31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Identification of markers of Cardiovascular Disease in women and the reconstruction of its corresponding Protein Interaction network

Camargo, A. Kim, J. T.

Abstract — **Cardiovascular disease is the second most prevalent cause of morbidity and mortality in women of developed countries. Although it is well established that gender is a risk factor for cardiovascular disease, most gene expression analysis studies favour the identification of disease bio-markers and potential drug targets over combined populations. This study integrates genomic and systems approaches to identify a female-related set of genes that intervene in signal and metabolic pathways leading to cardiovascular disease**.

I. INTRODUCTION

Cardiovascular disease (CVD) is the second most prevalent cause of morbidity and mortality in women of developed countries [1],[2],[3]. Although it is well established that gender is a risk factor for cardiovascular disease, most gene expression analysis studies tend to use combined populations of samples. These studies have applied different methodologies to gain insights into the aetiology of this disease and have produced a myriad of valuable data. In the past, we integrated gene expression analysis on some of those datasets, a Protein-Protein Interaction Network (PPI), and a machine learning methodology to investigate biological responses in experimental Dilated Cardiomyophaty [4],[6]. Because we had seen a gap regarding gender-based analysis, this study integrates genomic and systems approaches to identify a female-related set of genes that intervene in signal and metabolic pathways leading to CVD. However, this time we have used a more constrained, yet more robust, approach due to the scarcity of the data and the goals of this study.

II. METHODS

A. Gene expression analysis

A microarray dataset generated by a study on cardiovascular disease and composed of 172 male samples and 50 female samples was obtained from the Gene expression Omnibus [6], accession number GSE12288

Data were pre-processed, log transformed and split according to gender; the female subset was composed of 28 Non-CVD samples and 22 CVD samples. Probes with little variation among experiments were filtered out through coefficient of variability ($p > 0.20$), leaving 4443 probes for further analysis. Statistically significant changes in expression of these genes' expression were determined by applying a t-Test over a permutation test which is performed as follows: first, test statistic and corresponding P-value are calculated on the original data set; second, data are permuted at random B times and test statistics are calculated on each permuted data set; third, permuted distribution is calculated by counting the times (K) the statistic value obtained in the original data set was smaller than the statistic value obtained from the permuted data sets, and dividing that value by the number of random permutations i.e. K/B. In this study B was set to 5000 and the significance level to reject the null hypothesis was set to 0.05.

B. Network analysis

A PPI network composed of all validated interactions in humans was assembled and visualised according to a colour scheme [4]. In the network, nodes representing proteins encoded by genes associated with cardiovascular disease were coloured in yellow, nodes representing proteins encoded by genes differentially expressed were coloured in light purple, and other nodes were coloured in cyan. Potential protein complexes were identified through the MCL and the MCODE algorithms. The detection of densely connected regions in MCL is based on the simulation of stochastic flows, whereas in MCODE is based on the use of vertex weights that represent local neighbourhoods [7].

A metasearch tool that queries the scientific literature was used to identify genes associated with cardiovascular disease. PPIs were retrieved from the HPRD. The HUGO nomenclature standard was used to define unique gene identifiers.

C. Gene Ontology analysis

Biological enrichment of group of genes against whole genome was assessed under hypergeometric test ($p < 0.05$).

D. Classification model

Gene expression profiles of candidate genes were evaluated in the context of machine learning which led to the identification of the best classification model (i.e. best discrimination trade-off between CVD and Non-CVD samples). Correlation-based feature subset selection [8] and Support Vector Machine were used for attribute selection - 8-fold cross-validation was used in both cases, groups were split at random 10 times. Radial Basis Function network was used as a machine learning method in which 66% and 34% of the whole dataset was used for training and testing purposes.

