the batch effect. Nygaard et. al [51] suggested that, when possible, the batch variable should be included in the statistical analysis. Therefore, when using the *limma* package we included the platform type as a covariate. Another approach is to adjust the data for batch effects before using the dataset and that is what we have done before using the random forest algorithm. We used the ComBat method [52] implemented in the *sva* package [53] to batch-adjust the gene expression data of the merged dataset.

### Random forest

The next step after batch-adjustment is to select from the various sets of features the one with the best predictive accuracy. Accuracy is the fraction of correct predictions and is determined as  $\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of prediction made}}$ . We used the random forest algorithm to evaluate the predictive accuracy of the various sets of features and implemented it using the R package *caret* [54]. *Caret's* random forest implementation has two parameters that can be fine-tuned, namely the number of trees in the forest (*ntree*) and the number of features in the random subset used in each split (*mtry*). To select the best parameters for every set of features, we used the training sets and repeated 10-fold cross-validation. The models were fine-tuned to maximize the accuracy. Combinations of the following values for each parameter were used:

- *ntree*: 125, 250, 375, 500, 625, 750, 875, 1000;
- *mtry*: 2, 3, ..., *n*, where *n* represents the total number of features in a set.

The evaluation of the performance of the tuned models was done using the test sets. Besides accuracy, other metrics can be used to evaluate the model's performance, namely balanced accuracy, specificity, precision, recall or sensitivity, the F1 Score, the Matthews correlation coefficient (MCC), the area under the ROC (receiver operating characteristic) curve (AUC) and the area under the precision-recall curve (AUCPR). These measurements are determined as:

$$\text{Balanced Accuracy} = \frac{1}{2} \times \left( \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP represents the number of diseased samples correctly classified; TN represents the number of control samples correctly classified; FP represents the number of control samples wrongly classified as diseased samples; and FN represents the number of diseased samples wrongly classified as control samples.

**Table 2** Number of GenBank identifiers remaining after the removal of repeated GenBank identifiers

Platform	GPL96	GPL570	GPL6244	GPL11532
No. of GenBank identifier	20079	48100	121142	57714

The ROC curve plots the recall as a function of 1-specificity at all classification thresholds and the AUC is the area under the ROC curve [55].

The precision-recall (PR) curve plots the recall as a function of precision at all classification thresholds. PR curves are more accurate in presenting the performance of models than ROC curves when the datasets used are unbalanced. The AUCPR is also known as the average precision [56]. Both AUC and AUCPR are summary metrics of the respective curves.

### Functional analysis

To investigate the functional meaning of the genes obtained with our approach, we performed a gene ontology (GO) enrichment analysis on the genes of the selected feature sets. We analyzed the up-regulated and down-regulated genes separately, using a statistical over-representation test (Fisher's exact, False Discovery Rate correction) in the PANTHER Classification System <sup>1</sup> [57]. Next, we obtained the networks of protein-protein interactions (PPIs) of the up-regulated and down-regulated genes from the gene set with the best overall results. We used the STRING database <sup>2</sup> [58], and searched for data from text-mining, experiments, databases and co-expression, with a default median confidence level. Finally, we retrieved information from the DisGeNet database <sup>3</sup> [59] to assess which genes had previously been associated with disease, specifically cardiac-related diseases. From the file with all gene disease associations (GDA), we filtered "diseaseSemanticType" by "Disease or Syndrome" and then "diseaseName" by all containing "cardio\*" or "cardiac" or "heart".

### Results

To identify a GES common in heart disease by merging microarray studies, we used a methodology composed of several steps. The first step consisted of identifying the studies to merge. We identified 9 datasets, whose data were obtained using four different microarray platforms. Using different platforms involves pre-processing of the datasets before they can be merged. We used the GenBank accession identifier to identify the probes and exclude the probes corresponding to multiple or no GenBank identifier. Table 2 presents the number of GenBank identifiers remaining in each dataset, according to the corresponding platform.

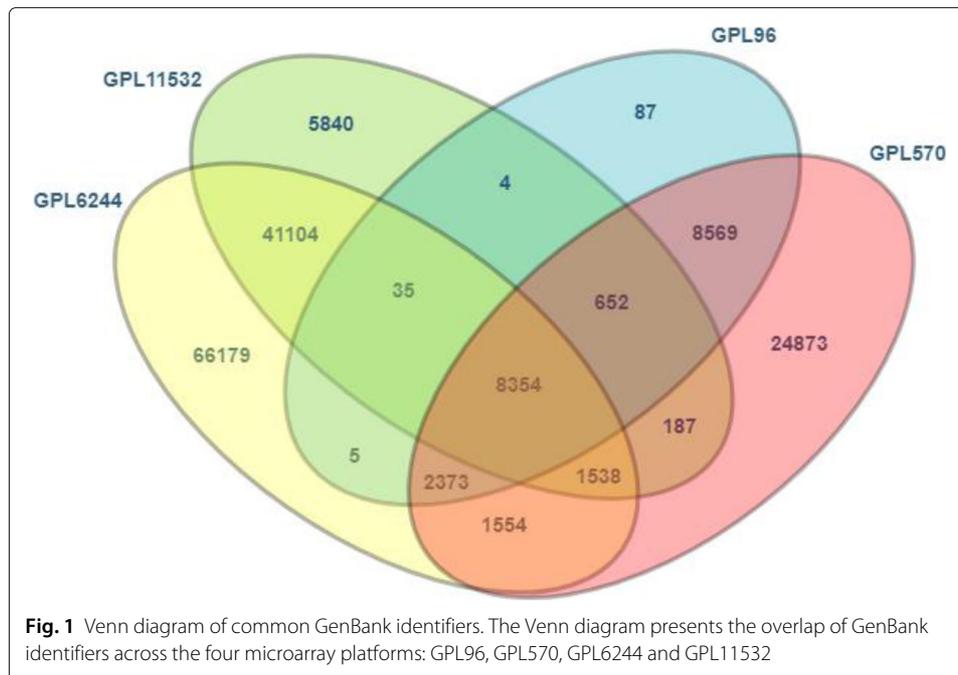
In platforms GPL6244 and GPL11532, the number of unique GenBank identifiers is higher than in the other platforms because in these two platforms the probes are associated with a list of GenBank identifiers.

