Identification of dilated cardiomyopathy signature genes through gene expression and network data integration

Camargo-Rodriguez, Anyela Velentine; Azuaje, Francisco

Published in:
Genomics

DOI:
10.1016/j.ygeno.2008.05.007

Publication date:
2008

Citation for published version (APA):
Camargo-Rodriguez, A. V., & Azuaje, F. (2008). Identification of dilated cardiomyopathy signature genes through gene expression and network data integration. *Genomics*, *92*(6), 404-413. https://doi.org/10.1016/j.ygeno.2008.05.007

Genomics 92 (2008) 404–413

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification of dilated cardiomyopathy signature genes through gene expression and network data integration

Anyela Camargo [a], Francisco Azuaje [b,*]

[a] School of Computing and Mathematics, University of Ulster at Jordanstown, Shore Road, Newtownabbey, County Antrim BT37 0QB, Northern Ireland, UK
[b] Laboratory of Cardiovascular Research, CRP-Santé, L-1445, Luxembourg

ARTICLE INFO

Article history:
Received 13 February 2008
Accepted 8 May 2008
Available online 1 July 2008

Keywords:
Biodata mining and integration
Heart failure
Dilated cardiomyopathy
Protein networks
Diagnostic systems
Gene expression data

ABSTRACT

Dilated cardiomyopathy (DCM) is a leading cause of heart failure (HF) and cardiac transplantations in Western countries. Single-source gene expression analysis studies have identified potential disease biomarkers and drug targets. However, because of the diversity of experimental settings and relative lack of data, concerns have been raised about the robustness and reproducibility of the predictions. This study presents the identification of robust and reproducible DCM signature genes based on the integration of several independent data sets and functional network information. Gene expression profiles from three public data sets containing DCM and non-DCM samples were integrated and analyzed, which allowed the implementation of clinical diagnostic models. Differentially expressed genes were evaluated in the context of a global protein–protein interaction network, constructed as part of this study. Potential associations with HF were identified by searching the scientific literature. From these analyses, classification models were built and their effectiveness in differentiating between DCM and non-DCM samples was estimated. The main outcome was a set of integrated, potentially novel DCM signature genes, which may be used as reliable disease biomarkers. An empirical demonstration of the power of the integrative classification models against single-source models is also given.

Dilated cardiomyopathy (DCM) is a leading cause of heart failure (HF) and cardiac transplantations in Western countries [1,2]. In DCM, the heart muscle becomes enlarged, making the pumping of blood less efficient to vital organs. Gene expression studies have offered insights into the etiology of this disease [2–5]. For example, Barth et al. [2] analyzed gene expression patterns related to DCM and identified specific gene regulatory relationships relevant to this disease condition. King et al. [3] analyzed the gene expression of samples with distinctive HF histological grading and identified genes, as well as specific biological pathways, associated with the disease. Wittchen et al. [4] analyzed gene expression profiles of human inflammatory cardiomyopathy and suggested novel therapeutic targets. Camargo and Azuaje [6] integrated gene expression analysis with a protein–protein interaction (PPI) network in human heart failure to investigate biological responses in experimental DCM. However, it remains uncertain whether the integration of those independent data sets may improve systems-level knowledge and support potential clinical applications. Moreover, there are concerns in connection to the reproducibility of prediction results based on single-source analyses. Furthermore, there is a need to exploit available public data to aid in the translation of biomedical research from bench to bedside.

The main hypothesis of this study is that the integration of publicly available data and other information sources may support the identification of common, potentially relevant (and reproducible) disease biomarkers [7,8]. Barth et al. [2] addressed a similar concern. However, in their study data were not integrated but analyzed separately. Zhan et al. [9] integrated three data sets from breast cancer. The results of the latter study led to the conclusion that data integration increases predictive analysis and reduces the number of false positives.

To probe our hypothesis, the following sequence of analyses was carried out: (a) three published microarray data sets, containing human DCM and non-DCM samples, were integrated; (b) significant expression patterns were analyzed; and (c) potential candidate disease signature genes were identified. In parallel, for comparison purposes, known associations with heart failure (KHF) were retrieved from public databases [9]. The query constraint was based on the keywords previously used by King et al. [3] on studies of heart failure. Candidate genes and KHF were analyzed in the context of a global human heart failure PPI network to discover responses not identified by the expression analysis alone. The potential novelty and relevance of the key genes identified from the previous analyses were estimated by large-scale text-mining analyses. In addition, automated

Abbreviations: DCM, dilated cardiomyopathy; HF, heart failure; PPI, protein–protein interaction; SDE, significantly differentially expressed; CP, class predictor; PAM, prediction analysis of microarray; IP, interacting proteins; HPRD, human protein reference database; SVM, support vector machine; IEA, inferred automatically; LOO, leave-one-out; RBF, radial basis function; SAM, significance analysis of microarray.

* Corresponding author.
E-mail address: fj.azuaje@ieee.org (F. Azuaje).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.05.007

classification models were built to distinguish between DCM and non-DCM samples. Statistical comparisons between integrated vs single-source classification models were implemented. Finally, novel disease signature genes were selected by evaluating outputs produced by each of the procedures previously mentioned. Based on an integrative ranking of the preceding outcomes, a list of potentially novel signature genes is presented, together with specific biological processes that appear to be significantly altered during DCM development.

Results

This study identified novel DCM signature genes by performing a twofold procedure. First, candidate genes were selected from the expression profile analysis of three microarray data sets from DCM and non-DCM patients. Second, from this list, potentially significant genes were selected by searching the scientific literature and by performing data-mining analyses.

Three heterogeneous, lab-independent data sets (from now on referred to as D1, D2, and D3) with samples from DCM and non-DCM patients were standardized and normalized. Gene expression analyses were carried out on individual data sets and on integrated data sets. The latter sets consisted of the pairwise combination of the original data sets. Significance analysis of microarray (SAM) [11] was applied to the genes shared by all the data sets (i.e., 5651 in total). Prediction analysis of microarray (PAM) [12] analyzed pairwise combinations of D1, D2, and D3 data sets and identified class predictor genes. These analyses produced a list of candidate genes that were used to build different classification models (see Methods). Similar analyses were also performed on each data set independently (single-source analyses).

Following the selection of candidate genes, a PPI network was assembled by including validated gene (or protein)–disease and protein–protein associations, from Entrez and the Human Protein Reference Database (HPRD) [13]. The resulting network represents a knowledge base of HF-relevant interactions. In addition, biological pathways associated with these genes were mapped onto the network. A color labeling scheme was used to distinguish between the different types of proteins and biological pathways that each node represented in the network. In addition, nodes were classified according to the degree of connectivity (see Methods). The network analysis allowed us to evaluate significant quantitative relationships between network connectivity and significantly differentially expressed patterns.

Potentially significant and novel genes were selected by searching the scientific literature. Analyses in the context of Gene Ontology (GO) [14] were also implemented to identify significant biological processes altered in DCM (see Methods).

