BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 16 2006, pages 2005–2011
doi:10.1093/bioinformatics/btl343

Systems Biology

Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation

Je-Gun Joung[1,†], Dongho Shin[3,†], Rho Hyun Seong[3,*] and Byoung-Tak Zhang[1,2,*]

[1]Center for Bioinformation Technology, [2]School of Computer Science and Engineering and [3]Research Center for Functional Cellulomics, Institute of Molecular Biology and Genetics and Department of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

Received on February 26, 2006; revised on May 10, 2006; accepted on June 20, 2006

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: An important issue in stem cell biology is to understand how to direct differentiation towards a specific cell type. To elucidate the mechanism, previous studies have focused on identifying the responsible gene regulators, which have, however, failed to provide a systemic view of regulatory modules. To obtain a unified description of the regulatory modules, we characterized major stem cell species by employing a co-clustering latent variable model (LVM). The LVM-based method allowed us to elucidate the cell type-specific transcription factors, using genomic sequences as well as expression profiles.

Results: We used a list of genes enriched in each of 21 stem cell subpopulations, and their upstream genomic sequences. The LVM-based study allowed us to uncover the regulatory modules for each stem cell cluster, e.g. GABP and E2F for the proliferation phase, and Ap2α and Ap2γ for the quiescence phase. Furthermore, the identities of the stem cell clusters were well revealed by the constituent genes that were directly targeted by the modules. Consequently, our analytical framework was demonstrated to be useful through a detailed case study of stem cell differentiation and can be applied to problems with similar characteristics.

Contact: btzhang@bi.snu.ac.kr, rhseong@snu.ac.kr

Supplementary Information: Supplementary data are available at http://bi.snu.ac.kr/Publications/LVM_SC/.

1 INTRODUCTION

To make good use of stem cells in clinical applications, it is necessary to comprehend the mechanisms by which stem cells operate. Among the diverse investigations into this, it is crucial to identify core stem cell regulators. Stem cells are quite different from other cell types in nature, especially in the aspect of pluripotency and self-renewing capability. Thus, the transcriptional profiles of stem cells are expected to provide the molecular evidence that may account for stem cell character. The three most intensely studied stem cell species [embryonic (ESCs), neural (NSCs) and hematopoietic stem cells (HSCs)] were previously analyzed for gene transcription (Ivanova *et al*., 2002; Ramalho-Santos *et al*., 2002; Venezia *et al*., 2004). The data provide useful source material for identifying stem cell regulatory networks.

Since the core properties of stem cells are likely to be shared by various stem cell species, a circuitry of global regulation as well as the gene regulators specific to individual stem cell species may also exist. An underlying assumption is that stem cell genes are controlled by the gene regulators that mediate phenotypic changes. Among them, transcription factors (TFs) have been reported to play a major part in lineage commitment and stage progression of stem cells, by directly modulating patterns of gene expression (Reid, 1990). Furthermore, several master regulators turn on additional transcription factors that are responsible for activating entire networks of genes necessary for generating many different specialized cells and tissues (Boyer *et al*., 2005).

As high-throughput technologies such as microarrays are introduced, it is possible to measure the abundance of mRNA on a whole-genome scale. Previously, molecular studies for understanding the stem cell character have usually depended on the data obtained from large-scale gene expression analysis. However, most of the results were confined to the own interests of the researchers. In addition, they barely provided critical evidence for the gene regulation mechanisms. This may be attributed to the lack of a decisive method of identifying genuine regulators out of a number of candidate genes.

We propose an approach based on co-clustering latent variable models (LVMs) to identify stem-cell-specific regulatory modules from integrated experimental datasets. The LVMs have been quite successful in detecting hidden patterns in biological profiling data (Zhang *et al*., 2003; Flaherty *et al*., 2005). We adapted the probabilistic latent semantic model (PLSA) (Hofmann, 2001), which is one of the LVMs, to cluster simultaneously both rows and columns of a subpopulation-TF binding site (TFBS) matrix. Compared with the standard clustering algorithms such as *k*-means and hierarchical clustering (Eisen *et al*., 1998), the co-clustering LVMs can reveal more flexibly the association between two objects (i.e. rows and columns) (Bishop, 1999). Moreover, since most co-clustering algorithms known as hard clustering techniques (Madeira *et al*., 2004) work on the basis of mutual exclusivity (Flaherty *et al*., 2005), they seem inappropriate to represent the biological regulatory systems that frequently share the core elements. On the other hand, the co-clustering LVM is an effective algorithm in that it does not only permit an element to belong to several different clusters but also finds the modular structure that is constituted by a highly

*To whom correspondence should be addressed.
†The authors wish it to be known that 'in their opinion' the first two authors should be regarded as joint First Authors.