Before merging the datasets we determined the common GenBank identifier across the four platforms. Figure 1 presents a Venn diagram of the common GenBank identifier across the four platforms. The four platforms have 8354 GenBank identifiers in common

<sup>1</sup><http://pantherdb.org>, Version 15.0

<sup>2</sup><https://string-db.org>, Version 11.0

<sup>3</sup><https://www.disgenet.org>, Version 6.0



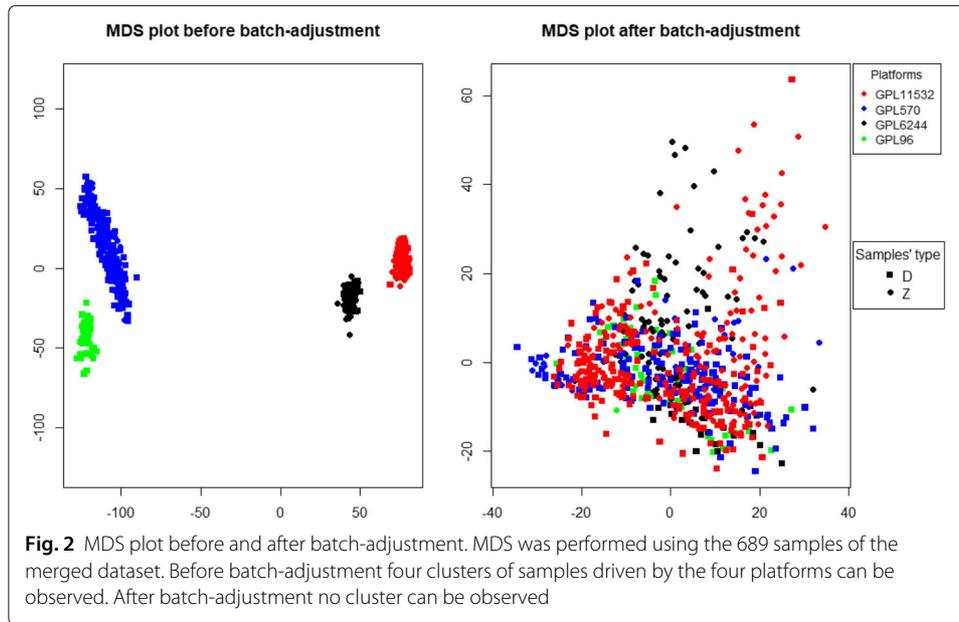
and all the probes not corresponding to these GenBank identifiers were removed from the nine datasets.

As can be observed in Fig. 1, platforms GPL6244 and GPL11532 are more similar to each other than to the other two, having 51031 GenBank identifiers in common. Platforms GPL96 and GPL570 are also more similar to each other, having 19948 GenBank identifiers in common.

After merging the various datasets to form a common dataset with 689 samples and 8354 features, we randomly divided the merged dataset into a training set and a test set and repeated this procedure 30 times, obtaining 30 different training and 30 different test sets. Then, we used the R/Bioconductor software package `limma` to obtain several sets of features for every training set. To get a unique feature set for every combination of adjusted  $p$ -value and fold change cutoff, we intersect the 30 features sets obtained using the 30 training sets. We observed that for a fold change within the range 1.5–3, we obtained the same sets of features using a cutoff for the adjusted  $p$ -value of 0.01 and of 0.05. Additionally, for fold changes within the range 2.7–3 the features sets had a very small number of genes (less than five) and therefore we choose not to use these feature sets.

For each fold change threshold used and a  $p$ -value of 0.01 we obtained a set of different features, resulting in a total of 12 different sets. As the fold change threshold increases, the number of features decreases. So for a fold change cutoff of 1.5 we obtained a set of 95 features and for a fold change cutoff of 2.6 we obtained a set of 7 features.

To evaluate the predictive accuracy of each set of features we used the random forest algorithm. However, before applying the random forest algorithm, a batch-adjustment to the data was required. Figure 2 presents the multidimensional scaling (MDS) plot showing the distribution of the merged dataset before and after the batch-adjustment. As can be observed before the batch-adjustment there are four clusters corresponding to the four platforms which disappear after the batch-adjustment.



The parameters *ntree* and *mtry* of the random forest were fine-tuned using a set of different values and using the training sets. To evaluate the performance of the tuned models we used the test sets.

Table 3 presents for each set of features, the number of features, the correspondent number of genes, the mean and the 95% confidence interval (95%-CI) of the accuracy and balanced accuracy of the model when applied to the test sets.

Table 4 presents for every set of features the mean and the 95% confidence interval of the specificity, precision, and recall for the model when applied to the test sets. Table 5 presents the mean and the 95% confidence interval of the F1 score, the MCC, and the mean and the 95% confidence interval of the AUC and the AUCPR are presented in Table 6.

**Table 3** For every fold change, the number of features, the number of genes, the accuracy and the balanced accuracy of the classifier

Fold change	No. of features	No. of genes	Accuracy		Balanced Accuracy	
			mean	95%-CI	mean	95%-CI
1.5	95	90	0.9468	(0.9409, 0.9526)	0.9382	(0.9311, 0.9454)
1.6	65	62	<b>0.9492</b>	<b>(0.9439, 0.9545)</b>	<b>0.9407</b>	<b>(0.9341, 0.9472)</b>
1.7	48	46	0.9476	(0.9419, 0.9532)	0.9388	(0.9318, 0.9458)
1.8	31	29	0.9474	(0.9419, 0.9530)	0.9388	(0.9321, 0.9456)
1.9	23	21	0.9461	(0.9404, 0.9518)	0.9375	(0.9309, 0.9442)
2.0	20	18	0.9426	(0.9365, 0.9487)	0.9337	(0.9266, 0.9409)
2.1	16	14	0.9424	(0.9363, 0.9485)	0.9336	(0.9265, 0.9407)
2.2	14	12	0.9416	(0.9350, 0.9481)	0.9326	(0.9252, 0.9401)
2.3	13	11	0.9388	(0.9320, 0.9457)	0.9294	(0.9216, 0.9372)
2.4	10	9	0.9309	(0.9245, 0.9373)	0.9221	(0.9152, 0.9290)
2.5	8	7	0.9299	(0.9222, 0.9377)	0.9215	(0.9133, 0.9297)
2.6	7	6	0.9293	(0.9217, 0.9368)	0.9211	(0.9131, 0.9290)

**Table 4** For every fold change the mean and the 95% confidence interval of the specificity, precision, and recall of the classifier

