Revealing Exclusive Usage Of T-BOX Family Paralogous Transcription Factors Through Identifying Diversity In Expression Profiles During hiPSC-Derived Cardiomyocytes Generation

Anirban Bhar
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
anirban.bhar@bioinf.med.uni-goettingen.de

Martin Haubrock
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
martin.haubrock@bioinf.med.uni-goettingen.de

Edgar Wingender *
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
edgar.wingender@bioinf.med.uni-goettingen.de

MOTIVATION

Functional genomics aims to understand dynamic features encoded in the genome such as transcription of genes, thereby frequently using results from high throughput approaches. Transcription, RNA splicing and translation are the key steps in the process of gene expression. Production of a specific gene product can be increased or decreased by regulation of any of these steps. DNA microarrays are used to measure expression levels of a large number of genes simultaneously over a set of experimental conditions. In recent years, expression levels of thousands of genes are not only measured over sets of experimental conditions but also across many time points. To analyze such high throughput 3D datasets we need computational approaches. Coexpression analysis helps to retrieve functionally coherent group of genes that are often coregulated by a common transcription factor. Clustering, one of the unsupervised learning approaches can retrieve a group of genes having similar expression profiles over all experimental conditions. But it has been observed that genes are not necessarily to be coexpressed over all samples in a gene expression dataset, i.e.- genes can have similar expression profiles over a subset of samples. To simultaneously group genes and samples, biclustering or subspace clustering methods are used. However, biclustering algorithms fail to cluster genes, samples and time points simultaneously in a time series gene expression data. To cope with that problem triclustering algorithms are used. Zhao et al. proposed a triclustering algorithm TRICLUSTER to find groups of coexpressed genes in such time-series gene expression data set [1]. Tchagang et. al. recently proposed OPTricluster algorithm that is also able to cluster genes, samples and time points simultaneously [2]. One of the limitations of OPTricluster is that it can only cope with short time series gene expression datasets. In our previous work we have proposed triclustering algorithm δ-TRIMAX to mine such 3D gene expression datasets by introducing a novel definition of mean squared residue score for mining 3D datasets [3]. The goal of δ-TRIMAX is to retrieve maximal triclusters having mean squared residue score below a threshold δ. The limitations of δ-TRIMAX is that it is unable to extract overlapping triclusters. As δ-TRIMAX replaces each element of tricluster found in one iteration by random numbers, it can affect the originality of the dataset. In this paper we introduce the triclustering algorithm EMOA-δ-TRIMAX that can retrieve a group of genes that are coexpressed and coregulated over a subset of samples across a subset of time points. Here we have used Non-dominated Sorting Genetic Algorithm-II (NSGA-II) to balance the trade-off between the aforementioned conflicting objectives i.e. minimizing mean squared residue score, maximizing volume of the triclusters and generate pareto optimal solutions that are equally distributed in the objective space [4]. Additionally we have also maximized Spearman correlation coefficient of resultant triclusters. Our proposed algorithm also effectively deals with the drawbacks of our previously proposed algorithm δ-TRIMAX.

Regulation of transcription by transcription factors (TFs) can be initiated through binding to defined cis-regulatory elements in promoters. For accomplishing the function as an activator or inhibitor, TFs must recognize the regions where they should bind to and they do so through DNA-binding domains (DBD) [5]. A systematic classification of TFs according to their DBDs can help to predict the DNA-binding specificity of TFs with as yet ill-characterized DNA-binding properties. Paralogous transcription factors may have derived from a common ancestor by a gene duplication event and these transcription factors are assumed to participate in a novel function or some specialized ones of their original functions. Many of them still share major properties of their DBD and, thus, bind to identical or highly related cis-regulatory elements [5]. Mutation of the activation domain of paralogous transcription factors may yield alteration of their interacting partners in spite of having similar DNA-binding domains. Divergence of

