Clustering based identification of co-expressed genes and their associated transcription factor enrichment network in Cardiomyopathy

Dr. Freddy AJ[1] , Beutline Malgija M[*]

[1] Department of Zoology, Madras Christian College, Chennai 600059.

* Corresponding Author: Computational Science Laboratory, MCC-MRF Innovation Park, Madras Christian College, Chennai 600059

ABSTRACT

Regulation of gene expression is vital to cell growth, differentiation and disease, which is achieved largely by transcription factors (TFs), the key players in transcriptional regulation, which have pulled in great experimental consideration, yet the elements of most human TFs are not well understood. The group of samples, gene clusters and the relationship between the samples and differentially expressed genes were analysed using principal Component Analysis (PCA) and cluster analysis and were explored to gain insights on the co-expressed genes. Cluster estimations were calculated based on Silhouette and Calinski criterion and further validated clusters for each datasets were generated using internal clustering algorithm. Further network analysis of the TFs revealed the importance of some hub TFs to be associated with cardiomyopathy.

Keywords:

Cardiomyopathy, Transcription factor, Differentially expressed genes, Clustering, Gene expression, Network analysis

1. INTRODUCTION

Cardiac development is a fine-tuned process governed by complex transcriptional networks mediated by the interaction of transcription factors with other regulatory elements. Cardiac TFs are key players that regulate inducible gene expression in cardiac myocytes. A large number of studies have revealed a set of myogenic TFs that forms the hub of an evolutionary conserved regulatory network defining cardiac morphogenesis, cardiac cell fate and myogenic differentiation. Cross-regulation or auto regulation of core cardiac TFs activates one or a few factors within the network which may finally activate them all, as well as common sets of downstream targets (Kathiriya et al., 2015). One of the first large-scale studies on cardiac transcription networks used a combinatorial approach to predict regulatory subnetworks in patients with varied cardiac

abnormalities. Large-scale high throughput gene expression studies and genome sequencing projects are providing vast amounts of information that can be used to identify or predict cellular regulatory processes. The differentially expressed genes (DEGs) screened from such gene expression data, needs to be analysed for their similarity in expression pattern (co-expressed genes), because the functionally related genes are co-expressed and such co-expression may reveal insights into the genes regulatory network. Genes can be clustered on the basis of the similarity of their expression profiles or function and these clusters are likely to contain genes that are regulated by the same transcription factors. This can be achieved by means of clustering techniques which plays an important role in the analysis of microarray data (Slonim 2002). Clustering analysis remains a significant tool for unsupervised learning, i.e., the problem of finding the clusters in data without the use of a response variable. A key challenge in cluster analysis is the estimation of the optimal number of 'clusters' (Tibshirani*et al*., 2001) and the methods used for clustering. Hence this study applied different approaches for cluster estimation and clustering to attain accurate results. Binding of transcription factors to transcription factor binding sites (TFBSs) is important in the process of transcriptional regulation. Statistical models were constructed in several earlier studies to understand the regulatory functions of TF based on gene expression and TF-binding data (Chen and Gerstein, 2011; Ouyang et al., 2009). The DNA-binding transcription factors (DbTFs) play an essential role in transcription of a gene, as they can guide the transcriptional machinery of various target genes by their binding to specific regulatory regions that determine which part of the gene is to be transcribed (Kadonaga, 2004). Identification of over-represented TFBSs from a list of significant genes might provide insights into the molecular mechanism relevant to particular biological context i.e., normal functioning or development of a disease (Sui *et al*., 2007). Hence this study aims at the screening of regulators of gene expression that induces altered gene expression in the pathologic mechanism of HF.

2. MATERIALS AND METHODS

2.1 Data Retrieval

Expression data of DEGs screened from our previous study (Malgija and Shanmughavel, 2015) was used for clustering. This study identified the altered expression of 15 transcription factors (8 up regulated and 7 down regulated) and 10 transcription factors (5 up regulated and 5

down regulated) in DCM and HCM respectively. All the deregulated genes were subjected to clustering.

2.1 Data Retrieval

Normalized expression data of DEGs screened from our previous study (Malgija et al., 2018) was used for clustering. All the deregulated genes were subjected to clustering.

2.2 Principal Component Analysis (PCA)

A data matrix with columns representing the samples and the rows corresponding to genes was given as input and the principal components were generated using prcomp() function in R.

2.3 Cluster Analysis

Clustering by K-means and Fuzzy C-means (FCM) algorithms requires the number of clusters to be predefined since it varies based on the input dataset. The number of clusters was calculated based on silhouette (Rousseuw, 1987) and Calinski criterion (Calinski and Harabasz, 1974) for each of the datasets separately. The significant genes were clustered using different clustering algorithms namely Hierarchical (Szekely and Rizzo, 2005), K-means (Hartigan and Wong, 1979) and Fuzzy C means (FCM; Bezdek, 1981) clustering using R. The expression matrix was in the dimension of 242 x 12 and 756 x 16 for DCM and HCM respectively. We took the average of all the probes for the same mRNA in order to deal with interrogated genes (Wang *et al*., 2015). We input the expression matrix of both DCM and HCM to calculate distances and the distance matrix was constructed based on Euclidean distance (Deza*et al*., 2009), which computes the distance between two points in Euclidean space. The fuzziness parameter for FCM was set to 2, which is the default value in most cases. Both the k-means and FCM algorithm was run for 1000 iterations, in order to increase the accuracy of P-value.