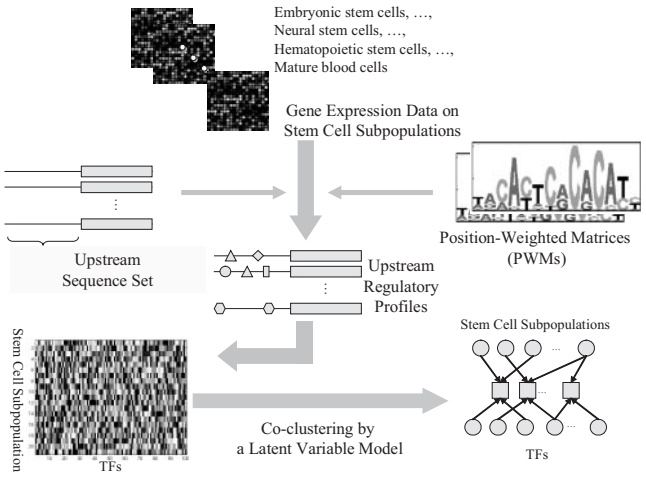

Fig. 1. Schematic flow diagram of co-clustering stem cell subpopulations and transcriptional regulators.

probable relationship between objects. Here we define the regulatory module of stem cells as a set of transcriptional regulators specified to the individual stem cell species.

By the integrative analysis of multiple experiments, our work will contribute to effectively identifying stem cell regulatory modules. Several representative regulators were retrieved in each stem cell cluster, and their relationships showed high relevance to the biological literature. The Gene Ontology of the regulated genes also supported the predicted relationships. In addition, the modularity was validated by the expression coherence of the regulated genes. In this report, we provide a comprehensive map of regulatory mechanisms of the major stem cells.

2 METHODS

The overall scheme of the strategy is illustrated in Figure 1. We attempt to cluster TFBSs and stem cell subpopulations simultaneously. First, the gene sets representing major stem cell populations were collected from microarray data (which will be described in detail). Next, the dataset of upstream sequences was extracted from murine genomic archives. The dataset was searched for *cis*-acting elements that were in the form of position-weighted matrices (PWMs). Then, a stem cell subpopulation-TFBS matrix was generated and clustered by the latent variable models.

2.1 Collection of gene sets representing major stem cell populations

Our collection of gene sets representing various stem cell populations was gathered from gene sets selected previously by three research groups based on a significant fold change from expression profiles of major murine stem cell populations (Table 1; Ramalho-Santos, 2002; Ivanova, 2002; Venezia, 2004). Details for each selected gene set are addressed in the supplementary data of the three references. There are 21 subpopulations categorized by cell phenotypes or development stages, which comprise three major stem cell sets, HSC maturation sets and HSC cell cycle sets. Two stemness sets were also obtained from the datasets of murine ESC, NSC and HSC, which were produced by two prominent research groups. Depending on the maturation status, the HSCs were subdivided into several subsets. In addition, the genes activated during the proliferation phase of HSCs were comparatively analyzed with the deactivated genes. The former belong to the subpopulations FL, ST and PG, the latter to BM, LT and QG.

Table 1. Stem cell subpopulations

Subpop.	Description	#Gene
SE	Embryonic stem cells (Ramalho-Santos, 2002)	1465
SN	Neural stem cells (Ramalho-Santos, 2002)	1930
SH	Hematopoietic stem cells (Ramalho-Santos, 2002)	1731
IE	Embryonic stem cells (Ivanova, 2002)	1459
IN	Neural stem cells (Ivanova, 2002)	1274
IH	Hematopoietic stem cells (Ivanova, 2002)	1484
SS	Stemness (Ramalho-Santos, 2002)	206
IS	Stemness (Ivanova, 2002)	97
SU	Stemness (Ramalho-Santos and Ivanova, 2002)	422
LT	Long-term HSCs (Ivanova, 2002)	578
ST	Short-term HSCs (Ivanova, 2002)	226
BM	Bone marrow stem cells (Venezia, 2004)	922
FL	Fetal liver stem cells (Venezia, 2004)	749
QG	Quiescence group (Venezia, 2004)	718
PG	Proliferation group (Venezia, 2004)	609
QS	Quiescence signature (Venezia, 2004)	271
PS	Proliferation signature (Venezia, 2004)	304
EP	Early progenitors (Ivanova, 2002)	498
IP	Intermediate progenitors (Ivanova, 2002)	170
LP	Late progenitors (Ivanova, 2002)	549
MBC	Mature blood cells (Ivanova, 2002)	488

There are 21 subpopulations categorized by cell phenotypes or development stages, which comprise three major stem cell sets, mHSC maturation sets and HSC cell cycle sets.

2.2 Screening transcription factor binding motifs

To extract the dataset of upstream regulatory sequences, murine genomic archives (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm5/) were retrieved. They contain 17 848 of the murine RefSeq and 41 208 entries for the KnownGene. After subsequent refining processes, we obtained 23 346 independent entries. With transcription start sites (TSSs) of these genes, upstream sequences were extracted from the mouse genome (assembly mm5). The 5 kb upstream regions of the 23 346 genes were extracted using a standalone version of BLAT (Kent, 2002).

We used 360 PWMs of the TRANSFAC *r*8.3 (Matys *et al*., 2003) to extract TFBSs in mouse upstreams. The putative TFBSs on each sequence were scanned by the program Patser (Hertz and Stormo, 1999), and matching positions were returned and scored. Patser was run with the following command line options: '–A a:t 0.275 c:g 0.225 –c –lp –13.0'. Here, the –A was used to provide the following background frequencies: A/T = 0.275, G/C = 0.225. –c is for scoring the complementary strand, and –lp is to determine the lower threshold score from a maximum ln (*p*-value). A detailed procedure is given in Supplementary Material.

