Statistical Analysis of Multi-Dimensional, Temporal Gene Expression of Stem Cells to Elucidate Colony Size-Dependent Neural Differentiation

Journal:	*Molecular Omics*
Manuscript ID	MO-RES-01-2018-000011.R1
Article Type:	Research Article
Date Submitted by the Author:	13-Feb-2018
Complete List of Authors:	Joshi, Ramila; The University of Akron, Biomedical Engineering Fuller, Brendan; The University of Akron Li, Jun; Kent State University, Mathematical Sciences Tavana, Hossein; The University of Akron, Biomedical Engineering

Statistical Analysis of Multi-Dimensional, Temporal Gene Expression of Stem Cells to Elucidate Colony Size-Dependent Neural Differentiation

Ramila Joshi[1], Brendan Fuller[1], Jun Li[2], Hossein Tavana[1*]

[1]Department of Biomedical Engineering, The University of Akron, Akron, Ohio 44325, USA

[2]Department of Mathematical Sciences, Kent State University, Kent, Ohio 44242, USA

[*]Corresponding author:

Hossein Tavana, Ph.D., P. Eng.

Department of Biomedical Engineering

The University of Akron,

260 S. Forge St., Akron, OH 44325

Tel: (330) 972-6031, E-mail: tavana@uakron.edu

Abstract

High throughput gene expression analysis using qPCR is commonly used to identify molecular markers of complex cellular processes. However, statistical analysis of multi-dimensional, temporal gene expression data is complicated by limited biological replicates and large number of measurements. Moreover, many available statistical tools for analysis of time series data assume that the data sequence is static and does not evolve over time. With this assumption, the parameters used to model the time series are fixed and thus, can be estimated by pooling data together. However, in many cases, dynamic processes of biological systems involve abrupt changes at unknown time points, making the assumption of stationary time series break down. We addressed this problem using a combination of statistical methods including hierarchical clustering, change point detection, and multiple testing. We applied this multi-step method to multi-dimensional, temporal gene expression data that resulted from our study of colony size-dependent neural cell differentiation of stem cells. The gene expression data were time series as the observations were recorded sequentially over time. Hierarchical clustering segregated the genes into three distinct clusters based on their temporal expression profiles; change point detection identified specific time points at which the entire dataset was divided into several homogenous subsets to allow a separate analysis of each subset; and multiple testing procedure identified the differentially expressed genes in each cluster within each subset of data. We established that our multi-step approach pinpoints specific sets of genes that underlie colony size-mediated neural differentiation of stem cells and demonstrated its advantages over conventional parametric and non-parametric tests that do not take into account temporal dynamics of the data.

Importantly, our proposed approach is broadly applicable to any multivariate data sets of limited sample size from high throughput and high content screening such as in drug and biomarker discovery studies.

Key words: multivariate data analysis, stem cell colony size, neural differentiation, change point detection, qPCR data analysis

Introduction

Technological capabilities to produce large arrays of biological data of genes, transcripts, proteins, metabolites, and other biomolecules such as non-coding, regulatory RNAs provide an unprecedented opportunity to identify molecular markers of complex cellular processes such as stem cell differentiation.[1,2] Understanding complex molecular events at gene, transcript, or protein levels allows the discovery of biomarkers associated with normal physiological processes and disease conditions. Biomarker discovery helps to improve early disease detection, determine disease prognosis, monitor the response to therapy, and select promising treatments.[3–5] However, extracting meaningful information from complex datasets that result from high throughput gene, transcript, or protein level studies is not trivial. The first challenge is multi-dimensionality of the data. Genomic and proteomic studies typically involve measuring hundreds of factors at a time leading to the known issue of “large p, small n”, where the number of measurements, p, is far greater than the number of independent samples, n.[1,2,6] The analysis of hundreds to thousands of measurements without appropriate statistical methods often leads to results with poor biological interpretability and plausibility. Although there are some recent developments to analyze very large data sets, the performance of these methods relies on a key assumption that the data sequence is static to test parameters of interest by pooling together the data. However, this can lead to inaccurate results because biological processes are very often dynamic and involve temporal changes at unknown time points. Another challenge of analyzing time series data is to identify a proper statistical tool that takes the changes

into account. Therefore, innovative approaches to integrate statistical tools and expert knowledge-based methods are crucial to analyze multi-dimensional time series data.

qPCR is a widely-used technique to quantify gene expression patterns and changes in cells. Similar to the challenges of analyzing data resulting from other high throughput techniques, testing statistical significance of qPCR data is often complicated by limited biological sample replicates, and lack of normal distribution and inherent variations of the data.[7,8] As such, conventional statistical tests that rely on large sample sizes often do not result in statistically significant differences between different experimental conditions, despite clear and visible differences in the data. Moreover, the challenges of limited sample size, large variance, and dynamic data structure also hamper the performance of conventional statistical tests. Thus, these tests may not elicit statistically significant differences among experimental conditions. More rigorous data analysis using reliable and suitable statistical tests are required to analyze such qPCR data.

Our goal was to explain the colony size-dependent neural differentiation of stem cells using a gene expression profiling. This study was motivated by our finding that increasing the size of mouse embryonic stem cell (mESC) colonies disproportionately increased the expression of various neural cell proteins and enhanced the efficiency of deriving neural cells.[9,10] The main objectives were to (i) identify genes with distinct expression levels among different colony sizes that mediate colony size effect on neural differentiation of mESCs, and (ii) examine temporal differences in the expression of these genes among different colony sizes of mESCs. We performed temporal qPCR to assess differences in neural differentiation among three different colony sizes of mESCs. We used three biological replicates for each colony size (sample) and

measured fold change values of 28 genes over 15 time points (days), resulting in 420 data points for each of the three colony sizes. Therefore, a robust and efficient statistical tool to analyze the multivariate, time series gene expression data was required to overcome the obstacles discussed above.

We used a combination of hierarchical clustering, change point detection, and multiple *t*-tests to statistically analyze the data. We first implemented an agglomerative hierarchical clustering approach to segregate genes representing pluripotent cells, neural stem cells, and specific neural cells into three distinct clusters based on temporal expression profiles of the genes. Then, we applied an E-divisive change point detection method to the clusters of genes.[12] This identified specific time points at which the entire dataset partitioned into several homogeneous subsets, each having its own constant parameters. Finally, within each homogeneous subset, we identified the differentially expressed genes in each cluster using false discovery rate (FDR)-controlled multiple *t*-tests. Combining the change point detection method with multiple *t*-tests pinpointed specific sets of genes that determine colony size-mediated neural differentiation of mESCs. We also discuss the advantages of our robust multi-step statistical approach over conventional parametric and non-parametric statistical methods that do not consider the temporal dynamics of the data.

Materials and Methods

Mouse embryonic stem cells (mESCs) culture and differentiation

Stromal PA6 cells (Riken) and undifferentiated mESCs (EB5, Riken) were maintained separately on 0.1% gelatin coated dishes as described before.[11] For preparation of the

stromal layer, PA6 cells were grown to a confluent monolayer on a gelatin coated 35 mm Petri dish and mitotically inactivated with 10 µg/ml mitomycin-c (Sigma) for 2 hrs. PA6 cells were washed and then incubated overnight at 37ºC and 5% CO2 in a differentiation medium. A polymeric aqueous two-phase system was used to microprint mESCs in size controlled niche over stromal cells as described before.[12] Briefly, a defined number of mESCs was suspended in a 6.4%(w/v) aqueous solution of dextran (DEX, Mw: 500 kDa; Pharmacosmos) prepared in a differentiation medium. Regular culture medium in Petri dishes containing stromal cells was replaced with the 5.0%(w/v) aqueous polyethylene glycol (PEG, Mw: 35 kDa; Sigma) solution prepared with the differentiation medium. A liquid handling robot equipped with 50 nl hydrophobic slot pins aspirated the mESCs suspension in the aqueous DEX phase from a 384-well source plate. Inserting the pins into the Petri dishes containing the aqueous PEG phase resulted in autonomous dispensing of the content of pins on the stromal layer without contacting the stromal cells (Fig 1a). mESCs remained confined within the DEX phase drops (Fig 1b) and adhered to the PA6 cell layer. The two-phase solution was replaced with fresh differentiation medium after incubating the cells for 3 hours (Fig 1c). Printed mESCs proliferated to form individual colonies of controlled size determined by the density of mESCs within each drop.