2.4 Cluster Validation

A numerous number of clustering algorithms were developed and many of which have shown promise in high throughput data analysis (Fu and Medico 2007; Dembele and Kastner 2003; Herrero*et al*. 2001) and deciding which clustering algorithm to be used is important. Hence we validated the clusters using the clValid package (Brock *et al*., 2008). This package upholds functions for validating the results generated from clustering algorithms, using any of the validation measures "internal", "stability" and "biological". We performed internal clustering based on the expression profiles of different disease conditions using the internal measures

connectivity, Dunn, and Silhouette. The neighborhood size for connectivity was set as 10 by default, with the methods hierarchical, K-means and fanny (for FCM).

2.5 Identification of Transcription factor Binding sites

Each of the clusters was further analyzed for their TFBS using oPOSSUM server (Sui et al., 2007). oPOSSUM combines a pre-computed database of conserved TFBS in human promoters with statistical methods for identification of over-represented sites in a set of co-expressed genes. It predicts the over-represented transcription factors based on two complementary scoring methods Z-score and F-score. Z-score is based on a binomial distribution that measures the change in the relative number of TFBS motifs in the input gene set compared with the background set, whereas Fisher score is based on a one-tailed Fisher exact probability scanning the number of genes with the TFBS motifs in the given gene list Vs. the background list. TFs with Z-score>10 and F-score>7 were considered, as this is the empirically selected thresholds. Information of TFBS with no hits is searched through Genecards (Safran et al., 2010). Information regarding the selected transcription factors was obtained from Human TFDB (Transcription Factor Database).

2.6 Network Analysis

The neighbour interactive partners of the enriched TFs were obtained from STRING database. The network files were created based on these partners and network was constructed using cytoscape. MCODE module identified the highly interconnected regions in the network.

3.RESULTS AND DISCUSSION

The expression values of the differentially expressed genes (242 in DCM and 758 in HCM) were extracted from the normalized data for further clustering analysis.

3.1 Principal Component Analysis

PCA based on correlation matrix with samples as variables were performed for 242 and 758 genes respectively, concerning both DCM and HCM. The first two PCs from the table 1 explained more than 90% of the total variations in both cases. The eigen values are extracted for the dimensions from the analysis to create a customised scree plot. The contributions of the variables to dimensions one and two are plotted using the fviz_famd_var() function and the percentage contribution of each variable to each dimension's explanatory power is given as a scree plot(Figure 1).

.

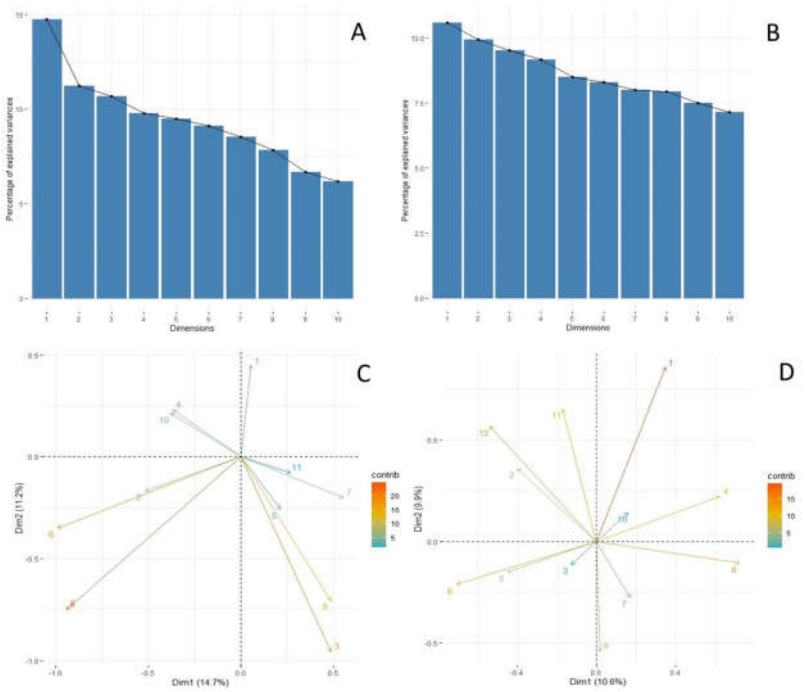

Figure 3. 1. Scree plot and Biplot of the variables.

A &B) Screeplot of DCM and HCM showing contributions of variable to dimensions. In DCM the first two components are sufficient whereas HCM shows only small variations. C & D) Biplot of variables plotted against different dimensions. Positive correlated variables point to the same side of the plot. Negative correlated variables point to opposite sides of the graph.

3.2 Estimation of cluster number

Clustering techniques such as K-means and FCM requires the number of clusters to be defined before clustering. Though several statistical criteria and cluster validity indices are available to select the number of clusters automatically, it is important for the user to specify the number of clusters. Because the number of clusters varies based on the dataset and such cluster estimation has a deterministic effect on the clustering results. The calculated number of clusters for differentdataset is displayed in figure 2.