Finally, based on a support vector machine (SVM) [15], data classification models were assessed. This allowed us to demonstrate the power of integrative classifiers, as well as the identification of optimal combinations of genes for distinguishing between DCM and non-DCM samples. These tasks provided the basis for the definition of novel DCM signature genes for potential diagnostic or drug target prediction purposes.

Gene expression analysis

Gene expression analysis of DCM and non-DCM samples, from the combination of the D1 and D2 data sets, identified 444 significantly differentially expressed (SDE) genes; 207 were up-regulated and 237 were down-regulated in DCM. From the combination of the D1 and D3 data sets, 408 SDE genes were identified; 199 were up-regulated and 209 were down-regulated in DCM. From the combination of the D2 and D3 data sets, 668 SDE genes were identified; 255 were up-regulated and 413 were down-regulated in DCM. In the up-regulated and down-regulated categories 53 and 32 genes, respectively, overlapped in the three analyses. In addition, expression analysis of single sources reported that 22 genes were SDE in D1, 1129 genes were SDE in D2, and 16 genes were SDE in D3. Only genes HMGN2 and ODC1 were shared by the single-source analyses.

Following gene expression analysis, the PAM technique was used to identify significant class predictor (CP) genes, i.e., genes whose expression profile vector showed remarkable discrimination capability between DCM and non-DCM samples. After cross-validation, PAM identified 73 CP genes when the D1 and D2 data sets were combined. Similarly, in the integrated analysis of the D1 and D3 data sets, 53 CP genes were identified. In the integration of the D2 and D3 data sets, 55 CP genes were identified. When results were compared, it was observed that the genes HMGN2 (high-mobility group nucleosomal binding domain 2), HTRA1 (HtrA serine peptidase 1), MDFIC (MyoD family inhibitor domain-containing), ODC1 (ornithine decarboxylase 1), and SEC31A (SEC31 homolog A) were shared by the outcomes of the three analyses. PAM of single sources identified 47 CP genes from D1, whose classification performance was 100%; 3 CP genes from D2, whose classification performance was 96%; and 36 CP genes from D3, whose classification performance was 96%. These single-source analyses did not report common CP genes. Regarding the HMGN2 and ODC1 genes, the expression analysis performed above showed that they were also up-regulated in DCM.

Once CP genes were identified, their prediction strength was statistically validated further by using the corresponding data set not included in the integrative PAM as an independent evaluation set. Table 1 shows improvement in classification performance whenever the expression vector of CP genes was used as model input only. When the classifier was built based on the pairwise combination of the D1 and D3 data sets, the overall classification accuracy (non-DCM vs DCM) on the D2 data (used as an independent evaluation) was higher than when all the genes were involved (89% vs 64%). Regarding the evaluation of a single source, a similar evaluation procedure was performed, i.e., the classifier was composed of one data set and the remaining data sets were used for independent evaluation, one at a time. Results of these analyses showed that, when all the genes were used as model inputs, the overall classification accuracy (non-DCM vs DCM) was below 64% on the corresponding data set used for independent evaluation. When CP genes were used as model inputs only, the results differed as follows. When the classifier was built based on the D1 or D3 data sets, classification accuracy on the data sets D3 and D1, respectively, was above 93%. When the classifier was built based on D2, classification accuracy on the corresponding evaluation set, D1 or D3, was below 70%.

Gene Ontology analysis

The list of common SDE genes, which also contain CP genes, was analyzed in the context of GO. This analysis identified the following GO biological processes as significantly overrepresented: phosphorylation, protein amino acid phosphorylation, positive regulation of epithelial cell proliferation, epithelial cell proliferation, and regulation of epithelial cell proliferation. Other studies have identified the alteration of these processes whenever there is an ongoing risk of

Table 1
Classification accuracy from independent evaluation when the expression vectors of all genes or of CP genes only were used as model inputs

Learning set	Evaluation set	% Classification accuracy	
		All genes	CP
D1–D2	D3	81	89
D1–D3	D2	64	89
D2–D3	D1	100	100

Classifiers were built based on pairwise combination of the D1, D2, and D3 data sets. The data set not involved in the combination was used for independent evaluation.

heart failure [17,18]. GO analysis of CP genes reported no significantly overrepresented biological processes.

PPI network analysis

The PPI network (Fig. 1A) consisted of nodes representing proteins, their interaction partners, and the biological pathways associated with their encoding genes. Some of the nodes represented proteins encoded by significantly differentially expressed genes obtained from the previous expression pattern analyses. In total, the network includes 2227 protein–protein interactions and 1515 nodes corresponding to proteins. In the above section on gene expression analysis of pairwise combinations of data sets, we reported the overlap of 85 SDE genes. Of this total, 53 genes were up-regulated in DCM and 32 genes were down-regulated in DCM. However, when they were mapped onto the network, only 45 genes (27 up-regulated and 18 down-regulated genes in DCM) could be represented by a node. This is because the proteins encoded by some of those genes were not associated with at least one interacting protein partner. In the case of CP genes, only 3 of them were represented by a node in the network: HMGN2, HTRA1, and ODC1. The network also contained 71 nodes representing proteins encoded by KHF genes. How the KHF genes were obtained is summarized under Methods.

According to the node categorization scheme described under Methods, 2.1% of genes were represented by network hubs or superhubs (only seven hubs represented 7 significantly differentially expressed genes). In contrast, 97.8% of the genes were represented by either network peripheral-A or network peripheral-B nodes (30 peripheral-A and 9 peripheral-B nodes represented 39 significantly differentially expressed genes). Details are shown in Table 2.

Examples of potentially relevant associations are described as follows. DYNLL1 (dynein, light chain, LC8-type 1) is a protein represented by a network hub (Fig. 1B), and TERF1 (telomeric repeat binding factor) is represented by a network peripheral-A. These two proteins interact through ZHX1 (zinc fingers and homeoboxes 1). According to gene expression analysis, both DYNLL1 and TERF1 encode significantly differentially expressed genes (up-regulated in DCM). Also, PICK1 (protein interacting with PRKCA1) represents a network hub (Fig. 1C). Currently, PICK1 is not known to be associated with HF. However, it intervenes in the protein amino acid phosphorylation and protein kinase C activation biological processes [16], which are involved in the progression of heart failure [3,4].

Only 33 KHF genes had a corresponding transcript in the gene expression data set, and only 1 of these genes, CCL2, was significantly differentially expressed (down-regulated in DCM). In addition, 656 genes that encoded other proteins' interaction partners in the network had a corresponding transcript in the gene expression data set and were not significantly differentially expressed.