2.3 Co-clustering latent variable models

Our goal was to cluster stem cell subpopulations and TFBSs from the given matrix simultaneously. We assume that the matrix consists of weights of m TFBSs in n subpopulations. When the dataset is represented by a set of m of TFBSs, $T = \{ t_1, t_2, \ldots, t_m \}$, and a set of n subpopulations, $S = \{ s_1, s_2, \ldots, s_n \}$, it can be viewed as an subpopulation-TFBS matrix, $\mathrm{ST} = [w(t_i, s_j)]$. Here $w(t_i, s_j)$ denotes the weight of the i-th TFBS in the j-th subpopulation. Each weight indicates the measure that the i-th TFBS affects j-th subpopulation. We calculate the ratio between the occurrence probability of the subpopulation and that of the total set using the following equation:

$e_{ij} = (f_{ij}^{\mathrm{Sub}}/N^{\mathrm{Sub}})/(f_{ij}^{\mathrm{Total}}/N^{\mathrm{Total}})$. Here $\#N^{\mathrm{Total}}$ and $\#N^{\mathrm{Sub}}$ are the number of total genes and subpopulation genes, respectively. f_{ij}^{Total} and f_{ij}^{Sub} are the frequencies of TFBSs observed in the entire set and the subpopulation, respectively. The matrix has values from 0 to 6.7.

We assume that there exists a set of hidden (unobserved) factors underlying the co-occurrences among sets of subpopulation and TF. Introducing latent factors $\#Z=\{\#z_1, \#z_2, \ldots, \#z_l\}$, the model measures the relationships between TFBSs and hidden factors, as well as between subpopulations and hidden factors. We use a modified version of the PLSA model to identify these relationships (Hofmann, 2001). First we describe the following probability definition for a generative model: (1) $\#P(\#t_i)$ is the probability that a TFBS will be observed in $\#T$. (2) $\#P(\#z_k \mid t_i)$ is a TFBS-specific probability distribution on latent factor $\#z_k$. (3) $\#P(\#s_j \mid z_k)$ is the probability of subpopulation over latent factor $\#z_k$. Based on these definitions, we obtain the probability of an observed pair ($\#t_i$, s_j) by adopting the latent factor $\#z_k$ as:

$$P(t_i, s_j) = P(t_i)P(s_j \mid t_i),$$

where

$$P(s_j \mid t_i) = \sum_{k=1}^{l} P(s_j \mid z_k)P(z_k \mid t_i).$$

Using Bayes' rule, the joint probability can be rewritten as

$$P(t_i, s_j) = \sum_{k=1}^{l} P(z_k)P(s_j \mid z_k)P(t_i \mid z_k).$$

In order to find the above parameters, we maximize the total likelihood of observations:

$$L(T, S) = \sum_{i=1}^{m} \sum_{j=1}^{n} w(t_i, s_j) \log P(t_i, s_j).$$

The standard procedure to estimate maximum likelihood parameters is the Expectation–Maximization (EM) algorithm. The EM algorithm starts with random initial parameter values of $P(z_k)$, $P(s_j \mid z_k)$ and $P(t_i \mid z_k)$. Then the algorithm iterates both an expectation step (E-step) and a maximization step (M-step) alternately until a certain convergence criterion is satisfied. In the E-step we compute:

$$P(z_k \mid t_i, s_j) = \frac{P(z_k)P(t_i \mid z_k)P(s_j \mid z_k)}{\sum_{k'=1}^{l} P(z_{k'})P(t_i \mid z_{k'})p(s_j \mid z_{k'})}$$

and the M-step is as follows:

$$P(z_k) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} w(t_i, s_j)P(z_k \mid t_i, s_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k'=1}^{l} w(t_i, s_j)P(z_{k'} \mid t_i, s_j)},$$

$$P(t_i \mid z_k) = \frac{\sum_{j=1}^{n} w(t_i, s_j)P(z_k \mid t_i, s_j)}{\sum_{i'=1}^{m} \sum_{j=1}^{n} w(t_{i'}, s_j)P(z_k \mid t_{i'}, s_j)},$$

$$P(s_j \mid z_k) = \frac{\sum_{i=1}^{m} w(t_i, s_j)P(z_k \mid t_i, s_j)}{\sum_{i=1}^{m} \sum_{j'=1}^{n} w(t_i, s_{j'})P(z_k \mid t_i, s_{j'})}.$$

After the total likelihood $L(S,T)$ of the observation data increases monotonically by E-step and M-step, it converges to a local optimum solution.

3 RESULTS

3.1 Co-clustering: stem cell subpopulations and transcriptional regulators

We obtained a co-clustering profile from the input dataset by co-clustering LVM. The input dataset was generated in the form of a [292 × 21] matrix that is referred to subpopulation/TFBS vectors. Of 360 PWMs, 292 were found to have at least one putative binding site on the whole promoters.

