# TRANSCRIPTIONAL PROFILING OF DEFINITIVE ENDODERM DERIVED FROM HUMAN EMBRYONIC STEM CELLS

Huiqing Liu[1,2], Stephen Dalton[1], Ying Xu[1,2,*]

[1]*Dept. of Biochemistry and Molecular Biology and* [2]*Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA*

[*]*Email: xyn@bmb.uga.edu*

Definitive endoderm (DE), the inner germ layer of the trilaminar embryo, forms gastrointestinal tract, its derivatives, thyroid, thymus, pancreas, lungs and liver. Studies on DE formation in *Xenopus*, zebrafish and mouse suggest a conserved molecular mechanism among vertebrates. However, relevant analysis on this activity in human has not been extensively carried out. With the maturity of the techniques for monitoring how human embryonic stem cells (hESCs) react to signals that determine their pluripotency, proliferation, survival, and differentiation status, we are now able to conduct a similar research in human. In this paper, we present an analysis of gene expression profiles obtained from two recent experiments to identify genes expressed differentially during the process of hESCs differentiation to DE. We have carried out a systematic study on these genes to understand the related transcriptional regulations and signaling pathways using computational predictions and comparative genome analyses. Our preliminary results draw a similar transcriptional profile of hESC-DE formation to that of other vertebrates.

## 1. INTRODUCTION

During gastrulation, three primary germ layers (endoderm, mesoderm and ectoderm) are derived from the epiblast of human embryonic stem cells (hESCs). From these initial embryo layers, all the other somatic tissue types will develop. For instance, endoderm (the inner layer, also called as definitive endoderm, DE) forms gastrointestinal tract, its derivatives, thyroid, thymus, pancreas, lungs, and liver. Therefore, investigation of the biological mechanisms that occur during the hESCs differentiation will help us understand the developmental pathways involved in the formation of a mature organ.

In this study, we attempt to investigate transcriptional regulation and associated signaling pathways related to DE formation from hESCs. Although studies on DE formation in *Xenopus*, zebrafish, and mouse suggested a conserved molecular mechanism among vertebrates [7], relevant analysis on this activity in human has not been extensively done. Recently, several new methods for directing the differentiation of hESC towards DE have been investigated and two techniques were reported to have successfully directed DE formation [1, 5]. The core part of one technique, Ref. 1, is to first treat hESCs with Activin A in a low FCS (fetal calf serum) condition, and then enrich the culture by the DE cell surface marker CXCR4. Another technique, as described in our previous publication (Ref. 5), is to grow hESCs in mouse embryonic fibroblast conditioned medium (MEF-CM) under feeder free conditions with phosphatidylinositol 3-kinase signaling being suppressed. After five days, about 70-80% of the hESC culture is converted into DE. To compare the DE generated by the two techniques and to obtain an overall gene expression profile of this cell line, RNA samples from these two experiments are hybridized to the Affymetrix HG-U133 Plus 2.0 oligonucleotide microarray, which contains more than 54,000 probe sets, representing 38,500 human genes [5].

Studies on other vertebrates suggest that DE formation requires first the Activin/Nodal signaling of TGF$\beta$ (transforming growth factor $\beta$ super-family), followed by the activation of a set of downstream transcription factors (TFs) such as SOX17 of Sox (SRY-related HMG-box) family, FOXA2 (HNF3$\beta$) of Forkhead family and a number of TFs from the Gata family [7]. Manipulation of the Activin/Nodal ligands is done through a few transcription factors of the Smad family that lie at the core of the TGF$\beta$ pathway. Current understanding of this process is that, when the Activin/Nodal signaling protein meets its receptor, the highly homologous SMAD2 and SMAD3 intracellular mediators get phosphorylated on their conserved C-terminal motif and translocate

[*]Corresponding author.

with SMAD4 into the nucleus. SMAD2 or SMAD3 associates with SMAD4 to form a Smad complex which incorporates an additional DNA-binding cofactor to activate or repress the expression of the regulated genes [4].