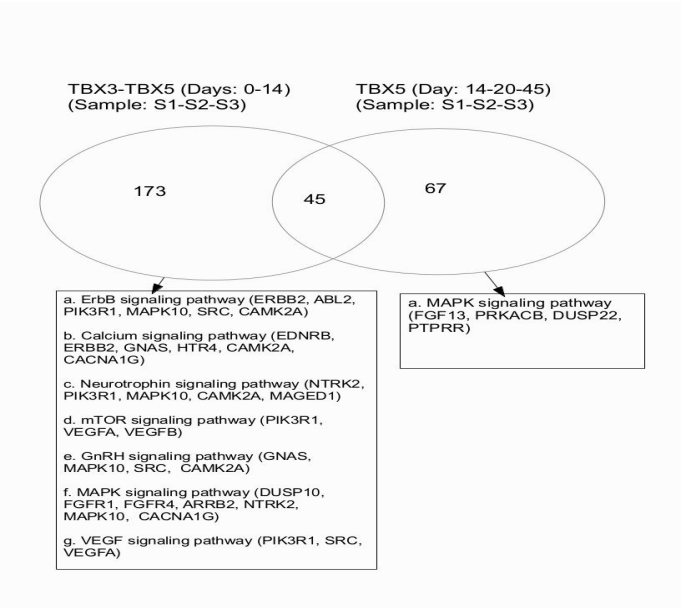

Figure 1: Differentially expressed targets of paralogous transcription factor across different subset of time points

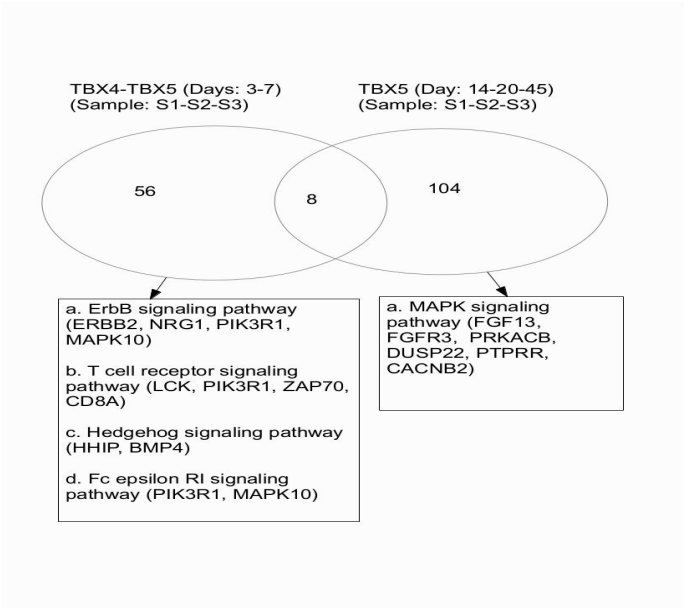

Figure 2: Differentially expressed targets of paralogous transcription factor across different subset of time points

expression profiles of paralogous transcription factors across tissues or time points can be a cause for participating in distinct pathways or regulating the same genes across different tissues or time points. For instance it has been previously reported that two paralogous transcription factors *Pax2* and *Pax3* regulate the gene c-Ret in kidney and neural crest, respectively [6]. Though recent works reveal roles of cardiac transcription factors in molecular regulation of pluripotent stem cell derived cardiomyocytes differentiation, the roles of cardiac paralogous T-Box family transcription factors are still poorly understood during different stages of cardiac differentiation.

RESULT

In this work we have applied our proposed EMOA-δ-TRIMAX algorithm on a time series gene expression dataset that contains mRNA expression profiles during differentiation of human induced pluripotent stem cell (hiPSC) derived cardiomyocytes. This dataset contains 48803 Illumina probe ids, 12 time points (day 0, 3, 7, 10, 14, 20, 28, 35, 45, 60, 90, 120) and 3 samples (GEO accession number GSE35671). Expression values at each time point were generated by three independent runs (Run 1-3) [7]. Our algorithm results in 100 triclusters that cover 88.14% of all probe-ids, 100% of all time points and 100% of all samples. We could show that EMOA-δ-TRIMAX outperforms other triclustering algorithms. It has been reported in the original work that the differentiation of hiPSCs to cardiomyocytes was observed during days 0, 3, 7, 10, 14, 20, 28 and on day 14 heart beating was first perceived. Days 35, 45, 60, 90 and 120 are reported as post-differentiation time points [7]. To establish biological significance of group of co-expressed genes, we checked for KEGG pathway and transcription factor binding site (TFBS) enrichment, the latter by using the TRANSFAC library (version 2012.2) [8]. We used an internal database of around 52 million TFBS predictions that have high affinity scores and are conserved between human, mouse, dog and cow [9]. Out of these 52 million conserved TFBSs we have selected the best 1% for each TRANSFAC matrix individually to select the most specific regulator (transcription factor) - target interactions. We have observed KEGG pathway and TFBS enrichment for 100% and 98% of resultant triclusters, respectively. Through our analysis we identified similar expression profiles of paralogous TFs *TBX3* and *TBX5* across days 0, 14 but divergence in their expression profiles across days 14, 20, 45 over all samples. Figure 1 shows that at early time points both *TBX3*, *TBX5* and at later time points only *TBX5* regulate target genes that participate in distinct sets of pathways. Additionally we observed that both *TBX3*, *TBX5* and only *TBX5* regulate MAPK signaling pathways through binding promoter regions of different target genes at early and later time points, respectively. We also observed similar expression profiles of paralogous transcription factors *TBX4* and *TBX5* across days 3, 7 but divergence in their expression profiles across days 14, 20, 45 over all samples. In Figure 2 we can observe that at early time points both *TBX4*, *TBX5* and at later time point only *TBX5* regulate distinct sets of genes that participate almost different signaling pathways. It has been reported in previous studies that ErbB, calcium, neurotrophin, VEGF, hedgehog signaling pathways play critical roles in cardiac differentiation and development [10–14]. It has been revealed in a previous study that *TBX5* plays a crucial role in