A.C. is with the School of Computing, University of East Anglia, Norwich, NR4 7TJ, England, UK (phone:+44(0)1603292912; e-mail: a.camargo-rodriguez@uea.ac.uk)

J.T.K. is with the School of Computing, University of East Anglia, Norwich, NR4 7TJ, England, UK (e-mail: j.kim@uea.ac.uk)

III. RESULTS

A. Candidate genes

Gene expression analysis between CVD and Non-CVD female samples identified 173 Significantly Differentially Expressed (SDE) probes. The same analysis on male samples identified 516 SDE genes. Between these two groups, a total of 26 genes symbols overlapped and were withdrawn from the analysis, one of them was the COL3A1 (collagen, type III, alpha 1). This gene whose p-value was one of the lowest ($P < 0.0002$) in the female dataset has been associated with the risk of coronary artery disease. In addition, biological enrichment analysis of discarded genes suggested that they were more likely to participate in processes such as "Steroid hormone receptor signalling pathway", "intracellular receptor-mediated signalling pathway" and "protein complex assembly". Other expression profiles whose p-values were low ($P < 0.0005$) corresponded to genes *ARHGEF9*, *B4GALNT1*, *LPAR4*, *PKP4*, *ETAA1*, *ZNF236*, and *WNT16*. Gene Ontology (GO) analysis identified no over-expressed biological processes, which suggest the analysis of these genes in the context of gene-gene association.

B. Network analysis

Once candidate genes were identified, their possible roles in the context of a Protein-Protein Interaction Network (PPI) were investigated. For this purpose, we assembled a PPI network of 9021 nodes representing proteins and 34119 edges representing confirmed *in vivo*, or *in vitro*, interactions among them. In the network, nodes representing proteins encoded by genes associated with cardiovascular disease (65) were coloured in yellow, nodes representing proteins encoded by SDE genes (132) were coloured in light purple, and other nodes were coloured in cyan. Note that the number of SDE is down because some gene's encoded proteins have no validated interactions according to the Human Protein Reference Database (HPRD) [6].

Molecular Clustering Detection (MCODE) analysis of the whole PPI network's topology extracted six highly scored clusters where at least one of our candidate genes was present (i.e. *MET*, *CLTB*, *TRPC5*, *PSMD5*, *CNKSR2*, *RAD51*, *ESR2*, *PRKCI*, *MDM2* and *BCR*) - Table 1 shows main details about each of these clusters. In particular, one of these clusters contained genes *ESR2*, *PRKCI*, *MDM2* (already associated with CVD) and BCR. Biological significance of this gene group identified processes not frequently associated with CVD such as "Biopolymer biosynthetic process" and "Neutrophil chemotaxis", and processes commonly associated with CVD such as "Phosphorylation". In the same group we also found processes such as "Embryonic development" and "Estrogen receptor signalling pathway", which makes us suspect the correlation between those genes and some metabolic and signalling female-associated processes. Fig 1 is an excerpt of the global PPI (see methods for colour scheme). Nodes representing proteins encoded by SDE genes and also present in highly scored clusters were coloured grey. According to metasearch queries, genes *MDM2* and *BCR* were associated with atherosclerosis and CVD, respectively, which might suggest the involvement of ESR2 and *P*RKCI in the development of CVD. Further queries on the National Centre for Biotechnology Information (NCBI) database suggested that MET and TRPC5 are involved in the endothelial cell growth factor which is a process commonly linked to the onset of CVD.

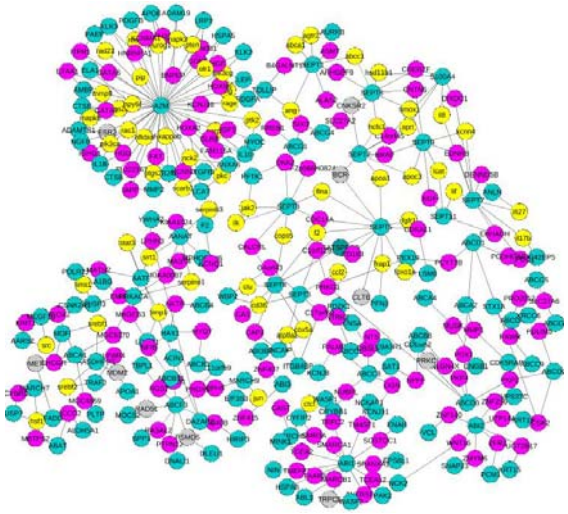

Fig. 1. Excerpt from global PPI network showing proteins encoded by SDE genes and snapshots of highly ranked clusters where SDE gene/proteins were present. In the network, nodes representing proteins encoded by genes associated with cardiovascular disease were coloured in yellow, nodes representing proteins encoded by genes differentially expressed were coloured in light purple, and other nodes were coloured in cyan. Nodes in grey are SDE genes that were found in highly scored clusters identified by the MCODE algorithm.