Immunofluorescence and neural protein expression analysis

mESC colonies were fixed in 3.7% formaldehyde on day 8 of co-culture with stromal cells. Immunocytochemistry was performed for a neural progenitor cell marker, TuJ, with a rabbit monoclonal class III β-tubulin primary antibody (Biolegend). Expression was visualized using a rhodamine red conjugated rabbit monoclonal secondary antibody

(Jackson Immunoresearch). Each colony was imaged in sections, which were merged using Photoshop CS (Adobe) to generate a single image of the entire colony. Neural differentiation of mESC colonies stained with TUJ was quantified using an adaptive thresholding plugin in ImageJ and our previously defined method.[10]

Gene expression analysis

Total mRNA was isolated daily from mESC colonies for a duration of two weeks using an RNA isolation kit. cDNA was synthesized from 1 µg of total RNA using random hexamer primers post removing DNase using RNase-free DNase kit. Real time qPCR was performed with a Lightcycler 480 II instrument using a SYBR Green Master Mix using predefined protocol.[11] Expression levels of mRNA for different marker genes were calculated relative to GAPDH and β-actin using the $\Delta\Delta C_t$ method. The fold change in mRNA expression of all genes was determined according to the $2^{-\Delta\Delta Ct}$ method. All experiments were performed in triplicates.

Statistical Analysis

Out of 28 genes analyzed, 22 pluripotency and neural cell marker genes were considered in three distinct clusters according to our previous study.[13] Because the focus of this study was to identify differentially expressed neural genes, two clusters (labeled clusters 1 and 2) comprising of 17 neural lineage marker genes were subjected to further statistical tests. For each colony size, two vectors whose elements were the temporal $\Delta\Delta C_t$ values of genes in clusters 1 and 2, were defined. Then, delta vectors (δ_t) were calculated as the differences between respective vectors for each two colony sizes. Statistical tests were performed to identify the differentially expressed genes from

the δ_t vectors in two steps. In the first step, an E-divisive change point detection method was applied to the δ_t vectors that represented the differences in - $\Delta\Delta C_t$ values of genes within one cluster, to identify time points at which a population mean shift occurs for at least one gene in the δ_t vector.[14] The change point divided the temporal trajectory of δ_t into two homogenous subsets, each having a population counterpart of δ_t that is static within each subset. Then, as the second step, by pooling together the observations within each subset, the standard one-sample *t*-test with Benjamini-Hochberg's false discovery rate (FDR) control method was used to identify the statistically significant elements of δ_t in each population.[15] The nominal significance level was chosen at 0.01 to control the FDR.

To compare the results from the above analyses with conventional statistical tests without taking the temporal dynamics of the data into account, the $-\Delta\Delta C_t$ values obtained daily from samples of each colony size were considered as a separate sample. Then, student's *t*-test and Mann-Whitney U test were performed to test the difference between samples representing two different colony sizes. Benjamini-Hochberg's false discovery rate (FDR) control method was applied to both tests to assess the statistical significance of the difference observed after one-to-one comparison among the three colony sizes for each gene.[15] The nominal significance level was chosen at 0.01 to control the FDR.

Result and Discussion

Colony size effect on neural differentiation of mESCs: Protein expression

mESCs within the printed DEX phase drop adhered to the stromal layer and proliferated to generate a single colony during incubation (Fig 1d-e). The mitotically arrested PA6 cells remained intact for 2 weeks while the proliferating mESCs differentiated into neural cells as shown by thick neurite processes extending out from the colonies (Fig 1e-f). The size of mESC colonies was controlled through the density of printed mESCs in the 50 nl DEX phase drops. Densities of 100, 250, and 500 cells per 50 nl drops yielded individual mESC colonies with average diameters of 1.00±0.05 mm, 1.35±0.04 mm, and 2.20±0.10 mm by day 8 (Fig 1g). We label these colonies small, medium, and large.

Culturing mESCs with stromal PA6 cells induces neural differentiation via intercellular signaling.[16,17] We recently showed that spatial organization of mESCs can further regulate this process.[10] Using the two-phase system cell printing technique, we generated mESCs colonies of three different sizes on stromal cells and immunostained the differentiating cells in the colonies for a neural cell protein marker, TuJ, at regular intervals for 2 weeks. Fig 1h-j show colonies of three different sizes stained for TuJ on day 8 of culture on stromal PA6 cells. Longer and denser neural processes extended out from the periphery of the large colonies compared to the medium colonies, and from the medium colonies compared to the small colonies. When protein expression data were normalized to the colony size, the large mESC colonies yielded disproportionately greater expression of TuJ (Fig 1k) and other neural cell proteins,[10] indicating the role of endogenous factors of mESCs to self-regulate their own neural differentiation.

Colony size effect on neural differentiation of mESCs: Gene expression

To elucidate this finding, we performed a comprehensive temporal gene expression profiling study to identify major genes and transcription factors that mediate colony size effect on neural differentiation of mESCs. Out of 28 genes analyzed, we conducted a time-course gene expression profiling of 22 stage-specific gene markers of pluripotency, neural progenitors, specific neuronal and glial cells, as well as the transcription factors and regulators of neurogenic pathways in colonies of three different sizes for two weeks. Selection of this set of genes was based on a comprehensive literature review. To ensure the specificity of neural differentiation of mESCs in our engineered niches, we also performed qPCR analysis of the mesodermal marker genes NKX 2.5, GATA4, FLK1, and PECAM. Table S1 lists all the genes and their primers sequences.

On each day, we performed qPCR on three experimental replicates for each colony size. And for each experiment, we obtained 15 samples over the two-week culture. Without taking into account that each qPCR reaction was run in duplicates, this gave a total of 3780 qPCR reactions (including the two reference genes and four mesodermal markers). We used GAPDH and β-actin as reference genes and undifferentiated mESCs as the negative control. Fig 2 schematically shows the experimental workflow. We calculated $\Delta\Delta C_t$ values and represented fold change as $2^{-\Delta\Delta Ct}$. Consistent with the protein expression study,[10] a majority of the neural genes showed the highest mRNA fold change in the large colonies, followed by the medium, and then the small colonies.

Fig 3 represents the temporal mRNA fold change for five pluripotency marker genes. Expression of these genes, such as Oct4 and Nanog, steadily decreased and remained low over time irrespective of the colony size. Fig 4 shows temporal mRNA fold change

for six marker genes of neural stem and progenitor cells, whereas Fig 5 shows mRNA fold change of 11 marker genes for specific neuronal and glial cells. Results in Fig 4 and Fig 5 suggest the following key conclusions. First, with increase in the colony size, there was a greater mRNA fold change for neural stem and progenitor cell markers as well as specific neuronal cells and astrocyte markers. Second, the temporal expression of neural genes showed a colony size-dependent effect where the expression levels increased earlier in the large colonies followed by the medium and the small colonies. Third, the gene expression trajectories were similar in all three colony sizes. Expression of neural stem cells marker genes such as Sox1 and Pax6, rose to a peak level around days 4-7 and declined thereafter, whereas the gene expression of markers of specific neural cell lineages such as GAP43 and GFAP continuously increased throughout the 14 days of culture. These results are consistent with the role of these markers in terms of loss of pluripotency (Oct4 and Nanog), commitment to a neural cells (Sox1 and Pax6), followed by differentiation into specific neural cells (GFAP and GAP43).[18–21]

Statistical analysis of multivariate, temporal gene expression data

Although the gene expression profiles displayed a clear colony size-dependent effect, evaluating the statistical significance of the expression differences was challenging. As seen in Fig 3-5, the gene expression levels in the three colony sizes were not steady over the two-week culture and transiently varied. For example, the highest difference in expression levels among the three colony sizes in Fig 4 was observed between days 4-6 for Pax6 gene, and between days 8-10 for Nestin and Wnt1. This indicates that in addition to differences in the expression levels of these specific genes between colonies of different sizes, the time points at which the differences occur is a key factor

underlying colony size-mediated neural differentiation of mESCs. Below, we present a multi-step statistical approach to analyze the multivariate, temporal qPCR data.