Network connectivity versus significant gene expression patterns

This section of the study integrated gene expression analysis results, in the form of *di* values, obtained from the SAM, with the PPI network. The aim was to describe potential significant relationships between network connectivity and gene expression patterns (as described in Methods). Results were as follows: when the focus was on significantly differentially expressed genes only, we found that genes represented by network superhubs and hubs tend to have lower range of *di* values. In Fig. 2A genes with those characteristics are shown on the far right of the plot. Furthermore,

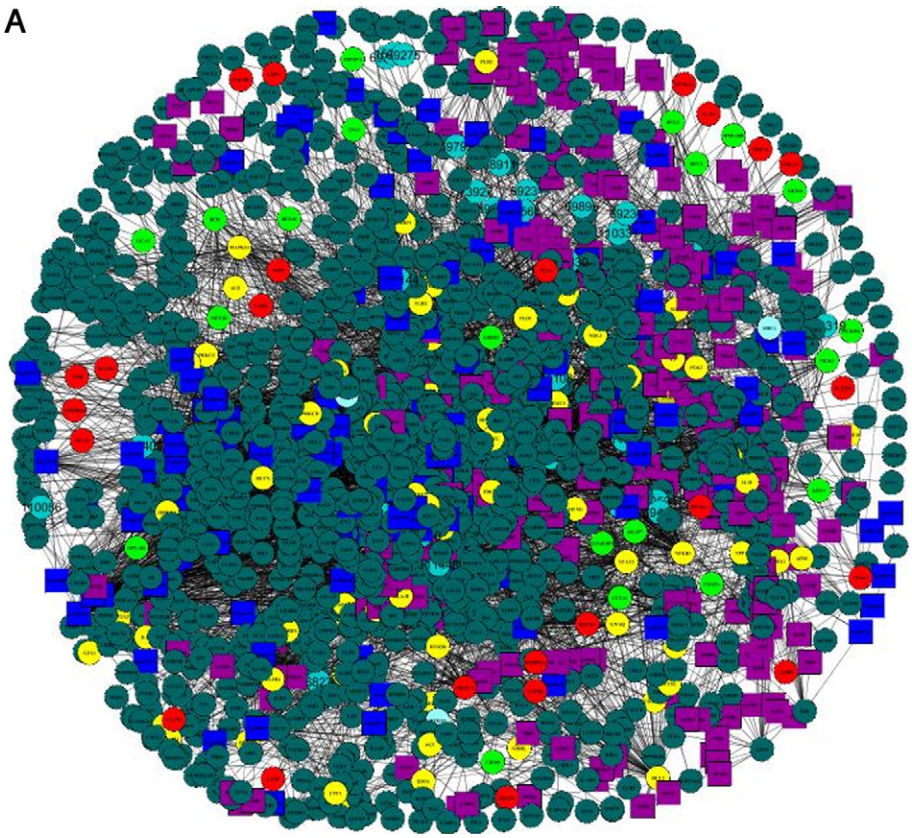

Fig. 1. PPIs corresponding to (A) the global network, (B) DYNLL1's network, and (C) PICK1's network. All PPIs were retrieved from the HPRD. DCM up-regulated genes are represented by red nodes. Down-regulated genes are represented by green nodes. KHF genes are represented by nodes in yellow. Other genes encoding interacting partner proteins are represented by nodes in purple, if they were present across D1, D2, and D3 data sets, and in teal, if they were not common across D1, D2, and D3 data sets.

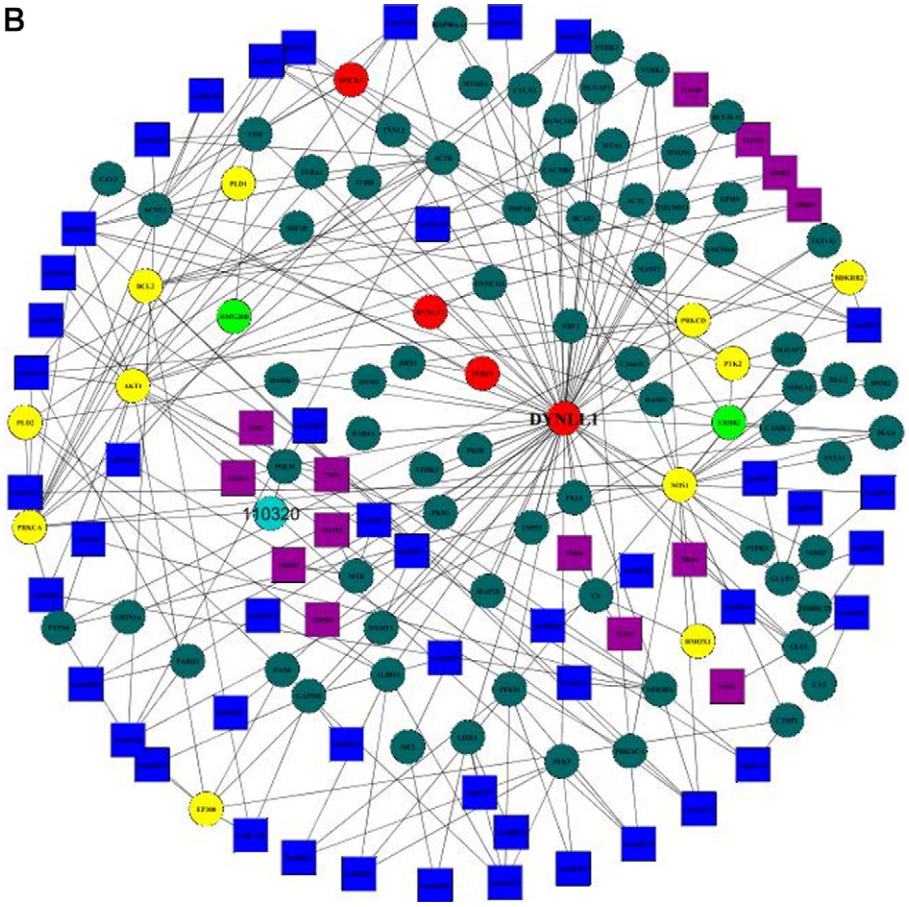

Fig. 1 (*continued*).

genes represented by network peripherals-A and -B tend to have higher range of *di* values than hubs and superhubs. When proteins encoded by non-significantly differentially expressed genes were assessed, we found similar results (Fig. 2B). However, neither superhubs nor hubs were found. In fact, only two genes (EGFR and SRC) were represented by nodes with 11 and 12 interactions, respectively. The other genes were represented by nodes that had fewer than 10 interactions. Interestingly, EGFR is involved in more than 10 biological pathways; some of them are the MAPK signaling pathway and hemostasis, which play important roles in the development of DCM [19].

Extracting significant associations based on large-scale literature mining

Multiple searches on the scientific literature were performed to identify known associations with DCM or HF. The SDEs, including CP genes, were part of the query list, together with those found to be relevant in the PPI network analysis. Therefore, each of these genes' IDs was used to query the PubMed information source, through the Agilent literature search tool [20]. The search was constrained by the keywords "heart failure," "dilated cardiomyopathy," and "cardiovascular diseases." According to these parameters, Table 3 shows a list of gene IDs that, to date, have not been associated with the disease target. Additional details about the genes are also provided, such as IPs, node hierarchy, and average "*d*" value. The genes DYNLL1, TERF1, and PICK1 are again highlighted, as well as the CP genes HMGN2, HTRA1, and ODC1. These three genes have also been associated with tumor-related disorders [16]. In future research, it may be interesting to investigate why these genes show evidence of overregulation in samples obtained from patients with a DCM condition.

