Discovering monotonic stemness marker genes from time-series stem cell microarray data

Hsei-Wei Wang\textsuperscript{1,2,3,4\#}, Hsing-Jen Sun\textsuperscript{1\#}, Ting-Yu Chang\textsuperscript{3\#}, Hung-Hao Lo\textsuperscript{5}, Wei-Chung Cheng\textsuperscript{6}, George C. Tseng\textsuperscript{7}, Chin-Teng Lin\textsuperscript{8,9}, Shing-Jyh Chang\textsuperscript{10}, Nikhil Ranjan Pal\textsuperscript{11\$} and I-Fang Chung\textsuperscript{1,2\$}

\$ Correspondence: nikhil@isical.ac.in and ifchung@ym.edu.tw

\# The software is available online http://microarray.ym.edu.tw/tools/MFSelector/

1. Microarray data collection and preprocessing
2. Results of ESCV
3. Biological relevance of other monotonic genes in the ESCV data set
4. Biological processes involved in monotonic genes in the ESCV data set
5. Supplemental Figures
6. Supplemental References
1. Microarray data collection and preprocessing

All data sets were downloaded from the NCBI GEO public archive, generated in Affymetrix Human Genome U133 Plus 2.0 platform with 54675 probe sets on chips, and the raw data (CEL file) normalized by RMA algorithm using the ‘affy’ package of the Bioconductor (http://www.bioconductor.org) [1] software suite in the R Project for statistical Computing (http://www.r-project.org). The default RMA settings are used for background correction, normalization and summarization of all expression values. The processed data matrices can be downloaded from the website: http://microarray.ym.edu.tw/tools/MFSelector/. In addition, the expression values of each gene are normalized in the range from zero to one across samples. This step preserves the richness in the original expression values for each gene among the samples, and helps us easily visualize the distribution of expression values for the significant genes. The PCA plots (generated by the statistical software, Partek® Genomics Suite™) are used for helping us understand the sample groupings.

1.1 Embryonic Stem Cell Neurogenesis data set (ESCN)

This data set contains 27 samples over five periods of human embryonic development: three embryonic stem cell (ESC) samples, three embryoid body (EB) samples, six primitive ectoderm cell (PEL) samples, six neural tube-like rosette cell samples, and nine post-natal neural stem cell (NSC) samples. This ESCN data set was downloaded from the NIH Neuroscience Microarray Consortium (GEO accession numbers are GSE9940 and GSE13307 [2]). In this data set, Zhang et al. (2008) first re-created in culture the developmental events of the first two to three weeks of human embryonic development during which embryonic stem (ES) cells were differentiated through the stages of embryoid bodies (EBs), primitive ectoderm cells (PEL), and neural tube-like rosette cells. The stage-specific events were then defined by Affymetrix Human Genome U133 Plus 2.0 array analysis along with the characteristic morphologic changes. Total RNAs were extracted from cells at the following neural specification developmental stages: ES cells (ESC, three samples), EBs grown in suspension (day 6, three samples), PEL stage (day 10, six samples), and neural rosettes (day 17, six samples). The post-natal neural stem cell (NSC, nine samples) array data are obtained from GSE13307 and our own home-made data [3].

1.2 Embryonic Stem Cell Vasculogenesis data set (ESCV)

In this data set, there are 13 samples over four periods of human embryonic stem cell differentiation into human mature (vascular) endothelial cells: three undifferentiated embryonic stem cell (ESC) samples, three mesodermal progenitor cell (MPC) samples, four embryoid body (EB) samples and three human mature vascular endothelial cell (VEC) samples. The ESCV data set is obtained by combining the following 2 differentiation data sets: GSE19735 [4] and GSE21668 [5]. The ESC samples are obtained from GSE21668. The three MPC samples (also from GSE21668) are extracted from the day 3.5 mesodermal progenitor (CD326neg CD56+) population. Mesoderm induction from human embryonic stem cells is initiated with combination of morphogens and growth factors including activin A, bone morphogenic protein 4, basic fibroblast growth factor and vascular endothelial growth factor. The mesodermal progenitor population is isolated by fluorescence-activated cell sorting (FACS) on day 3.5 of the culture based on the presence of CD56 expression and the absence of CD326 expression. The four EB samples (from GSE19735) are taken from two parts: one from human ESC EBs differentiated for 14 days in pro-angiogenic cytokines and the other from purified human ESC-derived durable endothelial cells isolated at day 14th of differentiation in the presence of TGFbeta inhibition.
2. Results of ESCV

The ESCV data set shows that vasculogenesis involves the differentiation of embryonic stem cells into mesodermal progenitor cells, which thereafter differentiate into embryoid bodies. The 18,046 probe sets (obtained using t-test, \( q<0.01 \)) distinguishing embryonic stem cells (ESCs) from mature VECs show that mesodermal progenitor cells (MPC) are also differentiated along the vasculogenic lineage into embryoid bodies (EB), as verified by the development of these ESCs toward mature VECs (shown by PCA plot in Fig. S11).

