devgea: A Software Suite Allowing the Robust Analysis of Developmental Gene Signatures During Human Cardiomyocyte Differentiation

Isaiah Bartlett, Dr. Jeff Kinne[1,2,3], and Dr. Kris Schwab[1,2,4]

[1]The Center for Genomic Advocacy, Indiana State University, Terre Haute, Indiana, [2]Rich and Robin Porter Cancer Research Center, Indiana State University, Terre Haute, Indiana, [3]Department of Mathematics and Computer Science, Indiana State University, Terre Haute, Indiana, United States of America, [4]Department of Biology, Indiana State University, Terre Haute, Indiana

Introduction

The development of the heart involves the commitment of pluripotent embryonic cells to the cardiac cell lineage pathway followed by the differentiation of these cells into cardiomyocytes, the contractile heart muscle cells. Many important cardiac-lineage genes possessing a distinct temporal expression pattern have been identified and incorporated into the developmental cardiac regulatory network; however, the identification, classification, and functional investigation of novel and putative cardiac genes using high-throughput gene expression analysis will expand our knowledge of this complex network.

An ISU-developed bioinformatic software tool was used to classify the major gene expression signatures of human *in vitro* cardiomyocyte differentiation using previously published high-throughput RNA-seq data sets to classify differentially expressed genes into cardiac differentiation expression signatures (stages). This analysis identified over three-thousand differentially expressed genes classified into four/five cardiac developmental signatures. Several gene encoding, gene regulatory, and signaling pathway components from each signature were selected for validation of our classification procedure. Our classification procedure successfully assigned a majority of these components into their stages with high accuracy, validating our overall classifications. Therefore, our clustering, and consequently the software, is effective in classifying differentially expressed genes into specific developmental signatures thereby identifying putative cardiac regulatory genes based upon their expression signature.

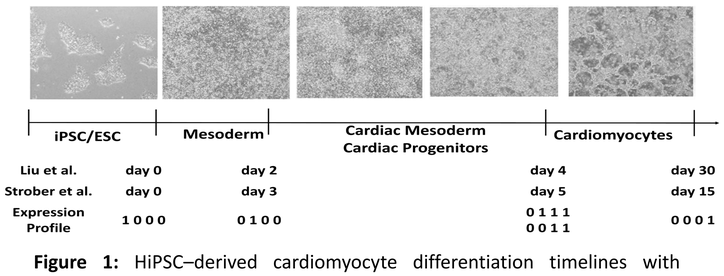

Figure 1: HiPSC–derived cardiomyocyte differentiation timelines with developmental stages and their selected correlational expression profile

Materials & Method

Data preparation

Raw count RNA-seq data sets were downloaded from the Gene Expression Omnibus (GEO) for the Liu et al. (2017) and Strober et al. (2019) publications (see references). Due to the differences in methods, it became necessary to reduce the Strober et al. publication to one more comparable to the Liu et al. publication using the differentiation timeline from each **(**Figure 1**).** Each dataset contains over 40,000 rows with each row corresponding to a different gene transcript measured and each column a biological sample micro–dissected at a specific time point. A bioinformatic software tool, developed in R by ISU, called devGEA (Figure 2) and additional R programming scripts were utilized to complete the analysis.

Data filtering

The data sets are filtered using statistical and fold change analysis. Differentially expressed genes are identified between the conditions using ANOVA analysis in each data set. The ANOVA statistical test with FDR (P <.001) and a fold change criteria cutoff of >= +/-1 of log2 values (fold change of 2) must be present for each gene.

Correlational clustering

In order to categorize the gene expression profiles from the candidates in the data filtering process, a correlational clustering function was performed using devGEA (Figure 3). Each distinct stage of differentiation was assigned a basic expression profile that was determined to be representative of that stage during each time point (Figure 1, Expression Profile). devGEA then uses these user–inputted "correlation vectors" to correlate the medians of each row in the dataset to one of the given vectors. A table was created to store the genes from the analyses during filtering, the stage they were correlated to, the strength of that correlation, and the maximum fold change between their expressions in log2 scale (Table 1).

References

Liu et al. (2017). Genome-Wide Temporal Profiling of Transcriptome and Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. Circulation research, 121, 376–391.