Evaluation of diagnostic models

This part of the study investigated the optimal combination of gene expression patterns to differentiate between DCM and non-DCM samples. As described in Methods, the SVM technique was chosen to construct these classifiers and study optimal combinations of genes as model inputs. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve was used to estimate class prediction performance. Different comparative schemes for testing and independent evaluations were implemented. In the pairwise combination of individual data sets for building and cross-validating classifiers (i.e., D1–D2, D1–D3, and D2–D3), the best class prediction performance was obtained when CP genes were used as the inputs to the classifier only. Using leave-one-out (LOO) cross-validation, estimated AUCs for each pairwise combination were 0.64, 0.48, and 0.57 (Table 4). In contrast, when other classification models (other groups of genes, such as SDE, were used as the inputs to the classifier) were assessed, prediction performance declined. For instance, when the genes listed in Table 3, including three CP genes, were used as the inputs to the classifier, prediction performance declined. AUCs for each pairwise combination were 0.51, 0.23, and 0.28. When all SDE genes were used as inputs, estimated (LOO) AUCs per pairwise combination of data sets were 0.46, 0.11, and 0.17. We also found that in general, the classification accuracy of non-DCM and DCM samples was above 90%. Classification models were also independently evaluated on the data set left out from each data integration experiment. Table 4 shows results from these analyses, which in general confirm the predictive ability of the CP genes. For example, without exception, when CP genes were the inputs to the classifiers, estimated (LOO) AUCs were above 0.99. Surprisingly, when KHF genes were the inputs to the classifier, estimated (LOO) AUCs were below 0.8; recall that KHF genes

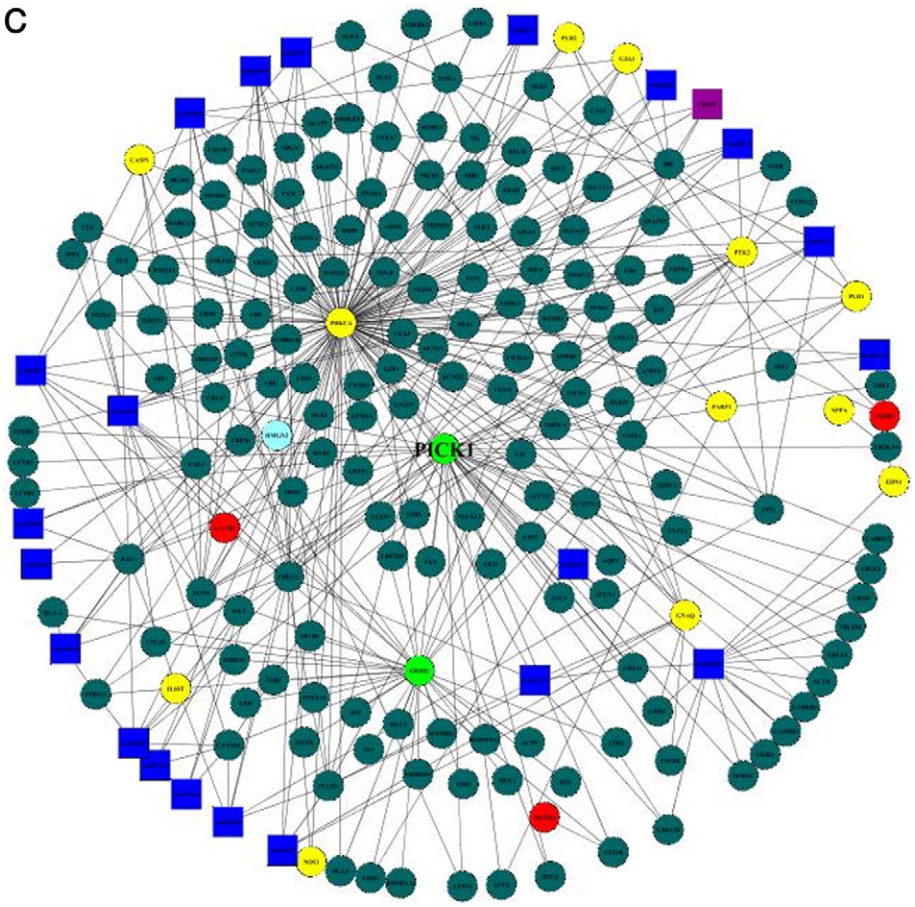

Fig. 1 (*continued*).

have been previously associated with HF. Fig. 3A shows ROCs corresponding to the evaluation of the pairwise integration of the D1 and D2 data sets (LOO cross-validation results). The ROC in Fig. 3B corresponds to the evaluation of the integrated classification model D1–D3 on an independent evaluation set, D3.

Single-source classifiers were also evaluated independently (i.e., models trained and tested using D1, D2, and D3 independently). Table 5 reviews results obtained when expression profiles of SDE, CP, or KHF genes from each data set were used as inputs to the SVM classifier. In general, class prediction performance was weaker than that achieved by the integrated data sets. For example, Table 5 and Fig. 3B show ROCs and AUCs respectively, for each classification model, when the classifier was built on the D2 data set (corresponding evaluation sets were D1 and D2). Low performance may be explained by the relative lack of data within the single-source models in comparison to the integrated models. For example, the relation of samples between non-DCM and DCM in D3 was 25% (7) to 75% (20), respectively. Estimated (LOO) AUCs were lower than in the other two groups. Conversely, the relation of samples between non-DCM and DCM in D3 was 54% (15) to 13% (20), respectively. Estimated (LOO) AUCs were higher than in the other two groups. Note that according to the results obtained from PAM, on the analysis of a single-source data set, class prediction performance of CP genes was on average above 96%. This may seem to contradict the results obtained with the SVM.

As in the previous integrative analysis, when model evaluation was based on a single source, estimated (LOO) AUCs improved (Table 5). However, here, some classification models based on CP genes could not be evaluated (indicated by the word NAN). This happened because CP genes were not common among learning and evaluation sets. In the case of classification models based on SDE genes, the average estimated (LOO) AUC of the corresponding independent evaluations was above 0.9. As for KHF genes, the average estimated (LOO) AUC of the corresponding independent evaluations was below 0.8. These results confirmed that estimated AUCs improved whenever the number of samples in the learning set increased. In other words, single-data-source analysis may produce several false positives owing to evaluation methodology bias and overall lack of data. Results also suggest that measures such as classification accuracy may convey unreliable results and that other measures that estimate false discovery rates (FDR) and AUCs should be favored. Therefore, it is usual to find that genes regarded as significant biological markers may be restricted or biased to a particular microarray data set that was hybridized under special conditions. This problem was highlighted early on by Ancona et al. [21], who analyzed cancer-related microarray

Table 2
Summary of node populations according to connectivity

Hierarchy	SDE	NSDE	KHF	Total
Superhubs	0	0	3	3
Hubs	7	0	22	29
Peripheral-A	30	412	43	485
Peripheral-B	9	986	3	998
Total	46	1398	71	1515

The number of nodes within each category present in the interaction network is shown. SDE, significantly differentially expressed genes; NSDE, non-significantly differentially expressed genes; KHF, known HF genes retrieved from the Entrez database [9]. PPIs were obtained from the HPRD [13].