Application of multi-step statistical analysis

Step 1: Hierarchical clustering

With cluster analysis, data points are placed in discrete sets according to a similarity measure and a grouping algorithm.[22] The reduction in the data results from forming *g* groups out of *n* data points, where $g < n$. The most common method of cluster analysis is agglomerative hierarchical cluster analysis. This analysis first merges the two closest data points into a single group. Then, similarities of this group to all other groups are calculated. Repeatedly, the two closest groups are combined until only a single group remains. The results are usually expressed in a dendrogram, a two-dimensional hierarchical tree diagram representing the complex multivariate relationships among the objects. In past studies focused on biomarker discovery and gene expression profiling, clustering was used as an initial screen to separate the samples into different classes based on gene/marker expression profiles. After clustering of data, further analysis identified the markers that best characterized the segregation of the data.[23,24] Clustering has also been used after biomarker discovery to verify that the expression levels of identified markers can help separate two samples effectively into distinct clusters.[24,25]

In our previous study using co-cultures of mESCs-stromal cells, applying agglomerative hierarchical clustering on a set of 22 genes of pluripotency and neural cells segregated them into three distinct clusters based on their temporal expression trajectories.[13] Genes in these clusters represented distinct phases of transition of mESCs from a

pluripotent stage to neural progenitor cells, and to terminally-differentiated neuronal or glial cells. Considering our interest in identifying genes responsible for colony size-mediated neural differentiation, in this study, we focused on two clusters of genes that represent markers of neural progenitors and specific neuronal and glial cells. The cluster containing the pluripotency marker genes (Oct4, Nanog, Wnt8a, Notch2, and Notch3) was not considered for further statistical analysis because the expression of these genes quickly downregulated and showed minimal differences among the three colony sizes. The first cluster consists of genes that were highly expressed between days 4 and 11 and then either downregulated or leveled off. Genes in this cluster (Fig 4) mark the onset of neural differentiation: Sox1, CDH2, Wnt1, Notch1, Nestin, and Pax6. The second cluster represents a group of genes whose expression increased steadily as the culture progressed in time (Fig 5). This cluster contained several specific neuronal and glial markers and included TuJ, NCAM, TH, GAD1, Synaptophysin, ChAT, MAP2, Olig1, GFAP, NeuN, and GAP43.

Step 2: E-divisive change point detection

For each colony size, we defined two vectors whose elements are the temporal $\Delta\Delta C_t$ values of genes in clusters 1 and 2. That is, $Y_{it} = (Y_{it,1}, Y_{it,2}, \ldots, Y_{it,6})^T$ for cluster 1, and $Y_{it} = (Y_{it,1}, Y_{it,2}, \ldots, Y_{it,11})^T$ for cluster 2, over time t ($t \in \{0, 1, 2, \ldots, 14\}$) for colony sizes small (i=1), medium (i=2), and large (i=3). For example, a vector representing the first cluster for the small colony contained $\Delta\Delta C_t$ values of six genes at 15 time points. Then, delta vectors (δ_t) were calculated as the differences between respective vectors for each two colony sizes (i.e., large vs. medium, medium vs. small, and small vs. large): $\delta Y_t^{32} = Y_{3t} - Y_{2t}$, $\delta Y_t^{21} = Y_{2t} - Y_{1t}$, $\delta Y_t^{31} = Y_{3t} - Y_{1t}$, over time t ($t \in \{0, 1, 2, \ldots, 14\}$).

Then, we used a hierarchical E-divisive method to identify time points at which at least one of the genes in the δ_t vector had a statistically significant mean shift at the identified change point. For example, if the change point was detected on day T_q when comparing small and medium colonies, then $\delta Y_{T0}^{21} = \delta Y_{T1}^{21} = \ldots = \delta Y_{Tq}^{21} \neq \delta Y_{Tq+1} = \ldots = \delta Y_{T14}^{21}$.[21] In other words, the average of δ_t values on days 0-T_q was significantly different from the average of δ_t values on days T_q -14 for at least one gene ($p < 0.01$). Thus, all observations before T_q belonged to one population. Similarly, all observations after T_q represented a second population.

Applying this method, we identified a change point of day 6 for cluster 1 genes among all three colony sizes compared pairwise (Fig 6a). The constituent genes of this cluster showed a steady expression increase during the first few days when mESCs undergo differentiation to neural stem cells, followed by a steady decline during the differentiation of neural stem cells to specific neuronal and glial progenitors. Therefore, the activity of these genes was predominantly confined to a few days towards the end of the first week of culture, validating the detected day 6 change point. Comparing the expression of the cluster 2 genes for each two colony sizes gave a change point of day 6 between large and small colonies, and day 7 between large and medium as well as between medium and small colonies (Fig 6a). Considering that the genes in this cluster had insignificant expression during the first week of culture but showed high activities as neural stem cells differentiated to distinct neuronal and glial lineage cells, these change points indicate emergence of differences in gene markers of specific neuronal and glial cells among the three colony sizes.

As a non-parametric statistic, the E-divisive change point detection method does not rely on assumptions that the data are derived from a certain probability distribution, does not require data pre-processing, and is computationally efficient compared to other parametric methods. A similar application of the non-parametric change point statistic was was reported for microarray data to detect differentially expressed genes among different tumor samples.[26] An important outcome of applying change point detection to our multivariate qPCR data was that it divided the time series expression data of each gene from 14 days into two statistically distinct homogeneous populations based on the temporal changes in the gene expression. This step accounted for the variability due to temporal dynamics in the gene expression data and enabled the use of standard parametric tests (*t*-tests) separately on each of the new homogenous populations of data as described below.

Step 3: Multiple t-tests

Next, using the homogeneous subsets of data, we identified the genes with significant expression differences among the three colony sizes before and after the change points. In our experimental setup (Fig 2), the expression levels of multiple genes were quantified for each experimental group at 15 time points (i.e., days 0 - 14). The gene expression data from different time points are independent as the experimental samples were terminated to prepare RNA for the qPCR experiments on each day. As a result, the δ_t elements are independently and identically distributed within each homogenous subset formed after applying the change point detection method. We therefore applied multiple *t*-tests to assess the statistical significance of δ_t. To avoid finding a random false positive in the multiple comparisons problem, we compensated for the false

discovery rate (FDR) using a correction method known as the Benjamini–Hochberg method.[27] This method arranges the individual *p* values in an ascending order and ranks them according to their position. Then, it compares each individual, ranked *p* value, $p_{(i)}$, obtained from *t*-test to its Benjamini-Hochberg critical value, $(i/m)\times Q$, where *i* is the rank, *m* is the total number of tests, and *Q* is the selected false discovery rate. The largest *p* value that has $p_{(i)} < (i/m)\times Q$, and all of smaller *p* values denote statistical significance.[15]

The detection of specific change points (e.g., day 6) implied that the δ_t values before the change point (i.e., days 1-6) and after the change point (i.e., days 7-14) represent two different populations. Therefore, we conducted multiple one sample *t*-tests with controlled FDR for each set of δ_t values before and after the change points to identify differentially-expressed genes in each set. Since all the gene expression data had a common starting point of $-\Delta\Delta C_t = 0$, the δ_t values (i.e., the differences between the $-\Delta\Delta C_t$ values of each two colony sizes) were also equal to zero in the beginning. The second step of statistical analysis was therefore to identify the components (i.e., genes) of the δ_t vectors that were significantly different from zero, before and/or after the identified change points (the nominal significance level, Q= 0.01). The results are summarized in Venn diagrams for both gene clusters, and between each pair of colonies within each cluster (Fig 6b). We note that this analysis only revealed significant differentially expressed genes between each two colony sizes. The actual temporal mRNA fold change values are represented in Fig 4 and Fig 5.

