Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Woo Jeong December 9, 2019

Gene Expression Analysis of Endothelial Cells Derived from Human Induced Pluripotent Stem Cells

by

Woo Jeong

Dr. Young-sup Yoon
Adviser

Department of Quantitative Theory and Methods

Dr. Young-sup Yoon
Adviser

Dr. Jeremy Jacobson
Committee Member

Dr. Nicole Gerardo
Committee Member

Dr. Leonard Carlson
Committee Member

2019

Gene Expression Analysis of Endothelial Cells Derived from Human Induced Pluripotent Stem Cells

By

Woo Jeong

Dr. Young-sup Yoon
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Quantitative Theory and Methods

2019

Abstract

Gene Expression Analysis of Endothelial Cells Derived
from Human Induced Pluripotent Stem Cells

By Woo Jeong

Blood vessels play an essential role in transporting oxygen and nutrients to tissues, leading to tissue homeostasis. Cardiovascular ischemic diseases, such as peripheral artery disease, are highly linked with damaged and dysfunctional blood vessels. The damaged blood vessels restrict efficient blood supply to tissues, leading to shortages of oxygen and nutrients and ultimately to dysfunctional tissues. Human induced pluripotent stem cells (hiPSCs), which have an unlimited proliferation capacity to differentiate into any type of somatic cells without ethical issues, were cultured under a fully defined and clinically compatible system to differentiate hiPSCs into endothelial cells (ECs). The resultant hiPSC-derived ECs (hiPSC-ECs) showed highly enriched and genuine EC characteristics and proangiogenic properties. However, the gene expression profile that facilitates endothelial and proangiogenic characteristics has been only partially explored. To expand understanding, we used RNA sequencing to identify differentially expressed (DE) genes and enriched pathways that significantly contribute to genuine EC features and proangiogenic attributes. Total RNA of hiPSC-ECs were marked at 6 EC development timepoints: Day0, Day2, Day4, Day8, Day14 before sorting, and Day14 after sorting, and we compared all later timepoints to Day0. We discovered that gene ontology (GO) terms for biological processes were enriched in hiPSC-ECs in EC differentiation, EC proliferation, EC migration, and positive regulation of angiogenesis. Furthermore, we identified 28 DE genes among 1252 DE genes that significantly contribute to the endothelial and proangiogenic characteristics by comparing hiPSC-ECs (Day14 after sorting) to hiPSC (Day0). The results provide new insight into a transcriptomic understanding of hiPSC-ECs, and the identified DE genes may serve as therapeutic markers of hiPSC-ECs.

Gene Expression Analysis of Endothelial Cells Derived from Human Induced Pluripotent Stem Cells

By

Woo Jeong

Dr. Young-sup Yoon

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Quantitative Theory and Methods

2019

Acknowledgements

First and foremost, I wish to thank Dr. Young-sup Yoon for providing an opportunity for me to participate in the Honors Thesis Program. Second, I wish to express appreciation to Dr. Young-sup Yoon and Dr. Shin-Jeong Lee for providing the RNA sequencing data for analysis. Third, I would like to express the sincerest gratitude to Dr. Young-sup Yoon, Dr. Kyung Hee Kim, and Dr. Brandon Johnson for making this study possible and providing consistent support, guidance, and inspiration throughout my time at Yoon Lab since the freshman year. Fourth, I would like to express immense gratitude to the committee members, Dr. Jeremy Jacobson, Dr. Nicole Gerardo, and Dr. Leonard Carlson, for serving on my thesis committee and providing insightful feedback. Last but not least, I wish to show my deepest appreciation to my family for always providing me with unconditional love, support, and trust.

Table of Contents

List of Figures and Tables

Introduction

Dysfunctional blood vessels increase morbidity and mortality. The cardiovascular ischemic diseases, myocardial infarction (MI) and peripheral arterial disease (PAD), are leading causes of mortality worldwide.[1] PAD has reached about 12% and affects more than 25 million people in North America and Europe. PAD clinically results in claudication, soreness with walking, and critical limb ischemia (CLI), soreness at rest in the limb.[2] Hence, innovative regenerative medicine, particularly pluripotent stem cells, that is capable of curing ischemic diseases is receiving remarkable attention.

Pluripotent stem cells, which possess the capability to grow infinitely and differentiate into all types of somatic cells, are classified into two types: embryonic stem cells (ESCs) and human induced pluripotent stem cells (hiPSCs). However, due to the destruction of blastocytes of human embryos, use of ESCs raises ethical and moral concerns. hiPSCs are simply generated by combining Yamanaka transcription factors in isolated adult somatic cells and resemble ESCs in gene expression, growth properties, and morphology.[3] Therefore, hiPSCs are widely acclaimed as a way to generate vascular cells and are used largely for applications in regenerative medicine.[3,4] To treat cardiovascular diseases, cell therapy using cellular components of blood vessels derived from hiPSCs has emerged as a promising candidate for vascular regeneration therapy as the loss of vascular supply is a main pathophysiologic feature of ischemic diseases.

Blood vessels have two types of main cellular components: endothelial cells (ECs) and vascular smooth muscle cells (VSMCs). Both ECs and VSMCs have vital roles in blood pressure control, interactions with immune cells, and the uptake of nutrients.[2] Between the two, ECs refer to a thin layer of cells in the endothelium that lines the interior surface of blood vessels generating interface between circulating blood and the vessel walls.[5] ECs construct a single cell layer that

covers all blood vessels and facilitates exchanges between circulating blood and the surrounding tissues. Furthermore, ECs modulate the growth and reactivity of the underlying smooth muscle, control the interaction of the vessel wall with circulating blood elements, and regulate vascular responses to hemodynamic forces.[6] ECs can be differentiated from hiPSCs in chemically defined conditions.[7] Lee *et al.* developed a clinically compatible protocol to generate ECs derived from hiPSCs (hiPSC-ECs).[8] In clinical settings, they have demonstrated high potential to treat ischemic conditions by cell therapy through the angiogenic therapeutic potential, which is the capability to generate new blood vessels.[2,8]

The functionality of cells depends on the expression of different genes and proteins induced. High-throughput RNA sequencing (RNA-seq) is a technology that enables sequencing a large amount of transcriptome from mRNA to cDNA to measure the expression of isoforms and unknown transcripts with a better sequencing quality, cheaper cost, and shorter time.[9] Moreover, RNA-seq maintains accuracy and high correlation with PCR and has high coverage of the transcriptome in the discovery of differentially expressed (DE) genes.[10] Hence, gene expression levels generated by RNA-seq data can be analyzed for differential expression (DE) to produce gene expression profiles that are distinct for different conditions. DE analysis is particularly useful in investigating gene expression patterns over time and in identifying enriched pathways during EC differentiation.[11,12]

In the EC differentiation protocol from hiPSCs, there are three distinct stages of differentiation from hiPSCs to ECs: mesoderm induction, EC differentiation, and EC enrichment. On Day0, initial hiPSCs started to be cultured. On Day2, mesoderm was induced from hiPSCs. On Day4 and Day8, hiPSCs differentiated into ECs. On Day14, ECs derived from hiPSCs (hiPSC-ECs) were enriched. To further enrich endothelial lineage cells, hiPSC-ECs were sorted for CDH5 by the magnetic-labeled cell separation system on Day14. Furthermore, Lee *et al.* demonstrated

that hiPSC-ECs generated by this highly efficient EC differentiation system showed proangiogenic potential and direct vessel-forming effects in a hindlimb ischemic mouse model.[8] The therapeutic angiogenetic potential expressed from hiPSC-ECs triggers curiosity to identify genes that significantly contribute to those functions. In this study, we investigate differentially expressed (DE) genes during EC differentiation, identify gene ontology (GO) terms that DE genes during EC differentiation and enrichment belong to, and identify promising DE genes that are highly associated with endothelial and proangiogenic therapeutic properties.

Methods

Dataset. We acquired a dataset of hiPSC-ECs for six sample groups, Day0, Day2, Day4, Day8, Day14 before sorting, and Day14 after sorting, in FASTQ format from Dr. Shin-Jeong Lee at Emory University School of Medicine. Each group has a total of 4 samples. Each group contains two biological replicates, each of which has two technical replicates. Therefore, we have a total of 24 samples for the six sample groups.

Raw Data Generation. We uploaded the 24 files in FASTQ format on Galaxy, which is open web-based software for genomic analysis. We executed quality control on the RNA-seq data using FastQC software to assess the sequencing quality of the data. To compare the replicates within each condition group, we used MultiQC software to merge the results of the quality control by FastQC software.[13] Next, we aligned data that had passed the quality control to the reference genome, GRCh38.p12 (GCA_000001405.27), which is the most recently updated *Homo sapiens* genome in the UCSC Genome Browser. In this alignment step, we determined the location of the genome where the reads originated from the reference genome by using HISAT2 software.[14,15] The outputs of alignment generated Sequence Alignment Map (SAM) format files were converted to BAM format files, which are much smaller than SAM format files in size. The

BAM format files included the genome mapping information and alignment quality.[16] Following alignment, we counted the reads aligning to exons of each gene using ht-seq software with the general feature format (GFF) file gen3code_v29_annotation_gff3.gz, which depicts genes and other features of DNA, RNA and protein of the reference genome, human GRCh38.12. We incorporated union mode if any form of interaction with the genes to the reference genome was present. Each read quantified by ht-seq software yielded a count matrix in tabular format with an appropriate quality and subject.[15,17,18]

Sample Level Quality Control. We conducted computational analysis in R version 3.6.1, which is a programming language for statistical computing. We first converted the 12 tabular format files into text format files. Following the conversion, we imported the 12 text files into R and merged them as a single data frame with appropriate column names: “D.0_1”, “D.0_2”, “D.2_1”, “D.2_2”, “D.4_1”, “D.4_2”, “D.8_1”, “D.8_2”, “D14.not.sorted_1”, “D14.not.sorted_2”, “D14.sorted_1”, and “D14.sorted_2”. Then, we generated metadata that represented the corresponding conditions: “Day.0”, “Day.2”, “Day.4”, “Day.8”, “Day14 before sorting”, and “Day14 after sorting”. We confirmed that the column names of the raw data matched with the raw names of the metadata to proceed to the next procedure.

Normalization. As differential expression (DE) analysis for the RNA-seq samples was to compare the relative gene expression levels between the different condition groups, technical artifacts that affected the gene expression counts were normalized for a proper apple-to-apple comparison.[19] Among many normalization methods, we used the “median of ratio method of normalization” adopted by the DESeq2 package because this method normalizes RNA-seq data with high efficiency.[20,21] To account for the sequencing depth and RNA composition and eliminate technical artifacts other than RNA expression, the normalization method divided by sample-

specific normalization factors by median ratio gene counts relative to geometric mean per gene. First, DESeq2 generated a reference sample, which is the geometric mean of all the condition groups, across all genes. Second, DESeq2 measures the ratio of each sample to the reference by dividing the sample expression for each sample group of each gene by the reference sample created in the previous step. Third, DESeq2 determines the median of the ratio of the sample to the reference sample as the normalized factors for each sample. Lastly, DESeq2 generates normalized count values by dividing the original count values by the normalized factors across all genes.[11] This normalization method is highly appropriate for differential expression (DE) analysis as it accounts for library size by geometric mean and established resistance to a large number of DE genes by median values.[11,22]

Hierarchical Clustering Analysis. After normalization, we executed hierarchical clustering analysis to detect potential outliers and contamination within the data by exploring similarities among the conditions.[23] First, we conducted variance stabilized transformation to incorporate approximately stabilized variance and correction for normalized factors to mediate the heteroscedasticity of the data.[24] After the variance stabilized transformation, we calculated the Pearson correlation values for all pairwise combinations of the samples in the data. Then, we created a hierarchical clustering heatmap in which each box represented Pearson correlation values in colors from blue to red.[23]

Principal Component Analysis (PCA). We plotted the normalized count values in 12-dimensions as we had a total of 12 samples. After plotting, the most extensive spread in the data was signified as PC1, and the second-largest spread perpendicular to the PC1 in the data was present as PC2. After plotting the line for spread and establishing the amount of influence per gene, PCA computed a per sample score. The per sample PC value referred to the product of the

influence and the normalized read count and summation across all genes. This mechanism continued until the number of total samples were indicated.[23,24] However, using the dimension reduction mechanism, we projected the high-dimensional data into only two dimensions representing the two greatest spreads.

Gene-level Quality Control. By gene-level quality control, we elevated the chance to detect DE genes more accurately and efficiently. The DESeq2 package eliminated the following three gene groups: genes with zero counts in all samples, outlier genes with extreme count, and genes with low mean normalized counts.[23]

DESeq2 Generalized Linear Model. RNA-seq data typically consist of a small number of replicates due to the relatively expensive cost.[25] Therefore, the RNA-seq data demonstrate a robust mean-variance relationship that is not statistically inferred by normal-based parametric hypothesis testing, such as Student t-test, to assess the statistical significance. Instead, we applied the generalized linear model of the DESeq2 package to fit the data to the negative binomial distribution and used Wald Test to draw statistical inference.[25]

$$\begin{aligned} \lambda_i &= np_i \\ Y_i \mid \lambda_i &\sim Poisson(\lambda_i) \end{aligned} \tag{1}$$

RNA-seq gene count data are modeled by the binomial distribution, the probability of getting success based on the number of trials.[26] However, not all the count data are modeled by the binomial distribution, which is only for discrete events. When the probability of an event is minuscule with the large number of trials, the Poisson distribution is appropriate to fit the gene expression levels as the Poisson distribution is for continuous events (1).[26]

$$\begin{aligned} Y_{ij} \mid \lambda_i &\sim Poisson(\lambda_i) \\ \lambda_i &\sim Gamma(\alpha, \beta) \end{aligned} \tag{2}$$

As the chance of selecting a specific transcript from the vast number of RNA is minimal, the Poisson distribution is an appropriate model to fit RNA-seq gene count data.[26] However, the Poisson distribution has a particular property that the mean and variance are identical. In reality, biological variation across the samples is always present in RNA-seq data. Genes with larger mean expression levels are more likely to have substantial variances across replicates. Under the assumption that data are appropriately normalized, the counts for a specific gene *i* and a specific replicate *j* is modeled by the following hierarchical model (2).[26]

$$Y_{ij} \sim NegBin(\alpha, \beta) \tag{3}$$

Marginally, the Gamma-Poisson compound distribution is the over-dispersed Poisson distribution, which is the Negative Binomial distribution.[26,27] Therefore, the gene counts for multiple replicates were fit to the Negative Binomial distribution with the variance more significant than the mean (3).[26,27] To confirm whether the gene count data fit the DESeq2, we generated two diagnostic plots: mean-variance plot and dispersion-mean plot. We counted the mean and variance value for each gene across all conditions and generated a new data frame with the mean and variance values. We plotted the mean and variance and incorporated a logarithmic scaling on both the x and y-axis. We then drew a linear regression line by the ordinary least-square method for all the genes plotted.[23]

$$Var = \mu + \alpha\mu^2 \tag{4}$$

To investigate more in-depth about the difference in spread among different the mean estimates, we plotted the dispersion-mean plot. Dispersion is a measure of spread used explicitly in the DESeq2 package by variance for a specific mean estimate (4). Due to the small number of samples of RNA-seq data, the DESeq2 package incorporated a shrinkage method to generate more accurate estimates of variability.[28,29]

Differential Expression (DE) Analysis. To execute differential expression (DE) analysis, we used the DESeq2 package from Bioconductor. First, normalized counts for each sample and one dispersion estimate for each gene were used as inputs to calculate a gene count in each sample group. Second, we determined the gene-wise dispersions and shrank the dispersion estimates to the fitted estimates of dispersion for more accurate maximum likelihood estimates of dispersion. Third, we used the generalized linear model to gene counts with the negative binomial distribution.[23] Fourth, we performed Wald test for all possible pairwise comparisons with Day0 as the base-level treatment to examine the null hypothesis that there was no differential expression across the two sample groups (LFC=0) with an alpha of 0.05. The analysis induced a collection of DE genes with the following outputs: "baseMean" for average normalized counts for all samples, "log2FoldChange" for the gene expression difference between the two selected conditions, "lfcSE" for fold change standard error, "stat" for the Wald Test statistics, "pvalue" for p-values for the Wald Test, and "padj" for Benjamin-Hochberg (BH) multiple testing adjusted p-values.[23] We mainly used padj instead of regular p-value due to an issue of false discovery that would inhibit us from accurately identifying the true positive genes by generating a 5% chance that the gene was DE when it was actually not.[30] To remedy this issue, we implemented multiple test correction by BH to control the rate of false positives relative to the true, inducing padj. In addition to the padj threshold of 0.05, we incorporated log2 fold change thresholds of 1 for the up-regulated genes and -1 for the down-regulated genes. Therefore, if genes have padj smaller than 0.05 and log2 fold change either greater than 1 or smaller than -1, we rejected the null hypothesis and identified the genes as DE. We performed the Wald test for all pairwise comparisons with Day0 as the base-line sample.