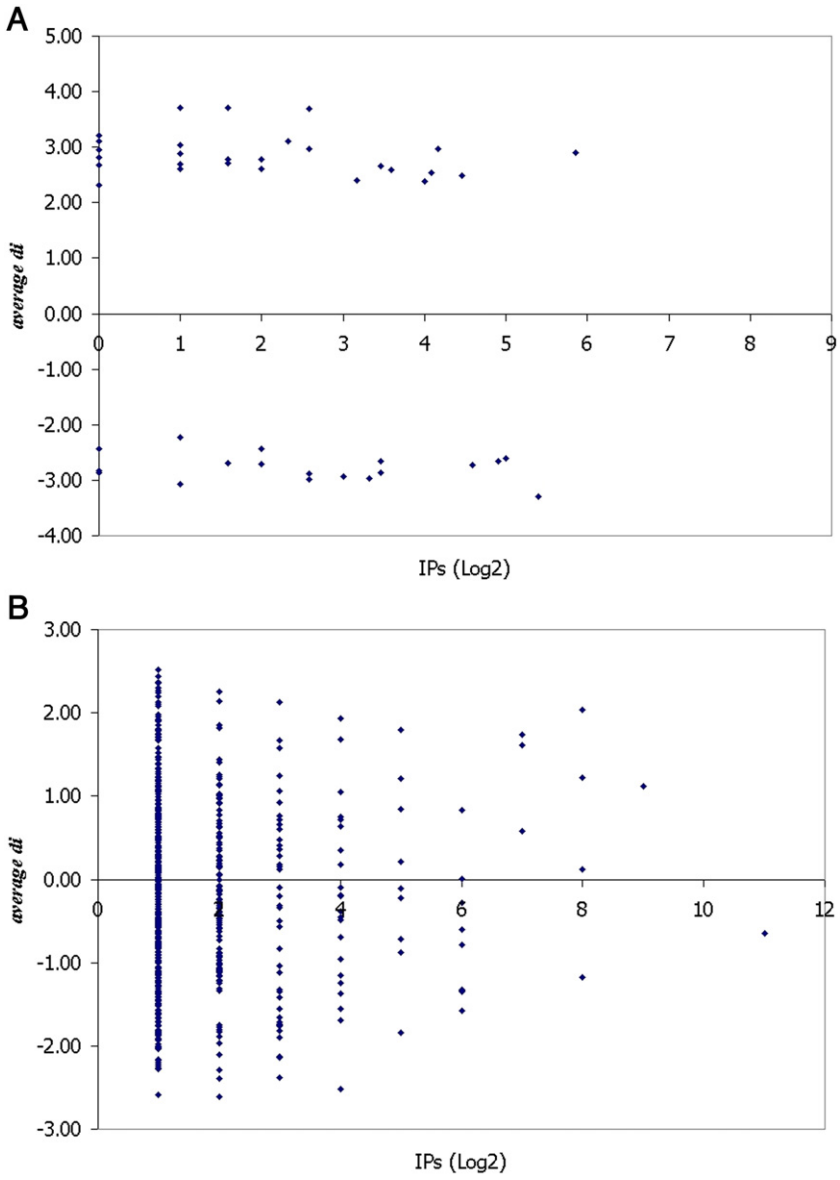

Fig. 2. Plots of *t* statistic (*di*) representing the score for gene *i* vs the number of interacting partners (IPs), expressed on the $\log_2$ scale, associated with the protein encoded by gene *i*. The plots show the average of *di* values obtained in each of the three gene expression analyses. (A) Significantly differentially expressed genes. (B) Non-significantly differentially expressed genes.

data and found that the expression profile of previously reported CP genes could not be replicated in their own study. This underscores the need for integrating several data sources for potentially rendering more accurate and less biased results.

Fig. 3C shows ROCs corresponding to the evaluation of models based on a single source. The ROCs in Fig. 3D were obtained when the classifier was built using the D1 data set, and for independent evaluation the sets D2 and D3 were used.

Potential most significant biomarkers

The preceding evaluation procedure helped to identify a list of candidate genes that could be seen as possible disease signatures. The overall results suggested that the following genes could be novel DCM signature genes, which were also significantly differentially expressed in the integrative analysis: PICK1, DYNLL1, ODC1, HTRA1, and HMGN2. ODC1, HTRA1, and HMGN2 were identified as class predictor genes, which were the inputs to the most successful disease classification model. However, the proteins these three genes encode were associated with a few IPs. Although not selected as CP genes, PICK1 and DYNLL1 encode proteins associated with several IPs. Finally, the search performed on public data sources suggested that none of these five genes has been previously associated with heart failure. Other genes that are suggested as potentially interesting for future experimental and computational analyses are MORF4L2, MEF2D, TMEM66, USP11, and TRIP12. These genes were all significantly differentially expressed. In addition, searches in the scientific literature suggested that none of these five genes has been previously associated with heart failure.

Discussion

Heart failure, one of the main causes of morbidity and mortality in the world [1], is a polygenic disease whose etiology stems from complex genetic, environmental, and lifestyle factors. Various research efforts have aimed to dissect the molecular mechanisms of this disease [7,8]. To date, a number of disease signature genes have been reported as the result of (single-source) gene expression analyses [2–6,8], and a fair amount of microarray expression data sets have been

Table 3
Potential novel gene–HF associations

Gene	IPs	Hierarchy	Average *d*
DYNLL1	5.86	Hub	2.8998
PICK1	5.00	Hub	−2.6123
COPB1	4.09	P-A	2.5357
TERF1	4.00	P-A	2.3909
MEF2D	3.32	P-A	−2.9648
DYNLT1	3.17	P-A	2.3975
TFE3	3.00	P-A	−2.9414
MORF4L2	2.58	P-A	2.9716
NR2F6	2.58	P-A	−2.8883
HMGN2[a]	2.58	P-A	3.6840
TRIP12	2.32	P-A	3.1072
MATR3	2.00	P-A	2.7881
HMG20B	2.00	P-A	−2.4366
ZBTB7A	2.00	P-A	−2.7172
HTRA1[a]	1.58	P-A	3.7141
PUM1	1.00	P-A	2.6043
STK38L	1.00	P-A	2.8791
USP11	1.00	P-A	3.0334
ODC1[a]	1.00	P-A	3.7095
GOLGA7	0.00	P-B	2.3191
MKRN1	0.00	P-B	2.6822
TMEM66	0.00	P-B	3.2034
BCKDK	0.00	P-B	−2.8296
EIF4EBP2	0.00	P-B	−2.4417
GCAT	0.00	P-B	−2.8648

IPs ($\log_2$). Node hierarchy according to the PPI network (P-A, peripheral-A; P-B, peripheral-B). The average of (*d*) values obtained in each of the three gene expression analyses performed is given.
[a] CP genes.

uploaded into public databases. Despite these efforts, results from each study may be biased and may vary widely due to the independent experimental conditions being defined. Therefore, a key challenge is to find ways to integrate such information to facilitate the discovery of disease-specific knowledge and more powerful predictive targets. In a recent paper, Zhan et al. [10] held that the integration of microarray data gives more analytical power and reduces the false discovery rate. As a result, the probability of identifying more biologically meaningful patterns is high. This study, which focused on DCM, integrated independent microarray data sets and other information sources to add power to the data-mining techniques implemented. The methodology aimed at the identification of potentially novel and powerful DCM signature genes.