Differentially-expressed genes among different mESC colony sizes: Cluster 1

For cluster 1 that contained genes representing specific functions in differentiation of ESCs and early patterning of the nervous system, our analysis detected a change point of day 6 among all three colony sizes. This agrees with our protein expression data that showed the greatest activity of neural stem and progenitor markers around days 4 to 8.[28,29] As seen in panels i – iii of Fig 6b, Pax6 expression was different only before the change point in large vs. small colonies, and in medium vs. small colonies. In large vs. small colonies, the expression levels of Sox1, Nestin, and CDH2 were different both before and after the change point. Both Pax6 and Sox1 are active in neural stem cells and their differentiation to specific neuronal and glial progenitors.[30,31] Therefore, higher expression of these genes in larger colonies indicates greater activities in differentiating mESCs to generate more neural progenitors. Moreover, Pax6 expression was different only before the change point where the neural stem cell state precedes the appearance of post mitotic neurons, whereas Sox1 expression remained different even after the change point. This indicates a temporary role for Pax6 and but a more persistent role for Sox1 in driving neural stem cells to specific neuronal and glial precursors. Our result is consistent with previous reports that increasing Pax6 levels was sufficient to drive neural stem cells toward neuronal cells by upregulating Neurogenin 2, and that neuronal precursors maintaining a prolonged expression of Pax6 failed to become neuronal cells.[32] On the other hand, Sox1 marks the cells with a neurogenic potential, has an initial role of self-renewal of neural stem cells pool, and its persistent expression leads to neuronal cells.[20] Greater expression of Sox1 with increase in the colony size indicates larger yield of neuronal cells from mESCs.

The genes that were expressed differently between each pair of colony sizes only after the change points are shown in the blue circles, whereas those that were expressed differently throughout the culture are located at the interface of the orange and blue circles. From cluster 1 genes, Wnt1 and Notch1 levels were different only after the change point when comparing the large and small colonies. Nestin, CDH2, Wnt1, and Sox1 expression was different between the large and medium colonies only after day 6. And only CDH2 (N-cadherin) showed different expression levels between the medium and small colonies. Wnt1, which had different expression levels between each pair of the colony sizes after the change point, belongs to the canonical Wnt family that promotes proliferation of cells during central nervous system development and helps induce sensory and midbrain dopaminergic neurons.[33] Therefore, increased Wnt1 expression and very high levels of this marker with increase in colony size (Fig 4) suggest that Wnt1 augments neural cell commitment of mESCs and promotes dopaminergic neuron and astrocyte differentiation in a colony size-mediated manner.[34] CDH2 (N-cadherin) is a key marker that showed differential expression levels among all three colony sizes. CDH2 is an intercellular junctional protein that maintains β-catenin signaling during cortical development, regulates Wnt signaling, and induces radial glial progenitor cells.[35] Moreover, CDH2 supports neuronal circuit maturation through axonal extension and regulates neurites outgrowth through fibroblast growth factor receptor signaling.[36] As shown in Fig 4, higher fold change of CDH2 in the large colonies correlates well with the longer neurites length in this niche observed in our previous study.[10]

Differentially-expressed genes among different mESC colony sizes: Cluster 2

For the cluster 2 genes, our analysis identified day 6 as the change point when comparing the large and small colonies, and day 7 when comparing the other two pairs of colony sizes. This cluster contains genes whose expression continuously increased throughout the two-week culture, or increased and then leveled off at some time point. These genes are associated with growth and development of neuronal and glial cells, or with specific cell types such as dopaminergic, GABAergic, and cholinergic neurons. A change point of day 6 or 7 is reasonable as previous studies showed that in the presence of stromal cells, neural stem cells differentiate into specific neuronal and glial precursors around this time point.[37,38] This is also consistent with our study that showed a substantial increase in the expression of the genes during the second week of culture (Fig 5). Additionally, an earlier change point between the large and small colonies (Fig 6a) together with greater fold change of the genes in the large colonies (Fig 5) imply earlier commitment of stem cells to neural cells in the large colonies, in agreement with the protein expression results that showed the largest difference between these two colony sizes (cf. Fig 1h and 1j).[10]

From the cluster 2 genes, GAP43, Synaptophysin, MAP2, and Olig1 showed expression differences between the large and small colonies prior to the change point (Fig 6b-iv), and GAP43, Synaptophysin, NeuN, and GFAP had different levels between the large and medium colonies. Genes in this cluster had a similar expression in the medium and small colonies prior to the change point. Difference in the expression of these genes before the change point is in part due to the differences already present in their progenitor cells. For example, higher expression of MAP2 was seen in the large colony

as soon as the Nestin-positive cells, which are already present in greater quantities in the large colonies, differentiated into MAP2-positive neurons. Closer scrutiny of results in Fig 4 and Fig 5 showed that except for Olig1, other neural factors had greater fold change with increase in the size of colonies.

After the change point, all 11 genes of cluster 2 showed significant expression differences between the large and small colonies (Fig 6b-iv). Comparing the large and medium colonies also showed that eight markers were statistically different between the two configurations (Fig 6b-v). And between the medium and small colonies, nine genes from this cluster showed significant differences after the change point (Fig 6b-vi). For cluster 2, detection of a change point and significant expression differences of the same gene before and after the change point (e.g., GAP43 and Synaptophysin) indicates that the colonies with a richer content of neural stem cells lead to more specific neural cells. A closer observation of temporal trajectories of such genes in Fig 5 reveals that the differences in gene expression levels steadily increased after the change point and throughout the culture.

Greater expression of growth- and development-associated genes such as NCAM, GAP43, and MAP2 in large colonies implies greater density and longer neurite processes.[39,40] This is consistent with the higher neurites length and density measurements performed at a protein level in our previous study.[10] Larger mRNA fold change values for Synaptophysin in the large colony than in the medium and small colonies indicate greater neuronal maturation and synaptic development. Higher expression of specific neuronal cell markers such as GAD1, ChAT, and TH implies potentially greater number of functional neurons generated in the large colony. Again,

this is consistent with our previous protein expression analysis that showed higher levels of GFAP expression and greater number of TH-positive neuronal cells with increase in the colony size.[10]

To ensure the specificity of neural differentiation of mESCs in these niches, we performed qPCR analysis on the mesodermal marker genes NKX 2.5, GATA4, FLK1, and PECAM. Interestingly, the small colony configuration had higher fold change of mesodermal markers than the medium and large colonies (Fig S1). This, along with the results presented above, indicates that larger colonies are more efficient in restricting the fate of differentiating mESCs toward neural cell lineages, whereas the small colonies provide a more supportive niche for mesodermal cell differentiation. This is consistent with the results obtained from previous studies that showed decreasing the size of mESC embryoid bodies (EBs) enhanced their response to BMP4 signal to induce mesodermal differentiation.[41,42] EBs of ~450 μm diameter, which are smaller than the small colony in our study, also showed maximum mesodermal differentiation.[43]

Performance of conventional statistical tests to analyze the multivariate, temporal gene expression data

To demonstrate the significance of selecting the multi-step statistical tests to analyze our multivariate, temporal gene expression data, we performed a parametric Student's *t*-test and a non-parametric Mann Whitney U-test on all pooled data without taking temporal dynamics of the data into account. Below, we summarize the results.

