

Dwayne Tally, Laura Cochran, Hayden Fell, Garrett Oxford, Joseph Djaloui, Andrew Williamson, Rusty Gonser, Jeff Krime and Kris Schwab
 Department of Biology*, Department of Mathematics and Computer Science†, and The Center for Genomic Advocacy, Indiana State University, Terre Haute, IN 47809

Introduction

Mammalian heart development is regulated by an evolutionarily conserved genetic network that has been elucidated from numerous studies of model organisms and the genetic investigation of human congenital heart defects. Published high-throughput gene expression data sets of normal and abnormal mammalian heart development provide the research community with the opportunity to investigate important biological processes and develop new hypotheses and experiments evaluating the author's original findings. The goal of our bioinformatic investigations is to identify and validate unique developmentally-conserved developmental gene expression profiles of mammalian cardiac (heart) developmental in which orthologous cardiac genes can then be investigated using multiple *in vivo* and *in vitro* experimental models, such as pluripotent stem cells, zebrafish, and mouse experimental systems.

Our first step in completing this unique experimental approach is to identify previously published data sets deposited in the Gene Expression Omnibus (GEO) studying developmental stages of the mouse embryonic heart yielded a robust, gene expression microarray data set that investigated over 45,000 transcripts. The Li et al. (2014) heart developmental atlas possessed a number of replicates per condition (minimum n=3), and included mouse embryonic stem cells, the early mouse embryo, several embryonic heart stages, fetal heart stages, and postnatal stages. Our bioinformatic investigations of this data set encouraged us to develop a set of simple tools that can streamline gene expression data analysis in the hand of the investigator.



Figure 1. The Li et al. gene expression microarray data set investigates several important developmental stages from early embryonic to adult heart. Adapted from X. Li et al., Transcriptional atlas of cardiogenesis maps congenital heart disease intradome. *Physical Genomics* 46, 482-495 (2014).

Methods

Our research goal is to develop an easy, user-friendly applet utilizing R and Bioconductor packages to evaluate high-throughput gene expression data that can be used by scientists interested in generating gene lists describing a specific series of sample comparisons that represent a developmental stage or timeline - *without previous extensive R programming experience*. To accomplish this task, a bioinformatic analysis approach was developed using R and Bioconductor generating a gene list representing early cardiac developmental genes by comparing ESC, day 7 whole embryo, heart tissue from days 8 / 9 / 12, and adult heart tissue. Once this was accomplished, the bioinformatic pipeline was translated into a Shiny applet - allowing the pre-processing and analysis to be completed *using little to no R programming*.

Raw data was downloaded from the Gene Expression Omnibus (GSE51483 [Li et al.]). Each dataset contains over 40,000 rows with each row corresponding to a different gene transcript measured; columns represent different biological sample microdissected at different time points. Metadata for the genes were downloaded using the R packages `mouse4302d`. Differential expression calculations were performed using the `limma` package functions `lmFit`, `treat`, and `log2FC`. For analysis, the data from GSE51483 mouse development were grouped based on age - ESC, day 7 whole embryo, heart tissue from days 8/9/12, adult heart tissue - with each age containing either three or six replicates (columns) in the data. The following statistical tests were performed - comparison of day 8/9/12 samples versus ESC and adult heart samples. For each comparison, gene transcripts were selected as having statistically significant differential expression using the criteria - *t*-fold change of at least 2.0 and an adjusted *p*-value of at most 0.01. This resulted in 4,004 significant transcripts. Gene transcripts that were selected as significant were further grouped by clustering with the `kmeans` R function and visualized using the heatmap2 function from the R package `gplots`. Gene Ontology (GO) and GO Enrichment Analysis was performed on each cluster using the tools available at `geneontology.org`.

References

X. Li et al., Transcriptional atlas of cardiogenesis maps congenital heart disease intradome. *Physical Genomics* 46, 482-495 (2014).

Acknowledgments

This work was supported by: The Center for Genomic Advocacy (Krime, Schwab), University Research Council (Schwab), Indiana Academy of Sciences (Schwab), ISU Summer Undergraduate Research Experience (SURE), and NIH #5R25MD011712 (Gonser, Krime)