The first gene expression analysis phase of this study identified a total of 85 SDE genes, common to all the single-source analyses. Two of those genes, ODC1 and HMGN2, were also previously identified by Barth et al. [2] as significantly differentially expressed in the analysis of independent data sets. However, a further analysis carried out by them found that HMGN2 could not be regarded as a class predictor gene. The second stage of our study selected CP genes, i.e., those SDE genes whose expression vector showed significant differentiation between DCM and non-DCM samples based on PAM. This analysis identified 5 genes, HMGN2, ODC1, HTRA1, MDFIC, and SEC31A, whose expression patterns showed significant differences between the two experimental conditions being evaluated (non-DCM and DCM). Results were supported by further testing procedures. Using a network-based analysis strategy that we published elsewhere, we found additional relevant biomarkers when we assessed microarray data in the context of a global HF PPI network [6]. Also we found that some of our candidate CP genes encoded proteins associated with no interacting partners (i.e., MDFIC and SEC31A). Similarly, the remaining CP genes encoded proteins associated with just a few interacting partners (network peripheral-B). Also, we found non-CP genes encoding proteins associated with several other interacting partners (e.g., DYNLL1, ERBB2, and PICK1), i.e., network hubs. We confirmed the potential novelty of some of these biomarkers, such as DYNLL1 and PICK1, when querying public information sources to identify known associations with heart failure or DCM. Various classification models were also implemented, and their classification performance was estimated. Results led to the conclusion that whenever CP genes were the inputs to the classifier, a classifier with high AUCs was obtained. Moreover, whenever KHF genes were the inputs to the classifier, the estimated AUCs were in general low. Gathering and evaluating all these results led us to select a pool of genes and suggest them as possible novel DCM signature genes. It also confirmed that KHF genes may not be so interesting in other independent studies, because these studies may have used other experimental settings (i.e., type of tissue samples, data-mining approach, and microarray technology used to hybridize RNA). We probed the latter when in the evaluation of models based on a single source, we built the classifier on D3 and used D1 as an independent evaluation set. Because these two data sets were obtained from similar microarray technologies, CP genes overlap. Therefore, learning and evaluation of the classification model based on CP genes were possible. What it is more, the best estimated AUC (0.99) was obtained here.

This study confirms the importance of integrating independent data sets and their subsequent evaluation against other information sources. In this way more meaningful information can be discovered as opposed to that obtained from single-source and traditional analyses. This methodology may lead to the identification of potentially influential genes not identified when single studies were carried out in the first place. For example, the gene HTRA1 was not identified as an SDE by any of the expression analyses performed on each single source. However, this gene was selected as a key biomarker based on the combination of gene expression and network analysis. HTRA1 is a regulator of cell growth and intervenes in the process that regulates the availability of insulin-like growth factors, which has already been associated with HF. In addition, data integration may contribute to lessening the degree of false positives that are bound to be found when single data sets are assessed. Note, for instance, that in the independent evaluation of single sources, the PAM class prediction rate, based on CP genes, was on average above 96%. However, when the same group of CP genes was the input of independent classification models and evaluated on independent data sets, class prediction rate was severely degraded in most cases.

The proposed prediction models were evaluated using model cross-validation, as well as published literature and external knowledge bases. These strategies have become standard in the initial phases of model design and evaluation of biomarkers and diagnostic models. However, in the longer term, the potential clinical relevance of biomarker discovery and new diagnostic models will be truly estimated by

Table 4
AUCs of ROC curves representing class prediction performance of different classification models

No.	Classification model	Learning (AUC)			Independent evaluation (AUC)		
		D1–D2	D1–D3	D2–D3	E: D3 based on D1–D2	E: D2 based on D1–D3	E: D1 based on D2–D3
1	CP	0.64	0.48	0.57	0.99	1.00	1.00
2	KHF	0.51	0.10	0.19	0.79	0.66	0.51
3	SDE	0.46	0.11	0.17	0.93	0.98	1.00
4	SDEnKHF	0.51	0.23	0.28	0.94	0.99	1.00
5	SDEKHF	0.52	0.28	0.38	0.94	0.95	1.00
6	All	0.46	0.11	0.17	0.93	0.98	1.00

The classifier was built on the pairwise combinations of the individual D1, D2, and D3 data sets. Classification performance was evaluated through LOO cross-validation. From those results, ROC curves were plotted and AUCs were estimated. Classification models were also independently evaluated on the data set left out from each data integration experiment (E). Classification models according to the group of genes used as inputs to the classifier: CP, class predictor genes; KHF, known HF genes; SDE, significantly differentially expressed genes; SDEnKHF, significantly differentially expressed + non-KHF genes; SDEKHF, significantly differentially expressed + KHF genes.