TABLE I. DENSELY CONNECTED REGIONS DETECTED BY MCODE ALGORITHM OVER WHOLE PPI NETWORK. CLUSTERS THAT INCLUDED AT LEAST ONE SDE GENE AND WHOSE SCORES WERE ABOVE 1 WERE SELECTED.

Cluster	Score	No of genes	SDE
1	3.6	18	1
2	2.2	65	4
3	1.7	128	2
4	1.6	129	1
5	1.6	20	1
6	1.5	36	1

Further to the analysis of the whole PPI network, we used the Markov Cluster Algorithm (MCL) to identify potential dense regions over the excerpt of the network that pertained to proteins - and their interactions – encoding SDE genes (268 nodes, 250 edges) (Fig 1). This time however, we discarded clusters whose biological enrichment assessment was considered insignificant. The output of this analysis reported 33 significant clusters (composed of at least three nodes), some of them showed strong evidence of their

association with CVD. For example, the second highest scored cluster, were candidate genes *ALDH5A1*, *MCCC2*, *ABAT* and *MBTPS2* were present was enriched in the "Glutamate metabolism" metabolic pathway whose role has been associated with hypoglycaemia autonomic failure in type 1 diabetes. The third highest scored cluster, composed of 128 genes, contained two candidate genes *CLTB* and *TRPC5*. This cluster of genes was enriched in biological processes such as "Apoptosis", "Neutrophil activation" and "Positive regulation of caspase activity". These processes are commonly linked with cardiovascular-disease-liked disorders.

Another interesting finding of the network analysis process was that in general proteins encoded by SDE genes were neither hubs nor superhubs. Instead, they were what are known as peripheral nodes (i.e. few edges) [4] suggesting the role of these genes in very specific signalling and metabolic processes.

C. Class predictor genes

Gene expression analysis singled out genes whose expression profiles showed significant differences in relation to the average, and showed clear differentiations between case and control samples. Having identified 146 SDE genes, the next step was to search for a set of genes that together could make up for the best classification model. In the past we have successfully used Support Vector Machines (SVM) for attribute selection [6], however this time we added the correlation-based feature subset selection as a confirmation method. The combination of these two techniques led to the selection of 14 class predictor genes (*ARHGEF9*, *PCSK2*, *EHHADH*, *EGF*, *LPAR4*, *OBSL1*, *ITPR1*, *ETAA1*, *SLC27A6*, *WNT16*, *TAAR3*, *NTRK2*, *PDLIM5* and *ZNF236*). These genes's expression vectors were used as inputs to a Radial Basis Function network classifier whose classification accuracy was of 100%. In addition, classification accuracy of 92% was obtained when cross-validation was used as validation method. In the past, research studies have used several other data mining approaches to analyse meta-datasets (some of them documented in [5]) and arrived to same as conclusion as we did, they found that the expression profile of a small proportion of genes could be used as the class predictor set.

Biological enrichment analysis of class predictor genes showed no over-expressed biological process. The same gene symbols were assessed against the scientific literature to confirm any gene-disease target association. Results of this search reported that except from *EGF*, *ITPR1* and *NTRK2* genes, most of our class predictor genes were not associated with CVD. In addition, searches on the NCBI suggest that genes *PCSK2, SLC27A6* and *WNT16* might be linked to processes that are involved in the development of CVD such as "proinsulin-processing" and "fatty acid transport".