Fold change	Specificity		Precision		Recall	
	mean	95%-CI	mean	95%-CI	mean	95%-CI
1.5	0.9030	(0.8894, 0.9166)	0.9432	(0.9357, 0.9507)	0.9734	(0.9685, 0.9784)
1.6	<b>0.9056</b>	<b>(0.8927, 0.9184)</b>	<b>0.9447</b>	<b>(0.9376, 0.9519)</b>	<b>0.9758</b>	<b>(0.9714, 0.9801)</b>
1.7	0.9026	(0.8889, 0.9162)	0.9431	(0.9355, 0.9506)	0.9750	(0.9706, 0.9794)
1.8	0.9034	(0.8905, 0.9164)	0.9434	(0.9363, 0.9506)	0.9742	(0.9694, 0.9790)
1.9	0.9021	(0.8901, 0.9142)	0.9426	(0.9359, 0.9492)	0.9729	(0.9676, 0.9782)
2.0	0.8974	(0.8843, 0.9106)	0.9399	(0.9327, 0.9471)	0.9701	(0.9646, 0.9755)
2.1	0.8974	(0.8846, 0.9102)	0.9398	(0.9328, 0.9469)	0.9698	(0.9641, 0.9755)
2.2	0.8957	(0.8823, 0.9091)	0.9389	(0.9316, 0.9463)	0.9695	(0.9627, 0.9764)
2.3	0.8906	(0.8771, 0.9041)	0.9360	(0.9285, 0.9434)	0.9682	(0.9620, 0.9745)
2.4	0.8859	(0.8735, 0.8983)	0.9328	(0.9260, 0.9396)	0.9583	(0.9500, 0.9667)
2.5	0.8868	(0.8738, 0.8997)	0.9330	(0.9257, 0.9403)	0.9563	(0.9476, 0.9649)
2.6	0.8872	(0.8743, 0.9000)	0.9332	(0.9261, 0.9404)	0.9549	(0.9456, 0.9643)

As can be observed in Tables 3, 4, 5, and 6 the feature set obtaining the best mean accuracy (approximately 95%), mean specificity, mean precision, mean F1 score, mean MCC, mean AUC and mean AUCPR is the one using a fold change cutoff of 1.6.

The feature set using a fold change cutoff of 1.6 is composed of 65 GenBank identifiers which correspond to 62 genes. 37 are up-regulated (fold change > 1.6) in heart disease and 25 of these genes are down-regulated (fold change <  $\frac{1}{1.6} \simeq 0.625$ ).

Table 7 presents the 37 up-regulated genes from the feature set obtained using a fold change cutoff of 1.6, as well as the respective mean adjusted  $p$ -value, the mean fold change and the mean variable importance obtained. Table 8 presents the 25 down-regulated genes along with the respective mean adjusted  $p$ -value, the mean fold change and the mean variable importance obtained. The mean adjusted  $p$ -values and the mean fold changes presented in Tables 7 and 8 were obtained by averaging the values obtained using the 30 training sets. The variable importances are determined by measuring the mean decrease accuracy using the out-of-bag samples. Variables with a larger importance measurement are more important for classification [60]. The mean variable importances

**Table 5** For every fold change the mean and the 95% confidence interval of the F1 score and MCC of the classifier

Fold change	F1 Score		MCC	
	mean	95%-CI	mean	95%-CI
1.5	0.9579	(0.9534, 0.9625)	0.8869	(0.8744, 0.8994)
1.6	<b>0.9599</b>	<b>(0.9557, 0.9640)</b>	<b>0.8921</b>	<b>(0.8808, 0.9034)</b>
1.7	0.9586	(0.9542, 0.9630)	0.8887	(0.8767, 0.9007)
1.8	0.9584	(0.9541, 0.9628)	0.8883	(0.8766, 0.9001)
1.9	0.9574	(0.9529, 0.9619)	0.8855	(0.8733, 0.8976)
2.0	0.9546	(0.9498, 0.9593)	0.8779	(0.8649, 0.8908)
2.1	0.9544	(0.9496, 0.9592)	0.8775	(0.8646, 0.8905)
2.2	0.9538	(0.9486, 0.9590)	0.8760	(0.8622, 0.8898)
2.3	0.9517	(0.9463, 0.9571)	0.8699	(0.8554, 0.8845)
2.4	0.9451	(0.9399, 0.9503)	0.8533	(0.8397, 0.8669)
2.5	0.9443	(0.9381, 0.9505)	0.8511	(0.8347, 0.8674)
2.6	0.9437	(0.9376, 0.9498)	0.8499	(0.8339, 0.8660)

**Table 6** For every fold change the mean and the 95% confidence interval of the AUC and the AUCPR of the classifier

Fold change	AUC		AUCPR	
	mean	95%-CI	mean	95%-CI
1.5	0.9391	(0.9322, 0.9459)	0.9392	(0.9321, 0.9462)
1.6	<b>0.9407</b>	<b>(0.9341, 0.9472)</b>	<b>0.9402</b>	<b>(0.9332, 0.9473)</b>
1.7	0.9388	(0.9318, 0.9458)	0.9384	(0.9310, 0.9459)
1.8	0.9388	(0.9321, 0.9456)	0.9387	(0.9316, 0.9458)
1.9	0.9375	(0.9309, 0.9442)	0.9376	(0.9309, 0.9443)
2.0	0.9337	(0.9266, 0.9409)	0.9344	(0.9272, 0.9416)
2.1	0.9336	(0.9265, 0.9407)	0.9348	(0.9276, 0.9420)
2.2	0.9326	(0.9252, 0.9401)	0.9334	(0.9259, 0.9408)
2.3	0.9795	(0.9734, 0.9856)	0.9042	(0.8842, 0.9242)
2.4	0.9221	(0.9152, 0.9290)	0.9253	(0.9186, 0.9320)
2.5	0.9215	(0.9133, 0.9297)	0.9252	(0.9176, 0.9328)
2.6	0.9211	(0.9131, 0.9290)	0.9251	(0.9178, 0.9324)

presented in Tables 7 and 8 were obtained by averaging the values obtained using the 30 test sets.

As can be observed in Tables 7 and 8, the values of the adjusted  $p$ -value are greatly reduced, with  $4.5375 \times 10^{-05}$  being the highest value.

It is worth noticing that the feature set obtained using a fold change cutoff of 2.6 still achieves an accuracy of approximately 93%. This set is composed of 7 GenBank identifiers which correspond to 6 genes. 3 genes are up-regulated (ASPN, SFRP4 and NPPA) and 3 are down-regulated (CD163, IL1RL1 and SERPINA3). We can, also, observe that in Table 7 the mean fold change of gene EIF1AY is higher than 2.6 and that in Table 8 the mean fold changes of genes FCN3 and PLA2G2A are lower than  $\frac{1}{2.6}$ . However, these genes are not included in the feature set obtained using fold change 2.6 since it is sufficient that a gene is not included in one of the thirty feature sets obtaining using the thirty training sets, to be excluded from the intersection set. We analyzed the thirty feature sets and observed that in four of them gene EIF1AY had a fold change lower than 2.6, in fourteen of them and in three of them genes FCN3 and PLA2G2A, respectively, had a fold change greater than  $\frac{1}{2.6}$ .

To further analyze the feature sets obtained using a fold change cutoff of 1.6, we performed a GO analysis on the up-regulated and down-regulated genes (see Fig. 3 and Additional file 1). Using the PANTHER GO-slim annotation datasets, we observed that the up-regulated genes had the same result, for both the molecular function and the biological process categories, which was related to the Wnt signaling pathway. The down-regulated genes had no statistically significant results. Using the PANTHER complete annotation datasets, we observed an enrichment in up-regulated genes in processes related to tissue regeneration and development and with structural components from the extracellular matrix. The complete results of the GO analysis are presented in Fig. 3 (see Additional file 1 for details on the fold enrichment and FDR values).