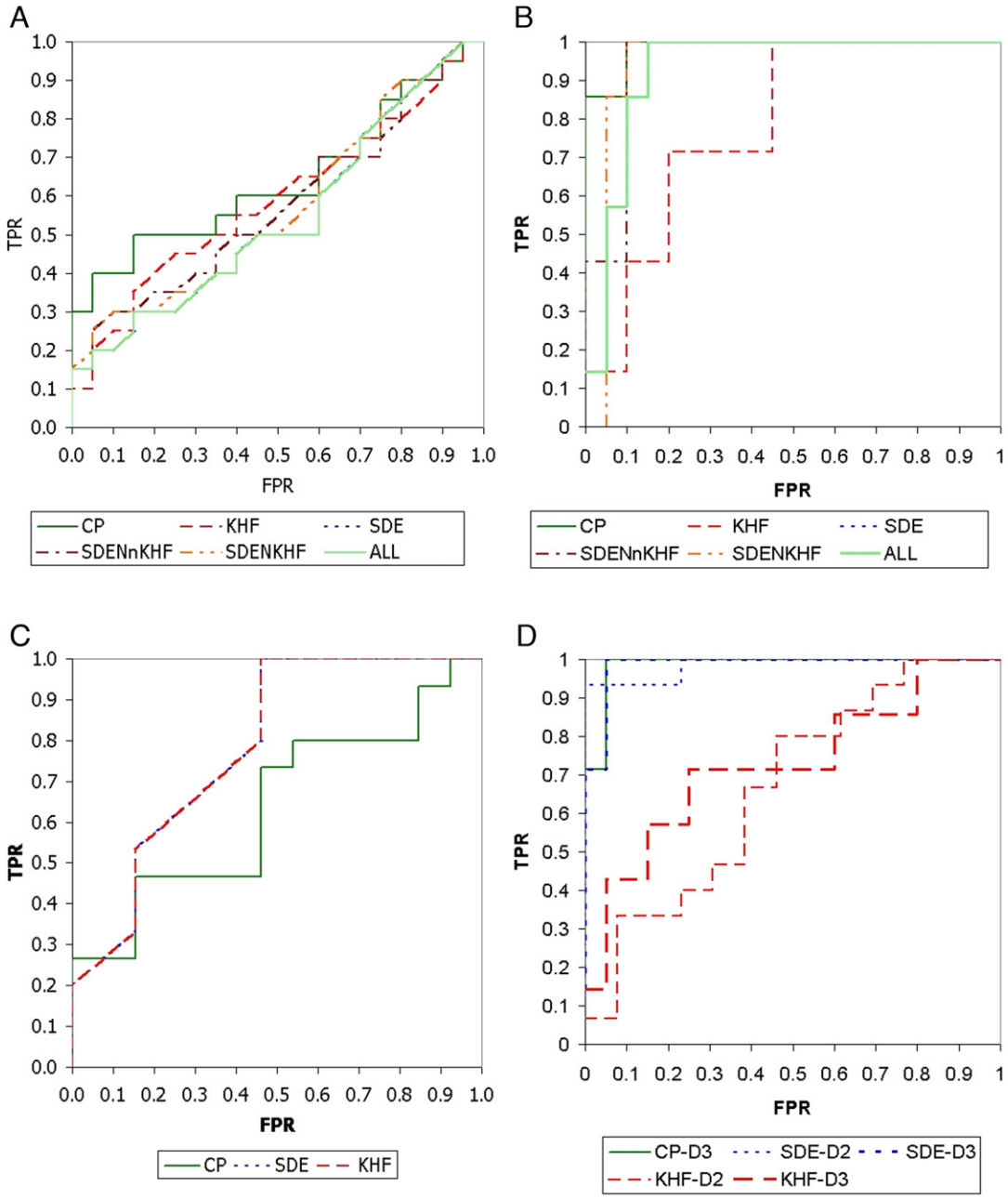

Fig. 3. ROC curves derived from LOO cross-validation and corresponding independent evaluation of (A) the model based on pairwise combination of D1 and D2; (B) the model evaluated on D3; (C) the model based on the single-source D1; and (D) the model evaluated on D2 and D3, independently.

incorporating gene expression data independently generated by other research groups. This paper reports new testable hypotheses, in the form of biomarkers and computational models, which may be independently evaluated. Furthermore, it offers a new prediction model methodology, which may be adapted to other clinical domains.

One aspect to be covered in the future is the analysis of responses on HF-specific biological pathways, on a PPI network. This type of analysis may help us to visualize and interpret interesting stimulus–response activities that take place in HF-associated biological pathways.

Methods

Microarray data analysis

Three microarray data sets generated by independent studies on DCM were obtained from the Gene Expression Omnibus [22], Accession Nos. GDS2205, GDS2206, and GDS1362 (referred as to D1, D2, and D3 herein). D1, oligo array, was composed of 12 samples: 5 and 7 samples were obtained from nonfailing hearts and DCM heart patients, respectively. D2, cDNA array, was composed of 28 samples: 15 and 13 samples were obtained from nonfailing hearts and DCM heart patients, respectively. D3, oligo array, was composed of 37 samples: 7, 20, and 10 samples were obtained from nonfailing hearts, DCM heart, and ischemic cardiomyopathy (ICM) patients, respectively. The latter ICM pool of samples was discarded as the disease condition was not involved in this study. Data sets were originally available in log scale.

Data preprocessing

Probe sets with absent calls in more than 50% of their transcripts were discarded. Transcripts of probe sets corresponding to similar Gene IDs were averaged. Common probe sets among D1, D2, and D3 data sets were selected. In total, 5651 were present. Data sets were normalized per chip and then per gene. Values were transformed using the mean and standard deviation of the row (per gene) or column (per chip).

Gene expression analysis

Differential gene expression was measured by performing SAM [11]. The algorithm computes a *t* statistic (*di*) representing the score of class differentiation for gene *i*. Gene expression differences were considered significant if FDR was <0.05 and fold change

Table 5
AUCs representing class prediction performance of different classification models

No.	Classification model	LOO evaluation of model based on a single source			Independent evaluation					
					Model based on D1 and independently evaluated on:		Model based on D2 and independently evaluated on:		Model based on D3 and independently evaluated on:	
		D1	D2	D3	D2	D3	D1	D3	D2	D3
1	CP	0.37	0.64	0.45	NAN	0.99	NAN	NAN	1.00	NAN
2	KHF	0.40	0.92	0.29	0.67	0.71	0.71	0.39	0.74	0.51
2	SDE	0.48	0.92	0.36	0.98	0.99	1.00	0.99	1.00	1.00

Evaluation was based on a single source, and performance was measured through LOO cross-validation. From those results, ROC curves were plotted and AUCs were estimated. The data sets not included in the construction and cross-validation of each classification model were used as evaluation sets. AUC was estimated. Classification models, according to group of genes used as inputs to the classifier: CP, class predictor genes; KHF, known HF genes; SDE, significantly differentially expressed genes.

was >1.2. As three data sets were involved in the study, pairwise combinations of D1, D2, and D3 were created; DCM and non-DCM samples were separated; and SAM was applied to each integrated data set. As a measure for comparison, this study analyzed the expression profile of single data sources. SAM was also used in this part of the analysis.

PAM [12], a statistical technique for class prediction from gene expression data that uses nearest shrunken centroids, identified class predictor genes, i.e., DCM and non-DCM. For learning purposes, based on 10-fold cross-validation to select optimal threshold, PAM analyzed the expression profile of DCM versus non-DCM samples of the pairwise combinations of D1, D2, and D3 data sets. For evaluation purposes, the data set not involved in each combination was used as the evaluation set. As a measure for comparison, single data sources were analyzed independently. PAM was also applied in this part of the analysis.

A human HF interaction network

A PPI network was assembled by including validated interactions reported for KHF genes, for proteins encoded by genes included in the gene expression data sets, and for the biological pathways associated with them. This network is offered as a public resource of the current status of human HF-relevant interactions (network is provided on request). The list of KHF genes was obtained from the Entrez database [9]. Entrez query was restricted by the same set of keywords used in King et al. [3] (i.e., smooth muscle, endothelial cell, apoptosis, cytokine, and adhesion molecule) and within the context of human HF. PPIs were retrieved from the HPRD [13]. The HUGO nomenclature standard was used to define unique ID identifiers.

The PPI network was assembled using a routine written in JAVA, and its structure was encoded in the SIF format, which can be used by well-known network visualization tools (e.g., Cytoscape). The product of this assembly was a network composed of 1515 nodes and 4452 interactions. The number of interacting partners ranges from 1 to more than 100. A color labeling scheme was used to distinguish between the types of proteins each node represented. Proteins encoded by up- and down-regulated genes (as predicted in the gene expression data) were represented by nodes colored red and green, respectively. Proteins encoded by class predictor genes were represented by nodes colored light blue. Proteins encoded by KHF genes were represented by yellow nodes. Proteins encoded by nonsignificantly differentiated genes were represented by blue nodes. Proteins encoded by genes whose expression pattern was not present in the data set, but which encoded relevant interacting partners in the HF network, were also represented by blue nodes. Kegg and Reactome pathways were represented by blue and purple square nodes, respectively.