Data Visualization. We visualized the results of DE analysis by Wald Test in several graphics. First, we created MA plots for a global view of the distribution of DE genes in fold change relative

to the normalized mean counts. Since low mean counts are vulnerable to imprecisely captured log fold changes (LFC), we additionally incorporated LFC shrinkage that more accurately captured DE genes.[29] We compared the number of DE genes for both up- and down-regulated across all the condition samples. Second, we created expression heatmaps to detect gene expression similarities and differences among the samples. To generate expression heatmaps, we subset the significant DE genes of Day14 after sorting based on padj. The expression heatmaps scaled by row and induced Z-scores, where the normalized count values for each sample and each gene is subtracted by the mean and is divided by the standard deviation. Third, we generated volcano plots to visualize statistical significance relative to the biological significance of all the genes analyzed. To create volcano plots, we generated a binary variable on the column of the data to capture the significance of genes in two different colors. All the genes tested by DE analysis were plotted with the corresponding padj scaled by negative logarithm and the log2 fold change.

Functional Analysis. To investigate enriched functionalities during EC differentiation and enrichment, we used the three results of DE analysis: “Day0 vs. Day8”, “Day0 vs. Day14 before sorting”, and “Day0 vs. Day14 after sorting”. To illustrate the distribution of up- and down-regulated genes among the timepoints, we generated two triple Venn Diagrams. Next, we used the clusterprofiler package from Bioconductor to implement GO enrichment analysis to understand the functions of the collections of up- and down-regulated genes.[31] We first converted the DE gene symbols into Ensembl IDs while eliminating duplicate IDs. Then, we generated a background gene list for all the genes tested for each DE analysis. For analysis, the number of genes associated with a category that overlap with the set of DE genes follow hypergeometric distribution.[32] Under the Hypergeometric Test, we estimated the probability of genes with the specific category in the gene list according to the chance of genes in the background set.[33] For GO enrichment analysis, we focused on the biological process out of the three categories, biological process, molecular function, and cellular component.[34] After we acquired the results,

we extracted the significantly enriched GO terms of our interest and compared statistical significance among Day8, Day14 before sorting, and Day14 after sorting.

Gene Ontology (GO) Analysis. We acquired GO terms we were interested in from Broad Institute (http://software.broadinstitute.org/gsea/index.jsp): GO EC development, GO EC differentiation positive, GO EC proliferation positive, GO EC migration positive, GO angiogenesis, and GO blood vessel remodeling. For genes in these GO terms, we identified the trends of gene expressions over time. Concentrating on those upregulated on Day14 after sorting but downregulated on Day0, we confirmed via the result of DE analysis whether those genes were differentially expressed under the statistical test. Then, we researched the biological functionality of genes through Uniprot (https://www.uniprot.org/).

Results

Understanding of general workflow of RNA-seq data analysis

There are two main stages of RNA-seq pipelines: first to generate gene count data and second to perform DE analysis on the raw count data (Figure 1). The first part to induce gene count data is comprised of six stages. First, we conducted biological sample library preparation. Second, we generated sequence reads. Third, we executed quality control to assess the quality of RNA-seq. Fourth, we aligned the RNA-seq data to the reference genome. Fifth, we quantified reads associated with genes from the alignment. Now, we acquired gene count data.

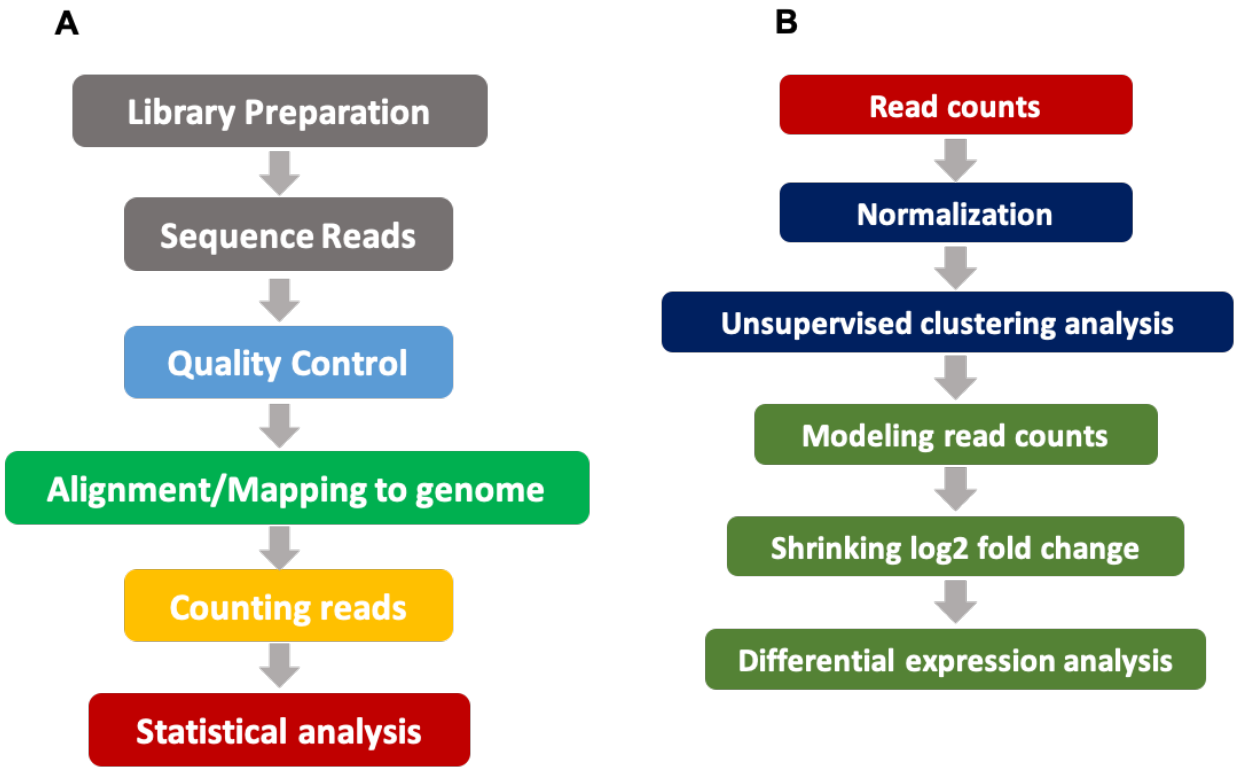


Figure 1. Overview of RNA-Seq Pipeline. A. The 6 stages of RNA-Seq pipelines. Stage 1: biological sample library preparation. Stage 2: sequence reads. Stage 3: quality control. Stage 4: splice-aware mapping to genome, Stage 5: counting reads associated with genes. Stage 6: statistical analysis to identify differentially expressed genes. **B.** The 6 stages of computational analysis of RNA-Seq of hPSC-ECs. Stage 1: normalization to remove technical variations. Stage 2: unsupervised clustering analyses to show pairwise correlation between samples and variation present in the data. Stage 3: modelling raw counts for each gene to fit the read counts into the DESeq2 generalized linear model. Stage 4: shrinking log2 fold changes to improve gene estimates with low mean counts. Stage 5: testing for differential expression to identify differentially expressed genes between the selected conditions.

The second part to perform DE analysis is composed of five stages. First, we normalized the count data to remove technical variations that affect the count reads. Second, we performed unsupervised clustering analysis to explore similarities among the samples and distinguish the source of variation present in the data. Third, we model read counts for each gene to fit the reads into the DESeq2 generalized linear model. Fourth, we shrank log2 fold changes to improve gene estimates with low mean counts that are vulnerable to dispersions. Lastly, we executed DE analysis to identify DE genes during the course of EC differentiation.

Quality control of RNA-seq data

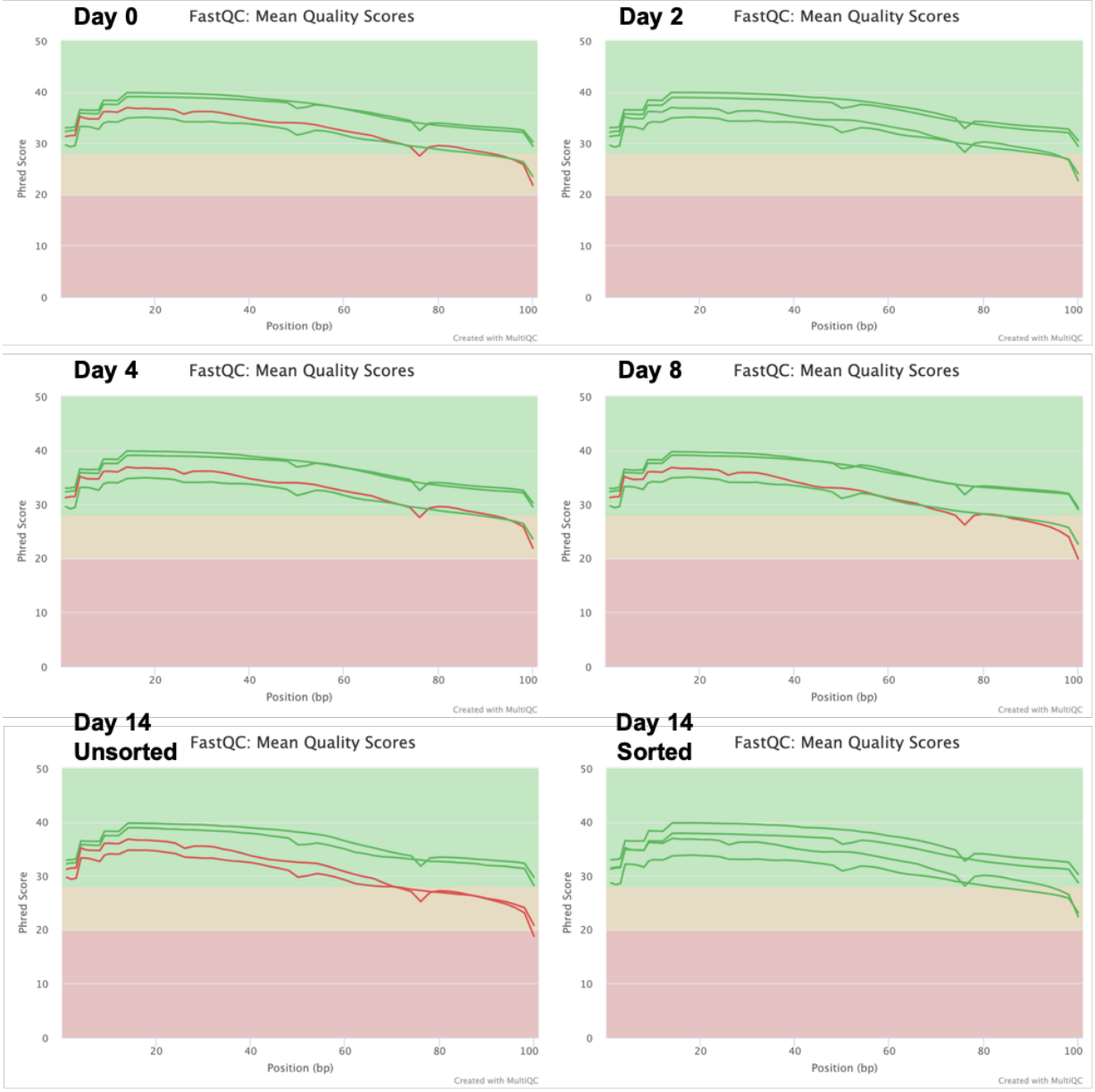


Figure 2. Quality control of RNA-Seq of ECs derived from hPSCs. Mean quality score assesses the quality of the sequencings. Phred score (y-axis) is plotted relative to nucleotide positions (x-axis). The background signifying the quality of the sequencing: very good quality (green), reasonable quality (yellow), and poor quality (red). The lines in green pass the quality test, whereas the lines in red fail the quality test. The six condition groups are assessed: Day0, Day2, Day4, Day8, Day14 before sorting, and Day14 after sorting.

To ensure that the RNA-seq procedure has been appropriately conducted with no contamination, we conducted the quality control test, using the FastQC and MultiQC software (Figure 2).[13] The quality was examined by the accumulation of a Phred score based on the corresponding

nucleotide position and resulted in either “pass” or “fail.”[35] We anticipated that most of the data would pass the quality control test under the assumption that RNA-seq was appropriately performed. We observed that Day0 had three passing and one failing samples, Day2 had four passing samples, Day4 had three passing and one failing samples, Day8 had three passing and one failing samples, Day14 before sorting had two passing and two failing samples, and Day14 after sorting had four passing samples. Most of the samples in the conditions passed the quality score except for Day14 before sorting. Two technical replicates of only one biological replicate from Day14 before sorting passed the quality control test. However, since those failing had similar patterns with those passing and it was integral to have replicates from two different biological samples in this particularly small sample nature of RNA-seq data, we proceeded to the next procedure with one each sample for all the conditions including Day14 before sorting for proper differential expression (DE) analysis.