To construct the PPI networks, we considered the up-regulated genes from the feature set with a fold change cutoff of 1.6, comprised of 37 genes. The genes DDX3Y, EIF1AY, KDM5D, RPS4Y1, UPS9Y and UTY had no protein match on STRING and were excluded. Out of the 31 proteins found and using only the queried proteins, we retrieved a final

**Table 7** Common up-regulated genes in heart disease obtained when using a fold change cutoff of 1.6

Gene Symbol	GenBank identifier	Adjusted <i>p</i> -value	Fold change	Variable importance
ASPN	NM_017680	$9.8581 \times 10^{-39}$	3.1542	3.0923
CRYM	NM_001888	$1.1463 \times 10^{-24}$	1.7969	3.2103
DDX3Y	NM_004660	$1.3868 \times 10^{-05}$	2.1572	1.6515
DSC1	NM_004948	$3.0749 \times 10^{-14}$	2.0620	1.7366
ECM2	NM_001393	$5.1541 \times 10^{-41}$	2.1289	3.9118
EIF1AY	NM_004681	$1.8035 \times 10^{-06}$	3.0338	1.9761
FMOD	NM_002023	$2.4210 \times 10^{-18}$	1.8844	3.313
FRZB	NM_001463	$3.4459 \times 10^{-43}$	2.5309	4.0964
GATM	NM_001482	$1.9088 \times 10^{-23}$	1.9125	2.4652
HAPLN1	NM_001884	$1.6750 \times 10^{-14}$	1.7568	1.4900
IFI44L	NM_006820	$2.5184 \times 10^{-26}$	1.9858	3.031
ISLR	NM_005545	$6.5434 \times 10^{-39}$	1.8439	7.1403
ITIH5	NM_030569	$1.4377 \times 10^{-41}$	1.7718	4.1672
KDM5D	NM_004653	$5.4325 \times 10^{-07}$	1.9162	1.3922
LRRC17	NM_005824	$7.9437 \times 10^{-18}$	1.7555	2.2458
LTBP2	NM_000428	$2.7206 \times 10^{-20}$	1.9063	4.0368
LUM	NM_002345	$7.7796 \times 10^{-39}$	2.3008	4.9349
MATN2	NM_002380	$2.6707 \times 10^{-18}$	1.7430	0.6137
MME	NM_007287	$3.4768 \times 10^{-28}$	1.8158	3.1662
MNS1	NM_018365	$3.1674 \times 10^{-41}$	1.9615	8.5390
MXRA5	AF245505	$5.9001 \times 10^{-25}$	2.2967	3.4605
NAP1L3	NM_004538	$4.7889 \times 10^{-15}$	1.8298	1.9093
NPPA	M30262	$4.5989 \times 10^{-12}$	3.4108	4.9190
OGN	NM_014057	$1.5746 \times 10^{-28}$	2.4848	3.2359
OMD	NM_005014	$1.2118 \times 10^{-20}$	1.9088	2.0563
PDE5A	NM_001083	$1.7283 \times 10^{-31}$	1.9547	7.1236
PROM1	NM_006017	$6.1354 \times 10^{-18}$	1.8225	1.7519
PTN	BC005916	$1.7766 \times 10^{-24}$	1.8927	2.3117
RASL11B	NM_023940	$3.6655 \times 10^{-21}$	1.8148	3.0627
RPS4Y1	NM_001008	$4.5375 \times 10^{-05}$	1.9747	1.0981
SFRP1	NM_003012	$6.8682 \times 10^{-22}$	2.2164	2.8539
SFRP4	NM_003014	$6.1063 \times 10^{-40}$	2.8962	5.9620
STAT4	NM_003151	$3.0232 \times 10^{-17}$	2.1129	4.0354
THBS4	NM_003248	$1.9041 \times 10^{-13}$	1.8075	1.3539
TLL2	NM_012465	$1.9225 \times 10^{-32}$	1.7020	3.1669
USP9Y	NM_004654	$1.2498 \times 10^{-08}$	2.3565	3.8957
UTY	AF000994	$8.8337 \times 10^{-07}$	1.8757	2.8492
UTY	NM_007125	$6.5398 \times 10^{-07}$	1.8808	2.8435

network, with 17 nodes (proteins) and 25 edges (interactions), which is shown in Fig. 4. Among the 17, eleven proteins had more than one interaction and 2 were still present in the feature set with a fold change cutoff of 2.6. For the 25 down-regulated genes, we obtained a final network with 15 nodes and 21 interactions (Additional file 2). We found that 10 proteins had more than one interaction.

As a final functional validation step, we used the DisGeNet database to evaluate the disease association status of the genes. We found a total of 7351 associations between 2757 genes and 241 diseases. From the 62 genes of the feature set with a fold change cutoff of 1.6, twenty-four genes were found to be associated with heart related diseases (see

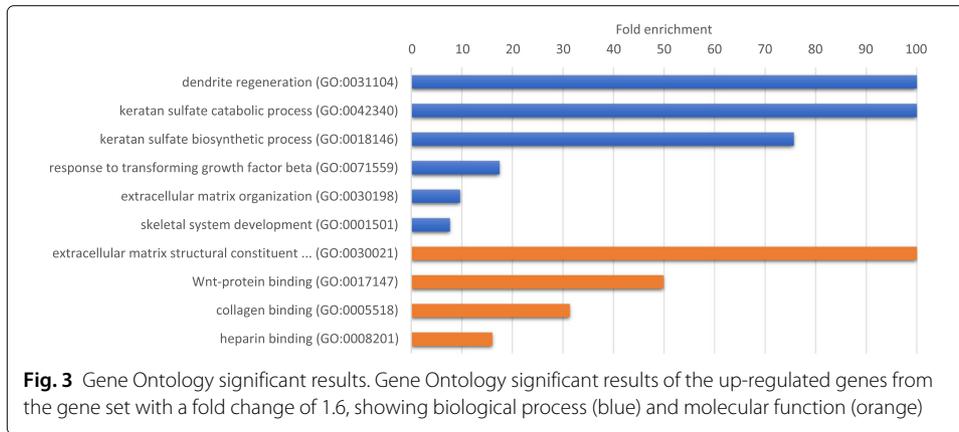
**Table 8** Common down-regulated genes in heart disease obtained when using a fold change cutoff of  $\frac{1}{1.6}$ 