IV. CONCLUSION

The post-genome era has made available vast amounts of genomic, proteomic and metabolomic data to the scientific community. Traditional gene-expression-based analyses have led to the identification of biomarkers of CVD whose roles in disease development have been confirmed later through *in vivo* and *in vitro* tests. As a result, the expression levels of some of those gene markers are regularly assessed as a first check for disease presence, or as a tool to measure up disease evolution. Omoigui [3] and Braunwald [10] have succeeded in describing a set of genes that are commonly used as disease markers. They have also made clear that genes' behaviour must be seen within the regulatory network context to identify gene's interactions and their role and impact in disease onset and development.

The importance of integrative approaches lies in that they can help identify candidate genes whose relevance and interaction among their encoded proteins can be then confirmed or ruled out by wet-lab. With this hope, this study on CVD in women integrated multiple information sources and used several data mining techniques to identify potentially novel and relevant signature genes.

The integrative study reported here highlights three important aspects. First, the approaches implemented here helped identify potentially influential genes[1] (e.g. class predictor genes). Second, it represents a powerful methodology to trace biological processes that may outline potential clinical biomarkers or therapeutic targets and their corresponding interacting partners. The latter can be used as tool to reconstruct metabolic and signalling networks.

This study also evidence that it is very difficult to gather women-related data. We queried well known data repositories such as Entrez and GeneExpress and found only one dataset that provided enough samples to perform this study.

ACKNOWLEDGMENT

A.C. thanks Dr. Francisco Azuaje at the Laboratory of Cardiovascular Research, Luxembourg, for his feedback at the very start of this research study.

REFERENCES

[1] S. Davison, R.S. Davis, New Markers for Cardiovascular Disease Risk in Women: Impact of Endogenous Estrogen Status and Exogenous Postmenopausal Hormone Therapy," J. Clin. Endocrinol. Metab., vol. 88, pp. 2470-2478, 2003.

[2] K.K. Berg, H.O. Madsen, P. Garred, R. Wiseth, S. Gunnes, V. Videm, "The additive contribution from

[1] To get access to the complete list of candidate genes, refer to: http://www2.cmp.uea.ac.uk/~jtk/embc2009/

inflammatory genetic markers on the severity of cardiovascular disease". Scand J Immunol., vol. 69, no. 1, pp. 36-42, 2009.

[3] S. Omoigui, "The Interleukin-6 inflammation pathway from cholesterol to aging – Role of statins, bisphosphonates and plant polyphenols in aging and age-related diseases," Immunity & Ageing, vol. 4, no. 1, 2007.

[4] A Camargo, F. Azuaje, "Linking gene expression and functional network data in human heart failure," PLoS ONE,vol. 2, no. 12, pp. e1347, 2007.

[5] A. Camargo, F. Azuaje, "Identification of dilated cardiomyopathy signature genes through gene expression and network data integration," Genomics, vol. 9, no. 6, 2008.

[6] Gene Expression Omnibus (GEO). (2009, June). [Online]. Available: http://www.ncbi.nlm.nih.gov.geo/geo.

[7] G.D. Bader, C.W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics, vol. 13, no. 4, 2003.

[8] M. Hall, "Feature Subset Selection: A Correlation Based Filter Approach," 1997.

[9] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, "Development of human protein reference database as an initial platform for approaching systems biology in humans," Genome, vol. 13, no. 10, pp. 2363-71, 2003

[10] E.Braunwald, "Biomarkers in heart failure," N. Engl J Med., vol. 358, no. 20, pp. 2148-2159, 2008.