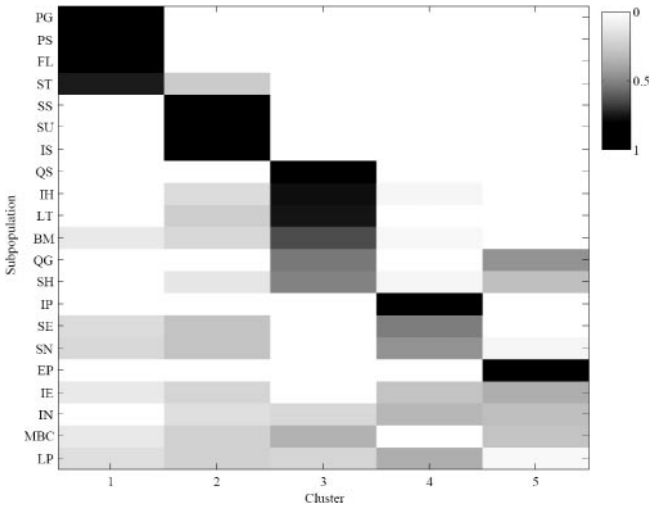

Fig. 2. Clusters based on latent (hidden) variables. The cell subpopulation is assigned by probability belonging to several clusters. Each block represents the normalized fraction in the cluster. The darker block indicates a higher probability. Each cluster exhibits representative subpopulations: Cluster 1 is closely related to subpopulations associated with mHSC proliferation; Cluster 2, stemness related subpopulations; Cluster 3, quiescence phase of mHSC; Clusters 4 and 5, hematopoietic progenitors.

Figure 2 shows the result obtained by running the co-clustering LVM with the number of clusters set to 5. Here the number of clusters was determined so that it satisfies a necessary and sufficient condition for distinguishing each characteristic cell status, namely three major murine undifferentiated stem cells, stemness, proliferation and quiescence phase of HSCs, and the HSC maturation processes. Cluster 1 represents the subpopulations associated with proliferation of HSCs that includes the PS, PG, FL and ST. This result is perfectly consistent with the previous observation that the four subpopulations were characterized by the genes that were activated during HSC proliferation (Venezia *et al*., 2004). Stemness subpopulations are all allocated in Cluster 2. It may suggest the possibility that the two stemness gene sets may be regulated by common transcriptional regulators, although the two sets share only a few out of hundreds of genes in common in the microarray data.

Cluster 3 contains those related to quiescence in cell cycle of HSCs, namely the QS, LT, BM and QG. Besides, this cluster also has two subpopulations representing whole HSCs, suggesting that the general characteristics of HSCs may mainly display quiescence rather than their proliferation properties. Clusters 4 and 5 represent two hematopoietic progenitors. IP is assigned to Cluster 4 with two non-HSC subpopulations from the Ramalho-Santos data. This means that the transcriptional regulators may be shared between the unrelated adult stem cells, possibly suggesting their *trans*-differentiation potential.

Table 2 shows the top-ranked 5% of TFBSs for each cluster sorted by decreasing probability. Each cluster shows the representative TFBS profile, with some TFBSs being found in more than one cluster. Five TFBSs are assigned with probability >0.08 in three

Table 2. List of TFBSs ranked within top 5% in cluster

	I	II	III	IV	V
1	*E2F*	*GATA6*	*Pax6*	*Tax/CREB*	*Ncx*
2	*E2F1*	*CREB*	*Pax8*	*RFX1*	*GATA6*
3	*GATA6*	*AhR:Arnt*	*AP2α*	*HIF1*	*c-Myc:Max*
4	*HSF1*	*HSF2*	*AP2γ*	*MEF2*	*HNF1*
5	*GABP*	*Pax5*	*Whn*	*ATF4*	*STAT5A*
6	*Pax5*	*Egr2*	*NFκB*	*IPF1*	*Ebox*
7	*c-Ets1(p54)*	*c-Myc:Max*	*Oct1*	*E2F*	*USF*
8	*Tst1*	*ATF1*	*TATA*	*TATA*	*c-Myb*
9	*Pax8*	*HIF1*	*N-Myc*	*c-Myb*	*GCM*
10	*NF-Y*	*Pax8*	*Pax1*	*CREB*	*E12*
11	*CEBP*	*AP2*	*c-Myb*	*HSF*	*NF1*
12	*USF*	*AhR*	*Arnt*	*C/EBPα*	*SRF*
13	*HES1*	*GATA1*	*AP1*	*CRE-BP1*	*CREB*
14	*c-Myc:Max*	*Oct1*	*CREB*	*N-Myc*	*CRE-BP1*
15	*FOXJ2*	*c-Myb*	*AP2*	*GABP*	*p53*

The corresponding TFs affect gene expression in each cluster. Each cluster shows a distinct TFBS representation.