Gene Symbol	GenBank identifier	Adjusted <i>p</i> -value	Fold change	Variable importance
ALOX5AP	NM_001629	$2.4434 \times 10^{-22}$	0.5147	2.1998
ANKRD2	NM_020349	$9.1377 \times 10^{-15}$	0.5051	2.4536
ANPEP	NM_001150	$8.1714 \times 10^{-23}$	0.5709	1.7386
AOX1	NM_001159	$6.0757 \times 10^{-18}$	0.5363	2.5075
CD14	NM_000591	$8.2159 \times 10^{-20}$	0.5454	5.6363
CD163	NM_004244	$8.6407 \times 10^{-36}$	0.2996	4.7795
CD163	Z22970	$2.0766 \times 10^{-36}$	0.3315	4.1402
CD53	NM_000560	$1.7938 \times 10^{-15}$	0.5862	2.1460
CTSC	NM_001814	$7.2224 \times 10^{-18}$	0.5549	2.8080
ETNPPL	NM_031279	$5.0566 \times 10^{-12}$	0.5824	5.701
F13A1	NM_000129	$5.1104 \times 10^{-15}$	0.5540	3.0357
FCER1G	NM_004106	$1.2145 \times 10^{-19}$	0.5216	3.0903
FCN3	NM_003665	$1.4659 \times 10^{-40}$	0.3841	6.3081
HMGCS2	NM_005518	$2.9399 \times 10^{-08}$	0.5035	1.8650
IL1R2	NM_004633	$3.0359 \times 10^{-22}$	0.5477	2.2061
IL1RL1	AB012701	$2.5061 \times 10^{-29}$	0.3773	3.5570
IL1RL1	NM_003856	$3.5562 \times 10^{-36}$	0.3192	2.8270
LMCD1	NM_014583	$1.2320 \times 10^{-14}$	0.5711	2.2943
MYOT	NM_006790	$2.4705 \times 10^{-22}$	0.4674	2.5393
PLA2G2A	NM_000300	$1.1862 \times 10^{-17}$	0.3579	1.8016
PLIN2	BC005127	$3.3583 \times 10^{-25}$	0.5383	4.2095
PTX3	NM_002852	$4.2469 \times 10^{-09}$	0.5383	0.7522
RNASE2	NM_002934	$1.0328 \times 10^{-31}$	0.5119	3.0843
S100A8	NM_002964	$8.1619 \times 10^{-12}$	0.4876	1.5052
SERPINA3	NM_001085	$5.0459 \times 10^{-71}$	0.1907	13.2424
SERPINE1	NM_000602	$9.2699 \times 10^{-18}$	0.3793	1.3912
VSIG4	NM_007268	$4.5242 \times 10^{-33}$	0.4330	3.6720

Additional file 3). From these 24 genes, five were also in the feature set with a fold change cutoff of 2.6.

## Discussion

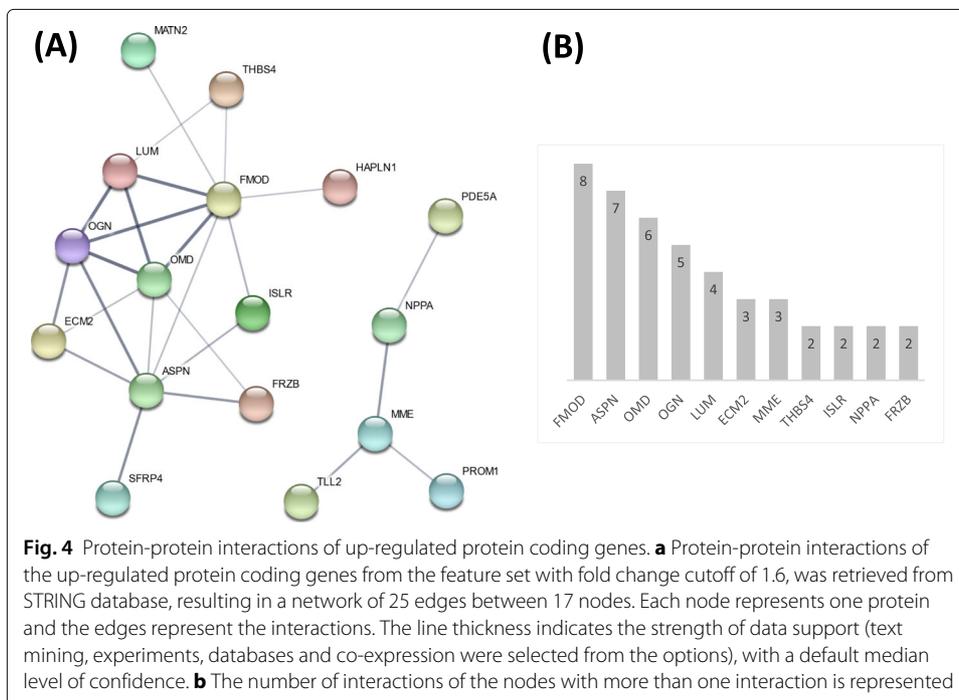
In this study, nine microarray datasets were merged in order to identify a GES common to different heart diseases. The term “heart diseases” is often incorrectly used as a synonym of cardiovascular diseases. According to the World Health Organization, as published in the “Global atlas on cardiovascular disease prevention and control” [61], cardiovascular diseases include diseases of the heart, vascular diseases of the brain and diseases of blood vessels. In our study, we focused on heart diseases that affect the structure of the organ (muscle and valves), leaving out all types of vascular diseases, as well as diseases that affect the function or rhythm of the heart (arrhythmias). As depicted in Table 1, the data sets analyzed included different types of cardiomyopathies (diseases of the heart muscle) and diseases of the heart valves, which are often grouped under the term “structural heart diseases”. Genes expressed in atrial and ventricular myocardial tissues from failing and non-failing hearts were analyzed. After applying the methodology described above, we obtained a set of 62 genes with altered expression levels by a fold change cutoff of 1.6, which best discriminates diseased from control samples. To evaluate the performance of



the model we used 30 tests sets. Besides this approach, we also evaluated the leave one study out procedure, where the datasets of 8 of the 9 studies are used for selecting the differentially expressed genes and training the model, and the dataset of the remained study is used to evaluate the model's performance. This process was repeated nine times so that the dataset of each study is once used for independent evaluation. However, the results obtained were over-optimistic, making this procedure unfeasible for our study.

To explore the biological meaning of the results, we performed a functional analysis comprised of GO analysis and PPIs network. Our goal was to understand if the function of the genes obtained was relevant in heart disease and therefore validate our approach to obtain a GES.

The GO analysis of the up-regulated genes revealed an over-representation of several biological processes and molecular functions related to the development of heart tissue and cardiac remodeling after injury (see Fig. 3 and Additional file 1), specifically the



involvement of the keratan sulfate metabolic process, which will be discussed later. Interestingly, enrichment in genes that code for proteins involved in the Wnt protein binding and in the regulation of the Wnt signaling pathway was a result highlighted by the PANTHER GO-slim annotation datasets, which are selected by curation. The Wnt-protein is a secreted growth factor involved in signaling. The Wnt signaling pathway, mostly via the beta-catenin pathway, also known as the canonical pathway, has a well-known role in cell proliferation and cell differentiation in tissue development and homeostasis [62]. The role of this pathway in the context of cardiovascular disease has been recently reviewed by Foulquier et. al. [63]. Specifically, the role of secreted frizzled-related proteins (SFRPs), which are a family of Wnt modulators, has been studied in this context [64], but remains largely unknown. Considering the down-regulated genes, most of the results are related to the immune system and inflammatory processes. Though the primary event of heart tissue damage is an inflammatory response, the cardiac repair is dependent on the suppression of inflammation to ensure the formation of a scar in the post-infarction response [65–67]. The down-regulation of genes involved in immune and inflammatory response had been observed in a similar study in dilated cardiomyopathy [8].