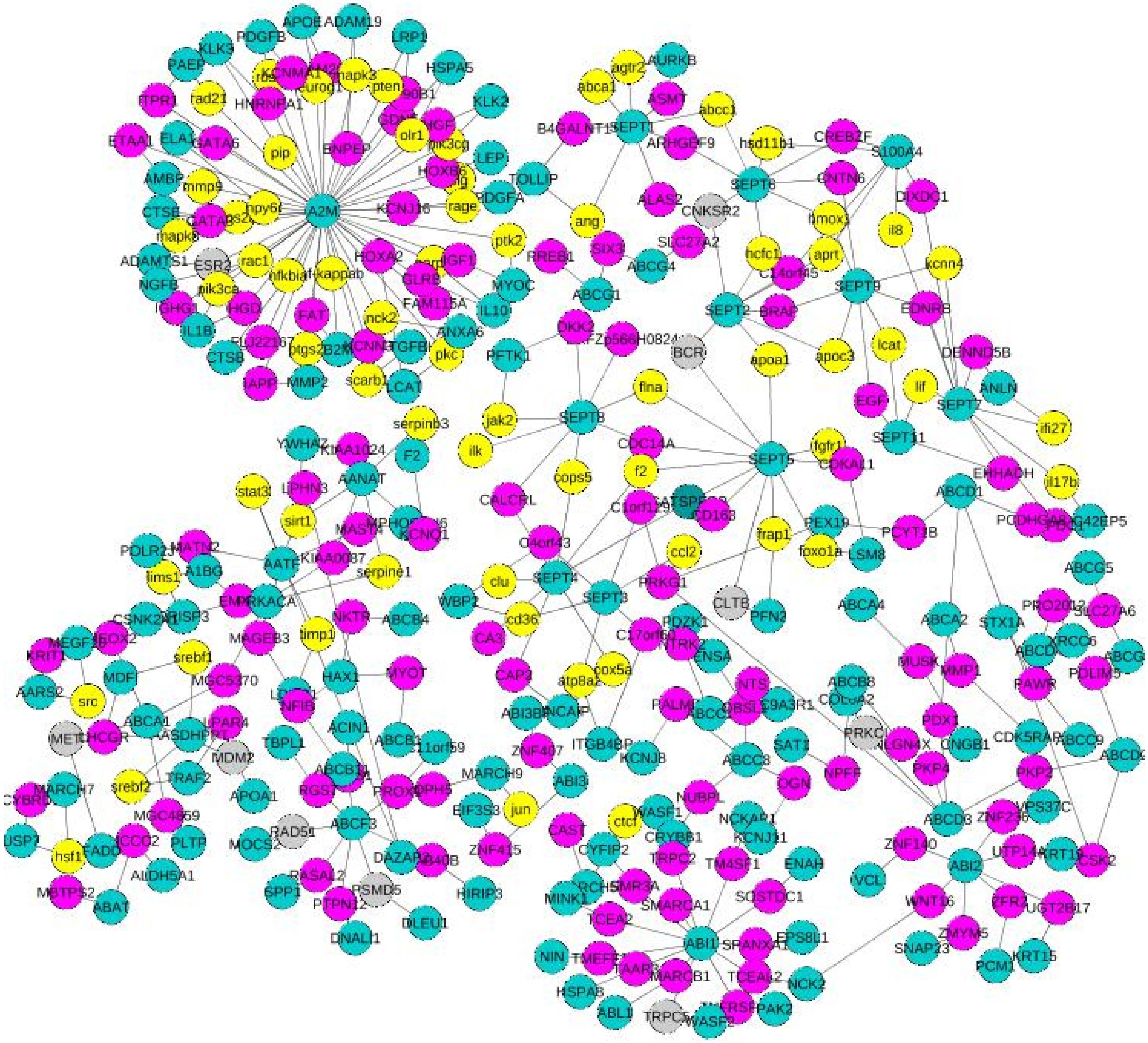

Fig. 2. Excerpt from global PPI network showing proteins encoded by SDE genes and snapshots of highly ranked clusters where SDE gene/proteins were present. In the network, nodes representing proteins encoded by genes associated with cardiovascular disease were coloured in yellow, nodes representing proteins encoded by genes differentially expressed were coloured in light purple, and other nodes were coloured in cyan. Nodes in grey are SDE genes that were found in highly scored clusters identified by the MCODE algorithm.

TABLE I. DENSELY CONNECTED REGIONS DETECTED BY MCODE ALGORITHM OVER WHOLE PPI NETWORK. CLUSTERS THAT INCLUDED AT LEAST ONE SDE GENE AND WHOSE SCORES WERE ABOVE 1 WERE SELECTED.

Cluster	Score	No of genes	SDE
1	3.6	18	1
2	2.2	65	4
3	1.7	128	2
4	1.6	129	1
5	1.6	20	1
6	1.5	36	1