Detection of potential sample outliers and the sources of variation

To detect potential sample outlier, we performed hierarchical clustering analysis (Figure 3).[36,37] We calculated Pearson correlation values for all pairwise combinations. The biological replicates were highly similar to each other, whereas the replicates that belonged to different groups clustered separately (Figure 3). Furthermore, regardless of the colors, the degree of correlation was larger than at least 0.9 due to no contamination nor outliers in the data. This phenomenon was evident as the proportion of DE genes was only a small fragment of all the genes of the human genome. Overall, the result of the hierarchical clustering analysis demonstrated that the data were clean from contamination or outliers.

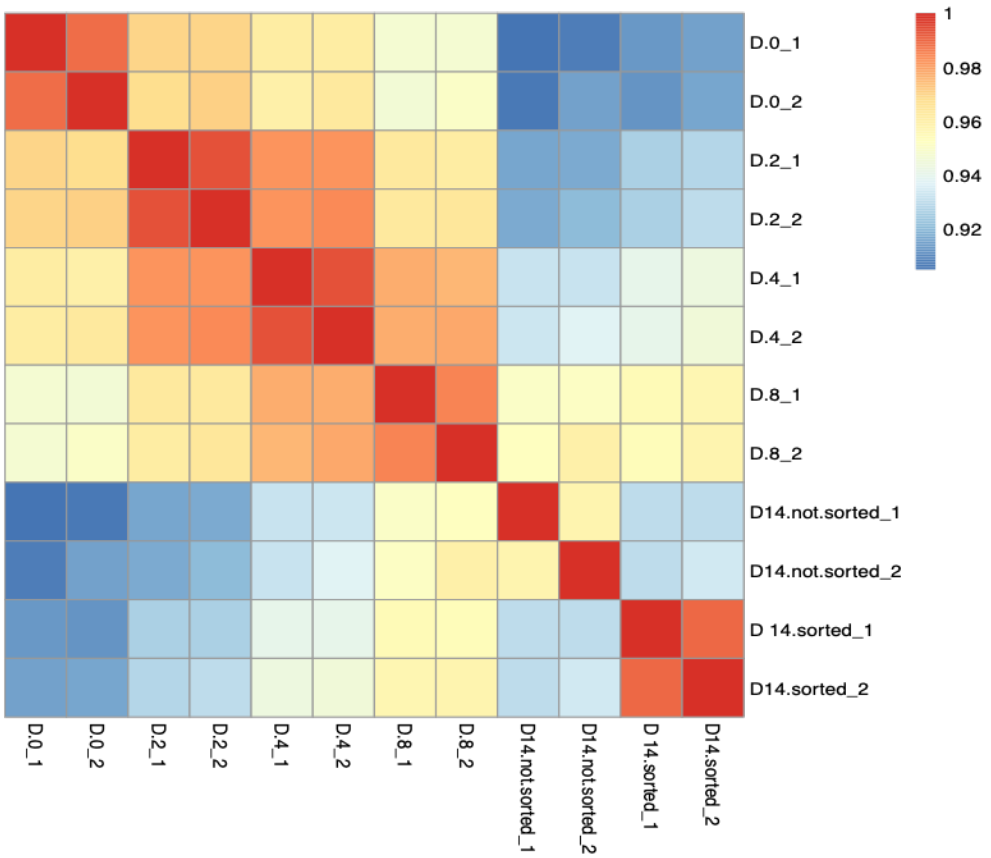


Figure 3. Hierarchical Clustering Analysis. Hierarchical heatmap shows the global gene expression similarity among Day0, Day2, Day4, Day8, and Day14 before sorting, and Day14 after sorting. The color from blue to red represents Pearson correlation for all pairwise combination of the samples. Red represents the perfect correlation, whereas blue represents a relative low correlation.

To explore the source of variation in the data, we conducted principal component analysis (PCA) (Figure 4).[24] Through PCA, we investigated whether "timepoint" signified the primary source of variation in the data. As expected, there was a moderate variation in gene expression caused by the condition, "timepoint" (Figure 4). Specifically, "timepoint" corresponding to PC1, represented 65% of variation, and PC2 showed 16% variation. This result indicates that the differences in timepoint could explain a considerable amount of variation in gene expression. By identifying "timepoint" as the major source variation, we could capture DE genes more accurately and efficiently across the timepoints.

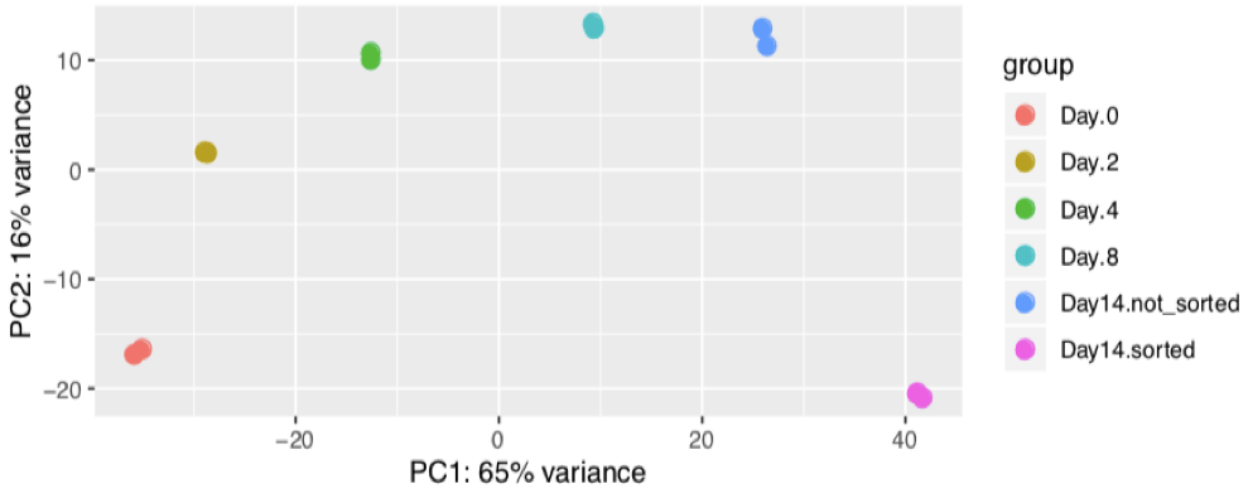


Figure 4. Identification of the Source of Variation. The Principal Component Analysis (PCA) plot depicts the source of variation present in the dataset. PC1 (x-axis) represents the greatest amount of variance in the data. PC2 (y-axis) shows the second largest variation, which is perpendicular to PC1 and is not described by PC1, in the data. PC value is calculated by the product of the influence and normalized read count for each gene and summing across all genes. Color distinguishes the type of the group.

Application of DESeq2 generalized linear model to fit the RNA-seq data

To determine if data fit the generalized linear model of the DESeq2 package, we designed two diagnostic plots: mean-variance and dispersion-mean plots (Figure 5).[11] We anticipated a robust relationship between mean and variance due to the small sample size. We detected a positive linear relationship between the mean and variance on the logarithmic scale across all genes (Figure 5A).[11] However, variance increased more rapidly than the mean did. This observation rejected the assumption of the Poisson distribution (blue line) that the mean was equal to the variance. Disregarding biological variability leads to more robust false-discovery rates due to the underestimation of sampling error.[11] Instead, the RNA-seq data were well represented by the negative binomial model due to the greater variance than the mean. Furthermore, we observed the low mean estimates with a higher degree of spread.

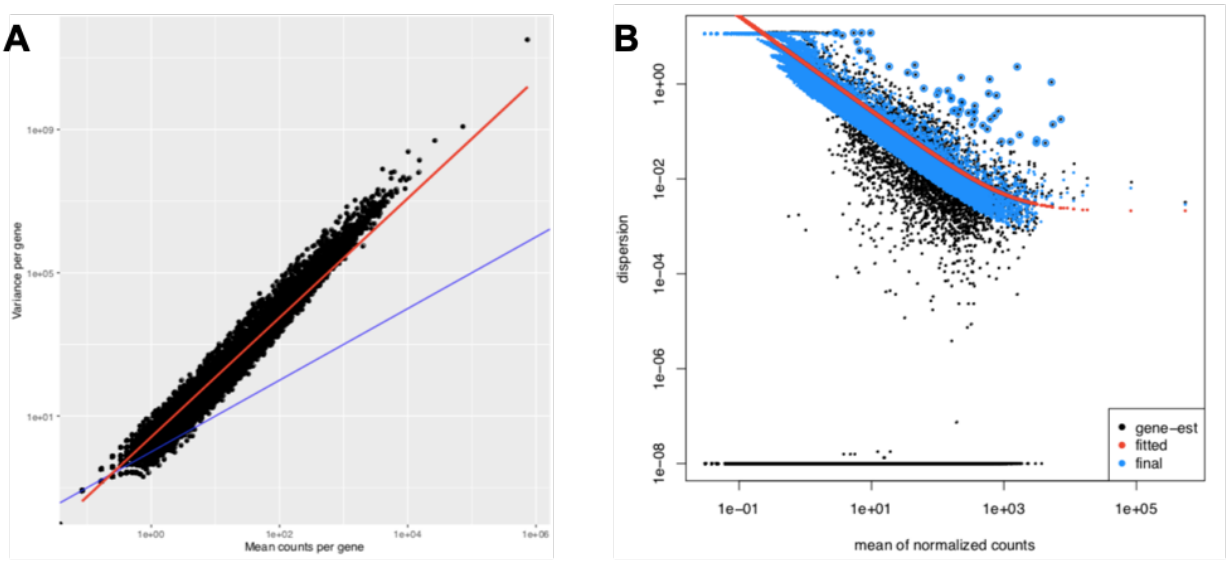


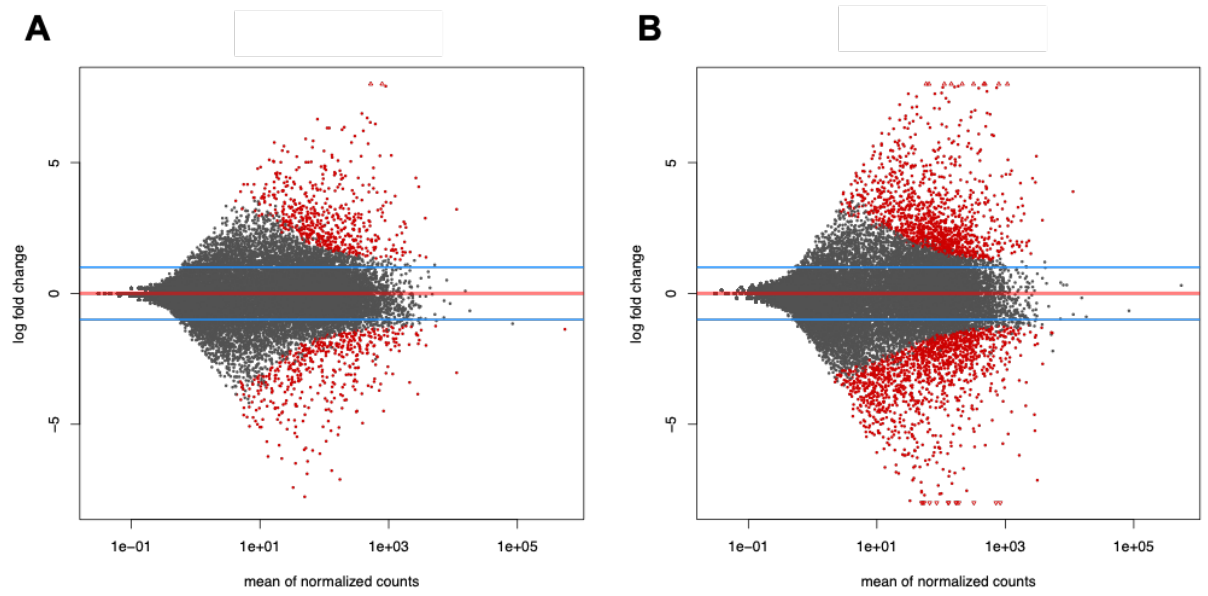
Figure 5. Application of DESeq2 Generalized Linear Model to Fit the Raw Counts. A. The variance-mean plot shows the relationship between variance per gene (y-axis) and the mean counts per gene (x-axis) on a logarithmic scale. The blue line represents Poisson distribution, a theoretically appropriate model if no biological variation present. The red line signifies the linear regression line that fits a linear equation to the observed data using the ordinary least-squared method. **B.** The dispersion-mean plot shows dispersion estimates (y-axis) over the average mean counts (x-axis). First, the maximum likelihood estimates of dispersion for each gene is plotted as a black dot. A fitted estimates of dispersion curve (red) is fit to the MLEs to capture the overall trend of dispersion-mean dependence. Genes shrunken (blue dots) towards the curve (red line) shows more accurate estimate of dispersions. The black dots circled in blue are dispersion outliers and not shrunk toward the red line.

To investigate the low mean estimates with more substantial variability, we examined the relationship between the dispersion and the mean (Figure 5B).[12,29] This plot showed the overall distribution of the dispersion estimates associated with the corresponding mean. This plot examined whether the dispersion estimates were feasible with the biological spread of each gene. According to the mathematical equation, we hypothesized that the distribution of the dispersion estimates would increase as the mean counts decreased, and the maximum likelihood estimates (MLEs) of dispersion would shrink toward the fitted dispersion estimates. We observed that a dispersion estimate of each gene, shown in a black dot, had an overall distribution to increase as the mean counts decreased. Moreover, those estimates were shrunken (blue dots) toward to the fitted dispersion estimate line (red line). This shrinkage method did not deliberately account for outlier genes because they hardly followed the DESeq2 generalized linear model due to spread

from other than biological and technical aspects.[29] The two diagnostic plots conveyed that our data fit the DESeq2 generalized linear model since variance estimates increased more significantly than the mean, dispersion estimates scattered and shrank toward the fitted maximum likelihood dispersion estimates curve.

Summarization of DE analysis

To examine differential expression profiles of hiPSC-ECs, we compared all timepoints to Day0 by generating MA plots showing the relationship between fold-change and mean counts (Figure 6).[29] Through MA plots, we first confirmed that the data were normalized appropriately because the fold-changes of genes were clustered around zero. We second validated that the data were shrunk properly since genes with low mean counts were clustered together. To identify any patterns of DE genes across the timepoints, we summarized the DE analysis of all pairwise comparisons for both up- and down-regulated DE genes (Figure 6C). We anticipated that the number of DE genes would increase over time because ECs are enriched from hiPSCs over time; it is what we observed for both up- and down-regulated DE genes over time (Figure 6C). For the up-regulated genes, there were 137, 290, 550, 1117, and 1252 DE genes on Day2, Day4, Day8, Day14 before sorting, and Day14 after sorting, respectively. For the down-regulated genes, there were 340, 536, 478, 748, and 1398 DE genes on Day2, Day4, Day8, Day14 before sorting, and Day14 after sorting, respectively. This positive trend in both up- and down-regulated DE genes indicates that hiPSCs were properly turned into ECs and enriched under the EC differentiation protocol.