In order to attain confident results, we used two algorithms to calculate the number of clusters. Both the Silhoutee and Calinski criterion decided the number of clusters to be 2 for DCM

and 3 for HCM. A silhouette close to 1 implies their assignment in the appropriate cluster, whereas that with the value close to -1 implies the datum in the wrong cluster.

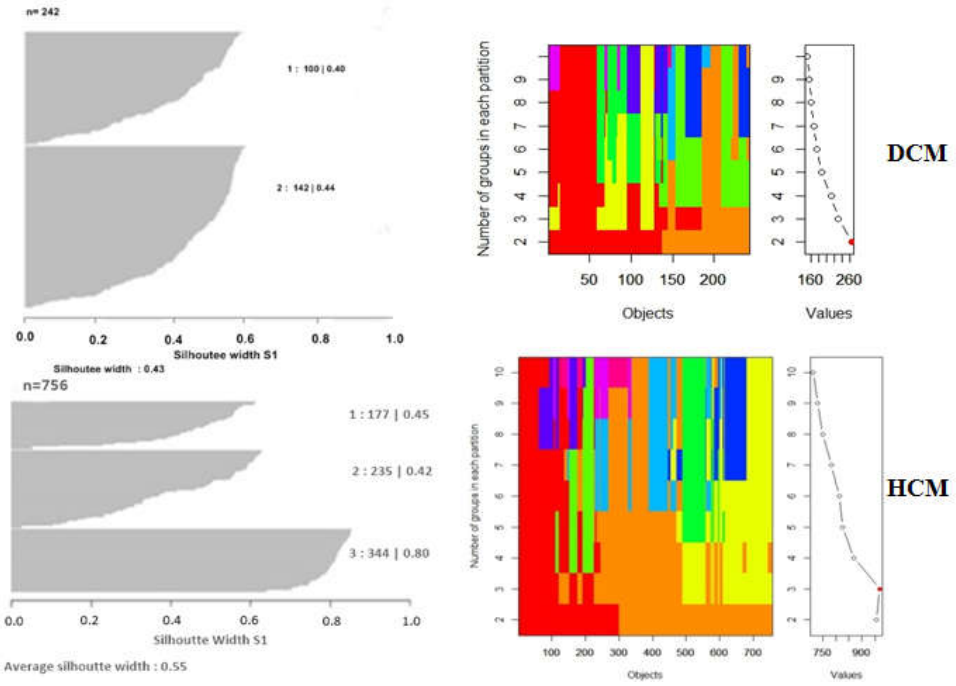

Figure 3.2 Number of clusters by Silhouette and Calinski criterion for DCM and HCM. Thickness of the silhouette plot specifies the size of clusters. Red dot in calinski plot represents the number of clusters (i.e., 2 for DCM and 3 for HCM).

3.3 Clustering

Choosing the right clustering technique is a critical step in clustering. We observed slight variations in genes occupying various clusters using different algorithms. Since some of the genes were repeated, we calculated averages of their expression values. Clustering by the hierarchical method found that some genes occupy different clusters based on their expression values in different datasets of DCM and HCM. The distances for hierarchical clustering were calculated using Euclidean distance. The number of genes generated in different clusters by different methods for DCM and HCM is shown in supplementary file 1.

In hierarchical clustering, the resulted tree after clustering was further cut based on the preferred number of clusters, whereas in K-means and FCM the cluster number was given initially. By the implementation of different clustering approaches, we observed mere similar results from

FCM and K-means for different disease conditions. For DCM, hierarchical clustering generated 235 genes in the first cluster and 7 in the second cluster (Supplementary file 2).

Expression values of HCM produced 232, 523 and 1 gene in first, second and third clusters respectively as depicted in supplementary file 1. In FCM, the silhouette values for each gene were computed, indicating that when the value is higher than 0, the corresponding gene is classified to the correct cluster.

3.4 Cluster Validation

The quality of clusters generated by the above clustering methods can be assessed by cluster validation procedures. We performed internal validation using R based on the three internal measures connectivity, Dunn, and Silhouette, which measured the quality of the clusters based on intrinsic properties of the expression data. The dataset with their expression values was given as input for validation. DCM revealed hierarchical algorithm and k-means with two clusters were found to be a good cluster, among which hierarchical is preferred as it shows good score. Similarly Hierarchical with 3 clusters was preferred for HCM (Table 1). By performing various clustering approaches we observed that the genes responsible for related functions are found to occupy the same cluster. Hierarchical clustering showed the genes ACTA1, APOD, FN1, HSPB1, HSPB3, NPPA, and NPPB were found to occupy the second cluster of DCM. Extracellular matrix (ECM) proteins, collagen genes, growth factors (CTGF, LTBP1), Chemokines, Serpin family proteins occupy the first cluster. In contrast, K-means cluster found collagen genes and chemokines in different clusters. The same microarray data set can lead to very different conclusions by using different data analysis techniques and different clustering algorithms [17]. In HCM dataset, hierarchical clustering found a single gene LOC285556 (Uncharacterised LOC285556), a protein-coding gene in a separate cluster. Growth factors, regulatory genes and apoptosis related genes are clustered together in first cluster, whereas Zinc finger genes, ubiquitin related genes, transmembrane, inflammatory and immune related genes occupy the second cluster.