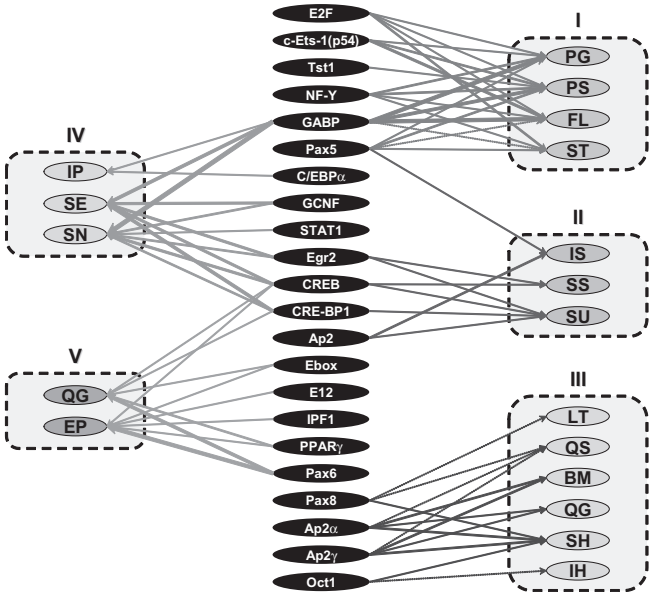

Fig. 3. Stem cell regulatory modules. The links between stem cell subpopulations and TFBSs in clusters are made according to *p*-value, which was calculated by hypergeometric probability law. Links of *p*-value < 0.015 were selected to depict the diagram. Out of 75 TFBSs showing high probability ($P(t_i \mid z_k) > 0.008$) 22 have *p*-value < 0.015. The strength of the links is indicated by the relative thickness of the lines. Based on the thickness, *p*-values for the solid lines correspond to $p < 10^{-2}$, $p < 10^{-3}$ and $p < 10^{-5}$ respectively and the dotted line to $p > 10^{-2}$.

clusters: *Egr2, Gata6, Pax8, CREB* and *CRE-BP1*. A total of 14 TFBSs are in two clusters: *Oct1, FoxO1, BSAP, Egr1, E2F, IPF1, STAT3, c-Myb, HIF1, GATA1, GABP, N-Myc* and *c-Myc:Max*.

3.2 Identifying stem cell regulatory modules

Using the resultant clustering information, we constructed a relational network for TFBSs and stem cell subpopulations. Conditional dependencies between the stem cell subpopulations and TFBSs were incorporated in the model. Significant links were selected by *p*-value cutoff: only 22 of 75 TFBSs, showing high probability ($P(t_i \mid z_k) > 0.008$, the minimum cutoff for allocating at least 10 TFBSs to each cluster), have *p*-value < 0.015 (the minimum cutoff for at least one link to each subpopulation). Generally, a few key TFs are believed to have a major influence on controlling cell fate, though more TFs should be identified to fully characterize a cell state. As each cluster represents a specific status of stem cell differentiation, a candidate pool comprising more than 10 TFs will be sufficient to explain a status of stem cells. At the selected threshold, the regulatory modules were well illustrated as in Figure 3. It depicts stem cell regulatory modules that were reconstructed based on the statistical significance of the co-clustering data. This network may provide core TFs representing or regulating stem cell subpopulations. Table 3 presents the summarized descriptions on the representative TFs regulating each cluster.

Cluster 1: proliferation phase in HSC maturation. *GABP* showed relatively higher significance in this cluster. Embryos with null *GABPα* allele die before implantation, being consistent with the broad expression throughout embryogenesis and in ESCs (Ristevski *et al*., 2004). During liver regeneration, the expression of *GABPα*/*GABPβ* heterodimer increased considerably (Du *et al*., 1998). These results provide a clue that *GABPα* may play essential roles in the proliferation of diverse tissues including HSCs. *E2F* family TFs are well known to play essential and redundant roles in the proper coordination of cell-cycle progression. The combined loss of *E2F1* and *E2F2* in mice leads to profound cell-autonomous defects in the hematopoietic development of multiple cell lineages (Li *et al*., 2003). *NF-Y* is essential for the recruitment of RNA polymerase II onto *E2F1* promoter (Kabe *et al*., 2005). From its *E2F1* activating potential, it is likely that *E2F* is activated by *NF-Y* during embryogenesis or tissue development.

Cluster 2: stemness. The pooled stemness set SU is shown to be regulated by *Egr2, CREB, CRE-BP1* and *Ap2*. A signal from insulin/IGFs to *CREB* determines cell size and animal size during embryogenesis (Sordella *et al*., 2002). *CRE-BP* is crucial for HSC self-renewal. Meanwhile, its paralogue p300 is essential for proper hematopoietic differentiation (Rebel *et al*., 2002). Interactions between *Ap2α* and p300/ *CRE-BP* are necessary for *Ap2α*-mediated transcriptional activation (Braganca *et al*., 2003). *Ap2* is responsible for maintaining proliferative and undifferentiated states of cells, which are important for embryonic development and in tumorigenesis (Jager *et al*., 2003).

Cluster 3: quiescence phase in HSC maturation. *Oct1*-deficient embryos die during gestation, frequently appear anemic and suffer from a lack of erythroid precursor cells (Wang *et al*., 2004). On the other hand, tissue-specific expression of *Oct1* isoforms in lymphocytes may be related to B- and T-cell differentiation and expression of the immunoglobulin genes (Pankratova *et al*., 2001). This evidence indirectly supports the possibility that *Oct1* may function specifically during the differentiation of HSCs to a variety of sublineages, not in HSC repopulations.