C

	Day 2	Day 4	Day8	Day14 before sorting	Day14 after sorting
LFC > 1	137 (0.4%)	290 (0.9%)	550 (1.8%)	1117 (3.6%)	1252 (4.1%)
LFC < 1	340 (1.1%)	536 (1.7%)	478 (1.5%)	748 (2.4%)	1398 (4.5%)

Figure 6. Distribution of DE Genes through MA Plots. The MA plot shows the relationship between the log fold change and average expression strength. Genes (Wald Test padj < 0.05 and |log2 fold change| > 1) are marked in red, whereas genes that are insignificant are plotted in gray. The upper blue line is the log2 fold change threshold of 1, and the lower blue line is the log2 fold change threshold of -1. **A.** The DE result of Wald Test compares Day8 to Day0. **B.** The DE result of Wald Test compares Day14 after sorting to Day0. **C.** The column represents the DE results of Wald Test comparing the timepoint indicated to Day0. The first row represents up-regulated DE genes, and the second row represents down-regulated DE genes.

Detection of differential gene expression levels during EC differentiation

To compare relative gene expression levels across the timepoints, we measured normalized gene expression for all the timepoints (Figure 7).[38] We subset DE genes for the pairwise comparison: Day0 vs. Day14 after sorting. We extended the heatmap to include other timepoints—Day2, Day4, Day8, and Day14 before sorting—to facilitate a better gene expression level comparison across all the timepoints and scaled by row. Z-scores represented the gene expression levels in colors, where green and red signified lower and higher Z-scores, respectively. As hiPSCs and ECs are distinctively different cell types, we hypothesized that the gene expression levels for the significant DE genes clustered by only the same sample group and clustered apart by different sample groups. As the hypothesis, those up-regulated on Day0 were lowly expressed at other timepoints,

particularly on Day14 after sorting, and those highly expressed on Day14 after sorting were down-regulated at other timepoints. This observation demonstrates that different genes were turned on at each timepoint during the course of EC differentiation and suggests that hiPSCs and hiPSC-ECs are distinctively different cell types.

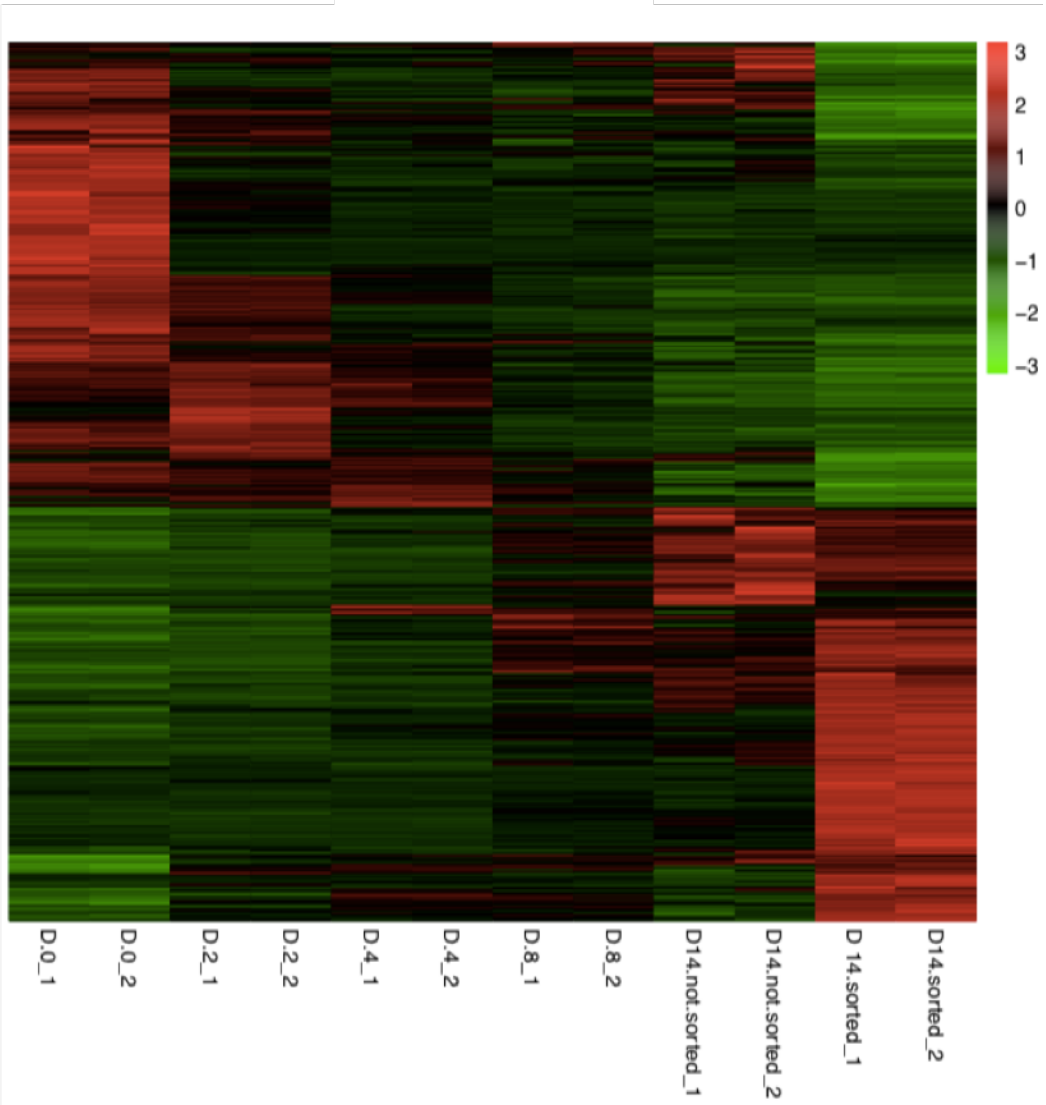


Figure 7. Detection of Gene Expression Levels across Timepoints. Significant DE genes in Day0 vs. Day14 after sorting (Wald Test padj < 0.05 and |log2 fold change| > 1) are filtered. The gene expressions of all samples including all other conditions after sorting for the corresponding DE genes are shown. They are scaled per row so that the colors represent Z-scores.

Distribution of DE genes during EC differentiation and enrichment

To enhance the understanding of the distribution of DE genes during EC differentiation and enrichment, we generated two Venn Diagrams for Day8, Day14 before sorting, and Day14 after sorting (Figure 8).[39] For the up-regulated genes, 337 genes were in the intersection of the three sets, 94 genes were in the intersection of Day8 and Day14 before sorting, 70 genes were in the intersection of Day8 and Day14 after sorting, and 216 genes were common between Day14 before sorting and Day14 after sorting. 49, 470, and 629 genes were solely up-regulated only on Day8, Day14 before sorting, and Day14 after sorting, respectively. For the down-regulated genes, 296 genes were in the intersection of the three sets. 57 genes were common between Day8 and Day14 before sorting, 85 genes were in the intersection of Day8 and Day14 after sorting, and 159 genes were common between Day14 before sorting and Day14 after sorting. 40, 236, and 858 genes were exclusive on Day8, Day14 before sorting, and Day14 after sorting, respectively.

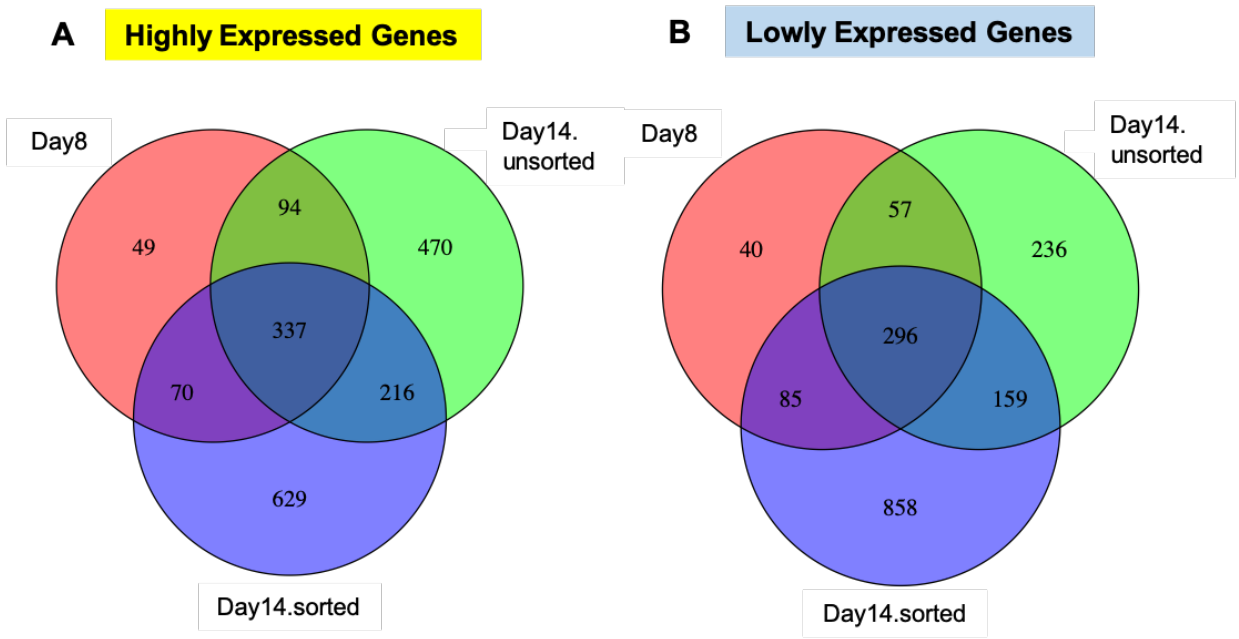


Figure 8. Distribution of Up- and Down-regulated Genes in EC Differentiation. A. Overlap of highly-expressed genes (log2 fold change > 1) among Day8, Day14 before sorting, and Day14 after sorting compared to Day0. Red, green, and purple represent Day8, Day14 before sorting, and Day14 after sorting, respectively. **B.** Overlap of lowly-expressed genes (log2 fold change < -1) among Day8, Day14 before sorting, and Day14 after sorting compared to Day0. Red, green, and purple represent Day8, Day14 before sorting, and Day14 after sorting, respectively.

Identification of enriched gene ontology (GO) terms based on the DE genes

Within the GO terms associated with EC functionality, we compared the statistical significance among Day8, Day14 before sorting, and Day14 after sorting (Figure 9). This analysis highlighted that most of the GO terms (EC migration, EC differentiation, EC apoptotic process, and angiogenesis positive) were more statistically significant at the EC enrichment phase (Day0 vs. Day14) than at the EC differentiation phase (Day0 vs. Day8). This observation validates that enriched ECs on Day14 are more likely to display the genuine EC characteristics than ECs on Day8.

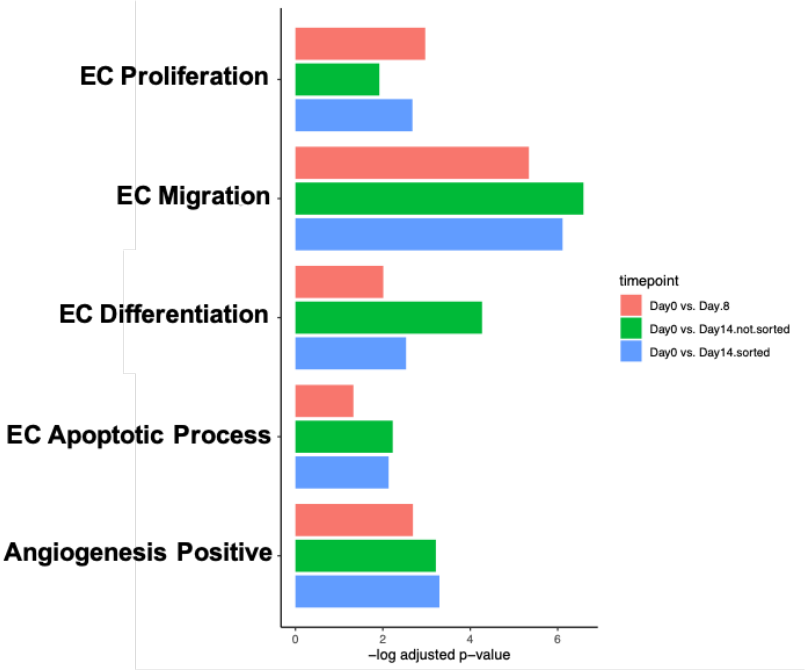


Figure 9. The Comparison of Significant Enriched Gene Ontology (GO) Terms among Day8, Day14 before sorting, and Day14 after sorting. The enriched GO terms (y-axis) across all the three timepoints were indicated with the corresponding padj scaled by –log (x-axis). Red, green, and blue represent Day8, Day14 before sorting, and Day14 after sorting, respectively.

To determine the functions of the up-regulated DE genes for enriched ECs, we performed GO enrichment over-expression analysis on the DE genes of Day0 vs. Day14 after sorting (Figure 10A).[31] We hypothesized that $CDH5^+$ cells would have a significant number of GO terms that are highly involved with EC biological processes because $CDH5^+$ cells should show proangiogenic potentials according to the previous study.[8] We observed that GO terms for biological processes

were significantly enriched in CDH5^{+} cells in EC migration, positive regulation of angiogenesis, EC proliferation, blood vessel EC migration, EC differentiation, positive regulation of EC proliferation, EC apoptotic process, branching involved in blood vessel morphogenesis, and EC morphogenesis. This result supports the previous finding that CDH5^{+} cells significantly contribute to endothelial and proangiogenic properties of hiPSC-ECs.