Table 3. 1. Internal validation measures of DCM and HCM

Validation Measures	DCM			HCM		
	Score	Method	**Cluster**	**Score**	**Method**	**Cluster**
Connectivity	16.6083	hierarchical	2	4.3825	hierarchical	3
Dunn	0.2548	k-means	2	0.1655	hierarchical	2
Silhouette	0.4495	K-means	2	0.5448	hierarchical	3

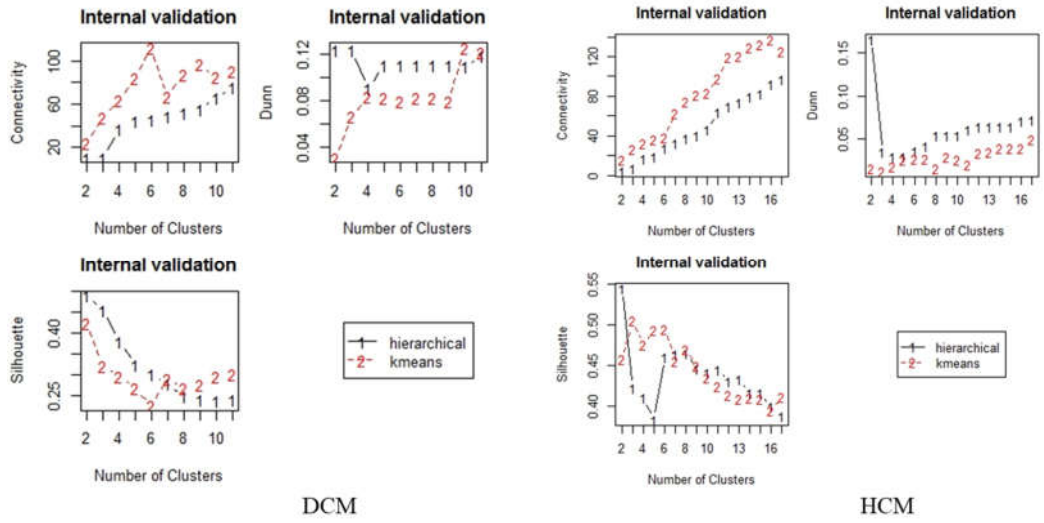

Figure 3.4 Validation of hierarchical, K-means and FCM clustering

This shows the validated clusters generated by internal measures connectivity, Dunn and Silhouette.

3.4 Prediction of TFBS

In the present study, we found the common TFBS patterns in promoter sequences of co-expressed genes. TFBS analysis was employed to find the common TFs controlling the CVD genes. The regions spanning 2000 bps prior to the transcription start site of the over and under-expressed genes were searched for the presence of TFBS using the oPOSSUM database. This predicted the enrichment of different transcription factors and their binding sites. The transcription factors enriched with a significant p-value for overrepresentation were considered and those enriched in different clusters are depicted in Table 2. We found the overlap of transcription factors SRY, Nkx2-5 and HOXA5 in DCM and HCM.

Table 3.2 TFs Enriched in different clusters of CM

Disease	Cluster 1	Cluster 2	Cluster 3	Deregulated TFs among DEGs
DCM	SRF, FOXD1, Nkx2-5, HOXA5, SRY	NHLH1, TAL1::TCF 3	-	↑AEBP1, ↑MEOX2, ↑PLAGL1, ↑ETV1, ↑ZBTB16, ↑ETV5, ↑EGR1, ↑MYBL1, ↓STAT3, ↓NR4A3, ↓ATF3, ↓NFIL3, ↓NKX2-5, ↓PLSCR1, ↓CEBPD
HCM	Nkx2-5, FOXD3, SRY, ARID3A, HOXA5	Zfx, Klf4, SP1, Pax5	RELA, NF-1, GR, Olf-1	↑MEOX2, ↑AEBP1, ↑HEY1, ↑RORA, ↑PLAGL1, ↓STAT3, ↓BCL6, ↓ATF4, ↓TFE3, ↓DIDO1

The Table represents the transcription factors predicted based on the evidence of their regulatory regions in the significant genes, predicted by oPOSSUM server. (-) indicates that there are only two clusters. Fifth column shows the TFs deregulated among the DEGs: ↑ indicated up regulated and ↓ indicates down regulated TFs.