Cluster 4: intermediate progenitors in HSC maturation. HSC repopulating and self-renewal capacity is enhanced in the absence of *C/EBPα*. Disruption of *C/EBPα* blocks the transition from the common myeloid to granulocyte or monocyte progenitors (Zhang *et al*., 2004). It also results in hyperproliferation of hematopoietic

Table 3. The representative TFBSs allocated in each cluster

TF	RefSeq ID	Reported biological process
Cluster 1: proliferation phase in mHSC maturation		
GABP	NM_008065	embryogensis; liver regeneration
E2F	NM_007891	hematopoietic development; S phase progression
Ets1	NM_011808	UV-induced apoptosis in ESCs
NF-Y	NM_010913	recruits RNA pol II onto E2F1 promoter
Cluster 2: stemness		
CREB	NM_009952	determines cell size and animal size during embryogenesis
CRE-BP1	NM_001025432	HSC self-renewal
Ap2	NM_011547	orchestrates embryonic development by influencing the differentiation, proliferation, and survival of cells
Cluster 3: quiescence phase in mHSC maturation		
Oct1	NM_011137	differentiation of erythroid lineage, B- and T-cell
Cluster 4: intermediate progenitors in mHSC maturation		
C/EBPα	NM_007678	transition from the common myeloid to the granulocyte/ monocyte progenitor
GCNF	NM_010264	early mouse embryogenesis; differentiation and maturation of neuronal precursor cells
STAT1	NM_009283	protects against IFNα-mediated injury in CNS
CREB	NM_009952	maintenance of neural cells
Cluster 5: early progenitors in mHSC maturation		
E2A	NM_011548	B-cell development beyond the progenitor cell stage

The stem-cell-related biological processes are specified here.

Table 4. Featured biological processes enriched in each cluster

GO ID	Adjusted *p*-value	Biological process
Cluster I: PG/PS/FL/ST		
GO:0007059	3.53E−02	Chromosome segregation
GO:0006263	3.19E−05	DNA-dependent DNA replication
GO:0006270	7.30E−05	DNA replication initiation
Cluster III: LT/QS/BM/QG/SH/IH		
GO:0030154	1.82E−02	Cell differentiation
GO:0007275	1.82E−02	Development
Cluster IV: IP/SE/SN		
GO:0051297	7.08E−04	Centrosome organization and biogenesis
GO:0006454	4.96E−03	Translational initiation
GO:0000279	1.37E−03	M phase
GO:0051301	2.13E−03	Cell division
GO:0007098	7.08E−04	Centrosome cycle
GO:0008104	9.77E−06	Protein localization
GO:0045184	3.22E−06	Establishment of protein localization
GO:0031023	1.46E−03	Microtubule organizing center organization
GO:0006997	3.18E−02	Nuclear organization and biogenesis
GO:0007000	1.97E−02	Nucleolus organization and biogenesis
GO:0007100	1.97E−02	Mitotic centrosome separation
GO:0007067	4.17E−03	Mitosis
Cluster V: QG/EP		
GO:0045321	3.87E−02	Immune cell activation
GO:0046650	3.71E−02	Lymphocyte differentiation
GO:0046649	2.02E−02	Lymphocyte activation
GO:0042089	1.75E−02	Cytokine biosynthesis

The GO terms were searched for the target genes with the binding sites for the corresponding transcriptional regulators. The overrepresented terms were selected by the hypergeometric test and multiple testing adjustment using FDR procedure ($p < 0.05$).

progenitor cells (Heath *et al*., 2004). Thus, these results support a role for *C/EBPα* in the differentiation of cells in the IP stage. Meanwhile, *GABP, GCNF, Egr2, CREB* and *CRE-BP1* are enriched in ESC and NSC. *GABP*, which was emphasized in Cluster 1, is also associated with ESC and NSC. This may imply its global function in various stem cell species. *GCNF, STAT1* and *CREB* are widely known to be involved in neural development. The level of *GCNF* is critical for differentiation and maturation of neuronal precursor cells (Sattler *et al*., 2004). The brain in GFAP-IFNα mice lacking *STAT1* had neurodegeneration, inflammation and calcification with apoptosis (Wang *et al*., 2002). Mice lacking *CREB* in the CNS during development show extensive apoptosis of postmitotic neurons (Mantamadiotis *et al*., 2002).

Cluster 5: Early progenitors in HSC maturation. E12 is a member of the E 2A TF family. *E2A*-deficient hematopoietic progenitor cells reconstitute the T, NK, myeloid, dendritic and erythroid lineages but fail to develop into mature B cells. *E2A*-deficient hematopoietic progenitor cells remain pluripotent after long-term culture *in vitro*, and *E2A* proteins play a critical role in B-cell commitment (Ikawa *et al*., 2004). This suggests that the upregulated *E2A* in the early progenitor stage may be responsible for leading the repopulating HSCs to the B-cell differentiation pathway.

3.3 Functional correlation and expression coherence of target genes

If stem cell subpopulations and gene regulators are closely co-clustered, the target genes controlled by their corresponding regulators may reflect the relevance of the co-clustering data. We extracted GO terms for target genes (Table 4) using BiNGO (Maere *et al*., 2005). As a whole, the target genes in each cluster apparently belong to characteristic functional categories.