## 2. DATA

Temporal gene expression profiles from the two experiments described in Ref. 1 and Ref. 5 are collected. Data are scaled to a median intensity with target setting of 500 and CEL files were normalized using probe quantile [5]. In both experiments, only those transcripts differentially expressed during DE formation are kept for further study. Table 1 shows the number of genes with certain fold changes on their expression levels during DE development in the experiments (data set I is from Ref. 5 and data set II is from Ref. 1). Seventy-five genes are selected, which exhibit substantial changes in both experiments are categorized in Table 1 of Ref. 5 according to their biological functions. Throughout the rest of this paper, we use this gene set for our data analysis studies unless stated otherwise. By looking at the functional (biological process) assignments of these genes according to the Gene Ontology (GO) database (http://www.geneontology.org/), we found that, 70% of these genes have GO term "cellular physiological process" (level 3 biological process), 50% have term "cell communication" and 48% are involved in "regulation of cellular process". In addition to the microarray data, we have also collected gene expression information using the quantitative polymerase chain reaction (Q-PCR) under the same protocol described in Ref. 5.

Table 1. Number of genes differentially expressed in DE formation from two microarray data sets. nf means n-fold change of expression level (n=2,4,6,8,10).

| Data Set | 2f | 4f | 6f | 8f | 10f |
|---|---|---|---|---|---|
| I | 3360 | 901 | 475 | 325 | 236 |
| II | 4168 | 1088 | 527 | 339 | 232 |
| Both | 1926 | 389 | 190 | 113 | 75 |

## 3. METHODS AND PRELIMINARY RESULTS

To analyze DE genes in a systematic manner, we present a number of studies in this section. For each study, we report the preliminary results that we have obtained as of now.

### 3.1. Markers of hESC-DE

Among the seventy-five genes identified above, we found all previously known markers of DE, namely SOX17, CXCR4, GSC, CER1, HHEX, FOXA2, GATA4 and GATA6. All these genes are upregulated during the differentiation. On the other hand, different from these genes, three indicators of the mesoderm (ME) patterning from mesendoderm, namely Brachyury (T), MIXL1 and FOXC1, all have their expression levels stop increasing in the middle of the DE formation (at ~24 hour or 36 hour; both microarray and Q-PCR data) and then drop sharply immediately after that turning point. This confirms that the differentiation is to DE, not to ME.

### 3.2. Transcription factor identification

To investigate the functional roles of the Smad family members and other TFs during the human DE formation, we analyzed the promoter regions of the obtained hESC-DE genes using computational tools and comparative genomics approach. Figure 1 describes the workflow of our procedure.

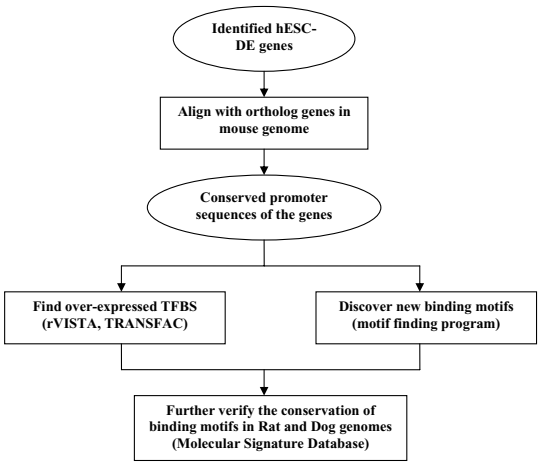


Fig. 1. A workflow for identifying transcription factors in hESC-DE.