Student's t-test

Student's *t*-test is a conventional method of choice to analyze statistical difference between two experimental groups. Therefore, we conducted three separate *t*-tests to compare gene expression differences of large colony and small colony, large colony and medium colony, and medium colony and small colony. This allowed us to compare the results from the *t*-tests with those from the multi-step method. Student's *t*-test provides a test to determine whether the means of the groups are statistically the same.[44] Because the experimental samples from different days were independent, we pooled the temporal $\Delta\Delta C_t$ data from days 0-14 for each colony size into a single population and conducted student's *t*-tests for 17 neural genes. We subjected the resulting *p* values to the Benjamini–Hochberg FDR correction to avoid the random false positive identification of significant genes. The results are included in Table 1. After applying FDR corrections, the expression of none of the genes was significantly different between medium and small colonies. Only CDH2 expression was different between large and medium colonies. And eight genes, CDH2, Pax6, Notch1, ChAT, Synaptophysin, Wnt1, Sox1, and GAP43, were also different between large and small colonies.

It is important to note that there were genes such as NCAM, GFAP, TH, and MAP2 that had hundreds of fold-change differences among different colony sizes (Fig 4 and Fig 5). However, *t*-tests performed ignoring the temporal heterogeneity of the data failed to elicit statistically significant differences in the expression of these genes among the three colony sizes. Student's *t*-test functions on the assumption that the sequence of observations is independent and normally distributed.[45] However, with the dynamic temporal profile of gene expression and large differences in expression levels between

days 0 and 14 within each group, the assumption is not fulfilled. Therefore, simply applying *t*-tests without taking into account the temporal dynamics of the data is insufficient to establish the significance of some of the key genes in colony size-mediated neural differentiation of mESCs.

Mann-Whitney U-test

We also considered another conventional statistical technique, a non-parametric equivalent of *t*-test known as Mann-Whitney U-test or Wilcoxon rank-sum test. This test converts original quantities to ranks (e.g., the smallest quantity has rank=1, the second smallest quantity has rank=2, etc.) before calculating the statistical parameter U and its significance.[46,47] Due to performing multiple tests on each experimental group, we applied FDR correction to avoid false positive results. The genes that were identified statistically different between different colony sizes are listed in Table 2 with their respective adjusted *p*-values post FDR correction. While the expression of none of the genes was statistically different between medium and small colonies, only four genes, Nestin, CDH2, Sox1, and ChAT, showed differential expression between large and medium colonies, and only 11 genes were identified different between large and small colonies. In contrast, our multi-step analysis using change point detection showed that all 17 neural genes were differentially expressed between large and small colonies at some point over the 14-day period.

Unlike *t*-test, Mann-Whitney U-test does not assume normal distribution of data. However, it still assumes that the sequence of data is independent and identically distributed. Besides, an inherent disadvantage of this test is substituting ranks for the original experimental quantities and loss of information on magnitude of differences

between the experimental groups. Although Mann-Whitney U-test identified more differentially-expressed genes between pairs of different colony sizes, it still was not as powerful as the multiple tests following the change point detection analysis.

Both student's *t*-test and Mann-Whitney U-test are widely used statistical techniques, but they failed to elicit all differentially-expressed genes identified by our multi-step method. This indicates that without taking the change points into account, these methods are inadequate to analyze dynamic and heterogeneous data such as our gene expression time series data. Temporal heterogeneity in the expression of each gene increased the variance of each population, resulting in wide overlap in the gene expression data distribution between each two colony sizes. For example, the NCAM gene had a few folds difference between large and small colonies during days 0-6, but more than 100 folds after day 7 (Fig 5a). The *t*-test performed on the sub-population (from day 7-14) obtained after applying change point detection identified this difference as statistically significant. However, the *t*-test performed on the entire population (from day 0-14) failed to identify NCAM as a statistically significant differentially-expressed gene between large and small colonies because very low gene expression differences during days 0-6 masked the significant differences during days 7-14.

Another shortcoming of conventional *t*-test and Mann-Whitney U-test is disregarding the temporal dynamics of gene expression, making it difficult to know at what time points gene expression differences arise or reach a maximum. On the contrary using the change point detection method, we could temporally correlate the phenotypic changes observed in differentiating cells in colonies of different sizes, with the expression of significant genes at different time points.

The change point detection step can be particularly helpful in cases of multivariate problems such as genomic and proteomic studies concerned with the analysis of data consisting of "large p (i.e., the number of factors or measurements), and small n (i.e., the number of independent samples)".[2] Although the commonly used conventional statistical tests perform well when sufficient experimental replicates are available, adding an experimental replicate in such experimental setting involves additional cell culture and/or preparation steps, and intensive genomic or proteomic experiments. The change point detection method addresses this issue by grouping the measurements into homogenous populations and increasing the sample size to enable the use of conventional statistical tools to reliably identify molecular markers of interest.

Conclusions

We modulated the intercellular interactions of differentiating stem cells by controlling the size of stem cell colonies and evaluated the resulting changes in the efficiency of neural differentiation through a temporal gene expression study and a multi-step biostatistical analysis approach. This enabled capturing and comparing the evolution of gene markers of neural cells in differentiating stem cell colonies of different sizes. The results elucidated molecular markers of size-disproportionate enhanced neural differentiation of stem cells by increase in colony size. We established that applying a multi-step statistical analysis consisting of hierarchical clustering, E-divisive change point detection, and traditional multiple testing to temporal gene expression data of different experimental groups (colony sizes) identify major genes that underlie differences in the neural differentiation efficiency of stem cells. Beyond this work, our proposed approach

will benefit any high throughput and high content screening in drug and biomarker discovery that deals with temporally varying multivariate data analysis.

Acknowledgment

This research is supported by grants 1264562 from National Science Foundation, CA182333 from National Institutes of Health, and TECG20140954 from Ohio Third Frontier.

Disclosure of Potential Conflict of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1 R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan and Y. Wang, *Nat. Rev. Cancer*, 2008, **8**, 37–49.

2 J. E. Mcdermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey and K. D. Rodland, *Expert Opin. Med. Diagnosis*, 2013, **7**, 37–51.

3 G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati and E. Abadie, *Clin. Cases Miner. Bone Metab.*, 2008, **5**, 149–154.

4 R. P. Horgan and L. C. Kenny, *Obstet. Gynaecol.*, 2011, **13**, 189–195.

5 J. A. Ludwig and J. N. Weinstein, *Nat. Rev. Cancer*, 2005, **5**, 845–856.

6 Å. M. Wheelock and C. E. Wheelock, *Mol. Biosyst.*, 2013, **9**, 2589–2596.

7 R. Goni, P. García and S. Foissac, *Integromics White Pap.*, 2009, **1**, 1–9.

8 R. R. Kitchen, M. Kubista and A. Tichopad, *Methods*, 2010, **50**, 231–236.

9 H. Tavana, B. Mosadegh and S. Takayama, *Adv. Mater.*, 2010, **22**, 2628–2631.

10 R. Joshi, P. S. Thakuri, J. C. Buchanan, J. Li and H. Tavana, *Adv. Healthc. Mater.*, 2017, 1700832.

11 R. Joshi, J. C. Buchanan and H. Tavana, *Integr. Biol.*, 2017, **5**, 26.

12 H. Tavana, A. Jovic, B. Mosadegh, Q. Y. Lee, X. Liu, K. E. Luker, G. D. Luker, S. J. Weiss and S. Takayama, *Nat. Mater.*, 2009, **8**, 736–741.

13 R. Joshi, J. C. Buchanan, S. Paruchuri, N. Morris and H. Tavana, *PLoS One*, 2016, **11**, e0166316.

14 D. S. Matteson and N. A. James, *J. Am. Stat. Assoc.*, 2014, **109**, 334–345.

15 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. B*, 1995, **57**, 289–300.

16 H. Kawasaki, K. Mizuseki, S. Nishikawa, S. Kaneko, Y. Kuwana, S. Nakanishi, S. I. Nishikawa and Y. Sasai, *Neuron*, 2000, **28**, 31–40.

17 H. Tavana, B. Mosadegh, P. Zamankhan, J. B. Grotberg and S. Takayama, *Biotechnol. Bioeng.*, 2011, **108**, 2509–2516.