Nodes in the network were also classified according to the degree of connectivity, based on a scheme similar to that used in Lu et al. [23]. Superhubs are represented by nodes with connectivity degree greater than 100, hubs refer to nodes with connectivity degree greater than 20 and lower than 100, peripheral-A are nodes with connectivity greater than 2 and lower than 20, and peripheral-B nodes represent proteins with one interacting partner only.

Cytoscape version 2.4 [24] was used for network visualization.

Network analysis

Relationships between PPI's network topology and expression profile patterns were investigated. First analysis focused on genes present in the data sets. Therefore, the average d value of each gene was compared against the degree of connectivity associated with the protein they encoded. The second analysis focused on genes not present in the data sets (but present in the PPI network).

Gene Ontology analysis

Interaction network topology was analyzed in the context of Gene Ontology [14]. Cytoscape-BiNGO [24] was applied to detect significantly overrepresented GO biological processes. Benjamini and Hochberg multiple-test corrections adjusted raw P values at a significance level of <0.05. To increase the level of stringency, GO-IEA terms were discarded. GO annotations with IEA evidence code refer to annotations inferred from sequence-based similarity searches, which have not been reviewed by curators.

Extracting significant associations based on large-scale literature mining

To determine which genes, from a group of candidates, were potentially novel in the area of HF and DCM, the Agilent Literature Search plug-in in Cytoscape [25] was used. Agilent queries multiple text-based search engines, such as PubMed and OMIM, to find document-based associations between each gene and a keyword (i.e., heart failure and DCM). When a query is processed, Agilent implements information retrieval and knowledge extraction and returns the documents (from papers) that match the query. The retrieval is based on comparing the search term, i.e., gene and disease target, against abstracts and keywords.

Evaluation of diagnostic models

This section of the study evaluated classification models based on two different schemes: first, the classifier was built on the pairwise combinations of D1, D2, and D3 data, and the model was evaluated on the set not involved in the combination. The second scheme built the classifier on a single data set, and the model was evaluated on the remaining two data sets, one by one. As for the classifier, the SVM [26] data classification technique was used for this purpose. SVM trained on several classification models used in this study and classification performance, through cross-validation, was estimated. Results led to the selection of the best model. Because of the nonlinearity of the data, radial basis function (RBF) was used as the kernel. The LOO cross-validation technique was applied to select optimal RBF parameters and to train classification models, for each testing scheme involved in the study. ROC curves were used to measure class prediction rate according to the classification model being evaluated. ROC plots the false positive rate against the true positive rate and the AUC is estimated. AUC ranges from 0 to 1, which gives an indication of how good a classifier is. LibSVM [26] was used to implement the SVM classifier.

Acknowledgment

This work was supported in part by a grant from the EU-FP6, CARDIOWORKBENCH project, to F.A.

References

[1] American Heart Association, January 30, 2007. URL: http://www.americanheart.org.
[2] A. Barth, R. Kuner, A. Buness, M. Ruschhaupt, S Merk, L. Zwermann, et al., Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies, J. Am. Coll. Cardiol. 48 (2006) 1610–1617.
[3] J.Y. King, R. Ferrara, R. Tabibiazar, J.M. Spin, M.M. Chen, A. Kuchinsky, et al., Pathway analysis of coronary atherosclerosis, Physiol. Genomics 23 (2005) 103–108.
[4] F. Wittchen, L. Suckau, H. Witt, C. Skurk, D. Lassner, H. Fechner, et al., Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets, J. Mol. Med. 85 (2007) 257–271.
[5] M.M. Kittleson, M.M. Khalid, A.I. Rafael, Q.Y. Shui, E. Gina, B. Elayne, et al., Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure, Physiol. Genomics 21 (2005) 299–307.
[6] A. Camargo, F. Azuaje. Linking gene expression and functional network data in human heart failure, PLoS ONE 2(12) e1347.
[7] A.P.J.J. Bijnens, E. Lutgens, T. Ayoubi, J. Kuiper, A.J. Horrevoets, M.J.A.P. Daemen, Genome-wide expression studies of atherosclerosis: critical issues in methodology, analyses, interpretation of transcriptomics data, Arterioscler. Thromb. Vasc. Biol. 26 (2006) 1226–1235.
[8] G.S. Ginsburg, D. Seo, C. Frazier, Microarrays coming of age in cardiovascular medicine: standards, predictions, and biology, J. Am. Coll. Cardiol. 48 (2006) 1610–1617.

[9] Entrez Gene, January 30, 2007. URL: http://www.ncbi.nlm.nih.gov/entrez.
[10] Z. Zhang, D. Chen, D.A. Fenstermacher, Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome, BMC Genomics 20 (8) (2007) 331.
[11] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.
[12] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. USA 99 (2002) 6567–6572.
[13] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, et al., Development of human protein reference database as an initial platform for approaching systems biology in humans, Genome Res. 13 (2003) 2363–2371.
[14] Gene Ontology, January 30, 2007. URL: http://www.geneontology.org.
[15] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge Univ. Press, Cambridge, UK, 2000.
[16] Genecards, January 30, 2007. URL: http://www.genecards.org.
[17] R.M. Bostick, L. Fosdick, G.A. Grandits, P. Grambsch, M. Gross, T.A. Louis, Effect of calcium supplementation on serum cholesterol and blood pressure: a randomized, double-blind, placebo-controlled, clinical trial, Arch. Fam. Med. 9 (2000) 31–38.
[18] K.C. Bilchick, J.G. Duncan, R. Ravi, E. Takimoto, H.C. Champion, W.D. Gao, L.B. Stull, D.A. Kass, A.M. Murphy, Heart failure-associated alterations in troponin I phosphorylation impair ventricular relaxation-afterload and force-frequency responses and systolic function, Am. J. Physiol. Heart Circ. Physiol. 292 (2007) H318–325.
[19] A.I. Malinin, C.M. O'Connor, A.I. Dzhanashvili, D.C. Sane, V.L. Serebruany, Platelet activation in patients with congestive heart failure: do we have enough evidence to consider clopidogrel? Am. Heart J. 145 (2003) 397–403.
[20] Agilent Literature Search Software, January 2007. URL: http://www.agilent.com/labs/research/litsearch.html.
[21] N. Ancona, R. Maglietta, A. Piepoli, A. D'Addabbo, R. Cotugno, M. Savino, et al., On the statistical assessment of classifiers using DNA microarray data, BMC Bioinform. 7 (2006) 387.
[22] Gene Expression Omnibus, January 2007. URL: http://www.ncbi.nlm.nih.gov/geo/geo.
[23] X. Lu, V.V. Jain, P.W. Finn, D.L. Perkins, Hubs in biological interaction networks exhibit low changes in expression in experimental asthma, Mol. Syst. Biol. 3 (2007) 98.
[24] S. Maere, K. Heymans, M. Kuiper, BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks, Bioinformatics 21 (2003) 3448–3449.
[25] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[26] C.C. Chang, C.J. Lin. LIBSVM: a library for support vector machines, January 30, 2007. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.