Strober et al.. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, *364*, 1287–1290.

Genes	Expected Class	Observed Class	Correlation	Fold Change
NANOG	Pluripotent	Pluripotent	0.696401262	1.559722992
POU5F1	Pluripotent	Pluripotent	0.521434922	9.226836719
SOX2	Pluripotent	Pluripotent	0.645439879	6.866911853
FOXC1	Mesodermal	Cardiac Mesoderm	0.942100869	5.757407526
MIXL1	Mesodermal	Mesodermal	0.97857647	7.734963971
MSX1	Mesodermal	Cardiac Mesoderm	0.89263496	7.588834181
TBXT	Mesodermal	Mesodermal	0.965854329	8.180423387
WNT3A	Mesodermal	Mesodermal	0.8353497	6.690775522
MESP1	Cardiac Mesoderm	Mesodermal	0.931988375	7.61851596
PDGFRA	Cardiac Mesoderm	Cardiac Mesoderm	0.97264427	8.077236749
GATA4	Cardiac Mesoderm	Cardiac Mesoderm	0.976053175	8.189103932
HAND2	Cardiac Progenitor	Cardiac Progenitor	0.955722353	8.88947048
TBX2	Cardiac Progenitor	Cardiac Progenitor	0.94375206	8.897951336
TBX5	Cardiac Progenitor	Cardiac Progenitor	0.953671731	7.26619871
MEF2C	Cardiomyocyte	Cardiac Progenitor	0.939977058	6.776389094
MYH6	Cardiomyocyte	Cardiomyocyte	0.956006548	8.632881373
MYH7	Cardiomyocyte	Cardiomyocyte	0.975302328	9.105411846
MYL6	Cardiomyocyte	Cardiomyocyte	0.944675754	1.79547228
NKX2-5	Cardiomyocyte	Cardiomyocyte	0.903885783	8.408610023
NPPA	Cardiomyocyte	Cardiomyocyte	0.965296979	7.943103807
PLN	Cardiomyocyte	Cardiomyocyte	0.968283104	9.40545134
RYR2	Cardiomyocyte	Cardiomyocyte	0.889339526	5.411513596
TNNI2	Cardiomyocyte	Cardiac Progenitor	1	1.087817512
TNNT2	Cardiomyocyte	Cardiomyocyte	0.936451985	8.133848937

Table 1: Results of the analyses and correlational function from devGEA and R scripts

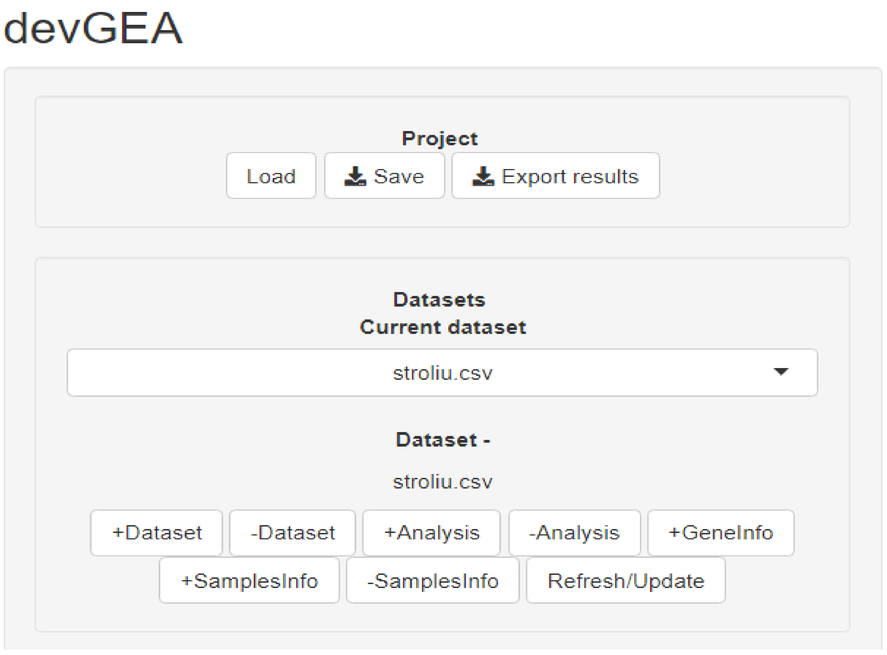

Figure 2: devGEA project panel where users can load/save full devGEA sessions (projects), add/remove datasets, and modify the currently loaded dataset