We used the (whole genome) rVISTA tool to assist our work in this step, which is one of VISTA computational tools developed at Lawrence Berkeley National Laboratory (`http://genome.lbl.gov/vista/index.shtml`). By checking against the known transcription factor binding sites (TFBS) stored in TRANSFAC database, rVISTA is able to identify TFBS that are enriched in the promoter regions of a group of input genes and are conserved between pairs of species. In our study, we chose to scan 5,000 bps upstream region of each of our genes for possible conserved TFBS in the human genome against its counterpart in the mouse genome. The enrichment is measured by a $p$-value taking all upstream regions of human RefSeq5 genes as the background and the output is the corresponding TFs. The two top enriched TFBS returned by rVISTA are for Forkhead family members FOXO1 ($p < 10^{-22}$) and FOXO4 ($p < 10^{-21}$). FOXO is one of the identified DNA-binding cofactors of SMAD2/3-SMAD4 [4]. Furthermore, by scanning MSigDB (Molecular Signature Database; `http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html`), 25 out of the 75 hESC-DE specific genes are reported to have FOXO motifs conserved across the human, mouse, rat and dog genome. MSigDB stores all conserved transcription factor binding motifs derived from a recent comparative analysis of the four genomes [9].

Other top TFs whose binding sites are over-represented include E2F1DP1 ($p < 10^{-19}$), LEF1TEF1 ($p < 10^{-13}$), PITX2 ($p < 10^{-12}$), and TCF4 ($p < 10^{-10}$), suggesting the involvement of Wnt/$\beta$-catenin and Nodal signaling pathways in the DE formation. We observed that the Smad binding element (SBE) is not enriched in our data set. This is not surprising since the SBE sequence, 5'-GTCT-3' or its complement 5'-AGAC-3', is too short to be identified alone in the background with a large population of these 4-mers, by chance, in the genome.

To discover new binding motifs or those not captured by the existing popular TFBS databases, we applied CUBIC, a motif finding program developed by our group, to the relevant promoter regions (conserved between human and mouse) of the identified hESC-DE genes. CUBIC is an efficient tool to identify transcription factor binding sites via data clustering [6]. One motif identified by CUBIC is similar to the previously reported FOXH (FAST) binding site (CAATxxACA) [3]. FOXH is another known important cofactor of SMAD2/3-SMAD4 in response to the TGF$\beta$ signaling [4]. A sequence logo of this motif is given in Figure 2 drawn by Weblogo (`http://weblogo.berkeley.edu/`). In addition, several GC-rich motifs are also identified, which is consistent with the previously reported fact that Smad complexes recognize GC-rich regions in certain promoters [4]. Our results support the current general belief that in SMAD2/3-SMAD4 regulations, DNA-binding partners determine the choice of the target genes [4].

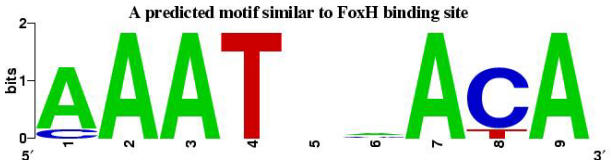


Fig. 2. A motif identified by CUBIC. It is similar to the reported binding site of FOXH.

### 3.3. Transcriptional regulation in the different phases of DE formation

Smad regulation is the first response to the TGF$\beta$ signaling in the DE formation. In order to further direct the differentiation to DE, several downstream transcription factors are also required. SOX17 and FOXA2 are among the ones that have been previously reported [8]. To identify the transcription factors that function in different phases during the DE formation, we have clustered genes based on the similarities of their expression profiles and attempted to find the "dominant" regulator(s) for each gene cluster. Figure 3 shows that the genes are grouped into five clusters. Most genes in Cluster 1 start to get up-regulated from ~72hour. Genes in Cluster 2 get up-regulated around 48hour while genes in Cluster 4 start from ~24/36hour. Besides, genes in Cluster 3 represent "early response" genes, which start to increase their expression levels in the very early phase, and most of them have a decreasing pattern in the later phases during the DE formation. We notice that two of them are previously known ME markers, i.e. MIXL1 and FOXC1. Different from these four clusters, Cluster 5 consists of genes that are down-regulated during the DE differentiation.