18 Z. Wang, E. Oron, B. Nelson, S. Razis and N. Ivanova, *Cell Stem Cell*, 2012, **10**, 440–454.

19 N. Osumi, H. Shinohara, K. Numayama-Tsuruta and M. Maekawa, *Stem Cells*, 2008, **26**, 1663–1672.

20 M. Venere, Y.-G. Han, R. Bell, J. S. Song, A. Alvarez-Buylla and R. Blelloch,

Development, 2012, **139**, 3938–49.

21 N. Rosskothen-Kuhl and R.-B. Illing, *PLoS One*, 2014, **9**, e92624.

22 R. Nugent and M. Meila, *Methods Mol. Biol.*, 2010, **620**, 369–404.

23 H. Ji and X. S. Liu, *Nat. Biotechnol.*, 2010, **28**, 337–340.

24 X. Wang, A. Zhang, Y. Han, P. Wang, H. Sun, G. Song, T. Dong, Y. Yuan, X. Yuan, M. Zhang, N. Xie, H. Zhang, H. Dong and W. Dong, *Mol. Cell. Proteomics*, 2012, **11**, 370–380.

25 O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein and R. B. Altman, *Bioinformatics*, 2002, **18**, 1454–1461.

26 Y. Wang, C. Wu, Z. Ji, B. Wang and Y. Liang, *PLoS One*, 2011, **6**, e20060.

27 A. Alyass, M. Turcotte and D. Meyre, *BMC Med. Genomics*, 2015, **8**, 33.

28 R. Joshi and H. Tavana, *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, 3557–3560.

29 R. Joshi, J. Buchanan and H. Tavana, *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, 4173–4176.

30 M. Parmar and M. Li, *BMC Dev. Biol.*, 2007, **7**, 86.

31 S. N. Sansom, D. S. Griffiths, A. Faedo, D. J. Kleinjan, Y. Ruan, J. Smith, V. Van Heyningen, J. L. Rubenstein and F. J. Livesey, *PLoS Genet.*, 2009, **5**, e1000511.

32 S. Bel-Vialar, F. Medevielle and F. Pituello, *Dev. Biol.*, 2007, **305**, 659–673.

33 A. L. Perrier, V. Tabar, T. Barberi, M. E. Rubio, J. Bruses, N. Topf, N. L. Harrison

and L. Studer, *Proc. Natl. Acad. Sci.*, 2004, **101**, 12543–12548.

34 P. B. Kuegler, B. Zimmer, T. Waldmann, B. Baudis, S. Ilmjärv, J. Hescheler, P. Gaughwin, P. Brundin, W. Mundy, A. K. Bal-Price, A. Schrattenholz, K.-H. Krause, C. van Thriel, M. S. Rao, S. Kadereit and M. Leist, *ALTEX*, 2010, **27**, 17–42.

35 J. Zhang, J. R. Shemezis, E. R. McQuinn, J. Wang, M. Sverdlov and A. Chenn, *Neural Dev.*, 2013, **8**, 7.

36 S. Alimperti and S. T. Andreadis, *Stem Cell Res.*, 2015, **14**, 270–282.

37 L. Kan, A. Jalali, L.-R. Zhao, X. Zhou, T. McGuire, I. Kazanis, V. Episkopou, A. G. Bassuk and J. A. Kessler, *Dev. Biol.*, 2007, **310**, 85–98.

38 D. Park, A. P. Xiang, F. F. Mao, L. Zhang, C.-G. Di, X.-M. Liu, Y. Shao, B.-F. Ma, J.-H. Lee, K.-S. Ha, N. Walton and B. T. Lahn, *Stem Cells*, 2010, **28**, 2162–2171.

39 N. I. Perrone-Bizzozero and D. C. Tanner, in *Handbook of Neurochemistry and Molecular Neurobiology*, 2006, 315–329.

40 F. Walsh and P. Doherty, *Annu. Rev. Cell Dev. Biol.*, 1997, **13**, 425–456.

41 M. a. Kinney, R. Saeed and T. C. McDevitt, *Integr. Biol.*, 2012, **4**, 641–650.

42 M. A. Kinney, R. Saeed and T. C. McDevitt, *Sci. Rep.*, 2014, **4**, 4290.

43 Y.-S. Hwang, B. G. Chung, D. Ortmann, N. Hattori, H.-C. Moeller and A. Khademhosseini, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 16978–16983.

44 T. K. Kim, *Korean J. Anesthesiol.*, 2015, **68**, 540–546.

45 E. Whitley and J. Ball, *Crit. care*, 2002, **6**, 424–8.

46 R. M. Conroy, *Stata J.*, 2012, **12**, 182–190.

47 N. Nachar, *Tutor. Quant. Methods Psychol.*, 2008, **4**, 13–20.

Figures

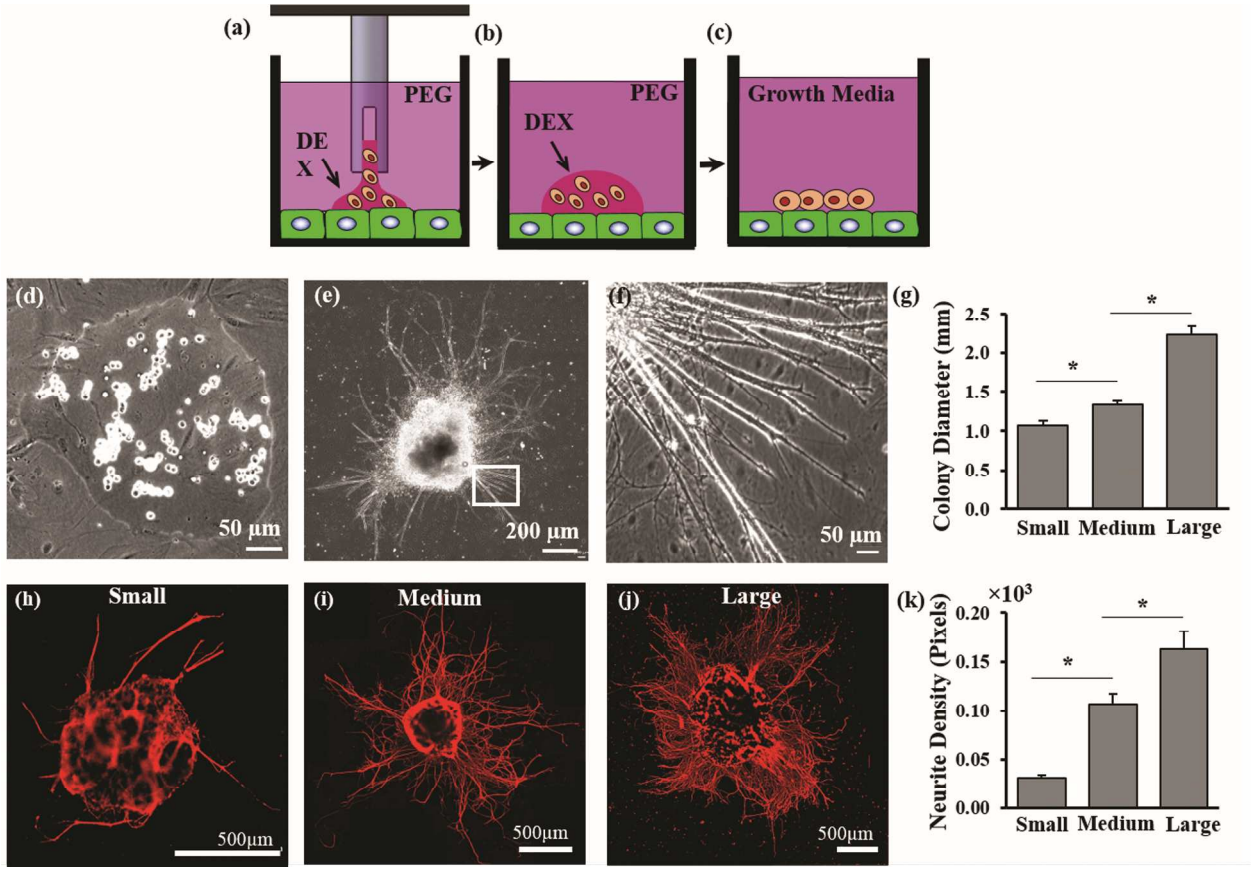

Figure 1. Aqueous two-phase system mediated cell printing generates controlled size colonies. (a) Slot pins loaded with mESCs (beige) in the DEX phase is lowered on to PA6 cells (green) immersed in the PEG phase. (b, d) Pins content is autonomously dispensed to form isolated drops confining mESCs. (c) Printed mESCs attach to the stromal layer and proliferate to form a single colony. (e) A colony on day 8 of culture. (f) Neurite processes extend out from differentiating cells in mESC colonies. (g) Day 8 measured diameter of colonies formed with a density of 100, 250, and 500 mESCs. (h-j) Immunostained images of TuJ-positive colonies of 3 different sizes on day 8 of culture. (k) Day 8 measured total neurites density normalized with colony perimeter. * $p < 0.01$. n=18