Table 3.3 Details of TFs enriched by the DEGs

TF	Full name	Family	category
NKX2-5	NK2 homeobox 5	Homeobox	Helix-turn-helix
SRF	Serum response factor	SRF	Helix-turn-helix
FOXD1	Forkhead box D1	Fork_head	Helix-turn-helix
HOXA5	Homeobox A5	Homeobox	Helix-turn-helix
ARID3A	AT-rich interaction domain 3A	ARID	Helix-turn-helix
SRY	Sex determining region Y	HMG	Helix-turn-helix
FOXD3	Forkhead box D3	Fork_head	Helix-turn-helix
NHLH1	Nescient helix-loop-helix	bHLH	Basic Domains Group
ZFX	Zinc finger protein, X-linked	zf-C2H2	Zinc-Coordinating Group
KLF4	Kruppel like factor 4	zf-C2H2	Zinc-Coordinating Group
SP1	Sp1 transcription factor	zf-C2H2	Zinc-Coordinating Group
PAX5	Paired box 5	PAX	Helix-turn-helix

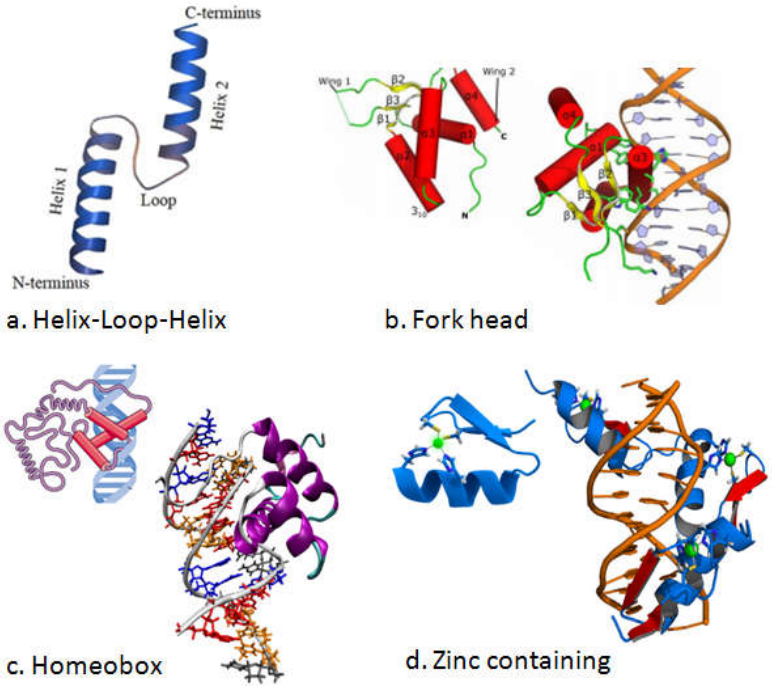

Figure 3.4 Structure of motifs highly enriched in the DEGs. a) The basic Helix-Loop-Helix TFs have two highly conserved and functionally distinct domains. At the amino terminal end is the basic domain, which binds the TF to the DNA at a consensus sequence known as the E-box. b)Fork head / winged family of TFs is characterized by a conserved DNA binding domain known as fork head box (fox) that targets a DNA binding sequence TTGTTTAC. c) Homeo box domain TFs are characterized by the presence of an evolutionarily conserved 60 amino acid domain (Home domain), encoded by an 180bp-DNA sequence (home box). d) Zinc containing DNA binding proteins is characterised by the presence of one or more zinc atoms which stabilizes the modular structure of the domain by coordinating with amino acids, usually cysteine or histidine in the appropriate spatial orientation.

3.4.1 TFs for DCM

Based on the criteria genes are found to be co-expressed depending on their regulatory regions, genes occupy different clusters in respect to specific sites for the binding of TFs that can regulate their expression. Only two TFs NHLH1 and TAL1::TCF3 are overrepresented in cluster 2. NHLH1 and TAL1::TCF3 are shared by 23 genes. NHLH1 showed no overexpressed genes and all genes with TAL1::TCF3 binding sites except DDX3Y and PDK4 are under expressed. Similarly, cluster 2 shows the enrichment of TFs SRF, FOXD1, Nkx2-5, HOXA5 and SRY possessing a number of targets genes 13, 94, 109, 114 and 100 respectively. These TFs were recognized to be significantly overrepresented amongst the up-regulated genes.

3.4.2 TFs for HCM

We observed the overrepresentation of different TFs in different HCM clusters. Homeo TFs Nkx2-5 and HOXA5, SRY, ARID3A and forkhead transcription factor Foxd3 are highly enriched in the first cluster. Over-representation of Zinc-coordinating TFs belonging to Beta-beta-alpha zinc finger motif family of TFs Zfx, Klf4 and SP1 and a homeo transcription factor Pax5 were noticed in cluster 2. As single gene LOC285556 in cluster 3 had not shown any hits, surfing through gene cards revealed the binding site for RELA, NF-1, GR and Olf-1. Both the under expressed and overexpressed genes were highly enriched by the overrepresented TFs.

3.5 Network Analysis

Based on the neighbor interactive partners of the enriched TFs from STRING database, a network with 121 nodes and 128 edges was constructed. This identified some of the deregulated genes from the study to be associated with the enriched TFs as depicted in figure 3 (Red and pink are up regulated and green and cyan down regulated). The constructed network revealed the interaction of various cardiac TFs including GATA TF family (GATA4), MEF-2 family (MEF2C), NKX2-5, SRF and HAND TFs (HAND1, HAND2).