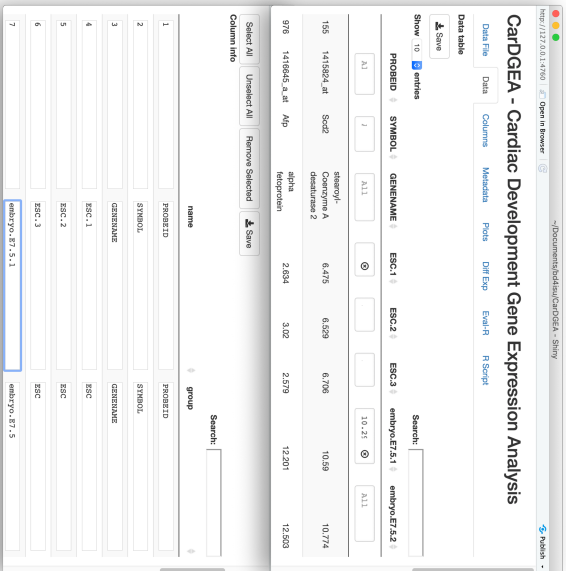
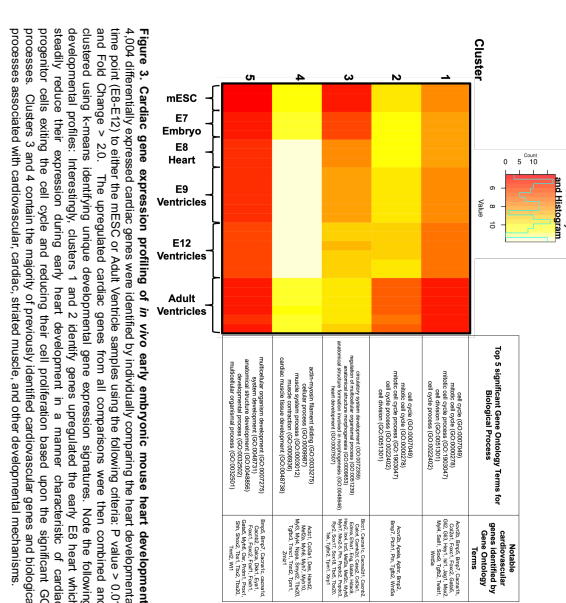
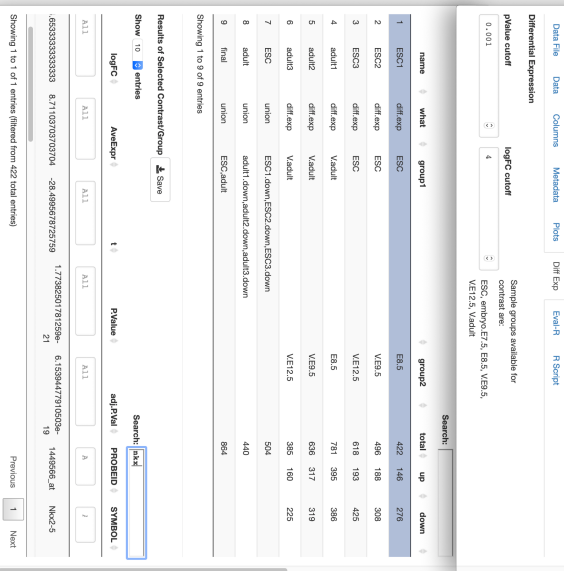


Figure 2. Workflow of CardGEA. The screenshots above and below demonstrate key portions of the R Shiny app that we developed. Data is imported (Data File tab) and examined (Data tab, above top, and Data tab). Metadata is imported (Data tab). Samples are grouped for analysis (Columns tab, above top, and below top), and results are visualized in table format (Diff Exp tab, below bottom) and graphically (Heat tab).

CardGEA - Cardiac Development Gene Expression Analysis



Results and Discussion

The experimental pipeline that we implemented (Using R, Bioconductor, and Shiny) allows the stepwise import and analysis of gene expression data. Furthermore, our application streamlines the investigation of complex gene expression data sets involving multiple conditions such as multiple developmental stages, multiple drug treatments, or time course series allowing the investigator to complete additional bioinformatic queries using the data set. Figure 2 describes a brief walk through of the application.

The gene expression analysis pipeline enabled us to complete a preliminary analysis of *in vivo* early mouse embryonic heart development using the Li et al. (2014) cardiac developmental atlas data set. This preliminary analysis identified approximately 4,000 differentially expressed genes that are upregulated in the early stages of heart development by combining multiple independent comparisons between the early mouse heart samples (E8.5, E9.5, and E12.5) and mESC or adult heart samples. Once the 4,000 upregulated early cardiac genes were identified, the gene list was clustered based on expression expression profile using k-means (Figure 3). Each cluster identifies a unique cardiac developmental expression pattern and important cardiac regulatory genes (identified in bold). Gene Ontology enrichment analysis of each cluster identifies a unique dichotomy between the clusters. The top GO Biological Process terms associated with Clusters 1 and 2 identify cell cycle and cell proliferation processes. Interestingly, within the adult heart profile of clusters 1 and 2 identify genes that are downregulated with the adult heart which is mitotically senescent. The remaining clusters (cluster 3, 4, and 5) identified genes associated with general developmental, striated muscle, or cardiovascular processes. Finally, many notable cardiovascular regulatory genes were identified in our early cardiac developmental gene list that validate our analysis. Several important genes necessary for normal heart development and proper specification, differentiation, and proliferation of the cardiac mesoderm and cardiac progenitor genes were identified, such as *Trpc7*, *Isl1*, *Mef2c*, *Nkx2-5*, *Tbx5*, and *Tbx20*. Also, important cardiomyocyte genes encoding contractile components were found including both calcium transporters and myofibril components, such as *Myl3*, *Myl4*, *Myl6*, *Myl7*, *Ry2*, *Tnni3*, and *Tnni2*.

This analysis provides a focused investigation of gene expression profiles of the early developing mouse heart. This data set has identified both the unique expression profiles of known cardiac regulatory genes as well as genes that remain uncharacterized in cardiac development and cardiac function. We intend to extend this work using our CardGEA pipeline characterizing the gene expression profiles of the pre-cardiac mesoderm and other early embryonic tissues. Furthermore, the 4,000 genes upregulated in early heart development produces an "early cardiac reference gene set" for the evaluation of gene function in previously published work and our future genetic loss-of-function experiments.