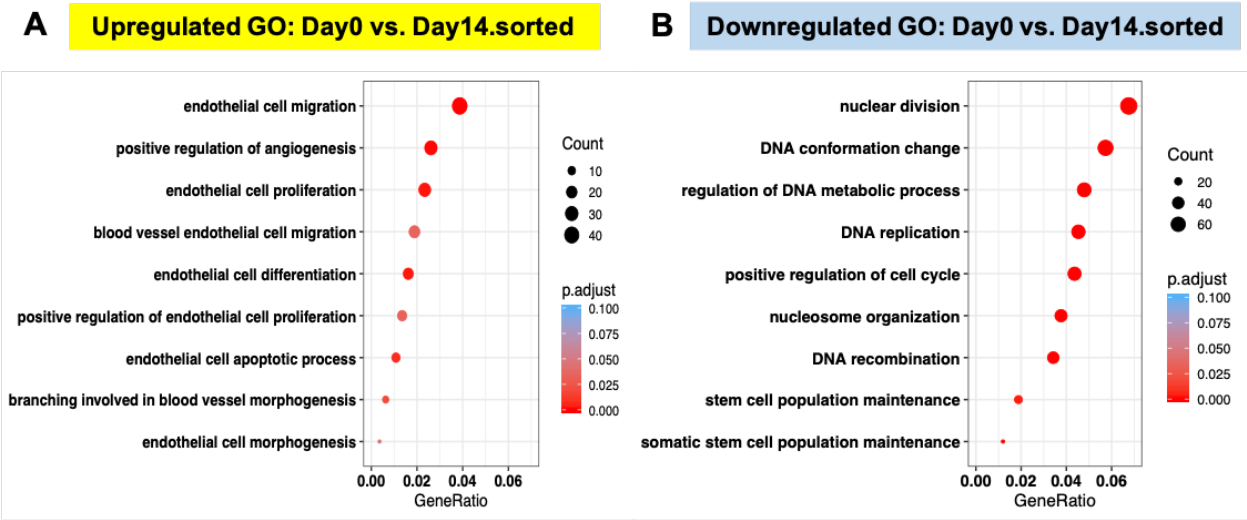


Figure 10. Identification of Enriched Gene Ontology (GO) Terms for Day0 vs. Day14 after sorting. A. Nine statistically significant GO terms are indicated. The GO terms (y-axis) are indicated with the corresponding GeneRatio (x-axis). The size of circle represents the number of count, and the color of the circle represents the adjusted p-values. **B.** Nine statistically significant GO terms are indicated. The GO terms (y-axis) are indicated with the corresponding GeneRatio (x-axis). The size of circle represents the number of count, and the color of the circle represents the adjusted p-value.

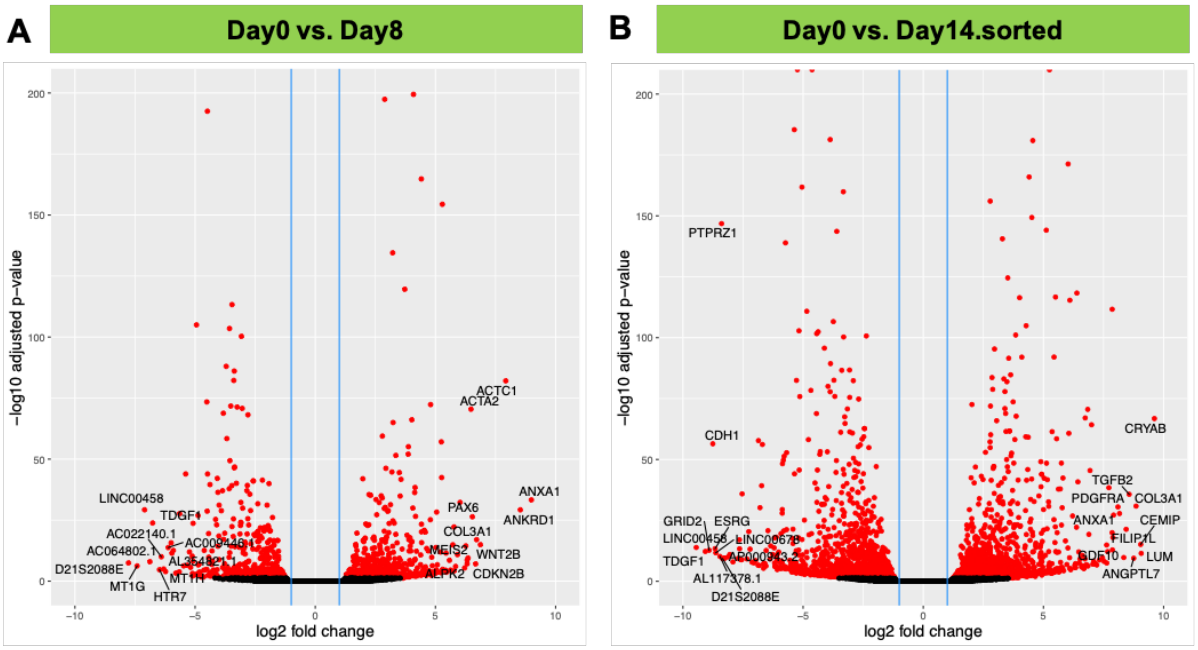
To identify the roles of the down-regulated DE genes for enriched ECs, we performed GO enrichment under-expression analysis on the list of DE genes of Day0 vs. Day14 after sorting (Figure 10B).[31] As hiPSCs and ECs are distinctively different cell types, we hypothesized that CDH5^{+} cells would lose the pluripotent characteristic of hiPSCs. We observed that GO terms for biological processes in CDH5^{+} cells were significantly down-regulated in nuclear division, conformation change, regulation of DNA metabolic process, DNA replication, positive regulation of cell cycle, nucleosome organization, DNA recombination, stem cell population maintenance, and somatic stem cell population maintenance. The result conveys that CDH5^{+} cells underwent the degradation of pluripotency because rapid DNA replication is an intrinsic characteristic of stem

cells that contributes to pluripotency maintenance.[40] This loss of pluripotency supports the genuine differentiation from hiPSCs to ECs and leads ECs to maintain the EC lineage.

Identification of biological and statistical significance of DE genes

To determine the biological and statistical significance of gene expression profiles, we compared statistical significance (padj) to biological significance (log2FoldChange) by generating volcano plots (Figure 11).[41] Fold-change, which refers to the changes in gene expression levels between the treatment and base-line treatment, represents biological significance; padj signifies statistical significance. As the population of ECs was enriched over time, we hypothesized that a greater number of DE genes would have larger differences in gene expression levels over time. We observed that $CDH5^+$ cells displayed a substantial fold-change for both up- and down-regulated genes with large padj scaled by negative logarithm (more volcano-shaped plot), compared to Day8. This observation indicates a greater number of genes altered more dramatically from hiPSCs to hiPSC-ECs.

In large samples, p-values rapidly converge to zero, and therefore depending exclusively on p-values causes results with no practical significance.[42] Within genes identified as statistically DE, we instead referred to log2 fold change to draw biological interpretation. Hence, we labeled 10 up- and down-regulated DE genes in the order of the largest and smallest log2 fold change, respectively. In the pairwise comparison Day0 vs. Day8, the representative up-regulated genes are *ANXA1, ANKRD1, ACTC1, WNT2B, COL3A1, CDKN2B, PAX6, ACTA2, MEIS2,* and *ALPK2*, and the representative down-regulated genes are *MT1G, TDGF1, HTR7,* and *MT1H*. In the pairwise comparison Day0 vs. Day14 after sorting, the representative up-regulated genes are *CRYAB, LUM, CEMIP, COL3A1, ANGPTL7, TGFB2, FILIP1L, GDF10, ANXA1,* and *PDGFRA*, and the representative down-regulated genes are *TDGF1, GRID2, CDH1, ESRG,* and *PTPRZ1*.



C

Down-regulated	Up-regulated
D21S2088E	ANXA1
MT1G	ANKRD1
LINC00458	ACTC1
AC064802.1	WNT2B
TDGF1	COL3A1
HTR7	CDKN2B
AC022140.1	PAX6
AL354821.1	ACTA2
MT1H	MEIS2
AC009446.1	ALPK2

D

Down-regulated	Up-regulated
LINC00458	CRYAB
TDGF1	LUM
GRID2	CEMIP
CDH1	COL3A1
ESRG	ANGPTL7
LINC00678	TGFB2
AL117378.1	FILIP1L
PTPRZ1	GDF10
D21S2088E	ANXA1
AP000943.2	PDGFRA

Figure 11. Identification of Biological and Statistical Significance of DE Genes. Volcano plot shows biological and statistical significance of all the tested genes. The color distinguishes significantly differentially expressed genes (red) from insignificant (gray) using Wald Test adjusted p-value < 0.05, log2 fold change > 1, and log2 fold change < -1. For the up-regulated genes, 10 genes are labeled in the order of the highest log2 fold change. For the down-regulated genes, 10 genes are labeled in the order of the smallest log2 fold change. **A.** The DE result of Wald Test compares Day8 to Day0. **B.** The DE result of Wald Test compares Day14 after sorting to Day0. **C.** 10 representative up-regulated and down-regulated DE genes on Day8 in the order of the highest and lowest log2 fold change, respectively. **D.** 10 representative up-regulated and down-regulated DE genes on Day14 after sorting in the order of the highest and lowest log2 fold change, respectively.

For instance, *ANXA1*, a regulator of the innate immune response, regulates the inflammatory process. *PDGFRA*, a cell-surface receptor for *PDGFA, PDGFB,* and *PDGFC*, significantly regulates embryonic development, cell proliferation, and survival. Among the 1252 up-regulated DE genes, the 10 genes with the largest fold-change were associated with cellular functions but were hardly directly involved in endothelial and proangiogenic properties. Therefore,

to identify significantly meaningful genes in the list of DE genes, we further implemented gene ontology (GO) analysis.

Process to identify notable genes

To identify genuinely significant genes that contribute to endothelial and proangiogenic effects, we performed relative gene expression level analysis based on the selected GO terms. As the objective of the EC differentiation is to generate ECs for cell therapy to improve ischemic conditions, we selected GO terms involved in endothelial biological processes: GO EC development, GO EC differentiation positive, GO EC proliferation positive, GO EC migration positive, GO angiogenesis, and GO blood vessel remodeling.[47] For each selected GO term, we compared relative gene expression levels across the timepoints by creating expression heatmaps (Figure 12). To determine those highly involved with endothelial biological processes, we mainly focused on the up-regulation in the $CDH5^{+}$ cell population (Day14 after sorting) compared to the initial hiPSCs (Day0). After we confirmed notably up-regulated genes, we revisited the DE analysis result on Day0 vs. Day14 after sorting to examine whether those genes were statistically significant as well.

GO EC development refers to the progression of ECs over time.[47] We anticipated that genes contributing to the endothelial formation would start to be expressed at the early phase of EC differentiation (Day8), and endothelial maturation would be highly expressed, particularly at the EC enrichment phase (Day14 after sorting). In GO EC development (Figure 12A; 42 genes), we observed 7 up-regulated genes—*RAPB1, RDX, HEG1, GSTM3, MET, ID1,* and *RAP2B*—on Day14 after sorting from the heatmap. However, from the DE result, only 6 genes were identified as DE: *RDX, HEG1, GSTM3, RAP2B, MET,* and *PDE4D*; among them, *HEG1* and ID1 are notable. *HEG1*, a receptor of the *CCM* signaling pathway, regulates heart and vessel formation by stabilizing EC junctions. *ID1*, a transcriptional regulator inhibiting DNA binding, regulates

cellular processes, including cellular growth, differentiation, and angiogenesis. The result indicates that the synthesis of the genes highly expressed facilitates to develop ECs over time.

GO EC differentiation positive refers to the process that elevates the rate of EC differentiation.[47] We hypothesized that genes would increase the rate of EC differentiation at the EC differentiation phase (Day8). In GO EC differentiation positive (Figure 12B; 14 genes), we observed 10 up-regulated genes—*CTNNB1, TMEM100, BMP4, ETV2, CDH5, S1PR2, PROC, ACVRL1, ATOH8,* and *BTG1*—on Day8 and Day14 after sorting. However, from the DE result of Day8, only *BMP4* was identified as DE. *BMP4* plays a role in mesoderm induction and limb formation. *BMP4* underwent a sudden upregulation on Day8 and Day14 before sorting. The result displays that most genes that belong to GO EC differentiation positive were highly expressed during EC differentiation and enrichment. Although only one gene was actually DE from the list, the up-regulated genes in this GO term promotes EC differentiation.

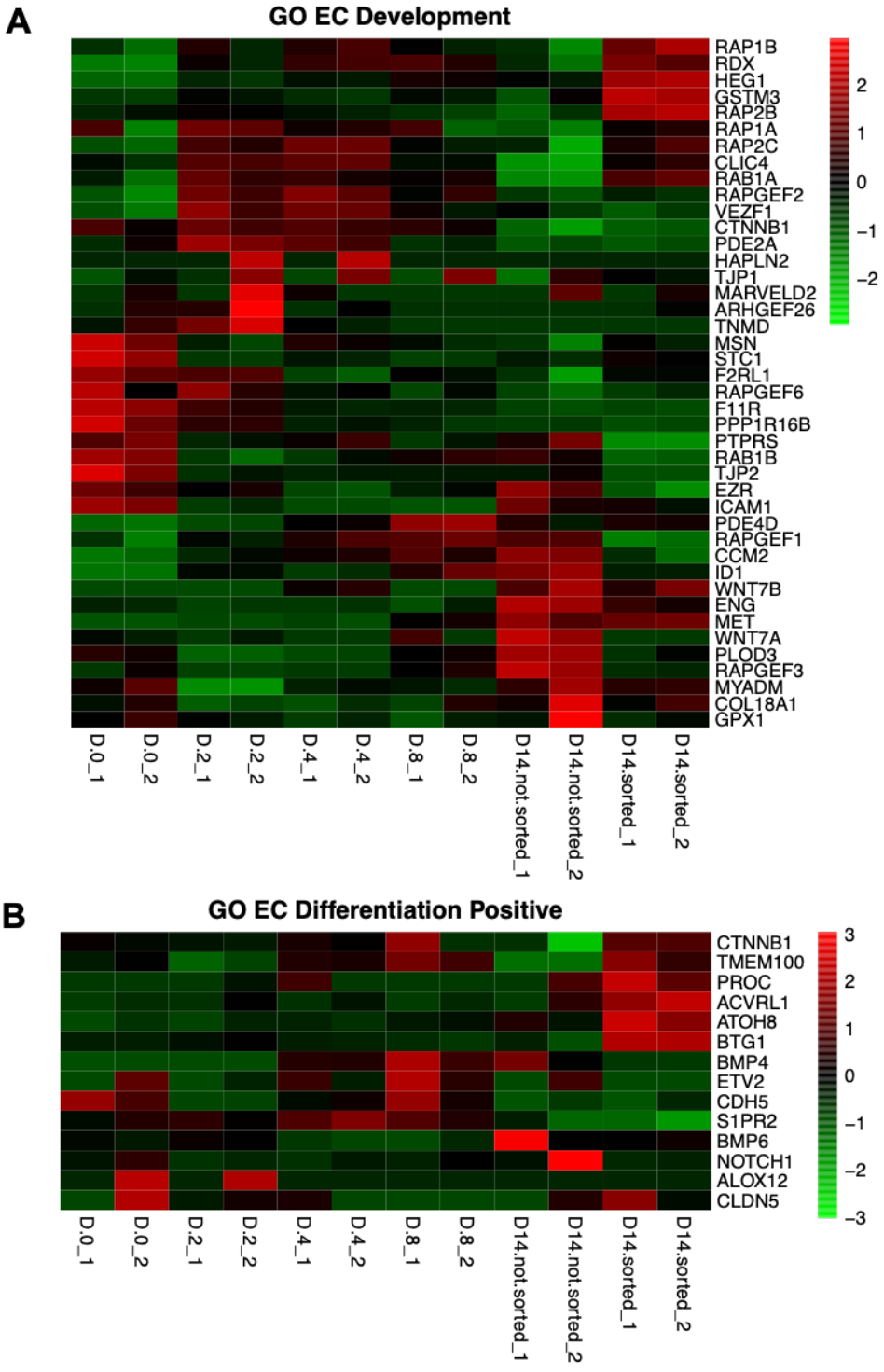


Figure 12. Detection of Gene Expression Level Changes within Specific GO Terms. GO terms of our interest were selected, and gene expression trends were analyzed across the timepoints. The gene expression scaled by Z-scores (color) of genes that belong to the GO terms (y-axis) are indicated to the corresponding timepoints (x-axis). **A.** GO EC Development. **B.** GO EC Differentiation Positive.