Applying our algorithm to the ESCV data set, we obtain the 216 ascending and 563 descending monotonic genes using \( \text{DE}_{\text{total}}=0 \) with the constraint \( p\text{-value}<1.0\times10^{-5} \) and with \( N-1 \) distinct discriminating lines (here \( N=4 \)). These genes are also included in Table S1 (in Additional file 4). Figure S12(A) displays the heatmap of the top 216 ascending (monotonic) genes with \( \text{DE}_{\text{total}}=0 \). Similarly, the gene expressions of the top 563 descending monotonic genes (\( \text{DE}_{\text{total}}=0 \)) are shown as a heatmap in Fig. S12(B). These figures, like the ESCN data set, reveal the quality of the monotonic genes identified by our method.

One of the top 216 ascending (monotonic) genes for this data set is \( JAK1 \) (Fig. S13(A)), which has been reported to be significantly involved in the interferon-alpha/-beta and -gamma signal transduction pathways [6]. Some of the other top 216 ascending monotonic genes, such as \( PDE8A \) [7], \( ACTR2 \) [8], \( TPM4 \) [9-11], \( ANXA2 \) [12], \( PON2 \) [13, 14] and \( ADAM9 \) [15, 16] have been reported to be expressed in vascular endothelial cells, vascular smooth muscle proliferation, neovascularization or have effect on heart/cardiovascular tissue.

On the other hand, one of the top 563 monotonically descending genes with \( \text{DE}_{\text{total}}=0 \) during ESC vasculogenesis is \( DDR1 \) (Fig. S13(B)). \( DDR1 \) is one of the cell adhesion molecules and it plays a key role in the cell-cell interactions [17]. In addition, \( FOXC2 \) transcription factor (forkhead box C2, mesenchyme forkhead 1; \( \text{DE}_{\text{total}}=0 \)) is crucial for the induction of in vivo endothelial gene expression during endothelial cell differentiation from the primitive mesodermal cells [18]. And \( FOXC2 \) is also a crucial regulator involved in lymphangiogenesis [19]. Some of the other top 563 descending monotonic genes, such as \( RBM25 \) [20], \( CTBP2 \) [21], \( PABPN1 \) [22], \( RB1CC1 \) [23], \( FAN1 \) [24], are shown to be related to key biological processes such as alternative pre-mRNA splicing/transcript, cell death, DNA replication or DNA repair. The genes \( TACC3 \) [25], \( HSPA9 \) [26], \( TCOF1 \) [27] and \( CHAF1A \) [28], which are also in the list of top 563 descending genes, have correlation with cell growth/proliferation, cell differentiation or embryonic development.

For this data set also we demonstrate the effectiveness of SVDE on four genes with \( \text{DE}_{\text{total}}=0 \). These genes are depicted in Fig. S14. These four genes are distinguished based on the SVDE values. Inspection of Fig. S14(A) suggests that \( CAPNS1 \) has the highest degree of monotonicity compared to the other three genes in Fig. S14, and this is indeed conformed by the SVDE values. Figure S14(B) shows that \( NNMT \) with \( \text{SVDE}=0.96 \) is the second most monotonic gene. This agrees with our expectation from visual inspection. \( NNMT \) is less monotonic than \( CAPNS1 \) because the expression values of two samples, the third sample from Stage Three and the first sample from Stage Four, are very close. Based on the SVDE values we find that \( GUK1 \) (Fig. S14(C)) is the third most monotonic one while \( TMEM45B \) (Fig. S14(D)) is least monotonic of the four. This conclusion based on SVDE agrees with our visual assessment of the figures.