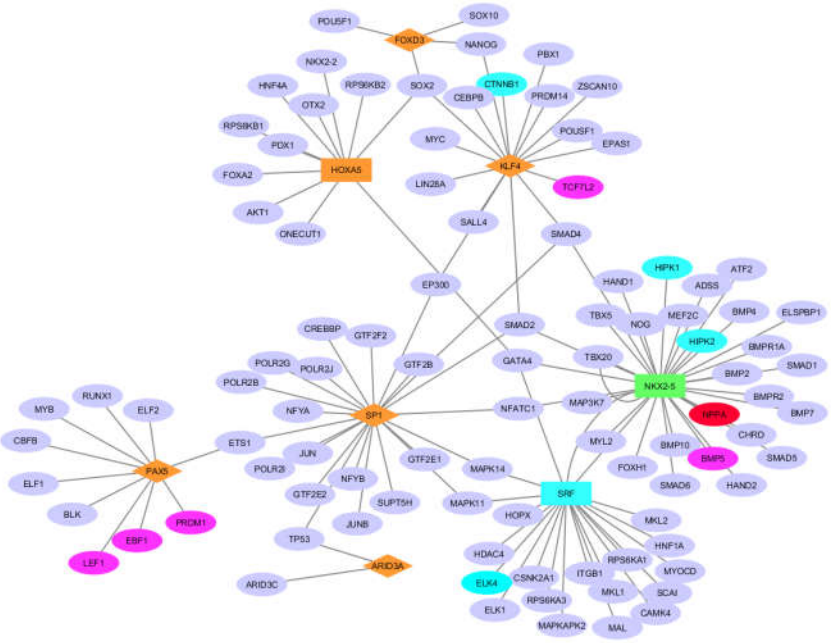

Figure 3.5 Transcription factor interaction Network.

Red and pink represents the up regulated genes in DCM and HCM respectively, whereas Green and cyan are their respective down regulated genes. Diamond and rectangle denotes the

transcription factors enriched in different clusters, in which rectangle represents those which are common in both HCM and DCM clusters. Other nodes (lavender) represent the other interactive partners predicted by STRING.

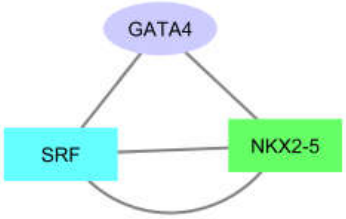

Figure 3.6 Highly interconnected regions generated by MCODE module.

The genes GATA, SRF and Nkx-5 form the highly interconnected regions in the network. SRF and Nkx2-5 are down regulated in DCM and HCM respectively.

4. DISCUSSION

Clustering based on the gene expression profile of the significant genes identified two clusters each for DCM and MI and three clusters for HCM. Homeobox-containing genes Nkx2-5 and HOXA5 and a high mobility group transcription factor, SRY were found to be common in both DCM and HCM. Csx/ Nkx2-5 are a homeobox-containinggene, functions in heart formation and development. It is implicated in commitment and differentiation of the myocardial lineage and has enhanced activity under hypertrophic conditions which regulate the cardiac gene program in hypertrophied hearts (Kohli et al., 2011). HOXA5 encodes a DNA-binding TF; methylation of this gene may result in loss of its gene expression since the encoded protein up regulates the tumor suppressor p53. Zhang *et al*., (2001) reported that the overexpression of Serum response factor (SRF), a cardiac transcription factor involved in cell growth, differentiation and control of a number of cardiac genes altered the expression of the genes regulated by it, leading to muscular dysfunction, a characteristic feature in cardiomyopathy. The protein product of SRY (Sex-determining region Y protein) gene acts as a transcription factor, involved in male sexual development. Forehead/ wing family of TFs, FOXD1 and FOXD3, enriched by the deregulated genes of DCM and HCM respectively modulate response to stress conditions, cell cycle progression, protein degradation and apoptosis (Accili et al., 2004). Ni et al., (2006) reported the role of FOXO TFs in regulating hypertrophic growth in cardiomyocytes, by inhibiting calcineurin signaling. The nucleo-cytoplasmic shuttling mechanism of FOXO is controlled by PI3-kinase-Akt

signaling, whereby PI3-kinase phosphorylation causes Akt to translocate to the nucleus where it phosphorylates FOXO (Van Der Heide and Hoekman, 2004).

Apart from these TFs, DCM and HCM showed the overrepresentation of SRF, NHLH1, TAL1::TCF3 and ARID3A, Zfx, Klf4, Sp1, pax5 respectively. NHLH1 (Nescient Helix-loop-helix 1), located at chromosomal position 1q23.2 belonging to the basic helix-loop-helix (bHLH) family of transcription factors was reported to play significant role in growth and development of a wide variety of tissues and species. bHLH is a protein structural motif that represents one of the largest family of dimerizing TFs. Different subgroups that share particular aminoacid homology within the bHLH domain can be defined within the bHLH family. Members of such subgroups like NHLH1 and NHLH2 shares very high aminoacid homology, showing 98% identity with the bHLH region (Cogliati et al., 2002). Studies reported the over expression of NHLH2 in DCM (Farooqi and O'Rahilly, 2006).A single gene (LOC285556) found in a separate cluster for HCM, specifies the divergent function and regulatory regions of the gene.