Next, we obtained a network of PPIs from the STRING database, using the 37 up-regulated genes from the feature set with a fold change cutoff of 1.6 (Fig. 4). We considered that the analysis of the up-regulated genes could be more interesting from the clinical point of view. The analysis of the PPIs network constructed with the down-regulated genes (Additional file 2) showed that the proteins with more interactions were mostly involved in processes related to both the inflammatory response and the immune system (FCER1G, CD14, CD163, and S100A8), which is in agreement with the GO results.

Among the proteins with the highest number of interactions in the up-regulated genes network, we found ASPN and NPPA, which are also found in the smallest feature set with a fold change cutoff of 2.6.

Natriuretic Peptide Precursor A (NPPA) gene encodes the precursor for the hormone atrial natriuretic peptide (ANP). ANP is synthesized and secreted by cardiac muscle cells from the atria in the heart and is a well-established biomarker for cardiovascular disease [68]. According to a recent review [69], it plays a key role in the regulation of cardiovascular volume and pressure homeostasis by inducing natriuresis, diuresis and vasodilation. Over the last four decades, studies have shown that the phenotype associated with NPPA genetic variants and the changes in the circulating levels of ANP reflect its value as a potential therapeutic target for cardiometabolic diseases, including heart failure [70, 71]. Therefore, adding to our study, NPPA gene has been previously identified in GES of cardiac diseases, such as heart failure [72] and dilated cardiomyopathy [8, 73].

Asporin, encoded by ASPN gene, is a glycoprotein from the family of the small leucine rich proteoglycans (SLRP) present in the cartilage tissue. It is a known negative regulator of osteoblast differentiation and might be involved in development of the heart valves [74], being among the structural and extracellular matrix proteins that have putative roles in mitral valve degeneration [75]. Nevertheless, to our knowledge, only a few studies have found an alteration in ASPN gene expression in the context of cardiovascular diseases. In the two studies performed in humans [8, 75], its importance has not been properly discussed. A very recent study performed in mice by Wang et. al. [76] has reported ASPN among the up-regulated genes in a GES of cardiac remodeling. Following those studies, we have identified ASPN as one of the most significant genes altered in heart disease,

since it was found among the highest fold change cutoff used and prominent in the PPI network. Together with another group of four proteins from the SLRP family, namely fibromodulin (FMOD), osteomodulin (OMD), osteoglycin (OGN) and lumican (LUM), the importance of extracellular remodeling processes in heart disease is emphasized. These FMOD, OMD, OGN and LUM proteins, also underlined by the PPIs network, are involved in keratan sulfate metabolic, catabolic and biosynthetic processes. Keratan sulfate is a glycosaminoglycan, a structural molecule, mostly found in the extracellular matrix. Keratan sulfate metabolism is involved in the development of heart tissue and has been implicated in heart disease. FMOD and LUM, for example, are increased in heart failure as a response to inflammation and play a role in cardiac remodeling [77, 78]. The findings highlighted by the PPIs network analysis agree with the GO analysis, where the biological processes of the biosynthesis and catabolism of the keratan sulfate was one of the main results.

After the functional study using GO and PPIs network analysis, we searched the DisGeNet database and found that approximately 39% of the genes studied had previously been associated with cardiac-related diseases, including NPPA, SERPINA3, SFRP4, IL1RL1 and CD163, still present in the feature set with a fold change cutoff of 2.6. This final observation strongly validates our results.

## Conclusion

With this study, we were able to successfully identify a GES common to different cardiac diseases, mainly structural heart diseases, and supported our findings by showing their involvement in the pathophysiology of the heart.

According to our findings and given its structural function, we suggest that asporin is likely to be involved in a cardiac tissue mechanism that is up-regulated in response to disease development, rather than having a causal effect. Although this has not yet been demonstrated, it should be further studied.

Regardless of advantages of having a GES, we also consider here that having a small set of markers to distinguish normal from diseased samples can ease their use as a panel for diagnosis or screening. Additionally, such genes can be further investigated in the context of new therapeutic approaches.

Finally, the approach used in this study is suitable for the identification of gene expression signatures and can be extended to different diseases.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s13040-020-00217-8>.

**Additional file 1:** gO results. Gene Ontology significant results of the up-regulated and down-regulated genes from the gene set with a fold change cutoff of 1.6. Only results for  $p$ -value < 0.05 are displayed.

**Additional file 2:** Protein-protein interactions of Down-regulated protein coding genes. **Figure (A)** Protein-protein interactions of the down-regulated protein coding genes from the feature set with fold change cutoff of 1.6, was retrieved from STRING database, resulting in a network of 21 edges between 15 nodes. Each node represents one protein and the edges represent the interactions. The line thickness indicates the strength of data support (text mining, experiments, databases and co-expression were selected from the options), with a default median level of confidence.

**Figure (B)** The number of interactions of the nodes with more than one interaction is represented.

**Additional file 3:** Table with the association status of the 24 genes which were found, using disGeNet, associated with heart related disease.

### Abbreviations

AUC: Area under the receiver operating characteristic curve; AUCPR: Area under the precision-recall curve; EMBL-EBI: European Bioinformatics Institute; GDA: Gene disease association; GEO: Gene Expression Omnibus; GES: Gene expression signature; GO: Gene ontology; MCC: Matthews correlation coefficient; MDS: Multidimensional scaling; NCBI: national Center for Biotechnology Information; PPI: Protein-protein interactions; PR: Precision-recall; RMA: Robust multichip average; RNA-seq: RNA-Sequencing; ROC: Receiver operating characteristic; VAD: Ventricular assist device

### Acknowledgments

Not applicable.

### Authors' contributions

All authors designed the study. OF and SDP selected the datasets supervised by RMS. OF implemented the code to obtain the results, supervised by JLO. SDP and OF analyzed the results supervised by RMS and JLO. OF and SDP wrote the manuscript and RMS and JLO revised it. All authors read and approved the final manuscript.

### Funding

This work was funded by the NETDIAMOND project, grant number POCI-01-0145-FEDER-016385. iBiMED is supported by UIDB/04501/2020, project POCI-01-0145-FEDER-007628. SDP is supported by FCT - Foundation for Science and Technology (national funds), grant SFRH/BD/108890/2015.