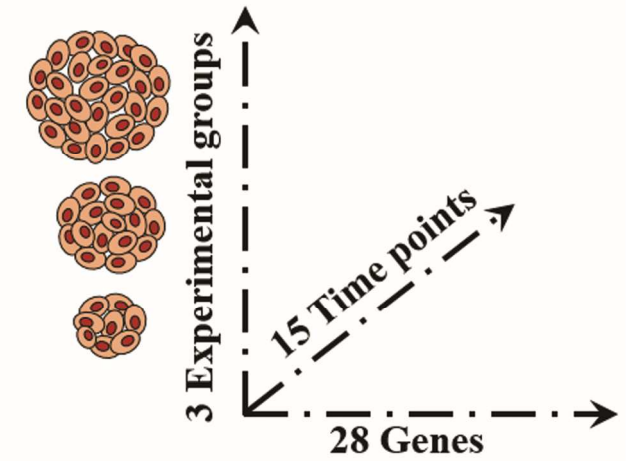

Figure 2. Schematic representation of the experimental setup.

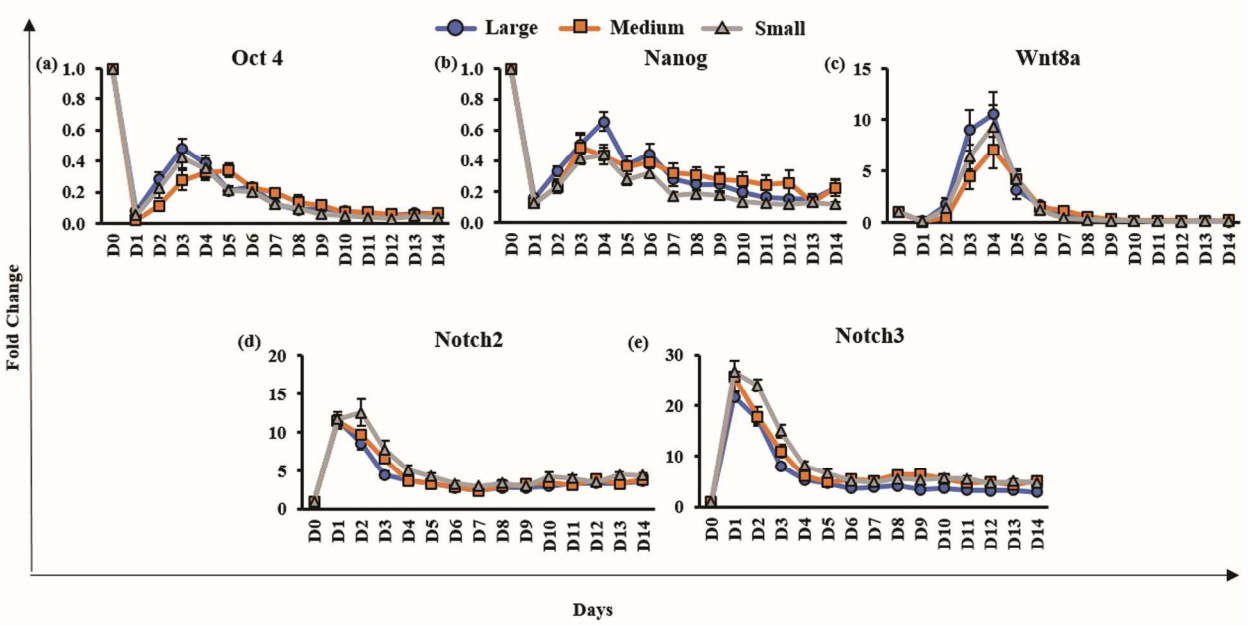

Figure 3. mRNA fold change values of pluripotency marker genes during 2-week culture period. n=3

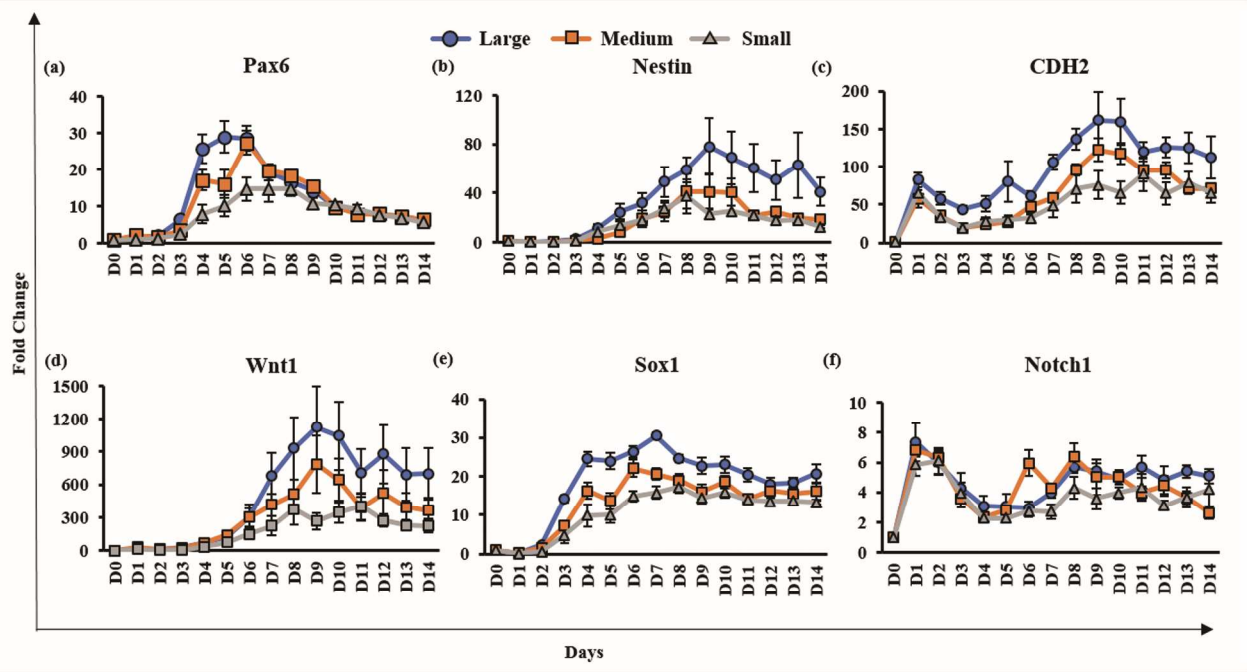

Figure 4. mRNA fold change values of neural stem cell genes during 2-week culture period. n=3