3. Biological relevance of other monotonic genes in the ESCV data set

Some genes with significantly descending characteristics are the subject of interest because of their rules and/or regulation during stem cell differentiation and vasculogenesis. For example, \( CBX3 \) (\( \text{DE}_{\text{total}}=1 \)) is a monotonic gene with a strong descending attribute in the ESCV data set. As described earlier, \( CBX3 \) plays a significant role in
DNA repair and cell lineage differentiation. CBX3 directly promotes stem cell differentiation into smooth muscle cells, which participates in the development of the vascular system in the embryonic stage [29]. Furthermore, two descending genes, SMAD4 (DE<sub>total</sub>=1) and MDM4 (DE<sub>total</sub>=1) are responsible for stemness in the early stage of differentiation. SMAD4 is considered a factor for stabilizing the state of undifferentiated embryonic stem cells and the reduction of SMAD4 affects the stability of stem cells and drives the differentiation of stem cells into a particular cell lineage [30]. SMAD4 also regulates self-renewal and pluripotency through a complicated mechanism via the TGF-β signaling pathway. This signaling maintains the stemness of embryonic stem cells in an early stage with an undifferentiated state. Although the role of MDM4 (another monotonic gene identified by MFSelector) in embryonic stem cells remains unclear, MDM4 is known as a negative regulator of p53 protein [31]. MDM4 can directly interact with the transactivation domain of p54 for inhibiting the function of p53. Genetic defects in MDM4 leads to embryonic lethality, which is mediated by abolishing the regulatory mechanism to p53 [32]. Recent studies also reported that MDM4 directly down-regulates the transcriptional activity of p53 during the process of embryonic stem cell differentiation [33]. Through the regulation of p53, MDM4 modulates cell survival during the differentiation of embryonic stem cells. Noticeably, NCL (DE<sub>total</sub>=1) also shows down-regulation through the progress of vasculogenesis. NCL is a transcription activator of VEGF promoter [34]. As described above, VEGF is a well-known growth factor for inducing and stimulating vasculogenesis. Expression of NCL in early stage of embryonic stem cell is a crucial switch for activating VEGF expression. Whenever the level of VEGF is elevated, which directly promotes cell differentiation and processing vasculogenesis, NCL will no longer express with high level and will show a descending pattern.

On the other hand, some of the ascending genes, found by MFSelector, have also been proven to have functions in vascular development. For example, SOX17 (DE<sub>total</sub>=1), a transcription factor, is a differentially expressed temporal specification of blood progenitors from mouse embryonic stem cells and induced pluripotent stem cells (iPS cells) [35]. PECAM1 (platelet endothelial cell adhesion molecule; DE<sub>total</sub>=1), also known as CD31, has been widely reported to be involved in vascular endothelial cell integrity/formation [36-38]. PGF (DE<sub>total</sub>=2), also named PIGF, is a member of VEGF family. PGF acts as an agonist for VEGFR1, which is ubiquitously expressed in the endothelial cell lineage. Several studies have reported significant role of PGF in vessel formation or tumor vasculogenesis [39, 40]. Another important gene found by MFSelector is RhoB (DE<sub>total</sub>=1), which plays a unique role in several cellular processes, such as vesicle trafficking, Akt control, and cell survival. Through the regulation of Akt, RhoB maintains the stability and survival of endothelium during the process of vascular development. A recent study has demonstrated that the role of RhoB in the regulation of endothelial cell survival during vasculogenesis is in a stage-specific manner. The function of RhoB has also been evaluated through the RhoB knockout mice, which results in defects during the development of the vascular system. Moreover, another characteristic of RhoB is to maintain the angiogenesis progress through suppression of the anti-angiogenic agents. The stage-specific functions of RhoB in vasculogenesis make it a meaningful marker for the progression of vasculogenesis [41]. Epidermal growth factors (EGFs) are the major contributors to the growth and survival of endothelial cells, including embryonic endothelial progenitor cells. Another ascending gene, NRG1 (DE<sub>total</sub>=1), is a ligand of EGF family and it directly mediates the activity of endothelial progenitor cells and promotes angiogenesis [42]. In addition to EGF family, another growth factor IGF-1 (DE<sub>total</sub>=1) also has an ascending profile. As a paracrine and autocrine, IGF-1 and IGF-II modulate the proliferation and differentiation during the process of embryogenesis [43]. IGF-1/IGF-I receptor systems are involved in vasculogenesis during the development of lung and other organogenesis. Through maintenance of the endothelial cell population as a survival factor, IGF-1 promotes the development of the vascular system during the embryonic stage of rats. Higher expression of IGF-1
plays a crucial role in the process of vasculogenesis [44]. Notably, **TEK** (*TIE-2; DE_{total}=2*) is a receptor tyrosine kinase that is known to function as a molecule of vascular endothelial cells. **TEK** is found to be expressed specifically in the endothelial lineage and is required for the normal development of vascular structures during embryogenesis [45, 46]. In addition, **TEK** is also uniformly expressed in the endothelial lining of the vasculature [47]. Hence, we infer that **TEK** is usually expressed in endothelial cells of blood vessels during embryonic development.