Network analysis showed few cardiac TFs including GATA TF family (GATA4), MEF-2 family (MEF2C), NKX2-5, SRF and HAND TFs (HAND1, HAND2) to be important. MEF2C have been found to have an important role in differentiation of myocardial cells and postnatal growth of myocardium (Desjardins and Naya, 2016).The highly interconnected regions of the network include the TFs Nkx2-5 (down regulated in DCM), SRF (down regulated in HCM) and GATA-4. GATA-4 and Nkx2-5 homeo domain protein are two early markers of precardiac cells, necessary for heart formation, but neither can initiate cardiogenesis. Over expression of GATA-4 or Nkx2-5 increases cardiac development in committed precursors, indicating each interacts with a cardiac cofactor. The current network identifies the interconnection of SRF with Nkx2-5 and GATA4; suggesting SRF may be a cofactor. Nkx2-5 regulates multiple characteristics of cardiac cell structure, function and development (Kasahara et al., 2003; Harvey et al., 2002). Mutations in Nkx2-5 mainly functions in a dominant-negative fashion and cause various congenital heart malformations including cardiomyopathy, septal defects, outflow tract defects, hypoplastic left heart and associated arrhythmias (Elliott et al., 2003; Gutierrez-Roelens et al., 2002). A model proposed by Riazi et al., (2009) demonstrated the positive regulation of β-catenin (CTNNB1) and negative regulation of GATA4 by Nkx2-5 in cardiac myocytes. Nkx2-5, SRF and CTNNB1 were found to be decreased in our study, whereas GATA5, another zinc finger transcriptional regulator like GATA4 was increased. Both the GATA TFs were reported to have their role in heart

development (Haworth et al., 2008; Srivastava and Olson, 2000). GATA4 is an important TF expressed in cardiac cells that plays an important role in cardiac development and growth as wells as in cardiac hypertrophy and heart failure (Brewer and Pizzey, 2006; Pikkarainen et al., 2004).

CONCLUSION

Microarray experiments that measure the fluorescence intensity using two colour comparisons has potential pitfall for data analysis. Many factors influence the intensity and produce multicentric effect, creating a need for bias correction or normalisation between two colour systems. Functionally related genes were clustered together. Although internal validation suggested k-means and hierarchical clustering to be good clusters for DCM dataset, we observed that hierarchical clustering showed biologically relevant results. Hence for our datasets, hierarchical clustering showed biologically relevant clusters. Also the use of biological validation instead of internal validation alone might give more accurate results. The identified hub TFs including GATA TF family (GATA4), MEF-2 family (MEF2C), NKX2-5, SRF and HAND TFs (HAND1, HAND2)were found to be associated with heart development and functioning. Abnormalities in expression pattern of these genes might causes disturbance in normal processes, thereby linked to the consequences in cardiomyopathy. Analysing the regulatory regions of over and under dominant genes of HCM showed different binding patterns in up and down-regulation in opposition to DCM.Moreover, TFBS analysis revealed the grouping of genes with similar regulatory regions in the same cluster. Hence in this study, we conclude that the altered gene expression may be co-regulated directly or indirectly by a limited number of transcription factors related to cardiomyopathy. Further study is needed to validate the transcription factors scanned by our analysis.

REFERENCES

1. Kathiriya, I.S., Nora, E.P. and Bruneau, B.G., 2015. Investigating the transcriptional control of cardiovascular development. Circulation research, 116(4), pp.700-714.
2. Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nature genetics, 32(4), pp.502-508.
3. Tibshirani, R., Walther, G. and Hastie, T., 2001.Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), pp.411-423.
4. Cheng, C. and Gerstein, M., 2012.Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic acids research, 40(2), pp.553-568.

5. Ouyang, Z., Zhou, Q. and Wong, W.H., 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences, 106(51), pp.21521-21526.
6. Kadonaga, J.T., 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell, 116(2), pp.247-257.
7. Sui, H.S.J., Fulton, D.L., Arenillas, D.J., Kwon, A.T. and Wasserman, W.W., 2007. oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic acids research, 35(suppl_2), pp.W245-W252.
8. Malgija, B., and Shanmughavel, P., 2015. Differential Gene Expression Analysis of Hypertrophic and Dilated Cardiomyopathy Signature Genes, IJARCSSE, 5, 537-543.
9. Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.
10. Caliński, T. and Harabasz, J., 1974.A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), pp.1-27.
11. Szekely, G.J. and Rizzo, M.L., 2005. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. Journal of classification, 22(2).
12. Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society.Series C (Applied Statistics), 28(1), pp.100-108.
13. Bezdek, J.C., 1981. Objective function clustering. In Pattern recognition with fuzzy objective function algorithms (pp. 43-93). Springer, Boston, MA.
14. Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N.A., Muth, K., Palmer, J., Qiu, Y., Wang, J. and Lam, D.K., 2015. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell stem cell, 16(4), pp.386-399.
15. Deza, M.M. and Deza, E., 2009. Encyclopedia of distances.In Encyclopedia of distances (pp. 1-583).Springer, Berlin, Heidelberg.
16. Fu, L. and Medico, E., 2007. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, 8(1), p.3.
17. Dembélé, D. and Kastner, P., 2003. Fuzzy C-means method for clustering microarray data. bioinformatics, 19(8), pp.973-980.
18. Herrero, J., Valencia, A. and Dopazo, J., 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, 17(2), pp.126-136.
19. Brock, G., Pihur, V., Datta, S. and Datta, S., 2011.clValid, an R package for cluster validation. Journal of Statistical Software (Brock et al., March 2008).
20. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H. and Sirota-Madi, A., 2010.GeneCards Version 3: the human gene integrator. Database, 2010.
21. Kohli, S., Ahuja, S. and Rani, V., 2011. Transcription factors in heart: promising therapeutic targets in cardiac hypertrophy. Current cardiology reviews, 7(4), pp.262-271.
22. Zhang, X., Azhar, G., Chai, J., Sheridan, P., Nagano, K., Brown, T., Yang, J., Khrapko, K., Borras, A.M., Lawitts, J. and Misra, R.P., 2001.Cardiomyopathy in transgenic mice with cardiac-specific overexpression of serum response factor. American Journal of Physiology-Heart and Circulatory Physiology, 280(4), pp.H1782-H1792.
23. Accili, D. and Valenti, L., 2004. Turning up the heat in the fat cell. Nature medicine, 10(11), pp.1168-1169.