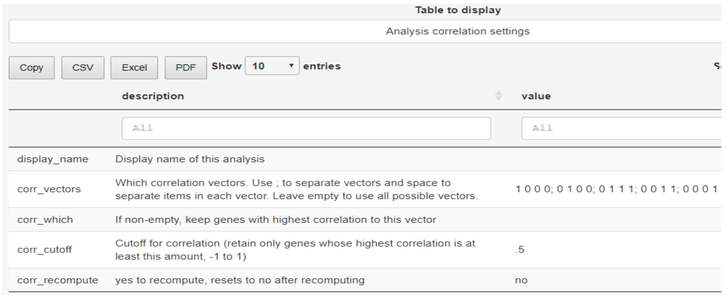

Figure 3: devGEA correlation settings, currently setup using our vectors, classifying each gene's timepoint medians into the strongest matching stage.

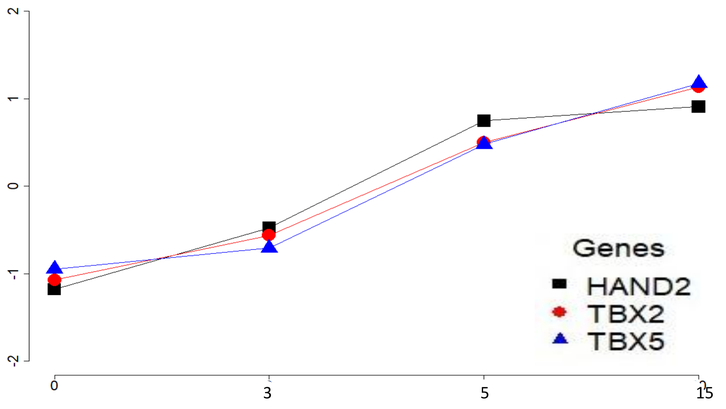

Figure 4: Cardiac progenitor medians taken for correlational clustering

Results And Discussion

The study of heart development and disease has undergone a paradigm shift in the last decade using HiPSCs to model both heart development and disease *in vitro*. Furthermore, breakthroughs in HiPSC procedures have allowed scientists to study the differentiation and physiology of the functional cells of the heart, the cardiomyocyte. In this procedure, pluripotent HiPSCs undergo the process of mesoderm, cardiac mesoderm specification, cardiac progenitor proliferation, and functional cardiomyocyte differentiation (Figure 1). High-throughput gene expression allows the measurement of gene expression changes through the cardiomyocyte differentiation process. Previous studies of mammalian heart development and HiPSC-directed cardiomyocyte differentiation profiling have identified several important genes that are differentially expressed during cardiomyocyte differentiation. Using bioinformatic analysis tools, we sought to classify differentially expressed genes identified in directed-cardiomyocyte differentiation of HiPSC data sets (Liu et al. 2017 and Strober et al. 2019) according to the stages within the differentiation process. Each gene was classified based on the strength of their correlation to our selected expression profiles for each stage. 24 transcriptional or cell signaling regulatory genes were selected for the validation of our correlation function. Our function impressively validated a majority of these genes which was found to be consistent across each dataset (Table 1). Over 7,500 differentially expressed genes were classified into 5 distinct cardiomyocyte differentiation signatures with a minimum of a 50% correlation strength. Each gene was found to have a significant expression profile via ANOVA and statistical analysis thresholds. Given the rigor and stringency of this analysis, we have not only identified new putative developmental functions for genes identified within these classifications within cardiomyocyte differentiation but have found potential developmental markers for these processes which will aid future experimental investigations.

Acknowledgements

This work was supported by: Indiana State University - The Center for Genomic Advocacy (Gonser, R), and the National Institutes of Health (NIH) – Big Data 4 ISU (Kinne, J and Rusty Gonser).