embryonic cardiac cell cycle progression and depletion of *TBX5* leads to cardiac programmed cell death [15]. Interestingly through our analysis we also observed that *TBX5* is expressed in both early and later time points.

CONCLUSION

Our integrated systems biology approach reveals exclusive usage of paralogous transcription factors of the T-BOX family through identifying diversity of their expression profiles and provides new insights into their roles in regulating cardiac differentiation.

REFERENCES

1. L. Zhao and M. J. Zaki, "Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 694–705, 2005. ISBN:1-59593-060-4.

2. A. Tchagang, S. Phan, F. Famili, H. Shearer, P. Fobert, Y. Huang, J. Zou, D. Huang, A. Cutler, Z. Liu, and Y. Pan, "Mining biological information from 3d short time-series gene expression data: the optricluster algorithm," *BMC Bioinformatics*, vol. 13, April 4 2012.

3. A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell," *Algorithms for Molecular Biology*, vol. 8, March 23 2013.

4. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

5. E. Wingender, T. Schoeps, and J. Dnitz, "Tfclass: an expandable hierarchical classification of human transcription factors," *Nucleic Acids Research*, vol. 41, no. D1, pp. D165–D170, 2013.

6. L. Singh and S. Hannenhalli, "Functional diversification of paralogous transcription factors via divergence in dna binding site motif and in expression," *PloSOne*, vol. 3, no. 6, 2008.

7. B. JE, R. M, S. S, R. P, S. B, B. H, W. T, C. E, C. U, and K. KL, "Determination of the human cardiomyocyte mrna and mirna differentiation network by fine-scale profiling," *Stem Cell Development*, vol. 21, pp. 1956–1965, July 20 2012.

8. E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhuser, M. Prss, F. Schacherer, S. Thiele, , and S. Urbach, "The transfac system on gene expression regulation," *Nucleic Acids Research*, vol. 29, pp. 281–283, January 1 2001.

9. X. Xie, J. Lu, E. Kulbokas, T. Golub, V. Mootha, K. Lindblad-Toh, E. Lander, , and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3 utrs by comparison of several mammals," *Nature*, vol. 434, no. 7031, pp. 338–345, 2005.

10. R. LF, "Neurotrophin-regulated signalling pathways," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, pp. 1545–1564, September 29 2006.

11. M. Bekhite, A. Finkensieper, S. Binas, J. Mller, R. Wetzker, H. Figulla, H. Sauer, and M. Wartenberg, "Vegf-mediated pi3k class ia and pkc signaling in cardiomyogenesis and vasculogenesis of mouse embryonic stem cells," *Journal of Cell Science*, vol. 124, pp. 1819–1830, June 1 2011.

12. J. Fu, H. Yu, R. Wang, J. Liang, and H. Yang, "Developmental regulation of intracellular calcium transients during cardiomyocyte differentiation of mouse embryonic stem cells," *Acta Pharmacologica Sinica*, vol. 27, pp. 901–910, 2006.

13. F. Rochais, K. Mesbah, and R. Kelly, "Signaling pathways controlling second heart field development," *Circulation Research*, vol. 104, pp. 933–942, 2009.

14. W. Zhu, Y. Xie, K. Moyes, J. Gold, B. Askari, and M. Laflamme, "Neuregulin/erbb signaling regulates cardiac subtype specification in differentiating human embryonic stem cells," *Circulation Research*, vol. 107, pp. 776–786, September 17 2010.

15. S. Goetz, D. Brown, and F. Conlon, "Tbx5 is required for embryonic cardiac cell cycle progression," *Development*, vol. 133, pp. 2575–2584, July 2006.