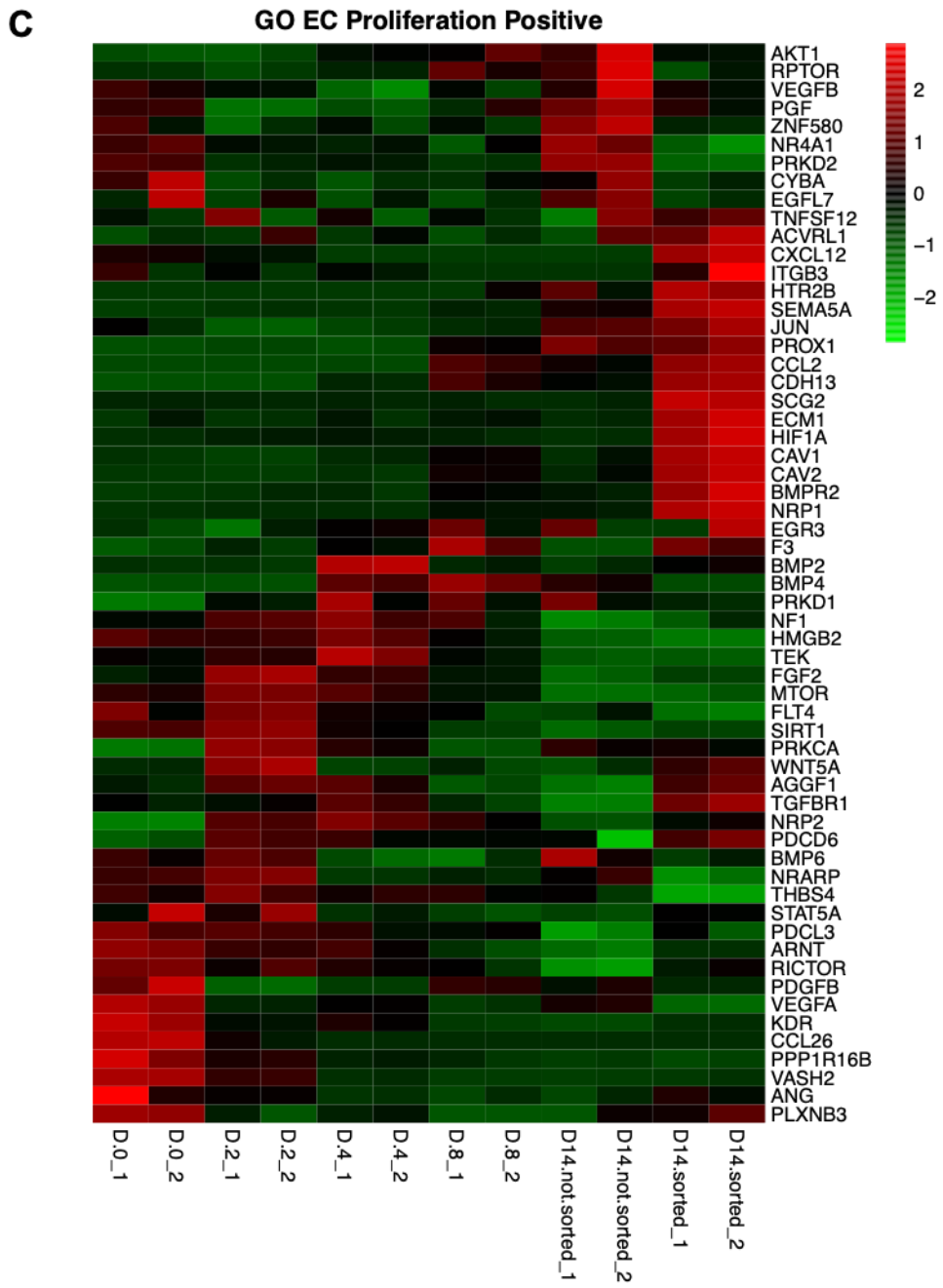


Figure 12 (cont.). Detection of Gene Expression Level Changes within Specific GO Terms. C. GO EC Proliferation Positive.

GO EC proliferation positive refers to the process that elevates the rate of EC proliferation.[47] We expected that the proliferation of ECs would significantly increase at the EC enrichment phase

(Day14 after sorting). In GO EC proliferation positive (Figure 12B; 59 genes), we observed 17 up-regulated genes—*ACVRL1, CXCL12, ITGB3, HTR2B, SEMA5A, JUN, PROX1, CCL2, CDH13, SCG2, ECM1, HIF1A, CAV1, CAV2, BMPR2, NRP1,* and *EGR3*— on Day14 after sorting from the heatmap. However, among them, only 13 genes were identified as DE: *CXCL12, SEMA5A, JUN, PROX1, CCl2, CDH13, SCG2, ECM1, HIF1A, CAV1, CAV2, BMPR2,* and *NRP1*. *CXCL12* proliferates bone marrow-derived B-cell progenitors during embryonic development. Along with GO *SEMA5A* facilitates EC proliferation, migration, and angiogenesis. *PROX1*, a transcription factor, plays a vital role in embryonic development and functions. *ECM1* promotes angiogenesis by triggering EC proliferation. *HIF1A* contributes to embryonic vascularization, tumor angiogenesis, and pathophysiology of ischemic disease. *NRP2*, which binds to the VEGF165 isoform of *VEGFA* and *VEGFB*, regulates VEGF-induced angiogenesis. *CAV1* negatively regulates TGFB1-mediated activation of SMAD2/3 by mediating the internalization of *TGFBR1*. *CAV2* drives caveolae formation and modulates mitosis in ECs. The result indicates that the synthesis of the up-regulated genes proliferates ECs.

GO EC migration positive refers to a process that increases the rate of the orderly movement of ECs into the extracellular matrix to form an endothelium.[47] Cell migration is a fundamental functionality of physiological and pathological processes. In particular, EC migration restores vessel integrity in a damaged vessel and promotes angiogenesis.[43] We hypothesized that EC migration would be highly activated at the EC enrichment phase (Day14 after sorting) because ECs should demonstrate the most genuine EC characteristics at this timepoint. In GO EC migration positive (Figure 12F; 26 genes), we observed 14 up-regulated genes—*MET, PROX1, ITGB1BP1, AGT, ATOH8, FOXP1, SPARC, VEGFC, SRPX2, SEMA5A, BMPR2, NRP1, ETS,* and *ITGB3*—on Day14 after sorting from the heatmap. However, among them, 10 genes were identified as DE: *MET, PROX1, ITGB1BP1, FOXP1, SPARC, SRPX2, SEMA5A, BMPR2,* and *NRP1*.

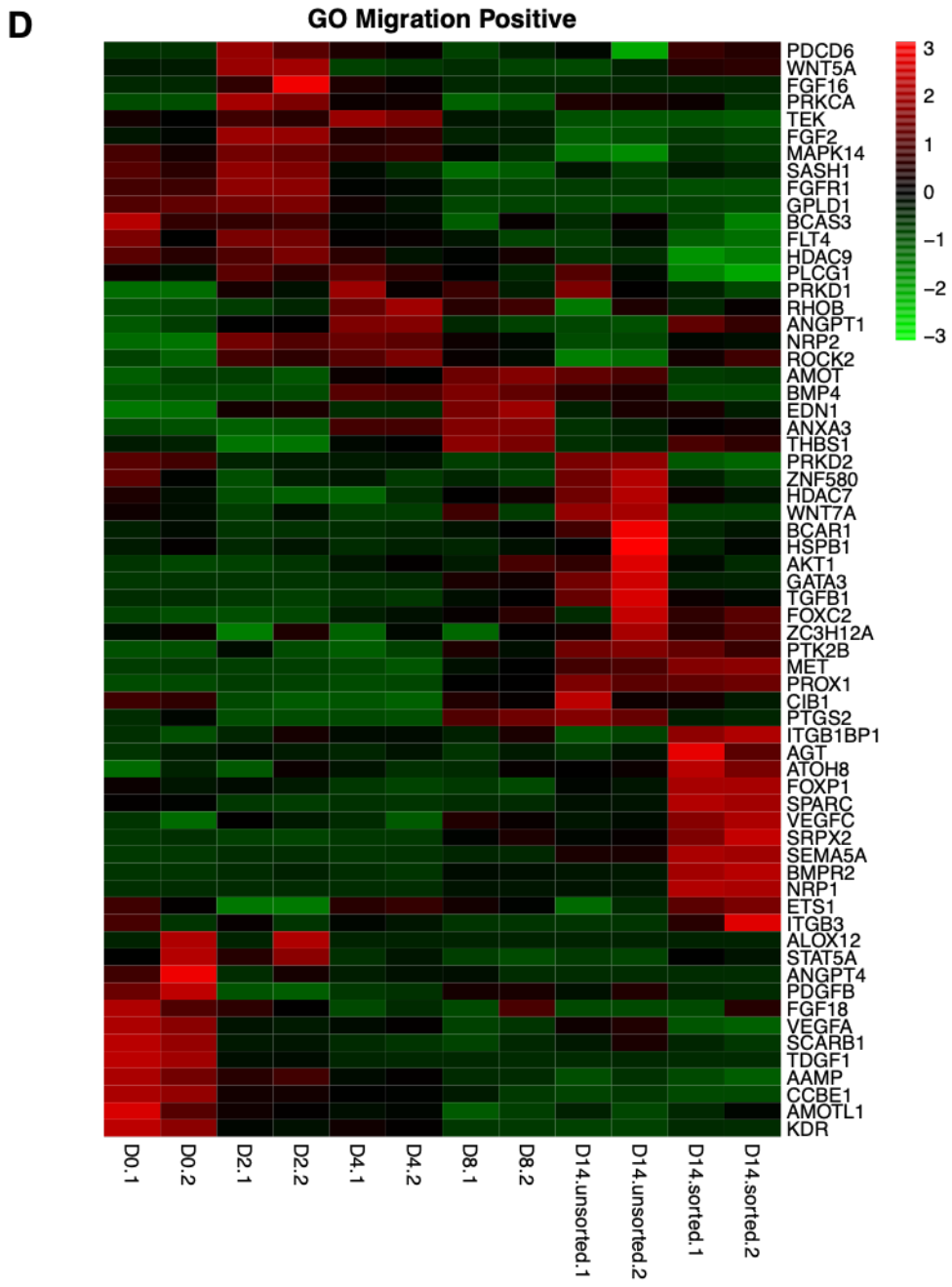


Figure 12 (cont.). Detection of Gene Expression Level Changes within Specific GO Terms. D. GO Migration Positive.

MET promotes angiogenesis and would healing. *ITGB1BP1* plays a role in cell proliferation, differentiation, spreading, adhesion, and migration in the context of angiogenesis. *SRPX2* promotes angiogenesis by inducing EC migration. *SEMA5A* stimulates angiogenesis by

increasing EC proliferation and migration and inhibiting apoptosis. *NRP1,* a regulator of VEGF-induced angiogenesis, is highly involved in the development of the cardiovascular system in angiogenesis. As a consequence, the reconciliation of the up-regulation of the genes promote the movement of ECs to form an endothelium and therefore contribute to angiogenesis.

GO angiogenesis refers to blood vessel formation when new vessels emerge from the proliferation of pre-existing blood vessels.[47] As hiPSC-ECs demonstrated direct vessel-forming effects in the previous study, we anticipated that genes involved with proangiogenic potentials to be highly expressed, particularly on Day14 after sorting. Due to the large size of GO angiogenesis (270 genes), we only displayed DE genes. In GO angiogenesis (Figure 12E, 270 genes), we observed genes—*ANGPT1, TGFBR2, ECM1, NRP1, VAV3, NOV, TGFBI, SCG2, CYP1B1, HIF1A, ANGPT2, PDGFRA, SEMA5A, TGFB2, ARHGAP24, FMNL3, RSPO3, HEY1, SRPX2, CAV1, MMP14, COL8A1, COL4A1, ITGAV, CCL2, CDH13, ANXA2,* and *PARVA*—were up-regulated DE on Day14 after sorting. *ECM1* stimulates EC proliferation and angiogenesis. *NRP1*, with the expression of *KDR*, regulates VEGF-induced angiogenesis. *VAV3* promotes EC migration and angiogenesis. *NOV*, a regulator of hematopoietic stem and progenitor cell function, plays an essential role in EC cell adhesion, cell migration, and cell survival. *ANGPT2*, with the expression of *VEGF*, stimulates EC migration and proliferation. *SEMA5A* facilitates EC proliferation, migration, and angiogenesis. *SRPX2* promotes the formation of vascular networks and angiogenesis by stimulating EC migration. *CAV1* negatively regulates *TGFB1*, which promotes mesenchymal and smooth muscle functions. *COL8A1* is the main component of corneal ECs and endothelial blood vessels. As a result, the reconciliation of the up-regulation of the genes promotes the blood vessel formation by the previously stimulated EC proliferation and migration. The activation of angiogenesis positively influences blood vessel remodeling.