### Availability of data and materials

The datasets used during the current study are available from the Gene Expression Omnibus repository and their accession numbers are listed in Table 1. The R code used to obtain the results presented in the paper is available at <https://github.com/olgafajarda/MergingHD>.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal. <sup>2</sup>Department of Medical Sciences and iBiMED-Institute of Biomedicine, University of Aveiro, 3810-193 Aveiro, Portugal. <sup>3</sup>Current Address: Universidade Católica Portuguesa, Faculdade de Medicina Dentária, CIIS-Centro de Investigação Interdisciplinar em Saúde, Campus de Viseu, 3504-505 Viseu, Portugal.

Received: 18 February 2020 Accepted: 5 June 2020

Published online: 08 July 2020

### References

1. Mathers CD, Boerma T, Ma Fat D. Global and regional causes of death. *Br Med Bull.* 2009;92(1):7–32.
2. Murphy SL, Xu J, Kochanek KD, Arias E. Mortality in the united states, 2017. In: *NCHS Data Brief*; 2018. p. 1–8.
3. Townsend N, Wilson L, Bhatnagar P, Wickramasinghe K, Rayner M, Nichols M. Cardiovascular disease in europe: epidemiological update 2016. *Eur Heart J.* 2016;37(42):3232–45.
4. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet.* 2018;392(10159):1736–88.
5. Chang JT, Gatz ML, Lucas JE, Barry WT, Vaughn P, Nevins JR. Signature: a workbench for gene expression signature analysis. *BMC Bioinformatics.* 2011;12(1):443.
6. Sithara S, Crowley TM, Walder K, Aston-Mourney K. Gene expression signature: a powerful approach for drug discovery in diabetes. *J Endocrinol.* 2017;232:131–39.
7. Chibon F. Cancer gene expression signatures—the rise and fall?. *Eur J Cancer.* 2013;49(8):2000–9.
8. Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, Käbb S, Kreuzer E, Steinbeck G, Mansmann U, et al. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J Am Coll Cardiol.* 2006;48(8):1610–7.
9. Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, Conte JV, Tomaselli G, Garcia JG, Hare JM. Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. *Physiol Genomics.* 2005;21(3):299–307.
10. Tan F-L, Moravec CS, Li J, Apperson-Hansen C, McCarthy PM, Young JB, Bond M. The gene expression fingerprint of human heart failure. *Proc Natl Acad Sci.* 2002;99(17):11387–92.
11. Babu MM. Introduction to microarray data analysis. *Comput Genomics Theory Appl.* 2004:225–49. UK: Horizon Bioscience.
12. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10(1):57.
13. Raddatz BB, Spitzbarth I, Matheis KA, Kalkuhl A, Deschl U, Baumgaertner W, Ulrich R. Microarray-based gene expression analysis for veterinary pathologists: A review. *Vet Pathol.* 2017;54(5):734–55.
14. Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and rna-seq data for machine learning applications. *PeerJ.* 2016;4:1621.

15. Clough E, Barrett T. The gene expression omnibus database. In: Mathé E. DSe, editor. *Statistical Genomics*. New York: Springer; 2016. p. 93–110.
16. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, et al. Arrayexpress update—from bulk to single-cell expression data. *Nucleic Acids Res*. 2018;47(D1):711–5.
17. Soto J, Ortuño F, Rojas I. Integrative gene expression analysis of lung cancer based on a technology-merging approach. In: *IEEE EUROCON 2015-International Conference on Computer as a Tool (EUROCON)*. New Jersey: IEEE; 2015. p. 1–5.
18. Wang J, Do KA, Wen S, Tsavachidis S, McDonnell TJ, Logothetis CJ, Coombes KR. Merging microarray data, robust feature selection, and predicting prognosis in prostate cancer. *Cancer Informat*. 2006;2:117693510600200009.
19. Kumar Sarmah C, Samarasinghe S. Microarray data integration: frameworks and a list of underlying issues. *Curr Bioinforma*. 2010;5(4):280–9.
20. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solis DY, Duque R, Bersini H, Nowé A. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinforma*. 2012;14(4):469–90.
21. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC Med Genomics*. 2012;5(1):23.
22. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003;4(4):210.
23. Draghici S. Statistical intelligence: effective analysis of high-density microarray data. *Drug Discov Today*. 2002;7(11):55–63.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):80.
25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res*. 2015;43(7):47.
26. Chrominski K, Tkacz M. Comparison of high-level microarray analysis methods in the context of result consistency. *PLoS ONE*. 2015;10(6):0128845.
27. Jeanmougin M, De Reynies A, Marisa L, Paccard C, Nuel G, Guedj M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS ONE*. 2010;5(9):12336.
28. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3(1). <https://doi.org/10.2202/1544-6115.1027>.
29. Nishiwaki K, Kanamori K, Ohwada H. Finding a disease-related gene from microarray data using random forest. In: *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI\* CC)*. New Jersey: IEEE; 2016. p. 542–6.
30. Verhagen LM, Zomer A, Maes M, Villalba JA, Del Nogal B, Eleveld M, van Hijum SA, de Waard JH, Hermans PW. A predictive signature gene set for discriminating active from latent tuberculosis in warao amerindian children. *BMC Genomics*. 2013;14(1):74.
31. Wang D, Li J-R, Zhang Y-H, Chen L, Huang T, Cai Y-D. Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes*. 2018;9(3):155.
32. Yan Z, Li J, Xiong Y, Xu W, Zheng G. Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data. *Oncol Rep*. 2012;28(3):1036–42.
33. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen V, Duque R, Bersini H, Nowe A. *IEEE/ACM Trans Comput Biol Bioinforma (TCBB)*. 2012;9(4):1106–19.
34. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
35. Cutler A, Cutler DR, Stevens JR. Random forests. In: *Ensemble Machine Learning*. Boston: Springer; 2012. p. 157–75.
36. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh: ACM; 2006. p. 161–8.
37. Paul A, Schinke M, Brown J, Riggi LE, Izumo S, Bartunek J, Allen P, Tsubakihara M. Changes in cardiac transcription profiles brought about by heart failure. In: *Cardiogenomics* HU, editor. Bauer Center for Genomic Research, NCBI, Gene Expression Omnibus; 2004.
38. Barth AS, Merk S, Arnoldi E, Zwermann L, Kloos P, Gebauer M, Steinmeyer K, Bleich M, Kääh S, Hinterseer M, et al. Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature. *Circ Res*. 2005;96(9):1022–9.
39. Ameling S, Herda LR, Hammer E, Steil L, Teumer A, Trimpert C, Dörr M, Kroemer HK, Klingel K, Kandolf R, et al. Myocardial gene expression profiles and cardiodepressant autoantibodies predict response of patients with dilated cardiomyopathy to immunoadsorption therapy. *Eur Heart J*. 2012;34(9):666–75.
40. Schwientek P, Ellinghaus P, Steppan S, D'Urso D, Seewald M, Kassner A, Cebulla R, Schulte-Eistrup S, Morshuis M, Röfe D, et al. Global gene expression analysis in nonfailing and failing myocardium pre- and postpulsatile and nonpulsatile ventricular assist device support. *Physiol Genomics*. 2010;42(3):397–405.
41. Pilbrow AP, Folkersen L, Pearson JF, Brown CM, McNoe L, Wang NM, Sweet WE, Tang WW, Black MA, Troughton RW, et al. The chromosome 9p21.3 coronary heart disease risk allele is associated with altered gene expression in normal heart and vascular tissues. *PLoS ONE*. 2012;7(6):39574.
42. Molina-Navarro MM, Roselló-Lleti E, Ortega A, Tarazón E, Otero M, Martínez-Dolz L, Lago F, González-Juanatey JR, España F, García-Pavía P, et al. Differential gene expression of cardiac ion channels in human dilated cardiomyopathy. *PLoS ONE*. 2013;8(12):79792.
43. Liu Y, Morley M, Brandimarto J, Hannehalli S, Hu Y, Ashley EA, Tang WW, Moravec CS, Margulies KB, Cappola TP, et al. RNA-seq identifies novel myocardial gene expression signatures of heart failure. *Genomics*. 2015;105(2):83–9.
44. Deniz GC, Durdu S, Özdağ, Akar RA. Gene expression data from human left and right atrial tissues in patients with degenerative MR in SR and AFib. *Stem Cell Inst NCBI Gene Expr Omnibus*. 2019.
45. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010;26(19):2363–7.

46. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
47. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci*. 2003;100(16):9440–5.
48. Zhao B, Erwin A, Xue B. How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances. *Genomics*. 2018;110(1):67–73.
49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol)*. 1995;57(1):289–300.
50. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733.
51. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29–39.
52. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*. 2007;8(1):118–27.
53. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–3.
54. Kuhn M, et al. Building predictive models in r using the caret package. *J Stat Softw*. 2008;28(5):1–26.
55. Cortes C, Mohri M. AUC optimization vs. error rate minimization. In: *Advances in Neural Information Processing Systems*. Cambridge: MIT Press; 2004. p. 313–20.
56. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The binormal assumption on precision-recall curves. In: 2010 20th International Conference on Pattern Recognition. New Jersey: IEEE; 2010. p. 4263–6.
57. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. 2018;47(D1):419–26.
58. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2018;47(D1):607–13.
59. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2019;48(D1):845–55.
60. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal*. 2008;52(4):2249–60.
61. Mendis S, Puska P, Norrving B, Organization WH, et al. *Global Atlas on Cardiovascular Disease Prevention and Control*. Geneva: World Health Organization; 2011.
62. Steinhart Z, Angers S. Wnt signaling in development and tissue homeostasis. *Development*. 2018;145(11):146589.
63. Foulquier S, Daskalopoulos EP, Lluri G, Hermans KC, Deb A, Blankesteijn WM. Wnt signaling in cardiac and vascular disease. *Pharmacol Rev*. 2018;70(1):68–141.
64. Huang A, Huang Y. Role of sfrps in cardiovascular disease. *Ther Adv Chronic Dis*. 2020;11:2040622320901990.
65. Chen L, Deng H, Cui H, Fang J, Zuo Z, Deng J, Li Y, Wang X, Zhao L. Inflammatory responses and inflammation-associated diseases in organs. *Oncotarget*. 2018;9(6):7204.
66. Frangogiannis NG. Regulation of the inflammatory response in cardiac repair. *Circ Res*. 2012;110(1):159–73.
67. Sattler S, Fairchild P, Watt FM, Rosenthal N, Harding SE. The adaptive immune response to cardiac injury—the true roadblock to effective regenerative therapies? *NPJ Regen Med*. 2017;2(1):19.
68. Dhingra R, Vasan RS. Biomarkers in cardiovascular disease: Statistical assessment and section on key novel heart failure biomarkers. *Trends Cardiovasc Med*. 2017;27(2):123–33.
69. Cannone V, Cabassi A, Volpi R, Burnett JC. Atrial natriuretic peptide: A molecular target of novel therapeutic approaches to cardio-metabolic disease. *Int J Mol Sci*. 2019;20(13):3265.
70. Idzikowska K, Zielińska M. Midregional pro-atrial natriuretic peptide, an important member of the natriuretic peptide family: potential role in diagnosis and prognosis of cardiovascular disease. *J Int Med Res*. 2018;46(8):3017–29.
71. Song W, Wang H, Wu Q. Atrial natriuretic peptide in cardiovascular biology and disease (nppa). *Gene*. 2015;569(1):1–6.
72. Kääb S, Barth AS, Margerie D, Dugas M, Gebauer M, Zwermann L, Merk S, Pfeufer A, Steinmeyer K, Bleich M, et al. Global gene expression in human myocardium—oligonucleotide microarray analysis of regional diversity and transcriptional regulation in heart failure. *J Mol Med*. 2004;82(5):308–16.
73. Newman MS, Nguyen T, Watson MJ, Hull RW, Yu H-G. Transcriptome profiling reveals novel bmi-and sex-specific gene expression signatures for human cardiac hypertrophy. *Physiol Genomics*. 2017;49(7):355–67.
74. Chakraborty S, Cheek J, Sakthivel B, Aronow BJ, Yutzey KE. Shared gene expression profiles in developing heart valves and osteoblast progenitor cells. *Physiol Genomics*. 2008;35(1):75–85.
75. Tan HT, Lim TK, Richards AM, Kofidis T, Teoh KL-K, Ling LH, Chung MC. Unravelling the proteome of degenerative human mitral valves. *Proteomics*. 2015;15(17):2934–44.
76. Wang H-B, Huang R, Yang K, Xu M, Fan D, Liu M-X, Huang S-H, Liu L-B, Wu H-M, Tang Q-Z. Identification of differentially expressed genes and preliminary validations in cardiac pathological remodeling induced by transverse aortic constriction. *Int J Mol Med*. 2019;44(4):1447–61.
77. Andenaes K, Lunde IG, Mohammadzadeh N, Dahl CP, Aronsen JM, Strand ME, Palmero S, Sjaastad I, Christensen G, Engebretsen KV, et al. The extracellular matrix proteoglycan fibromodulin is upregulated in clinical and experimental heart failure and affects cardiac remodeling. *PLoS ONE*. 2018;13(7):0201422.
78. Engebretsen KV, Lunde IG, Strand ME, Waehre A, Sjaastad I, Marstein HS, Skrbic B, Dahl CP, Askevold ET, Christensen G, et al. Lumican is increased in experimental and clinical heart failure, and its production by cardiac fibroblasts is induced by mechanical and proinflammatory stimuli. *FEBS J*. 2013;280(10):2382–98.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.