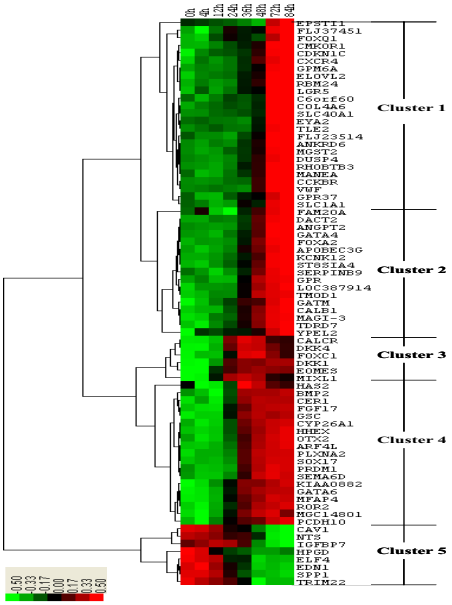


Fig. 3. Clustering on temporal expression profile of hESC-DE genes. Hierachical clustering are performed by Cluster3.0 on log transformed and normalized data. This software can be accessed from http://bonsai.ims.u-tokyo.ac.jp/ mdehoon/software/cluster/software.htm.

When we applied rVISTA to each cluster of genes, we found additional over-represented TFBS. Particularly, the binding sites of FOXO1 and FOXO4 are ranked as the top TFBS in Clusters 3 and 4 genes, confirming the Smad regulation in the early phases during the DE formation. Binding sites of MEIS1 (Meis homeobox 1) ($p < 10^{-3}$) and TITF1 (thyroid transcription factor) ($p < 10^{-3}$) are the two top TFBS for genes in Cluster 5. Although the expression profile of MEIS1 does not show substantial changes in our data, its overall trend is clearly decreasing (data not shown). In addition, previous studies in vertebrates have shown the involvement of TITF1 in the organogenesis of thyroid, lung and some areas of the forebrain [2]. The best TFBS from Cluster 2 genes is *muscle initiator sequence 20* ($p < 10^{-5}$) and for Cluster 1 is POU6F1 (POU domain, class 6, transcription factor 1) ($p < 10^{-3}$). This may imply the involvement of these two TFs in later phase of DE formation.

## 4. ON-GOING WORK AND CONCLUSION

Since SOX17 is not recorded in TRANSFAC, we cannot get any TFBS information of this transcription factor by using rVISTA. We are currently employing other computational methods to identify its target genes and the potential binding sites from the transcripts in Cluster 1 and Cluster 2 of Figure 3. Meanwhile, some of our discoveries reported in this paper are under wet lab verification. After a clear picture of transcriptional regulations is disclosed, we will investigate more on signaling pathways related to hESC-DE. In summary, we have applied computational tools and comparative genomics approach to analyze temporal gene expression data of hESC-DE formation. Our preliminary results demonstrate that the biological mechanism of DE differentiation in human is similar to that of the other vertebrates.

## Acknowledgment

HL and YX's work is supported in part by Georgia Cancer Coalition and NSF IIS-0407204. SD's work is supported by Georgia Research Alliance

## References

1. D'Amour KA, et al. (2005) Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat Biotechnol*, **23**(12):1534-4.
2. Devriendt K, et al. (1998) Deletion of thyroid transcription factor-1 gene in an infant with neonatal thyroid dysfunction and respiratory failure. *(Letter) New Eng. J. Med.* **338**: 1317-1318.
3. Labbe E, et al. (1998) Smad2 and Smad3 positively and negatively regulate TGF beta-dependent transcription through the forkhead DNA-binding protein FAST2. *Mol Cell.* **2**(1):109-20.
4. Massagu J, Seoane J, Wotton D. (2005) Smad transcription factors. *Genes Dev.* **19**(23):2783-2810.
5. McLean AB, et al (2007) Activin A efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3-kinase signaling is suppressed. *Stem Cells*, **25**(1):29-38
6. Olman V, Xu D, Xu Y. (2003) CUBIC: identification of regulatory binding sites through data clustering, *JBCB* **1**(1): 21-40.
7. Stainier DY. (2002) A glimpse into the molecular entrails of endoderm formation. *Genes Dev.* **16**(8):893-907.
8. Sinner D, et al. (2004) Sox17 and $\beta$-catenin cooperate to regulate the transcription of endodermal genes. *Development*, **131**(13):3069-80.
9. Xie X, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**(7031): 338-345.