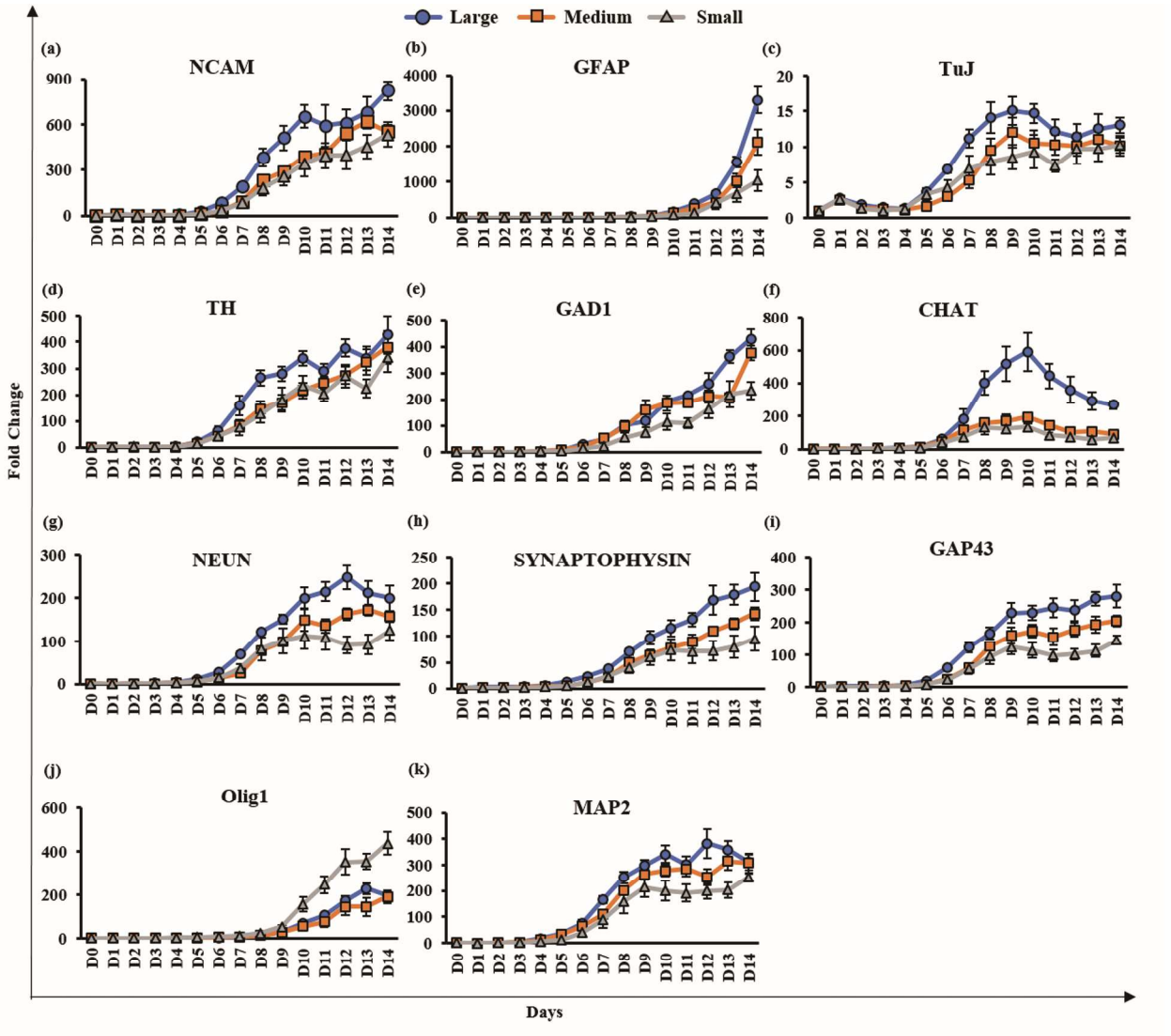

Figure 5. mRNA fold change values of specific neuronal and glial cell genes during 2-week culture period. n=3

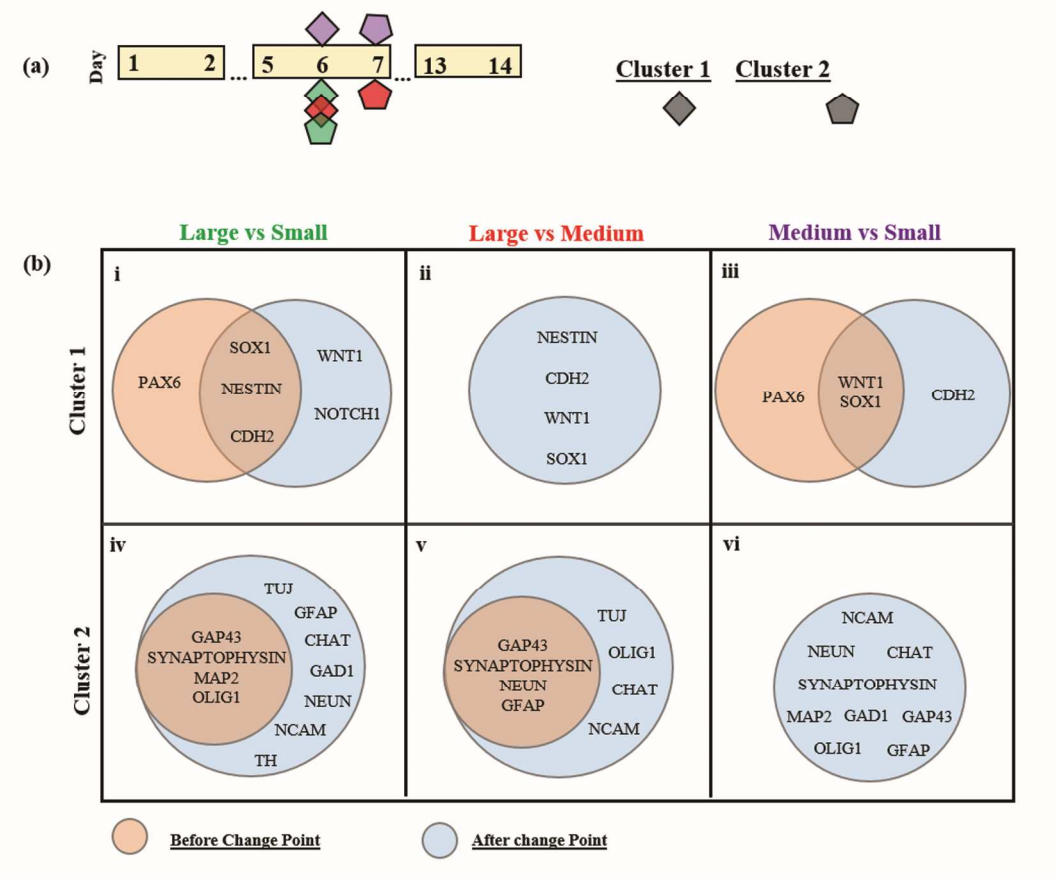

Figure 6. (a) Detected gene expression change points among three different colony sizes. (b) Identification of genes with distinct expression patterns between the colonies before and after the change point. To generate the Venn diagrams, the nominal significance level was chosen at 0.01 to control the FDR.

Table 1. Differentially expressed genes identified by *t*-test with the false discovery rate (FDR) controlled at the nominal significance level of 0.01.

Large vs Small

Genes	p-value
CDH2	$<10^{-4}$
PAX6	0.003
NOTCH1	0.003
CHAT	0.003
SYNAPTOPHYSIN	0.007
WNT1	0.009
SOX1	0.009
GAP43	0.009

Large vs Medium

Genes	p-value
CDH2	$<10^{-4}$

Table 2. Differentially expressed genes identified by Mann-Whitney U test with the false discovery rate (FDR) controlled at the nominal significance level of 0.01.

Large vs Small

Genes	p-value
NESTIN	$<10^{-4}$
CDH2	$<10^{-4}$
PAX6	0.004
TUJ	0.003
SOX1	$<10^{-4}$
WNT1	0.003
SYNAPTOPHYSIN	0.002
CHAT	$<10^{-4}$
NEUN	0.004
MAP2	0.007
GAP43	$<10^{-4}$

Large vs Medium

Genes	p-value
NESTIN	$<10^{-4}$
CDH2	$<10^{-4}$
SOX1	$<10^{-4}$
CHAT	0.007