24. Ni, Y.G., Berenji, K., Wang, N., Oh, M., Sachan, N., Dey, A., Cheng, J., Lu, G., Morris, D.J., Castrillon, D.H. and Gerard, R.D., 2006.Foxo transcription factors blunt cardiac hypertrophy by inhibiting calcineurinsignaling. Circulation, 114(11), p.1159.
25. Van Der Heide, L.P., Hoekman, M.F. and Smidt, M.P., 2004. The ins and outs of FoxO shuttling: mechanisms of FoxO translocation and transcriptional regulation. Biochemical Journal, 380(2), pp.297-309.
26. Cogliati, T., Good, D.J., Haigney, M., Delgado-Romero, P., Eckhaus, M.A., Koch, W.J. and Kirsch, I.R., 2002.Predisposition to arrhythmia and autonomic dysfunction in Nhlh1-deficient mice. Molecular and cellular biology, 22(14), pp.4977-4983.
27. Farooqi, I.S. and O'Rahilly, S., 2006. Genetics of obesity in humans. Endocrine reviews, 27(7), pp.710-718.
28. Desjardins, C.A. and Naya, F.J., 2016. The function of the MEF2 family of transcription factors in cardiac development, cardiogenomics, and direct reprogramming. Journal of cardiovascular development and disease, 3(3), p.26.
29. Kasahara, H., Ueyama, T., Wakimoto, H., Liu, M.K., Maguire, C.T., Converso, K.L., Kang, P.M., Manning, W.J., Lawitts, J., Paul, D.L. and Berul, C.I., 2003. Nkx2. 5 homeoprotein regulates expression of gap junction protein connexin 43 and sarcomere organization in postnatal cardiomyocytes. Journal of molecular and cellular cardiology, 35(3), pp.243-256.
30. Harvey, R.P., Lai, D., Elliott, D., Biben, C., Solloway, M., Prall, O., Stennard, F., Schindeler, A., Groves, N., Lavulo, L. and Hyun, C., 2002, January.Homeodomain factor Nkx2-5 in heart development and disease. In Cold Spring Harbor symposia on quantitative biology (Vol. 67, pp. 107-114).Cold Spring Harbor Laboratory Press.
31. Elliott, D.A., Kirk, E.P., Yeoh, T., Chandar, S., McKenzie, F., Taylor, P., Grossfeld, P., Fatkin, D., Jones, O., Hayes, P. and Feneley, M., 2003. Cardiac homeobox gene NKX2-5mutations and congenital heart disease: Associations with atrial septal defect and hypoplastic left heart syndrome. Journal of the American College of Cardiology, 41(11), pp.2072-2076.
32. Gutierrez-Roelens, I., Sluysmans, T., Gewillig, M., Devriendt, K. and Vikkula, M., 2002. Progressive AV-block and anomalous venous return among cardiac anomalies associated with two novel missense mutations in the CSX/NKX2-5 gene. Human mutation, 20(1), pp.75-76.
33. Riazi, A.M., Takeuchi, J.K., Hornberger, L.K., Zaidi, S.H., Amini, F., Coles, J., Bruneau, B.G. and Van Arsdell, G.S., 2009. NKX2-5 regulates the expression of β-catenin and GATA4 in ventricular myocytes. PloS one, 4(5).
34. Haworth, K.E., Kotecha, S., Mohun, T.J. and Latinkic, B.V., 2008. GATA4 and GATA5 are essential for heart and liver development in Xenopus embryos. BMC developmental biology, 8(1), p.74.
35. Srivastava, D. and Olson, E.N., 2000. A genetic blueprint for cardiac development. Nature, 407(6801), pp.221-226.
36. Brewer, A. and Pizzey, J., 2006. GATA factors in vertebrate heart development and disease. Expert reviews in molecular medicine, 8(22), pp.1-20.
37. Pikkarainen, S., Tokola, H., Kerkelä, R. and Ruskoaho, H., 2004. GATA transcription factors in the developing and adult heart. Cardiovascular research, 63(2), pp.196-207.