Clusters 1 and 4 are similar in that they cover relatively large numbers of target genes involved in cell cycle progression. Meanwhile, the two clusters seem to have differential features, namely, the former is related to chromosome duplication and the latter to mitotic cell division including cytokinesis. SE and SN may be most responsible for the cell cycle properties. On the other hand, Clusters 3 and 5 show quite different characteristics. Although Cluster 3, as well as Cluster 1, also comprises HSC subpopulations, the GO terms 'cell differentiation' and 'development' may clearly distinguish its character from that of Cluster 1 representing the self-replenishing property of HSCs. According to the data presented in Table 4, 'lymphocyte differentiation and activation' in Cluster 5 may be induced by 'cytokine production' during hematopoiesis. This assumption is strongly supported by the 'cell differentiation' property of the QG and EP (early hematopoietic progenitor) in Cluster 3.

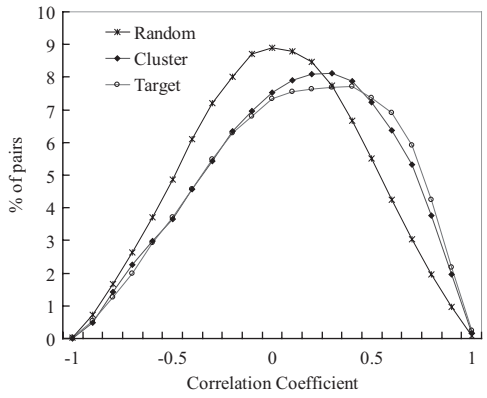

Fig. 4. The expression coherence. The distribution of the correlation coefficients for the Random (randomly selected genes), Cluster (genes in each cluster) and Target (the target genes containing the corresponding TFBSs) are shown in the graph. The *x*-axis indicates the bin intervals of 0.1 and the *y*-axis the percentage of the gene pairs in each bin.

We could not find any significant terms for Cluster 2 satisfying the *p*-value threshold ($p < 0.05$).

To validate further the modularity, we examined whether the targeted gene group has coherent gene expression. Figure 4 shows the 'bin' distribution of the correlation coefficients for Random, Cluster and Target gene sets. In the result, the correlative associations were stronger between the target genes within an individual module than between the non-modularized genes. As shown in the figure, the correlation curve of the target genes is shifted to the right compared with the others. This indicates that genes in a module show higher co-expression behavior.

4 DISCUSSION AND CONCLUSION

Stem cells are regarded as the cutting edge with regard to their practical application. However, their clinical efficacy still seems far away according to the current scientific information. Stem cells have been intensely studied to identify the novel gene functions underlying the cell character. Nevertheless, the accumulated evidence is insufficient to understand stem cell characteristics. This is probably because stem cells may have more unique and complex nature with unusual players or their non-redundant functions. To resolve multi-factorial characteristics, large-scale gene expression analyses have been employed, producing a large amount of expression data. Many experiments to study stem cells have been performed for different purposes, and their results have brought about different interpretations. However, we believe that the integration of the individual data can provide another comprehensive view. In the current study, we tried to find TFs characterizing stem cell subpopulations, using datasets adopted from more than two individual data sources. We could extract refined information, trimming noise by integrating the original dataset.

Motivated by the presence of the high-quality data, we examined the relationship between stem cell subpopulations and their corresponding gene regulators. As an appropriate model, an LVM was applied to co-clustering, which grouped highly correlated subpopulations and TFBSs simultaneously using latent variables. From the result, the regulatory module was defined, based on the significance of associations between two objects. The GO analysis showed an obvious bias between the modules and the biological functions of target genes. The lack of GO terms representing Cluster 2 suggests that genes belonging to a few specific functional categories and also several genetic factors involved in various biological functions may be responsible for the entire stemness property.

As shown previously, each of the five clusters represents distinct characteristics, which consist of one proliferation phase, one quiescence phase, stemness function and two progenitor stages. They have connections to several distinct transcriptional regulators, being partially overlapped among clusters. Though well separated, some TFBSs belong to more than one cluster. *CREB* is shared by three clusters, suggesting its ubiquitous function. *GABP, Egr2* and *Pax5* appear to have an influence over two different clusters. This observation supports their global functions in cell proliferation.

Recently co-clustering methods have been noticed in various biological issues, since many computational approaches have to deal with high-throughput biological datasets that are generated in the form of a two-dimensional matrix. These include microarray data, bionetworks and sequence motif sets. Previously, several co-clustering-based studies have successfully identified groups of genes in microarray datasets that show correlation between their expression patterns and the biological conditions (Cheng *et al*. 2000; Kluger *et al*., 2003; Madeira *et al*., 2004). Though our method also shares this aspect, the distinctiveness lies in its hidden variables. In this paper, the hidden variables flexibly and indirectly capture the relationship between stem cell species and the gene regulators. The number of hidden variables (i.e. the number of clusters) was determined in a heuristic manner by the reiterated tests. The larger the number of cluster is, the more clusters become redundant, which may result in lower generalization performance. Thus, consideration of prior knowledge (i.e. the biological context of stem cell subpopulations) will help determine an appropriate number of clusters.