### 4. Biological processes involved in monotonic genes in the ESCV data set

For the ESCV dataset, when the top 216 ascending monotonic genes (DE_{total}=0) found by MFSelector are subjected to GO analysis, few genes involved in ESCV-related processes, such as blood vessel morphogenesis (6 genes, \(p\)-value=0.035), blood vessel development (6 genes, \(p\)-value=0.06), vasculature development (6 genes, \(p\)-value=0.065), and striated muscle tissue development (5 genes, \(p\)-value=0.019) are found to be induced during ESC endothelial differentiation. These ascending genes directly guide the process of vasculogenesis and develop the basic structure of the vascular system. In addition to the functional genes related to vasculogenesis, other signal transducers such as VEGF or PGH are also involved during vascular development. As a ligand for VEGFR, VEGF transduces a significant signal for progression of angiogenesis. Hence, the genes related to cellular signaling transduction may sustain the function and activity of endothelial cells for angiogenesis or vasculogenesis. GO analysis reveals several genes with ascending profile during vasculogenesis for the signaling transduction system, such as cellular protein localization (11 genes, \(p\)-value=0.0279), cellular macromolecule localization (11 genes, \(p\)-value=0.0294), intracellular transport (14 genes, \(p\)-value=0.004), intracellular protein transport (10 genes, \(p\)-value=0.004), protein localization (16 genes, \(p\)-value=0.008), RAS protein signal transduction (5 genes, \(p\)-value=0.012) and RHO protein signal transduction (3 genes, \(p\)-value=0.041).

On the other hand, some of the top 563 descending monotonic genes with DE_{total}=0 are involved in DNA/RNA or other basic biological processes, such as DNA metabolic process (37 genes, \(p\)-value=2.9E-8), DNA replication (21 genes, \(p\)-value=9.1E-8), cell cycle (41 genes, \(p\)-value=0.00092), and chromosome organization (30 genes, \(p\)-value=0.00021). These genes with basic functions are responsible for the proliferation and replication of stem cells. Without any lineage differentiation and further stimulation, stem cells in an early stage do not possess specific functions and characteristics. Maintenance of high proliferation rate and chromosome stability is the most significant function in stem cells or precursor cells. Therefore, genes related to proliferative and self-renewal activities for maintaining the pluripotency are supposed to play important roles in undifferentiated cells, especially in stem cells. On the other hand, stimulation from circumstances may cause damage to genome structures. Hence we can find those genes with response to DNA damage stimulus (24 genes, \(p\)-value=9.62E-5) and DNA repair (20 genes, \(p\)-value=1.37E-4) as expressed in an early stage for stabilizing the genome structure.
Supplemental Figures

Fig. S11. A three-dimensional scatter plot of the ESCV data set analyzed by principal component analysis.

Figure S12. Heatmaps of the two monotonic gene sets of the ESCV data set. The gene expression values change gradually from the blue band (low expression values) into the red band (high expression values) and vice versa. (A) 216 monotonic genes with ascending profiles and their DEtotal values are zero; (B) 563 monotonic genes with descending profiles and their DEtotal values are zero.
Fig. S13. Scatter plots of 1552610_a_at (*JAK1*) and 1007_s_at (*DDR1*) of the ESCV data set with ascending profile and descending profile respectively. (A) This is one of the top 216 monotonically ascending genes with $\text{DE}_{\text{total}}=0$; (B) This is one of the top 563 monotonically descending genes with $\text{DE}_{\text{total}}=0$. 

(A) ID: 1552610_a_at (*JAK1*), $\text{DE}=0$, $p: 6.6\text{e}{-7}$, $q: 1.6\text{e}{-4}$, SDVE: 0.11

(B) ID: 1007_s_at (*DDR1*), $\text{DE}=0$, $p: 5.9\text{e}{-7}$, $q: 6.7\text{e}{-5}$, SDVE: 0.36
Fig. S14. Scatter plots of the four ascending monotonic genes of the ESCV data set, whose DE_{total} values all are equal to one, illustrate the sample variance for discriminating error by adding noise to each sample for 100 simulations. (A) 200001_at (CAPNS1) with SVDE=0 (1st); (B) 202237_at (NNMT) with SVDE=0.96 (2nd); (C) 200075_s_at (GUK1) with SVDE=1.3 (3rd); (D) 226226_at (TMEM45B) with SVDE=2.75 (4th).
Supplemental References


15. Guaiquil V, Swendeman S, Yoshida T, Chavala S, Campochiaro PA, Blobel CP: ADAM9 is involved in pathological