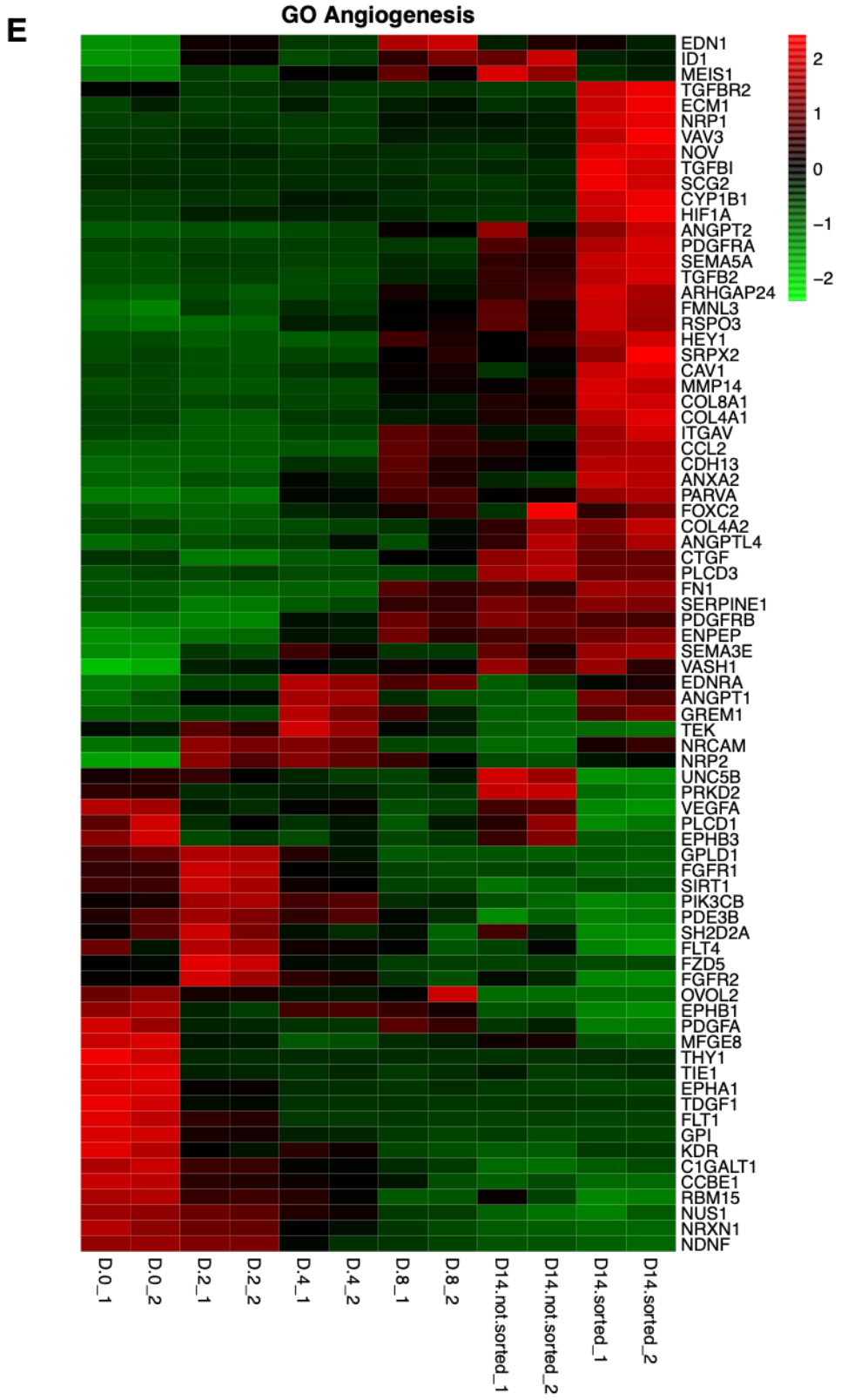


Figure 12 (cont.). Detection of Gene Expression Level Changes within Specific GO Terms. E. GO Angiogenesis.

GO blood vessel remodeling refers to the reorganization of existing blood vessels.[47] We hypothesized that genes promoting neovascularization in the context of blood vessel remodeling would be up-regulated, mainly at the EC enrichment phase (Day14 after sorting). In GO blood vessel remodeling (Figure 12F, 28 genes), we observed 9 up-regulated genes—*MEF2C, RSPO3, TGFB2, ACVRL1, MDM2, BMPR2, SEMA3A, AGT,* and *ATP7A*—on Day14 after sorting from the heatmap. However, among them, only 8 genes were identified as DE: *ATP7A, SEMA3C, BMPR2, MDM2, TGFB2, RSPO3, MEF2C,* and *FOXC2*. Among them, *SEMA3C, BMPR2,* and *RSPO3* are are directly associated with endothelial and proangiogenic functionalities. *SEMA3C* promotes

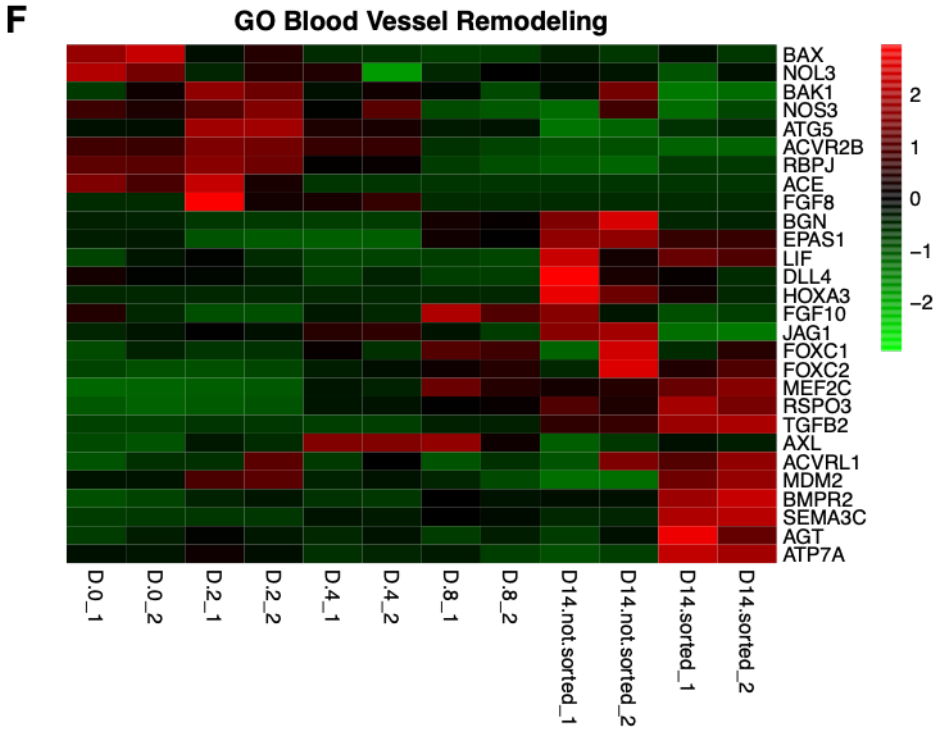


Figure 12 (cont.). Detection of Gene Expression Level Changes within Specific GO Terms. F. GO Blood Vessel Remodeling.

cardiovascular development during embryogenesis. *BMPR2* contributes to blood vessel remodeling, EC apoptotic process, and EC proliferation. *RSPO3* regulates angiogenesis by acting as a ligand for LGR4-6 receptors. Furthermore, we observed 7 up-regulated genes—*BGN, EPAS1, DLL4, HOXA3, JAG1, FOXC1,* and *FOXC2*—on Day14 before sorting. However, among them, only 4 genes were identified as DE: *BGN, EPAS1, JAG1,* and *FOXC2. EPAS1*, a

transcription factor, regulates the formation of the endothelium that gives rise to the blood brain barrier.

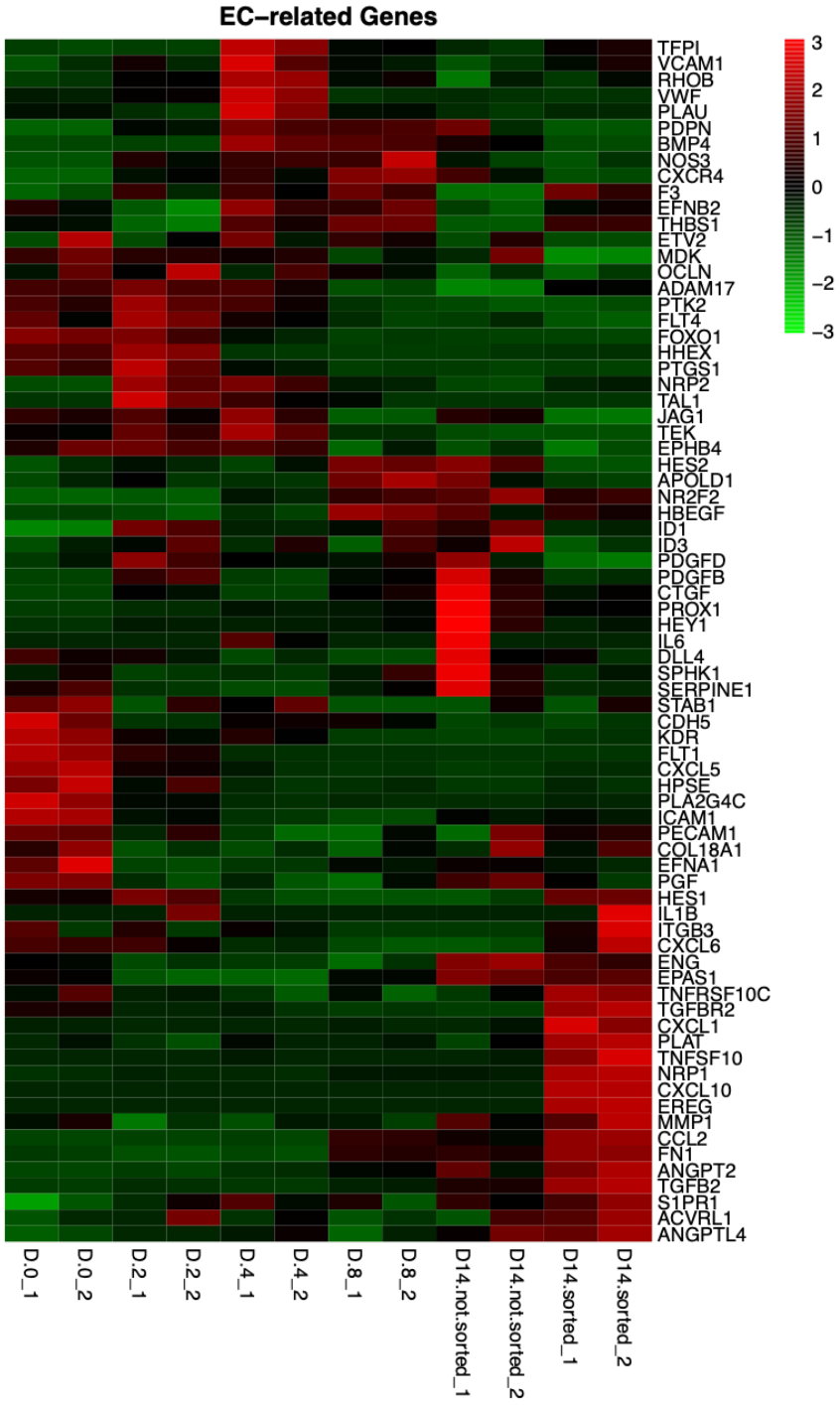


Figure 13. Detection of Gene Expression Level Changes within EC-related Genes. The collection of EC-related genes were found from Lee et al. The expression heatmap compares the gene expressions of the genes belonging to the list. Genes (y-axis) are plotted the normalized gene expressions scaled by Z-scores (color) to the corresponding timepoints (x-axis).

To explore genes that contribute to endothelial characteristics more in-depth, we referred to Lee *et al.* to acquire a list of EC-related genes.[44] We expected those genes involved in EC functions would be highly expressed at the EC enrichment phase (Day14 after sorting). In this collection of EC-related genes (Figure 13; 77 genes), we observed 17 up-regulated genes—*IL1B, ITGB3, CXCL6, TNFRSF10C, CXCL1, PLAT, TNFSF10, NRP1, CXCL10, EREG, MMP1, CCL2, FN1, ANGPT2, TGFB2, S1RP1,* and *ACVRL1*—on Day14 after sorting. However, among them, 8 genes were identified as DE: *CXCL1, PLAT, TNFSF10, NRP1, CCl2, FN1, ANGPT2,* and *TGFB2.* In addition, there were 6 up-regulated DE genes on Day14 before sorting: *CTGF, PROX1, HEY1, DLL4, SPHK1,* and *SERPINE1*. Furthermore, there was only one up-regulated DE gene on Day8: *F3*.

Determination of meaningful genes

According to the GO analysis, we confirmed whether the notable genes were truly DE in hPSCs in the course of EC differentiation. We identified 28 genes that are notably contribute to endothelial and proangiogenic characteristics of hPSC-ECs (Table 1): *TGFB2, COL8A1, SEMA3C, ANGPT2, NRP1, NOV, CCL2, CXCL1, CAV2, RSPO3, TNFSF10, PROX1, FN1, SRPX2, SEMA5A, CAV1, HEG1, VAV3, HIF1A, ECM1, BMPR2, ID1, ENG, EPAS, CCM2, PLAT, CXCL12,* and *NRP2*.

Gene	Description	GO Terms	log2FoldChange	padj
TGFB2	transforming growth factor beta 2	EC-related from Lee *et al.*	8.567231319	1.96177E-36
COL8A1	collagen type VIII alpha 1 chain	angiogenesis	7.000642063	5.59855E-65
SEMA3C	semaphorin 3C	blood vessel remodeling	6.39097491	4.8176E-119
ANGPT2	angiopoietin 2	angiogenesis/EC-related	5.63503329	5.61592E-06
NRP1	neuropilin 1	angiogenesis/EC-related	5.437310741	4.83632E-26
NOV	nephroblastoma overexpressed	angiogenesis	5.310836095	0.000500583
CCL2	C-C motif chemokine ligand 2	EC-related from Lee *et al.*	5.220807279	1.65873E-31
CXCL1	C-X-C motif chemokine ligand 1	EC-related from Lee *et al.*	4.944339643	0.000350778
CAV2	caveolin 2	EC proliferation positive	4.870395383	9.96176E-15
RSPO3	R-spondin 3	blood vessel remodeling	4.650810521	0.00101545
TNFSF10	TNF superfamily member 10	EC-related from Lee *et al.*	4.316726193	0.012095993
PROX1	prospero homeobox 1	EC proliferation positive	3.945652206	2.85341E-42
FN1	fibronectin 1	EC-related from Lee *et al.*	3.90044152	5.4555E-233
SRPX2	sushi repeat containing protein, X-linked 2	angiogenesis	3.899271969	0.000607875
SEMA5A	semaphorin 5A	EC proliferation positive/angiogenesis	3.84510724	8.1224E-102
CAV1	caveolin 1	EC proliferation positive/angiogenesis	3.761968748	1.00345E-19
HEG1	heart development protein with EGF like domains 1	EC development	3.674782478	2.48076E-53
VAV3	vav guanine nucleotide exchange factor 3	angiogenesis	3.593296532	1.73349E-29
HIF1A	hypoxia inducible factor 1 alpha subunit	EC proliferation positive	3.293394644	2.5967E-141
ECM1	extracellular matrix protein 1	EC proliferation positive/angiogenesis	2.940030177	0.005120679
BMPR2	bone morphogenetic protein receptor type 2	blood vessel remodeling	2.671957417	3.3499E-30
ID1	inhibitor of DNA binding 1, HLH protein	EC development	2.666505231	1.84941E-08
ENG	endoglin	EC development	2.509607576	0.048352492
EPAS	endothelial PAS domain protein 1	blood vessel remodeling	2.253576194	0.002685469
CCM2	CCM2 scaffolding protein	EC development	2.182355727	0.020492555
PLAT	plasminogen activator, tissue type	EC-related from Lee *et al.*	2.139072752	7.31996E-05
CXCL12	C-X-C motif chemokine ligand 12	EC proliferation positive	1.925501081	8.35216E-08
NRP2	neuropilin 2	EC proliferation positive	1.872754096	1.98617E-13

Table 1. Notable Genes. Genes that are highly involved in angiogenesis and endothelial functionality are filtered based on the results of GO analysis.