Obviously, the accuracy of the clustering will be improved with more available source data. Moreover, to disclose the conserved modules on the regulatory network, it will be a decisive factor in the comparative analyses to test a greater variety of biological contexts including stem cells from various species and differentiation status. Consequently, the comparative study of diverse stem cell species will contribute to elucidating core mechanisms of stem cell regulation.

ACKNOWLEDGEMENTS

This research was supported in part by the National Research Laboratory Program of the Korea Ministry of Science and Technology (MOST) to B.T.Z. and in part by a grant from the Stem Cell Research Center of the 21st Century Frontier Research Program funded by the MOST and by a grant from KOSEF, through RCFC to R.H.S.

Conflict of Interest: none declared.

REFERENCES

Bishop,C.M. (1999) Latent variable models. In Jordan,M.I. (ed.), *Learning in Graphical Models*. The MIT Press, Cambridge, MA, pp. 371–404.

Boyer,L.A. *et al*. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Braganca,J. *et al*. (2003) Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2. *J. Biol. Chem*., **278**, 16021–16029.

Cheng,Y. and Church,G.M. (2000) Biclustering of Expression Data. In *Proceedings of the Eighth Interanational Conference on Intelligent Systems for Molecular Biology (ISMB '00)*. La Jolla, CA, pp. 93–103.

Du,K. *et al*. (1998) Transcriptional up-regulation of the delayed early gene HRS/SRp40 during liver regeneration. Interactions among YY1, GA-binding proteins, and mitogenic signals. *J. Biol. Chem.*, **273**, 35208–35215.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Flaherty,P. *et al*. (2005) A latent variable model for chemogenomic profiling. *Bioinformatics*, **21**, 3286–3293.

Heath,V. *et al*. (2004) C/EBPalpha deficiency results in hyperproliferation of hematopoietic progenitor cells and disrupts macrophage development *in vitro* and *in vivo*. *Blood*, **104**, 1639–1647.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hofmann,T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, **42**, 177–196.

Ikawa,T. *et al*. (2004) Long-term cultured E2A-deficient hematopoietic progenitor cells are pluripotent. *Immunity*, **20**, 349–360.

Ivanova,N.B. *et al*. (2002) A stem cell molecular signature. *Science*, **298**, 601–604.

Jager,R. *et al*. (2003) Transcription factor AP-2gamma stimulates proliferation and apoptosis and impairs differentiation in a transgenic model, *Mol. Cancer Res.*, **1**, 921–929.

Kabe,Y. *et al*. (2005) NF-Y is essential for the recruitment of RNA polymerase II and inducible transcription of several CCAAT box-containing genes. *Mol. Cell Biol.*, **25**, 512–522.

Kent,W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kluger,Y. *et al*. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**, 703–716.

Li,F.X. *et al*. (2003) Defective gene expression, S phase progression, and maturation during hematopoiesis in E2F1/E2F2 mutant mice. *Mol. Cell Biol.*, **23**, 3607–3622.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 24–45.

Maere,S. *et al*. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Mantamadiotis,T. *et al*. (2002) Disruption of CREB function in brain leads to neurodegeneration. *Nat. Genet.*, **31**, 47–54.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Pankratova,E.V. *et al*. (2001) Tissue-specific isoforms of the ubiquitous transcription factor Oct-1. *Mol. Genet.Genomics*, **266**, 239–245.

Ramalho-Santos,M. *et al*. (2002) 'Stemness': transcriptional profiling of embryonic and adult stem cells. *Science*, **298**, 597–600.

Rebel,V.I. *et al*. (2002) Distinct roles for CREB-binding protein and p300 in hematopoietic stem cell self-renewal. *Proc. Natl. Acad. Sci. USA*, **99**, 14789–14794.

Reid,L. (1990) From gradients to axes, from morphogenesis to differentiation. *Cell*, **63**, 875–882.

Ristevski,S. *et al*. (2004) The ETS transcription factor GABPalpha is essential for early embryogenesis. *Mol. Cell Biol.*, **24**, 5844–5849.

Sattler,U. *et al*. (2004) The expression level of the orphan nuclear receptor GCNF (germ cell nuclear factor) is critical for neuronal differentiation. *Mol. Endocrinol.*, **18**, 2714–2726.

Sordella,R. *et al*. (2002) Modulation of CREB activity by the Rho GTPase regulates cell and organism size during mouse embryonic development. *Dev. Cell*, **2**, 553–565.

Venezia,T.A. *et al*. (2004) Molecular signatures of proliferation and quiescence in hematopoietic stem cells. *PLoS Biol.*, **2**, e301.

Wang,J. *et al*. (2002) STAT1 deficiency unexpectedly and markedly exacerbates the pathophysiological actions of IFN-alpha in the central nervous system. *Proc. Natl Acad.Sci.USA*, **99**, 16209–16214.

Wang,V.E. *et al*. (2004) Embryonic lethality, decreased erythropoiesis, and defective octamer-dependent promoter activation in Oct-1-deficient mice. *Mol. Cell Biol.*, **24**, 1022–1032.

Zhang,P. *et al*. (2004) Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha. *Immunity*, **21**, 853–863.

Zhang,B.-T *et al*. (2003) Self-organizing latent lattice models for temporal gene expression profiling. *Mach. Learn.*, **52**, 67–89.