Confirmation of enriched EC characteristics of CDH5^{+} Cells

To further identify enriched EC characteristics of CDH5^{+} cells, we subset the 11 genes directly involved with endothelial and proangiogenic properties, which belong to the intersection of the DE results of Day8 and Day14 after sorting. Then, we compared the log2 fold-change of the genes between Day8 and Day14 after sorting to examine how the gene expression levels altered during EC enrichment. As enriched ECs demonstrated distinct neovascularization capability in the previous study, we hypothesized that CDH5^{+} sorted cells would show much higher fold changes for the genes involved in proangiogenic functions than Day8. We observed that all the 14 genes exhibited larger fold changes on Day14 after sorting than Day8. This result indicates that EC characteristics start to be expressed at the early differentiation phase; however, the genuine EC characteristics are much enhanced during EC enrichment.

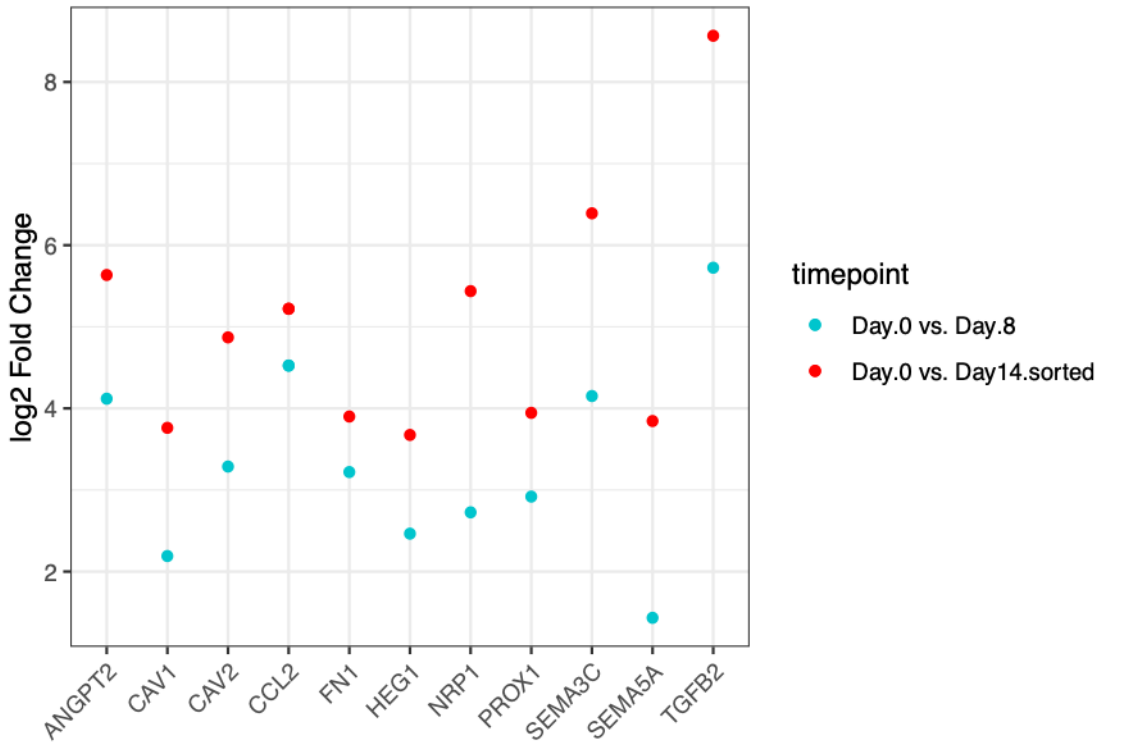


Figure 14. Identification of Enriched EC Characteristics of Day14 after sorting Compared to Day8. Genes involved in EC functions in the intersection of Day8 and Day14 after sorting are filtered. The DE genes (x-axis) are plotted to the corresponding log2Fold Change (y-axis) based on the timepoints (color).

Discussion

Human iPSC-ECs have shown angiogenic therapeutic potential to treat cardiovascular ischemic diseases in a hindlimb mouse model. However, a transcriptomic analysis to discover genes that significantly contribute to proangiogenic effects on hiPSC-ECs under the protocol developed by Lee *et al.* has not yet been performed; the gene expression and molecular pathways influencing endothelial and proangiogenic characteristics of hiPSCs have been partially explored. This comparative analysis highlights promising genes for proangiogenic properties and suggests the functionality of the potential genes during the course of EC differentiation and enrichment on a genome-wide scale.

In this study, we have exhibited 1252 DE genes that show an increase in transcript abundance in hPSC-ECs (Table 1). This extensive change depicts the substantial transition of cell type from hiPSCs to ECs under this EC differentiation protocol. Based on the results of GO analysis, we subset the 28 DE genes that significantly contribute to endothelial and proangiogenic characteristics of hiPSC-ECs. *COL8A1* (7-log2FoldChange) is a major component of basement membrane of corneal ECs and the endothelial of blood vessels. *ANGPT2* (5.6-log2FoldChange), with *VEGF*, facilitates EC migration and proliferation and hence serves as a permissive angiogenic signal. *NRP1* (5.4-log2FoldChange), a regulator of VEGF-induced angiogenesis, is highly involved in the development of the cardiovascular system in angiogenesis. *NOV* (5.3-log2FoldChange), also known as *CCN3*, regulates cell adhesion, migration, and survival in ECs. *SEMA5A* (3.8-log2FoldChange) promotes angiogenesis by increasing EC proliferation and migration and inhibiting apoptosis. *RSPO3* (4.7-log2FoldChange), an activator of Wnt signaling pathway in ECs, regulates angiogenesis. *ID1* (2.7-log2FoldChange), a transcription factor, regulates a variety of cellular processes: cellular growth, senescence, differentiation, apoptosis, angiogenesis, and neoplastic transformation. The expressions of these genes translate into proteins necessary for endothelial and proangiogenic functionalities.

In order to understand the biological functions of the DE genes of hiPSC-ECs, we conducted GO enrichment analysis. We discovered that GO terms for biological processes were significantly enriched in EC differentiation, EC proliferation, EC migration, and positive regulation of angiogenesis (Figure 10A). In contrast, GO terms for biological processes were significantly down-regulated in nuclear division, DNA replication, positive regulation of cell cycle, and stem cell population maintenance (Figure 10B). This result indicates that endothelial and proangiogenic functions are enhanced during the course of EC differentiation and enrichment as hiPSC-ECs lose the genuine pluripotent characteristics of hiPSCs.

However, the gene expression levels are not merely cumulative, suggesting that gene expression levels high at the initial phase tend to decrease over time. This indicates that EC specific genes are not highly expressed in the $CDH5^+$ samples all the time. Initially, we hypothesized that the gene expression level of *VEGFA* and *KDR*, an essential endothelial growth factor and receptor, would significantly increase over time. Contrary to the hypothesis, their gene expression levels reached a peak during the phase of mesoderm induction and decreased throughout the EC enrichment. This finding indicates that the gene turned on in the early phase of EC differentiation is likely to translate into proteins with specific functions in the early phase of EC differentiation. Therefore, the gene is no longer necessary to be highly expressed during the later phase of EC differentiation. The flow cytometry data from Lee *et al.* supports this idea by showing a high protein expression level of *KDR* in $CDH5^+$ cells after sorting.[8]

Although the analysis has presented a number of DE genes involved with endothelial and proangiogenic characteristics of hiPSC-ECs, mesenchymal genes—*PDGFRA, IGFBP7, ALPK2, SFRP4, MEIS2, SIX,* and others—were also up-regulated in $CDH5^+$ cells. ECs have an innate characteristic to lose the endothelial properties and express mesenchymal cell markers, called endothelial-to-mesenchymal transition (EndMT).[45] The up-regulation of those mesenchymal genes leads hiPSC-ECs to EndMT, where the morphology and property of ECs change into mesenchymal and smooth muscle cells.[46] Therefore, further investigation about potential EndMT markers may grant scientists a better protocol to generate ECs with prolonged EC characteristics with the minimized EndMT properties.

Through the identification of genes promoting endothelial and proangiogenic effects, this transcriptome study supports the findings from Lee *et al.* that hiPSC-ECs exhibit genuine EC characteristics and angiogenetic therapeutic potentials. This study further accentuates that Lee *et al.* developed a fully defined, clinically compatible cell culture system that generates purified,

functional, and therapeutically effective ECs. We have now expanded our understanding of genes that activate or inhibit angiogenic therapeutic potentials during EC differentiation from hiPSCs. In the future, we may strengthen the transcriptomic comprehension of hiPSC-ECs by comparing $CDH5^+$ population to EC positive control, such as human lung microvascular endothelial cells (HMVECs) and human umbilical vein endothelial cells (HUVECs), and a EC negative control, such as human dermal fibroblasts (HDF). This gene expression profiling may help to more accurately characterize ECs derived from hiPSCs in the contexts of angiogenic potentials and prolong the genuine EC attributes.

Conclusion

Through this transcriptome study, we have identified notable genes of hiPSC-ECs that significantly contribute to the endothelial functionality and vessel formation. This study may serve as a useful analysis to support the findings from Lee *et al.* to treat cardiovascular ischemic conditions that hiPSC-ECs highly promote neovascularization.[8] The results may provide new insights into EC generation from hiPSCs. Furthermore, this study could facilitate a development in regenerative medicine with ECs in the context of cardiovascular regeneration.

References

1. Lozano, R.*, et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2095-2128 (2012).
2. Patsch, C.*, et al.* Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. *Nat Cell Biol* **17**, 994-1003 (2015).
3. Lee, J., Park, Y.J. & Jung, H. Protein Kinases and Their Inhibitors in Pluripotent Stem Cell Fate Regulation. *Stem Cells Int* **2019**, 1569740 (2019).
4. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
5. Kim, K.L., Song, S.H., Choi, K.S. & Suh, W. Cooperation of Endothelial and Smooth Muscle Cells Derived from Human Induced Pluripotent Stem Cells Enhances Neovascularization in Dermal Wounds. *Tissue Eng Pt A* **19**, 2478-2485 (2013).
6. White, M.P.*, et al.* Limited gene expression variation in human embryonic stem cell and induced pluripotent stem cell-derived endothelial cells. *Stem Cells* **31**, 92-103 (2013).
7. Park, I.H.*, et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146 (2008).
8. Lee, S.J.*, et al.* Enhanced Therapeutic and Long-Term Dynamic Vascularization Effects of Human Pluripotent Stem Cell-Derived Endothelial Cells Encapsulated in a Nanomatrix Gel. *Circulation* **136**, 1939-1954 (2017).
9. Kukurba, K.R. & Montgomery, S.B. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015**, 951-969 (2015).
10. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
11. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
12. Anders, S.*, et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765-1786 (2013).
13. Patel, R.K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7**, e30619 (2012).
14. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
15. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
16. Grant, G.R.*, et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518-2528 (2011).
17. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
18. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017 (2012).
19. Maza, E., Frasse, P., Senin, P., Bouzayen, M. & Zouine, M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Commun Integr Biol* **6**, e25849 (2013).
20. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
21. Dillies, M.A.*, et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14**, 671-683 (2013).

22. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).
23. Love, M.I., Anders, S., Kim, V. & Huber, W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res* **4**, 1070 (2015).
24. Varet, H., Brillet-Gueguen, L., Coppee, J.Y. & Dillies, M.A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS One* **11**, e0157022 (2016).
25. McCarthy, D.J., Chen, Y. & Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288-4297 (2012).
26. Robinson, M.D. & Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887 (2007).
27. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* **38**, e170 (2010).
28. Yu, D., Huber, W. & Vitek, O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29**, 1275-1282 (2013).
29. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
30. Burger, T. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *J Proteome Res* **17**, 12-22 (2018).
31. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
32. Young, M.D., Wakefield, M.J., Smyth, G.K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14 (2010).
33. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
34. Pomaznoy, M., Ha, B. & Peters, B. GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics* **19**, 470 (2018).
35. Brown, J., Pirrung, M. & McCue, L.A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* (2017).
36. Benedito, V.A.*, et al.* A gene expression atlas of the model legume Medicago truncatula. *Plant J* **55**, 504-513 (2008).
37. Severin, A.J.*, et al.* RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol* **10**, 160 (2010).
38. Pan, Y.*, et al.* Whole tumor RNA-sequencing and deconvolution reveal a clinically-prognostic PTEN/PI3K-regulated glioma transcriptional signature. *Oncotarget* **8**, 52474-52487 (2017).
39. Nookaew, I.*, et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **40**, 10084-10097 (2012).
40. Matson, J.P.*, et al.* Correction: Rapid DNA replication origin licensing protects stem cell pluripotency. *Elife* **8**(2019).
41. McDermaid, A., Monier, B., Zhao, J., Liu, B. & Ma, Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform* (2018).
42. Greenland, S.*, et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* **31**, 337-350 (2016).
43. Michaelis, U.R. Mechanisms of endothelial cell migration. *Cell Mol Life Sci* **71**, 4131-4148 (2014).

44. Lee, S.*, et al.* Direct Reprogramming of Human Dermal Fibroblasts Into Endothelial Cells Using ER71/ETV2. *Circ Res* **120**, 848-861 (2017).
45. Cho, J.G., Lee, A., Chang, W., Lee, M.S. & Kim, J. Endothelial to Mesenchymal Transition Represents a Key Link in the Interaction between Inflammation and Endothelial Dysfunction. *Front Immunol* **9**, 294 (2018).
46. Pinto, M.T., Covas, D.T., Kashima, S. & Rodrigues, C.O. Endothelial Mesenchymal Transition: Comparative Analysis of Different Induction Methods. *Biol Proced Online* **18**, 10 (2016).
47. Retrieved from http://software.braodinstitute.org/gsea/index.jsp
48. Retrieved from http://uniprot.org